PENENTUAN SEGMENTASI KONSUMEN PADA MARKETING DATA *IFOOD* MENGGUNAKAN METODE *K – MEANS CLUSTERING*

TUGAS AKHIR

Karya tulis sebagai salah satu syarat
untuk memperoleh gelar Sarjana Teknik dari
Program Studi Teknik Industri
Fakultas Teknik Universitas Pasundan

Oleh

MUHAMMAD IRSYAD NUR RAIS

NRP : 203010158

PROGRAM STUDI TEKNIK INDUSTRI

FAKULTAS TEKNIK

UNIVERSITAS PASUNDAN

2022

PENENTUAN SEGMENTASI KONSUMEN PADA MARKETING DATA *IFOOD* MENGGUNAKAN METODE *K – MEANS CLUSTERING*

Muhammad Irsyad Nur Rais

NRP : 203010158

ABSTRAK

Perusahaan IFood merupakan perusahaan besar di Brazil yang bergerak dibidang teknologi yang menjual kebutuhan rumah tangga. IFood sendiri mempunyai beberapa gerai toko dan jasa pemesanaan dan pengantaran secara online yang beroperasi di Brazil sebagai pusatnya

Kendala yang terdapat pada perusahaan IFood terjadi pada tahun 2020 yaitu berupa campaign yang belum berhasil dan belum tercapai 100% jadi tingkat keberhasilan campaign hanya mencapai 15%. Jika melihat dari sisi manajemen pemasaran dikatakan perusahaan IFood belum memenuhi atau menentukan strategi pemasaran STP (Segmenting, Targeting, Positioning). Strategi pemasaran pertama segmenting atau penentuan segmentasi yaitu mengelompokkan target pasar berdasarkan karakteristik yang sama untuk dikelola secara efektif dan tepat agar mencapai tujuan bisnis yang telah ditetapkan, dalam hal ini mencapai campaign yang sukses.

Berdasarkan latar belakang tersebut dapat disimpulkan penentuan segmentasi konsumen perlu dilakukan pada perusahaan IFood. Hasil penentuan segmentasi tersebut menggunakan metode K-Means Clustering, untuk variabel jumlah pembelian produk per kategori terdapat 2 kluser, lalu untuk variabel metode pembelian terdapat 2 kluster dan yang terakhir tingkatan pendidikan konsumen terdapat 5 tingkata. Dengan penerapan metode K-Means Clustering diharapkan juga bisa memudahkan perusahaan untuk menentukan rencana promosi atau marketing pada campaign fase selanjutnya yang bisa mencapai 100%.

Kata kunci : Segmentasi, Konsumen, K-Means Clustering, Strategi Pemasaran

DETERMINATION OF CONSUMER SEGMENTATION IN *IFOOD* MARKETING DATA USING K – MEANS CLUSTERING METHOD

Muhammad Irsyad Nur Rais

NRP : 203010158

ABSTRACT

IFood company is a large company in Brazil engaged in technology that sells household needs. IFood itself has several store outlets and online ordering and delivery services operating in Brazil as its center

Constraints encountered by IFood companies occurred in 2020, namely in the form of campaigns that had not been successful and had not been achieved 100% so the campaign success rate only reached 15%. If you look at the marketing management side, it is said that the IFood company has not fulfilled or determined the STP (Segmenting, Targeting, Positioning) marketing strategy. The first marketing strategy is segmenting or determining segmentation, namely grouping the target market based on the same characteristics to be managed effectively and precisely in order to achieve the business goals that have been set, in this case achieving a successful campaign.

Based on this background, it can be concluded that the determination of consumer segmentation needs to be done in IFood companies. The results of the segmentation determination use the K-Means Clustering method, for the variable number of product purchases per category there are 2 clusters, then for the purchase method variable there are 2 clusters and the last level of consumer education is 5 levels. With the application of the K-Means Clustering method, it is hoped that it can also make it easier for companies to determine promotion or marketing plans in the next phase of the campaign which can reach 100%.

Keywords: Segmentation, Consumers, K-Means Clustering, Marketing Strategy

PENENTUAN SEGMENTASI KONSUMEN PADA MARKETING DATA *IFOOD* MENGGUNAKAN METODE *K – MEANS CLUSTERING*

Oleh

Muhammad Irsyad Nur Rais

NRP : 203010158

Menyetujui

Tim Pembimbing

Tanggal 26 Agustus 2022

Pembimbing	Penelaah
(Dr. Ir. M. Nurman Helmi, DEA)	(Dr. Drs. Iman Firmansyah, M.Sc)

Mengetahui,

Ketua Program Studi

(Dr. Ir. M. Nurman Helmi, DEA)

PEDOMAN PENGGUNAAN TUGAS AKHIR

Tugas Akhir Sarjana yang tidak dipublikasikan terdaftar dan tersedia di perpustakaan Universitas Pasundan, dan terbuka untuk umum dengan ketentuan bahwa hak cipta ada pada pengarang dengan mengikuti aturan HaKI yang berlaku di Universitas Pasundan. Referensi Kepustakaan diperkenankan dicatat, tetapi pengutipan atau peringkasan hanya dapat dilakukan seizin pengarang dan harus disertai dengan kebiasaan ilmiah untuk menyebutkan sumbernya.

Memperbanyak atau menerbitkan sebagian atau seluruh Tugas Akhir haruslah seizin Program Studi Teknik Industri, Fakultas Teknik, Universitas Pasundan.

PERNYATAAN

Dengan ini Saya menyatakan bahwa Judul Tugas Akhir :

PENENTUAN SEGMENTASI KONSUMEN PADA MARKETING DATA *IFOOD* MENGGUNAKAN METODE *K – MEANS CLUSTERING*

Adalah hasil kerja saya sendiri, kecuali beberapa kutipan dan ringkasan yang masing-masing disebutkan sumbernya dengan cara penelitian referensi yang sesuai.

Pernyataan ini saya buat dengan sebenar-benarnya dan jika pernyataan ini tidak sesuai dengan kenyataan maka saya bersedia menanggung sanksi yang akan dikenakan sesuai dengan ketentuan yang berlaku.

Bandung, 26 Agustus 2022

Muhammad Irsyad Nur Rais

NRP 203010158

KATA PENGANTAR

. Segala puji peneliti panjatkan kepada Allah SWT yang maha pemberi kasih, pelimpah karunia kepada hambaNya dan penabur ilmu, karena izinya peneliti dapat menyelesaikan laporan tugas akhir dengan judul "Penentuan Segmentasi Konsumen Pada Marketing Data *IFood* Menggunakan Metode *K-Means Clustering"* dengan tepat waktu meskipun peneliti sadar bahwa masih terdapat banyak kekurangan yang masih harus di perbaiki, demi perbaikan selanjutnya, saran dan kritik yang membangun akan peneliti terima dengan senang hati. Peneliti ucapkan terimakasih atas bantuan, bimbingan dan semangat kepada peneliti dalam mengerjakan tugas akhir ini sehingga tugas akhir ini dapat diselesaikan dengan baik, ucapan rasa syukur dan terimakasih ini peneliti ucapkan kepada :

1. Orang tua tercinta yang telah membesarkan dan selalu memberikan do'a, dukungan dan motivasi sehingga peneliti dapat menyelesaikan tugas akhir ini.
2. Dr. Ir. M. Nurman Helmi, DEA. selaku dosen pembimbing yang senantiasa memberikan saran dan memberikan waktu luangnya untuk proses bimbingan dan membantu peneliti sehingga peneliti dapat dilancarkan dalam proses pengerjaan tugas akhir ini.
3. Dr. Drs. Iman Firmansyah, M.Sc. selaku dosen penelaah yang senantiasa memberikan saran, arahan dan nasehatnya dalam proses pengerjaan tugas akhir ini.
4. Seluruh dosen yang mengajar di Program Studi Teknik Industri Universitas Pasundan Bandung yang telah memberikan ilmu, nasehat dan bimbingan selama proses perkuliahan.
5. Teman-teman seperjuangan mahasiswa teknik industri non regular F yang selalu memberikan semangat dan dukungan kepada peneliti.
6. Teman-teman alumni Polman Bandung teknik pengecoran logam yang selalu memberikan motivasi kepada peneliti agar tepat waktu.

7. Vira Rizkia S.Ikom yang selalu memberikan do’a dan semangat setiap peneliti ingin mengerjakan tugas akhir
8. Teman-teman diluar kampus yang selalu memberikan do’a dan dukungan kepada peneliti.
9. Semua pihak yang terlibat dalam pengerjaan tugas akhir ini yang tidak bisa peneliti ucapkan satu-persatu.

Akhir kata, peneliti berharap semoga Laporan Tugas Akhir ini dapat bermanfaat dan dapat diterima serta direspon oleh seluruh pihak. Semoga Allah SWT membalas kebaikan kepada semua pihak yang telah memberikan do’a, semangat, dan dorongan kepada peneliti. Terimakasih.

Bandung, 26 November 2022

(Muhammad Irsyad Nur Rais)

Kata Pengantar

Daftar Gambar

Daftar Tabel

Bab I
Pendahuluan

1. 1. Latar Belakang Masalah

Kebutuhan perusahaaan untuk mengembangkan dan menyelaraskan sesuai dengan tujuan utama perusahaan semakin sulit. Hal ini mendorong beberapa perusahaan melakukan *tracking* atau analisis secara mendalam dari penjualan sebelumnya. Berhasil tidaknya analisa tersebut sangat akan berpengaruh pada perkembangan penjualan atau hasil setiap bulannya.

Banyak faktor yang bisa membuat sebuah perusahaan mengembangkan produk setiap bulannya dengan melihat *demand* atau *consumer behavior*. *Consumer behavior* adalah studi yang dilakukan kepada konsumen dan bagaimana mereka melakukan pembelian dari sutau produk/layanan. Hal ini bisa sangat membuat suatu perusahaan mengerucutkan kepada beberapa konsumen secara spesifik.

Manajemen pemasaran yang diterapkan pada perusahaan akan efektif jika dilakukan berdasarkan persepsi dan prefensi konsumen. Persepsi yaitu tindakan menyusun, mengenali, dan menafsirkan informasi sensoris yang berguna memberikan gambaran dan pemahaman tentang lingkungan, sedangkan prefensi adalah pilihan-pilihan yang dibuat oleh para konsumen atas produk-produk yang dikonsumsi.

Sebuah perusahaan bisa berkembang dengan bantuan sudut pandang dari beberapa hal seperti penjualan, operasional, periklanan dan masih banyak lagi. Untuk mendukung hal tersebut sebuah perusahaan membutuhkan data yang aktual untuk membantu melihat sudut pandang tersebut dari berbagai sisi. Dalam hal ini penggunaan *dataset* yang sudah dirangkum oleh perusahaan sangatlah berpengaruh pada penentuan langkah selanjutnya dalam menentukan strategi pemasaran.

IFood merupakan perusahaan besar di Brazil yang bergerak dibidang teknologi yang menjual kebutuhan rumah tangga. *IFood* sendiri mempunyai beberapa gerai toko dan jasa pemesanaan dan pengantaran secara online yang beroperasi di Brazil sebagai pusatnya dan juga memiliki dibeberapa negara lainnya seperti Australia, Kanada, Jerman, India, Montenegro, Afrika Selatan, Spanyol dan Amerika Serikat.

Pada tahun 2020 *IFood* sudah melakukan *pilot campaign* atau kampanye percontohan yang melibatkan 2.240 pelanggan. Total biaya *campaign* sampel adalah 6.720MU atau sama dengan 147,20 dollar. Namun pendapatan yang dihasilkan oleh pelanggan yang menerima tawaran itu adalah 3.674MU atau sama dengan 80,48 dollar. Secara global *campaign* memiliki kerugian -3.046MU atau sama dengan -66.74 dollar dan tingkat keberhasilan *campaign* hanya 15%.

Dalam tawaran kepada konsumen yang diberikan hampir 50% belum menerima tawaran tersebut. Bisa terlihat jika perusahaan *IFood* pada *campaign* tersebut belum melakukan perencanaan strategis yang tepat atau STP (*Segmenting Targeting Positioning)* sehingga *campaign* tersebut belum berhasil 100%. Strategi pemasaran STP ini berarti proses mengkategorikan, membidik pasar yang diinginkan, lalu memposisikan pemasaran bisnis dibandingkan pesaing.

Dalam STP atau *Segmenting Targeting Positioning* terdapat strategi pemasaran yang pertama yaitu *Segmenting* atau segmentasi. Segmentasi merupakan strategi pemasaran yang mengelompokkan target pasar berdasarkan karakteristik yang sama untuk dikelola secara efektif dan tepat agar mencapai tujuan bisnis yang telah ditetapkan. Karakteristik segmentasi berdasarkan pada usia, jenis kelamin, pekerjaan, jumlah keluarga, ketertarikan, frekuensi pembelian dan lainnya. Tujuan dari segmentasi adalah untuk mengenali konsumen yang berpotensi untuk melakukan pembelian kembali pada perusahaan dan memahami kebutuhan setiap kelompok konsumen.

Diharapkan dengan adanya penentuan segmentasi konsumen perusahaan *IFood* dapat dengan mudah membidik konsumen sesuai target pasar dan melakukan *campaign* pada fase berikutnya secara tepat. Pentingnya melakukan segmentasi konsumen pada perusahaan karena akan memudahkan perusahaan untuk memasarkan produk secara efektif dan efisien baik secara biaya maupun waktu secara tepat kepada konsumen, sehingga penjualan produk dapat lebih meningkat dan membuat perusahaan lebih berkembang.

1. 2. Rumusan Masalah

Penentuan dan penyusunan strategi pemasaran terutama segmentasi konsumen yang efektif dilakukan dengan mempertimbangkan persepsi dan preferensi konsumen, dengan demikian dapat dirumuskan permasalahannya sebagai berikut:

Bagaimana mengelompokkan segmentasi konsumen pada perusahaan *IFood* menggunakan metode *K-Means Clustering* berdasarkan beberapa kategori ini:

A. Jumlah pembelian produk per kategori

B. Metode pembelian

C. Tingkatan pendidikan konsumen

1. 3. Tujuan Penelitian

Dari penjelasan latar belakang dan perumusan masalah sebelumnya, maka tujuan dari penelitian ini adalah mengelompokkan segmentasi konsumen perusahaan *IFood* dengan menggunakan metode *K-Means Clustering* dari beberapa kategori yang sudah ditentukan yaitu:

A. Jumlah pembelian produk per kategori

B. Metode pembelian yang dipakai konsumen

C. Tingkatan pendidikan konsumen

1. 4. Pembatasan Dan Asumsi

Pembatasan masalah pada lingkup penelitian ini adalah:

1. Peneliti hanya melakukan analisa pada hasil penjualan hingga tahun 2020 sesuai dengan data yang tersedia.
2. Peneliti melakukan penelitian menggunakan analisis visualisasi dan metode *K-Means Clustering* untuk menentukan segmentasi konsumen

Asumsi yang digunakan pada penelitian ini adalah:

1. Lokasi konsumen berada dalam 8 negara yaitu Australia, Kanada, Jerman, India, Montenegro, Afrika Selatan, Spanyol dan Amerika Serikat yang menggunakan layanan *IFood*
2. Konsumen yang berusia 15 sampai 45 tahun yang pernah melakukan pembelian pada perusahaan *IFood*

1. 5. Lokasi Penelitian

Penelitian ini dilakukan secara online dengan menggunakan data yang disediakan oleh Kaggle yang diambil langsung dari perusahaan *IFood*.

1. 6. Sistematika Penulisan

Penulisan tugas akhir ini terdiri dari 6 bab dan setiap bab terdiri dari sub-sub pembahasan dengan sistematika penulisan sebagai berikut:

1. Bab pertama pendahuluan, menguraikan tentang latar belakang masalah, rumusan masalah, tujuan, pembatasan dan asumsi, lokasi penelitian dan sistematika penulisan
2. Bab kedua tinjauan pustaka, menguraikan tentang landasan teori dan konsep-konsep yang relevan dengan permasalahan yang dikaji juga mengemukakan dan membandingkan masalah yang pernah dilakukan oleh peneliti lain yang terkait metode yang akan dikaji dalam penulisan tugas akhir ini.
3. Bab ketiga metodologi penelitian, menguraikan tentang penjelasan secara rinci data yang akan digunakan, metode yang akan dipakai sampai tahapan yang akan dilakukann dalam melakukan analisis data.
4. Bab keempat pengumpulan dan pengolahan data, menguraikan pengumpulan data-data yang diperlukan yang disesuaikan dengan metode yang digunakan. serta pengolahan data yang akan menunjukkan hasil optimal yang didapat sesuai dengan tujuan penelitian ini.
5. Bab kelima analisis dan pembahasan, menguraikan hasil kajian dari masalah yang akan dibahas. Dalam bab ini juga dikemukakan pendapat atau ide gagasan yang

sesuai dengan rumusan masalah dan tujuan yang berlandaskan pada informasi serta teori-teori yang ada.

6. Bab keenam adalah kesimpulan dan saran, merupakan bab penutup yang berisi tentang uraian kesimpulan dan saran yang akan diperoleh perusahaan dari hasil tugas akhir ini.

Bab II
Tinjauan Pustaka

Ada beberapa bahasan yang perlu dikembangkan dan dianalisis lebih dalam untuk mengetahui hasil dari tujuan penelitian tersebut. Bahasan dibawah ini merupakan ilmu yang wajib diketahui dan dikembangkan untuk memudahkan mengolah data atau mengambil kesimpulan untuk perusahaan tersebut.

2.1 Pemasaran, Segmentasi & Proses Bisnis *E-Commerce*

Pemasaran adalah salah satu rumpun ilmu yang terdapat pada mata kuliah Teknik Industri yang berhubungan dengan topik penelitian kali ini. Berikut adalah penjelasan lengkap tentang pemasaran

2.1.1 Definisi Pemasaran

Definisi manajemen pemasaran menurut Kotler dan Keller (2016) adalah penganalisaan, perencanaan, pelaksanaan, dan pengawasan program-program yang bertujuan menimbulkan pertukaran dengan pasar yang dituju dengan maksud untuk mencapai tujuan perusahaan. Untuk mencapai strategi pemasaran, setiap pemasar harus memahami perilaku konsumen sehingga bisnis yang dijalankan dapat meraih kesuksesan.

Dalam teori pemasaran terdapat pula strategi pemasaran yang mempunyai peranan penting untuk keberhasilan usaha perusahaan. Strategi pemasaran merupakan rencana menyeluruh, terpadum dan menyatu di bidang pemasaran, yang memberikan panduan tentang kegiatan yang akan dijalankan untuk dapat tercapainya tujuan pemasaran suatu perusahaan (Assauri, 2017).

Oleh karena itu, penentuan strategi pemasaran harus didasarkan atas analisis lingkungan dan eksternal perusahaan melalui analisis keunggul-an, kelemahan perusahaan, serta analisis kesempatan dan ancaman yang dihadapi perusahaan dari lingkungannya. Termasuk kedalamnya yaitu *historical purchase* atau pembelian yang sudah dilakukan oleh konsumen pada suatu perusahaan.

2.1.2 Segmentasi Konsumen

Segmentasi konsumen dapat digunakan untuk mengidentifikasi kelompok konsumen secara alami dan manfaat lainnya adalah untuk memahami motif masing-masing segmen, karakteristik, dan kebutuhan (Hijrah, 2017). Dengan adanya informasi ini maka perusahaan dapat dengan mudah untuk menentukan dan merancang strategi pemasaran agar sesuai dengan target pasar, juga mendapat profit sesuai tujuan perusahaan tersebut.

Mengidentifikasi dan menentukan sesuai kelompok – kelompok tersebut iasanya dikumpulkan berdasarkan demografis, kebutuhan, atau perilaku konsumsi mereka. Setiap kelompok mempunyai nilainya masing – masing yang sudah disesuaikan dengan karakteristik konsumen. Segmentasi konsumen memiliki beberapa jenis seperti berikut ini (Sodexo, 2019):

A. Segmentasi Menurut Demografis

Segmentasi menurut demografis ini adalah mengelompokkan konsumen menurut data demografis atau kependudukan. Pengelompokan ini bisa berdasarkan usia, jenis kelamin, profesi, tingkat pendidikan hingga status pernikahan. Segmentasi konsumen ini adalah yang paling sering untuk dipakai berbagai perusahaan karena kemudahannya.

B. Segmentasi Menurut Geografi

Segmentasi menurut geografis adalah mengelompokkan konsumen berdasarkan kondisi geografis tempat mereka tinggal. Bisa mempertimbangkan iklim, jarak atau ukuran lokasi pada daerah konsumen yang akan menjadi target pasar. Biasanya segmentasi ini dilakukan yang memang menggunakan strategi pemasaran menggunakan jarak tempuh konsumen terhadap toko tersebut.

C. Segmentasi Menurut Perilaku Konsumsi

Segmentasi perilaku konsumsi yaitu menjadikan perilaku konsumen menjadi indikator utama pengelompokan. Perilaku yang dimaksudkan adalah bagaimana cara konsumen berinteraksi dengan barang atau jasa yang akan ditawarkan. Segmentasi ini

banyak dipakai oleh perusahaan yang memiliki berbagai metode pembelian seperti *online, offline store, application mobile* dan lainnya.

D. Segmentasi Menurut Siklus Hidup

Segmentasi menurut siklus hidup adalah segmentasi yang mementingkan siklus hidup konsumen atau *customer journey*. Siklus hidup konsumen ini menunjukkan sedang ada di tahap pembelian manakah mereka. Segmentasi jenis ini biasanya bertujuan untuk mempertahankan pembelian dari suatu konsumen atau menghadirkan sifat loyalitas pada konsumen terhadap produk perusahaan tersebut. Selain itu, segmentasi ini terjadi berulang atau terus menerus seperti ada seorang konsumen yang lebih memilih belanja *online* daripada datang langsung ke toko.

2.1.2 Proses Bisnis *E-Commerce*

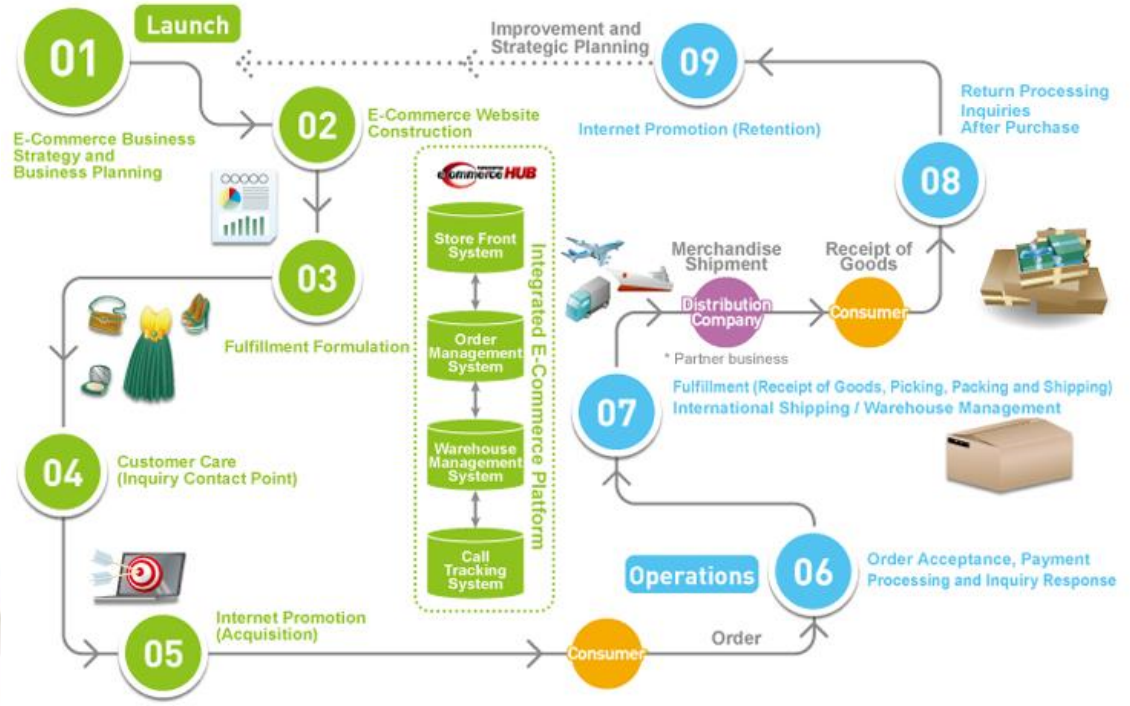
Pada dasarnya *E-Commerce* merupakan dampak dari berkembangnya teknologi informasi dan telekomunikasi sehingga secara signifikan mengubah cara manusia melakukan interaksi dengan lingkungannya, yang dalam hal ini adalah mekanisme dagang.

E-Commerce di era sekarang memang sedang memasuki puncak dan keberadaannya sangat terlihat yang membuat kemudahan seorang pembeli untuk melakukan pembelian barang yang mereka inginkan hanya dengan melalui smartphonenya.

Banyak keuntungan dan kerugian yang bisa didapatkan dari pihak perusahaan yang akan dirasakan setelah beralih menggunakan *E-Commerce.* Keuntungan bagi perusahaan adalah memperpendek jarak, perluasan pasar, perluasan mitra bisnis dan efisiensi, dengan kata lain mempercepat pelayanan ke konsumen. Pelayanan yang responsive ini juga yang menjadi diandalkan dalam proses bisnis *E-Commerce*.

Biaya untuk keperluan ATK atau alat tulis kantor pun akan berkurang karena semua dilakukan secara online dan digital jadi tidak perlu kesulitan kehilangan data yang diinginkan karena sudah tersimpan pada server. Kerugian untuk perusahaan adalah pembebanan dan perawatan secara server dan juga penyimpanan data secara digital yang membutuhkan *space* secara online yang terintegrasi dan *size* nya besar. Perawatannya pun harus dilakukan dan harus sangat diperhatikan.

Berikut adalah penggambaran flow dari salah satu *E-Commerce* yang sudah berkembang dan terintegrasi secara benar (Trans Cosmos, 2013).



Gambar II. 1. *Flow Business E-Commerce*

2.1.4 Klasifikasi *E-Commerce*

Secara umum *E-Commerce* dapat dibagi menjadi dua jenis yaitu:

A. *Business to Business* (B2B)

B2B atau *Business to Business* adalah system komunikasi bisnis online antar pelaku bisnis. Dalam B2B pada umumnya transaksi dilakukan oleh para trading partners yang sudah saling mengenal dengan format data yang telah disepakati (Pradana, 2015). Karateristiknya seperti dibawah ini

- Trading partners yang sudah diketahui atau umumnya memiliki hubungan yang cukup lama. Informasi pun hnaya dipertukarkan dengan partner tersebut atau bisa dikatakan lebih *private* atau rahasia.

- Pertukaran data *(Data Exhange)* berlangsung berulang-ulang dan secara berkala seperti setiap hari dengan format data yang sudah disepakati bersama. Dengan kata lain, servis yang digunakan sudah tertentu.

- Salah satu pelaku dapat melakukan inisiatif untuk mengirimkan data, tidak harus menunggu partner

- Model yang umum digunakan adalah *peer to peer* dimana seorang *processing intelligence* dapat didistribusikan di kedua pelaku bisnis.

B. *Business to Customers* (B2C)

Merupakan mekanisme toko online yaitu transaksi antara *e-merchant* dengan *e-customer*. Sedangkan dalam *Business to Customer* sifatnya terbuka untuk public, sehingga setiap individu dapat mengaksesnya melalui suatu web server. Berikut adalah karakteristik dari B2C:

- Terbuka untuk umum, dimana informasi disebarkan ke umum

- Servis yang diberikan bersifat umum dengan mekanisme ayng dapat digunakan oleh khalayak ramai.

- Servis diberikan berdasarkan permohonan (*On Demand).* Konsumen melakukan inisiatif dan produser harus siap memberikan respon sesuai dengan permohonan.

- Pendekatan client/server sering digunakan dimana diambil asumsi client menggunakan system minimal dan processing diletakkan disis server.

2.2 Unsupervised Machine Learning

Unsupervised machine learning adalah tipe algoritma hasil pengembangan dari machine learning. Selain *unsupervised machine learning* ada juga *supervised machine learning*. Namun secara teknis dua metode tersebut sangat lah berbeda.

2.2.1 Pengertian Unsupervised Machine Learning

Pada algoritma *unsupervised learning*, data tidak memiliki label secara eksplisit dan model mampu belajar dari data dengan menemukan pola yang implisit. Sangat berbeda dengan *supervised learning. Unsupervised machine learning* merupakan jenis *machine learning* yang hanya mempunyai variabel input tapi tidak mempunyai variabel output yang berhubungan (Roihan, Sunarya, & Rafika, 2020). Tujuan dari *machine learnin*g ini adalah untuk memodelkan struktur data dan menyimpulkan fungsi yang mendeskripsikan data tersebut.

Unsupervised machine learning adalah salah satu tipe algoritma machine learning yang digunakan untuk menarik kesimpulan dari dataset. Metode ini hanya akan mempelajari suatu data berdasarkan kedekatannya saja atau yang biasa disebut dengan clustering. Metode *unsupervised machine learning* yang paling umum adalah analisis cluster, yang digunakan pada analisa data untuk mencari pola-pola tersembunyi atau pengelompokan dalam data.

Cara kerja algoritma ini yaitu akan mencari pola tersembunyi (eksplisit) dari dataset yang diberikan. *Unsupervised machine learning* bekerja dengan menganalisis data yang tidak berlabel untuk menemukan pola tersembunyi dan menentukan korelasinya. Pendekatan ini tidak menggunakan data training dan data test untuk melakukan prediksi maupun klasifikasi dengan tujuan mengelompokkan objek yang hampir sama dalam suatu area tertentu. Beberapa contoh algoritma yang dapat digunakan dalam unsupervised learning seperti, *K-Means, Hierarchical clustering, DBSCAN, dan Fuzzy C-Means.*

2.2.2 Metode Pengelompokkan (*Clustering)*

Metode *clustering* merupakan proses pengelompokan sejumlah besar data menjadi beberapa kelas sesuai dengan ciri khasnya masing-masing. Algoritma clustering yang paling efisien untuk menentukan cluster pada data dengan kepadatan yang berbeda.

A. Metode *K-Means Clustering*

K-means clustering adalah salah satu algoritma analisis klaster (*cluster analysis*) non hirarki. Analisis klaster merupakan salah satu alat untuk mengelompokkan data berdasarkan variabel atau *feature*. Tujuan dari *k-means clustering*, seperti metode klaster lainnya, adalah untuk mendapatkan kelompok data dengan memaksimalkan kesamaan karakteristik dalam klaster dan memaksimalkan perbedaan antar klaster. (Pradana, 2015)

Algoritma *K-means clustering* mengelompokkan data berdasarkan jarak antara data terhadap titik centroid klaster yang didapatkan melalui proses berulang. Analisis perlu menentukan jumlah K sebagai input algoritma. Contoh pengaplikasiannya ada beberapa hal seperti :

- Segmentasi Pasar
- Segmentasi Citra
- Kompresi Gambar
- Klasifikasi Citra Pengindraan Jauh

Untuk tujuan-tujuan eksploratori, *K-Means* dapat dimanfaatkan untuk melengkapi proses *Exploratory Data Analysis* atau EDA, selain menggunakan analisis statistik deskriptif dan visualisasi data.

Sedangkan dalam proses confirmatori dan eksplanatori, *K-means clustering* dapat digunakan untuk melakukan konfirmasi terhadap teori-teori yang sudah ada. Selain itu, algoritma ini juga digunakan untuk melakukan identifikasi jika tiba-tiba terjadi perubahan cluster setelah data baru masuk. Algoritma atau metode *K-Means Clustering* ini dapat dilakukan dengan beberapa langkah seperti dibawah ini:

1. Tentukan jumlah cluster (k). dalam contoh ini, kita tetapkan bahwa k =3

2. Pilih titik acak sebanyak k. titik ini merupakan titik seed dan akan menjadi titik centroid proses pertama. titik ini tidak harus titik data kita
3. Label semua data berdasarkan titik centroid terdekat. semua data diberikan label mengikuti titik centroid dari setiap klaster. perhitungan jarak ini bisa menggunakan algoritma jarak tertentu, secara *default* dilakukan dengan *euclidean distance*
4. Tentukan titik centroid baru berdasarkan cluster yang terbentuk. titik centroid selanjutnya "berpindah" ke lokasi centroid setiap cluster yang telah terbentuk.
5. Label ulang data berdasarkan jarak terdekat terhadap centroid baru. langkah ini merupakan langkah yang sama dengan langkah ketiga. perhatikan titik data yang diberikan tanda panah, berubah dari cluster merah ke cluster biru.
6. Ulangi langkah 4 dan langkah 5 sampai tidak ada pergerakan lagi. secara berulang, algoritma akan mencari lokasi centroid baru dan melabel data berdasarkan centroid tersebut sampai mendapat hasil final, yaitu tidak ada lagi perpindahan centroid di setiap cluster.

B. Metode *Hierarchical Clustering*

Hierarchical clustering adalah metode pengelompokan dengan menggabungkan dua cluster terdekat. Algoritma clustering ini akan berakhir ketika hanya ada satu cluster yang tersisa. Biasanya, metode ini digunakan pada data yang jumlahnya tidak terlalu banyak dan jumlah cluster yang akan dibentuk belum diketahui.

Terdapat dua jenis metode *hierarchical clustering* yaitu *Agglomerative* atau strategi pengelompokan hirarki yang dimulai dengan setiap objek dalam satu cluster yang terpisah kemudian membentuk cluster yang semakin membesar. Jadi, banyaknya cluster awal adalah sama dengan banyaknya objek.

Yang kedua adalah *Divisive* atau strategi pengelompokan hirarki yang dimulai dari semua objek dikelompokkan menjadi cluster tunggal kemudian dipisah sampai setiap objek berada dalam cluster yang terpisah.

C. Metode *DB Scan*

DBSCAN adalah salah satu contoh pelopor perkembangan teknik pengelompokan berdasarkan kepadatan atau yang biasa dikenal dengan sebutan density based clustering. *Density-Based Spatial Clustering of Application with Noise* (DBSCAN) merupakan sebuah metode clustering yang membangun area berdasarkan kepadatan yang terkoneksi (densityconnected). Setiap objek dari sebuah radius area (cluster) harus mengandung setidaknya sejumlah minimum data. Semua objek yang tidak termasuk di dalam cluster dianggap sebagai *noise*.

D. Metode *Fuzzy C-Means*

Fuzzy clustering adalah salah satu teknik untuk menentukan cluster optimal dalam suatu ruang vektor yang didasarkan pada bentuk normal Euclidian untuk jarak antar vektor. Fuzzy C-Means (FCM) adalah suatu teknik pengklusteran data yang mana keberadaan tiap-tiap titik data dalam suatu cluster ditentukan oleh derajat keanggotaan.

Fuzzy C-Means (FCM) ini adalah untuk meminimalisasikan objective function yang diset dalam proses clustering, yang ada pada umumnya berusaha meminimalisasikan variasi di dalam suatu cluster dan memaksimalkan variasi antar cluster.

2.3 Bahasa Pemrograman

Banyak sekali bahasa pemrograman yang sudah berkembang dan memiliki fitur yang lengkap dan sangat bisa membantu dalam memudahkan dalam melakukan analisis data atau memudahkan dalam membuat sebuah aplikasi. Namun dalam menganalisis data kita membutuhkan bahasa pemrograman ini untuk melakukan konversi dari suatu data mentah (*raw data)* yang jumlahnya banyak menjadi suatu *insight.* Bahasa pemrograman peneliti pakai adalah bahasa pemrograman *R*.

2.3.1 Sejarah dan Fitur Bahasa Pemrograman *R*

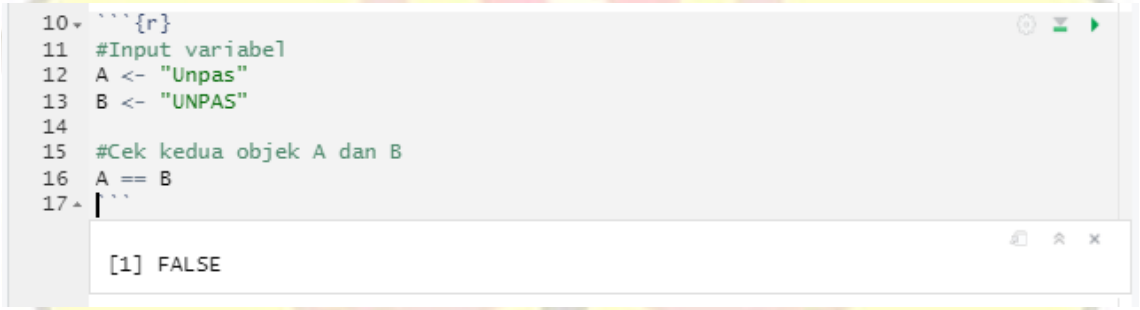
R merupakan bahasa yang digunakan dalam komputasi statistic yang pertama kali dikembangkan oleh Ross Ihaka dan Robert Gentlement di University of Auckland New Zealend yang merupakan akronim dari nama depan kedua pembuatnya. R dapat

dibilang merupakan aplikasi system statistic yang kaya, karena disebabkan banyaknya paket yang dikembangkan oleh pengembang dan komunitas untuk keperluan analisis statistik seperti *liner regression, clustering, statistical test* dan masih banyak lagi (Rosidin, 2019).

Sebagai sebuah bahasa pemrograman yang banyak digunakan untuk keperluan analisa data, R dapat dioperasikan pada berbagai system operasi pada computer. Adapun system operasi yang didukung antara lain *Unix, Linus, Windows* dan *MacOs*. R memiliki karakteristik yang beda dengan bahasa pemrograman lain, berikut beberapa ciri dan fitur pada R antara lain:

A. Bahasa R bersifat *case sensitive*

Maksudnya adalah dalam proses input huruf besar dan kecil sangat diperhatikan. Sebagai contoh adalah seperti berikut ini:



Gambar II. 2. Penggambaran *Case Sensitive* pada *RStudio*

Jika dilihat pada gambar hasil *capture* dari *RStudio* tersebut untuk isi yang memiliki kata yang sama namun dengan perbedaan huruf kapital maka tetap tidak bisa dikatakan sama atau disebut *False*.

B. Segala sesuatu yang ada pada program R akan dianggalp sebagai objek.

Konsep objek ini sama dengan bahasa pemrograma berbasis objek lainnya seperti *Java, C++*. Perbedaannya R ini relatif lebih sederahana dibandingkan bahasa pemrograman berbasik objek lainnya.

C. R sebagai *Interpreted Language* atau *script*

Bahasa R memungkinkan pengguna untuk melakukan kerja pada R tanpa perlu kompilasi kode program menjadi bahasa mesin.

D. R mendukung proses loop, decision making dan proses aritmatika

Bahasa R mendukung proses *loop* (pengulangan), *decision making* (penentuan keputusan) dan proses aritmatika seperti penjumlahan, pengurangan dan yang lainnya.

E. Mendukung ekspor dan impor berbagai format file

Selain itu R juga mendukung ekspor dan impor berbagai format file seperti *TXT, CSV, XLS,* dan yang lainnya. Dengan ini sangat memudahkan untuk melakukan impor dari berbagai data yang umum digunakan

F. Menyediakan berbagai fungsi untuk keperluan visualisasi data

Visualisasi data pada R dapat menggunakan paket bawaan yang sudah tersedia atau paket lain yang dapat didownload secara gratis pada website resmi R. Ini juga yang menjadi alasan peneliti menggunakan bahasa pemrograman R, karena mudah untuk membuat visualisasi data secara langsung dari suatu *dataset.*

2.3.2 Sejarah dan Fitur Bahasa Pemrograman *R*

Python dibuat dan dikembangkan oleh Guido Van Rossum, yaitu seorang programmer yang berasal dari Belanda. Pembuatannya berlangsung di kota Amsterdam, Belanda pada tahun 1990. Pada tahun 1995 Python dikembangkan lagi agar lebih kompatibel oleh Guido Van Rossum. Selanjutnya pada awal tahun 2000, terdapat pembaharuan versi Python hingga mencapai Versi 3 sampai saat ini.

Python sendiri adalah salah satu bahasa pemrograman yang dapat melakukan eksekusi sejumlah instruksi multi guna secara langsung (interpretatif) dengan metode orientasi objek (*Object Oriented Programming*) serta menggunakan semantik dinamis untuk memberikan tingkat keterbacaan syntax. Kelebihan Python sebagai bahasa pemrograman adalah sebagai berikut ini:

A. Mudah dipelajari terutama untuk pemula

Python menggunakan kode-kode yang ada mudah dibaca dan dipahami dengan menggunakan logic basic dari seorang pemula yang ingin mempelajari

B. Dapat menjalankan banyak fungsi kompleks

Python memiliki banyak standard library, sehingga penggunannya sangat efisien jika dilakukan.

C Mendukung multi platform dan multi system

Python mendukung multi platform seperti IOS, Windows atau Linux serta memiliki sistem pengelolaan memori otomatis seperti Java.

2.4 Jurnal Pembanding

Jurnal pembanding adalah hasil jurnal international yang berhubungan dengan penelitian yang dilakukan untuk memposisikan penelitian yang dilakukan berbeda dengan penelitian sebelumnya.

2.4.1 Jurnal berkaitan dengan Metode K-Means Clustering

Berikut beberapa jurnal yang didapatkan yang bisa menjadi pembanding dan referensi dalam penggunaan metode K-Means Clustering:

1. Penelitian Berbagai Tingkat Emisi Gas Rumah Kaca di Benua Eropa Dengan Menggunakan Metode K-Means.

Jurnal ini dibaut oleh Anna Kijewska dan Anna Bluszcz dari Departemen Manajemen Pertambangan dan Keselamatan, *Silesian University of Technology Poland.* Pada penelitian ini melibatkan 28 negara pada Benua Eropa sebagai objek penelitian dan memiliki 4 variabel yang terbagi atas beberapa jenis gas seperti karbon dioksida, metana, nitrogen oksida dan dinitrogen oksida. Hasil yang didapat yaitu menjelaskan pada Benua Eropa terdapat 4 cluster yang membagi tingkatan terhadap gas emisi rumah tangga (Kijewska & Bluszcz, 2016)

2. Implementasi *Clustering* dengan Metode K-Means untuk Menentukan Status Nutrisi.

Jurnal ini dibuat oleh Stefannny Surya Nagari dan Lilik Inayati dari departemen biostatistika dan populasi, fakultas kesehatan masyarakat, Universitas Airlangga. Penelitian ini dilakukan untuk mengelompokan status gizi anak usia dibawah 60 bulan yang dilakukan dengan metode C-Means Clustering. Penelitian ini adalah non-reaktif, menggunakan data sekunder di Ponkesdes Mayangrejo Kabupaten Bojonegoro, tanpa interaksi langsung dengan subjek. Penelitian ini menyimpulkan bahwa dapat dilakukan proses pengelompokan status gizi menggunakan K-Means dengan hasil 4 kluster yang terbentuk, terdiri dari 23 balita gizi buruk, 17 balita gizi kurang, 7 balita gizi baik dan 10 balita gizi lebih (Nagari & Inayati, 2020).

3. Implementasi K-Means Clustering Ujian Nasional Sekolah Menengah Pertama di Indonesia Tahun 2018/2019.

Jurnal ini dibuat oleh Agil Aditya, Ivan Jovian, Betha Nurina Sari dari fakultas ilmu kompuer, Teknik informatika, Universitas Singaperbangsa Karawang. Pada penelitian ini dilakukan clustering dengan menggunakan algoritma K-Means menggunakan data capaian Ujian Nasional Sekolah Menengah Pertama pada tahun 2018/2019 yang diperoleh dari website resmi Pusat Penilaian Pendidikan dan Kebudayaan Kementrian Pendidikan dan Kebudayaan Republik Indonesia.

Hasil cluster didapatkan untuk cluster 1 terdapat 14 provinsi, cluster 2 terdapat 5 provinsi, dan cluster 3 terdapat 15 provinsi. Tingkatan cluster 1 adalah cluster dengan nilai ujian nasional Tinggi, cluster 2 adalah cluster dengan nilai ujian nasional Rendah dan cluster 3 adalah cluster dengan nilai ujian nasional Sedang (Aditya, Jovian, & Sari, 2020)

4. Analisis Clustering Menggunakan Metode K-Means dalam Pengelompokkan Penjualan Produk pada Swalayan Fadhila.

Jurnal ini dibuat oleh Benri Melpa Metisen dan Herlina Latipa Sari dari program studi Teknik informatika, fakultas ilmu komputer, Universitas Dehasen Bengkulu. Dalam penelitian ini, digunakan penerapan clustering dengan menggunakan algoritma K-means. Dari data yang diolah dengan sampel data yang diambil di Swalayan Fadhilla Bengkulu, maka menghasilkan dua jenis kelompok data yaitu data penjualan rendah dan data penjualan tinggi sehingga dengan adanya pengelompokkan data ini pihak Swalayan Fadhilla dapat mengetahui jenis barang yang laris terjual dan tidak sehingga barang yang ada di gudang tidak menumpuk (Metisen & Sari, 2015).

5. Penerapan Metode Clustering K-Means untuk Menentukan Kategori Stok Barang

Jurnal ini dibuat oleh Elly Muningsih dari program studi manajemen informatika, AMIK BSI Yogyakarta. Penelitian ini menghasilkan 3 kelompok produk paling diminati untuk jumlah stok banyak, jumlah stok sedang untuk produk diminati dan jumlah stok sedikit untuk produk yang kurang atau tidak diminati. Pengolahan data dilakukan

menggunakan metode clustering yaitu metode K-Means berdasarkan data historis penjualan yang memuat kode produk, jumlah transaksi, volume penjualan dan rata-rata penjualan. Dari penelitian dihasilkan 3 anggota kelompok produk untuk stok banyak, 11 anggota kelompok untuk jumlah stok sedang, dan 17 anggota kelompok stok sedikit (Muningsih, 2014).

6. Penerapan *Data Mining* dalam Meningkatkan Mutu Pembelajaran pada Instansi Perguruan Tinggi Menggunakan Metode K-Means Clustering (Studi Kasus di Program Studi TKJ Akademi Komunitas Solok Selatan)

Jurnal ini dibuat oleh Koko Handoko dari Universitas Putera Batam. Penelitian ini menerapkan Data Mining dengan menggunakan metode clustering untuk meningkatkan mutu pembelajaran instansi perguruan tinggi di Program Studi TKJ Akademi Komunitas Solok Selatan. Pengujian dilakukan dengan aplikasi *RapidMiner 5.3* dan menggunakan 4 variabel yaitu IP mahasiswa, jarak tempuh mahasiswa, jumlah kehadiran dan penghasilan orang tua. Dimana akan mempresentasikan data mahasiswa dengan mutu pembelajaran sangat baik, baik, cukup baik dan kurang baik (Handoko, 2016).

2.4.2 Jurnal berkaitan dengan metode *Clustering* lainnya

Berikut beberapa jurnal yang didapatkan yang bisa menjadi pembanding dan referensi dalam penggunaan metode *Clustering* lainnya:

1. Implementasi Metode Clustering DBSCAN pada Proses Pengambilan Keputusan.

Jurnal ini dibuat oleh Ni Made Anindya Santika Devi, I Ketut Gede Darma Putra dan I Made Sukarsa dari Jurusan Teknologi Informasi, Universitas Udayana. Pada penelitian ini menggunakan *Spatial Data Clustering* yang merupakan salah satu teknik penting pada data mining yang digunakan untuk mendapatkan informasi atau pengetahuan pada data spasial dalam jumlah yang besar dari berbagai aplikasi. Hasil uji coba pada penelitian ini menunjukkan bahwa Metode DBSCAN telah berhasil melakukan proses clustering untuk membantu proses pengambilan keputusan dalam penentuan konsumen

potensial dengan membentuk sejumlah cluster (Santika Devi, Sukarsa, & Darma Putra, 2015).

2. Penerapan Data Mining ntuk Evaluasi Data Penjualan Menggunakan Metode *Clustering* dan Algoritma *Hirarki Divisive.*

Jurnal ini dibuat oleh Yuda Irawan dari Sistem Informasi, STMIK Hang Tuah Pekanbaru. Dalam penelitian ini digunakan Algoritma Hirarki Divisive untuk membentuk klaster-klaster. Dari pola yang diperoleh diharapkan dapat memberikan pengetahuan untuk perusahaan Media World Pekanbaru sebagai alat pendukung untuk mengambil kebijakan. Hasil analisa dan pengujian data menytakan bahwa data mining bermanfaat untuk menghasilkan pengetahuan berupa customer loyal yang ada di perusahaan Media World Pekanbaru dan setelah dievaluasi dengan menggunakan algoritma hirarki divisive serta pengolahan data dengan menggunakan software rapid miner maka ditemukan bahwa customer loyal berada pada cluster 3 dengan range 117-358 (Irawan, 2019).

3. Perbandingan Algoritma Fuzzy C-Means (FCM) dan Algoritma Mixture Dalam Penclusteran Data Curah Hujan Kota Bengkulu.

Jurnal ini dibuat oleh Herlina Latipa Sari dan Dewi Suranti sebagai Dosen Tetap Program Studi Teknik Informatika, Universitas Dehasen Bengkulu. Penelitian ini dilakukan untuk mendesain Fuzzy Clustering menggunakan algoritma C-Means dan algoritma Mixture dalam penclusteran data curah hujan Kota Bengkulu, membandingkan algoritma C-Means dan Algoritma Mixture dalam menghasilkan performasi algoritma C-Means dan lagoritma Mixture dalam menghasilkan tingkat keakuratan lokasi perkiraan curah hujan bulanan stasiun Klimatologi Pulau Baii Bengkulu. Hasil penelitian ini Algoritma FCM dapat membantu Badan Metereologi, Klimatologi dan Geofisika Stasiun Klimatologi Pulau Baii Bengkulu dalam mengelompokkan atau mengclusterkan data berdasarkan dengan sifat hujan (Sari & Suranti, 2016).

2.4.3 Penelitian yang akan dilakukan

Penentuan Segmentasi Konsumen Pada Marketing Data *IFood* Menggunakan Metode K – Means Clustering.

Penelitian yang dilakukan saat ini untuk menentukan segmentasi konsumen pada perusahaan *IFood* menggunakan metode K-Means Clustering dengan jumlah data konsumen sebanyak 2.240. Metode yang dilakukan ada visualisasi bar dan K-Means Clustering dengan berdasarkan beberapa variable yaitu jumlah pembelian produk per kategori, metode pembelian, dan tingkatan pendidikan konsumen.

2.4.4 Perbedaan dengan Jurnal Pembanding

Dari berbagai jurnal yang sudah peneliti cantumnkan pada sub bab sebelumnya terdapat beberapa perbedaan dengan penelitian pada jurnal ini seperti:

- Perbedaan objek penelitian yang digunakan
- Perbedaan variabel pada data yang digunakan
- Perbedaan aplikasi pengolahan data yang digunakan

Dari jurnal yang telah dibandingkan tersebut, peneliti dapat memahami konsep, ide, atau temuan utama yang terkait dengan penelitian yang akan dilakukan. Bagian diatas merupakan *literatur review* yang sudah dilakukan peneliti. Beberapa jurnal yang sudah dibahas akan dijelaskan menggunakan tabel berikut:

Tabel II. 1. Perbandingan Referensi Jurnal

Nama peneliti & tahun jurnal	Parameter			
	Permasalahan	Metoda yang digunakan	Bahasa pemprograman	Hasil penelitian/ kesimpulan
Elly Muningsih, 2014	Bagaimana pengelompokkan kategori stok barang pada toko online Ragam Jogja	K – Means Clusterin	RapidMiner	Terdapat 3 kluster yaitu 3 anggota kelompok stok banyak, 11 anggota kelompok untuk jumlah stok sedang dan 17 anggota

				kelompok stok sedikit
Benri Melpa Metisen & Herlina Latipa Sari, 2015	Bagaimana pengelompokkan penjualan produk pada swalayan Fadhilla Bengkulu untuk	K – Means Clustering	Tanagra	Terdapat 2 kluster yaitu data penjualan rendah dan data penjualan tinggi
Ni Made Anindya Santika Devi , I Ketut Gede Darma Putra , I Made Sukarsa, 2015	Bagaimana implementasi metode DBSCAN pada proses pengambilan keputusan untuk membantu perusahaan menentukan konsumen potensialnya	DBScan	Tidak dijelaskan	Menunjukkan Metode DBScan telah berhasil melakukan proses clustering untuk membantu proses pengambilan keputusan dalam penentuan konsumen potensial dengan membentuk sejumlah cluster
Anna Kijewska & Anna Bluszcz, 2016	Bagaimana pengelompokkan negara di Benua Eropa berdasarkan emisi gas rumah tangga yang	K – Means Clustering	Tidak dijelaskan	Terdapat 4 kluster/kelompok pada Benua Eropa berdasarkan 4 jenis gas emisi rumah tangga

	dihasilkan			
Koko Handoko, 2016	Bagaimana meningkatkan mutu pembelajaran pada Instansi Perguruan Tinggi di Program Studi TKJ Akademi Komunitas Solok Selatan	K – Means Clustering	RapidMiner	Terdapat 4 kluster yaitu mahasiswa dengan mutu pembelajaran sangat baik, baik, cukup baik dan kurang baik
Herlina Latipa Sari & Dewi Suranti, 2016	Bagaimana perbandingan data curah hujan Kota Bengkulu dengan 2 metode *Fuzzy C-Means* dan Algoritma *Mixture*	*Fuzzy C-Means* dan Algoritma *Mixture*	Matlab & SOCR	Didapatkan algoritma Fuzzy Clustering Means (FCM) dan Gaussian Mixture Modelling (GMM) adalah alternatif yang terbaik, yang dapat digunakan untuk memecahkan masalah dalam mengelompokkan data yang memiliki kesamaan jumlah data curah hujan yang sama atau

				mendekati, sehingga dapat digunakan sebagai pendukung pengambilan keputusan dalam mengelompokkan data.
Stefannny Surya Nagari & Lilik Inayati, 2019	Bagaimana pengelompokkan status gizi anak usia dibawah 60 bulan di Kabupaten Bojonegoro	*K – Means Clustering*	Tidak dijelaskan	Terdapat 4 kluster yang terbentuk, terdiri dari 23 balita gizi buruk, 17 balita gizi kurang, 79 balita gizi baik dan 10 balita gizi lebih
Yuda Irawan, 2019	Bagaimana pengelompokkan *customer loyal* pada Media World Pekanbaru	Algoritma *Hirarki Divisive*	Accurate & RapidMiner	Terdapat 3 kluster dengan range 117-358 yaitu Mater Q, Cash, Istana Print, Sahabat Adv, Jasa Reklame, Brilian Adv, Kreasi Adv, Family Print, Dian Print, Banner Teddy, Galaxi, Multi

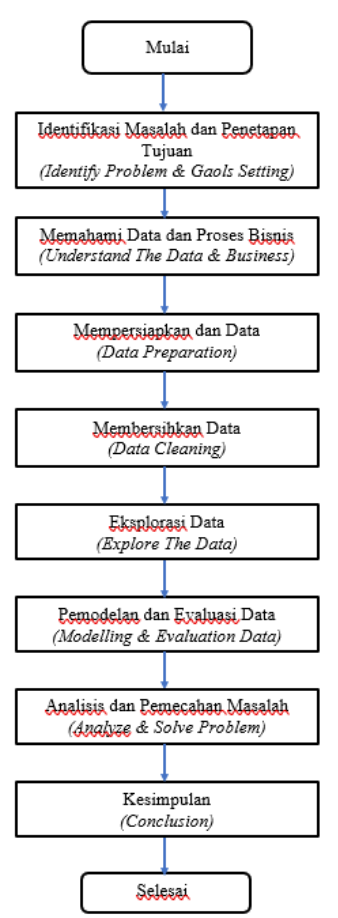
				Baru, Print Art, Master Print, Zoom Reklame, Media World Jambi, Image Print dan WSN Adv
Agil Aditya, Ivan Jovian, Betha Nurina Sari, 2020	Bagaimana pengelompokkan nilai ujian nasional sekolah menengah pertama di Indonesia Tahun 2018/2019	K – Means Clustering	R Studio	Terdapat 3 kluster yang terdiri dari 14 provinsi dengan niai ujian nasional tinggi, 5 provinsi dengan nilai ujian rendah dan 15 provinsi dengan nilai ujian sedang

Bab III
Metodologi Penelitian

Metodologi penelitian berisi tahap-tahap yang akan dilakukan dalam proses penelitian. Metodologi penelitian yang akan dilakuka yaitu metodologi penelitan kuantitatif.

3.1 Langkah – Langkah Penelitian

Dalam memecahkan masalah yang ada, maka perlu disusun secara sistematis langkah – langkah dari pemecahan masalah tersebut agar terarah. Adapun langkah – langkah usulan pemecahan masalah yang dilakukan dalam penelitiann di perusahaan *IFood* digambarkan pada flowchart dibawah ini:



Gambar III. 1. *Flow Chart* Langkah - Langkah Penelitian

3.1.1 Identifikasi Masalah dan Penetapan Tujuan

Mengidentifikasi masalah adalah hal pertama yang harus dilakukan dan sangat penting karena ini adalah upaya untuk mendefinisikan masalah pada penelitian tersebut. Masalah yang terjadi pada perusahaan *IFood* adalah belum adanya pengelompokkan konsumen sehingga strategi pemasaran yang dilakukan belum sesuai dengan target pasar

Setelah mengidentifikasi masalah lalu ditetapkannya tujuan yang bisa dilakukan pada sebuah penelitian. Tujuan yang akan dilakukan pada penelitian ini adalah mengelompokkan *(clustering)* konsumen perusahaan *IFood* dengan menggunakan metode *K-Means Clustering* dari beberapa kategori yang sudah ditentukan, seperti jumlah pembelian produk per kategori, metode pembelian yang dipakai konsumen, dan tingkatan pendidikan konsumen

3.1.2 Memahami Data dan Proses Bisnis

Langkah selanjutnya yaitu memahami data yang akan digunakan dan proses bisnis yang digunakan. Langkah ini akan memahami variable apa saja yang akan dianggap penting untuk bisa mendapatkan tujuan awal yaitu mendapatkan segmentasi konsumen. Identifikasi proses bisnis yang digunakan untuk mengetahui seberapa pengaruh dari satu variable ke variable lainnya, agar terjadinya kesinambungan untuk mengolah data sesuai tujuan peneliti.

A. Jenis dan Sumber Data

Sumber data dalam penelitian ini didapatkan dari *Kaggle.com* yang merupakan website penyedia data yang bisa digunakan oleh seorang *data analyst* atau *data science*. Data yang sudah berlisensi *CCO: Public Domain* oleh *Creative Common Org* sehingga bisa digunakan oleh umum atau siapa saja.

Data ini berisi tentang data penjualan dari salah satu startup besar di Brazil yaitu *IFood* yang merupakan penyedia jasa pemesanan dan pengantaran makanan secara online yang langsung beroperasi di Brazil dan Colombia. Data yang telah tersedia ini sudah berbentuk dokumen excel atau biasa disebut *raw data* (data kasar). *Raw data*

adalah data kasar yang berbentuk *original* atau asli sesuai dari hasil yang didapatkan dari survei pengguna tanpa dirubah sedikit pun.

Data yang dipakai ini memiliki 2.240 konsumen yang sudah diisi oleh konsumen dari perusahaan *Ifood* tersebut. Selain itu juga data tersebut memiliki 28 variabel yang berisi tentang profil konsumen, jumlah pembelian per kategori barang dan juga ketepatan *campaign* yang dilakukan perusahaan tersebut. Data tersebut diterbitkan pada tahun 2020 dan berisi penjualan dari tahun 2018 sampai tahun 2020.

B. Variabel Data

Terdapat 28 variabel yang ada pada data perusahaan *IFood* dan dijelaskan sebagai tabel dibawah ini:

Tabel III. 1. Variabel Pada Data

Atribut	**Tipe data**	**Keterangan**
ID	Numerik	Nomor konsumen
Year_Birth	Numerik	Tahun lahir konsumen
Education	Karakter	*Level* pendidikan konsumen
Marital_Statis	Karakter	Status konsumen
Income	Karakter	Penghasilan konsumen per tahun
KidHome	Numerik	Jumlah anak kecil yang dimiliki dalam keluarga
TeenHome	Numerik	Jumlah remaja yang dimiliki dalam keluarga
Dt_Customer	Karakter	Tanggal konsumen melakukan pendaftaran dengan perusahaan
Recency	Numerik	Jumlah hari sejak terakhir pembelian
MntWines	Numerik	Jumlah pembelian produk wines
MntFruits	Numerik	Jumlah pembelian produk buah
MntMeatProducts	Numerik	Jumlah pembelian produk daging
MntFishProducts	Numerik	Jumlah pembelian produk ikan
MntSweetProducts	Numerik	Jumlah pembelian produk manis
MntGoldProds	Numerik	Jumlah pembelian produk emas
NumDealsPurchases	Numerik	Jumlah pembelian menggunakan diskon
NumWebPurchases	Numerik	Jumlah pembelian pada website

NumCatalogPurchases	Numerik	Jumlah pembelian pada katalog/aplikasi
NumStorePurchases	Numerik	Jumlah pembelian pada toko
NumWebVisitsMonth	Numerik	Jumlah kunjungan pada website
AcceptedCmp1	Numerik	Konsumen menerima penawaran pada *campaign* pertama
AcceptedCmp2	Numerik	Konsumen menerima penawaran pada *campaign* kedua
AcceptedCmp3	Numerik	Konsumen menerima penawaran pada *campaign* ketiga
AcceptedCmp4	Numerik	Konsumen menerima penawaran pada *campaign* keempat
AcceptedCmp5	Numerik	Konsumen menerima penawaran pada *campaign* kelima
Complain	Numerik	Jumlah komplain yang diberikan
Country	Karakter	Negara asal konsumen
Age	Numerik	Umur konsumen

3.1.3 Mempersiapkan Data

Mempersiapkan data yang akan dilakukan peneliti adalah mempersiapkan mulai dari format data yang digunakan yaitu excel untuk mempermudah menginput kedalam *R program*, bahasa pemrograman *R program* yang nanti akan digunakan, *function* dan *library* pada bahasa pemrograman yang akan dipakai hingga bentuk penyajian yang perlu dirancang oleh peneliti.

3.1.4 Membersihkan Data

Membersihkan data yaitu mencari *gap* atau data yang kosong dan tidak terbaca oleh *R program* dan juga tidak bisa dikarakteran secara variable. Setelah itu data yang dibersihkan akan dilakukan persamaan secara format agar bisa terbaca jelas oleh *R program* sehingga valuenya tetap bisa terdefinisi.

Secara garis besar nantinya proses ini akan dibantu *function* yang sudah disediakan *R Studio* sehingga memudahkan untuk melakukan proses ini. Ada 3 pembagian dalam bahasa pemrograman yang akan dilakukan pada data berikut yaitu:

- Mengatur data yang hilang (*handling missing value).*
- Mengatur tipe data kepada setiap kategori (*handling data type).*

– Mengatur data yang tidak normal secara nilai atau biasa disebut sebagai *outliers* (*handling outliers).*

3.1.5 Eksplorasi Data

Eksplorasi data yang berarti langkah untuk memahami data sebelum dilakukan diproses atau biasa disebut *Explore Data/Data Mining.* Pemahaman terhadap data yang akan dieksplorasi ini dapat membantu dalam menentukan teknik-teknik pra-proses dan analisis data terhadap data awal tersebut.

Dalam eksplorasi data peneliti akan mendapatkan analisis dari setiap variable yang sudah dipilih sebelumnya oleh peneliti yang dianggap berkaitan dengan tujuan dari penelitian ini, seperti jumlah pembelian produk per kategori, metode pembelian, dan tingkatan pendidikan konsumen.

3.1.6 Pemodelan dan Evaluasi Data

Pemodelan data adalah menghubungkan berbagai elemen data berbeda untuk mengetahui informasi yang dibutuhkan sesuai tujuan. Pemodelan data ini menekankan apa saja yang akan dibutuhkan dan mencari korelasi yang besar antar setiap elemen data. Tujuan utama dari pemodelan data ini adalah untuk menciptakan metode penyimpanan informasi yang paling efisien serta menyedikaan akses dan pelaporan yang lengkap.

Pemodelan bisa digunakan dengan berbagai metode salah satunya *K-Means Clustering* yang akan digunakan pada penelitian ini. *K-Means Clustering* merupakan bagian dari *unsupervised learning* yang merupakan jenis *machine learning* yang hanya mempunyai variabel input tapi tidak mempunyai variabel output yang berhubungan. *K-Means Clustering* bertujuan mendapatkan kelompok data dengan memaksimalkan kesamaan karakteristik dalam klaster dan memaksimalkan perbedaan antar klaster.

F. Analisis dan Kesimpulan

Langkah terakhir ini berisi tentang hasil analisis dari pemodelan dan pengolahan data yang sudah dilakukan yang berbentuk *bar chart* dan *plot chart*. Dilanjutkan dengan menarik kesimpulan dari hasil yang didapatkan agar bisa menentukan segmentasi konsumen pada perusahaan *IFood* tersebut.

3.1.7 Analisis dan Pemecahan Masalah

Analisis ini dilakukan terhadap hasil dari pemodelan dan evaluasi yang sudah dilakukan dengan metode *K-Means Clustering.* Berdasarkan hasil dari metode tersebut, akan diperoleh kluster konsumen berdasarkan variable – variable yang sudah ditentukan dalam tujuan awal yaitu jumlah pembelian produk per kategori, metode pembelian, dan tingkatan pendidikan konsumen. Sehingga perusahaan *IFood* nantinya akan dengan mudah menentukan strategi pemasaran yang tepat yang bisa dilakukan untuk menambah *profit* perusahaan tersebut.

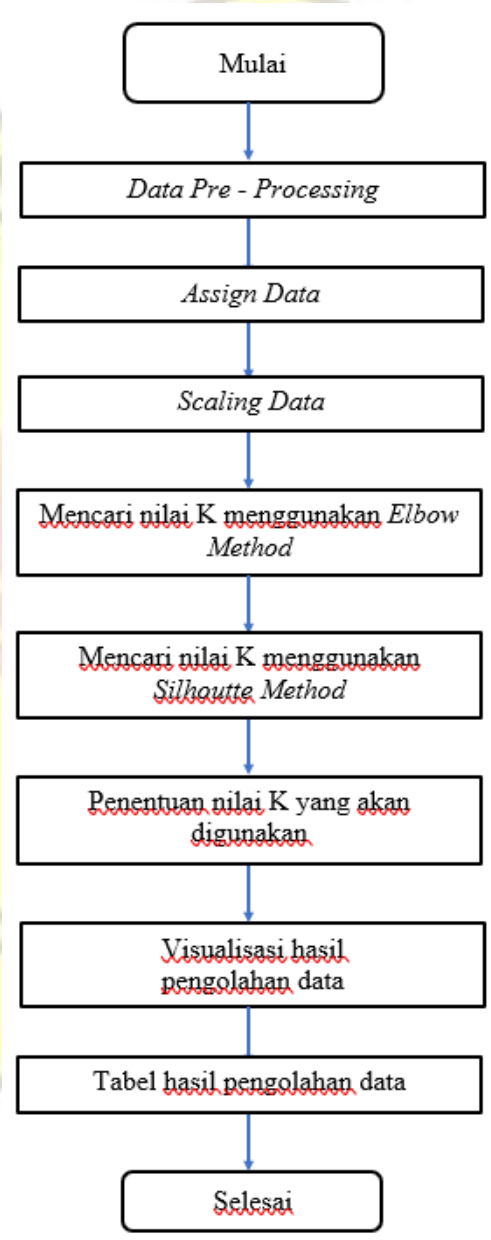
Hasil analisis tersebut akan digambarkan dengan visualisasi data berupa *bar plot, cluster plot* dan *line plot*, yang dimaksudkan untuk memudahkan perusahaan *IFood* membaca hasil analisis dan peneliti dapat memecahkan masalah yang ditetapkan.

3.1.8 Kesimpulan

Pada tahap ini merupakan langkah akhir berupa kesimpulan hasil pemecahan masalah yang mencerminkan jawaban atas tujuan penelitian yang telah ditetapkan sebelumnya.

3.2 Model Pemecahan Masalah

Model pemecahan masalah yang dilakukan disesuaikan dengan metode *K-Means Clustering*, sehingga didapatkan hasil akhir yang akan menentukan beberapa kluster konsumen terhadap variable yang ditentukan diawal. Berikut ini adalah langkah -langkah secara detail untuk pemodelan dan evaluasi *K-Means Clustering* yang dilakukan:



Gambar III. 2. *Flow Chart* Model Pemecahan Masalah

3.2.1 Data Pre-Processing

Dalam penelitian ini tidak terdapat *null* atau data yang tidak bernilai sama sekali maka bisa dilakukan langkah selanjutnya. Namun pada data ini terdapat *missing values* maka perlu dilakukan pemrosesan menghilangkan nilai tersebut pada data yang akan dipakai. Selain itu juga dalam *data pre-processing* ini dilakukan pengubahan tipe data untuk yang memiliki tipe karakter menjadi faktor. Pengubahan ini dilakukan untuk melakukan pengolahan dan visualisasi data, ditambah lagi ketentuan untuk pemodelan perlu tipe data yang berjenis numerik.

Faktor adalah tipe data yang memasukkan nilai kepada setiap angka sehingga didapatkan variabel tersebut berkategori numerik jika dilakukan pemrosesan selanjutnya. Contohnya adalah kategori *Marital_Status* yang diubah menjadi faktor sehingga untuk *divorced* bernilai 1, *married* bernilai 2, *single* bernilai 3, *together* bernilai 4, dan *widow* bernilai 5.

Variabel yang akan dirubah tipe datanya menjadi faktor adalah *Education*, *Marital_Status*, dan *Country*. Selain itu variabel lainnya yang dilakukan pengubahan tipe data adalah *dt_customer* menjadi tipe data *date* yang merupakan tipe data penunjukkan tanggal untuk memudahkan melakukan *data mining.*

Dalam *data pre-processing* juga terdapat pemilihan data yang merupakan tipe numerik untuk menuju pemodelan dengan metode *K-Means Clustering.* Setiap variabel yang sudah ditentukan pada tujuan akan melalui tahap ini jika belum berbentuk tipe data numerik.

3.2.2 Assign Data

Langkah ini merupakan penentuan atau pemilihan variabel terkait yang sudah ditentukan pada tujuan awal untuk memasukkan nilainya pada 1 *data frame* yang bertujuan untuk lebih memilih secara detail setiap variabel yang akan dilakukan pemodelan. Pada tahap ini akan terdapat 3 *data frame* baru yang mewakili setiap variabel tersebut yaitu *data frame* terkait pembelian produk per kategori, *data frame* terkait metode pembelian, dan *data frame* terkait tingkatan pendidikan konsumen.

Secara pemrograman dalam *R Studio* akan dimasukkan dengan *function* yaitu *Select(data frame)*. Akan digambarkan seperti dibawah ini untuk *assign* beberapa variabel menjadi data frame baru yaitu *numeric_data_k2*

```
#K Means - Total Amount
```{r}
#Assign by amount of items
numeric_data_k2 <- numeric_data %>%
  select(c(MntWines, MntFruits, MntMeatProducts, MntFishProducts, MntSweetProducts, MntGoldProds))
```
```

Gambar III. 3. *Assign* Beberapa Variabel Menjadi Data Frame *numeric_data_k2*

3.2.3 Scaling Data

Scaling data bertujuan untuk mencegah *outliers* atau nilai yang nilainya *abnormal* dari data biasanya. *Scaling data* ini juga biasanya disebut normalisasi data atau menormalkan suatu *data frame* yang sudah ditentukan. Proses *scale data* ini pada *R Studio* hanya tinggal memasukkan *function* yaitu *Scale(data frame).* Adapun tahapan proses yang dilakukan yaitu:

a. Mencari nilai maksimum dan minimum dari masing – masing variable

   Contoh peneliti akan mengambil nilai pada variabel *MntWines*

   Nilai maksimum (Xmaks)= 769

   Nilai minimum (Xmin) = 2

b. Menghitung nilai normalisasi menggunakan persamaan:

$$\text{Normalisasi} = \frac{(\text{Nilai awal} - \text{Nilai minimal})}{(\text{Nilai maksimal} - \text{Nilai minimal})}$$

c. Menghitung nilai normalisasi menggunakan persamaan diatas

   Contoh: *MntWines* = (J2 – Xmin) / (Xmaks – Xmin)

   = (189 – 2) / (769 – 2)

   = 0.243

Setelah itu memunculkan bagian yang hanya mewakili dari hasil tersebut, peneliti menggunakan *function head(data frame)* untuk memunculkan dan membuktikan data tersebut sudah hasil *scaling*.

3.2.4 Mencari nilai K menggunakan *Elbow Method*

Elbow Method adalah metode yang digunakan untuk menginterpretasikan dan uji performa tingkat konsistensi jumlah cluster yang tepat dengan melihat nilai SSE (Aditya, Jovian, & Sari, 2020). Pada titik tertentu akan terjadi grafik penurunan secara drastis dengan sebuah lekukan yang disebut dengan kriteria siku. Setelah itu nilai tersebut kemudian menjadi nilai K atau jumlah cluster yang terbaik.

Untuk mendapatkan perbandingannya adalah dengan menghitung SSE (*Sum of Square Error*) dari masing-masing nilai *cluster*. Karena semakin besar jumlah *cluster* K maka nilai SSE akan semakin kecil. SSE sendiri memiliki persamaan seperti ini:

$$\text{SSE} = \sum^{n}_{i=1}(Xi - \bar{X})^2$$

Namun jika dalam *R Studio* dapat dibantu dengan *function* yang sudah disediakan yaitu *fviz_nbclust (data frame, kmeans, method = 'wss').* Setelah itu digambarkan langsung menjadi *line plot*.

3.2.5 Mencari nilai K menggunakan *Silhouette Method*

Sebagai pembanding lain maka terdapat metode pendukung lain yang bisa dijadikan pilihan yaitu *Silhouette Method.* Metode ini merupakan menghitung rata-rata nilai setiap titik pada himpunan data. Perhitungan nilai setiap titik adalah selisih nilai *separation* dan *compactness* yang dibagi dengan maksimum antara keduanya. Jumlah klaster yang terbaik ditunjukkan dengan nilai *Silhouette* yang semakin mendekati 1.

Pada *R Studio* untuk mendapatkan nilai dari *Silhouette Method* ini bisa menggunakan *function fviz_nbclust (data frame, kmeans, method = 'silhouette').* Setelah itu digambarkan langsung menjadi *line plot* untuk memudahkan menentukan nilai K yang optimal.

3.2.6 Penentuan nilai K yang akan digunakan

Penentuan nilai K yang optimal ini menggunakan hasil dari kedua metode yang sudah dilakukan sebelumnya yaitu *Elbow Method* dan *Silhouette Method*. Jika nilai dari K tersebut sama atau mendekati maka dinyatakan nilai tersebut adalah yang paling optimal pada data frame tersebut.

Pada *R Studio* perlu dilakukannya *assign value* terhadap nilai K yang optimal kepada *data frame* yang sebelumnya sudah digunakan. Penggunaan *function* untuk proses ini adalah *data frame = kmeans(data frame, centers = K optimum, nstart = 50).* Pada *function* yang telah digunakan maka akan mendapat *value* baru yang bisa dilakukan visualisasi untuk memudahkan analisis hasilnya.

3.2.7 Visualisasi hasil menggunakan *Cluster Plot*

Visualisasi digunakan untuk memudahkan peneliti membaca hasil dari pemodelan tersebut. Pada *R Studio* bisa menggunakan *function* yang sudah terdapat sebelumnya dengan memasukkan *library* visualisasi data. *Function* tersebut adalah *fviz_cluster = (data frame, data = data frame).*

Hasil yang akan muncul adalah *cluster plot* dengan jumlah kluster yang sudah didapatkan dalam 2 metode sebelumnya. Namun dalam *cluster plot* tersebut masih bisa melihat nilainya secara detail dengan menggambarkan pada tabel.

3.2.8 *Clustering* dalam tabel

Clustering dalam tabel ini berfungsi untuk melihat nilai yang muncul dari nilai K optimal yang sudah ditentukan sebelumnya. Pada proses ini akan didapat nilai yang mewakili dari setiap variabel yang sudah ditentukan sebelumnya dan bisa dilihat perbedaan nilainya untuk setiap kluster tersebut.

Untuk melihat *clustering* dalam tabel bisa dilakukkan dengan membuat *data frame* baru untuk memudahkan dan memisahkan dari *data frame* sebelumnya yang sudah diberi *function* lainnya. Untuk melihat hasilnya peneliti hanya perlu melihat secara *means* atau rata – rata saja dan dimunculkan secara tabel dalam *R Studio* dengan bantuan *function* berikut:

Data frame(, related variabel) %>%

Mutate(cluster = data frame table$data frame cluster) %>%

Group_by(cluster) %>%

Summarise_all(“means”)

Bab IV
Pengumpulan dan Pengolahan Data

4.1 Pengumpulan Data

Pengumpulan data merupakan langkah awal sebagai input untuk meyelesaikan penelitian menentukan segmentasi konsumen pada marketing data *IFood* dengan menggunakan metode *K-Means Clustering.* Adapun data-data yang diperlukan adalah data-data tentang jumlah pembelian produk per kategori, jumlah pembelian konsumen berdasarkan metode pembelian, dan juga tingkat pendidikan konsumen. Data tersebut didapatkan dari *Kaggle* atau website yang menyediakan data yang bisa digunakan untuk analisis dan pemodelan dalam bidang *data science* dan *data mining*.

4.1.1 Profil Perusahaan

IFood merupakan perusahaan penyedia dan pengiriman makanan dan kebutuhan lainnya secara online yang beroperasi di Brazil. Produk yang disediakan pada perusahaan ini ada beberapa kategori yaitu: jenis - jenis minuman anggur *(wines),* jenis - jenis daging *(meat),* jenis – jenis ikan *(fish),* jenis – jenis makanan manis *(sweet products),* jenis – jenis produk emas *(golds products).* Perusahaan ini juga sudah memiliki beberapa cabang diberbagai negara seperti: Australia, Kanada, Jerman, India, Montenegro, Afrika Selatan, Spanyol dan Amerika Serikat.

4.1.2 Visi dan Misi Perusahaan

Adapun visi dan misi *IFood* adalah sebagai berikut:

1. Visi:

a. Menjadi perusahaan teknologi terbesar di Amerika Latin

b. Menjadi perusahaan terdepan yang menghubungkan dunia makanan dengan jutaan konsumen

c. Menjadi perusahaan yang memimpin digitalisasi dalam pemesanan makanan ehingga bisa membantu eksosistem manusia

2. Misi:

a. Mengutamakan kualitas pada ekosistemnya yaitu: konsumen, petugas pengiriman dan restoran

b. Produktif dan Inovatif

4.1.3 Alur Pembelian Produk pada Perusahaan

Pada perusahaan *IFood* untuk membeli sebuah produk yang sudah disediakan konsumen bisa melakukan beberapa metode pembelian seperti:

1. Pembelian secara langsung ke toko

Konsumen bisa melakukan pembelian produk yang diinginkan secara langsung ke toko yang sudah disediakan oleh perusahaan. Konsumen akan memilih langsung produknya sesuai dengan sistem *supermarket* yaitu mengambil dan membayar kepada kasir secara langsung.

2. Pembelian melalui website

Metode pembelian yang kedua adalah melalui website resmi *IFood* yaitu *https://www.ifood.com.br/*. Konsumen dapat mengunjungi website tersebut dan melakukan registrasi terkait tempat dan profil konsumen, lalui bisa memilih secara langsung produk yang akan dipesan. Setelah itu produk akan diantarkan sekaligus oleh pengantar produk menuju lokasi pengantaran sesuai dengan waktu yang sudah diberikan dan harga yang sudah ditentukan oleh perusahaan.

3. Pembelian melalui aplikasi

Metode ketiga adalah pemesanan produk melalui aplikasi *IFood* yang bisa diunduh secara langsung melalui *App Store* dari masing-masing ponsel konsumen. Secara singkat konsumen pertama-tama harus mengunduh aplikasi *IFood*, lalu konsumen mengisi data diri terkait profil konsumen dan lokasi pengantaran. Setelah itu konsumen dapat memilih produk yang akan dipesan, dan algoritma aplikasi akan memunculkan harga yang sudah disesuikan dengan jarak dari tempat/restoran tersebut sebagai biaya pengiriman.

Setelah makanan dipesan maka algoritma aplikasi tersebut akan meneruskan kepada tempat/restoran yang akan dipesan juga tempat pengantar makanan yang berada disekitar tempat/restoran tersebut. Tempat dan juga pengantar makanan akan

mengkonfirmasi, lalu setelah itu produk tersebut diantarkan ke lokasi tempat pengantaran yang sudah ditentukan diawal.

4.1.4 Data yang diperoleh

Data yang diperoleh merupakan data marketing perusahaan *IFood* yang disediakan oleh *Kaggle.* Data yang digunakan sudah berlisensi *public data* sehingga bisa digunakan oleh siapa saja (Creative Commons, 2017). Data yang sudah dikumpulkan berisi 2.240 data, dengan 28 variabel, berbentuk *raw data* atau data mentah, dan juga menggunakan format *excel* sehingga perlu dilakukan pengolahan data. Berikut adalah beberapa variabel dari data tersebut:

Tabel IV. 1. Variabel dan Penjelasan Pada Data

Atribut	**Tipe data**	**Keterangan**
ID	Numerik	Nomor konsumen
Year_Birth	Numerik	Tahun lahir konsumen
Education	Karakter	*Level* pendidikan konsumen
Marital_Statis	Karakter	Status konsumen
Income	Karakter	Penghasilan konsumen per tahun
KidHome	Numerik	Jumlah anak kecil yang dimiliki dalam keluarga
TeenHome	Numerik	Jumlah remaja yang dimiliki dalam keluarga
Dt_Customer	Karakter	Tanggal konsumen melakukan pendaftaran dengan perusahaan
Recency	Numerik	Jumlah hari sejak terakhir pembelian
MntWines	Numerik	Jumlah pembelian produk wines
MntFruits	Numerik	Jumlah pembelian produk buah
MntMeatProducts	Numerik	Jumlah pembelian produk daging
MntFishProducts	Numerik	Jumlah pembelian produk ikan
MntSweetProducts	Numerik	Jumlah pembelian produk manis
MntGoldProds	Numerik	Jumlah pembelian produk emas
NumDealsPurchases	Numerik	Jumlah pembelian menggunakan diskon
NumWebPurchases	Numerik	Jumlah pembelian pada website

NumCatalogPurchases	Numerik	Jumlah pembelian pada katalog/aplikasi
NumStorePurchases	Numerik	Jumlah pembelian pada toko
NumWebVisitsMonth	Numerik	Jumlah kunjungan pada website
AcceptedCmp1	Numerik	Konsumen menerima penawaran pada *campaign* pertama
AcceptedCmp2	Numerik	Konsumen menerima penawaran pada *campaign* kedua
AcceptedCmp3	Numerik	Konsumen menerima penawaran pada *campaign* ketiga
AcceptedCmp4	Numerik	Konsumen menerima penawaran pada *campaign* keempat
AcceptedCmp5	Numerik	Konsumen menerima penawaran pada *campaign* kelima
Complain	Numerik	Jumlah komplain yang diberikan
Country	Karakter	Negara asal konsumen
Age	Numerik	Umur konsumen

4.2 Pengolahan Data

Pada pengolahan data dilakukan dengan mengolah data yang telah dikumpulkan sebelumnya dengan metode *K-Means Clustering* dengan bantuan bahasa pemrograman *R* dengan aplikasi *RStudio*. Tahapan pengolahan data sudah dijelaskan pada sub bab 3.3.2.

4.2.1 Memasukkan *library* pada *RStudio*

Tujuan memasukkan library adalah untuk membantu dalam pengolahan dan pemodelan data dengan *K-Means Clustering* dikarenakan secara *default* bahasa pemrograman *R* sendiri tidak memiliki banyak *function* yang bisa dipakai. Berikut adalah beberapa *library* yang akan dipakai beserta dengan *code* pada *RStudio*:

Tabel IV. 2. Daftar *Library* yang Akan Dipakai Pada *RStudio*

Library	*Code*	Keterangan
readr	*library*(readr)	Digunakan untuk pemanggilan atau input data dalam format csv.
ggplot2	*library*(ggplot2)	Digunakan untuk menghasilkan/menampilkan data dan visualisasi grafik
tidyverse	*library*(tidyverse)	Digunakan untuk melakukan pengolahan data
visdat	*library*(visdat)	Digunakan untuk memberikan visualisasi pada seluruh *data frame* pada satu waktu
tidyr	*library*(tidyr)	Digunakan untuk merapikan data

assertive	*library*(assertive)	Digunakan untuk melakukan pemeriksaan *function* untuk memastikan integritas *code*
scales	*library*(scales)	Digunakan untuk membuat panduan grafik, membaca grafik dan menerapkan sistem grafis pada suatu *function*
dplyr	*library*(dplyr)	Digunakan untuk meanipulasi/transformasi data
cluster	*library*(cluster)	Digunakan untuk melakukan proses *clustering*
factoextra	*library*(factoextra)	Digunakan untuk melakukan proses *clustering* dan visualisasi

4.2.2 *Input* dan Membaca *Dataset*

Proses selanjutnya adalah *input*, membaca dan melihat secara *general* seperti apa marketing data *IFood* yang berformat .csv tersebut menggunakan *RStudio*. Sintaks yang digunakan adalah sebagai berikut ini:

```
data <- read_csv("marketing_data.csv")

str(data)
```

Penggunaan *library (readr)* yaitu untuk melakukan pemanggilan atau *input* data, lalu dilanjutkan dengan sintaks *read_csv("marketing_data.csv")* yang merupakan nama file data yang sudah tersimpan pada sistem komputer. *Str(data)* digunakan untuk melihat secara *general* berupa tipe data dan beberapa nilai awal dari setiap variabel. Hasil dari sintaks tersebut seperti gambar dibawah ini,

```
spec_tbl_df [2,240 x 28] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
 $ ID                 : num [1:2240] 1826 1 10476 1386 5371 ...
 $ Year_Birth         : num [1:2240] 1970 1961 1958 1967 1989 ...
 $ Education          : chr [1:2240] "Graduation" "Graduation" "Graduation" "Graduation" ...
 $ Marital_Status     : chr [1:2240] "Divorced" "Single" "Married" "Together" ...
 $ Income             : chr [1:2240] "$84,835.00" "$57,091.00" "$67,267.00" "$32,474.00" ...
 $ Kidhome            : num [1:2240] 0 0 0 1 1 0 0 0 0 0 ...
 $ Teenhome           : num [1:2240] 0 0 1 1 0 0 0 1 1 1 ...
 $ Dt_Customer        : chr [1:2240] "6/16/14" "6/15/14" "5/13/14" "5/11/14" ...
 $ Recency            : num [1:2240] 0 0 0 0 0 0 0 0 0 0 ...
 $ MntWines           : num [1:2240] 189 464 134 10 6 336 769 78 384 384 ...
 $ MntFruits          : num [1:2240] 104 5 11 0 16 130 80 0 0 0 ...
 $ MntMeatProducts    : num [1:2240] 379 64 59 1 24 411 252 11 102 102 ...
 $ MntFishProducts    : num [1:2240] 111 7 15 0 11 240 15 0 21 21 ...
 $ MntSweetProducts   : num [1:2240] 189 0 2 0 0 32 34 0 32 32 ...
 $ MntGoldProds       : num [1:2240] 218 37 30 0 34 43 65 7 5 5 ...
 $ NumDealsPurchases  : num [1:2240] 1 1 1 1 2 1 1 1 3 3 ...
 $ NumWebPurchases    : num [1:2240] 4 7 3 1 3 4 10 2 6 6 ...
 $ NumCatalogPurchases: num [1:2240] 4 3 2 0 1 7 10 1 2 2 ...
 $ NumStorePurchases  : num [1:2240] 6 7 5 2 2 5 7 3 9 9 ...
 $ NumWebVisitsMonth  : num [1:2240] 1 5 2 7 7 2 6 5 4 4 ...
 $ AcceptedCmp3       : num [1:2240] 0 0 0 0 1 0 1 0 0 0 ...
 $ AcceptedCmp4       : num [1:2240] 0 0 0 0 0 0 0 0 0 0 ...
 $ AcceptedCmp5       : num [1:2240] 0 0 0 0 0 0 0 0 0 0 ...
 $ AcceptedCmp1       : num [1:2240] 0 0 0 0 0 0 0 0 0 0 ...
 $ AcceptedCmp2       : num [1:2240] 0 1 0 0 0 0 0 0 0 0 ...
 $ Response           : num [1:2240] 1 1 0 0 1 1 1 0 0 0 ...
 $ Complain           : num [1:2240] 0 0 0 0 0 0 0 0 0 0 ...
 $ Country            : chr [1:2240] "SP" "CA" "US" "AUS" ...
```

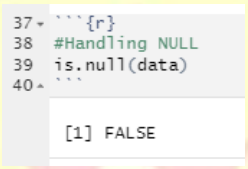
Gambar IV. 1. Hasil *Input* Data ke *RStudio*

4.2.3 Data Pre-Processing

Pada proses ini melakukan pembersihan data dari nilai-nilai yang bisa membuat pemodelan tidak bisa dilakukan, seperti: menghilangkan nilai *null*, menghilangkan nilai yang kosong, menyamakan tipe data agar bisa dilakukan pemodelan *K-Means Clustering*, dan juga menghilangkan *outliers.*

A. Pemeriksaan dan menghilangkan nilai *null*

Null merupakan suatu nilai yang tidak tersedia, tidak terdefinisikan, tidak diketahui dan tidak bisa diimplementasikan (RDocumentation, 1970). Fungsi yang akan dipakai *is.null(data)* dan berikut hasil pemeriksaan nilai NULL tersebut.



Gambar IV. 2. Hasil Pemeriksaan Nilai NULL

Setelah dilakukan pemeriksaan tidak ada nilai NULL pada data tersebut dengan dihasilkan *FALSE* yang berarti tidak terdapat nilai NULL, sehingga bisa dilanjutkan ke proses selanjutnya.

B. Menghilangkan nilai yang kosong

Nilai yang kosong atau biasa disebut *missing value* sering ditemukan pada *big data*, sehingga harus dilakukannya pemeriksaan untuk menghilangkan nilai yang kosong atau hilang tersebut, agar pemodelan dan analisis data akurat sesuai persebarannya. Fungsi yang akan dipakai ada beberapa seperti:

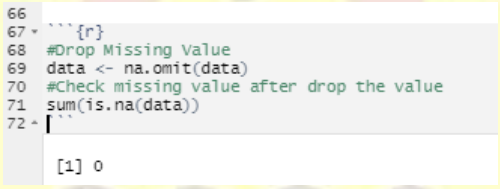
- *sum(is.na(data))* yang berfungsi untuk melihat total nilai yang kosong
- *ggplot()* yang berfungsi untuk membuat visualisasi data
- *na.omit(data)* yang berfungsi untuk menghilangkan semua nilai yang kosong

Secara jelas *pseudocode* untuk proses pemeriksaan dan menghilangkan nilai yang kosong seperti dibawah ini.

```
#Handling Missing Value
#Count Missing Value
sum(is.na(data))
#Visualization Missing Value
missing.values <- data %>%
    gather(key = "key", value = "val") %>%
    mutate(is.missing = is.na(val)) %>%
    group_by(key, is.missing) %>%
    summarise(num.missing = n()) %>%
    filter(is.missing==T) %>%
    select(-is.missing) %>%
    arrange(desc(num.missing))
#Visualization Total Missing Value
missing.values %>%
  ggplot() +
    geom_bar(aes(x=key, y=num.missing), stat = 'identity') +
    labs(x='variable', y="number of missing values", title='Total Missing Value') +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
#Drop Missing Value
data <- na.omit(data)
#Check missing value after drop the value
sum(is.na(data))
```

Gambar IV. 3. Hasil Pemeriksaan Nilai yang Kosong

Hasil yang didapat setelah pemeriksaaan nilai kosong atau hilang berjumlah 24 data yang hanya berada pada kolom *income.* Setelah itu dilakukan proses selanjutnya dengan sintaks dibawah ini yang berfungsi untuk menghilangkan nilai yang kosong tersebut. Lalu dilanjutkan dengan pemeriksaan kembali untuk memastikan tidak ada nilai kosong.



Gambar IV. 4. Menghilangkan Nilai yang Kosong dan Pemeriksaan Kembali

C. Menangani *Outliers*

Outliers adalah data yang muncul dengan nilai-nilai ekstrim atau nilai yang jauh atau nilai beda sama sekali dengan sebagian besar nilai lain dalam kelompoknya. Langkah yang akan dilakukan adalah memilih tipe data yang berbentuk *numerical* dikarenakan untuk tipe *character* tidak bisa dilakukan pemrosesan. Langkah selanjutnya membuat visualisasi untuk melihat *outliers* berdasarkan data yang sudah dipilih.

Terdapat 22 variabel yang akan dilakukan visualisasi menggunakan *box plot* yaitu variabel *AcceptedCmp1, AcceptedCmp2, AcceptedCmp3, AcceptedCmp4,*

AcceptedCmp5, Complain, Response, Year_birth, kidhome, teenhome, recency, MntWines, Mntfruits, MntMeatProducts, MntFishProducts, MntSweetProducts, MntGoldprods, NumDealsPurchases, NumWebPurchases, NumCatalogPurchases, NumStorePurchases, NumWebVisitsMonth. Fungsi yang akan dipakai ada beberapa macam dan akan digambarkan seperti *pseudocode* dibawah ini:

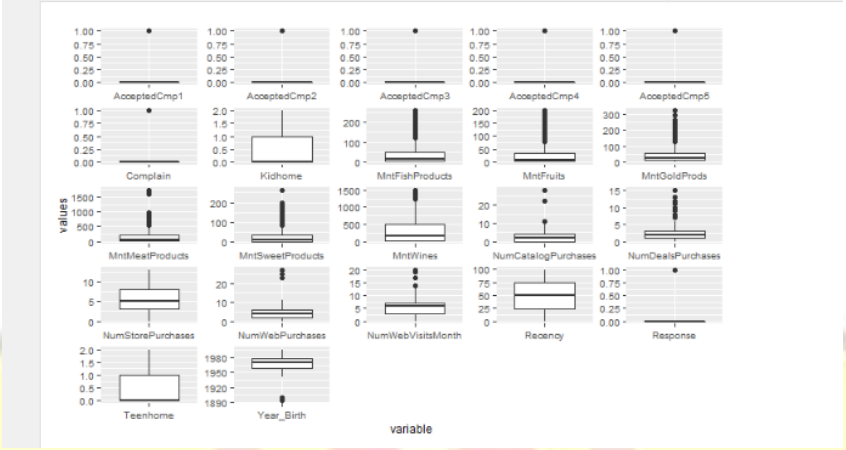
```
#Handling Outliers
#Assign Related Variable
unselect <- c("ID", "Income", "Education","Marital_Status", "Dt_Customer", "Country")
#Filter main dataframe
outliers <- data %>%
  select(-unselect)
outliers <- outliers %>%
  gather(variable,values,1:22)
#Plot Outliers unless Char Type
outliers %>%
  ggplot()+
  geom_boxplot(aes(x=variable,y=values)) +
  facet_wrap(~variable,ncol=,scales="free") +
  theme(strip.text.x = element_blank(),
```

```r
#Handling Outliers
unselect <- c("ID", "Income", "Education","Marital_Status", "Dt_Customer", "Country")

outliers <- data %>%
  select(-unselect)

outliers <- outliers %>%
  gather(variable,values,1:22)

#Plot Outliers unless Char
outliers %>%
  ggplot()+
  geom_boxplot(aes(x=variable,y=values)) +
  facet_wrap(~variable,ncol=,scales="free") +
  theme(strip.text.x = element_blank(),
        text = element_text(size=9))
```

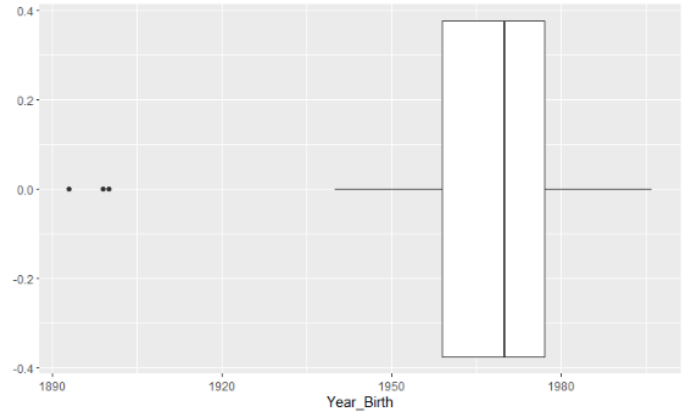


Gambar IV. 5. Hasil *Box Plot* dari Variabel yang Dipilih

Didapatkan hasil dari visualisasi menggunakan *box plot* pada beberapa variabel yang sudah dipilih, didapatkan terdapat 1 variabel yang terlihat memiliki *outliers* yaitu variabel *year_birth* yang merupakan variabel tahun lahir dari konsumen. *Pseudocode* berikut akan hanya akan memunculkan variabel *year_birth* untuk melihat secara rinci variabel tersebut:

```
#Outliers at year_birth is looks abnormal. Some data are year of birth 1890
#Plot outliers year_birth
data %>%
  ggplot(aes(x = Year_Birth)) +
  geom_boxplot()
#Use data with Year Birth > 1920
data <- data %>%
  filter(Year_Birth > 1920)
```

```r
#Outliers at year_birth is looks abnormal. Some data are year of birth 1890
#Plot outliers year_birth
data %>%
  ggplot(aes(x = Year_Birth)) +
  geom_boxplot()
```



Gambar IV. 6. Hasil *Box Plot* dari ariabel *Year_Birth*

Terlihat hasil dari *box plot* variabel *Year_birth* terdapat *outliers* berupa tahun lahir konsumen yang berada pada *range* tahun 1890 hingga tahun 1920, sehingga diperlukan *filterisasi* untuk data utama yang akan dipakai hanya menggunakan yang tahun lahir konsumennya melebihi tahun 1920 agar tidak ada *outliers*.

D. Menangani *Unique Value*

Unique value adalah nilai pada suatu variabel yang tidak sesuai dengan nilai lainnya, sehingga bisa dilakukan persamaan menjadi sesuai dengan nilai yang lainnya yang lebih relevan. Langkah pertama harus melihat *unique value* nya menggunakan fungsi *unique()* dan dimasukkan data utama tersebut dengan variabel yang sudah diketahui yaitu *marital_status*. Berikut adalah hasil yang didapat:

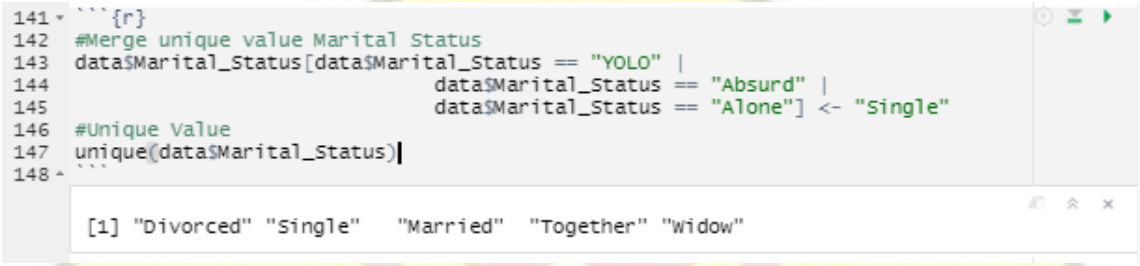
```r
#Handling Data Type
#Find Unique Value
unique(data$Marital_Status)
```

```
[1] "Divorced" "Single"   "Married"  "Together" "Widow"    "YOLO"     "Alone"
[8] "Absurd"
```

Gambar IV. 7. *Unique Value* yang Terdapat pada Variabel *marital_status*

Terdapat 8 nilai pada variabel *marital_status*, namun terdapat 3 variabel yang disebut *unique value* sehingga dirubah menjadi nilai yang lebih relevan, yaitu nilai *YOLO*, *absurd*, dan *alone.* Nilai tersebut akan diubah menjadi *single* karena lebih relevan dengan yang dimaksudkan, berikut *pseudocode* yang dipakai untuk proses ini:

```
#Merge unique value Marital Status
data$Marital_Status[data$Marital_Status == "YOLO" |
                    data$Marital_Status == "Absurd" |
                    data$Marital_Status == "Alone"] <- "Single"
#Unique Value
unique(data$Marital_Status)
```



Gambar IV. 8. Hasil Pengubahan *Unique Value*

Setelah dilakukan proses menyatukan 3 nilai tersebut didapatkan hasilnya akan menjadi 5 nilai saja yaitu *divorced, single, married, together,* dan *widow.*

E. Menangani Tipe Data

Data yang akan dipakai untuk pemodelan harus memiliki tipe data yang sesuai dengan isinya agar mudah untuk diidentifikasi dan dianalisis menggunakan *RStudio*. Ada beberapa variabel yang belum sesuai sehingga perlu dilakukan proses pengubahan tipe data, namun disini peneliti akan membagi 2 cara penanganan tipe data yang telah disesuaikan jenis variabel yang ada untuk memudahkan pengolahan data.

- Mengubah menjadi tipe data *factor*

Tipe data *factor* adalah tipe data berupa numerik atau karakter, namun secara proses pengolahan dalam *RStudio* akan terbaca sebagai angka (Rosidi, 2020). Contohnya adalah jika dalam satu variabel terdapat warna merah, biru, kuning yang merupakan tipe data karakter, namun secara pengolahan data *RStudio* akan bernilai 1, 2, 3 sehingga didapatkan untuk warna merah terbaca 1, biru terbaca 2, dan kuning terbaca 3 dan seterusnya yang akan disesuaikan dengan banyaknya jenis pada data tersebut.

Dalam data yang sudah dikumpulkan mengharuskan proses mengubah tipe data dari karakter menjadi *factor* dalam beberapa variabel yaitu *marital_status, education*

dan *country*. Dalam *RStudio* menggunakan fungsi *class(data$variable)* untuk memeriksa tipe data, dan berikut hasil dari pemeriksaan 3 variabel tersebut

```
149 ```{r}
150 #Handling Character Data into Factor
151 #Checking data type
152 class(data$Marital_Status)
153 class(data$Education)
154 class(data$Country)
155 ```
```

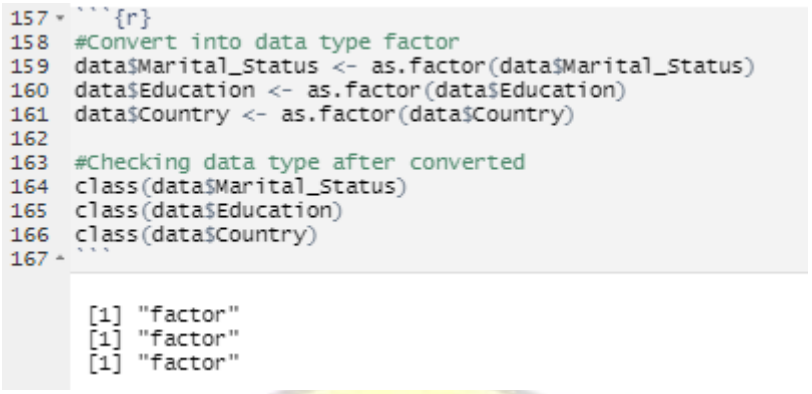
```
[1] "character"
[1] "character"
[1] "character"
```

Gambar IV. 9. Pemeriksaan Awal Tipe Data dari 3 Variabel

Didapatkan hasil dari ketiga variabel tersebut adalah karakter yang berarti perlu dilakukan pengubahan kedalam bentuk tipe data *factor* untuk dapat mudah melakukan pengolahan data dan visualisasi data. Secara mudah dapat menggunakan fungsi *as.factor(data$variable)* dan setelah itu data langsung terubah menjadi tipe data *factor*. Berikut adalah *pseudocode* yang digunakan dan hasil yang didapat setelah perubahan tipe data

```
#Convert into data type factor
data$Marital_Status <- as.factor(data$Marital_Status)
data$Education <- as.factor(data$Education)
data$Country <- as.factor(data$Country)


#Checking data type after converted
class(data$Marital_Status)
class(data$Education)
class(data$Country)
```



```r
#Convert into data type factor
data$Marital_Status <- as.factor(data$Marital_Status)
data$Education <- as.factor(data$Education)
data$Country <- as.factor(data$Country)

#Checking data type after converted
class(data$Marital_Status)
class(data$Education)
class(data$Country)
```

```
[1] "factor"
[1] "factor"
[1] "factor"
```

Gambar IV. 10. Mengubah dan Memeriksa Tipe Data Menjadi *Factor*

Hasil dari pengubahan tersebut untuk 3 variabel sudah menjadi *factor* secara algoritma dalam *RStudio.*

- Mengubah menjadi tipe data numerik

Tipe data numerik adalah data yang memiliki nilai segala jenis angka. Pada data yang sudah dikumpulkan, 22 variabel memiliki tipe data numerik. Namun terdapat 1 variabel yang seharusnya diubah, variabel tersebut yaitu *income* yang seharusnya secara arti penghasilan itu bernilai angka. Berikut adalah pemeriksaan tipe data awal menggunakan fungsi *str(data$variable)* untuk sekaligus melihat beberapa nilai diawal data,

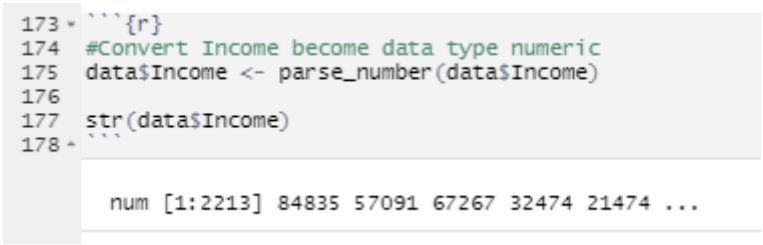
```r
#Checking data type
str(data$Income)
```

```
chr [1:2213] "$84,835.00" "$57,091.00" "$67,267.00" "$32,474.00" "$21,474.00" "$71,691.00" "$63,564.00" "$44,931.00" "$65,324.00" "$65,324.00" "$81,044.00" "$62,499.00" ...
```

Gambar 16. Pemeriksaan awal tipe data dari variabel *income*

Setelah itu dilakukan pengolahan data agar mengubah menjadi bentuk numerik menggunakan fungsi *parse_number(data$income).* Berikut adalah *pseudocode* dan hasil yang didapatkan setelah diubah menjadi tipe data numerik,

```
#Convert Income become data type numeric
data$Income <- parse_number(data$Income)
str(data$Income)
```

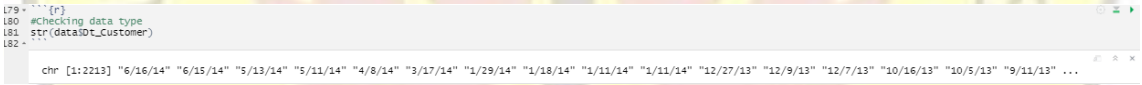


Gambar 17. Mengubah dan memeriksa tipe data menjadi *numeric*

Hasil setelah proses pengubahan tersebut menjadi tipe data numerik yang berarti semua huruf dan tanda baca menghilang, dan hanya menyisakan seluruh angka saja.

- Mengubah menjadi tipe data *date*

Tipe data *date* adalah tipe data yang memiliki nilai tanggal. Fungsi dari tipe data ini akan berfungsi untuk menyusun berdasarkan tanggal atau pemilihan tanggal yang diinginkan. Pada data yang sudah dikumpulkan variabel *dt_customer* memiliki tipe data awal karakter yang berarti sulit untuk mengurutkan atau memilih sesuai tanggal yang diinginkan. Berikut adalah pemeriksaan untuk variabel *dt_customer* tersebut yang merupakan variabel yang menjelaskan tanggal konsumen melakukan pendaftaran pada perusahaan tersebut.



Gambar 18. Pemeriksaan awal tipe data pada variabel *dt_customer*

Proses yang akan dilakukan yaitu mengubah menjadi tipe data *date* menggunakan fungsi *as.date(data$variable, "format")*. Pada fungsi tersebut peneliti bisa memilih format apa yang akan dipakai, seperti ingin menyusun dari tanggal, bulan, tahun atau malah sebaliknya. Peneliti memilih ingin menyusun bulan, tanggal dan tahun, maka berikut adalah *pseudocode* yang akan dipakai beserta hasil pemeriksaan kembali variabel *dt_customer* tersebut

#Convert become data type date

data$Dt_Customer <- as.Date(data$Dt_Customer, "%y/%m/%d")

#Checking data type

str(data$Dt_Customer)

```r
#Handling income data
data$Dt_Customer <- as.Date(data$Dt_Customer, "%y/%m/%d")

str(data$Dt_Customer)
```

```
 Date[1:2213], format: "2014-06-16" "2014-06-15" "2014-05-13" "2014-05-11" "2014-04-08" "2014-03-17" "2014-01-29" "2014-01-18" "2014-01-11" "2014-01-11" "2013-12-27" "2013-12-09" "2013-12-07" ...
```

Gambar 19. Mengubah dan memeriksa tipe data menjadi *date*

Hasilnya untuk seluruh variabel *dt_customer* sudah menggunakan format yang diingkan dan memiliki tipe data *date*. Hasil akhir dari proses *data pre-processing* adalah sebagai berikut ini,

```r
str(data)
```

```
tibble [2,213 x 28] (S3: tbl_df/tbl/data.frame)
 $ ID                 : num [1:2213] 1826 1 10476 1386 5371 ...
 $ Year_Birth         : num [1:2213] 1970 1961 1958 1967 1989 ...
 $ Education          : Factor w/ 5 levels "2n Cycle","Basic",..: 3 3 3 3 3 5 1 3 5 5 ...
 $ Marital_Status     : Factor w/ 5 levels "Divorced","Married",..: 1 3 2 4 3 3 2 4 2 2 ...
 $ Income             : num [1:2213] 84835 57091 67267 32474 21474 ...
 $ Kidhome            : num [1:2213] 0 0 0 1 1 0 0 0 0 0 ...
 $ Teenhome           : num [1:2213] 0 0 1 1 0 0 0 1 1 1 ...
 $ Dt_Customer        : Date[1:2213], format: "2014-06-16" "2014-06-15" "2014-05-13" "2014-05-11" ...
 $ Recency            : num [1:2213] 0 0 0 0 0 0 0 0 0 0 ...
 $ MntWines           : num [1:2213] 189 464 134 10 6 336 769 78 384 384 ...
 $ MntFruits          : num [1:2213] 104 5 11 0 16 130 80 0 0 0 ...
 $ MntMeatProducts    : num [1:2213] 379 64 59 1 24 411 252 11 102 102 ...
 $ MntFishProducts    : num [1:2213] 111 7 15 0 11 240 15 0 21 21 ...
 $ MntSweetProducts   : num [1:2213] 189 0 2 0 0 32 34 0 32 32 ...
 $ MntGoldProds       : num [1:2213] 218 37 30 0 34 43 65 7 5 5 ...
 $ NumDealsPurchases  : num [1:2213] 1 1 1 1 2 1 1 1 3 3 ...
 $ NumWebPurchases    : num [1:2213] 4 7 3 1 3 4 10 2 6 6 ...
 $ NumCatalogPurchases: num [1:2213] 4 3 2 0 1 7 10 1 2 2 ...
 $ NumStorePurchases  : num [1:2213] 6 7 5 2 2 5 7 3 9 9 ...
 $ NumWebVisitsMonth  : num [1:2213] 1 5 2 7 7 2 6 5 4 4 ...
 $ AcceptedCmp3       : num [1:2213] 0 0 0 0 1 0 1 0 0 0 ...
 $ AcceptedCmp4       : num [1:2213] 0 0 0 0 0 0 0 0 0 0 ...
 $ AcceptedCmp5       : num [1:2213] 0 0 0 0 0 0 0 0 0 0 ...
 $ AcceptedCmp1       : num [1:2213] 0 0 0 0 0 0 0 0 0 0 ...
 $ AcceptedCmp2       : num [1:2213] 0 1 0 0 0 0 0 0 0 0 ...
 $ Response           : num [1:2213] 1 1 0 0 1 1 1 0 0 0 ...
 $ Complain           : num [1:2213] 0 0 0 0 0 0 0 0 0 0 ...
 $ Country            : Factor w/ 8 levels "AUS","CA","GER",..: 7 2 8 1 7 7 3 7 8 4 ...
 - attr(*, "na.action")= 'omit' Named int [1:24] 135 263 395 450 526 591 900 998 1097 1186 ...
  ..- attr(*, "names")= chr [1:24] "135" "263" "395" "450" ...
```

Gambar 20. Hasil *Data Pre-Processing*

4.2.4 Assign Data

Assign data merupakan proses selanjutnya yang akan dilakukan yaitu menetapkan beberapa variabel pilihan menjadi 1 data frame yang akan disesuaikan dengan kebutuhan. Terdapat 4 data frame baru yang ditetapkan yaitu dengan nama *numeric_data, numeric_data_k2, numeric_data_k3* dan *education_data,* yang disesuaikan dengan tujuan penelitian agar memudahkan untuk melakukan pemodelan dan analisisnya.

- *Assign data* untuk data frame *numeric_data*

Pada data frame pertama ini akan dipilih hanya variabel bersangkutan sesuai tujuan awal yaitu variabel dengan jumlah pembelian produk per kategori dan metode pembelian. Pada *RStudio* yang akan dilakukan adalah menghilangkan variabel yang tidak dibutuhkan dan langsung menetapkannya kepada data frame yang baru. Proses ini akan dibantu dengan fungsi *select(-c(variables)* untuk menghilangkan variabel yang tidak dibutuhkan, berikut adalah *pseudocode* yang akan dipakai,

```
#Assign data

#Select relevant variable to assign

numeric_data <- data %>%

    select(-c("AcceptedCmp1", "AcceptedCmp2", "AcceptedCmp3",
"AcceptedCmp4", "AcceptedCmp5", "Recency", "Complain", "ID",
"Response",  "Year_Birth",  "Dt_Customer", "Country"))
```

```r
#Assign data
#Select relevant variable to assign
numeric_data <- data %>%
                select(-c("AcceptedCmp1", "AcceptedCmp2","AcceptedCmp3","AcceptedCmp4","AcceptedCmp5",
                          "Recency","Complain","ID", "Response", "Year_Birth",
                          "Dt_Customer", "Country"))

str(numeric_data)
```

```
tibble [2,213 x 16] (S3: tbl_df/tbl/data.frame)
 $ Education          : Factor w/ 5 levels "2n Cycle","Basic",..: 3 3 3 3 3 5 1 3 5 5 ...
 $ Marital_Status     : Factor w/ 5 levels "Divorced","Married",..: 1 3 2 4 3 3 2 4 2 2 ...
 $ Income             : num [1:2213] 84835 57091 67267 32474 21474 ...
 $ Kidhome            : num [1:2213] 0 0 0 1 1 0 0 0 0 0 ...
 $ Teenhome           : num [1:2213] 0 0 1 1 0 0 0 1 1 1 ...
 $ MntWines           : num [1:2213] 189 464 134 10 6 336 769 78 384 384 ...
 $ MntFruits          : num [1:2213] 104 5 11 0 16 130 80 0 0 0 ...
 $ MntMeatProducts    : num [1:2213] 379 64 59 1 24 411 252 11 102 102 ...
 $ MntFishProducts    : num [1:2213] 111 7 15 0 11 240 15 0 21 21 ...
 $ MntSweetProducts   : num [1:2213] 189 0 2 0 0 32 34 0 32 32 ...
 $ MntGoldProds       : num [1:2213] 218 37 30 0 34 43 65 7 5 5 ...
 $ NumDealsPurchases  : num [1:2213] 1 1 1 1 2 1 1 1 3 3 ...
 $ NumWebPurchases    : num [1:2213] 4 7 3 1 3 4 10 2 6 6 ...
 $ NumCatalogPurchases: num [1:2213] 4 3 2 0 1 7 10 1 2 2 ...
 $ NumStorePurchases  : num [1:2213] 6 7 5 2 2 5 7 3 9 9 ...
 $ NumWebVisitsMonth  : num [1:2213] 1 5 2 7 7 2 6 5 4 4 ...
 - attr(*, "na.action")= 'omit' Named int [1:24] 135 263 395 450 526 591 900 998 1097 1186 ...
  ..- attr(*, "names")= chr [1:24] "135" "263" "395" "450" ...
```

Gambar 21. Data frame *numeric_data*

Hanya tersisa 16 variabel yang berfungsi untuk memudahkan pemilihan variabel untuk dilakukan proses modelling dan tidak mengacaukan data frame utama yang sudah dilakukan *data pre-processing*.

- *Assign data* untuk data frame *numeric_data_k2*

Data frame selanjutnya adalah *numeric_data_k2* yang akan berisi variabel yang ingin dilakukan pemodelan metode *K-Means Clustering*. Data frame ini akan berisi variabel yang relevan dengan jumlah pembelian produk per kategori, seperti *pseudocode* dan hasilnya berikut ini,

#Assign by amount of items per categories

numeric_data_k2 <- numeric_data %>%

select(c(MntWines, MntFruits, MntMeatProducts, MntFishProducts, MntSweetProducts, MntGoldProds))

```r
#Assign by amount of items per categories
numeric_data_k2 <- numeric_data %>%
  select(c(MntWines, MntFruits, MntMeatProducts, MntFishProducts, MntSweetProducts, MntGoldProds))

str(numeric_data_k2)
```

```
tibble [2,213 x 6] (S3: tbl_df/tbl/data.frame)
 $ MntWines        : num [1:2213] 189 464 134 10 6 336 769 78 384 384 ...
 $ MntFruits       : num [1:2213] 104 5 11 0 16 130 80 0 0 0 ...
 $ MntMeatProducts : num [1:2213] 379 64 59 1 24 411 252 11 102 102 ...
 $ MntFishProducts : num [1:2213] 111 7 15 0 11 240 15 0 21 21 ...
 $ MntSweetProducts: num [1:2213] 189 0 2 0 0 32 34 0 32 32 ...
 $ MntGoldProds    : num [1:2213] 218 37 30 0 34 43 65 7 5 5 ...
 - attr(*, "na.action")= 'omit' Named int [1:24] 135 263 395 450 526 591 900 998 1097 1186 ...
  ..- attr(*, "names")= chr [1:24] "135" "263" "395" "450" ...
```

Gambar 22. Hasil data frame *numeric_data_k2*

Pada data frame ini tersisa 6 variabel yang berisi pembelian produk per kategori yang sesuai dengan tujuan pada awal penelitian ini. Variabel tersebut diambil dari data frame *numeric_data* untuk menghindari pengulangan pembuatan data frame dari awal jika pemodelan gagal dilakukan pada data frame tersebut.

- *Assign data* untuk data frame *numeric_data_k3*

Data frame ketida adalah *numeric_data3* yang berisi variabel yang relevan dengan tujuan kedua penelitian ini yaitu metode pembelian. Data frame ini juga akan dipakai untuk pemodelan metode *K-Means Clustering*, berikut *pseudocode* yang akan dilakukan beserta hasil dari *assign data* untuk *numeric_data_k3*

#Assign by purchase method

numeric_data_k3 <- numeric_data %>%

select(c(NumWebPurchases, NumCatalogPurchases,

```r
#Assign by method purchase
numeric_data_k3 <- numeric_data %>%
  select(c(NumWebPurchases, NumCatalogPurchases, NumStorePurchases))

str(numeric_data_k3)
```

```
tibble [2,213 x 3] (S3: tbl_df/tbl/data.frame)
 $ NumWebPurchases    : num [1:2213] 4 7 3 1 3 4 10 2 6 6 ...
 $ NumCatalogPurchases: num [1:2213] 4 3 2 0 1 7 10 1 2 2 ...
 $ NumStorePurchases  : num [1:2213] 6 7 5 2 2 5 7 3 9 9 ...
 - attr(*, "na.action")= 'omit' Named int [1:24] 135 263 395 450 526 591 900 998 1097 1186 ...
  ..- attr(*, "names")= chr [1:24] "135" "263" "395" "450" ...
```

Gambar 23. Hasil data frame *numeric_data_k3*

Hasil yang didapat pada data frame berikut ini hanya 3 variabel yang sudah disesuaikan dengan tujuan kedua penelitian ini. Data frame terbaru ini diambil dengan melakukan pemilihan variabel secara langsung dari data frame *numeric_data*.

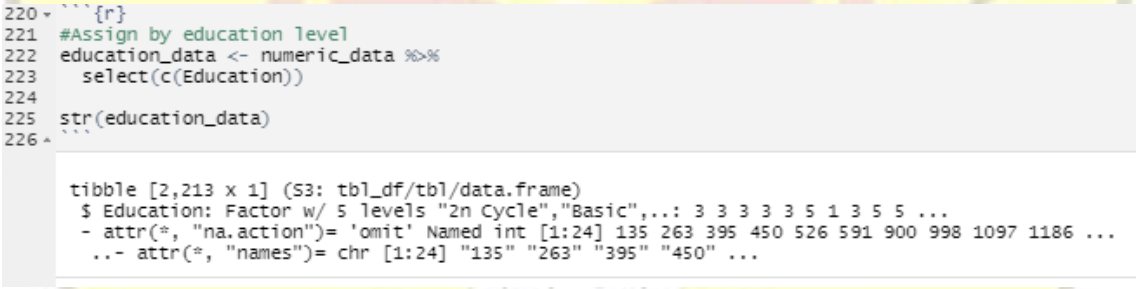
- *Assign data* untuk data frame *education_data*

Data frame terakhir yang akan ditetapkan adalah *education_data* yang akan berisi 1 variabel yang merupakan variabel satu-satunya yang relevan dengan tingkatan edukasi. Data frame ini tidak akan memakai metode *K-Means Clustering* karena data bersifat factor dan tidak memiliki nilai angka. Berikut adalah *pseudocode* dan hasil dari data frame *education_data* tersebut,

#Assign by education level

education_data <- numeric_data %>%

select(c(Education))



```
```{r}
#Assign by education level
education_data <- numeric_data %>%
  select(c(Education))

str(education_data)
```
```

```
tibble [2,213 x 1] (S3: tbl_df/tbl/data.frame)
 $ Education: Factor w/ 5 levels "2n Cycle","Basic",..: 3 3 3 3 3 5 1 3 5 5 ...
 - attr(*, "na.action")= 'omit' Named int [1:24] 135 263 395 450 526 591 900 998 1097 1186 ...
  ..- attr(*, "names")= chr [1:24] "135" "263" "395" "450" ...
```

Gambar 24. Hasil data *education_data*

Dipilih hanya 1 variabel *education* yang diambil langsung dari data frame *numeric_data* agar tidak mengganggu data frame utama yang telah dibuat diawal.

4.2.5 Scaling Data

Proses berikutnya adalah *scaling data* atau biasa disebut normalisasi data yang bertujuan mengurangi dampak *outliers* dan memungkinkan untuk membandingkan pengamatan tunggal terhadap setiap rata-rata (*mean*). Secara matematis dapat dijelaskan dengan mencari nilai maksimum dan minimum dari masing-masing variabel dengan contoh perhitungan seperti berikut ini,

Contoh peneliti akan mengambil nilai pada variabel *MntWines* dengan nilai maksimum (Xmaks)= 769, lalu nilai minimum (Xmin) = 2. Menghitung nilai normalisasi menggunakan persamaan:

$$Scale\ variable = \frac{(\text{Nilai awal} - \text{Nilai minimal})}{(\text{Nilai maksimal} - \text{Nilai minimal})}$$

$$MntWines = (J2 - Xmin) / (Xmaks - Xmin)$$

$$= (189 - 2) / (769 - 2)$$

$$= 0.243$$

Perhitungan tersebut akan digunakan pada setiap nilai pada data frame tersebut yang disesuaikan dengan setiap variabel/kolomnya. Hasil *scaling data* tersebut dapat dilihat pada lampiran 1.1. Pada *RStudio* fungsi yang akan dipakai adalah *scale(data)* yang berasal dari *library(dplyr)* dengan menggunakan data frame *numeric_data_k2* yang sebelumnya sudah ditetapkan. Berikut adalah *pseudocode* dan hasil dari peerhitungan dengan memperlihatkan 6 data pertama dari setiap variabel,

#Scaling the data (amount of items per categories)

numeric_data_k2 <- scale(numeric_data_k2)

head(numeric_data_k2)

```
229 #Scaling the data
230 numeric_data_k2 <- scale(numeric_data_k2)
231 head(numeric_data_k2)
232 ```
```

```
        MntWines  MntFruits MntMeatProducts MntFishProducts MntSweetProducts MntGoldProds
[1,] -0.34435739  1.9548052       0.9456412       1.3396690        3.9421565   3.36730023
[2,]  0.47092730 -0.5366426      -0.4591903      -0.5594139       -0.6580141  -0.13368407
[3,] -0.50741433 -0.3856458      -0.4814893      -0.4133306       -0.6093351  -0.26908125
[4,] -0.87503360 -0.6624733      -0.7401566      -0.6872367       -0.6580141  -0.84935489
[5,] -0.88689229 -0.2598151      -0.6375816      -0.4863722       -0.6580141  -0.19171143
[6,]  0.09144933  2.6091248       1.0883542       3.6952621        0.1208508  -0.01762934
```

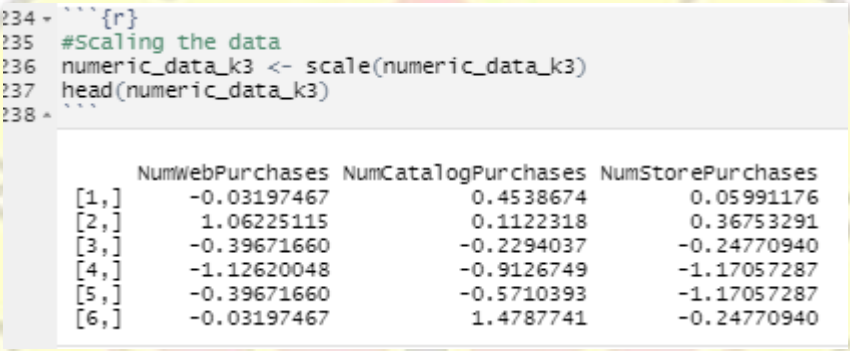
Gambar 25. Hasil data *Scaling data numeric_data_k2*

Dapat terlihat dari 6 variabel yang digunakan semua nilainya sudah disesuaikan dengan perhitungan, disini peneliti hanya memunculkan 6 data pertama pada *RStudio* karena hanya ingin memastikan data frame tersebut sudah ter-*scaling*. Langkah selanjutnya adalah melakukan *scaling data* pada data frame *numeric_data_k3*. Pada data frame selanjutnya terdapat 3 variabel yang akan dilakukan proses *scaling data*, berikut adalah *pseudocode* dan hasilnya,

#Scaling the data (purchase method)

numeric_data_k3 <- scale(numeric_data_k3)

head(numeric_data_k3)



```
234 ```{r}
235 #Scaling the data
236 numeric_data_k3 <- scale(numeric_data_k3)
237 head(numeric_data_k3)
238 ```
```

```
     NumWebPurchases NumCatalogPurchases NumStorePurchases
[1,]      -0.03197467           0.4538674        0.05991176
[2,]       1.06225115           0.1122318        0.36753291
[3,]      -0.39671660          -0.2294037       -0.24770940
[4,]      -1.12620048          -0.9126749       -1.17057287
[5,]      -0.39671660          -0.5710393       -1.17057287
[6,]      -0.03197467           1.4787741       -0.24770940
```

Gambar 26. Hasil data *Scaling data numeric_data_k3*

Nilai pada data frame tersebut dimunculkan dengan fungsi *head(data)* untuk melihat 6 data pertama untuk memastikan setiap nilai sudah ter-*scaling* dengan benar. Pada data frame yang ketiga yaitu *numeric_data_k3* tidak dilakukan *scaling data* dikarenakan tipe data tersebut faktor dan bernilai *character*, juga tidak akan dilakukan metode *K-Means Clustering* sehingga tidak dilakukan proses ini.

4.2.6 Mencari nilai *K* menggunakan *Elbow Method*

Elbow Method adalah metode yang digunakan untuk menginterpretasikan dan uji performa tingkat konsistensi jumlah cluster yang tepat dengan melihat nilai SSE *(Sum of Square Error).* Berikut adalah rumus SSE pada *K-Means* yang dipakai dan contoh perhitungannya,

Contoh peneliti akan mengambil nilai pada variabel *MntWines* dengan nilai awal (Xi)= 189, lalu nilai rata-rata pada variabel *MntWines* (X rata-rata) = 305,1536. Menghitung salah satu nilai dengan *Elbow Method* menggunakan persamaan:

$$\text{SSE} = \sum(\text{Xi} - \text{X rata- rata})^2$$

$$\textit{MntWines} = (189 - 305{,}1536)^2$$

$$\textit{MntWines} = 13.491{,}65879296$$

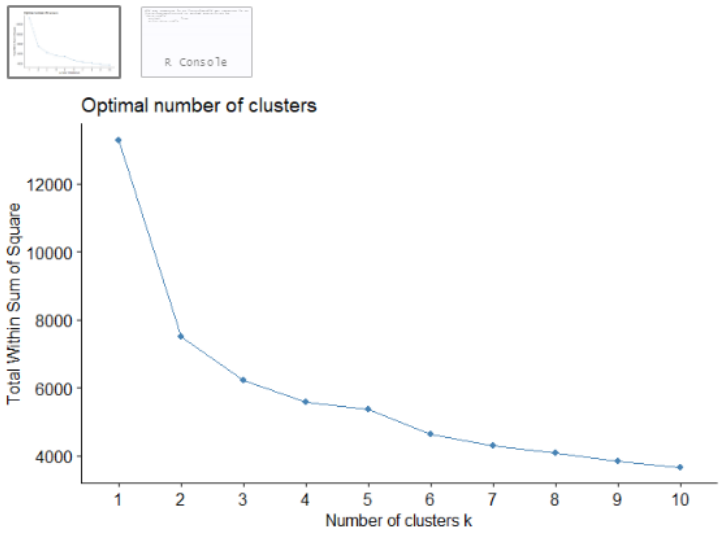
Selanjutnya persamaan tersebut akan dilakukan perhitungan pada semua nilai yang terdapat pada data frame. Nilai *K* yang optimal dapat terlihat setelah dilakukan proses ini dengan melihat penurunan garis pada grafik. Jika pada grafik terlihat mengalami penurunan secara drastis diikuti penurunan yang stabil, maka nilai tesebut menjadi acuan sebagai nilai *K*. Pada *RStudio* dapat menggunakan fungsi *fviz_nbclust(data, kmeans, method = 'wss')* yang terdapat pada *library(factoextra).* Ada 2 data frame yang akan dilakukan proses ini yaitu *numeric_data2* dan *numeric_data3* yang sebelumnya sudah ditetapkan

- *Elbow Method* untuk *numeric_data_k2*

Data frame *numeric_data2* akan dilakukan proses ini dengan sebelumnya sudah dilakukan proses *scaling*. Berikut adalah *pseudocode* yang akan dilakukan beserta dengan hasilnya,

```
# K- Means with Elbow Method (amount of items per categories)
fviz_nbclust(numeric_data_k2, kmeans, method = 'wss')
```

```r
# K- Means with Elbow Method
#Best K = 2
fviz_nbclust(numeric_data_k2, kmeans, method = 'wss')
```



Gambar 27. Hasil *elbow method* pada *numeric_data_k2*

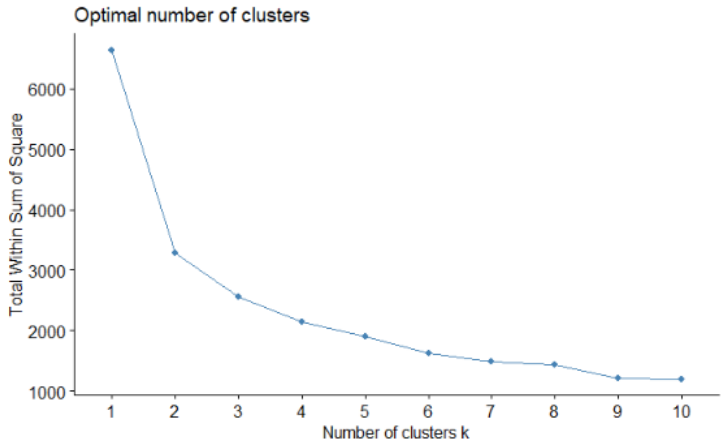
Hasil dari *Elbow Method* pada *numeric_data_k2* berupa grafik, dan cara menentukan nilai *K* yang terbaik atau optimal adalah dengan penurunan grafik yang drastis dan diikuti dengan penurunan yang stabil. Pada grafik tersebut didapatkan nilai *K* yang optimal adalah *K* = 2 karena terlihat grafik menurun secara drastis lalu diikuti dengan penurunan stabil pada nilai *K* setelahnya.

- *Elbow Method* untuk *numeric_data_k3*

Data frame selanjutnya adalah *numeric_data3* akan dilakukan proses ini dengan sebelumnya sudah dilakukan proses *scaling*. Berikut adalah *pseudocode* yang akan dilakukan beserta dengan hasilnya,

 # K- Means with Elbow Method (purchase method)
fviz_nbclust(numeric_data_k3, kmeans, method = 'wss')

```r
# K- Means with Elbow Method
fviz_nbclust(numeric_data_k3, kmeans, method = 'wss')
```



Gambar 28. Hasil *elbow method* pada *numeric_data_k3*

Hasil dari *elbow method* pada *numeric_data_k3* berupa grafik, dan cara menentukan nilai *K* yang terbaik atau optimal adalah dengan penurunan grafik yang drastis dan diikuti dengan penurunan yang stabil. Pada grafik tersebut didapatkan nilai *K* yang optimal adalah *K* = 2 karena terlihat grafik menurun secara drastis lalu diikuti dengan penurunan stabil pada nilai *K* setelahnya.

4.2.6 Mencari nilai *K* menggunakan *Silhouette Method*

Silhouette Method merupakan metode dengan menghitung rata-rata nilai setiap titik pada himpunan data. Perhitungan nilai setiap titik adalah selisih nilai *separation* dan *compactness* yang dibagi dengan maksimum antara keduanya. Jumlah klaster yang terbaik ditunjukkan dengan nilai *Silhouette* yang semakin mendekati 1. Tahapan perhitungan *Silhouette Method* sebagai berikut ini:

A. Hitung rata-rata jarak dari suatu objek misalkan i dengan semua objek lain yang masih berada dalam satu cluster

$$a(i) = \frac{1}{|A|-1} \sum j \in_{a,j=1} d(I,j)$$

dengan j adalah objek lain yang berada dalam satu cluster A dan d(I,j) adalah jarak antara objek i dengan j

B. Hitung rata-rata jarak dari objek i tersebut dengan semua objek yang berada di cluster lain, dan diambil nilai paling minimumnya

$d(i, C) = \frac{1}{|A|} \sum j \in C\ d(I,j)$

d(i,C) adalah jarak rata-rata objek i dengan semua objek pada cluster lain C dimana A ≠ C

d(i, C) = min C ≠ Ad(i, j)

C. Nilai *Silhouette Coefficient* nya adalah :

$s(i) = \frac{b(i)-a(i)}{\max\ (a(i).b(i))}$

Nilai yang didapat dari metode silhouette terletak pada kisaran nilai -1 hingga 1. Jika nilai *silhouette* mendekati nilai 1, maka semakin baik pengelompokkan objeknya dalam 1 kluster dan sebaliknya jika mendekati nilai -1, maka semakin buruk pengelompokkan objeknya dalam 1 kluster tersebut. Berikut adalah contoh perhitungannya:

Tabel IV. 3. Contoh Perhitungan *Silhouette Method* Menggunakan Data Frame *numeric_data_k3*

Data	NumWeb	NumCatalog	NumStore	Kluster
1	-0,03	0,45	0,05	1
2	1,06	0,11	0,36	1
3	-0,39	-0,22	-0,24	2
4	-1,12	-0,91	-1,17	2

- Menghitung nilai a_i masing-masing data menggunakan rumus A

A. Data 1 cluster 1

d(data 1, data 2) :

$\sqrt{(-0,03 - 1,06)^2 + (0,45 - 0,11)^2 + (0,05 - 0,36)^2}$ = 1,183131438

B. Data 2 cluster 1

d(data2, data 1) :

$\sqrt{(1,06 - (-0.03))^2 + (0,11 - 0,45)^2 + (0,36 - 0,05)^2}$ = 1,183131438

C. Data 3 cluster 2

d(data 3, data 4):

$\sqrt{(-0{,}39-1{,}12)^2+(-0{,}22-(-0{,}91))^2+(-0{,}24-(-1{,}17))^2}$ = 1,368904672

D. Data 4 cluster 2

d(data 4, data 3):

$\sqrt{(1{,}12-(-0{,}39))^2+(-0{,}91-(-0{,}22))^2+(-1{,}17-(-0{,}24))^2}$ = 1,368904672

- Menghitung nilai b_1 masing-masing data menggunakan rumus B

A. Data 1 cluster 1

d(data 1, data 3):

$\sqrt{(-0{,}03-(-0{,}39))^2+(0{,}45-(-0{,}22))^2+(0{,}05-(-0{,}24))^2}$ = 0,814002457

d(data 1, data 4):

$\sqrt{(-0{,}03-(-1{,}12))^2+(0{,}45-(-0{,}91))^2+(0{,}05-(-1{,}17))^2}$ = 2,127463278

Nilai b_1 = (2,127463278 - 0,814002457)/2 = 0,656730411

B. Data 2 cluster 1

d(data 2 , data 3):

$\sqrt{(1{,}06-(-0{,}39))^2+(0{,}11-(-0{,}22))^2+(0{,}36-(-0{,}24))^2}$ = 1,603558543

d(data 2 , data 4):

$\sqrt{(1{,}06-(-1{,}12))^2+(0{,}11-(-0{,}91))^2+(0{,}36-(-1{,}17))^2}$ = 2,851964235

Nilai b_1 = (2,851964235 - 1,603558543)/2 = 0,624203

C. Data 3 cluster 2

d(data 3 , data 1):

$\sqrt{(-0{,}39-(-0{,}03))^2+(-0{,}22-0{,}45)^2+(-0{,}24-0{,}05)^2}$ = 0,814002457

d(data 3 , data 2):

$\sqrt{(-0{,}39-1{,}06)^2+(-0{,}22-0{,}11)^2+(-0{,}24-0{,}36)^2}$ = 1,603558543

Nilai b_1 = (1,603558543 - 0,814002457)/2 = 0,394778

D. Data 4 cluster 2

d(data 4 , data 1):

$\sqrt{(1{,}12-(-0{,}03))^2+(-0{,}91-0{,}45)^2+(-1{,}17-0.05)^2}$ = 2,127463278

d(data 4 , data 2):

$\sqrt{(1{,}12-1{,}06)^2+(-0{,}91-0{,}11)^2+(-1{,}17-0.36)^2}$ = 2,851964235

Nilai b_1 = (2,851964235 - 2,127463278)/2 = 0,36225

- Menghitung nilai SI sesuai dengan rumus C

A. SI data 1 = (0,656730411 - 1,183131438)/ 0,656730411 = -0,80155
B. SI data 2 = (0,624203 - 1,183131438)/ 0,624203 = -0,89543
C. SI data 3 = (0,394778 - 1,368904672)/ 0,394778 = -2,46753
D. SI data 4 = (0,36225 - 1,368904672)/ 0,36225 = -2,77889

- Menghitung nilai SI setiap *cluster*

A. SI *cluster* 1 = (-0,80155 + (-0,89543)) = -1,69698
B. SI *cluster* 2 = (-2,46753 + (-2,77889)) = -5,24642

- Menghitung nilai Si global

SI global = (-1,69698 + (-5,24642)) = -6,9434

Hasil dari contoh perhitungan diatas didapatkan nilai SI -6,9434, maka dikatakan buruk pengelompokkan objeknya dalam 1 kluster. Namun tidak bisa dikatakan benar dikarenakan hanya mengambil 4 data saja per variabel dari 2.240, maka dari itu digunakan pengolahan menggunakan *RStudio* dengan menggunakan fungsi *fviz_nbclust(data, kmeans, method = 'silhouette')* pada data data frame yang sudah ditentukan. Terdapat 2 data frame yang akan digunakan seperti berikut ini:

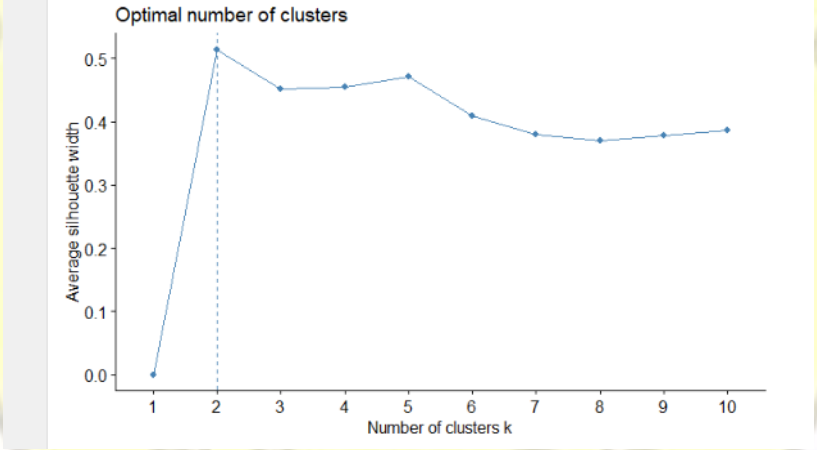
- *Silhouette Method* untuk *numeric_data_k2*

Data frame *numeric_data_k2* akan dilakukan proses dengan *silhouette method* dengan sebelumnya sudah dilakukan proses *scaling* sama seperti menggunakan *elbow method*. Berikut adalah *pseudocode* yang akan dilakukan beserta dengan hasilnya,

#K - Means with Silhoutte Method (amount of items per categories)

fviz_nbclust(numeric_data_k2, kmeans, method = 'silhouette')

```
253 ```{r}
254 #K - Means with Silhoutte Method
255 fviz_nbclust(numeric_data_k2, kmeans, method = 'silhouette')
256 ```
```



Gambar 29. Hasil *silhouette method* pada *numeric_data_k2*

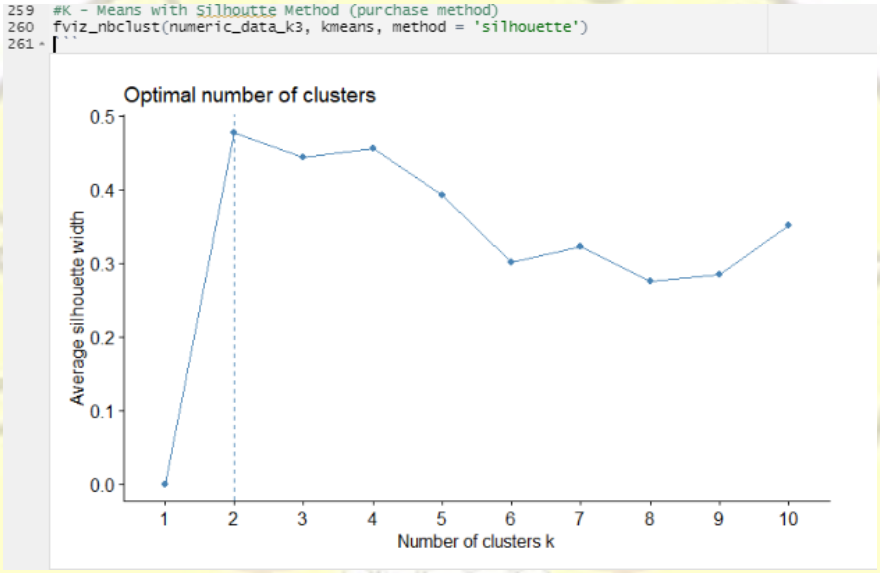
Didapatkan hasilnya berupa grafik dan cara menentukan *K* yang optimal adalah yang mengalami grafik meningkat secara drastis dan diikuti penurunan pada nilai selanjutnya. Terlihat pada hasil tersebut terdapat nilai *K* yang optimal berada di *K* = 2.

- *Silhouette Method* untuk *numeric_data_k3*

Data frame selanjutnya adalah *numeric_data_k3* akan dilakukan proses ini dengan sebelumnya sudah dilakukan proses *scaling* dengan yang sama dilakukan untuk proses *Elbow Method*. Berikut adalah *pseudocode* yang akan dilakukan beserta dengan hasilnya,

#K - Means with Silhoutte Method (purchase method)

fviz_nbclust(numeric_data_k3, kmeans, method = 'silhouette')



Gambar 30. Hasil *silhouette method* pada *numeric_data_k3*

Didapatkan hasilnya berupa grafik *line plot* dan cara menentukan *K* yang optimal adalah yang mengalami grafik meningkat secara drastis dan diikuti penurunan pada nilai selanjutnya. Terlihat hasil tersebut terdapat nilai *K* yang optimal berada pada *K* = 2.

4.2.7 Penentuan nilai *K* yang akan digunakan

Setelah dilakukan 2 metode yaitu *Elbow Method* dan *Silhouette Method*, proses selanjutnya adalah menentukan nilai *K* yang harus digunakan oleh setiap data frame dalam metode *K-Means Clustering* agar bisa diketahui nilai *K* yang optimalnya seperti apa. Ada 2 data frame yang akan peneliti bagi dalam penentuan nilai *K* tersebut berdasarkan variabel awal yang sudah ditentukan yaitu berdasarkan jumlah pembelian produk per kategori dan metode pembelian. Namun untuk variabel tingkatan edukasi tidak bisa dilakukan metode *K-Means Clustering* dikarenakan tipe data yang bersifat karakter.

- Nilai *K* untuk data frame *numeric_data_k2*

Dari hasil grafik dalam proses mencari nilai *K* dengan menggunakan *Elbow Method* didapatkan nilai *K* yang optimal adalah *K = 2.* Setelah itu dilanjutkan dengan menggunakan *Silhouette Method* untuk memastikan nilai *K* yang didapat sama sehingga bisa dilanjutkan ke pemodelan *K-Means Clustering*, dan didapatkan nilai grafik yang optimal berada pada *K = 2.* Jadi untuk nilai *K* yang akan digunakan pada data frame *numeric_data2* yang berisi variabel jumlah pembelian produk per kategori yaitu *K = 2.*

Dilanjutkan dengan proses selanjutnya yaitu proses pemodelan *K-Means Clustering* berdasarkan hasil pencarian nilai *K*. Proses ini menggunakan *RStudio* dengan fungsi *kmeans(data, centers = K value, nstart = best setting configure).* Berikut adalah *pseudocode* dalam proses ini,

```
#K-Means using K = 2 (amount of items per categories)
k_amount_2 = kmeans(numeric_data_k2, centers = 2, nstart = 50)
```

- Nilai *K* untuk data frame *numeric_data_k3*

Dari hasil grafik dalam proses mencari nilai *K* dengan menggunakan *Elbow Method* didapatkan nilai *K* yang optimal adalah *K = 2.* Setelah itu dilanjutkan dengan menggunakan *Silhouette Method* untuk memastikan nilai *K* yang didapat sama sehingga bisa dilanjutkan ke pemodelan *K-Means Clustering*, dan didapatkan nilai grafik yang optimal berada pada *K = 2.* Jadi untuk nilai *K* yang akan digunakan pada data frame *numeric_data2* yang berisi variabel jumlah pembelian produk per kategori yaitu *K = 2.*

Dilanjutkan dengan proses selanjutnya yaitu proses pemodelan *K-Means Clustering* berdasarkan hasil pencarian nilai *K*. Proses ini menggunakan *RStudio* dengan fungsi *kmeans(data, centers = K value, nstart = best setting configure).* Berikut adalah *pseudocode* dalam proses ini,

```
#K-Means using K = 2 (purchase method)
k_amount_3 = kmeans(numeric_data_k3, centers = 2, nstart = 50)
```

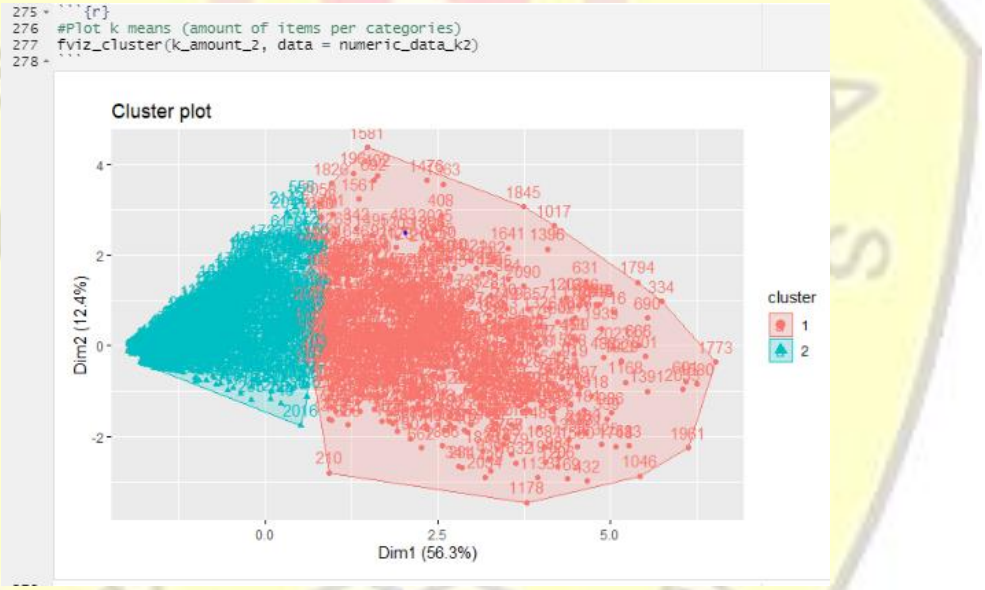
4.2.8 Visualisasi Hasil Pengolahan Data

Pada proses ini akan ditampilkan hasil pengolahan *K-Means Clustering* dan pengolahan data lainnya sehingga bisa dilihat secara lebih jelas gambaran klustering dari beberapa variabel yang sudah ditentukan. Terdapat 3 visualisasi yang akan dimunculkan seperti dibawah ini:

- *K- Means Clustering* pada *numeric_data_k2*

Memunculkan hasil dari *K-Means Clustering* menggunakan *RStudio* dengan menggunakan fungsi *fviz_cluster(data K-means, data = main data).* Secara *pseudocode* akan seperti dibawah ini beserta dengan hasil dari grafik tersebut,

```
#Plot k means (amount of items per categories)
fviz_cluster(k_amount_2, data = numeric_data_k2)
```



Gambar 31. Hasil *K-Means Clustering* pada *numeric_data_k2*

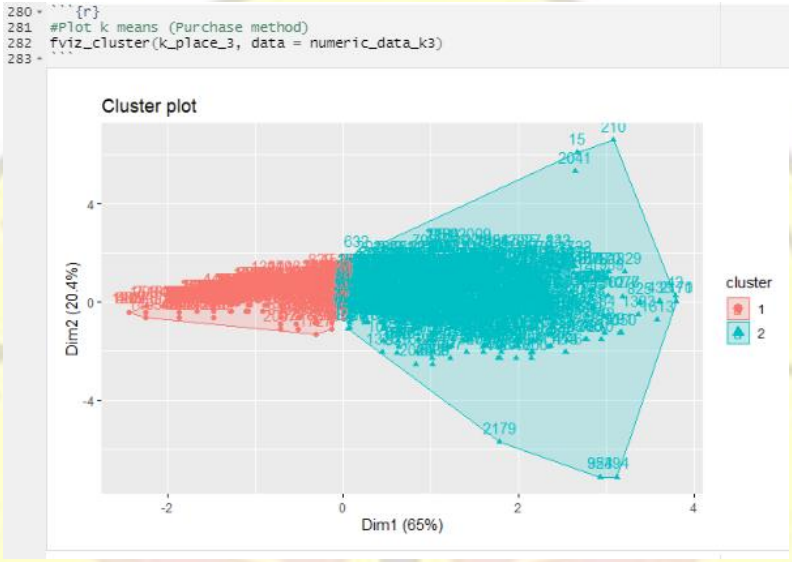
Didapatkan hasil secara *cluster plot* seperti gambar diatas dengan menggunakan *K = 2,* yang didapat dari penentuan 2 metode sebelumnya. Digambarkan untuk kluster 1 menggunakan warna merah dan kluster 2 menggunakan warna biru. Pada hasil tersebut dilihat persebaran untuk kluster 1 lebih besar dan meluas dibandingkan dengan kluster 2.

- *K- Means Clustering* pada *numeric_data_k3*

Memunculkan hasil dari *K-Means Clustering* selanjutnya pada data frame *numeric_data_k3* yang berisi variabel metode pembelian. Berikut adalah *pseudocode* dan hasil dari visualisasi datanya,

#Plot k means (Purchase method)

fviz_cluster(k_place_3, data = numeric_data_k3)



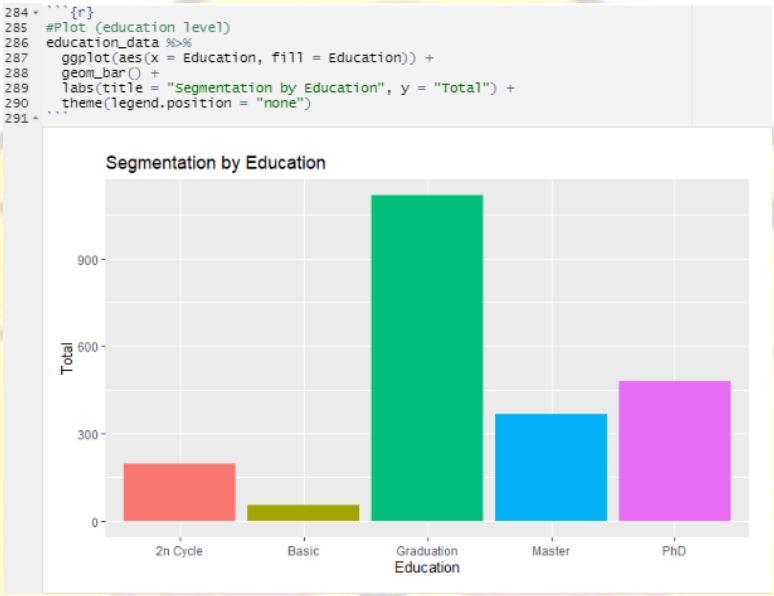
Gambar 32. Hasil *K-Means Clustering* pada *numeric_data_k3*

Didapatkan hasil secara *cluster plot* seperti gambar diatas dengan menggunakan *K = 2,* yang didapat dari penentuan 2 metode sebelumnya. Digambarkan untuk kluster 1 menggunakan warna merah dan kluster 2 menggunakan warna biru. Pada hasil tersebut dilihat persebaran untuk kluster 1 lebih sempit jika dibandingkan dengan kluster 2.

- *Plotting* pada *education_data*

Pada data frame berikut ini akan memunculkan grafik *bar plot* atau grafik yang berbentuk bar berisi dari total dari setiap gelar atau tingkatan edukasi pada data yang digunakan. Pada *RStudio* dapat menggunakan *library ggplot2* yang berisi beragam visualisasi yang bisa ditampilkan. Fungsi yang akan dipakai adalah *ggplot()* dengan mengisi dengan data, bentuk *plot*, warna dan lain – lain. Berikut adalah *pseudocode* beserta hasilnya,

```
#Plot (education level)
education_data %>%
  ggplot(aes(x = Education, fill = Education)) +
  geom_bar() +
  labs(title = "Segmentation by Education", y = "Total") +
  theme(legend.position = "none")
```



Gambar 33. Hasil *plotting* pada *education_data*

Hasil dari visualisasi data diatas ini menggambarkan seluruh tingkatan edukasi yang terdapat pada data tersebut dan juga perbandingan antara setiap *level* edukasinya. Didapatkan untuk hasil tertinggi pada *graduation* atau yang sudah lulus, dan untuk tingkat terendah pada *basic* atau jika di Indonesia digambarkan dengan SD dan SMP (bersekolah selama 9 tahun).

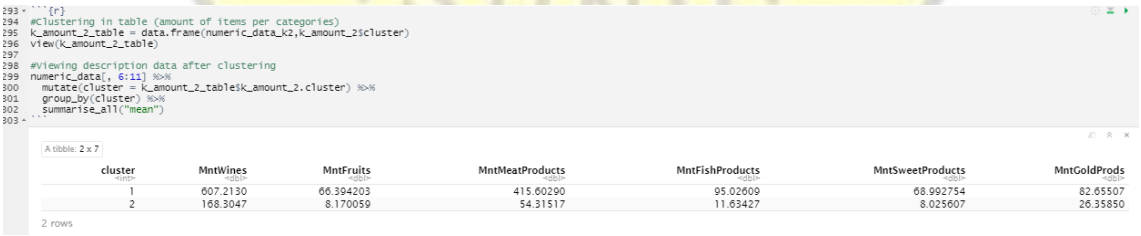
4.2.9 Tabel Hasil Pengolahan Data

Secara harfiah *K-Means Clustering* memang hanya menggambarkannya melalui penentuan nilai *K* yang sesuai untuk dipakai dalam suatu data, namun dari hasil pengolahan data yang besar ini peneliti akan memunculkan tabel untuk setiap nilai variabel untuk hasil *K-Mean Clustering* yang sudah divisualisasikan sebelumnya. Terdapat 2 tabel yang akan dimunculkan terutama hasil dari *K-Means Clustering* pada data frame *numeric_data_k2* dan *numeric_data_k3.*

- Tabel Hasil *K-Means Clustering* pada *numeric_data_k2*

Hasil berikut ini dipakai untuk melihat setiap nilai yang pada setiap kluster agar terlihat dalam kluster tersebut termasuk kemanakah setiap variabel tersebut. Sebelumnya hasil data *k-means* yang sudah dibuat harus dirubah menjadi data frame menggunakan fungsi *data.frame(data, data k-means).* Setelah itu menggunakan fungsi *mutate()* dan *group_by()* untuk membagi nilai ke setiap kluster per variabel bersangkutan. Berikut adalah *pseudocode* dan hasil berupa tabelnya,

```
#Clustering in table (amount of items per categories)
k_amount_2_table = data.frame(numeric_data_k2,k_amount_2$cluster)
view(k_amount_2_table)
#Viewing description data after clustering
numeric_data[, 6:11] %>%
  mutate(cluster = k_amount_2_table$k_amount_2.cluster) %>%
  group_by(cluster) %>%
  summarise_all("mean")
```



A tibble: 2 x 7

cluster	MntWines	MntFruits	MntMeatProducts	MntFishProducts	MntSweetProducts	MntGoldProds
1	607.2130	66.394203	415.60290	95.02609	68.992754	82.65507
2	168.3047	8.170059	54.31517	11.63427	8.025607	26.35850

2 rows

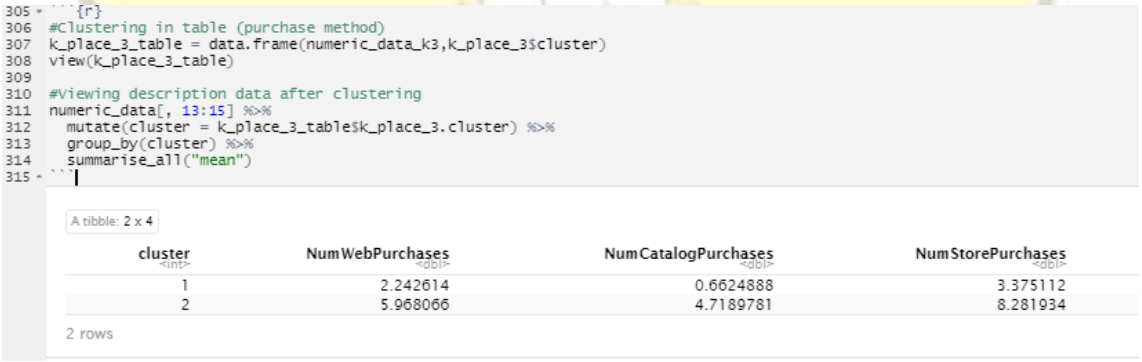
Gambar 34. Tabel Hasil *K-Means Clustering* pada *numeric_data_k2*

Hasil dari setiap nilai untuk kluster 1 lebih besar dibandingkan dengan setiap nilai kluster 2 dengan perbandingan rata-rata 6:1. Hasil nilai pada tabel tersebut menggambarkan dari nilai pada *cluster plot* pada proses sebelumnya.

- Tabel Hasil *K-Means Clustering* pada *numeric_data_k3*

Proses pada data frame juga dilakukan sama seperti sebelumnya yang bertujuan untuk melihat dalam berbentuk tabel kluster termasuk kemanakah setiap variabel tersebut. Sebelumnya hasil data *k-means* yang sudah dibuat harus dirubah menjadi data frame menggunakan fungsi *data.frame(data, data k-means).* Setelah itu menggunakan fungsi *mutate()* dan *group_by()* untuk membagi nilai ke setiap kluster per variabel bersangkutan. Berikut adalah *pseudocode* dan hasil berupa tabelnya,

```
#Clustering in table (purchase method)
k_place_3_table = data.frame(numeric_data_k3,k_place_3$cluster)
view(k_place_3_table)
#Viewing description data after clustering
numeric_data[, 13:15] %>%
  mutate(cluster = k_place_3_table$k_place_3.cluster) %>%
  group_by(cluster) %>%
  summarise_all("mean")
```



```
```{r}
#Clustering in table (purchase method)
k_place_3_table = data.frame(numeric_data_k3,k_place_3$cluster)
view(k_place_3_table)

#Viewing description data after clustering
numeric_data[, 13:15] %>%
  mutate(cluster = k_place_3_table$k_place_3.cluster) %>%
  group_by(cluster) %>%
  summarise_all("mean")
```
```

A tibble: 2 x 4

cluster <int>	NumWebPurchases <dbl>	NumCatalogPurchases <dbl>	NumStorePurchases <dbl>
1	2.242614	0.6624888	3.375112
2	5.968066	4.7189781	8.281934

2 rows

Gambar 35. Tabel Hasil *K-Means Clustering* pada *numeric_data_k3*

Hasil dari setiap nilai untuk kluster 1 lebih kecil jika dibandingkan dengan setiap nilai kluster 2 dengan perbandingan rata-rata 1:3. Hasil nilai pada tabel tersebut menggambarkan dari nilai pada *cluster plot* pada proses sebelumnya.

Bab V
Analisis dan Pembahasan

5.1 Analisis Urgensi Penentuan Segmentasi Konsumen

Penentuan segmentasi konsumen termasuk dalam tipe strategi pemasaran yang disebut STP atau *Segmenting, Targeting, Positioning.* Penentuan segmentasi konsumen ini adalah langkah pertama untuk perusahaan menentukan jenis atau kelompok dalam pasar atau *market* yang akan dituju dari perusahaan *IFood.* Penentuan segmentasi ini berperan dalam pengelompokkan konsumen yang disesuaikan dengan beberapa hal seperti demografis, geografis, tingkah laku dan siklus hidup konsumen.

Dengan adanya penentuan segmentasi konsumen secara benar akan memberikan kemudahan dan keuntungan perusahaan *IFood* dalam membuat *campaign* atau pemasaran pada fase tersebut atau selanjutnya. Selain itu dengan adanya penentuan segmentasi konsumen ini perusahaan *IFood* bisa lebih fokus dalam mengalokasikan sumber daya berdasarkan segmen-segmen yang akan dilayani atau dijadikan target.

Disisi lain segmentasi merupakan faktor utama untuk mengalahkan *competitor* atau saingan dalam bidang bisnis yang sama dengan melihat dari sudut pandang yang berbeda dari perusahaan saingan tersebut.

5.2 Analisis Variabel yang Dipilih Terhadap Jenis Segmentasi

Variabel yang sudah dipilih dan menjadi tujuan tugas akhir ini ada 3 variabel yaitu jumlah pembelian produk per kategori, metode pembelian dan tingkatan edukasi konsumen. Secara teori yang sudah dijelaskan terdapat 4 jenis segmentasi yaitu segmentasi berdasarka demografis, geografis, perilaku dan siklus hidup konsumen. Jika dikaitkan terhadap jenis segmentasi tersebut maka 3 variabel tersebut bisa masuk kedalam beberapa jenis segmentasi seperti dibawah ini:

- Jumlah pembelian produk per kategori

Pada variabel yang pertama ini untuk variabel jumlah pembelian produk per kategori berada pada segmentasi berdasarkan perilaku konsumsi, karena variabel ini berisi jumlah pembelian setiap produk dari setiap konsumen yang sudah membeli pada

perusahaan *IFood* tersebut. Segmentasi berdasarkan perilaku konsumsi adalah segmentasi berdasarkan perilaku konsumen berinteraksi dengan produk yang ditawarkan pada perusahaan *IFood*. Jadi bisa dikatakan konsumen berinteraksi dengan setiap produk yang ditawarkan perusahaan dengan cara membeli sejumlah yang mereka inginkan.

- Metode pembelian

Variabel yang kedua ini variabel metode pembelian yang berada pada segmentasi berdasarkan perilaku konsumsi, karena variabel ini juga berisi jumlah berdassarkan metode pembelian dari setiap konsumen yang tersedia pada perusahaan *IFood* tersebut. Berdasarkan segmentasi tersebut perilaku konsumsi yang berkaitan yaitu konsumen melakukan pembelian yang menggunakan beberapa metode yang sudah disediakan perusahaan, seperti metode secara *offline* dengan datang ke toko, *online* melalui website atau melalui aplikasi yang sudah disediakan.

- Tingkatan edukasi konsumen

Variabel ini merupakan segmentasi berdasarkan demografis, karena secara teori segmentasi demografis merupakan mengelompokkan konsumen menurut data demografis atau kependudukan seperti umur, jenis kelamin, profesi dan lain-lain, juga termasuk didalamnya berdasarkan tingkatan edukasi konsumen. Tingkatan edukasi konsumen merupakan *level* atau tingkatan pendidikan terakhir dari setiap konsumen.

5.3 Analisis Menggunakan Jenis *Unsupervised Machine Learning*

Dalam *unsupervised machine learning* terdapat banyak metode termasuk didalamnya *K-Means Clustering*, namun semua memilki fungsi yang sama yaitu untuk menarik kesimpulan dari dataset dengan algoritma yang mempelajari suatu data berdasarkan kedekatannya. Disimpulkan jenis *machine learning* ini memodelkan struktur yang mendeskripsikan data tersebut.

Unsupervised machine learning yang paling sering digunakan dalam bidang bisnis untuk menentukan segmentasi konsumen terutama, selain karena bisa mencari algoritma secara mandiri dengan memberikan pola yang tidak dikenal dalam data. Jenis *machine learning* ini pula bisa mendapatkan hasil yang tidak biasa dikarenakan bisa

menggunakan berbagai variabel asalkan berbentuk *nilai* yang bisa dilakukan pemrosesan. Jenis *machine learning* ini optimal dilakukan dalam penentuan segmentasi terutama jika sudah memiliki data dari beberapa variabel sebelumnya. Jenis *machine learning* ini juga bisa melakukan pemrosesan banyak data secara cepat. Jadi untuk jumlah data 2.240 konsumen dengan 28 variabel yang terdapat pada perusahaan *IFood* bisa optimal untuk digunakan.

5.4 Analisis Pemilihan Metode *K-Means Clustering* Sebagai Usulan

Pemilihan metode *K-Means Clustering* yang termasuk dalam *unsupervised machine learning* pada penentuan segmentasi konsumen di perusahaan *IFood* dapat dinilai bisa berjalan optimal karena metode ini dapat melihat dari sisi lain dari sebuah variabel yang memiliki banyak nilai, dimana pada variabel jumlah pembelian produk per kategori dan metode pembelian setiap konsumen memiliki nilai atau jumlah yang bervariasi mulai dari nilai yang terkecil (0) hingga terbesar (>10). Inilah yang menjadi keunggulan metode *K-Means Clustering* karena dengan data yang bervariasi dan jumlahnya banyak metode ini bisa mencari *insight* atau kesempatan lain dari variabel tersebut.

Selain itu metode *K-Means Clustering* sangat mudah dijalankan dan dipahami karena bentuk yang disajikan secara umum berupa *cluster plot* atau grafik cluster sehingga cocok untuk digunakan dalam bidang bisnis atau persentasi antara divisi. Hasil dari setiap klusternya pun sangat meluas sehingga jika didukung dengan analisis deskriptif lain akan sangat terbantu.

Waktu proses untuk metode ini bisa terbilang cepat, terlebih lagi dibantu dengan aplikasi pengolahan data atau bahasa pemrograman, seperti yang peneliti lakukan yaitu menggunakan *RStudio. K-Means Clustering* bisa didapatkan oleh beberapa aplikasi lain seperti SPSS, bahasa pemrograman *python, RStudio,* hingga *javascript.* Implementasi yang mudah pun menjadi keunggulan metode ini terutama dalam bidang pemasaran yang memanfaatkan era teknologi dan *big data*. Bahkan untuk sekelas UMKM baru pun mudah untuk diimplementasi asalkan sudah mempunyai data yang dibutuhkan.

Namun variabel terakhir yaitu tingkatan pendidikan tidak bisa digunakan metode ini dikarenakan bentuk data yang dimiliki yaitu *character* atau bukan bersifat nilai yang bisa diukur jadi hanya menggunakan metode statistic deskriptif.

5.5 Analisis Pemilihan Nilai *K* yang Optimal dalam Metode *K-Means Clustering*

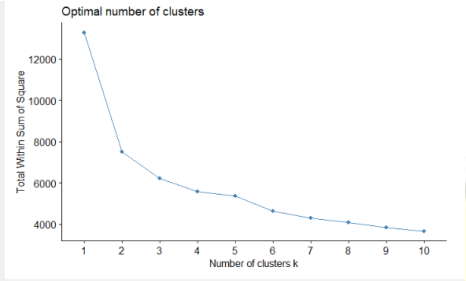
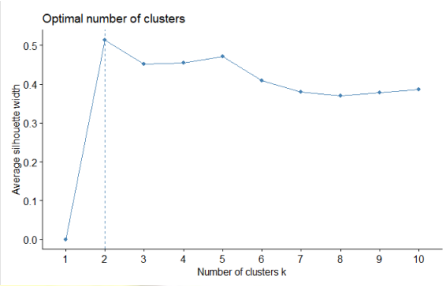
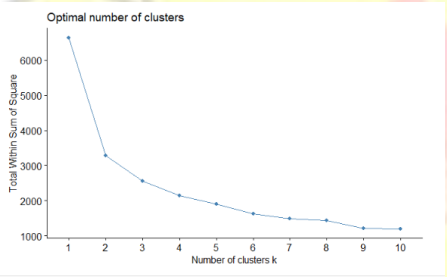
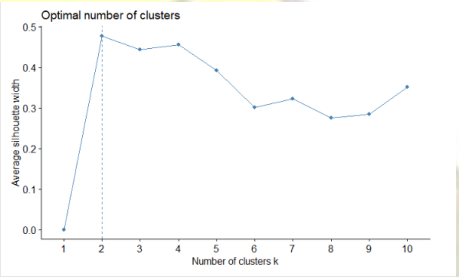
Dalam metode *K-Means Clustering* terdapat beberapa metode evaluasi yang digunakan untuk mendapatkan nilai yang kluster (*K)* yang optimal dari setiap variabel. Dalam Bab 4 sudah dijelaskan terdapat 2 metode evaluasi yang digunakan yaitu *elbow method* dan *silhouette method*.

Elbow method merupakan metode evaluasi yang menginterpretasikan dan uji performa tingkat konsistensi jumlah kluster yang tepat dengan melihat SSE atau *Sum of Square Error*. Dalam metode ini, peneliti memilih rentang nilai kandidat *K* dengan mencari jarak rata-rata setiap titik atau data dalam sebuah kluster (*K)* ke centroid-nya atau titik pusatnya, dan dengan mudah dinyatakan kedalam sebuah plot. Dalam plot tersebut dinyatakan dengan grafik yang turun dengan drastis dan sudah dijelaskan secara perhitungann pada Sub bab 4.2.5.

Selanjutnya metode kedua yaitu *silhouette method* yang merupakan metode dengan menghitung koefisien *silhouette* dari setiap titik yang mengukur seberapa mirip suatu titik dengan klasternya sendiri dibandingkan dengan klaster lainnya. dengan memberikan representasi grafis ringkas tentang seberapa baik setiap objek telah diklasifikasikan. Metode evaluasi tersebut sudah dijelaskan secara perhitungannya dalam Sub bab 4.2.6.

Dua variabel yang menggunakan metode *K-Means Clustering* menggunakan 2 metode evaluasi ini, yaitu variabel jumlah pembelian produk per kategori dan variabel metode pembelian yang akan dijelaskan dengan tabel dibawah ini:

Tabel 5. 1. Tabel Pemilihan *K* yang optimal

Variabel	*Elbow Method*	*Silhouette Method*
Jumlah pembelian produk per kategori	 K = 2	 K = 2
Metode pembelian	 K = 2	 K = 2

Didapatkan hasil dari setiap variabel tersebut memiliki nilai *K* =2 yang optimal dan akan dijadikan dalam *K* dalam metode proses *K-Means Clustering* terhadap masing-masing variabel.

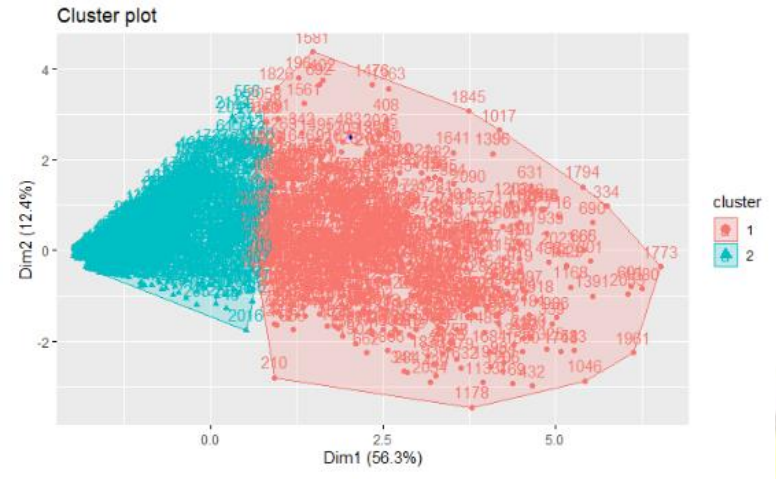
5.6 Analisis Hasil Setiap Variabel

Hasil analisis metode *K-Means Clustering* pada setiap variabel yang sudah ditentukan sesuai tujuan memiliki peran penting dalam penentuan segmentasi konsumen. Berikut adalah hasil dan interpretasi metode *K-Means Clustering* dari variabel yang telah ditentukan sesuai tujuan dan digambarkan dengan tabel dibawah :

- Analisis hasil metode *K-Means Clustering* pada variabel jumlah pembelian produk per kategori

Didapatkan hasil kluster yang optimal setelah menggunakan metode *K-Means Clustering* adalah 2 kluster yaitu kluster pertama menyatakan konsumen dengan

pembelian produk yang banyak dan kluster kedua menyatakan konsumen dengan pembelian sedikit.



Hasil *cluster plot* tersebut didapatkan untuk kluster pertama yang menyatakan konsumen dengan pembelian produk yang banyak dan terlihat lebih meluas areanya dibandingkan dengan kluster kedua yang menyatakan konsumen dengan pembelian produk sedikit. Terlihat dalam kluster pertama lebih meluas karena terdapat data yang jaraknya jauh namun jika dibaca algoritma data tersebut akan menentukan pada titik centroid terdekat yaitu kluster pertama.

A tibble: 2 x 7

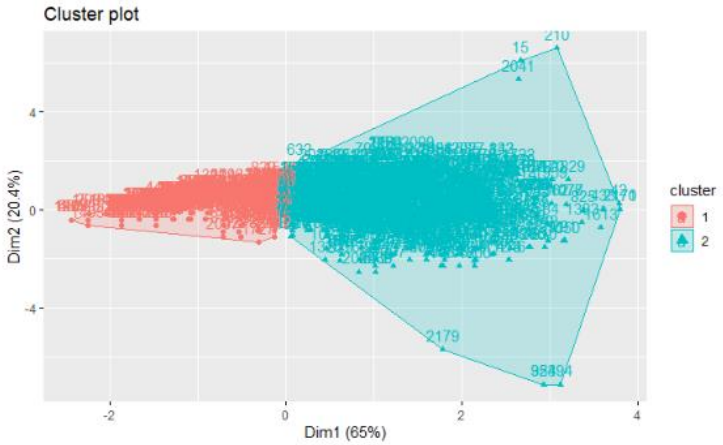
cluster	MntWines	MntFruits	MntMeatProducts	MntFishProducts	MntSweetProducts	MntGoldProds
<int>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	607.2130	66.394203	415.60290	95.02609	68.992754	82.65507
2	168.3047	8.170059	54.31517	11.63427	8.025607	26.35850

2 rows

Jika melihat hasil menggunakan tabel dari variabel tersebut didapatkan untuk kategori paling banyak diminati adalah *Wines Product* atau produk minuman anggur dan dilanjut dengan *Meats, Fish, Gold, Sweet* dan *Fruits*. Jadi dengan ini perusahaan diharapkan bisa menentukan *campaign* selanjutnya dengan mempertahankan produk yang paling banyak pembeliannya yaitu *Wines.*

- Analisis hasil metode *K-Means Clustering* pada variabel metode pembelian

Didapatkan hasil kluster yang optimal setelah menggunakan metode *K-Means Clustering* adalah 2 kluster yaitu kluster pertama menyatakan konsumen dengan pembelian menggunakan metode pembelian sedikit dan kluster kedua menyatakan konsumen dengan pembelian menggunakan metode pembelian yang banyak.



Melihat hasil *cluster plot* tersebut didapatkan untuk kluster pertama yang menyatakan metode pembelian sedikit memiliki area yang mengkerucut dan lebih padat dibandingkan dengan kluster kedua yang menyatakan konsumen dengan pembelian yang banyak yang lebih meluas. Terlihat dalam kluster kedua lebih meluas karena terdapat data yang jaraknya jauh namun jika dibaca algoritma data tersebut akan menentukan pada titik centroid terdekat yaitu kluster kedua.

A tibble: 2 x 4

cluster <int>	NumWebPurchases <dbl>	NumCatalogPurchases <dbl>	NumStorePurchases <dbl>
1	2.242614	0.6624888	3.375112
2	5.968066	4.7189781	8.281934

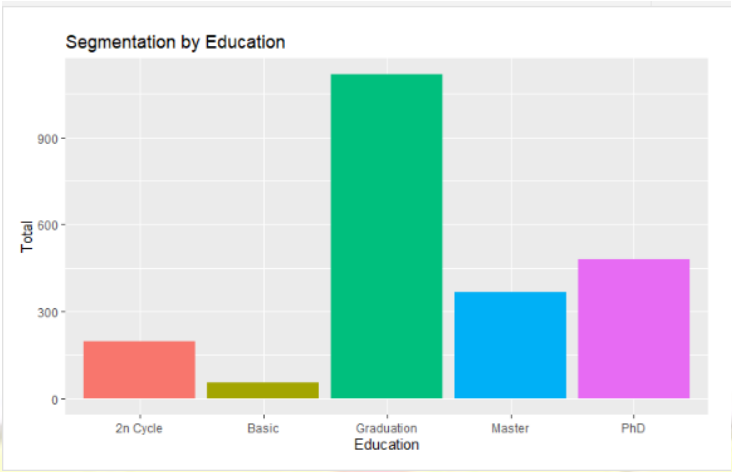
2 rows

Jika melihat hasil menggunakan tabel didapatkan untuk metode pembelian terbanyak yaitu *Store Purchases* atau pembelian ke toko langsung. Selanjutnya ada pembelian menggunakan *Web* dan *Catalog* atau aplikasi yang telah disediakan perusahaan. Jadi perusahaan bisa melakukan *campaign* untuk mempertahankan pelanggan pada pembelian ke toko, dan juga bisa menambahkan *revenue* dengan menambah *campaign* pada pembelian menggunakan *catalog* atau aplikasi yang lebih sesuai dengan era teknologi sekarang.

- Analisis hasil metode *Static Descriptif* pada variabel tingkatan edukasi

Didapatkan hasil berdasarkan variabel tingkatan edukasi terdapat 5 jenis konsumen dengan berbeda tingkatan edukasi yaitu tingkat *graduation* atau mahasiswa strata 1 yang sudah lulus yang memiliki perolehan tertinggi, setelah itu dilanjut dengan *PhD* atau mahasiswa strata 3, *master* atau mahasiswa strata 2, *2n Cycle* atau program

gelar multidisiplin, dan yang terakhir dengan tingkatan sangat sedikit yaitu *basic* atau jika di Indonesia digambarkan dengan SD dan SMP (bersekolah selama 9 tahun).



Hasil analisis untuk variabel ini tidak memiliki kluster karena tidak bisa menggunakan metode *K-Means Clustering* dikarenakan tipe data yang diperoleh berupa karakter sehingga tidak memiliki nilai. Pada tingkatan edukasi konsumen dengan jumlah tertinggi yaitu *graduation* dikarenakan perusahaan *IFood* merupakan perusahaan teknologi baru yang umumnya konsumen yang sudah memiliki gelar memiliki pemahaman terhadap teknologi yang tinggi sehingga lebih ingin melakukan pembelian diperusahaan teknologi terbaru ini pada tahun tersebut.

5.7 Analisis Segmentasi Konsumen dan Contoh Implementasinya

Setelah dilakukannya penentuan jumlah kluster untuk setiap variabel yang ditentukan pada awal, maka segmentasi untuk jumlah produk pembelian per kategori dan metode pembelian merupakan termasuk segmentasi menurut perilaku konsumsi dikarenakan perilaku konsumen menjadi indikator utama pengelompokkan atau kluster. Dalam segmentasi perilaku konsumsi tersebut dijelaskan jika setiap konsumen memiliki perilaku yang berbeda terhadap produk pada perusahaan *IFood* tersebut.

Jadi dikatakan 2 variabel tersebut masuk kedalam jenis segmentasi ini sehingga perusahaan bisa memaksimalkan *campaign* mereka dengan memanfaatkan dari perilaku konsumsi konsumennya. Perusahaan *IFood* bisa mempertahankan perilaku konsumsi dari konsumen mereka yang sudah memiliki banyak pembelian dan menggunakan pembelian dengan banyak metode seperti menambahkan fitur *membership* agar konsumen bisa tetap loyal untuk membeli diperusahaan tersebut. Sebaliknya perusahaan *IFood* juga bisa menambahkan *campaign* yang berhubungan dengan peningkatan jumlah pembelian dan jumlah penggunaan metode pembelian seperti diskon pada beberapa barang yang menjadi *top sales* dari campaign sebelumnya dengan memberikan diskon tambahan jika membeli melalui website atau aplikasi yang sudah disediakan.

Sedangkan untuk tingkatan edukasi termasuk segmentasi menurut demografis karena pengelompokkan berdasarkan data kependudukan atau bisa disebut *profil* konsumen dengan 5 tingkatan edukasi pada perusahaan *IFood.* Terdapat banyak *campaign* untuk mendapatkan *revenue* sesuai yang diingkan jika melihat variabel ini seperti perusahaan bisa memberikan potongan harga jika menunjukkan *id card* dari mahasiswa yang telah lulus atau sedang berkuliah dikarenakan jenis tingkatan ini memiliki nilai tertinggi pada tingkatan tersebut. Selain itu juga bisa menambahkan *campaign* pada hasil terendah yaitu *basic* seperti memberikan promosi beli satu gratis satu untuk beberapa produk yang banyak diminati oleh kalangan tersebut contohnya *sweet product* atau produk manis yang banyak digemati, sehingga diharapkan tingkatan edukasi *basic* tersebut ingin membeli dengan jumlah banyak dari kategori produk tersebut.

Bab VI
Kesimpulan dan Saran

6.1 Kesimpulan

Berdasarkan tujuan penelitian, hasil pengolahan data, analisis dan pembahasan mengenai "Penentuan Segmentasi Konsumen Menggunakan Metode *K-Means Clustering"* peneliti mengambil kesimpulan sebagai berikut ini:

1. Terdapat 2 kluster atau pengelompokkan konsumen berdasarkan jumlah pembelian produk per kategori yaitu konsumen dengan jumlah pembelian produk yang banyak dan konsumen dengan jumlah pembelian produk sedikit, dan untuk variabel ini termasuk dalam kategori segmentasi menurut perilaku konsumsi dengan produk kategori terbanyak dibeli yaitu *wines products* dan produk kategori paling sedikit dibeli yaitu *fruits product*.

2. Terdapat 2 kluster atau pengelompokkan konsumen berdasarkan metode pembelian yaitu konsumen dengan metode pembelian yang sedikit dan konsumen dengan metode pembelian banyak, dan untuk variabel ini termasuk dalam kategori segmentasi menurut perilaku konsumsi dengan metode pembelian terbanyak yaitu *store purchases* dan metode pembelian paling sedikit digunakan yaitu *catalog* atau aplikasi perusahaan.

3. Terdapat 5 tingkatan edukasi konsumen yang telah melakukan pembelian pada perusahaan *IFood* yaitu yang tertinggi *graduation* dilanjut dengan *PhD, Master, 2n Cycle* dan tingkatan aterakhir adalah *basic* yang memiliki perolehan paling sedikit dalam variabel tersebut. Untuk variabel ini termasuk kedalam kategori segmentasi menurut demografis.

Jika secara menyeluruh dari metode *K-Means Clustering* yang sudah dilakukan terdapat kelemahan dari metode tersebut yaitu kluster yang dihasilkan masih terlalu *general* atau meluas sehingga perusahaan *IFood* belum maksimal melakukan strategi pemasaran selanjutnya yaitu *targeting* secara mendetail. Perlu dilakukannya proses lanjutan yaitu dengan menganalisis dari variabel lainnya dan mengkaitkan dengan hasil kluster dari metode *K-Means Clustering* tersebut agar lebih maksimal. Secara kompleksitas metode ini memiliki tingkat kompleksitas yang rendah dikarenakan mencari secara *general* kluster yang terdapat pada setiap variabel yang ditentukan,

terlebih lagi dengan bantuan *RStudio* tentunya dapat memudahkan perusahaan untuk melakukan metode ini dengan hanya menentukan variabel yang *eligiable* atau layak secara metode tersebut namun dengan tingkat ketelitian yang rendah.

6.2 Saran

Berdasarkan analisis yang telah dilakukan maka peneliti memberikan saran yang diharapkan dapat menjadi masukan bagi perusahaan dan sebagai bahan pertimbangan yaitu sebagai berikut:

1. Perusahaan dapat mengaplikasikan dalam *campaign* selanjutnya hasil dari penelitian ini dengan menambahkan 3 variabel tersebut untuk dilakukan proses strategi pemasaran selanjutnya yaitu *targeting* dan *positioning*.
2. Perusahaan sebaiknya membuat *campaign* untuk fase berikutnya yang lebih berhubungan dengan hasil analisis berikut seperti yang sudah dicontohkan pada Sub Bab 5.7.
3. Perusahaan dapat memilih analisis dari ketiga variabel tersebut dengan menyesuaikan kondisi perusahaan terutama biaya *campaign* atau marketing pada fase berikutnya.
4. Perusahaan dapat menambahkan metode analisis statistik deskriptif dengan variabel lain selain ketiga variabel tersebut, sehingga bisa lebih terperinci produk atau metode pembelian manakah yang lebih cocok untuk diimplementasikan dalam *campaign* selanjutnya.

DAFTAR PUSTAKA

Aditya, A., Jovian, I., & Sari, B. N. (2020). Implementasi K-Means Clustering Ujian Nasional Sekolah Menengah Pertama di Indonesia Tahun 2018/2019. *Jurnal Media Informatika Budidarma Vol. 4 No. 1*, 51-58. Diunduh pada 8 Juni 2022 pukul 21.24

Assauri, S. (2017). *Manajemen Pemasaran.* Jakarta: FT RajaGrafindo Persada. Diunduh pada 16 Mei 2022 pukul 20.09

Creative Commons. (2017, November 7). *Public Domain*. Retrieved from CC0 1.0 Universal: https://creativecommons.org/publicdomain/zero/1.0, diakses pada 31 Juli 2022 pukul 20.41

Handoko, K. (2016). Penerapan Data Mining Dalam Meningkatkan Mutu Pembelajaran Pada Instansi Perguruan Tinggi Menggunakan Metode K-Means Clustering (Studi Kasus Program Studi TKJ Akademi Komunitas Solok Selatan). *TEKNOSI Vol. 2 No. 3*, 31-40. Diunduh pada 8 Juni 2022 pukul 21.55

Hijrah, L. (2017). Segmentasi Konsumen Pada Pasar Online di Indonesia. *Forum Ekonomi Vol. 19 No.2* , 210-19. Diunduh pada 12 Juli 2022 21.24

Irawan, Y. (2019). Penerapan Data Mining Untuk Evaluasi Data Penjualan Menggunakan Metode Clustering dan Algoritma Hirarki Divisive. *JTIULM Vol. 4 No. 1*, 13-20. Diunduh pada 8 Juni 2022 pukul 22.06

Kijewska, A., & Bluszcz, A. (2016). Research of Varying Levels of Greenhouse Gas Emissions in European Countries Using the K-Means Method. *Atmospheric Pollution Research 7*, 935-944. Diunduh pada 8 Juni 2022 pukul 22.35

Kotler, K. (2016). *Marketing Management.* England: Pearson Education. Diunduh pada 31 Juli 2022 pukul 20.17

Leite, N. B. (2020). *IFood CRM Data Analyst Case.* Github & Ifood. Diunduh pada 10 Agustus 2022 pukul 21.10

Metisen, B. M., & Sari, H. L. (2015). Analisis Clustering Menggunakan Metode K-Means Dalam Pengelompokkan Penjualan Produk Pada Swalayan Fadhila. *Jurnal Media Infotama Vol. 11 No. 2*, 110-118. Diunduh pada 8 Juni 2022 pukul 21.36

Muningsih, E. (2014). Penerapan Metode Clustering K-Means Untuk Menentukan Kategori Stok Barang. *Prosiding - Seminar Nasional Ilmu Komputer* , 1-6. Diunduh pada 8 Juni 2022 pukul 21.43

Nagari, S. S., & Inayati, L. (2020). Implementation of Clustering Using K-Means Method to Determine Nutritional Status. *Jurnal Biometrika dan Kependudukan Vol. 9 Issue 1 July*, 62-68. Diunduh pada 24 Juli 2022 pukul 19.58

Pradana, M. (2015). Klasifikasi Jenis-Jenis Bisnis E-Commerce di Indonesia. *Neo-bis Vol. 9 No. 2*, 32-40. Diunduh pada 3 Agustus 2022 pukul 18.49

RDocumentation. (1970, January 1). *R packages*. Retrieved from NULL : The Null Object: https://www.rdocumentation.org, diakses pada 31 Juli 2022 pukul 17.42

Roihan, A., Sunarya, P., & Rafika, A. S. (2020). Pemanfaatan Machine Learning dalam Segala Bidang. *IJCIT 5(1)*, 75-82. Diunduh pada 3 Agustus 2020 pukul 19.00

Rosidin, M. (2019). *Metode Numerik Menggunakan R untuk Teknik Lingkungan.* Diunduh pada 31 Juli 2022 pukul 17.44

Santika Devi, N. A., Sukarsa, I., & Darma Putra, I. G. (2015). Implementasi Metode Clustering DBScan pada Proses Pengambilan Keputusan. *Lontar Komputer Vol. 6 No. 3*, 185-191. Diunduh pada 7 Juni 2022 pukul 22.38

Sari, H. L., & Suranti, D. (2016). Perbandingan Algoritma Fuzzy C-Means (FCM) dan Algoritma Mixture Dalam Penclusteran Data Curah Hujan Kota Bengkulu. *ISSN: 1907-5022*, 7-15. Diunduh pada 7 Juni 2022 pukul 22.51

Sodexo. (2019, November 28). *Apa itu Segmentasi Pelanggan dan Dampaknya bagi Bisnis Anda?* Retrieved from Sodexo Corporation Website: https://www.sodexo.co.id/segmentasi-pelanggan-dan-dampaknya-bagi-bisnis-anda, diakses pada 27 Juli 2022 pukul 23.46

Trans Cosmos. (2013). *Global E-Commerce One Stop Service*. Retrieved from Trans Cosmos Web Site: https://www.trans-cosmos.co.id/service/ec, diakses pada 16 Maret 2022 pukul 20.37