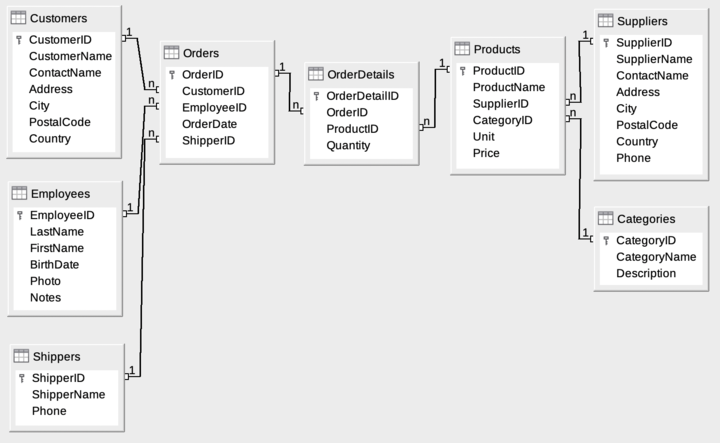
UTS IBDA 3122 Knowledge Discovery

Dibuat Oleh: Stefannus Christian 202000138 (NIM Genap)

Text

Description automatically generated



Persiapan Data

Hal pertama yang saya lakukan adalah *remove* semua *unicode* yang tidak dapat dibaca oleh *read\_*csv pandas. Contohnya adalah untuk customer dengan nama Bólido Comidas preparadas akan saya ubah menjadi Blido Comidas preparadas. Hal kedua yang saya lakukan mengekstrak Tanggal Bulan Tahun dari Kolom Order Date. Hal ini saya untuk mempermudah proses groupby yang nantinya akan dilakukan untuk mencari informasi – informasi penting pada database Northwind. Contoh untuk Order Date 1-Jan-94 akan saya ekstrak 1 (tanggal), 1 (bulan), dan 1994 (tahun), kemudian akan saya tambahkan ke kolom paling akhir tabelnya untuk mempermudah proses groupby.

Kode process date yang digunakan.

Text

Description automatically generated

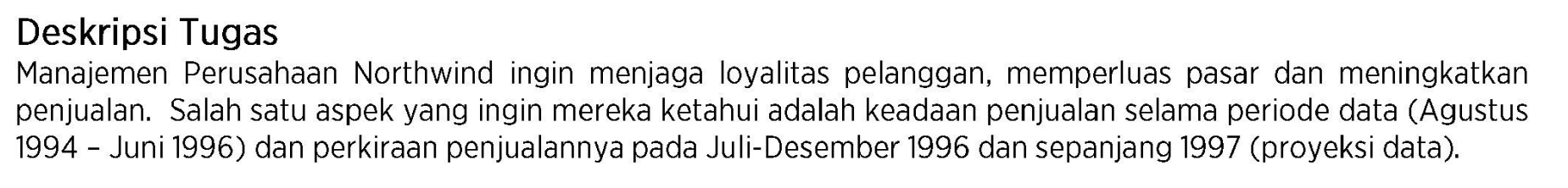
Langkah kedua yang dilakukan adalah memproses kolom diskon dan unit price. Kode yang digunakan adalah seperti ini.

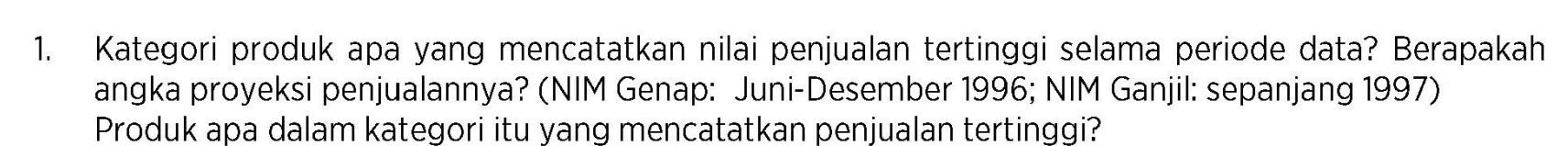
Text

Description automatically generated

Hal yang dilakukan kode diatas hanya mengubah unit price dan discount menjadi float dengan slicing string sehignga symbol $ (untuk unit price) dan % (untuk diskon) tidak diambil karena pada dataset yang diberikan, tipe data kedua kolom ini adalah string.

Langkah ketiga yang dilakukan setelah membersihkan datanya adalah adalah membuat kolom baru yaitu kolom nilai penjualan pada tabel order details. Kolom ini digunakan untuk mencari kategori apa atau produk apa yang memiliki nilai penjualan tertinggi dll. Cara untuk membuat kolom tabel penjualan adalah mengalikan Unit Price dengan Quantity dan Diskon. Jadi Nilai Penjualan = Unit Price \* Quantity \* (1 – Diskon). Contohnya untuk unit price 33.25, quantity 2 dengan discount 3% maka Nila penjualan nya adalah 64.505 yang didapatkan dari (33.25) \* 2 \* 0.97 = 64.505





Untuk menjawab pertanyaan ini, hal yang harus dilakukan adalah menggabungkan tabel order, order details dengan products. Kemudian langkah selanjutnya adalah groupby tabel yang sudah digabungkan based on Category dan di sort berdasarkan Nilai Penjualan. Berikut hasil tabelnya.

Graphical user interface, text, application

Description automatically generated

Dapat dilihat dari tabel diatas bahwa kategori *Beverages* memiliki nilai penjualan tertinggi selama periode Agustus 1994 – Juni 1996 dengan nilai penjualan sebesar $267868.

Pihak manajemen perusahaan juga ingin mengetahui produk apa memiliki nilai penjulaan terbesar di kategori dengan penjualan terbesar. Dengan kata lain produk beverages apa yang memiliki nilai penjualan terbesar. Cara yang saya lakukan adalah groupby tabel diatas dengan memfilter Category == “Beverages” dan di sort lagi berdasarkan nilai penjualan. Berikut hasil tabelnya untuk top 5 Beverages dengan nilai penjualan tertinggi.

A picture containing graphical user interface

Description automatically generated

Dapat dilihat dari tabel diatas bahwa beverages dengan niali penjualan tertinggi adalah Cote De Blaye dengan nilai penjualan sebesar $141396.

Salah satu aspek yang diinginkan manajemen perusahaan northwind adalah proyeksi data pada Juli – Desember 1996 untuk kategori dengan nilai penjualan tertinggi yaitu Beverages. Saya menggunakan Linear Regression dan Random Forest untuk melakukan hal ini. Random Forest akan memberikan akurasi yang jauh lebih tinggi dibandingkan dengan Linear Regression, tetapi untuk melihat apakah proyeksi data nya naik atau turun jauh lebih mudah dilihat menggunakan Linear Regression. Hal pertama yang saya lakukan adalah groupby tabel beverages berdasarkan Category, Bulan, dan Tahunnya. Berikut merupakan hasil tabel groupby nya.

A black screen with white text

Description automatically generated with low confidence

Setelah itu saya mengambil Tahun dan bulan sebagai input variable (X) dan Nilai Penjualan sebagai ouput variable (y)

Text

Description automatically generated

Kemudian data yang ingin di prediksi disini adalah data dari Juli 1996 – Desember 1996. Berikut merupakan kode plotting prediksi nya.

Text

Description automatically generated

Berikut merupakan grafik proyeksi nilai penjualan Beverages untuk proyeksi menggunakan linear regression. Proyeksi nilai penjualan Beverages dari Juli 1996 – Desember 1996 adalah sebagai berikut.

Chart, line chart

Description automatically generated

Graphical user interface, application

Description automatically generated

Berdasarkan data pada kolom "Proyeksi Nilai Penjualan", terlihat bahwa terdapat tren penurunan pada nilai penjualan yang diproyeksikan dari bulan Juli 1996 hingga Desember 1996. Hal ini dapat disebabkan oleh berbagai faktor seperti perubahan permintaan pasar, musiman, atau persaingan.

Berikut merupakan grafik proyeksi nilai penjualan Beverages untuk proyeksi menggunakan random forest. Proyeksi nilai penjualan Beverages dari Juli 1996 – Desember 1996 adalah sebagai berikut.

Chart, line chart

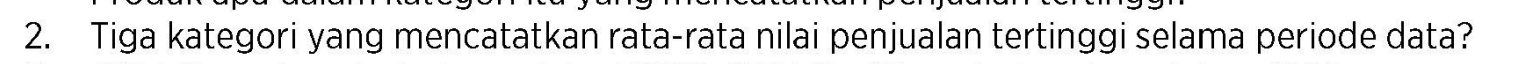
Description automatically generated

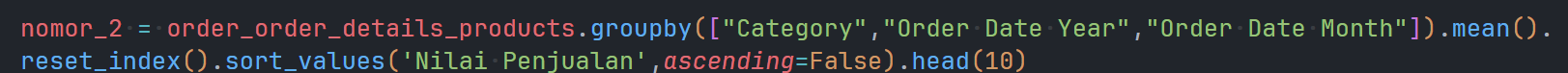
Graphical user interface

Description automatically generated with low confidence

Dengan menggunakan random forest, proyeksi penjualan beverages dari Juli 1996 - Desember 1996 cenderung meningkat setiap bulannya. Pada bulan Juli, proyeksi penjualan beverages sebesar 10755.57650. Kemudian pada bulan Agustus proyeksi penjualan meningkat menjadi 11315.04975, dan pada bulan September proyeksi penjualan sedikit menurun menjadi 11298.93255. Namun, proyeksi penjualan beverages kembali meningkat pada bulan Oktober menjadi 11173.18100 dan terus meningkat pada bulan November menjadi 11965.21925. Pada bulan Desember, proyeksi penjualan beverages mencapai puncaknya dengan nilai 12300.99240. Oleh karena itu, proyeksi penjualan beverages dari Juli 1996 - Desember 1996 mengalami kecenderungan yang meningkat.

Terdapat perbedaan prediksi menggunakan 2 model yang berbeda. Dengan menggunakan model linear regression, proyeksi data cenderung menurun, tetapi menggunakan Random Forest, proyeksi data cenderung mengingkat. Hal ini terjadi karena linear regression mengasumsikan bahwa hubungan antara variabel independen dan variabel dependen bersifat linear, sehingga jika terdapat tren atau pola tertentu dalam data yang cenderung menurun, maka proyeksi data yang dihasilkan oleh linear regression juga cenderung menurun. Sedangkan pada metode random forest, proyeksi dilakukan dengan menggabungkan hasil dari beberapa model pohon keputusan yang berbeda. Random forest tidak asumsi hubungan antara variabel independen dan variabel dependen seperti yang dilakukan pada linear regression. Sehingga, jika ada pola non-linear atau kompleksitas dalam data, random forest lebih mampu menangkapnya dan menghasilkan proyeksi data yang lebih baik dan akurat.



Cara untuk mencari Tiga Kategori dengan rata-rata nilai penjualan tertinggi sama seperti nomor 1 hanya bedanya adalah disini yang dicari adalah rata – rata nilai penjualan. Berikut kode pandas python yang saya gunakan. 

Berikut merupakan hasil dari Tiga Kategori yang mencatatkan rata – rata nilai penjualan tertinggi.

Chart, bar chart

Description automatically generated

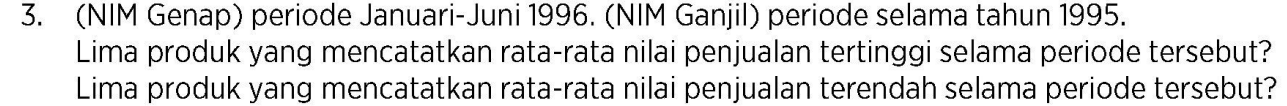
Tiga Kategori itu ialah Beverages pada Bulan Kedua Tahun 1995, Meat / Poultry Bulan Maret 1996, dan Produce pada Bulan 11 Tahun 1995.

Graphical user interface, application

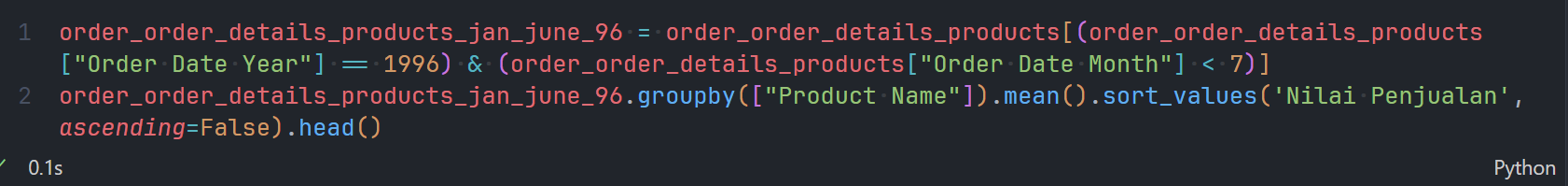
Description automatically generated

Dapat dilihat bahwa Beverages sebagai kategori yang memiliki nilai penjualan tertinggi juga memiliki rata – rata penjualan yang tertinggi. Hal ini menunjukkan bahwa kategori beverages sangat sukses dalam hal penjualan dan merupakan contributor utama utama untuk pendapatan bisnis. Kategori Beverages tidak hanya memiliki nilai penjualan tertinggi, tetapi juga rata-rata penjualan tertinggi, yang menunjukkan bahwa bisnis memiliki kesuksesan yang konsisten dalam menjual produk-produk yang termasuk dalam kategori tersebut. Hal ini dapat menjadi fokus bisnis untuk mempertahankan keberhasilan kategori ini dan mengoptimalkan strategi pemasaran dan penjualan untuk produk-produk Beverages.

Top 3 kategori dengan nilai penjualan tertinggi dapat dilihat pada tabel nomor 1 adalah Beverages, Dairy Products dan Confections. Dapat dilihat bahwa Meat / Poultry dan Produce tidak ada dalam top 3 nilai penjualan tertinggi melainkan Meat / Poultry ada di posisi ke 4 dan Produce berada di posisi ke 7. Dari informasi tersebut, dapat disimpulkan bahwa kategori Meat/Poultry masih cukup berhasil dalam hal penjualan dan merupakan kontributor signifikan terhadap pendapatan bisnis. Meskipun Meat/Poultry tidak memiliki nilai penjualan tertinggi, namun rata-rata penjualan tertinggi di antara kategori lainnya, yang menunjukkan bahwa produk-produk dalam kategori Meat/Poultry biasanya memiliki nilai penjualan yang lebih tinggi daripada kategori lainnya. Hal ini dapat menjadi fokus bisnis untuk mengevaluasi kinerja kategori ini dan mencari cara untuk meningkatkan penjualan produk-produk yang termasuk dalam kategori Meat/Poultry. Selain itu, juga dapat disimpulkan bahwa kategori Produce mungkin memiliki beberapa produk yang memiliki nilai penjualan yang tinggi, namun secara keseluruhan kategori ini belum berhasil secara maksimal dalam menghasilkan penjualan. Walaupun nilai penjualan dari kategori Produce tidak terlalu tinggi, rata-rata penjualan produk-produk dalam kategori ini relatif lebih tinggi dibandingkan kategori lainnya, yang menunjukkan bahwa produk-produk dalam kategori Produce memang memiliki potensi untuk menghasilkan nilai penjualan yang lebih tinggi. Oleh karena itu, bisa menjadi fokus bisnis untuk melakukan analisis lebih lanjut terhadap kategori Produce untuk memperbaiki strategi pemasaran dan meningkatkan penjualan produk-produk dalam kategori tersebut.



Untuk mencari periode 5 produk yang mencatatkan rata – rata nilai penjualan tertinggi selama periode tertinggi selama periode Januari – Juni 1996 saya hanya filter dan groupby menggunakan kode ini.



5 produk yang mencatatkan rata – rata nilai penjualan tertinggi selama periode Januari – Juni 1996 dapat dilihat pada tabel dibawah ini.

Graphical user interface, text

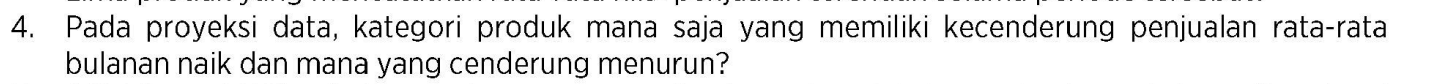
Description automatically generated

5 produk yang mencatatkan rata – rata nilai penjualan terendah selama periode Januari – Juni 1996 dapat dilihat pada tabel dibawah ini.

Graphical user interface, text

Description automatically generated with medium confidence

Dari informasi ini, dapat disimpulkan bahwa Cote De Blaye adalah produk yang sangat sukses dalam hal penjualan, baik dalam hal nilai penjualan maupun rata-rata penjualan. Kategori Beverages, di mana produk ini termasuk, juga terbukti sebagai kategori dengan nilai penjualan tertinggi dan rata-rata penjualan tertinggi selama periode Agustus 1994 - Juni 1996. Ini menunjukkan bahwa produk-produk di kategori Beverages memiliki potensi penjualan yang besar dan strategi pemasaran dapat difokuskan pada kategori ini. Selain itu, Cote De Blaye mungkin merupakan produk yang penting bagi bisnis ini dan dapat menjadi fokus pemasaran dan pengembangan di masa depan.



Untuk menjawab pertanyaan ini, hal yang saya lakukan sama persis untuk menjawab proyeksi data pada nomor 1. Bedanya hanyalah saya melakukan hal ini untuk semua kategori. Berikut merupakan hasil visualisasinya menggunakan model Linear Regression.

Chart, line chart

Description automatically generated

Chart, line chart

Description automatically generated

Chart, line chart

Description automatically generated

Chart, line chart

Description automatically generated

Chart, line chart

Description automatically generated

Chart, line chart

Description automatically generated

Chart, line chart

Description automatically generated

Chart, line chart

Description automatically generated

Informasi yang bisa didapatkan dari grafik diatas adalah sebagai berikut:

1. Beverages memiliki nilai penjualan dan rata – rata penjualan tertinggi, namun penjualan rata-rata bulanannya cenderung menurun pada proyeksi data.
2. Kategori produk yang memiliki kecenderungan penjualan rata-rata bulanan naik pada proyeksi data adalah Condiments, Produce, Dairy Products, Meat / Poultry (sangat kecil), Seafood, dan Grain / Cereals.
3. Kategori produk yang memiliki kecenderungan penjualan rata-rata bulanan turun pada proyeksi data adalah Beverages dan Confections.

Berikut merupakan hasil visualisasinya menggunakan model Random Forest.

Chart, line chart

Description automatically generated

Chart, line chart, histogram

Description automatically generated

Chart, line chart, histogram

Description automatically generated

Chart, line chart

Description automatically generated

Chart, line chart

Description automatically generated

Chart, line chart, histogram

Description automatically generated

Chart, line chart, histogram

Description automatically generated

Chart, line chart

Description automatically generated

Terdapat perbedaan jawaban jika menggunakan Linear Regression dan Random Forest. Hal ini sudah saya jelaskan pada poin nomor 1. Ada tujuh kategori produk yang dianalisis, yaitu Beverages, Condiments, Produce, Meat/Poultry, Seafood, Dairy Products, dan Confections.

1. Berdasarkan plot, kategori produk Beverages, Dairy Products, dan Meat/Poultry menunjukkan kecenderungan proyeksi data yang meningkat. Ini berarti bahwa penjualan produk-produk tersebut diproyeksikan akan meningkat seiring berjalannya waktu.
2. Di sisi lain, kategori produk Confections, Grain/Cereals, dan Seafood menunjukkan kecenderungan proyeksi data yang menurun. Ini berarti bahwa penjualan produk-produk tersebut diproyeksikan akan menurun seiring berjalannya waktu.
3. Kategori produk Condiments dan Produce tidak menunjukkan kecenderungan arah tren yang jelas. Ini berarti bahwa proyeksi penjualan produk-produk tersebut cenderung stabil atau tidak banyak berubah seiring berjalannya waktu.

Dengan demikian, toko dapat mempertimbangkan untuk meningkatkan stok dan fokus pada penjualan produk-produk dalam kategori yang memiliki kecenderungan penjualan rata-rata bulanan naik pada proyeksi data, dan memperhatikan penjualan produk-produk dalam kategori yang memiliki kecenderungan penjualan rata-rata bulanan turun pada proyeksi data. Hal ini dapat membantu toko untuk mengoptimalkan penjualan dan meningkatkan keuntungan.

Text

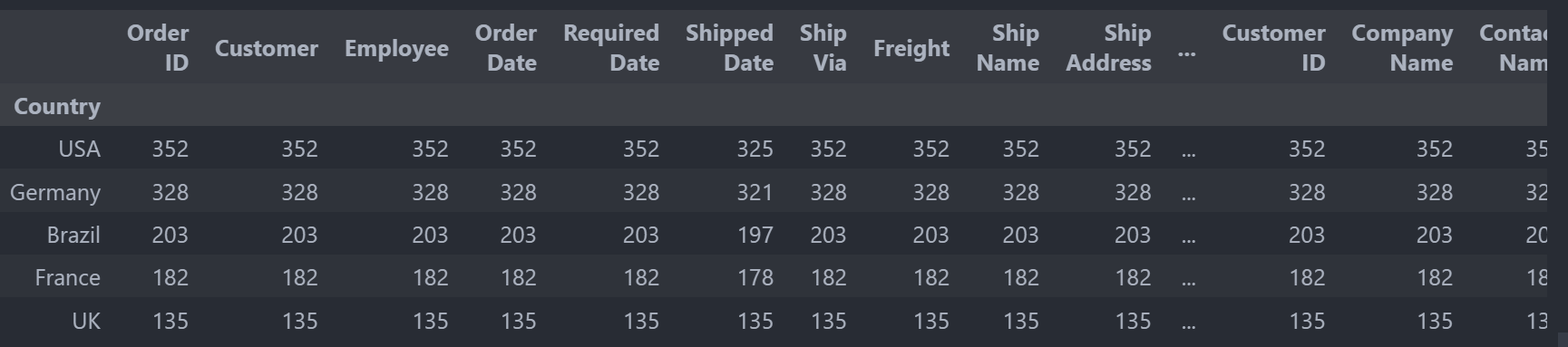
Description automatically generated

Untuk menjawab country mana yang paling banyak melakukan order (count) dan paling tinggi nilai order (sum) hal yang saya lakukan adalah menggabungkan tabel order, order details dan customers kemudian saya groupby berdasarkan country dan saya count kemudian saya sort berdasarkan nilai penjualan. Berikut merupakan kode yang saya gunakan.

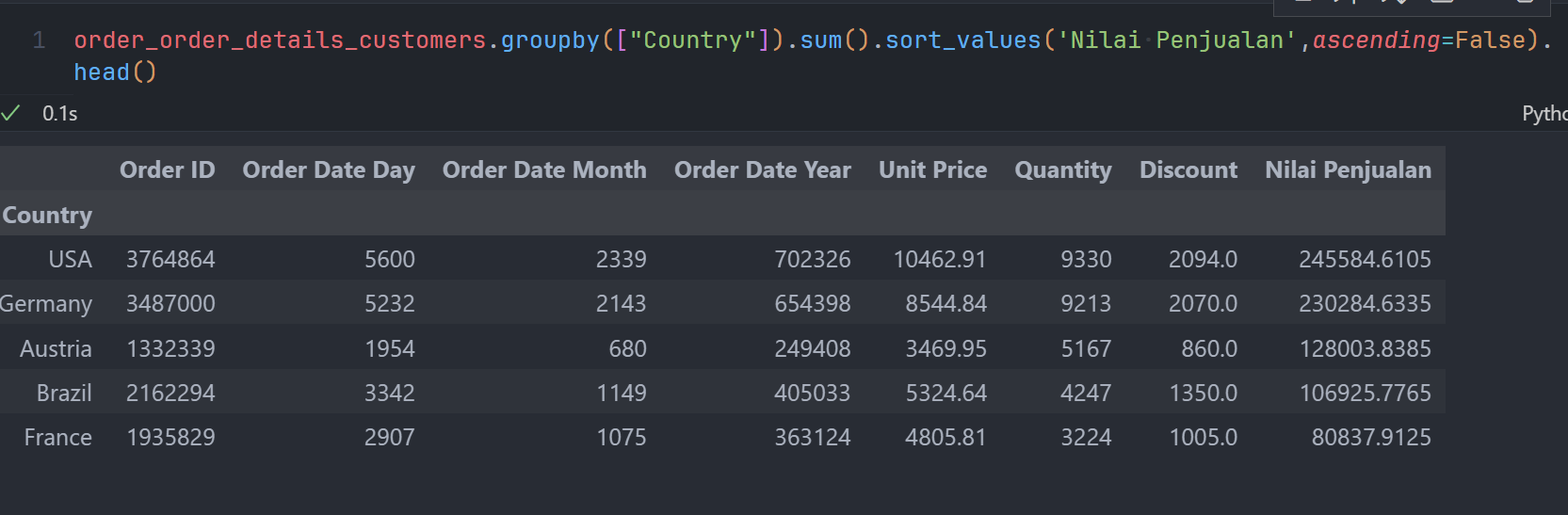
Text

Description automatically generated

Berikut merupakan top 5 country yang paling banyak melakukan order dengan USA sebagai country yang paling banyak melakukan order. Jumlah order negara USA adalah 352 order dari total 2153 Order.



Untuk mencari country yang melakukan nilai order paling tinggi, caranya adalah sama tetapi perbedaan nya ketika di groupby itu si sum bukan di count. Berikut merupakan hasilnya.



Dapat dilihat dari tabel ini bahwa USA dan Germany tetap menjadi Top 2 country yang memiliki nilai order penjualan tertinggi. Hal ini artinya kedua negara ini memiliki market share besar dalam penjualan produk toko Northwind. Brazil dan France merupakan negara dengan peringkat ke-3 dan ke-4 dalam jumlah order terbanyak, namun tidak masuk dalam urutan top 3 nilai order terbanyak. Hal ini menunjukkan bahwa meskipun negara-negara tersebut memiliki banyak pembeli, namun nilai pembelian mereka relatif lebih rendah dibandingkan negara-negara lain seperti Austria. Austria merupakan negara yang masuk dalam urutan top 5 nilai order terbanyak, namun tidak masuk dalam urutan top 5 jumlah order terbanyak. Hal ini menunjukkan bahwa meskipun jumlah pembeli dari Austria tidak sebanyak negara-negara seperti Brazil dan France, namun nilai pembelian mereka cukup tinggi.



Untuk mencari relasi antara kategori produk dengan asal customer, hal yang saya lakukan adalah menggabungkan tabel categories, products, order details, orders, dan customers. Saya melakukan 4 teknik untuk menemukan apakah ada kaitan / relasi antara kategori produk dengan asal customer.

Metode 1 ( .corr() python )

Graphical user interface, text

Description automatically generated

Saya menggunakan fungsi built in .corr() python untuk mencari korelasi antar untuk 2 kolom. Fungsi .corr() ini hanya bisa mencari korelasi jika input yang dimasukkan memiliki tipe data numerik sedangkan fitur country dan kategori produk dua duanya memiliki fitur categorical (string) sehingga perlu diganti dulu menjadi representasi numeriknya. Hal ini saya capai menggunakan ordinal encoder. Ordinal encoder akan mengubah seluruh representasi kategorikal menjadi numerik. Misalnya suatu kolom memiliki 3 kategori yaitu “Red”, “Green”, dan “Blue” maka setelah dimasukkan ordinal encoder outputny adalah 0,1,2. Hasil korelasi yang didapatkan adalah 0.02760358198744432. Hal ini menunjukkan adanya korelasi positif yang lemah antara kedua variabel. Dengan kata lain, tidak ada hubungan yang kuat antara jenis produk yang dibeli oleh pelanggan dan asal negara pelanggan tersebut. Meskipun demikian, informasi ini tetap dapat digunakan oleh toko untuk menyusun strategi pemasaran yang lebih tepat. Toko dapat memperhatikan preferensi konsumen dari masing-masing negara dan mencoba untuk menyesuaikan jenis produk yang ditawarkan dan strategi pemasaran sesuai dengan preferensi tersebut, tanpa perlu mempertimbangkan korelasi antara kategori produk dan asal customer secara signifikan.

Metode 2 ( Chi Squared)

Metode chi squared adalah salah satu metode statistik yang digunakan untuk mengukur korelasi antara dua variabel kategori. Dalam konteks ini, metode chi squared digunakan untuk mencari korelasi antara kategori produk dan asal customer.

Text

Description automatically generated

Hasil chi squared statistic yang saya dapatkan (145.27062658850997) menunjukkan bahwa terdapat perbedaan yang signifikan antara jumlah observasi yang diharapkan dan jumlah observasi yang diamati dalam tabel kontingensi kategori produk dan asal customer. Namun, nilai P-value yang diperoleh (0.36285296961683283) menunjukkan bahwa tidak terdapat korelasi yang signifikan antara kategori produk dan asal customer.

Metode 3 (Cramer’s V Test)

Cramer's V test adalah metode statistik yang digunakan untuk mengukur kekuatan hubungan antara dua variabel kategori. Hasil Cramer's V statistic yang saya dapatkan (1.6106266891920293) menunjukkan bahwa terdapat hubungan yang kuat antara Asal Customer dengan Kategori Produk. Namun, nilai P-value yang diperoleh (0.36285296961683283) menunjukkan bahwa hubungan ini tidak signifikan secara statistik.

Metode 4 (Analisis Manual)

Cara analisis manual yang saya lakukan adalah melihat jumlah kategori produk apa yang dipesan pada setiap negara. Berikut merupakan Analisa manual saya dengan format <negara> = <Category yang paling banyak di pesan>, <Category yang paling sedikit di pesan>

Argentina = Seafood, Grains/Cereal

Austria = Dairy, Beverage

Belgium = Dairy, Beverage

Brazil = Seafood, Beverage

Canada = Seafood, Beverage

Denmark = Seafood, Grain/Cereal

Finland = Seafood, Beverage

France = Seafood, Condiments

German = Seafood, Beverage

Ireland = Seafood, (Confections, Beverage)

Italy = Seafood, Beverage

Mexico = Seafood, Condiments

Norway = Seafood, Condiments

Poland = Produce, Condiments

Portugal = Grains/Cereal, Produce

Spain = Seafood, Produce

Sweden = Seafood, Condiments

Swiss = Seafood, Condiments

UK = Seafood, Condiments

USA = Seafood, Beverage

Venezuela = Seafood, Condiments

Hasil analisis manual menunjukkan bahwa kategori yang paling banyak dan paling sedikit dibeli bervariasi di setiap negara, menunjukkan bahwa negara pelanggan bukanlah faktor yang signifikan dalam menentukan kategori produk.

Kesimpulan

Berdasarkan hasil analisis, nilai korelasi antara kategori produk dan asal customer sangat rendah, yaitu sebesar 0.027. Hal ini menunjukkan bahwa tidak terdapat korelasi yang signifikan antara kategori produk dengan asal customer. Selain itu, hasil uji chi-squared juga menunjukkan nilai p-value yang cukup besar, yaitu 0.3628, sehingga tidak ada bukti yang cukup untuk menolak hipotesis nol bahwa tidak ada korelasi antara kategori produk dan asal customer. Hasil uji Cramer's V juga menunjukkan nilai yang rendah, yaitu 1.6106266891920293, yang mengindikasikan bahwa hubungan antara kategori produk dan asal customer lemah.

Namun, berdasarkan hasil analisis secara manual terhadap setiap negara, terlihat beberapa negara memiliki pola pembelian tertentu terkait kategori produk. Sebagai contoh, banyak negara yang membeli seafood sebagai produk yang paling banyak, sedangkan minuman (beverage) menjadi produk yang sering dibeli di banyak negara. Namun, hal ini tidak cukup untuk menyimpulkan adanya korelasi antara kategori produk dan asal customer, karena bisa saja faktor lain seperti budaya dan preferensi masyarakat mempengaruhi pola pembelian tersebut.

Secara keseluruhan, dapat disimpulkan bahwa berdasarkan nilai korelasi yang rendah dan hasil uji chi-squared serta Cramer's V yang tidak signifikan, tidak terdapat korelasi yang kuat antara kategori produk dengan asal customer. Namun, hasil analisis manual terhadap setiap negara menunjukkan adanya pola pembelian tertentu terkait kategori produk, namun hal ini tidak cukup untuk menyimpulkan adanya korelasi.



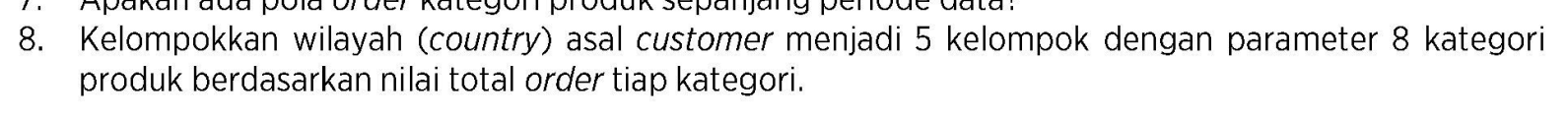
Cara yang saya lakukan untuk menjawab pertanyaan ini adalah mengumpulkan seluruh jumlah order kategori produk dan nilai penjualan masing – masing kategori produk kemudian di plot menjadi satu plot kemudian melihat dan menganalisa apakah ada pola *order* nya atau tidak. Berikut merupakan plot dan hasil temuan saya.

Chart, line chart

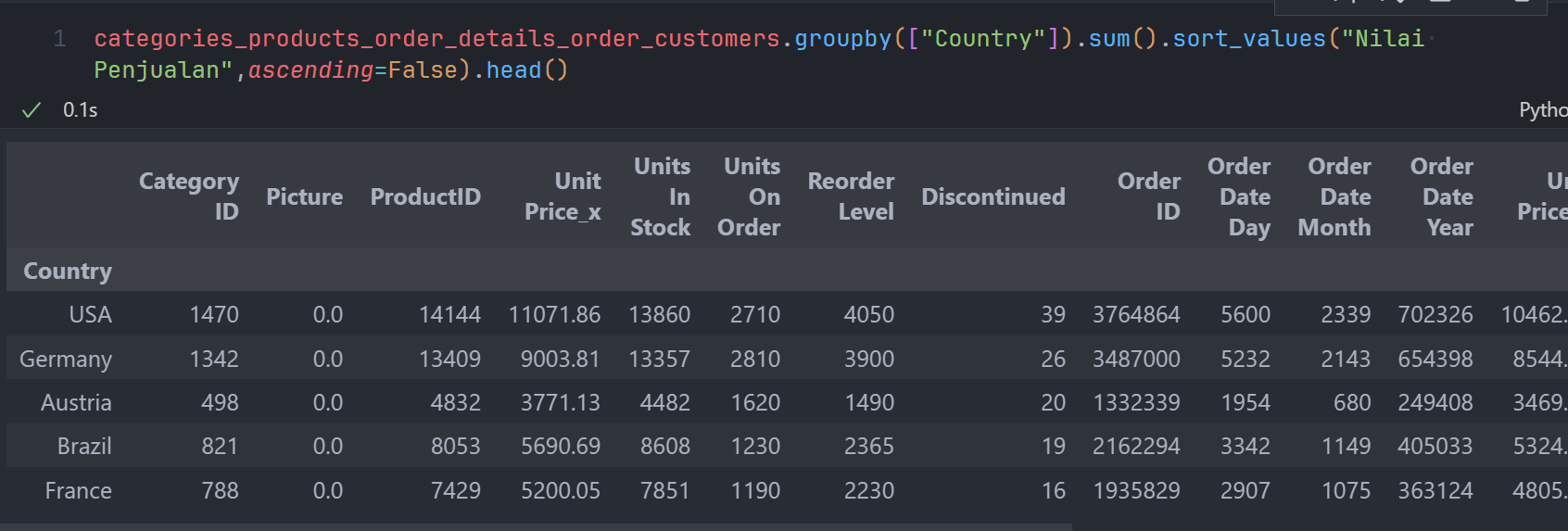
Description automatically generated

Dari data yang diberikan, terlihat bahwa jumlah order tiap bulan pada kategori produk tertentu naik dan turun dalam pola yang berbeda-beda. Namun, secara umum, terlihat bahwa jumlah order pada setiap kategori produk cenderung naik dari waktu ke waktu. Berikut merupakan temuan yang saya dapatkan dari grafik diatas.

1. Beverages adalah kategori produk dengan jumlah order terbanyak pada setiap bulannya, dengan jumlah order tertinggi terjadi pada bulan Mei 1996.
2. Seafood adalah kategori produk dengan jumlah order terbanyak kedua pada setiap bulannya, dengan jumlah order tertinggi terjadi pada bulan Desember 1995.
3. Produce adalah kategori produk dengan jumlah order terendah pada setiap bulannya, dengan jumlah order terendah terjadi pada bulan September 1994.
4. Kategori produk yang menunjukkan kenaikan yang signifikan dalam jumlah order dari waktu ke waktu adalah Beverages dan Seafood. Keduanya menunjukkan tren kenaikan yang stabil hingga akhir periode data.
5. Kategori produk lainnya seperti Condiments dan Meat/Poultry cenderung menunjukkan fluktuasi dalam jumlah order mereka dari waktu ke waktu, meskipun secara umum jumlah order mereka cenderung naik.
6. Selama dua tahun periode data, terdapat peningkatan yang signifikan dalam jumlah order pada bulan-bulan akhir tahun seperti November, Desember dan Mei. Ini dapat mengindikasikan bahwa bulan-bulan ini adalah bulan yang lebih sibuk bagi toko tersebut, mungkin terkait dengan musim liburan atau perayaan.



Saya melakukan k-means Clustering untuk menjawab soal ini. Di sini saya asumsikan bahwa nilai total order == nilai penjualan. Hal pertama yang saya lakukan adalah groupby berdasarkan country kemudian di sort berdasarkan Nilai Penjualan.



Hal berikut yang saya lakukan adalah normalisasi data dengan standard scaler. Untuk melakukan clustering dengan metode k-means, normalisasi data seringkali perlu dilakukan sebelumnya. Hal ini disebabkan karena k-means menggunakan jarak euclidean untuk menghitung jarak antar titik data. Jika variabel-variabel pada data memiliki skala yang berbeda-beda, maka variabel dengan skala besar akan mendominasi dalam perhitungan jarak, sementara variabel dengan skala kecil mungkin tidak signifikan dalam perhitungan jarak. Oleh karena itu, normalisasi data dapat membantu menyamakan skala variabel-variabel pada data sehingga setiap variabel memiliki kontribusi yang seimbang dalam perhitungan jarak. Teknik normalisasi yang saya gunakan adalah standard scaler. Standard scaler adalah teknik normalisasi yang mentransformasi data sedemikian rupa sehingga nilai rata-rata variabel menjadi 0 dan standar deviasi menjadi 1. Teknik ini dilakukan dengan mengurangi nilai setiap data dengan nilai rata-rata variabel dan kemudian membaginya dengan standar deviasi variabel. Hal ini dilakukan untuk setiap variabel pada data.

Mengapa teknik normalisasi ini harus dilakukan sebelum melakukan clustering? Hal ini dikarenakan jika variabel-variabel pada data tidak memiliki skala yang seimbang, k-means dapat menghasilkan klaster yang tidak optimal atau bahkan salah. Selain itu, normalisasi data juga dapat meningkatkan efisiensi dan performa algoritma clustering karena mengurangi jumlah dimensi data dan memperbaiki kemungkinan numerik.

Berikut merupakan kode pythonnya.

Text

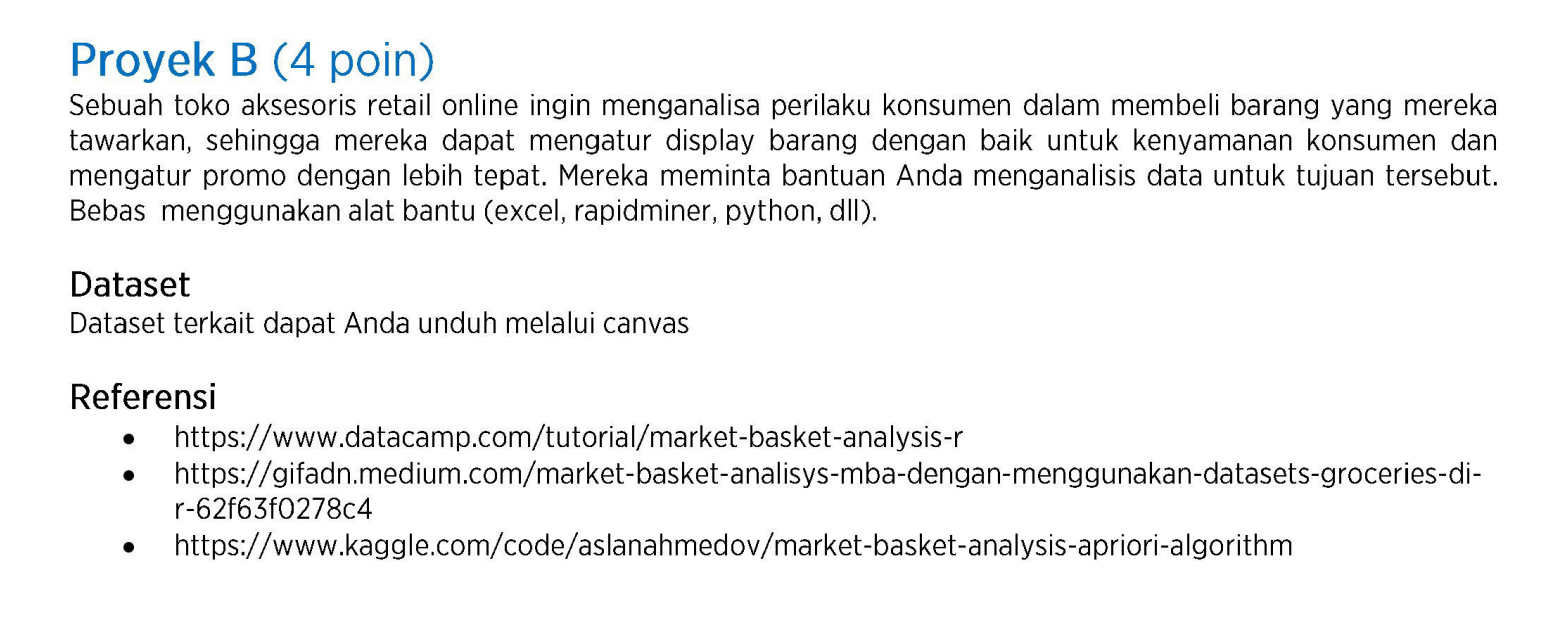
Description automatically generated

Berikut merupakan hasil clusteringnya.

Index Country Cluster

|  |  |  |
| --- | --- | --- |
| 0 | Argentina | 0 |
| 1 | Austria | 0 |
| 2 | Belgium | 0 |
| 3 | Brazil | 0 |
| 4 | Canada | 0 |
| 5 | Denmark | 0 |
| 6 | Finland | 0 |
| 7 | France | 0 |
| 8 | Germany | 0 |
| 9 | Ireland | 0 |
| 10 | Italy | 0 |
| 11 | Mexico | 0 |
| 12 | Norway | 0 |
| 13 | Poland | 0 |
| 14 | Portugal | 0 |
| 15 | Spain | 0 |
| 16 | Sweden | 0 |
| 17 | Switzerland | 0 |
| 18 | UK | 0 |
| 19 | USA | 0 |
| 20 | Venezuela | 0 |
| 21 | Austria | 1 |
| 22 | Brazil | 1 |
| 23 | Canada | 1 |
| 24 | Germany | 1 |
| 25 | Ireland | 1 |
| 26 | Sweden | 1 |
| 27 | Switzerland | 1 |
| 28 | USA | 1 |
| 29 | Argentina | 2 |
| 30 | Austria | 2 |
| 31 | Belgium | 2 |
| 32 | Brazil | 2 |
| 33 | Canada | 2 |
| 34 | Denmark | 2 |
| 35 | Finland | 2 |
| 36 | France | 2 |
| 37 | Germany | 2 |
| 38 | Ireland | 2 |
| 39 | Italy | 2 |
| 40 | Mexico | 2 |
| 41 | Portugal | 2 |
| 42 | Spain | 2 |
| 43 | Sweden | 2 |
| 44 | Switzerland | 2 |
| 45 | UK | 2 |
| 46 | USA | 2 |
| 47 | Venezuela | 2 |
| 48 | Brazil | 3 |
| 49 | Denmark | 3 |
| 50 | Germany | 3 |
| 51 | Ireland | 3 |
| 52 | USA | 3 |
| 53 | Austria | 4 |
| 54 | Belgium | 4 |
| 55 | Brazil | 4 |
| 56 | Canada | 4 |
| 57 | Denmark | 4 |
| 58 | France | 4 |
| 59 | Germany | 4 |
| 60 | Ireland | 4 |
| 61 | Mexico | 4 |
| 62 | Norway | 4 |
| 63 | Spain | 4 |
| 64 | Sweden | 4 |
| 65 | Switzerland | 4 |
| 66 | UK | 4 |
| 67 | USA | 4 |
| 68 | Venezuela | 4 |

Dari hasil clustering tersebut dapat dilihat bahwa terdapat cluster yang memiliki jumlah negara yang sangat banyak seperti cluster 0 dan cluster 2, sedangkan cluster yang lainnya hanya memiliki beberapa negara. Selain itu, terdapat negara yang termasuk ke dalam lebih dari satu cluster seperti Austria, Brazil, Canada, Denmark, Germany, Ireland, Sweden, Switzerland, UK, dan USA. Hal ini menunjukkan bahwa nilai penjualan negara-negara tersebut cukup beragam di setiap kategori produk. Dalam praktiknya, hasil clustering ini dapat membantu dalam mengelompokkan negara-negara yang memiliki karakteristik serupa dalam hal nilai penjualan masing-masing kategori produk. Hal ini dapat membantu perusahaan dalam mengambil keputusan terkait strategi pemasaran dan penjualan di setiap negara yang berbeda. Namun, perlu diingat bahwa hasil clustering ini hanya dapat menjadi panduan dan tidak sepenuhnya dapat dijadikan acuan dalam pengambilan keputusan karena masih dapat terdapat perbedaan yang signifikan dalam hal demografi, budaya, dan kebiasaan konsumen antara negara satu dengan yang lainnya.

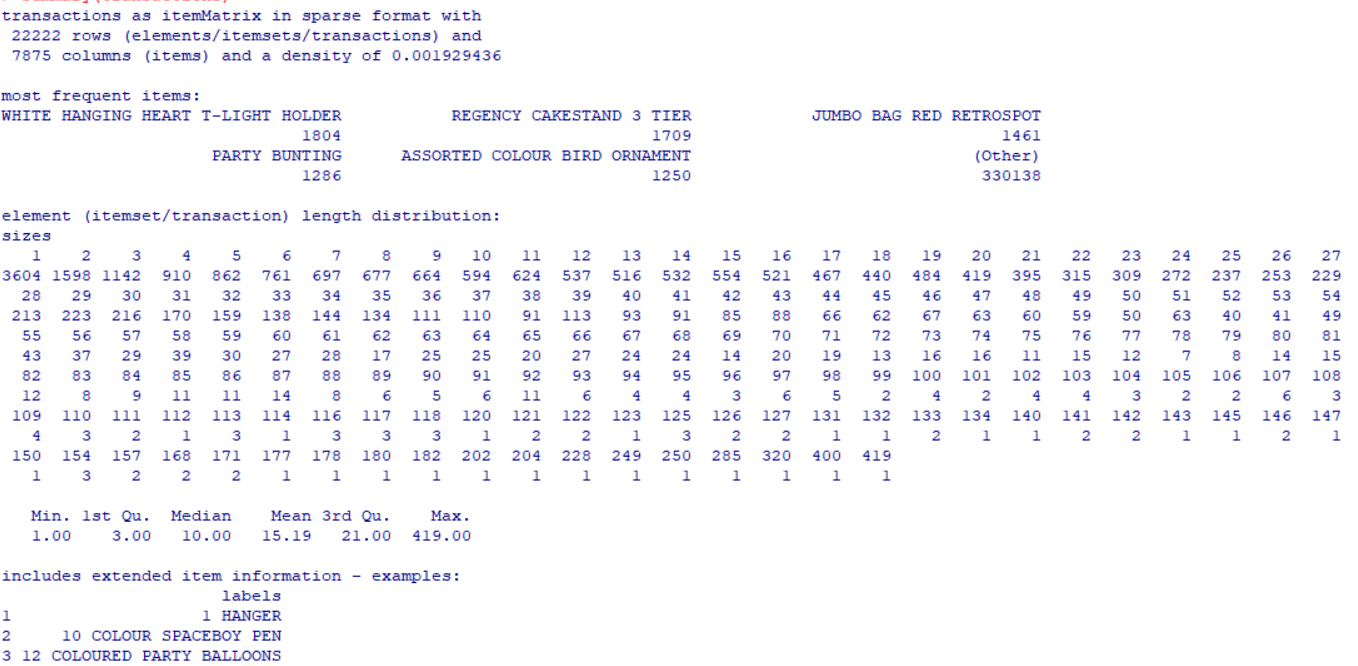


Saya melakukan Algoritma Apriori dengan Bahasa pemgrograman R untuk mengerjakan projek B. Algoritma Apriori adalah salah satu algoritma dalam data mining yang digunakan untuk menemukan pola-pola asosiasi antar-item dalam data transaksional atau data berbasis transaksi. Algoritma Apriori menggunakan pendekatan bottom-up (dari bawah ke atas) dalam mencari pola asosiasi yang paling sering muncul dalam data. Secara umum, algoritma Apriori bekerja dengan cara melakukan scan awal terhadap dataset untuk mengidentifikasi item-item yang paling sering muncul dalam transaksi. Setelah itu, algoritma ini akan membangun sebuah struktur yang disebut "itemset" yang terdiri dari kombinasi beberapa item yang sering muncul bersama dalam satu transaksi. Setelah dibangun, itemset-itemset ini akan diuji apakah memenuhi suatu batas minimum support (minimum frequency threshold). Itemset yang memenuhi batas minimum support tersebut kemudian akan dianggap sebagai itemset yang signifikan, dan dapat digunakan untuk menghasilkan aturan asosiasi.

Hal pertama yang saya lakukan adalah menganalisa Most Frequent Items, Element (itemset/transaction) length distribution, dan Extended Item information.Berikut merupakan tabel dan visualisasi bar plot untuk menjelaskan ketiga hal ini.

A picture containing text

Description automatically generated



Berdasarkan tabel diatas, terdapat berikut analisisnya:

* WHITE HANGING HEART T-LIGHT HOLDER adalah item yang paling sering muncul dalam transaksi sebanyak 1804 kali.
* Disusul oleh REGENCY CAKESTAND 3 TIER dengan frekuensi muncul 1709 kali dan JUMBO BAG RED RETROSPOT dengan frekuensi 1461 kali.
* Item yang paling jarang muncul adalah item yang termasuk dalam kategori "Other", dengan frekuensi hanya 330138 kali.
* Sebagian besar transaksi dalam data terdiri dari 1-10 item (sekitar 77% dari total transaksi).
* Rata-rata panjang transaksi adalah 15.19, dengan transaksi terpanjang terdiri dari 419 item dan transaksi terpendek terdiri dari 1 item.

Dari analisis di atas, dapat disimpulkan bahwa item-item tertentu memiliki frekuensi yang lebih tinggi dalam transaksi, sehingga item-item tersebut dapat menjadi fokus strategi pemasaran. Selain itu, sebagian besar transaksi terdiri dari jumlah item yang relatif sedikit, sehingga mungkin ada peluang untuk mempromosikan pembelian beberapa item sekaligus atau menawarkan diskon jika pembelian mencapai jumlah tertentu.

Hal kedua yang saya lakukan adalah menerapkan algoritma Apriori untuk dataset Online Retail. Berikut merupakan hasil dan penjelasannya.

Graphical user interface, text, application

Description automatically generated

A picture containing timeline

Description automatically generated

Terdapat 31.8 aturan asosiasi dengan support rata-rata 0.001431 dan confidence rata-rata 0.8854 yang dihasilkan dengan menggunakan algoritma Apriori. Aturan yang dihasilkan memiliki panjang rata-rata 5.458 itemset dan lift rata-rata 68.609. Jumlah aturan terbesar terdapat pada aturan dengan panjang 5 itemset, dan kriteria yang digunakan untuk menentukan aturan tersebut adalah confidence minimal 0.8. Selain itu, terdapat nilai maksimum untuk lift dan support yang cukup tinggi, yaitu 716.839 dan 0.015975, masing-masing.

Berikut merupakan generated rules (aturan asosiasi) yang didapatkan dari algortima apriori.

Text

Description automatically generated

Berikut merupakan penjelasan masing-masing aturan asosiasi

1. {WOBBLY CHICKEN} => {DECORATION}. Aturan ini memiliki support 0.00126, artinya 0.126% transaksi di dalam dataset mengandung item-item WOBBLY CHICKEN dan DECORATION secara bersamaan. Confidence-nya adalah 1, yang berarti setiap transaksi yang mengandung WOBBLY CHICKEN juga pasti mengandung DECORATION. Aturan ini memiliki coverage 0.00126, artinya 0.126% transaksi di dalam dataset mengandung item WOBBLY CHICKEN.
2. {WOBBLY CHICKEN} => {METAL}. Aturan ini memiliki support, confidence, dan coverage yang sama dengan aturan-asosiasi pertama. Hal ini menunjukkan bahwa item-item WOBBLY CHICKEN, DECORATION, dan METAL cenderung dibeli bersamaan.
3. {DECOUPAGE} => {GREETING CARD}. Aturan ini memiliki support 0.00104, confidence 1, dan coverage 0.00104. Hal ini menunjukkan bahwa setiap transaksi yang mengandung item DECOUPAGE juga pasti mengandung GREETING CARD.
4. {BILLBOARD FONTS DESIGN} => {WRAP}. Aturan ini memiliki support 0.00131, confidence 1, dan coverage 0.00131. Hal ini menunjukkan bahwa setiap transaksi yang mengandung item BILLBOARD FONTS DESIGN juga pasti mengandung WRAP.
5. {WOBBLY RABBIT} => {DECORATION}. Aturan ini memiliki support 0.00153, confidence 1, dan coverage 0.00153. Hal ini menunjukkan bahwa setiap transaksi yang mengandung item WOBBLY RABBIT juga pasti mengandung DECORATION.
6. {WOBBLY RABBIT} => {METAL} dengan nilai support sebesar 0.001530015, confidence sebesar 1, coverage sebesar 0.001530015, dan lift sebesar 444.44000. Artinya, dari seluruh transaksi yang terdapat dalam dataset, 0.15% transaksi membeli WOBBLY RABBIT dan METAL secara bersamaan, dan dari seluruh transaksi yang membeli WOBBLY RABBIT, 100% di antaranya juga membeli METAL.
7. {ART LIGHTS} => {FUNK MONKEY} dengan nilai support sebesar 0.001710017, confidence sebesar 1, coverage sebesar 0.001710017, dan lift sebesar 584.78947. Artinya, dari seluruh transaksi yang terdapat dalam dataset, 0.17% transaksi membeli ART LIGHTS dan FUNK MONKEY secara bersamaan, dan dari seluruh transaksi yang membeli ART LIGHTS, 100% di antaranya juga membeli FUNK MONKEY.
8. {FUNK MONKEY} => {ART LIGHTS} dengan nilai support sebesar 0.001710017, confidence sebesar 1, coverage sebesar 0.001710017, dan lift sebesar 584.78947. Artinya, dari seluruh transaksi yang terdapat dalam dataset, 0.17% transaksi membeli FUNK MONKEY dan ART LIGHTS secara bersamaan, dan dari seluruh transaksi yang membeli FUNK MONKEY, 100% di antaranya juga membeli ART LIGHTS.
9. {BLACK TEA} => {SUGAR JARS} dengan nilai support sebesar 0.002070021, confidence sebesar 1, coverage sebesar 0.002070021, dan lift sebesar 238.94624. Artinya, dari seluruh transaksi yang terdapat dalam dataset, 0.21% transaksi membeli BLACK TEA dan SUGAR JARS secara bersamaan, dan dari seluruh transaksi yang membeli BLACK TEA, 100% di antaranya juga membeli SUGAR JARS.
10. {BLACK TEA} => {COFFEE} menunjukkan bahwa setiap kali ada pembelian black tea, maka pasti juga dibeli coffee. Hal ini dibuktikan dengan nilai confidence sebesar 1, yang menandakan bahwa aturan ini selalu benar. Selain itu, nilai lift sebesar 69.44375 menunjukkan bahwa pembelian black tea dan coffee terkait secara positif, yaitu pembelian black tea akan meningkatkan kemungkinan pembelian coffee sebesar 69.44 kali dari yang diharapkan secara acak.

Berikut merupakan visualisasi tambahan.

Chart

Description automatically generatedChart, scatter chart

Description automatically generated

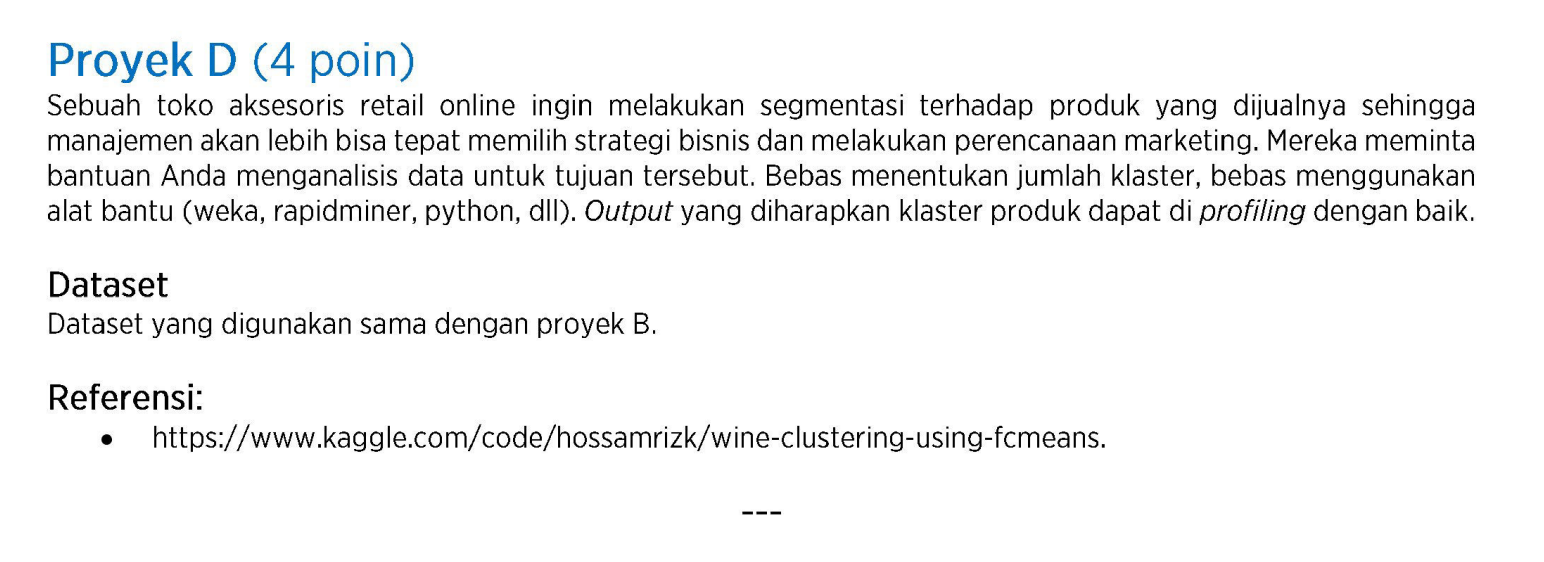
Berdasarkan data yang sudah diberikan dan dianalisa di atas, terdapat beberapa informasi yang dapat membantu dalam menganalisa perilaku konsumen dalam membeli barang yang ditawarkan oleh toko aksesoris retail online.

Pertama, terdapat beberapa produk yang sering dibeli bersama-sama oleh konsumen, seperti WOBBLY CHICKEN dan METAL, WOBBLY RABBIT dan DECORATION, serta BLACK TEA dan SUGAR JARS. Hal ini menunjukkan bahwa ada kecenderungan konsumen untuk membeli produk-produk ini secara bersamaan, sehingga dapat diatur display barang yang berdekatan untuk meningkatkan kemungkinan penjualan produk-produk tersebut secara bersamaan.

Kedua, terdapat beberapa aturan asosiasi dengan nilai confidence yang tinggi, seperti aturan {DECOUPAGE} => {GREETING CARD} dan {BILLBOARD FONTS DESIGN} => {WRAP}. Hal ini menunjukkan bahwa konsumen yang membeli produk DECOUPAGE cenderung juga membeli GREETING CARD, begitu pula dengan konsumen yang membeli produk BILLBOARD FONTS DESIGN cenderung juga membeli WRAP. Informasi ini dapat membantu dalam menentukan produk mana yang harus ditempatkan berdekatan di display barang, serta mempertimbangkan penempatan produk-produk yang memiliki hubungan erat dalam promo yang dilakukan.

Ketiga, terdapat produk yang memiliki nilai lift yang tinggi, seperti FUNK MONKEY dan ART LIGHTS, serta BLACK TEA dan COFFEE. Hal ini menunjukkan bahwa produk-produk ini memiliki keterkaitan yang tinggi dalam pembelian oleh konsumen, sehingga dapat dipertimbangkan untuk ditempatkan berdekatan di display barang dan digabungkan dalam promo-promo tertentu.

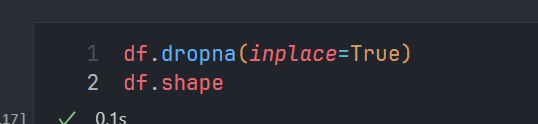
Dengan mempertimbangkan informasi-informasi tersebut, toko aksesoris retail online dapat mengatur display barang dengan baik, serta melakukan promo-promo yang lebih tepat dan efektif. Hal ini dapat meningkatkan kenyamanan konsumen dalam berbelanja dan memaksimalkan penjualan produk yang ditawarkan oleh toko.



Saya menggunakan python untuk projek D ini. Saya menggunakan metode klister k-means. Metode K-Means adalah salah satu metode dalam clustering atau pengelompokkan data yang paling populer digunakan. Metode ini digunakan untuk mengelompokkan data menjadi beberapa kelompok atau klaster berdasarkan jarak antara titik data dalam suatu kelompok dengan titik data yang lain. Algoritma K-Means bekerja dengan cara menghitung jarak antara setiap titik data dengan pusat klaster yang ditentukan, kemudian mengelompokkan titik-titik data tersebut ke dalam klaster yang paling dekat. Selanjutnya, algoritma akan memperbarui posisi pusat klaster dengan menghitung rata-rata dari seluruh titik data yang ada di dalam klaster tersebut, dan kemudian menghitung kembali jarak antara titik data dengan pusat klaster yang baru.

Proses Klastering.

Membuang data Null.



Mengubah Invoice Date menjadi datetime dan melakukan LabelEncoder dan melakukan normalisasi.

A screenshot of a computer

Description automatically generated with medium confidence

Text

Description automatically generated

Berikut merupakan hasil klastering yang didapatkan

Graphical user interface, text

Description automatically generated

Saya menggunakan 5 klaster untuk output.

Komitmen Integritas

“Di hadapan TUHAN yang hidup, saya menegaskan bahwa saya tidak memberikan maupun menerima bantuan apapun—baik lisan, tulisan, maupun elektronik—di dalam ujian ini selain daripada apa yang telah diizinkan oleh pengajar, dan tidak akan menyebarkan baik soal maupun jawaban ujian kepada pihak lain.”

