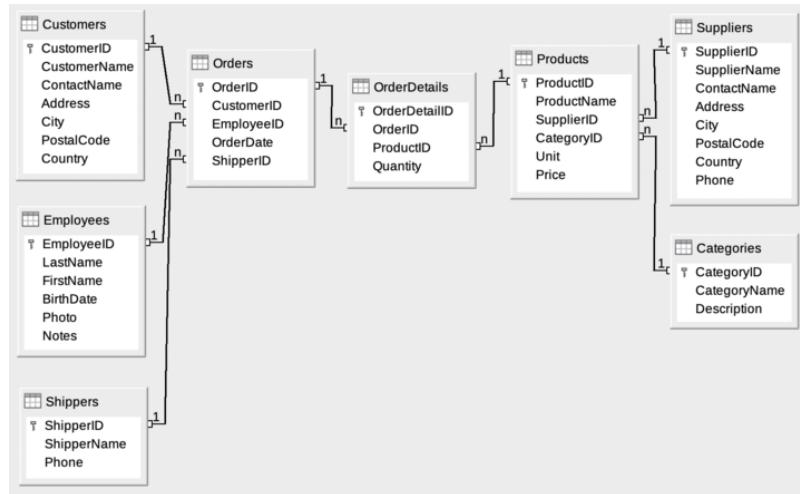


UTS IBDA 3122 Knowledge Discovery

Dibuat Oleh: Stefannus Christian 202000138 (NIM Genap)

Proyek A (16 poin)

Database Northwind adalah database sampel, yang berisi data penjualan untuk perusahaan fiktif bernama "Northwind Traders", yang mengimpor dan mengekspor makanan khusus dari seluruh dunia. Database Northwind memiliki data sampel *customers*, *orders*, *product*, *suppliers*, *shipping*, dan *employees*. Hubungan tabel ditampilkan dalam diagram E-R berikut.



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.

```
def convert_date(date: str, info):
    d,m,y = date.split('-')
    if info == 'day':
        return int(d)
    elif info == 'month':
        return dates[m]
    elif info == 'year':
        return int('19'+y)
```

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

```

1 def clean_unit_price(unit_price):
2     unit_price = unit_price[1:]
3     unit_price = unit_price.replace('$', ',')
4     return float(unit_price)
5
6 def clean_discount(discount):
7     discount = discount[:-1]
8     discount = discount.replace('%', ',')
9     return float(discount)
10

```

Hal yang dilakukan kode diatas hanya mengubah unit price dan discount menjadi float dengan slicing string sehingga 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 penjualannya adalah 64.505 yang didapatkan dari $(33.25) * 2 * 0.97 = 64.505$

Deskripsi Tugas

Manajemen Perusahaan Northwind ingin menjaga loyalitas pelanggan, memperluas pasar dan meningkatkan penjualan. Salah satu aspek yang ingin mereka ketahui adalah keadaan penjualan selama periode data (Agustus 1994 – Juni 1996) dan perkiraan penjualannya pada Juli-Desember 1996 dan sepanjang 1997 (proyeksi data).

- Kategori produk apa yang mencatatkan nilai penjualan tertinggi selama periode data? Berapakah angka proyeksi penjualannya? (NIM Genap: Juni-Desember 1996; NIM Ganjil: sepanjang 1997)
Produk apa dalam kategori itu yang mencatatkan penjualan tertinggi?

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.

Category	Order ID	Order Date Day	Order Date Month	Order Date Year	Unit Price	Quantity	Discount	Nilai Penjualan	ProductID	Units In Stock	Units On Order	Reorder Level	Discontinued
Beverages	4312144	6346	2371	806085	11811.65	9532	2500.0	267868.1800	16356	18111	2430	6715	51
Dairy Products	3894474	6016	2248	730230	9875.80	9149	1956.0	234507.2850	17806	13424	5310	3640	0
Confections	3557446	5334	1945	666404	7549.30	7906	1902.0	167357.2250	12147	7944	3580	2895	0
Meat/Poultry	1839680	2947	1028	345164	7417.33	4199	1115.0	163022.3595	6951	4696	0	1020	104
Seafood	3523066	5253	2022	658433	6290.78	7681	1988.0	131261.7375	11139	21178	1280	3790	0
Condiments	2303505	3258	1352	430965	4605.30	5298	1137.0	106047.0850	9557	10416	1640	2405	10
Produce	1450501	2147	755	271356	4786.45	2990	618.0	99984.5800	4386	2895	260	745	33
Grains/Cereals	2090139	2731	1171	391069	4164.30	4562	888.0	95744.5875	9567	7096	2900	4435	30

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 penjualan 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.

	Order ID	Order Date Day	Order Date Month	Order Date Year	Unit Price	Quantity	Discount	Nilai Penjualan	ProductID	Units In Stock	Units On Order	Reorder Level	Discontinued
Product Name													
Cte de Blaye	255583	410	120	47887	5902.4	623	110.0	141396.735	912	408	0	360	0
Ipo Coffee	298531	401	169	55866	1205.2	580	140.0	23526.700	1204	476	280	700	0
Chang	470965	696	245	87795	786.6	1057	450.0	16355.960	88	748	1760	1100	0
Lakkalikri	415841	554	260	77812	662.4	981	205.0	15760.440	2964	2223	0	780	0
Steeleye Stout	383553	552	200	71829	612.0	883	170.0	13644.000	1260	720	0	540	0

Dapat dilihat dari tabel diatas bahwa beverages dengan nilai 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.

Category	Order Date Year	Order Date Month	Order ID	Order Date Day	Unit Price	Quantity	Discount	Nilai Penjualan	ProductID	Units In Stock	Units On Order	Reorder Level	Discontinued
0 Beverages	1994	8	112857	205	134.4	272	75.0	3182.50	383	336	140	180	
1 Beverages	1994	9	164550	299	201.0	347	55.0	4866.88	656	827	20	220	
2 Beverages	1994	10	123689	215	222.2	285	10.0	5088.40	529	526	100	250	
3 Beverages	1994	11	113604	138	371.4	286	85.0	7971.36	423	433	130	215	
4 Beverages	1994	12	165680	264	764.4	347	90.0	17378.06	636	913	0	195	

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

```

1 X = beverages_products_groupby[["Order Date Month", "Order Date Year"]].values
2 y = beverages_products_groupby["Nilai Penjualan"].values
3 reg = LinearRegression().fit(X,y)
4 prediction_score = reg.score(X,y)*100

```

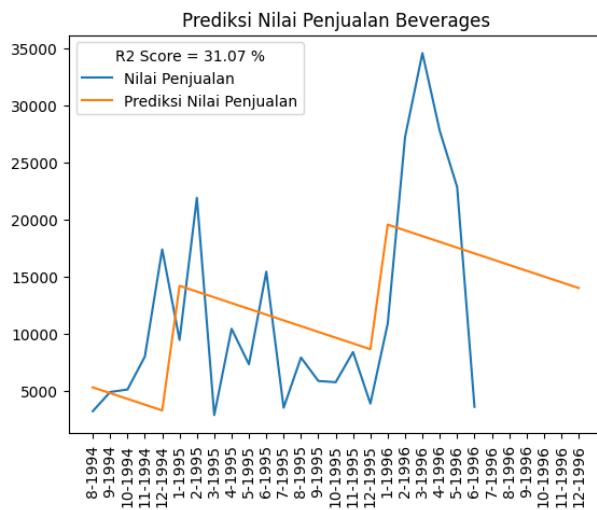
Kemudian data yang ingin di prediksi disini adalah data dari Juli 1996 – Desember 1996. Berikut merupakan kode plotting prediksi nya.

```

1 x_predict = np.array([
2     [7,1996],
3     [8,1996],
4     [9,1996],
5     [10,1996],
6     [11,1996],
7     [12,1996]
8 ])
9 x_concat = np.concatenate((X,x_predict),axis=0)
10 x_label = [f'{i}-{j}' for i,j in x_concat]
11 X_label = [f'{i}-{j}' for i,j in X]
12 prediction = reg.predict(x_concat)
13 plt.plot(X_label,y,label="Nilai Penjualan")
14 plt.plot(x_label,prediction,label="Prediksi Nilai Penjualan")
15 plt.xticks(rotation=90)
16 plt.legend()
17 plt.title("Prediksi Nilai Penjualan Beverages")
18 plt.legend(title=f'R2 Score = {round(prediction_score,2)}%')
19 plt.show()

```

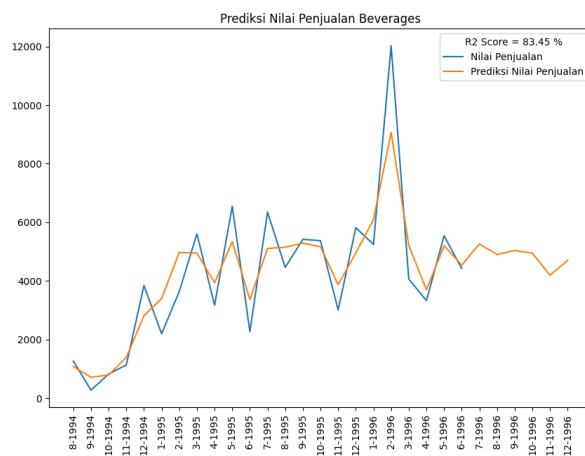
Berikut merupakan grafik proyeksi nilai penjualan Beverages untuk proyeksi menggunakan linear regression. Proyeksi nilai penjualan Beverages dari Juli 1996 – Desember 1996 adalah sebagai berikut.



	Proyeksi Nilai Penjualan	Tanggal
0	16518.334696	July 1996
1	16012.658593	August 1996
2	15506.982489	September 1996
3	15001.306385	October 1996
4	14495.630281	November 1996
5	13989.954178	December 1996

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.



	Proyeksi Nilai Penjualan	Tanggal
0	10755.57650	July 1996
1	11315.04975	August 1996
2	11298.93255	September 1996
3	11173.18100	October 1996
4	11965.21925	November 1996
5	12300.99240	December 1996

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.

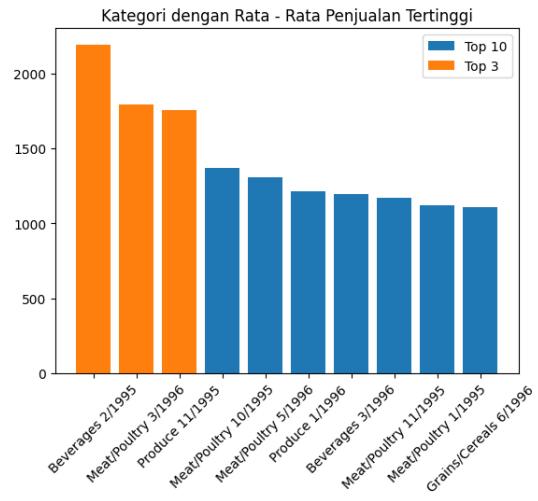
Terdapat perbedaan prediksi menggunakan 2 model yang berbeda. Dengan menggunakan model linear regression, proyeksi data cenderung menurun, tetapi menggunakan Random Forest, proyeksi data cenderung mengingkat. Hal ini terjadi karena linear regression mengasumsikan bahwa hubungan antara variabel independen 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.

2. Tiga kategori yang mencatatkan rata-rata nilai penjualan tertinggi selama periode data?

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.

```
nomor_2 = order_order_details_products.groupby(["Cat", "Month", "Year"], as_index=False).reset_index().sort_values('Nilai Penjualan', ascending=False).head(3)
```

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



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

Category	Order Date Month	Order Date Year	Nilai Penjualan
Beverages	2	1995	2190.416000
Meat/Poultry	3	1996	1794.926923
Produce	11	1995	1754.350000

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.

3. (NIM Genap) periode Januari-Juni 1996. (NIM Ganjil) periode selama tahun 1995.
 Lima produk yang mencatatkan rata-rata nilai penjualan tertinggi selama periode tersebut?
 Lima produk yang mencatatkan rata-rata nilai penjualan terendah selama periode 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.

```
1 order_order_details_products_jan_june_96 = order_order_details_products[(order_order_details_products
                           ["Order Date Year"] == 1996) & (order_order_details_products["Order Date Month"] < 7)]
2 order_order_details_products_jan_june_96.groupby(["Product Name"]).mean().sort_values('Nilai Penjualan',
                           ascending=False).head()
3 0.1s
```

Python

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

Product Name	Order ID	Order Date Day	Order Date Month	Order Date Year	Unit Price	Quantity	Discount	Nilai Penjualan	ProductID	Units In Stock	Units On Order	Reorder Le
Cte de Blaye	10873.727273	15.545455	2.636364	1996.0	263.50	25.000000	0.909091	6479.704545	38.0	17.0	0.0	1
Thringer Rostbratwurst	10896.500000	17.785714	2.928571	1996.0	123.79	27.857143	11.785714	2983.339000	29.0	0.0	0.0	
Raclette Courdavault	10919.062500	16.187500	3.312500	1996.0	55.00	36.125000	3.437500	1867.421875	59.0	79.0	0.0	
Sir Rodney's Marmalade	10952.857143	13.142857	4.000000	1996.0	81.00	15.857143	2.000000	1243.465714	20.0	40.0	0.0	
Schoggi Schokolade	10881.750000	15.000000	2.750000	1996.0	43.90	26.250000	0.000000	1152.375000	27.0	49.0	0.0	

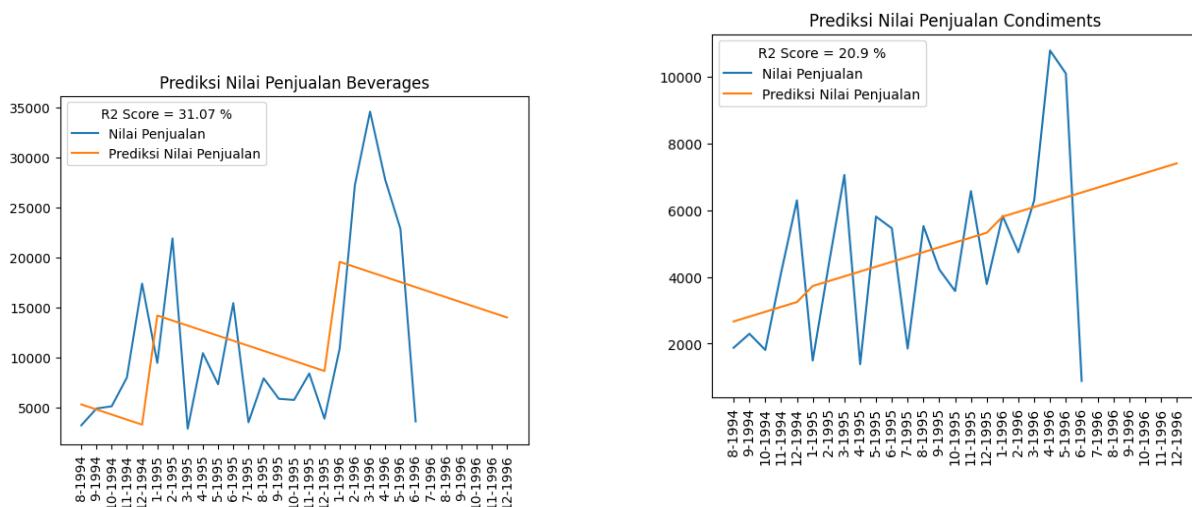
5 produk yang mencatatkan rata – rata nilai penjualan terendah selama periode Januari – Juni 1996 dapat dilihat pada tabel dibawah ini.

	Order ID	Order Date Day	Order Date Month	Order Date Year	Unit Price	Quantity	Discount	Nilai Penjualan	ProductID	Units In Stock	Units On Order	Reorder Level
Product Name												
Rd Kaviar	10979.000000	9.000000	4.500000	1996.0	15.00	6.000000	0.500000	89.850000	73.0	101.0	0.0	5.0
Tourtire	10880.333333	23.555556	2.444444	1996.0	7.45	14.000000	12.777778	86.875278	54.0	21.0	0.0	10.0
Chocolade	10814.000000	5.000000	2.000000	1996.0	12.75	8.000000	15.000000	86.700000	48.0	15.0	70.0	25.0
Filo Mix	10926.076923	11.153846	3.615385	1996.0	7.00	12.461538	5.769231	78.319231	52.0	38.0	0.0	25.0
Geitost	10940.000000	19.000000	3.600000	1996.0	2.50	20.200000	4.500000	47.662500	33.0	112.0	0.0	20.0

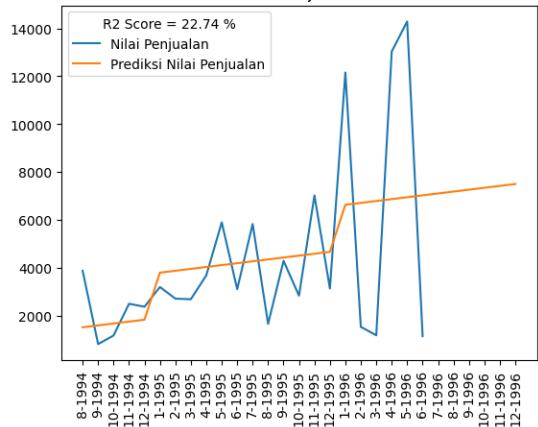
Dari informasi ini, dapat disimpulkan bahwa Cote De Blaye adalah produk yang sangat sukses dalam hal penjualan, baik dalam hal nilai penjualan maupun rata-rata penjualan. Kategori Beverages, di mana produk ini termasuk, juga terbukti sebagai kategori dengan nilai penjualan tertinggi dan rata-rata penjualan 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.

4. Pada proyeksi data, kategori produk mana saja yang memiliki kecenderungan penjualan rata-rata bulanan naik dan mana yang cenderung menurun?

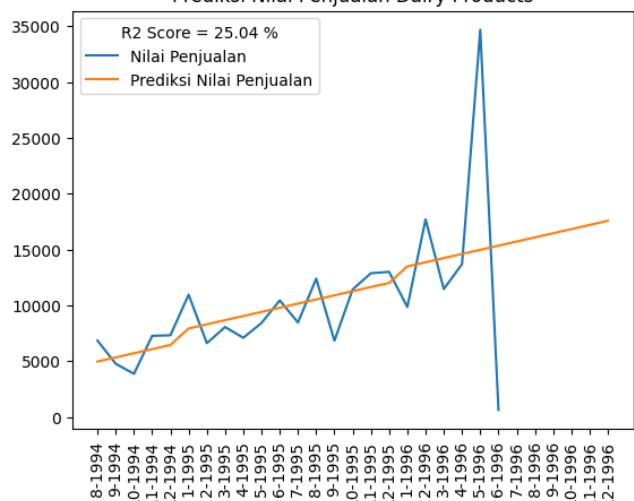
Untuk menjawab pertanyaan ini, hal yang saya lakukan sama persis untuk menjawab proyeksi data pada nomor 1. Bedanya hanyalah saya melakukan hal ini untuk semua kategori. Berikut merupakan hasil visualisasinya menggunakan model Linear Regression.



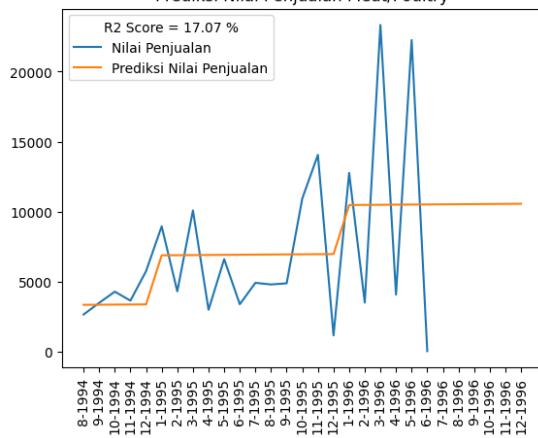
Prediksi Nilai Penjualan Produce



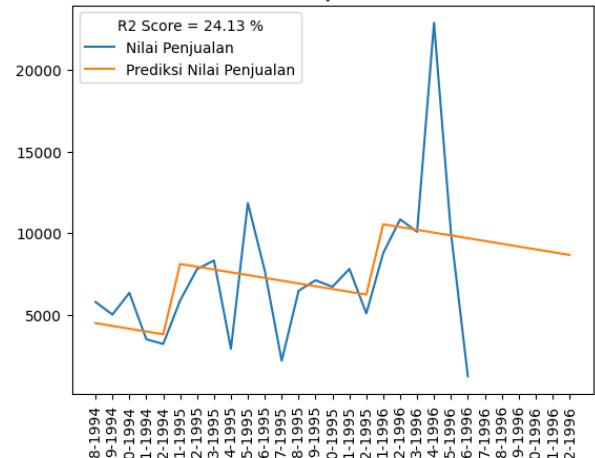
Prediksi Nilai Penjualan Dairy Products



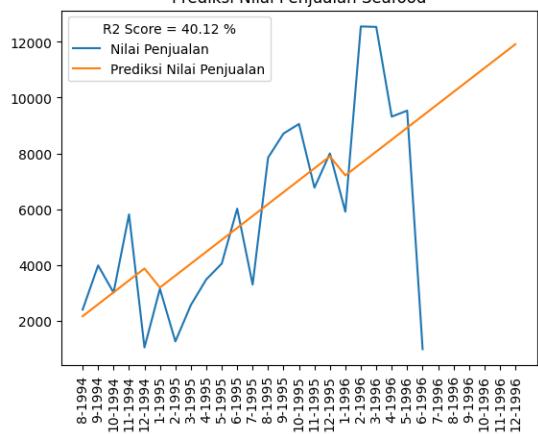
Prediksi Nilai Penjualan Meat/Poultry



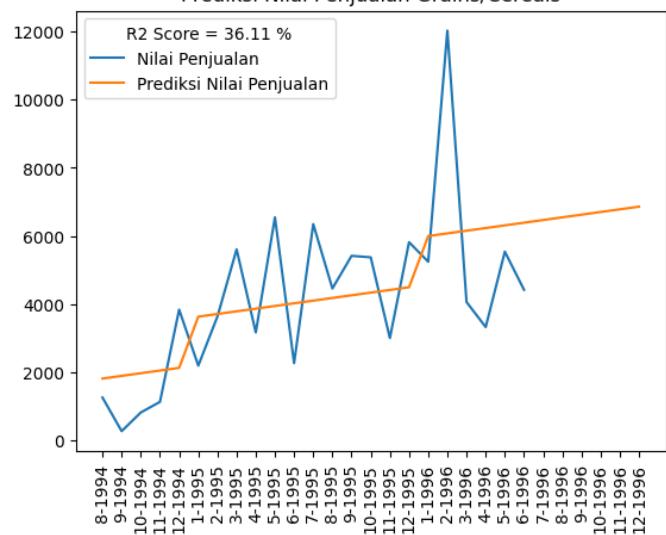
Prediksi Nilai Penjualan Confections



Prediksi Nilai Penjualan Seafood



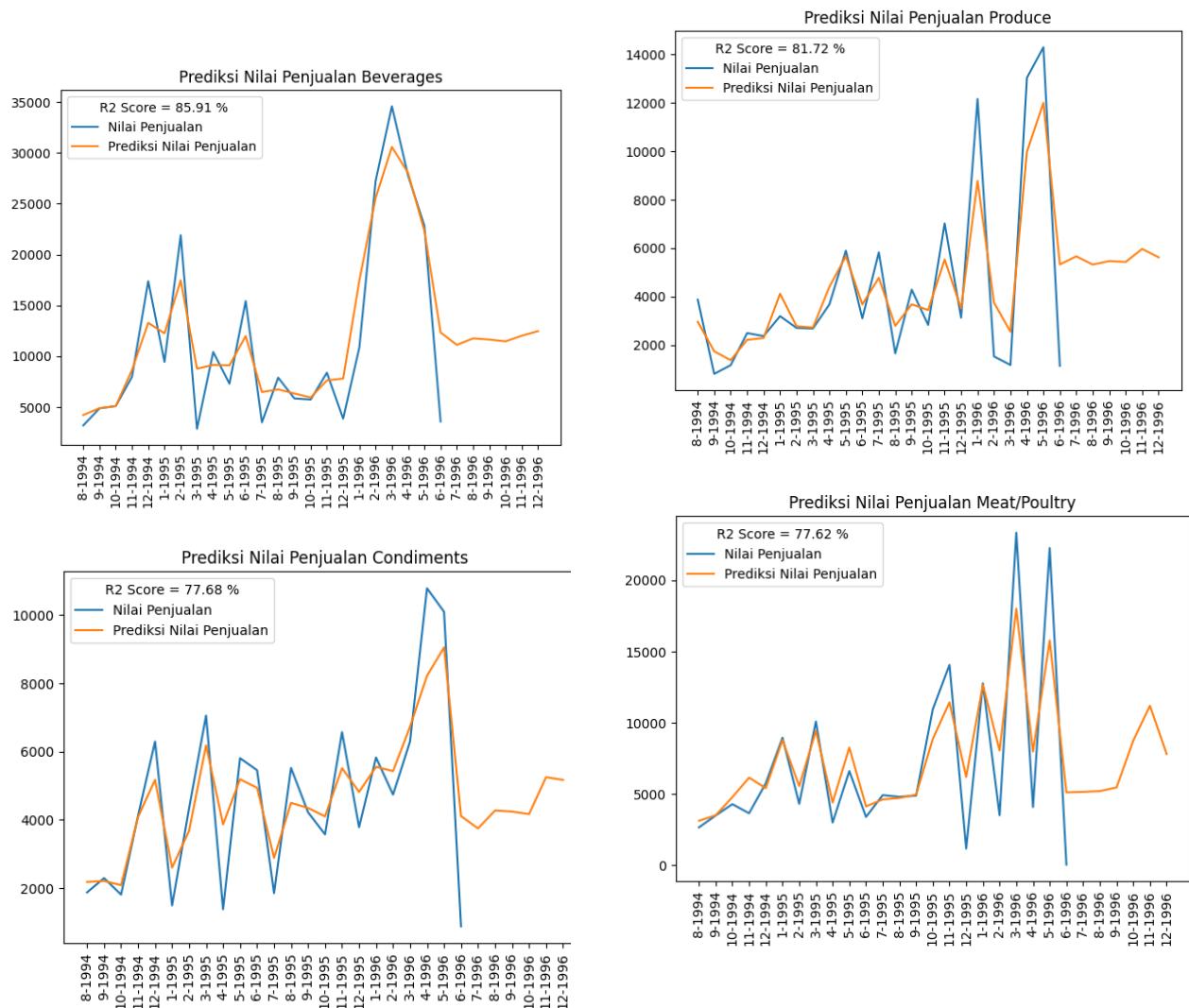
Prediksi Nilai Penjualan Grains/Cereals

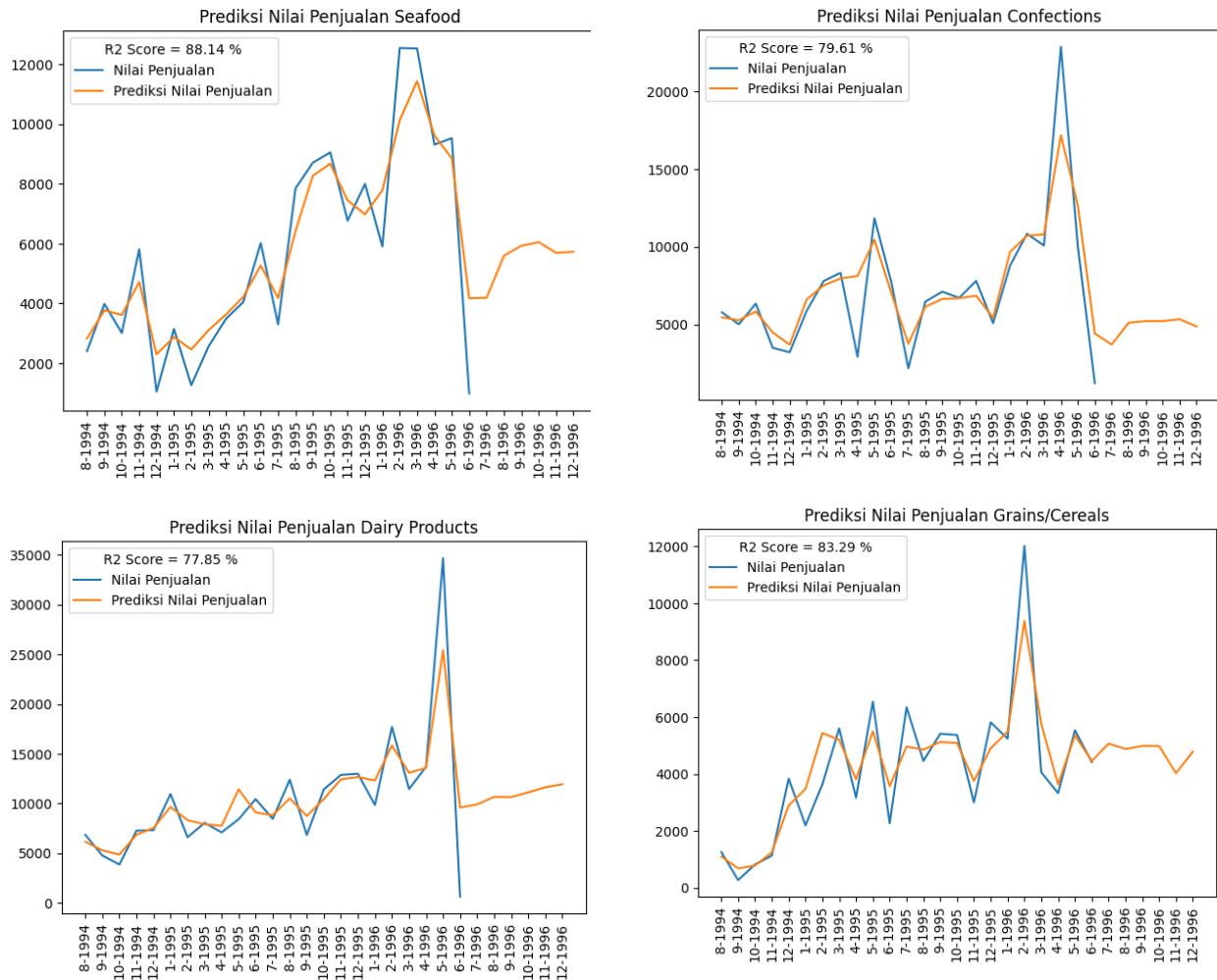


Informasi yang bisa didapatkan dari grafik diatas adalah sebagai berikut:

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.





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.

5. (NIM Genap) Wilayah (*country*) mana yang paling banyak melakukan order (*count*) dan paling tinggi nilai order (*sum*) selama periode data?
- (NIM Ganjil) Wilayah (*country*) mana yang memiliki trend order (*count*) meningkat dan menurun selama periode data?

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.

```
1 order_order_details_customers = order_order_details.merge(customers,
  left_on="Customer", right_on="Company Name")
2 order_order_details_customers.groupby(["Country"]).count().sort_values('Nilai Penjualan', ascending=False)
```

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.

	Order ID	Customer	Employee	Order Date	Required Date	Shipped Date	Ship Via	Freight	Ship Name	Ship Address	...	Customer ID	Company Name	Contact Name
Country														
USA	352	352	352	352	352	325	352	352	352	352	...	352	352	35
Germany	328	328	328	328	328	321	328	328	328	328	...	328	328	32
Brazil	203	203	203	203	203	197	203	203	203	203	...	203	203	20
France	182	182	182	182	182	178	182	182	182	182	...	182	182	18
UK	135	135	135	135	135	135	135	135	135	135	...	135	135	13

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.

	Order ID	Order Date Day	Order Date Month	Order Date Year	Unit Price	Quantity	Discount	Nilai Penjualan
Country								
USA	3764864	5600	2339	702326	10462.91	9330	2094.0	245584.6105
Germany	3487000	5232	2143	654398	8544.84	9213	2070.0	230284.6335
Austria	1332339	1954	680	249408	3469.95	5167	860.0	128003.8385
Brazil	2162294	3342	1149	405033	5324.64	4247	1350.0	106925.7765
France	1935829	2907	1075	363124	4805.81	3224	1005.0	80837.9125

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.

6. Apakah ada kaitan/relasi antara kategori produk (8 kategori) dengan asal customer (21 country).

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)

```
1 import numpy as np
2 ordinalEncoder = OrdinalEncoder()
3
4 country_df = np.array(categories_products_order_details_order_customers["Country"]).reshape(-1,1)
5 country_ordinal_encoder = ordinalEncoder.fit_transform(country_df)
6
7 categories_df = np.array(categories_products_order_details_order_customers["Category"]).reshape(-1,1)
8 categories_ordinal_encoder = ordinalEncoder.fit_transform(categories_df)
✓ 0.8s
```



```
1 categories_products_order_details_order_customers["Country Ordinal"] = country_ordinal_encoder
2 categories_products_order_details_order_customers["Category Ordinal"] = categories_ordinal_encoder
3 categories_products_order_details_order_customers["Category Ordinal"].corr
(categories_products_order_details_order_customers["Country Ordinal"])
✓ 0.9s
```

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 outputnya 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.

```
1 # Create a contingency table of Product Category and Customer Country
2 cont_table = pd.crosstab(categories_products_order_details_order_customers['Category'], -
   categories_products_order_details_order_customers['Country'])
3
4 # Apply the chi-squared test to the contingency table
5 chi2, pval, dof, expected = stats.chi2_contingency(cont_table)
6
7 # Print the results
8 print('Chi-squared statistic:', chi2)
9 print('P-value:', pval)
✓ 0.1s

Chi-squared statistic: 145.27062658850997
P-value: 0.36285296961683283
```

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.

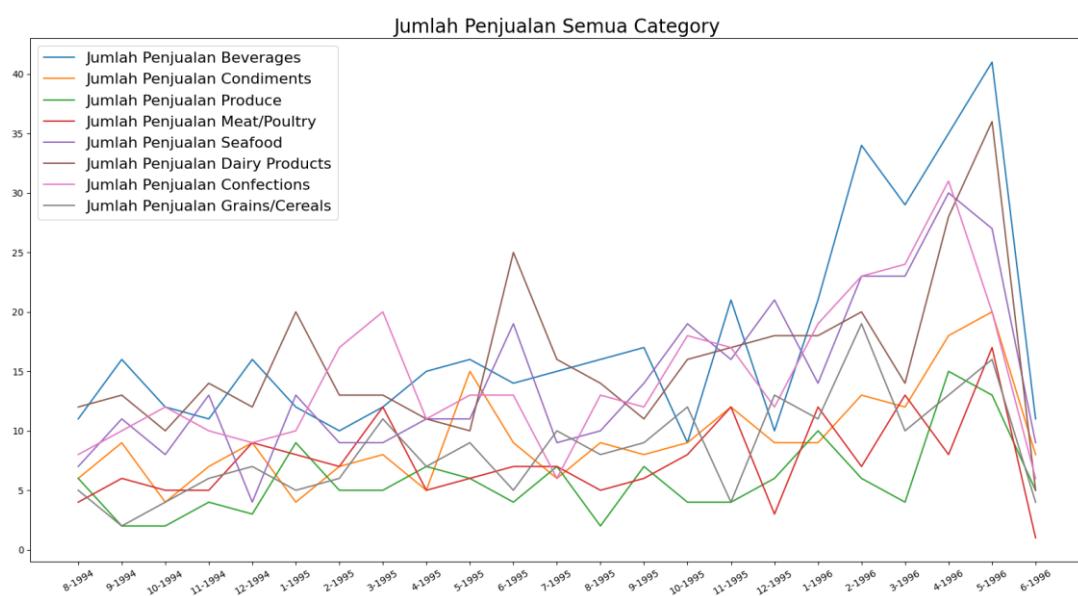
Namun, berdasarkan hasil analisis secara manual terhadap setiap negara, terlihat beberapa negara memiliki pola pembelian tertentu terkait kategori produk. Sebagai

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.

7. Apakah ada pola *order* kategori produk sepanjang periode data?

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.



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.
-
8. Kelompok wilayah (*country*) asal *customer* menjadi 5 kelompok dengan parameter 8 kategori produk berdasarkan nilai total *order* tiap kategori.

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.

```
1 categories_products_order_details_order_customers.groupby(["Country"]).sum().sort_values("Nilai Penjualan", ascending=False).head()
✓ 0.1s
```

Python

Category ID	Picture	ProductID	Unit Price_x	Units In Stock	Units On Order	Reorder Level	Discontinued	Order ID	Order Date Day	Order Date Month	Order Date Year	Unit Price
Country												
USA	1470	0.0	14144	11071.86	13860	2710	4050	39	3764864	5600	2339	702326
Germany	1342	0.0	13409	9003.81	13357	2810	3900	26	3487000	5232	2143	654398
Austria	498	0.0	4832	3771.13	4482	1620	1490	20	1332339	1954	680	249408
Brazil	821	0.0	8053	5690.69	8608	1230	2365	19	2162294	3342	1149	405033
France	788	0.0	7429	5200.05	7851	1190	2230	16	1935829	2907	1075	363124

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.



```
1 # Normalisasi data dengan StandardScaler
2 scaler = StandardScaler()
3
4 normalized_df = pd.DataFrame(scaler.fit_transform(categories_products_order_details_order_customers
[[['Nilai Penjualan']])))
5
6 # Melakukan k-means clustering dengan jumlah kluster = 5
7 kmeans = KMeans(n_clusters=5, random_state=0).fit(normalized_df)
8
9 # Menambahkan kolom kluster pada DataFrame
10 categories_products_order_details_order_customers['Cluster'] = kmeans.fit_predict(normalized_df)
11 categories_products_order_details_order_customers.groupby(['Cluster', "Country"]).sum().reset_index()
[1] ✓ 0.3s
```

Berikut merupakan hasil clusteringnya.

Index Country Cluster

0	Argentina	0	12	Norway	0	24	Germany	1
1	Austria	0	13	Poland	0	25	Ireland	1
2	Belgium	0	14	Portugal	0	26	Sweden	1
3	Brazil	0	15	Spain	0	27	Switzerland	1
4	Canada	0	16	Sweden	0	28	USA	1
5	Denmark	0	17	Switzerland	0	29	Argentina	2
6	Finland	0	18	UK	0	30	Austria	2
7	France	0	19	USA	0	31	Belgium	2
8	Germany	0	20	Venezuela	0	32	Brazil	2
9	Ireland	0	21	Austria	1	33	Canada	2
10	Italy	0	22	Brazil	1	34	Denmark	2
11	Mexico	0	23	Canada	1	35	Finland	2

36	France	2	47	Venezuela	2	58	France	4
37	Germany	2	48	Brazil	3	59	Germany	4
38	Ireland	2	49	Denmark	3	60	Ireland	4
39	Italy	2	50	Germany	3	61	Mexico	4
40	Mexico	2	51	Ireland	3	62	Norway	4
41	Portugal	2	52	USA	3	63	Spain	4
42	Spain	2	53	Austria	4	64	Sweden	4
43	Sweden	2	54	Belgium	4	65	Switzerland	4
44	Switzerland	2	55	Brazil	4	66	UK	4
45	UK	2	56	Canada	4	67	USA	4
46	USA	2	57	Denmark	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.

Proyek B (4 poin)

Sebuah toko aksesoris retail online ingin menganalisa perilaku konsumen dalam membeli barang yang mereka tawarkan, sehingga mereka dapat mengatur display barang dengan baik untuk kenyamanan konsumen dan mengatur promo dengan lebih tepat. Mereka meminta bantuan Anda menganalisis data untuk tujuan tersebut. Bebas menggunakan alat bantu (excel, rapidminer, python, dll).

Dataset

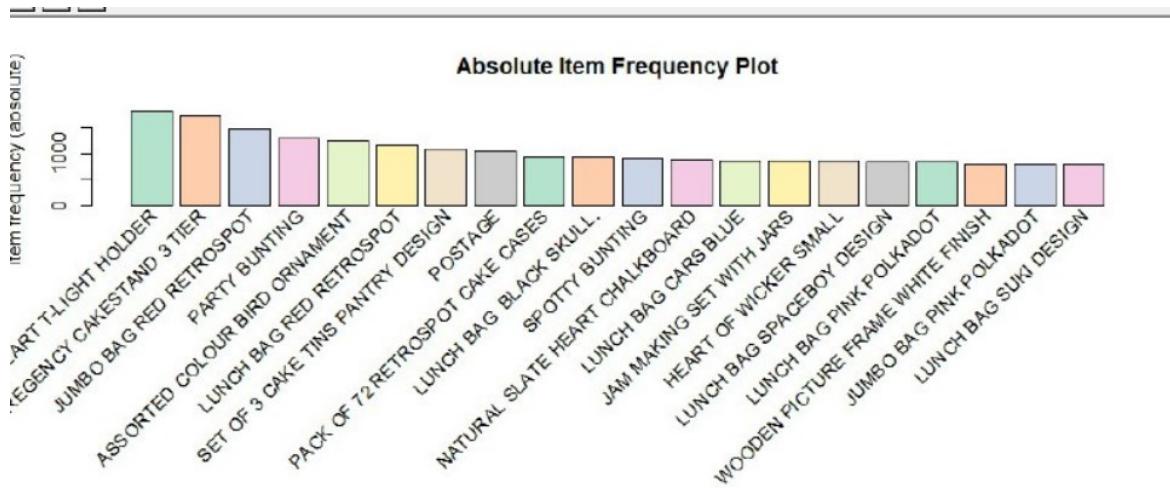
Dataset terkait dapat Anda unduh melalui canvas

Referensi

- <https://www.datacamp.com/tutorial/market-basket-analysis-r>
- <https://gifadn.medium.com/market-basket-analisis-mba-dengan-menggunakan-datasets-groceries-di-r-62f63f0278c4>
- <https://www.kaggle.com/code/aslanahmedov/market-basket-analysis-apriori-algorithm>

Saya melakukan Algoritma Apriori dengan Bahasa pemrograman 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.



```

transactions as itemMatrix in sparse format with
2222 rows (elements/itemsets/transactions) and
7875 columns (items) and a density of 0.001929436

most frequent items:
WHITE HANGING HEART T-LIGHT HOLDER           REGENCY CAKESTAND 3 TIER           JUMBO BAG RED RETROSPOT
1804                                         1709                                         1461
PARTY BUNTING          ASSORTED COLOUR BIRD ORNAMENT
1286                                         1250                                         (Other)
                                               
element (itemset/transaction) length distribution:
sizes
 1   2   3   4   5   6   7   8   9   10  11  12  13  14  15  16  17  18  19  20  21  22  23  24  25  26  27
3604 1598 1142 910 862 761 697 677 664 594 624 537 516 532 554 521 467 440 484 419 395 315 309 272 237 253 229
 28  29  30  31  32  33  34  35  36  37  38  39  40  41  42  43  44  45  46  47  48  49  50  51  52  53  54
213  223  216  170  159  138  144  134  111  110  91  113  93  91  85  88  66  62  67  63  60  59  50  63  40  41  49
 55  56  57  58  59  60  61  62  63  64  65  66  67  68  69  70  71  72  73  74  75  76  77  78  79  80  81
 43  37  29  39  30  27  28  17  25  25  20  27  24  24  14  20  19  13  16  16  11  15  12  7  8  14  15
 82  83  84  85  86  87  88  89  90  91  92  93  94  95  96  97  98  99  100 101 102 103 104 105 106 107 108
 12  8   9   11  11  14  8   6   5   6   11  6   4   4   3   6   5   2   4   2   4   4   3   2   2   6   3
109  110 111 112 113 114 116 117 118 120 121 122 123 125 126 127 131 132 133 134 140 141 142 143 145 146 147
 4   3   2   1   3   1   3   3   3   1   2   2   1   3   2   2   1   1   1   2   1   1   1   1   1   1
150  154 157 168 171 177 178 180 182 202 204 228 249 250 285 320 400 419
 1   3   2   2   2   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
Min. 1st Qu. Median Mean 3rd Qu. Max.
 1.00  3.00 10.00 15.19 21.00 419.00
includes extended item information - examples:
      labels
1      1 HANGER
2      10 COLOUR SPACEBOY PEN
3      12 COLOURED PARTY BALLOONS

```

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.

```

Apriori

Parameter specification:
confidence minval smax arem  aval originalSupport maxtime support minlen maxlen target ext
        0.8    0.1     1 none FALSE           TRUE       5  0.001     1    10 rules TRUE

Algorithmic control:
filter tree heap memopt load sort verbose
  0.1 TRUE TRUE FALSE TRUE     2    TRUE

Absolute minimum support count: 22

```

```

rule length distribution (lhs + rhs):sizes
  2   3   4   5   6   7   8   9   10
105 2092 6680 15176 12673 5120 1871 604 119

Min. 1st Qu. Median Mean 3rd Qu. Max.
2.000 5.000 5.000 5.458 6.000 10.000

summary of quality measures:
      support      confidence      coverage      lift      count
Min. :0.001035  Min. :0.8000  Min. :0.001035  Min. : 9.854  Min. : 23.0
1st Qu.:0.001125 1st Qu.:0.8333 1st Qu.:0.001260 1st Qu.:22.505 1st Qu.: 25.0
Median :0.001260 Median :0.8788 Median :0.001440 Median :29.623 Median : 28.0
Mean   :0.001431 Mean   :0.8854 Mean   :0.001623 Mean   :68.609 Mean   : 31.8
3rd Qu.:0.001530 3rd Qu.:0.9286 3rd Qu.:0.001755 3rd Qu.:77.344 3rd Qu.: 34.0
Max.   :0.015975 Max.   :1.0000  Max.   :0.019080 Max.   :716.839 Max.   :355.0

```

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 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.

	lhs	rhs	support	confidence	coverage	lift	count
[1]	{WOBBLY CHICKEN}	=> {DECORATION}	0.001260013	1	0.001260013	444.44000	28
[2]	{WOBBLY CHICKEN}	=> {METAL}	0.001260013	1	0.001260013	444.44000	28
[3]	{DECOUPAGE}	=> {GREETING CARD}	0.001035010	1	0.001035010	389.85965	23
[4]	{BILLBOARD FONTS DESIGN}	=> {WRAP}	0.001305013	1	0.001305013	716.83871	29
[5]	{WOBBLY RABBIT}	=> {DECORATION}	0.001530015	1	0.001530015	444.44000	34
[6]	{WOBBLY RABBIT}	=> {METAL}	0.001530015	1	0.001530015	444.44000	34
[7]	{ART LIGHTS}	=> {FUNK MONKEY}	0.001710017	1	0.001710017	584.78947	38
[8]	{FUNK MONKEY}	=> {ART LIGHTS}	0.001710017	1	0.001710017	584.78947	38
[9]	{BLACK TEA}	=> {SUGAR JARS}	0.002070021	1	0.002070021	238.94624	46
[10]	{BLACK TEA}	=> {COFFEE}	0.002070021	1	0.002070021	69.44375	46

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

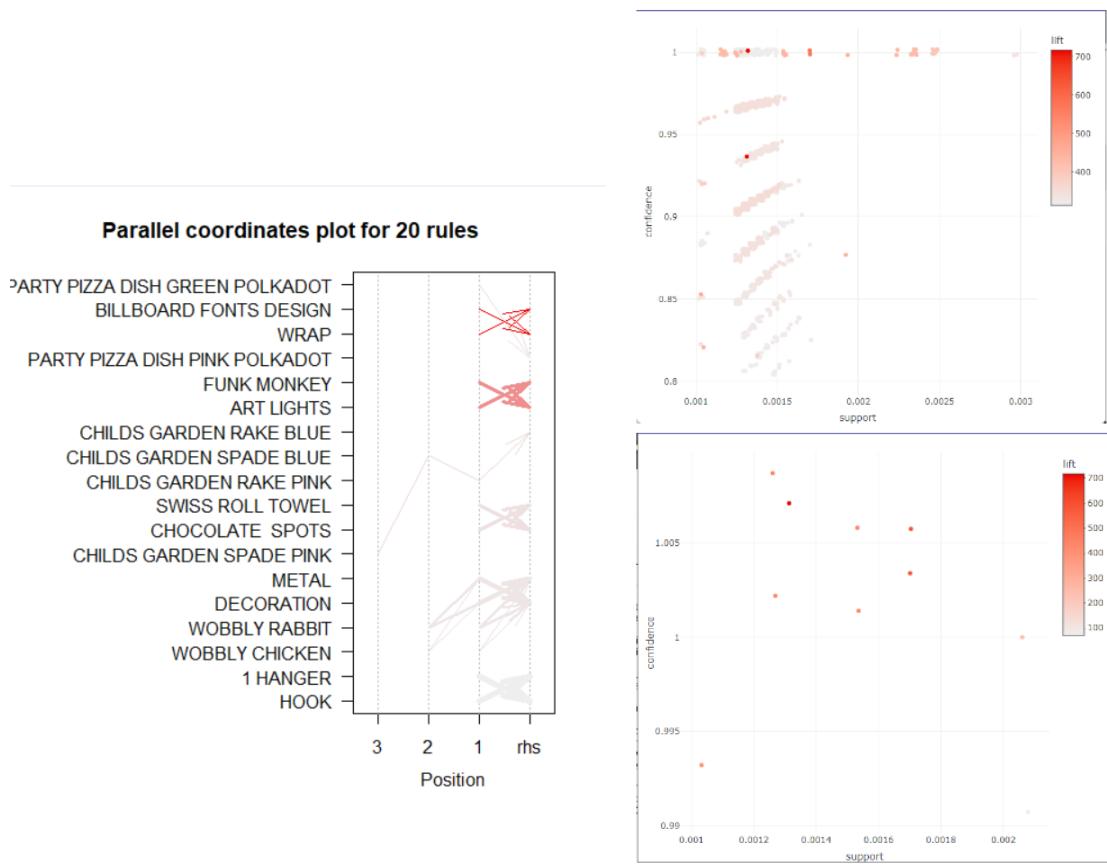
transaksi yang mengandung item DECOUPAGE juga pasti mengandung GREETING CARD.

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.



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 {DECOPAGE} => {GREETING CARD} dan {BILLBOARD FONTS DESIGN} => {WRAP}. Hal ini menunjukkan bahwa konsumen yang membeli produk DECOPAGE 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.

Proyek D (4 poin)

Sebuah toko aksesoris retail online ingin melakukan segmentasi terhadap produk yang dijualnya sehingga manajemen akan lebih bisa tepat memilih strategi bisnis dan melakukan perencanaan marketing. Mereka meminta bantuan Anda menganalisis data untuk tujuan tersebut. Bebas menentukan jumlah klaster, bebas menggunakan alat bantu (weka, rapidminer, python, dll). *Output* yang diharapkan klaster produk dapat di *profiling* dengan baik.

Dataset

Dataset yang digunakan sama dengan proyek B.

Referensi:

- <https://www.kaggle.com/code/hossamrizk/wine-clustering-using-fcmeans>.

Saya menggunakan python untuk projek D ini. Saya menggunakan metode kluster k-means. Metode K-Means adalah salah satu metode dalam clustering atau pengelompokan 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 Klustering.

Membuang data Null.

```
1 df.dropna(inplace=True)
2 df.shape
171 ✓ 0.1s
```

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

```
> ~
1 df['InvoiceDate'] = pd.to_datetime(df['InvoiceDate'])
118] ✓ 0.6s

1 from sklearn.preprocessing import LabelEncoder
2
3 label_encoder = LabelEncoder()
4
5 df['StockCode'] = df['StockCode'].astype(str)
6 encoded_stock_code = pd.DataFrame(label_encoder.fit_transform(df['StockCode']))
7 encoded_description = pd.DataFrame(label_encoder.fit_transform(df['Description']))
119] ✓ 0.6s

1 from sklearn.preprocessing import MinMaxScaler
2
3 scaler = MinMaxScaler()
4
5 df['Quantity'] = scaler.fit_transform(df[['Quantity']])
6 df['UnitPrice'] = scaler.fit_transform(df[['UnitPrice']])
120] ✓ 0.3s

1 from sklearn.cluster import KMeans
2 kmeans = KMeans(n_clusters=5, random_state=0).fit(X_baru)
3 labels = kmeans.labels_
4 X_baru['Cluster'] = labels
```

Berikut merupakan hasil klustering yang didapatkan

InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country	Cluster
77373	542720	21231 SWEETHEART CERAMIC TRINKET BOX	0.500081	2011-01-31 14:35:00	0.000729	12778.0	Netherlands	0
86879	543610	21212 PACK OF 72 RETROSPOT CAKE CASES	0.500162	2011-02-10 14:30:00	0.000321	16425.0	United Kingdom	0
31230	538907	21843 RED RETROSPOT CAKE STAND	0.500040	2010-12-15 10:40:00	0.006382	15373.0	United Kingdom	0
31231	538907	22179 SET 10 LIGHTS NIGHT OWL	0.500027	2010-12-15 10:40:00	0.003934	15373.0	United Kingdom	0
86873	543610	16161P WRAP ENGLISH ROSE	0.500168	2011-02-10 14:30:00	0.000245	16425.0	United Kingdom	0
...
75722	542608	21870 I CAN ONLY PLEASE ONE PERSON MUG	0.500013	2011-01-30 13:51:00	0.000729	16770.0	United Kingdom	4
75721	542608	21873 IF YOU CAN'T STAND THE HEAT MUG	0.500007	2011-01-30 13:51:00	0.000729	16770.0	United Kingdom	4
26454	538513	22585 PACK OF 6 BIRDY GIFT TAGS	0.500013	2010-12-12 14:21:00	0.000729	15454.0	United Kingdom	4
26456	538513	22241 GARLAND WOODEN HAPPY EASTER	0.500027	2010-12-12 14:21:00	0.000729	15454.0	United Kingdom	4
50944	540568	22414 DOORMAT NEIGHBOURHOOD WITCH	0.500007	2011-01-10 11:22:00	0.004633	15039.0	United Kingdom	4

65085 rows × 9 columns

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.”

