Laporan Ujian Tengah Semester IBDA3122 Knowledge Discovery

PROYEK A

Saya mengerjakan **PROYEK A** menggunakan **python** dengan *environment* **Anaconda** (python 3.9.13) dan beberapa *Library* dari python, yaitu:

1. pandasql
2. numpy
3. pandas
4. matplotlib
5. sklearn
6. seaborn

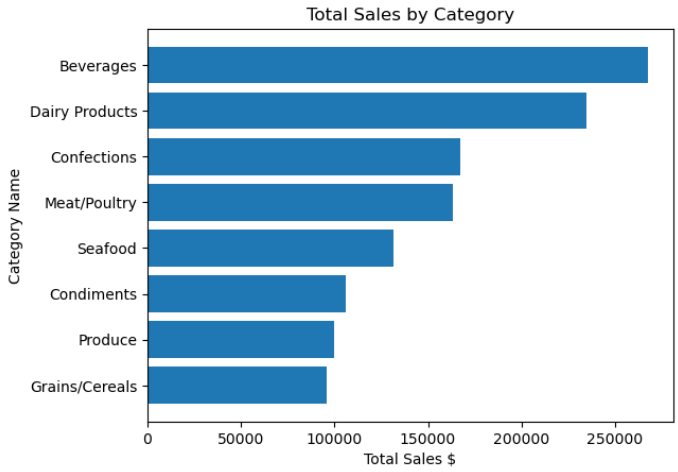
Karena dataset yang diberikan adalah excel dan dataset terbagi menjadi 8 tabel/file, maka saya mengerjakannya menggunakan SQL. Karena data yang diberikan terbagi-bagi dan bersifat database, sehingga sangat cocok menggunakan SQL.

1. **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?**
   a. Kategori produk yang mencatat nilai penjualan tertinggi selama periode data adalah 'Beverages'.
   Dengan hasil *query* dan grafik data sebagai berikut:

	Category Name	Total Penjualan	Total Sales
0	Beverages	9532	267868.1800
1	Dairy Products	9149	234507.2850
2	Confections	7906	167357.2250
3	Meat/Poultry	4199	163022.3595
4	Seafood	7681	131261.7375
5	Condiments	5298	106047.0850
6	Produce	2990	99984.5800
7	Grains/Cereals	4562	95744.5875

```
Kategori produk dengan nilai penjualan tertinggi      = Beverages
Dengan total penjualan dan total sales berturut-turut = 9532 produk - $267,868.18
```

Dengan total penjualan 9532 produk dan total sales $267,868.18
Grafik:



b. Berapakah angka proyeksi penjualannya? (NIM Genap: Juli-Desember 1996)
Pertama saya melakukan query SQL dengan syntax berikut:

```
_query = """
    SELECT
            products.`Category Name`,
            strftime('%Y-%m', orders.`Order Date`) AS 'Year-Month',
            order_details.`Quantity` AS 'Total Penjualan',
            SUM(order_details.`Unit Price` * order_details.`Quantity` * (1 - order_details.`Discount`)) AS `Total Sales`
    FROM products
    INNER JOIN order_details ON order_details.`Product Name` = products.`Product Name`
    INNER JOIN orders ON orders.`Order ID` = order_details.`Order ID`
    WHERE products.`Category Name` = 'Beverages'
    GROUP BY `Year-Month`
    ORDER BY `Year-Month` ASC;
"""
```

	Category Name	Year-Month	Total Penjualan	Total Sales
0	Beverages	1994-08	20	3182.500
1	Beverages	1994-09	45	4866.880
2	Beverages	1994-10	20	5088.400
3	Beverages	1994-11	25	7971.360

Setelah query di atas, saya memisahkan year dan month menjadi kolom baru dan menghilangkan kolom 'Total Penjualan', karena yang akan di prediksi adalah 'Total Sales'. Menjadi seperti di bawah ini.

	Category Name	Month	Year	Total Sales
0	Beverages	08	1994	3182.500
1	Beverages	09	1994	4866.880
2	Beverages	10	1994	5088.400
3	Beverages	11	1994	7971.360

Setelah itu, saya melakukan modeling. Di sini saya menggunakan Random Forest Regressor sebagai model karena setelah saya mencoba model yang lain, seperti Linear Regression dan hasilnya underfit. Sehingga model yang cocok adalah RFR.

```
X_test = np.array([
    [7, 1996],
    [8, 1996],
    [9, 1996],
    [10, 1996],
    [11, 1996],
    [12, 1996],
])
```

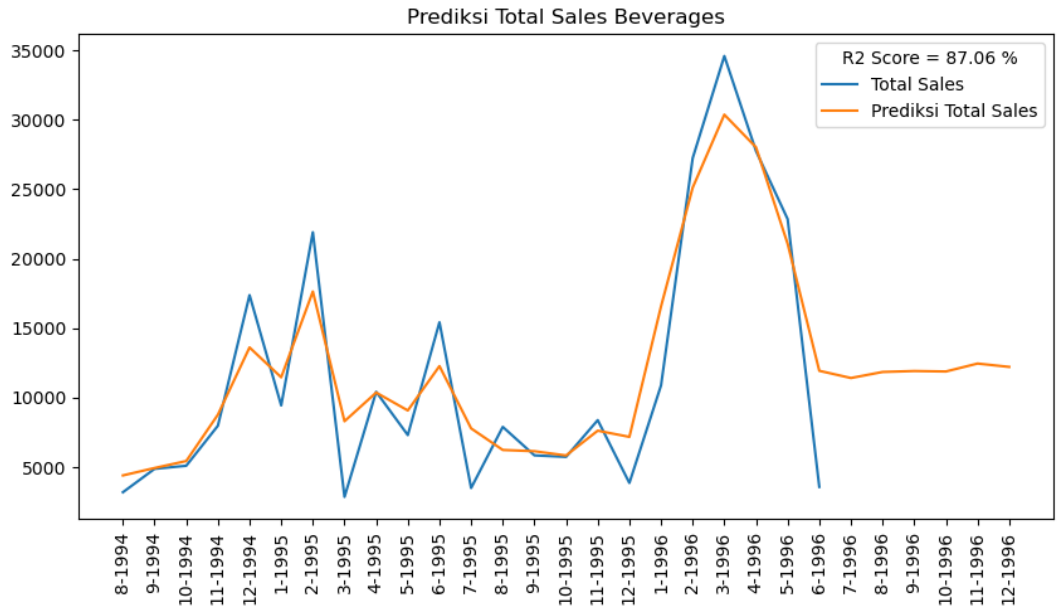
Saya menambahkan bulan dan tahun yang akan diprediksi yaitu dari bulan Juli-Des 1996.

Sehingga hasil prediksi menggunakan Random Forest Regressor untuk periode Juli-Des 1996 adalah sebagai berikut:

	Month	Year	Prediction Total Sales
0	7	1996	11408.65450
1	8	1996	11843.69450
2	9	1996	11912.62830
3	10	1996	11878.63325
4	11	1996	12449.79035
5	12	1996	12208.07355

```
Rata-rata prediksit Total Sales: 11950.245741666668
```

Grafik prediksi Total Sales Beverages



c. Produk apa dalam kategori itu yang mencatatkan penjualan tertinggi?

	Product Name	Category Name	Total Penjualan	Total Sales
0	Côte de Blaye	Beverages	623	141396.735
1	Ipoh Coffee	Beverages	580	23526.700
2	Chang	Beverages	1057	16355.960
3	Lakkalikööri	Beverages	981	15760.440
4	Steeleye Stout	Beverages	883	13644.000
5	Chai	Beverages	828	12788.100
6	Chartreuse verte	Beverages	793	12294.540
7	Outback Lager	Beverages	817	10672.650
8	Rhönbräu Klosterbier	Beverages	1155	8177.490
9	Sasquatch Ale	Beverages	506	6350.400
10	Guaraná Fantástica	Beverages	1125	4504.365
11	Laughing Lumberjack Lager	Beverages	184	2396.800

```
Produk dengan penjualan tertinggi                    = Côte de Blaye
Dengan total penjualan dan total sales berturut-turut = 623 produk - $141,396.735
```

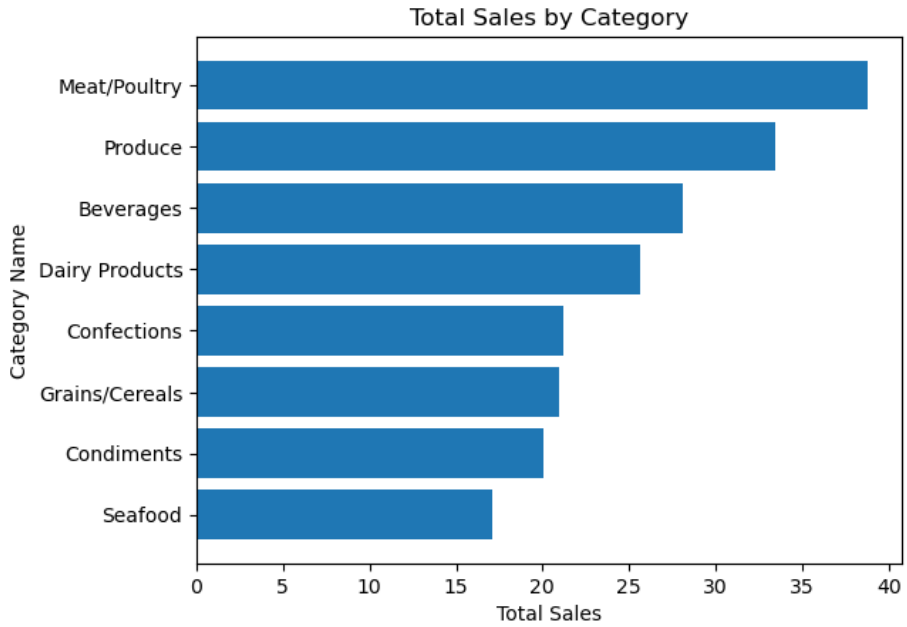
Produk kategori Beverages dengan penjualan tertinggi adalah ‘Côte de Blaye’
Dengan total penjualan 623 produk dan total sales $141,396.735

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

	Category Name	Total Penjualan	Total Sales	Rata-rata
0	Meat/Poultry	4199	163022.3595	38.824091
1	Produce	2990	99984.5800	33.439659
2	Beverages	9532	267868.1800	28.101991
3	Dairy Products	9149	234507.2850	25.632013
4	Confections	7906	167357.2250	21.168382
5	Grains/Cereals	4562	95744.5875	20.987415
6	Condiments	5298	106047.0850	20.016437
7	Seafood	7681	131261.7375	17.089147

Berdasarkan hasil query di samping, 3 kategori yang mencatat rata-rata nilai penjualan/Total Sales adalah ‘Meat/Poulty’, ‘Produce’, dan ‘Beverages’.

Dengan grafik di bawah ini.



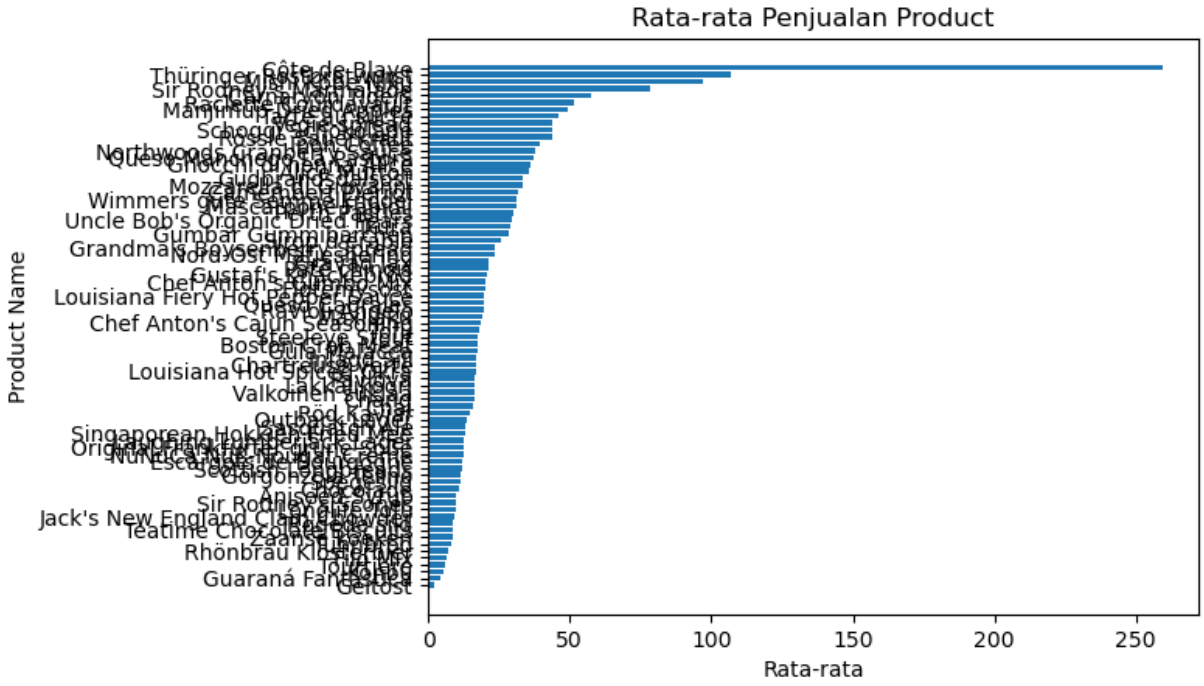
3. (NIM Genap) periode Januari-Juni 1996.
   Lima produk yang mencatatkan rata-rata nilai penjualan tertinggi selama periode tersebut?
   Lima produk yang mencatatkan rata-rata nilai penjualan terendah selama periode tersebut?
   a. Lima produk dengan nilai rata-rata penjualan/Total Sales tertinggi adalah sebagai berikut:

	Product Name	Order Date	Total Penjualan	Total Sales	Rata-rata
0	Côte de Blaye	1996-01-18 00:00:00.000000	275	71276.750	259.188182
1	Thüringer Rostbratwurst	1996-01-10 00:00:00.000000	390	41766.746	107.094221
2	Mishi Kobe Niku	1996-02-23 00:00:00.000000	3	291.000	97.000000
3	Sir Rodney's Marmalade	1996-02-13 00:00:00.000000	111	8704.260	78.416757
4	Carnarvon Tigers	1996-01-22 00:00:00.000000	181	10371.875	57.303177

   b. Lima produk dengan nilai rata-rata penjualan/Total Sales terendah adalah sebagai berikut:

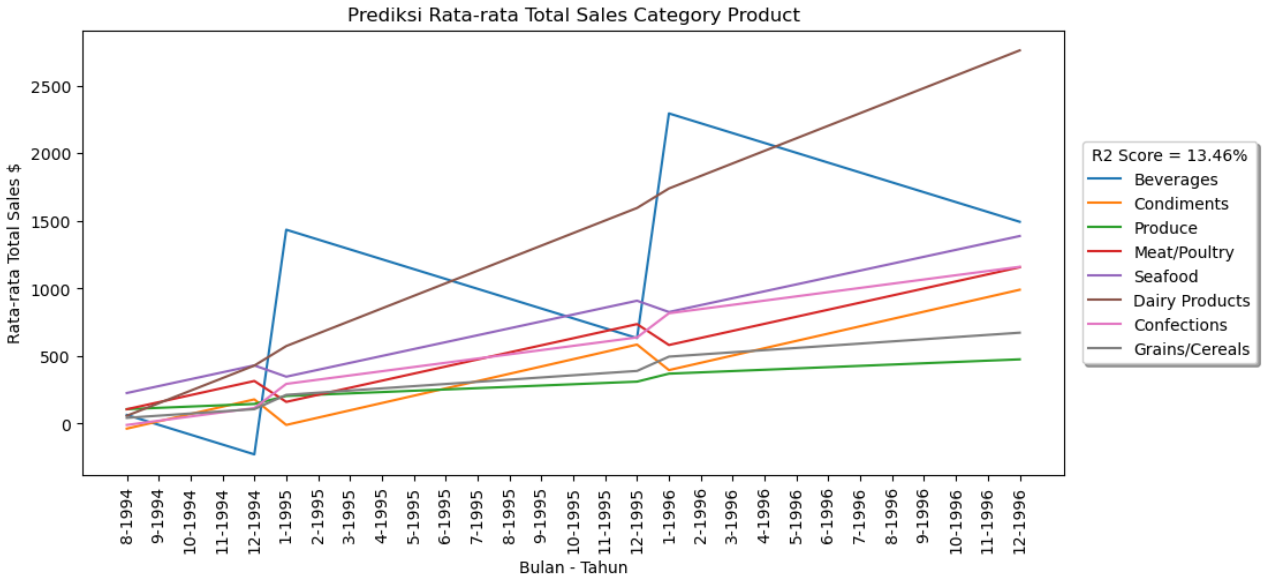
	Product Name	Order Date	Total Penjualan	Total Sales	Rata-rata
0	Geitost	1996-02-05 00:00:00.000000	202	476.6250	2.359530
1	Guaraná Fantástica	1996-01-03 00:00:00.000000	586	2484.0000	4.238908
2	Konbu	1996-01-26 00:00:00.000000	679	3827.1000	5.636377
3	Tourtière	1996-01-17 00:00:00.000000	126	781.8775	6.205377
4	Filo Mix	1996-01-08 00:00:00.000000	162	1018.1500	6.284877

   Dengan grafik sebagai berikut:



4. **Pada proyeksi data, kategori produk mana saja yang memiliki kecenderung penjualan rata-rata bulanan naik dan mana yang cenderung menurun?**
   Untuk memproyeksi data, saya menggunakan Linear Regression karena dari namanya 'Linear' yang mana dapat memudahkan apakah tren linenya naik atau turun. Dengan menggunakan Linear Regression sebagai model, berikut adalah hasil plotting visualisasi dari rata-rata total sales di kategori produk selama periode data.

Grafik keseluruhan



Dari grafik yang tersedia, kategori produk yang cenderung memiliki nilai rata-rata Total Sales naik adalah

1. Condiments
2. Produce
3. Meat/Poultry
4. Seafood
5. Dairy Products
6. Confections
7. Grains/Cereals

Dan nilai rata-rata Total Sales yang cenderung menurun adalah Beverages.

5. **(NIM Genap) Wilayah (Country) mana yang paling banyak melakukan order (count) dan paling tinggi nilai order (sum) selama periode data?**
   a. Negara yang paling banyak melakukan order adalah USA
   Dengan jumlah transaksi 352 kali transaksi dan

	Country	Jumlah Transaksi	Jumlah Barang Terjual
0	USA	352	9330
1	Germany	328	9213
2	Brazil	203	4247
3	France	184	3254
4	UK	135	2742
5	Austria	125	5167
6	Venezuela	118	2936
7	Sweden	97	2235
8	Canada	75	1984
9	Mexico	72	1025

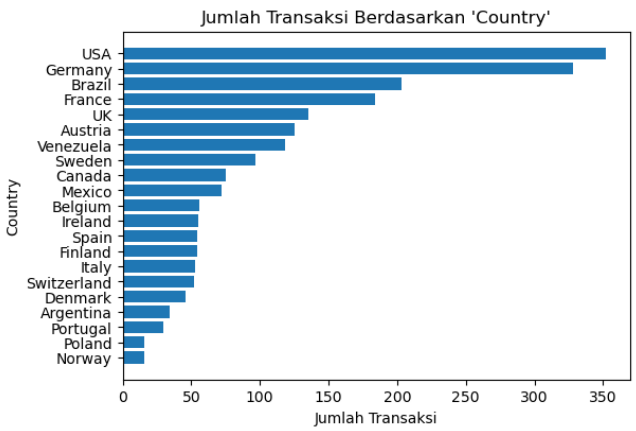
```
Wilayah atau Negara/(Country) yang paling banyak melakukan order adalah: USA
Dengan total order/jumlah transaksi: 352 dari negara 'USA'
```

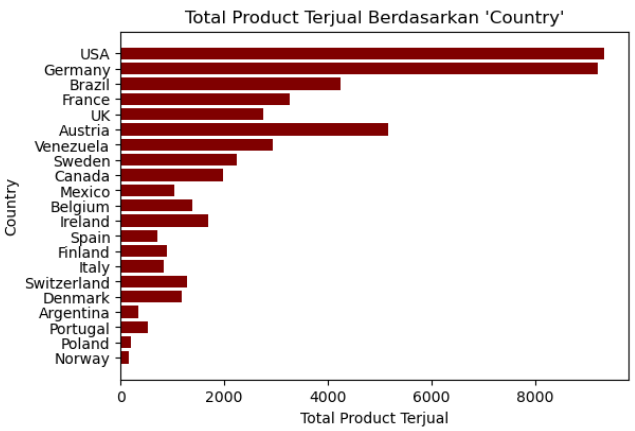
Untuk memperoleh query seperti gambar di samping, kita perlu menggabungkan 3 data dari tiga tabel yaitu order_details, orders, dan customers menggunaka INNER JOIN.

Setelah itu, samakan Order IDE dari tabel orders dan order_detail. Untuk menghitung jumlah transaksi berdasarkan negara, digunakan fungsi COUNT() dan GROUP BY berdasarkan negara yaitu Country. Kemudian di SORT/ORDER BY Jumlah Transaksi nya secara terbalik/DESC.

Dengan begitu, query ini dapat menampilkan daftar negara pelanggan yang melakukan transaksi paling banyak dan total produk yang terjual di setiap negara.

```
_query1 = """
    SELECT
        customers.Country,
        COUNT(order_details.`Order ID`) AS 'Jumlah Transaksi',
        SUM(order_details.Quantity) AS 'Total Product Terjual'
    FROM order_details
    INNER JOIN orders ON orders.`Order ID` = order_details.`Order ID`
    INNER JOIN customers ON customers.`Company Name` = orders.`Customer Name`
    GROUP BY customers.Country
    ORDER BY `Jumlah Transaksi` DESC;
"""
```

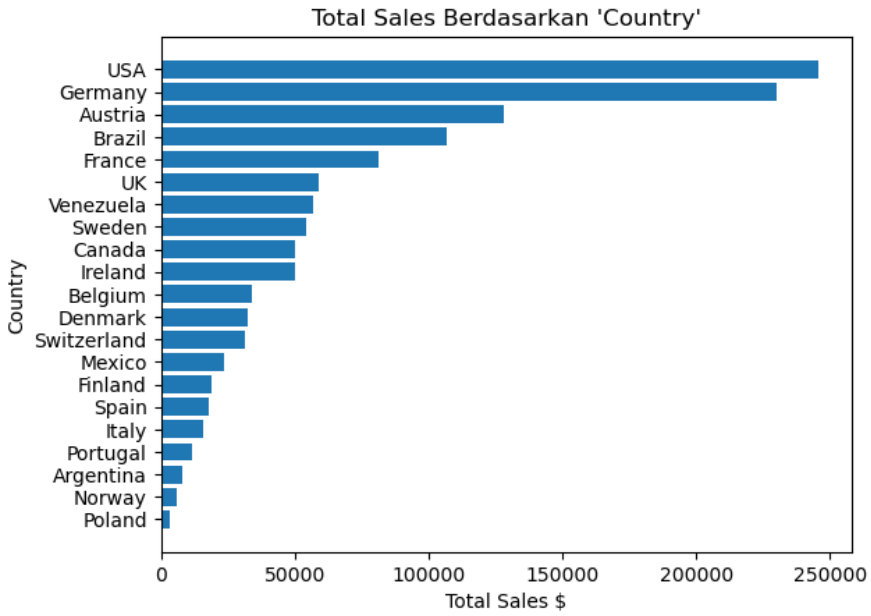




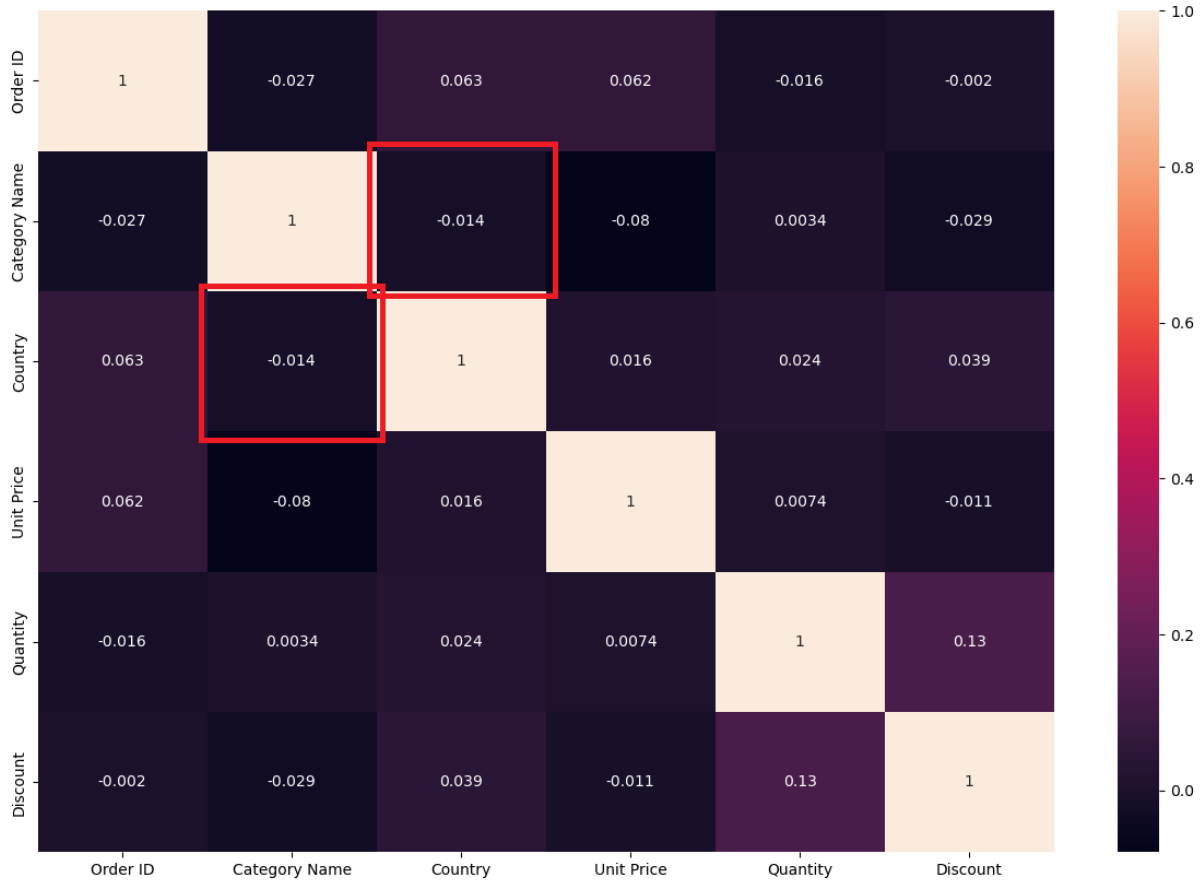
b. Negara yang paling tinggi nilai salesnya adalah USA
Dengan total sales $245,584.6105

	Jumlah Transaksi	Country	Total Sales
0	352	USA	245584.6105
1	328	Germany	230284.6335
2	125	Austria	128003.8385
3	203	Brazil	106925.7765
4	184	France	81358.3225
5	135	UK	58971.3100
6	118	Venezuela	56810.6290
7	97	Sweden	54495.1400
8	75	Canada	50196.2900
9	55	Ireland	49979.9050

```
Wilayah atau Negara/(Country) yang paling tinggi nilai sales adalah: USA
Dengan total sales: $245,584.61050
```

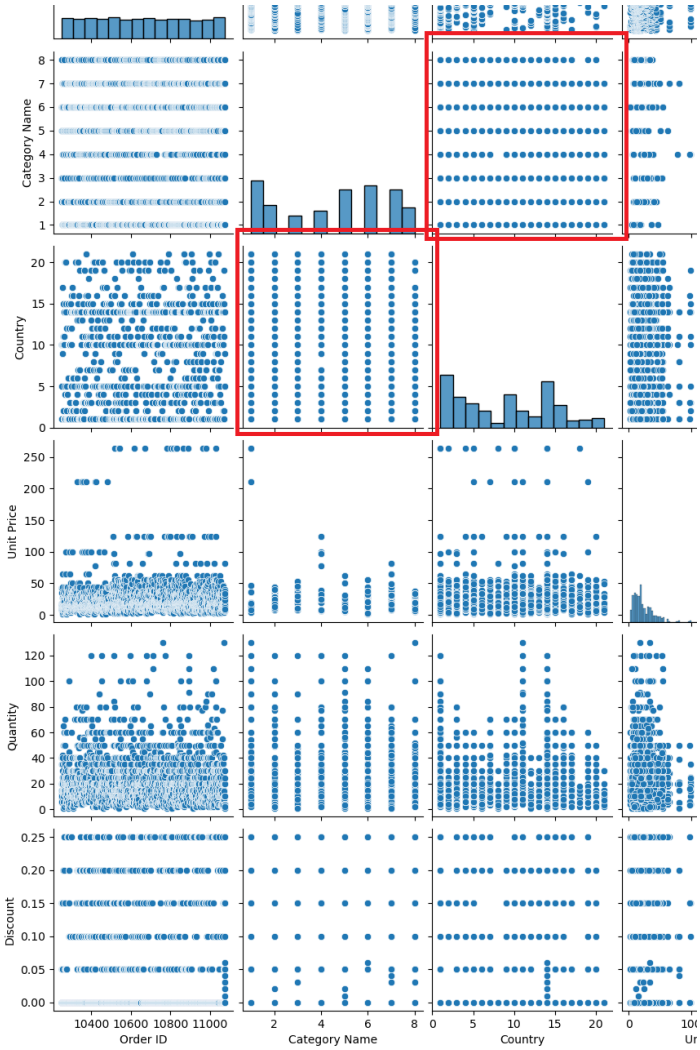


6. Apakah ada kaitan/relasi antara kategori produk (8 kategori) dengan asal customer (21 country).
Jawabannya adalah tidak. Tidak ada korelasi antar kategori produk dengan asal negara customer. Dengan menggunakan sns.heatmap (menampilkan matriks korelasi), hal ini dapat dilihat dari diagram korelasi berikut.




```
Nilai korelasi (`Category Product` - `Country`): -0.014107924449096861
```

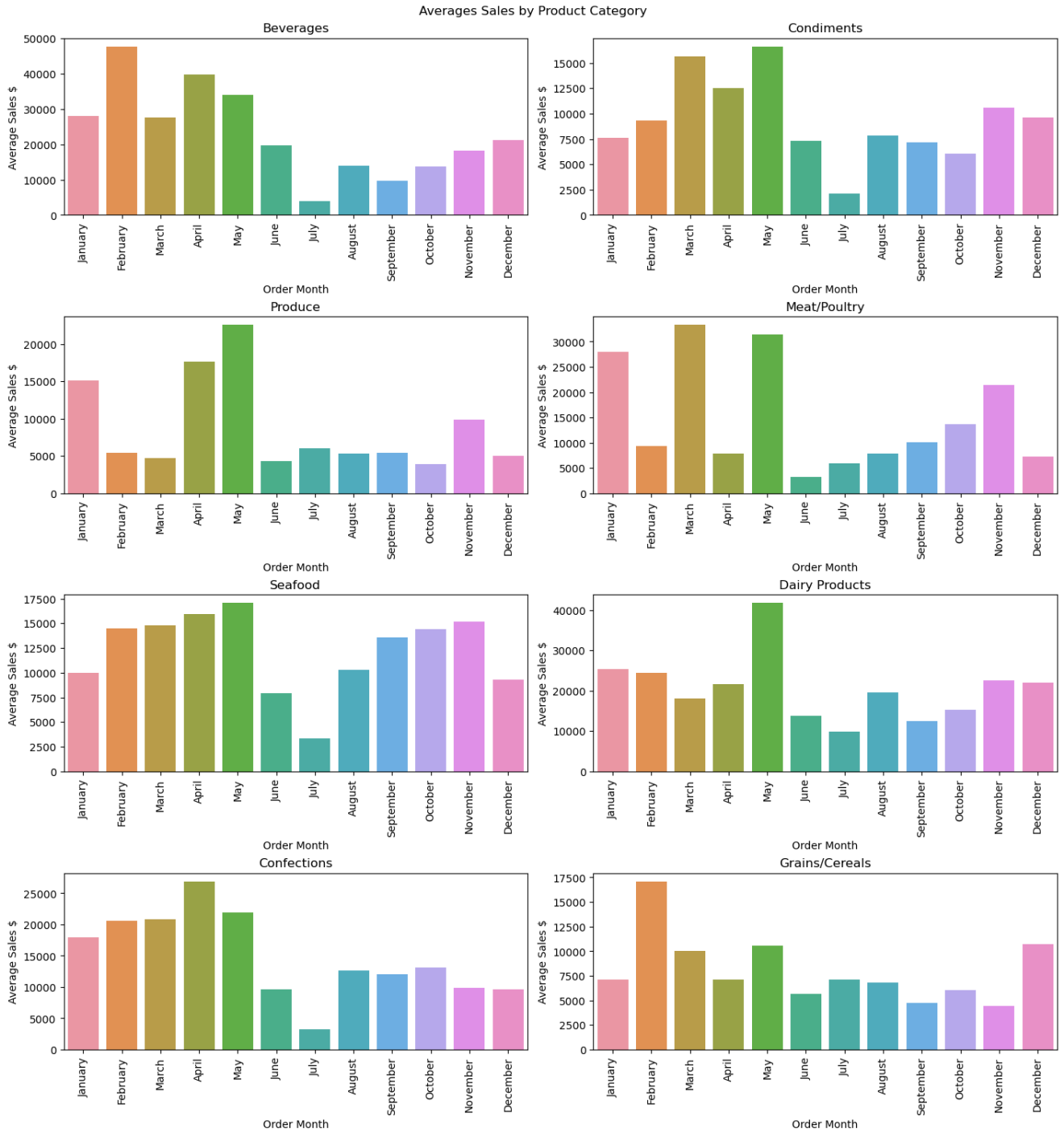
Jika nilainya mendekati 1, artinya kolom tersebut sangat berelasi/berkaitan. Sehingga dapat dikatakan kategori produk dan country tidak ada relasinya.



Hasil plotting terhadap semua kolomnya juga menggambarkan ketidakberhubungannya dua kategori tersebut.

7. **Apakah ada pola order kategori produk sepanjang periode data?**
   Ya, terdapat pola order kategori produk di sepanjang periode data, yaitu apa pun kategori produknya, rata-rata nilai penjualan terendah berada di tengah-tengah tahun, yaitu di rentang bulan Juni-Agustus dan rata-rata pangkal terbawah nilai penjualan berada di bulan Juni-Juli.
   Sedangkan, rata-rata nilai penjualan tertinggi berada di awal tahun, yaitu di rentang bulan Maret-Mei, dan rata-rata puncak nilai penjualan adalah di bulan Mei.

   Untuk membuktikannya, berikut adalah hasil visualisasinya,



Visualisasi di atas, diperoleh menggunakan SQL dengan menggabungkan 3 tabel, yaitu tabel Product, Orders, dan Order Details. Kemudian menghitung rata-rata dari kolom Unit Price dari tabel Order Details. Lalu menjumlah (SUM) kolom Quantity dari tabel Order Details. Dan terakhir mengkalikan kolom hasil rata-rata dari Unit Price dengan kolom hasil SUM Quantity menjadi kolom baru seperti gambar di bawah.

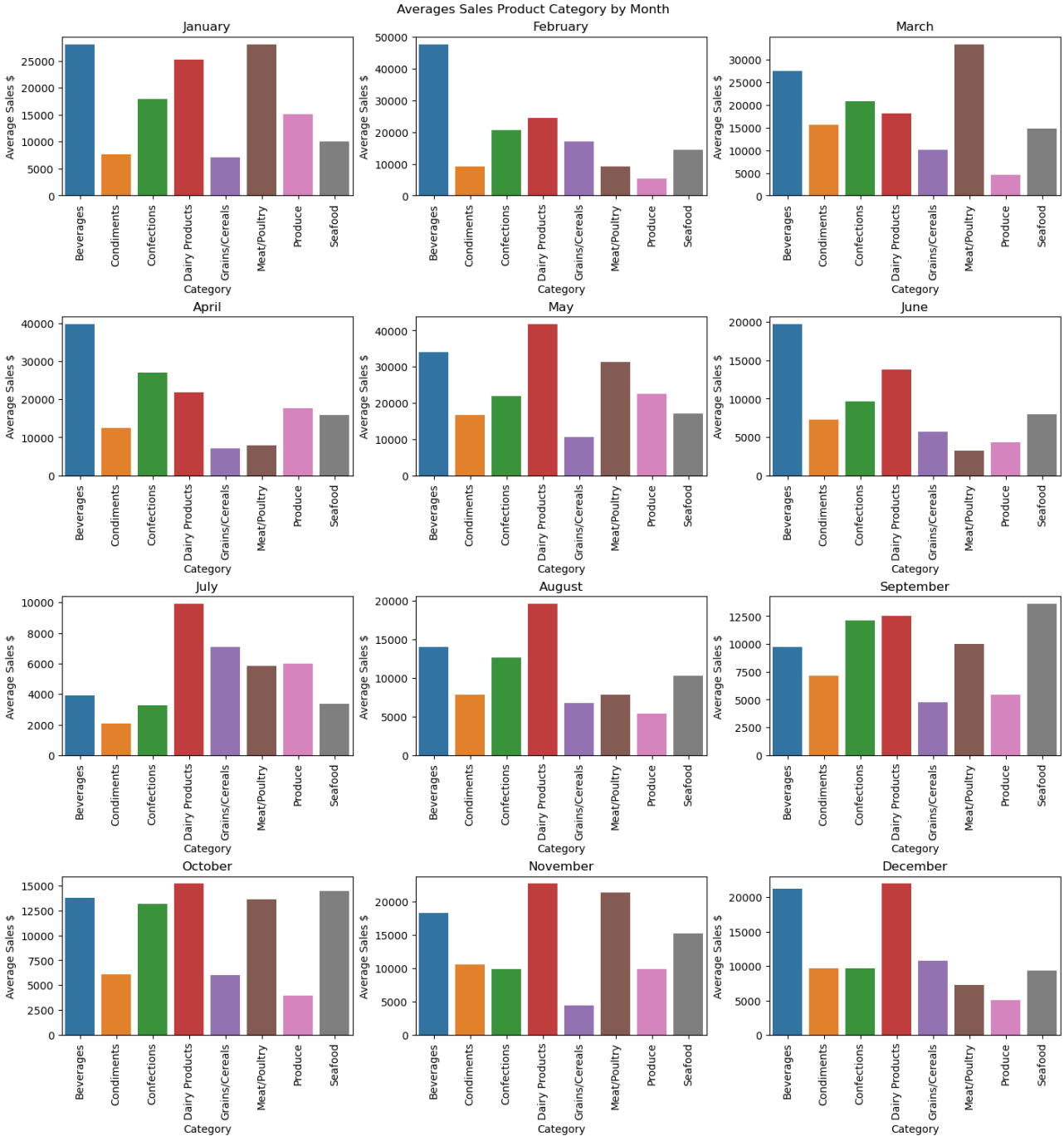
	Order Month	Category Name	Rata-rata Harga Unit	Total Penjualan	Rata-rata Penjualan
0	January	Beverages	34.913636	803	28035.650000
1	January	Condiments	22.557692	339	7647.057692
2	January	Confections	25.176897	711	17900.773448
3	January	Dairy Products	23.878947	1059	25287.805263
4	January	Grains/Cereals	21.590625	330	7124.906250
5	January	Meat/Poultry	49.343000	568	28026.824000
6	January	Produce	29.539474	510	15065.131579
7	January	Seafood	17.762222	562	9982.368889

Rata-rata Harga Unit: dapat diartikan sebagai harga rata-rata dari produk berkategori 'Beverages' (contoh) per-unit yang terjual pada bulan Januari/bulan itu.
Total Penjualan: dapat diartikan total penjualan produk berkategori 'Condiments' (contoh) pada bulan Januari/bulan itu.
Rata-rata Penjualan: kolom ini diperoleh dengan mengkalikan 2 kolom di atas, sehingga dapat dikatakan sebagai rata-rata nilai penjualan produk yang berkategori 'Seafood' (contoh) pada bulan Januari/bulan itu.

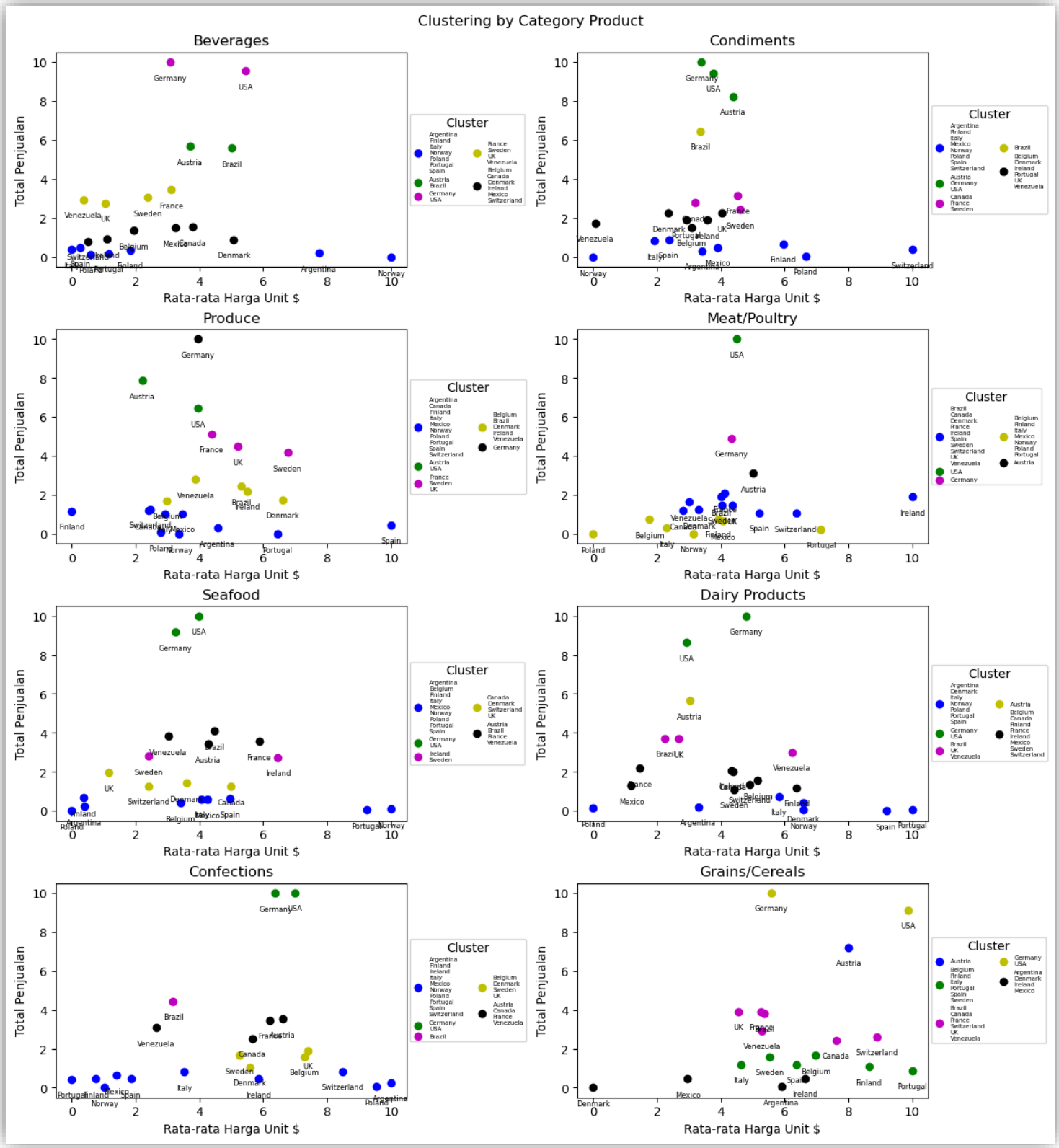
Dengan adanya kolom baru dari hasil perkalian 2 kolom 'Rata-rata Harga Unit' dengan 'Total Penjualan', menjadi kolom 'Rata-rata Penjualan', ini dapat mewakili harga unit & kuantiti penjualan produk berkategori tersebut. Sehingga ketika bar plotting, hanya menggunakan kolom 'Category Name' dan 'Rata-rata Penjualan'.

Menurut saya, jika menggunakan kolom 'Total Penjualan' saja sebagai tolak ukurnya, kita tidak akan mendapatkan pola order kategori dengan baik, karena banyaknya penjualan produk dapat bergantung dari harga per-unitnya. Tidak menentu 'Total Penjualan' saja.

Saya melakukan 2 visualisasi, namun untuk visualisasi di bawah ini, sulit untuk melihat polanya, sehingga lebih bagus melihat visualisasi yang di atas.


Averages Sales Product Category by Month
January
February
March
April
May
June
July
August
September
October
November
December
Average Sales $
Category
Beverages
Condiments
Confections
Dairy Products
Grains/Cereals
Meat/Poultry
Produce
Seafood

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



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).

Saya mengerjakan **PROYEK B** ini menggunakan *python,* sama seperti **PROYEK A** di atas. Solusi yang tepat untuk menganalisa data yang diberikan adalah menggunakan teknik *data mining*, yaitu ***Market Basket Analysis***. **MBA** merupakan teknik *data mining* yang mana cocok untuk mencari tahu tentang pola atau perilaku dari sebuah database transaksi. Metode ini memungkinkan untuk menemukan hubungan antara item-item yang dibeli oleh customer dalam suatu transaksi. Adapun algoritma yang dipakai dalam mengerjakan proyek ini adalah **Algoritma Apriori** yang mana merupakan salah satu teknik dalam ***Market basket analysis***. Library untuk melakukan **Apriori** adalah 'mlxtend'.

Hal pertama yang saya lakukan adalah mengekstrak `Date` `Time` `PeriodDay` `Day` `Month` `Hour` dari kolom `InvoiceDate`. Hal ini untuk mempermudah melakukan visualisasi dan menunjukkan hasil analisa yang lebih dalam seperti rata-rata per-hari, per-bulan, per-jam, dan lain-lain.

(Tabel mula-mula)

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
112890	545901	20713	JUMBO BAG OWLS	2	2011-03-07 17:52:00	1.95	15834.0	United Kingdom
14972	537627	21187	WHITE BELL HONEYCOMB PAPER GARLAND	48	2010-12-07 14:58:00	1.45	14739.0	United Kingdom
256185	559462	23202	JUMBO BAG VINTAGE LEAF	30	2011-07-08 13:12:00	2.08	16523.0	United Kingdom
431859	573763	23354	6 GIFT TAGS 50'S CHRISTMAS	12	2011-11-01 09:56:00	0.83	17139.0	United Kingdom
188801	553061	23201	JUMBO BAG ALPHABET	4	2011-05-13 11:12:00	2.08	17238.0	United Kingdom

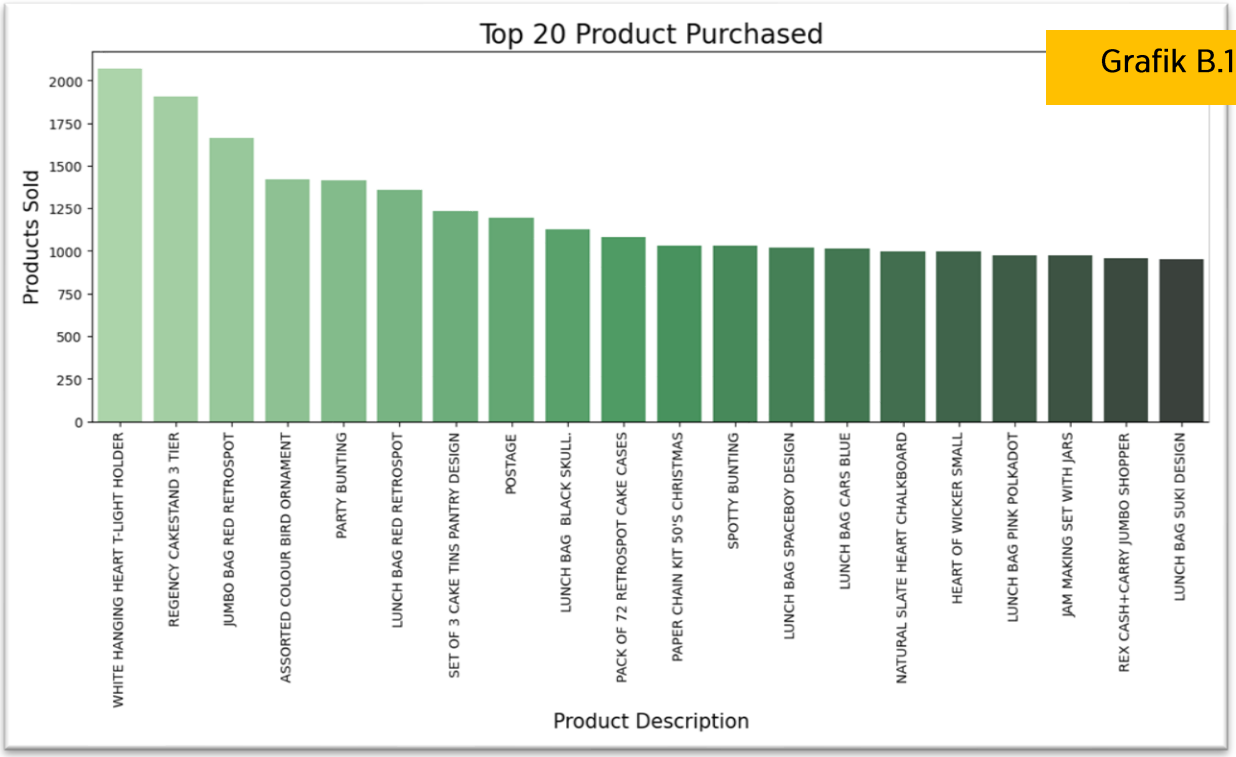
```
(541909, 8)
```

(Tabel akhir setelah diekstrak)

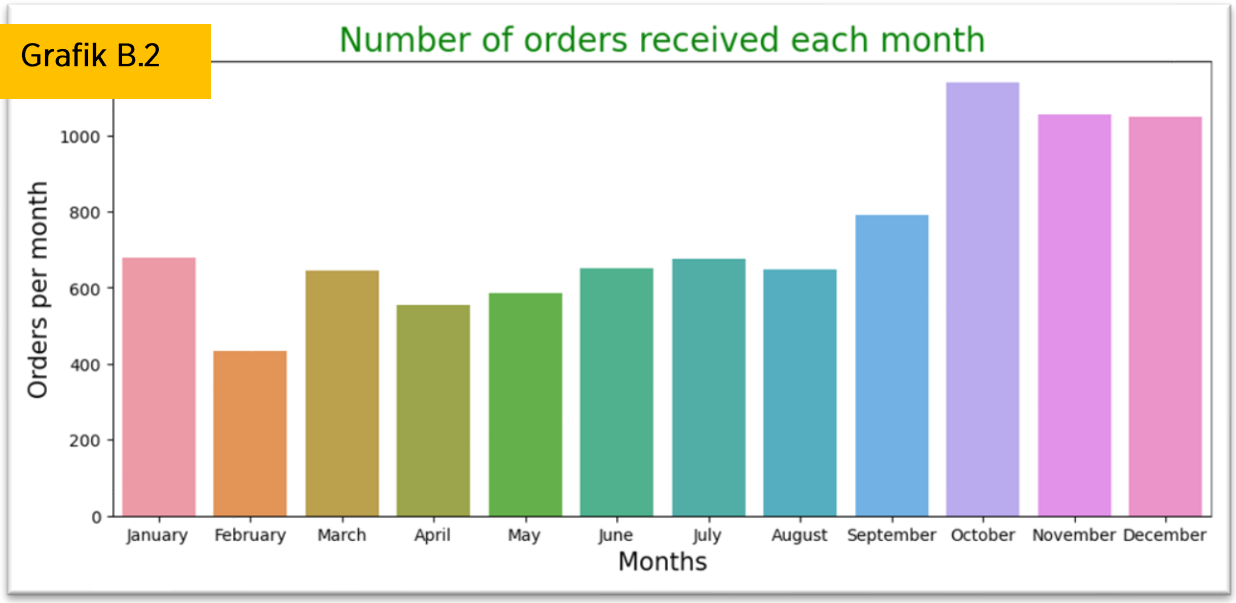
	InvoiceNo	StockCode	Description	Quantity	UnitPrice	CustomerID	Country	Date	Time	PeriodDay	Day	Month	Hour
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2.55	17850.0	United Kingdom	2010-12-01	08:26:00	Morning	Wednesday	December	8-9
1	536365	71053	WHITE METAL LANTERN	6	3.39	17850.0	United Kingdom	2010-12-01	08:26:00	Morning	Wednesday	December	8-9
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2.75	17850.0	United Kingdom	2010-12-01	08:26:00	Morning	Wednesday	December	8-9
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	3.39	17850.0	United Kingdom	2010-12-01	08:26:00	Morning	Wednesday	December	8-9
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	3.39	17850.0	United Kingdom	2010-12-01	08:26:00	Morning	Wednesday	December	8-9

Berikut beberapa hasil visualisasi.

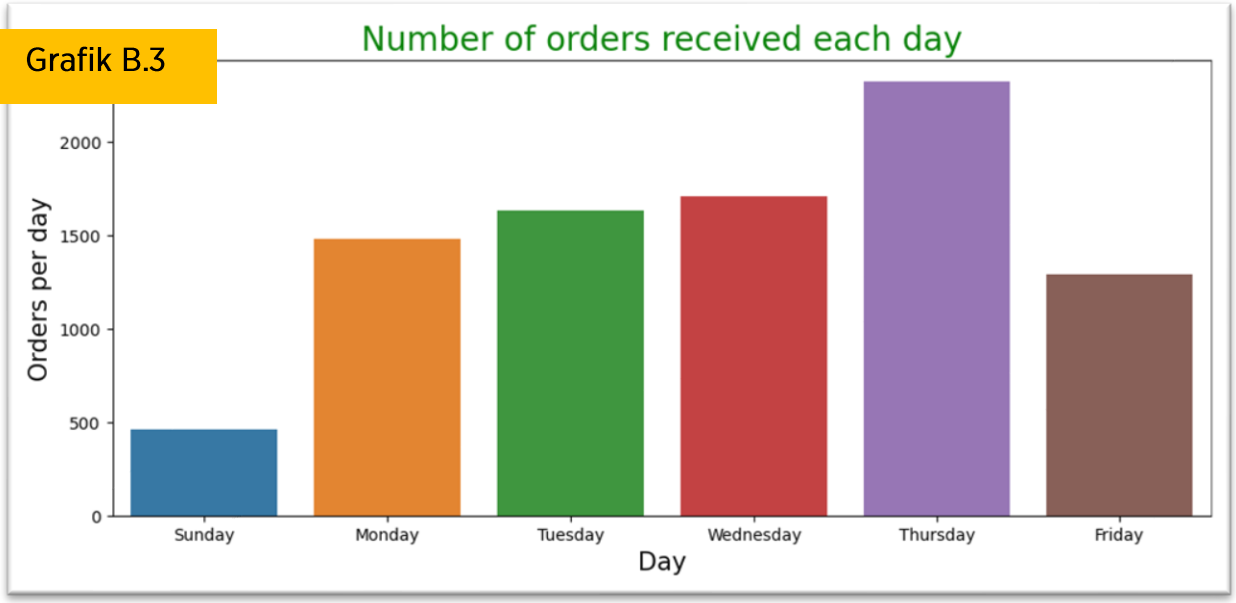
Grafik B.1

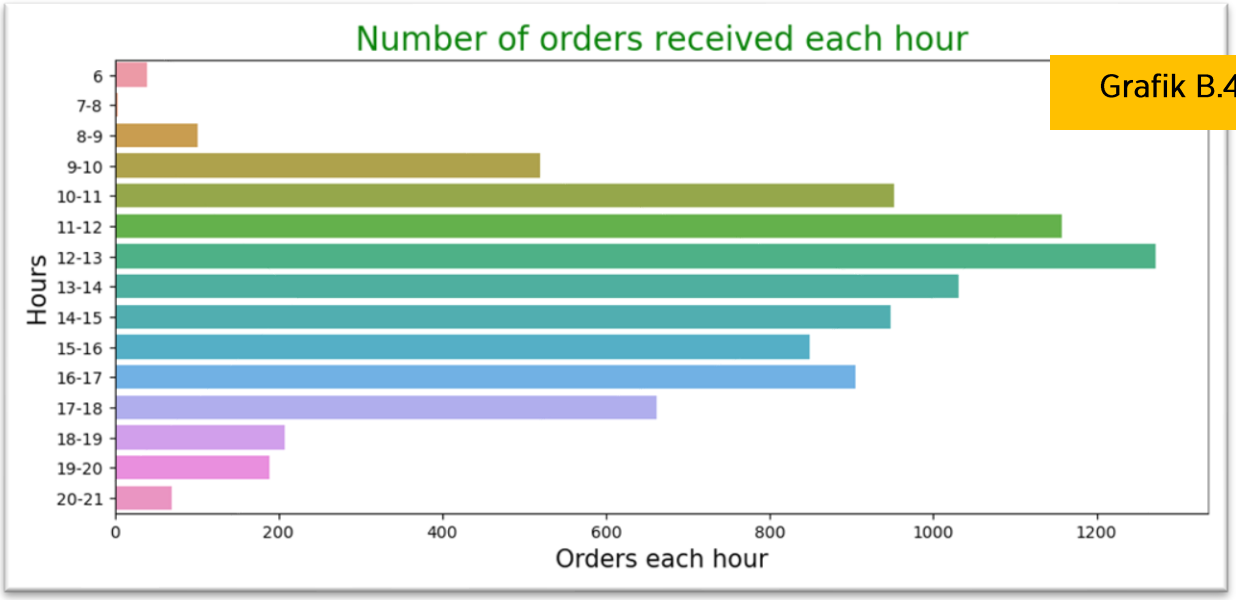


Grafik B.2

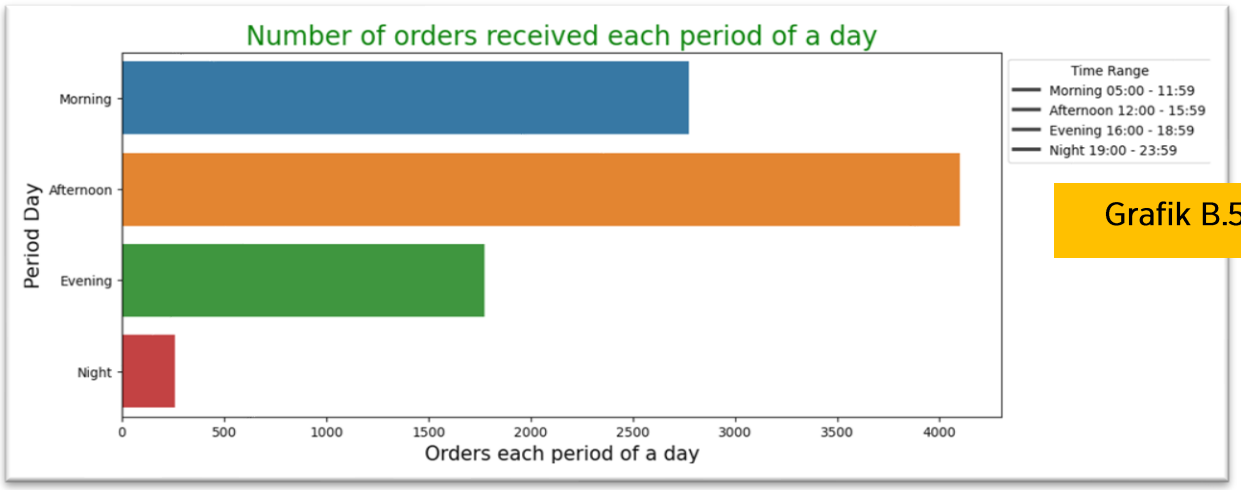


Grafik B.3





Grafik B.4



Grafik B.5

Dari beberapa visualisasi yang ditampilkan di atas, dapat kita simpulkan sebagai beriku:

1. (Grafik B.1) Produk yang paling laris sepanjang record data adalah 'WHITE HANGING HEART T-LIGHT HOLDER'.
2. (Grafik B.2) Order/transaksi terbanyak di setiap bulannya adalah pada bulan Oktober. Dan order terendah ada pada bulan Februari.
3. (Grafik B.3) Order/transaksi terbanyak berada di hari Kamis, dan order terendah ada pada hari Minggu. Berdasarkan visualisasi, tidak ada transaksi pada hari Sabtu. Artinya Toko tersebut tidak buka/melayani pada hari sabtu.
4. (Grafik B.4) Toko tersebut buka mulai dari jam 06.00 – 21.00. Selama toko tersebut buka, order mulai ramai pada jam 11.00 – 14.00, dengan Puncak order berada pada jam 12.00-13.00.
5. (Grafik B.5) Order terbanyak berada di siang hari, sedangkan order paling sedikit berada di malam hari.

Dengan menggunakan **Algoritma Apriori,** diperoleh tabel sebagai berikut:

Tabel B.1

antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
(GIRAFFE WOODEN RULER)	(REVOLVER WOODEN RULER)	0.001368	0.001916	0.001368	1.000000	522.000000	0.001366	inf
(ALARM CLOCK BAKELIKE ORANGE, ALARM CLOCK BAKE...	(ALARM CLOCK BAKELIKE RED)	0.001368	0.008210	0.001368	1.000000	121.800000	0.001357	inf
(LUNCH BAG RED RETROSPOT, LUNCH BAG WOODLAND)	(LUNCH BAG CARS BLUE)	0.001642	0.006294	0.001368	0.833333	132.391304	0.001358	5.962233
(SET OF TEA COFFEE SUGAR TINS PANTRY, SET OF 3...	(SET OF 3 REGENCY CAKE TINS)	0.001642	0.010126	0.001368	0.833333	82.297297	0.001352	5.939245
(LUNCH BAG RED RETROSPOT, LUNCH BAG BLACK SKU...	(JUMBO BAG RED RETROSPOT)	0.001642	0.011768	0.001368	0.833333	70.813953	0.001349	5.929392

Kolom 'antecendents' dan 'consequents' artinya jika customer membeli produk yang ada pada kolom antecendents, maka kemungkinan pelanggan tersebut juga akan membeli produk pada kolom consequents. Hal ini diperkuat oleh nilai pada kolom confidence, yang artinya dalam persen. Sebagai contoh, jika nilai confidence-nya 1.0, artinya 100% kombinasi produk di kolom antecedents dan consequents akan terjadi.

Nilai lift juga artinya menunjukkan bahwa nilai confidence dapat dipercaya.

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.

Victor Chendra