Capstone Project - Module 2

Diberikan dataset 'Northwind' kepada siswa yang bertujuan untuk melakukan analisa pada dataset tersebut, dimana para siswa dimintai Penjelasan tentang data dan konteks permasalahan yang hendak dianalisis. Dalam projek ini, siswa akan memposisikan diri sebagai data analyst dalam sebuah perusahaan. Siswa akan membuat serangkaian pertanyaan sesuai dengan fokus analisis, pertanyaan dibuat berdasarkan keperluan bisnis dan masalah yang mungkin dihadapi oleh perusahaan terkait sesuai dengan fokus analisisnya.

Database Information

Sumber Database :
https://drive.google.com/drive/folders/1fTHrwh_gcLsOFKXHnUzUGEu_APxLoD9i

General Question

1. Bagaimana konteks bisnis berdasarkan dengan data yang telah diberikan?

Sebuah perusahaan fiktif "**Northwind**", bergerak di bidang ekspor-impor makanan khusus dari seluruh dunia ingin mengetahui gambaran secara umum tentang bisnis yang mereka jalankan. Dimana perusahaan ini memberikan database mereka yang menunjukan detail dari proses transaksi dengan skema bisnis yang baik dengan basis data penjualan mereka.

2. Ada berapa banyak tabel yang disediakan oleh database? Jabarkan setiap tabelnya!

Database ini mempunyai 13 Tabel, yaitu :

- **categories** : Menyimpan informasi tentang kategori makanan
- **customercustomerdemo** : Sub-relations dari tabel customers
- **customerdemographics** : Sub-relations dari tabel customercustomerdemo
- **customers** : Menyimpan informasi pelanggan yang membeli produk dari Northwind
- **employees** : Menyimpan detail informasi karyawan dari Northwind
- **employeeterritories** : Sub-relations dari tabel employees dan territories
- **orderdetails** : Sub-relations dari tabel orders dan tabel products, berisikan detail dari pesanan
- **orders** : Menyimpan informasi detail tentang transaksi antara pelanggan dan perusahaan
- **products** : Menyimpan informasi detail tentang produk
- **region** : Sub-relations dari tabel territories
- **shippers** : Menyimpan detail informasi dari pengiriman
- **suppliers** : Menyimpan informasi Suppliers dan Vendors dari Northwind
- **territories** : Sub-relations dari tabel Employee territories

Setiap tabel yang tertera pada database dapat terhubung, baik secara langsung maupun tidak langsung, sehingga setiap informasi dari database ini akan dapat saling berkaitan.

SQL

1. Apakah tabel customers, orders, ordersdetail, products dan categories dapat digabungkan menjadi 1 tabel? Jika memungkinkan, tampilkan tabel yang memuat informasi transaksi yang melibatkan kelima tabel tersebut.

Ya, kelima table tersebut dapat digabungkan dengan menjadi 1 tabel dikarenakan mempunyai Primary Key yang saling berkaitan antar table

```
    '''
select c.CustomerID, c.CompanyName, c.ContactName, c.ContactTitle, c.City, c.Region, c.PostalCode, c.Country,
o.OrderID, o.OrderDate, o.RequiredDate, o.ShippedDate, o.Freight,
od.ProductID, od.Quantity,
p.ProductName, p.UnitPrice, p.UnitsInStock, p.UnitsOnOrder, p.ReorderLevel, p.Discontinued,
ca.CategoryName,
(timediff(o.ShippedDate,o.OrderDate))ProcessingDate,
(p.UnitsInStock - p.UnitsOnOrder) Restock
from customers c
left join orders o on c.CustomerID = o.CustomerID
left join orderdetails od on o.OrderID = od.OrderID
left join products p on od.ProductID = p.ProductID
left join categories ca on p.CategoryID = ca.CategoryID;
    ''')
```

CustomerID	CompanyName	ContactName	ContactTitle	City	Region	PostalCode	Country	OrderID	OrderDate	...	Quantity	ProductName	UnitPrice	UnitsInStock	Units
ALFKI	Alfreds Futterkiste	Maria Anders	Sales Representative	Berlin	None	12209	Germany	10643.0	1997-08-25	...	15.0	Rssle Sauerkraut	45.6000	26.0	
ALFKI	Alfreds Futterkiste	Maria Anders	Sales Representative	Berlin	None	12209	Germany	10643.0	1997-08-25	...	21.0	Chartreuse verte	18.0000	69.0	
ALFKI	Alfreds Futterkiste	Maria Anders	Sales Representative	Berlin	None	12209	Germany	10643.0	1997-08-25	...	2.0	Spegesild	12.0000	95.0	
ALFKI	Alfreds Futterkiste	Maria Anders	Sales Representative	Berlin	None	12209	Germany	10692.0	1997-10-03	...	20.0	Vegie-spread	43.9000	24.0	
ALFKI	Alfreds Futterkiste	Maria Anders	Sales Representative	Berlin	None	12209	Germany	10702.0	1997-10-13	...	6.0	Aniseed Syrup	10.0000	13.0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
WOLZA	Wolski Zajazd	Zbyszek Piestrzeniewicz	Owner	Warszawa	None	01-012	Poland	10998.0	1998-04-03	...	12.0	Guaran Fantstica	4.5000	20.0	
WOLZA	Wolski Zajazd	Zbyszek Piestrzeniewicz	Owner	Warszawa	None	01-012	Poland	10998.0	1998-04-03	...	7.0	Sirop d'rable	28.5000	113.0	
WOLZA	Wolski Zajazd	Zbyszek Piestrzeniewicz	Owner	Warszawa	None	01-012	Poland	10998.0	1998-04-03	...	20.0	Longlife Tofu	10.0000	4.0	
WOLZA	Wolski Zajazd	Zbyszek Piestrzeniewicz	Owner	Warszawa	None	01-012	Poland	10998.0	1998-04-03	...	30.0	Rhnbru Klosterbier	7.7500	125.0	
WOLZA	Wolski Zajazd	Zbyszek Piestrzeniewicz	Owner	Warszawa	None	01-012	Poland	11044.0	1998-04-23	...	12.0	Tarte au sucre	49.3000	17.0	

ows × 24 columns

2. Berapa banyak jenis line produk yang tersedia? Berapa banyak jumlah line produk untuk setiap jenis produknya?

Ada 8 jenis produk line yang tersedia. Berikut untuk tampilan jumlah line produk setiap produknya:

	CategoryID	CategoryName	Banyak_Produk
0	1	Beverages	12
1	2	Condiments	12
2	3	Confections	13
3	4	Dairy Products	10
4	5	Grains/Cereals	7
5	6	Meat/Poultry	6
6	7	Produce	5
7	8	Seafood	12

Data Manipulation

1. Apakah terdapat anomaly berupa missing values, data duplicate, kesalahan data formatting, dan/atau inconsistency typing? Jika ada tunjukan serta lakukan penanganan pada anomaly tersebut.

Ya, terdapat anomaly berupada missing values, kesalahan data formatting dan inconsistency typing dalam data. Berikut Penanganan pada anomaly data tersebut :

```
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   CustomerID      2159 non-null   object
 1   CompanyName     2159 non-null   object
 2   ContactName     2159 non-null   object
 3   ContactTitle    2159 non-null   object
 4   City            2157 non-null   object
 5   Region          826 non-null    object
 6   PostalCode      2102 non-null   object
 7   Country         2157 non-null   object
 8   OrderID         2155 non-null   float64
 9   OrderDate       2155 non-null   datetime64[ns]
 10  RequiredDate    2155 non-null   datetime64[ns]
 11  ShippedDate     2082 non-null   datetime64[ns]
 12  Freight         2155 non-null   object
 13  ProductID       2155 non-null   float64
 14  Quantity        2155 non-null   float64
 15  ProductName     2155 non-null   object
 16  UnitPrice       2155 non-null   object
 17  UnitsInStock    2155 non-null   float64
 18  UnitsOnOrder    2155 non-null   float64
 19  ReorderLevel    2155 non-null   float64
 20  Discontinued    2155 non-null   float64
 21  CategoryName    2155 non-null   object
 22  ProcessingDate  2082 non-null   timedelta64[ns]
 23  Restock         2155 non-null   float64
dtypes: datetime64[ns](3), float64(8), object(12), timedelta64[ns](1)
memory usage: 404.9+ KB
```

...n	PostalCode	Cou
...e	12209	Germ
...e	12209	Germ
...e	12209	Germ
...e	12209	Germ
...	...	
...e	01-012	Po
...e	01-012	Po
...e	01-012	Po
...e	01-012	Po
...e	01-012	Po

Mari sejenak melihat apa yang ditampilkan pada general info yang disajikan di atas. Terlihat bahwa secara keseluruhan terdapat 2159 baris data dengan total 21 kolom. Setiap kolomnya memiliki tipe data yang berbeda-beda. Ada object, datetime64, timedelta64 dan float.

1. Oke, mari sejenak mengesampingkan tipe data. Mari berforkus pada non-null values atau data yang tersedia pada setiap kolomnya. Jika melihat informasi tersebut, tidak semua kolom atau feature yang memiliki data lengkap. Yang paling terlihat jomplang adalah pada `Region`. Features tersebut kehilangan lebih dari 50% data. Selain dari itu, ada beberapa feature yang datanya juga missing, yang nantinya akan ditampilkan pada bagian berikutnya. **Kesimpulan pertama adalah bahwa terdapat *missing value* yang harus ditanggulangi.**

2. Fokus berikutnya adalah perhatikan pada features berikut ini:

   - Freight
   - UnitPrice

   Yang kedua, yaitu pada features `Freight` dan juga `UnitPrice`. Kedua feature ini sama seperti sebelumnya, masih dibaca sebagai object yang seharusnya merupakan tipe data `Integer`. Oleh karena itu, kedua features ini juga harus ditanggulangi dengan cara mengubah tipe datanya. Dari kedua penjelasan tersebut, maka **kesimpulan keduanya adalah terdapat features yang memiliki tipe data yang salah dan harus diubah sesuai dengan tipe data seharusnya.**

3. Fokus berikutnya adalah pada features `PostalCode`. Pada Postal Code / Zip Code pada sebuah daerah adalah 6 series number yang memudahkan untuk proses sorting mailing address. Pada data di tabel1, kolom PostalCode mempunyai beragam cara penulisan yang salah. Hal ini mungkin disebabkan oleh Human Error dsb. Untuk mengatasinya, maka kita akan melakukan drop (delete) kolom PostalCode pada tabel1. **Kesimpulan ketiga adalah bahwa terdapat anomali data yang mungkin disebabkan oleh Human Error dsb, yang harus ditanggulangi**

2. Apakah terdapat tipe data yang berupa datetime pada data? Jika iya, apakah tipe data yang berupa datetime tersebut dapat dicari tahu selisihnya? Silahkan tampilkan hasilnya, berikan insight yang sesuai (jika memungkinkan).
Iya, terdapat tipe data yang berupa datetime pada data. Dimana kita menggunakan data datetime ini untuk menganalisa insight berapa lama waktu prossess

- ProcessingDate : selisih antara OrderDate dan ShippedDate

```
tabel1['ProcessingDate'].value_counts()
```

```
7 days 00:00:00     271
6 days 00:00:00     241
3 days 00:00:00     187
9 days 00:00:00     184
5 days 00:00:00     177
4 days 00:00:00     176
8 days 00:00:00     174
2 days 00:00:00     149
10 days 00:00:00    129
12 days 00:00:00     45
```

3. Ada berapa banyak transaksi untuk setiap status barang yang di berhentikan produksinya? Lalu, bagaimana rata-rata dari UnitPrice untuk setiap status pengiriman tersebut? Jabarkan dan berikan insight.

```
Aggregate = tabel1[['UnitPrice','Discontinued']].groupby('Discontinued').describe()
Aggregate
```

	UnitPrice							
	count	mean	std	min	25%	50%	75%	max
Discontinued								
0.0	1861.0	26.666926	30.823924	2.5	12.5	19.0	33.25	263.50
1.0	221.0	40.830226	38.656280	4.5	14.0	32.8	45.60	123.79

Ada 221 produk yang di stop dari semua total transaksi. Dapat disimpulkan bawa secara garis besar, masih banyak kategori produk dengan sub produk tertentu yang masih di produksi oleh perusahaan dari semua total penjualan. Namun jika kita melihat UnitPrice, secara rata-rata nilai beli barang pada produk yang di stop produksi (discontinued) lebih besar dibandingkan dengan yang masih diproduksi (continued) yakni 40.83 juta US Dollar. Bisa dibayangkan bahwa hanya dengan 8 produk yang discontinued dan total order produk sebanyak 221 pieces, perusahaan akan mengurangi penjualan bersih mereka sebesar 40.83 juta US Dollar (hampir 2/3 nilai dari rata-rata nilai beli barang keseluruhan).

4. Buatlah tabel yang melihat keadaan berapa banyak product yang di Discontinued oleh perusahaan beserta dengan kolom UnitsOnOrder dan ReorderLever. Berikan Insight tentang tabel tersebut!

```
tabel3 = sql_table(
    '''
    select ProductID, UnitsInStock, UnitsOnOrder, ReorderLevel, Discontinued
    from products
    where Discontinued = 1.0
    '''
)
tabel3
```

0]:

	ProductID	UnitsInStock	UnitsOnOrder	ReorderLevel	Discontinued
0	5	0	0	0	1
1	9	29	0	0	1
2	17	0	0	0	1
3	24	20	0	0	1
4	28	26	0	0	1
5	29	0	0	0	1
6	42	26	0	0	1
7	53	0	0	0	1

Walaupun secara aggregate rata-rata nilai beli barang pada produk yang di stop produksi (discontinued) lebih besar dibandingkan dengan yang masih diproduksi (continued), kita dapat mengetahui kenapa 8 produk tersebut di discontinued (lihat tabel3). Pada tabel3, kita dapat mengambil kesimpulan bahwa 8 produk tersebut di discontinued dikarenakan sudah tidak adanya lagi customers yang melakukan reorder (ReorderLevel) bahkan sedang melakukan order (UnitsOnOrder) dan stock unitnya (UnitsInStock) pun kebanyakan kosong atau menipis. Kemungkinan besar perusahaan melakukan discontinued pada produk ini for good reason.

5. Restock adalah sebuah variabel baru dimana merupakan hasil dari selisih antara UnitsInStock dan UnitsOnOrder. Apakah terdapat anomali data pada variabel tersebut? Berikan cara penanganan/insight.

```
tabel1['Restock'].value_counts()
```

```
 26.0     132
 20.0     122
 0.0      107
 17.0      72
 112.0     61
          ...
 52.0       9
 49.0       9
-96.0       7
-55.0       6
-39.0       6
Name: Restock, Length: 61, dtype: int64
```

Melihat output unique values beserta dengan banyaknya data di setiap unique values tersebut, terdapat stock yang menunjukan nilai minus (-96, -55 dan -39) dan terdapat 19 data di dalamnya. Asumsinya adalah kemungkinan murni ada kesalahan input oleh user saat memasukan ke dalam database (Human Error)

Dari asumsi tersebut, cara mengatasinya cukup dengan mengabaikan (dipertahankan) karena bisa dilakukan analisis lebih lanjut untuk mengetahui letak permasalahannya

6. Apakah terdapat outlier pada variable UnitsInStock? Jika ada, tunjukan di data keberapa.

Melihat Data Outlier pada feature `UnitsInStock`

```
# Outlier Check With Function

Q1_amount = tabel1['UnitsInStock'].describe()['25%']
Q3_amount = tabel1['UnitsInStock'].describe()['75%']
iqr = Q3_amount - Q1_amount

outlier_index = tabel1[(tabel1['UnitsInStock'] < Q1_amount - (1.5 * iqr)) | (tabel1['UnitsInStock']> Q3_amount + (1.5 * iqr))
not_outlier_index = tabel1[(tabel1['UnitsInStock'] > Q1_amount - (1.5 * iqr)) & (tabel1['UnitsInStock']< Q3_amount + (1.5 *
tabel1.loc[outlier_index]
```

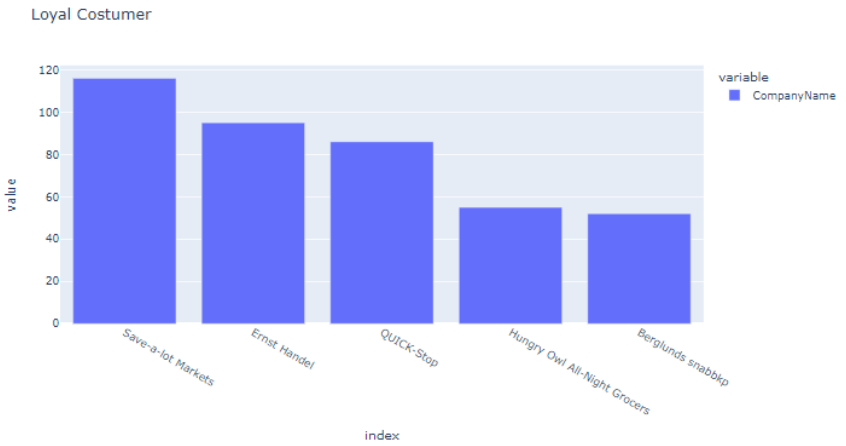
302]:

	CustomerID	CompanyName	ContactName	ContactTitle	City	Country	OrderID	OrderDate	RequiredDate	ShippedDate	...	Quantity	Proc
8	ALFKI	Alfreds Futterkiste	Maria Anders	Sales Representative	Berlin	Germany	10952.0	1998-03-16	1998-04-27	1998-03-24	...	16.0	Bo
26	ANTON	Antonio Moreno Taquera	Antonio Moreno	Owner	Mxico D.F.	Mexico	10535.0	1997-05-13	1997-06-10	1997-05-21	...	10.0	Bo
36	ANTON	Antonio Moreno Taquera	Antonio Moreno	Owner	Mxico D.F.	Mexico	10682.0	1997-09-25	1997-10-23	1997-10-01	...	30.0	
75	BERGS	Berglunds snabbkp	Christina Berglund	Order Administrator	Lule	Sweden	10280.0	1996-08-14	1996-09-11	1996-09-12	...	30.0	
90	BERGS	Berglunds snabbkp	Christina Berglund	Order Administrator	Lule	Sweden	10572.0	1997-06-18	1997-07-16	1997-06-25	...	50.0	Bo
...	...	...	...	...	...	...	...	...	...	...	...	...	
2004	WANDK	Die Wandernde Kuh	Rita Mller	Sales Representative	Stuttgart	Germany	10301.0	1996-09-09	1996-10-07	1996-09-17	...	10.0	Bo
2009	WANDK	Die Wandernde Kuh	Rita Mller	Sales Representative	Stuttgart	Germany	10312.0	1996-09-23	1996-10-21	1996-10-03	...	10.0	
2101	WHITC	White Clover Markets	Karl Jablonski	Owner	Seattle	USA	10596.0	1997-07-11	1997-08-08	1997-08-12	...	30.0	
2135	WILMK	Wilman Kala	Matti Karttunen	Owner/Marketing Assistant	Helsinki	Finland	10879.0	1998-02-10	1998-03-10	1998-02-12	...	12.0	Bo
2157	WOLZA	Wolski Zajazd	Zbyszek Piestrzeniewicz	Owner	Warszawa	Poland	10998.0	1998-04-03	1998-04-17	1998-04-17	...	30.0	

96 rows × 22 columns

Data Visualisation & Statistics

1. Apakah terdapat pelanggan yang loyal? Jika ada, bagaimana treatment dan strategi yang bisa dimanfaatkan dengan melihat adanya pelanggan setia tersebut?



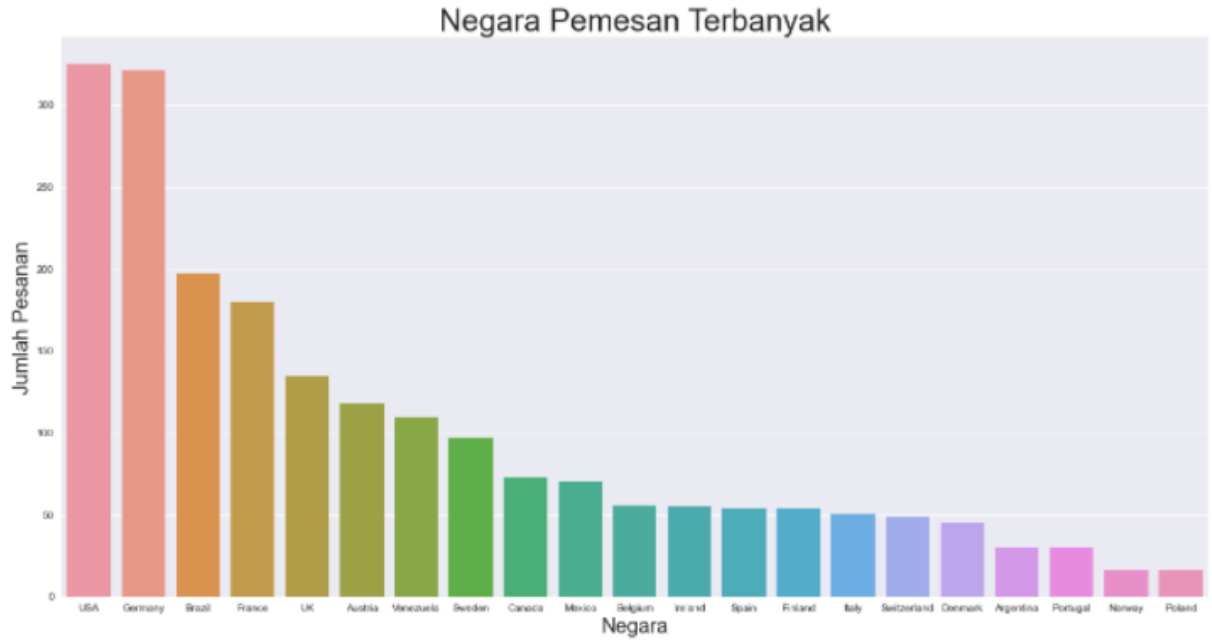
Dari grafik diatas dapat kita lihat bahwa dari sekitar 2000an transaksi yang terjadi, hanya terdapat satu perusahaan yang memiliki data transaksi lebih dari 100 kali yakni Save-a-lot Markets. Ini menandakan bahwa Save-alot Markets merupakan perusahaan yang paling loyal dengan total

transaksi lebih dari 100 kali. Adapun perusahaan-perusahaan yang masuk ke dalam top 5 loyal costumers antara lain : Save-a-lot Markets, Ernst Handel, QUICK-Stop, Hungry Owl All-Night Grocers dan Berglunds snabbkp.

Melihat dengan adanya segmentasi customers yang cukup loyal di dalam transaksi ini, menandakan bahwa terdapat peluang untuk menawarkan barang-barang baru kepada mereka. Dimana sebagai produsen, kita dapat menawarkan promosi atau special offer kepada customers yang cukup loyal membeli barang. Dimana hal ini dilakukan agar customers yang sudah memiliki kepercayaan terhadap perusahaan akan semakin bisa lebih sering melakukan transaksi.

Selain itu, kita juga dapat melihat data transaksi yang dibeli oleh customer lainnya dengan melihat features UnitsInOrder dan ReorderLevel sehingga kita mengetahui jenis barang yang banyak di order kembali oleh customers

2. Negara mana saja yang menjadi tujuan pemesan produk terbanyak? Tampilkan jabarkan dan buatlah analisisnya.

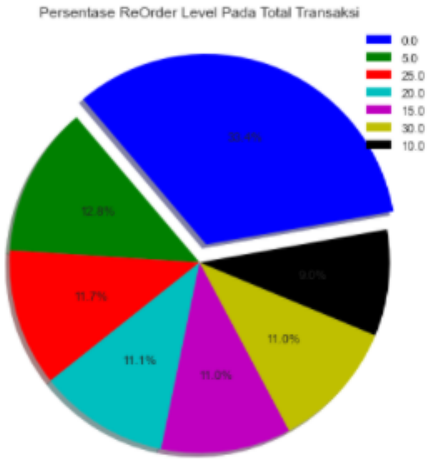


Setelah sebelumnya telah mengetahui pelanggan atau customer yang loyal, berikutnya mari kita lihat negara mana saja yang menjadi pemesan terbanyak. Jika melihat data dari grafik di atas, terlihat jelas bahwa USA dan Germany menjadi negara tujuan pemesan terbanyak. Dimana kedua negara ini memiliki tingkat pembelian mencapai lebih dari 300. Jika kita melihat sesuai dengan grafik diatas secara seksama, maka dapat dilihat bahwa Negara Pemesan Terbanyak berasal dari negara-negara Benua Amerika dan Benua Eropa.

Dari kesimpulan diatas, kita dapat mempertimbangkan untuk membangun kantor pusat pada Benua Amerika dan Benua Eropa. Dimana negara yang dapat dijadikan kantor pusat adalah USA dan Germany (sebagai negara pemesan terbanyak). Hal ini bertujuan agar proses produksi dan distribusi dari negara tujuan pembeli dapat dilakukan dalam waktu cepat sehingga akan meningkatkan keuntungan jika pasar dari pelanggan ini tidak bergeser.

Sebenarnya tidak ada masalah bagi perusahaan untuk mencoba pangsa pasar di benua lainnya (Asia, Afrika, Australia). Namun perusahaan mungkin harus menghadirkan produk makanan yang sesuai dengan makanan khas pada setiap benua tersebut atau dapat menyuplai produk ke perusahaan-perusahaan makanan barat di benua-benua tersebut.

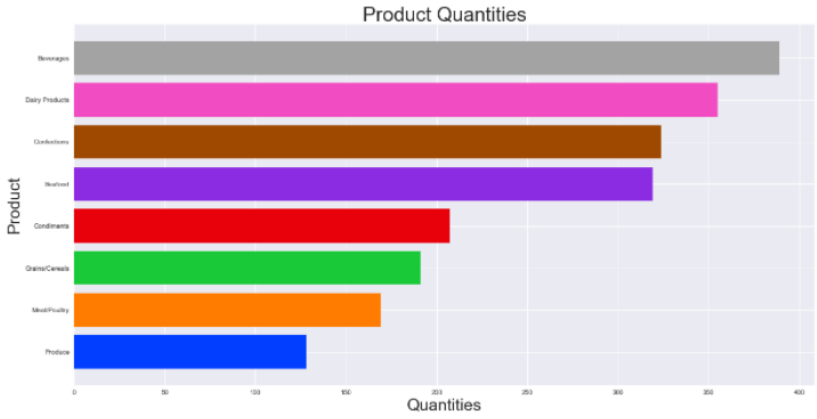
3. Bagaimana persentase status Level Reorder dengan Produksi Barang? Apakah strategi yang diambil oleh perusahaan sudah menempuh jalan yang benar?



Telah kita ketahui sebelumnya bahwa ada beberapa barang yang Discontinued atau tidak di produksi kembali dengan maksud yang telah disebutkan. Dimana kemungkinan alasan Discontinued barang tersebut tidak terlepas dari dikitnya ReorderLevel pada barang tersebut (pangsa pasar pada produk tersebut mengalami penurunan). Pada pie chart pertama kita dapat melihat bahwa banyak produk pada data transaksi tidak mengalami reorder (pemesanan kembali) pada produk tersebut, dimana didalamnya terdapat 8 produk yang semuanya berstatus Discontinued oleh perusahaan (*lihat tabel 3).

Jika melihat nilai pada kedua pie chart tersebut, bisa disimpulkan bahwa perusahaan mungkin benar untuk men-cut 8 buah produk tersebut apabila perusahaan ingin memperbesar persentase reorder level transaksi dan berfokus untuk memaksimalkan profit dengan memproduksi dan menjual barang yang masih memiliki status reorder level yang masih tinggi.

4. Jenis product apa yang paling laku dan paling banyak terjual? Berikanlah analisisnya jika ada

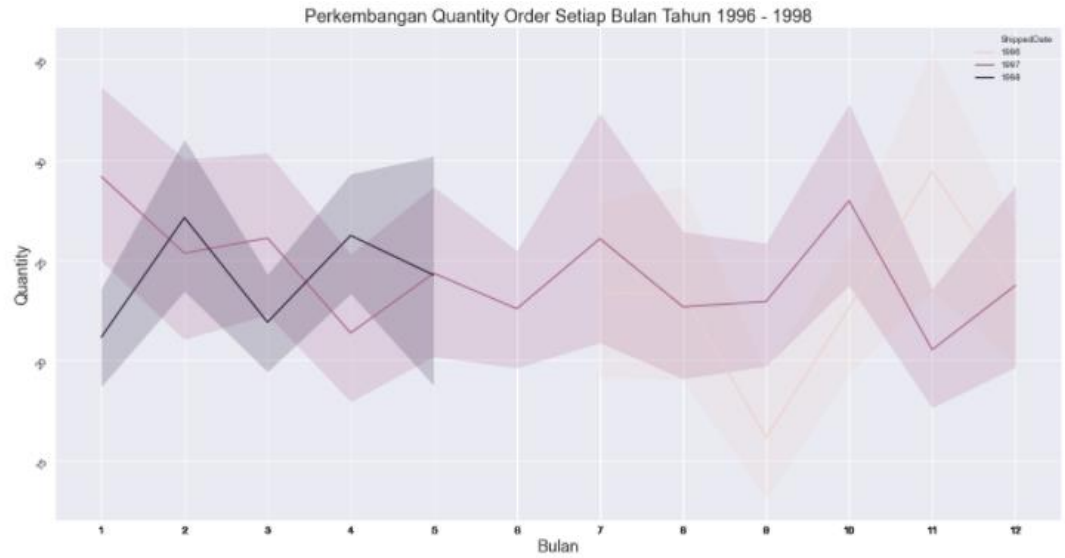


Pada grafik diatas, kita alihkan fokus pada produk yang dijual dan ditawarkan. Terlihat top 3 product yang paling banyak dipesan adalah Beverages, Dairy Products dan Confections. Jika melihat jenis makanannya (Soft drinks, coffees, teas, beers, ales, Cheeses, Sweet and savory sauces, relishes, spreads, seasonings,), ketiga product line yang menjadi top 3 ini adalah jenis makanan yang normal atau pada umumnya dibeli dan digunakan pada Benua Amerika dan Benua Eropa. Hal tersebut bisa menjadi alasan wajar kenapa ketiga line product ini menjadi yang paling banyak dibeli.

Melihat keadaan tersebut, pihak perusahaan dapat menjadikan produksi Beverages, Dairy Products dan Confections untuk menjadi line produksi yang di prioritaskan. Hal ini dilakukan karena masih banyak demand pada ketiga line products ini di kedua benua tersebut. Namun mungkin perusahaan harus meakukan riset terlebih dahulu di setiap negara untuk mementukan stok pasaran yang bisa dibuat sehingga stock akan berbanding lurus dengan demand.

5. Pada tahun berapa quantity order terbesar perusahaan didapatkan?

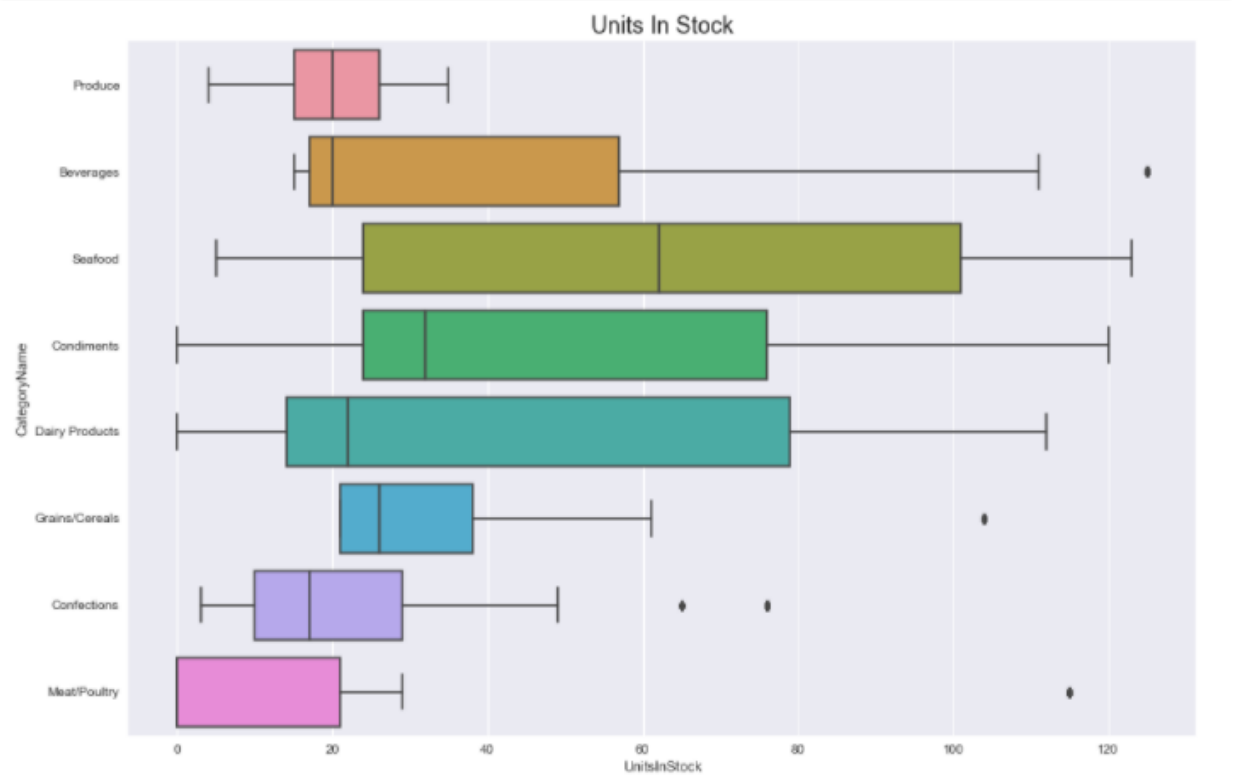
OrderDate	Quantity
1996	9581.0
1997	25489.0
1998	15049.0



Database yang dimiliki saat ini adalah database transaksi penjualan kendaraan dalam rentan tahun 1996 hingga tahun 1998. Mari coba kita lihat line graphic quantity order yang dipesan oleh customers dari tahun ke tahun.

Terlihat jelas pada grafik, penjualan terbesar secara quantity dalam 1 tahun terjadi pada tahun 1997 yakni sebanyak 25489 pieces. Secara grafik diatas, dapat dilihat bahwa pada tahun 1996 dan 1998 nilai penjualan cukup menurun. Namun hal tersebut dikarenakan dari database yang dimiliki, tahun 1996 hanya memiliki range data 6 bulan terakhir dan tahun 1998 hanya memiliki range data 5 bulan terakhir.

6. Tampilkanlah Visualisasi Data Outliers dari variabel UnitsInStock.



7. Ujilah Perbedaan Quantity tiap Product pada dataset. Apakah data berdistribusi normal atau tidak?

```
# Uji Perbandingan Jumlah Quantity Antar Setiap Product Line (Normalitas)

from scipy.stats import shapiro

norm, pval = shapiro(tabel1['Quantity'])

if pval < 0.05 :
    print (f'Tolak H0 Karena P-Value ({pval} < 5%)')
    print ('DATA TIDAK BERDISTRIBUSI NORMAL')
else :
    print (f'Gagal Tolak H0 Karena P-Value ({pval} > 5%)')
    print ('DATA BERDISTRIBUSI NORMAL')
```

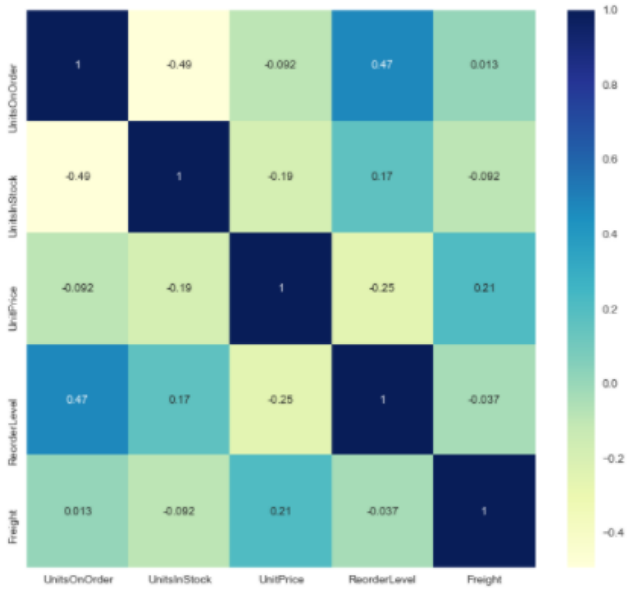
```
Tolak H0 Karena P-Value (3.3119829334163484e-40 < 5%)
DATA TIDAK BERDISTRIBUSI NORMAL
```

8. Ujilah perbandingan jumlah quantity antar setiap product, serta berikan hipotesisnya!

```
# Uji Perbandingan Jumlah Quantity Antar Setiap Product Line (Kruskal Wallis)

from scipy.stats import kruskal

krus, pvalkrus = kruskal(
    tabel1[tabel1['CategoryName'] == 'Beverages']['Quantity'],
    tabel1[tabel1['CategoryName'] == 'Condiments']['Quantity'],
    tabel1[tabel1['CategoryName'] == 'Confections']['Quantity'],
    tabel1[tabel1['CategoryName'] == 'Dairy Products']['Quantity'],
    tabel1[tabel1['CategoryName'] == 'Grains/Cereals']['Quantity'],
    tabel1[tabel1['CategoryName'] == 'Meat/Poultry']['Quantity'],
    tabel1[tabel1['CategoryName'] == 'Produce']['Quantity'],
    tabel1[tabel1['CategoryName'] == 'Seafood']['Quantity']
)

if pvalkrus < 0.05 :
    print (f'Tolak H0 Karena P-Value ({pval} < 5%)')
    print ('Terdapat Perbedaan Nilai Median Quantity pada Setiap Product Yang Ada')
else :
    print (f'Gagal Tolak H0 Karena P-Value ({pval} > 5%)')
    print ('Tidak Terdapat Perbedaan Nilai Median Quantity pada Setiap Product Yang Ada')
```

```
Gagal Tolak H0 Karena P-Value (3.3119829334163484e-40 > 5%)
Tidak Terdapat Perbedaan Nilai Median Quantity pada Setiap Product Yang Ada
```

9. Apakah terdapat hubungan antara 'UnitsOnOrder', 'UnitsInStock', 'UnitPrice', 'ReorderLever' dan Freight? Jika iya, variable mana saja yang paling mempengaruhi satu sama lain? Gambarkan bentuk hubungannya.



Dari heatmap diatas dapat kita lihat bahwa yang memiliki korelasi paling tinggi adalah antara feature Reorder Level dan feature UnitsOnOrder. Hal ini menunjukkan bahwa unit yang sedang di order di dalam dataset kemungkinan besar merupakan produk yang di reorder oleh customers. Oleh karena itu perusahaan sangat disarankan untuk berfokus atau meningkatkan produksi pada produk yang sering dibeli dah dilakukan reorder oleh customers *(untuk melihat line produk apa saja yang sering di dilakukan reorder oleh customers kita sudah membahasnya sebelumnya)*.

Kita juga dapat melihat bahwa adanya sedikit korelasi antara feature UnitPrice dan feature Freight. Hal ini mungkin disebabkan oleh perbedaan penetapan harga pada biaya pengiriman (freight) pada setiap unit makanan. Sehingga pada produk tertentu, UnitPrice akan menyesuaikan harga sesuai dengan biaya pengiriman freight