E-Commerce Shipping Data set

Getting Started :

- Import pandas,numpy,matplotlib and seaborn library
  - Info and describe data set
  - Model Validation
  - Model Visualization

How to import library

Import library to google collabs or you can use Jupyter notebook.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from scipy.stats import kruskal
```

Import Rcparams for default settings

```
from matplotlib import rcParams
rcParams['figure.figsize'] = 12, 4
rcParams['lines.linewidth'] = 3
rcParams['xtick.labelsize'] = 'x-large'
rcParams['ytick.labelsize'] = 'x-large'
```

Importing data

Importing data will be the first step of your project. Whether you have an excel file or csv file, they can easily be imported into pandas. For importing comma separated value (CSV) files, use this line of code.

```
Import dataset
df = pd.read_csv('Train.csv')
```

How to get Informations from your dataset

Provides a summary of the data including the index data type, column data types, non-null values and memory usage.

```
df.info()
```

How to Describe dataset

Provides descriptive statistics that summarizes the central tendency, dispersion, and shape.

```
df.describe()
```

Returns the first 5 rows of the dataframe.

you may insert a value between the parenthesis to change the number of rows returned.
Example: df.head(10) will return 10 rows

```
df.head()
```

Returns the last 5 rows of the dataframe.

you may insert a value between the parenthesis to change the number of rows returned.
Example: df.head(10) will return 10 rows

```
df.tail()
```

How to Model Visualization

Histogram Plot

```
plt.figure(figsize=(20,12))
for i, column in enumerate(df.columns, 1):
    plt.subplot(4,3,i)
    sns.histplot(df[column])
    plt.tight_layout()
```

Box Plot

```
plt.figure(figsize=(20,12))
nums = df.select_dtypes(exclude='object')
for i, column in enumerate(nums.columns, 1):
    plt.subplot(4,3,i)
    sns.boxplot(y = nums[column])
    plt.tight_layout()
```

Kdeplot

```
features = numerical
plt.figure(figsize=(15, 5))
for i in range(0, len(features)):
  plt.subplot(2, 4, i+1)
  sns.kdeplot(x=df[features[i]], color='green')
  plt.xlabel(features[i])
  plt.tight_layout()
```

Violin Plot

```
plt.figure(figsize=(15,5))
for i in range(0, len(numerical)):
    plt.subplot(1, len(numerical), i+1)
    sns.violinplot(y=df[numerical[i]], color='green', orient='v')
    plt.tight_layout()
```

Pair Plots (Hue + Numeric)

```
plt.figure(figsize=(10, 10))
sns.pairplot(dfcorr, diag_kind='kde', hue='Late_shipment(cat)')
```

Source

Source	Link
Kaggle Dataset Download	[https://www.kaggle.com/datasets/prachi13/customer-analytics]
Google Collabs	[https://colab.research.google.com/drive/1NUEws22AMUXb8YdyrinVXXZ8o1EyRV0usp=sharing#scrollTo=KM_JhBnGhpvf]

Summary Insight

Descriptive Statistics

- A. Pada kolom Discount_offered kami mengubah tipe data dari int64 menjadi float
- B. Tidak memiliki nilai kosong pada data set
- C. Pada kolom discount_offered minimal nilainya adalah 1 sehingga nilai std jauh lebih besar dibanding nilai mean,
- D. Nilai median lebih kecil dibanding dengan nilai mean sehingga memiliki distribusi positif.

Univariate Analysis

Pada Box Plot A. Customer care call pada boxplot menandakan bahwa pelanggan lebih banyak menelpon ke customer call 3 sampai 5 kali dan kolom customer care call ini tidak memiliki outlier

B. Persebaran data dari kolom Cost of the product pada boxplot berada di harga 180 hingga 250 dan cost of the product ini tidak memiliki outlier

C .Prior purchase pada boxplot pelanggan melakukan repeat order, kebanyakan 3 sampai 4 kali order dan di kolom prior purchase ini terdapat outlier yang cukup extreme di 10 kali order

D. Discount offered pada boxplot pelanggan lebih banyak mengambil discount 1% sampai dengan 10% dan di kolom discount offered ini memiliki outlier yang sangat banyak.

E. Weight in gms pada boxplot berat paket kebanyakan berada di 1900 sampai 5000 dan kolom tersebut tidak memiliki outlier

F. Customer rating pada boxplot perusahaan menilai pelanggan kebanyakan di rating 2 sampai 4 dan pada kolom ini tidak memiliki outlier

G. Late shipment pada boxplot nya terlihat penuh dikarenakan lateshipment ini hanya ada 2 pilihan yaitu 0 dan 1 sehingga late shipment ini bisa dimasukan kedalam categorical

Kdeplot A. Customer care call pada kdeplot kolom tersebut memiliki bentuk multimodal yang dimana kolom tersebut bisa dimasukan kedalam categorikal B.Cost of the product pada kdeplot kolom tersebut memiliki bentuk hampir normal

C.Prior purchase pada kdeplot kolom tersebut memiliki bentuk multimodal yang dimana kolom tersebut bisa dimasukan kedalam categorikal

D. Discount offered pada kdeplot kolom tersebut memiliki bentuk positif skewed dimana nilai mean lebih kecil dari pada median hingga memiliki longtial yang panjang

E. Weight in gms pada kdeplot kolom tersebut memiliki bentuk negative skewed yang tidak sempurna dikarenakan tingginya lonjakan di angka 1900 membuat bentuk kolom tersebut seperti bimodal

F. Customer rating pada kdeplot kolom tersebut memiliki bentuk multimodal yang dimana kolom tersebut bisa dimasukan kedalam categorikal

G. Late shipment pada kdeplot kolom tersebut memiliki bentuk bimodal yang dimana kolom tersebut bisa dimasukan kedalam categorikal

Count Plot A. Warehouse block pada countplot dapat dilihat bahwa warehouse F merupakan gudang dengan jumlah penyimpanan barang terbanyak dibanding warehouse di block lain seperti A, B, C dan D

B. Mode_of_shipment pada countplot dapat dilihat bahwa Mode pengiriman menggunakan kapal menjadi mode pengiriman terbanyak, sementara mode menggunakan jalur darat dan udara memiliki nilai yang hampir sama

C. Product_importance pada countplot dapat dilihat Pelanggan kebanyakan membeli tipe barang low priority, dan ada ketimpangan pada high priority

D. Gender pada countplot dapat dilihat pelanggan paling banyak merupakan wanita namun perbandingannya tidak terlalu jauh dengan pria

E. Reached_on_Time_Y.N pada countplot dapat dilihat bahwa barang saat ini sering kali telat dibanding tepat waktu

F. Customer Rating pada countplot dapat dilihat bahwa rating dari 1 hingga 5 perbandingannya tidak jauh berbeda hanya rating 3 lebih dominan dibandingkan rating yang lain

Multivariate Analysis

A. **Hasil berdasarkan metode heatmap dan kendall tau:**

- Daftar fitur yang berkorelasi positif dengan Late_shipment(num)= Customer_rating(num) dan Discount_offered.
- Daftar fitur yang berkorelasi negatif dengan Late_shipment(num)= Customer_care_calls, Cost_of_the_product, Prior_purchases, dan Weight_in_gms
- Dua fitur terkuat untuk dipasangkan dengan Late_shipment(num), yaitu Discount_offered dan Weight_in_gms.
- Berdasarkan metode heatmap dan kendall tau, Late_shipment(num) dengan Discount_offered memiliki korelasi positif. Semakin tinggi diskon yang ditawarkan pada suatu produk, maka semakin tinggi potensi pengirimannya mengalami keterlambatan.
- Berdasarkan metode heatmap,Late_shipment(num) dengan Weight_in_gms memiliki korelasi negatif. Semakin ringan suatu produk, maka semakin tinggi potensi pengirimannya mengalami keterlambatan.
- Berdasarkan metode heatmap, tidak ditemukan fitur yang berkorelasi kuat dengan nilai di atas 0.7. Meskipun demikian, dua fitur yang berkorelasi mendekati kuat

untuk dipasangkan dengan Late_shipment(num), yaitu Discount_offered dan Weight_in_gms. Berdasarkan metode kendall tau, fitur dengan korelasi terkuat untuk dipasangkan dengan Late_shipment(num) adalah Discount_offered. Meskipun demikian, nilai tersebut masih jauh mendekati satu.

Pendapat Tambahan: **Penggunaan heatmap kurang cocok dipakai dalam kasus ini karena sifat Late_shipsment(num) yang sebenarnya adalah tipe data kategori.**

B. **Kesimpulan**

- Scatter plot dimana kedua warna dapat terpisah dengan baik dan memiliki pola yang jelas sehingga menarik untuk diperhatikan terletak pada kombinasi variabel Late_shipment(num) dengan Discount_offered dan Weight_in_gms.
- Discount_Offered lebih dari 10% mengalami keterlambatan. Hipotesis: Diskon produk di atas 10% mengakibatkan tingginya pembelian dan pengiriman suatu produk, sedangkan kuota pengiriman barang tidak berubah sehingga pengiriman mengalami keterlambatan. Kasus semacam ini disebut dengan overload. Hal ini sering dialami oleh berbagai ekspedisi di dunia.
- Weight_in_gms direntang masa 2000 - 4000 gram mengalami keterlambatan.

Business Insight

Dari analisis yang telah kami lakukan, berikut beberapa business insight yang telah di tentukan :

- Melihat metode pengiriman apa yang sering terlambat Pengiriman menggunakan kapal cenderung mengalami keterlambatan dalam pengiriman. Hal itu wajar dikarenakan pengiriman paling banyak dengan moda kapal. Dengan itu, perusahaan bisa lebih mengurangi pengiriman moda kapal dengan menggunakan moda pengiriman lainnya. Sehingga, meminimalisir penumpukan dan keterlambatan dalam pengiriman barang. Pengiriman barang berdasarkan prioritas dan status pengiriman
- Berdasarkan dari product important, dapat dilihat pengiriman dengan tipe low priority lebih banyak mengalami keterlambatan dalam pengiriman. Dikarenakan, customer lebih banyak melakukan pembelian dengan tipe barang low priority. Untuk itu, perusahaan dapat mengatur pengiriman berdasarkan prioritas dan status pengiriman yang mengalami penumpukan.
- Hubungan berat dengan keterlambatan dan merekomendasikan pengiriman berdasarkan berat barang Pengiriman barang dengan berat < 4000 gram cenderung mengalami keterlambatan, dengan itu perlu digali lebih lanjut data pengiriman barang dengan berat < 4000 gram dan 2000-4000 gram. Sehingga, kita bisa merekomendasikan pengiriman berdasarkan berat barang.
- Membuat diskon jadi lebih menarik, gunakan diskon 5%, 10%, dan 15%
- Memberikan diskon kepada customer dengan prior purchase lebih dari 3, dengan menggunakan diskon 5% 10% dan 15% agar lebih menarik dan customer kemungkinan untuk repeat order lagi akan tinggi.