Exploratory Data Analysis (EDA)

Widia Mulya Hartanti

Portofolio

01 - Memasukkan dataset

02 - Mengecek data yang hilang (missing value)

03 - Mengecek dan mengatasi data duplicate

01 - Memasukkan dataset

```
[1]  # Mengimport library pandas untuk membaca dataset
     import pandas as pd

     # Membaca dataset
     df = pd.read_csv('/content/FMCG_2022_2024.csv')
```

```
[2]  df
```

	date	sku	brand	segment	category	channel	region	pack_type	price_unit	promotion_flag	delivery_days	stock_available	delivered_qty	units_sold
0	2022-01-21	MI-006	MiBrand1	Milk-Seg3	Milk	Retail	PL-Central	Multipack	2.38	0.0	1.0	141.0	128.0	9.0
1	2022-01-21	MI-006	MiBrand1	Milk-Seg3	Milk	Retail	PL-North	Single	1.55	1.0	3.0	0.0	129.0	0.0
2	2022-01-21	MI-006	MiBrand1	Milk-Seg3	Milk	Retail	PL-South	Carton	4.00	0.0	5.0	118.0	161.0	8.0
3	2022-01-21	MI-006	MiBrand1	Milk-Seg3	Milk	Discount	PL-Central	Single	5.16	0.0	2.0	81.0	114.0	7.0
4	2022-01-21	MI-006	MiBrand1	Milk-Seg3	Milk	Discount	PL-North	Single	7.66	0.0	4.0	148.0	204.0	12.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
45145	2023-04-03	MI-023	MiBrand3	Milk-Seg3	Milk	Retail	PL-Central	Carton	1.59	0.0	4.0	114.0	180.0	12.0
45146	2023-04-03	MI-023	MiBrand3	Milk-Seg3	Milk	Retail	PL-North	Single	8.44	0.0	2.0	220.0	225.0	20.0
45147	2023-04-03	MI-023	MiBrand3	Milk-Seg3	Milk	Retail	PL-South	Single	2.34	1.0	4.0	150.0	212.0	31.0
45148	2023-04-03	MI-023	MiBrand3	Milk-Seg3	Milk	Discount	PL-Central	Multipack	2.13	0.0	3.0	81.0	117.0	8.0
45149	2023-04-03	M	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

45150 rows × 14 columns

Sumber: https://www.kaggle.com/datasets/beatafaron/fmcg-daily-sales-data-to-2022-2024/data

Ringkasan data

Menggunakan syntax info() untuk memahami struktur dan konten DataFrame dalam membantu eksplorasi dan persiapan data.

- DataFrame terdiri dari 45150 entri dengan indeks dari 0 sampai 45149.
- 8 kolom data bertipe object: date, sku, brand, segment, category, channel, region, dan pack_type.
- 6 kolom data bertipe float64: price_unit, promotion_flag, delivery_days, stock_available, delivered_qty, dan units_sold.

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 45150 entries, 0 to 45149
Data columns (total 14 columns):
 #   Column           Non-Null Count  Dtype  
---  ------           --------------  -----  
 0   date             45150 non-null  object 
 1   sku              45150 non-null  object 
 2   brand            45149 non-null  object 
 3   segment          45149 non-null  object 
 4   category         45149 non-null  object 
 5   channel          45149 non-null  object 
 6   region           45149 non-null  object 
 7   pack_type        45149 non-null  object 
 8   price_unit       45149 non-null  float64
 9   promotion_flag   45149 non-null  float64
 10  delivery_days    45149 non-null  float64
 11  stock_available  45149 non-null  float64
 12  delivered_qty    45149 non-null  float64
 13  units_sold       45149 non-null  float64
dtypes: float64(6), object(8)
memory usage: 4.8+ MB
```

02 - Mengecek missing value

Berdasarkan informasi tersebut, didapatkan beberapa data yang hilang pada kolom brand, segment, category, channel, region, pack_type, price_unit, promotion_flag, delivery_days, stock_available, delivered_qty, dan units_sold

```
# mengecek missing value
df.isna().sum()
```

	0
date	0
sku	0
brand	1
segment	1
category	1
channel	1
region	1
pack_type	1
price_unit	1
promotion_flag	1
delivery_days	1
stock_available	1
delivered_qty	1
units_sold	1

dtype: int64

Mengatasi missing value

```python
# Mengatasi missing value
for column in df.columns:
    if df[column].dtype == 'object':
        # Jika kolom bertipe object, isi dengan mode
        df[column].fillna(df[column].mode()[0], inplace=True)
    else:
        # Jika kolom bertipe numerik, isi dengan mean
        df[column].fillna(df[column].mean(), inplace=True)
```

Untuk mengatasi missing value di mana data bertipe object akan diisi dengan modus/nilai yang sering muncul. Jika data bertipe numerik akan diisi dengan mean/nilai rata-rata.

```python
# Cek kembali missing value
df.isna().sum()
```

	0
date	0
sku	0
brand	0
segment	0
category	0
channel	0
region	0
pack_type	0
price_unit	0
promotion_flag	0
delivery_days	0
stock_available	0
delivered_qty	0
units_sold	0

dtype: int64

Hasil penyelesaikan missing value

Observasi data

```
# Cek statiscial summary
df.describe()
```

	price_unit	promotion_flag	delivery_days	stock_available	delivered_qty	units_sold
count	45149.000000	45149.000000	45149.000000	45149.000000	45149.000000	45149.000000
mean	5.246357	0.149549	3.008992	158.064210	179.441272	21.001728
std	2.171060	0.356633	1.414201	52.761184	40.159331	12.596743
min	1.500000	0.000000	1.000000	0.000000	1.000000	0.000000
25%	3.360000	0.000000	2.000000	124.000000	152.000000	13.000000
50%	5.240000	0.000000	3.000000	156.000000	180.000000	18.000000
75%	7.130000	0.000000	4.000000	192.000000	207.000000	26.000000
max	9.000000	1.000000	5.000000	358.000000	366.000000	133.000000

Output syntax df.describe() menunjukkan bahwa data di atas memiliki nilai minimum dan maksimum untuk semua kolom yang masuk akal. Kolom delivery_days memiliki distribusi normal sementara kolom price_unit, stock_available, units_sold, dan delivered_qty memiliki skewness yang jelas yaitu right-skewed. Kolom promotion_flag adalah kolom boolean/biner karena nilainya 0 atau 1, sehingga tidak perlu menyimpulkan kesimetrisannya, hanya perlu memeriksa balance level.

03 - Mengecek data duplicate

```
# Mengecek apakah ada duplicate di seluruh kolom
check_duplicate = df.duplicated().sum()

print(f"Jumlah data yang duplikat = {check_duplicate}")
```

```
Jumlah data yang duplikat = 0
```

Berdasarkan gambar tersebut diketahui bahwa tidak ada data duplikat

Mengatasi data duplicate

```
# Handling duplicate
df = df.drop_duplicates()
```

```
[ ] # Mengecek duplicate setelah di-handle
handle_duplicate = df.duplicated().sum()

print(f"Jumlah data yang duplikat = {handle_duplicate}")
```

Data duplicate dapat diatasi dengan syntax tersebut, di mana keseluruhan kolom data akan dicek lalu menghapus data yang duplikat.

Thanks

Widia Mulya Hartanti