

# Exploratory Data Analysis

## Part II

| Bootcamp Data Science



# Kurnia Anwar Ra'if

A Highly-motivated Data Scientist

Senior Data & AI Platform @ PT. Mastersystem Infotama

Data Scientist @ PT. KitaLulus International

Data Scientist @ PT. Sharing Vision– BRI Consultant

Software Engineering @ PT. AILIMA Geothermal

Mentor & Instructor DS/BI/AI ML @ dibimbing.id



# Outline



## Outline:

1. Sampling and Randomization
2. Feature Engineering
3. Data Manipulation \*

\* Combine with visualizations



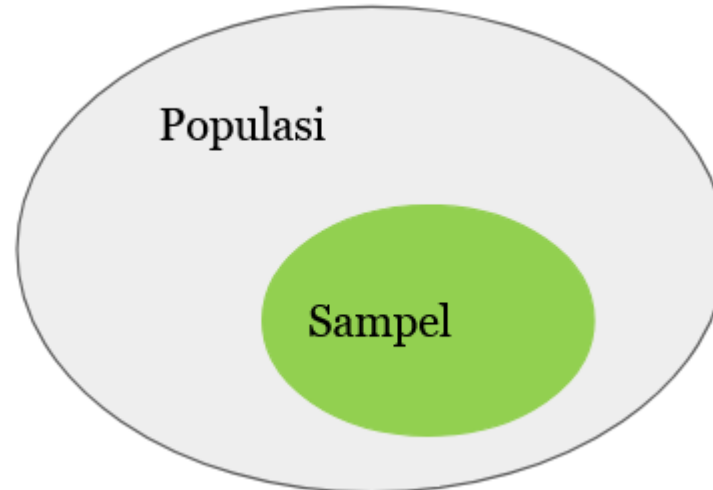
# Sampling & Randomization



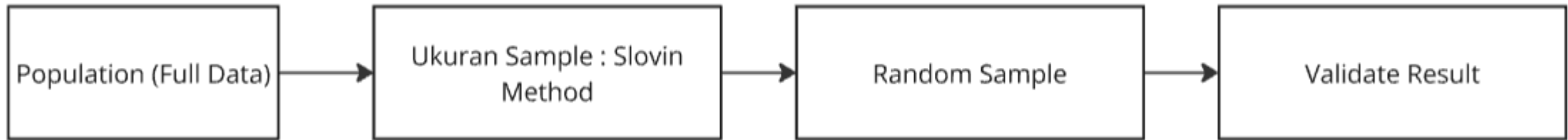
# Definisi Populasi dan Sampel

Populasi adalah **seluruh kumpulan data atau objek** yang menjadi fokus dari suatu penelitian atau studi. Dalam kasus Big Data, Populasi merupakan sekumpulan data dari database pada periode tertentu.

Sampel adalah **sebagian kecil atau subset dari populasi** yang dipilih untuk mewakili populasi



# Flowchart Sampling & Randomization



Proses sampling ini bisa dilakukan **sebelum melakukan exploratory data analysis**. Hal ini bertujuan untuk menghindari *out of memory pada server*.

# Bagaimana metode untuk mencari jumlah sample (*Sample size*) ?

Dalam mengukur jumlah sampel yang ingin diambil kita bisa gunakan persamaan pada metode Slovin :

Rumus Slovin's :

$$n = \frac{N}{(1 + N * error^2)}$$

keterangan:

- n = Jumlah sample
- N = Jumlah population
- error = error yang masih diterima (pilih 1%, 2.5%, atau 5%)

# Random Sampling (Randomization)

## **random sample (randomization)**

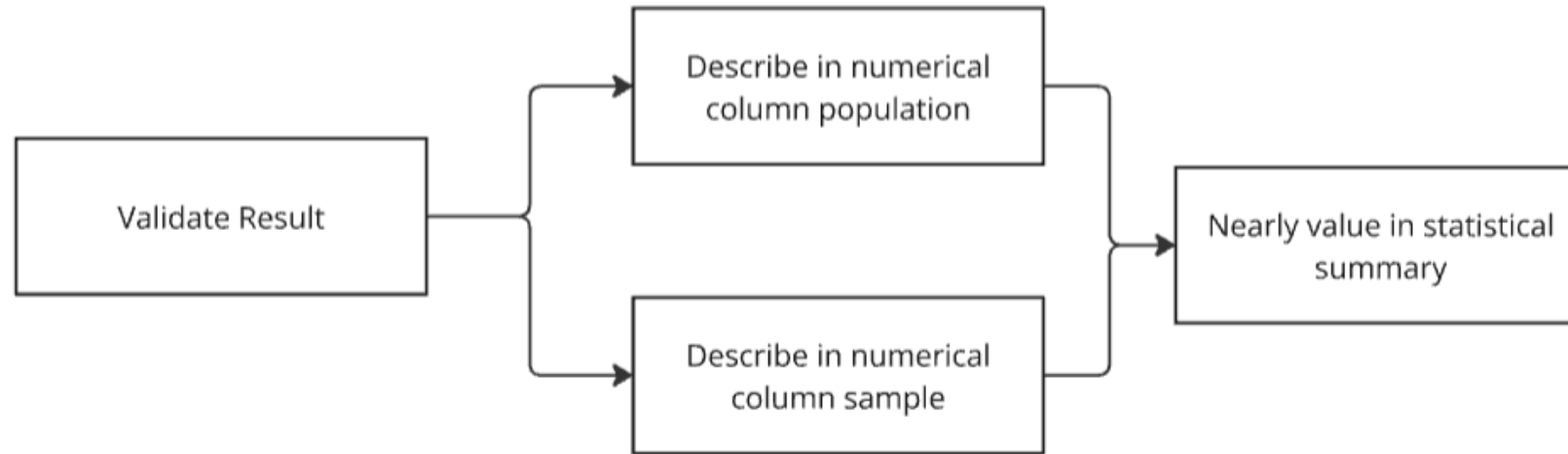
**Setiap anggota dan setiap kelompok anggota memiliki kesempatan yang sama untuk dimasukkan dalam sampel dan dipilih secara acak**

- Contoh Penerapan: Mengambil sampel acak dari kumpulan karyawan di sebuah perusahaan di mana setiap karyawan memiliki kesempatan yang sama untuk dipilih.



# Validate Result

We can check the result with checking in numerical value by describe()



Population Statistical Summary

	Total	cogs	gross income	Unit price
count	1000.000000	1000.00000	1000.000000	1000.000000
mean	322.966749	307.58738	15.379369	55.672130
std	245.885335	234.17651	11.708825	26.494628
min	10.678500	10.17000	0.508500	10.080000
25%	124.422375	118.49750	5.924875	32.875000
50%	253.848000	241.76000	12.088000	55.230000
75%	471.350250	448.90500	22.445250	77.935000
max	1042.650000	993.00000	49.650000	99.960000

Sample Statistical Summary

	Total	cogs	gross income	Unit price
count	616.000000	616.000000	616.000000	616.00000
mean	324.823074	309.355308	15.467765	55.68625
std	246.352673	234.621594	11.731080	26.70841
min	10.678500	10.170000	0.508500	10.13000
25%	125.695500	119.710000	5.985500	33.27500
50%	257.911500	245.630000	12.281500	54.61000
75%	478.584750	455.795000	22.789750	78.17500
max	1042.650000	993.000000	49.650000	99.96000

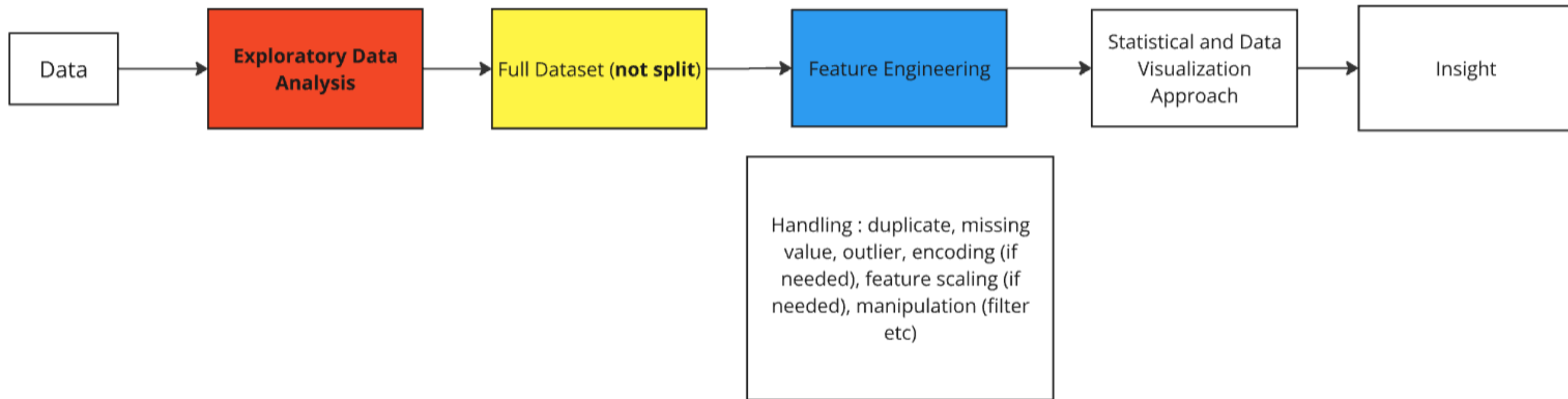
A decorative graphic in the top-left corner consisting of a 3x3 grid of squares. The top row has a yellow square, a grey square, and a dark blue square. The middle row has a grey square, a dark blue square, and a grey square. The bottom row has an orange square, a grey square, and a yellow square.

# Hands on

Balik ke Collabs lagi yuks 😊



# Remember Feature Engineering



A decorative graphic in the top-left corner consisting of a 3x3 grid of squares. The top row has a yellow square, a grey square, and a dark blue square. The middle row has a grey square, a dark blue square, and a grey square. The bottom row has an orange square, a grey square, and a yellow square.

# Hands on

Balik ke Collabs lagi yuks 😊



# Data Manipulation + Visualization



# Remember, EDA Question Below based on course Data Manipulation


**Notes :** harap bisa mengecek kembali pada sesi hands on code Data manipulation

## **EDA Question :**

1. Apa saja produk yang memiliki harga unit di atas rata-rata? **(in Part Filtering)**
2. Produk mana yang memiliki Total penjualan tertinggi ? **(in Group By 1 Kolom)**
3. Cabang mana yang memiliki total penjualan tertinggi ? **(in Group By 2 Kolom)**
4. Bagaimana jumlah maksimal, minimal, jumlah dari barang yang terjual di tiap Branch ? **(in Multiple aggregations)**
5. Bagaimana total penjualan setiap kategori produk di setiap branch ? **(in Pivoting) \***
6. Metode pembayaran apa yang paling sering digunakan oleh customer ? **(in Crosstab) \***
7. Kapan waktu dengan penjualan tertinggi dalam sehari ? **(In working with date)**

\* Additional Pivoting vs Melt

# Code tips Understanding in EDA

- 1. Buat pertanyaan analisis (EDA Questions)**
  - 2. Jawablah pertanyaan EDA dengan pendekatan Filter, Group By, Pivot, dsb (disesuaikan)**
  - 3. Lakukan reset index dari step 2 diatas, supaya outputnya menjadi dataframe**
  - 4. Gunakan plot visualisasi yang tepat (biasanya barplot atau lineplot, atau disesuaikan saja)**
  - 5. Tambahkan text angka (opsional) pada tiap barplot atau titik pada lineplot untuk mempermudah Analisa**
  - 6. Buatlah insight observasi dari output yang didapat.**
- 

# Code tips Understanding in EDA

1. **Pertanyaan : Produk mana yang memiliki Total penjualan tertinggi ?**
2. **Jawab dengan data manipulation ? (Group By)**

```
penjualan_produk = data.groupby('Product line')['Total'].sum()  
penjualan_produk
```

```
Product line  
Electronic accessories    54337.5315  
Fashion accessories       54305.8950  
Food and beverages        56144.8440  
Health and beauty         49193.7390  
Home and lifestyle        53861.9130  
Sports and travel         55122.8265  
Name: Total, dtype: float64
```

3. **Lakukan reset\_index() supaya menjadi dataframe**

```
penjualan_produk = data.groupby('Product line')['Total'].sum().sort_values(ascending=False).reset_index()  
penjualan_produk
```

	Product line	Total
0	Food and beverages	56144.8440
1	Sports and travel	55122.8265
2	Electronic accessories	54337.5315
3	Fashion accessories	54305.8950
4	Home and lifestyle	53861.9130
5	Health and beauty	49193.7390

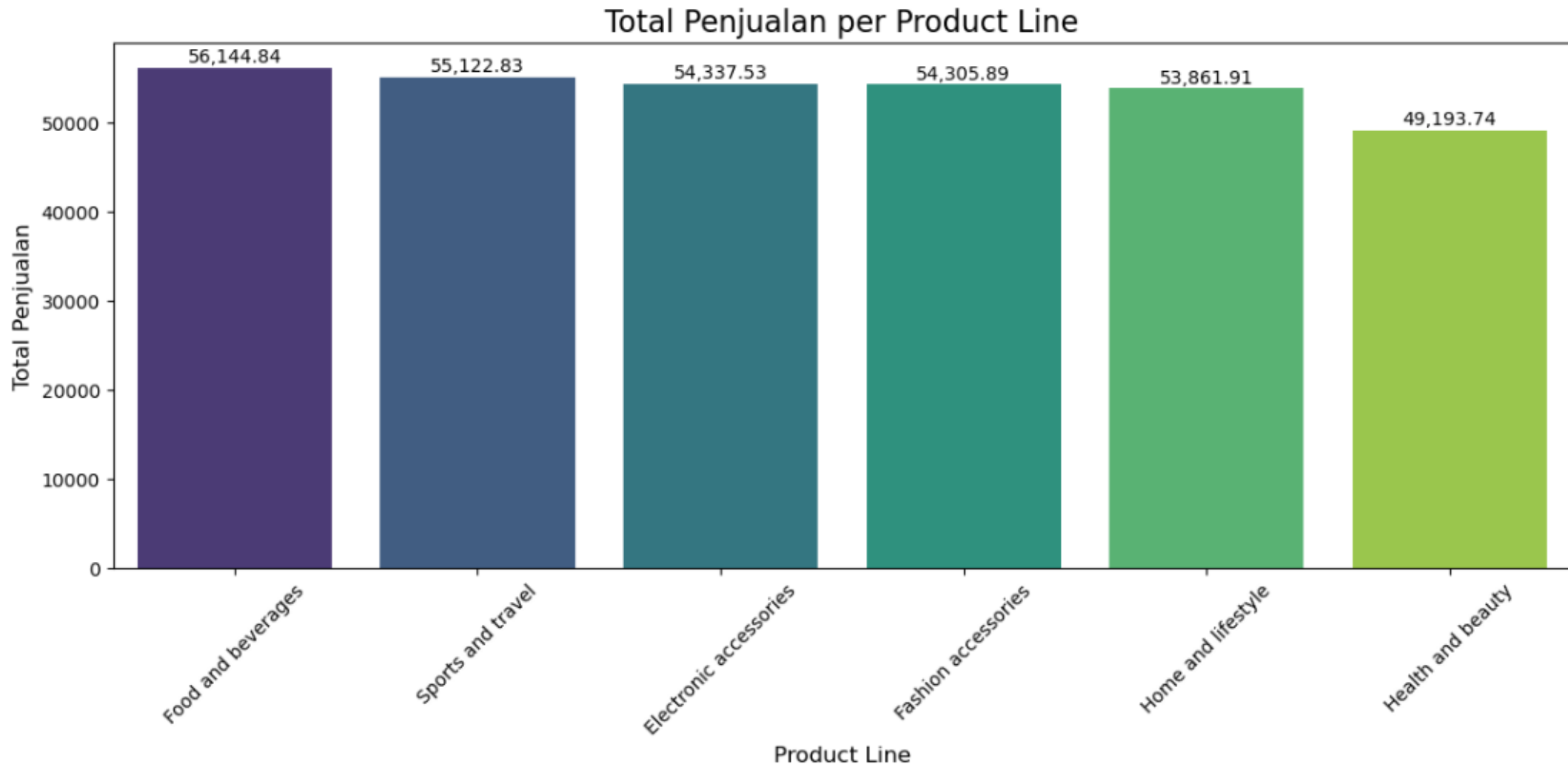


# Code tips Understanding in EDA

## 4. Gunakan plot visualisasi yang tepat :

### 4A. Tips Barplot :

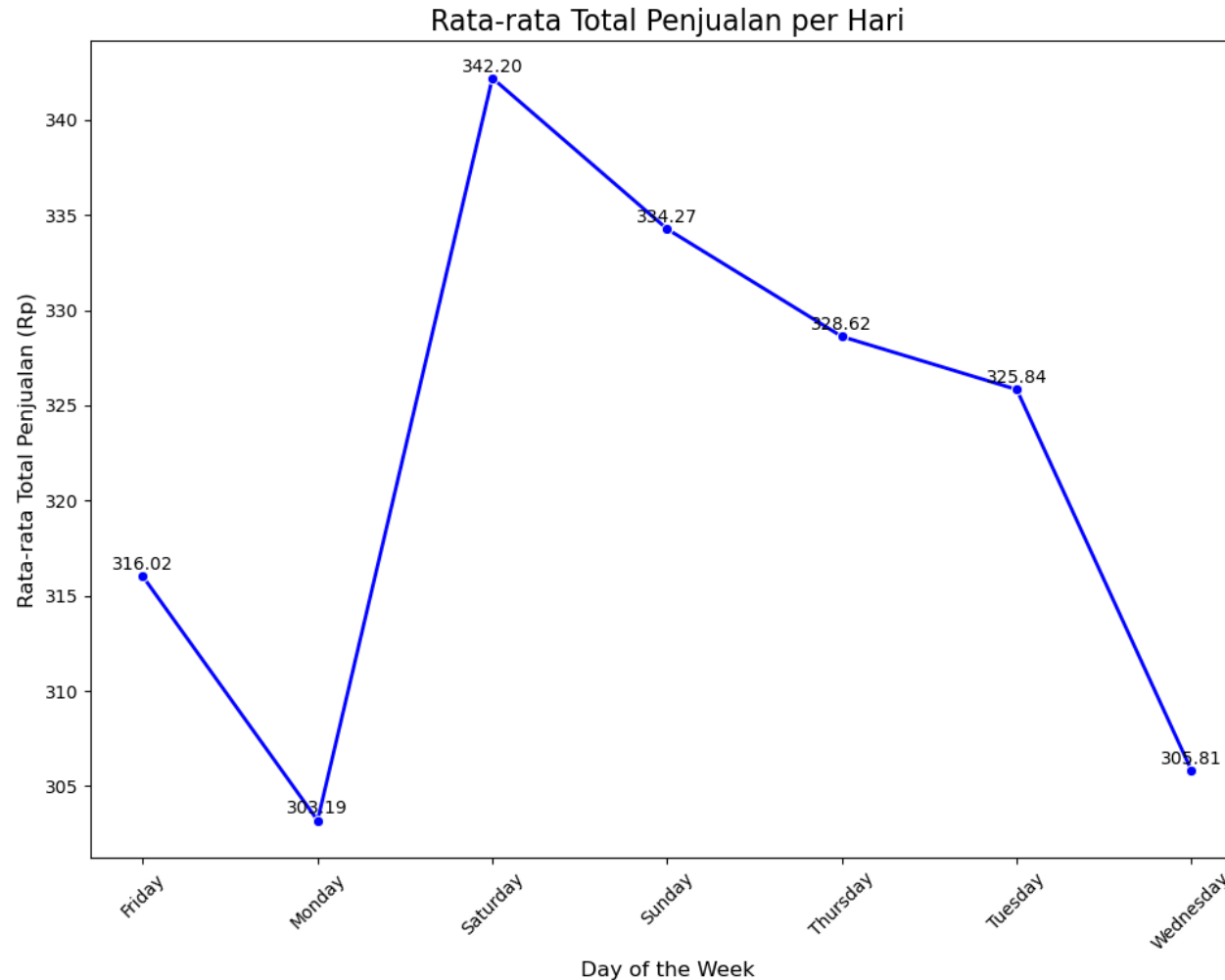
#### - Value vs Categorical (Value in each Categorical)



# Code tips Understanding in EDA

## 4B. Lineplot :

- **Datetime vs Value**
- Datetime can be : day, month, week, year, time (hours, minutes, etc)



# Code tips Understanding in EDA

## 5. Tambahkan text angka (opsional) pada tiap barplot atau titik pada lineplot untuk mempermudah Analisa

	Product line	Total
0	Food and beverages	56144.8440
1	Sports and travel	55122.8265
2	Electronic accessories	54337.5315
3	Fashion accessories	54305.8950
4	Home and lifestyle	53861.9130
5	Health and beauty	49193.7390

**Notes :** Jika barplot terlalu banyak maka tidak perlu ada nilai teks yang ditampilkan.

```
# Visualisasi barplot
plt.figure(figsize=(12, 6))
sns.barplot(x='Product line', y='Total', data=penjualan_produk, palette='viridis')

# Tambahkan angka di atas bar
for i, total in enumerate(penjualan_produk['Total']):
    plt.text(i, total + (0.01 * penjualan_produk['Total'].max()), f"{total:,.2f}",
             ha='center', fontsize=10, color='black')
```

Data suatu kolom

Letak di tiap bar

Ukuran font

Jarak dengan barplot

2 decimal saja

A decorative graphic in the top-left corner consisting of a 3x3 grid of squares. The top row has a yellow square, a grey square, and a dark blue square. The middle row has a grey square, a dark blue square, and a grey square. The bottom row has an orange square, a grey square, and a yellow square.

# Hands on

Balik ke Collabs lagi yuks 😊



# **Additional : Pivoting VS Melt**



# Pivot Table

Konsep Pivot table mirip dengan group by, tetapi pivot table memerlukan **index, colom, value dan fungsi aggregate**

Jika kita memiliki table berikut :

```
# Membuat pivot table
tabel_pivot = pd.pivot_table(df, values='Total', index='Product line', columns='Branch', aggfunc='sum')
tabel_pivot
```

Branch	A	B	C
Product line			
Electronic accessories	18317.1135	17051.4435	18968.9745
Fashion accessories	16332.5085	16413.3165	21560.0700
Food and beverages	17163.1005	15214.8885	23766.8550
Health and beauty	12597.7530	19980.6600	16615.3260
Home and lifestyle	22417.1955	17549.1645	13895.5530
Sports and travel	19372.6995	19988.1990	15761.9280

**index:** Kolom yang akan dijadikan indeks dalam pivot table.

**columns:** Kolom yang akan dijadikan kolom dalam pivot table.

**values:** Kolom yang akan dihitung (dalam contoh ini, 'Nilai').

**aggfunc:** Fungsi agregasi yang akan digunakan. Dalam kasus ini, rata-rata (mean) digunakan, tetapi Anda juga bisa menggunakan fungsi lain seperti 'sum', 'count', 'max', 'min', dll.

# Melt

- Melt is the **reverse** of pivot table
  - I.e. you have “wide” formatted dataframe and want to make it “long”
- This is sometimes useful to prepare data for visualization
  - Which requires “long” formatted data, e.g. seaborn package
- General syntax format

```
df.melt(id_vars,      _____→ Column(s) that remain as index (rows)
        var_name,    _____→ Resulting column name of melted columns
        value_name  _____→ Name of the values column
        )
```

Dari table pivot sebelumnya, kita akan mengubah bentuknya kedalam format melt berikut :

## Pivot Table

Branch	A	B	C
Product line			
Electronic accessories	18317.1135	17051.4435	18968.9745
Fashion accessories	16332.5085	16413.3165	21560.0700
Food and beverages	17163.1005	15214.8885	23766.8550
Health and beauty	12597.7530	19980.6600	16615.3260
Home and lifestyle	22417.1955	17549.1645	13895.5530
Sports and travel	19372.6995	19988.1990	15761.9280



Reset index

```
tabel_pivot.reset_index()
```

Branch	Product line	A	B	C
0	Electronic accessories	18317.1135	17051.4435	18968.9745
1	Fashion accessories	16332.5085	16413.3165	21560.0700
2	Food and beverages	17163.1005	15214.8885	23766.8550
3	Health and beauty	12597.7530	19980.6600	16615.3260
4	Home and lifestyle	22417.1955	17549.1645	13895.5530
5	Sports and travel	19372.6995	19988.1990	15761.9280



Melt

```
# Mengubah pivot table menjadi format long untuk visualisasi
tabel_long = tabel_pivot.reset_index().melt(id_vars='Product line', var_name='Branch', value_name='Total')
tabel_long
```

	Product line	Branch	Total
0	Electronic accessories	A	18317.1135
1	Fashion accessories	A	16332.5085
2	Food and beverages	A	17163.1005
3	Health and beauty	A	12597.7530
4	Home and lifestyle	A	22417.1955
5	Sports and travel	A	19372.6995
6	Electronic accessories	B	17051.4435
7	Fashion accessories	B	16413.3165
8	Food and beverages	B	15214.8885
9	Health and beauty	B	19980.6600
10	Home and lifestyle	B	17549.1645
11	Sports and travel	B	19988.1990
12	Electronic accessories	C	18968.9745
13	Fashion accessories	C	21560.0700
14	Food and beverages	C	23766.8550
15	Health and beauty	C	16615.3260
16	Home and lifestyle	C	13895.5530
17	Sports and travel	C	15761.9280

## Melt Table



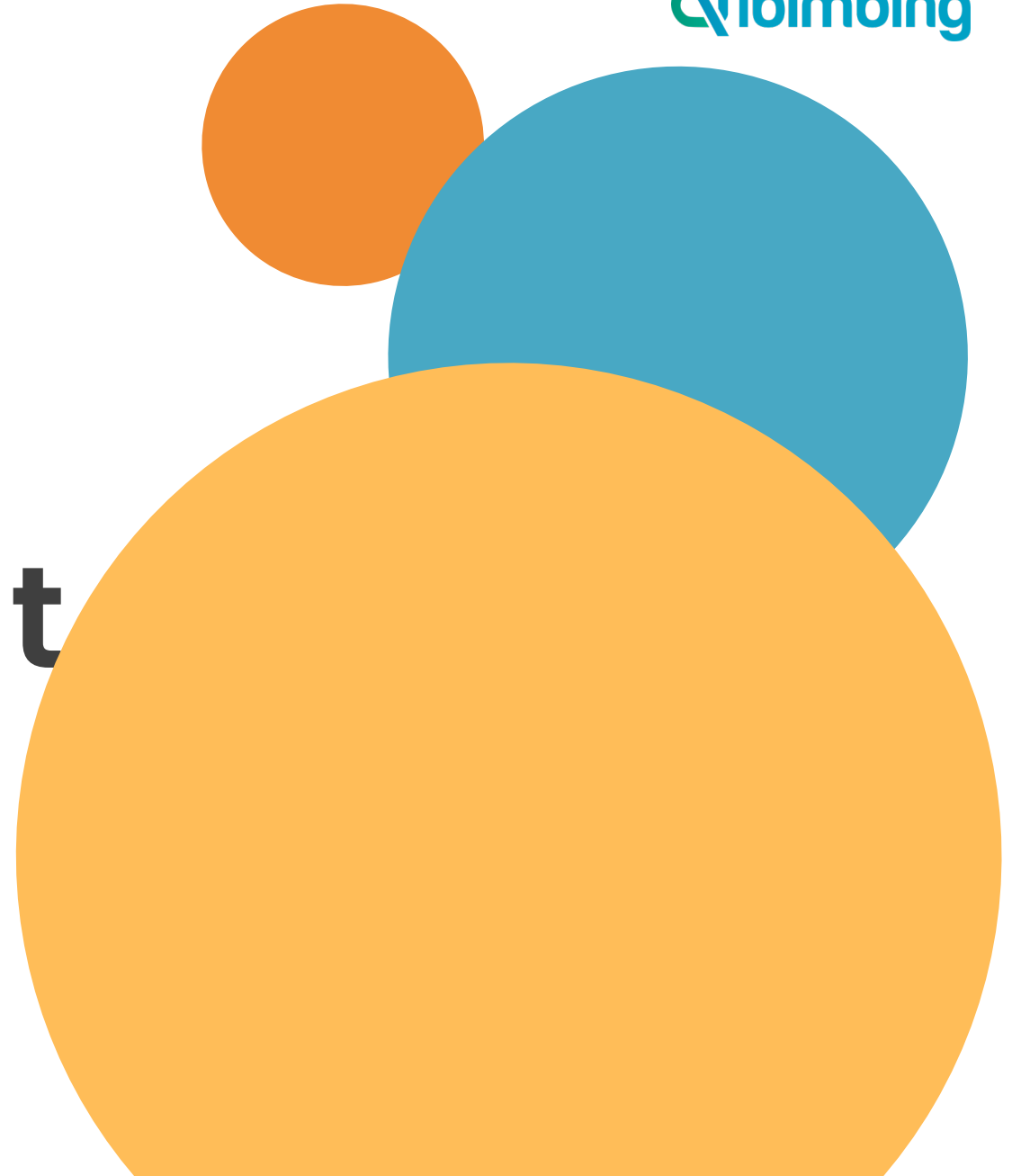
A decorative graphic in the top-left corner consisting of a 3x3 grid of squares. The top row has a yellow square, a grey square, and a dark blue square. The middle row has a grey square, a dark blue square, and a grey square. The bottom row has an orange square, a grey square, and a yellow square.

# Hands on

Balik ke Collabs lagi yuks 😊



# **Additional : Crosstab VS Melt**



# Crosstab (Cross Tabulation)

Crosstabulation, atau crosstab, adalah sebuah metode statistik yang digunakan untuk menganalisis hubungan antara **dua atau lebih variabel kategorikal**. **Value didalamnya bermakna frekuensi**.

Jika kita memiliki table berikut :

```
crosstab = pd.crosstab(df['Branch'], df['Payment'])
crosstab
```

Payment	Cash	Credit card	Ewallet
Branch			
A	110	104	126
B	110	109	113
C	124	98	106

Cross tab



Reset index

```
crosstab.reset_index()
```

Payment	Branch	Cash	Credit card	Ewallet
0	A	110	104	126
1	B	110	109	113
2	C	124	98	106



Melt

```
crosstab_long = crosstab.reset_index().melt(id_vars='Branch', var_name='Payment', value_name='Jumlah Pembayaran')
crosstab_long
```

	Branch	Payment	Jumlah Pembayaran
0	A	Cash	110
1	B	Cash	110
2	C	Cash	124
3	A	Credit card	104
4	B	Credit card	109
5	C	Credit card	98
6	A	Ewallet	126
7	B	Ewallet	113
8	C	Ewallet	106

Melt

A decorative graphic in the top-left corner consisting of a 3x3 grid of squares. The top row has a yellow square, a grey square, and a dark blue square. The middle row has a grey square, a dark blue square, and a grey square. The bottom row has an orange square, a grey square, and a yellow square.

# Hands on

Balik ke Collabs lagi yuks 😊



A large graphic on the left side of the slide, consisting of a blue circle containing a large orange circle, which in turn contains a smaller orange circle, creating a stylized '8' or '0' shape.

# Thank you



<https://www.linkedin.com/in/anwaraif/>



[kurniafreelancer@gmail.com](mailto:kurniafreelancer@gmail.com)