Eksplorasi Data

(Boxplot, Histogram, Scatter Plot. Dan Heat Map)

BoxPlot

Percentiles

- For continuous data, the notion of a percentile is more useful.

Given an ordinal or continuous attribute x and a number p between 0 and 100, the p^{th} percentile is a value x_p of x such that p%of the observed values of x are less than x_p

- For instance, the 50th percentile is the value $x_{50\%}$ such that 50% of all values of x are less than $x_{50\%}$

persentil ke-i dari data bergolong dirumuskan sebagai berikut.

$$P_i = b + l\left(\frac{\frac{i \cdot n}{100} - F}{f}\right)$$

Keterangan:

Pi = persentil ke-i

b = tepi bawah

n = banyaknya data

F = frekuensi kumulatif kelas sebelum kelas persentil

f = frekuensi kelas persentil

l = lebar kelas

- Contoh :

Contoh soal

Diketahui: 9, 10, 11, 6, 8, 7, 7, 5, 4, 5, tentukan persentil ke-30 dan persentil ke-75.

Penyelesaian

Data diurutkan: 4, 5, 5, 6, 7, 7, 8, 9, 10, 11

Letak persentil ke-30 di urutan data ke- 30(10 +1)/100 = 330/100 = 3,3.

P30 = x3 + 0,3 (x4 – x3) = 5 + 0,3 (6 – 5) = 5,3

Jadi, P30 = 5,3.

Letak persentil ke-75 di urutan data ke- 75(10 +1)/100 = 8,25.

P75 = x8 + 0,25 (x9 – x8) = 9 + 0,25 (10 – 9) = 9,25

Jadi, P75 = 9,25.

Contoh:

x	f
41 – 45	3
46 – 50	6
51 – 55	16
56 – 60	8
61 – 65	7

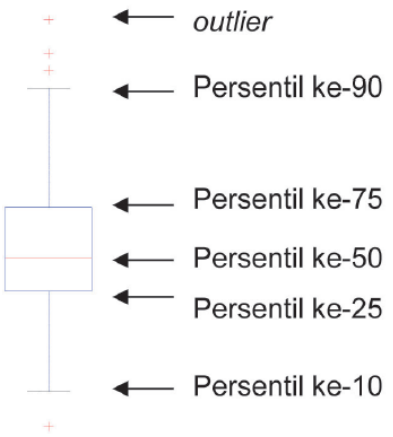
Penyelesaian:

x	f	F kumulatif
41 – 45	3	3
46 – 50	6	9
51 – 55	16	25
56 – 60	8	33
61 – 65	7	40

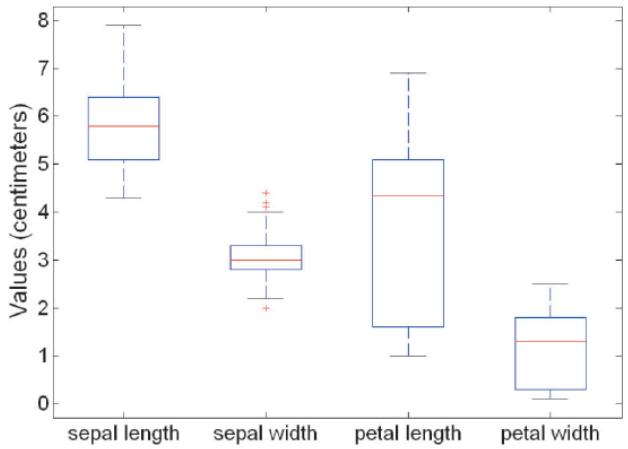
Dari data di atas tentukan persentil ke-25

Letak P25 = · (25/100). 40 = 10, yaitu data ke-10 dan kelas P25 = 51 – 55 sehingga diperoleh:

$$P_{25} = 50{,}5 + \left(\frac{\frac{25 \cdot 40}{100} - 9}{16} \right) 5 = 50{,}5 + \left(\frac{10-9}{16} \right) 5$$
$$= 50{,}5 + 0{,}31$$
$$= 50{,}81$$



Gambar 3.5 Deskripsi *box plot*.



Contoh Soal:

CONTOH TEKNIK EKSPLORASI DATA

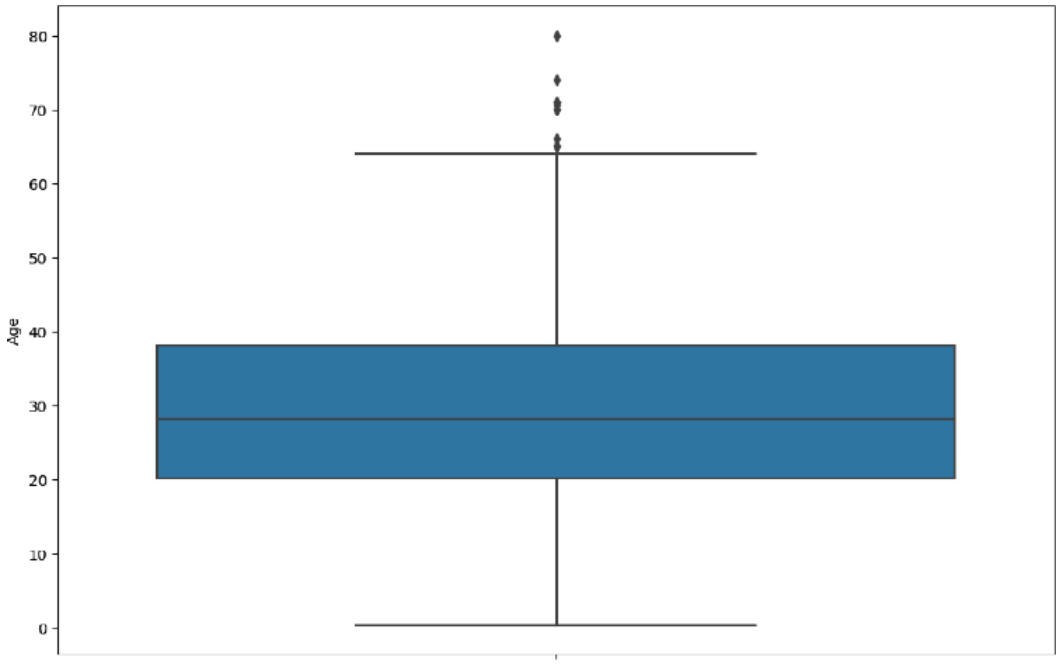
Berikut ini adalah contoh teknik perhitungan eksplorasi data.

1. Terdapat nilai statistik dari 15 mahasiswa sebagai berikut.
   45, 85, 41, 50, 46, 56, 80, 67, 70, 73, 85, 82, 43, 47, 50.

   Tentukan:

   a. Nilai modus, median, *mean*.
   b. Nilai persentil 10, persentil 25, persentil 50, persentil 90!
   c. Visualisasi *box plot*.

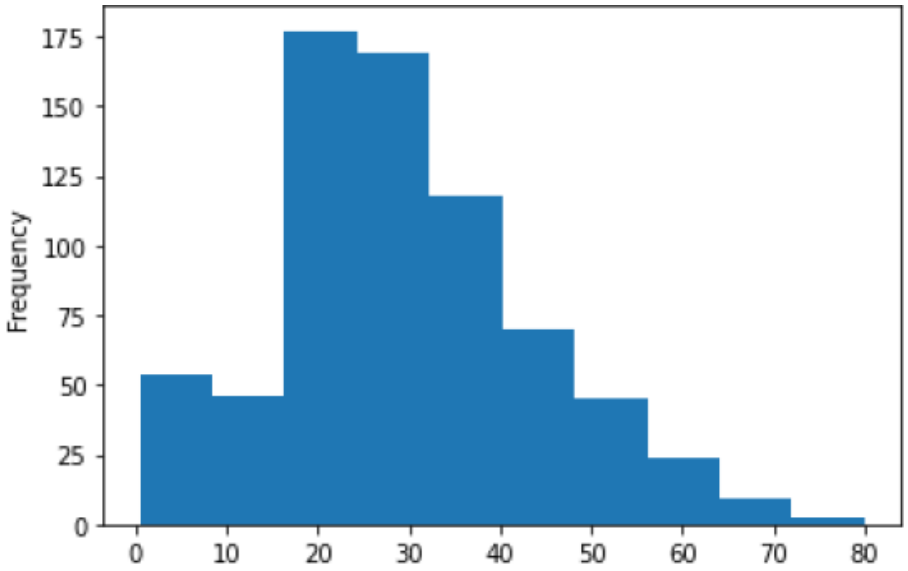
```
import seaborn as sns

plt.subplots(figsize=(12,8), dpi=100)
sns.boxplot(y=df['Age'])
plt.show()
```



Jika data terletak diatas atau dibawah garis maka terdapat outlier. Data Outlier ini adalah suatu data hasil observasi yang memunculkan nilai-nilai yang berlebihan atau melebihi batas dan jauh berbeda dengan data-data yang masih masuk dalam satu sub set data.

Histogram

```
import matplotlib.pyplot as plt
import pandas as pd

df_train=pd.read_csv('/content/sample_data/train.csv')

df_train['Age'].plot.hist()
```



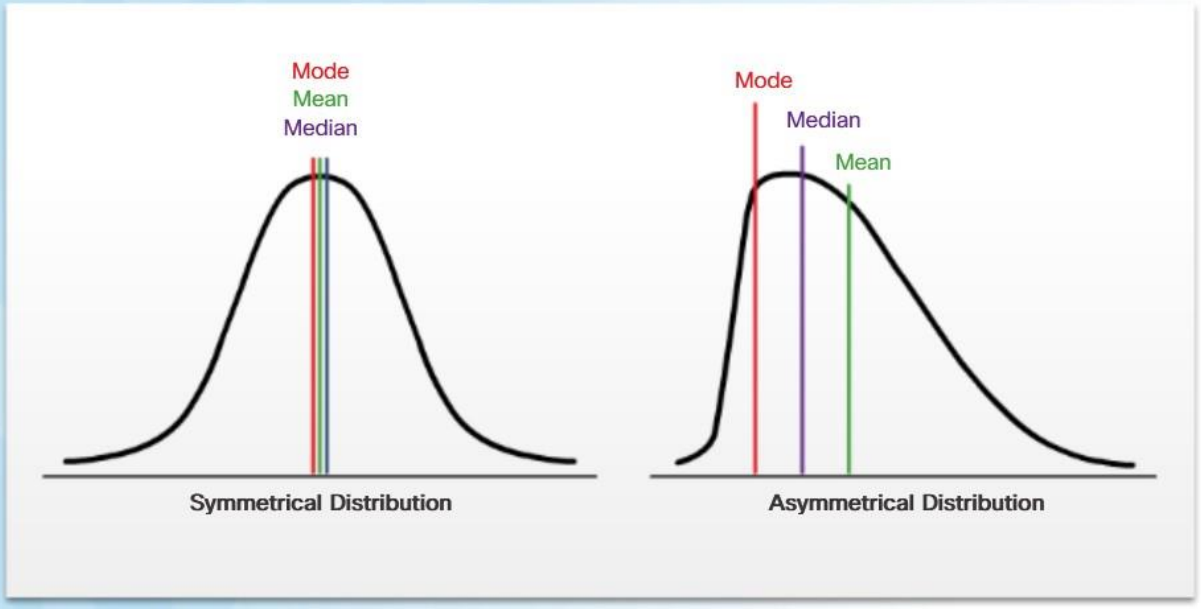
Histogram adalah tampilan bentuk grafis untuk menunjukkan distribusi data secara visual atau seberapa sering suatu nilai yang berbeda itu terjadi dalam suatu kumpulan data.

Analisis

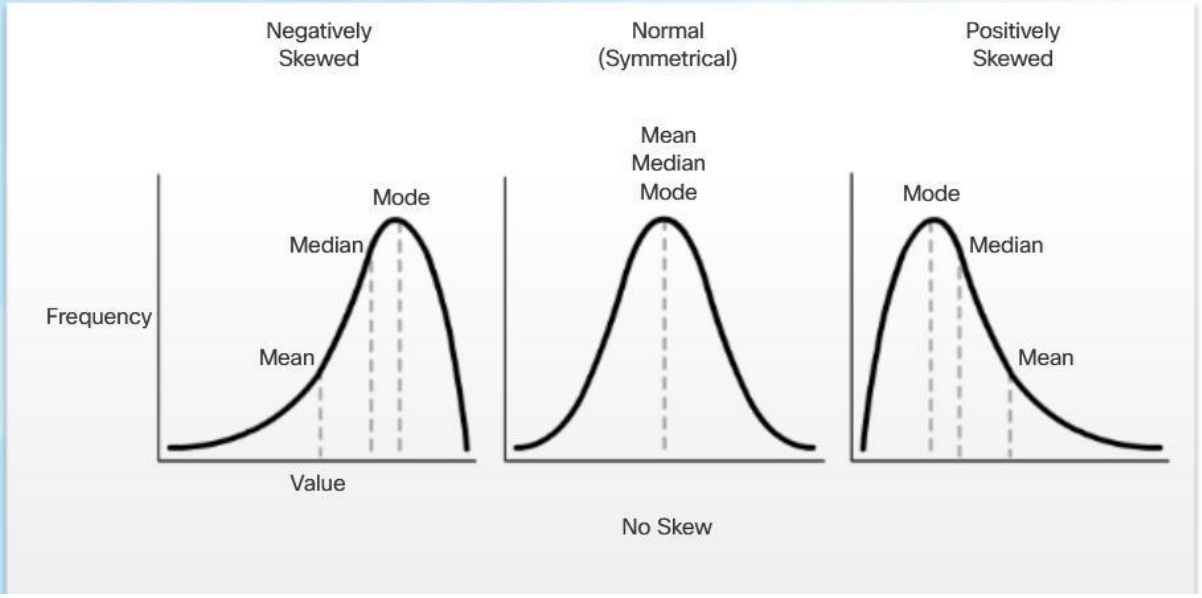
Bentuk histogram yang pada kedua sisi (kiri dan kanan) dari kelas yang tertinggi adalah simetris (seperti lonceng), nilai rata-rata range data berada ditengah yang berarti proses berjalan secara konsisten.

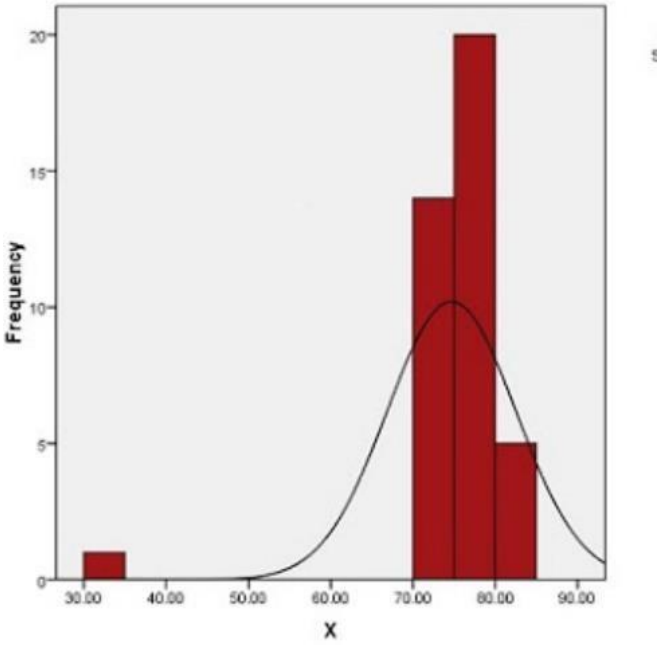
Bentuk histogram, lebih tinggi sisi kiri atau kanan, cenderung sama tinggi atau muncul beberapa titik tertinggi berselang kelas satu atau dua kelas interval. Kelainan ini kemungkinan terjadi karena jumlah data yang tidak menentu pada masing-masing kelas, ada kecenderungan pengumpulan / pembulatan data yang kurang tepat atau ketidaktepatan dalam pengukuran sehingga berpengaruh pada penetapan batas-batas kelas.

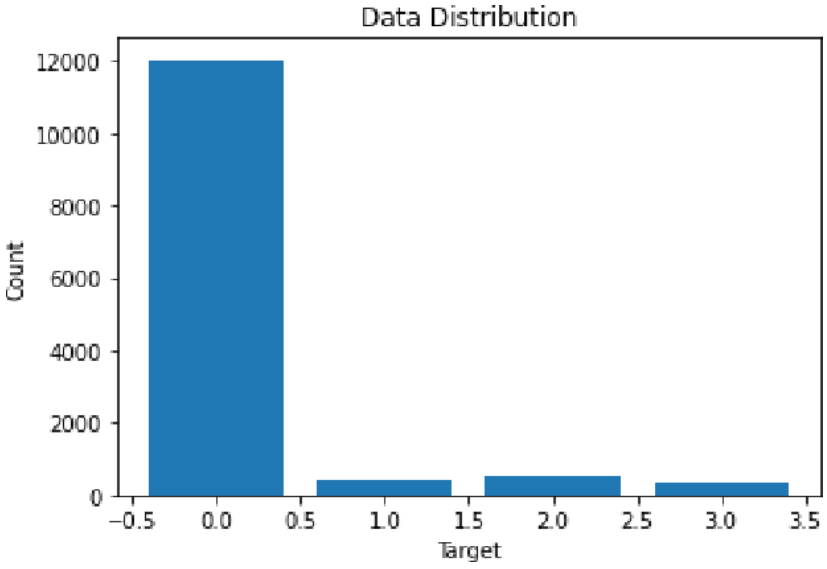
Mean, Mode, and Median in Normal and Skewed Distributions



Direction of Skew

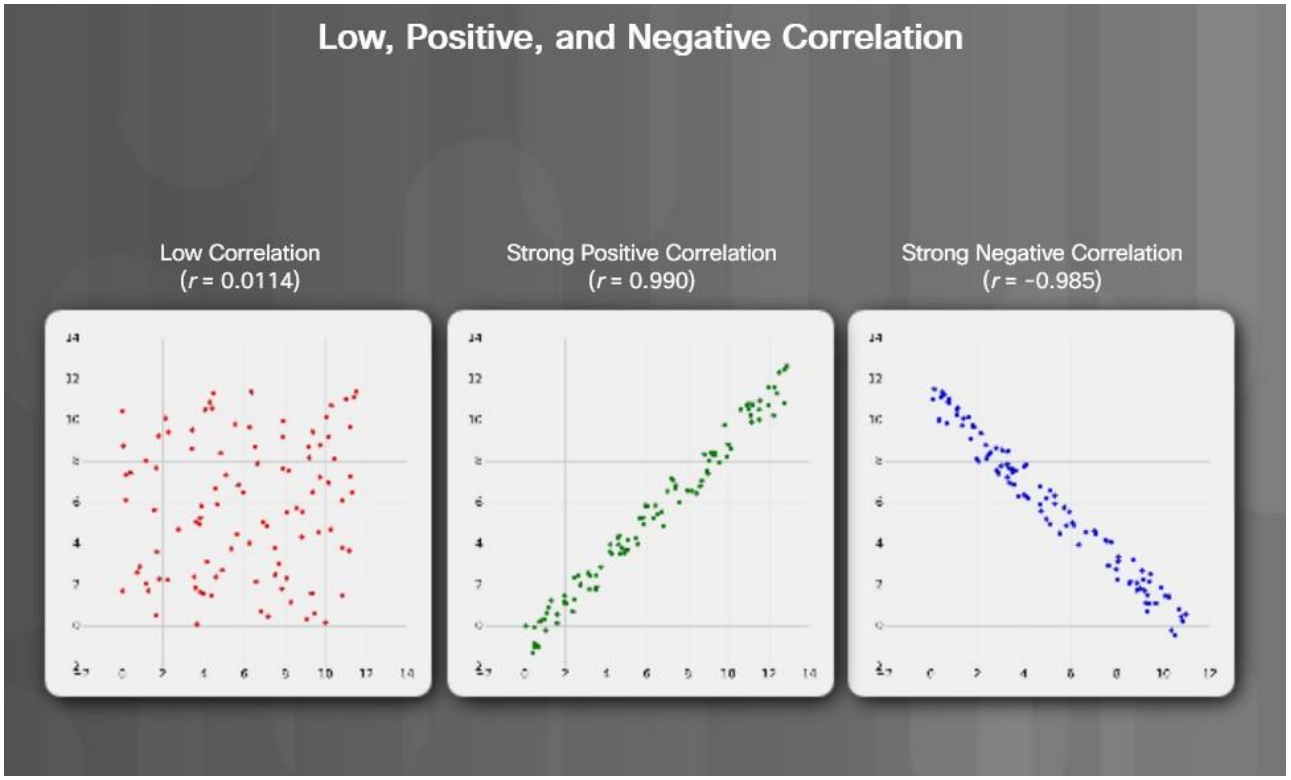



Frequency
20
15
10
5
0
30.00
40.00
50.00
60.00
70.00
80.00
90.00
X


Data Distribution
Count
12000
10000
8000
6000
4000
2000
0
−0.5
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
Target

Korelasi

Korelasi dipakai untuk mengukur hubungan antara dua variable. Analisis korelasi dibagi menjadi dua jenis hubungan positif dan negatif. Arah positif terjadi jika nilai satu variabel dinaikkan maka menyebabkan naiknya variabel lain. Sedangkan koefisien korelasi negatif berarti nilai satu variabel dinaikkan, menyebabkan naiknya nilai variabel lain begitu juga sebaliknya. Besarnya koefisien korelasi nilainya antara -1 sampai 1.



The Scenario

Tom wants to sell his car

Tom

How much money should he sell his car for?

The price he sets should not be too high, but not too low either.

Understanding the Data

Here's the documentation on what each of the 26 columns represent. https://archive.ics.uci.edu/ml/datasets/Automobile

Each of the attributes in the dataset

No.	Attribute name	attribute range	No.	Attribute name	attribute range
1	symboling	-3, -2, -1, 0, 1, 2, 3.	14	curb-weight	continuous from 1488 to 4066.
2	normalized-losses	continuous from 65 to 256.	15	engine-type	dohc, dohcv, l, ohc, ohcf, ohcv, rotor.
3	make	audi, bmw, etc.	16	num-of-cylinders	eight, five, four, six, three, twelve, two.
4	fuel-type	diesel, gas.	17	engine-size	continuous from 61 to 326.
5	aspiration	std, turbo.	18	fuel-system	1bbl, 2bbl, 4bbl, idi, mfi, mpfi, spdi, spfi.
6	num-of-doors	four, two.	19	bore	continuous from 2.54 to 3.94.
7	body-style	hardtop, wagon, etc.	20	stroke	continuous from 2.07 to 4.17.
8	drive-wheels	4wd, fwd, rwd.	21	compression-ratio	continuous from 7 to 23.
9	engine-location	front, rear.	22	horsepower	continuous from 48 to 288.
10	wheel-base	continuous from 86.6 120.9.	23	peak-rpm	continuous from 4150 to 6600.
11	length	continuous from 141.1 to 208.1.	24	city-mpg	continuous from 13 to 49.
12	width	continuous from 60.3 to 72.3.	25	highway-mpg	continuous from 16 to 54.
13	height	continuous from 47.8 to 59.8.	26	price	continuous from 5118 to 45400.

Descriptive Statistics – Scatter Plot

- Often, we tend to see continuous variables in our data. Each observation represented as a point. These data points are numbers contained in some range
- For example, in our dataset, price and engine size are continuous variables. What if we want to understand the relationship between "engine size" and "price"?
- One good way to visualize this is using a scatter plot. **Scatter plot show the relationship between two variables :**

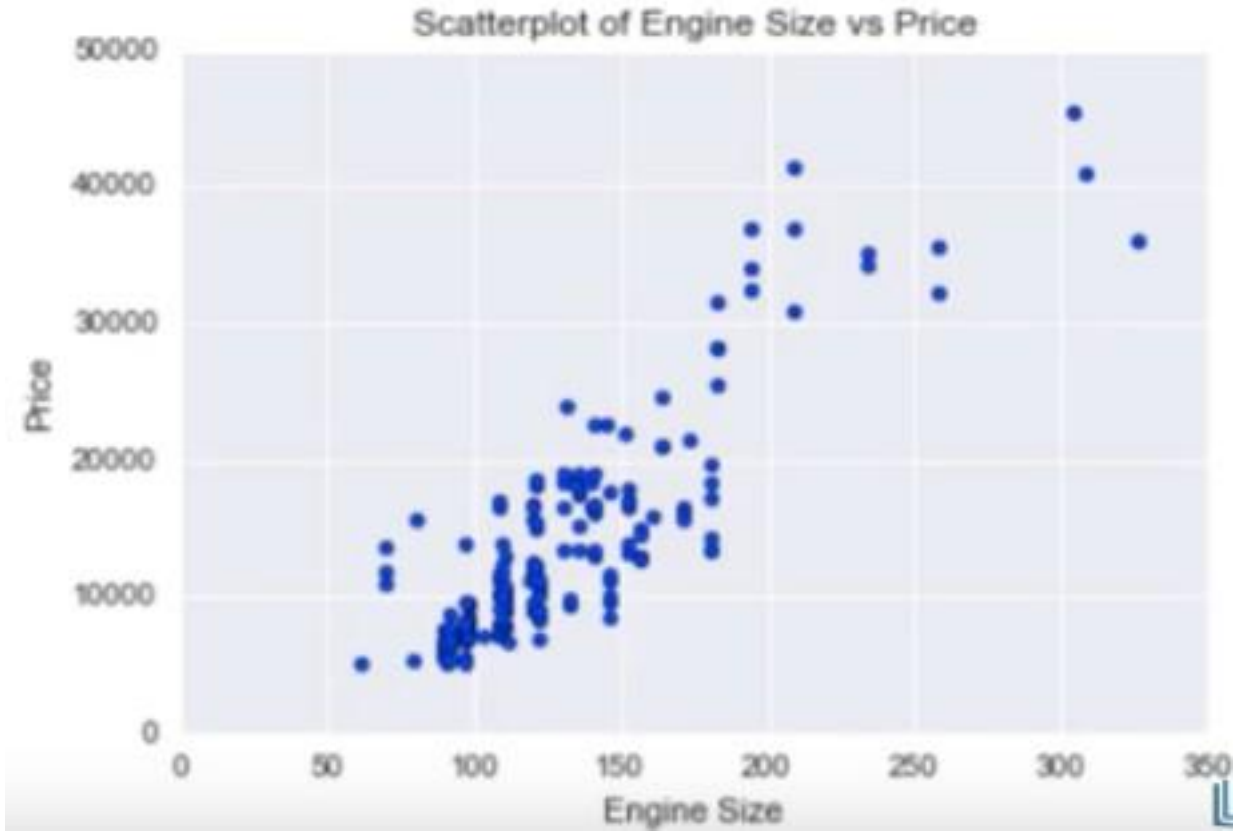
- **The predictor variable** on x-axis. Variable that you are using to predict an outcome. In this case, our predictor variable is the engine size.
- **The target variable** on y-axis. Variable that you are trying to predict. In this case, our target variable is the price, since this would be the outcome.

- In this case, we will thus plot the engine size on the x-axis and the price on the y-axis. **We are using the** Matplotlib function "scatter" here, taking in x and a y variable.

```
y=df["engine-size"]
x=df["price"]
plt.scatter(x,y)
```

Descriptive Statistics – Scatter Plot

- From the scatterplot we see that as the engine size goes up, the price of the car also goes up.
- This is giving us an initial indication that there is a positive linear relationship between these two variables.



Correlation

- Correlation is a statistical metric for measuring to what extent different variables are interdependent. In other words, when we look at two variables over time, **if one variable changes how does this affect change in the other variable**?
- For example, smoking is known to be correlated to lung cancer. Since you have a higher chance of getting lung cancer if you smoke.
- In another example, there is a correlation between umbrella and rain variables where more precipitation means more people use umbrellas. Also, if it doesn't rain people would not carry umbrellas. Therefore, we can say that umbrellas and rain are interdependent, and they are correlated.
- It is important to know that correlation doesn't imply causation.
- In fact, we can say that umbrella and rain are correlated but we would not have enough information to say whether the umbrella caused the rain, or the rain caused the umbrella.

Correlation

Let’s look between two features from our car dataset (engine-size and price)

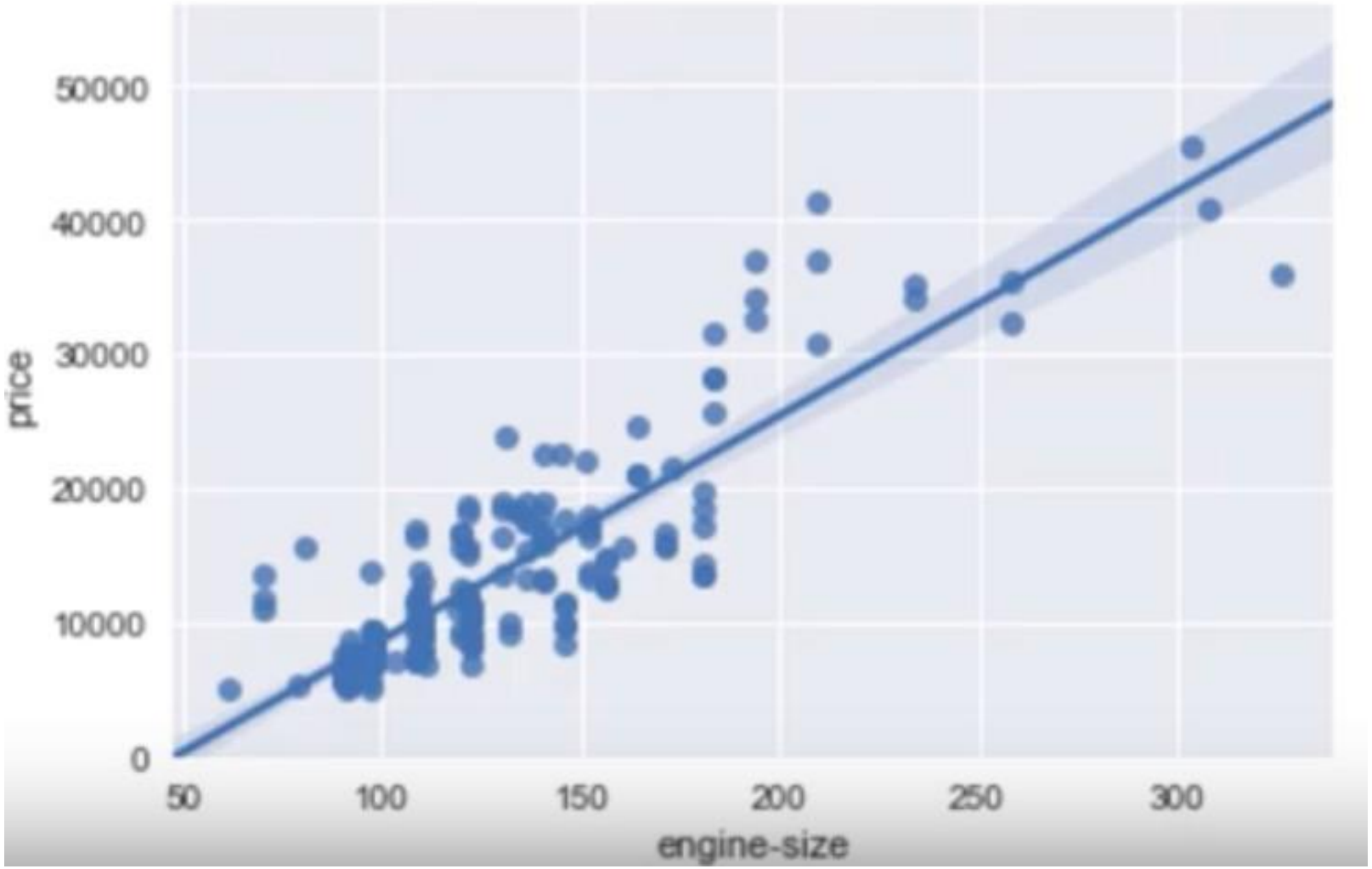
We'll visualize these two variables using a scatter plot and an added linear line called a regression line, which indicates the relationship between the two.

```
sns.regplot(x="engine-size", y="prices", data=df)
plt.ylim(0,)
```

The main goal of this plot is to see whether the engine size has any impact on the price.

Correlation

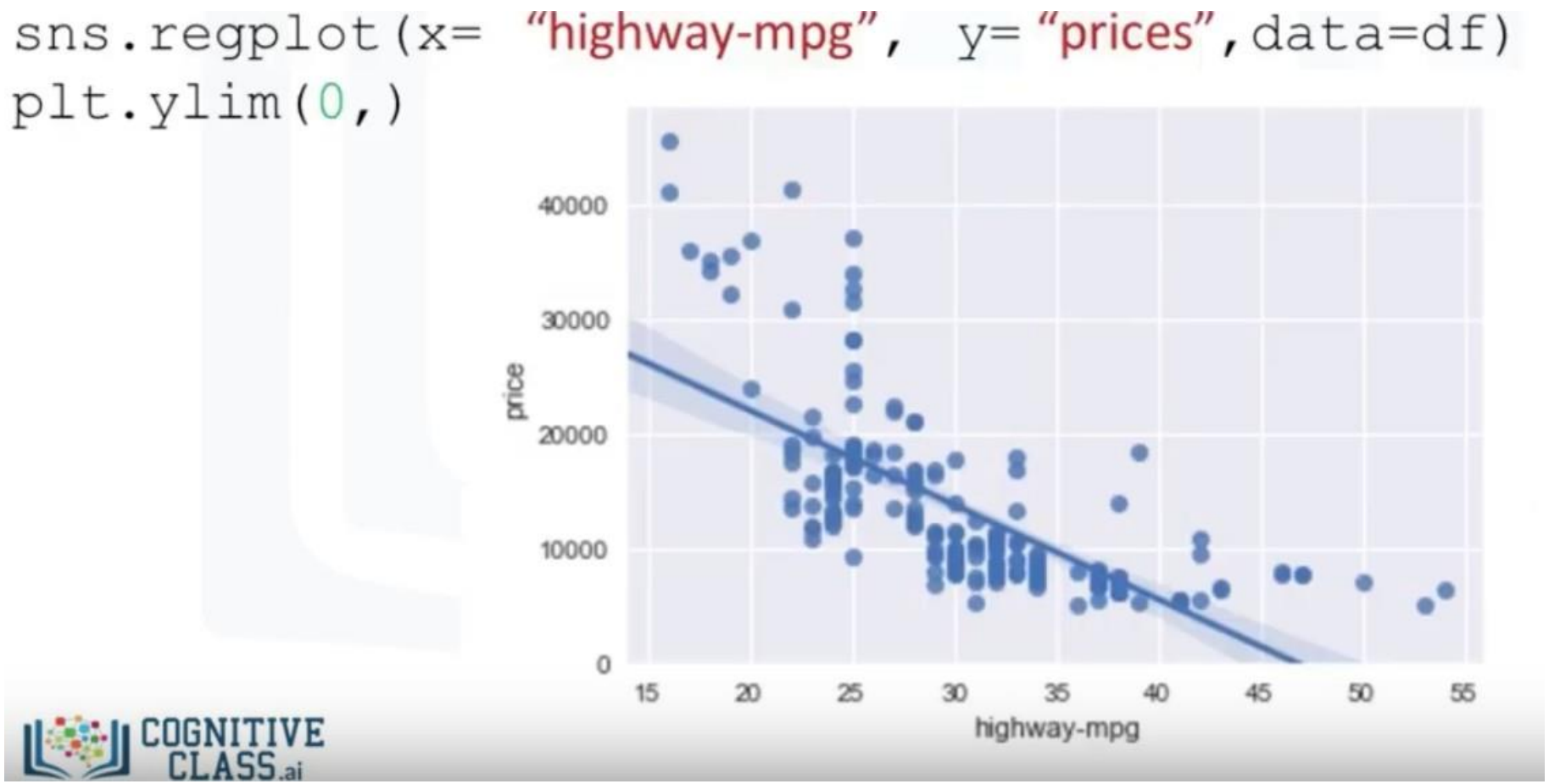
- In this example, you can see that there's a positive linear relationship between the two variables
- With increase in values of engine size, values of price go up as well. So there is a positive correlation between engine size and price.



Correlation

Let's look correlation between highway miles per gallon and price

```
sns.regplot(x= "highway-mpg",  y= "prices",data=df)
plt.ylim(0,)
```



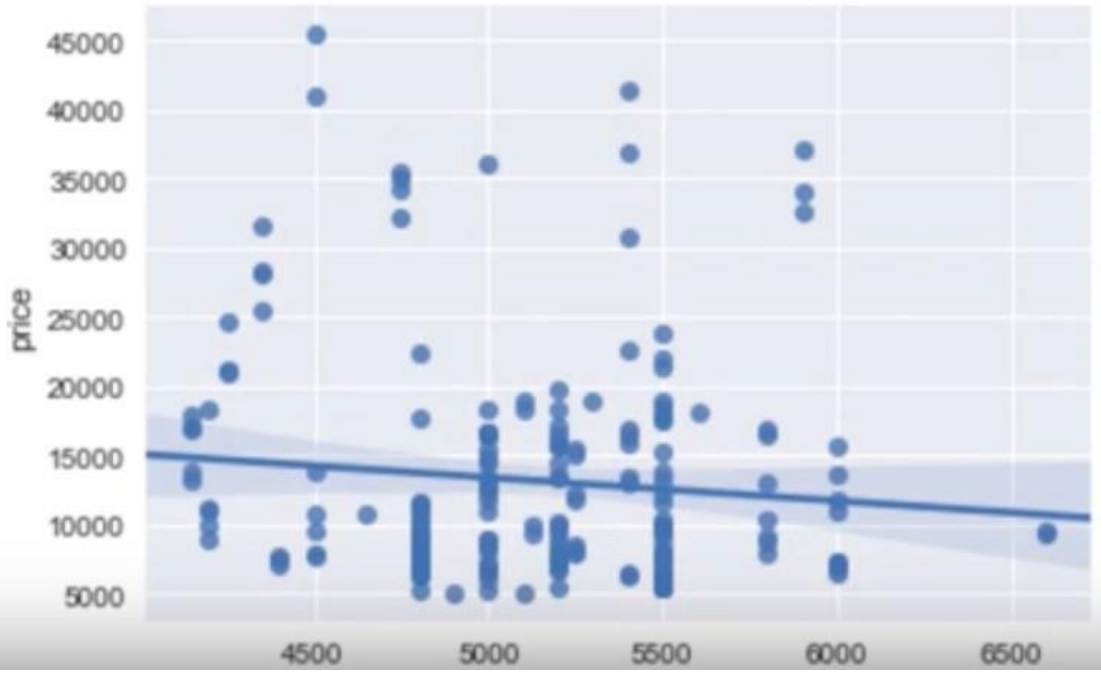
Correlation

- As we can see from previous plot, when highway miles per gallon value goes up the value price goes down.
- Therefore there is a negative linear relationship between highway miles per gallon and price.
- Although this relationship is negative the slope of the line is steep which means that the highway miles per gallon is still a good predictor of price.
- These two variables are said to have a negative correlation.

Correlation

- Weak correlation between peak-rpm and price
- Both low peak RPM and high values of peak RPM have low and high prices. Therefore, we cannot use RPM to predict the values.

```
sns.regplot(x="peak-rpm", y="prices", data=df)
plt.ylim(0,)
```



Heatmap

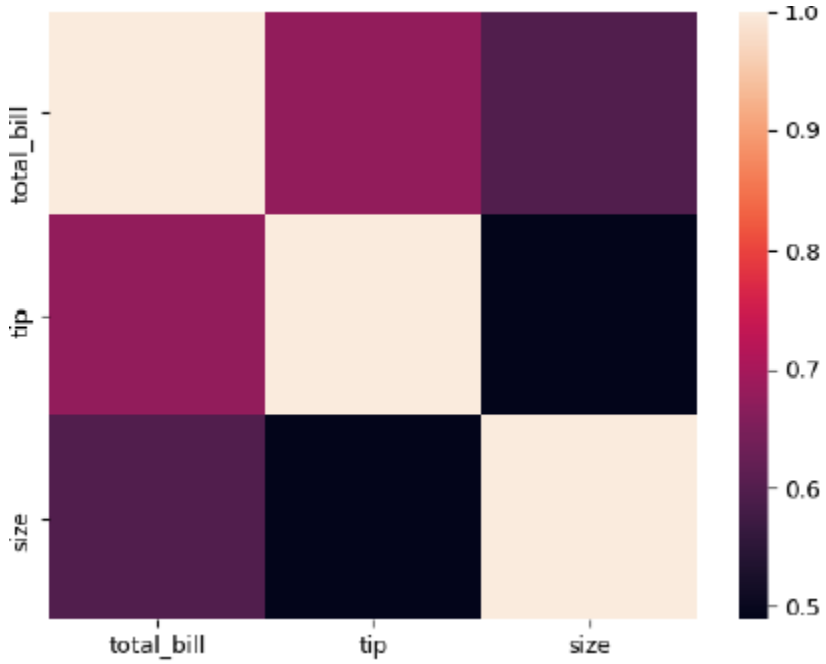
Guna memvisualisasikan data dengan menggunakan *heatmap*, kita memerlukan data yang berformat matriks. Artinya adalah jumlah peubah pada indeks sesuai dengan jumlah peubah pada kolom data. Untuk membentuk bentuk data yang sesuai, kita umumnya menggunakan metode korelasi data atau tabel pivot. Berikut kita akan mencoba dulu dengan menggunakan metode korelasi data pada dataset `tips`:

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

```
df = sns.load_dataset('tips')

df = df.corr()

print(df)

sns.heatmap(df)

plt.show()
```

Hasil :

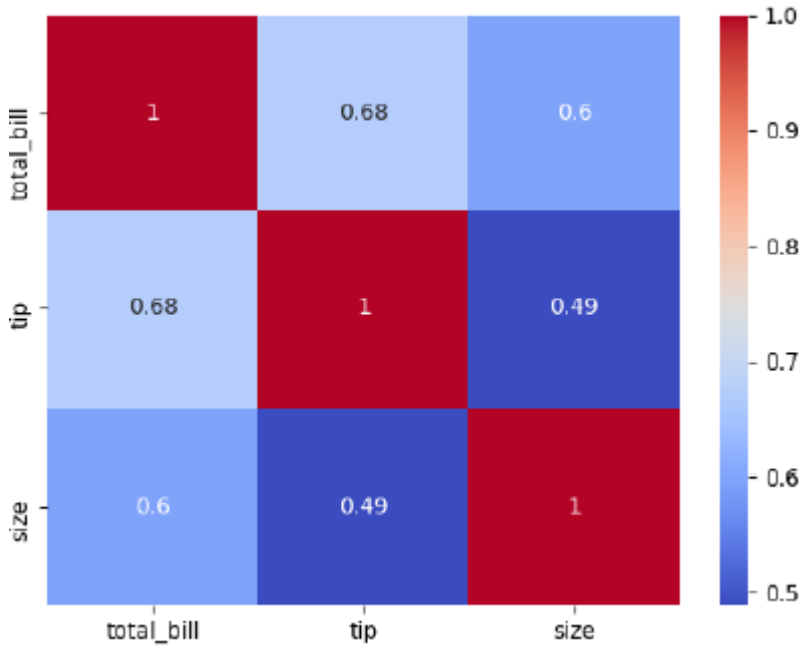
```
total_bill       tip      size
total_bill    1.000000  0.675734  0.598315
tip           0.675734  1.000000  0.489299
size          0.598315  0.489299  1.000000
```



Kita juga dapat menampilkan nilai setiap elemen dalam matriks dengan menambahkan argumen **annot=True**, serta mengganti warna pada plot sebagai berikut ini:

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

```
df = sns.load_dataset('tips').corr()

sns.heatmap(df, annot=True, cmap='coolwarm')

plt.show()
```



TUGAS :

1. Kerjakan tugas secara individu
2. Carilah 1 dataset
3. Cek apakah terdapat Outlier, analisa dan visualisasikan dengan boxplot
4. Tentukan distribusi dari dataset tersebut, analisa dan visualisasikan dengan Histogram
5. Carilah korelasi antar variabel dengan variabel output, analisa dan visualisasikan dengan Scatter plot dan Heatmap
6. File yang dikumpulkan : File python, data set, laporan.pdf
7. Laporan terdiri penjelasan dan screen shoot program
8. Penamaan File : Ekspolorasi Data.Zip
9. Tugas dikummpulkan hari Kamis / 21 September 2023 pukul 23.59 Wib