Kampus Merdeka
X
MyEduSolve

DESCRIPTIVE STATISTICS

Week #2

DATA SCIENCE – TEAM 3

Outline

- Probability and Probability Distribution
- Inferential Statistics
- Non Parametric Test
- Machine Learning
- Missing Values
- Handling Missing Values

1. Probability

Probabilitas (peluang) adalah pengukuran terhadap suatu kemungkinan dan suatu kejadian. Pemahaman terkait peluang merupakan dasar untuk materi Inferential Statistics.

TERMINOLOGI

Experiment adalah percobaan yang menghasilkan suatu perhitungan, pengukuran, atau respon.

Outcome adalah hasil dari suatu percobaan.

Sample Space adalah himpunan dari seluruh kemungkinan outcome pada suatu probability experiment.

Event adalah bagian dari sample space, bisa berupa satu atau lebih outcomes.

IMPORT LIBRARY

```
# import library
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from scipy import stats
import scipy.stats as ss
from scipy.stats import norm

from collections import Counter
import math

import warnings
warnings.filterwarnings('ignore')
```

2. Conditional Probability

adalat probabilitas kemunculan suatu event. dengan mengetahui bahwa event lain sudah muncul atau terjadi.

Independent Events

Jika kemunculan dari event yang satu tidak mempengaruhi probability kemunculan event kedua.

$$P(B|A) = P(B)$$
$$P(A|B) = P(A)$$

Dependent Events

Jika kemunculan dari event yang satu mempengaruhi probability kemunculan kedua.

$$P(B|A) \neq P(B)$$
$$P(A|B) \neq P(A)$$

MULTIPLICATION RULE, MUTUALLY EXCLUSIVE EVENTS, NON-MUTUALLY EXCLUSIVE EVENTS

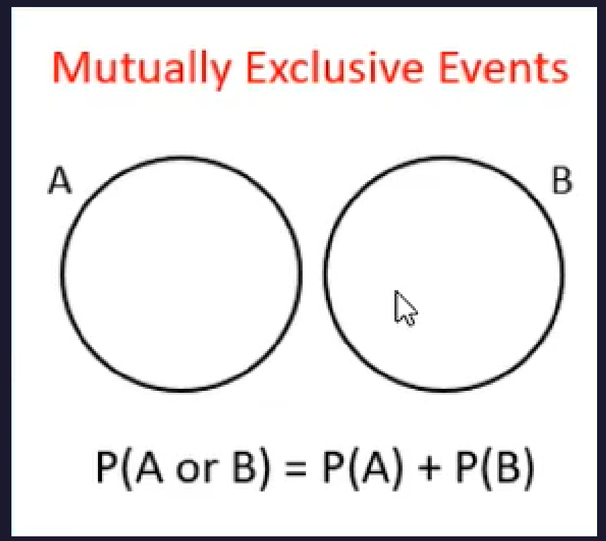
MULTIPLICATION RULE

Probability untuk dua buah events A dan B untuk muncul secara berurutan :

1. Jika kedua events A dan B tersebut dependent maka bisa disederhanakan menjadi P(A dan B)=P(A).P(B|A)
2. Jika kedua events A dan B tersebut independent maka bisa disederhanakan menjadi P(A dan B)=P(A).P(B)

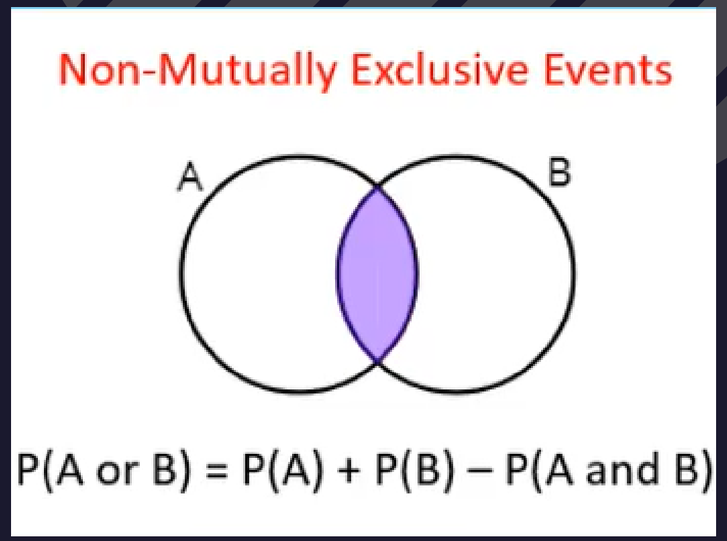
Mutually Exclusive Events

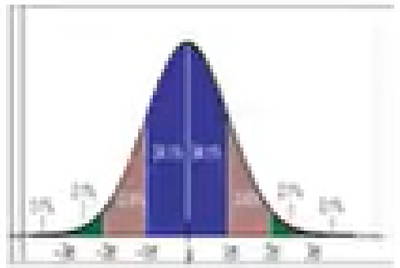
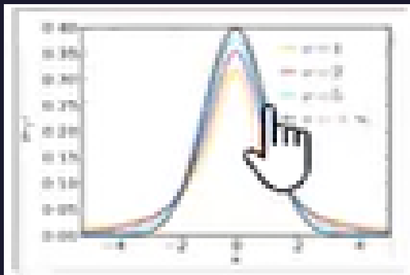
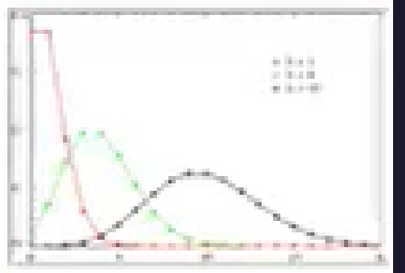
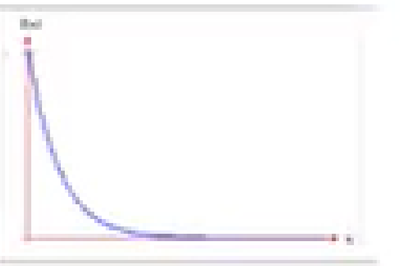
Kondisi ketika kejadian A dan B tidak dapat muncul pada waktu bersamaan.



Non- Mutually Exclusive Events

Kondisi ketika kejadian A dan B dapat muncul pada waktu bersamaan.



	Distribusi Normal	Distribusi Students T	Distribusi Binomial	Distribusi Poisson	Distribusi Eksponensial
What does it look like?					
Defining Characteristics	Distinctive Bell Shape	Shorter, fatter than the normal distribution.	Two outcomes: Success/Failure	Various shapes, but valid only for integers on the x-axis.	Models Time Between Events
Example of When to Use It	Modeling natural phenomena (height, weight, IQ, test scores etc.)	When you have small samples or don't know the population variance (σ^2).	Coin Toss Probability (Heads, Tails)	Gives probability of number of events in a fixed interval.	"How much time will go by before a major hurricane hits the Atlantic Seaboard?"
Example of DS Application	Least squares fitting or propagation of uncertainty.	Unknown σ^2 is common in real life data, you you'll have to use the T instead of the normal in that case.	Anywhere where binary (yes/no, black/white, vote/don't vote) data is used.	Anywhere there is a waiting time between events.	Building continuous-time Markov chains.

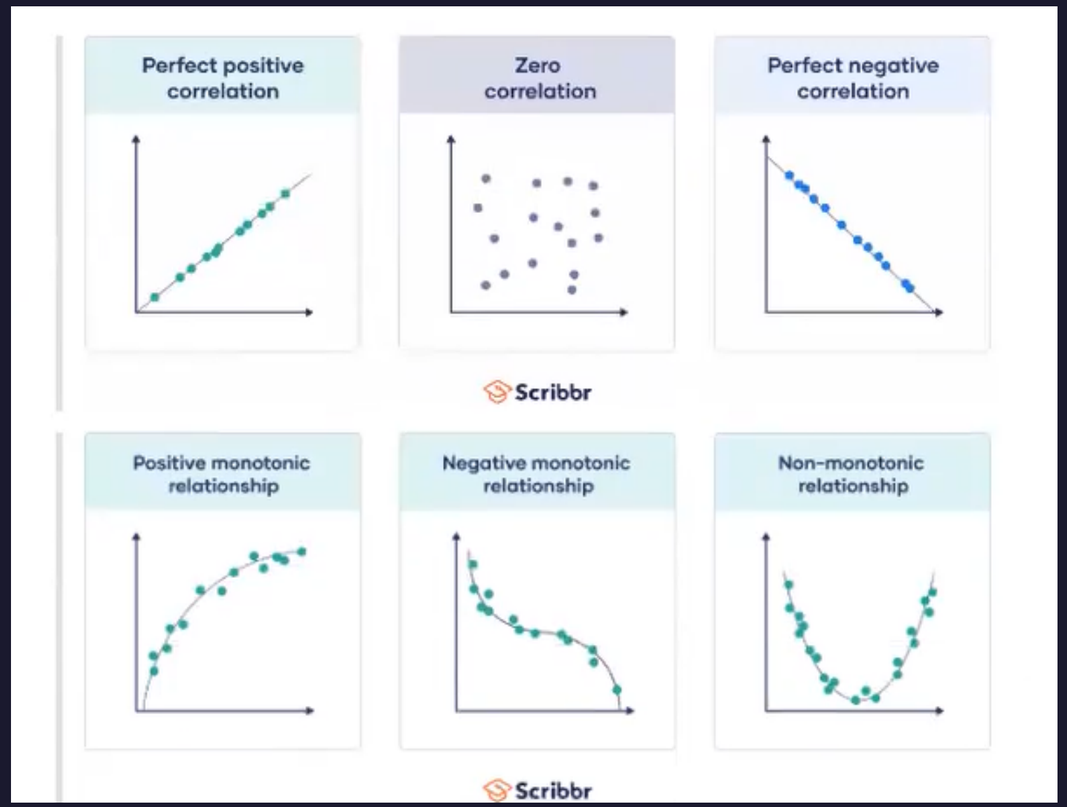
Correlation Coefficient

Continuous Feature

terdiri dari Pearson, Spearman, Kendall

Categorical Future

terdiri dari Cramer's V, Theil's U, Point-Biserial Corellation

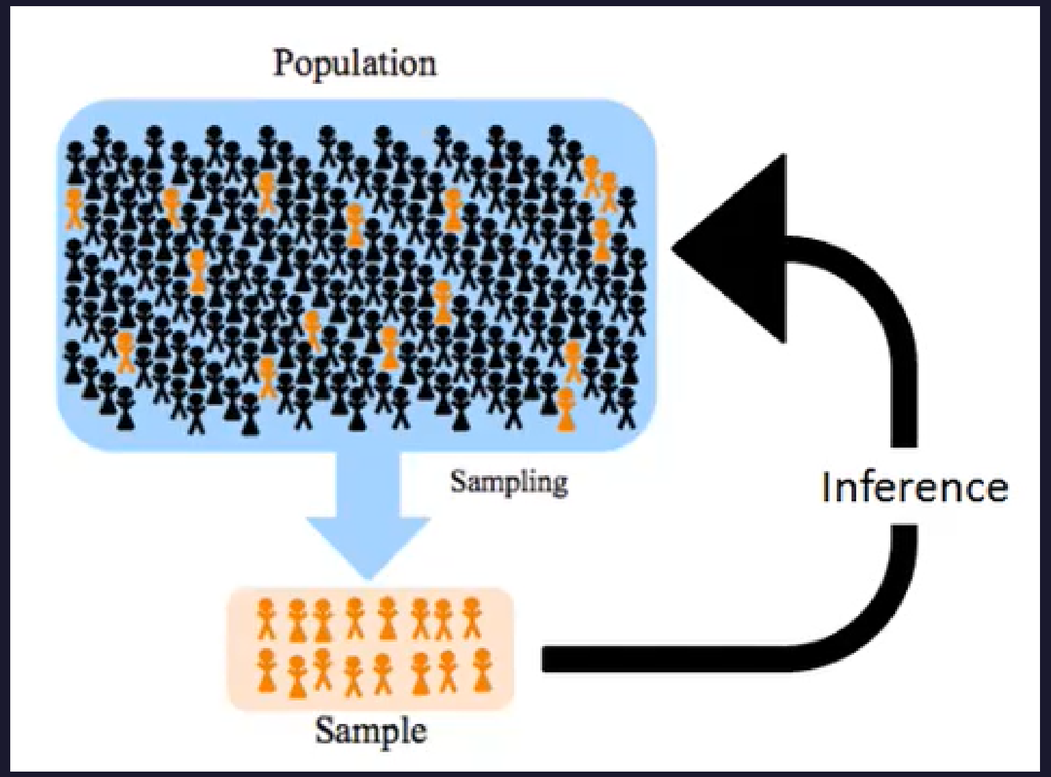


INFERENTIAL STATISTICS

Pengertian Inferential Statistics

Ilmu statistik yang bertugas mempelajari tata cara penarikan kesimpulan mengenai keseluruhan populasi berdasarkan data hasil penelitian pada sampel.

1. Populasi adalah seperangkat unit analisa yang lengkap yang sedang diteliti
2. Sampel adalah sub dari seperangkat elemen yang dipilih untuk dipelajari



3. PROBABILITY DISTRIBUTION

bertujuan untuk mendeskripsikan semua kemungkinan nilai dan juga kemungkinan yang dapat diambil dari berbagai variabel acak pada rentang tertentu.

TIPE DISTRIBUSI TERGANTUNG SIFAT

- **BINOMIAL DISTRIBUTION**
- **POISSON DISTRIBUTION**
- **UNIFORM DISTRIBUTION**

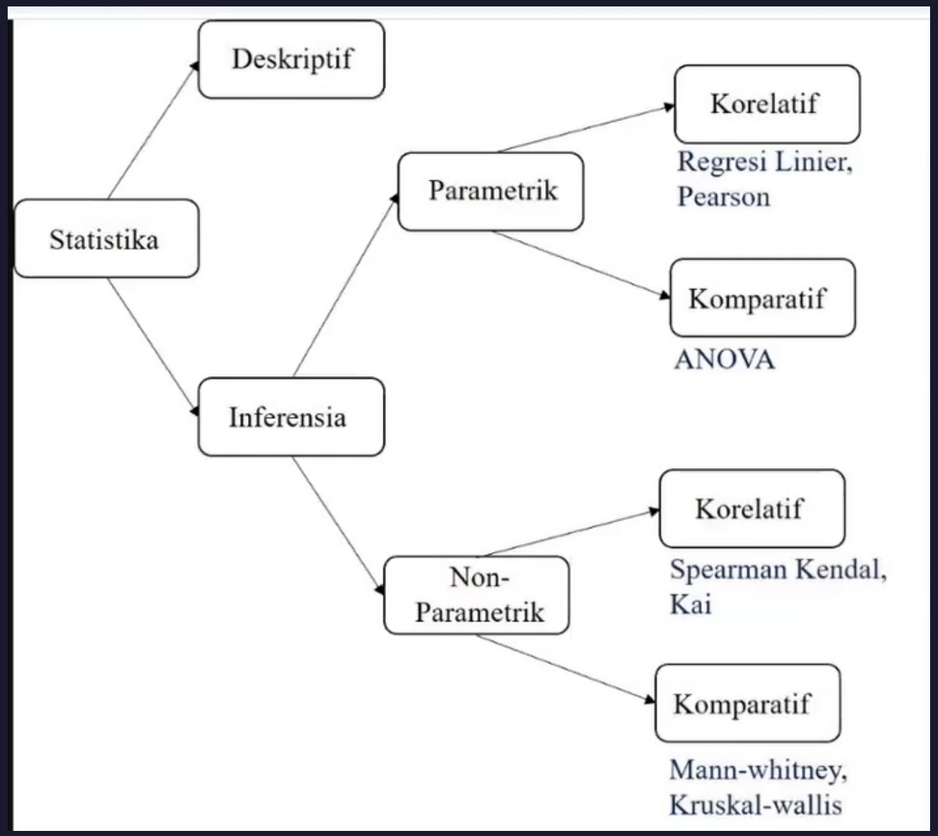
Jenis Statistik Inferensial

01 Statistik Parametrik

pendugaan dan uji hipotesis dari parameter populasi didasarkan anggapan bahwa skor-skor yang dianalisis telah ditarik dari suatu populasi dengan distribusi tertentu.

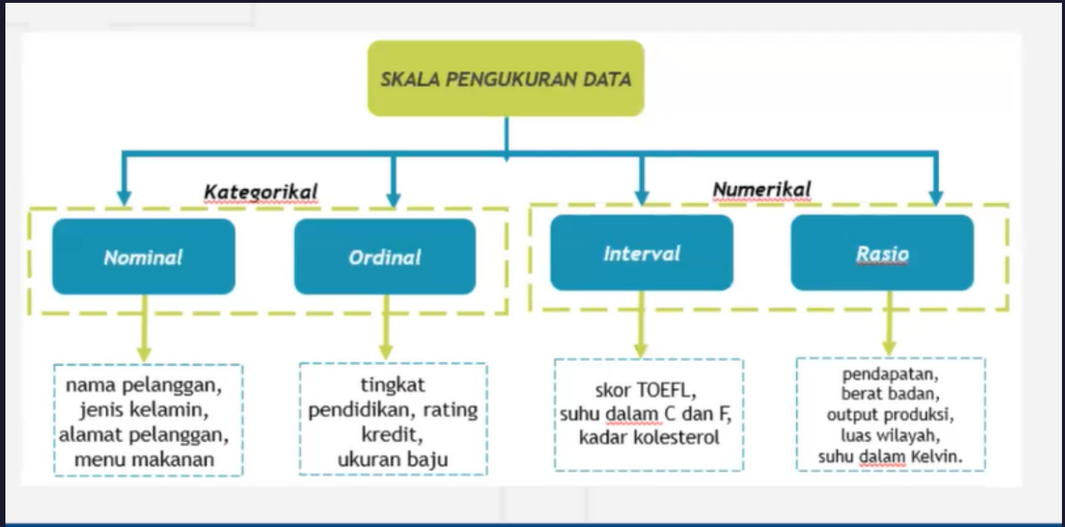
02 Statistik Non-Parametrik

pendugaan dan uji hipotesis dari parameter populasi didasarkan anggapan bahwa skor-skor yang dianalisis telah ditarik dari suatu populasi dengan bebas sebaran (tidak mengikuti distribusi tertentu).



Skala Pengkuran, Hypothesis Testing, Null VS Alternatif Hypothesis

Skala Pengukuran



Hypothesis

pernyataan dugaan sementara terhadap suatu masalah penelitian yang kebenarannya masih harus diuji secara empiris.

Null Hypothesis (HO)

hipotesis yang menyatakan tidak adanya saling hubungan antara dua variabel atau lebih.

Alternatif Hypothesis (Ha)

hipotesis yang menyatakan adanya saling hubungan antara dua variabel atau lebih.

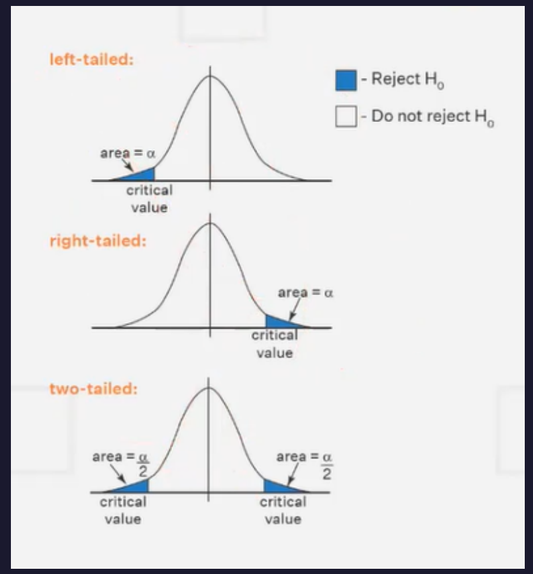
Significance Level		Specification
$p > 0.05$		not significant
$p \leq 0.05$	(5%)	significant
$p \leq 0.01$	(1%)	very significant
$p \leq 0.001$	(0.1%)	highly significant

KRITERIA Ha DITERIMA, H0 DITOLAK ADALAH KETIKA PVALUE < ALPHA

Z-Test, T-Test

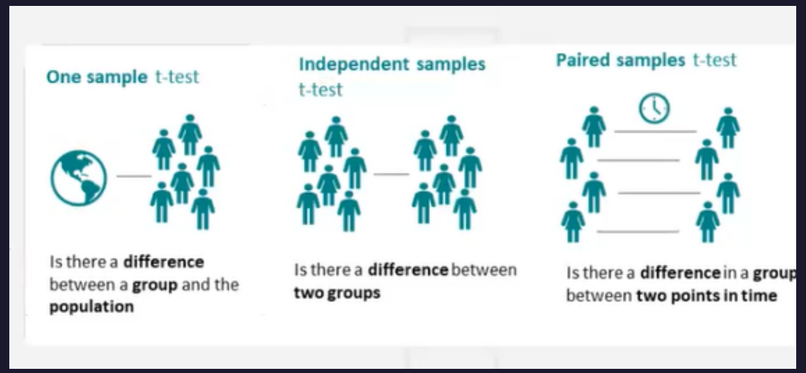
Z-Test

Uji statistik yang digunakan untuk mengetahui apakah suatu populasi memiliki rata-rata yang sama dengan, lebih kecil atau lebih besar dari suatu nilai rata-ratatertentu sesaui dengan hipotesis yang telah ditetapkan.



T-Test

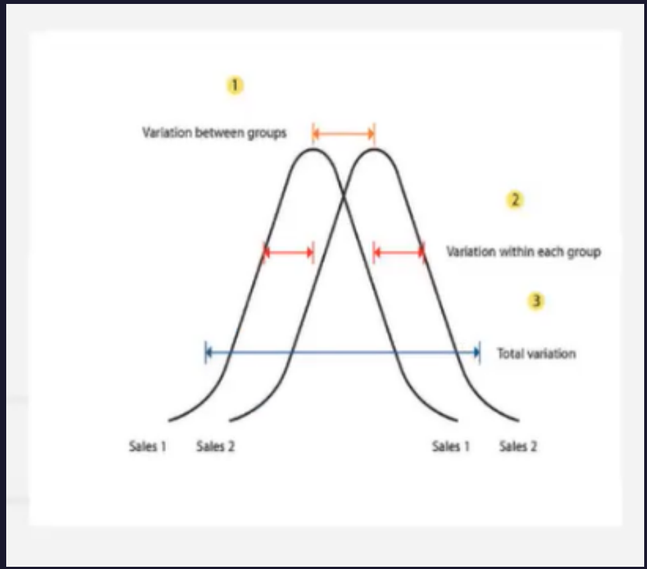
Uji statistik yang digunakan untuk mengetahui kebenaran hipotesis yang diajukan oleh peneliti dalam membedakan rata-rata pada dua populasi.



Anova

Pengertian

Anova merupakan alat uji statistik yang digunakan untuk menguji apakah lebih dari 2 populasi yang inndepeenden, memiliki rata-rata yang berbeda atau sama.




```
# contoh dengan f_oneway
a = df[df["Neighborhood"]=="CollgCr"]["SalePrice"]
b = df[df["Neighborhood"]=="Veenker"]["SalePrice"]
c = df[df["Neighborhood"]=="Crawfor"]["SalePrice"]

# f_oneway
f_stats, p_value = f_oneway(a,b,c)
```

4. Non-Parametric Test

Mann Withney Wilcoxon Test

```
# contoh mannwhitney wilcoxon dengan scipy
a = df[df["Neighborhood"]=="CollgCr"]["SalePrice"]
b = df[df["Neighborhood"]=="Veenker"]["SalePrice"]

# dengan ranksum
u_stats, p_value = stats.ranksums(a, b)

# nilai alpha
alpha =0.05
```

Wilcoxon Signed Rank Test

```
# signed rank test
control = [8.0, 7.1, 6.5, 6.7, 7.2, 5.4, 4.7, 8.1, 6.3, 4.8]
treatment = [9.9, 7.9, 7.6, 6.8, 7.1, 9.9, 10.5, 9.7, 10.9, 8.2]

w_stats, p_value = stats.wilcoxon(control, treatment)
```

Kruskal Wallis

```
# contoh dengan kruskal
a = df[df["Neighborhood"]=="CollgCr"]["SalePrice"]
b = df[df["Neighborhood"]=="Veenker"]["SalePrice"]
c = df[df["Neighborhood"]=="Crawfor"]["SalePrice"]

# kruskal
h_stats, p_value = kruskal(a,b,c)
```

Goodness Of Fit Test

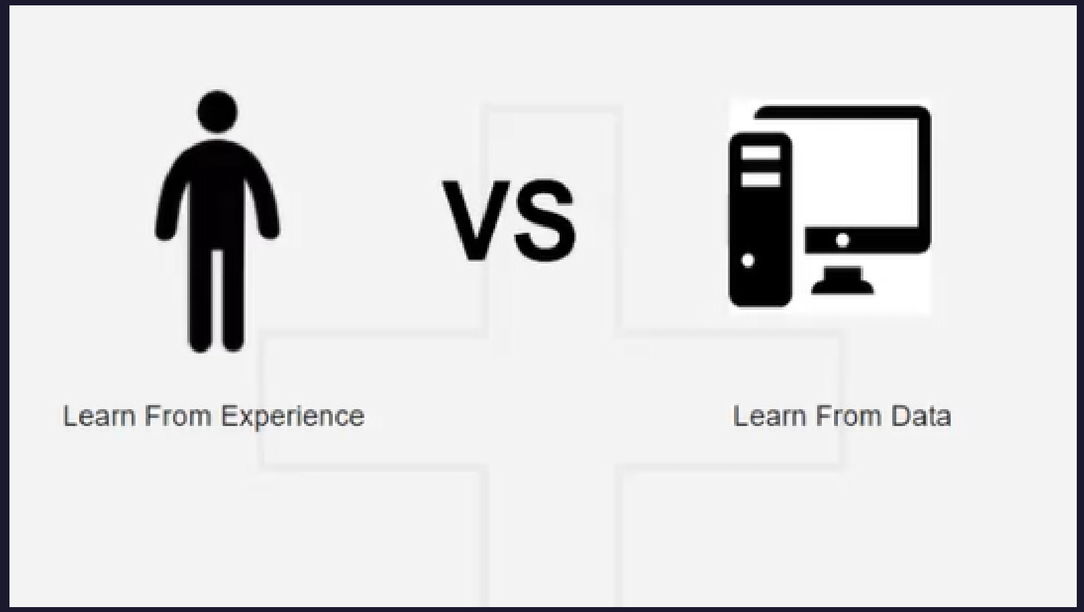
```
# contoh dengan d'agostino
k2, p_value = stats.normaltest(df["SalePrice"])
alpha = 0.05
```

```
# contoh dengan shapiro
shapiro, p_value = stats.shapiro(df["SalePrice"])
alpha = 0.05
```

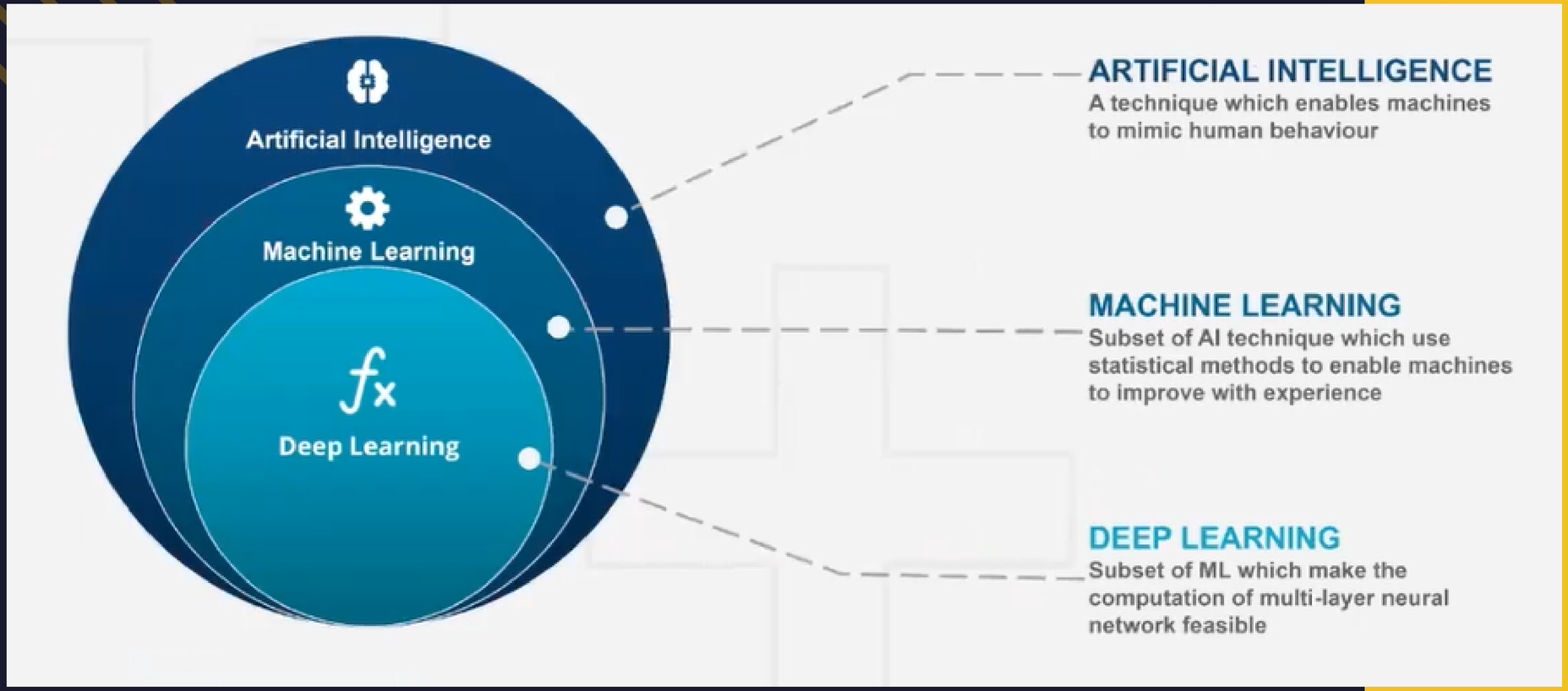
```
# QQplot
sm.qqplot(df["SalePrice"], line = '45')
plt.show()
```

Machine Learning

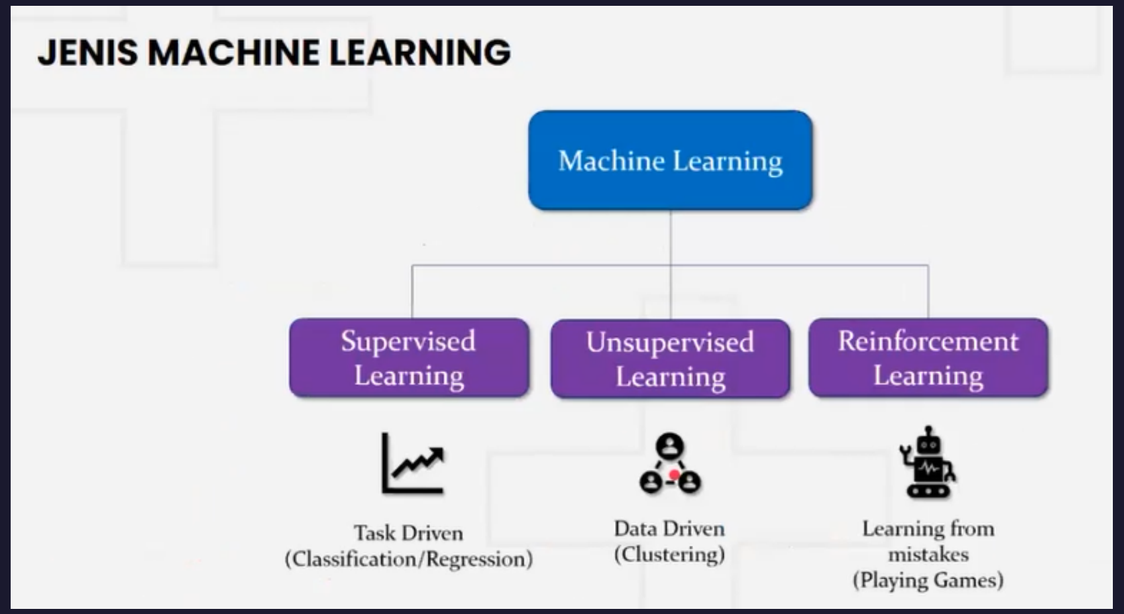
Suatu algoritma yang dikembangkan untuk bisa mempelajari pola dari data dan melakukan suatu task tertentu.

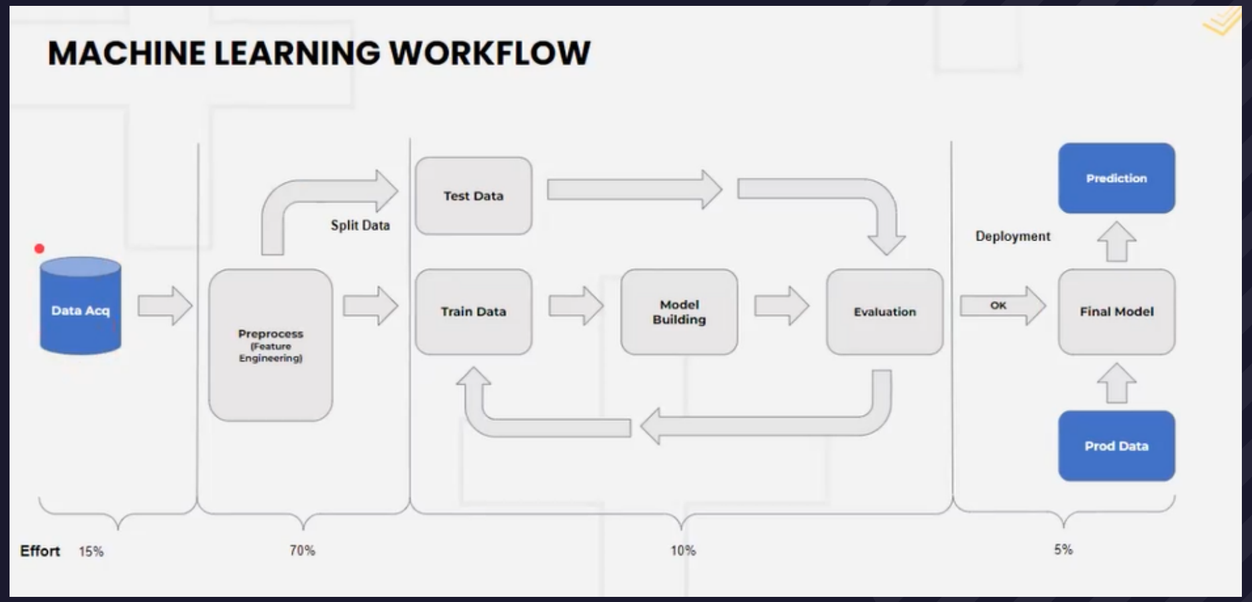


MACHINE LEARNING, ARTIFICIAL INTELLIGENCE, DEEP LEARNING



Jenis Machine Learning, Machine Learning Workflow



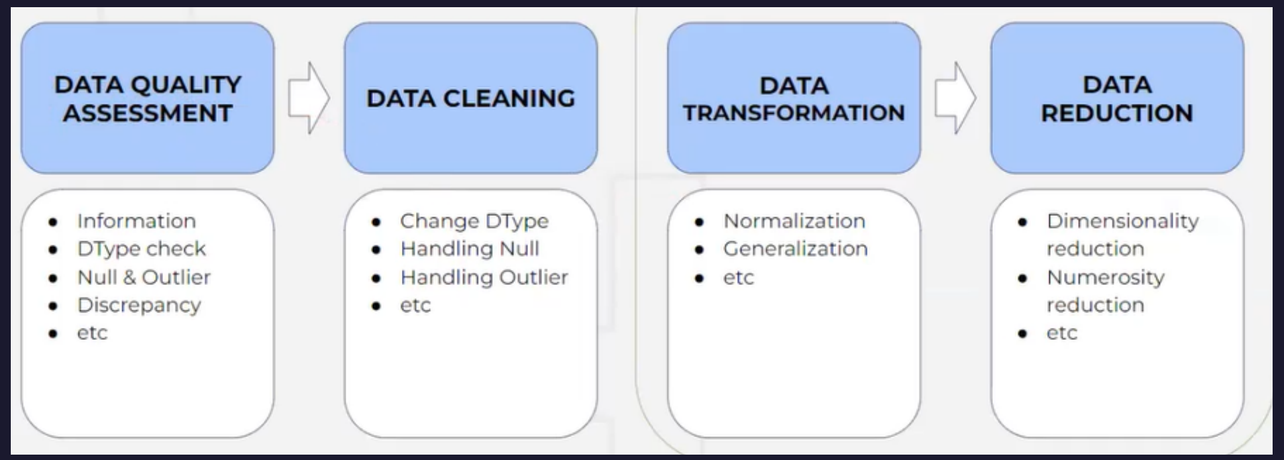


Preprocessing, Data Quality

Preprocessing

Data Preprocessing adalah teknik yang digunakan untuk mengubah raw kedalam digestible data.

Preprocessing Steps



Data Quality Assessment

Proses melakukan evaluasi pada initial data.

Data Cleaning

Proses membersihkan data.

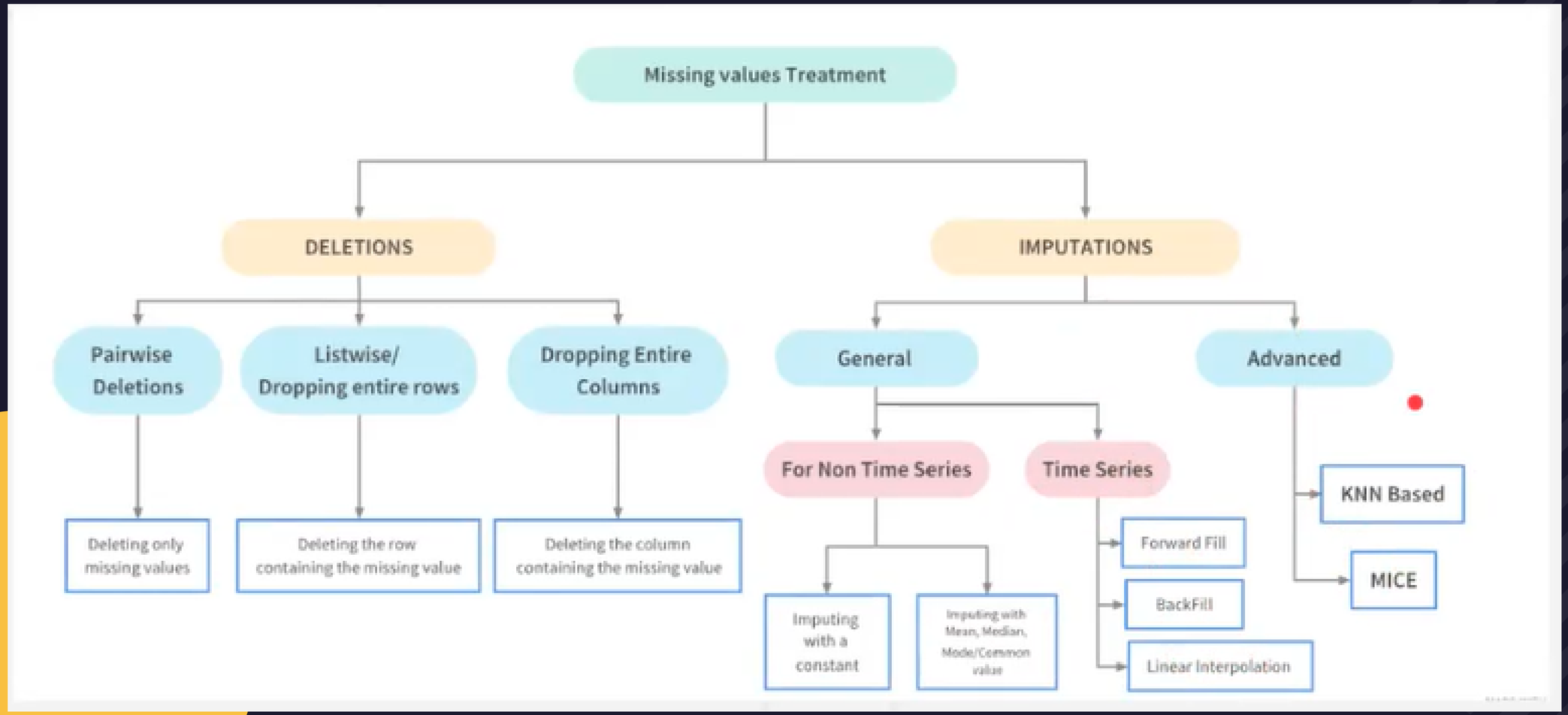
Missing Values

Missing Values Type

Import Library

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from scipy import stats
import warnings
warnings.filterwarnings("ignore")
```

Handling Missing Values



Outliers

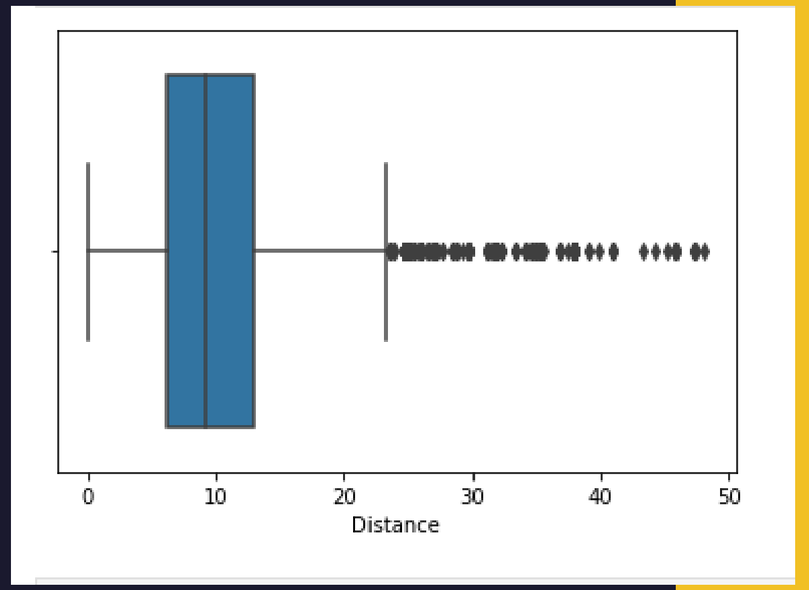
Detecting Outlier

Beberapa cara yang digunakan untuk mendekteksi outliers :

1. Boxplot
2. Scatterplot
3. Z-Score
4. Interquartile Range (IQR)

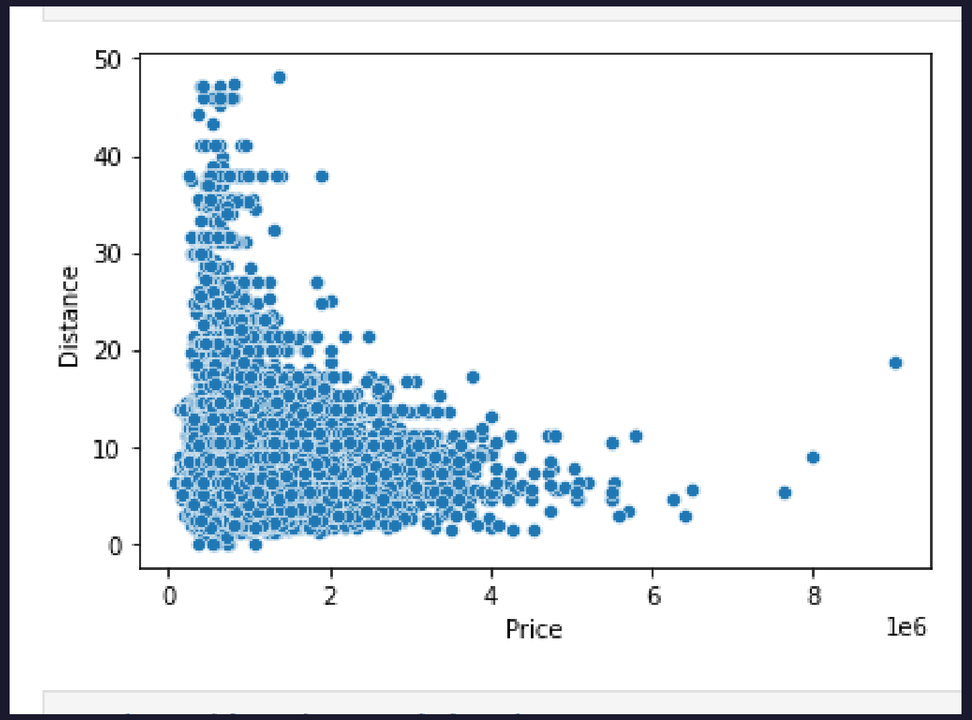
1. Boxplot

```
#cek outlier dengan boxplot
sns.boxplot(df['Distance'])
plt.show()
```



2. Scatterplot

```
# cek outlier dengan scatterplot
sns.scatterplot(data=df ,x='Price', y='Distance')
plt.show()
```



3. Z-Score

```
# cek outlier dengan zscore
z_score = np.abs(stats.zscore(df['Distance']))

# threshold 3
out_z = np.where(z_score > 3)
print("outlier", out_z)
```

4. IQR

```
# menemukan outlier
def find_outlier(data):
    # hitung nilai Q1 dan Q3
    Q1 = np.quantile(data, .25)
    Q3 = np.quantile(data, .75)

    # hitung nilai IQR
    IQR = Q3 - Q1
    min_IQR = Q1 - 1.5 * IQR
    max_IQR = Q3 + 1.5 * IQR
```

THANKYOU