Data Science | 30 Days of Machine Learning | Day - 19

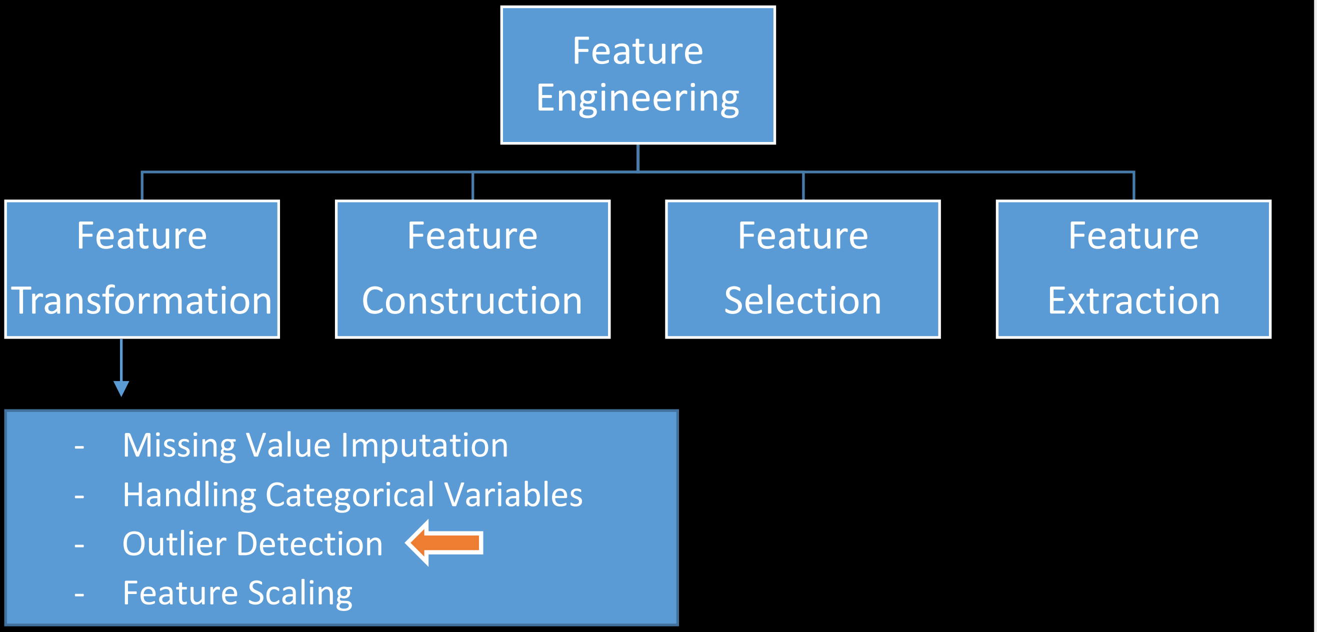
Educator Name: Nishant Dhote
Support Team: **+91-7880-113-112**

----Today Topics | Day 19----

Outliers: IQR (Interquartile Range) technique

- **What is Boxplot Distribution?**
- **What are the first quartile and third quartile in the box plot?**
- **What is the five-number summary in the box plot?**
- **What is Interquartile Range IQR?**
- **IQR Technique used for skewed distribution?**
- **IQR Percentile Rule.**

- Dataset Link GitHub: https://github.com/TheiScale/30_Days_Machine_Learning/



Feature Engineering

Feature Transformation

Feature Construction

Feature Selection

Feature Extraction

- Missing Value Imputation
- Handling Categorical Variables
- Outlier Detection
- Feature Scaling

Techniques to detect & remove outliers: -

Z-score treatment: - This technique assumes that the column follows a normal distribution.

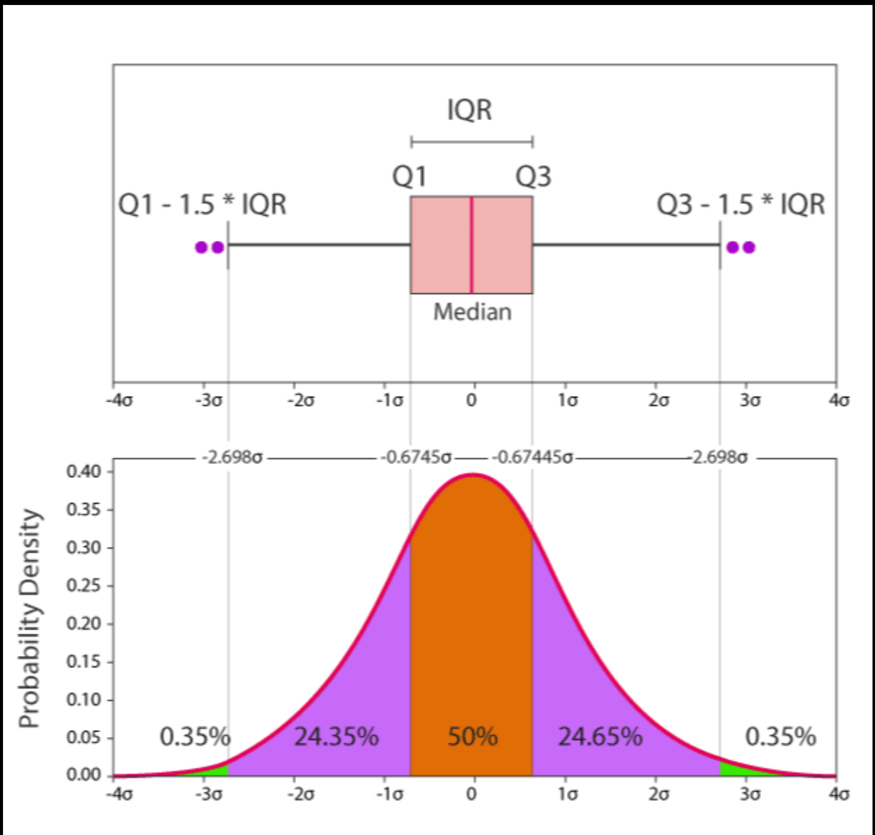
IQR (Interquartile Range) based filtering: - The IQR method involves calculating the range between the first quartile (Q1) and the third quartile (Q3).

Percentile method: - In this approach, a threshold is set based on percentiles. For example, if the threshold is set at 5%, any data point above the 95th percentile or below the 5th percentile is considered an outlier. These outliers can be removed or handled accordingly.

Winsorization:- Winsorization involves replacing outliers with values at a certain percentile, rather than removing them completely.

- What is Boxplot Distribution?

A box plot is a special type of diagram that shows the quartiles in a box and the line extending from the lowest to the highest value.



Boxplot on a normal distribution

- What are the first quartile and third quartile in the box plot?

The first quartile is the middle value of the lower half of the data, and it is represented by Q1.

The third quartile is the middle value of the upper half of the data and is represented by Q3.

- What is the five-number summary in the box plot?

The five-number summary in the box plot is minimum, maximum, median, first quartile, and third quartile.

- What is Interquartile Range IQR?

IQR is used to measure variability by dividing a data set into quartiles. The data is sorted in ascending order and split into 4 equal parts. Q1, Q2, Q3 called first, second and third quartiles are the values which separate the 4 equal parts.

Q1 represents the 25th percentile of the data.

Q2 represents the 50th percentile of the data.

Q3 represents the 75th percentile of the data.

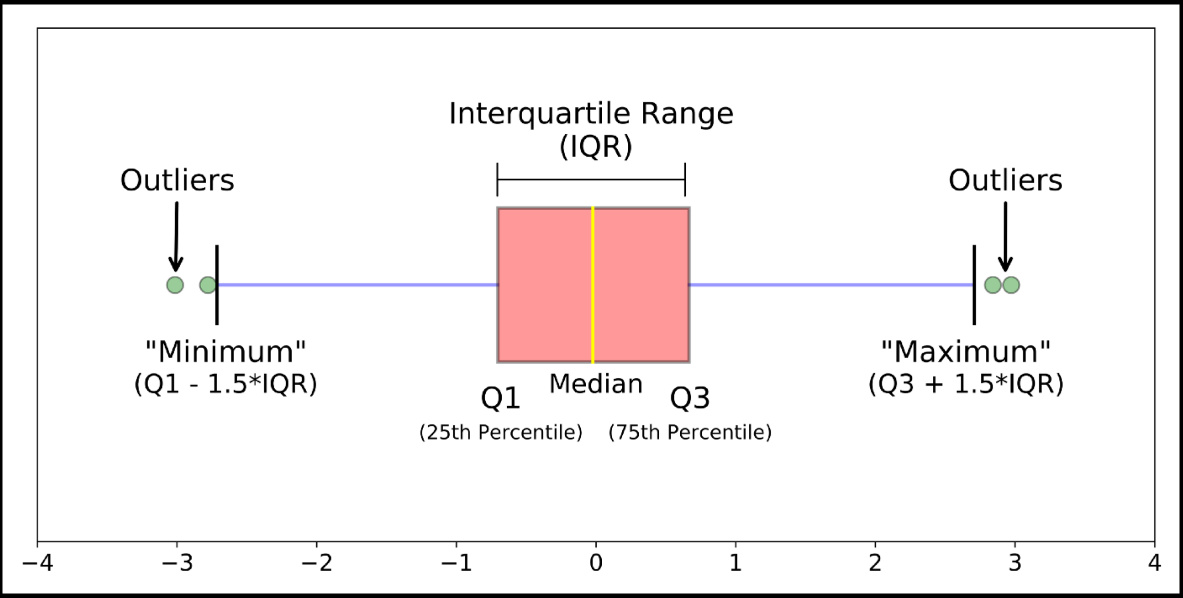
If a dataset has 2n or 2n+1 data points, then

Q2 = median of the dataset.

Q1 = median of n smallest data points.

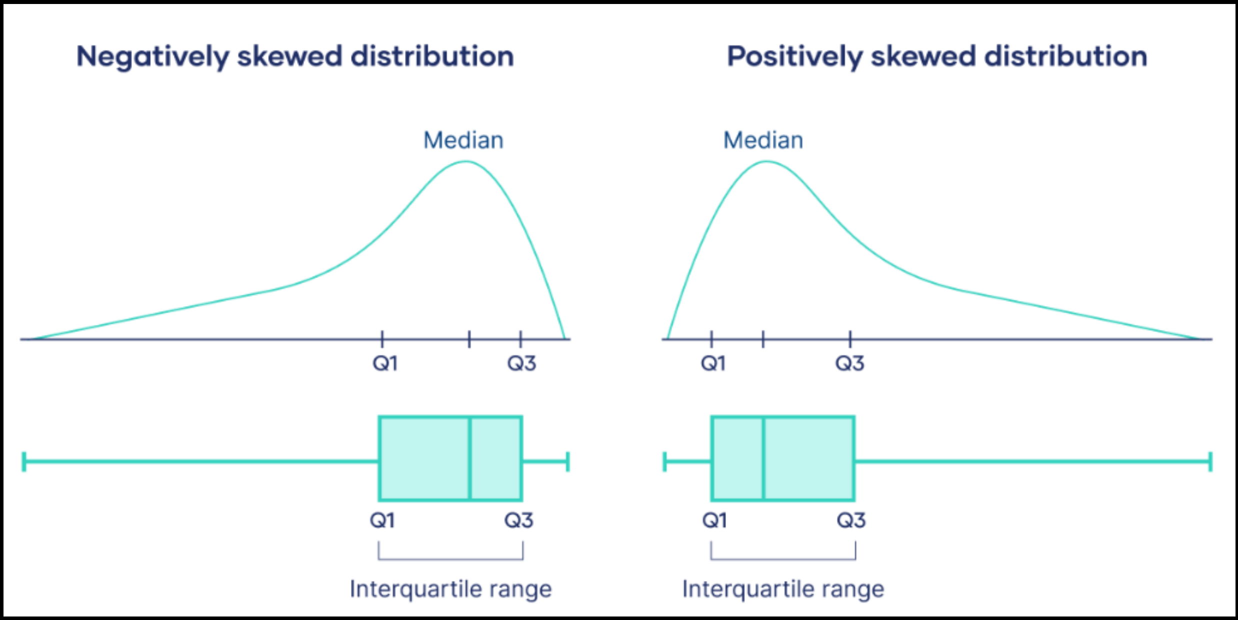
Q3 = median of n highest data points.

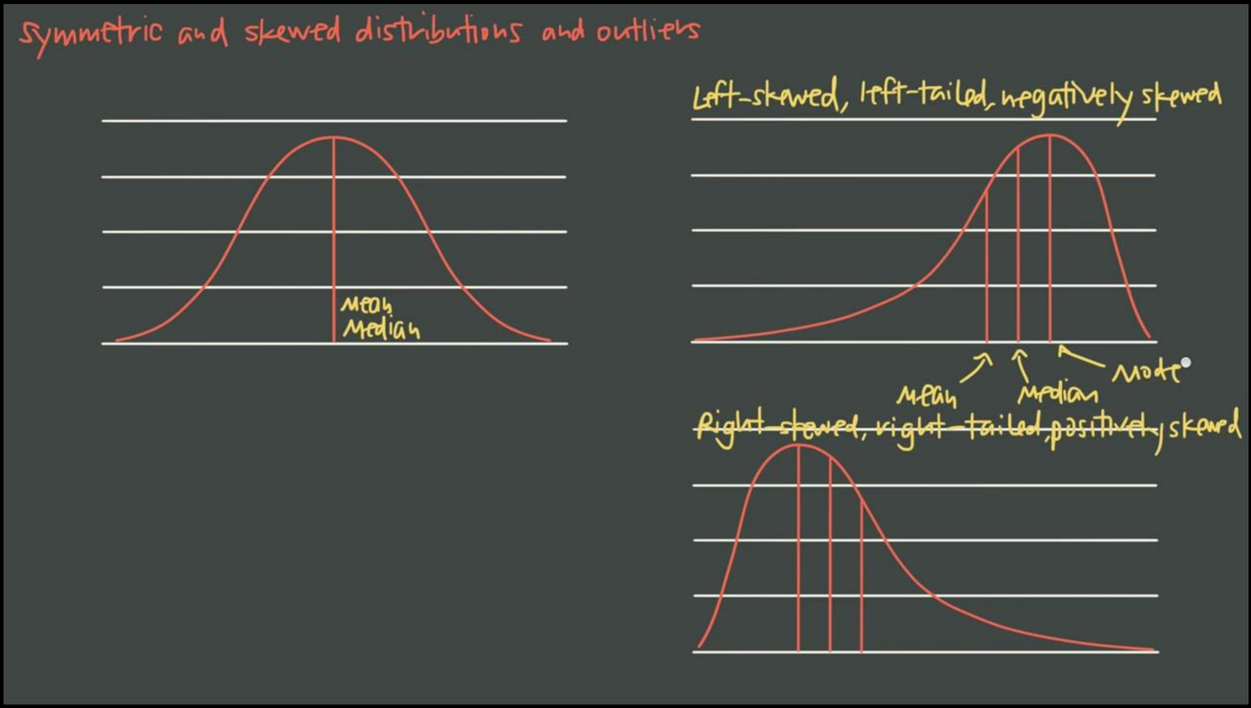
IQR is the range between the first and the third quartiles namely Q1 and Q3: IQR = Q3 – Q1. The data points which fall below Q1 – 1.5 IQR or above Q3 + 1.5 IQR are outliers.



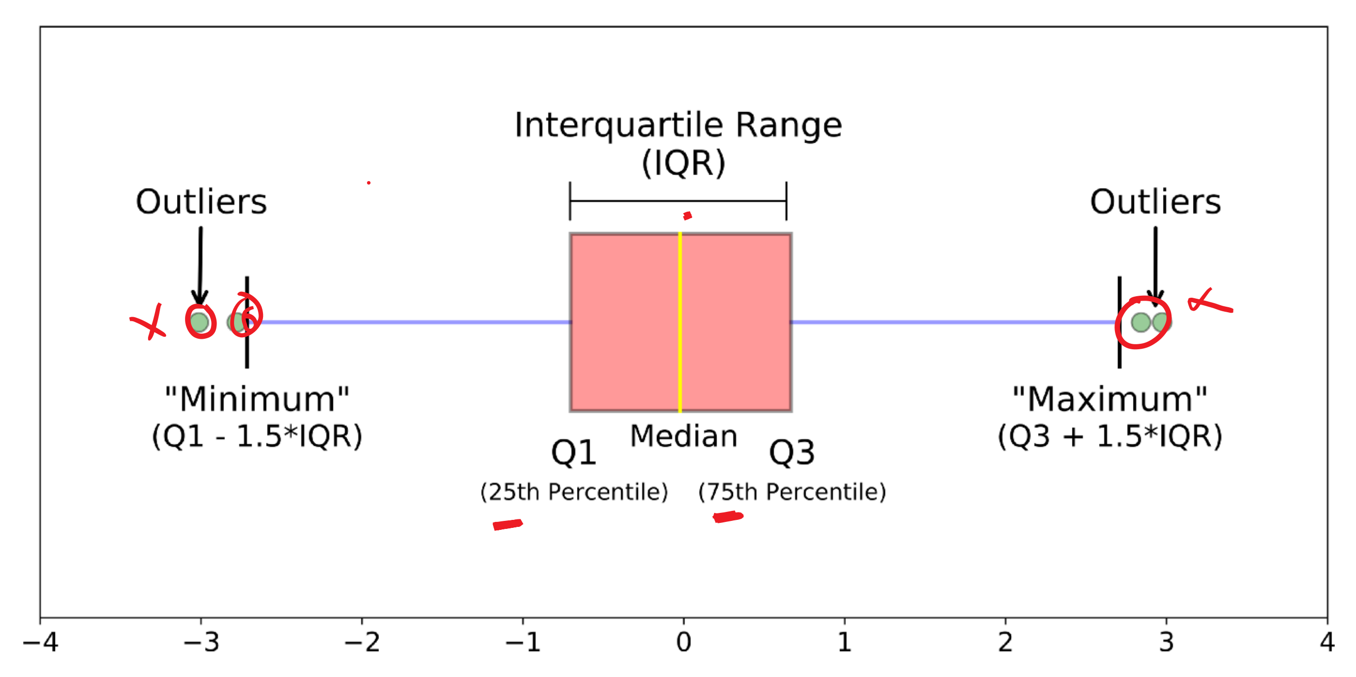
- IQR Technique used for skewed distribution?

The interquartile range is the best measure of variability for skewed distributions or data sets with outliers.




Symmetric and skewed distributions and outliers
Mean
Median
Left-skewed, left-tailed, negatively skewed
Mean
Median
Mode
Right-skewed, right-tailed, positively skewed

- IQR Percentile Rule:



Example:

Age	Gender
78	M
59	M
86	F
24	M
22	F
32	F

86 → 75%

86 — 100%

24 — 25%

IQR
- 25
- 50
- 75
- 100

86 → max
↓
75%

min ← 0
-1

<Start Coding>

```
#Import Library

import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns



#Import Dataset

df = pd.read_csv('placement.csv')

----
df

----
df.shape

----
df.sample(5)


 #Plot Show in CGPA and Placement Marks

plt.figure(figsize=(16,5))
plt.subplot(1,2,1)
sns.distplot(df['cgpa'])

plt.subplot(1,2,2)
sns.distplot(df['placement_exam_marks'])

plt.show()
```

```
#Describe placement marks

df['placement_exam_marks'].describe()


#Draw Box Plot

sns.boxplot(df['placement_exam_marks'])


#Finding IQR Value

percentile25 =
df['placement_exam_marks'].quantile(0.25)
percentile75 =
df['placement_exam_marks'].quantile(0.75)

----

percentile25

------

Percentile75


#Calculate IQR (Q3-Q1)


iqr = percentile75 - percentile25

----
iqr
```

```
#Calculate Upper and Lower Limit:


upper_limit = percentile75 + 1.5 * iqr
lower_limit = percentile25 - 1.5 * iqr

----
print("Upper limit",upper_limit)

print("Lower limit",lower_limit)


#Finding Outliers in Upper Limit:

df[df['placement_exam_marks'] > upper_limit]


----

df[df['placement_exam_marks'] > upper_limit].shape


#Finding Outliers in Lower Limit:

df[df['placement_exam_marks'] < lower_limit]



#Apply Trimming Method - 1:


new_df = df[df['placement_exam_marks'] <
upper_limit]

----

new_df.shape
```

```
#Compare Before and After (After Trimming):



plt.figure(figsize=(16,8))
plt.subplot(2,2,1)
sns.distplot(df['placement_exam_marks'])

plt.subplot(2,2,2)
sns.boxplot(df['placement_exam_marks'])

plt.subplot(2,2,3)
sns.distplot(new_df['placement_exam_marks'])

plt.subplot(2,2,4)
sns.boxplot(new_df['placement_exam_marks'])

plt.show()
```

```
#Apply Capping Method - 2:

 new_df_cap = df.copy()


new_df_cap['placement_exam_marks'] = np.where(
    new_df_cap['placement_exam_marks'] >
upper_limit,
    upper_limit,
    np.where(
        new_df_cap['placement_exam_marks'] <
lower_limit,
        lower_limit,
        new_df_cap['placement_exam_marks']
    )
)


----
new_df_cap.shape
```

```
#Compare Before and After (After Capping):
plt.figure(figsize=(16,8))

plt.subplot(2,2,1)

sns.distplot(df['placement_exam_marks'])


plt.subplot(2,2,2)

sns.boxplot(df['placement_exam_marks'])


plt.subplot(2,2,3)

sns.distplot(new_df_cap['placement_exam_marks'])


plt.subplot(2,2,4)

sns.boxplot(new_df_cap['placement_exam_marks'])


plt.show()
```