Outliers

Outliers

- An outlier is a data point in a data set that is distant from all other observations.

What are the impacts of having outliers in a dataset?

1. It causes various problems during our statistical analysis

2. It may cause a significant impact on the mean and the standard deviation

What is Interquartile Range IQR?

IQR is used to **measure variability** by dividing a data set into quartiles. 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.**

Algorithm:

1. Calculate first(q1) and third(q3) quartile

2. Find interquartile range

IQR = (q3-q1)

3. Find lower bound

Lower bound = Q1 – 1.5 IQR

4. Find upper bound

Upper bound = Q3 + 1.5 IQR

Anything that lies below the lower bound and above upper bound is an outlier

You are given height_weight.csv file which contains heights and weights of 1000 people.

You need to do this,

(1) Load this csv in pandas dataframe and first plot histograms for height and weight parameters

(2) Using IQR detect weight outliers and print them

(3) Using IQR, detect height outliers and print them

Project 1:

1. Build linear regression model for the data set MBA Salary. Csv

2. Predict salary for the percentage 74.66

Project 2:

1. Build Multi-Linear regression model for the data set cars.csv

2. Predict CO2 based on volume and weight

3. Display coefficient and intercept values

4. Predict CO2 for the volume = 1650 and Weight = 1310