EXPLORATORY DATA ANALYSIS

1) Description of Data
2) Handling Missing Data
3) Handling Outliers
4) Understanding relationships & new insights through plots

Outliers - result of mistake during data collection

Methods to detect Outliers :-

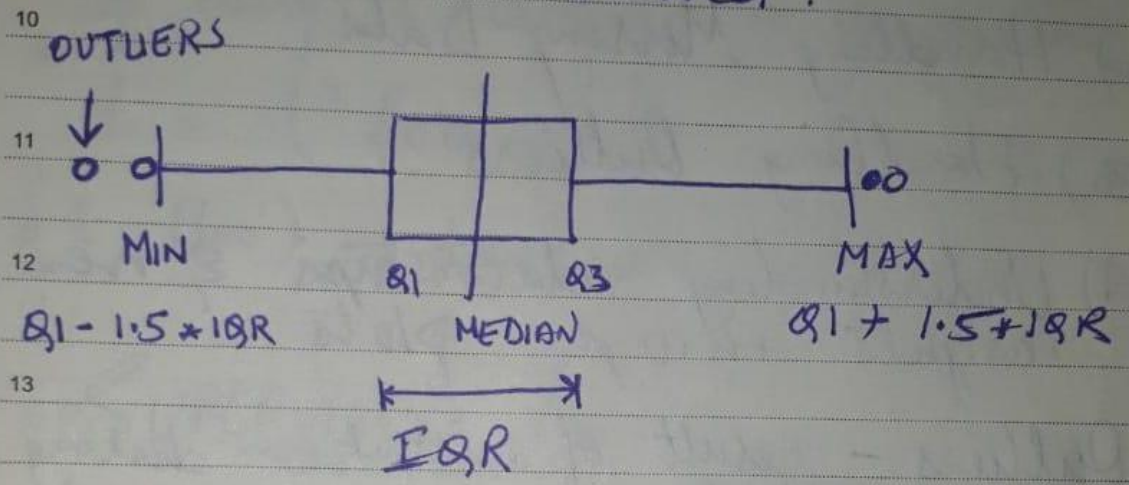
1) Boxplot
2) Scatterplot
3) Z-Score
4) IQR (Inter Quartile Range)

Notes

Boxplot :- way of displaying distribution of data based on 5 No. Summary

1) Min
2) Q_1 - 1st quartile
3) Median
4) Q_3
5) Max

It can tell you if data is
- symmetrical
- how tightly data is grouped
- how data is skewed.



notch = True ;

Z-Score :-

Signed No. of Std. Deviations by which value of an obs is above mean value of what is being observed.

05 Sunday

Data pts too far from 0 = Outliers

If Zscore > 3 or < -3

pt = Outliers

Heatmaps :- distribution of quantitative var over all combinations of 2 categorical factors

- Correlation b/w 2 random var is a no. that runs from $-1 \rightarrow 0 \rightarrow 1$ indicating strong inverse relationship, no relationship, strong direct.

Data

- Numerical
  - Discrete
  - Cont
- Categorical
  - Yes / No
- Ordinal (numerical + categorical)
  - Nos placed in categories have mathematical meaning.