Q) What do you mean by Uni-Variate Analysis? Explain in Brief.

Ans - Uni Means One, so One Variable Analysis.

① Measure of Central Tendency - Mean, Median, Mode.

② Measure of Data spread - Percentile, Range, IQR, Boxplot, Variance, Standard Deviation.

③ Variation between Variable - Covariance, Correlation Coefficient (Pearson, Spearman)

④ Measure of distribution - Skewness, Kurtosis.

① MEASURE OF CENTRAL TENDENCY

Example - Suppose take a list of value = 1, 2, 3, 4, 5, 100. (outliers) Mean = $\frac{115}{6} \approx 20$ (19.1)

Another list = a, a, b, b, b, c, d, d, d, d, mode = d. Median = 3.5

MEASURE	DEFINITION	VARIABLE TYPE	OUTLIER AFFECT
Mean	Sum of the data value divided by data count	Continuous	Yes
Median	It is an observation/value at the middle when data is sorted. If total observation is odd, we get exact value. If total is even, divide middle two numbers.	Continuous	Might/ Might not
Mode	Most frequently observed variable in a dataset.	Categorical	Might/ Might Not

② Measure of Data Spread

a) Percentile → N^{th} percentile of an observation variable is the value that cuts off first N elements of the data values when it is sorted in ascending order. Eg - 5% percentile, till 50 percentile what is the data.

Example - Consider a list → 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000

So 50% percentile will be 500.

b) Range → Measure how far apart the entire data is in terms of values.

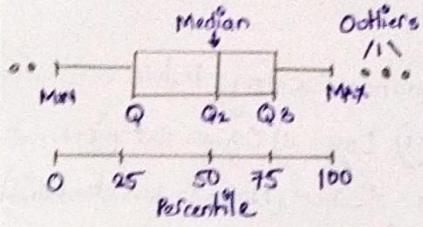
Range = Largest Value - Lowest Value.

c) Boxplot (IQR) → Graphical representation of i) Three quantile (First, Second, Third) ii) Smallest and largest value.

Example - Consider a list, 18, 34, 76, 29, 15, 41, 46, 25, 54, 38, 20, 32, 43, 22

Sort the data, 15, 18, 20 | 25, 29, 32 | 34, 38, 41 | 46, 54, 76

(Q_1 = 22, Q_2 = 33, Q_3 = 43; Min = 15, Max = 54, Outlier = 76)



Finding outliers, $Q_1 - 1.5(IQR)$ & $Q_3 + 1.5(IQR)$

IQR = Q3 - Q1 = 43 - 22 = 21

$= 22 - 1.5(21) = -9.5$

$= 43 + 1.5(21) = 74.5$.

So any data outside (-9.5, 74.5) is outliers.

Therefore 76 is an outlier.

d) Variance & Standard deviation (SD)

- Standard deviance is a square root of variance. = $\sqrt{\text{variance}}$
- Variance is the average of square difference from mean = $\frac{1}{n}\sum_{i=1}^{n}(x-\bar{x})^2$
  - we square in variance because if we don't then, when we sum up all $(x-\bar{x})$ this become zero (sum of all). That's why we square it.

→ Small SD means the values in the dataset are close to mean of dataset on average
Large SD means that values in the dataset are far away from mean of dataset.
In short, SD measures how concentrated the data is around the mean.
More concentration → Small SD.
SD is square root, so it will have same unit as original data, SD cannot be negative and lowest value is 0. And 0 is possible only if every entity is same.
Outliers affects both Variance & SD because of mean ($\bar{x}$).

→ Small SD can be goal in certain situation. For example, in manufacturing & Quality control, a car part is manufactured which is of 2cm in diameter to properly fit. So, if manufacture have high SD, then all material will not fit & will be wastage.

High SD reflect large variance in group. For example, if we look at salaries of companies from intern to CEO, SD may be very large (large variation).

③ VARIATION BETWEEN VARIABLES → Covariance, Correlation Coefficient

a) Variance → Covariance tells us how columns are related to each other.
3 types of Output, Covariance can generate.

i) +ve value → suggest variables positively related. Both variable tends to increase/decrease together.

ii) -ve value → suggest variables negatively related. If one variable increase then other variable decrease. or vice versa.

iii) 0 → Both variables are unrelated.

Problem → It give us sign (+ or −), not the strength of relationship. No upper/lower bound of the output value. [No range set for strength of relationship]

$$Cov(x, y) = \frac{1}{n}\sum_{x=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$$

b) Correlation Coefficient → Correlation find exact value of strength in the relationship and direction as well.

- Correlation Coefficient ranges from −1 to +1.

value tend close to +1 → Both variables are positively related.
value tend close to −1 → Both variables as negatively related.
value tend close to 0 → Both variables are unrelated.

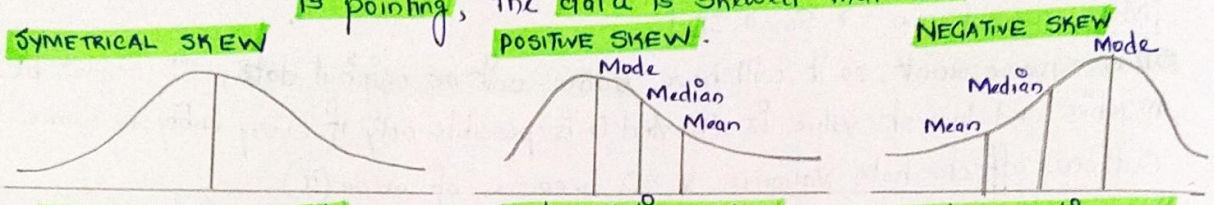
2 methods can be used in Correlation Coefficient.

i) Pearson correlation coefficient → It assume both variables are linear to each other.

ii) Spearman correlation coefficient → It does not assume (linear / non linear) among the variables.

④ Measure of distribution.

a) Skewness → Some modeling technique require normal distribution of data. So, sometimes we have skewed data. Which way the tail is pointing, the data is skewed that side.



Mean = Median = Mode (Skew = 0)

Mode > Median > Mean (Skew > 0)

Mode > Median > Mean (Skew < 0)

If skew value, ~~-0.5 < skew < 0.5~~

If Skewness = 0, data is perfectly symmetrical.

Skew less than -1 / greater than +1, then data is highly skewwed.

Skew is between -1 to -0.5 / 0.5 to 1, distribution moderately skewed.

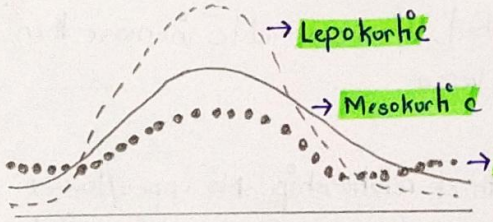
Skew is between -0.5 to 0.5, distribution is approximately symmetric.

b) Kurtosis – "Peakness" of the distribution.

~~Kurtosis could have positive value~~. Positive value means lots of data in the tails.

Negative value also means lots of data in the tail.

Standard Normal distribution has a Kurtosis of 3.



Kurtosis > 3, lepokurtic, highest peak

Kurtosis = 3, mesokurtic, normal peak.

Kurtosis < 3, platykurtic, lowest peak

- So in skewed data, the tail region may act as an outliers for statistical model and outliers adversely affect the model performance especially regression based model.
- If the data is skewed, then we use transformation (feature transformation) that will be covered in Step 4 (Feature Engineering)