INTERVIEW DAY - 006 EDA PART 2 (Bivariate Analysis)

Q) What do you mean by Bivariate Analysis ? Explain in Brief.

Ans - Bi means two so two variable analysis.
Majorly two types of data variables are there - Continuous & Categorical.
Possible combination - i) Continuous vs Continuous - Correlation Coefficient
ii) Categorical vs Categorical - Chi-Square test
iii) Continuous vs Categorical - a) T test
b) Z test
c) Anova test

1) CONTINUOUS VS CONTINUOUS DATA → 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 are negatively related.
  value tends close to 0 → Both variables are unrelated.
- 2 methods can be used in Correlation Coefficient →
  i) Pearson correlation coefficient → It assumes both variables are linear to each other.
  ii) Spearman correlation coefficient → It does not assume (linear / non-linear) among the variables.

So in short, when we have two independent continuous variables which are highly correlated, we should remove one of them because we don't want Multicollinearity issue.

Multicollinearity issue leads to regression coefficient become unreliable.
In short we are not adding incremental information but infusing the model with noise.

- If we want to keep highly correlated variable then we should use PCA.

2) CATEGORICAL VS CATEGORICAL DATA - CHI SQUARE TEST

- Chi square test determines the association between categorical variables
- Value = 0, shows complete dependency between two categorical variables
- Value = 1, shows categorical variable are completely independent.

3) CATEGORICAL VS CONTINUOUS DATA — T test, Z test, ANOVA.

- T test and Z test are basically the same.
- They assess whether the average of two groups are statistically different from each other. This analysis is appropiate from comparing the average of a numerical variables for two categories of a categorical variables.
- T test is used when $n <= 30$ and Z test is used when $n > 30$ where n is the number of samples.
- T or Z test work while dealing with two groups but ANOVA help us to compare more than two groups at a same time. (compare multiple group at a same time).