INTERVIEW DAY -

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 varables 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 numorical variables for two categories of a categorical variables.
- T test is used when n <=30 and Z test is used when ~~z~~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).

Q) What is Spurious Correlation?

Ans - Spurious correlation, or spuriousness is when two factors appear casually related but are not.

- The appearance of ~~caus~~ casual relationship is often due to similar movement which turns out to be coincidental or caused by a third "confounding factor".
- Spurious correlation can be often caused by small sample sizes or arbitrary endpoints.

Q) Give an example of Spurious Correlation.

Ans - During the festival month, Fixed Deposit sales goes high. It may seem due to bonus in festive season, people tend to invest in fixed deposit. But may be, due to tax saving (to show in March), people tend to invest in tax saving FD in later half of the year.

Q) How to spot Spurious Correlation.

i) Ensure proper representative sample ii) obtain adequate sample size.
iii) Controlling as many outside variable as possible.
iv) Use null hypothesis and check for strong p-value.