Day 5 of 100 Data Science Interview Questions Series!!

Q 21.) What if your dataset has 1 column related to multiple columns of your dataset, how will you handle your data?

In this case, if we draw a correlation heat map, we will see a lot of collinearity and won't be able to decide the columns we want to remove.

This problem is known as the MultiCollinearity Problem. One good way to deal with Multicollinearity is the Variance Inflation Factor (VIF).

Variance inflation factors (VIF) measure how much one column is related to all others in terms of their dependency. It is obtained by regressing each independent variable, say X on the remaining independent variables (say Y and Z) and checking how much of it (of X) is explained by these variables.

Intuitively, we just perform linear regression with all the columns as target variable one by one, in eat iteration we keep 1 column as target and the rest as features and see how it is explained by other variables. We do this with all the columns.

Hence,

In general, the Higher the VIF, the higher the R2 which means the variable X is collinear with Y and Z variables. Columns with values of VIF above 2 or 5 and in some cases 10, are removed.

Q 22.) How to handle skewness?]What are the mathematical properties of skewed data? How can you fix it?

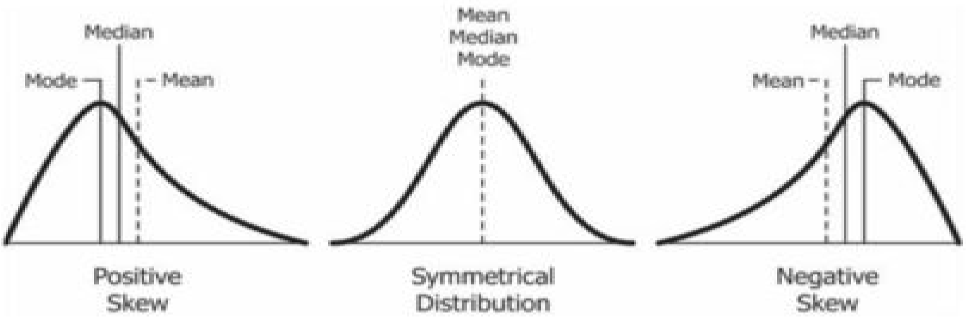
Skewness is when the distribution of data which is concentrated in 1 side(left or right) more, or it is sort of collected on one side. data can be left-skewed or right-skewed.

Right skewed is when data is concentrated on the left, and the tail is on right. This is also known as positive skewed. Left is -ve skewed when the tail is on left.

Skewness is caused by the presence of Outliers in the data.

In mathematical terms, the right skewness is when the mean is greater than the median.

We can remove Skewness by omitting Outliers or taking Log of the features.



Q 23.) Differentiate between univariate, bivariate, and multivariate analysis.

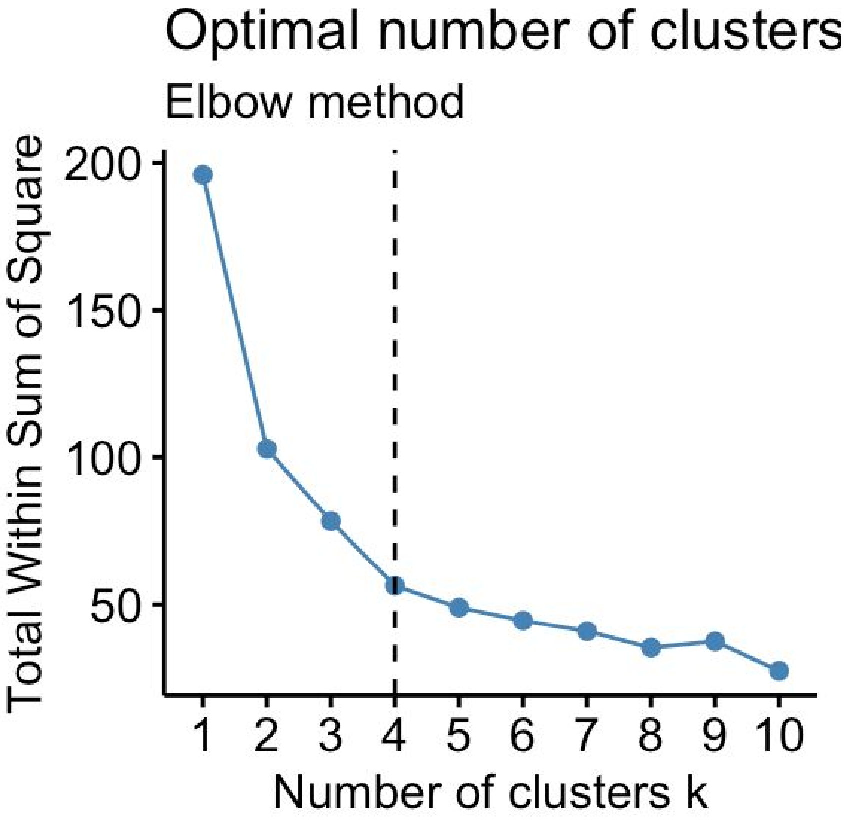
- Univariate data contains only one variable. The purpose of the univariate analysis is to describe the data and find patterns that exist within it.
- Bivariate data involves two different variables. The analysis of this type of data deals with causes and relationships and the analysis is done to determine the relationship between the two variables.
- Multivariate data involves three or more variables, it is categorized under multivariate. It is similar to a bivariate but contains more than one dependent variable.

Q 24.) In K means clustering, how will you decide the number of clusters to chose? How do you deal with the bad initialization problem?

In K-means clustering we find clusters of similar data points. If we chose more clusters than what are needed, it'll show wrong clustering.

One Brute-Force way to fix is trying all possible clusters from a given range and see what works. The way is do that is calculating the sq distance btw them. This is called the Elbow method, we use within the sum of squares (WSS) to see which clusters work best. WSS is defined as the sum of the squared distance between each member of the cluster and its centroid.

Less number of clusters means more WSS. We continue to increase our number of clusters to the point where our square distance is not decreasing much.



Q 25.) 'People who bought this also bought…' recommendations seen on Amazon are a result of which algorithm?

Recommendation systems are built by 2 methods. Content-based filtering and collaborative filtering. Collaborative filtering explains the behavior of other users and their purchase history in terms of ratings, selection, etc.

It takes into account all the choices of other users and features in order to recommend it. The engine makes predictions on what might interest a person based on the preferences of other users. In this algorithm, item features are unknown.

Learn more about recommendation systems as they are an important Unsupervised Machine Learning ALgorithms.

I will build a song recommendation project using RBM and AutoEncoder on my website https://www.aiunquote.com soon this week. Stay updated. :)

- Alaap Dhall

Follow Alaap Dhall on LinkedIn for more insights in Data Science and Deep Learning!!

Visit https://www.aiunquote.com for a 100 project series in Deep Learning.