Week 6 Daily Challenge

1. Hypothesis Testing is a statistical inference method that helps to draw conclusions and create insights about attributes of a data set by comparing it to an idealized dataset. Hypothesis testing is carried out by making an assumption concerning a dataset as compared to a null hypothesis which assumes no relationship between the two datasets. A significance level is then chosen above or below which the null hypothesis is rejected and the comparison is deemed statistically significant or vice versa.

In data science, hypothesis testing is useful at the initial stage in investigating connections and relationships between various parts of the data and guiding the data scientists on the most important feature variables to take into consideration while running analytics and even for prediction as well.

1. Cluster Analysis is an unsupervised data science tool in which related observations are classified into various groups based on how similar they are. Kmeans clustering is an example of a clustering algorithm. Outlier analysis on the other hand involves identifying observations that differ widely from the mean by a given amount of Standard deviation. Outlier analysis also helps to draw insights from the data by figuring out what may be responsible for the bizarre observations; perhaps an error in the data collection or measuring process or an anomaly that should be taken into consideration.
2. Data Preprocessing tasks:

**Feature Scaling/Normalization**: This involves scaling or normalizing the data such that all observations are within the same range (for example -1 to +1). This will prevent the prediction algorithms from assigning wrong weights to variables simply based on their size.

**Handling Missing Data:** There are various ways to fix the problem of missing data depending on the kind of data and the results of the EDA. In some cases, observations with missing data are simply deleted, while in other cases, a metric such as the mean of the whole feature variable or the mean of various segments are used to fill in for the missing data

**Dummy Variables:** This often occurs for categorical variables that are in string format or some other data type that may not be convenient for the predictive or analytic algorithm to handle. Dummy variables are numbers (such as 0, 1, 2, 3, etc) that are used to represent various categories in the categorical variable space.

**Feature Reduction:** In circumstances where there are so many feature variables, it is wise to either take the most important ones and discard variables that do not affect the target variable significantly. However in some cases, so as to preserve information, all of the feature variables are collapsed into one or two variables both for convenience and thoroughness. Principal Component Analysis is an example of a technique that does this collapse of feature variables into fewer numbers.