General Subjective Questions

1. Explain the linear regression algorithm in detail

Answer: Linear Regression algorithm is most popular form of Machine learning, which used for based on supervised learning. Linear Regression models a target prediction value based on independent variables. Linear regression algorithm finds the relationship between forecasting and variables.

Linear regression predicts a dependent variable (y) based on independent variable (x). This technique is used to find linear relationship between independent variable (x) and dependent variable.

2. Explain the Anscombe's quartet in details?

Answer: Anscombe's quartet is a group of four data sets which are identical in simple descriptive statistics but differently graphed. Each datasets consists of eleven (x,y) points. This was constructed in 1973 by statistician Francis Anscombe. This tells us about the importance of visualizing the data before using algorithms out .

3. What is pearson's R?

Answer: Pearson's R is a numerical summary of the strength of the linear association between variables. If variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the coefficient will be negative.

The Pearson's correlation coefficient varies between -1 and +1 where:

a. r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)
b. r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)
c. r = 0 means there is no linear association
d. r > 0 < 5 means there is a weak association
e. r > 5 < 8 means there is a moderate association
f. r > 8 means there is a strong association

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Answer: Scaling means you're transforming your data so that it fits within a specific scale. It is one type of data pre-processing step where we will fit data in specific scale and speed up the calculations in an algorithm.

Collected data contains features varying in magnitudes, units and range. If scaling is not performed than algorithm tends to weigh high values magnitudes and ignore other parameters which will result in incorrect modeling.

Difference between Normalizing Scaling and Standardize Scaling:

1. In normalized scaling minimum and maximum value of features being used whereas in Standardize scaling mean and standard deviation is used for scaling.
2. Normalized scaling is used when features are of different scales whereas standardized scaling is used to ensure zero mean and unit standard deviation.
3. Normalized scaling is scales values between (0,1) or (-1,1) whereas standardized scaling is not having or is not bounded a certain range.
4. Normalized scaling is affected by outlies whereas standardized scaling is not having any affect by outliers.
5. Normalized scaling is used when we don't know about the distribution whereas standardized scaling is used when distribution is normal.
6. Normalized scaling is called as scaling normalization whereas standardized scaling is called as Z-score Normalization.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

<u>Answer:</u> VIF is infinite when there is perfect correlation between two independent variables. Incase of perfect correlation , R^2=1 which lead to 1/(1-R^2) infinity. To solve this problem, we need to drop on of the variables which is causing the multicollinearity. An infinite VIF indicates the variable can be expressed by a linear combination of other variables.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

<u>Answer:</u> Q-Q plot stands for Quantile-Quantile plot. It is a graphical representing tool which is used to assess if a data is plausibly came from some theoretical distribution such as normal, exponential or uniform distribution.

Use of Q-Q plot: Q-Q plot is used to compare the shapes of distributions and provide graphical representation of how properties e.g., properties, scales etc. are similar or different in two distributions.

Importance of Q-Q plot: When there are 2 data sets, if desirable to know if assumption of a common distribution is justified. If so, location and scale estimators can pool both data sets to obtain estimates the common location and scale. If 2 data sets are different, it will provide some understanding on differences. The Q-Q plot can provide more insight on difference than the analytical methods chi-square etc.

Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Answer: Analysis of categorical variables done thru boxplot and bar chart. Please see below for some of the important points:

a. Fall season attract more bike rentals as compared to other seasons.
b. Bike rental increases in the year 2019 as compared to 2018.
c. Majority of bike rentals happen in the month of May, Jun, Jul, Aug and Sep.
d. Clear weather has more bike rentals booking.
e. Fri, sat and Sun has a greater number of bike rental bookings as compare to start of week.

2. Why is it important to use **drop_first=True** during dummy variable creation?

Answer: drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Answer: Temperature (temp) variables has highest correlation with the target variable.

4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Answer: Linear Regression Model based on below assumptions:

a. Normality of error.
b. Homoscedasticity
c. Multicolinearity

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Answer: Top 3 features that has significant impact towards explaining the demand of the shared bikes:

a. Temperature
b. Weather
c. Year