General Subjective Questions

1) Explain the linear regression algorithm in detail.

A) Linear regression is a statistical technique that is used to understand the linear relationship between a dependent variable and one or more independent variables. It helps in predicting the value of the dependent variable based on the values of the independent variables. To find the line of best fit, which can be used to predict the dependent variable, the sum of squared differences between the predicted values and the actual values of the dependent variable is minimized. This line is known as the regression line.

The equation for a simple linear regression model is: $y = b0 + b1 * x$

Here, y is the dependent variable, x is the independent variable, b0 is the intercept and b1 is the slope of the line. The values of b0 and b1 are chosen in a way that minimizes the sum of squared differences between the predicted values and the actual values of the dependent variable.

There are various methods to estimate the values of b0 and b1, such as the ordinary least squares method, gradient descent method, and least absolute deviations method.

2) Explain the Anscombe's quartet in detail.

Ans) Anscombe's quartet is a set of four datasets that have the same statistical properties but appear different when plotted. It was created to show the importance of visualizing data before analyzing it. All four datasets have the same mean, variance, correlation, and regression line.

3) What is Pearson's R?

Ans) Pearson's R quantifies the strength and direction of linear relationships between two variables. A value close to 0 indicates a weak or nonexistent relationship, while a value near 1 reflects a strong connection between the variables.

Pearson's R is calculated as:

$R = \sum(x - \bar{x})(y - \bar{y}) / \sqrt{(\sum(x - \bar{x})^2 \sum(y - \bar{y})^2)}$

Where x and y are the two variables being measured, $\bar{x}$ is the mean of the x values, and $\bar{y}$ is the mean of the y values. The numerator is the sum of the products of the deviations of the x and y values from their means, and the denominator is the product of the standard deviations of the x and y values.

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

Scaling involves transforming the values of a variable to a specific range or scale. There are several reasons for scaling:

1. Normalizing values: This scales the values of a variable to a specific range, usually between 0 and 1, to make sure all variables are on the same scale and can be compared directly.
2. Standardizing values: This scales the values of a variable to have a mean of 0 and a standard deviation of 1, often to ensure that the values of a variable follow a normal distribution for certain statistical tests.
3. Improving the performance of machine learning algorithms: Many machine learning algorithms work better when input variables are on the same scale, so scaling the variables can improve algorithm performance.

Normalizing and standardizing are different techniques. Normalizing scales values to a specific range, usually between 0 and 1, while standardizing scales values to have a mean of 0 and a standard deviation of 1. Both techniques may be useful depending on the needs of the analysis.

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

Ans) When two variables are associated with each other in a model, the variance of one variable is inflated by the presence of the other. This phenomenon occurs because both factors contribute to variability within that particular data set.

The model appears to be over-simplified, and may require further adjustment in order to improve its performance. One or more variables might need to be excluded from the equation for optimal results.

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

Ans) A Q-Q plot is a graphical tool used to test whether data falls within a specific distribution, such as the normal distribution. By comparing the dataset's observed values to theoretical values calculated from that particular distribution, it can help analysts determine if their data matches expected patterns.

One common way to check the normality of data is by using Q-Q plots. This plot can help determine if residuals are normally distributed and therefore meet the assumptions of linear regression. If residuals are not normal, this may mean that the model used in analysis isn't suitable for the data, which would result in inaccurate predictions.

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?

Ans) There are a few variables that have negative coefficients in this model. This means that as these variables increase, the dependent variable tends to decrease. These variables include spring, misty, cloudy, and light snow.

1. High demand in 2019
2. demand is more during spring and winter
3. month: jan,july,sep,nov, dec
4. If its a holiday
5. windspeed
6. Temp
7. snowrain
8. misty

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

Ans) It reduces the number of dummy variables and helps to avoid the dummy variable trap. This is why it is important to use "drop_first=True" when creating dummy variables.

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

Ans) temp, atemp, year

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

Ans) 1)Plot the residuals

2) Check the distribution of the residuals

3) Check for multicollinearity

4) Check for outliers

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

Ans) **temp, weathersit_Light_snowrain , yr_1(2019)**