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 : Season, yr, Months, holiday and weathersit have significant impact on dependent variable.

Or

Demands increases in the month of 3, 5, 6, 8 ,9, 7 , 10 and yr

Demand decreases if it is holiday , Spring, Light rain_Light snow_Thunderstorm, Mist_cloudy, Sunday.

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

Ans: 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?

Ans : temp/atemp

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

Ans : Checking ASSUMPTION OF NORMALITY:

Plot the histogram of the error terms using residuals(Actual-Predicted)

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

Ans : Light rain_Light snow_Thunderstorm, yr, spring.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans : Linear regression is a quiet and simple statistical regression method used for predictive analysis

and shows the relationship between the continuous variables. Linear regression shows the linear relationship between the independent variable (X-axis)

and the dependent variable (Y-axis), consequently called linear regression. If there is a single input variable (x),

such linear regression is called simple linear regression. And if there is more than one input variable,

such linear regression is called multiple linear regression.

The linear regression model gives a sloped straight line describing the relationship within the variables.

best-fit line linear regression uses a traditional slope-intercept form.

Linear Regression equation

y= Dependent Variable.

x= Independent Variable.

a0= intercept of the line.

a1 = Linear regression coefficient

A regression line can be a Positive Linear Relationship or a Negative Linear Relationship.

Positive Linear Relationship

If the dependent variable expands on the Y-axis and the independent variable progress on X-axis, then such a relationship is termed a Positive linear relationship.

Negative Linear Relationship

If the dependent variable decreases on the Y-axis and the independent variable increases on the X-axis,

such a relationship is called a negative linear relationship.

2. Explain the Anscombe's quartet in detail.

Ans : Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics,

but there are some peculiarities in the dataset that fools the regression model if built.

They have very different distributions and appear differently when plotted on scatter plots.

It was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting the graphs before analyzing and model building,

and the effect of other observations on statistical properties.There are these four data set plots which have nearly same statistical observations,

which provides same statistical information that involves variance, and mean of all x,y points in all four datasets.

This tells us about the importance of visualising the data before applying various algorithms out

there to build models out of them which suggests that the data features must be plotted in order to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data, etc.

Also, the Linear Regression can be only be considered a fit for the data with linear relationships and is incapable of handling any other kind of datasets.

3. What is Pearson's R

Ans : The Pearson product-moment correlation coefficient (or Pearson correlation coefficient, for short) is a measure of the strength of a linear association between two variables and is denoted by r. Basically, a Pearson product-moment correlation attempts to draw a line of best fit through the data of two variables, and the Pearson correlation coefficient, r, indicates how far away all these data points are to this line of best fit (i.e., how well the data points fit this new model/line of best fit).

The Pearson correlation coefficient, r, can take a range of values from +1 to -1. A value of 0 indicates that there is no association between the two variables. A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable. A value less than 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases

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

and standardized scaling?

Ans : Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.

Consider the two most important ones:

Min-Max Normalization: This technique re-scales a feature or observation value with distribution value between 0 and 1.

X_{\text {new }}=\frac{X_{i}-\min (X)}{\max (x)-\min (X)}

Standardization: It is a very effective technique which re-scales a feature value so that it has distribution with 0 mean value and variance equals to 1.

X_{\text {new }}=\frac{X_{i}-X_{\text {mean }}}{\text { Standard Deviation }}

Normalization vs. standardization is an eternal question among machine learning newcomers. Let me elaborate on the answer in this section.

Normalization is good to use when you know that the distribution of your data does not follow a Gaussian distribution. This can be useful in algorithms that do not assume any distribution of the data like K-Nearest Neighbors and Neural Networks.

Standardization, on the other hand, can be helpful in cases where the data follows a Gaussian distribution. However, this does not have to be necessarily true. Also, unlike normalization, standardization does not have a bounding range. So, even if you have outliers in your data, they will not be affected by standardization.

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

Ans : If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

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

Ans : Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.

This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.