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?**

   Most of the categorical variables like weathersit, season has not much impact on the dependent variable

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

   Drop_first = True drops the first column during dummy variable creation. It helps in reducing extra column creation during dummy value creation.

   Let's say there are 3 categories in a column. By creating 2 columns automatically infers the third column

   This is to reduce the redundancy of extra column

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

   temp and atemp variables

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

   Residual analysis -

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

   Holiday, Season and Year - All these variables have low p and low VIF (less than 5)

General Subjective Questions

6. **Explain the linear regression algorithm in detail**

   Linear regression can be simple linear regression or multiple linear regression

   Formula for simple linear regression

   $Y = B0 + B1x + E$

Where B0 – Intercept and B1 – Slope , x is the independent variable and Y is the dependent variable

E is the error

For multi linear regression

Y = B0 + B1X1 + B2X2 +.. BnXn + E

Based on the number variables the formula adds all the variable in the above formula

The coefficients (B1, B2, and so on) explain the correlation of the independent variables with the dependent variable. The sign of the coefficients (+/-) designates whether the variable is positively or negatively correlated.

2. **Explain the Anscombe's quartet in detail**

It comprises of 4 data sets that have nearly identical sample descriptive stastics but have very different distributions

Anscombe's quartet is used to illustrate the importance of plotting data before you analyze it and build your model.

This suggests the data features must be plotted to see the distribution of the samples that can help you identify the various anomalies present in the data (outliers, diversity of the data, linear separability of the data, etc.).

3. What is Pearson's R?

Pearson's correlation coefficient, R, is the most common way of measuring a linear correlation. It's between -1 and 1 that measures the strength and direction between two variables

Between 0 and 1 – Positive correlation – when one variable changes the other variable changes in the same direction

0 – No correlation – There is no relationship between variables

Between 0 and – 1 – Negative correlation – When one variable changes, the other variable changes in the opposite direction

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

Scaling is a process of bringing all variables to a comparable ranges. It is called as data normalization. It is performed during the data pre-processing to handle variance in the magnitude of the data

The machine learning models assign weights to the independent variables according to their data points and conclusions for output. In that case, if the difference between the data points is high, the model will need to provide more significant weight to the farther points, and in the final results, the model with a large weight value assigned to undeserving features is often unstable. This means the model can produce poor results or can perform poorly during learning.

Min-Max Normalization (or Min-Max scaling) - This technique rescales a feature with values between 0 and 1 or -1 to 1 if there are negative values in the dataset

Standardized scaling - This technique rescales a feature with mean value as 0 and variance equal to 1

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

If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables.

A large VIF on an independent variable indicates a highly collinear relationship to the other variables that should be considered or adjusted for in the structure of the model and selection of independent variables.

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

Q-Q plot compares distribution of two sets of data. Q-Q plot is a plot of the quantiles of the first data set against the quantiles of the second data set