**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? (3 marks)**

Ans=> Yes, we saw the effect of categorical varaible on the dependent varaible. We saw that season 1 tended to have a low count compared to the other season. We also saw that year one tends to have more ride than year one. The reason for the year having more rides on year one is because of the customer relationship. We also find that weather sit also impacts the independent variable.

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

Ans=> When creating dummy variables, it is important to drop the first level to avoid perfect multicollinearity. Perfect multicollinearity occurs when the dummy variables are perfectly correlated with each other. This can lead to problems with model estimation, and can make it difficult to interpret the results. Dropping the first level helps to avoid perfect multicollinearity and makes the results easier to interpret.

**3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)**

=> registered has the highest correlation with the target variable

**4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)**

=> After building the linear regression model on the training set, we can validate the assumptions of the model in a number of ways. Firstly, we can check the linearity assumption by plotting the predicted values against the independent variable. If the relationship is linear, we should see a straight line on the plot. Secondly, we can check the normality assumption by plotting a histogram of the residuals. The residuals should be normally distributed. Thirdly, we can check the homoscedasticity assumption by plotting the residuals against the predicted values. If the relationship is homoscedastic, we should see a uniform scatterplot. Finally, we can check the independence assumption by plotting the residuals against the order of the observations. If the residuals are independent, we should see a random scatterplot.

**5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)**

=> registered, casual, temp are the top 3 features contributing significantly towards explaining the demand of the shared bikes.

**General Subjective Questions**

**1. Explain the linear regression algorithm in detail. (4 marks)**

Ans=> Linear regression is a statistical technique that is used to predict future values of a dependent variable, based on values of an independent variable. The technique is based on the linear relationship between the dependent and independent variables.

The linear regression algorithm is used to calculate the values of the slope and intercept of the regression line. The slope is the rate of change of the dependent variable with respect to the independent variable. The intercept is the point on the y-axis where the regression line crosses.

The linear regression algorithm is used to minimize the sum of the squares of the residuals. The residual is the difference between the actual value of the dependent variable and the predicted value. The predicted value is the value of the dependent variable that is predicted by the regression line.

The linear regression algorithm is used to find the best fit line for the data. The best fit line is the line that minimizes the sum of the squares of the residuals.

**2. Explain the Anscombe’s quartet in detail. (3 marks)**

Ans=> The Anscombe’s quartet is a set of four datasets that have identical summary statistics but appear very different when graphed. The quartet was created by Francis Anscombe to illustrate the importance of graphical data analysis.

The first dataset consists of 11 points in a linear pattern. The second dataset consists of 11 points that are clustered around a horizontal line. The third dataset consists of 11 points that are clustered around a vertical line. The fourth dataset consists of 9 points that are clustered in a diagonal line.

Although the summary statistics for each dataset are identical, the datasets look very different when graphed. This illustrates the importance of graphical data analysis, as the graphs reveal patterns that are not immediately apparent from the summary statistics.

**3. What is Pearson’s R? (3 marks)**

Ans=> Pearson's R is a correlation coefficient that is used to measure the linear relationship between two variables. It is a statistical measure that is used to assess the strength of the relationship between two variables. The higher the Pearson's R value, the stronger the relationship between the two variables.

**4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)**

Ans=> Scaling is the process of transforming data so that it can be more easily analyzed. This is often done by converting data into a common format, such as converting all units to be the same. Scaling can also involve more complex transformations, such as changing the shape of data so that it can be more easily analyzed.

Scaling is often performed so that data can be more easily compared. For example, if data is collected from different geographical areas, it may be useful to scale the data so that all units are in the same format. This makes it easier to compare the data and look for patterns.

There are two main types of scaling: normalized scaling and standardized scaling. Normalized scaling involves transforming data so that it is between 0 and 1. Standardized scaling involves transforming data so that it has a mean of 0 and a standard deviation of 1.

**5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)**

Ans=> You might have observed that sometimes the value of VIF is infinite. There are a few reasons why this might happen:

1) One of the predictor variables is a perfect linear combination of the other predictor variables. In other words, one variable can be predicted perfectly by a linear combination of the other variables.

2) There is perfect multicollinearity among the predictor variables. This means that the predictor variables are highly correlated with each other, and there is very little information content in each individual predictor variable.

3) There is complete separation among the predictor variables. This means that there is very little overlap between the predictor variables, and they can be perfectly separated by a linear combination.

In all of these cases, the VIF will be infinite because the linear model cannot be fit properly.

**6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)**

Ans=> A Q-Q plot is a graphical tool that is used to assess whether or not a set of data is normally distributed. The plot compares the quantiles of the data to the quantiles of a standard normal distribution. If the data is normally distributed, the points on the plot will fall along a straight line. If the data is not normally distributed, the points on the plot will not fall along a straight line.

The Q-Q plot is a useful tool for assessing the normality of data because it can be used to identify outliers. Outliers can have a significant impact on the results of a linear regression analysis, so it is important to identify them prior to running the analysis. The Q-Q plot can also be used to assess the homogeneity of variance, which is another important assumption of linear regression.