Assignment-based Subjective Questions Answers

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

Ans:
The categorical variables from the dataset don't affect the dependent variable. These categorical variables are not associated with changes in the response variable, and they likely don’t contribute significantly to explaining the variability observed in the dependent variable.

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

Ans:
It is important to use drop_first=True during dummy variable creation to avoid the dummy variable trap and multicollinearity issues. When creating dummy variables for categorical features, omitting one level using drop_first=True prevents perfect multicollinearity between the dummy variables. Without dropping one dummy variable, the model may encounter issues during estimation due to the linear dependency among the dummy variables, leading to unreliable results.

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

Ans:
By examining the pair-plot among the numerical variables, the variable that exhibits the highest correlation with the target variable is the registered column and has the highest correlation of **0.945411** with the target variable cnt.

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

Ans:
To validate we examine the residuals (the differences between observed and predicted values). Plot a histogram of residuals to check for normality. The residuals should ideally follow a normal distribution.
Create a Q-Q plot to compare the distribution of residuals against a theoretical normal distribution.

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

Ans:
The top 3 features that contributed significantly towards explaining the demand of the shared bikes based on the final model are, **registered, casual, temp.**

General Subjective Questions Answers

1Q. Explain the linear regression algorithm in detail.

Ans:
Linear regression is a statistical method used for modeling the relationship between a dependent variable and one or more independent variables by fitting a linear equation to observed data. The linear equation is represented as:

$Y = \beta_0 + X_1\beta_1 + X_2\beta_2 + X_3\beta_3 + \ . \ . \ . \ + X_n\beta_n + \varepsilon$
Here:

Y is the dependent variable,

$X_1, X_2, X_3, ..., X_n$ are independent variables,

β_0 is the intercept,

$\beta_1, \beta_2, \beta_3, ..., \beta_n$ are the coefficients, and

ε is the error term.
The goal of linear regression is to find the values of the coefficients that minimize the sum of squared differences between the observed and predicted values. This is generally done by using the least squares method.

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

Ans:
Anscombe's quartet is a set of four datasets that have nearly identical simple descriptive statistics but differ significantly when graphed. Each dataset consists of 11 points and includes variables X and Y. The quartet was created by Francis Anscombe to emphasize the importance of graphing data before analyzing it and to show the impact of outliers on statistical properties.

Despite having similar means, variances, and correlation coefficients, the datasets can have different shapes when plotted. This highlights the limitations of relying solely on summary statistics and the importance of data visualization in understanding the underlying patterns in data.

3Q. What is Pearson's R?

Ans:
Pearson's correlation coefficient (r) is used to find the linear relationship between two variables. Where it has a range of -1 to 1, where:

r=1 indicates a perfect positive linear correlation,
r=−1 indicates a perfect negative linear correlation, and
r=0 indicates no linear correlation.
The formula for Pearson's correlation coefficient is:

$$r = \frac{\sum(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum(X_i - \overline{X})^2 \sum(Y_i - \overline{Y})^2}}$$

Here, X_i, Y_i and are individual data points, and $\overline{X}$ and $\overline{Y}$ are the means of X and Y, respectively.

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

Ans:
Scaling is the process of normalizing the range of independent variables or features of a dataset. It is performed to ensure that no variable dominates due to its scale and to bring all variables to a standard scale. The two common types of scaling are normalized scaling and standardized scaling.

Normalized Scaling:

Normalization scales the values of a variable between 0 and 1.
The formula is

$$X_{normalized} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Standardized Scaling:

Standardization transforms the data which will have 0 mean and 1 standard deviation.
The formula is

$X_{standardized} = \frac{X - \overline{X}}{\sigma}$, where is $\overline{X}$ the mean and σ is the standard deviation.

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

Ans:
Sometimes, VIF can become infinite when there is perfect multicollinearity in the data. Perfect multicollinearity occurs when one independent variable is a perfect linear function of others.
In such cases, the correlation matrix becomes singular, leading to the inversion problem in the calculation of VIF.

6Q. 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 to assess if a dataset follows a particular theoretical distribution. It compares the quantiles of the observed data against the quantiles of a theoretical distribution (e.g., normal distribution).
If the points on the Q-Q plot lie approximately along a straight line, it suggests that the data follows the assumed distribution.
In linear regression, Q-Q plots help to check the normality of residuals, which is an important assumption. If the residuals follow a normal distribution, the Q-Q plot will be roughly linear. Deviations from linearity indicate departures from normality.