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)

- There is more demand during the fall season as compared to other seasons.
- There was high demand of rental bikes in 2019 as compare to 2018
- There is not much of a significant difference when there is a holiday or not.
- There is not much of a significant difference when it's a working day or not.
- There is more demand when the weather is clear with few clouds as compared to other weather conditions.

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

The dummy variable trap occurs when one or more dummy variables are highly correlated or linearly dependent on each other. This means that one dummy variable can be predicted from the others with high accuracy, leading to multicollinearity in the dataset. Multicollinearity can cause problems like:

- When dummy variables are highly correlated, it becomes difficult to interpret the individual impact of each variable on the target variable. This can lead to misleading conclusions about the relationships between the features and the outcome.
- Multicollinearity can adversely affect the performance of certain machine learning algorithms, especially those that rely on matrix inversion, like linear regression. It can make the coefficient estimates unstable or lead to inflated standard errors.
- including all dummy variables can introduce redundant information, which increases the dimensionality of the dataset without providing additional useful information.

Using drop_first=True during dummy variable creation is essential to prevent multicollinearity, improve model interpretability, and enhance the performance of machine learning models that are sensitive to high-dimensional datasets. It ensures that the dummy variables convey the necessary information about the categorical feature without introducing redundant information or violating the assumptions of the models.

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

Temperature has the highest correlation with fall season, and later it has a high correlation with total count of rental bikes .

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, I have validated the assumptions of linear regression by plotting histogram of error terms and validated below that the residuals should be independent of each other. There should be no clear patterns or trends in the plot and the residuals should follow a normal distribution.

The relationship between the dependent variable and each independent variable should be linear. To validate this assumption, I have plotted the predicted values against the actual values and check if they form a reasonably straight line.

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

Below are the top 3 features that contribute significantly toward explaining the demand of shared bikes.

- Temperature
- Light snowfall
- Winter

<u>General Subjective Questions</u>

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

Linear regression is a fundamental statistical method used to model the relationship between a dependent variable and one or more independent variables. It is a simple but powerful technique for predicting numeric outcomes based on input features.

Linear regression relies on below key assumptions:

- ☐ Linearity: The relationship between the dependent variable and the independent variables should be approximately linear.
- ☐ Independence: The observations should be independent of each other.
- ☐ Homoscedasticity: The variance of the residuals should be constant across all levels of the independent variable.
- ☐ Normality: The residuals should follow a normal distribution.

Simple Linear Regression:
Simple linear regression involves only one independent variable and one dependent variable. The relationship between them is represented by a straight line equation i.e y = mX + b.

y is target variable

X is independent variable
m is the slope
b is y intercept

Multiple Linear Regression:
Multiple linear regression extends the concept to more than one independent variable.
y = b0 + b1X1 + b2X2 + b3X3..

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

Anscombe's quartet is a set of four datasets that was created by the statistician Francis Anscombe in 1973. The datasets have identical statistical properties, but when graphed, they appear very different, illustrating the importance of data visualization in understanding and interpreting data. Anscombe's quartet serves as a cautionary example against relying solely on summary statistics without visualizing the data. The main lesson from Anscombe's quartet is that relying solely on summary statistics (e.g., mean, variance, correlation) can obscure important patterns and outliers in the data.

3. What is Pearson's R? (3 marks)
Pearson's correlation coefficient, often denoted as "r" or "Pearson's r," is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables. It was developed by the British mathematician and statistician Karl Pearson and is widely used in various fields, including statistics, data analysis, and scientific research.

The Pearson correlation coefficient takes on values between -1 and +1:

If r = +1, it indicates a perfect positive linear relationship between the two variables. This means that as one variable increases, the other variable increases proportionally, and all the data points lie exactly on a straight line with a positive slope.

If r = -1, it indicates a perfect negative linear relationship between the two variables. This means that as one variable increases, the other variable decreases proportionally, and all the data points lie exactly on a straight line with a negative slope.

If r is close to 0, it indicates a weak or no linear relationship between the two variables. This means that there is little to no linear pattern in the data points.

The formula for calculating Pearson's correlation coefficient "r" between two variables X and Y, each with n data points, is as follows:

Pearson's correlation coefficient is commonly used to assess the linear association between two variables, to determine the strength and direction of the relationship, and to identify potential outliers that may be affecting the correlation. It is essential to remember that while a high correlation coefficient indicates a strong linear relationship, it does not necessarily imply causation between the two variables; other factors could be influencing their relationship.

.

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

Scaling, in the context of data preprocessing, refers to the process of transforming the features of a dataset to a standardized range. The goal of scaling is to bring all the features to a similar scale so that they are comparable and have equal influence on the analysis or model training. Scaling helps to improve the performance and efficiency of these methods.

Scaling is performed for
Equalizes Influence: Different features in a dataset can have varying units and scales. If these features are not scaled, features with larger scales may dominate or have more significant influence over the analysis or model training compared to features with smaller scales. Scaling ensures that all features contribute more evenly.

Gradient Descent Optimization: Many optimization algorithms used in machine learning converge faster on scaled data. Without scaling, some optimization algorithms may take much longer to find the optimal solution.

Distance-Based Algorithms: Algorithms that use distance measures, like k-nearest neighbors, are sensitive to the scales of the features. Scaling ensures that distance-based algorithms are not biased by features with larger scales.

Regularization: Regularization terms in some machine learning models penalize large coefficients. Scaling helps to balance the influence of different features on the regularization term.

Visualization: Scaling aids in visualizing data when plotting multiple features on the same graph, as all features are within a similar range.

Difference between normalized scaling and standardized scaling is below:

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

The Variance Inflation Factor (VIF) is a measure used in regression analysis to assess multicollinearity between predictor variables. Multicollinearity occurs when two or more predictor variables are highly correlated, which can lead to inflated standard errors and unstable coefficient estimates in the regression model.

There were cases where the VIF was infinite. This happens when there is perfect multicollinearity between one or more predictor variables. Perfect multicollinearity occurs when two or more predictor variables are linearly dependent on each other. In such cases, the VIF becomes undefined or infinite because the R^2 term in the formula approaches 1.

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

A Q-Q plot, short for Quantile-Quantile plot, is a graphical tool used to assess the distributional similarity between a sample of data and a theoretical probability distribution, such as the normal distribution. It is a scatter plot that compares the quantiles of the sample data against the quantiles of the theoretical distribution.

In a Q-Q plot, the x-axis represents the quantiles of the theoretical distribution, and the y-axis represents the quantiles of the sample data. If the data follows the theoretical distribution closely, the points in the Q-Q plot will fall approximately along a straight line. Deviations from a straight line indicate departures from the assumed distribution.

Use and Importance of a Q-Q plot in linear regression:

Checking Normality Assumption: In linear regression, it is often assumed that the residuals (the differences between the observed values and the predicted values) follow a normal distribution. A Q-Q plot is a valuable tool to assess whether this assumption holds. If the residuals approximate a straight line in the Q-Q plot, it suggests that the normality assumption is reasonable. On the other hand, significant deviations from a straight line indicate that the residuals might not be normally distributed, which can impact the validity of the regression results.

Identifying Outliers: Q-Q plots can also help identify outliers in the data. Outliers are data points that deviate significantly from the rest of the data and can have a substantial impact on the regression model. Outliers can manifest as points that deviate from the straight line in the Q-Q plot, indicating that these data points do not conform to the assumed distribution.

Assessing Distributional Assumptions: In addition to normality, linear regression might have other distributional assumptions, depending on the specific regression model used. Q-Q plots

can be employed to verify whether the residuals or other variables adhere to these distributional assumptions. This helps to understand the suitability of the chosen regression model and make informed decisions about data transformations if needed.

In summary, Q-Q plots are valuable diagnostic tools in linear regression analysis. They aid in evaluating the normality assumption, detecting outliers, and assessing distributional assumptions, helping researchers and data analysts ensure the validity and reliability of their regression models.