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)

The analysis indicates that bike rental rates tend to be higher during the summer and fall seasons, particularly in September and October. Additionally, rentals are more frequent on specific days like Saturday, Wednesday, and Thursday, and were notably higher in 2019. The data also suggests that bike rentals increase on holidays.

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

Using drop_first=True eliminates one of the dummy variables for each categorical feature, reducing multicollinearity and avoiding redundancy. This prevents the "dummy variable trap," where having a full set of dummy variables can cause issues in regression models due to perfect correlation.

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

The variable temp shows the strongest 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)

The assumptions of linear regression were validated by examining the Variance Inflation Factor (VIF) to check for multicollinearity, analyzing the distribution of residuals to ensure they follow a normal distribution, and confirming a linear relationship between the dependent variable and each feature variable.

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

The top three features significantly contributing to the demand for shared bikes are temperature, the year, and whether the day is a holiday.

General Subjective Questions

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

Linear Regression is a supervised machine learning algorithm used for predicting a dependent variable, also known as the target, based on one or more independent variables. The primary objective of this regression technique is to establish a linear relationship between the dependent variable and the independent variables.

There are two main types of linear regression:

- **Simple Linear Regression: This type is utilized when a single independent variable is used to predict the target variable. The relationship is represented by a straight line, known as the regression line, which minimizes the difference between the predicted and actual values.**
- **Multiple Linear Regression: This approach involves two or more independent variables to predict the target variable. It establishes a linear relationship while accounting for the influence of multiple predictors.**

The regression line represents the best fit for the data points in the scatter plot of the independent and dependent variables. A positive linear relationship occurs when the dependent variable increases as the independent variable increases, resulting in an upward slope on the graph. Conversely, a negative linear relationship is observed when the dependent variable decreases as the independent variable increases, producing a downward slope.

Key components of linear regression include:

- **Equation of the Line: The relationship can be mathematically expressed as Y equals b zero plus b one times X one plus b two times X two plus ... plus b n times X n plus epsilon, where Y is the predicted value of the dependent variable, b zero is the intercept of the regression line, b one, b two, ..., b n are the coefficients for each independent variable X one, X two, ..., X n, and epsilon represents the error term.**
- **Assumptions: Linear regression relies on several assumptions, including linearity, independence of errors, homoscedasticity or constant variance of errors, and normality of error distribution.**
- **Evaluation Metrics: The performance of a linear regression model can be assessed using various metrics such as R-squared, Mean Squared Error, and Adjusted R-squared, which provide insights into how well the model fits the data.**

Linear Regression is a foundational algorithm in statistics and machine learning, widely used due to its simplicity, interpretability, and effectiveness in various predictive modeling scenarios.

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

Anscombe's quartet consists of four distinct datasets that share nearly identical simple descriptive statistics but exhibit significantly different distributions and graphical representations. Each dataset contains eleven data points. The primary objective of Anscombe's quartet is to emphasize the importance of visually examining data before conducting any analytical processes. Relying solely on statistical measures can be misleading, as they may fail to capture the underlying patterns or relationships within the data. By comparing the four datasets, one can observe how different distributions can lead to the same summary statistics, highlighting the necessity of graphical analysis in data interpretation.

3. What is Pearson's R? (3 marks)

Pearson's Correlation Coefficient, often denoted as Pearson's R, is a statistical measure used to determine the strength and direction of a linear relationship between two variables. The coefficient value ranges from -1 to +1. A value of +1 indicates a perfect positive correlation, meaning that as one variable increases, the other variable also increases proportionally. A value of -1 indicates a perfect negative correlation, meaning that as one variable increases, the other decreases proportionally. A value of 0 indicates no correlation. Pearson's R is widely used in various fields to assess relationships between continuous variables.

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

Scaling is a preprocessing technique used in machine learning to standardize the independent feature variables in a dataset to a fixed range. This is important because datasets often contain features with varying magnitudes and units. Without scaling, the model may give undue importance to features with larger ranges, leading to incorrect modeling and biased results.

The difference between normalization and standardization lies in their approaches. Normalization transforms the data to a range between 0 and 1, making it suitable for algorithms sensitive to the scale of data. In contrast, standardization transforms the data into Z-scores, which indicate how many standard deviations a value is from the mean, resulting in a distribution with a mean of zero and a standard deviation of one.

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

The value of Variance Inflation Factor (VIF) becomes infinite when there is perfect multicollinearity between two independent variables, meaning they are perfectly correlated. In such cases, the R-squared value for the regression of one variable on the other is equal to 1. Since VIF is calculated using the formula VIF equals 1 divided by 1 minus R-squared, an R-squared value of 1 leads to division by zero, resulting in an infinite VIF. This indicates a multicollinearity problem, suggesting that one of the correlated variables should be removed to create a more effective regression model.

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

A quantile-quantile (Q-Q) plot is a graphical tool used to compare the quantiles of a sample distribution against the quantiles of a theoretical distribution, such as a normal, uniform, or exponential distribution. This plot helps to visually assess whether the data follows a specific distribution by plotting the quantiles of the sample on one axis and the quantiles of the theoretical distribution on the other.

In the context of linear regression, Q-Q plots are particularly important for validating the assumption of normality of the residuals. By examining the Q-Q plot of the residuals, we can determine if they follow a normal distribution, which is a key assumption for the validity of many statistical tests associated with linear regression. If the points on the Q-Q plot align closely along a straight line, it suggests that the residuals are normally distributed. Deviations from this line may indicate that the residuals are not normally distributed, potentially impacting the reliability of the regression model's conclusions. Thus, Q-Q plots serve as a valuable diagnostic tool in regression analysis.