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)

From an analysis of categorical variables in the dataset, we can infer the following effects on the dependent variable (bike demand):

Season: Demand may fluctuate based on the season, with higher demand likely during warmer seasons (spring and summer) due to more favorable weather conditions for biking, and lower demand in colder seasons (fall and winter).

Holiday: Bike demand could vary significantly on holidays, potentially increasing as more people are free for leisure activities, or decreasing if fewer people need transportation for work or errands.

Working Day: On working days, demand might be influenced by commuting patterns, with higher bike usage for work-related travel. Non-working days may show different patterns due to recreational use.

Weather Type: Weather conditions (sunny, cloudy, rainy) could significantly impact bike demand. Poor weather, such as rain, may decrease demand, while favorable conditions could boost it.

Weekday vs. Weekend: Bike demand may vary between weekdays and weekends, with higher commuter usage during the week and more recreational use on weekends.

These categorical variables help in understanding the factors influencing bike demand and are essential for building an accurate predictive model.

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

Using drop_first=True during dummy variable creation is important because it helps to:

Avoid Multicollinearity: It removes one of the dummy variables for each categorical feature, preventing the "dummy variable trap." This trap occurs when there is redundancy in the data, leading to multicollinearity, where one variable can be

predicted from others.

Ensure Efficient Modeling: By dropping the first category, we reduce the number of columns in the dataset, improving the model's efficiency without losing any information, as the dropped category can be inferred from the remaining variables.

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

To identify the numerical variable with the highest correlation to the target variable using a pair-plot, you would visually assess the scatter plots between each numerical variable and the target. However, typically this task is better suited for a correlation matrix that quantifies these relationships.

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

Check for a linear relationship between the independent variables and the target variable by plotting the residuals (errors) against the predicted values. If the residuals are randomly scattered around zero, this indicates a linear relationship.

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

Based on the final model, the top 3 features contributing significantly toward explaining the demand for shared bikes are typically:

Temperature (or Normalized Temperature): Warmer temperatures often lead to increased bike usage, making this one of the most significant predictors of demand.

Season: Different seasons (spring, summer) generally have a strong impact on demand, with more usage during favorable weather periods.

Working Day: This feature reflects commuter patterns, with higher demand on working days compared to weekends or holidays.

General Subjective Questions

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

A fundamental statistical method used to model the relationship between a dependent variable (target) and one or more independent variables (features). The goal is to find the best-fitting linear relationship that predicts the target variable based on the

features.

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

Anscombe's quartet is a collection of four datasets, each containing 11 data points, designed by the British statistician Francis Anscombe in 1973. The quartet is used to demonstrate the importance of visualizing data before analyzing it. Despite the four datasets having nearly identical simple statistical properties (e.g., mean, variance, correlation), their underlying distributions are vastly different when plotted, emphasizing that summary statistics alone can be misleading

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

Pearson's R, also known as Pearson's correlation coefficient, is a measure of the linear relationship between two continuous variables. It quantifies the strength and direction of the linear association between the 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 in data analysis and machine learning used to adjust the range and distribution of features in a dataset. The goal is to make sure that features contribute equally to the analysis or model, especially when they are on different scales.

Normalized Scaling (Min-Max Normalization):

Standardized Scaling (Z-score Normalization):

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

(3 marks)

The Variance Inflation Factor (VIF) measures how much the variance of an estimated regression coefficient is increased due to multicollinearity among independent variables. An infinite VIF occurs under specific conditions related to multicollinearity.

Causes of Infinite VIF

Perfect Multicollinearity:

Description: Infinite VIF arises when there is perfect multicollinearity among independent variables. This means that one or more independent variables are exact

linear combinations of other variables.

Example: If x3 = f(x1, x2) then the VIF for x3 becomes infinite.

Redundant Variables:

*Description***:** If variables are redundant or duplicate each other

Example: Including both x and **2**x in a regression model would cause multicollinearity issues leading to infinite VIF.

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

(3 marks)

Definition: A Q-Q plot compares the quantiles of a dataset's distribution against the quantiles of a theoretical distribution (e.g., normal distribution). It plots the quantiles of the observed data on the x-axis and the quantiles of the theoretical distribution on the y-axis.

Interpretation: If the points lie approximately along a straight line: The data follows the theoretical distribution well. If the points deviate significantly from the line: The data does not follow the theoretical distribution.