Assignment-based Subjective Questions

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

Based on the linear regression model and dataset analysis, here is a detailed explanation of how the categorical variables affect the dependent variable (bike demand):

1. **Year (yr)**:
    - **Effect**: Positive
    - **Inference**: The demand for bikes increases with the year. This could be due to growing awareness and acceptance of bike-sharing services over time.
2. **Holiday**:
    - **Effect**: Negative
    - **Inference**: Bike demand decreases on holidays. This might be because people are less likely to commute on holidays and may prefer other leisure activities instead.
3. **Season (Spring)**:
    - **Effect**: Negative
    - **Inference**: There is a decrease in bike demand during the spring season. This could be due to weather conditions that are not ideal for biking, such as rain or allergies.
4. **Weather Conditions**:
    - **Mist/Cloudy**:
        - **Effect**: Negative
        - **Inference**: Misty or cloudy weather reduces bike demand. Poor visibility and potentially uncomfortable conditions deter people from using bikes.
    - **Light Rain/Light Snow/Thunderstorm**:
        - **Effect**: Significant Negative
        - **Inference**: Adverse weather conditions like light rain, light snow, or thunderstorms significantly reduce bike demand. These conditions make biking difficult and unsafe, leading to a drop in usage.
5. **Specific Months**:
    - **March (3)**:
        - **Effect**: Positive
    - **May (5)**:
        - **Effect**: Positive
    - **June (6)**:
        - **Effect**: Positive
    - **July (7)**:
        - **Effect**: Positive

- **August (8)**:
  - **Effect**: Positive
- **September (9)**:
  - **Effect**: Positive
- **October (10)**:
  - **Effect**: Positive
- **Inference**: Demand increases in these specific months. This trend could be due to favorable weather conditions, vacation periods, and outdoor activity trends during these months.

6. **Day of the Week (Sunday)**:
   - **Effect**: Negative
   - **Inference**: Bike demand decreases on Sundays. This may be because fewer people commute to work or school on Sundays, and they may engage in other leisure activities instead.

Overall Inferences:

- **Positive Effects**: Year, specific months (March, May, June, July, August, September, October) suggest a higher demand for bikes.
- **Negative Effects**: Holidays, spring season, adverse weather conditions (mist/cloudy, light rain/light snow/thunderstorm), and Sundays lead to a decrease in bike demand.

Understanding these effects helps BoomBikes strategize their operations, marketing, and bike availability to maximize demand and improve customer satisfaction.

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

When creating dummy variables for categorical data, it is important to use **drop_first=True** to avoid the problem of multicollinearity

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

atemp has the highest correlation. (feeling temperature in Celsius)

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

We used VIF, Variance Inflation Factor (VIF), to check if independent variables are highly correlated with each other. Then we used the prediction of the test data on the validation data.

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

yr , holiday ,Spring are the top 3 features

General Subjective Questions

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 achieves this by fitting a linear equation to the observed data, represented as $y=\beta 0+\beta 1x1+\beta 2x2+\ldots+\beta nxn+\epsilon y=\beta 0+\beta 1x1+\beta 2$ $x2+\ldots+\beta nxn+\epsilon$, where yy is the dependent variable, $x1,x2,\ldots,xnx1,x2,\ldots,xn$ are the independent variables, $\beta 0\beta 0$ is the intercept, $\beta 1,\beta 2,\ldots,\beta n\beta 1,\beta 2,\ldots,\beta n$ are the coefficients, and $\epsilon\epsilon$ is the error term. The objective is to minimize the sum of squared residuals (differences between observed and predicted values) to find the best-fitting line. This is typically done using the Ordinary Least Squares (OLS) method. The model's performance is assessed through metrics such as R-squared and Mean Squared Error (MSE), ensuring the validity and reliability of predictions

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

Anscombe's Quartet is a collection of four datasets that have nearly identical simple descriptive statistics—such as means, variances, and correlation coefficients—but reveal strikingly different distributions and relationships when graphed. Created by statistician Francis Anscombe in 1973, the quartet underscores the importance of visualizing data before performing statistical analysis. While all four datasets share the same linear regression line and summary statistics, their plots display a variety of relationships: a simple linear trend, a clear non-linear pattern, an influential outlier, and a vertical stack with a single outlier. This illustrates that relying solely on summary statistics can be misleading and emphasizes the need for graphical exploration to understand data comprehensively.

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

Pearson's R, also known as the Pearson correlation coefficient, is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables, ranging from -1 to +1. A value of +1 indicates a perfect positive linear correlation, -1 indicates a perfect negative linear correlation, and 0 indicates no linear correlation. It is calculated by dividing the covariance of the two variables by the product of their standard deviations, thus standardizing the measure. Pearson's R is widely used in various fields to assess the degree of association between variables, though it only captures linear relationships and can be sensitive to outliers.

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

What is Scaling?

Scaling is the process of transforming the features of a dataset so that they fall within a similar range. This transformation is essential because many machine learning algorithms are sensitive to the scale of input data, meaning they perform better or converge faster when the data is scaled appropriately. By bringing all the features to a comparable level, scaling ensures that no single feature dominates the model due to its magnitude.

Why is Scaling Performed?

Scaling is performed for several reasons. Firstly, it improves the performance and accuracy of machine learning algorithms that rely on distance metrics, such as K-Nearest Neighbors and Support Vector Machines. These algorithms can be heavily influenced by the scale of the

input features, and scaling helps mitigate this effect. Secondly, scaling can speed up the convergence of gradient-based optimization algorithms like those used in neural networks and logistic regression. When features are on a similar scale, these algorithms can find the optimal solution more efficiently. Lastly, scaling enhances the interpretability of the model by ensuring that the importance assigned to each feature is not skewed by their varying magnitudes.

Difference Between Normalized Scaling and Standardized Scaling

Normalized Scaling (Min-Max Scaling): Normalized scaling adjusts the data to fit within a specific range, typically between 0 and 1. This method is useful when the distribution of the data is not Gaussian and when the algorithm requires features to be bounded within a fixed range. Normalization is especially common in image processing and neural networks, where pixel values or activation inputs need to be within a consistent range.

Standardized Scaling (Z-score Scaling): Standardized scaling, on the other hand, transforms the data so that it has a mean of zero and a standard deviation of one. This method is beneficial when the data follows a normal distribution or when the algorithm assumes normally distributed data, such as in linear regression, logistic regression, and SVMs. Standardization helps in centering the data and providing unit variance, making it easier to compare different features on the same scale without being affected by outliers as much as normalization.

In summary, while both normalized and standardized scaling aim to bring features onto a similar scale, they do so in different ways suited to different types of data and algorithms. Normalization confines data within a specified range, making it suitable for bounded input scenarios, while standardization standardizes the data based on its statistical properties, catering to algorithms that benefit from normally distributed data.

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

The value of the Variance Inflation Factor (VIF) becomes infinite when there is perfect multicollinearity among the independent variables in a regression model. This means that one predictor variable can be perfectly predicted using a linear combination of the other predictor variables, resulting in an exact correlation of 1 (or -1). In such cases, the denominator in the VIF calculation, which involves the determinant of the correlation matrix, becomes zero, causing the VIF to approach infinity. Perfect multicollinearity indicates redundant information in the model, making it impossible to isolate the individual effect of each predictor variable.

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

A Q-Q (Quantile-Quantile) plot is a graphical tool used in linear regression to assess if the residuals follow a normal distribution, a key assumption for valid statistical tests and confidence intervals. By plotting the quantiles of the residuals against the quantiles of a normal distribution, we can visually inspect if they lie along a straight line, indicating normality. Deviations from this line suggest issues like skewness or outliers, which can affect model performance. Thus, the Q-Q plot helps validate the normality assumption, identify data issues, and inform potential data transformations to improve model accuracy.