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 categorical variables in the dataset, such as season, yr, mnth, holiday, weekday, workingday, and weathersit, can significantly impact the dependent variable cnt (total bike rentals).

- **Season and Month:** These can influence rental patterns due to weather conditions, with peaks in summer and spring.
- **Year (yr):** Rentals may increase over years due to growing popularity.
- **Holiday and Workingday:** Rentals likely decrease on holidays and weekends compared to working days.
- **Weekday:** Variations may exist, with weekends showing different trends.
- **Weathersit:** Bad weather (rain/snow) usually leads to fewer rentals.

Analyzing these variables can reveal patterns and help predict bike rental trends.

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

Using drop_first=True in dummy variable creation prevents multicollinearity, which occurs when one dummy variable is linearly dependent on others. By dropping the first category, we avoid redundancy and ensure that each category is compared to a baseline (the dropped category). This makes the design matrix full rank, allowing for unique and stable parameter estimates in models like linear regression. It simplifies interpretation by showing the effect of each category relative to the baseline. Thus, drop_first=True is crucial for maintaining model accuracy and interpretability when dealing with categorical variables.

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

From the pair plot, temp (temperature) appears to have the highest correlation with the target variable cnt (total bike rentals). The scatter plot between temp and cnt shows a clear positive linear trend, suggesting that higher temperatures are associated with increased bike rentals. Similarly, atemp (feels-like temperature) also shows a strong positive correlation with cnt, nearly mirroring temp. In contrast, other variables like hum (humidity) and windspeed exhibit weaker correlations, with hum showing a slight negative relationship and windspeed showing more scattered points without a clear linear trend. Thus, temp is likely the most influential.

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

To validate the assumptions of Linear Regression after building the model on the training set, we typically follow these steps:

1. Linearity: Check if the relationship between predictors and the target variable is linear. This is often done using scatter plots or residual plots. The sns.regplot(y_train, y_pred) can help visualize if the relationship is approximately linear.
2. Homoscedasticity: Ensure that the residuals (errors) have constant variance. Plotting residuals against predicted values using sns.regplot(y_pred, residual) can help check for patterns. Residuals should be randomly scattered without any specific pattern.
3. Independence of Errors: Confirm that the residuals are independent of each other. This can be assessed using residual plots or by checking autocorrelation functions.
4. Normality of Residuals: Validate that residuals are normally distributed. This is often assessed using a Q-Q plot or a histogram of residuals. If residuals roughly follow a straight line in the Q-Q plot, the normality assumption is likely satisfied.

By evaluating these aspects, we can determine if our linear regression model meets the required assumptions and is likely to provide reliable predictions.

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

   - **Examine Coefficients**: Review the coefficients from your linear regression model. Features with the largest absolute coefficients are most influential.
   - **Check Feature Importance**: If using models like Decision Trees or Random Forests, use feature importance scores to identify key features.
   - **Consider Statistical Significance**: Look at p-values for each feature. Features with p-values below 0.05 are statistically significant.
   - **Identify Top Features**: Focus on features with the highest absolute coefficients or greatest importance scores to determine the top 3 features influencing bike demand.
     Correlation,P-Value,VIF

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**

   - It is one of the most-used regression algorithms in Machine Learning. A significant variable from the data set is chosen to predict the output variables (future values). Linear regression algorithm is used if the labels are continuous, like the number of flights daily from an airport, etc. The representation of linear regression is y = b*x + c.
   - In the above representation, 'y' is the independent variable, whereas 'x' is the dependent variable. When you plot the linear regression, then the slope of the line that provides us the output variables is termed 'b', and 'c' is its intercept. The linear regression algorithms assume that there is a linear relationship between the input and the output. If the dependent and independent variables are not plotted on the same line in linear regression, then there will be a loss in output. The loss in output in linear regression can be calculated as:
   - Loss function: $(\text{Predicted output} - \text{actual output})^2$.

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

   Anscombe's quartet comprises a set of four datasets, having identical descriptive statistical properties in terms of means, variance, R-squared, correlations, and linear regression lines but having different representations when we scatter plots on a graph.The datasets were created by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data and to show that summary statistics alone can be misleading.The four datasets that make up Anscombe's quartet each include 11 x-y pairs of data. When plotted, each dataset seems to have a unique connection between x and y, with unique variability patterns and distinctive correlation strengths. Despite these variations, each dataset has the same summary statistics, such as the same x and y mean and variance, x and y correlation coefficient, and linear regression line.

   Purpose of Anscombe's Quartet

   Anscombe's quartet is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statistics. It also emphasizes the importance of using data visualization to spot trends, outliers, and other crucial details that might not be obvious from summary statistics alone.

3. **What is Pearson's R?**

   The Pearson correlation measures the strength of the linear relationship between two variables. It has a value between -1 to 1, with a value of -1 meaning a total negative linear correlation, 0 being no correlation, and + 1 meaning a total positive correlation.

Pearson's product moment correlation coefficient (r) is given as a measure of linear association between the two variables: r^2 is the proportion of the total variance (s^2) of Y that can be explained by the linear regression of Y on x.
The limitations of the Pearson correlation coefficient include its assumption of linearity, bivariate normal distribution, and the presence of outliers and restricted range of data.

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

Scaling:

- Scaling is a geometric change that linearly enlarges or reduces things. A property of objects or rules known as scale invariance is that they remain unchanged when scales of length, energy, or other variables are multiplied by a common factor.
- Scaling law, a law that explains how many natural phenomena exhibit scale invariance.

scaling performed because:

It is a data pre-processing procedure used to normalize data within a specific range by applying it to independent variables. Additionally, it aids in accelerating algorithmic calculations. The majority of the time, the obtained data set includes characteristics that vary greatly in magnitudes, units, and range.

the difference between normalized scaling and standardized scaling

The values of a normalized dataset will always fall between 0 and 1. A standardized dataset will have a mean of 0 and a standard deviation of 1, but the maximum and minimum values are not constrained by any specified upper or lower bounds.

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

A large VIF on an independent variable indicates a highly collinear relationship to the other variables that should be considered or adjusted for in the structure of the model and selection of independent variables.

The variance inflation factor (VIF) quantifies the extent of correlation between one predictor and the other predictors in a model. It is used for diagnosing collinearity/multicollinearity. Higher values signify that it is difficult to impossible to assess accurately the contribution of predictors to a model.

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

Quantile-Quantile (**Q-Q**) **plot**, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution.

Q–Q plots are commonly used to compare a data set to a theoretical model. This can provide an assessment of goodness of fit that is graphical, rather than reducing to a numerical summary statistic. Q–Q plots are also used to compare two theoretical distributions to each other.

A variance inflation factor (VIF) is a measure of the amount of multicollinearity in regression analysis. Multicollinearity exists when there is a correlation between multiple independent variables in a multiple regression model. This can adversely affect the regression results. Thus, the variance inflation factor can estimate how much the variance of a regression coefficient is inflated due to multicollinearity.

Key Takeaways

- A variance inflation factor (VIF) provides a measure of multicollinearity among the independent variables in a multiple regression model.

- Detecting multicollinearity is important because while multicollinearity does not reduce the explanatory power of the model, it does reduce the statistical significance of the independent variables.
- A large VIF on an independent variable indicates a highly collinear relationship to the other variables that should be considered or adjusted for in the structure of the model and selection of independent variables.