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?

Answer: Both "season_spring" and "season_winter" are significant predictors based on their low p-values ($p < 0.05$).

The coefficients for these variables indicate that during spring, there's a decrease in the count compared to other season, while during winter, there's an increase in the count.

Both "mnth_jul" and "mnth_sept" are significant predictors ($p < 0.05$).

"mnth_jul" has a negative coefficient, suggesting a decrease in count compared to the other months.

"mnth_sept" has a positive coefficient, indicating an increase in count compared to the other month.

"weekday_sun" is a significant predictor ($p < 0.05$).

It has a negative coefficient, indicating a decrease in count compared to other weekdays.

Both "weathersit_bad" and "weathersit_moderate" are significant predictors ($p < 0.05$).

"weathersit_bad" has a more negative coefficient compared to "weathersit_moderate," indicating a more significant decrease in count during bad weather compared to moderate weather.

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

Answer: When creating dummy variables it is really important to avoid multicollinearity in the dataset. Multicollinearity occurs when one or more predictor variables in a regression model are highly correlated. So to avoid multicollinearity we use drop_first=True.

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

Answer: According to pair plots temp and atemp are highly correlated with target variable (cnt)

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

Answer:

i. No perfect multicollinearity: The independent variables are not perfectly correlated i.e VIF<10

ii. The residuals are normally distributed

iii. Plot the actual values vs predicted values

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

The top three significant features explaining bike demand are temperature (positive impact), year (positive trend), and bad weather conditions (negative impact).

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Answer: Linear regression is a supervised learning algorithm used for predictive analysis. It's a statistical method that models the relationship between a dependent variable (y) and one or more independent variables (x). The basic idea behind linear regression is to find the best-fitting linear relationship between the independent variables and the dependent variable.

The equation for simple linear regression is:

y=mx+c

where

y is the dependent variable.

x is the independent variable.

m is the slope of the line (the coefficient).

c is the intercept (the constant term).

Linear regression aims to minimize the difference between the actual values of the dependent variable and the values predicted by the linear model.

2. Explain the Anscombe's quartet in detail.

Anscombe's quartet comprises four datasets with nearly identical statistical properties but vastly different graphical representations. Created by statistician Francis Anscombe in 1973, it highlights the importance of visual exploration in data analysis. Despite identical summary statistics, the datasets exhibit diverse patterns, showcasing the limitations of relying solely on numerical summaries. This quartet emphasizes the necessity of data visualization for understanding the underlying structure of data and avoiding misinterpretation. It serves as a powerful reminder of the potential pitfalls of overlooking visual exploration in favor of numerical analysis when drawing conclusions from datasets.

3. What is Pearson's R?

Pearson's correlation coefficient, often denoted as

r, measures the strength and direction of the linear relationship between two continuous variables. It ranges from -1 to 1, where:

r=1 indicates a perfect positive linear relationship,

r=−1 indicates a perfect negative linear relationship,

r=0 indicates no linear relationship.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Scaling is the process of transforming data to a standardized range, typically between 0 and 1 or with a mean of 0 and a standard deviation of 1. It's performed to ensure that variables with different units or scales contribute equally to analyses like machine learning algorithms. Normalized scaling rescales the data to a range of 0 to 1, while standardized scaling transforms the data to have a mean of 0 and a standard deviation of 1. Normalization maintains the distribution of the data, while standardization centers the data around the mean, making it easier to compare across variables.

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

   The occurrence of infinite values for Variance Inflation Factor (VIF) typically arises when one or more independent variables are perfectly collinear, meaning they are perfectly linearly dependent on each other. This perfect multicollinearity results in an exact linear relationship between the independent variables, making the estimation of their coefficients impossible. Consequently, the VIF calculation involves dividing by zero, leading to infinite values. Perfect multicollinearity poses a serious problem in regression analysis as it destabilizes parameter estimation and inflates the standard errors, complicating the interpretation and reliability of the model.

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 to assess the normality of data by comparing the distribution of observed data to an expected

distribution, typically the normal distribution. In linear regression, Q-Q plots are vital for evaluating the assumption of normality of residuals. By plotting the quantiles of residuals against theoretical quantiles of a normal distribution, deviations from a straight line indicate departures from normality. Detecting non-normality in residuals is crucial, as it impacts the validity of statistical inference and the reliability of predictions. Thus, Q-Q plots serve as essential tools for ensuring the robustness of linear regression analysis.