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?

- Season: Bike demand is highest in summer and fall, lowest in winter.
- Year: There's a clear increase in bike demand from 2018 to 2019, indicating a growing trend.
- Month: Demand peaks in the middle months (summer) and is lowest in winter months.
- Weather Situation: Clear weather has the highest demand, while harsh weather significantly reduces demand.
- Holidays and weekends seem to have a slight negative effect on bike demand, possibly due to reduced commuter usage.

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

Using drop_first=True is important to avoid the dummy variable trap, which leads to perfect multicollinearity. By dropping one category, we prevent redundancy in the model and avoid singularity issues in matrix operations.

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

Based on the pairplot, temperature ('temp') shows the highest positive correlation with the target variable ('cnt').

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

- Linearity: We plotted actual vs. predicted values to check for a linear relationship.
- Homoscedasticity: We created a residual plot to check for constant variance of residuals.
- Normality of residuals: We used a Q-Q plot to check if residuals follow a normal distribution.
- Multicollinearity: We calculated VIF (Variance Inflation Factor) for each predictor variable.

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

- Year ('yr'): Indicating a strong positive trend over time.
- Temperature ('temp'): Higher temperatures are associated with increased bike demand.
- Fall season ('season_3'): The fall season shows a significant positive impact on bike demand.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression is a statistical method used to model the relationship between a dependent variable (Y) and one or more independent variables (X). The algorithm aims to find the best-fitting linear equation that minimizes the difference between predicted and actual values.

Key components of linear regression: a) Linear equation: $Y = \beta_0 + \beta_1X_1 + \beta_2X_2 + ... + \beta_nX_n + \varepsilon$ where β_0 is the y-intercept, β_1, β_2, ..., β_n are coefficients, and ε is the error term.

b) Ordinary Least Squares (OLS): This method minimizes the sum of squared residuals (differences between observed and predicted values).

c) Assumptions:

- Linearity: The relationship between X and Y is linear.
- Independence: Observations are independent of each other.
- Homoscedasticity: Constant variance of residuals.
- Normality: Residuals are normally distributed.
- No multicollinearity: Independent variables are not highly correlated.

d) Model evaluation: Using metrics like R-squared, adjusted R-squared, F-statistic, and p-values.

2. Explain the Anscombe's quartet in detail.

Anscombe's quartet is a set of four datasets created by statistician Francis Anscombe in 1973 to demonstrate the importance of graphing data before analyzing it and the effect of outliers on statistical properties.

Key points: a) The quartet consists of four datasets, each with 11 (x,y) pairs. b) All four datasets have nearly identical simple statistical properties:

- Same mean of x and y
- Same variance of x and y
- Same correlation between x and y
- Same linear regression line c) However, when plotted, the datasets look very different:
- Dataset 1: Shows a typical linear relationship
- Dataset 2: Shows a clear non-linear relationship
- Dataset 3: Shows a linear relationship with one outlier
- Dataset 4: Shows a case where one outlier determines the regression line

The quartet illustrates that:

- Relying solely on summary statistics can be misleading
- Visualization is crucial in data analysis
- The importance of checking assumptions in regression analysis

3. What is Pearson's R?

Pearson's R, also known as the Pearson correlation coefficient, is a measure of the linear correlation between two variables. It ranges from -1 to +1, where:

- +1 indicates a perfect positive linear correlation
- 0 indicates no linear correlation
- -1 indicates a perfect negative linear correlation

Key points: a) Formula: $R = Cov(X,Y) / (\sigma_x * \sigma_Y)$ where Cov(X,Y) is the covariance of X and Y, and σ_x and σ_Y are standard deviations. b) It measures the strength and direction of the linear relationship between two variables. c) It's sensitive to outliers and assumes a linear relationship. d) It doesn't imply causation, only correlation.

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 fit within a specific scale, like 0 to 1 or -1 to 1.

Why scaling is performed:

- To bring different features to a common scale
- To prevent features with larger magnitudes from dominating the model
- To improve the convergence of gradient descent algorithms
- To make feature coefficients comparable

Difference between normalized scaling and standardized scaling:

a) Normalized scaling (Min-Max scaling):

- Scales features to a fixed range, typically 0 to 1
- Formula: X_norm = (X - X_min) / (X_max - X_min)
- Preserves zero values and doesn't center the data

b) Standardized scaling (Z-score normalization):

- Transforms data to have a mean of 0 and standard deviation of 1
- Formula: $X_{std} = (X - \mu) / \sigma$
- Centers the data around zero and handles outliers better

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

Variance Inflation Factor (VIF) measures how much the variance of an estimated regression coefficient increases if your predictors are correlated. An infinite VIF occurs due to perfect multicollinearity.

Reasons for infinite VIF: a) Perfect linear relationship: One predictor is an exact linear combination of others. b) Dummy variable trap: Including all dummy variables for a categorical feature. c) Redundant features: Two or more features contain the same information. d) Sample size: Very small sample size compared to the number of predictors.

Consequences:

- Makes coefficient estimates unstable and unreliable
- Leads to singular matrix in regression calculations
- Indicates that the model is overspecified

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 if a dataset follows a particular theoretical distribution, most commonly the normal distribution.

Use in linear regression:

- To check the normality assumption of residuals
- Plots the quantiles of the residuals against the quantiles of a theoretical normal distribution

Importance: a) Assumption checking: Helps verify if residuals are normally distributed, a key assumption in linear regression. b) Outlier detection: Deviations at the tails can indicate outliers. c) Distribution shape: Shows skewness or heavy tails in the residual distribution. d) Model adequacy: Significant deviations from the diagonal line suggest the model may not be appropriate for the data.

Interpretation:

- If points roughly fall on a straight line, it suggests normality.
- Systematic deviations indicate non-normality and potential issues with the model.

By using Q-Q plots, analysts can visually assess the validity of their linear regression models and identify potential areas for improvement or further investigation.