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 the analysis of the categorical variables:
     - Season: Different seasons show varying levels of bike demand, with summer and fall typically having higher demand due to favorable weather.
     - Year: The variable yr indicates an increase in bike demand from 2018 to 2019, suggesting growth in popularity or market expansion.
     - Month: Certain months, such as July and September, show significant increases in bike demand, likely due to vacation periods and favorable weather.
     - Weather Situation: weathersit indicates that bad weather significantly reduces bike demand, as expected.
     - Day of the Week: Weekdays, particularly weekends, show variations in demand, with Sundays generally having lower demand compared to weekdays.

2. Why is it important to use drop_first=True during dummy variable creation? (2 marks)
   - Using drop_first=True during dummy variable creation is important because:
   - It prevents multicollinearity, which occurs when dummy variables are highly correlated. By dropping the first category, we avoid the dummy variable trap, ensuring that the model matrix is full rank and invertible.
   - It simplifies the model, reducing the number of predictor variables and improving computational efficiency without losing information.
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)
   - Based on the pair-plot, the variable temp (temperature) has the highest correlation with the target variable cnt (bike demand). This suggests that temperature plays a significant role in influencing bike demand.

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)
   - To validate the assumptions of Linear Regression:
     - **Linearity**: Checked scatter plots of residuals vs. predicted values to ensure there is no clear pattern.
     - **Independence**: Ensured data collection processes are independent.
     - **Homoscedasticity**: Used Breusch-Pagan test and visually inspected residual plots to check for constant variance.
     - **Normality**: Used Q-Q plots of residuals to verify that they are approximately normally distributed.
     - **Multicollinearity**: Checked VIF values to ensure no predictors have high multicollinearity.
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:

- **Temperature (**temp**)**: Highest coefficient, indicating significant influence on bike demand.
- **Year (**yr**)**: Indicates the overall growth in bike demand from 2018 to 2019.
- **Weather Situation (**weathersit_moderate**)**: Shows a significant impact, with moderate weather conditions influencing demand.

General Subjective Questions

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

- Linear regression is a statistical method for modeling the relationship between a dependent variable and one or more independent variables. It aims to fit a linear equation to observed data.
- Steps:
- **Formulate the Model**: The model is formulated as $y=\beta 0+\beta 1X1+\beta 2X2+\cdots+\beta nXn+\epsilon y=\beta 0+\beta 1X1+\beta 2X2+\cdots+\beta nXn+\epsilon$, where yy is the dependent variable, XiXi are independent variables, βiβi are coefficients, and ϵϵ is the error term.
- **Estimate Coefficients**: Use the least squares method to estimate the coefficients by minimizing the sum of squared residuals (differences between observed and predicted values).
- **Assess the Model**: Evaluate the fit of the model using metrics like R-squared, adjusted R-squared, F-statistic, and p-values for coefficients.
- **Validate Assumptions**: Check for linearity, independence, homoscedasticity, normality, and multicollinearity.

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

- Anscombe's quartet consists of four datasets that have nearly identical simple descriptive statistics but differ significantly when graphed. It illustrates the importance of graphing data before analyzing it and the limitations of relying solely on summary statistics.
- Key Points:
- Each dataset has the same mean, variance, correlation coefficient, and regression line.
- Visual inspection reveals different patterns: linear relationships, outliers, and non-linear relationships.
- Highlights the need for thorough exploratory data analysis (EDA) and visualization.

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

- Pearson's R, or the Pearson correlation coefficient, measures the linear relationship between two variables. It ranges from -1 to 1:
- **1**: Perfect positive linear relationship.
- **-1**: Perfect negative linear relationship.
- **0**: No linear relationship. It is calculated as the covariance of the variables divided by the product of their standard deviations, and is useful in identifying the strength and direction of linear relationships.

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

o Scaling transforms features to a common scale without distorting differences in the ranges of values. It improves the performance and convergence of machine learning algorithms.

o **Standardization (Z-score normalization)**: Transforms data to have zero mean and unit variance. Suitable for algorithms that assume Gaussian distribution, like linear regression.

o **Normalization (Min-Max scaling)**: Rescales data to a fixed range, typically [0, 1]. Useful for algorithms like neural networks and distance-based algorithms.

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

o An infinite VIF value occurs when there is perfect multicollinearity, meaning one predictor variable is an exact linear combination of one or more other predictor variables. This results in a singular matrix that cannot be inverted, causing VIF calculations to blow up. It indicates redundancy in predictors that must be addressed by removing or combining variables.

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

o A Q-Q (quantile-quantile) plot is a graphical tool to assess if a dataset follows a specified distribution, typically the normal distribution. It plots the quantiles of the data against the quantiles of the reference distribution.

o In linear regression:

o Used to check the normality of residuals.

o Points following a straight line indicate normality.

o Deviations from the line suggest departures from normality, which can affect model assumptions and inference.

o Q-Q plots are crucial for validating the assumption of normality, ensuring the reliability of statistical tests and confidence intervals in regression analysis.