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?

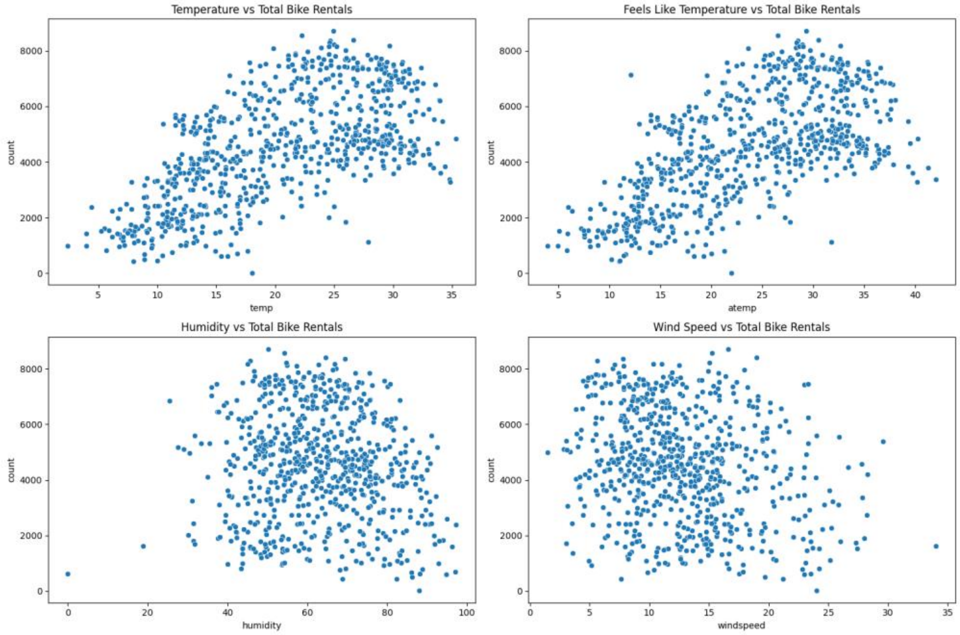
The categorical variables in our model include 'season', 'year', 'month', 'holiday', 'weekday', 'workingday', and 'weathersit'. Here's an interpretation of each:

- **season**: The coefficient for the 'season' variable is positive, indicating that as the season code increases (from spring to winter), the bike demand tends to increase. This suggests that certain seasons, like summer and fall, might have higher bike demands.
- **year**: The coefficient for 'yr' is positive and significant, suggesting that the demand for bikes is increasing year-over-year. This aligns with the growing popularity of bike-sharing systems.
- **month**: The coefficient for 'month' is negative, indicating that there might be a decrease in bike demand as the months progress within a year. This could be due to specific weather conditions or other factors affecting demand in later months.
- **holiday**: The negative coefficient for 'holiday' implies that bike demand might be lower on holidays compared to regular days. This could be due to reduced commuting needs on holidays.
- **weekday**: The positive coefficient for 'weekday' suggests that bike demand might increase during the weekdays. This aligns with the idea of people using bikes for commuting to work or school.
- **workingday**: The positive coefficient for 'workingday' indicates that bike demand might be higher on working days compared to weekends or holidays. However, this variable is borderline significant in the model, so the interpretation should be taken with caution.
- **weathersit**: The negative coefficient for 'weathersit' implies that as the weather situation worsens (from clear to heavy rain/snow), the bike demand decreases. This is logical, as adverse weather conditions might deter people from using bikes.

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

Using **drop_first=True** during dummy variable creation is important to address the "dummy variable trap," which refers to the multicollinearity introduced by dummy variables. Once you have values for n-1 dummy variables, the value of the nth can be inferred.

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



Temperature vs Total Bike Rentals

Feels Like Temperature vs Total Bike Rentals

Humidity vs Total Bike Rentals

Wind Speed vs Total Bike Rentals

Based on the above plot we can see that temp has the highest correlation with target variable.

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

- The relationship between the independent variables and the dependent variable should be linear.
- The residuals (errors) should be independent.
- The variance of the residuals should remain constant across different levels of the independent variables.
- Linear regression assumes that predictors are not highly correlated with each other.
- The residuals should be approximately normally distributed.

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

The top 3 factors contributing towards explaining the demand of the shared bikes are: temperature, climate and holiday.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear Regression is a supervised learning algorithm that models and analyzes the relationships between a dependent variable and one or more independent variables. The main goal is to find the best fit straight line that accurately predict the output values within a range.

Key Points:

- **Equation**:
  - Simple Linear Regression: y=β0+β1x+e
  - Multiple Linear Regression: y=β0+β1x1+β2x2+…..+βnXn+e
- **Objective**: Minimize the sum of squared differences (residuals) between the observed values and the values predicted by the model.
- **Technique**: Uses the "Least Squares Method" to find the best-fitting line.
- **Assumptions**: Assumes a linear relationship between variables, independent observations, constant variance of residuals (homoscedasticity), and no perfect multicollinearity among predictors.

2. Explain the Anscombe's quartet in detail

Anscombe's Quartet consists of four datasets that have nearly identical simple statistical properties (like mean, variance, and correlation), yet appear very different when graphed. Each dataset consists of eleven (x,y) points.

Key Insights from Anscombe's Quartet:

- **Statistics Similarity**: All four datasets have almost the same mean, variance for x and y, correlation between y and y, and linear regression line (slope and intercept).
- **Visual Dissimilarity**: When plotted, each dataset looks distinctly different. One appears linear, another like a curve, the third as a tight linear cluster with an outlier, and the last has x values mostly the same except for one outlier.

Anscombe's Quartet serves as a powerful reminder to always visually inspect your data, not just rely on summary statistics.

3. What is Pearson's R

Pearson's R, also known as the Pearson correlation coefficient or simply Pearson's correlation, measures the strength and direction of the linear relationship between two continuous variables.

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 into a specific range or scale, ensuring that certain features do not affect the final outcome disproportionately due to their larger or smaller numerical values

- **Normalized Scaling (Min-Max Scaling)**: Transforms features by scaling them in the range [0, 1].
- **Standardized Scaling (Z-score Normalization)**: Transforms features to have mean = 0 and variance = 1.

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

The VIF becomes infinite when R-square is exactly 1, which happens when:

- **Perfect multicollinearity**: One independent variable is a perfect linear combination of one or more other independent variables. In other words, one variable can be exactly predicted from the others.
- **Duplicate Variables**: If you mistakenly include the same variable twice in a regression, it's a special case of perfect multicollinearity.
- **Derived Variables**: One variable is a linear transformation of another, e.g., using both temperature in Celsius and Fahrenheit in the same regression.

5. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
Q-Q plot is a powerful diagnostic tool in linear regression (and other statistical methods) to visually assess the normality of residuals, identify outliers, and detect potential issues in the model's underlying assumptions.

Use and Importance of a Q-Q Plot in Linear Regression:

- **Normality of Residuals**:
  - In linear regression, one of the key assumptions is that the residuals (or errors) are normally distributed.
  - A Q-Q plot is used to visually assess this assumption by comparing the quantiles of the residuals to the quantiles of a standard normal distribution.
  - If the residuals are approximately normally distributed, the points on the Q-Q plot will lie close to the straight line.
- **Identifying Outliers**:
  - Deviations from the straight line at the ends of a Q-Q plot may indicate outliers in the data or heavy-tailed distributions.
- **Assessing Linearity**:
  - If certain non-linear patterns are visible in the Q-Q plot, it might suggest that the relationship between variables is not strictly linear, indicating potential transformations that might be necessary.
- **Homoscedasticity Check**:
  - While the primary tool for checking homoscedasticity (constant variance of residuals) is a residual plot, patterns in a Q-Q plot can also hint at issues with non-constant variance.

Importance:

- **Model Validity**:
  - Checking assumptions, such as the normality of residuals, is crucial to ensure the validity of the linear regression model. Violations can lead to biased or inefficient parameter estimates.
- **Informative Diagnostics**:
  - The Q-Q plot provides a visual diagnostic, which can be more intuitive and revealing than numerical tests for normality.
- **Versatility**:
  - Beyond linear regression, Q-Q plots are useful in various statistical modeling contexts to assess the fit of data to a given distribution.