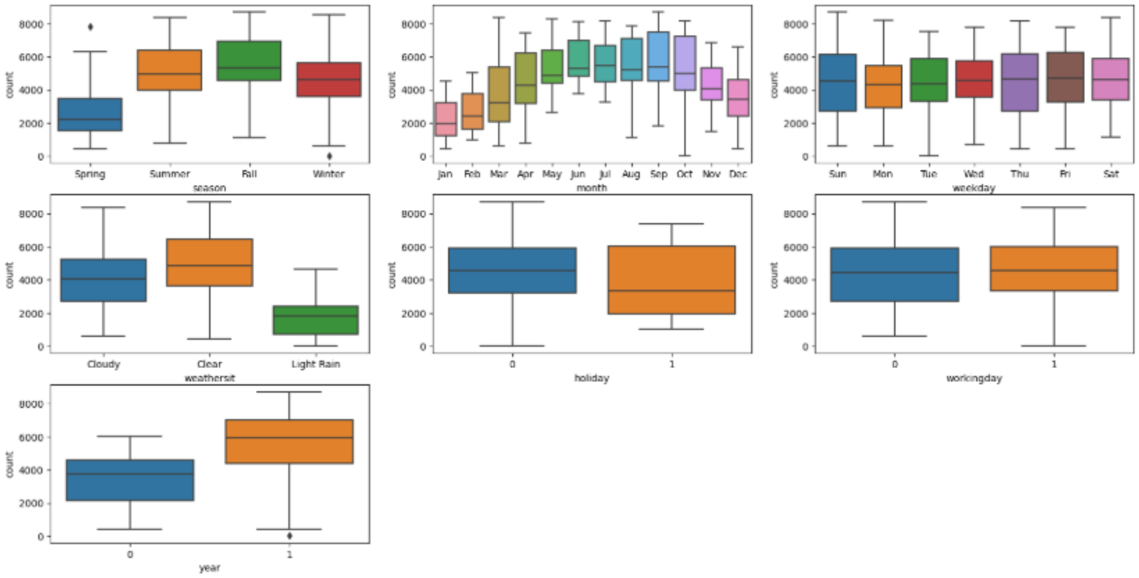
Assignment-based Subjective Questions

Question 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?   (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: <Your answer for Question 1 goes below this line> (Do not edit)

Please refer the below image:



Following is the observation of the effect of categorical variables on the dependent variable:

1. Highest demand for rentals in fall
2. Steady increase in demand until June, with a peak in September
3. No significant variation during weekdays, indicating a steady market
4. Demand spikes during clear weather
5. Rentals decrease during holidays
6. Noticeable growth in demand for the following year

Question 2. Why is it important to use **drop_first=True** during dummy variable creation? (Do not edit)

Total Marks: 2 marks (Do not edit)

Answer: <Your answer for Question 2 goes below this line> (Do not edit)

1. When creating dummy variables for a categorical variable, using drop_first=True helps avoid multicollinearity.
2. Multicollinearity happens when one dummy variable can be perfectly predicted by the others. If we include all the dummy variables (without dropping one), the model gets confused because the information from all the variables is redundant.
3. When we drop the first category, we choose it as the "reference" group. The other dummy variables show how each of the remaining categories is different from that reference group.
4. This helps keep the model clean and avoid any issues with multicollinearity.

Question 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?   (Do not edit)

Total Marks: 1 mark (Do not edit)

Answer: <Your answer for Question 3 goes below this line> (Do not edit)

Looking at the pair-plot analysis, variable ‘temp’ among the numerical variables has the highest correlation with the target variable ‘count’.

Question 4. How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: <Your answer for Question 4 goes below this line> (Do not edit)

To validate the assumptions of Linear Regression, the following steps were taken:

1. Linearity: Checked by plotting the residuals versus the fitted values to ensure there is no pattern.
2. Normality: Verified using a Q-Q plot to ensure the residuals are normally distributed.
3. Homoscedasticity: Ensured by plotting the residuals versus the fitted values to check for constant variance.
4. Multicollinearity: Assessed using Variance Inflation Factor (VIF) to ensure no high correlation among the independent variables.

Question 5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (Do not edit)
Total Marks: 2 marks (Do not edit)
Answer: <Your answer for Question 5 goes below this line> (Do not edit)

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

1. Temp
2. Year
3. Light Rain

General Subjective Questions

Question 6. Explain the linear regression algorithm in detail. (Do not edit)
Total Marks: 4 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 6 goes here>

Linear regression is a statistical method used to model the relationship between a dependent variable (also called the target or outcome variable) and one or more independent variables (also called predictors or features).

The goal of linear regression is to find the best-fitting line (in simple linear regression) or hyperplane (in multiple linear regression) that minimizes the difference between the actual data points and the predicted values.

For simple linear regression (one predictor), the model equation looks like:

$y = \beta 0 + \beta 1 x + \epsilon$ $y=\beta 0 \ +\beta 1 \ x+\epsilon$

y is the dependent variable (what you're trying to predict).

x is the independent variable (the predictor).

β_0 is the intercept of the line (the value of y when x = 0).

β_1 is the slope of the line (how much y changes when x increases by 1).

ε is the error term (the difference between the actual and predicted values).

For multiple linear regression (more than one predictor), the equation extends to:

$y = \beta 0 + \beta 1 x 1 + \beta 2 x 2 + . . . + \beta n x n + \epsilon$ $y=\beta 0 \ +\beta 1 \ x 1 \ +\beta 2 \ x 2 \ +...+\beta n \ x n \ +\epsilon$

Here, x_1, x_2, ..., x⍰ are the multiple predictors, and β_1, β_2, ..., β⍰ are their corresponding coefficients.

OLS is the most common way to find the values of β_0, β_1, ..., β⍰. It works by taking the derivative of the RSS with respect to each coefficient and setting it equal to zero to find the optimal values. Once we find the coefficients, we can use the equation to make predictions for new data.

For linear regression to work properly, there are some assumptions we need to check:

1. Linearity: The relationship between the predictors and the target is linear.
2. Independence: The residuals (errors) are independent of each other.
3. Homoscedasticity: The variance of residuals is constant across all levels of the independent variables. Normality of residuals: The residuals should be normally distributed.
4. No multicollinearity: The independent variables should not be highly correlated with each other.

Question 7. Explain the Anscombe’s quartet in detail. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 7 goes here>

Anscombe’s quartet is a set of four datasets that have nearly identical simple descriptive statistics but look very different when graphed. The purpose of Anscombe's quartet is to demonstrate the importance of visualizing data before making assumptions or drawing conclusions from summary statistics alone.

The four datasets in Anscombe's quartet all have the same:

1. Mean of x values
2. Mean of y values
3. Variance of x
4. Variance of y
5. Correlation between x and y

However, despite these identical statistical properties, when plotted, the datasets reveal very different relationships between x and y. This highlights how summary statistics alone (like mean, variance, and correlation) can be misleading without examining the data visually.

Datasets Overview:
Dataset I: Perfect linear relationship.
Dataset II: Curved (non-linear) relationship.
Dataset III: Linear relationship with a strong outlier.
Dataset IV: No linear relationship (vertical line with constant x values).

Even though all datasets have the same statistical properties, the visualizations reveal very different patterns, emphasizing the need for visualization to fully understand data, rather than relying solely on summary statistics

---

Question 8. What is Pearson's R? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 8 goes here>

Pearson's R is a measure of the linear relationship between two variables. It indicates how strongly two variables are related and the direction of their relationship.

Key Points:

- Range: Pearson's R values range from -1 to 1:
    1. +1 means a perfect positive linear relationship: As one variable increases, the other also increases in a perfectly straight line.
    2. -1 means a perfect negative linear relationship: As one variable increases, the other decreases in a perfectly straight line.
    3. 0 means no linear relationship: The variables are uncorrelated.

Interpretation:

- Positive values (0 to 1): A positive value indicates that as one variable increases, the other tends to increase as well (a direct relationship).
- Negative values (-1 to 0): A negative value indicates that as one variable increases, the other tends to decrease (an inverse relationship).
- Close to 0: Values close to 0 suggest a very weak linear relationship or no linear relationship at all.

Question 9. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 9 goes here>
Scaling is the process of transforming numerical features to a common range, 0 to 1 or -1 to 1. This will help improve model performance and stability.

Numeric input features will have values in different ranges. It can be in millions, thousands or even in fractions.

Scaling ensures no features dominate due to difference in magnitudes.
Reduces numerical instability in calculations.

Normalized scaling is also called min-max-scaling, which computes the values and converts them to range between 0 and 1.

X' = (X – min(X))/(max(X) – min(X)

Standardized scaling centers the data around mean 0 with standard deviation of 1, implying the values can range between -1 and +1

X' = (X – mean) / standard deviation

Normalized scaling is done when the data is not normally distributed and standardized scaling is done when the data is normally distributed

---

Question 10. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 10 goes here>

- The Variance Inflation Factor (VIF) measures how much the variance of a regression coefficient is inflated due to collinearity with other predictors in the model. A high VIF indicates that a predictor is highly correlated with one or more other predictors, which can lead to issues in the regression model, such as unstable coefficients.

- The VIF becomes infinite when two or more features in the model are perfectly correlated. This means that one feature can be exactly predicted from the others, causing the model to struggle in separating their individual effects.

- As a result, the variance (uncertainty) in estimating the coefficients of these features becomes very large, making the VIF go to infinity. Simply put, it happens when there's redundancy between the features, and there's no way to explicitly read which feature has better explanation for the target variable.

Question 11. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 11 goes here>

A Q-Q plot (quantile-quantile plot) is a graphical tool used to assess whether a dataset follows a particular theoretical distribution, typically the normal distribution. It plots the quantiles of the data against the quantiles of the theoretical distribution.

Use and Importance in Linear Regression:

- Checking Normality: In linear regression, one of the key assumptions is that the residuals (errors) are normally distributed. A Q-Q plot helps in visually assessing this assumption. If the residuals follow a straight line in the Q-Q plot, it indicates that they are normally distributed.
- Identifying Deviations: Deviations from the straight line in a Q-Q plot indicate departures from normality. This can help identify skewness, kurtosis, or other distributional issues in the residuals.
- Model Validity: Ensuring that the residuals are normally distributed is crucial for the validity of the regression model. It affects the accuracy of confidence intervals and hypothesis tests. A Q- Q plot provides a simple and effective way to check this assumption.

In summary, a Q-Q plot is an essential diagnostic tool in linear regression for verifying the normality of residuals, which is a critical assumption for the validity of the model.