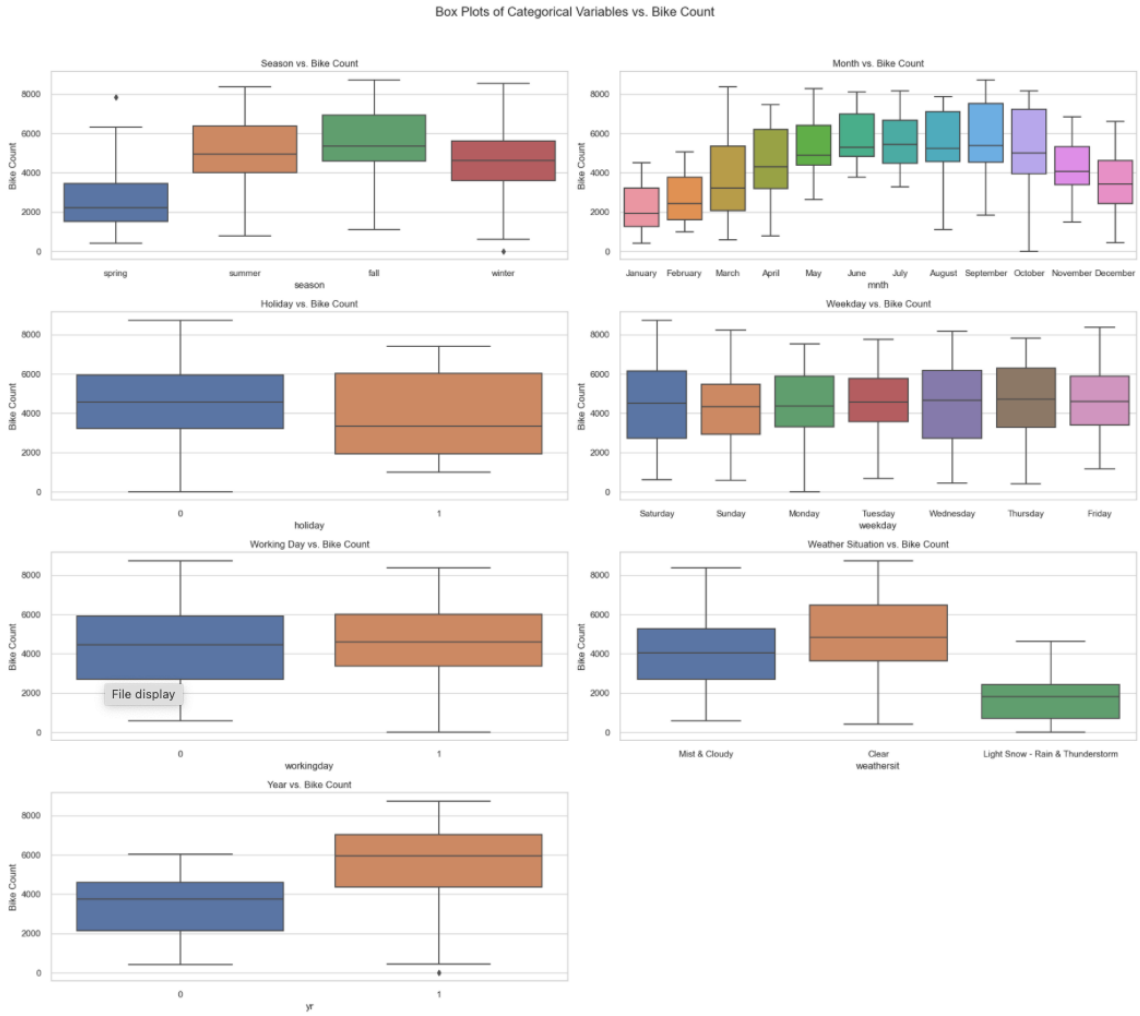
Assignment-based Subjective Questions

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)

This visualization provides a comprehensive analysis of how each categorical variable influences the target variable.



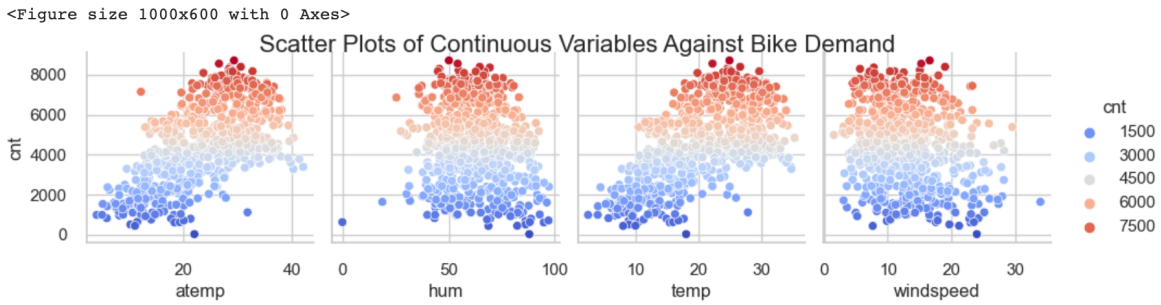
- **Seasonal Trends**: Bike rentals significantly decline in spring, with peaks in the fall and summer (June to September).
- **Monthly Patterns**: An upward rental trend is observed from January to September, with notable dips in November and December.
- **Holiday Impact**: Rentals are higher on non-holiday days, with minimal daily variation.
- **Weather Effects**: Clear weather conditions lead to increased rentals, while heavy rain and thunderstorms deter them.
- **Yearly Comparison**: Rentals in 2019 exceeded those in 2018.

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)

There are some reasons for which it becomes necessary to use drop_first=true while generating dummy variables. For instance, by removing one category, it helps in avoiding multicollinearity. Secondly, it makes the interpretation easier as the narrator has to define only how the remaining categories relate to the one that has been left out. Finally, it decreases the number of columns in the dataset which leads to a less complex model that has superior performance and a lower probability of being overfitted.

To conclude, drop_first=true is an optimization in the model, which does not lose sense and important information.

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)



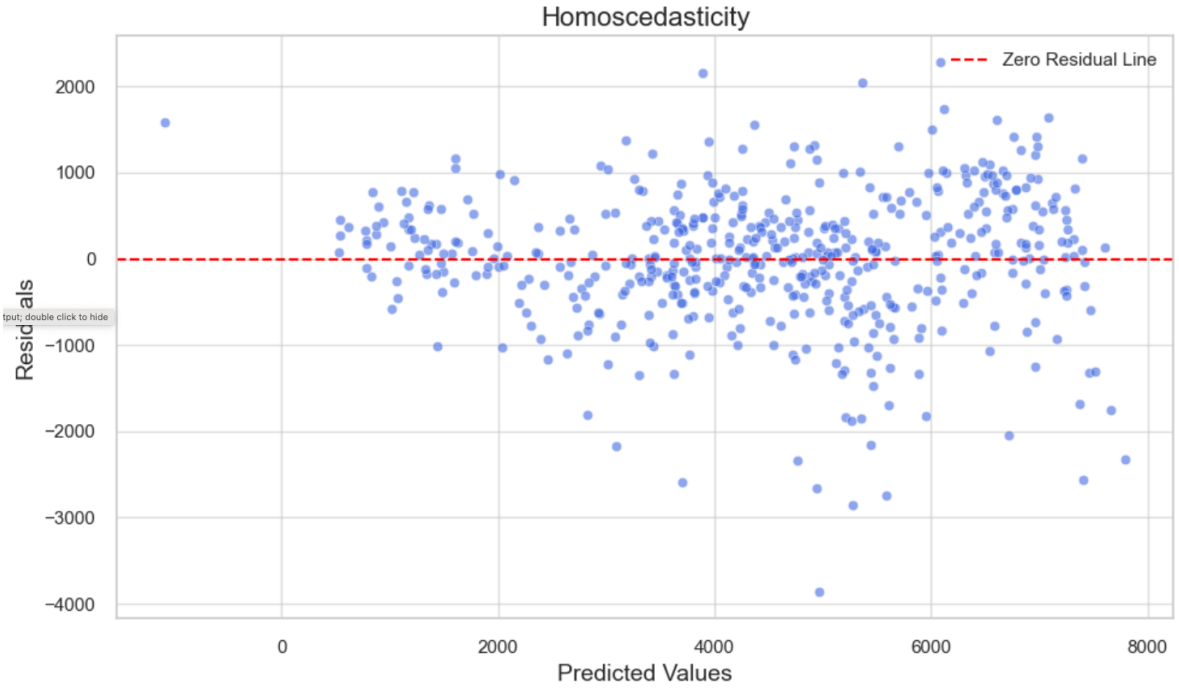
Based on the scatter plot, temperature (temp) appears to have the highest correlation with the target variable (bike demand).

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)

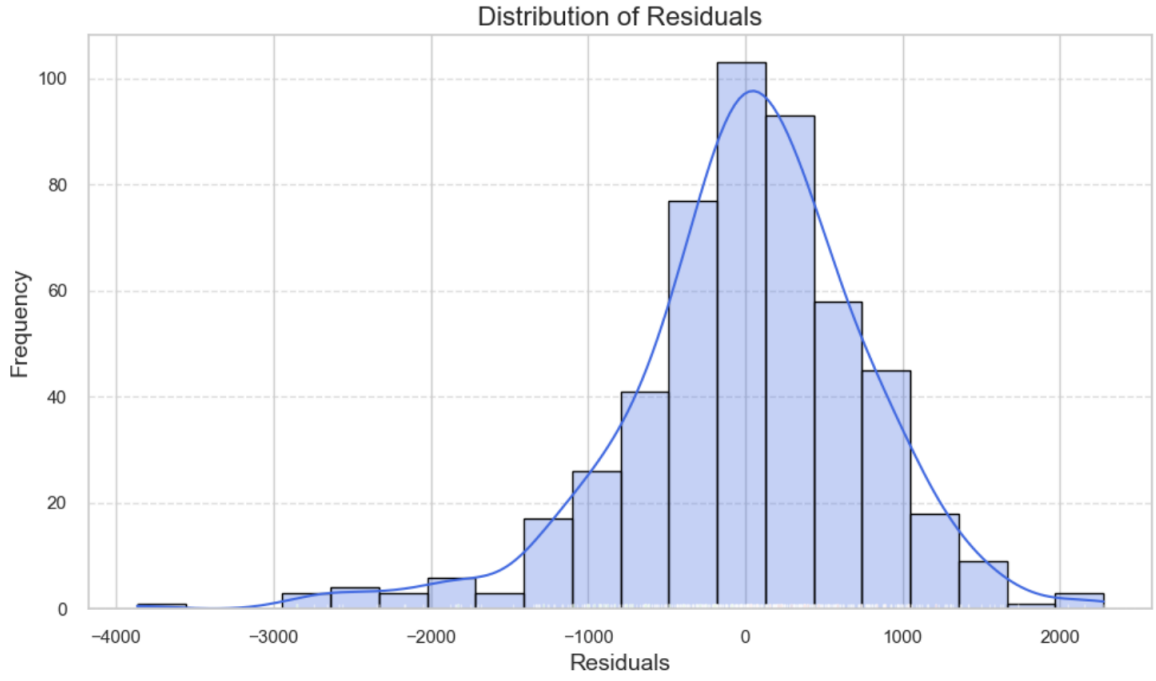
Post-model development, it is critical to validate the assumptions of linear regression. Recovery and maintenance of the integrity and dependability of the model is what is sought here. The following are the major assumptions supplemented by methods of their validation:

1. Linearity: Use scatter plots of predicted and actual outcomes as well as graphs of residuals to look for any trends.
2. Independence: Use Durbin-Watson test statistics (where 2 is indicative of independence) to test for independence of residuals and residual plots for time series residuals.
3. Homoscedasticity: Examine residual plots against predicted outcomes; uniform spread is

expected. If the form is a funnel, heteroscedasticity is in play.



4. Normality of Residuals: The residuals should be approximately normally distributed.



5. No Multicollinearity: Use Variance Inflation Factor (VIF) to assess predictors with values above 10 showing collinearity and correlation matrices to see highly correlated variables.

	Features	VIF
1	holiday	inf
2	workingday	inf
11	Saturday	inf
12	Sunday	inf
3	temp	3.44
6	spring	2.99
7	winter	2.09
4	hum	1.83
14	Mist & Cloudy	1.55
9	November	1.50
8	July	1.37
13	Light Snow - Rain & Thunderstorm	1.19
5	windspeed	1.17
10	September	1.13
0	yr	1.05

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)

The top three features that significantly influence the demand for shared bikes are

1. Temperature.
2. Year
3. Season

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>

A statistical technique used to represent the relationship between a dependent variable and one or more independent variables is called linear regression. We can make predictions and comprehend how independent variables affect the dependent variable because this relationship is represented as a linear equation.

1. **Conceptual Framework**:

The core idea of linear regression is to fit a straight line (or hyperplane in multiple dimensions) that best represents the relationship between the variables. This line helps us understand how changes in the independent variables affect the dependent variable.

---

2. **Assumptions**:

Linear regression relies on several assumptions:

a. Linearity: The relationship between the independent and dependent variables is linear, meaning that a straight line can represent the connection.
b. Independence: Observations are assumed to be independent from one another, meaning that the value of one observation does not influence another.
c. Homoscedasticity: The variability of the residuals (the differences between observed and predicted values) should remain constant across all levels of the independent variables.
d. Normality: The residuals should be approximately normally distributed, which is important for making inferences about the model.

3. **Training the Model:**

To create a linear regression model, we use a dataset where we know the outcomes (dependent variable) and the predictors (independent variables). The algorithm determines the best-fit line by finding the coefficients (weights) for each predictor that minimize the differences between the observed values and the values predicted by the model.

4. **Making Predictions:**

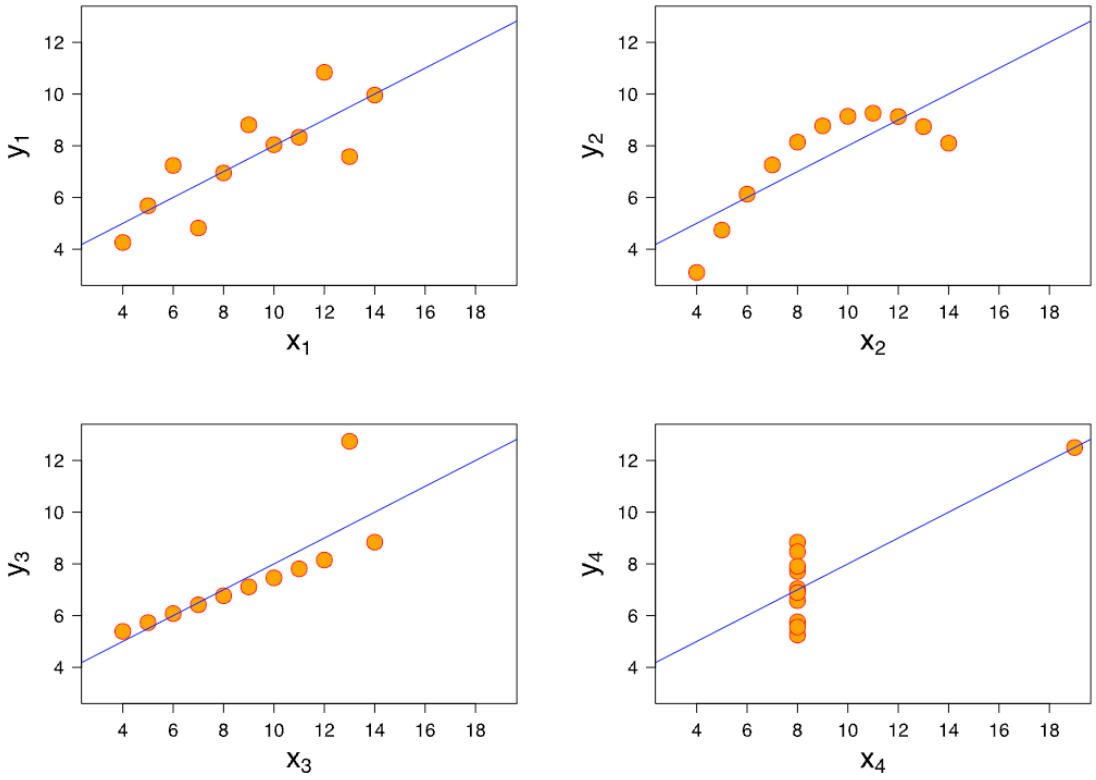
Once the model is trained, it can be used to make predictions on new data. For instance, if we have a model that predicts bike demand based on features like temperature and time of day, we can input new values for these features to estimate the expected demand.

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)

The four datasets that make up Anscombe's Quartet have the same descriptive statistics (mean, variance, and correlation), but their variable distributions and relationships are radically different. The quartet, which Francis Anscombe created in 1973, highlights the value of data visualisation in statistical analysis.

The Four Datasets:

1. Dataset 1: Displays a strong linear relationship, with points closely aligned along a straight line.
2. Dataset 2: Although it has the same statistics as Dataset 1, it reveals a parabolic relationship, indicating a non-linear trend.
3. Dataset 3: Features a vertical line, suggesting that the independent variable almost perfectly predicts the dependent variable, with minimal variability.
4. Dataset 4: Similar to Dataset 1 but includes a significant outlier that skews the summary statistics, affecting correlation and mean.



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)

Pearson's R, also known as the Pearson correlation coefficient, is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables. Here are three key points about Pearson's R:

1. **Value Range**: Pearson's R ranges from -1 to 1. A value of 1 indicates a perfect positive linear correlation, -1 indicates a perfect negative linear correlation, and 0 indicates no linear correlation between the variables.
2. **Interpretation of Strength**: The closer the value of Pearson's R is to 1 or -1, the stronger the linear relationship. Generally, values between 0.1 and 0.3 indicate a weak correlation, 0.3 to 0.5 indicate a moderate correlation, and above 0.5 indicate a strong correlation.
3. **Assumptions**: Pearson's R assumes that the relationship between the variables is linear, both variables are normally distributed, and the data points are independent of each other. It is sensitive to outliers, which can significantly affect the correlation coefficient.

The formula for Pearson's correlation coefficient (r) is as follows:

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

Where:

- n = number of data points (pairs of x and y)
- x = values of the first variable
- y = values of the second variable
- $\sum xy$ = sum of the product of paired scores
- $\sum x$ = sum of the values of the first variable
- $\sum y$ = sum of the values of the second variable
- $\sum x^2$ = sum of the squares of the values of the first variable
- $\sum y^2$ = sum of the squares of the values of the second variable

---

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)

Scaling is the process of adjusting the range or distribution of data to ensure equal contribution from each feature. It's important for algorithms sensitive to the scale, like distance-based models. There are two main types: normalization, which scales data to a specific range (usually 0 to 1), and standardization, which adjusts data to have a mean of 0 and a standard deviation of 1. Normalization is useful for bounding data, while standardization is better when data follows a normal distribution or requires equal variance.

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)

A Variance Inflation Factor (VIF) becomes infinite when perfect multicollinearity exists among two or more predictor variables within a regression model. This condition arises when one or more variables can be perfectly predicted by a linear combination of others.

In such a scenario, the VIF formula, which involves calculating the ratio of the variance of a model coefficient to its variance when the variable is perfectly uncorrelated with the others, results in a division by zero. Consequently, the VIF value becomes undefined or infinite, indicating a severe multicollinearity issue.

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)

A **Q-Q plot** (Quantile-Quantile plot) is a graphical tool used to assess if a dataset follows a specific theoretical distribution, such as a normal distribution. It compares the quantiles of the data to the quantiles of the theoretical distribution.

Use and Importance in Linear Regression:

- **Use:** In linear regression, a Q-Q plot is used to check the **normality of residuals** (the differences between observed and predicted values). If the residuals are normally distributed, they should align closely along a straight line in the Q-Q plot.
- **Importance:** Normality of residuals is a key assumption of linear regression. A Q-Q plot helps verify this assumption, ensuring that statistical tests and model predictions are valid. If the plot deviates significantly from a straight line, it indicates non-normal residuals, suggesting that the model may not be appropriate.