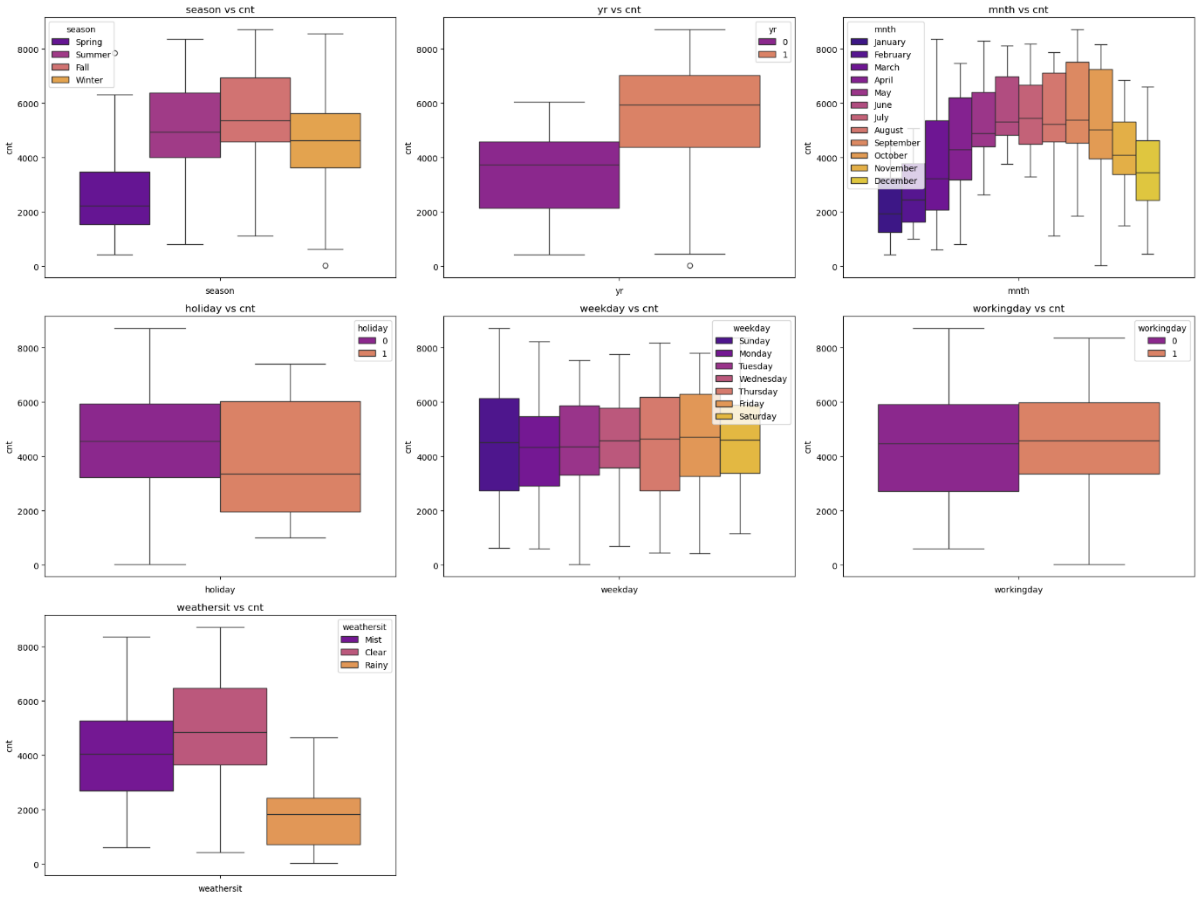
Assignment-based Subjective Questions

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)

- The Year has a strong positive impact on the demand for shared bikes. This suggests that the demand for shared bikes has been steadily increasing over the years from 2018 to 2019.
- Seasonality (spring, summer, fall, winter) also significantly influences demand. Summer and fall generally see higher demand compared to winter and spring.
- The day of the week impacts demand. Weekends (Saturday and Sunday) typically have higher demand compared to weekdays.
- The presence of holidays can significantly impact demand, usually people tend to stay at home in holidays and hence the booking count is low.
- Working Day and Non-Working Day are almost having same count values of bike booking and hence have not major impact.
- Majority of bookings were done in mid-year starting from June to Sept and then year-end the bike booking again decreases.



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)

By using drop_first=True, we drop the first level of the categorical variable. This effectively creates a reference category, and the coefficients of the remaining dummy variables represent the difference between that level and the reference category. This approach prevents redundancy and improves model interpretability.

For a categorical variable with n categories, there can be (n-1) dummy variables.

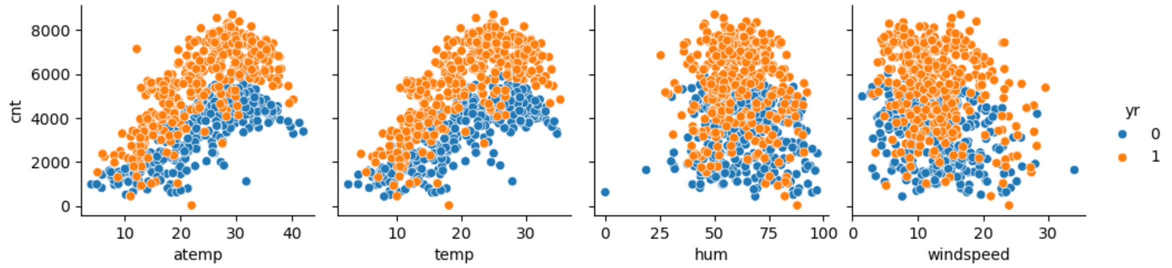
E.g. In Holiday variable there are 2 categorical values 0 and 1 but the dummy variable created for them is only one Holiday_1 which means if its value is 1 then it will be holiday and if 0 then non holiday

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 a pair plot, the dependent variable **temperature** (**temp**) has the highest correlation with the target variable (**cnt**) which is count. Given that 'temp' and 'atemp' are redundant variables

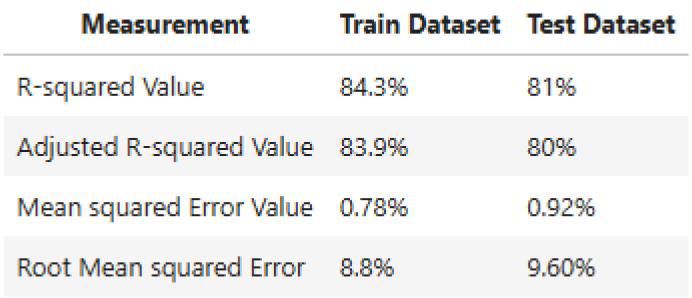


Question 4. How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)

Total Marks: 3 marks (Do not edit)

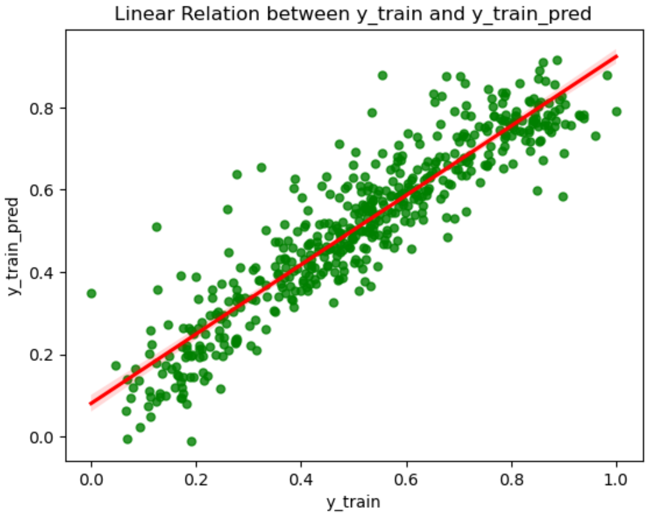
Answer: <Your answer for Question 4 goes below this line> (Do not edit)

Validating the assumptions of linear regression or creating Hypothesis is important for build a strong and good model. After building the train model there are below given steps which are followed for validating and unable to reject the hypothesis.

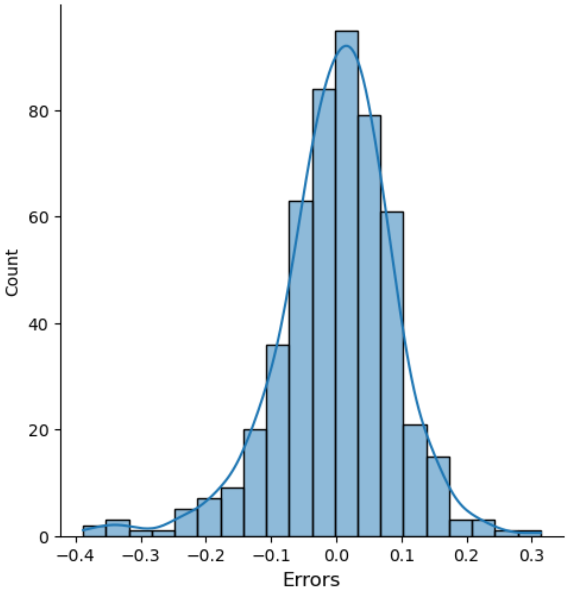


Measurement	Train Dataset	Test Dataset
R-squared Value	84.3%	81%
Adjusted R-squared Value	83.9%	80%
Mean squared Error Value	0.78%	0.92%
Root Mean squared Error	8.8%	9.60%

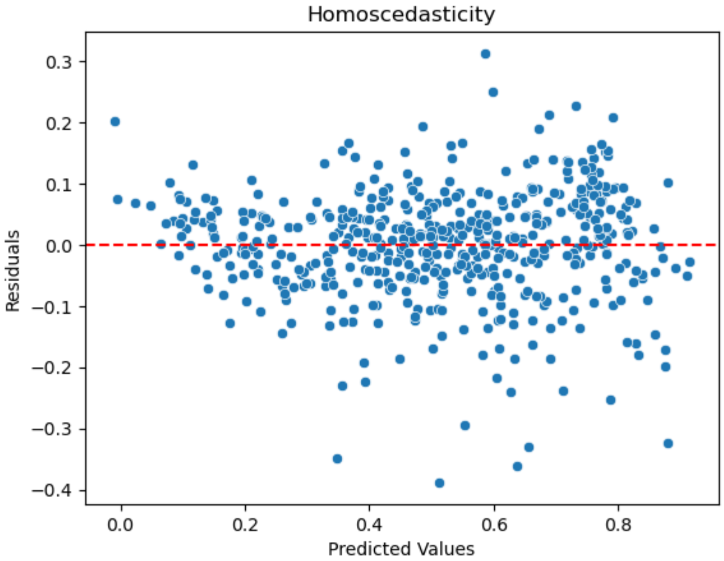
Linearity: Check the scatter plots between the independent variables and the dependent variable. The relationship should be approximately linear.



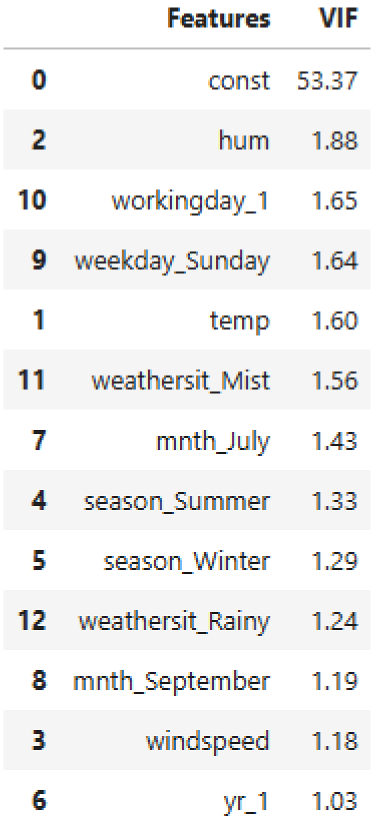
Normality of residuals: Create a histogram or Q-Q plot of the model residuals. They should be normally distributed.



Homoscedasticity (constant variance of residuals): Create a scatter plot of residuals against predicted values. The spread of residuals should be constant across all predicted values.



Independence of residuals: Check for autocorrelation in the residuals using the Durbin-Watson test. There should be no significant autocorrelation.
Multicollinearity: Calculate the Variance Inflation Factor (VIF) for each independent variable. High VIF values (typically above 5 or 10) indicate multicollinearity.



	Features	VIF
0	const	53.37
2	hum	1.88
10	workingday_1	1.65
9	weekday_Sunday	1.64
1	temp	1.60
11	weathersit_Mist	1.56
7	mnth_July	1.43
4	season_Summer	1.33
5	season_Winter	1.29
12	weathersit_Rainy	1.24
8	mnth_September	1.19
3	windspeed	1.18
6	yr_1	1.03

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)

As per the final equation for multiple linear regression
cnt = 0.18 + 0.60 x temp. - 0.17 x hum - 0.20 x windspeed + 0.08 x season_Summer + 0.14 x season_Winter + 0.23 x yr_1 - 0.04 x mnth_July + 0.09 x mnth_September + 0.06 x weekday_Sunday + 0.05 x workingday_1 - 0.05 x weathersit_Mist - 0.24 x weathersit_Rainy

The top 3 features contributing significantly towards explaining the demand of shared bikes (based on the model's coefficients or feature importance scores) are likely to be:

1. Season (Winter)
2. Temperature (temp)
3. Year (yr)

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)

Linear Regression is a statistical method used to model the relationship between a dependent variable (target) and one or more independent variables (predictors). It assumes a linear relationship between the variables, meaning the relationship can be represented by a straight line.

The core idea is to find the best-fitting line that minimizes the difference between the actual values of the dependent variable and the values predicted by the line. This is often done using the method of least squares, which aims to minimize the sum of the squared differences between the actual and predicted values.

Equation:

$Y = \beta 0 + \beta 1X1 + \beta 2X2 + ... + \beta n*Xn + \varepsilon$
where:
* Y is the dependent variable
* X1, X2, ..., Xn are the independent variables
* β0 is the intercept (the value of Y when all X's are 0)
* β1, β2, ..., βn are the coefficients (slopes) that represent the change in Y for a unit change in each X
* ε is the error term (the difference between the actual and predicted values)

Key Concepts:

- **Simple Linear Regression:** Involves one independent variable.
- **Multiple Linear Regression:** Involves two or more independent variables.
- **Assumptions:** Linearity, normality of residuals, homoscedasticity, independence of residuals, no multicollinearity.

Applications:

- Predicting stock prices
- Analyzing sales trends
- Determining the impact of factors on health outcomes
- Many other areas where relationships between variables need to be understood and quantified.

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)

Anscombe's Quartet is a set of four datasets that have nearly identical simple descriptive statistics (mean, variance, correlation) but appear very different when plotted. This demonstrates the importance of visualizing data beyond relying solely on summary statistics.

Key Points:

- **Dataset 1:** A typical linear relationship.
- **Dataset 2:** A non-linear relationship that appears linear when only looking at summary statistics.

- **Dataset 3:** A linear relationship with an outlier that significantly influences the regression line.
- **Dataset 4:** A horizontal line with a single point far from the others, resulting in a high correlation despite no real linear relationship.

Importance:

- **Visualizations are crucial:** They provide insights into the underlying data structure that summary statistics might miss.
- **Outliers can have a significant impact:** They can distort the results of statistical analyses.
- **Correlation does not imply causation:** High correlation does not necessarily mean a causal relationship exists.

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 linear relationship between two continuous variables.

- **Values:** It ranges from -1 to 1.
  - +1: Perfect positive linear correlation (as one variable increases, the other increases)
  - -1: Perfect negative linear correlation (as one variable increases, the other decreases)
  - 0: No linear correlation
- **Interpretation:**
  - A value close to 1 or -1 indicates a strong linear relationship.
  - A value close to 0 indicates a weak or no linear relationship.
- **Limitations:**
  - Only measures linear relationships.
  - Sensitive to outliers.

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

Scaling is a data preprocessing technique that transforms the features of a dataset to a common scale. This is often necessary because features in real-world datasets can have vastly different ranges (e.g., one feature might range from 0 to 1, while another might range from 1000 to 10000).

Why is Scaling Performed?

- **Improves model performance:** Many machine learning algorithms perform better when features are on a similar scale.
- **Prevents features with larger ranges from dominating others:** In distance-based algorithms (like k-Nearest Neighbors), features with larger ranges can have a disproportionate influence on the distance calculations.
- **Faster convergence:** Some optimization algorithms converge faster when features are scaled.

Normalized Scaling (Min-Max Scaling):

- Transforms features to a specific range, typically between 0 and 1.
- Formula:
  - (x - min(x)) / (max(x) - min(x))
- Sensitive to outliers.

Standardized Scaling (Z-score Scaling):

- Transforms features to have zero mean and unit variance.
- Formula: (x - mean(x)) / std(x)
- Less sensitive to outliers.

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)

The Variance Inflation Factor (VIF) becomes infinite when there is perfect multicollinearity among the independent variables in a regression model. Here's why:

Understanding VIF

- VIF quantifies how much the variance of a regression coefficient is inflated due to multicollinearity.
- It is calculated as:

  VIF=1/(1−Rj2)

  where Rj2 is the coefficient of determination for the regression of one independent variable on all other independent variables.

Why Infinite VIF?

- **Perfect Multicollinearity**: If one independent variable can be perfectly predicted as a linear combination of other variables, the Rj2R^2_jRj2 value for that variable becomes 1.
- Substituting (Rj2-1) into the VIF formula:

  VIF=1/(1−1)=∞. This results in an infinite VIF.

Causes of Perfect Multicollinearity

1. **Redundant Variables:**
   - Including variables that are linearly dependent (e.g., dummy variables for all categories of a categorical variable without dropping one baseline category).
2. **Duplicate Variables:**
   - If two variables are identical or highly correlated, one can be perfectly predicted by the other.
3. **Poor Data Scaling or Transformation:**
   - Improper handling of features (e.g., combining features in a way that creates perfect dependency).
4. **Insufficient Data:**
   - If the number of observations is less than or equal to the number of independent variables, perfect multicollinearity can arise.

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 whether a dataset follows a particular theoretical distribution, commonly a normal distribution. It compares the quantiles of the observed data against the quantiles of the theoretical distribution.

How a Q-Q Plot Works

1. **Observed Quantiles:** These are the quantiles from the dataset (e.g., residuals in linear regression).
2. **Theoretical Quantiles:** These are the quantiles from the chosen theoretical distribution (e.g., a standard normal distribution).
3. **Plotting:**
   - The x-axis represents the theoretical quantiles.
   - The y-axis represents the observed quantiles.
   - If the data follows the theoretical distribution closely, the points will lie approximately on a 45-degree line (diagonal).

Use of Q-Q Plot in Linear Regression

In the context of linear regression, Q-Q plots are primarily used to check the assumption of **normality of residuals**. The key assumptions of linear regression include:

1. **Linearity**: The relationship between predictors and the response is linear.
2. **Homoscedasticity**: The residuals have constant variance.
3. **Independence**: The residuals are independent.
4. **Normality of Residuals**: The residuals are normally distributed.

- **Why Check Normality?**

    - Normality of residuals ensures that hypothesis tests (like t-tests and F-tests) are valid.
    - It affects the accuracy of confidence intervals and predictions

Importance of Q-Q Plot in Linear Regression

1. **Assess Normality:**
    - A Q-Q plot helps visually assess whether the residuals are approximately normal.
    - Deviations from the diagonal line indicate departures from normality (e.g., skewness or kurtosis).
2. **Identify Patterns:**
    - Points systematically deviating from the line can reveal:
        - Heavy tails (leptokurtosis): Points curve away from the line at both ends.
        - Light tails (platykurtosis): Points deviate inward at the ends.
        - Skewness: Asymmetry in the distribution.
3. **Diagnostic Tool:**
    - If residuals are not normal, it might indicate model inadequacy, presence of outliers, or the need for transformations.
4. **Guide Model Refinement:**
    - Suggests remedial measures like:
        - Applying transformations (e.g., log or Box-Cox transformations) to the response variable.
        - Addressing outliers or influential points.
