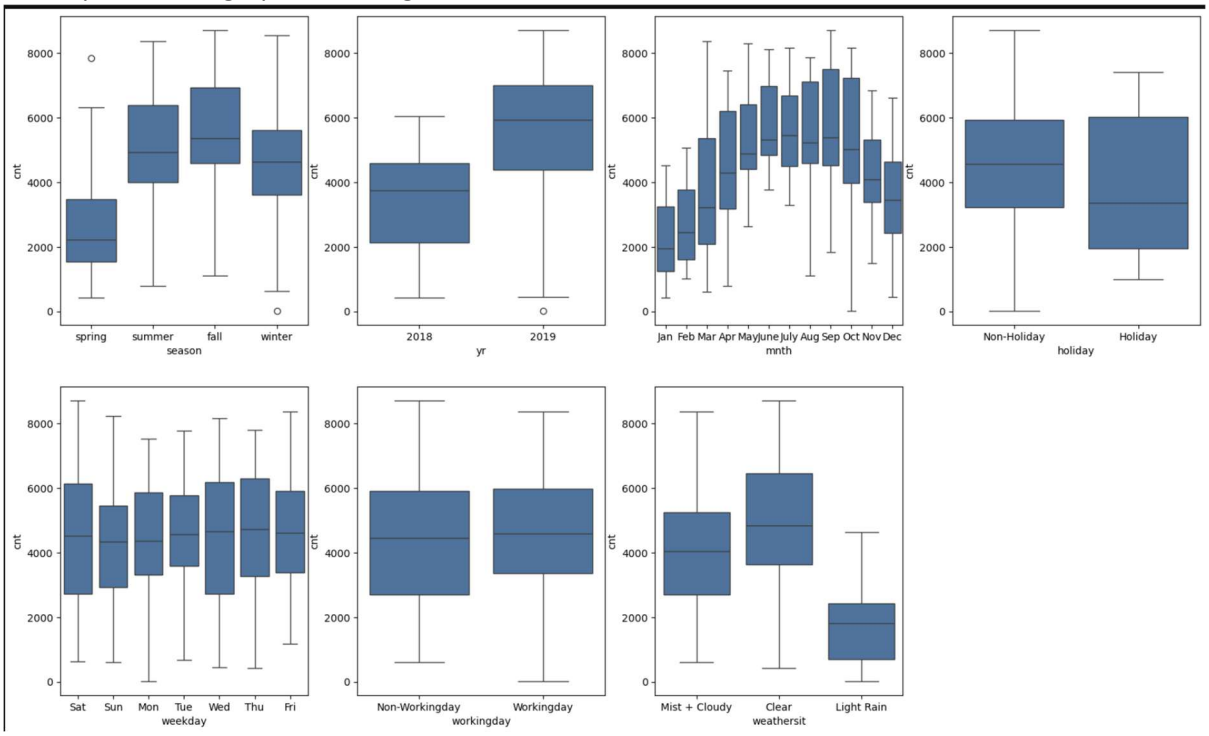
Assignment-based Subjective Questions

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)

I have plotted the graphs for categorical variables mentioned below:-



Here are the inferences about the effects of categorical variables on the dependent variable (cnt):

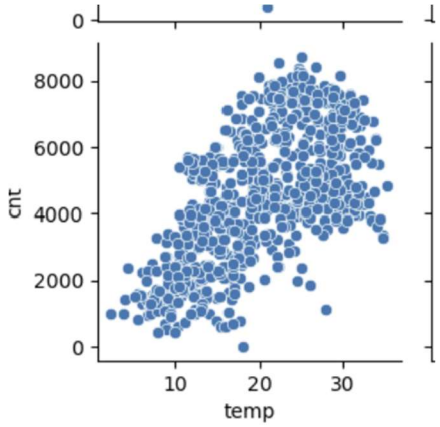
- Season (Plot-1): Bike counts are shown for four seasons: spring, summer, fall, and winter. Higher median counts are seen in summer and fall compared to spring and winter.
- Year (Plot-2): Comparison of bike counts between the years 2018 and 2019, showing a higher median count in 2019.
- Month (Plot-3): Bike counts for each month, from January to December. Counts generally increase from January to July, peak around August, and then decrease towards December.
- Holiday (Plot-4): Comparison of bike counts between non-holidays and holidays, with slightly higher median counts on holidays.
- Weekday (Plot-5): Bike counts for each day of the week, showing relatively consistent counts with slight variations.
- Working Day (Plot-6): Comparison of bike counts between non-working days and working days, with higher median counts on working days.
- Weather Situation (Plot-7): Bike counts for different weather conditions: Mist + Cloudy, Clear, and Light Rain. Clear weather has the highest median counts, while Light Rain has the lowest. Also, there is no data for Heavy Rain. It means bikes are not rented on days when there is heavy rain.

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)

Using drop_first=True during dummy variable creation avoids multicollinearity by dropping the first category, ensuring the model does not have redundant 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 pair-plot, the numerical variable with the highest correlation with the target variable (cnt) is temperature (temp). This strong correlation suggests that higher temperatures are associated with increased bike rentals.

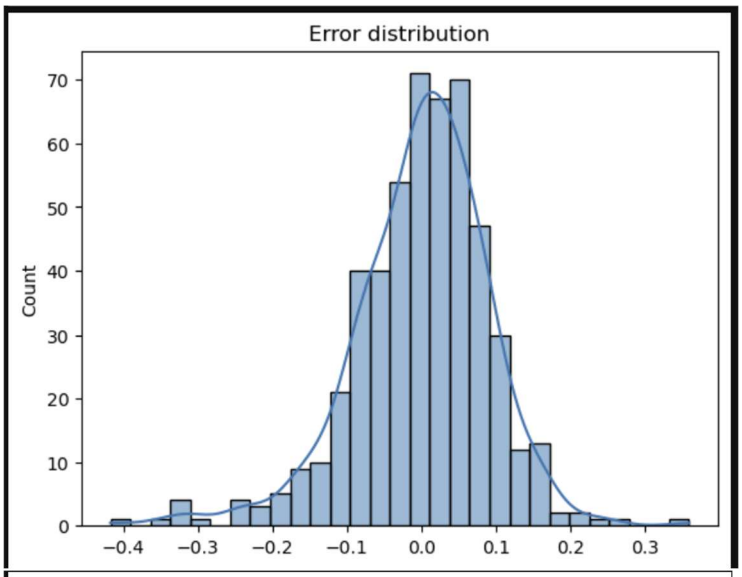


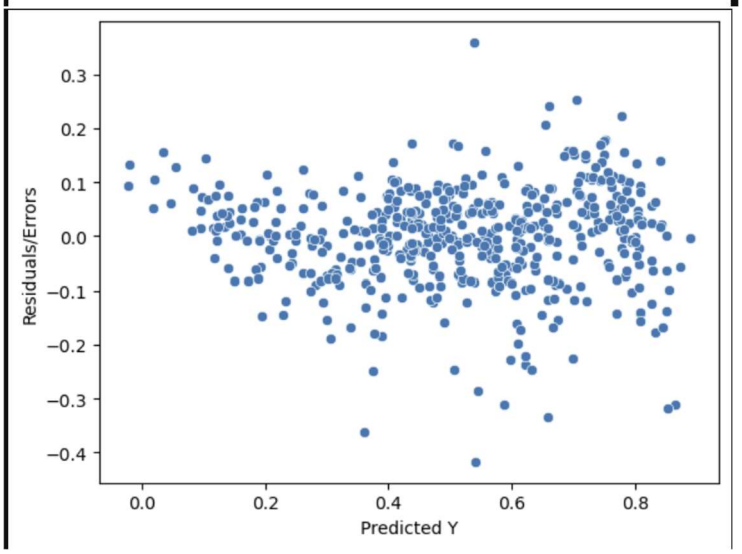
Question 4. How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: <Your answer for Question 4 goes below this line> (Do not edit)

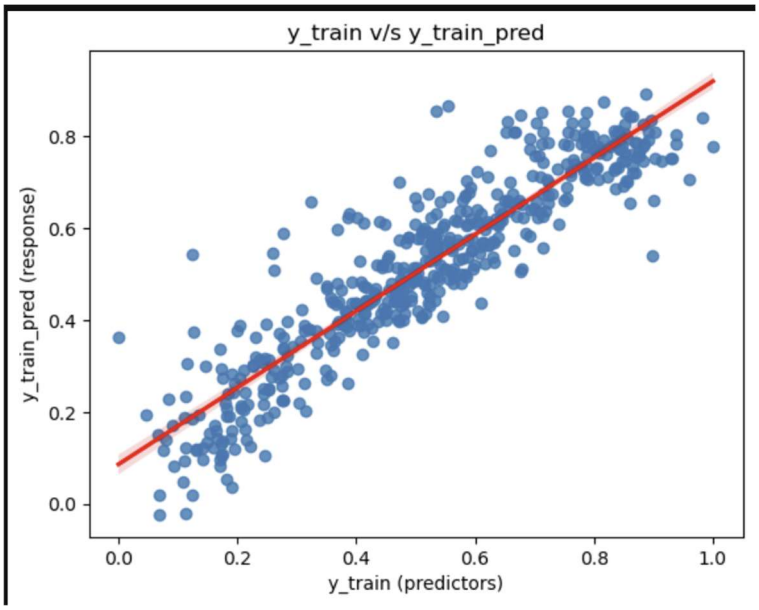
1. I have first checked Distribution of the error terms and checked if the error terms are also normally distributed and mean is centered at 0 (which is one of the major assumptions of linear regression)
2. Then I checked the **Independence Assumption**: The residuals were independent of each other. There was no correlation between them. Also the residuals have **constant variance.**
3. After I checked the **Linearity Assumption**: The relationship between the independent variables (predictors) and the dependent variable (response) was linear.

Also, for all selected features VIF was < 5 and p value was < 0.03.

Here are the 3 graphs I created.


Error distribution
Count
70
60
50
40
30
20
10
0
−0.4
−0.3
−0.2
−0.1
0.0
0.1
0.2
0.3


Residuals/Errors
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.4
0.0
0.2
0.4
0.6
0.8
Predicted Y


y_train v/s y_train_pred
y_train_pred (response)
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
y_train (predictors)

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 for shared bikes are:

1. Temperature (temp): With the highest positive coefficient (0.48), it has the most substantial impact on bike rentals.

2. Weather Condition (Light Rain): Negative coefficient (-0.29), significantly reducing bike rentals in adverse weather conditions.

3. Year (yr): Positive coefficient (0.23), indicating an increasing trend in bike rentals over the years.

The absolute value of the coefficient for Light Rain (0.29) is greater than that of yr (0.23), meaning that Light Rain has a more significant impact on bike rentals than the year variable, though its impact is negative.

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 (y) and one or more independent variables (x1,x2,x3…xn). The goal is to find the best-fitting line that describes the relationship.

Key Concepts:
Model Equation: The linear regression equation is given by:

$y = \beta 0 + \beta 1x1 + \beta 2x2 + \ldots + \beta nxn$
y is the dependent variable.
$\beta 0$ is the intercept.
$\beta 1, \beta 2, \ldots, \beta n$ are the coefficients of the independent variables also known as slopes.
$x1, x2, \ldots, xn$.

Assumptions: The algorithm relies on the following assumptions:

Linearity: The relationship between predictors and the response is linear.
Independence: Residuals (errors) are independent.
Homoscedasticity: Constant variance of residuals.
Normality: Residuals are normally distributed (for inference) and mean is at 0.

Objective: The objective is to minimize the sum of squared residuals (differences between observed and predicted values) to find the best-fitting line.

Training: The model is trained using methods like Ordinary Least Squares (OLS) to estimate the coefficients by minimizing the sum of squared residuals.

Prediction: Once trained, the model can predict the response variable for new data points using the estimated coefficients.

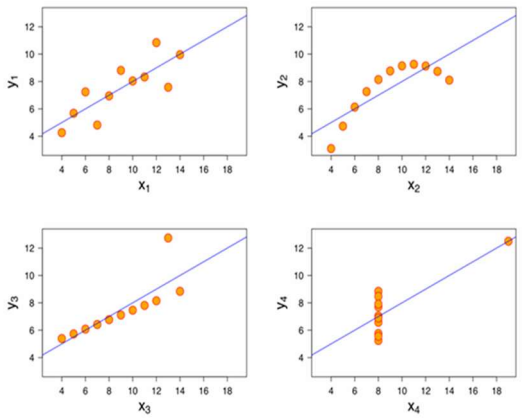
Evaluation:
The model's performance is assessed using metrics such as (R) (coefficient of determination), MSE, or other relevant metrics, depending on the context.

Linear regression is widely used for its simplicity, interpretability, and effectiveness in many applications, such as forecasting and understanding relationships between variables.

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 consists of four datasets created by statistician Francis Anscombe to emphasize the importance of visualizing data. Despite having nearly identical summary statistics (mean, variance, correlation, and linear regression line), the datasets appear very different when graphed.



Dataset I: Shows a clear linear relationship with consistent scatter around the regression line.
Dataset II: Suggests a nonlinear relationship, forming a curve.
Dataset III: Contains a prominent outlier that significantly influences the regression line.
Dataset IV: Displays a vertical line with a single horizontal outlier, indicating one influential point.

The quartet demonstrates that relying solely on statistical summaries can be misleading. Visualizing data reveals patterns, outliers, and relationships that are not evident from summary statistics alone.

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, also known as the Pearson correlation coefficient, measures the linear relationship between two variables.

Range: -1 to 1
1: Perfect positive correlation
0: No linear correlation
-1: Perfect negative correlation

Interpretation:
Positive Value: An increase in one variable corresponds to an increase in the other.
Negative Value: An increase in one variable corresponds to a decrease in the other.
Magnitude: Closer to 1 or -1 indicates a stronger linear relationship.

Pearson's R is widely used in statistics to assess and identify potential correlations in data.

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 features to a specific range to ensure uniformity and improve the performance of machine learning models.

Scaling is performed because of below reasons:

- Improves Model Performance: Many machine learning algorithms perform better when the features are on a similar scale.
- Accelerates Convergence: Scaling helps gradient-based algorithms converge faster by reducing the risk of numerical instabilities.
- Balances Feature Contributions: Ensures that no single feature dominates the model due to its magnitude.

Difference between Normalized Scaling and Standardized Scaling is mentioned below:-

Normalization (Min-Max Scaling):
Transforms features to a specific range, between [0, 1].
Formula: $x$norm= x− min(x) / (max(x)−min(x))

Standardization (Z-Score Scaling):
Transforms features to have a mean of 0 and a standard deviation of 1.
Formula:
$x$std=$(x-\mu)/\sigma$

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>

An infinite Variance Inflation Factor (VIF) value arises when there is perfect multicollinearity among the predictors in a regression model.
This condition means that one predictor is an exact linear combination of other predictors, making it impossible to uniquely estimate the regression coefficients. Essentially, perfect multicollinearity prevents the model from distinguishing the individual effects of the predictors, leading the VIF to become infinite.
To resolve this issue, you can remove one of the collinear predictors or combine them in a meaningful way.

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 (Quantile-Quantile) plot is a graphical tool used to assess if a dataset follows a particular theoretical distribution, commonly the normal distribution. It compares the quantiles of the sample data with the quantiles of a specified theoretical distribution.

Use
Assessing Normality: In linear regression, one key assumption is that the residuals (errors) are normally distributed. A Q-Q plot helps visualize if the residuals adhere to a normal distribution.

Identifying Deviations: Points on the Q-Q plot should ideally form a straight line if the residuals are normally distributed. Deviations from this line indicate departures from normality, such as skewness or kurtosis.

Importance
Model Validation: Ensuring the residuals follow a normal distribution validates the assumptions of

linear regression, contributing to the reliability of inference and predictions.

Diagnostic Tool: A Q-Q plot helps diagnose potential issues in the model, guiding necessary adjustments to improve model accuracy and adherence to assumptions.
