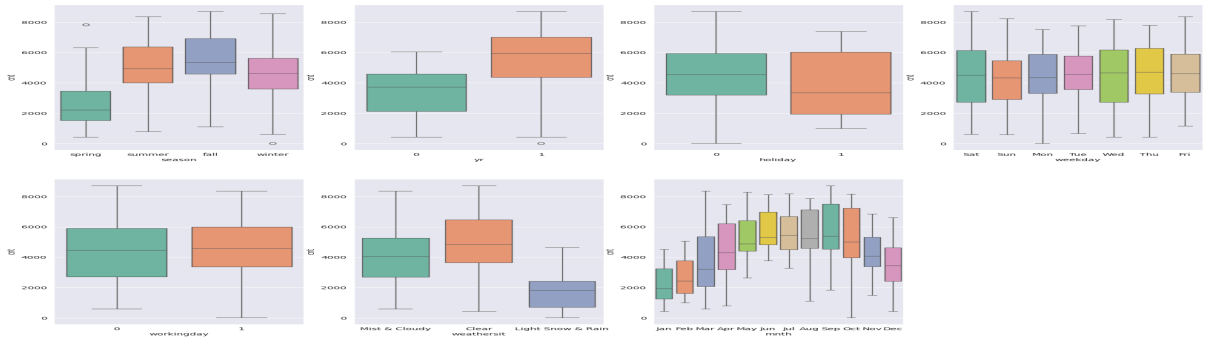
Assignment-based Subjective Questions

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)

- Seasonal analysis reveals that Fall (category 3) has the highest median demand, indicating peak rentals during this season, while Spring (category 1) has the lowest.
- The year 2019 had a higher user count compared to 2018.
- Rental distribution remains consistent throughout the weekday.
- No users count were recorded during heavy rain or snow, suggesting these weather conditions are highly unfavorable. The highest rental count occurred when the weather was clear or partly cloudy.
- The number of rentals peaked in September but declined in December, likely due to heavy snowfall, which affected demand.
- Holidays saw a lower user count compared to regular days.
- The "Workingday" boxplot shows that maximum bookings range between 4000 and 6000, with the median count remaining stable throughout the week, indicating little difference in bookings between working and non-working days.



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)

If we have a categorical variable with n categories and create n dummy variables, the variables are **perfectly correlated**(i.e., one can be derived from the others).
We only need the n-1 variable for coefficient calculation . **drop_first=True** avoids multicollinearity

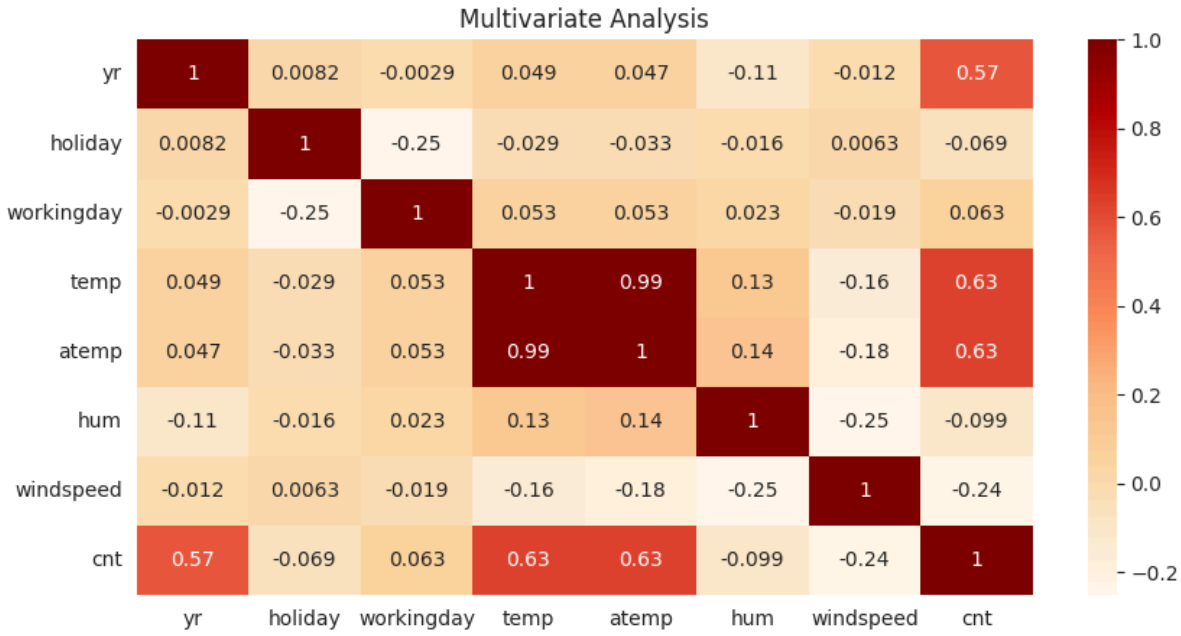
setting drop_first=True , helps us to have one less variable in our model and helps to reduce model complexity & overfitting.

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)

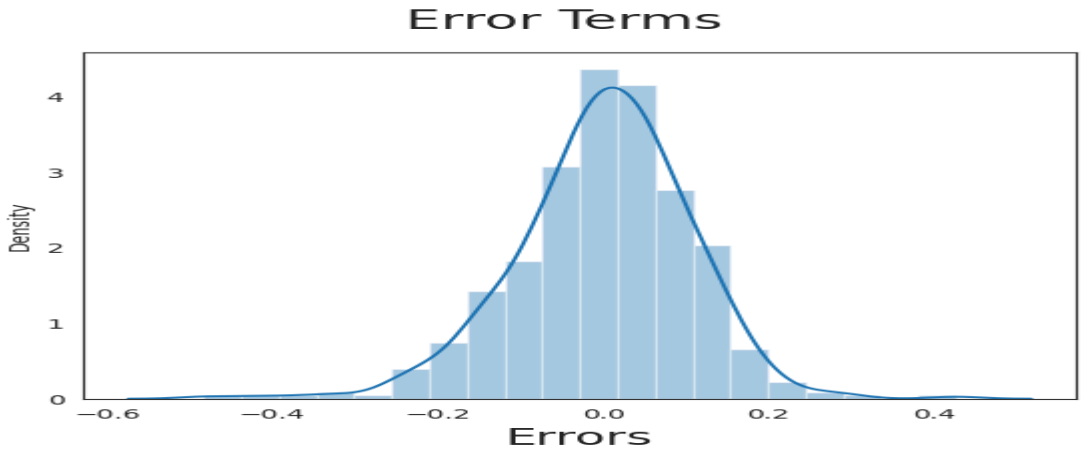
temp has the highest correlation (0.53) with the cnt target variable .



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)

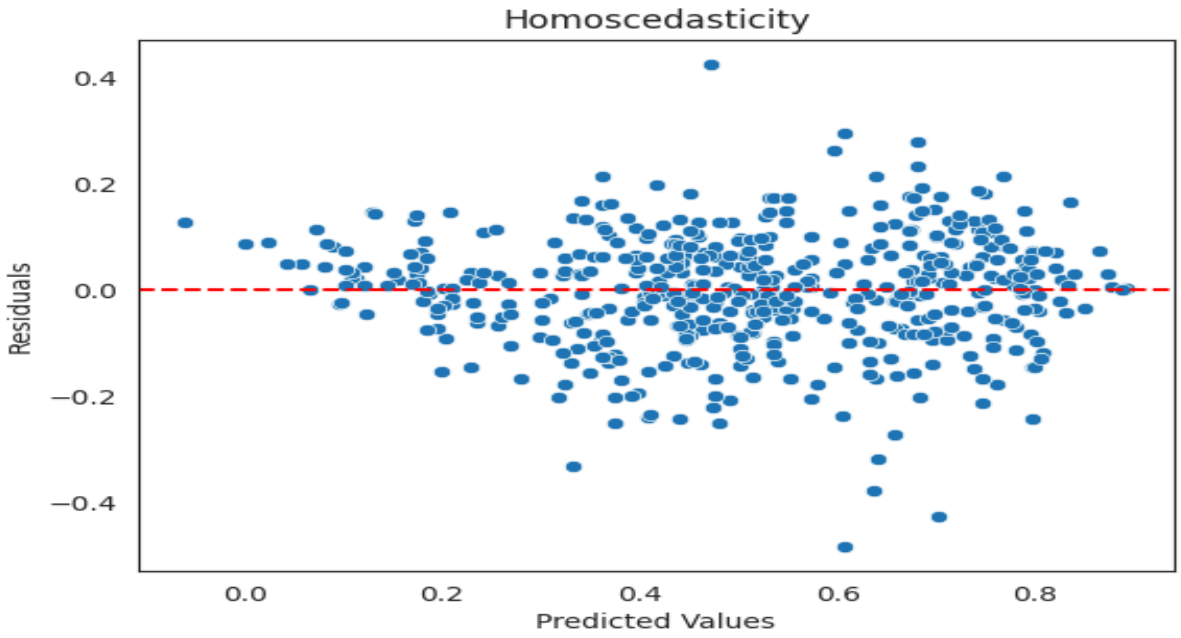
Following steps were used to validate the assumption of Linear Regression of the model.

1.Residual Analysis:
residual (difference between observed and predicted)



check - residual should be approximately distributed and there should not be any discernible pattern

2. **Homoscedasticity (constant variance):**



check : distribution is exclusively along the diagonal line

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)

yr - 0.25, weather_light_snow-rain , season_spring

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 fundamental machine learning algorithm used for predicting a continuous dependent variable (also called the output or target variable) based on one or more independent variables (predictors or features). The goal is to find the best-fitting straight line (or hyperplane in the case of multiple variables) that predicts the target variable as accurately as possible.

Linear regression is a supervised learning algorithm, which means that it is trained on labeled data (i.e., input-output pairs). It assumes a linear relationship between the input variables and the output.

Types of Linear Regression:

Simple Linear Regression – Involves one independent variable and one dependent variable.

The equation of a straight line is:

y=mx+c
Where:

- yy is the dependent variable (output),
- x is the independent variable (input),
- m (or β1) is the slope of the line (coefficient),
- cc (or β0) is the intercept (constant).

Multiple Linear Regression – Involves multiple independent variables and one dependent variable

The equation of a straight line is:

$$y = \beta 0 + \beta 1x1 + \beta 2x2 + \ldots + \beta nxn$$

Where:

- x1,x2,...,xnx1,x2,...,xn are independent variables,
- β1,β2,...,βnβ1,β2,...,βn are their respective coefficients,
- β0 is the intercept.

Objective of Linear Regression

The goal is to find the best-fitting line by minimizing the difference between predicted and actual values. This difference is measured using a cost function.

Cost Function (Mean Squared Error - MSE)

The Mean Squared Error (MSE) is used to measure the error in predictions:

MSE=1n∑i=1n(yi−y^i)2MSE=n1i=1∑n(yi−y^i)2

Where:

- yiyi is the actual value,
- y^iy^i is the predicted value,
- nn is the number of observations.

The objective is to minimize the MSE to find the best parameters (β).

Assumptions of Linear Regression

For linear regression to work effectively, the following assumptions must hold:

1. Linearity – The relationship between independent and dependent variables is linear.
2. Independence – Observations are independent of each other.
3. Homoscedasticity – Constant variance of residuals (errors).
4. Normality of Residuals – Errors should be normally distributed.
5. No Multicollinearity – Independent variables should not be highly correlated.

Evaluating Model Performance

After training the model, we evaluate its performance using:

1. R^2 (R-Squared Score) – Measures how well the independent variables explain the variance in the dependent variable.
   $R2 = 1 - SSres/SStot$
   Where:
   - SSres is the sum of squared residuals,
   - SStot is the total sum of squares.
2. Adjusted R^2 – Adjusts R^2 when adding more predictors.
3. RMSE (Root Mean Squared Error) – Square root of MSE.

Applications of Linear Regression

- Predicting house prices based on features like area, number of rooms, etc.
- Estimating sales revenue from advertising expenses.
- Forecasting stock prices based on historical data.
- Medical research, such as predicting disease risk from patient data.

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 identical simple descriptive statistics (mean, variance, correlation, etc.), yet they appear vastly different when graphed. This is to demonstrate the importance of visualizing data before making conclusions based solely on statistical metrics. Each dataset in the quartet has the following characteristics:

- Mean of x = 9 and mean of yy = 7.5.
- Variance of x = 11.0 and variance of yy = 4.12.
- Correlation between xx and yy = 0.82.
- Linear regression line for each dataset has a slope of approximately 0.5 and an intercept of around 3.

Anscombe's Quartet highlight

Importance of Data Visualization: While the descriptive statistics (mean, variance, correlation) may all look identical, the datasets vary widely in their structure and relationships. This shows that summary statistics can be misleading, and data visualization (such as scatter plots) is crucial for understanding the data.
Outliers Can Skew Interpretation: In the second dataset, the outlier affects the line of best fit, demonstrating how a single data point can distort statistical metrics.

Linear Regression Assumptions: The linear regression assumes a linear relationship, but datasets 3 and 4 reveal that the linear model may not always be the best fit, especially when relationships are quadratic or non-existent.
Correlation Does Not Imply Causation: The high correlation in all four datasets doesn't necessarily imply that one variable causes the other to change. For example, in Dataset 3, the relationship is clearly non-linear.

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 correlation coefficient is a statistical measure that describes the strength and direction of a linear relationship between two continuous variables

helps us with
Identifying Linear Relationships: It helps to determine if two variables are linearly related.
Predictive Modeling: In regression analysis, Pearson's R can help assess the relationship between independent and dependent variables.

The value of Pearson's R ranges from -1 to 1
Positive correlation – as one variable increases, the other variable increases proportionally.Indicates that as one variable increases, the other variable also increases

Negative correlation – as one variable increases, the other decreases proportionally.Indicates that as one variable increases, the other decreases

No correlation – non linear relationship between the variables.Weaker or no linear relationship

---

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 transforming numerical features in a dataset to a specific range or distribution. It ensures that features contribute equally to a regression model, preventing issues where large-scale features dominate smaller ones.

Advantage:

- Improves Model Performance – Regression algorithms perform better when features are on a similar scale.
- Prevents Dominance of Large Values – Some features may have much larger values than others, which can distort model training.
- Speeds Up Convergence – Models using gradient descent converge faster when features are properly scaled.
- Reduces Sensitivity to Measurement Units – Different units (e.g., kg vs. grams) can affect

model performance if not scaled properly.

Normalization (Min-Max Scaling):

- Transforms values to a specific range (usually 0 to 1, or -1 to 1).

  $Formula$: $X' = X - Xmin/Xmax - Xmin.$

- Sensitive to outliers (outliers can skew the min/max range).
- Used when data needs to be bounded within a fixed range (e.g., image processing, neural networks).
- Suitable when the data doesn't have a normal distribution.

Standardization (Z-score Scaling):

- Transforms values to have a mean of 0 and a standard deviation of 1.

  $X' = X - \mu/\sigma.$

- Less sensitive to outliers, but still impacted if outliers are extreme.
- Used when data follows a normal distribution (e.g., linear regression, PCA).
- Preferred for algorithms sensitive to the assumption of normally distributed data.

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)** is used to detect multicollinearity in regression models. A **VIF value can become infinite** (or extremely large) when there is **perfect multicollinearity** between independent variables.

VIF is calculated as:

$vif = 1/1 - R2$
where R2 is the coefficient of determination obtained by regressing one independent variable on the others.

- If there is perfect multicollinearity (i.e., one predictor is a perfect linear combination of others), then R2=1 hence vif become infinite

common scenario when it happen

1. Perfect Linear Dependence: If one variable is an exact multiple of another
2. Dummy Variable Trap: If dummy variables representing categories are not properly encoded
3. Redundant Features: If two or more features provide identical information (e.g., including both height in cm and height in meters).
4. Insufficient Data Points: If the number of observations is too small compared to the number of independent variables, leading to high correlation.

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>

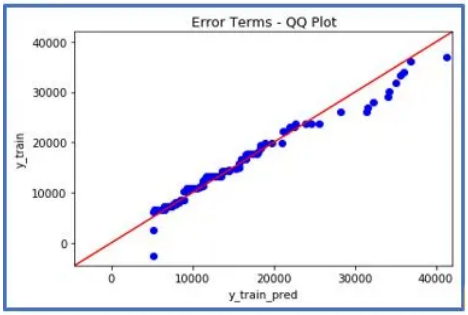
A Q-Q (Quantile-Quantile) Plot is used to assess whether a dataset follows a particular theoretical distribution, most commonly the normal distribution. It plots the quantiles of the sample data against the quantiles of the expected theoretical distribution.

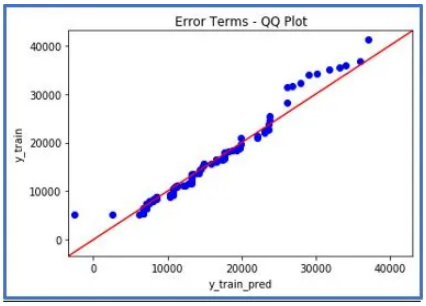
1. The x-axis represents the theoretical quantiles (from a normal distribution or another reference distribution).
2. The y-axis represents the actual quantiles from the dataset.
3. If the data follows the expected distribution, the points should roughly align along a straight diagonal line (45-degree line).

Q-Q Plot in Linear Regression

In linear regression, one key assumption is that the residuals (errors) are normally distributed. The Q-Q plot helps to visually check this assumption.

- If the residuals align closely to the 45-degree line, they are approximately normal, which validates the assumption.
- If the residuals deviate significantly, it indicates non-normality, which can affect statistical inference, such as confidence intervals and hypothesis testing.





Interpretations for two data sets.

a) Similar distribution: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x -axis
b) Y-values < X-values: If y-quantiles are lower than the x-quantiles.
c) X-values < Y-values: If x-quantiles are lower than the y-quantiles.
d) Different distribution: If all point of quantiles lies away from the straight line at an angle of 45 degree from x -axis

Importance of Q-Q Plot in Linear Regression

1. Checking Normality of Residuals: Ensures that the assumption of normality is met, which is required for accurate p-values and confidence intervals.
2. Detecting Skewness: If the points deviate systematically, it suggests skewness in the residuals.
3. Identifying Outliers: Extreme deviations from the line indicate potential outliers.
4. Improving Model Validity: If residuals are non-normal, transformations or alternative models (e.g., robust regression) may be required.

In summary, a Q-Q plot is a crucial diagnostic tool in regression analysis to check if the residuals are normally distributed and to determine whether assumptions are violated.