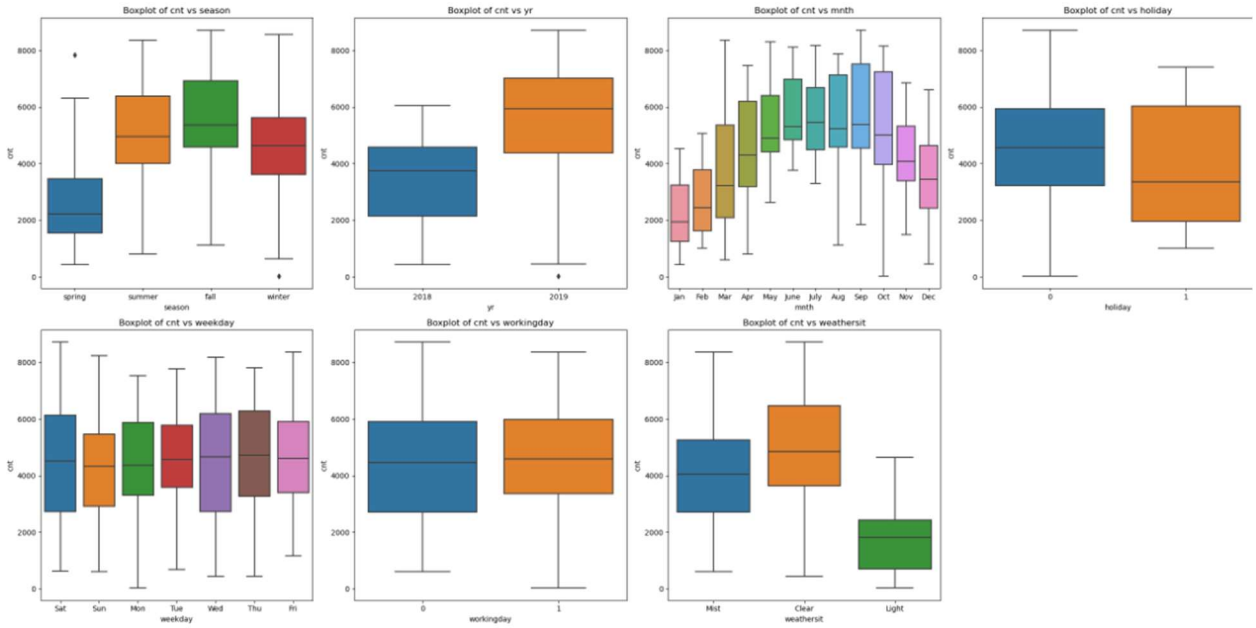
Assignment-based Subjective Questions

Question 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?



1.Seasonal Demand: The demand for shared bikes is highest in fall, followed by summer, as these seasons offer favorable riding conditions.

2.Yearly Increase: Bike demand was higher in 2019 than in 2018.

3.Monthly Trend: Demand peaks from May to September.

4.Holiday Effect: Bike rentals are higher on non-holidays.

5.Weekday Impact: Weekdays and working days have minimal impact on bike rental demand.

6.Weather Preference: Most rentals happen on days with clear skies or mild weather (few clouds or partly cloudy conditions).

7.Severe Weather: No bike rentals are recorded on days with heavy snow, or rainor fog.

Question 2. Why is it important to use drop_first=True during dummy variable creation?

If we use drop_first=True when creating dummy variables it helps for avoiding issues related to multicollinearity and for simplifying the interpretation of the model.

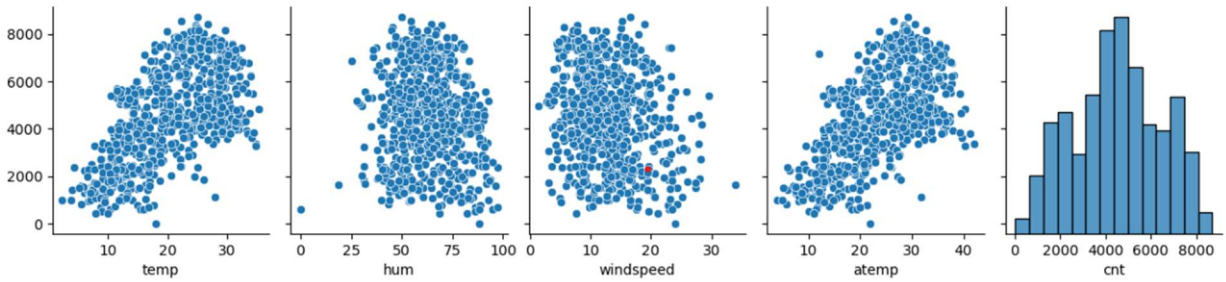
By setting drop_first=True, we drop the first dummy variable .This means the model will use the other categories to compare against the reference. This avoids multicollinearity and simplifies the model without losing any useful information.

For example, instead of creating dummy variables for all four seasons (Spring, Summer, Fall, Winter), we create dummies for only three of them (e.g., Spring, Summer, Fall), and use Winter

as the reference category. Now, the model knows that if the "Spring", "Summer", or "Fall" dummy is 1, the "Winter" is implicitly 0, and we avoid the redundant information.

Question 3.Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

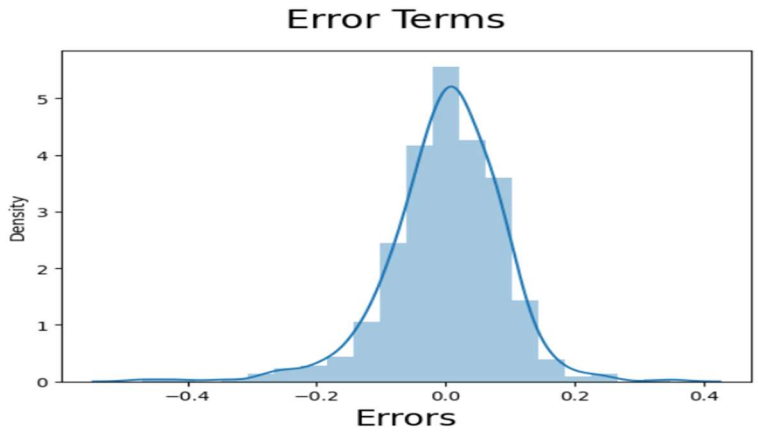
Temperature has highest the highest correlation with target variable (count)



Question 4.How did you validate the assumptions of Linear Regression after building the model on the training set?

1. Residual Analysis

- Take diff of actual and predicted values from training set.
- Plot the histogram of the diff values
- Residual should have normalized distribution.



2. Linearity

The relationship between the predictors and the target variable should remain linear.

Question 5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes

Top 3 Significant Features Contributing to Shared Bike Demand:

a.Temperature (temp) with a coefficient of 04201:

The positive coefficient for temperature indicates a strong, positive relationship with bike demand. As the temperature increases, the demand for shared bikes increases. This suggests that warmer weather encourages more people to rent bikes, likely because people are more willing to bike when the weather is favorable.

b. 2.Weather Situation:Light Snow/Rain(weathersit_Light) with a coefficient of -0.2742:

he negative coefficient for light snow, rain, or misty weather (represented by weathersit = 3) shows that adverse weather conditions reduce bike demand. People are less likely to rent bikes during rainy, snowy, or misty weather due to discomfort or potential safety concerns.

c. Year (yr) with a coefficient of 0.2373:

Interpretation: The positive coefficient for the year variable indicates that bike demand has increased significantly in 2019 compared to 2018.

General Subjective Questions

Question 6. Explain the linear regression algorithm in detail.

Regression is a supervised learning method where the machine learns to predict the value of a continuous variable (dependent variable) based on one or more input variables (independent variables).

A linear regression model tries to explain the relationship between the dependent variable (output variable) and independent variables (predictor variable) using a straight line.

- The main assumptions are:

  Linear dependence between the dependent and independent variables.

  Errors are normally distributed and centered around zero.

  No multicollinearity (i.e., no significant correlation between predictor variables).

Finding the Best Fit Line:

- The goal is to find the best fit line that minimizes the residuals and the Residual Sum of Squares (RSS).
  - Simple Linear Regression: Involves only one independent variable.
  - Multiple Linear Regression: Involves more than one independent variable.

Dataset Split:

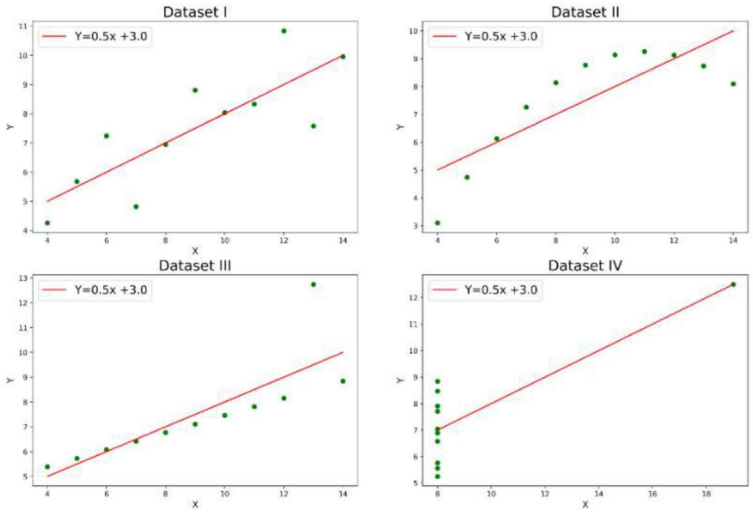
The dataset is usually divided into training and test sets (e.g., 70/30 or 80/20).

Training data is used to train the model and learn the relationships between variables.

Test data is used to evaluate the performance of the model and check its predictions on unseen data.

Question 7. Explain the Anscombe's quartet in detail?

Anscombe's Quartet is a set of four datasets created by the statistician Francis Anscombe in 1973. The datasets are specifically designed to demonstrate the importance of data visualization in understanding data, even when the summary statistics (like mean, variance, and correlation) are the same.



Dataset 1 looks like a straight line, where as x increases, y increases in a regular way. This follows the regression line.

Dataset 2 has a quadratic curve (a curve that goes up and down), so the relationship between x and y is not linear, even though the summary stats suggest it should be.

Dataset 3 has a strong outlier (a point that doesn't fit the trend of the rest of the data). This outlier is heavily affecting the summary statistics, making it look like there is a strong linear relationship, but the plot shows otherwise.

Dataset 4 has one outlier (the point at the end), which dramatically changes the regression line. The data appears to follow the same line as Dataset 1, but the outlier shifts the regression line and makes it misleading.

Summary statistics can be the same across different datasets, but the data can have very different distributions and patterns. Scatter plots or other visualizations help us see this clearly.

Question 8.What is Pearson's R?

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.

Range of Values:
Pearson's R ranges from -1 to 1, with the following interpretations:

- +1: Perfect positive linear correlation (as one variable increases, the other increases in a perfectly straight line).
- -1: Perfect negative linear correlation (as one variable increases, the other decreases in a perfectly straight line).
- 0: No linear correlation (the variables do not have any linear relationship).
- Between -1 and +1: A value closer to +1 or -1 indicates a stronger linear relationship, while a value closer to 0 indicates a weaker linear relationship.

Question 9. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Scaling refers to the process of transforming the features (variables) in your dataset into a specific range or distribution, making them easier to compare or interpret. The idea is to adjust the magnitude of the features so that they have a uniform scale.

Scaling is performed for several key reasons:

- I Better Model Performance:If one feature has a much larger range than others, the model may focus too much on that feature, causing biased or inaccurate predictions.
- Faster Convergence: When training models using algorithms like gradient descent, scaling helps the model learn more quickly. If the features are scaled similarly, it allows the model to treat them equally, speeding up the learning process and making the algorithm reach the optimal solution faster.

1. Types of Scaling

Two common methods of scaling are Normalization and Standardization.

---

1. Normalized Scaling (Min-Max Scaling)

Normalization or Min-Max Scaling scales the data to a fixed range, typically [0, 1]. It rescales each feature by subtracting the minimum value and dividing by the range (maximum - minimum)

2. Standardized Scaling (Z-score Scaling)

Standardization or Z-score Scaling transforms the data so that it has a mean of 0 and a standard deviation of 1. This is done by subtracting the mean and dividing by the standard deviation of the feature

Question 10. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

A VIF value becomes infinite when perfect multicollinearity is present, which means that one of the predictor variables is perfectly linearly related to one or more other predictor variables in the mode.

Question 11. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

A Q-Q plot (Quantile-Quantile plot) is a graphical tool used to assess if a dataset follows a specific theoretical distribution, typically the normal distribution. It compares the quantiles (i.e., specific points of data) of the observed data with the quantiles of a normal distribution.

Interpretation:

- Straight line: If the points on the Q-Q plot form a straight line, it suggests that the data is close to the theoretical distribution (usually normal).
- S-shaped curve: If the points bend upwards or downwards, it suggests that the data has heavy tails (more extreme values than a normal distribution) or is skewed.
- Outliers: Points that deviate significantly from the line represent outliers or observations that do not follow the expected distribution.