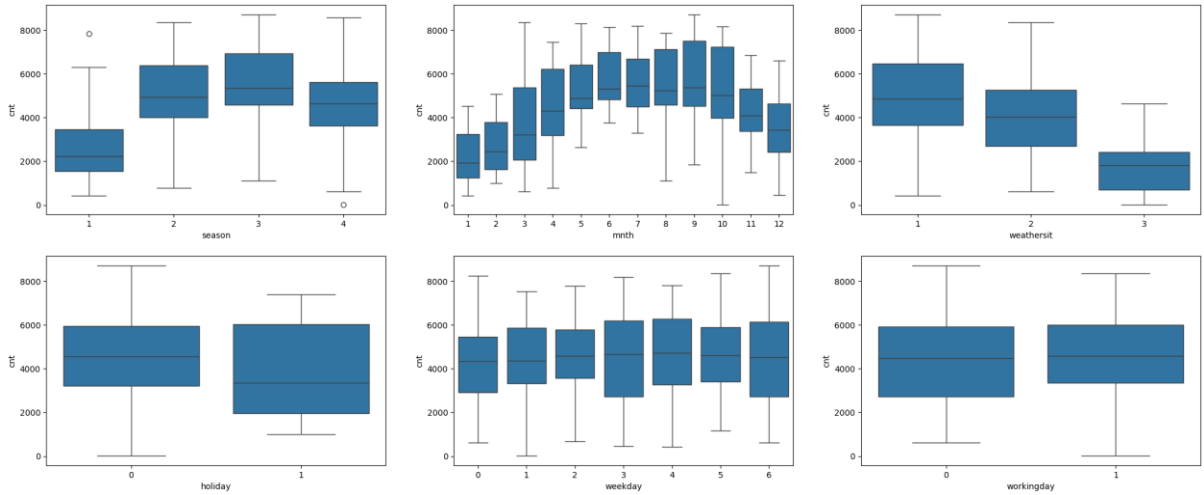
Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

It can be referred that the dependent variable cnt is significantly controlled by categorical variables like season, month, weather situation, type of days



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

When creating dummy variables from categorical features, it's often recommended to set the drop_first=True parameter in functions like pd.get_dummies(). This choice helps prevent multicollinearity and improves model interpretability.

Preventing Multicollinearity

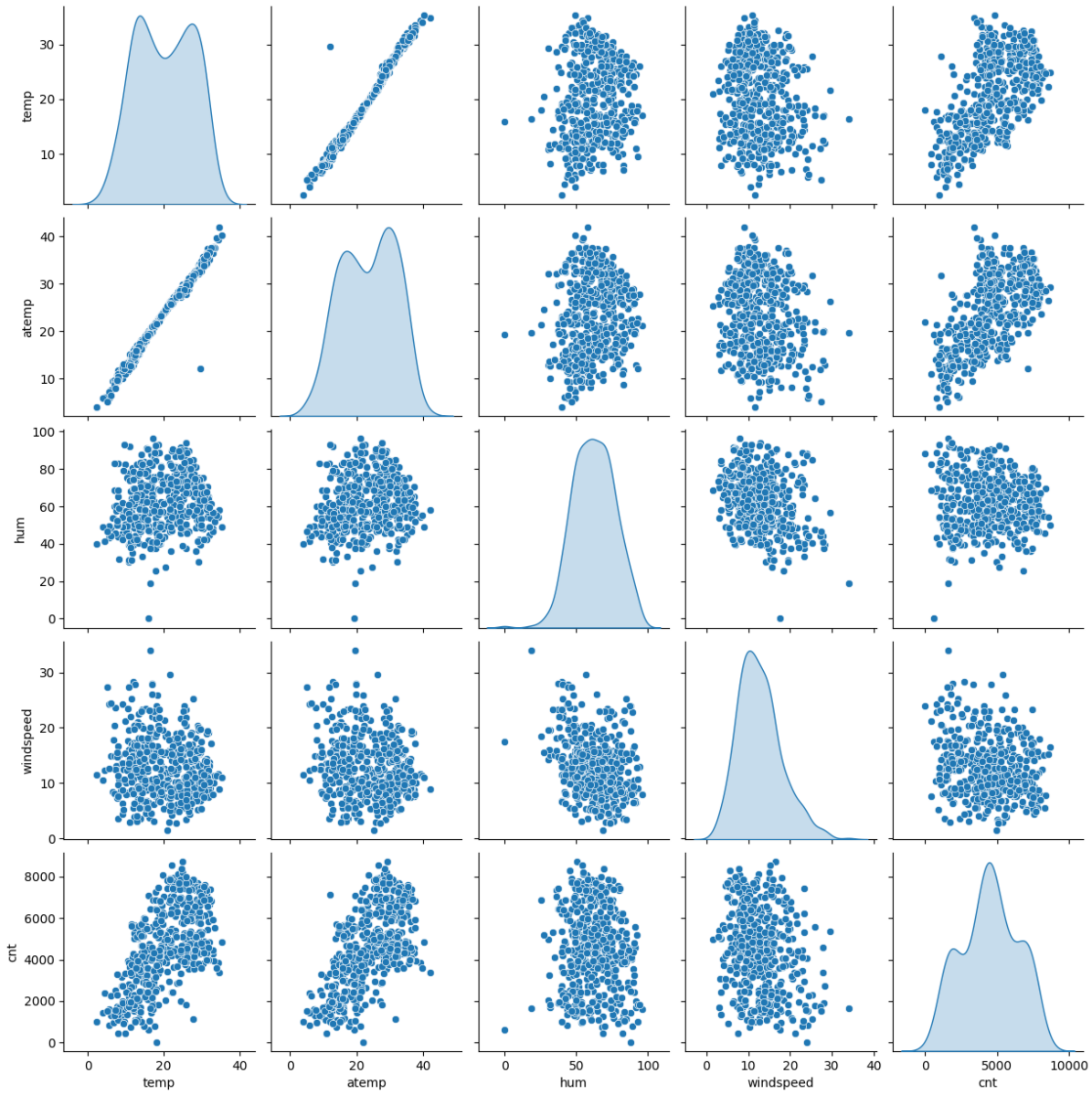
- **Redundancy:** Without dropping the first level, the dummy variables become redundant. For example, if a categorical feature has three levels (A, B, and C), creating three dummy variables (A, B, and C) would introduce redundancy. If A is 0 and B is 0, then C must be 1.
- **Multicollinearity:** This redundancy leads to multicollinearity, which can cause problems in statistical models. Multicollinearity can make it difficult to determine the individual impact of each predictor variable on the outcome.

Improving Interpretability

- **Reference Level:** By dropping the first level, one level becomes the reference level. The coefficients of the remaining dummy variables represent the difference in the outcome variable compared to this reference level. This makes the model's interpretation more straightforward.

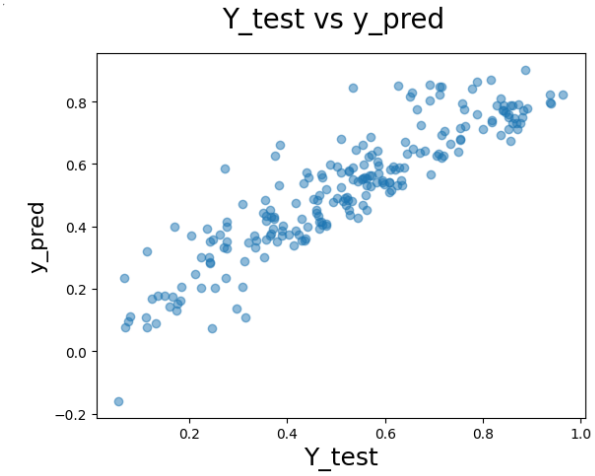
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Based on the pair plot, **temp** (temperature) appears to have the highest correlation with the target variable **cnt** (bike rentals). This is evident from the strong linear relationship shown in the scatter plot between these two variable

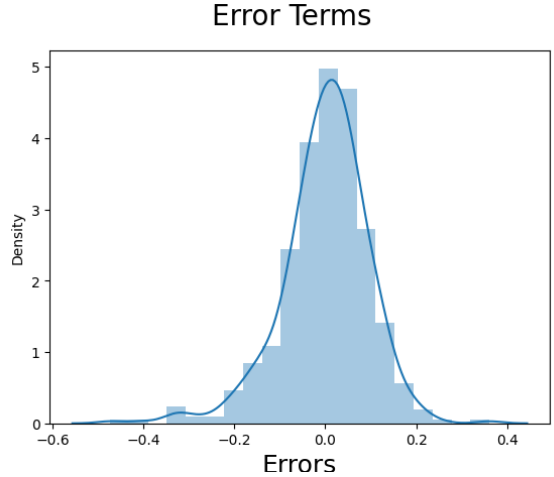


4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

- **Linearity:** The relationship between the independent and dependent variables is linear.



- **Homoscedasticity:** The variance of the residuals (errors) is constant.



- **No multicollinearity:** The independent variables are not perfectly correlated with each other.

	Features	VIF
3	season_Winter	2.31
1	temp	2.29
0	yr	2.02
6	mnth_nov	1.75
8	weathersit_Cloudy	1.50
2	season_Spring	1.42
4	mnth_dec	1.34
5	mnth_mar	1.21
7	mnth_sept	1.17
9	weathersit_Light Rain	1.05

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Here are the top 3 predictor variables and their coefficients:

weathersit_Light Rain: -0.329 temp: 0.415 yr: 0.231

Interpretation:

weathersit_Light Rain:

For each unit increase in this variable (i.e., more days with light rain), the dependent variable (cnt) is expected to decrease by 0.329 units, holding other variables constant. This suggests that light rain has a negative impact on cnt.

temp:

For each unit increase in temperature, cnt is expected to increase by 0.415 units, ceteris paribus. This indicates a positive relationship between temperature and cnt.

yr:

As the year increases by one unit, cnt is expected to increase by 0.231 units, controlling for other variables. This suggests a positive trend over time.

General Subjective Questions

1. **Explain the linear regression algorithm in detail. (4 marks)**

Linear regression is a statistical method used to model the relationship between a dependent variable (also known as the outcome variable or response variable) and one or more independent variables (also known as predictor variables or features). It assumes that the relationship between the variables is linear, meaning that the dependent variable can be expressed as a linear combination of the independent variables.

The Model Equation

The linear regression model is represented by the following equation:

$$y = \beta 0 + \beta 1x1 + \beta 2x2 + ... + \beta pxp + \varepsilon$$

where:

- **y:** is the dependent variable
- **β0:** is the intercept, representing the value of y when all independent variables are zero
- **β1, β2, ..., βp:** are the coefficients, representing the change in y for a unit increase in the corresponding independent variable, holding other variables constant
- **x1, x2, ..., xp:** are the independent variables
- **ε:** is the error term, representing the random variation that cannot be explained by the independent variables

The Goal of Linear Regression

The goal of linear regression is to estimate the coefficients (β0, β1, ..., βp) in a way that minimizes the sum of squared errors (SSE) between the predicted values and the actual values of the dependent variable. This is known as the **least squares method**.

Types of Linear Regression

There are two main types of linear regression:

- **Simple linear regression:** involves only one independent variable.
- **Multiple linear regression:** involves multiple independent variables.

Assumptions of Linear Regression

Linear regression makes several assumptions:

- **Linearity:** The relationship between the dependent and independent variables is linear.
- **Independence:** The observations are independent of each other.
- **Homoscedasticity:** The variance of the error term is constant across all values of the independent variables.
- **Normality:** The error term is normally distributed.

Applications of Linear Regression

Linear regression is widely used in various fields, including:

- **Statistics:** to model relationships between variables.
- **Economics:** to predict economic trends.
- **Finance:** to forecast stock prices.
- **Marketing:** to analyze the impact of marketing campaigns.
- **Engineering:** to model physical systems.

Limitations of Linear Regression

While linear regression is a powerful tool, it has some limitations:

- **Non-linear relationships:** If the relationship between the variables is non-linear, linear regression may not provide accurate predictions.
- **Outliers:** Outliers can have a significant impact on the results of linear regression.
- **Multicollinearity:** If the independent variables are highly correlated, it can make it difficult to interpret the results of linear regression.

2. Explain the Anscombe's quartet in detail. (3 marks)

Anscombe's quartet is a set of four small datasets that are statistically identical in terms of their mean, variance, correlation, and linear regression line. However, when visualized, they reveal strikingly different patterns, highlighting the importance of graphical data exploration.

The Four Datasets:

Dataset 1: A typical linear relationship with a positive correlation.

Dataset 2: A quadratic relationship with a positive correlation, but with a clear outlier that influences the regression line.

Dataset 3: A perfect linear relationship with a positive correlation, but with a constant x-value, making the relationship appear linear even though it's not.

Dataset 4: A linear relationship with a positive correlation, but with a vertical spread of points, indicating high variability.

3. What is Pearson's R? (3 marks)

Pearson's r is a statistical measure that quantifies the linear relationship between two numerical variables. It ranges from -1 to 1:

- **-1:** Indicates a perfect negative correlation, meaning that as one variable increases, the other decreases proportionally.
- **0:** Indicates no correlation between the variables.
- **1:** Indicates a perfect positive correlation, meaning that as one variable increases, the other increases proportionally.

```
r = (nΣxy - ΣxΣy) / sqrt((nΣx^2 - (Σx)^2)(nΣy^2 - (Σy)^2))
```

where:

- n is the number of data points
- Σx and Σy are the sums of x and y, respectively
- Σx^2 and Σy^2 are the sums of the squares of x and y, respectively
- Σxy is the sum of the product of x and y

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Scaling is a technique used in data preprocessing to transform numerical data to a common range or scale. This is often necessary because different features in a dataset may have vastly different ranges, which can skew the results of machine learning algorithms.
Scaling is performed to

- **Improve Algorithm Performance:** Many machine learning algorithms, especially distance-based algorithms like K-Nearest Neighbors or clustering algorithms, are sensitive to the scale of the data. Scaling can help these algorithms perform more accurately.
- **Prevent Domination:** Features with larger ranges can dominate the learning process, overshadowing the influence of other features. Scaling can help prevent this.
- **Regularization:** Some regularization techniques, like L1 and L2 regularization, work better when features are on a similar scale.

Difference between normalized scaling and standardized scaling:

- **Normalized Scaling (Min-Max Scaling):**
  - Rescales data to a specific range, typically between 0 and 1.
  - Formula: `(x - min(x)) / (max(x) - min(x))`
  - Preserves the relative distances between data points.
  - Sensitive to outliers.
- **Standardized Scaling (Z-score Standardization):**
  - Transforms data to have a mean of 0 and a standard deviation of 1.
  - Formula: `(x - mean(x)) / std(x)`

- Makes the data distribution more comparable across features.
- Less sensitive to outliers.

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

(3 marks)

When VIF is infinite, it indicates a perfect multicollinearity problem. This means that one of the independent variables can be perfectly predicted by a linear combination of the other independent variables. In such a case, the regression model becomes unstable and the coefficients of the correlated variables become indeterminate.

Here are some common reasons why VIF can be infinite:

- **Dummy Variable Trap:** When creating dummy variables for a categorical variable with k levels, it's essential to drop one of the categories to avoid perfect multicollinearity. If all k categories are included, the dummy variables will be perfectly correlated, leading to an infinite VIF.
- **Redundant Features:** If two or more features in the dataset are essentially the same or highly correlated, they can cause perfect multicollinearity. For example, including both "age" and "years of experience" might lead to redundancy.
- **Mathematical Errors:** In some cases, errors in data preparation or model specification can lead to perfect multicollinearity.

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

(3 marks)

What is a Q-Q plot?

A Q-Q plot (Quantile-Quantile plot) is a graphical technique used to compare the distribution of a sample data set to a theoretical distribution, typically the normal distribution. It plots the quantiles of the observed data against the quantiles of the theoretical distribution.

Use and Importance in Linear Regression

In linear regression, one of the key assumptions is that the residuals (errors) are normally distributed. A Q-Q plot is a valuable tool for assessing this assumption.

How to Interpret a Q-Q Plot:

- **Straight Line:** If the points on a Q-Q plot fall close to a straight line, it suggests that the observed data follows the theoretical distribution (in this case, the normal distribution).
- **Deviations from the Line:** Deviations from the straight line indicate departures from normality. For example, if the points in the tails of the plot deviate from the line, it might suggest a heavy-tailed distribution.

Importance of Q-Q Plots in Linear Regression:

1. **Assumption Check:** As mentioned, normality of residuals is a crucial assumption in linear regression. A Q-Q plot provides a visual check for this assumption.
2. **Identifying Outliers:** Outliers can significantly affect the normality of residuals. Q-Q plots can help identify outliers that might be skewing the distribution.
3. **Understanding Model Fit:** If the residuals are not normally distributed, it might indicate a problem with the model's specification or the underlying assumptions.
4. **Improving Model Performance:** By identifying deviations from normality, you can explore transformations or alternative models that might better fit the data.