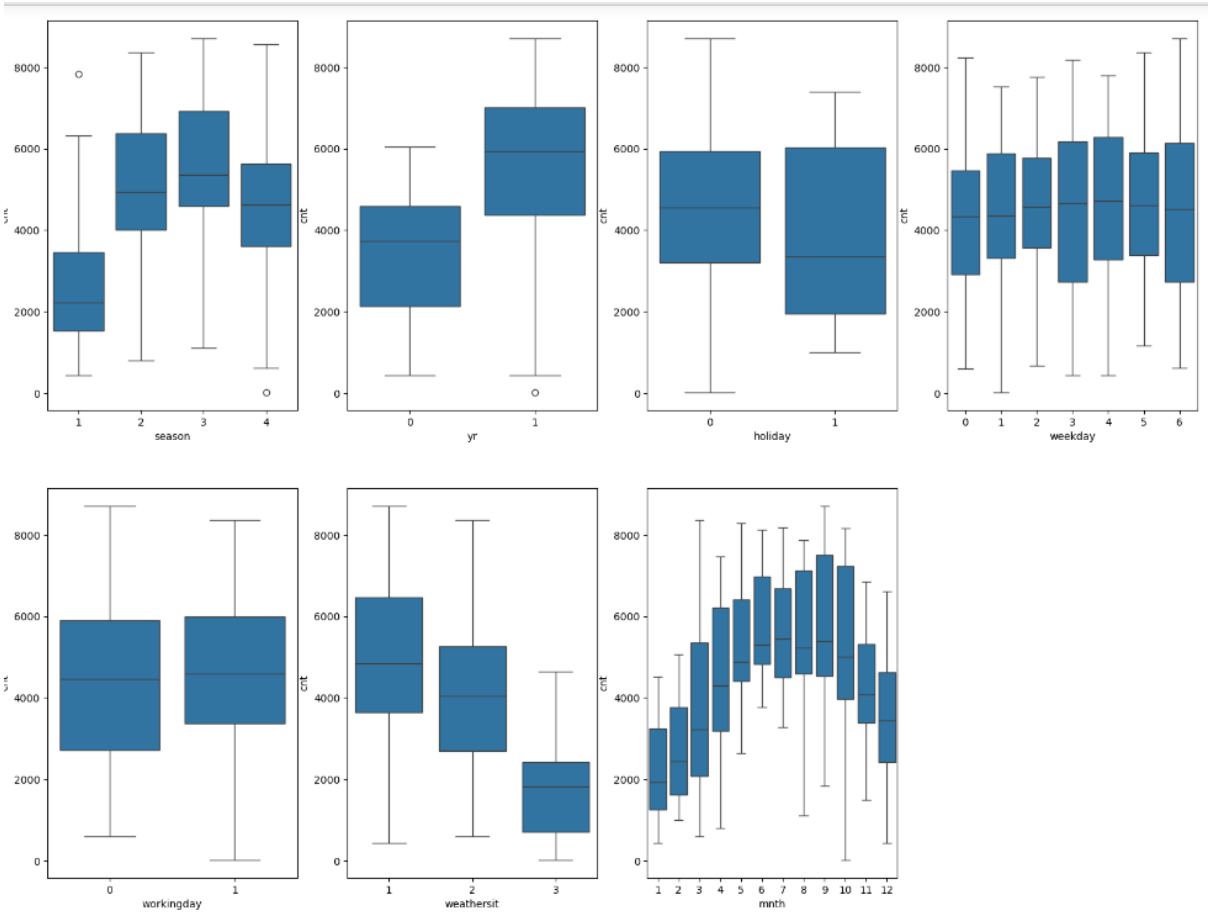
Assignment-based Subjective Questions

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



The dataset's categorical variables were analyzed using boxplots:

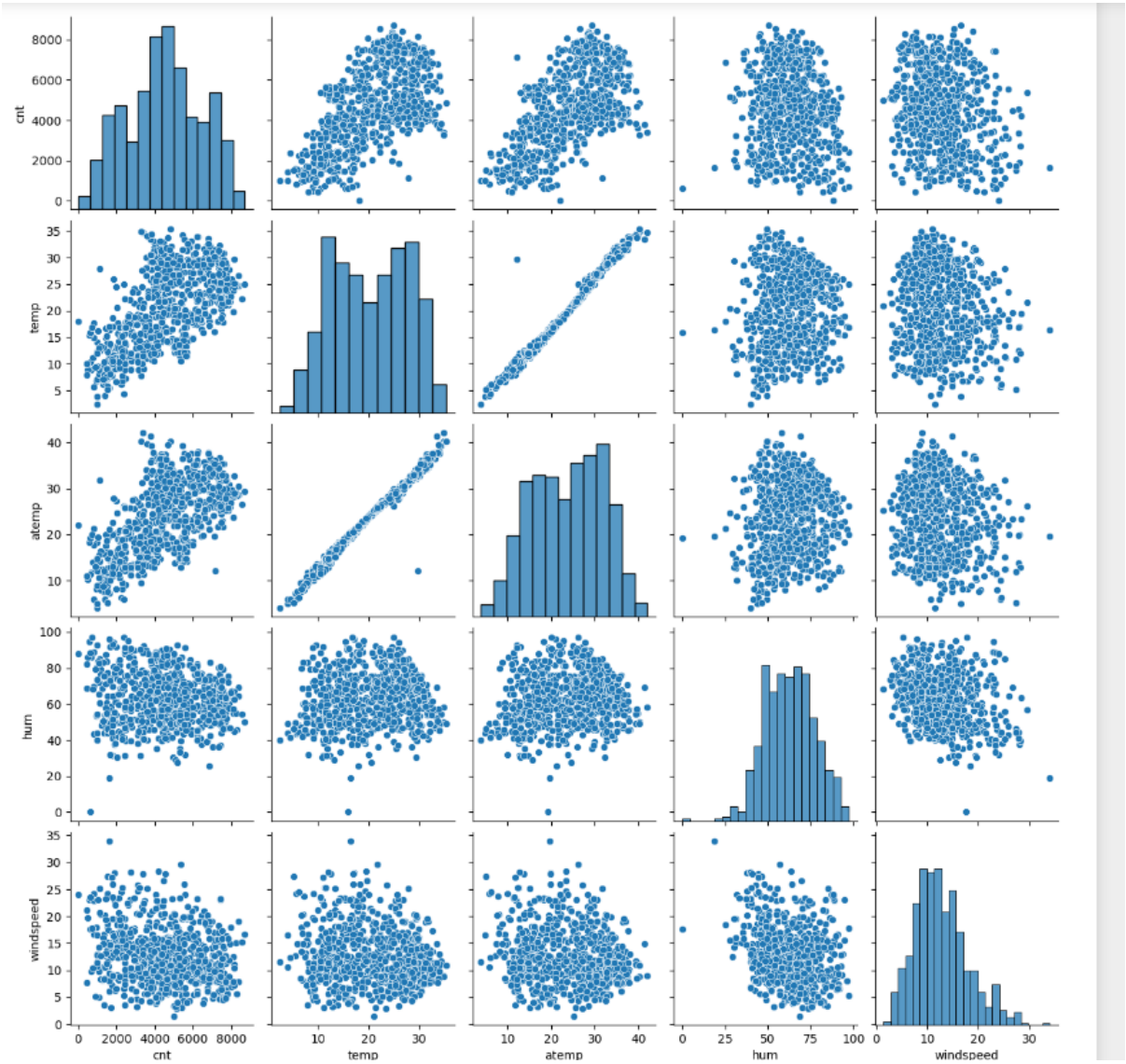
Season: Highest demand in Fall; lowest in Spring.
Year: Higher user count in 2019 than 2018.
Holiday: Rentals decreased.
Weekday: Demand was constant.
Working Day: Consistent bookings; little difference between working and non-working days.
Weather Situation: No rentals in heavy rain/snow; highest in clear/partly cloudy weather.
Month: Peak rentals in September; lowest in December due to snowfall.

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

Using drop_first=True in dummy variable creation is important as it reduces the extra column, thereby minimizing correlations among dummy variables. For a categorical variable with n levels, only n−1 columns are needed to represent the dummy variables.

For example, for a categorical column with three values (furnished, semi_furnished, unfurnished), creating dummy variables for furnished and semi_furnished is sufficient. If a variable is neither of these, it is implied to be unfurnished, so the third variable is unnecessary.

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



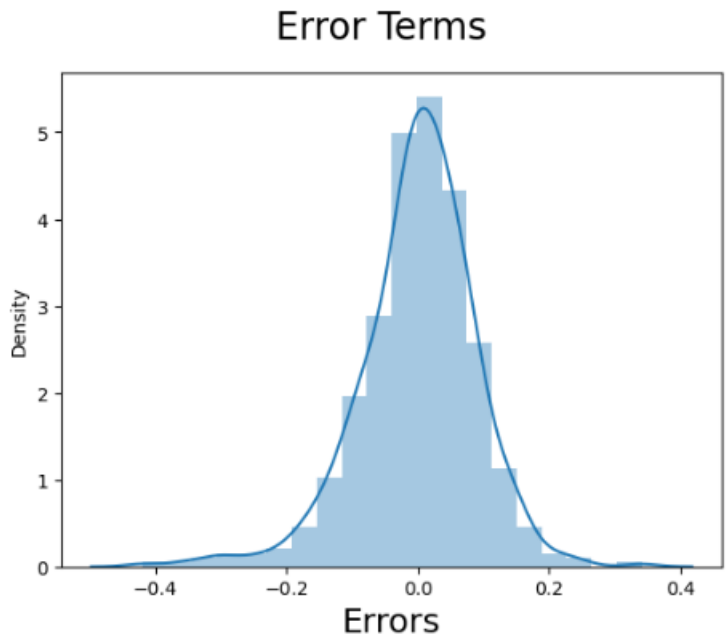
Column “temp” is highly correlated with “cnt”.
Column “atemp” is highly correlated with “cnt”

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

We validated the assumptions of Linear Regression through the following tests:

a. Linearity: Verified by visualizing numeric variables with a pairplot to check for linear relationships between independent and dependent variables.

b. Normal Distribution of Residuals: Checked by plotting a distplot of residuals to ensure they follow a normal distribution and are centered around zero.



c. Multicollinearity: Assessed by calculating the Variance Inflation Factor (VIF) to ensure independent variables are not too highly correlated.

	Features	VIF
11	temp	5.18
12	windspeed	4.63
1	season_Summer	2.24
0	season_Spring	2.13
9	yr	2.07
2	season_Winter	1.84
3	mnth_Jul	1.59
8	weathersit_Mist & Cloudy	1.56
4	mnth_Sep	1.34
5	weekday_Saturday	1.23
6	weekday_Sunday	1.22
7	weathersit_Light Snow & Rain	1.08
10	holiday	1.06

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

The top 3 features contributing significantly towards explaining the demand of the shared bikes are

1. temp 0.491531
2. yr 0.233727
3. weathersit_Light Snow & Rain -0.289513

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 (Y) and one or more independent variables (X). The primary goal is to predict the dependent variable based on the independent variables.

Simple Linear Regression: When there is only one independent variable, the relationship is modeled using the equation $Y=\beta0+\beta1X+\epsilon Y$ where:

- Y is the dependent variable.
- $\beta0$ is the intercept.
- $\beta1$ is the slope of the line.
- X is the independent variable.
- ϵ is the error term.

- Multiple Linear Regression: When there are multiple independent variables, the equation extends to $Y=\beta0+\beta1X1+\beta2X2+\ldots+\beta nXn+\epsilon Y$

  The algorithm involves the following steps:

1. Assumptions: Assumes a linear relationship between X and Y, independence of errors, homoscedasticity (constant variance of errors), normality of error terms, and no multicollinearity.
2. Estimation of Coefficients: Uses the method of least squares to find the best-fitting line by minimizing the sum of squared residuals (differences between observed and predicted values).
3. Model Evaluation: Assesses the model using metrics like R-squared, Adjusted R-squared, and p-values to determine the goodness-of-fit.
4. Prediction: Uses the estimated coefficients to predict the dependent variable for given values of independent variables.

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

Anscombe's quartet is a collection of four datasets that have nearly identical simple descriptive statistics (mean, variance, correlation, and linear regression line) but differ significantly when graphed. Created by Francis Anscombe in 1973, the quartet demonstrates the importance of graphing data before analyzing it.

- Dataset 1: Appears as a typical linear relationship.
- Dataset 2: Forms a clear curve, illustrating non-linearity.
- Dataset 3: Contains a single outlier influencing the linear trend.
- Dataset 4: Consists of vertical data points with one horizontal outlier, misleading the correlation and regression line.

These datasets underscore that numerical summaries alone can be deceptive, and visual examination is crucial for proper data analysis.

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

Pearson's R, or Pearson's correlation coefficient, is a measure of the linear correlation between two variables X and Y. It ranges from -1 to 1, where:

- 1 indicates a perfect positive linear relationship.
- -1 indicates a perfect negative linear relationship.
- 0 indicates no linear relationship.

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

Scaling refers to the process of transforming the features of a dataset to a similar scale, which is essential for algorithms that compute distances between data points, like k-nearest neighbors or gradient descent in linear regression.

- Why Scaling is Performed:
    1. To improve the performance and convergence of machine learning algorithms.
    2. To ensure that each feature contributes equally to the model.
    3. To prevent features with larger ranges from dominating the learning process.

- Normalized Scaling (Min-Max Scaling): Rescales the data to a fixed range, typically [0, 1]. The formula is: $X' = (X - Xmin)/(Xmax - Xmin)$
- Standardized Scaling (Z-score Scaling): Centres the data around the mean with a unit standard deviation. The formula is: $X' = (X - \mu)/\sigma$ where μ is the mean and σ is the standard deviation.

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

(3 marks)

A VIF (Variance Inflation Factor) value becomes infinite when there is perfect multicollinearity among the independent variables, meaning that one variable is a perfect linear combination of others. This leads to the denominator of the VIF calculation (1 - R-squared) becoming zero, causing the VIF to be undefined or infinite. This indicates that the regression model cannot estimate the coefficients due to exact collinearity.

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

(3 marks)

A Q-Q (Quantile-Quantile) plot is a graphical tool to compare the distribution of a dataset to a theoretical distribution, typically the normal distribution. It plots the quantiles of the data against the quantiles of the theoretical distribution.

- Use in Linear Regression:
    1. To assess whether the residuals of a regression model follow a normal distribution.
    2. Residuals should align closely with the reference line in a Q-Q plot if they are normally distributed.
- Importance:

1. Helps validate the assumption of normality of residuals, which is critical for hypothesis testing and constructing confidence intervals in linear regression.
2. Identifies deviations from normality, such as skewness or kurtosis, which can indicate model misspecification or the presence of outliers.