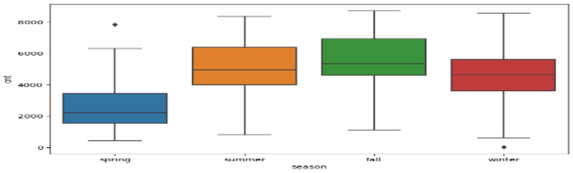
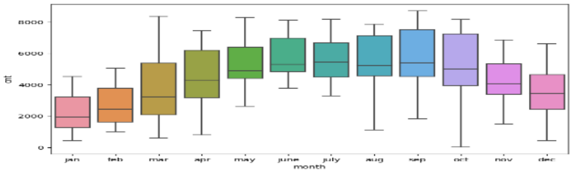
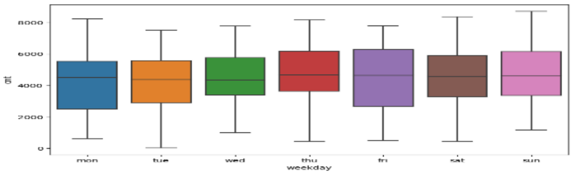
Assignment-based Subjective Questions

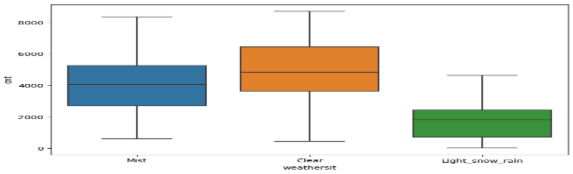
1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

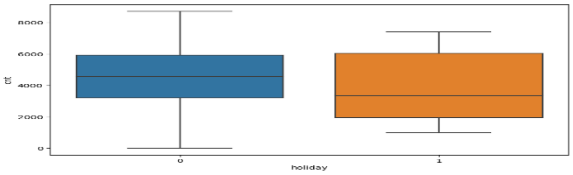
Ans:

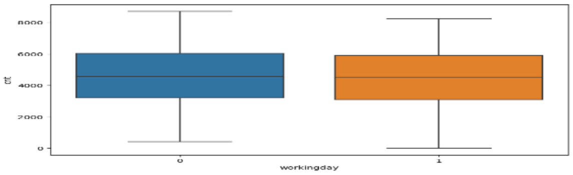


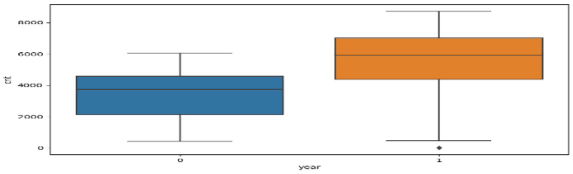


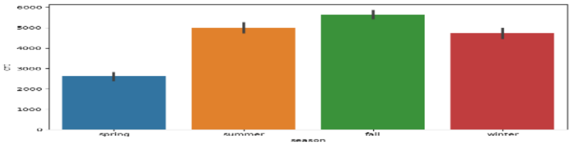


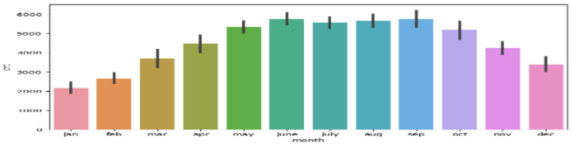


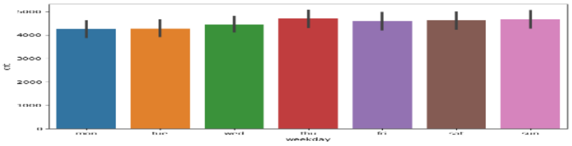


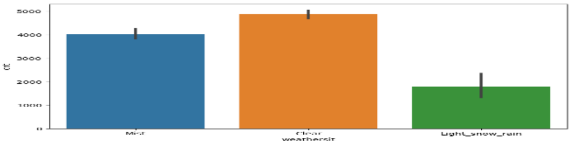


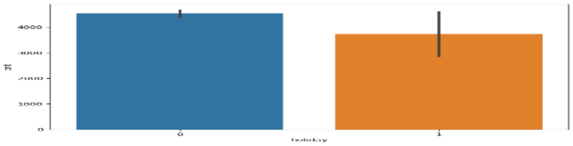


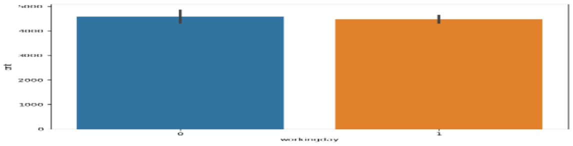


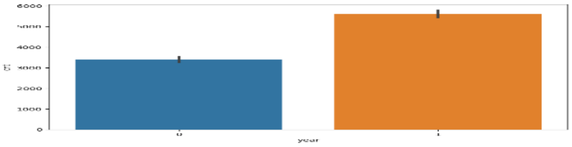












From plots we can draw many insights such as:

- The plots reveal that fall Season experiences the highest demand for rental bikes, with elevated median values during summer and fall indicating a surge in user numbers during these seasons.

- The number of users shows a gradual rise from January to July, with rental bike demand steadily increasing each month until June.
- September experiences the peak demand, followed by a decline in subsequent months.
- Demand remains relatively stable on weekdays and working days with minimal variation.
- Clear weathersit conditions correspond to the highest demand for rental bikes.
- Demand decreases on holidays.
- The year 2019 shows a significant rise in the number of users.

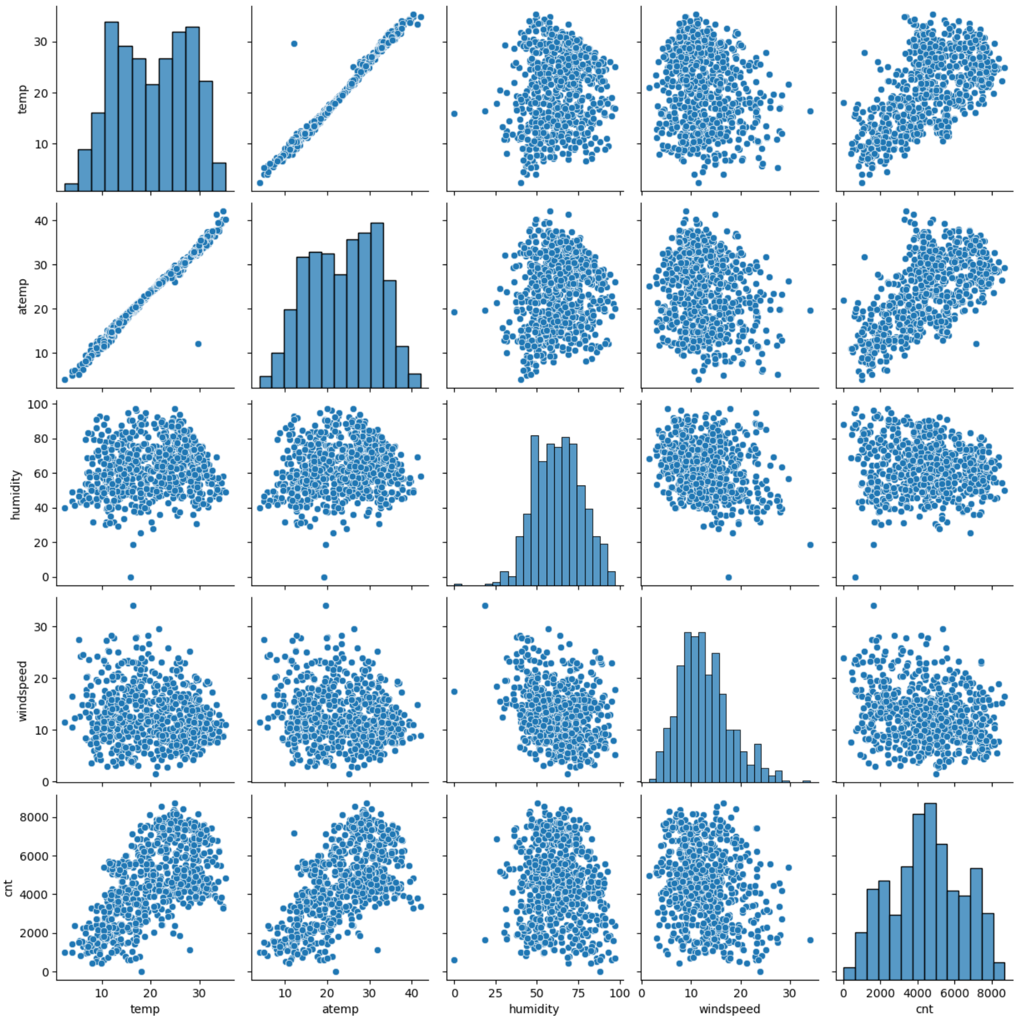
2. Why is it important to use **drop_first=True** during dummy variable creation? (2 mark)

Ans: Using drop_first=True is crucial as it eliminates the extra column generated during dummy variable creation, effectively reducing correlations among dummy variables. When dealing with a categorical variable with n levels, it's necessary to use n-1 columns to represent the dummy variables.

For instance, if we have a categorical column with 3 values and want to create dummy variables for it, if one variable is neither furnished nor semi-furnished, it is implicitly unfurnished. In this case, we don't need the third variable to identify the unfurnished category.

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

   Ans: The variable 'temp' exhibits the strongest correlation with the target variable.



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

Ans: I have validated the assumptions of the Linear Regression Model based on the following four criteria:

- Normality of Error Terms: Ensuring that the error terms are normally distributed.
- Multicollinearity Check: Verifying that there is no significant multicollinearity among variables.
- Linear Relationship Validation: Confirming the presence of linearity among variables.
- Homoscedasticity: Checking for the absence of discernible patterns in residual values. Independence of Residuals: Ensuring there is no autocorrelation present in the residuals.

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

Ans: The most influential factors in explaining the demand for shared bikes are:

- Temperature
- Year
- September

General Subjective Questions

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

Ans: Linear regression, a predictive modeling method, explores relationships between a dependent variable (target) and independent variables (predictors). It discerns how changes in independent variable values correspond to variations in the dependent variable, using either simple linear regression (single input) or multiple linear regression (multiple inputs). The goal is to establish a best-fit straight line, reflecting positive or negative linear relationships. This process involves finding optimal values (a0 and a1) by minimizing errors through techniques like RFE, MSE, or cost functions. The approach assumes a linear connection between predictors and outcomes. It entails data preparation with input (X) and output (Y) variables, forming the hypothesis equation $Y = \beta 0 + \beta 1 * X + \epsilon$. The objective is to minimize the difference between observed and predicted values, commonly achieved using methods like Ordinary Least Squares (OLS) or Gradient Descent. Model evaluation employs metrics like R-squared (R^2) and Mean Squared Error (MSE), ensuring a good fit. Linear regression's key assumptions include linearity, independence of observations, homoscedasticity, normality of residuals, and absence of perfect multicollinearity among predictors. Once trained, the model predicts outcomes for new data, serving as a foundational tool in statistics and machine learning.

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

Ans: Anscombe's quartet comprises four small datasets with nearly identical simple descriptive statistics but vastly different distributions and visual appearances. These datasets emphasize the crucial practice of visualizing data before analysis, revealing the limitations of relying solely on summary statistics.

Description of Datasets:

1. Dataset I:

- X: 10 distinct x-values ranging from 4 to 14.
- Y: Corresponding y-values form a linear relationship ($Y = 3 + 0.5X$).

2.Dataset II:

- X: Same as Dataset I.
- Y: Similar to Dataset I but exhibits a slight curve, indicating a non-linear relationship.

3.Dataset III:

- X: Same as Dataset I.
- Y: Comprises two distinct groups, creating a step function.

4. Dataset IV:

- X: Eight distinct x-values, some overlapping with Datasets I, II, and III.
- Y: Contains one outlier significantly deviating from the linear trend observed in Dataset I.

Implications and Significance:

- Visual Representation: Despite identical mean, variance, correlation, and regression line parameters, these datasets appear drastically different when graphed.
- Statistical Analysis Pitfalls: Relying solely on summary statistics can lead to erroneous interpretations about the underlying data patterns.
- Importance of Data Visualization: Anscombe's quartet highlights the critical role of data visualization in understanding inherent patterns, relationships, and outliers.
- Educational Tool: Often used in statistics education, it emphasizes the necessity of graphically exploring data before applying quantitative analysis methods.

In summary, Anscombe's quartet serves as a potent reminder that statistics can be deceptive without proper visualization, emphasizing the significance of graphical exploration to gain profound insights into data characteristics.

3. What is Pearson’s R? (3 marks)

Ans: Pearson's correlation coefficient (r) is a statistical metric gauging the strength and direction of the linear relationship between two variables. It quantifies how well a straight line can depict their relationship, ranging from -1 to 1. A value of 1 signifies a perfect positive correlation (both variables increase proportionally), -1 indicates a perfect negative correlation (one variable increases, the other decreases proportionally), and 0 suggests no linear relationship. The formula for sample correlation (r) calculates this relationship, utilizing individual data points (Xi, Yi) and means (X̄, Ȳ) of the variables. Widely used in fields like statistics, biology, and economics, Pearson's (r) specifically captures linear relationships and may not encompass nonlinear associations between variables.

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

Ans: Scaling, a crucial step in many machine learning algorithms, involves transforming numerical variables to a standard scale. This process ensures uniformity among features. The necessity for scaling arises from several factors:

Reasons for Scaling:

- Magnitude: Variables with diverse magnitudes could bias predictions, making scaling essential to avoid this bias.
- Algorithm Sensitivity: Numerous machine learning algorithms, especially distance-based ones like k-nearest neighbors, are sensitive to feature scales.
- Convergence: Optimization algorithms like gradient descent converge faster when features are scaled, ensuring quicker arrival at the optimal solution.

Normalized Scaling vs. Standardized Scaling:

1. **Normalized Scaling (Min-Max Scaling):**

- Formula: X_norm = (X - X_min)/(Xmax - Xmin)
- Range: Transforms data to a specific range, usually [0, 1].
- Advantage: Maintains the original distribution shape while constraining values within a defined range.

2. **Standardized Scaling (Z-score Standardization):**

  - Formula:X_std = (X - μ)/ σ
  - Range: Centers data around 0 with a standard deviation of 1.
  - Advantage: Particularly useful for data with outliers or a normal distribution, making data comparable after transformation.

Key Differences:

- Range: Normalized scaling confines data to a specific range, whereas standardized scaling centers data around zero with a standard deviation of 1.
- Outlier Sensitivity: Standardized scaling is less affected by outliers due to its reliance on mean and standard deviation, which are less influenced by extreme values.
- Interpretability: Normalized scaling retains the original data scale, enhancing interpretability, while standardized scaling standardizes data, simplifying comparisons.

- The choice between normalized and standardized scaling hinges on the dataset's characteristics and the specific requirements of the machine learning algorithm in use.

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

(3 marks)

Ans: The Variance Inflation Factor (VIF) gauges how much the variance of a regression coefficient increases due to correlated predictors. VIF values exceeding 10 or 5 often indicate multicollinearity, signifying high correlation between predictors. However, an infinite VIF signals perfect multicollinearity, where two or more independent variables are perfectly correlated, making one predictable from the others.

Perfect multicollinearity creates computational challenges as the matrix used in regression coefficient calculations becomes singular, lacking an inverse. Consequently, accurate coefficient estimation becomes impossible, leading to infinite VIF values. To manage perfect multicollinearity, identifying the problematic variables is crucial. Solutions include removing one variable, combining them to create a new one, or employing techniques like Ridge regression, which handles multicollinearity by penalizing large coefficients.

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

(3 marks)

Ans: A Q-Q plot is a graph used to see if data follows a specific pattern. In linear regression, it helps in several ways:

Checking if Data is Normal:

- Why: We need data to be normal for reliable results.
- How: Q-Q plots visually show if our data looks like a typical bell curve.

Spotting Outliers:

- Why: Outliers can mess up our predictions.
- How: Q-Q plots highlight unusual points, helping us see and handle them.

Understanding Data Shape:

- Why: Unusual shapes affect our predictions too.
- How: Q-Q plots reveal if our data is stretched or squished in odd ways.

Comparing Data Sets:

- Why: Sometimes we need to compare different sets of data.
- How: Q-Q plots show how similar or different two sets of data are.

Checking Prediction Assumptions:

- Why: We want to be sure our predictions are reliable.
- How: Q-Q plots help confirm if our predictions follow the expected pattern.

In short, Q-Q plots are like visual helpers in understanding our data better, ensuring our predictions are accurate, and guiding us in making the right decisions in our studies or analyses.