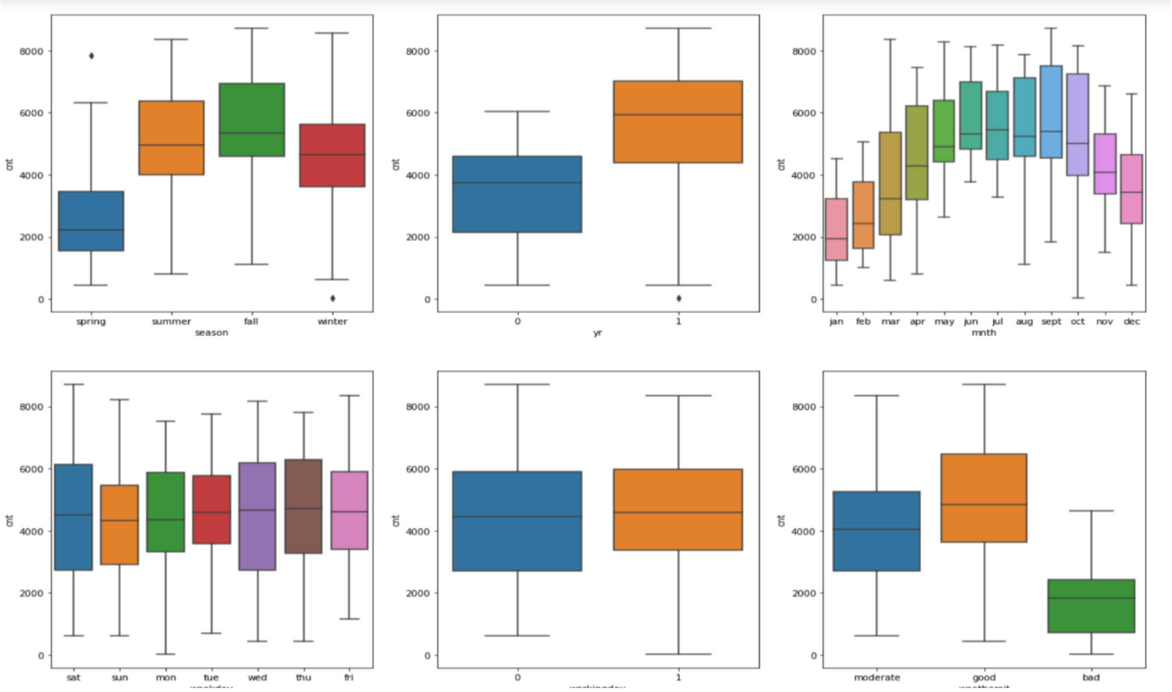
Assignment-Based Subjective Questions

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

   Several categorical variables, such as season, month, year, weekday, working day, and weather, significantly impact the dependent variable 'cnt.' The correlation among these variables is depicted in the following figure.

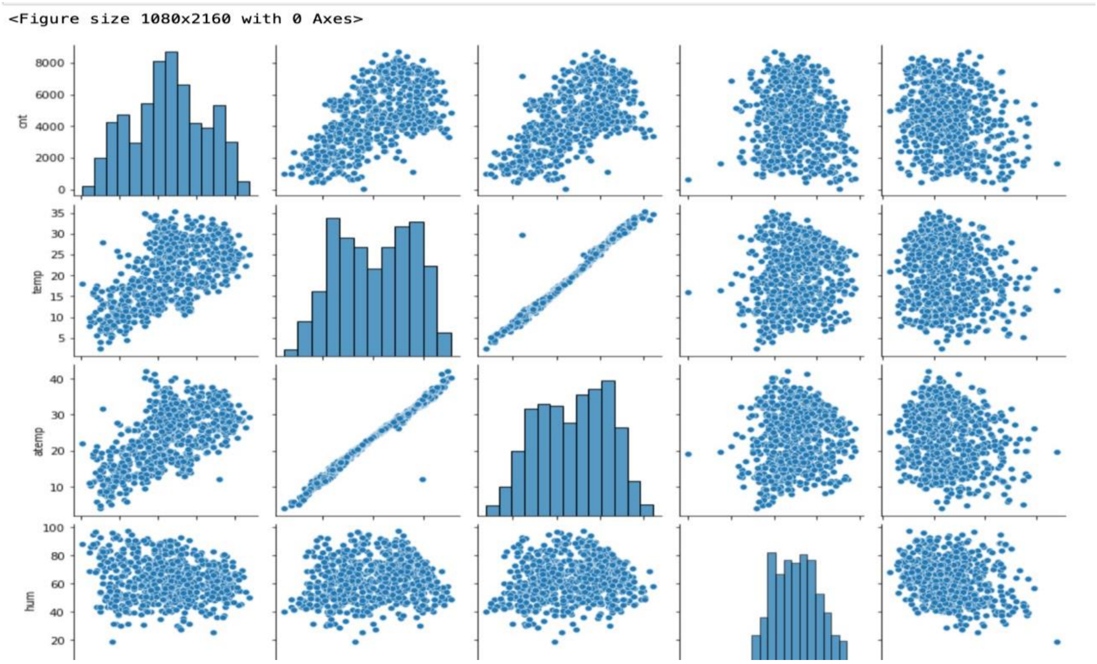


These variables are visualized using bar plot and Box plot both.

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

   Dummy variables are created for categorical variables with 'n' levels, forming 'n-1' new columns (0 or 1) to indicate the presence of each level. Setting drop_first=True ensures 'n-1' columns align with the levels, reducing correlation among dummy variables. For instance, if there are 3 levels, drop_first will exclude the first column.

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



```
<Figure size 1080x2160 with 0 Axes>
```

Among the variables, 'temp' and 'atemp' show the highest correlation with the target variable 'cnt.'

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

Linear regression models are validated using several criteria: Linearity, which checks if the relationship between variables is linear; No autocorrelation, ensuring errors are independent; Normality of errors, verifying if they follow a normal distribution; Homoscedasticity, ensuring consistent variability of errors; and Multicollinearity, assessing independence among predictor variables.

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

The three most significant features explaining shared bike demand are temperature, year, and season. These factors have a substantial impact on understanding and predicting the demand for shared bikes.

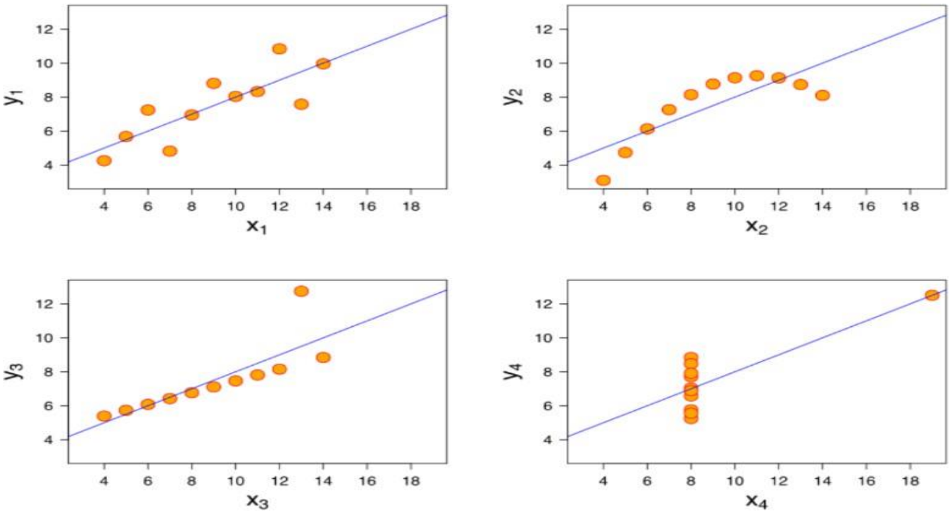
General Subjective Questions

1. Explain the linear regression algorithm in detail?

Linear regression is a predictive modeling technique showing the relationship between a dependent (target) and independent variables. It models this relationship as a straight line, with single input being simple linear regression and multiple inputs being multiple linear regression. The aim is to find the best-fit line with the least error. Techniques like RFE or Mean Squared Error help determine optimal values (a0 and a1) for the line, ensuring the best fit for the data points.

2. Explain the Anscombe's quartet in detail?

Anscombe's Quartet consists of four datasets that, despite having nearly identical simple descriptive statistics, reveal significant differences when visually inspected. These datasets have distinct distributions and scatter plot patterns. The quartet highlights the importance of graphing data before analysis and modeling, showcasing how diverse datasets can yield the same statistical properties, such as mean and variance. This emphasizes the impact of visualization on understanding data and the limitations of relying solely on summary statistics.



- The 1st dataset fits a linear regression model due to an apparent linear relationship between X and y.
- The 2nd dataset lacks a linear relationship between X and Y, making it unsuitable for a linear regression model.
- The 3rd dataset displays outliers that challenge a linear regression model's effectiveness.
- The 4th dataset has a high leverage point, leading to a high correlation coefficient.

  In conclusion, these scenarios highlight the need for data visualization before constructing machine learning models, as regression algorithms can be misled by the data's nuances.

3. What is Pearson’s R?

In statistics, the Pearson's Correlation Coefficient is commonly known as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or bivariate correlation. This statistic quantifies the linear correlation between two variables, indicating how much they vary together in a linear fashion.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

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

Scaling involves transforming data to fit within a specific scale, aiding algorithm efficiency. In data preprocessing, scaling is crucial as it ensures accurate modeling by preventing high-magnitude values from overshadowing others. There are two common scaling methods: Normalizing Scaling (using min-max values) and Standardize Scaling (using mean and standard deviation). Their differences include the range of scaled values, sensitivity to outliers, and applicability based on feature scales and distribution characteristics. Normalized scaling is suitable for diverse feature scales, while standardized scaling ensures zero mean and unit standard deviation.

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

The Variance Inflation Factor (VIF) reveals how one independent variable relates to all others. A VIF exceeding 10 indicates high correlation; values above 5 warrant inspection. Very high VIF suggests perfect correlation between two variables, leading to a mathematical issue. To address this, when facing perfect multicollinearity, one variable causing the issue is dropped from the dataset.

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

A Q–Q plot, or Quantile-Quantile plot, compares two probability distributions by plotting their quantiles against each other. It helps assess if a dataset aligns with a theoretical distribution like Normal or Exponential. Linearity in the plot suggests similarity between distributions, aiding in the verification of assumptions for linear regression. In linear regression, Q-Q plots are valuable for confirming that training and test datasets share the same distribution. Advantages include applicability to various sample sizes and the ability to detect shifts, changes in symmetry, and outliers in distributional aspects. The plot is used to check commonalities in location, scale, distribution shape, and tail behavior between two datasets.