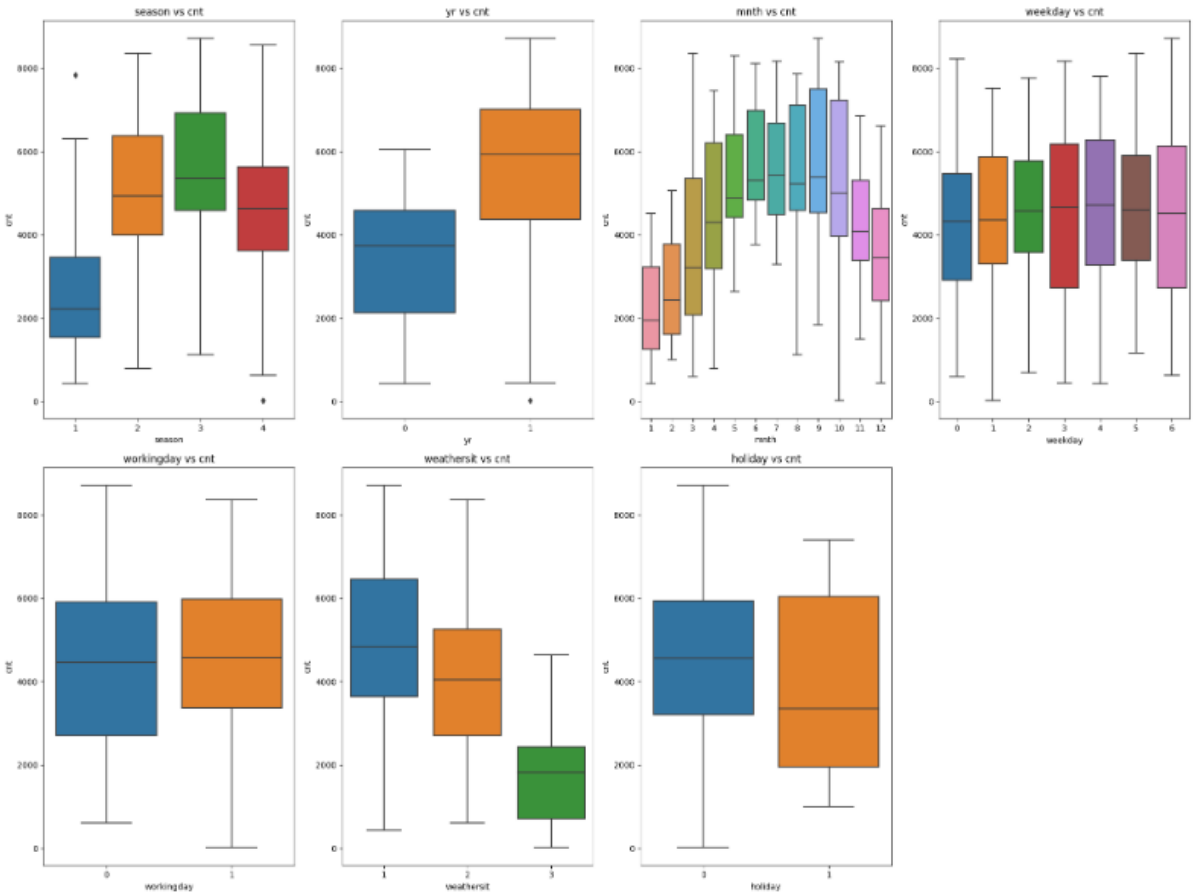
Assignment Questions

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?



Season: The boxplot indicates that bike rentals were lowest during the spring season and highest during fall. Summer and winter showed moderate rental counts.

Weathersit: Bike usage is notably absent during heavy rain or snow, suggesting these conditions are highly unfavorable. The highest rental counts occurred under 'Clear, Partly Cloudy' weather conditions.

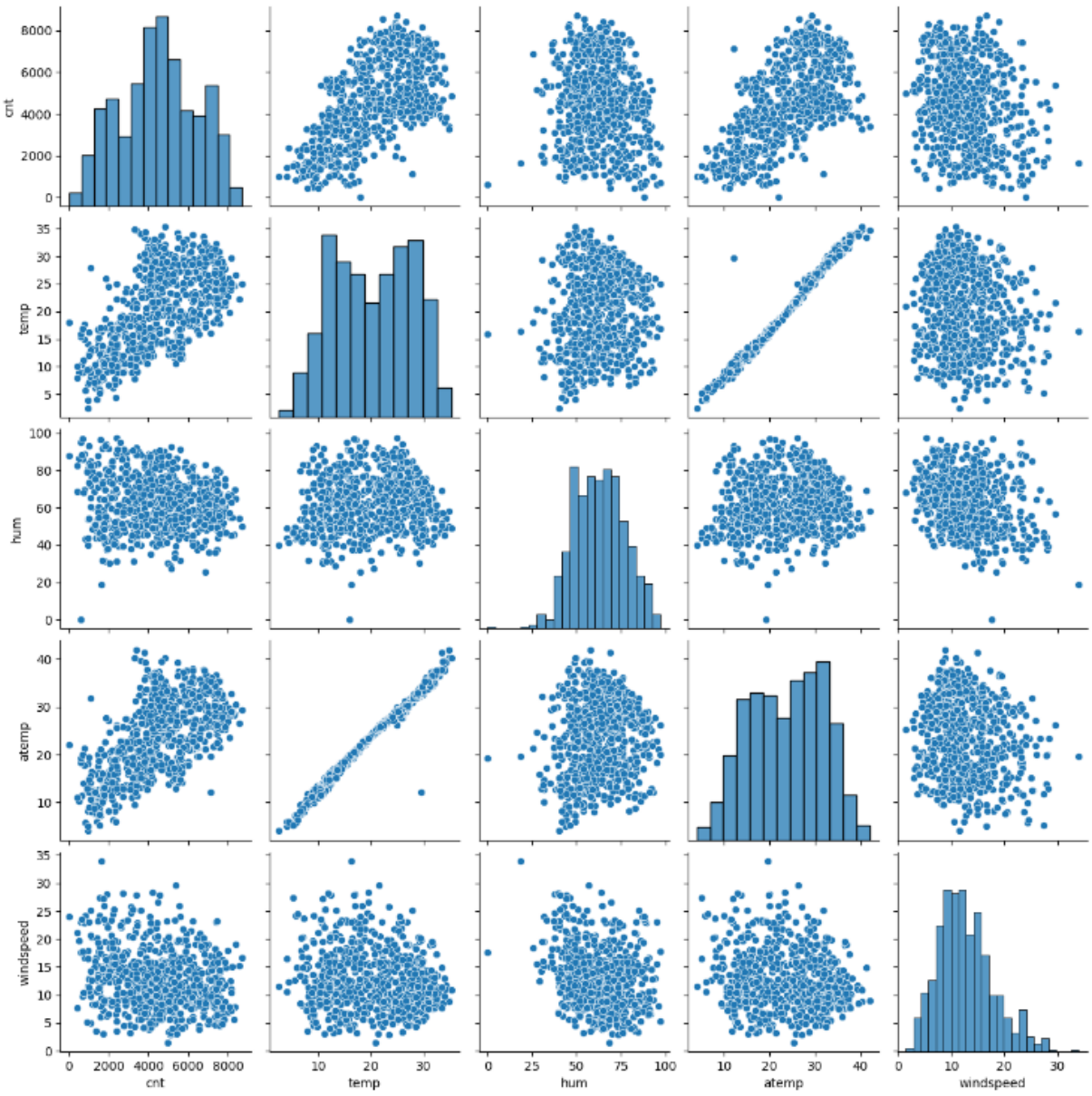
Mnth: September had the highest number of rentals, while December had the lowest. This observation aligns with the weather conditions noted, as December typically experiences heavy snow, likely reducing bike rentals.

Weekday: Bike rentals were consistent throughout the week.

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

   Using `drop_first=True` during dummy variable creation is crucial for several reasons. It ensures that the resulting set of dummy variables effectively represents categorical variables without introducing redundant information. By dropping the first column among the created dummies, which serves as a baseline or reference category, we mitigate multicollinearity issues that could adversely affect model performance. This approach helps maintain the independence of predictors in models, particularly those sensitive to correlated variables, thereby preserving the accuracy of variable importance metrics, and enhancing overall model interpretability. Moreover, reducing the number of dummy variables through `drop_first=True` improves computational efficiency by minimizing the number of features processed during model training, leading to faster convergence and reduced training times.

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



"temp" and "atemp" are the two numerical variables strongly correlated with the target variable (cnt).

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

   After constructing the linear regression model on the training set, several key assumptions were validated to ensure the model's reliability:
   Linearity: The relationship between independent variables and the dependent variable was assessed using a pairplot to visually inspect for linear patterns.

Normality of Residuals: To verify this assumption, a distribution plot (distplot) of the residuals was examined. It was observed that the residuals were approximately normally distributed around a mean of zero.
Multicollinearity: The presence of multicollinearity among independent variables was checked using the Variance Inflation Factor (VIF). This metric quantitatively assesses how much the variance of a regression coefficient is inflated due to collinearity with other predictors.
By confirming these assumptions, including checking for normality, linearity, and addressing multicollinearity, the model's validity, and suitability for making predictions were enhanced.

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

   Temperature (`temp`): This variable has the highest positive coefficient, indicating that higher temperatures are strongly associated with increased bike demand. Warmer weather likely encourages more people to rent bikes for outdoor activities.

   Year (`yr`): The year variable shows a significant positive impact, suggesting that demand has increased from one year to the next. This could reflect broader trends such as increased popularity of bike sharing over time or improvements in service offerings.

   Weather Situation (`weathersit`*: Specifically, the categories indicating favorable weather conditions (`weathersit_Clear, Partly Cloudy`) contribute positively to bike demand. This suggests that better weather conditions attract more riders, influencing rental patterns significantly.

   These features not only highlight the seasonal and weather-related factors impacting bike demand but also underscore the overall trend in increasing popularity of bike rentals over the studied period.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression is a fundamental statistical method used for modeling the relationship between a dependent variable (target) and one or more independent variables (predictors). Here's a detailed explanation of the linear regression algorithm:

Linear regression aims to fit a linear equation to observed data points. The equation takes the form:

$Y = \beta 0 + \beta 1x1 + \beta 2x2 + ... + \beta nxn + \epsilon y$

where:

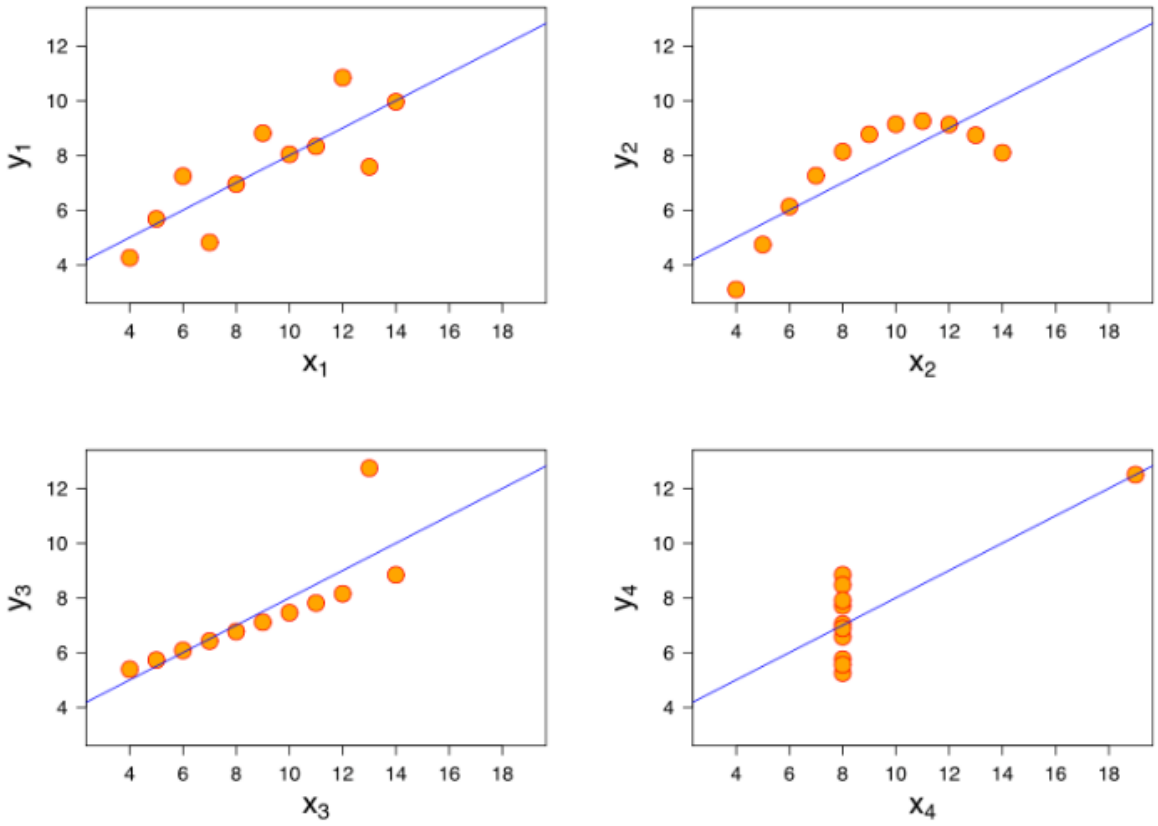
- y is the dependent variable (target) we want to predict.
- β0 is the intercept (constant term), which represents the predicted value of y when all independent variables xi are zero.
- βi (for i=1,2,...,n) are the coefficients of the independent variables xi, indicating the strength and direction of their impact on y.
- xi are the independent variables (predictors) that explain y.
- ϵ (epsilon) is the error term, representing the difference between the predicted and actual values of y.

The linear regression algorithm works by finding the coefficients βi (beta i) that minimize the sum of squared differences between the predicted and actual values (least squares method). This is achieved through techniques like Ordinary Least Squares (OLS) or gradient descent.

Assumptions of linear regression include:

1. **Linearity**: The relationship between xxx and y is linear.
2. **Independence**: Observations are independent of each other.
3. **Normality of Residuals**: Residuals (errors) should be normally distributed.
4. **Homoscedasticity**: Residuals should have constant variance across all levels of predictors.

2. Explain the Anscombe's quartet in detail.



Anscombe's quartet is a famous example in statistics that demonstrates the importance of visualizing data and checking assumptions before drawing conclusions based on statistical analysis. It consists of four datasets that have nearly identical simple descriptive statistics (mean, variance, correlation, regression line), yet are visually and analytically quite different.

Here is an explanation of each dataset in Anscombe's quartet:

1. Dataset I:
   - Characteristics: It forms a linear relationship where y increases linearly with x.
   - Statistics: The mean of x and y, variance, correlation coefficient, and the linear regression line fit are all typical and straightforward.

2. Dataset II:
   - Characteristics: It also shows a linear relationship between x and y, but with an outlier that significantly affects the regression line and correlation.
   - Statistics: The descriptive statistics like mean, variance, and correlation coefficient are very similar to Dataset I, but the presence of the outlier influences the interpretation.

3. Dataset III:

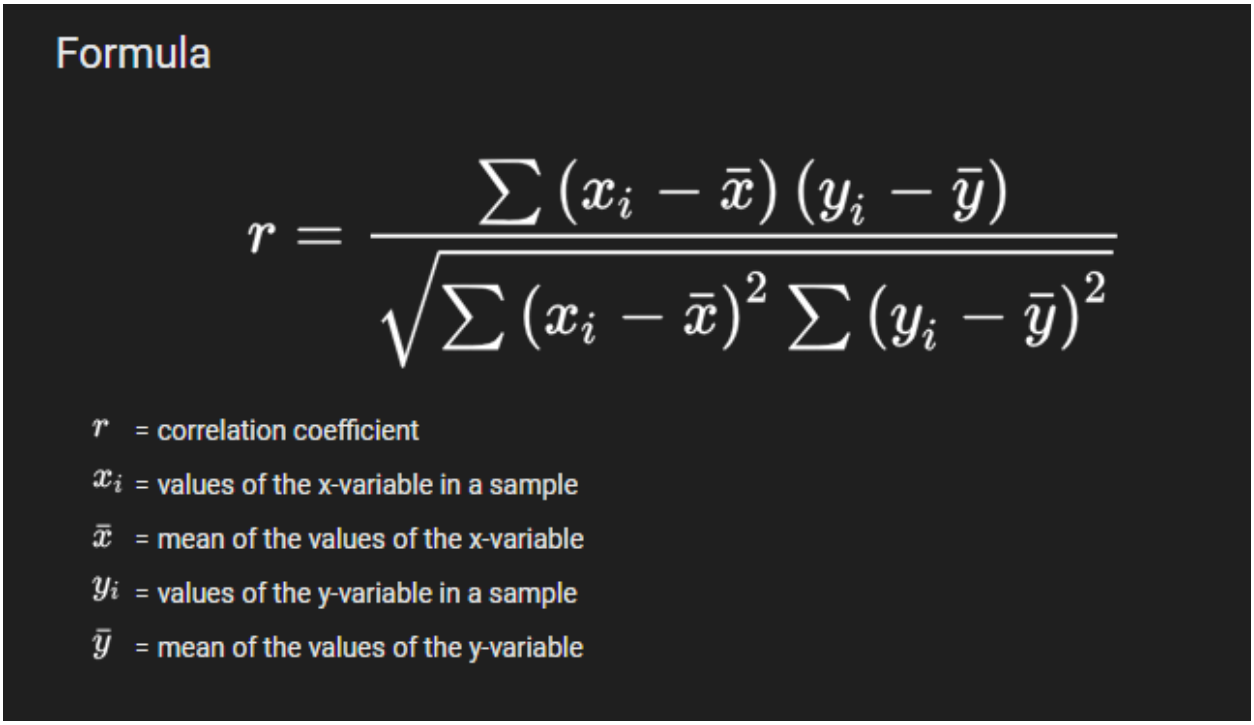
- Characteristics: This dataset appears to follow a quadratic relationship rather than a linear one. It shows an outlier that affects both the correlation and the linear regression model.
- Statistics: Despite the non-linear pattern, the mean, variance, and correlation coefficient are like those of the previous datasets, highlighting the limitation of relying solely on these summary statistics.

4. Dataset IV:
- Characteristics: It consists of three groups of data where x values are the same in each group, but y values vary significantly. Each group has a distinct linear relationship.
- Statistics: The summary statistics for each group, including mean, variance, and correlation, are nearly identical across all groups, but the actual distributions and relationships differ widely.

Anscombe's quartet serves as a cautionary tale in statistics, reminding researchers and analysts to approach data analysis with both numerical and visual methods to ensure robust and accurate conclusions.

3. What is Pearson's R?

Pearson's R (Pearson correlation coefficient) measures the linear correlation between two variables, X and Y, ranging from -1 (perfect negative correlation) to +1 (perfect positive correlation), with 0 indicating no linear correlation. It assesses the strength and direction of the relationship between variables based on their covariance and standard deviations.



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

   Feature scaling adjusts the range of independent variables or features in data to ensure fair comparison and optimal performance of machine learning algorithms. It is crucial during data preprocessing to handle varying value magnitudes. Without scaling, algorithms may prioritize larger values and disregard smaller ones, regardless of their inherent units.

   - **Normalization**: Typically used when data distribution is not Gaussian, like in K-Nearest Neighbors and Neural Networks. It transforms data to a common scale (0 to 1), preserving relative relationships between values.

   - **Standardization**: Suitable when data approximates a Gaussian distribution, though not strictly necessary. It centers data around 0 with a standard deviation of 1, making it less sensitive to outliers. Unlike normalization, it does not constrain values within a specific range.

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

   The occurrence of an infinite value of VIF (Variance Inflation Factor) typically happens when there is perfect multicollinearity among the predictor variables in the regression model. Perfect multicollinearity occurs when one or more independent variables can be exactly predicted from others with a linear combination. This situation leads to an infinite VIF because the correlation matrix of the predictors becomes singular or nearly singular, making it impossible to compute the inverse needed for VIF calculation.

   Key reasons for infinite VIF include:
   1. Redundant Predictors: One predictor can be expressed as a perfect linear combination of others.

   2. Dummy Variable Trap: When dummy variables are not properly handled (e.g., including all levels without dropping one as a reference category), it can lead to perfect multicollinearity among the dummy variables.

   3. Data Coding Issues: Incorrect coding or data errors can sometimes inadvertently create perfect multicollinearity.

   In practice, infinite VIF values indicate a severe issue that must be addressed, such as removing one of the correlated predictors or rethinking the model specification to avoid multicollinearity. Handling multicollinearity is crucial for the reliability and interpretability of regression models, as high VIF values indicate inflated standard errors and potentially misleading coefficient estimates.

6. 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 whether a dataset follows a particular distribution, such as the normal distribution. It compares the quantiles of the data against the quantiles of a theoretical distribution (usually the normal distribution).

- Use and Importance in Linear Regression:

- Distribution Assumption**: In linear regression, it is often assumed that the residuals (errors) follow a normal distribution. A Q-Q plot helps to visually inspect whether this assumption holds true. If the points on the Q-Q plot approximately fall along a straight line, it suggests that the residuals are normally distributed.

- Identifying Outliers: Q-Q plots can also reveal outliers or deviations from the expected distribution. Outliers appear as points that significantly deviate from the straight line, indicating potential issues with model assumptions or data quality.

- Model Validity: Ensuring that residuals are normally distributed is crucial for valid inference and prediction from a linear regression model. Non-normality can affect the accuracy of confidence intervals, hypothesis tests, and predictions based on the model.

- Interpretation: A clear Q-Q plot provides confidence that the assumptions of linear regression are met, enhancing the reliability of conclusions drawn from the model. It also guides potential adjustments or transformations to improve model performance if assumptions are violated.