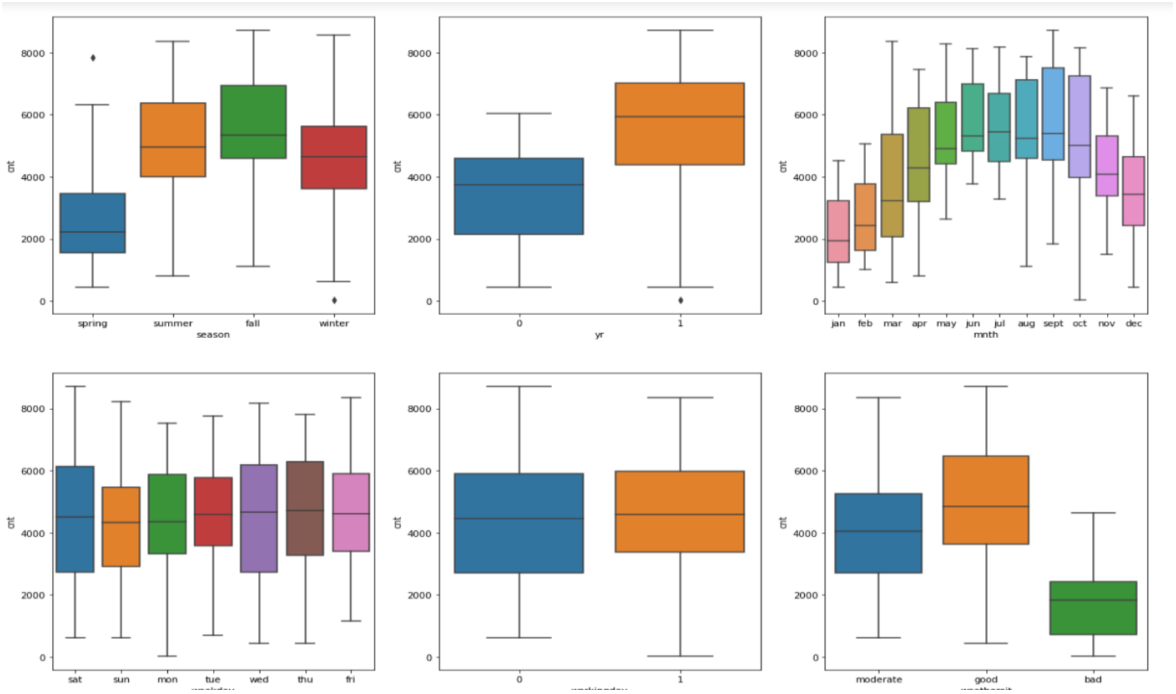
1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?**

   The following figure illustrates the correlation between several categorical variables—namely season, month, year, weekday, working day, and weather situation—and the dependent variable 'cnt'. These categorical variables significantly influence 'cnt'.

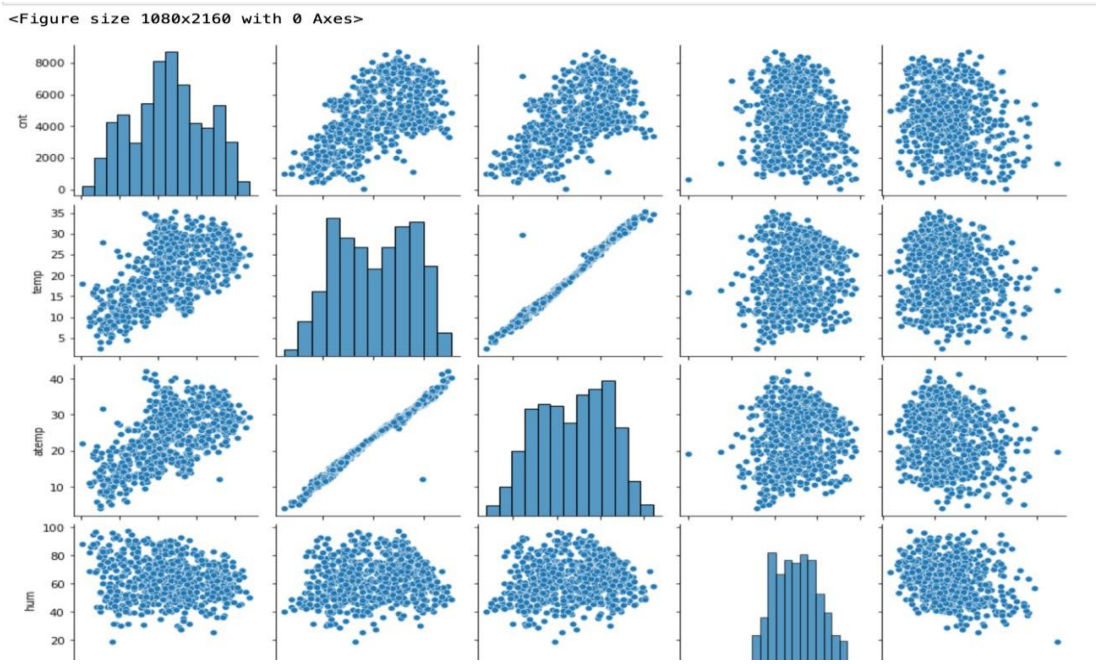


2. **Why is it important to use drop_first=True during dummy variable creation?**

Using drop_first=True is crucial to avoid multicollinearity by dropping the first category, thus preventing redundancy. For a categorical variable with 'n' levels, creating 'n-1' dummy variables ensures that the model avoids perfect multicollinearity, maintaining the integrity of the regression coefficients.

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

The 'temp' (temperature) and 'atemp' (feels-like temperature) variables show the highest correlation with the target variable 'cnt'. This suggests that higher temperatures lead to an increase in bike rentals, as people are more likely to ride bikes in pleasant weather.



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

Assumptions were validated by:

- **Linearity**: Scatter plots confirmed linear relationships between predictors and the target.
- **No auto-correlation**: The Durbin-Watson test indicated independence of residuals.
- **Normality of error**: Q-Q plots showed normally distributed residuals.
- **Homoscedasticity**: Residuals vs. fitted plots indicated constant variance.
- **Multicollinearity**: Variance Inflation Factor (VIF) values ensured no high correlations among predictors.

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

The top 3 features are temperature, year, and season, indicating that climatic conditions, yearly trends, and seasonal patterns significantly influence bike rental demand.

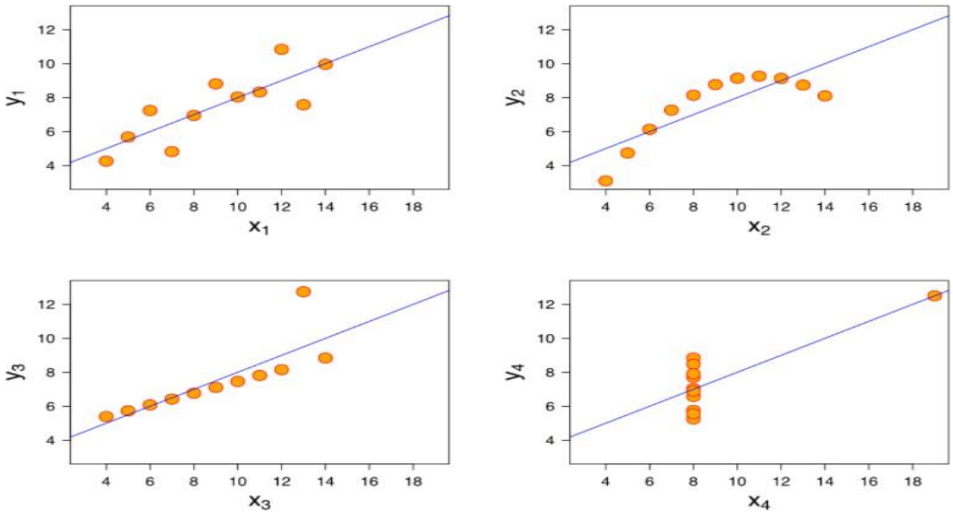
General Subjective Questions

1. **Explain the linear regression algorithm in detail.**

Linear regression models the relationship between a dependent variable and one or more independent variables using a linear equation y=a0+a1x1+a2x2+…+anxny = a_0 + a_1x_1 + a_2x_2 + \ldots + a_nx_ny=a0+a1x1+a2x2+…+anxn. The objective is to find the best-fit line that minimizes the prediction errors, typically using the Mean Squared Error (MSE). This algorithm can handle both simple (one predictor) and multiple (multiple predictors) linear regression, providing insights into how changes in predictors affect the response variable.

2. **Explain the Anscombe's quartet in detail.**

Anscombe's Quartet consists of four datasets with nearly identical statistical properties but different distributions and visual appearances. It demonstrates the importance of data visualization before analysis. These datasets highlight how different distributions, non-linear relationships, outliers, and leverage points can mislead regression models if not properly visualized.



3. **What is Pearson's R?**

Pearson's R, or Pearson's correlation coefficient, measures the linear relationship between two variables, ranging from -1 to 1. A value of 1 indicates a perfect positive linear relationship, -1 indicates a perfect negative linear relationship, and 0 indicates no linear relationship.

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

Scaling transforms data to fit within a specific range, crucial for algorithms reliant on distance metrics. It prevents features with larger magnitudes from dominating the model.

- **Normalized scaling**: Scales data to a range of [0, 1] or [-1, 1] using minimum and maximum values. Suitable for features with different scales but sensitive to outliers.
- **Standardized scaling**: Scales data to a mean of 0 and a standard deviation of 1, ensuring a normal distribution and less sensitivity to outliers.

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

An infinite VIF indicates perfect multicollinearity, where one independent variable is a perfect linear combination of others, resulting in an R-squared value of 1 and an undefined VIF. Removing one of the perfectly collinear variables resolves this issue.

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

A Q-Q (Quantile-Quantile) plot compares the quantiles of a sample distribution with a theoretical distribution, assessing normality. In linear regression, Q-Q plots verify if residuals are normally distributed. Linear patterns in Q-Q plots confirm normality, essential for valid regression models. Q-Q plots are advantageous for small samples and reveal distributional aspects like shifts, symmetry, and outliers. They ensure that training and testing datasets follow similar distributions, crucial for model reliability.