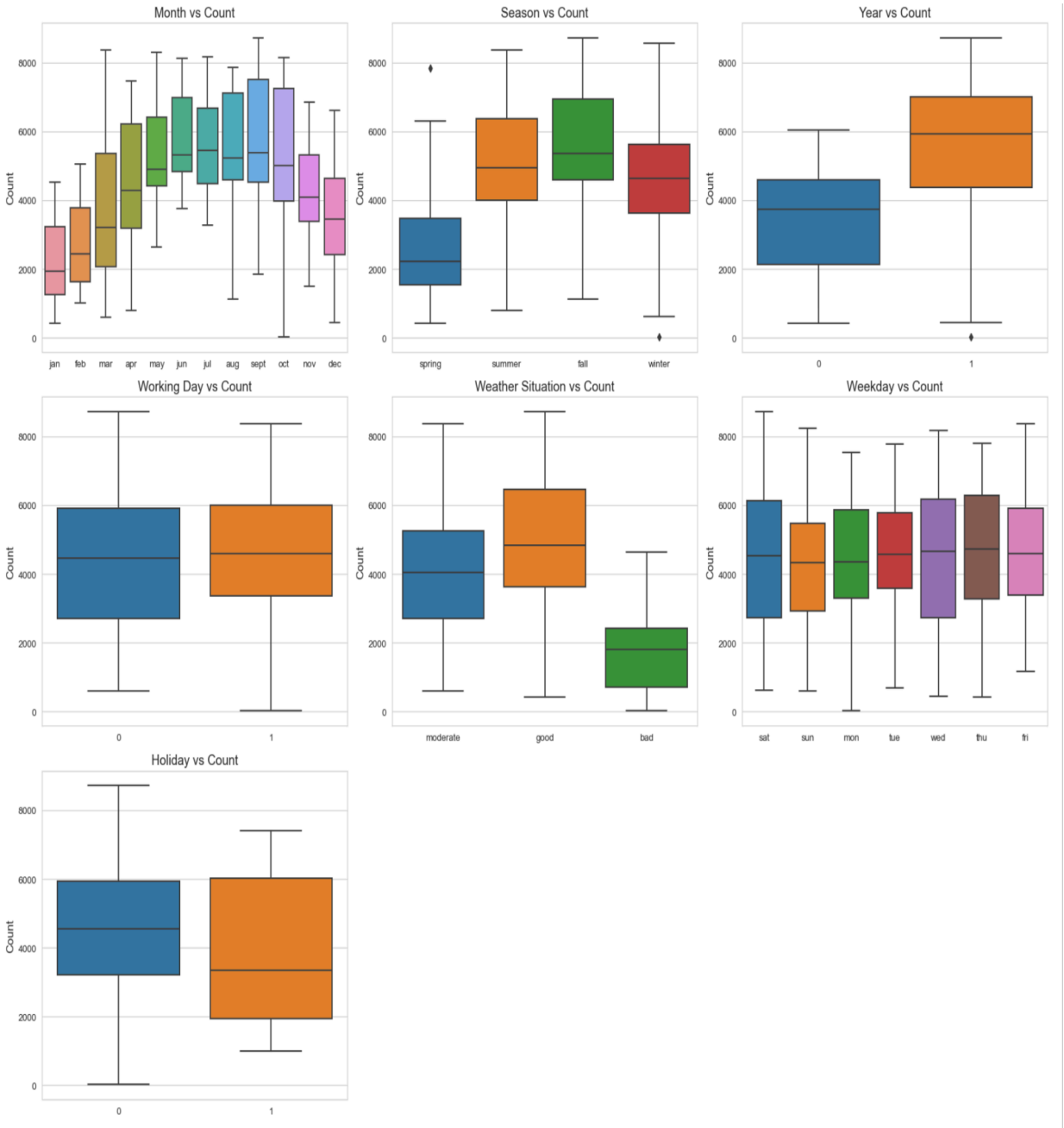
# Assignment-based Subjective Questions

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




- Bike sharing demand peaks in September and is lowest in January and December.
- The fall season has the highest bike sharing count, while spring has the lowest.
- The year 2019 saw more bike sharing than 2018, with higher demand on working days.
- Good weather conditions significantly increase bike sharing demand, whereas bad weather decreases it.

- Among weekdays, Saturday sees the highest bike sharing activity, while Sunday has the lowest, and holidays

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

We convert categorical variables into numerical variables by creating dummy variables, which are essential for model building. For a categorical variable with n levels, we create n-1 new columns by setting drop_first=True. This approach reduces correlation among the dummy variables and helps avoid multicollinearity by creating n-1 dummy variables instead of n.

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

| | cnt | temp | atemp | hum | windspeed |
|---|---|---|---|---|---|
| cnt | 1.000000 | 0.627044 | 0.630685 | −0.098543 | −0.235132 |
| temp | 0.627044 | 1.000000 | 0.991696 | 0.128565 | −0.158186 |
| atemp | 0.630685 | 0.991696 | 1.000000 | 0.141512 | −0.183876 |
| hum | −0.098543 | 0.128565 | 0.141512 | 1.000000 | −0.248506 |
| windspeed | −0.235132 | −0.158186 | −0.183876 | −0.248506 | 1.000000 |

atemp(0.630) & Temp (0.627) has the highest correlation with target variables

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

After building linear regression models, we validate the model by:

1. **Residual Analysis**: Compare the differences between actual and predicted values to check the pattern against the training model's pattern.
2. **Normality**: Ensure that the error terms are normally distributed with a mean of 0.
3. **Homoscedasticity**: Verify that the data points are spread out consistently and the error is consistent throughout the predictions.
4. **Multicollinearity**: Check for multicollinearity among predictor variables by calculating the Variance Inflation Factor (VIF). High VIF indicates multicollinearity, which can distort parameter effects and affect model interpretability. For a model to be considered good, the selected VIF should be less than 5.

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

## General Subjective Questions

**Explain the linear regression algorithm in detail**

Linear regression models the relationship between a dependent variable (target variable) and one or more independent variables by fitting a straight line through the data points.

1. **Assumptions**:
    - **Linearity**: The relationship between the independent variables and the dependent variable is assumed to be linear.
    - **Independence**: Observations are assumed to be independent of each other.
    - **Homoscedasticity**: The variance of the errors (residuals) is assumed to be constant across all levels of the independent variables.
    - **Normality**: The residuals are assumed to be normally distributed.
2. **Model Fitting**:
    - The model coefficients are estimated using a method called Ordinary Least Squares (OLS). In this assignment, OLS from the statsmodels API was used.
    - OLS minimizes the sum of the squared differences between the observed and predicted values.
    - It involves finding the values of coefficients that minimize the cost function.
3. **Gradient Descent**:
    - Gradient Descent is a general optimization algorithm used to minimize the cost function iteratively.
    - It updates the coefficients by taking steps proportional to the negative of the gradient of the cost function.
4. **Model Evaluation**:
    - After fitting the model, its performance is evaluated using metrics like R-squared, which measures the average squared difference between observed and predicted values.
    - The model is also evaluated by checking the coefficients.
5. **Feature Selection**:
    - The final model is selected based on criteria such as high R-squared value, P-value < 0.05, and VIF < 5.
6. **Residual Analysis**:
    - Analyze the residuals (differences between observed and predicted values) to check for patterns that might suggest problems with the model.
7. **Verifying the Assumptions**:
    - The assumptions of linearity, independence, homoscedasticity, and normality are verified to ensure the validity of the model.
8. **Model Validation**:

- Validate the model by comparing predictions on test data with actual values to ensure that the model generalizes well to unseen data.

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

Anscombe's quartet is a set of four datasets that have nearly identical simple descriptive statistics yet appear very different when graphed

## Datasets and Their Properties

**Dataset I**

- **Mean of x**: 9.0
- **Mean of y**: 7.5
- **Variance of x**: 11.0
- **Variance of y**: 4.127
- **Correlation between x and y**: 0.816
- **Linear regression line**: y = 3.00 + 0.500x

**Dataset II**

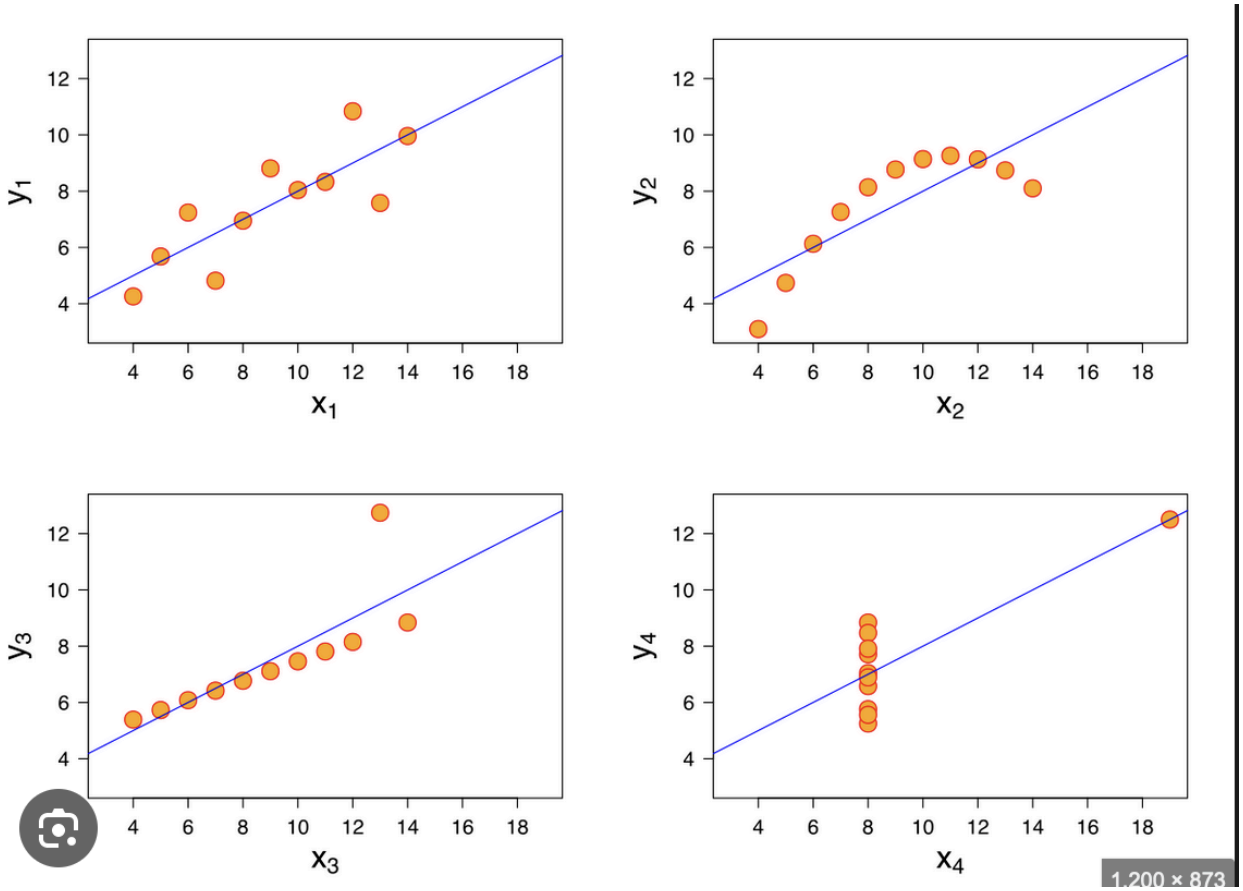
- **Mean of x**: 9.0
- **Mean of y**: 7.5
- **Variance of x**: 11.0
- **Variance of y**: 4.127
- **Correlation between x and y**: 0.816
- **Linear regression line**: y = 3.00 + 0.500x

**Dataset III**

- **Mean of x**: 9.0
- **Mean of y**: 7.5
- **Variance of x**: 11.0
- **Variance of y**: 4.127
- **Correlation between x and y**: 0.816
- **Linear regression line**: y = 3.00 + 0.500x

**Dataset IV**

- **Mean of x**: 9.0
- **Mean of y**: 7.5
- **Variance of x**: 11.0
- **Variance of y**: 4.127
- **Correlation between x and y**: 0.816
- **Linear regression line**: y = 3.00 + 0.500x




Anscombe's quartet serves as a critical reminder to always visualize data before analyzing it and to be cautious of relying solely on summary statistics for understanding data

**What is Pearson's R?**

Pearson's R, also known as Pearson's correlation coefficient, is a measure of the linear correlation between two variables. It quantifies the degree to which a relationship between two variables can be described by a straight line.

Key Characteristics of Pearson's R:

1. Range: Pearson's R ranges from -1 to +1.
    - +1: A perfect positive linear relationship between variables.
    - 0: No linear relationship between variables.
    - -1: A perfect negative linear relationship between variables.
2. Interpretation:
    - Positive Correlation (0 to +1): As one variable increases, the other variable also increases.

    - Negative Correlation (0 to -1): As one variable increases, the other variable decreases.
    - Zero Correlation: No linear relationship between the variables.
3. Symmetry: Pearson's R is symmetric, meaning the correlation between X and Y is the same as the correlation between Y and X.

## Calculation of Pearson's R:

The formula to calculate Pearson's correlation coefficient is:

$$r = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum (x_i - \overline{x})^2 \sum (y_i - \overline{y})^2}}$$

## Assumptions for Pearson's R:

1. Linearity: The relationship between the variables should be linear.
2. Homoscedasticity: The variance of the variables should be constant across the range of values.
3. Normality: The variables should be approximately normally distributed, especially when using Pearson's R for hypothesis testing.

## Usage of Pearson's R:

1. Correlation Analysis: Pearson's R is commonly used to measure the strength and direction of the linear relationship between two continuous variables.
2. Hypothesis Testing: It can be used to test hypotheses about the correlation between variables. For example, testing whether the correlation is significantly different from zero.
3. Regression Analysis: It provides an initial measure of the relationship strength before fitting a linear regression model.

## Conclusion:

Pearson's R is a widely used measure of linear correlation between two variables, providing insights into the strength and direction of their relationship. While it is a powerful tool, it is important to consider its assumptions and limitations, especially in the presence of outliers or non-linear relationships.

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

Scaling is the process of transforming the features of a dataset so that they fall within a specific range or follow a particular distribution. This transformation helps ensure that each feature contributes equally to the analysis and modeling process, particularly in machine learning algorithms where the scale of features can significantly impact model performance.

## Why is Scaling Performed?

1. Improves Model Performance:
    - Many machine learning algorithms, such as gradient descent-based methods, perform better when the input features are on a similar scale.
2. Speeds Up Convergence:
    - For algorithms that use gradient descent for optimization, scaling can speed up the convergence of the model by ensuring that all features are treated equally.
3. Reduces Numerical Instability:
    - Scaling helps prevent numerical issues that can arise from large differences in feature values, leading to more stable and accurate computations.
4. Enhances Interpretability:
    - When features are on similar scales, it is easier to interpret the importance and contribution of each feature to the model.

## Types of Scaling

There are two primary types of scaling: normalization (or min-max scaling) and standardization (or z-score scaling).

**1. Normalization (Min-Max Scaling)**

**Definition**:

- Normalization scales the features to a fixed range, typically [0, 1] or [-1, 1].

**Formula**:

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$$

**Characteristics**:

- **Range**: The transformed data will be within the specified range.
- **Sensitive to Outliers**: Outliers can significantly affect the min and max values, leading to a skewed transformation.

**2. Standardization**

**Definition:**

- Standardization transforms the features to have a mean of 0 and a standard deviation of 1.

**Formula**:

$$x' = \frac{x - \mu}{\sigma}$$

**Characteristics**:

- **Mean and Standard Deviation**: The transformed data will have a mean of 0 and a standard deviation of 1.
- **Less Sensitive to Outliers**: Compared to normalization, standardization is less affected by outliers.

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

Variance Inflation Factor (VIF) measures the extent to which the variance of an independent variable is explained by the linear relationship with other independent variables.

## Interpreting VIF Values:

- VIF > 10: Indicates very high multicollinearity. The variable should be considered for elimination.
- VIF > 5: May be acceptable but requires further investigation.
- VIF < 5: Indicates low multicollinearity. No need to eliminate the variable.

## Formula:

$$\text{VIF} = \frac{1}{1 - R^2}$$

If the R(square) value is 1 (indicating perfect correlation), the denominator becomes zero, resulting in an infinite VIF value.

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

Quantile-Quantile (Q-Q) Plot

The quantile-quantile (Q-Q) plot is a graphical technique used to determine if two datasets originate from populations with a common distribution, such as the normal distribution. It compares the quantiles of the dataset to the quantiles of a theoretical distribution, typically the normal distribution, on a scatter plot.

- Purpose: The Q-Q plot can be used to assess whether two distributions are similar. If the distributions are similar, the Q-Q plot will appear more linear. Linearity in the Q-Q plot indicates that the variables follow a common distribution.

Q-Q Plot for Linear Regression:

1. Check Normality of Residuals:
    - Q-Q plots are commonly used in linear regression to check the normality assumption of the residuals (errors).
2. Residual Distribution:
    - In linear regression, the residuals should ideally follow a normal distribution with a mean of zero.
3. Assessing Normality:
    - By examining the Q-Q plot of the residuals, you can assess whether the normality assumption is met.
4. Linearity Indication:
    - If the points on the Q-Q plot fall approximately along a straight line, it suggests that the residuals are normally distributed.
5. Deviation from Linearity:
    - Deviations from the straight line indicate departures from normality, which may affect the validity of statistical inference and confidence intervals in linear regression analysis.

Importance of Q-Q Plot:

1. Visual Assessment:
    - The Q-Q plot provides a visual assessment of the normality assumption, which is crucial for the validity of statistical inference in linear regression.
2. Complementary to Statistical Tests:
    - It complements formal statistical tests for normality by offering a graphical representation of the distributional properties of the residuals.
3. Identifying Departures from Normality:
    - Q-Q plots allow for easy identification of departures from normality, such as skewness or heavy tails, which may require further investigation or transformation of the data.
4. Informed Decisions:

- By examining the Q-Q plot, we can make informed decisions about the appropriateness of linear regression assumptions and take corrective actions if necessary.