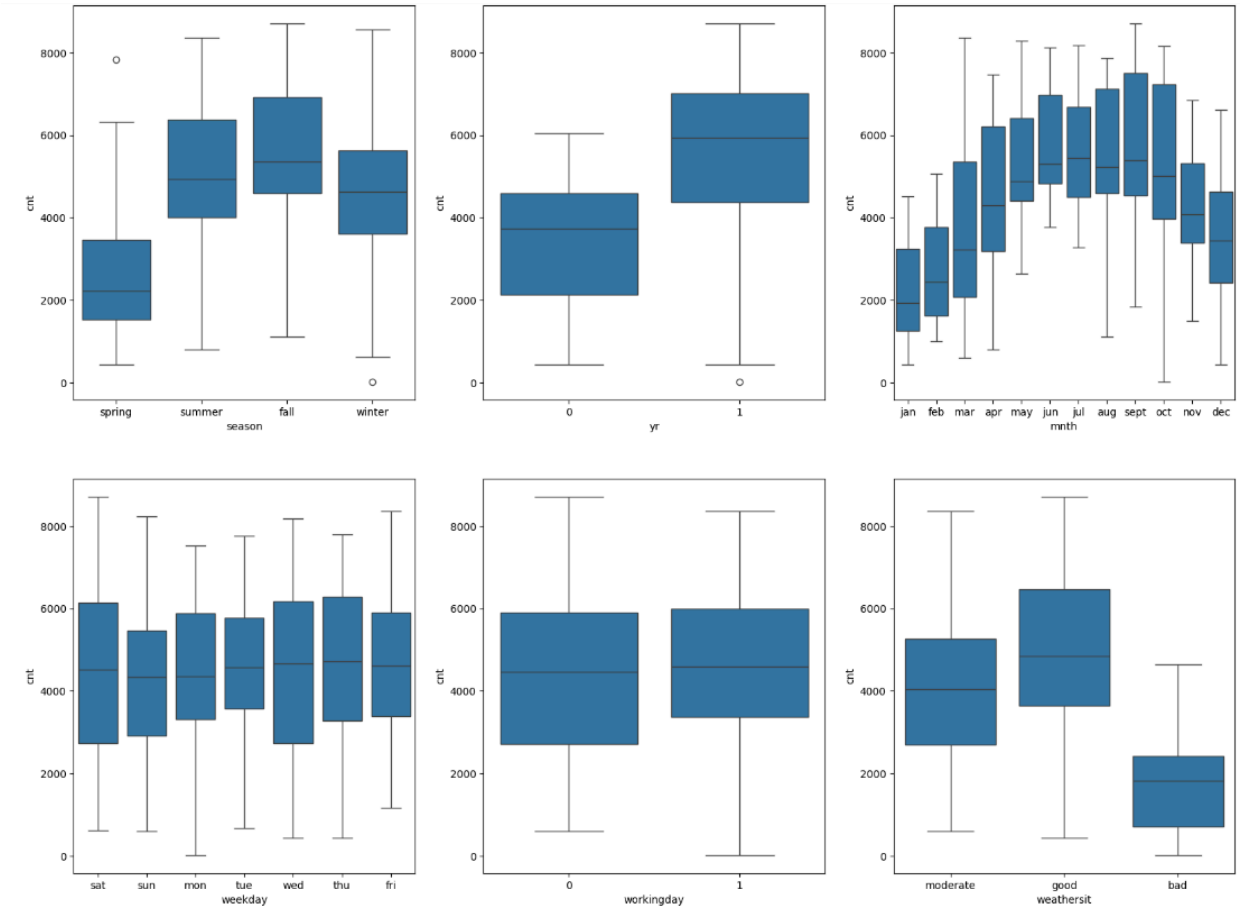
| | |
|---|---|
| Name | Swapnil William |
| Email ID | williamswapnil99@ gmail.com |

# 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?**

**Answer:** The categorical variables in the dataset show a significant impact on bike demand. Certain months like March and September positively influence demand, while July and November see a decline. Seasonal effects are evident, with spring reducing and winter increasing bike usage compared to fall. Weather conditions play a crucial role—bad weather significantly lowers demand, whereas good or mildly cloudy weather supports it. Overall, seasonality and weather patterns are strong indicators and must be considered when analyzing or forecasting bike demand

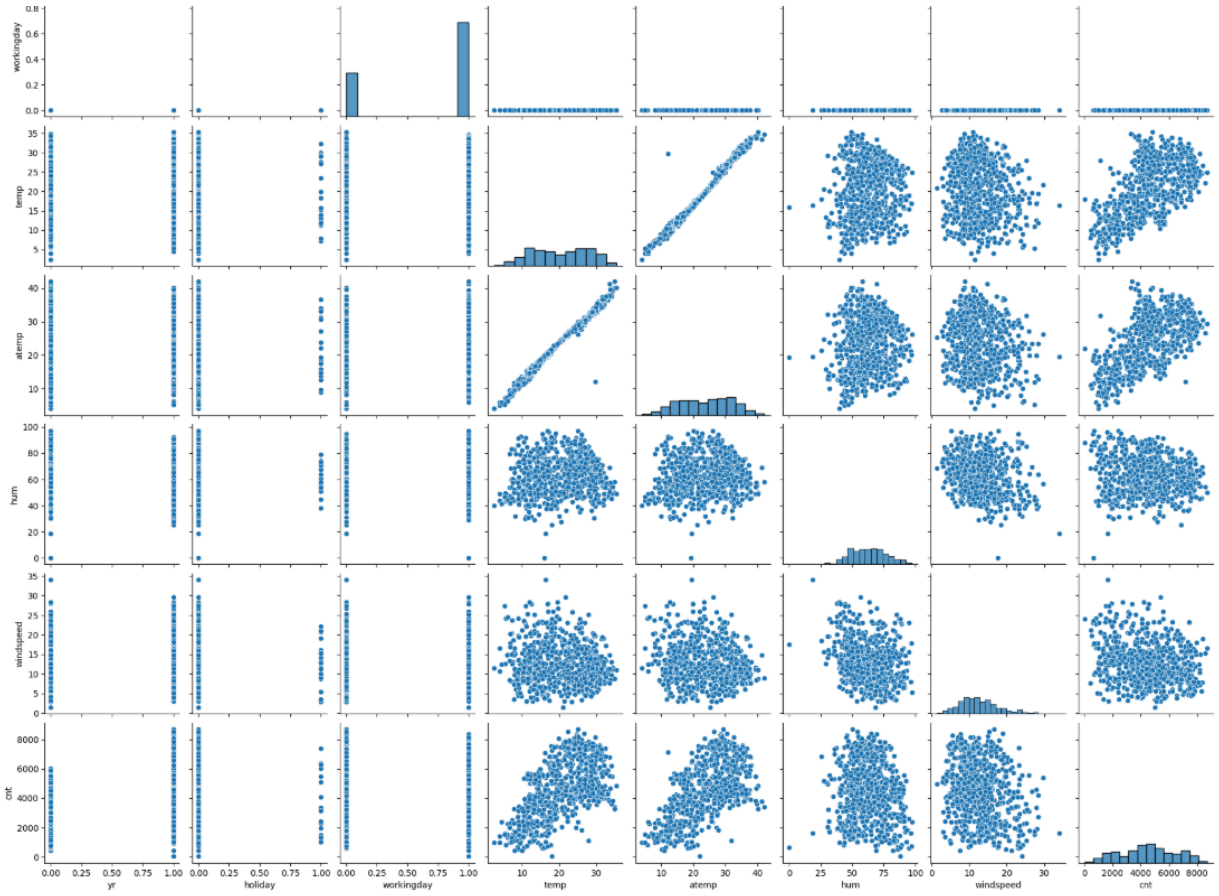


2. **Why is it important to use drop_first=True during dummy variable creation?**

**Answer:** Using drop_first=True when creating dummy variables is important to avoid the dummy variable trap, which occurs due to multicollinearity—a situation where one variable can be perfectly predicted from others. In dummy encoding, if all categories are included, one can be derived from the rest, leading to redundancy and unstable regression estimates. By dropping the first category, we remove this redundancy while still preserving all the necessary information, allowing the model to interpret the effects of each category relative to the one that was dropped.

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

**Answer:** When we observe the pair plot among the numerical variable, we find that 'temp' and 'atemp' have the highest correlation with the target variable.




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

**Answer:**

After building the linear regression model on the training set, I validated the assumptions using several diagnostic checks:

i. Linearity: Verified by plotting the actual vs. predicted values to check for a linear pattern.

ii. Normality of Residuals: Assessed using a histogram or Q-Q plot of residuals to ensure they follow a normal distribution.

iii. Homoscedasticity: Checked using a scatter plot of residuals vs. predicted values to ensure constant variance.

iv. Multicollinearity: Evaluated using the Variance Inflation Factor (VIF); variables with high VIF were considered for removal.

v. Independence of Errors: Confirmed using the Durbin-Watson statistic to ensure no autocorrelation in residuals.

These steps helped ensure that the linear regression model met key assumptions for reliable predictions.

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

**Answer:** The top 3 features contributing significantly towards influencing the demand are year, season, and temperature.

## General subjective Questions

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

**Answer:** Linear Regression is a supervised learning algorithm used for predicting a continuous dependent variable based on one or more independent variables. It assumes a linear relationship between the variables.

i. Equation of Line:

The model tries to fit a line:
$$\mathbf{y = \beta 0 + \beta 1 x1 + \beta 2 x2 + ... + \beta n xn + \varepsilon}$$

where:

- y: dependent variable (target)
- x1,x2,...,xnx independent variables (features)
- β0β: intercept
- β1,...,βnβ: coefficients (weights)
- ε: error term

ii. Objective:
Find the best-fitting line that minimizes the sum of squared residuals (differences between actual and predicted values). This is known as the Ordinary Least Squares

iii. Training the Model:

- During training, the algorithm calculates the optimal values for the coefficients (βs) by minimizing the cost function (mean squared error).
- These coefficients represent the change in the dependent variable for a unit change in the respective independent variable.

iv. Assumptions:
Linear Regression assumes:

- Linearity between dependent and independent variables
- Independence of errors
- Homoscedasticity (constant variance of errors)
- No multicollinearity among predictors
- Normality of residuals

Linear Regression is easy to interpret and widely used for its simplicity and effectiveness in identifying relationships between variables.

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

**Answer:** Anscombe's quartet is a set of four datasets that were constructed by statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data before analyzing it statistically. Each dataset in the quartet has nearly identical statistical properties, such as:

- The same mean and variance for both x and y
- The same correlation coefficient between x and y
- The same linear regression line and $R^2$ value

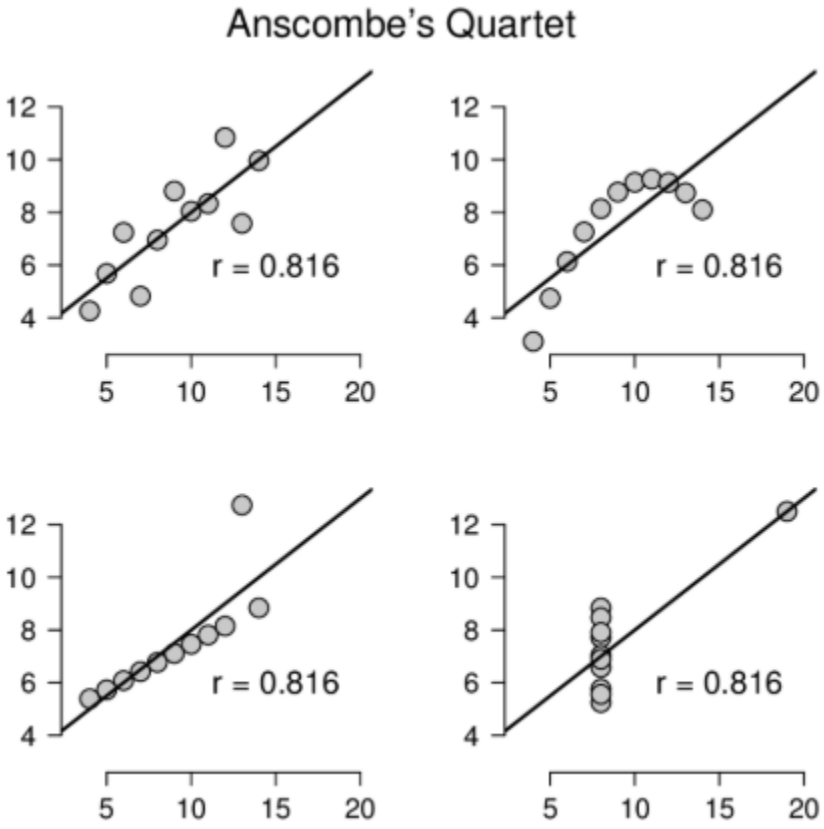
Despite these similarities, the four datasets are very different when graphed..

A. Dataset I shows a typical linear relationship between x and y, which is well modeled by linear regression.

B. Dataset II appears to follow a clear curve rather than a line, indicating a non-linear relationship.

C. Dataset III contains a linear relationship but with an outlier that strongly influences the regression line.

D. Dataset IV consists of many identical x-values with one extreme outlier that skews the regression results.




Anscombe's quartet emphasizes that relying solely on summary statistics or regression output can be misleading. Visual inspection of data (e.g., through scatter plots) is critical to understanding underlying patterns, spotting anomalies, and choosing appropriate models.

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

**Anwer:** Pearson's R, also known as the Pearson correlation coefficient, is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables. It ranges from -1 to +1:

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[\, n\Sigma x^2 - (\Sigma x)^2 \,][\, n\Sigma y^2 - (\Sigma y)^2 \,]}}$$

- A value of +1 indicates a perfect positive linear correlation: as one variable increases, the other increases proportionally.
- A value of -1 indicates a perfect negative linear correlation: as one increases, the other decreases proportionally.
- A value of 0 indicates no linear relationship between the variables.

Pearson's R is sensitive to outliers and assumes both variables are normally distributed and linearly related.

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

**Answer:** Scaling is the process of transforming numerical features in a dataset so they share a common scale without distorting differences in the range of values. It ensures that features contribute equally to model training, especially in algorithms like linear regression, KNN, and SVM that are sensitive to the magnitude of data.

Why Scaling is Performed:
Scaling helps improve model convergence, accuracy, and training efficiency. It prevents features with large values from dominating those with smaller ones.

Difference Between Normalization and Standardization:

i. **Normalized Scaling (Min-Max Scaling):** Transforms data to a fixed range, usually [0, 1], using the formula:
(x−min)/(max−min)(x - min) / (max - min)
It is sensitive to outliers.

ii. **Standardized Scaling (Z-score Scaling):** Transforms data to have a mean of 0 and standard deviation of 1 using:

(x−mean)/std(x - mean) / std

It retains the shape of the original distribution and is more robust to outliers.

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

**Answer:** A Variance Inflation Factor (VIF) becomes infinite when there is perfect multicollinearity between two or more independent variables. This means one variable can be expressed as an exact linear combination of others. In such cases, the $R^2$ value (used in VIF calculation) becomes 1, and the denominator of the VIF formula 1−R21 - R^2 becomes zero, causing the VIF to approach infinity. This typically indicates a severe redundancy in predictors, which can distort the regression model's estimates and should be addressed by removing or combining correlated variables.

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

**Answer:** A Q-Q plot (Quantile-Quantile plot) is a graphical tool used to compare the distribution of a dataset against a theoretical distribution—typically the normal distribution. In the context of linear regression, it is mainly used to check whether the residuals (errors) of the model follow a normal distribution, which is one of the key assumptions of linear regression.

In a Q-Q plot, if the residuals are normally distributed, the points will lie approximately along the 45-degree reference line. Significant deviations from this line suggest non-normality, indicating that the model may not meet the assumption of normally distributed errors. This can affect the reliability of confidence intervals and hypothesis tests in the regression model. Therefore, the Q-Q plot is important for validating model assumptions and ensuring the results of the regression analysis are trustworthy.