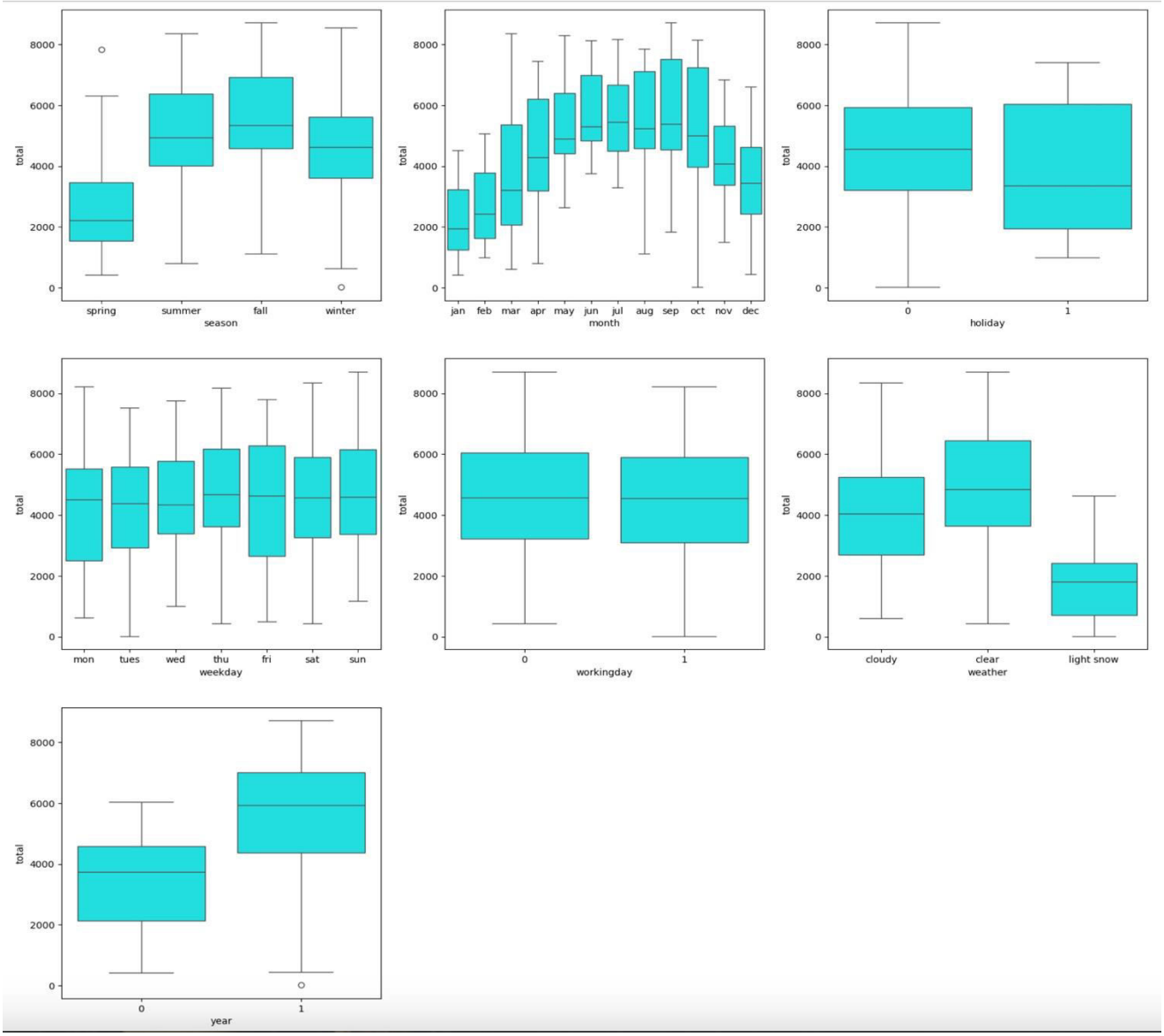
Assignment-based Subjective Questions

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

Answer:



From the above boxplot we can visualize that season, weather situation, holiday, month, working day, and weekday were the categorical variables in the dataset. These variables influenced our dependent variable in the following ways:

1. **Season**: The boxplot analysis revealed that the spring season exhibited the lowest rental count, whereas the fall season recorded the highest. Rentals during the summer and winter seasons fell within an intermediate range, positioned between the extremes of spring and fall.

2. **Weather Conditions**: No rentals were observed during periods of heavy rain or snow, suggesting that such weather conditions are deemed highly unfavourable for outdoor activities. In contrast, the highest rental counts were recorded under conditions of 'Clear' or 'Partly Cloudy' weather, indicating that users tend to prefer more favourable and stable weather.

3. **Holidays**: Rental activity was found to be lower during holiday periods, which may reflect changes in consumer behaviour or external factors such as travel patterns, with individuals possibly opting for vacation or other commitments during these times.

4. **Month**: September emerged as the month with the highest rental activity, while December saw the lowest. This pattern is consistent with typical weather trends, as December is characterized by colder, snowier conditions, which likely deter rentals compared to the more temperate conditions in September.

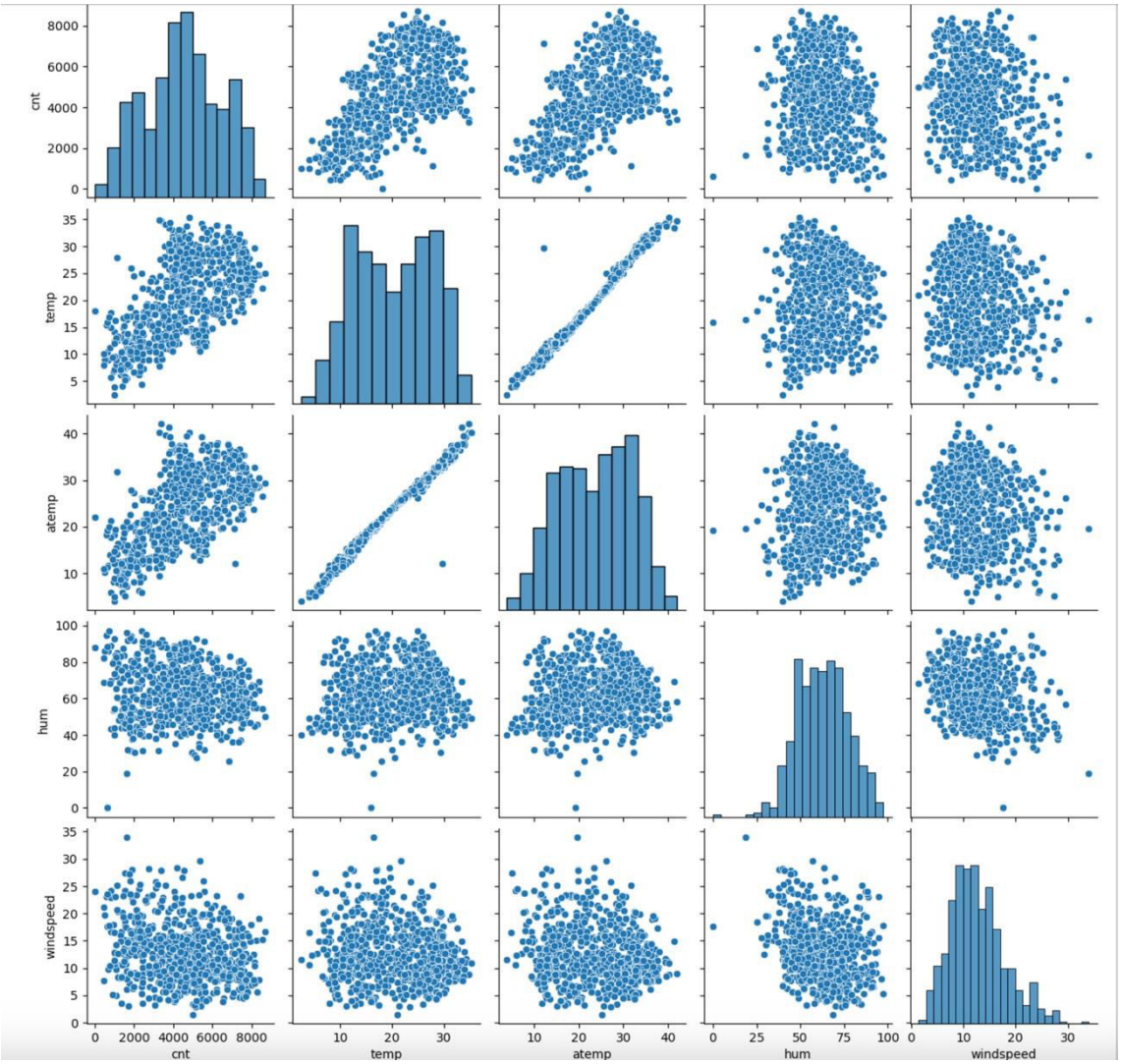
5. **Weekday**: A marked increase in rental bookings was observed on weekends compared to weekdays, suggesting that users are more inclined to make reservations during the latter part of the week, potentially due to greater leisure time or a preference for weekend activities.

6. **Working Day**: The variable "working day" appeared to have a negligible effect on rental behaviour, indicating that the distinction between weekdays and weekends had a more significant impact on rental activity than whether a day was classified as a typical workday.

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

Answer: Failing to remove the first dummy variable when creating categorical features can introduce correlation between the dummy variables, which may negatively impact certain models. This issue becomes particularly pronounced when the cardinality of the categorical variable is low. In iterative models, such as gradient-based algorithms, this multicollinearity can hinder the convergence of the algorithm, leading to instability during training. Additionally, including all dummy variables can distort the interpretation of variable importance, as the redundant information can cause misleading or inflated significance for certain features. By dropping the first dummy variable, we mitigate these issues and maintain the integrity of the model, ensuring a more stable and interpretable result.

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



The variables *'temp'* and *'atemp'* exhibit the highest correlation with the target variable *'cnt'* when compared to the other features in the dataset.

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

Answer: Linear regression models are assessed and validated based on several key assumptions: the **linearity** of the relationship between the predictors and the target variable, the **absence of autocorrelation** in the residuals, the **normality** of the residuals, the **homoscedasticity** (constant variance) of the residuals, and the absence of **multicollinearity** among the independent variables **(Low VIF)**. These assumptions are critical for ensuring the reliability, interpretability, and accuracy of the model's estimates and predictions.

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

Answer: Top 3 features that has significant impact towards explaining the demand of the shared are:

1. The coefficient for **Temperature (temp)** is **0.2360**, suggesting that a one-unit increase in temperature leads to an increase of **0.2360 units** in bike hire demand. This indicates that higher temperatures positively influence bike rentals, making it a key factor in predicting demand.

2. The coefficient for **Weather (light snow)** is **-0.2413**, implying that, relative to **clear weather**, a one-unit increase in the occurrence of light snow results in a decrease of **0.2413 units** in bike hire numbers. This suggests that adverse weather conditions, such as light snow, have a negative impact on bike demand.

3. The coefficient for **Year** is also **0.2360**, meaning that as the year progresses, bike hire numbers increase by **0.2360 units** per unit increase in the year variable. This could reflect growing demand over time, possibly due to increased awareness, availability, or infrastructure improvements.

Therefore, **Temperature**, **Weather (light snow)**, and **Year** emerge as the most significant predictors of bike hire demand, and should be prioritized when planning for bike rental operations to maximize bookings.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Answer: Linear regression is a statistical method used to model the relationship between a dependent variable (also known as the target or output) and one or more independent variables (also known as features or predictors).

The simplest form, **simple linear regression**, involves a single independent variable, while **multiple linear regression** involves two or more independent variables.

In simple linear regression, the model assumes the relationship between the dependent variable y and the independent variable xxx is linear and can be expressed by the equation:

$$y=\beta 0+\beta 1x+\epsilon$$

Here, y is the predicted output, β0 is the intercept (the value of y when x=0), β1 is the slope (the rate of change in y with respect to x), and ϵ represents the error term, which accounts for any deviations or noise in the data.

The goal of linear regression is to determine the values of β0 and β1 that minimize the difference between the predicted values y^ and the actual observed values of y. This is typically done using **Ordinary Least Squares (OLS)**, which minimizes the sum of squared residuals (the vertical distances between the actual and predicted points).

For multiple linear regression, the model extends to:

$$y=\beta0+\beta1x1+\beta2x2+\cdots+\beta nxn+\epsilon$$

Where x1,x2,….,xn are multiple independent variables. The parameters β0,β1,…,βn are again estimated using OLS.

Linear regression is widely used due to its simplicity and interpretability, though it assumes that the relationship between the variables is linear, the residuals are normally distributed, and the variance of the errors is constant (homoscedasticity).

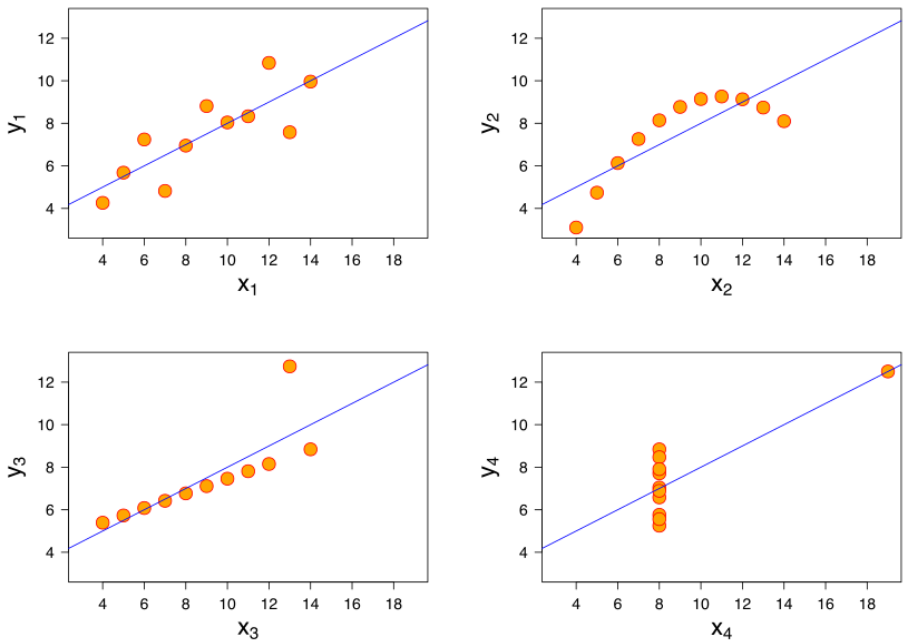
2. Explain the Anscombe's quartet in detail.

Answer: Anscombe's Quartet is a set of four datasets that are identical in key statistical properties but differ significantly in their visual appearance and underlying data patterns. The quartet, introduced by statistician Francis Anscombe in 1973, highlights the importance of visualizing data to better understand its structure, as statistics alone (such as means, variances, and correlation) can be misleading.

The four datasets in Anscombe's Quartet each consist of 11 data points with two variables, x and y. Despite having nearly

identical summary statistics - such as the same mean and variance for both x and y, as well as the same correlation coefficient (0.816) between x and y -the datasets differ dramatically in their distributions and relationships between the variables.

- **Dataset 1:** This dataset shows a strong linear relationship between x and y, as expected from the correlation, and is best fitted by a straight line.
- **Dataset 2:** Despite having the same correlation as Dataset 1, the data points appear to follow a nonlinear pattern, resembling a curve.
- **Dataset 3:** In this dataset, the relationship is also nonlinear, but with one extreme outlier that heavily influences the correlation and creates a misleading impression of a linear relationship.
- **Dataset 4:** This dataset features a perfect linear relationship with a significant outlier that skews the visual appearance of the data, even though the summary statistics are similar to the other datasets.



Anscombe's Quartet emphasizes the crucial point that summary statistics, such as mean, variance, and correlation, do not fully capture the complexities of the data. It underscores the importance of visualizing data through scatter plots to uncover patterns, outliers, and other nuances that may not be apparent from statistical analysis alone.

3. What is Pearson's R?

Answer: **Pearson's r** (also known as the **Pearson correlation coefficient**) is a statistical measure that evaluates the **strength** and **direction** of the **linear relationship** between two continuous variables. It provides a value between **-1 and 1**:

- **r = 1**: Perfect positive linear correlation. As xxx increases, y increases in a perfectly linear fashion.
- **r = −1**: Perfect negative linear correlation. As x increases, y decreases in a perfectly linear fashion.
- **r = 0**: No linear correlation. There is no consistent linear relationship between xxx and y.
- **0 < r < 1**: A positive linear correlation, with values closer to 1 indicating a stronger positive relationship.
- **−1 < r < 0**: A negative linear correlation, with values closer to -1 indicating a stronger negative relationship.

Formula:

The formula for Pearson's r is:

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

Where:

- xi and yi represent the individual data points for the two variables x and y,
- $\bar{x}$ and $\bar{y}$ are the means of the variables x and y, respectively.

The formula computes the **covariance** between the variables and then divides it by the **product of their standard deviations**. This normalization ensures that the correlation is always between -1 and 1, making it dimensionless and easy to interpret.

Limitations:

- **Linearity**: Pearson's r only detects **linear relationships**. It may not be effective if the data follows a **curved** or **nonlinear pattern**.
- **Outliers**: Pearson's r is sensitive to outliers, which can skew the correlation significantly, giving a misleading impression of the relationship.

Important Notes:

- **Linearity**: Pearson's r only measures linear relationships. It may not accurately represent associations that are nonlinear.
- **Sensitivity to Outliers**: Pearson's r is sensitive to outliers, which can significantly distort the correlation.
- **Interpretation**: While Pearson's r measures the strength and direction of a linear relationship, it does **not** imply causality between the variables.

In summary, Pearson's r is a widely-used statistic for evaluating the degree to which two variables are linearly

related, but its interpretation should be done cautiously, especially when the data exhibits nonlinearity or contains outliers.

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

Answer: Scaling refers to the process of adjusting or modifying something to accommodate changes in size, capacity, or scope. The term is used across various fields, such as business, technology, biology, and mathematics, but it generally involves increasing or decreasing the resources or capacity of a system in order to improve its performance, accommodate growth, or handle greater demand.

Key Aspects of Scaling:

1. **Capacity Adjustment**: Scaling involves increasing or decreasing the capacity of resources (e.g., hardware, software, workforce) to match the demands placed on them.
2. **Handling Growth**: Whether it's increased traffic to a website, more data to process, or higher production levels in a factory, scaling ensures the system or business can handle growth.
3. **Optimizing Performance**: Scaling may also be used to improve or maintain system performance as demand fluctuates.

Why is scaling performed? Scaling is performed to ensure that systems can handle increased demand or workload without performance degradation, downtime, or failure. As businesses

grow, websites get more traffic, or applications handle more data, scaling allows systems to accommodate these changes. It also ensures that resources are utilized efficiently, preventing both under-provisioning (leading to slowdowns) and over-provisioning (resulting in wasted costs). In essence, scaling supports growth, maintains system reliability, and optimizes resource use, enabling organizations to deliver seamless services to user.

Below is the difference between normalized scaling and standardized scaling:

S.NO.	Normalized scaling	Standardized scaling
1.	Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
2.	It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
3.	Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
4.	It is really affected by outliers.	It is much less affected by outliers.
5.	Scikit-Learn provides a transformer called MinMaxScaler for Normalization.	Scikit-Learn provides a transformer called StandardScaler for standardization.

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

Answer: The **Variance Inflation Factor (VIF)** is a measure used to assess how much the variance (or "spread") of a regression coefficient is inflated due to multicollinearity in the data. In simple terms, it tells us if any of the independent variables in a regression model are highly correlated with each other, which can cause problems in the analysis.

A **VIF can become infinite** when there is **perfect multicollinearity** between two or more independent variables. This means that one variable is an exact linear combination of others, so you can predict one variable perfectly using the others. When this happens, the regression model cannot distinguish between the variables, causing issues with estimating the coefficients.

1. VIF for a variable is calculated as:

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

where R2 is the coefficient of determination from a regression of that variable on all other independent variables. When R2 = 1, meaning that one variable can be exactly predicted from the others, the denominator becomes zero, causing the VIF to become infinite.

For example, imagine you have two variables, X1 and X2, and they are perfectly correlated (say, $X2 = 2 \times X1$). In this case, VIF for X1 or X2 can become infinite because the model cannot tell which of the two is driving the effect.

In practical terms, if VIF is infinite, it suggests that you have **perfect multicollinearity**, meaning the model has redundant predictors, and this should be addressed by removing one of the correlated variables to improve the model's stability and interpretability.

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

Answer: A **Q-Q (Quantile-Quantile) plot** is a graphical tool used to assess if a dataset follows a particular distribution, typically the **normal distribution**. It compares the quantiles of the data against the quantiles of the reference distribution (e.g., normal distribution). If the points on the plot fall approximately along a straight line, it indicates that the data likely follows the reference distribution.

Use and Importance in Linear Regression:

In Linear Regression when we have a train and test dataset then we can create Q-Q plot by which we can confirm that both the data train and test data set are from the population with the same distribution or not.

Advantages:

- It can be used with sample size also.
- Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot Q-Q plot use on two datasets to check.
- If both datasets came from population with common distribution.
- If both datasets have common location and common scale.
- If both datasets have similar type of distribution shape.
- If both datasets have tail behaviour.