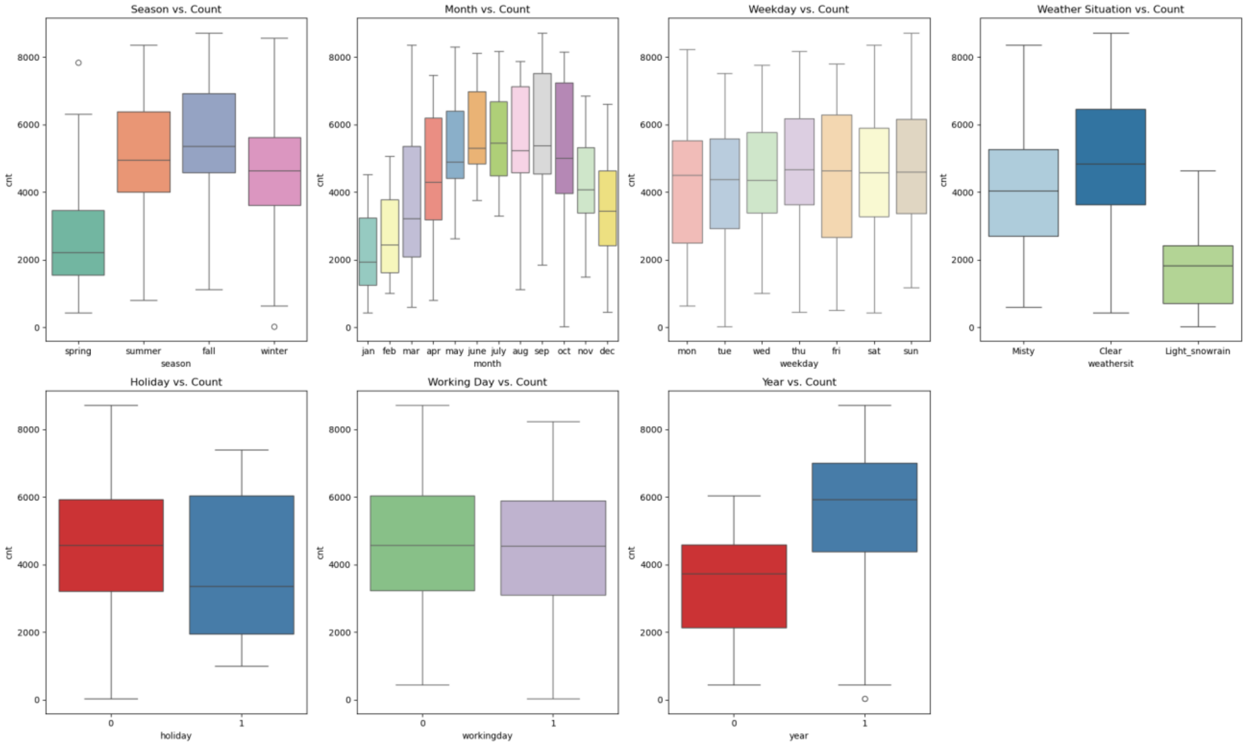
-Assignment-Based Subjective Questions-

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

The dataset includes categorical variables that significantly influence bike rental counts: **Season** and **Month** capture the impact of different times of the year, **Year** tracks trends over time, **Weekday** and **Working day** differentiate between work and leisure periods, and **Weathersit** describes how various weather conditions affect rentals. Each variable provides insights into the patterns and trends in bike usage.



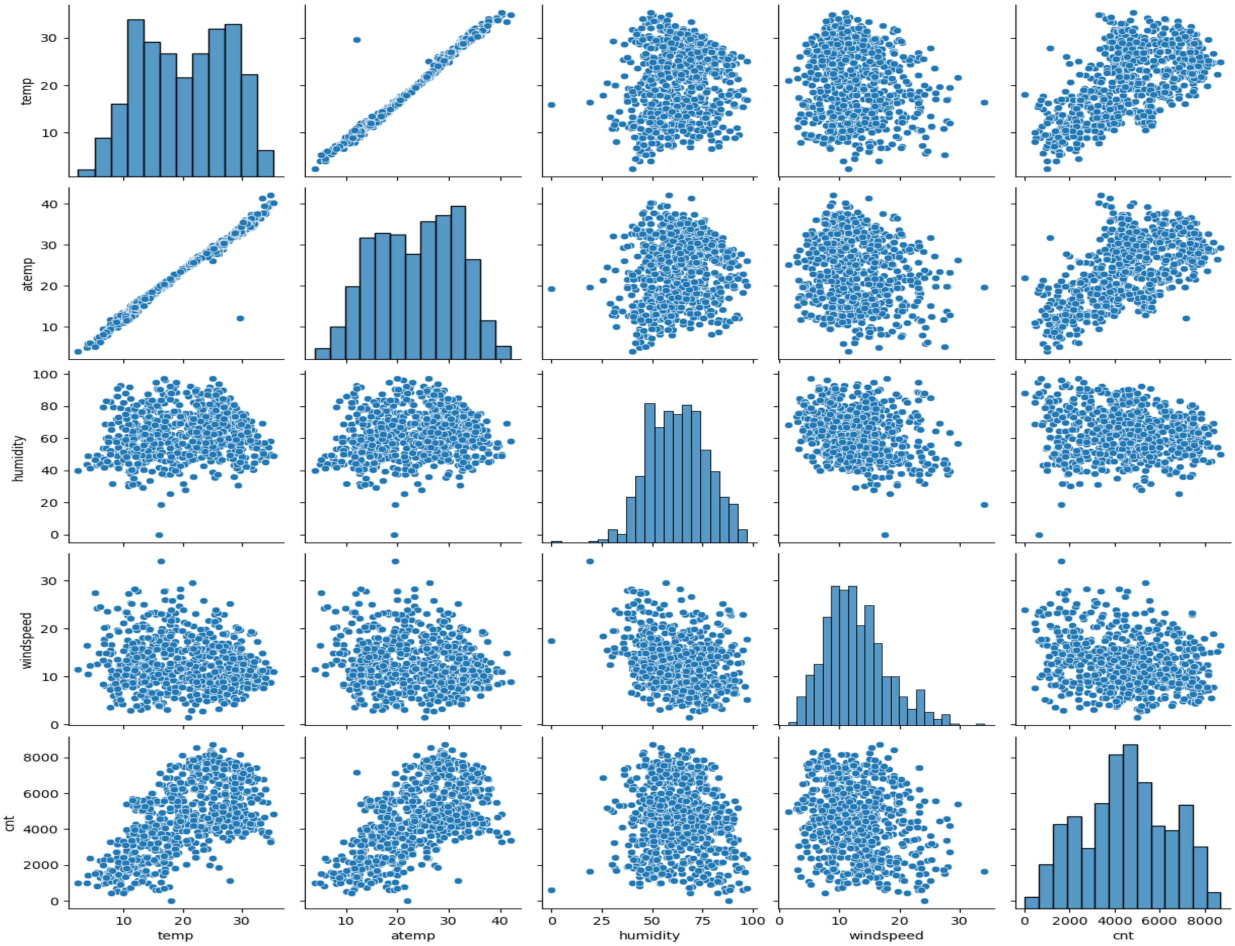
- These variables are visualized using bar plot and Box plot both.

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

Using `drop_first=True` during dummy variable creation is important to avoid the **dummy variable trap**, which occurs when including all dummy variables for a categorical feature leads to multicollinearity. Multicollinearity happens when one variable can be perfectly predicted from the others, making the regression coefficients unstable and difficult to interpret.

By dropping the first category, you create a reference category against which the other categories are compared. This ensures that the dummy variables are linearly independent and prevents redundancy in the regression model.

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



The ‘temp’ and ‘atemp’ variables have highest correlation when compared to the rest with target variable as ‘cnt’.

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

Linear regression models are validated by ensuring **linearity**, where the relationship between predictors and the dependent variable is linear; **no autocorrelation**, where residuals are independent; **normality of errors**, where residuals follow a normal distribution;

homoscedasticity, where residuals have constant variance; and **multicollinearity**, where predictors are not highly correlated with each other. Each assumption is checked through various diagnostic tests and plots to ensure the model's reliability and validity.

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

The top three features significantly impacting bike rental demand are **temperature**, which influences weather comfort; **year**, reflecting trends and changes over time; and **season**, which captures seasonal variations in bike usage. Each of these features plays a crucial role in explaining the variability in bike rental counts.

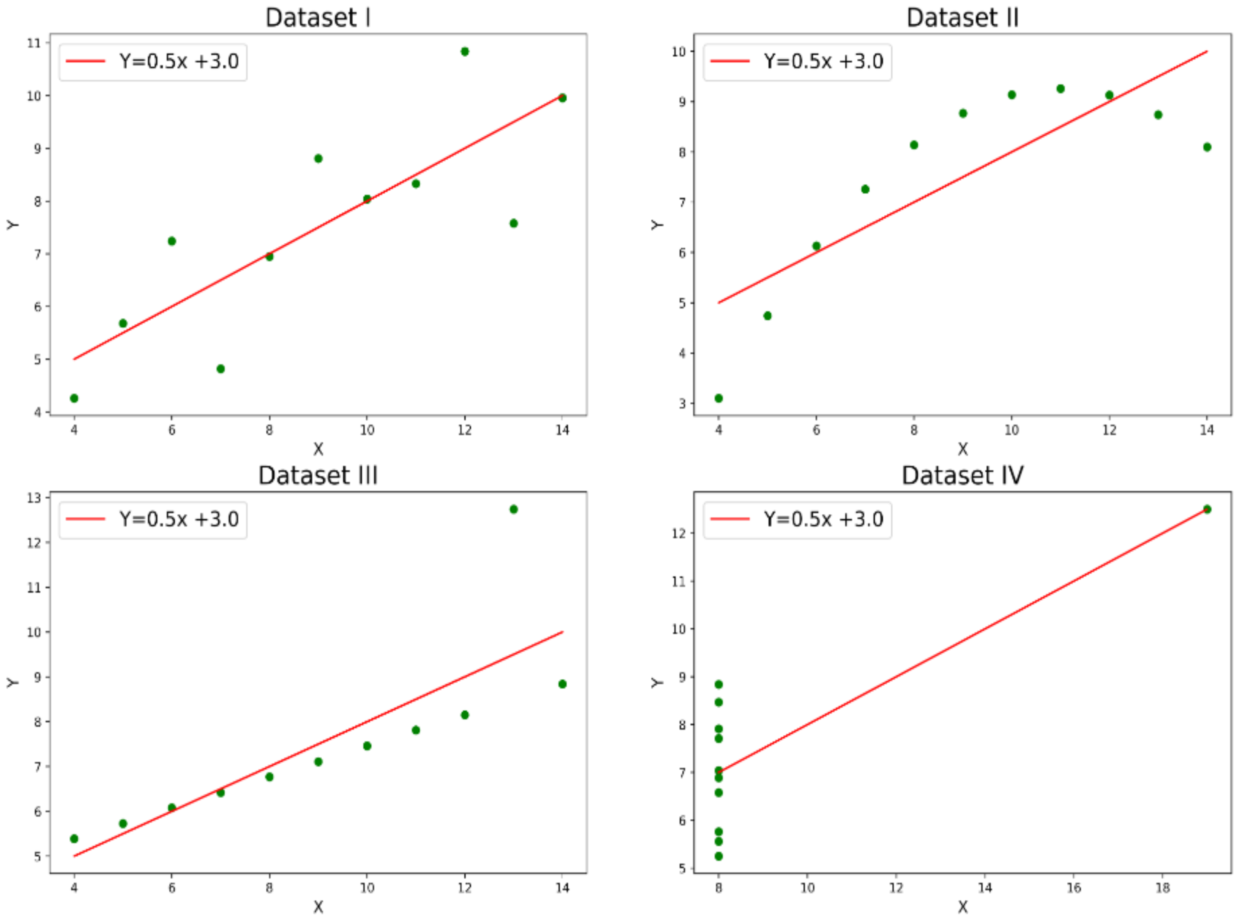
-General Subjective Questions-

1.Explain the linear regression algorithm in detail.

Linear regression models the relationship between a dependent variable YYY and independent variables X1,X2,…,XnX_1, X_2, \ldots, X_nX1,X2,…,Xn using a linear equation: Y=β0+β1X1+β2X2+…+βnXn+ϵY = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \epsilonY=β0+β1X1+β2X2+…+βnXn+ϵ. The objective is to estimate the coefficients (β\betaβ) that minimize the sum of squared residuals, achieved through the Ordinary Least Squares (OLS) method. Key assumptions include linearity, no autocorrelation, normality of residuals, homoscedasticity, and no multicollinearity. Model performance is assessed using metrics like R-squared and p-values, and diagnostic plots and tests are employed to validate assumptions and ensure model reliability.

2.Explain the Anscombe’s quartet in detail.

Anscombe’s Quartet consists of four datasets with identical statistical properties—mean, variance, correlation, and linear regression line—but distinct visual patterns when plotted. Created by Francis Anscombe in 1973, it demonstrates that relying solely on summary statistics can be misleading. For instance, one dataset shows a clear linear trend, another has a perfect line with an influential outlier, a third has a non-linear pattern, and the fourth features a linear relationship with a significant outlier. The quartet emphasizes the importance of visualizing data to accurately understand its underlying patterns and avoid misinterpretation.



Dataset I

Dataset II

Dataset III

Dataset IV

Explanation of this output:

- In the first one(top left) if you look at the scatter plot you will see that there seems to be a linear relationship between x and y.
- In the second one(top right) if you look at this figure you can conclude that there is a non-linear relationship between x and y.
- In the third one(bottom left) you can say when there is a perfect linear relationship for all the data points except one which seems to be an outlier which is indicated be far away from that line.
- Finally, the fourth one(bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient.

3.What is Pearson’s R?

Pearson’s R, also known as Pearson’s correlation coefficient, measures the strength and direction of the linear relationship between two continuous variables. It ranges from -1 to 1, where +1 indicates a perfect positive linear relationship, -1 signifies a perfect negative linear relationship, and 0 implies no linear relationship. The formula for Pearson’s R is:

r=Cov(X,Y)σXσYr = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}r=σXσYCov(X,Y)

where Cov(X,Y)\text{Cov}(X, Y)Cov(X,Y) is the covariance between variables XXX and YYY, and σX\sigma_XσX and σY\sigma_YσY are the standard deviations of XXX and YYY, respectively. Pearson’s R helps to assess how well the relationship between two variables can be described by a straight line.

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

Scaling refers to the process of adjusting the range and distribution of features in a dataset to make them comparable and suitable for analysis, particularly in machine learning models. Scaling is performed to ensure that features contribute equally to the model and to improve algorithm performance, especially for methods sensitive to feature magnitudes like gradient descent or distance-based algorithms.

Scaling is performed to ensure that all features in a dataset contribute equally to the model, particularly for algorithms sensitive to feature magnitudes like gradient descent or distance-based methods. It helps improve model performance and convergence speed by bringing features onto a comparable scale, preventing any one feature from dominating due to its magnitude. Additionally, scaling reduces numerical issues, such as instability or overflow, that can arise from features with vastly different ranges.

Normalized scaling (Min-Max scaling) adjusts feature values to fit within a specified range, typically 0 to 1, using the formula:

Xnorm=X−XminXmax−XminX_{\text{norm}} = \frac{X - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}}Xnorm=Xmax−XminX−Xmin

Standardized scaling (Z-score normalization) transforms feature values to have a mean of 0 and a standard deviation of 1, using the formula:

Xstd=X−μσX_{\text{std}} = \frac{X - \mu}{\sigma}Xstd=σX−μ

where μ\muμ is the mean and σ\sigmaσ is the standard deviation. The key difference is that normalization scales data to a fixed range, while standardization centers data around the mean and scales based on variance, making it useful when features have different units or magnitudes.

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

A Variance Inflation Factor (VIF) becomes infinite when there is **perfect multicollinearity** among predictors, meaning one or more predictors are exact linear combinations of others. This perfect collinearity causes the R2R^2R2 value for the predictor's regression on other predictors to be 1, leading to a zero denominator in the VIF formula:

VIFi=11−Ri2\text{VIF}_i = \frac{1}{1 - R^2_i}VIFi=1−Ri21

When Ri2=1R^2_i = 1Ri2=1, the denominator is zero, resulting in an infinite VIF. This indicates that the predictor is highly redundant and suggests the need to address multicollinearity by removing or combining variables.

A VIF becomes infinite due to **perfect multicollinearity**, where one predictor is a perfect linear combination of others, causing R2R^2R2 to be 1. This results in a zero denominator in the VIF formula, making the value infinite. This indicates the predictor is redundant and should be addressed.

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** is a graphical tool used to assess if a dataset follows a specified theoretical distribution, such as the normal distribution. It compares the quantiles of the dataset against the quantiles of the theoretical distribution.

1. **Plot Construction**: To create a Q-Q plot, the quantiles of the sample data are plotted against the corresponding quantiles of the theoretical distribution.
2. **Interpretation**: If the data follows the theoretical distribution, the points on the plot will approximately lie on a straight line (typically the line y=xy = xy=x).
3. **Deviations**: Deviations from the straight line indicate departures from the theoretical distribution. For instance, systematic curves may suggest a different distribution shape.

4. **Application**: Commonly used to assess normality in residuals for regression models, helping to validate assumptions about the data distribution.

A Q-Q (Quantile-Quantile) plot is used in linear regression to assess if the residuals follow a normal distribution, a key assumption for valid model results. By plotting the quantiles of residuals against the quantiles of a normal distribution, deviations from the straight line can reveal departures from normality. This helps diagnose potential issues with the model and ensures accurate hypothesis testing and confidence intervals. If significant deviations are observed, it may prompt model adjustments or data transformations to better meet assumptions.