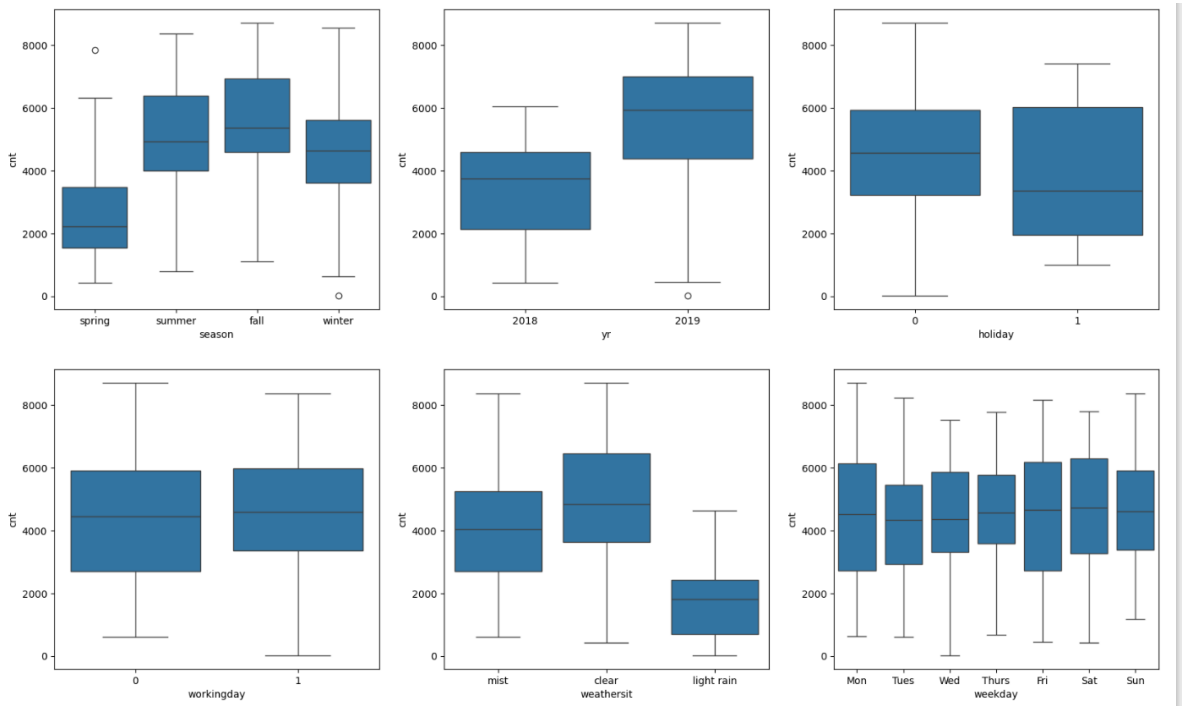
Assignment-based Subjective Questions

1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)**



The above is the image of the categorical relationships plotted with respect to the target variable cnt. Some of the main trends and observations inferred are :-

- Seasonal Effect: **Fall** shows the highest demand and **Spring** sees the lowest demand for bikes.
- Yearly Comparison: The year **2019** has significantly higher demand for bikes compared to **2018**
- Weather Conditions: **Clear weather** sees the highest bike demand and **Light rain** significantly reduces demand. This indicates a clear drop in usage during rainy conditions.

The other categories indicate little to no difference in bike demand between **working days** and **Non-holiday days** have slightly higher demand, though the difference is not very pronounced.

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

   A feature with n levels can be represented by n-1 number of dummy variables. For example, if you have a categorical variable "season" with four categories: spring, summer, fall, and winter, creating dummy variables would produce four columns, one for each season. However, if all four dummy variables are used in a model, they

introduce perfect multicollinearity because the value of one dummy variable can be inferred from the other three. For instance, if a row has 0 for spring, summer, and fall, it must be winter. **drop_first=True** removes one of the dummy columns (e.g., "spring"), effectively using it as the reference category. This eliminates multicollinearity. It reduces the number of features, which can improve model efficiency.

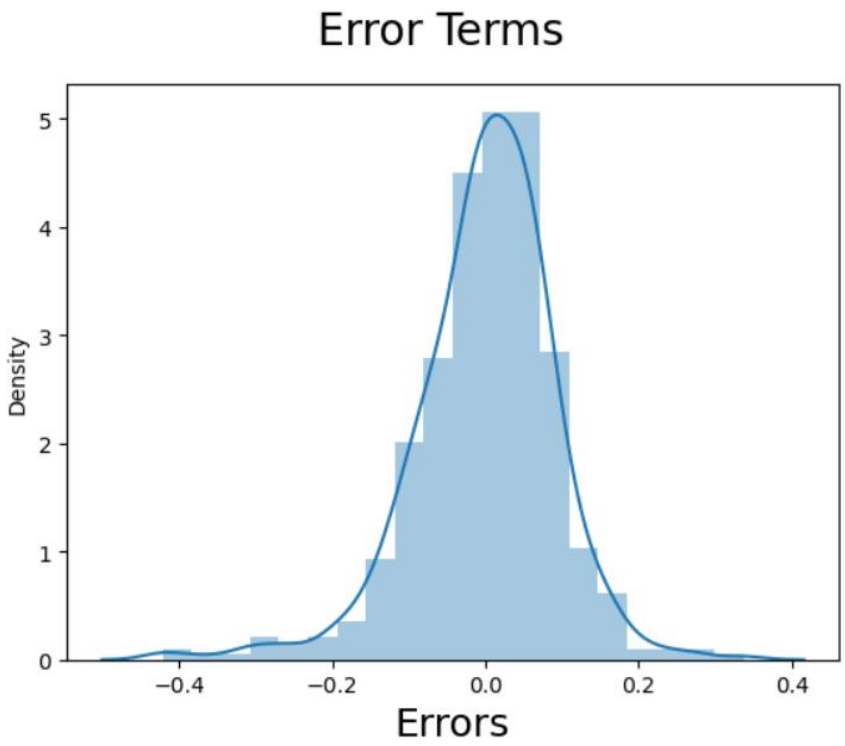
3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)**

Variable 'temp' and 'atemp' have the highest correlation with the demand of the bikes. With the correlation coefficient of 63%.

4. **How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)**

Linearity: Checked if the relationship between predictors and the target variable is linear by plotting predicted values vs. actual values. A random scatter of residuals indicates linearity.

Normality of Errors: Conducted residual analysis by plotting the residuals (error terms). The errors were verified to follow a normal distribution, ensuring the assumptions of linear regression were valid.



5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)**

- **Temperature (atemp)**: It has the highest positive coefficient (0.4461), indicating a strong positive relationship between temperature and bike demand.
- **Weather Condition (light rain)**: This has a significant negative impact, with a coefficient of -0.2842, showing that demand decreases sharply during light rain.
- **Year (2019)**: The positive coefficient (0.2365) suggests higher bike demand in 2019 compared to 2018, likely reflecting growth in bike-sharing usage.

These features play a major role in predicting the demand for shared bikes.

General Subjective Questions

1. **Explain the linear regression algorithm in detail. (4 marks)**
   Linear regression is a supervised learning algorithm used to predict a continuous target variable based on one or more predictor variables (features). The goal of the algorithm is to establish a linear relationship between the dependent variable (target) and independent variables (features).

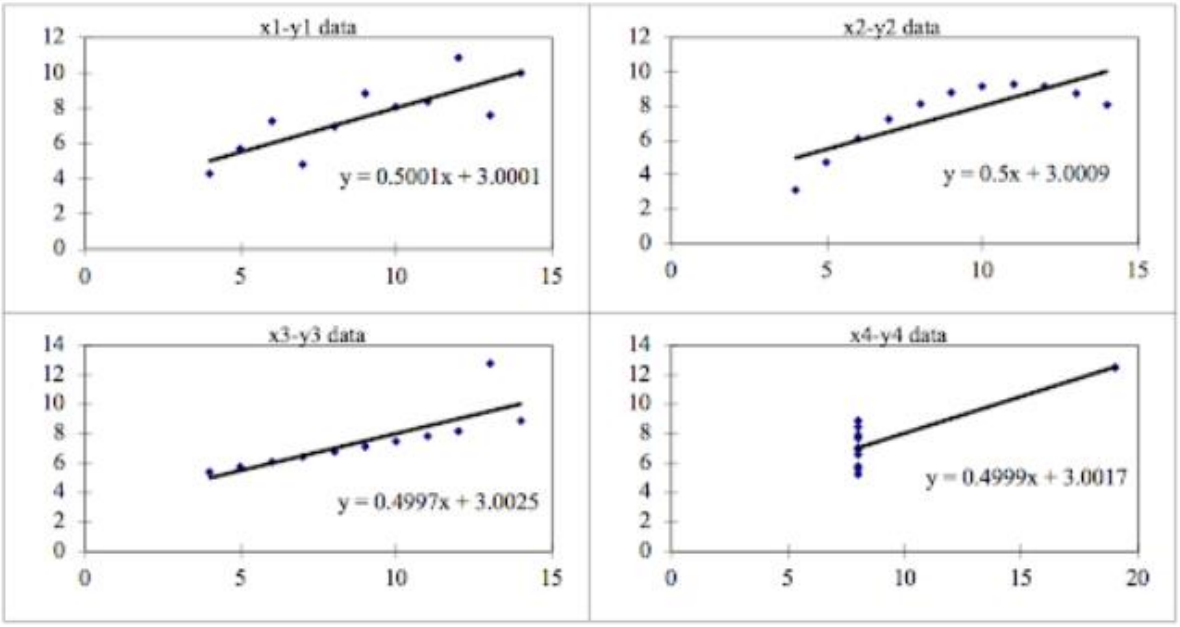
   In a **simple linear regression** model (with one predictor variable), the relationship between the dependent variable 'y' and the independent variable 'x' is modeled as: $y=\beta0+\beta1x+\epsilon$. Which is essentially the equation of a straight line.

   For **multiple linear regression**, the equation is extended to: $y=\beta0+\beta1x1+\beta2x2+\cdots+\beta nxn+\epsilon$. Here there are 'n' predictor variables.

   The algorithm aims to find the best-fitting line that minimize the distance (error) between the actual and predicted values by finding the optimal values of the coefficients $\beta0,\beta1,\ldots,\beta n$. This is done using a method **Ordinary Least Squares (OLS)** which directly computes the best coefficients by minimizing the residual sum of squares. It finds the values of the coefficients that minimize the total error.

   For linear regression to be reliable, certain assumptions must hold true:
   - The relationship between the independent variables and the dependent variable is linear.
   - The observations are independent of each other.
   - The variance of residuals (errors) is constant across all levels of the independent variables.
   - The residuals (errors) should follow a normal distribution.
   - Independent variables should not be highly correlated with each other.

2. **Explain the Anscombe's quartet in detail. (3 marks)**



Anscombe's quartet is a set of four datasets that have nearly identical simple descriptive statistics, such as the mean, variance, correlation, and linear regression line, yet appear very different when graphed. As you can see in the image above, the underlying relationships between the variables are very different in each dataset.

While the summary statistics suggest a linear relationship, a scatter plot reveals a clear non-linearity that would be missed without visualization in the 2nd data set. The outlier heavily influences the slope of the regression line, which would be misleading if the outlier were not visualized in the 3rd dataset. The 4th dataset shows that even with extreme outliers, statistical summaries might suggest a misleading linear relationship.

Anscombe's quartet is a powerful reminder that statistical analysis should not be based solely on numbers; visual exploration of data is crucial.

3. **What is Pearson's R? (3 marks)**

Pearson's R, also known as Pearson correlation coefficient, is a measure of the linear relationship between two variables. It quantifies how well the changes in one variable predict the changes in another. The value of Pearson's R ranges from -1 to 1:

- +1 indicates a perfect positive linear relationship (as one variable increases, the other increases proportionally).
- -1 indicates a perfect negative linear relationship (as one variable increases, the other decreases proportionally).
- 0 indicates no linear relationship.

Pearson's R only measures linear relationships and does not capture non-linear associations. It is commonly used in statistics to assess the strength and direction of relationships between variables.

4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)**
   Scaling is the process of adjusting the range of values in a dataset so that the data is comparable across different features.
   Scaling is important in machine learning because many algorithms are sensitive to the magnitude of input features. When features are on different scales, those with larger ranges can dominate the model's behavior, leading to biased results.
   Normalized Scaling (Min-Max Scaling): Compresses data to a specific range (e.g., 0 to 1), sensitive to outliers.

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

   Standardized Scaling: Centers data to have a mean of 0 and a standard deviation of 1, less sensitive to outliers.

$$X_{std} = \frac{X - \mu}{\sigma}$$

5. **You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)**
   The Variance Inflation Factor (VIF) is a measure used to quantify how much the variance of a regression coefficient is increased due to multicollinearity among the predictor variables.
   Infinite VIF values occur when there is perfect multicollinearity among the independent variables. This means that one or more independent variables can be expressed as an exact linear combination of other variables in the model.
   This can be rectified by removing redundant features or combining variables that are closely related into a single feature.

6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)**
   A Quantile-Quantile (Q-Q) plot is a graphical tool used to compare the quantiles of a sample distribution to the quantiles of a theoretical distribution (usually the normal distribution). In a Q-Q plot, each point represents a quantile from the sample data plotted against the corresponding quantile from the theoretical

distribution. If the data follows the theoretical distribution closely, the points will lie approximately along a straight diagonal line (y = x).

A Q-Q plot is essential for validating the normality assumption of residuals in linear regression, identifying outliers, and diagnosing the overall fit of the model. This helps ensure that the results and interpretations derived from the regression analysis are reliable and valid.