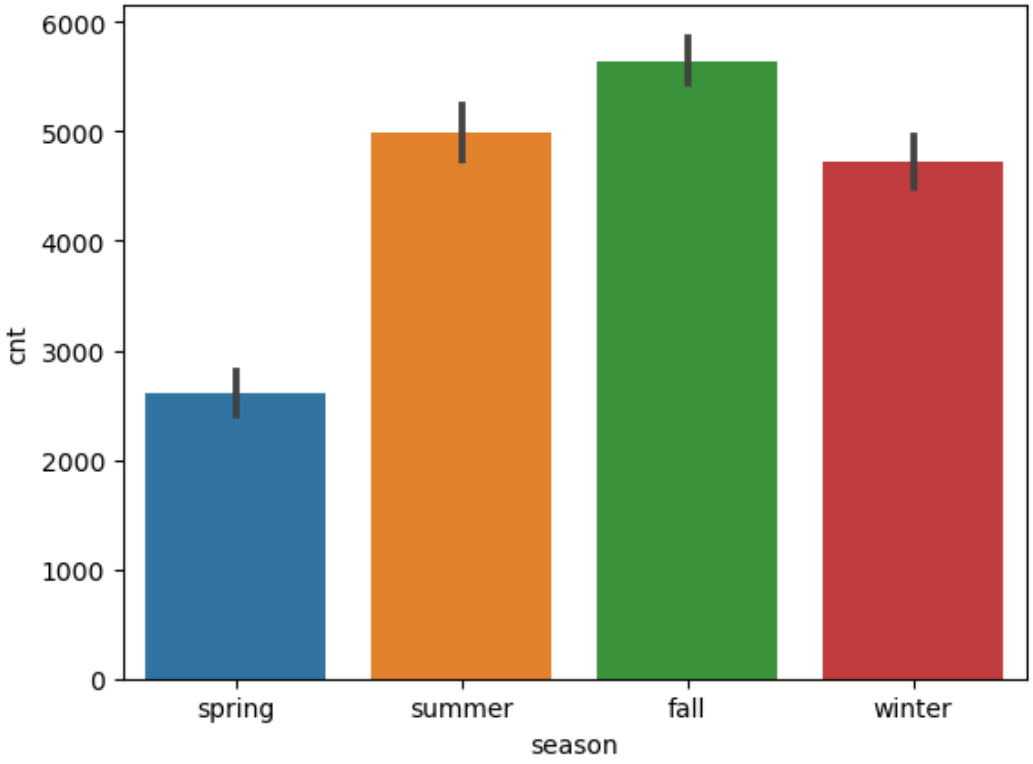
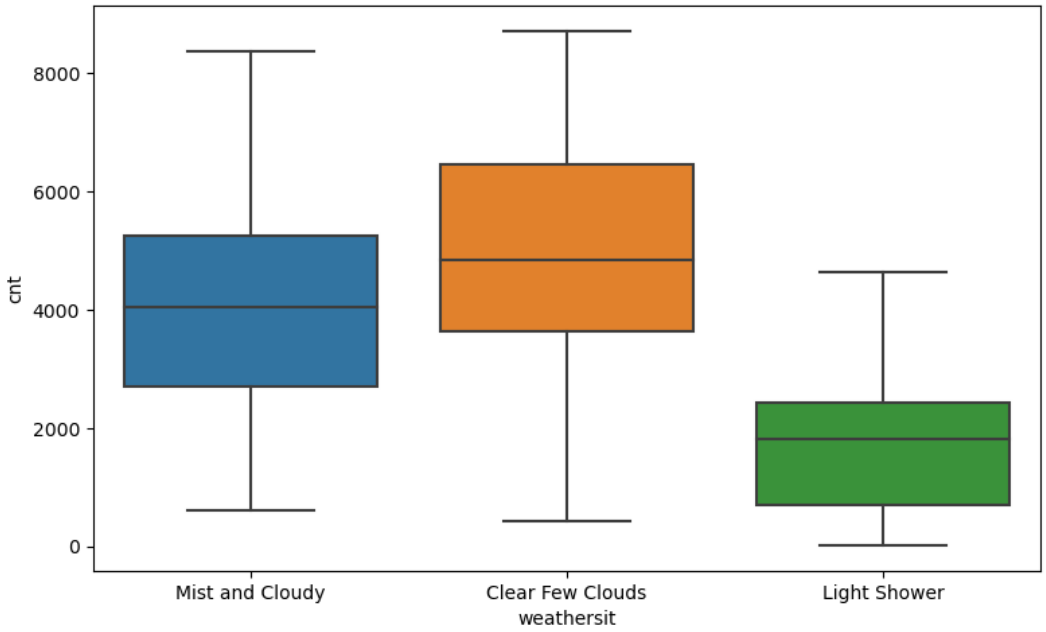
Assignment-based Subjective Questions

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

In the dataset given , we have **season** and **weathersit**, which were given as numeric values which we encoded using the data dictionary . After encoding the data we can say that When the weather is Clear Few Clouds and seasons – Fall and Spring having positive impact .

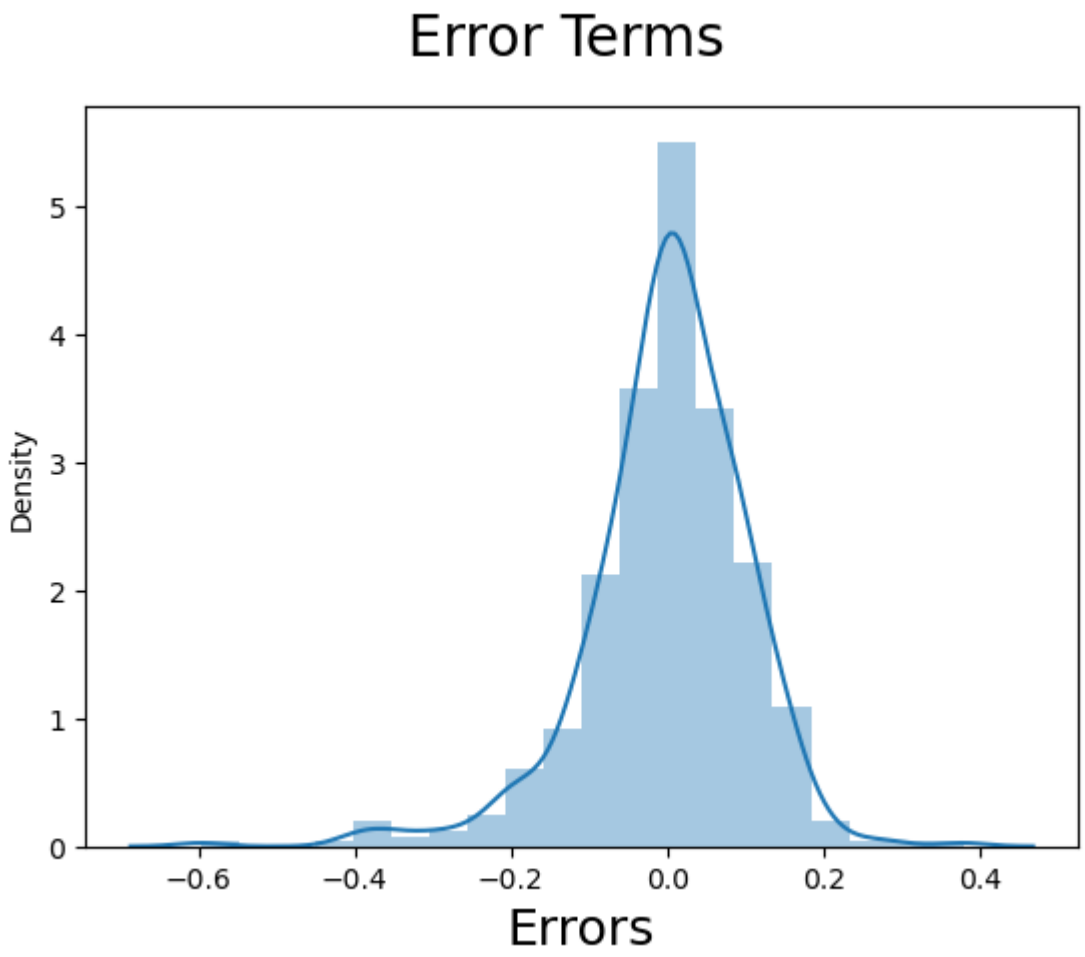




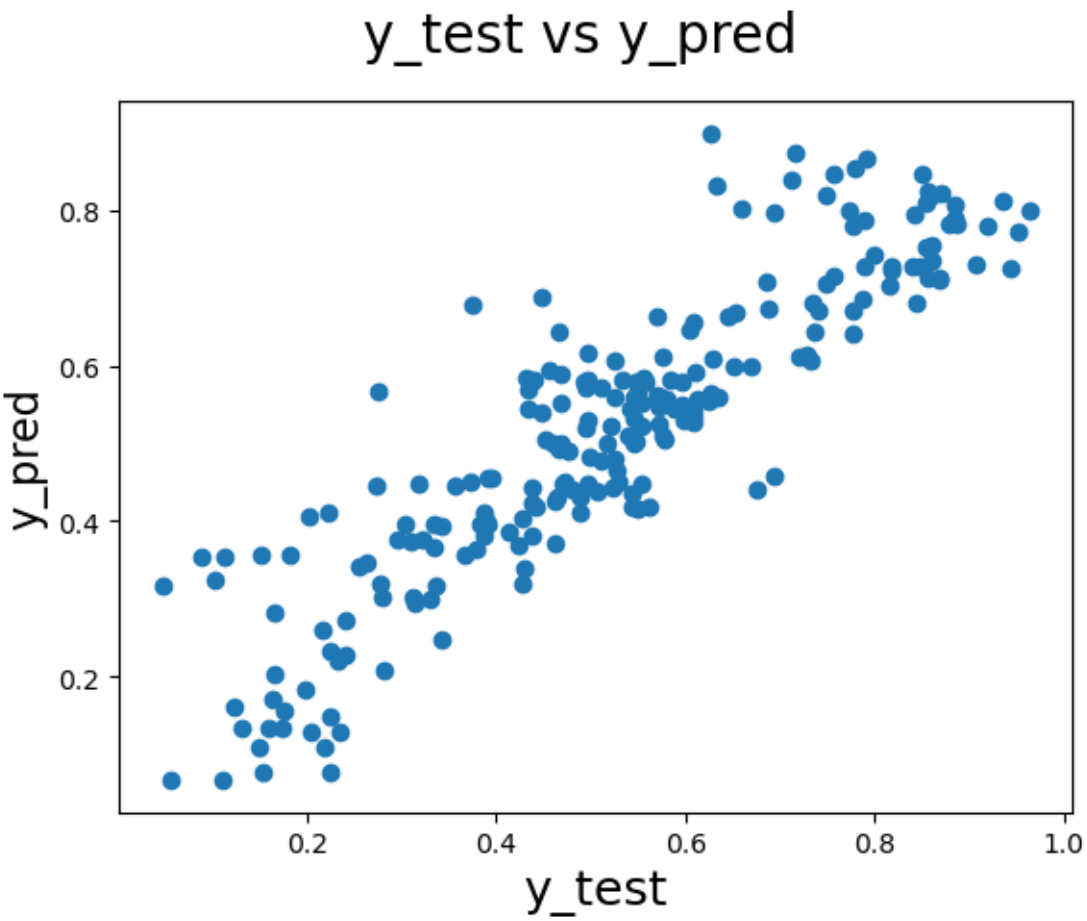
2. Why is it important to use drop_first=True during dummy variable creation?
    a. In general when we use drop_first to prevent multicollinearity and reducdency
    b. However I have not used drop_first as it was removing entire category and it was not giving any significate difference.
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

    Temp (Temparature) and atemp (feeling temperature in Celsius
) variables which are actually redundant are having highest correlation . I kept atemp and delete temp as I felt its how a person feels the temp is important than actual temperature

4. How did you validate the assumptions of Linear Regression after building the model on the training set?
    a. There is a linear correlation between dependent and independent variables. In our assignment Cnt (demand for bikes /no of bikes taken) is dependent variable and it has high correlation with atemp independent variable , which means if atemp increase demand also will increase
    b. Error Terms are Normally distributed with mean as 0



c. Error terms do not follow the pattern
d. Error Terms have constant variance – The R2 score is same ,Adjusted R2 is not having much difference



5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
    a. Atemp – Its having positive correlation indicating if the temperature increases demand also increase
    b. Clear Few Clouds – It also have the positive impact .If weather is good there is demand increase vise versa
    c. During the weekdays also there is higher demand than weekends and holidays.

    Apart from this we have seen positive correlation in the year also indicating the growth in demand compared to the previous it. As stated in the problem statement after the lockdown situation , as we seen increasing trend in demand year wise , we can predict increase in demand.

General Subjective Questions

1. **Explain the linear regression algorithm in detail. (4 marks)**
    a. Linear regression is supervised machine learning algorithm that computes the linear relationship between a dependent variable and one or more independent variables.
    b. In regression we will have lables and we will use continuous numeric values to create model basis which we can predict .

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

Anscombe's quartet is a set of four small datasets that have nearly identical simple descriptive statistics (mean, variance, correlation, etc.) but vastly different distributions and appear very different when graphed. It was created by the statistician Francis Anscombe in 1973 to emphasize the importance of graphing data before analyzing it and to demonstrate the impact of outliers and influential observations.

Its is often used to emphasize the dangers of relying solely on summary statistics and to promote the use of data visualization in exploring and understanding datasets.

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

Pearson's correlation coefficient, often denoted as "Pearson's R," is a statistical measure that quantifies the strength and direction of a linear relationship between two continuous variables. It assesses how well the relationship between the variables can be described using a straight line. The values of Pearson's R range from -1 to 1

i. -1 –perfect negative correlation – As one variable value increase other will decrease
ii. 0 - No linear correlation -- Variables are not linearly related
iii. 1 – perfect positive correlations – As one variable value increases other will also increase

Pearson's R is widely used in various fields such as psychology, sociology, economics, and natural sciences to analyze relationships between variables. However, it's important to note that Pearson's R measures only linear relationships and may not capture other types of relationships, such as non-linear or curvilinear associations.

4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)**

a. Scaling, in the context of data preprocessing in machine learning, refers to the process of transforming the features of a dataset so that they fall within a specific range or distribution. The goal of scaling is to make sure that each feature contributes equally to the computation of distances, similarities, and other mathematical operations in machine learning algorithms.

b. Scaling is performed for the following reasons:

i. **Algorithms Sensitivity:** Many machine learning algorithms are sensitive to the scale of the features. Algorithms like k-nearest neighbours, support vector machines, and neural networks can be influenced by the magnitude of the features, potentially leading to biased results.
ii. Convergence Speed: Some optimization algorithms converge faster when features are on a similar scale, ensuring that the algorithm reaches the optimal solution more efficiently.
iii. **Accuracy and Performance**: Scaling can improve the accuracy and performance of the model by ensuring that the features are comparable and contribute equally to the learning process.

c. **Normalized Scaling (Min-Max Scaling):**Range: Normalized scaling, also known as min-max scaling, scales the features to a specific range, typically [0, 1].

d. **Standardized Scaling (Z-score Normalization)**:Mean and Variance: Standardized scaling transforms the features to have a mean of 0 and a standard deviation of 1.

5. **You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)**
    a. When VIF (Variance Inflation Factor) is infinite due to perfect multicollinearity, it indicates a severe problem in the model. It makes the regression results unreliable and hard to interpret. To address this issue, you should identify and handle the linearly dependent variables by removing redundant variables or using alternative methods like dropping one level of a dummy variable set (dummy variable reference category) to avoid the dummy variable trap.
    b. We should always consider the VIF values <5

6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks**
    a. A Q-Q (Quantile-Quantile) plot is a graphical tool used to compare the distribution of a sample of data to a theoretical distribution, typically the normal distribution. It's a powerful visualization technique to assess the normality assumption of a dataset or to compare it against another specified distribution.
        i. The x-axis represents the quantiles of the theoretical distribution.
        ii. The y-axis represents the quantiles of the sample data.
    b. We can say a Q-Q plot is a valuable tool in linear regression for assessing the normality of residuals, identifying outliers, and comparing the distribution of data against a theoretical distribution. It provides a clear visual representation that aids in making informed decisions about the assumptions and validity of a linear regression model.