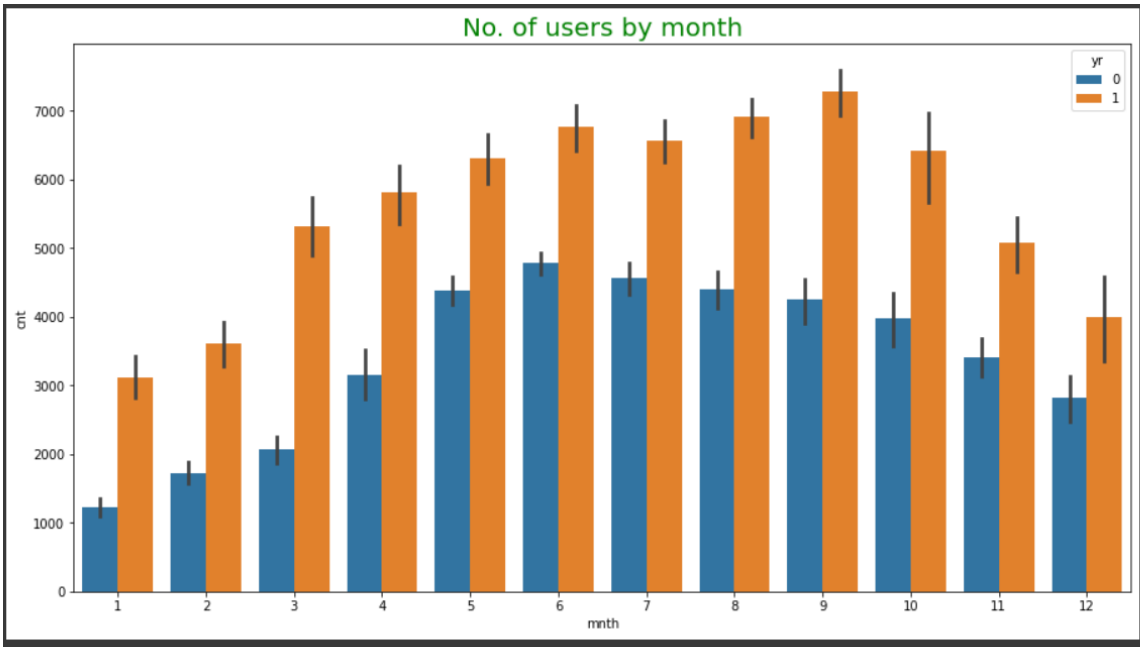
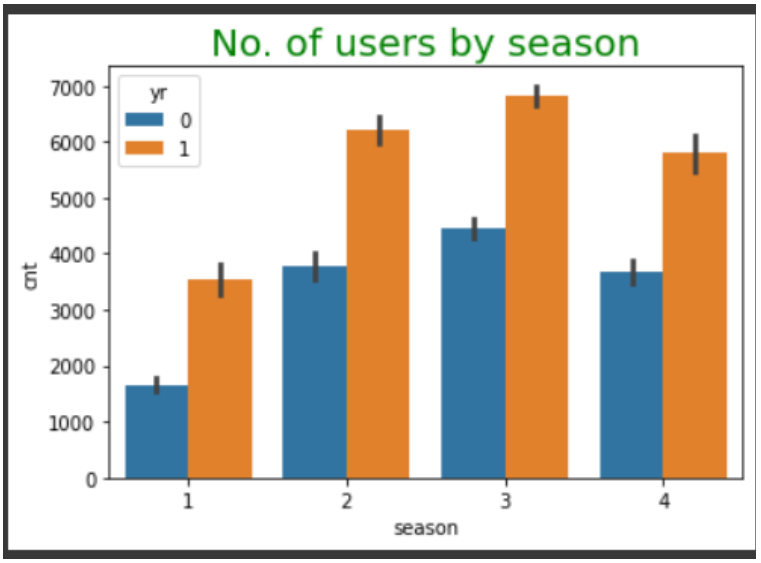
Assignment-based Subjective Questions

Q1: From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

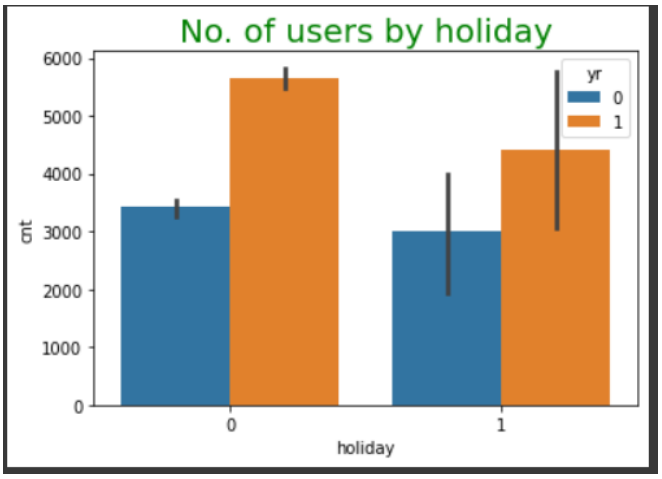
A: While analysing the month with the total count of customers, we can see that there is gradual increase in count as the year begins and a gradual decrease by the end of the year. We can also say that the customer count increases as summer comes up and decreases during winters.



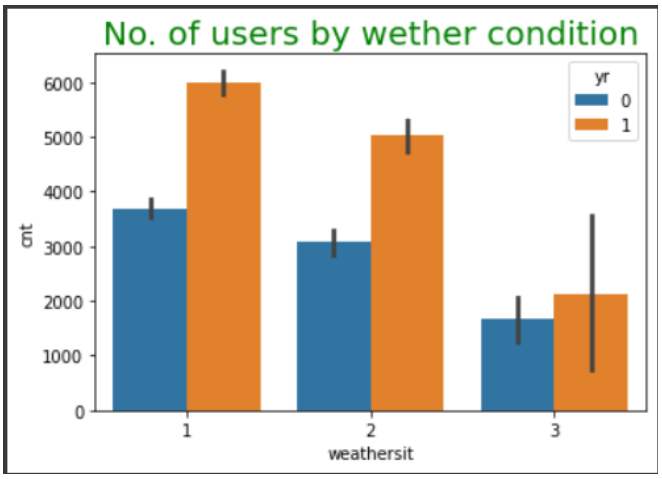
While analysing the season with the count of customers, we can see that most of the customers rented bikes in the fall and summer season.



While analysing holiday with the count of customers, we can see that most of the customers rented bikes when it's not a holiday.



While analysing weather with the count of customers, we can see that most of the customers rented bikes when it's Clear, Few clouds or Partly cloudy.



Q2: Why is it important to use drop_first=True during dummy variable creation?

A: We use binary numbers to create dummy variables.

Lets see with an example.

So suppose we have 50 **Fruits** of 3 different categories as **Apple, Banana, Orange** in a basket. When we pick a fruit we assign 1 for each category they fall in. So if i picked an **apple**, then apple will be **1** and the other two will be **0** for that fruit. Similarly if i pick a **banana**, then banana will be **1** and others will be **0**. So ultimately when I neither pick an **apple** nor a **banana**, it will be an **orange.** Orange will be **1** and the other two **0**.

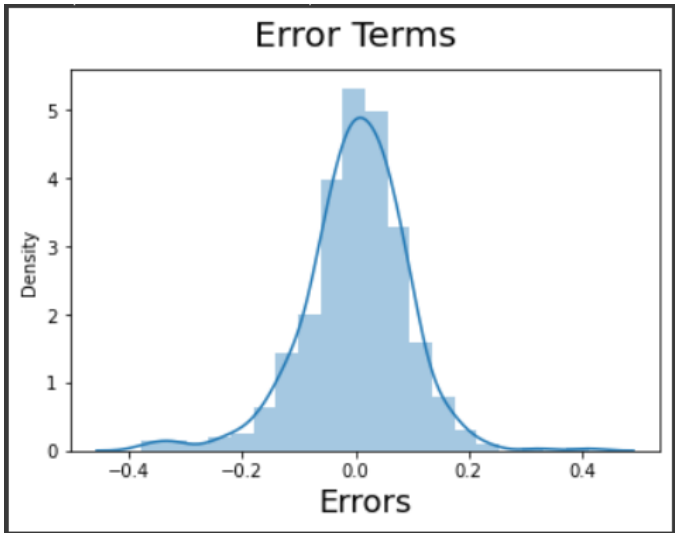
In a similar way we create dummy variables, we assign 1 to that category and 0 to the other categories and if all the categories are 0 then it will ultimately be the last category and hence we always create 1 less dummy variable than the categories in categorical variable and for that we need to pass drop_first=True while creating dummy variables.

Q3: Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

A: Looking at the pairplot, Variable atemp and temp has the highest correlation considering casual and registered as redundant variables as the target variable is sum of casual and registered.

Q4: How did you validate the assumptions of Linear Regression after building the model on the training set?

A: The assumptions of the Linear Regression model on the training set were validated by plotting the distribution graph of the errors. The errors are normally distributed with mean at approximately 0.



Q5: Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

A: Based on my final model, the top 3 features contributing significantly towards explaining the demand of the shared bikes are year and two dummy variables winter and september. Considering from which variable they came from the top 3 features are **Year, Month, Season.**

General Subjective Questions

Q1. Explain the linear regression algorithm in detail.

A:

- Linear Regression is a machine learning algorithm based on supervised learning.
- In this algorithm, the independent variables are used to predict the values of the dependent variables.
- The relationship between the independent and dependent variables are shown with the help of a straight line which is called the best fit straight line.
- The equation of this straight line is given as y=c+mx where m is the slope and c is the intercept.
- The equation of Linear Regression is given as y=c+m1x1+m2x2… and so on. Here m1 is the regression coefficient of the variable x1. Similarly is the case for m2x2 and all the variables in the dataset.

Q2. Explain the Anscombe's quartet in detail.

A:

- Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed.
- The purpose of Anscombe's quartet is to demonstrate both the importance of graphing data when analysing it, and the effect of outliers and other influential observations on statistical properties.
- Ansombe's quartet is the modal example to demonstrate the importance of data visualisation.

Q3. What is Pearson's R?

A: Pearson's R is a measure of linear correlation between two sets of data. It is the ratio between the covariance of two variables and the product of their standard deviation. It is essentially a normalised measurement of the covariance, such that the result always has a value between −1 and 1. As with covariance itself, the measure can only reflect a linear correlation of variables, and ignores many other types of relationships or correlations. Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations.
The formula of pearson's R is cov(X,Y)/std(X)*std(Y) where cov is covariance and std is standard deviation.

Q4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

A:

- Scaling is a technique to standardise the independent features present in the data in a fixed range.
- Scaling is performed to standardise all the continuous variable to same scale so the model does not create any bias due to higher values in some variables. This might even affect the accuracy as maybe some variables may have higher values than important variables.
- Normalized scaling scales the data to a value between 0 to 1 whereas Standardized scaling scales the data in such a way that the mean becomes 0 and standard deviation turns to be 1.

Q5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

A: VIF shows correlation between variables. A high VIF means the variables are highly correlated. VIF value to be infinite means that the variables are exactly same or in other words, a perfect correlation between variables.

Q6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

A: Q-Q plots are also known as Quantile-Quantile plots. As the name suggests, they plot the quantiles of a sample distribution against quantiles of a theoretical distribution.
QQ plots is very useful to determine

- If two populations are of the same distribution
- If residuals follow a normal distribution. Having a normal error term is an assumption in regression and we can verify if it's met using this.
- Skewness of distribution