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 categorical variables in the data set were season, weathersit_holiday, mnth,yr and weekday.

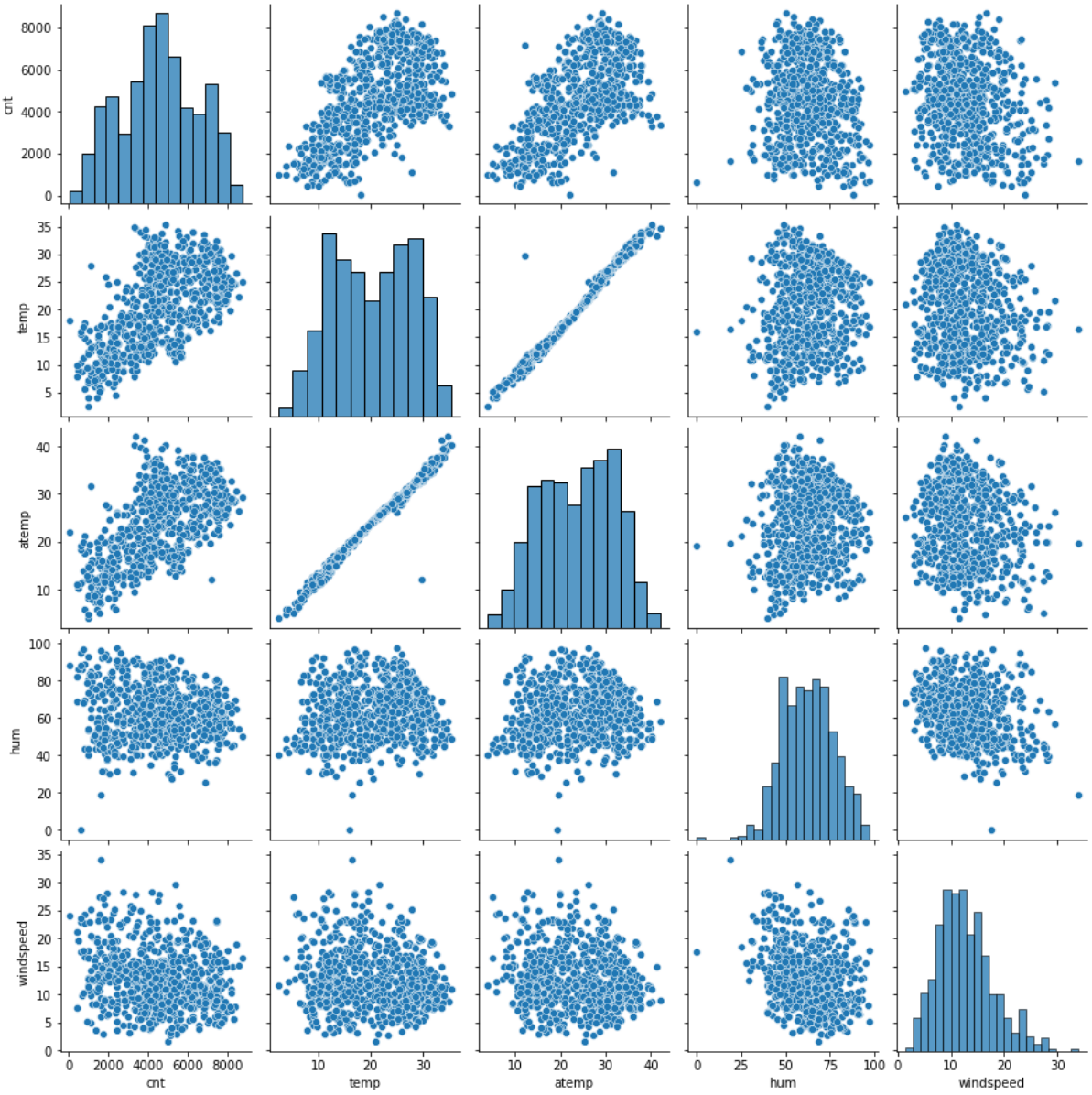
   1. Season - Using the plotted boxplot this variables had least value of cnt whereas fall had maximum value of cnt.Summer and winter had intermediate value of cnt.
   2. 2.Weathersit -The users are nil when the rain and snow is high.
   3. Holiday – Rentals reduce significantly during holiday
   4. Mnth – December had the least rentals & September ad no rentals. It could be due to the weather.
   5. Yr

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

   The dummy variables will be corelated and will affect the model.

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

   Temp & atemp are highly correlated.



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

The residual distribution shows a normal distribution & centred around 0. By plotting a distplot and see if residuals follow normal distribution.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes
   1.Temp – coefficient ; 0.491508
   2.yr = coefficient :0.233482
   3.weathersit_Light Snow & Rain – coefficient – 0.285155

General Subjective Questions

1. Explain the linear regression algorithm in detail.

   Linear Regression is a type of supervised Machine Learning algorithm that is used for the prediction of numeric values. Regression is the most basic for of regression analysis.

   We're trying to find the linear relation between depending & independatnt variables using a best fit line by minimizing the error terms using RSS respectively.

   There are 2 types

   Simple linear regression – one dependent variable & one independent variable.

   Multiple Linear regression – one dependent variable & multiple independent variables.

2. Explain the Anscombe's quartet in detail
3. What is Pearson's R?
4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling

   Scaling is used to normalize or standardize the range of independent variables or features of data. If scaling not done, ML algorithm tends to weigh greater values.

   Normalization is used when you know that the distribution of data doesn't follow Gaussian distribution. Standarization is used when the data follows Gaussian distribution.

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

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

   It's a plot of quantiles of the first data ser against the quantiles of the second data set. It is used to compare the shapes of distribution. It is a scatter plot that creates two sets of quantiles against one another.
   It can be used to determine the population within a common distribution, data sets having common location & scale.