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?
Ans- From the analysis it is clear that the variable **yr** share a very strong relationship with the dependent variable apart from this spring season has negative relation with the dependent variable which means if season is spring then cnt values are likely to reduce .Simlarly weathersit(Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds,Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist) negatively related to dependent variable

2. Why is it important to use drop_first=True during dummy variable creation?
Ans- Let's assume that we have a categorical variable with 3 unique values so this variable can be represent by 2 dummy variable not 3 the first one can be represented with 10 value ,second one with 01 and the remaining one can be represented with 00 .so we can represent the same amount information using n-1 dummy variable which help us to make data simpler .so when we create dummy variable it create a variable for each category in that column so we keep **drop_first=True** so that it create n-1 dummy variable for n category

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
Ans-temp and atemp both are highly correlated with the target variable

4. How did you validate the assumptions of Linear Regression after building the model on the training set?
Ans- We plot the histogram to check the distribution of error terms and then we plot the scatter plot to check the error distribution so the distribution come out to normal centered at 0 and the error term were randomly distributed and have constant variance

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
Ans- The top 3 features are atemp,yr and weathersit after that season is the next most important feature

General Subjective Questions

1.Explain the linear regression algorithm in detail.
Ans- Linear regression is an statistical method to predict the values for a continuous target variable .It is a supervised machine learning algorithm which has x values and corresponding ,y values by using this algorithm we try to find the linear relationship between the dependent variables and the independent variables .There are certain assumptions which are required to use the linear regression algorithm

1. Linear relationship between dependent variables and independent variables

2. Error terms are normally distributed
3. Error terms re independent of each other
4. Error term have constant variance or constant standard deviation

2.Explain the Anscombe's quartet in detail.
Ans- There are various distributions which are completely different but when we do a statistical analysis over it the statistical parameters come to be same for each of the distribution .so Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some dissimilarity in the dataset that **f**ools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots.

3.What is Pearson's R?
Ans- It is a statistical parameters which measures the linear correlation between two variables it has a numerical value that lies between -1.0 and +1.0. This help us to find the features during the feature selection process .if two features have high absolute value of pearson R then we can drop one of them .It not only give the direction of correlation but also the strength

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

Ans- It is a preprocessing step to bring all the variables in the same range which help the algorithm to reach to the final optimal point in less amount of time

Scaling is performed so that model can be interpreted in a better way and if we don't do scaling then the constants coming out from the model are very difficult to understand and their values depends on the variables magnitude

- **Normalized Scaling-***It brings all of the data in the range of 0 and 1. sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.*
- **standardized scaling-***Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).*

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
Ans- when the R value is 1 then VIF become infinite .This happens when the variable for which VIF is being calculated can be completely expressed with the help of other variables which means it is completely multicollinear with the other variables

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
Ans- A Q-Q plot is used to find the distribution of data i.e which distribution the data flows normal distribution or uniform distribution or any other distribution .It can be used to find the

distribution between 2 datasets also .so if 2 distributions are compared and plotted on Q-Q plot and if they are linearly dependent then they will fit to a straight line