**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?**

* here for weathersit with clear - Clear, few clouds, partly cloudy, partly cloudy as mentioned above highest demand for bikes.
* for weekdays almost all are nearly equal demand.
* And for seasons summer and fall are highest demand for bikes
* And for months May to September has high demand
* Weekdays are more in demand comparing to weekends

1. **Why is it important to use drop\_first=True during dummy variable creation?**

* it will create dummy variable one less than the categorical variable.
* Like if there are three values in a feature then it will create two dummy variables
* Like 00,01 and eliminate the third one.
* This helps in decreasing the redundancy.
* This helps in reducing the correlation among the dummy variables.

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

* 'temp' and 'atemp' has highest correlation with ‘cnt’ 0.63.
* ‘yr' has next highest 0.57

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

* We validate assumptions of linear regression on model by looking at the model summary.
* The variables should have the linear relation between them this can be validated by looking at the scatter plot.
* The p-values should be less than 0.05 and the more r2 value the more model is significant. we should also check for the VIF values for all the features of the model, they all should be below 5.

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

* As per the last model the features temp, yr and weathersit\_clear has three highest coefficients,
* Here weathersit\_clear is where weather is in following conditions:

“Clear, Few clouds, partly cloudy, partly cloudy”

* For 0.3709 unit increase in temp there will 0.3709 in target variable
* For 0.2342 increase in yr same increase in target var
* Same for weathersit\_clear there will be increase by 0.0924

**General subjective questions**

1. **Explain linear regression algorithm in detail?**

- Linear regression Algorithm is supervised Machine learning Algorithm. This is performed on the data where; we have Independent and dependent variable, and the dependent variable is continuous Variable.

- it determined it the two variables have best fit linear relation between them

* Here we will divide the data set to train and test data and crate model on train data and make predictions on test data.
* And we assume that there is kind of linear relationship between Independent and dependent variable for both the train and test data will similar significance
* There are two kinds of linear regression Algorithms

1. **Simple linear regression**: where there is one Dependent and one independent variable.

Formula is given by Y=β0+β1X1 +ϵ where β0 is intercept and β1 is slope

2. **Multiple linear regression:** here there can be multiple independent variables but only one dependent variable

Formula is given as : Y=β0+β1X1+β2X2+…+βpXp+ϵ

1. **Explain the Anscombe’s quartet in detail.**

* A group of data sets which seems to have same descriptive statistics can fool any machine learning or regression algorithm
* The datasets which seem to have same statistics but when the data is visualized and plotted on the scatter plots it shows the difference.
* Hence it is good to visualize any data before creating a machine algorithm to get a fit model.

1. **What is Pearson’s R?**

* It is one of the most important correlation coefficients in Linear Regression.
* Generally, correlation tells the relation between two variables in the data set , how change in one have the effect in the other.
* Pearson coefficient is given to define a linear relationship between two variables
* It is denoted by R
* When ever we talk about correlation coefficient in statistics mostly it is Pearson’s R.
* However, it can explain nonlinear elation between two variables and cannot differentiate between dependent and independent variables then
* Like all R value exists between -1 to +1
* If the value is +1 , then there is strong positive linear relation.
* If the value is negative, there is negative linear relation as in increase of one variable decreases other.
* 0 means there no relation between the variables.

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

* It is data Pre-Processing step in linear regression method to bring all the independent numerical variables to a comparable scale.
* It is important to bring all the variables to a same, because some will have higher coefficients, and some will have lower coefficients will affect the outcome of the module
* This will also affect the model evaluation
* Rescaling is done in mainly two methods
* 1. Normalization/(min-max)Scaling:
* This will bring all the values in the range of 0-1
* sklearn.preprocessing.MinMaxScaler is used for this process
* the formula is as below:

Minmax(x) = **x-min(x)/max(x)-x**

* 2. Standardization scaling:
* In this method the values are replaced by its Z-Scores
* sklearn.preprocessing.scale is used here
* the formula is given by

standardization x: **x-mean(x)/std(x)**

-- std(x) is standard deviation of x.

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

* VIF infinity as in R2(r2 is in R-squared value) is 1, it means 1/1-R2 is infinity.
* This happens when there is strong positive or perfect correlation between two variables
* As in increase in one unit will increase in one unit in another variable.

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

* Q-Q plots means quantile -quantile plots.
* A quantile can be described as percentage of values that are below the given value in the dataset.
* Q-Q plot is scatter pot between two sets of quantiles.
* A 45-degree reference line is drawn in the plot.
* Is says that if the two data sets come from same population with same distribution, they fall on the taken reference line
* **Uses:**

1. This helps in determining if the two data sets come from same population, like if we get train and test data sets separately, we can determine if they are from same population and have same common distribution.
2. If they have same location, scale, and tail behaviors.

* **Importance:**

1. With this we can take sample of different sizes
2. Many distributional features can be determined from this plot like presence of outliers, location shifts etc.,