**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 demand of bikes was decreasing in the spring season whereas it was highest in the fall season.
* Compared to 2018, the demand of bikes was more in the year 2019.
* The lowest and highest demand was in the month of January and September respectively.
* The demand of bikes is less on holidays compared to the normal days.
* The highest demand is on Mondays whereas it’s almost similar throughout the other days.
* As expected, the demand of bikes is more on days with a clear weather. During other days with weather like snow or rain, people usually stay at home or prefer to go out in a car.

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

* Our main objective to create a dummy variable is to convert the columns into binary format which is 0 and 1.
* When creating a dummy variable let’s say we have created 3 dummy variables. The 1st variable has the binary value 1, the 2nd and 3rd have 0. If we consider only the 2nd and 3rd variable, we can automatically say that the 1st variable is 1 because the other two variables are 0.
* Similarly, if the 3rd variable is 1, we can automatically conclude that the 1st and 2nd is 0.
* Hence, it’s very important to use drop\_first=True command to eliminate any one dummy variable.

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

* The pair-plot shows that temp and atemp has the highest correlation with the target variable cnt.

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

* After completing the model building on training set, the next comes the test set.
* We prepare the test data set by keeping the common columns between train and test set.
* Then we create a scatter plot for train and test dataset and check if the points fall on a straight line or not.

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

* The top 3 features contributing significantly towards explaining the demand of the shared bikes are:

1. Year
2. Temperature
3. Weather

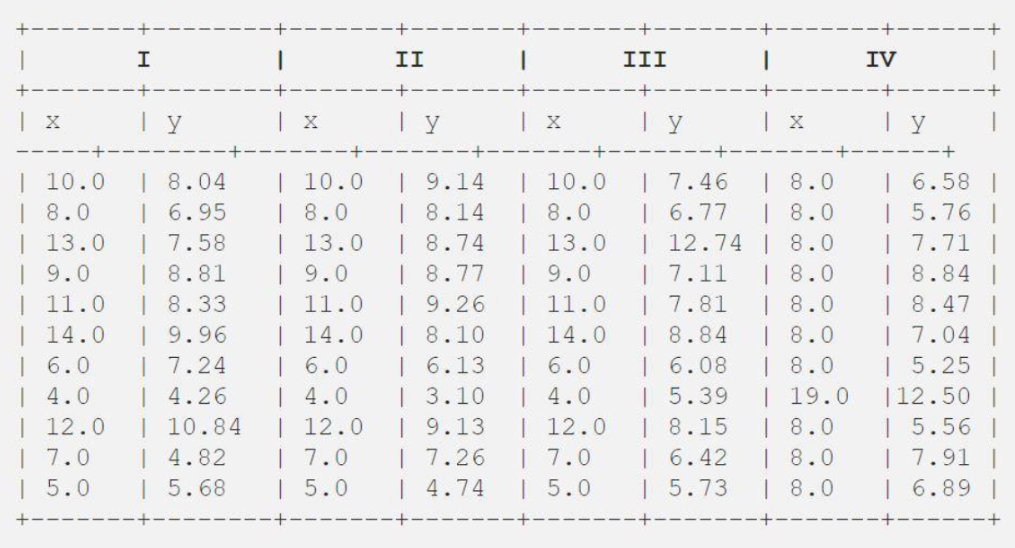
**General Subjective Questions**

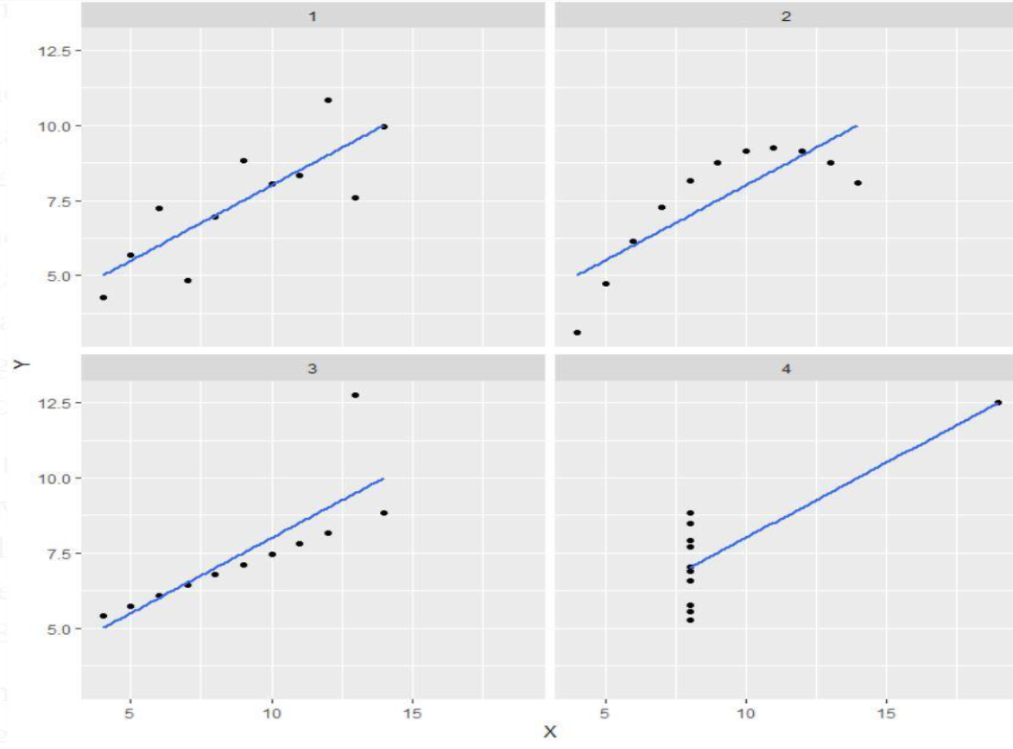
1. *Explain the linear regression algorithm in detail.*

* Linear regression is a model which is used to perform Machine Learning on different types of data sets to get a perfect model for the data.
* Linear regression is used to find out the relationship between two variables.
* Below are the steps which we follow in a linear regression algorithm:

1. Read and understand the data using the data dictionary.
2. Visualize the data using Exploratory Data Analysis.
3. After getting a clear understanding of the data using the above two steps, prepare the data by converting the categorical columns into binary columns using the dummy variables.
4. Split the data into Train and Test data.
5. Work on the train dataset and perform multiple linear regression tasks to get a perfect model that gives us the maximum percentage of Adjusted R-squared.
6. Similarly, work on test dataset where we keep the columns which are present in the train dataset, i.e., we keep only the common columns between the two datasets.
7. Compare the train and test data using scatter plot to check their linear regression line.
8. Also compare their R-square value to check if the train and test data is the best fit or not.
9. *Explain the Anscombe’s quartet in detail?*

* Anscombe’s quartet contains 4 datasets which have almost similar statistical values in the table.
* Also, when it is plotted on a scatter plot, it appears totally different. As you can see the below image for the reference:





* This tells us the importance of data visualization.
* With the help of data visualization only we can actually identify the abnormalities in the data.
* Anscombe’s quartet has set an example because data visualization is necessary to build any kind of model.

1. *What is Pearson’s R?*

* Pearson’s R is known as Pearson’s Correlation Coefficient ‘R’ in statistics.
* Pearson’s correlation coefficient is the covariance of the two variables divided by the product of their standard deviations.
* The formula to calculate Pearson’s R:

*r =*

Where:

N = the number of pairs of scores

∑xy = the sum of the products of paired scores

∑x = the sum of ‘x’ scores

∑y = the sum of ‘y’ scores

∑ = the sum of squared ‘x’ scores

∑ = the sum of squared ‘y’ scores

* The above formula finds out the relationship between the variables and returns the value between -1 and 1.

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

* Scaling is a pre-processing step where it is applied to categorically independent variables to normalize the data within a particular range.
* Scaling is performed on the collected data which contains features highly varying in magnitudes, units and range. If scaling is not done, the algorithm takes the magnitudes into account hence ignoring the units which will result in incorrect modelling.
* In normalized scaling, we bring the data into the range of 0 and 1.
* In standardized scaling, we replace the values by their z-scores.

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

* When VIF=∞, it means that there is a perfect correlation between two independent variables.
* It happens when there is a perfect multicollinearity between one of the variables.
* This can be solved by dropping one of the variables from the dataset.

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

* Q-Q plot is known as a Quantile-Quantile plot.
* It’s a scatter plot which is created by plotting 2 different quantiles against each other.
* The first quantile is for testing the hypothesis and the second quantile is for actual distribution we are testing against.
* It is used to determine weather two samples are from the same population. Also weather the two samples have the same tail, same distribution shape and common location behaviour.