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? (3 marks)
   Ans:
   - The demand of bike was falling in the spring season where as in the fall season the demand was the highest.
   - Compared to 2018 the demand of bike was increased in the year 2019.
   - The demand in the month of Jan is the lowest whereas the highest demand is in the month of Sept.
   - Demand of bike is less on holidays as compared to not being a holiday**.**
   - The demand of bike on Monday is more where throughout the week the demand is almost similar.
   - As expected, the demand of bikes on a clear weather is more compared to other weathers where there might be chance of snow or rain where during these weathers people prefer to stay at home or go out on a car.

2. Why is it important to use drop_first=True during dummy variable creation? (2 marks)
   Ans:
   - Our main motive 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 2^{nd} and 3^{rd} have 0. If we consider only the 2^{nd} and 3^{rd} variable, we can automatically say that the 1^{st} variable is 1 because the 2^{nd} and 3^{rd} are 0.
   - Similarly, if the 3^{rd} variable is 1, then we can the 1^{st} and 2^{nd} are automatically 0.
   - That is why it is very important to use drop_first=True command to eliminate any one dummy variable.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)
   Ans:
   - temp and atemp has the highest correlation with the target variable cnt.

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)
   Ans:
   - After completing the model building on training set the next comes is the test set.
   - We then prepared the test data set by keeping the common columns between train and test set.
   - Then we created a scatter plot for train and test dataset and check if the points fall on a straight line or not.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)
   Ans:
   - Top 3 features contributing significantly towards explaining the demand of the shared bikes are as below:
     a) Year
     b) Temperature
     c) Weather

General Subjective Questions

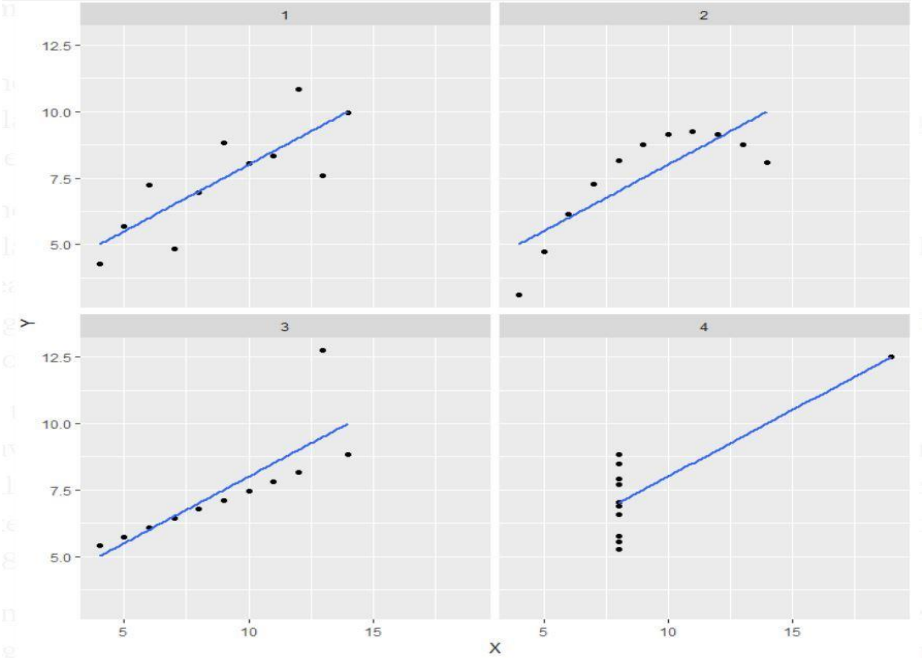
1. Explain the linear regression algorithm in detail. (4 marks)
   Ans:
   - Linear Regression is a model which is used to perform Machine Learning on different types of data sets to get a perfect model for the data.
   - Linear Regression is used to find out the relationship between 2 variables.
   - In linear regression algorithm there are certain steps that we follow, Steps are mentioned in details below:
     i. We read and understand the data using the data dictionary.
     ii. Then we visualize the data using Exploratory Data Analysis.
     iii. After getting a clear understanding of the data using the above 2 steps, we start preparing the data by converting the categorical columns into binary columns using dummy variables.
     iv. Then we split the data into Train and Test.
     v. We work on the train dataset and perform multiple linear regression tasks to get a perfect model that gives us the maximum percent of Adjusted R-squared.
     vi. Similarly, we 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.
     vii. Then we compare the train and test data using scatter plot to check their linear regression line
     viii. Also, we compare their R-square value to check if the train and test data is the best fit or not.

2. Explain the Anscombe's quartet in detail. (3 marks)

Ans:

- Anscombe's quartet contains 4 datasets which have almost similar statistical values in the table.
- But when the same is plotted on a scatter plot it appears totally different. As you can see the below image for reference.

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

- 

- 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 why data visualization is necessary to build any kind of model.

3. What is Pearson's R? (3 marks)

Ans:

- 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.
- Pearson correlation coefficient formula:

$$r = \frac{N\Sigma xy-(\Sigma x)(\Sigma y)}{\sqrt{[N\Sigma x^2-(\Sigma x)^2][N\Sigma y^2-(\Sigma y)^2]}}$$

- Where:
  N = the number of pairs of scores
  Σxy = the sum of the products of paired scores
  Σx = the sum of x scores
  Σy = the sum of y scores
  Σx2 = the sum of squared x scores
  Σy2 = the sum of squared y scores
- The above formula finds out the relationship between the variables and returns the value between -1 and 1.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)
   - Scaling is a pre-processing step where it is applied to categorical 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 then the algorithm takes the magnitudes in to account hence ignoring the units which will result in incorrect modelling.
   - In normalizing scaling, we bring the data into the range of 0 and 1.
   - In standardized scaling we replace the values by their z-scores.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)
   Ans:
   - When VIF=infinity 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.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)
   Ans:
   - Q-Q plot is known as Quantile-Quantile plot.
   - Q-Q plot is 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 you are testing against.
   - Q-Q plot is used to determine whether two samples are from the same population, also whether the two samples have the same tail, same distribution shape and common location behaviour.