Presented by – Saurav Roy (sauravroy.ciem@gmail.com)

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?**

   Answer: The categorical variables which are present in the dataset, using a Box plot and Bar plot following points are inferred:
   a. Fall season has witnessed more booking for both 2018 and 2019. However, there is an increase in number of booking for 2019 compared to 2018.
   b. In 2018, it is observed that from June to Dec, the demand decreased on monthly basis. However, in 2019 the overall demand increased. Also, there has been an increase in demand on a month on month basis from May,2019 to Sep,2019.
   c. There is an increase in demand from 2018 to 2019. In 2019, usually Thursday, Friday, Saturday and Sunday observed more demand compared to rest of the days.
   d. Bike sharing is more on holidays compared to working days. However the gap is very minimal.
   e. Riders preferred clear weather for booking the bikes.

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

   Answer: drop_first= True is used to drop the extra column created during dummy variable. Suppose there are 3 variables, if the first 2 are known then the last variable is also determined. Thus, (n-1) formula is used. Screenshot attached.

	furnished	semi-furnished	unfurnished
0	1	0	0
1	1	0	0
2	0	1	0
3	1	0	0
4	1	0	0

Now, you don't need three columns. You can drop the `furnished` column, as the type of furnishing can be identified with just the last two columns where —

- `00` will correspond to `furnished`
- `01` will correspond to `unfurnished`
- `10` will correspond to `semi-furnished`

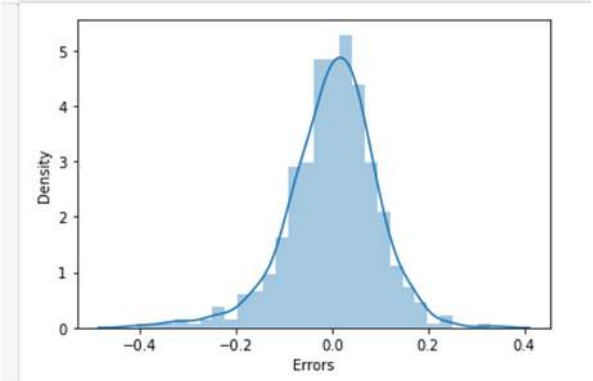
From, the above furnished can be easily determined if semi-furnished and un-furnished is known.

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

   Answer: Temp variable has the highest correlation with the target variable.

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

   Answer: The assumptions of Linear Regression after building the model on the training set is validated using Residual analysis. In the graph, it is observed that the error terms are normally distributed.



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

Answer: Top 3 features contributing significantly towards explaining the demand of the shared bikes are:

a. Temp
b. Yr
c. season_winter

General Subjective Questions

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

Answer: Linear Regression is a machine learning algorithm which is based on supervised learning. Here a model is trained to predict the behaviour of the data set based on the variables. Here, it performs a regression task. Regression models a target prediction value based on independent variables. In Linear regression the two variables i.e. X and y should be linearly corelated.

Mathematically, we can write a linear regression equation as:

$$y= a0+a1x+ \varepsilon$$

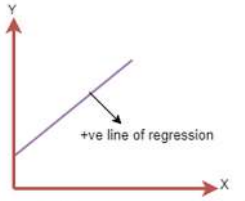
Here, x and y are two variables on the regression line.
Y= Dependent Variable (Target Variable)
X= Independent Variable (predictor Variable)
a0= intercept of the line (Gives an additional degree of freedom)
a1 = Linear regression coefficient (scale factor to each input value).
ε = random error

Types of Linear Regression:

1. Simple Linear Regression: Here a single independent variable is used to predict the value of numerical dependent variable.
2. Multiple Linear Regression: Here more than one independent variable is used to predict the value of numerical dependent variable.

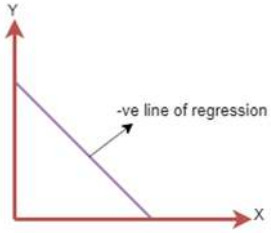
Liner Regression Line:

1. Positive Linear relationship: Both dependent and independent variable increases on Y-axis and X-axis respectively.



The line equation will be: $Y= a_0+a_1x$

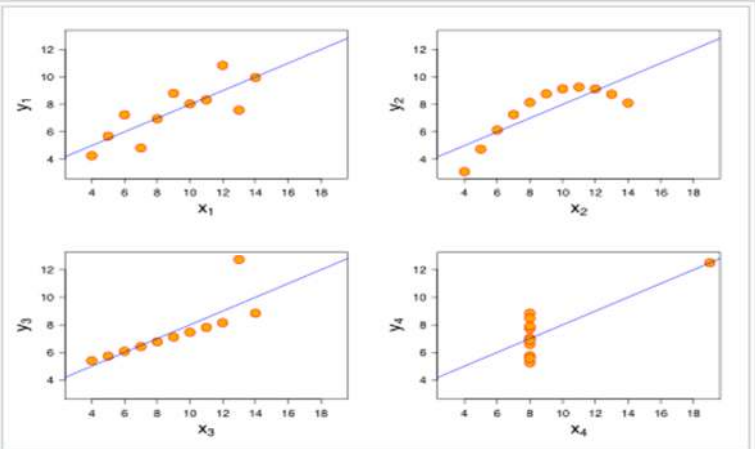
2. Negative Liner relationship: Here dependent variable decreases on y-axis and the independent variable increases on the X-axis.



The line of equation will be: $Y= -a_0+a_1x$

2. Explain the Anscombe's quartet in detail.

Answer: Anscombe's quartet is defined as a group of four data sets which are nearly identical in simple descriptive statistics , but there are some peculiarities in the data set that fools the regression model if built. Each data set consists of 11 x and y points, however when plotted graphically shows different representation.



All four sets are identical when examined using simple summary statistics, but vary considerably when graphed

3. What is Pearson's R ?

Answer: Pearson's R is also called the Pearson Correlation coefficient (PCC). It is measure of linear co-relation between two sets of data. It is the covariance of two variables , divided by the product of their standard deviation. Thus, it is essentially a normalised measurement of the covariance, such that the result always has a value between −1 and 1.

The Pearson's R varies between -1 and +1 where:
r=1 means that the data is perfectly linear with the positive slope.
r=-1 means data is perfectly linear with a negative slope.
r=0 means no linear relation.
r> 0 <5 means there is a weak association
r> 5 <8 means there is a moderate association
r>8 means there is a strong association.

Pearson r Formula

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

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

Answer: Rescaling the features: In MLR one of the challenge is the interpretation of the co-efficients after the model is trained. Eg. value of temp, atemp is different from workingday, holiday etc. The co-efficients of workingday will be higher and that of temp will be lower after modelling. So, to avoid the confusing we need to bring all the variables to comparable scale. Also, it makes the model fast. There are two ways of re-scaling

1. Min-Max scaling (normalization): Between 0 and 1
It is the simplest method and consists of rescaling the range of features to scale in the range [0,1]. The general formula for normalization is

2. Standardisation (mean-0, sigma $x' = \frac{x - \min(x)}{\max(x) - \min(x)}$

In feature standardisation, it makes the value of each feature in the data have zero mean and unit variance. The general method of calculation is to determine the distribution mean and standard deviation for each feature and calculate new data points using the formula:

$$x' = \frac{x - \mu}{\sigma}$$

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

Answer: A perfectly correlation results in VIF= infinity. This means that R^2 is 1. To solve this problem we need to drop one of the variables from the dataset which is causing the perfect multicollinearity. An infinite VIF indicates that the corresponding variable is exactly expressed by the linear combination of the other variable.

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

Answer: Q-Q plot is also known as Quantile- Quantile plots. These are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q Q plots is to find out if two sets of data come from the same distribution. A 45 degree angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.

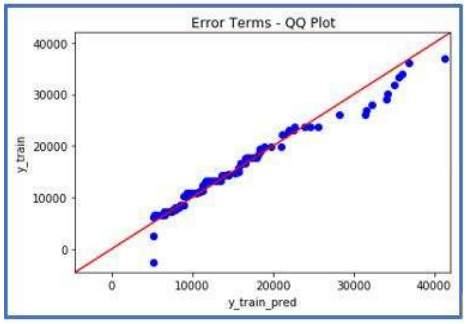
This helps in a scenario where we have training and test data set received separably and by Q-Q plot we can confirm whether they are from population with same distributions.

Interpretation:

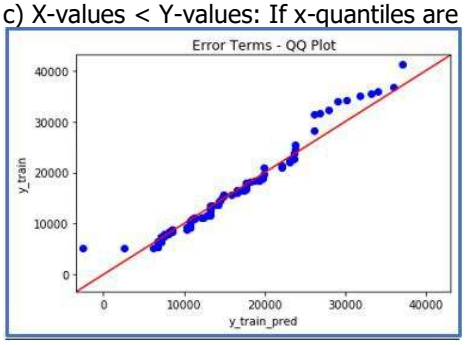
A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. Below are the possible interpretations for two data sets.

a) Similar distribution: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x -axis

b) Y-values < X-values: If y-quantiles are lower than the x-quantiles.



c) X-values < Y-values: If x-quantiles are lower than the y-quantiles.



d) Different distribution: If all point of quantiles lies away from the straight line at an angle of 45 degree from x -axis