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)

Answer: For the analysis I have used bagplots to get the visualization about the total data and divided the whole data by the year so that we will get to know about the business according to the year. Histograms helps us to visualize the distribution of the data.

a. we have a greater number of data points available for the year 2019 than year 2018 seams demands were increased in year 2019

b. bike demands are more on working day

c. most of the riders are interested in riding when weather is good, we have more than 2M+ records for that

   where,
   good weather = Clear, Few clouds, partly cloudy, partly cloudy
   Moderate weather = Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
   serve weather = Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog

   and approx. 1M records shows that people have also shown the interest while weather condition was moderate there were no customers who traveled when the weather condition was serve

d. Fall seems to be the best season for riding while summer comes on second

e. Almost all days are profitable as Thursday, Sunday and Saturday are the days of high demands

f. August, June, September and July are the months who shows majority demands

g. 2019 attracted a greater number of bookings from the previous year, which shows good progress in terms of business.

2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)
Answer:

a. drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

b. drop_first: bool, default False, is the syntax which implies whether to get Total -1 dummies out of total categorical levels by removing the first level.

c. Let's say we have 5 types of values in Categorical column and we want to create dummy variable for that column. If a variable is not in a,b,c,d and e then It is obvious f. So we do not need 5'th variable to identify that the variable belongs to the f category

d. Having many features can disturb the accuracy of the model and we may get bad results in the end so we need clean data with good features correlated with the target variable

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

Answer:

a. 'atemp' variable has the highest correlation with the target variable. Correlation of 'atemp' and 'cnt' comes around 0.6307

b. While 'temp' comes the second highest in the correlation with the target variable. 'temp' and 'cnt' have correlation around 0.627

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Answer: The assumptions to validate the Linear Regression after building the model on the training set are as follows:

a. The Dependent variable and independent variable must have a linear relationship.

b. No Autocorrelation in residuals.

c. No Heteroskedasticity meaning there should be no visible pattern in residual values.

d. No Perfect Multicollinearity there should be insignificant multicollinearity among variables.

e. Residuals must be normally distributed.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Answer:

a. Based on my final model Temperature, year and spring are top highly correlated features by person's correlation and contributing significantly towards explaining the demand of the shared bikes.

b. Final equation looks like:

```
cnt = 0.1776 + (yr X 0.2372)-(holiday X 0.0798)+(temp X 0.3919)-(mnth_dec X
0.0617)-(mnth_jan X 0.0479)-(mnth_jul X 0.0521)-(mnth_nov X 0.0641)+(mnth_sept
X 0.0536)-(season_spring X 0.1246)+(season_winter X 0.0682)-(weathersit_bad X
0.2253)+(weathersit_good X 0.0684)
```

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

Answer: Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables they are considering, and the number of independent variables getting used

a. Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression.

b. Linear regression is one of the very basic forms of machine learning where we train a model to predict the behaviour of your data based on some variables. In the case of linear regression as you can see the name suggests linear that means the two variables which are on the x-axis and y-axis should be linearly correlated.

c. An example is let's say you are running a sales promotion and expecting a certain number of count of customers to be increased now what you can do is you can look the previous promotions and plot if over on the chart when you run it and then try to see whether there is an increment into the number of customers whenever you rate the promotions and with the help of the previous historical data you try to figure it out or you try to estimate what will be the count or what will be the estimated count for my current promotion this will give you an idea to do the planning in a much better way about how many numbers of stalls maybe you need or how many increase number of employees you need to serve the customer. Here the idea is to estimate the future value based on the historical data by learning the behaviour or patterns from the historical data.

d. In some cases, the value will be linearly upward that means whenever X is increasing Y is also increasing or vice versa that means they have a correlation or there will be a linear downward relationship.

e. One example for that could be that the police department is running a campaign to reduce the number of robberies, in this case, the graph will be linearly downward.

f. Linear regression is used to predict a quantitative response Y from the predictor variable X.

g. Mathematically, we can write a linear regression equation as:

$$\underset{\substack{\uparrow \\ \text{Dependent} \\ \text{Variable}}}{Y_i} = \overset{\substack{\text{Constant/Intercept} \\ \downarrow}}{\beta_0} + \underset{\substack{\uparrow \\ \text{Slope/Coefficient}}}{\beta_1} \overset{\substack{\text{Independent} \\ \text{Variable} \\ \downarrow}}{X_i}$$

Where a and b given by the formulas:

$$b\,(slope) = \frac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{n\sum x^2 - \left(\sum x\right)^2}$$

$$a\,(intercept) = \frac{n\sum y - b\left(\sum x\right)}{n}$$

Here, x and y are two variables on the regression line.

b = Slope of the line.

a = y-intercept of the line.

x = Independent variable from dataset

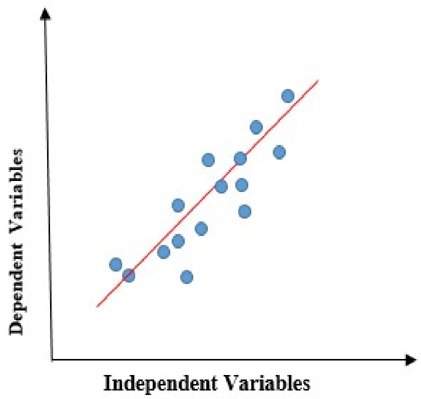
y = Dependent variable from dataset

Use Cases of Linear Regression:

a. Prediction of trends and Sales targets – To predict how industry is performing or how many sales targets industry may achieve in the future.

b. Price Prediction – Using regression to predict the change in price of stock or product.

c. Risk Management- Using regression to the analysis of Risk Management in the financial and insurance sector.



The above graph presents the linear relationship between the dependent variable and independent variables. When the value of x (independent variable) increases, the value of y (dependent variable) is likewise increasing. The red line is referred to as the best fit straight line. Based on the given data points, we try to plot a line that models the points the best.

To calculate best-fit line linear regression uses a traditional slope-intercept form.

$$y = mx+b \Longrightarrow y = a_0+a_1x$$

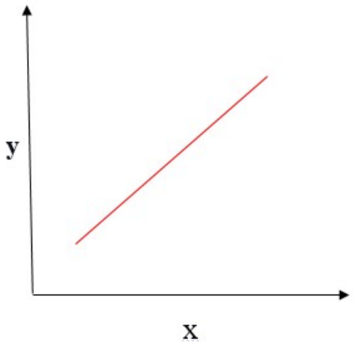
Where:

y= Dependent Variable.
x= Independent Variable.
a0= intercept of the line.
a1 = Linear regression coefficient.

A regression line can be a Positive Linear Relationship or a Negative Linear Relationship.

Positive Linear Relationship

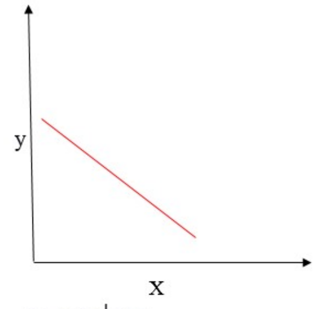
If the dependent variable expands on the Y-axis and the independent variable progress on X-axis, then such a relationship is termed a Positive linear relationship.



$y = a_0+a_1x$

Negative Linear Relationship

If the dependent variable decreases on the Y-axis and the independent variable increases on the X-axis, such a relationship is called a negative linear relationship.



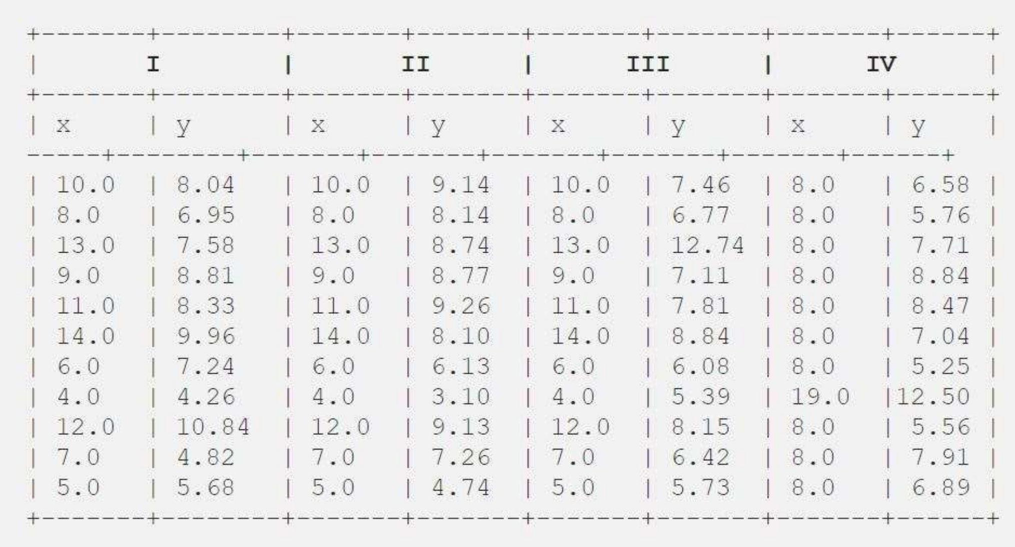
$y = -a_0 + a_1 x$

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

Answer: Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties.

Simple understanding:

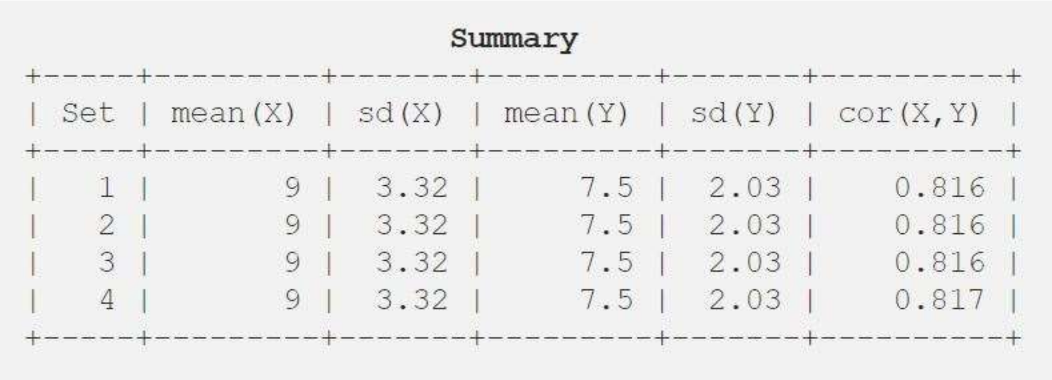
Once Francis John "Frank" Anscombe who was a statistician of great repute found 4 sets of 11 data-points in his dream and requested the council as his last wish to plot those points. Those 4 sets of 11 data-points are given below.



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

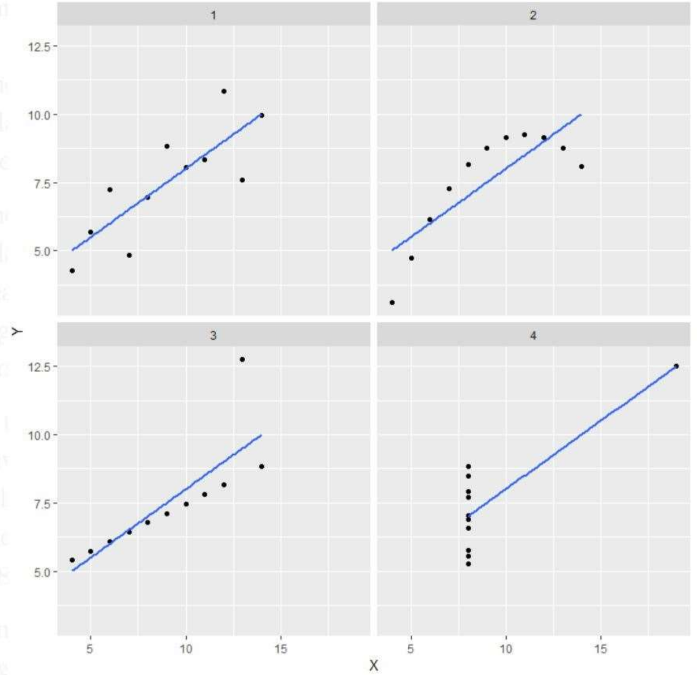
After that, the council analyzed them using only descriptive statistics and found the mean, standard deviation, and correlation between x and y.

So let me show you the result in a tabular fashion for better understanding.



```
                    Summary
+-----+---------+-------+---------+-------+----------+
| Set | mean(X) | sd(X) | mean(Y) | sd(Y) | cor(X,Y) |
+-----+---------+-------+---------+-------+----------+
|   1 |       9 |  3.32 |     7.5 |  2.03 |    0.816 |
|   2 |       9 |  3.32 |     7.5 |  2.03 |    0.816 |
|   3 |       9 |  3.32 |     7.5 |  2.03 |    0.816 |
|   4 |       9 |  3.32 |     7.5 |  2.03 |    0.817 |
+-----+---------+-------+---------+-------+----------+
```

<ins>Graphical representation:</ins>



<ins>Explanation of this output:</ins>

1. In the first one(top left) if you look at the scatter plot you will see that there seems to be a linear relationship between x and y.
2. In the second one(top right) if you look at this figure you can conclude that there is a non-linear relationship between x and y.
3. In the third one(bottom left) you can say when there is a perfect linear relationship for all the data points except one which seems to be an outlier which is indicated be far away from that line.
4. Finally, the fourth one(bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient.

<ins>Application:</ins>

The quartet is still often used to illustrate the importance of looking at a set of data graphically before starting to analyze according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets.

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

Answer:

1. Correlation is a statistic that measures the relationship between two variables in the finance and investment industries. It shows the strength of the relationship between the two variables as well as the direction and is represented numerically by the correlation coefficient. The numerical values of the correlation coefficient lies between -1.0 and +1.0.
2. A negative value of the correlation coefficient means that when there is a change in one variable, the other changes in a proportion but in the opposite direction, and if the value of the correlation coefficient is positive, both the variables change in a proportion and the same direction.
3. When the value of the correlation coefficient is exactly 1.0, it is said to be a perfect positive correlation. This situation means that when there is a change in one variable, either negative or positive, the second variable changes in lockstep, in the same direction.
4. A perfect negative correlation means that two assets move in opposite directions, while a zero correlation implies no linear relationship at all. We can determine the strength of the relationship between two variables by finding the absolute value of the correlation coefficient.
5. In Statistics, the Pearson's Correlation Coefficient is also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or bivariate correlation. It is a statistic that measures the linear correlation between two variables. Like all correlations, it also has a numerical value that lies between -1.0 and +1.0.
6. Whenever we discuss correlation in statistics, it is generally Pearson's correlation coefficient. However, it cannot capture nonlinear relationships between two variables and cannot differentiate between dependent and independent variables.
7. Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations. The form of the definition involves a "product moment", that is, the mean (the first moment about the origin) of the product of the mean-adjusted random variables; hence the modifier product-moment in the name.
8. Pearson's Correlation Coefficient is named after Karl Pearson. He formulated the correlation coefficient from a related idea by Francis Galton in the 1880s.
9. Using the formula proposed by Karl Pearson, we can calculate a linear relationship between the two given variables. For example, a child's height increases with his increasing age (different factors affect this biological change). So, we can calculate the relationship between these two variables by obtaining the value of Pearson's Correlation Coefficient r. There are certain requirements for Pearson's Correlation Coefficient:
    a. Scale of measurement should be interval or ratio
    b. Variables should be approximately normally distributed
    c. The association should be linear
    d. There should be no outliers in the data
10. the formula is given by:

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

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

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Answer:

1. It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.
2. Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.
3. It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.
4. Normalization/Min-Max Scaling:

   It brings all of the data in the range of 0 and 1. sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.

$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

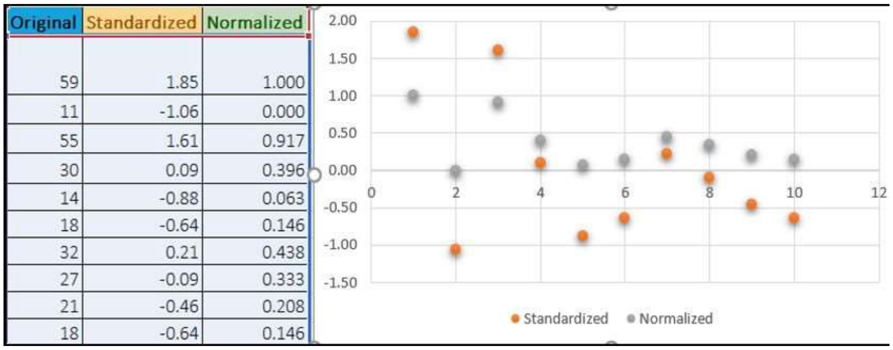
5. Standardization Scaling:
   a. Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).

$$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$$

   b. sklearn.preprocessing.scale helps to implement standardization in python.
   c. One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers.

6. Example:

   Below shows example of Standardized and Normalized scaling on original values.



Original	Standardized	Normalized
59	1.85	1.000
11	-1.06	0.000
55	1.61	0.917
30	0.09	0.396
14	-0.88	0.063
18	-0.64	0.146
32	0.21	0.438
27	-0.09	0.333
21	-0.46	0.208
18	-0.64	0.146

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

Answer:

i. If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

ii. An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

iii. In general, a VIF above 10 indicates high correlation and is cause for concern. Some authors suggest a more conservative level of 2.5 or above....A rule of thumb for interpreting the variance inflation factor:

   a. 1 = not correlated.

   b. Between 1 and 5 = moderately correlated.

   c. Greater than 5 = highly correlated.

iv. if VIF is large and multicollinearity affects your analysis results, then you need to take some corrective actions before you can use multiple regression. Here are the various options:

1. One approach is to review your independent variables and eliminate terms that are duplicates or not adding value to explain the variation in the model. For example, if your inputs are measuring the weight in kgs and lbs then just keep one of these variables in the model and drop the other one. Dropping the term with a large value of VIF will hopefully, fix the VIF for the remaining terms and now all the VIF factors are within the threshold limits. If dropping one term is not enough, then you may need to drop more terms as required.
2. A second approach is to use principal component analysis and determine the optimal set of principal components that best describe your independent variables. Using this approach will get rid of your multicollinearity problem but it may be hard for you to interpret the meaning of these “new” independent variables.
3. The third approach is to increase the sample size. By adding more data points to our model, hopefully, the confidence intervals for the model coefficients are narrower to overcome the problems associated with multicollinearity.
4. The fourth approach is to transform the data to a different space like using a log transformation so that the independent variables are no longer correlated as strongly with each other.
5. Finally, you can use a different type of model call ridge regression that better handles multicollinearity.
6. In conclusion, when you are building a multiple regression model, always check your VIF values for your independent variables and determine if you need to take any corrective action before building the model.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Answer:

1. Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.
2. This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.
3. <u>Few advantages:</u>

   a) It can be used with sample sizes also

b) Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.

4. <u>It is used to check following scenarios:</u>

If two data sets —

i. come from populations with a common distribution

ii. have common location and scale

iii. have similar distributional shapes

iv. have similar tail behavior

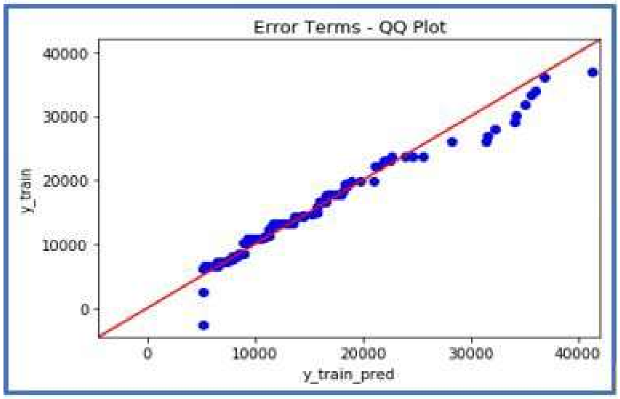
5. <u>Interpretation:</u>

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.

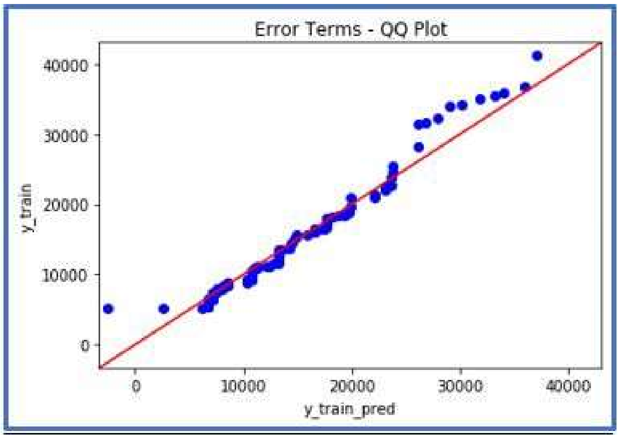
Below are the possible interpretations for two data sets.

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

6. Implementation in Python:

   statsmodels.api provide qqplot and qqplot_2samples to plot Q-Q graph for single and two different data sets respectively.