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)

<u>**Ans:**</u> There are different effects of each categorical variable, which are as mentioned below:

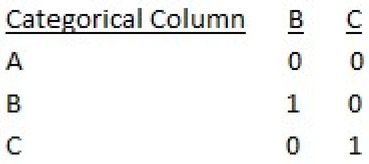
1. The demand for bikes is more for a weather situation, when '*weathersit*' is **partly_cloudy** ('cnt' increases for 'weathersit'=1)
2. When we see the yearly trend, we can see that demand was greater in the year **2019**. ('*yr*' = 1 for the year 2019)
3. For the year-wise and month-wise impact on '*cnt*', we can see that the months **May, June, July, August, September, October** have the highest demands. And the start and end months of the year have less demand.
4. When the day is a *holiday* the demand is less, compared to that on a *workingday.* hence we can say thay people opt to use bikes more on a working day.
5. The seasons **fall** and **summer** show the highest demands for the *season*, followed by **winter** in which the demand is 3rd highest.
6. The variables '*weekday*' and '*day*' doesn't seem to be of much significance for the inference of the dependent variable '*cnt*' .

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

<u>**Ans:**</u> While creating dummies if we don't use **drop_first=True,** then an extra redundant column will be created. Also the correlations among the independent dummy variables will increase, which might lead to higher VIF values.

A higher VIF is not suitable for a good model. Hence it is important to drop the first column while creating dummy variables.

Suppose we have 3 values for a categorical column, while creating dummies for this variable we can explain the 3 values in 2 dummy columns. If we add 3 dummy columns, then the last column would be of use as data can be explained with just 2.



Categorical Column	B	C
A	0	0
B	1	0
C	0	1

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

<u>Ans:</u> From the pairplot for numeric values, we can say that the variable '*registered*' has the highest correlation with the target variable '*cnt*' .

Followed by the variable '*casual*' having the 2nd highest correlation.

But also since these are the variables which makeup the target variable, we can say that '*temp*' has a high correlation considering that it has a good linear relationship with target variable '*cnt*'.

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

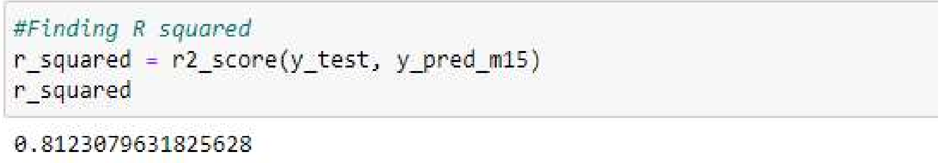
<u>Ans:</u> After finalizing the built model based on the p-values and VIFs of the features, we performed some residual analysis and evaluations on the model.

Firstly using the model,we made predictions of the target variable and used these values to find the error terms with respect to the actual value of these target variables for the train data set. Then plotted these error terms on a distribution plot to check if the terms are normally distributed and centered around mean=0.

The error terms being normally distributed and finding no pattern signifies that the built model is good.

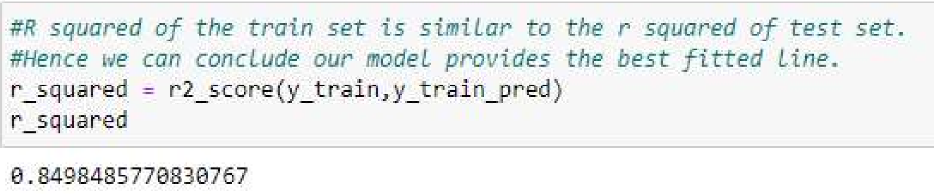
For the Model Evaluation, we found the R squared value for the train and test data, which was **0.850** and **0.812** respectively. As the values are close enough we can say that the model is not overfit. Also as the value of R squared is closer to 1, signfies that the model correctly shows the relation of predictor variables with the target variable.

Based on these observations we can validate our assumptions of Linear Regression for the built model to be correct and acceptable.



```
#Finding R squared
r_squared = r2_score(y_test, y_pred_m15)
r_squared
```

```
0.8123079631825628
```



```
#R squared of the train set is similar to the r squared of test set.
#Hence we can conclude our model provides the best fitted line.
r_squared = r2_score(y_train,y_train_pred)
r_squared
```

```
0.8498485770830767
```

5. Based on the final model, which are the top 3 features contributing significantly towards

explaining the demand of the shared bikes? (2 marks)

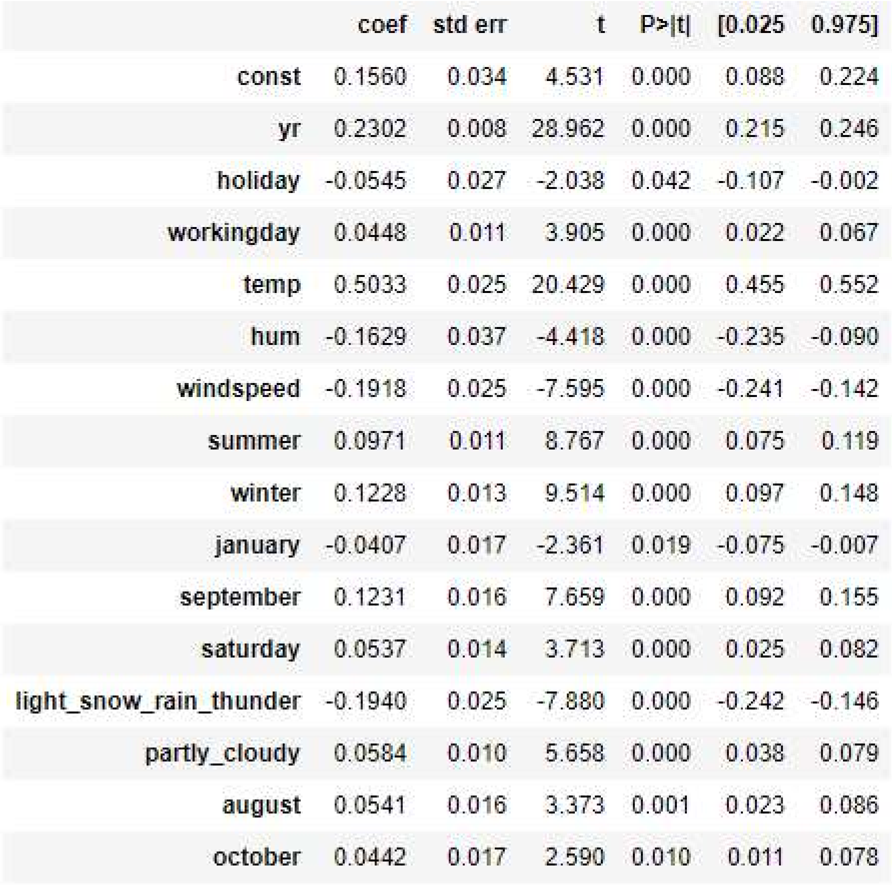
<u>Ans:</u> While determining the top predictor variables for a model, the general rule which is followed would be picking the features with the largest coeffient value.

According to our model here, the variable '*temp*' has the highest coefficent, hence it becomes our top feature for prediction.

The next feature with high coefficient value is '*yr*', which is the 2nd top feature for this model.

The 3rd top feature would be '*light_snow_rain_thunder*' based on the next highest value of coefficient irrespective of the sign, as the sign would just indicate a positive or negative correlation.

Hence the top 3 features of our model contibuting signficantly towards the increasing demands of bikes would be *'temp'*, *'yr'* and *'light_snow_rain_thunder*'.



	coef	std err	t	P>|t|	[0.025	0.975]
const	0.1560	0.034	4.531	0.000	0.088	0.224
yr	0.2302	0.008	28.962	0.000	0.215	0.246
holiday	-0.0545	0.027	-2.038	0.042	-0.107	-0.002
workingday	0.0448	0.011	3.905	0.000	0.022	0.067
temp	0.5033	0.025	20.429	0.000	0.455	0.552
hum	-0.1629	0.037	-4.418	0.000	-0.235	-0.090
windspeed	-0.1918	0.025	-7.595	0.000	-0.241	-0.142
summer	0.0971	0.011	8.767	0.000	0.075	0.119
winter	0.1228	0.013	9.514	0.000	0.097	0.148
january	-0.0407	0.017	-2.361	0.019	-0.075	-0.007
september	0.1231	0.016	7.659	0.000	0.092	0.155
saturday	0.0537	0.014	3.713	0.000	0.025	0.082
light_snow_rain_thunder	-0.1940	0.025	-7.880	0.000	-0.242	-0.146
partly_cloudy	0.0584	0.010	5.658	0.000	0.038	0.079
august	0.0541	0.016	3.373	0.001	0.023	0.086
october	0.0442	0.017	2.590	0.010	0.011	0.078

General Subjective Questions

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

<u>Ans:</u> Regression is the technique of predicting continuous set of values for a target variable. In linear regression the relationship between the independent variable used to predict the target or dependent variable should be linear. That is it should fit a line or follow a linear trend of distribution for the data points. Among the different techniques used forlinear regression the most common one is Ordinary Least Squares (OLS).

There are two types of Linear regressions:

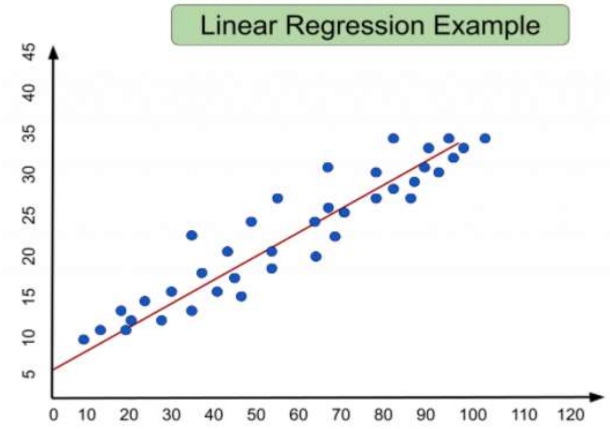


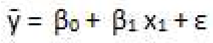
Image source: https://www.erp-information.com/regression-analysis.html

1. **Simple Linear Regression**:

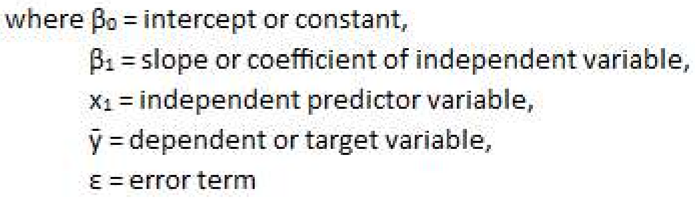
When an independent variable is used to estimate a dependent variable linearly, the regression method is called as simple linear regression.

Eg: Predicting CO_2 emissions based on the engine size of the cars.

It is described by the equation of line as:



$$\bar{y} = \beta_0 + \beta_1 x_1 + \varepsilon$$



where β_0 = intercept or constant,

β_1 = slope or coefficient of independent variable,

x_1 = independent predictor variable,

$\bar{y}$ = dependent or target variable,

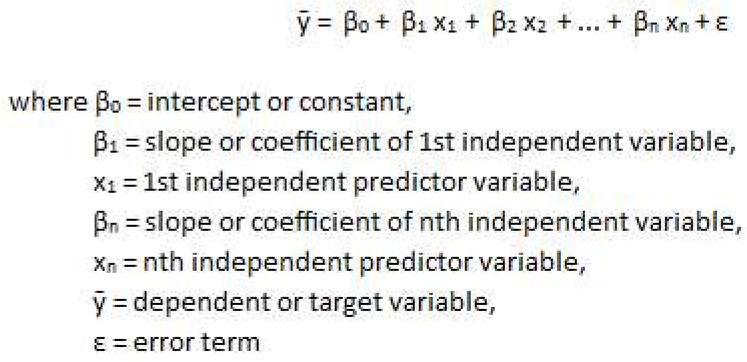
ε = error term

2. **Multiple Linear Regression**:

When more than one independent variables are used to estimate a dependent variable linearly, the regression method is called as multiple linear regression.

Eg: Predicting house prices based on area, locality, number of bedrooms, number of bathrooms.

It is described by the equation of line as:



$$\bar{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \varepsilon$$

where β_0 = intercept or constant,

β_1 = slope or coefficient of 1st independent variable,

x_1 = 1st independent predictor variable,

β_n = slope or coefficient of nth independent variable,

x_n = nth independent predictor variable,

$\bar{y}$ = dependent or target variable,

ε = error term

Applications of regression:

Price estimation, Sales forecasting, Emplyment income, Satisfaction analysis, etc

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

Ans: Anscombe's quartet was constructed by a statistician Francis Anscombe in the year 1973. It is used to illustrate how a data when viewed theoretically or through table values might not give a clear idea about the trend of the variables. Hence, it focuses on graph plotting to understand and get the gist of the data in a glance.

Francis used an example to describe and explain his findings. The observation being if we have 4 datasets having the same statistical distributions such as the values of x, y, mean, variance etc. Yet when we plot these on a plot we will get different equation to fit on a linear regression line.

This explains the importance of having visualizations on data before building models on them. It also helps in spotting the various anomalies and outliers in the data.

In the below plot, the datasets are describes as:

Dataset 1 having data that fits the linear regression model perfectly.

Dataset 2 shows a non-linear trend of points which does not fit well on the linear model line.

Dataset 3 shows that the linear model can handle the outliers present in the data set.

Dataset 4 shows that the linear model cannot handle the outliers present in the data set.

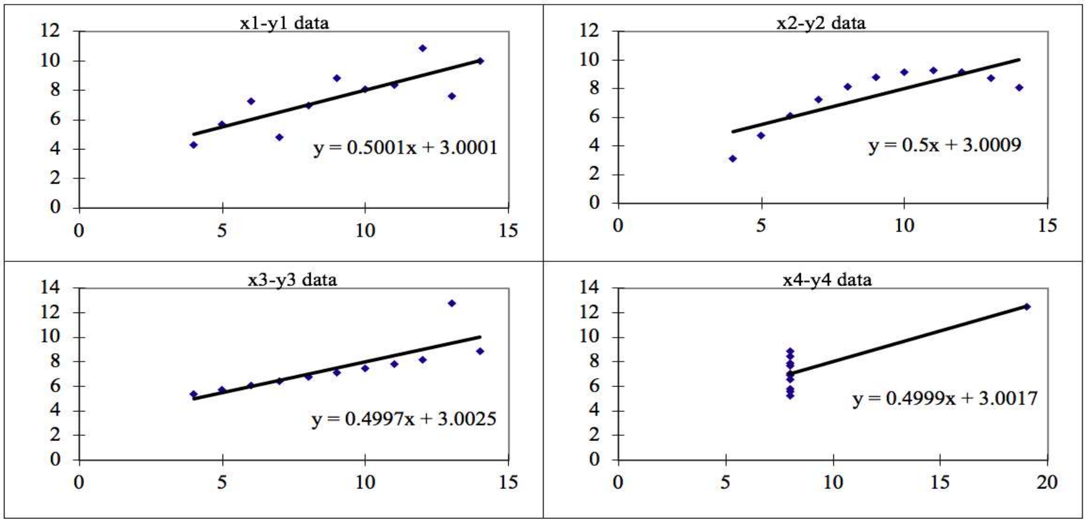
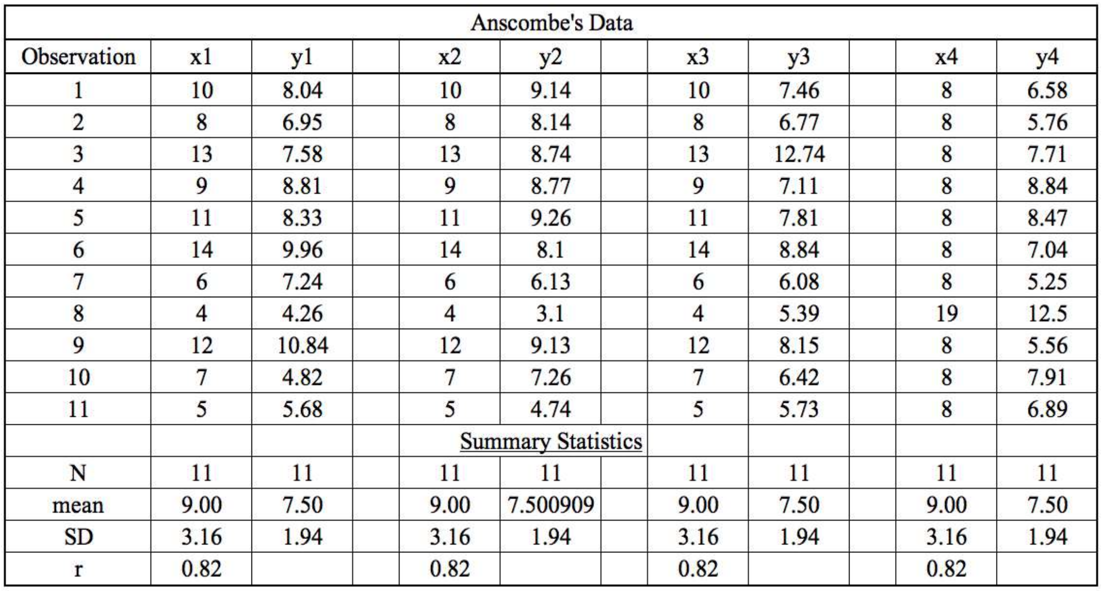


Image source: https://towardsdatascience.com/importance-of-data-visualization-anscombes-quartet-way-a325148b9fd2



Anscombe's Data

Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	5.39		19	12.5
9	12	10.84		12	9.13		12	8.15		8	5.56
10	7	4.82		7	7.26		7	6.42		8	7.91
11	5	5.68		5	4.74		5	5.73		8	6.89
					Summary Statistics						
N	11	11		11	11		11	11		11	11
mean	9.00	7.50		9.00	7.500909		9.00	7.50		9.00	7.50
SD	3.16	1.94		3.16	1.94		3.16	1.94		3.16	1.94
r	0.82			0.82			0.82			0.82	

Image source: https://towardsdatascience.com/importance-of-data-visualization-anscombes-quartet-way-a325148b9fd2

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

Ans: Pearson's R or the Pearson's Correlation Coefficient is a measure that explains the strength of linear relationship or association between two variables in a data set. It is denoted by the letter 'r'. It can be described as the change in one variable effecting the other variable to change linearly.

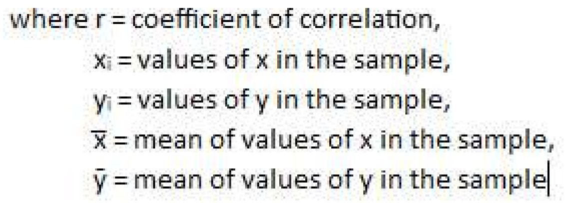
The linear association between these two variables can be postive or negative, that is the two variables can be positively or negatively correlated.

To put it in simpler terms, if an increase in one variable causes the other variable to change, then they are positively correlated. Similarly, if an increase in one variable causes the other variable to decrease, then thery are in negative correlation.

For a postive correlation the two variables move in the same direction whereas for a negative correlation the two variables move in opposite directions.

The Pearson's Correlation Coefficient has a value range for +1 to -1 where +1 indicates a perfect positive correlation and -1 indicates a perfect negative correlation and it is given by the formula,

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



where r = coefficient of correlation,

x_i = values of x in the sample,

y_i = values of y in the sample,

$\bar{x}$ = mean of values of x in the sample,

$\bar{y}$ = mean of values of y in the sample

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

Ans: Scaling is the technique of converting all the numeric predictor variables in the data to a standardized format based on a fixed range scale. It is used to handle high measures or units during the data pre-processing stage.

Sometimes it may happen that the units of measurement for 2 or more feature variables are measured on different scales. This will cause the model to return a higher value of coefficient for a smaller feature value and a small coefficient value for a larger feature. Thus for standardizing and bringing all the values on one scale we use the technique of feature scaling.

There are 2 types of commonly used feature scaling methods:

1. **MinMax or Normalized Scaling**:

This type of scaling normalizes the features between a range of 0 and 1. With the highest values in the dataset to be at 1 and the lowest to be at the value 0. It uses the below formula for scaling -

$$X_{normalized} = (X_i - min(X)) / (max(X) - min(X))$$

2. **Standardized Scaling**:

This type of scaling standardizes the values for the feature variables such that the rescaled distribution has a mean equal to 0 and the variance or standard deviation equal to 1. It uses the below formula for scaling -

$$X_{standardized} = (X_i - mean(X)) / (standard\ deviation)$$

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

Ans: The correlation between few feature variables is high sometimes with a perfect linear relation among those variables. It can be said that one variable is perfectly capable of describing the trend of the other variable. This causes the VIF to return a value close to infinity.

Mathematically explaining this theory, we know the formula for VIF is 1 divided by 1 minus square of correlation between the independent variables.

$$\mathbf{VIF = 1 / (1 - R_i^2)}$$

Now, if we assume a perfect correlation between two variables then the value of R_i^2 turns out to be 1. Hence the formula would give the value,

$$\begin{aligned} \mathbf{VIF} &= \mathbf{1 / (1 - 1)} \\ &= \mathbf{1 / 0} \\ &= \textbf{infinity} \end{aligned}$$

Thus, higher value of VIF and multicollinearity will affect the analysis and hence it is advised to drop the variables having multicollinearity and high correlation.

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

Ans: A Q-Q plot is a Quantile-Quantile plot in statistics. It is used to graphically plot the probability distributions of two variables. It will plot a fraction of distribution with respect to the other variables fraction of distribution or quantile of one variable versus the quantileof the other variable.

Q-Q plots are used to determine if two data sets have a same distribution of population or not. These plots are more widely used for plotting theoretical distributions like Uniform distribution or Normal exponential distribution.

For linear regression, we can check if the data sets which we get separately for a train and test data have similar population distribution. Also it is not mandatory to have two datasets with same sample size. We can use it even if the two datasets have different number of sample sizes.

It is also used to determine if two datasets have similar distributional shapes or tail behaviour. Or It can be used to determine if the two data sets have common scale and location.

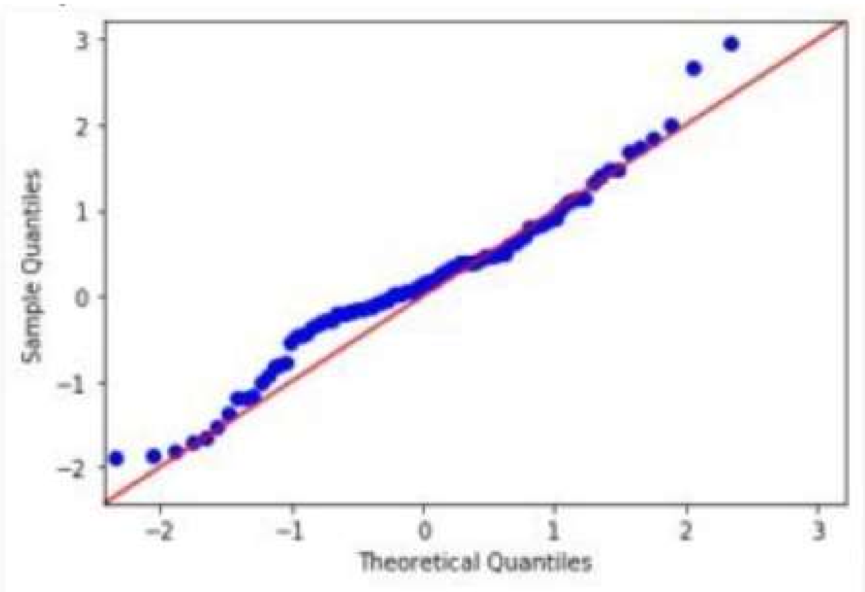


Image source: https://www.geeksforgeeks.org/qqplot-quantile-quantile-plot-in-python/