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?

   Ans:

   Below are some key observations made regarding the categorical variables against the target variable.

   We have used Box plot and Bar charts in our coding to understand the relationships between various variables.

   1. “season”
   a) number of rides requested gradually increases from spring to summer to fall and drops in winter.
   b) Highest rides are in the fall season

   2. “Mnth”
   a) There is a steady increase in the number of rides requested in the first 6 months of the year
   b) There is constant demand for the rides in the months of June, July, August and September – which is in line with the seasons data.
   c) Steady drop in demand is also seen in the last 3 months of the year too.
   3. “yr”
   a) There is an increase in the demand for the number of rides from year 2018 to year 2019

   4. "weathersit"
   a) Number of rides when the weather is “Clear/partly influenced by other factors were the highest in comparison to others.
   b) The drop in demand from first category to second category is not as significant as the drop we see between second category to third category.

   5. “holiday”
   a) On the days of the holiday, we see the number of rides are less in comparison to the non-holiday days.

   6. “workingday”
   a) the number of rides on the days which are not weekend or holiday is more in comparosi on to the rides requested on a holiday or a weekend put together. But the difference is n ot very
   significant

   7. “weekday”
   a) Demand raises consistently across the week and stabilizes in the second half of the we ek.

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

Ans:

get_dummies function is used to convert the categorical variables into numeric in order to have a refined dataset for the regression model to learn and predict the outcome successfully.

Syntax :

pandas.get_dummies(*data*, *prefix=None*, *prefix_sep='_'*, *dummy_na=False*, *columns=None*, *sparse=False*, *drop_first=False*, *dtype=None*)

As part of the same function, there is a parameter called – "drop_first" when marked as "True", would drop 1st dummy column created, by doing this we would have reduced the number of columns getting created and minimize their correlation effect on variables.

For a column with "n" categories, we will need "n-1" dummy columns

weathersit :

1: Clear, Few clouds, Partly cloudy, Partly cloudy

2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist

3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds

4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog

We are going to convert this categorical data into dummy columns

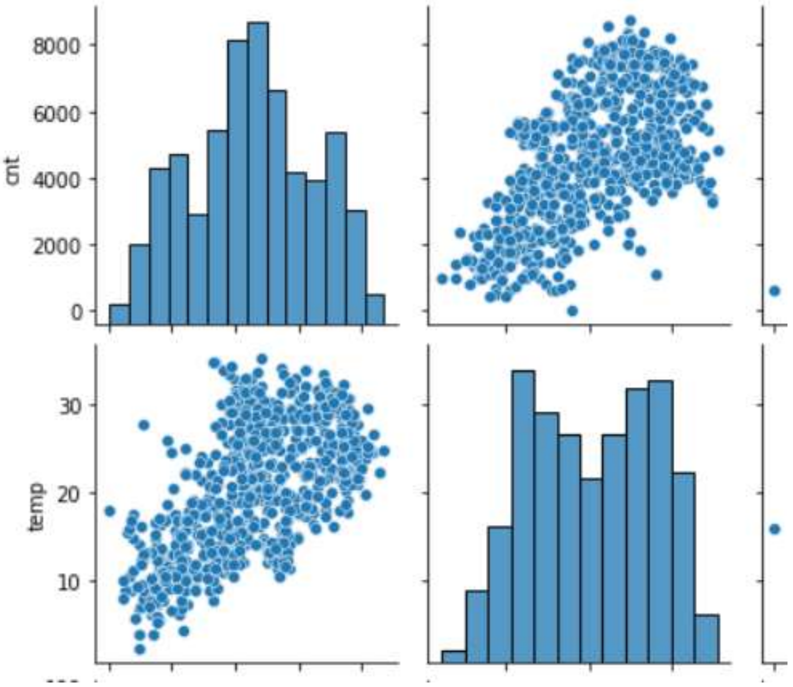
WeatherSit	D1	D2	D3	D4
1	1	0	0	0
2	0	1	0	0
3	0	0	1	0
4	0	0	0	1

Even if we drop the D1 columns, the presence of three "0" under the D2, D3, and D4, would imply that it's D1.

So, here n = 4 and we have n-1 dummy = 3

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

Ans:



Looking at the pair plot of the numeric variables, it's evident that Target Variable – "cnt" has greater correlation with "temp" – independent variable.

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

Ans:

Below are the various steps to check and verify the assumptions were proved correct for the train set

a) Whether the train set was normally distributed.
b) Verified with VIF values on multicollinearity and ensured that all the independent variables have values under 5.
c) We checked Homoscedasticity by plotting a scatter plot and the data did not show any trend as such and they were scattered randomly.
d) Verified autocorrelation as to whether it was within limits

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

Ans:

The equation for our model is

2067.99 + (2050.37 *yr) + (-1416.32 * windspeed) + (-547.16 * spring) + (427.66 * summer) + (849.43 * winter) + (-2522.33 * orange) + (-672.74 * orange) + (-395.33 * december) + (-469.60 * january) + (-402.95 * november) + (677.02 * september) + (198.32 * saturday) + (3681.92 * temp)

Top 3 variables which have the highest influence are

1) yr
2) windspeed
3) winter

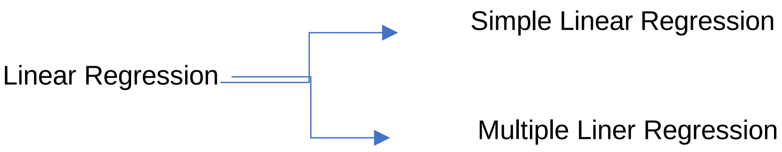
General Subjective Questions

1. Explain the linear regression algorithm in detail?

Ans:

In simple terms, linear regression is the linear relationship between independent variable and dependent variables and this relation is explained by a best fit straight line.

Linear Regression is part of supervised learning algorithm.



Linear Regression

Simple Linear Regression

Multiple Liner Regression

Simple Linear Regression – we only one independent variable and one target variable

Multiple Linear Regression – we have multiple independent variables and one target variable

Line of regression is defined as -> **y = mx + c**

y = dependant variable/label/target

x = independent variable/predictor/

m = slope of the line/co-efficient

c = intercept value/co-efficient

Cost Function: This is a very important concept; this is the function which tells how well the regression line fits the given set of data and it is represented by letter “J”.

And the cost function for linear regression is RSS (Residual Sum of Square) or also called as MSE (Mean squared error) represented by below formulae.

$$\boxed{RSS = \sum_{i=1}^{n}(y_i - (\alpha + \beta x_i))^2}$$

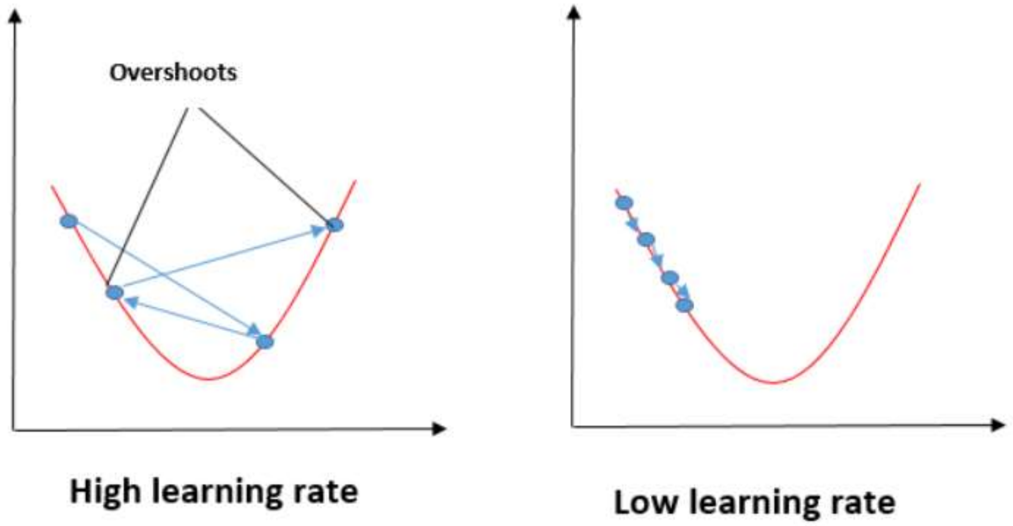
When the value of RSS is the least, we can say that we have the best fit line.

Gradient Descent:

This is the method which helps to find the least values of the coefficients to reduce the cost function which intern ensures to arrive at the best fit line that defines the regression line.

Mathematically this is achieved by applying concept of derivatives and these coefficients are randomly changed in a manner that our objective is to reach the min cost function. The value of the change is always kept low as we would like to approach the minimum value in a way that we do not miss it.

This approach of small changes to reach minimum cost function is also called as learning rate. Greater the learning rate, we have a chance of missing the minimum/optimum cost function.



2. Explain the Anscombe's quartet in detail?

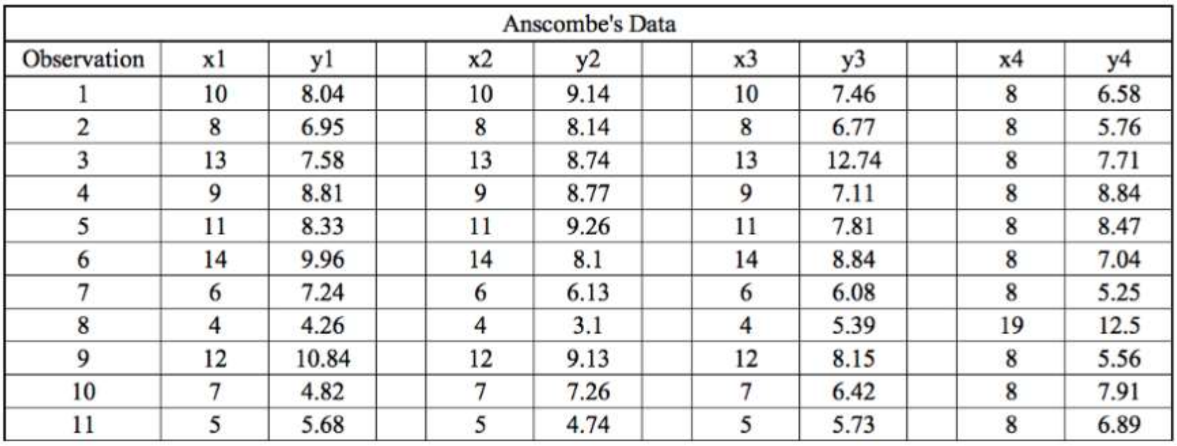
Ans:

Anscombe quartet was conceptualised by statistician Francis Anscombe.

In Anscombe quartet, we have 4 datasets which have similar statistic data points i.e. mean, median, variance etc but when they are plotted, they paint a different pictures.

So, he laid the emphasis that, statistical model building is not the ultimate method involved in analysing the data, but visualizing the data before building the model gives insights which would be missed in a statistical viewpoint.

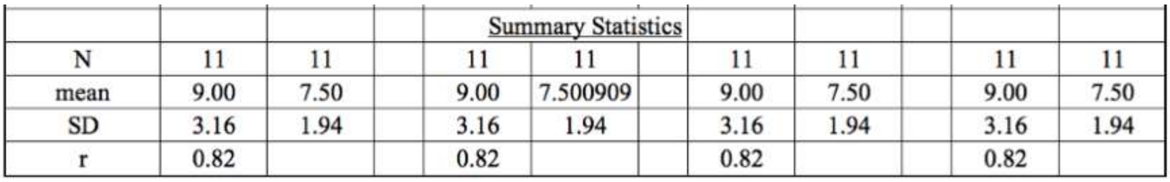
Anscombe Data:



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

Image by Author



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 by Author

Plot:

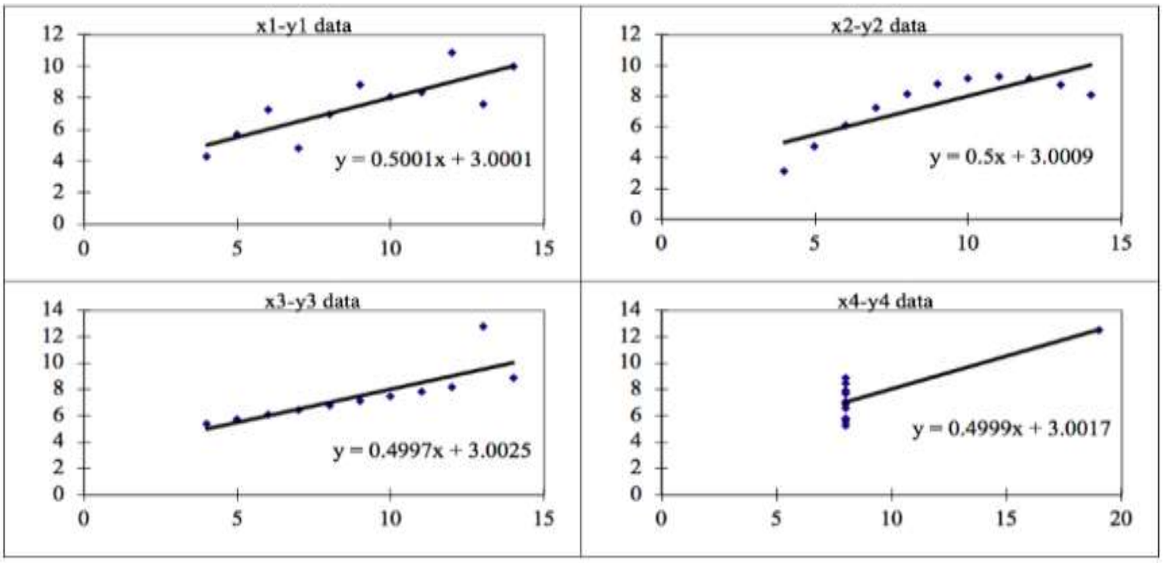


Image by Author

It is equally important to observe that, while the statistical datapoints are similar, the linear regression model is not able to explain for all the points.

3. What is Pearson's R?

Ans:

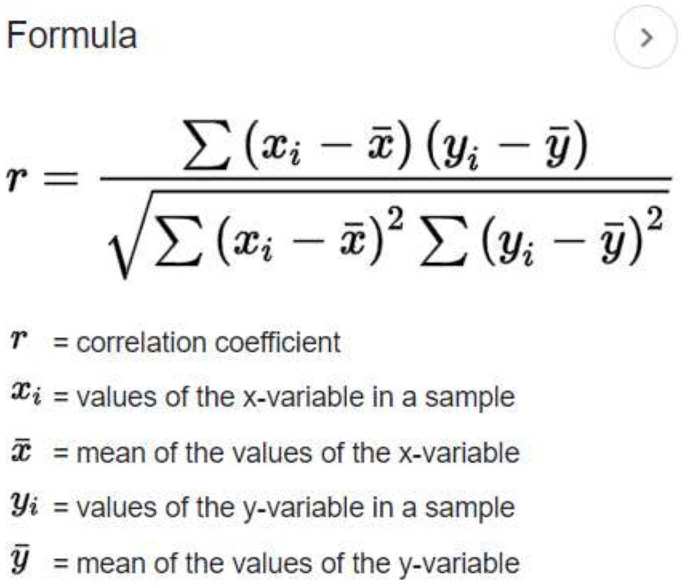
Pearson’s R – is also called as pearson’s correlation coefficient and was formulated by Karl Pearson from a related idea by Francis Galton in the 1880s.

Lets understand what is correlation – it’s the strength of association between two variables. It could either be negative or positive. So, in general a correlation could be ranging from -1 to +1 signifying the strength of association.

Pearson’s R or Pearson Correlation is one type of correlation calculation which tells us about the strength of influence of one variable over the other

Pearsons correlation can determine linear correlation between two variables and not non-linear once. It cannot differentiate between dependant and independent variables either.

Definition - **Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations**.



Formula

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

r = correlation coefficient

x_i = values of the x-variable in a sample

$\bar{x}$ = mean of the values of the x-variable

y_i = values of the y-variable in a sample

$\bar{y}$ = mean of the values of the y-variable

Assumptions for Pearson R to be accurate:

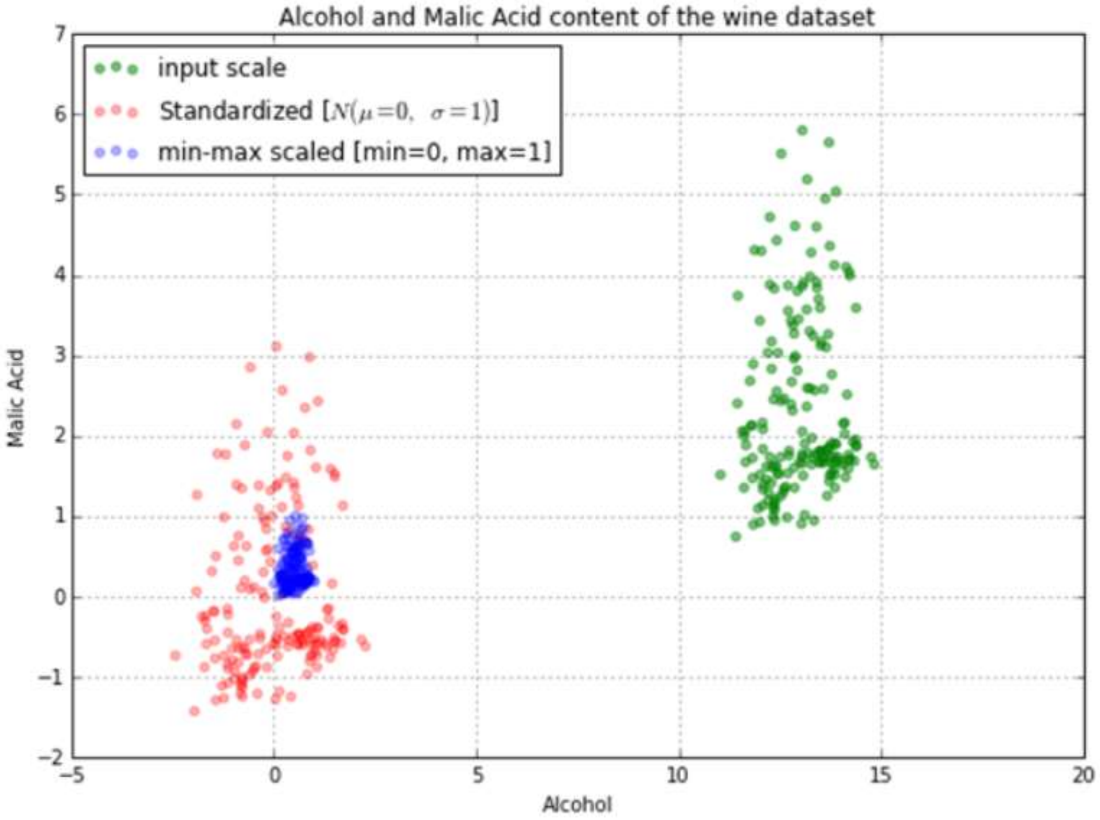
- Scale of measurement should be interval or ratio
- Variables should be normally distributed
- The association should be linear
- There should be no outliers in the data

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

Ans:

Scaling is the method to normalize the independent variable/feature into a certain range before we fit the data into the model.

By scaling we bring various independent variables within similar range. Below is an image which shows how the data looks post normalizing it with respective scaling method.



The impact of Standardization and Normalisation on the Wine dataset

Widely used scaling methods are

1) Normalization (Min Max scaling)
2) Standardization

1) Normalization:

In this method of scaling the values are scaled down with the range of 0 to 1 and it is achieved using below mathematical formulae

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Here the Max and Min values are of the respective feature or independent variable.

2) Standardization:

In this method the data is scaled in such a way that the mean is zero and the standard deviation is one.

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

Sigma is the standard deviation and x-bar is the mean, by calculating these values, we can substitute the same for each data point and find out the scaled value.

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

Ans:

VIF is the measure of multicollinearity.

Multicollinearity is the nothing but the linear correlation between two independent variables and when this happens, it becomes difficult to ascertain the impact of which variable is greater or lesser over the over all model. VIF is the measure of multicollinearity.

As a general rule:

VIF –

0 to 5 – Non-Collinear/Mild

5 to 10 – Moderate

>10 – Severe

Formula for VIF :

$$VIF = \frac{1}{1 - R^2}$$

When we see highest degree of collinearity, it means the R2 value is equal to 1.

Which makes the equations

VIF = 1/(1-1) = 1/0 as the denominator tends to zero, the value becomes infinity.

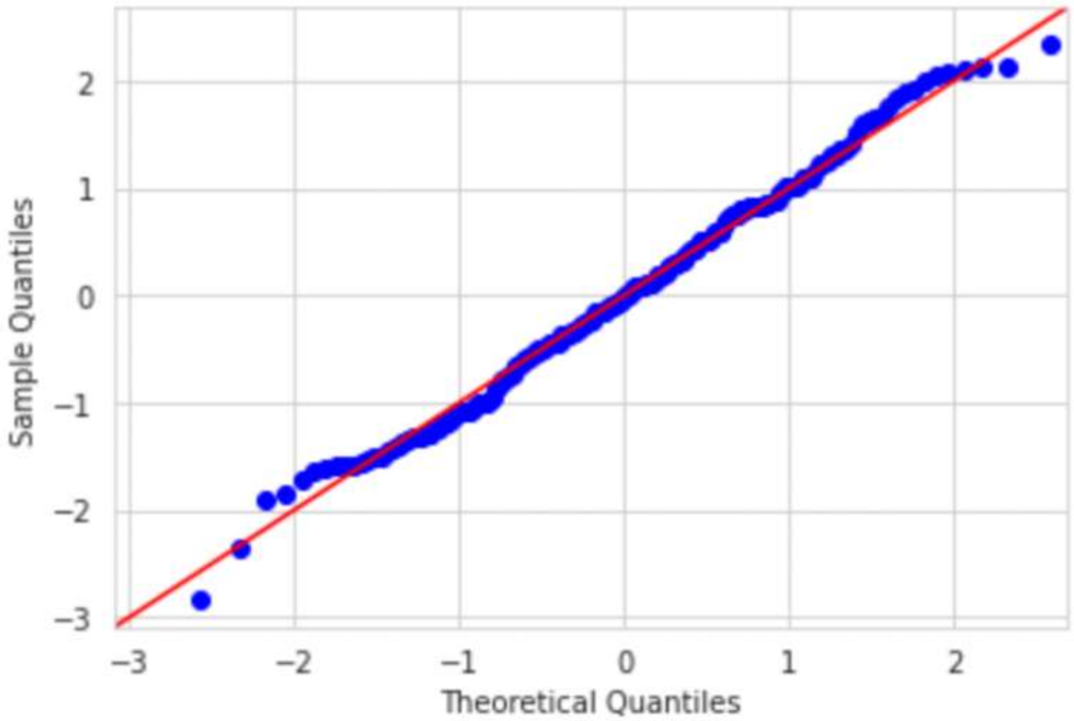
Indirectly, whenever VIF is infinity, it means the correlation is 1.

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

Ans:

Q-Q Plot is called as quantile quantile plot, and it basically plots the quatile level distribution between the sample distribution to that of theoretical distribution.

Its primary objectives is to see whether the 2 set of data have the same distribution or not.



Above image is of a normally distributed data.

If the data is linearly related then we will see the plots on the red line. Q-Q plot also helps us to find the skewness, similarity between the data distribution etc.

Q-Q plot not just checks for normal distributions, but you can also check for uniform distribution and exponential distributions too.