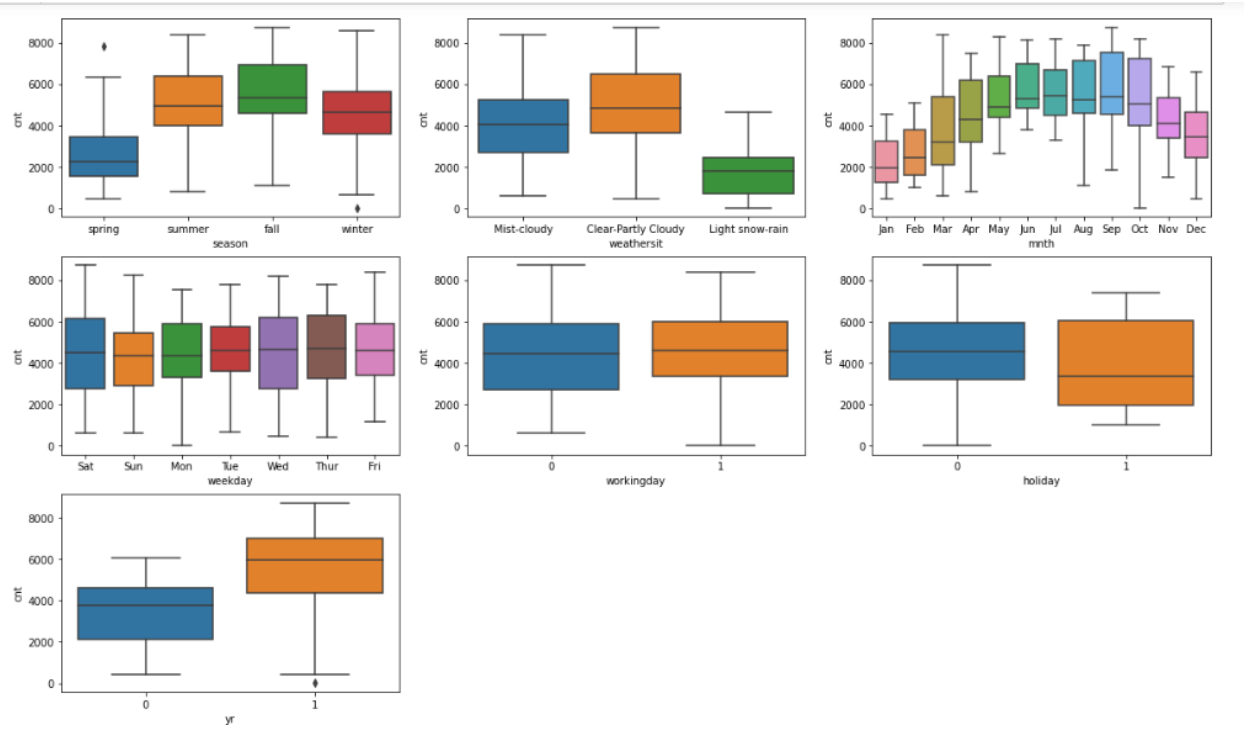
BIKE SHARING ASSIGNMENT SUBMISSION

- Siddhika Shetty

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)



Based on analysis of categorical variables we can infer the below effect on the dependent variable cnt (count):

Year: Bike hiring is more in 2019 than in 2018.

Season: Most of the bike hiring was done in fall and summer, while it is drastically low in spring

Weathersit: Bike rental count fell drastically in Light snow-rain

Month: Between Apr to Oct a huge number of bikes were hired, quite above the median of 4000.

Working Day, Holiday, Weekday: On basis of the graphs there is not much difference seen. (except for Saturday, but still I would not count that to be major, on the basis of the plotted graph)

Overall **year, season, month, weather** are major factors for the model (on the basis of initial review of the plotted graphs)

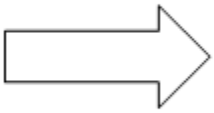
2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)

Drop first = True needs to be used to avoid the dummy variable trap. Dummy variable trap is a scenario in which independent variables are multicollinear / correlated to each other.

Say we have a categorical variable 'temperature' having possible value as High / Low / Normal.

Post creating dummy variables we get 3 new columns labeled 'High', 'Low', 'Normal'. Only one of these values will be 1 for a given record. Like if say temp = Normal in original dataset, then new column Normal will have value '1' and rest two columns will have '0'.

Normal	High	Low
1	0	0
0	1	0
0	0	1

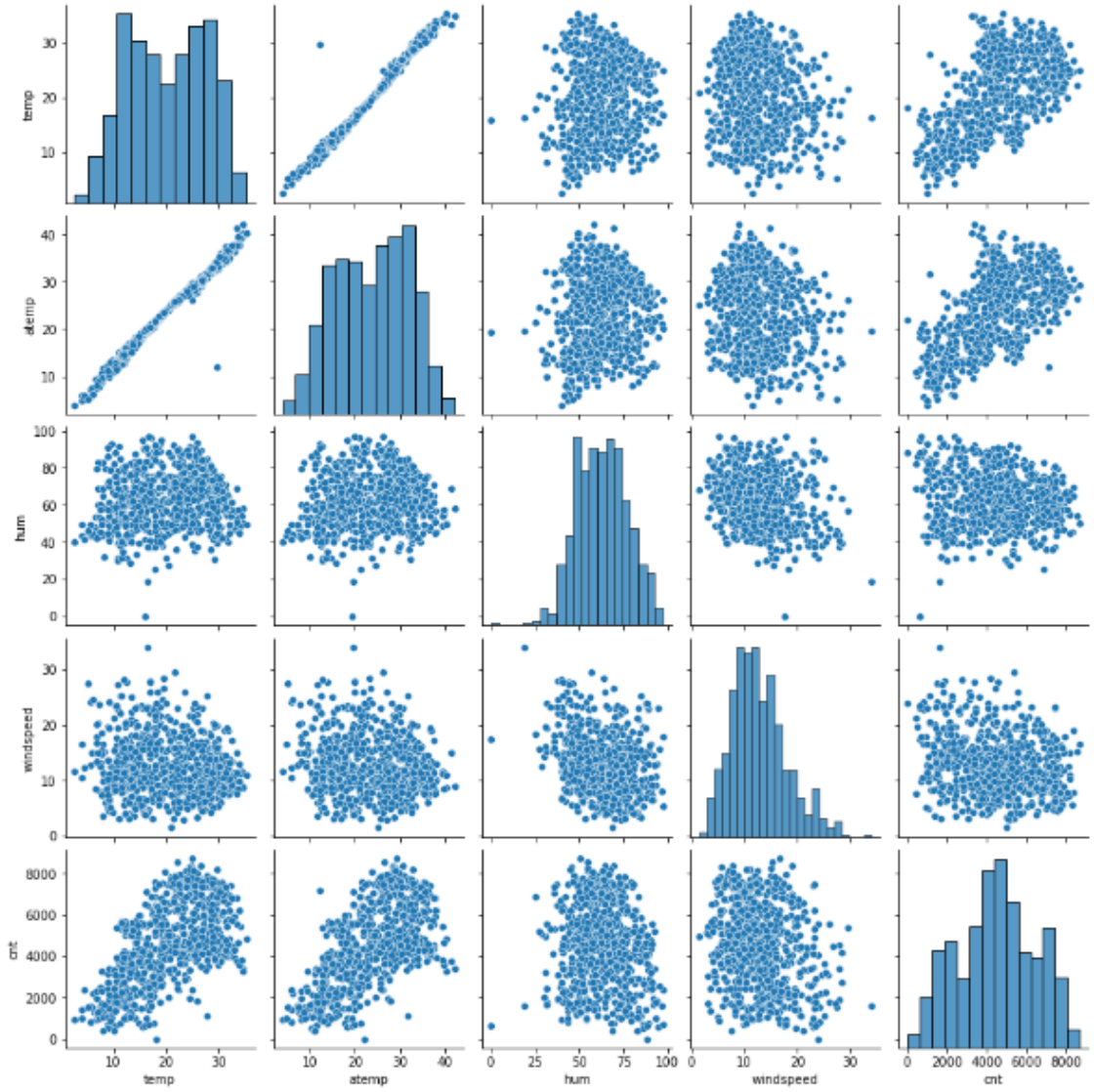


Normal	High
1	0
0	1
0	0

From the above figure it is obvious that the third column is correlated to the values of the other 2 columns. Hence it can be dropped to avoid dummy variable trap.

In a model it is advisable to avoid having highly correlated variables since it becomes difficult for the model to estimate the relationship between each independent variable and the dependent variable independently because the independent variables tend to change in unison.

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



Target variable cnt (Count) has the highest correlation with temp and atemp

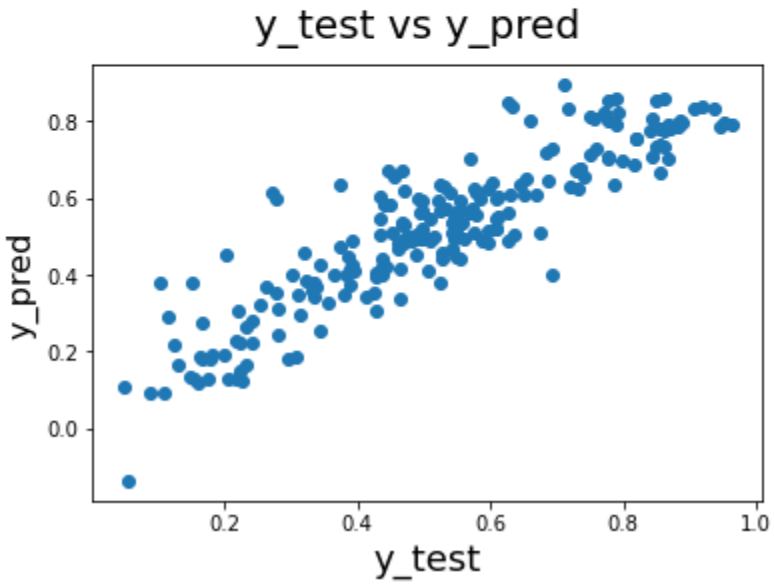
4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Below checks were done to validate the assumptions of linear regression:

a. For evaluating the model we plotted a scatter plot of y_test against y_pred.

```
]:  1  # Plotting y_test and y_pred to understand the spread.
    2  fig = plt.figure()
    3  plt.scatter(y_test,y_pred)
    4  fig.suptitle('y_test vs y_pred', fontsize=20)              # Plot heading
    5  plt.xlabel('y_test', fontsize=18)                          # X-label
    6  plt.ylabel('y_pred', fontsize=16)                          # Y-label
```

```
]: Text(0, 0.5, 'y_pred')
```



As its evident from the graph that it's a linear relation so this model is able to predict y values well.

b. Another check we did was to compare the R squared values of the test and train set:
While R squared of train set is 0.833

R squared of the test set came to be 0.803, which is well within the max difference of 5% and hence this is a good model

calculate the R-squared score on the test set

```
n [48]:  1  from sklearn.metrics import r2_score
         2  r2_score(y_test, y_pred)
```

```
ut[48]: 0.8035441330582012
```

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

Covariance Type: nonrobust

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.1910	0.030	6.456	0.000	0.133	0.249
yr	0.2341	0.008	28.246	0.000	0.218	0.250
holiday	-0.0969	0.026	-3.691	0.000	-0.148	-0.045
temp	0.4782	0.033	14.446	0.000	0.413	0.543
windspeed	-0.1482	0.025	-5.860	0.000	-0.198	-0.098
spring	-0.0551	0.021	-2.641	0.009	-0.096	-0.014
summer	0.0610	0.014	4.271	0.000	0.033	0.089
winter	0.0959	0.017	5.730	0.000	0.063	0.129
Light snow-rain	-0.2860	0.025	-11.492	0.000	-0.335	-0.237
Mist-cloudy	-0.0801	0.009	-9.090	0.000	-0.097	-0.063
Sep	0.0909	0.016	5.565	0.000	0.059	0.123

Omnibus:	63.599	Durbin-Watson:	2.076
Prob(Omnibus):	0.000	Jarque-Bera (JB):	143.759
Skew:	-0.674	Prob(JB):	6.07e-32
Kurtosis:	5.225	Cond. No.	17.2

Notes:

On the basis of the final model (looking at the coeff) we can say **temp, year and light snow – rain** are top 3 major features contributing significantly towards explaining bike hiring trend

General Subjective Questions

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

Machine learning algorithms are majorly classified as **Supervised** and **unsupervised**, based on the input (while supervised uses labeled input and output, unsupervised does not).

Now supervised algorithms are broadly further classified as Regression (where output variable to be predicted is a continuous variable) and classification (where output variable to be predicted is a categorical variable).

Machine Learning algoriths

- Supervised
  - **Regression**
  - Classification
- Unsupervised

Linear regression algorithm is a basic form of machine learning where we train a model to predict the behavior of data based on the input variables. As name suggests, the 2 variables which are on x and y axis should be linearly correlated and the output is a continuous variable.

Basic mathematical expression -

$$y = mx + c$$

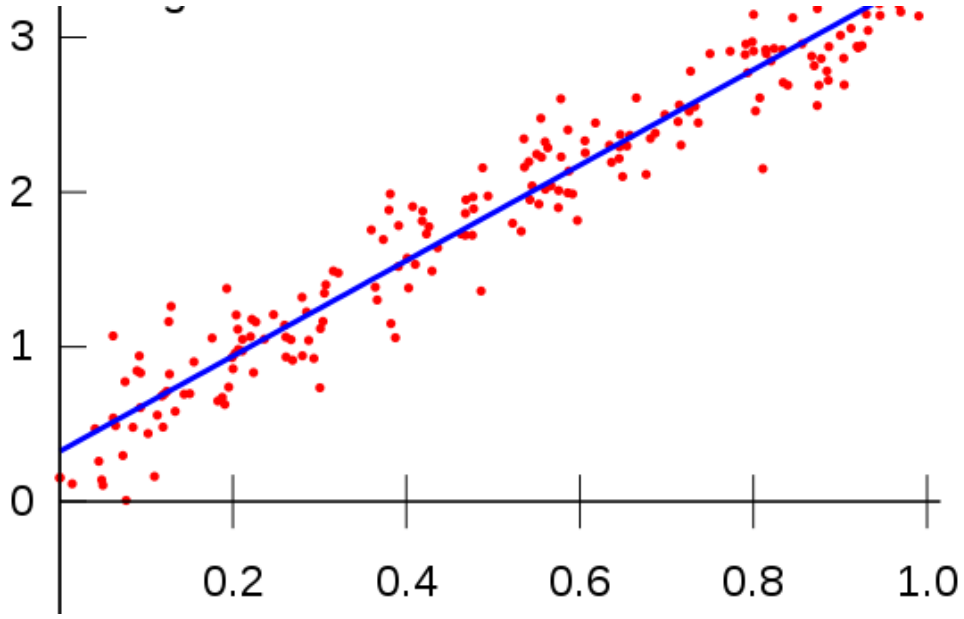
where

y -> dependent variable

x -> input independent variable x

c -> constant

m -> slope



Use Cases of Linear Regression:

1. Prediction of trends and Sales targets – To predict how industry is performing or how many sales targets industry may achieve in the future.
2. Price Prediction – Using regression to predict the change in price of stock or product.
3. Risk Management- Using regression to analyze Risk Management in financial and insurance sector.

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

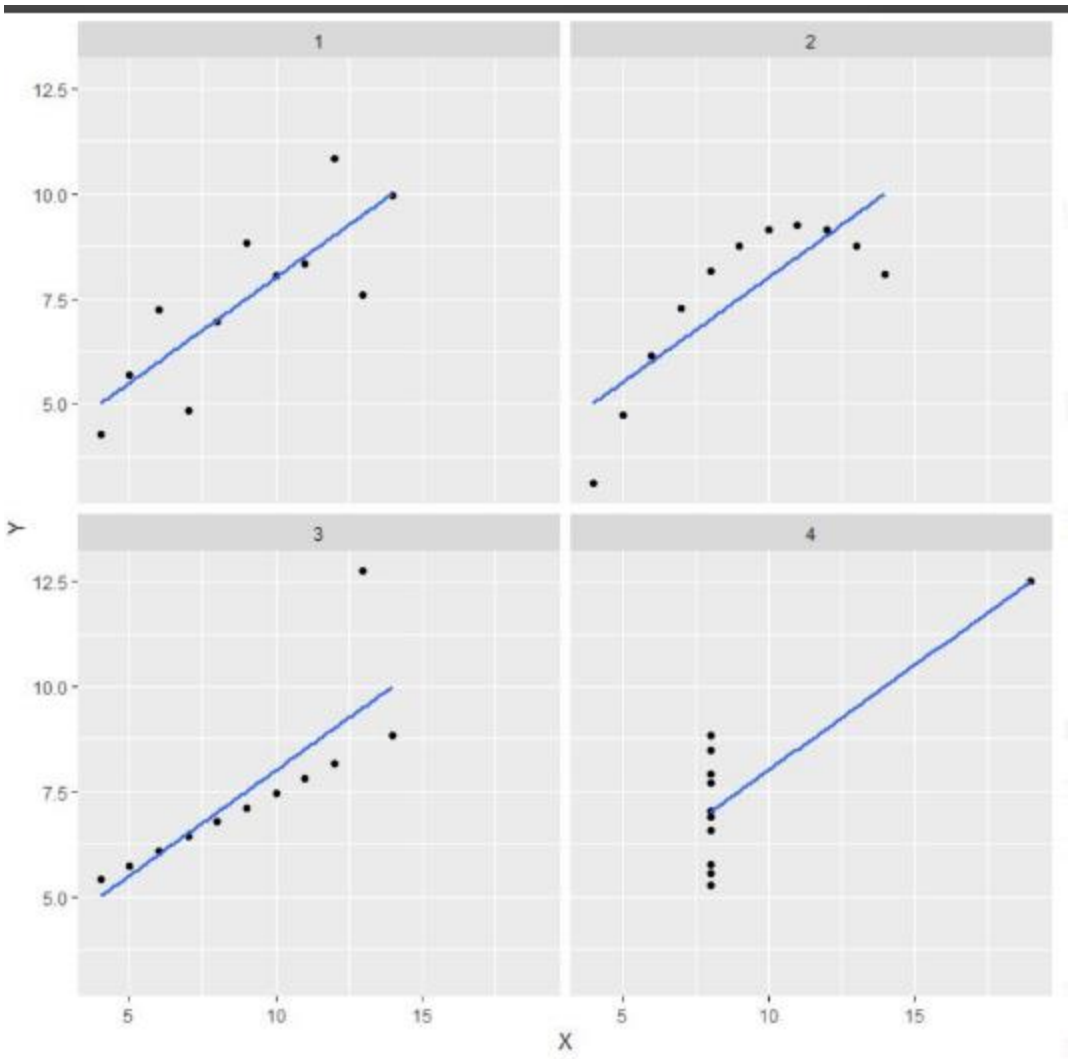
Anscombe's quartet (constructed by statistician Francis Anscombe) is a set of 4 datasets (of 11 daata points each) that highlights the importance of plotting data to confirm the validity of the model fit.

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

The 4 datasets appear similar when using summary statistics

- The average x value is 9 and average y value is 7.5 for each dataset
- The variance for x is 11 and the variance for y is 4.12
- The correlation between x and y is 0.816 for each dataset
- A linear regression (line of best fit) for each dataset follows the equation y = 0.5x + 3

However the graphical presentation shows that the patterns are very different from one another



Importance of Anscombe's quartet:

This quartet is used to highlight the importance of checking the graphical representation of the data before starting further analysis and the inadequacy of summary statistics for some scenarios.

This is not to say that summary statistics is useless, just that it can be misleading on its own.

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

Correlation coefficient formula is used to measure how strong a relationship is between two variables. It calculates the effect of change in one variable when the other changes.

Commonly used in linear regression. It normally returns value in range from -1 to 1 where

- 1 indicates strong positive correlation. For every positive increase in one variable, there is a positive increase of fixed proportion in the other
- -1 indicates strong negative correlation. For every positive increase in one variable, there is a decrease of fixed proportion in the other variable.
- 0 indicates no relationship at all.

There are several types of correlation coefficient formulas, but the most popular one is Pearson's correlation (referred to as Pearson's r). Formula for calculating Pearson's r

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

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

Scaling is a technique to normalize the independent variables into a fixed range.

Machine learning algorithm works on numbers and does not know what the numbers represent. So if there is a vast difference in the range of 2 different variables then the high ranging variable gets a more decisive role in training the model. Say for loans data loan amount ranges in lakhs while the loan period is between 1-20 years. While developing a machine learning model for loans if we don't scale these columns then loan amount will have a more decisive role in the final model.

Most common techniques of scaling are:

- Normalized scaling – where we bound the values between 0 & 1. Also referred as Min-Max scaling. Normalization is highly impacted by outliers. Used when data does not fit into Gaussian curve representation.

$$X_{new} = \frac{X_i - \min(X)}{\max(x) - \min(X)}$$

- Standardized scaling – is based on standard deviation. Here the data is transformed to range between mean 0 and have standard deviation as 1. Standardization is less impacted by outliers. Used when data represents Gaussian curve

$$X_{new} = \frac{X_i - X_{mean}}{\text{Standard Deviation}}$$

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

VIF for a regression model is equal to the ratio of the overall model variance to the variance of a model that includes only that independent variable.

VIF = 1 / (1 – R squared)

Where R squared = coefficient of determination of the regressive equation

R squared is a measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable / variables in a regression model.

Now when R squared reaches 1 it's obvious that VIF reaches infinity.

So high / infinite VIF indicates that the associated independent variable is highly collinear with the other variables in the model and can be expressed by a linear combination of other variables and hence is redundant in the model

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

Q – Q (Quantile – Quantile plot) is a probability plot created for comparing two probability distributions by plotting the quantile of first data set against quantile of second data set.

A 45 degree reference line (y = x) is also plotted. If the two sets have the same distribution then the points should fall nearly close to this reference line. Greater the departure from this reference line, more is the evidence that the 2 sets come with different distributions

One one axis you plot the data from the hypothesis and on another you plot the data that was returned in the prediction.

Use of QQ plot:

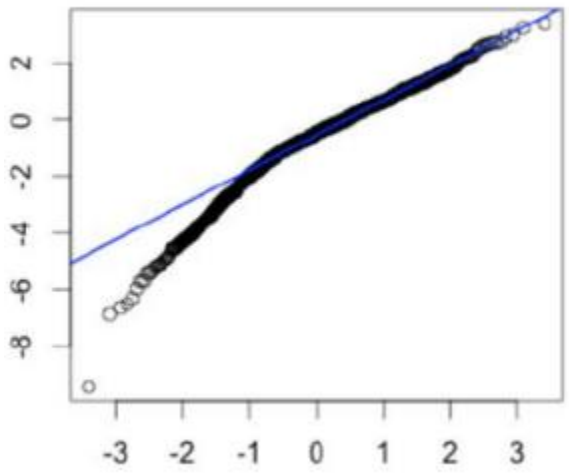
1) This plot is not for visualizing the data , but rather to help check if the data fits a particular model.

If all points plotted on the graph lies on the straight line then it's a **Normal distribution**

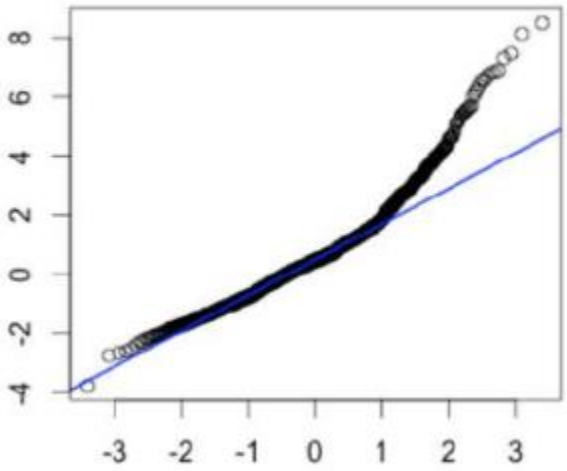
2) Another check is to find the Skewness of a distribution.

If the bottom end of the Q Q plot deviates from the straight line but the upper end does not, then we can say that the distribution is **left skewed or negatively skewed**, while for the reverse case we say its **right skewed or positively skewed**

Left skewed



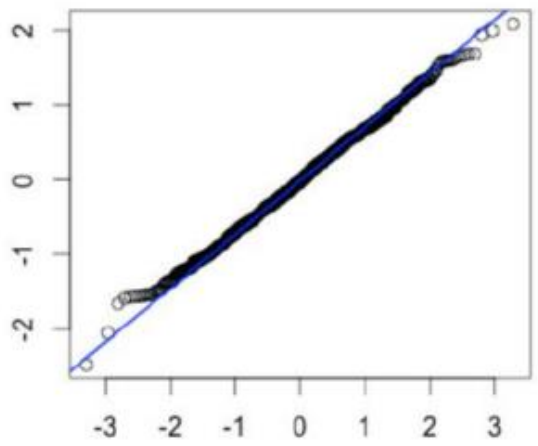
Right Skewed



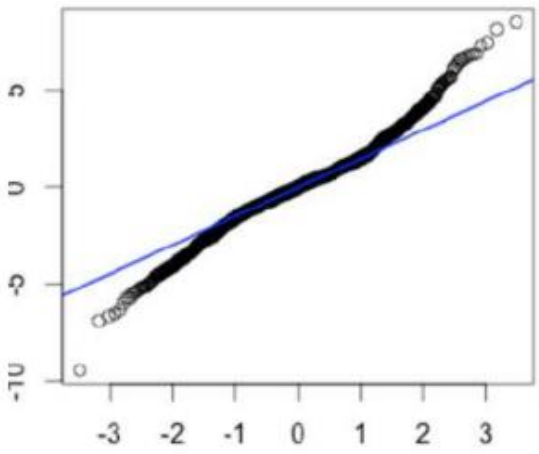
3) Kurtosis

If for a Q –Q plot there is very less deviation at the ends then it's a **thin tailed** distribution. Else if it deviates at the ends it's a fat tailed distribution

Thin tailed



Fat Tailed



Thus the Q – Q plot is used to answer the following queries:

- Are both the datasets from the same distribution
- Do they have similar shapes