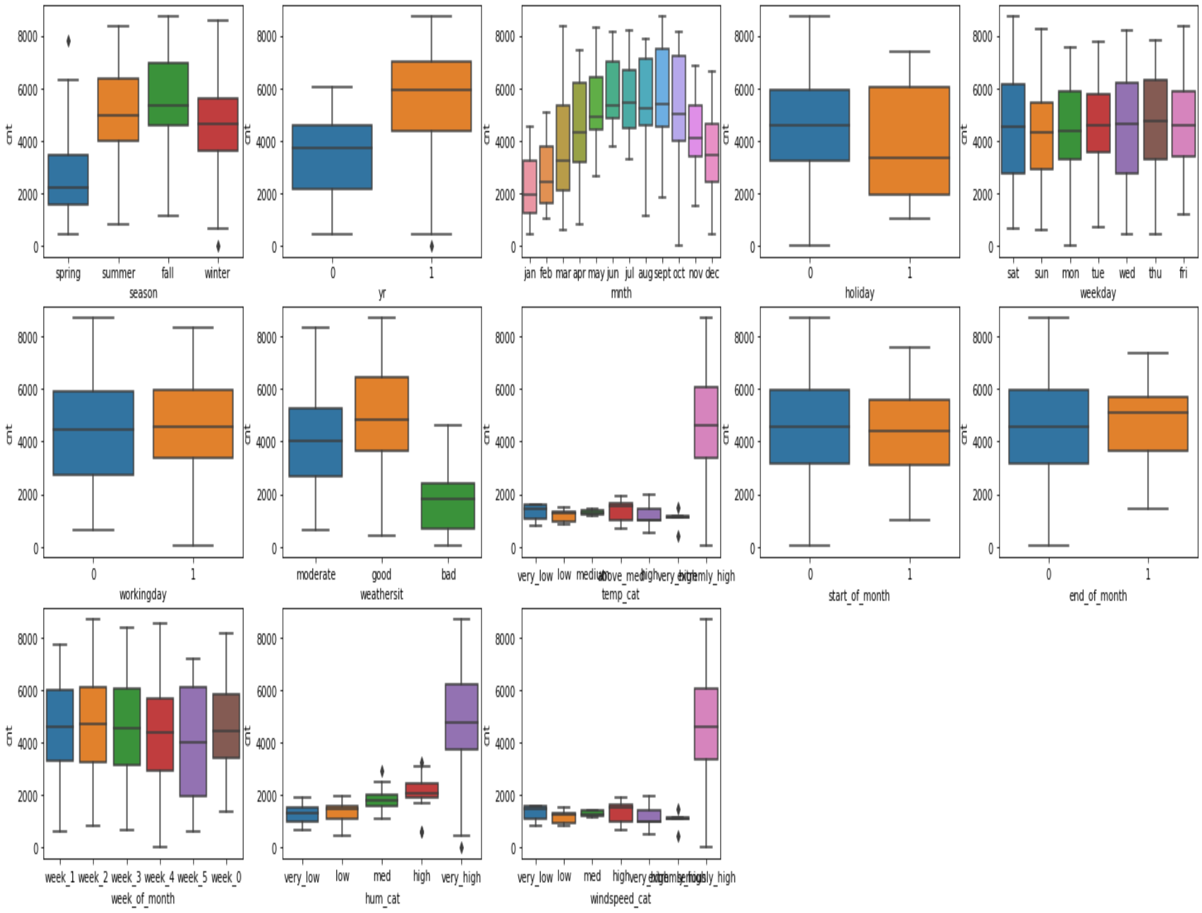
Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)
Ans:

Categorical values that are considered in the dataset, and few also derived are:
'season','yr','mnth','holiday','weekday','workingday','weathersit', 'temp_cat',
'start_of_month','end_of_month','week_of_month', 'hum_cat', 'windspeed_cat'



Note: Graph can be checked in notebook for better visualization

1. Fall has the Highest Demand, whereas spring has the lowest demand
2. Demand in bike sharing business has increased in 2019 compared to 2018
3. Demand keeps on increasing till September, with September Month being month of Highest demand. And there is decreased demand from Oct to Jan.
4. Higher demand when it is not a Holiday.
5. Weekends have slightly higher demand, but the mean is more or less the same. We can derive that there is not much impact of day of week in the data. And similar is with working day
6. Bad weather has the lowest demand, whereas when the weather is good, there is a high demand.
7. Start of the Month has less deviation in demand compared to the rest of the days.

8. End of the Month also has less deviation in demand compared to the rest of the days.
9. Week of the Month does not have much to say

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

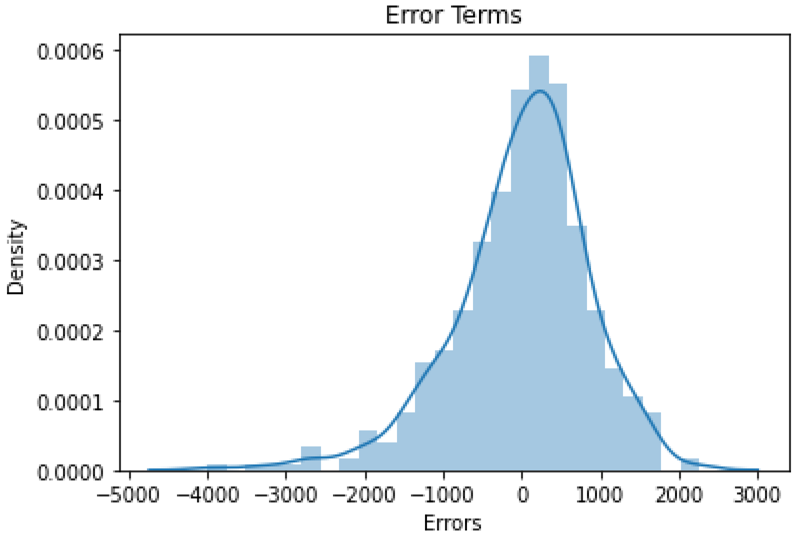
Ans:
drop_first=True parameter in the dummy variable creation is important in order to maintain the N-1 Columns,
Example:
If we have 12 Columns of the Months, then the final number of columns for dummy variable needs to be N-1 which is achieved by drop_first=True

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)
Ans:
Looking at the pair-plot , variable with highest correlation is variable “temp” with correlation of 0.63 (as seen in heatmap) with respect to “cnt”.

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

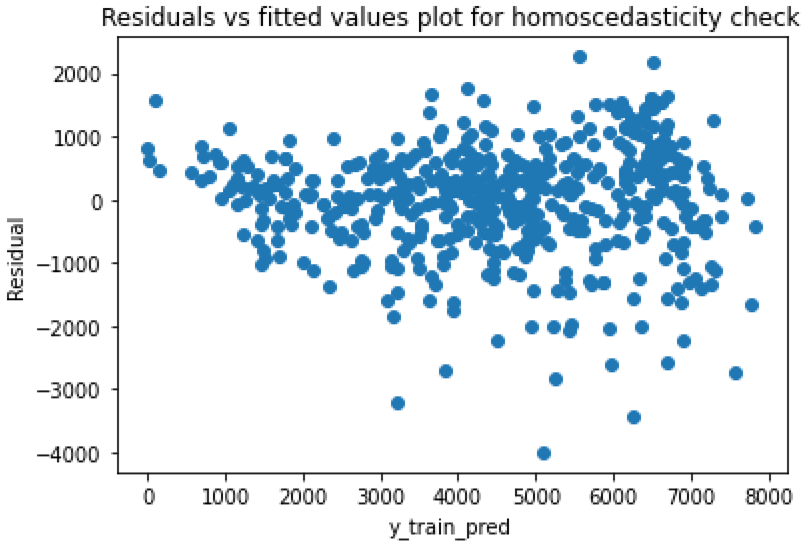
To Validate the assumptions of Linear Regression after Building, It was validated using three parameters:

1. Check for Normally Distributed Error Terms.
   As we can see all the error terms follow Normal Distribution with 0 as mean.

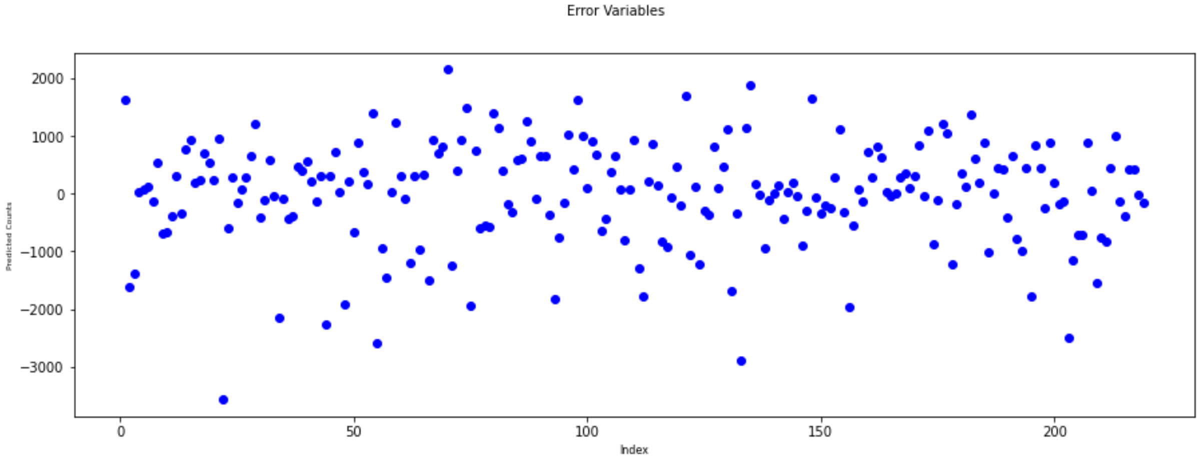


2. Homoscedasticity Check
   Homoscedasticity is also known as absence of heteroscedasticity. Which generally means that there is an impact of outliers in the model. The Homoscedasticity test

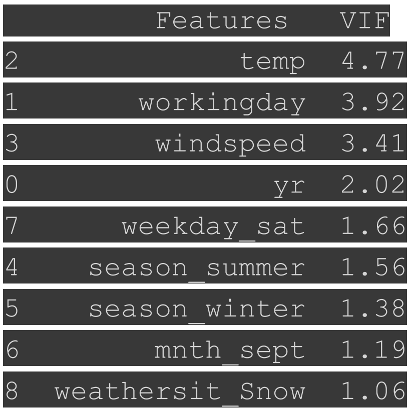
shows that there is no variance in the test variables



3. Randomly Distributed Error variables with respect to Predicted Counts
   The residuals are independent of each other. In the plot we can see that the variables are distributed randomly and do not have any correlationship.



4. Selecting a model with low VIF scores of Variable to Avoid Multicollinearity.
   The Independent Variables are not correlated to each other. This tends to keep the model to be built using independent variables and reduce the number of variables required to predict the model.
   As we can see below, all the VIFs in the final model are below 5.0



		Features	VIF
5.	2	temp	4.77
6.	1	workingday	3.92
7.	3	windspeed	3.41
8.	0	yr	2.02
9.	7	weekday_sat	1.66
10.	4	season_summer	1.56
11.	5	season_winter	1.38
12.	6	mnth_sept	1.19
13.	8	weathersit_Snow	1.06

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

On Based on Final Model, top 3 Features contributing significantly are:

1. *"Temp"* , with coefficient of 4960.5648

2. *"Weathersit category 3*"
   i.e. Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds (with coefficient of -2236.57)

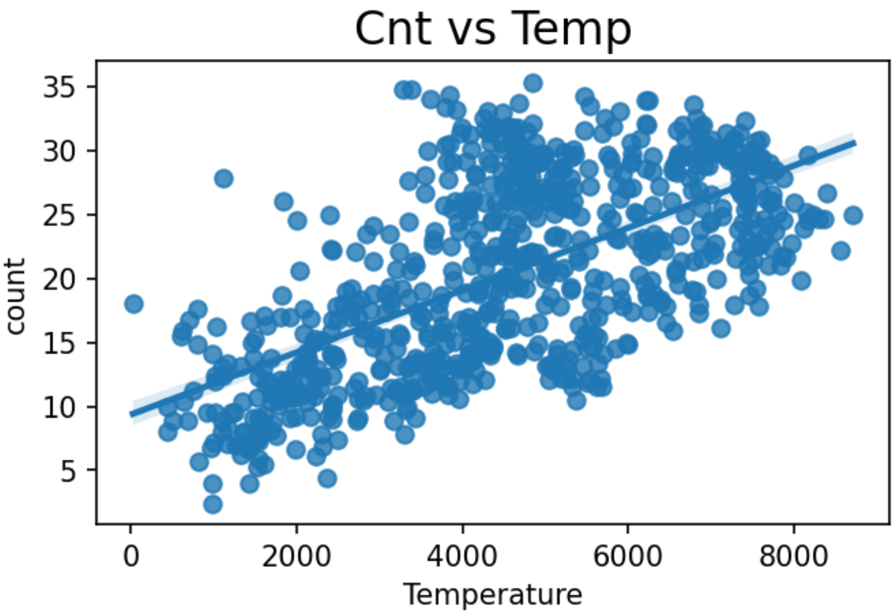
3. *"yr* ", this variable explains the demand based on 2018 and 2019. Here we can clearly see that demand is increasing in consecutive years. (with coefficient of 2037.00)

General Subjective Questions

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

Linear Regression is used to predict the target variable, given some Input data points. The basic principle of linear regression is to create a model which can use correlation of two or more variables and fit a line between the data points to calculate or predict target variable.

1. **Concept of Linearity:**
   When two variables are plotted on a graph, and we fit a line through the points, the resulting correlation between the given points is called linearity.



Here in the graph above, we can see that there is an upward trend in the fit line. This property can be further harnessed to create Multivariable linear regression.

2. **The Model:**
   The statistical/ML Model which is created is nothing but a mathematical equation which is similar to the equation of a line.
   Y = mx + c

   The Equation can be modified to get the the relation between two variables and generally denoted as:
   Y = β0+β1X+[ϵ]

   Here,
   Y is dependent Variable,
   X is Independent Variable
   ϵ is the error term

3. **Evaluation:**
   To Evaluate the Linear Regression, we use certain statistical metrics, like R-Squared Score, Adjusted R-Squared Score, Root Mean Square Error etc.

   R-Squared score is a statistical measure that determines how good the fit is between the variables. A R-Squared score lies between 0 to 1

$$R^2 = 1 - \frac{RSS}{TSS}$$

   R^2 = coefficient of determination
   RSS = sum of squares of residuals
   TSS = total sum of squares

   Adjusted R-Squared score is upgraded form of R-Squared Score such that also considers the number of variables which are used to build the model and in turn penalises the model's R-Square score

$$Adjusted\ R^2 = 1 - \frac{(1-\ R^2)(N-1)}{N-p-1}$$

   Where
   R^2 Sample R-Squared
   N Total Sample Size
   p Number of independent variable

   Root Mean Squared Error is nothing but standard deviation of the residuals of the predicted variables.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(Predicted_i - Actual_i)^2}{N}}$$

4. **Assumptions of Linear Regression**
   Linear relationship: This Assumption in Linear Regression is one of the most important assumptions. Where it states that there is some linear relationship between the dataset variables and the target variable.

No auto-correlation or independence: The residual terms are independent of each other. I.e. there is no correlation between the error terms. The Residual terms need to be randomly distributed.

No Multicollinearity: It is considered that there is low multicollinearity in the model. This can be achieved using the VIF.

Homoscedasticity: This assumption states that there is absence of heteroscedasticity. This ensures that the model is not influenced heavily by outlier terms in the data set

Normal distribution of error terms: This assumption states that the error terms are normally distributed with its mean at 0.

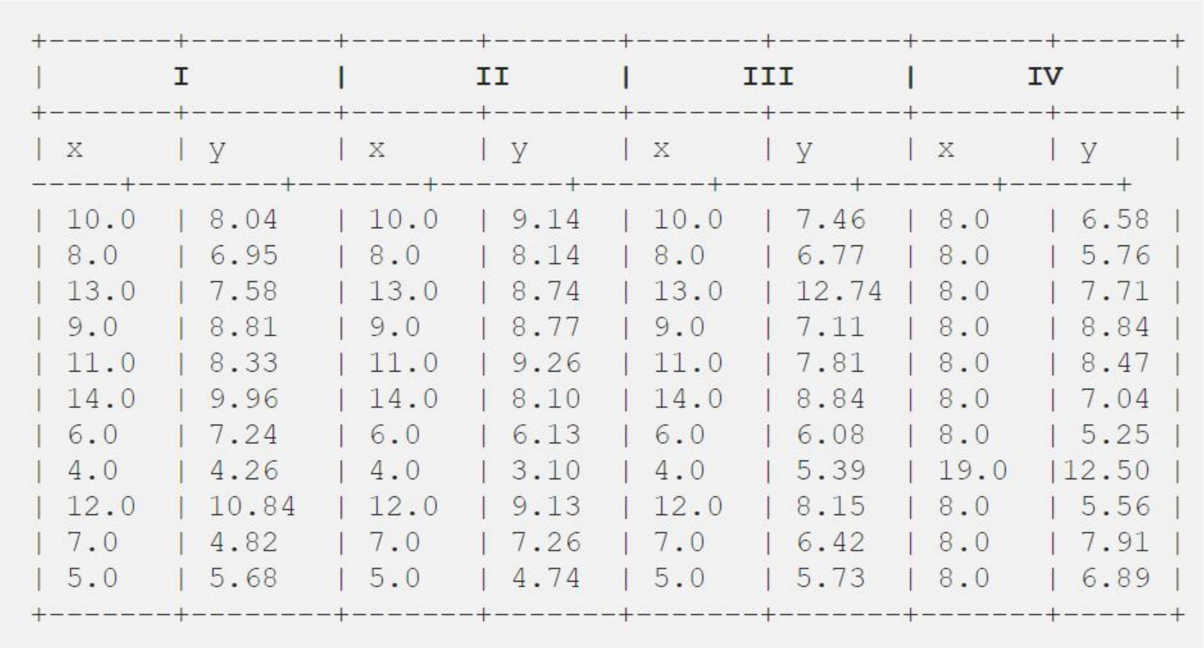
2. Explain the Anscombe's quartet in detail. (3 marks)
Ans:

Anscombe's quartet is a combination of four data sets such that they have nearly identical descriptive statistics and yet have very different distributions and look different when plotted.

This dataset was first constructed in 1973 by Francis Anscome. He was a statistician and his findings imposed the importance of visualizing the data as part of analyzing it. It also stated the effect of outliers on the statistical properties.

According to Anscome, he quotes *"Numerical Calculations are exact, but graphs are rough"*

Original dataset:

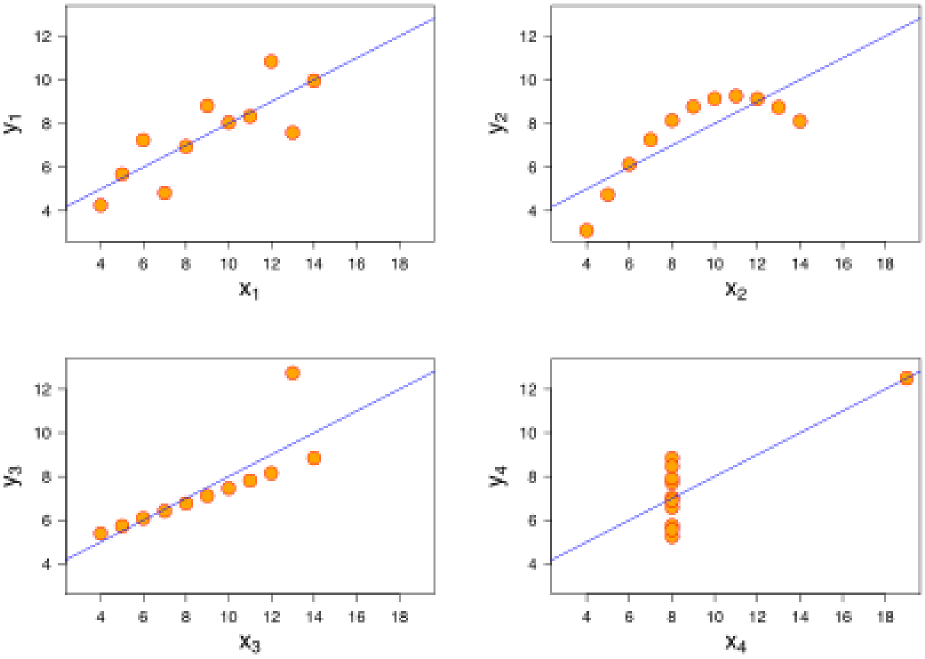


I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	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

Properties of Dataset:

Mean of x = 9, which are exact for all four dataset
Sample Variance of x = 11 exact for all four dataset
Mean of y = 7.50 Accurate to 2 decimals
Sample Variance of y = 4.125 +- 0.003

Correlation between x and y = 0.816, Accurate to 3 decimals
Linear Regression Line: y = 3.00 + 0.500x , accurate to 2 and 3 decimals respectively for x and y
Coefficient of determination of the linear regression : R^2 = 0.67, accurate to 2 decimals



All four Datasets when plotted together

3. What is Pearson’s R? (3 marks)
Ans:

1. Definition:
   Pearson’s R is the ratio of covariance of two variables and the product of their standard deviations. Hence, it has values between -1 to 1.

   It is also widely known as Pearson Product Moment Correlation (PPMC)

2. Formula:

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

   .

3. PPMC does not take Dependent and Independent Variables into account. This is one of the major drawbacks of PPMC.

4. Assumptions of PPMC:
   1: Two variables should be measured on a continuous scale (i.e., they are measured at the interval or ratio level). Here, we assume that the variables are not categorical and hence examples like , time series data, Stocks data etc.

2: Two continuous variables should be paired,such that each case has two values. Here "values" can also be defined as "data points".

3: There should be independence of cases, that is both observations for one case . If observations are not independent, that means the cases are related, and PPMC would not be an appropriate statistical test

There are other sets of Assumptions as well which state the importance of linearity between the variables, Not including outliers, and that there is a need for homoscedasticity.

5. Values:
The Pearson’s R lies between -1 to 1 and hence ,
Value of 1 indicates strong correlation,
Value of -1 indicates strong negative correlation
Whereas, Value of 0 indicates absolute no relation between the variables.

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

<u>Scaling:</u>
Scaling is a statistical method usually used to “scale” the data in the same range for each independent variable in the dataset.

Min-Max Normalization: This technique re-scales a feature or observation value with distribution value between 0 and 1.

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

Standardization: It is a very effective technique which re-scales a feature value so that it has distribution with 0 mean value and variance equals to 1.

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

<u>Reason for Scaling:</u>
Scaling is performed on data preprocessing steps while building a Model. The Major reason for the use of Scaling is to bring all the variables which lie on different scales to be brought under the same range of scales.

For example, while considering a linear regression model which has variables of temperature in Degree Celsius which normally lie between 0 to 35, and wind speeds in km/hr which ranges from 0 to 80 km/hr in all makes it difficult to use them together in the same model.

Here, practically to get the same range of coefficients of similar weight range, we can implement Scaling.

Normalized scaling vs Standardized scaling:

- Normalized Scaling has boundaries for Maximum and Minimum values in the scale, i.e from 0 to 1 or -1 to 1. Whereas in Standardized Scaling, Mean and Standard Deviation are used, and hence do not have well defined boundaries.
- Normalization is used when we require all the data on the same scale, whereas Standardization is used to ensure that all the data has a similar mean and standard deviation.
- Normalization is affected by outliers and hence it is important to take care of outliers before using Normalization. Whereas, In Standardization, Outliers does not make any concerning effects.
- Normalization is generally used when the distribution of the variables does not follow Normal Distribution, whereas In Standardization, the data is expected to be Normally distributed so as to have similar mean and Standard Distribution.

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

1. VIF stands for Variance Inflation Factor. VIF is a measure used to calculate the Multicollinearity in the Model.

2. Formula of VIF:

$$VIF_i = \frac{1}{1 - R_i^2}$$

   Here, R^2 is a statistical measure given for each variable.

   General Interpretation of VIF that is accepted:

   - 1 = not correlated.
   - Between 1 and 5 = moderately correlated.
   - Greater than 5 = highly correlated.

3. Reason for infinite VIF:

   As we can derive from the formula of VIF, that VIF is inversely correlated to R^2 value, i.e. if R^2 value is High which means that the given variable is almost perfectly correlated. This makes the VIF very high and hence the value tends to be towards infinity.

   We solve this problem by simply dropping the variable which is highly correlated to the base variable.

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

Q-Q plot stands for Quantile Quantile Plot.

Q–Q plot is a plot, which is a graphical method for comparing two probability distributions by plotting their quantiles against each other.

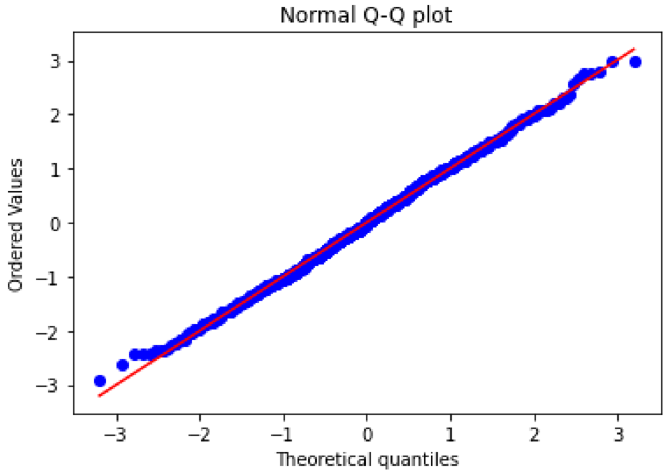
Formula:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

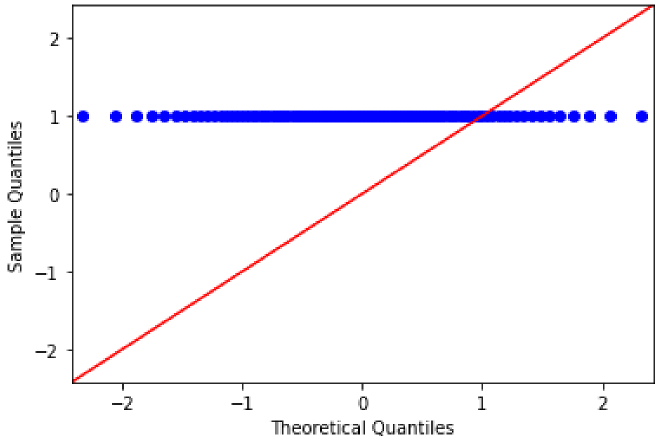
Where, u is mean
And p is sigma

Use Case:
Q-Q plots are used to find the type of distribution for a random variable for Gaussian Distribution, Uniform Distribution, Exponential Distribution or Pareto Distribution.
The probability plot determines if the data points are to be interpreted as which kind of distribution the data follows.

For example, To determine if the data points are Normally distributed, the probability graph would look like this:



Similarly, For data which is uniformly distributed, the Q-Q Plot would look something like this:



- To check whether two samples are from the same population.
- To check whether two samples have the same tail

- To check whether two samples have the same distribution shape.
- To check whether two samples have common location behavior.

Advantages of Q-Q plot

- Two datasets of the given sample size are not required to be equal.
- Dimensions of the dataset are not of much significance.

Q-Q Plot in Linear Regression:

Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.

A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of properties such as location, scale, and skewness that are similar or different in the two distributions.