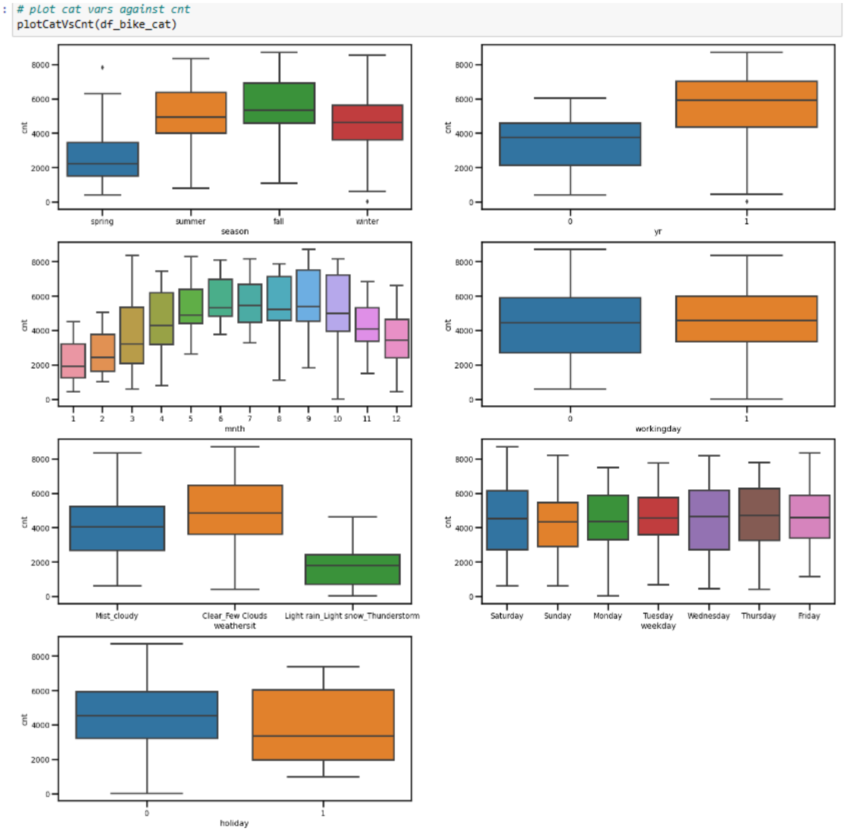
Linear Regression Assignment Subjective Questions:

As answered by Shailesh Kadre - Jan 11 2023 ML C-46 Batch

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?

The categorical variables identified are 'season', 'yr', 'mnth','workingday', 'weathersit', 'weekday', and 'holiday'. Following Box Plot shows their individual effect on bike demand ‘ cnt’.



```
# plot cat vars against cnt
plotCatVsCnt(df_bike_cat)
```

Following is the summary:

- Summer and winter season has highest effect on bike counts
- year 2019 had more counts as compared to year 2018
- April to Sept has maximum number of bike counts due to weather conditions
- working day really not much affect the bike counts
- For rainy days, the bike counts are minimum
- Working day has more bike counts than holidays

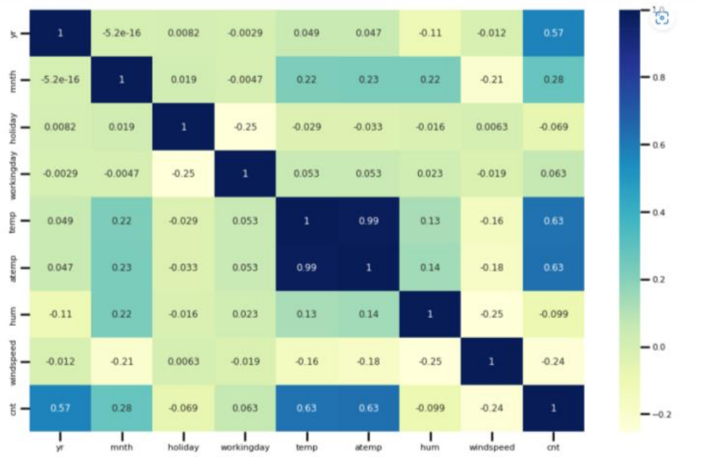
2. Why is it important to use **drop_first=True** during dummy variable creation?

During Dummy variable creation, option drop_first=True is useful as it helps us in the reduction of categorical variable by eliminating extra columns which ultimately reduces the correlations otherwise will be present in the dummy variables.

Let us take an example of days of the week. We know that there are 7 days a week. There is a need to define only 6 days of the week and one day can be skipped as it is not required to be defined.

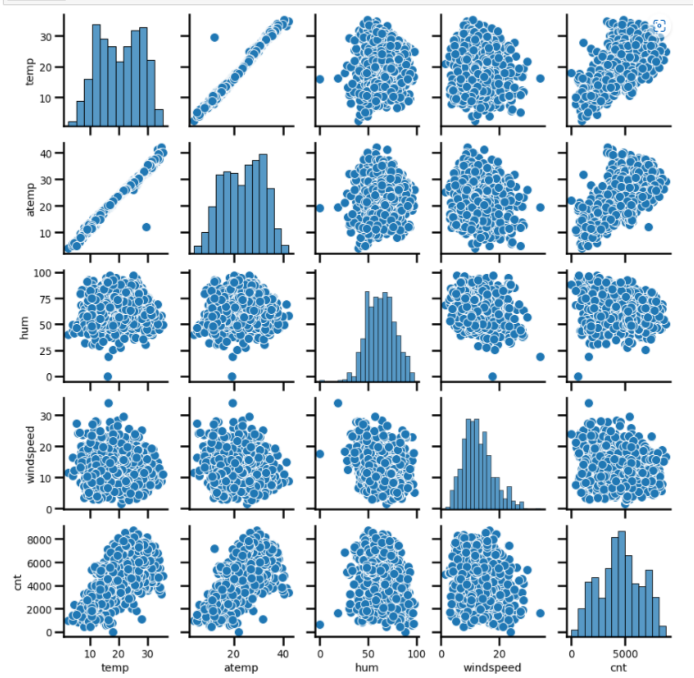
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

In the bike demand problem, the Numerical variables are 'temp','atemp','hum','windspeed', and 'cnt' . Following heat map provides their correlation matrix.



	yr	mnth	holiday	workingday	temp	atemp	hum	windspeed	cnt
yr	1	-5.2e-16	0.0082	-0.0029	0.049	0.047	-0.11	-0.012	0.57
mnth	-5.2e-16	1	0.019	-0.0047	0.22	0.23	0.22	-0.21	0.28
holiday	0.0082	0.019	1	-0.25	-0.029	-0.033	-0.016	0.0063	-0.069
workingday	-0.0029	-0.0047	-0.25	1	0.053	0.053	0.023	-0.019	0.063
temp	0.049	0.22	-0.029	0.053	1	0.99	0.13	-0.16	0.63
atemp	0.047	0.23	-0.033	0.053	0.99	1	0.14	-0.18	0.63
hum	-0.11	0.22	-0.016	0.023	0.13	0.14	1	-0.25	-0.099
windspeed	-0.012	-0.21	0.0063	-0.019	-0.16	-0.18	-0.25	1	-0.24
cnt	0.57	0.28	-0.069	0.063	0.63	0.63	-0.099	-0.24	1

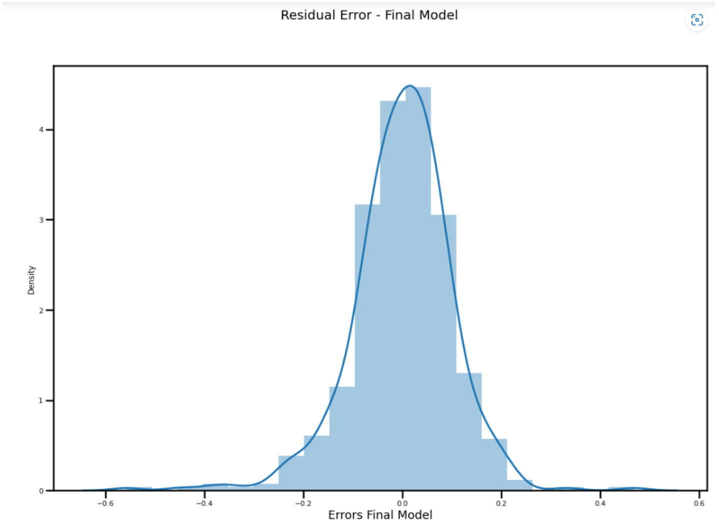
temp and atemp has highest correlation with the target variables It is evident from the following pair plots also.



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

a. Linear Relationship: Target and inputs to have linear fit

b. Multivariate Normality: Residuals have zero mean, normally distributed and constant variance as per below histogram of error residuals



c. No Multi-collinearity: all the variables should be linearly independent

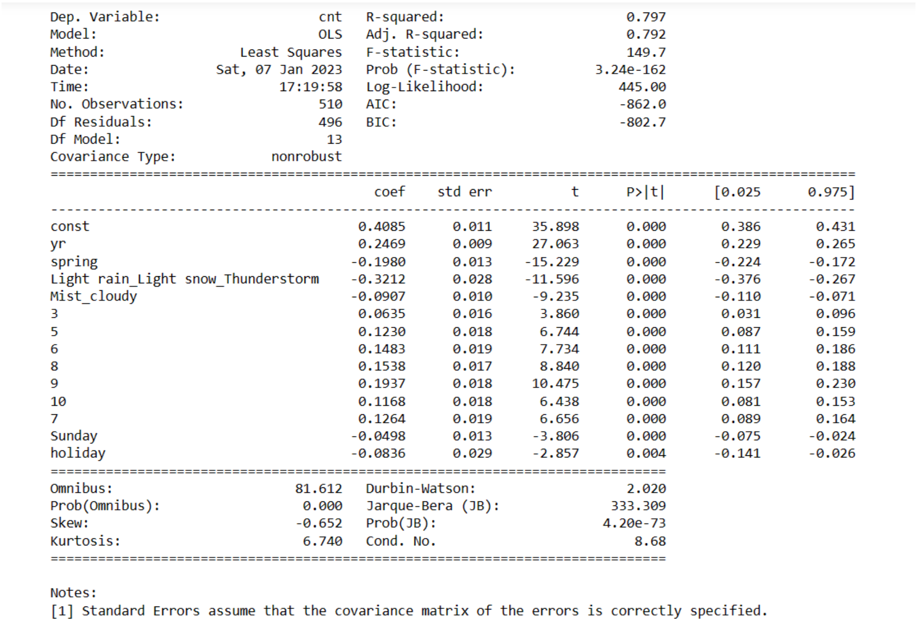
It is validated by followings in the final model:

- All the independent variables have P values less than 0.005
- The VIF values are below 5

d. Homoscedasticity- No patterns in the residuals vs independent variables

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

Following table provides the details of various model parameters .

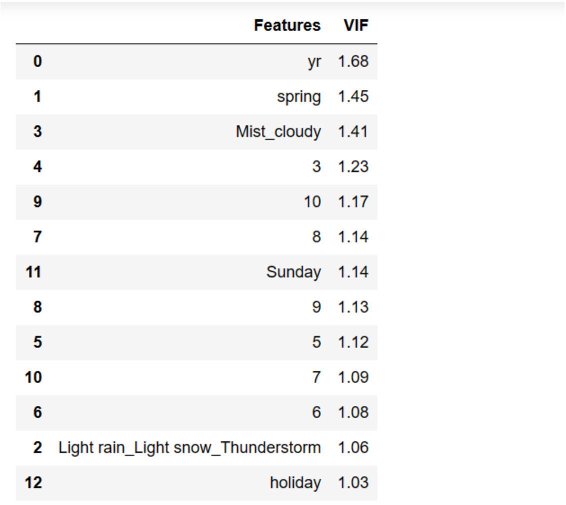


```
Dep. Variable:                    cnt   R-squared:                       0.797
Model:                            OLS   Adj. R-squared:                  0.792
Method:                 Least Squares   F-statistic:                     149.7
Date:                Sat, 07 Jan 2023   Prob (F-statistic):          3.24e-162
Time:                        17:19:58   Log-Likelihood:                 445.00
No. Observations:                 510   AIC:                            -862.0
Df Residuals:                     496   BIC:                            -802.7
Df Model:                          13
Covariance Type:            nonrobust
======================================================================================================
                                         coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------------
const                              0.4085      0.011     35.898      0.000       0.386       0.431
yr                                 0.2469      0.009     27.063      0.000       0.229       0.265
spring                            -0.1980      0.013    -15.229      0.000      -0.224      -0.172
Light rain_Light snow_Thunderstorm   -0.3212      0.028    -11.596      0.000      -0.376      -0.267
Mist_cloudy                       -0.0907      0.010     -9.235      0.000      -0.110      -0.071
3                                  0.0635      0.016      3.860      0.000       0.031       0.096
5                                  0.1230      0.018      6.744      0.000       0.087       0.159
6                                  0.1483      0.019      7.734      0.000       0.111       0.186
8                                  0.1538      0.017      8.840      0.000       0.120       0.188
9                                  0.1937      0.018     10.475      0.000       0.157       0.230
10                                 0.1168      0.018      6.438      0.000       0.081       0.153
7                                  0.1264      0.019      6.656      0.000       0.089       0.164
Sunday                            -0.0498      0.013     -3.806      0.000      -0.075      -0.024
holiday                           -0.0836      0.029     -2.857      0.004      -0.141      -0.026
==============================================================================
Omnibus:                       81.612   Durbin-Watson:                   2.020
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              333.309
Skew:                          -0.652   Prob(JB):                     4.20e-73
Kurtosis:                       6.740   Cond. No.                         8.68
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

From the above table, it can be observed that all the selected input variables have P value zero except holiday for which P value is 0.004 so holiday variable cannot be among the top 3 features which will affect bike demand.

The next criteria is VIF values for the selection of main features: Following table provides the summary of VIF values in the decreasing order and criteria of best feature is lower VIF is batter:



	Features	VIF
0	yr	1.68
1	spring	1.45
3	Mist_cloudy	1.41
4	3	1.23
9	10	1.17
7	8	1.14
11	Sunday	1.14
8	9	1.13
5	5	1.12
10	7	1.09
6	6	1.08
2	Light rain_Light snow_Thunderstorm	1.06
12	holiday	1.03

From the above VIF table, it can be concluded that following three features are most critical for bike demand

a. weathersit – Weather Situation
b. mnth – Month of the year
c. weekday – Days of the week

General Subjective Questions

1. Explain the linear regression algorithm in detail.

As we know in supervised learning, all the data points are labeled. Linear Regression is one of these types of algorithm. Linear regression is a type of predictive technique which enables to establish a input / output relationship between the independent and dependent variables and this is one of the key techniques of Regression Modeling.

In Regression modeling, we have mainly independent variables called as input variables and a dependent variables called as target variables. The main objectives of a linear regression algorithm are as follows:

a. To establish a mathematical relation between input and target variables
b. To determine the relative importance of each independent variable and find the effect of changes of the input variables on the target variables
c. The another key aspect is to make predictions for business problems, such as in our assignment problems where the main objective was to predict the bike demand with the change in inputs such as months, year, weather conditions etc.

Assumptions: The main assumption for linear regression is that the linear relation holds good between the input and target variables. In case it is not possible to fit between input and target advanced regression such as polynomial regression is used.

The multiple linear regression for n observations is having more than one independent variables and is formally defined by following mathematical equation:

$$y_i = \beta_o + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots \beta_p x_{ip} + \varepsilon_i \text{ , } \textit{for i =1,2,…n}$$

Here β_{i1} are the coefficients to be determined. The method of least square is used for the best fit lines for the above observations.

In the multi linear regression model, the fit is a Hyper plane. These coefficients are determined by minimization of sum of square error. It is assumed that that residuals have a zero mean, normally distributed errors and errors with constant variance.

Following are the key assumptions in the linear regression:

e. Linear Relationship: Target and inputs to have linear fit
f. Multivariate Normality: Residuals have zero mean, normally distributed and constant variance
g. No Multi-collinearity: all the variables should be linearly independent
h. Homoscedasticity- No patterns in the residuals vs independent variables

2. Explain the Anscombe's quartet in detail.

it is general perception that if two data sets have same descriptive statistics they are identical when their variables are plotted together. This is an myth and was proven wrong by a Statistician named Francis Anscombe. He developed four sets of data popularly known as Anscombe's Quartet. Each of these datasets have 11 data points (x, y). The main feature of these data sets to have exactly same descriptive statistics. But when these datapoints are plotted, they provide entirely different picture.

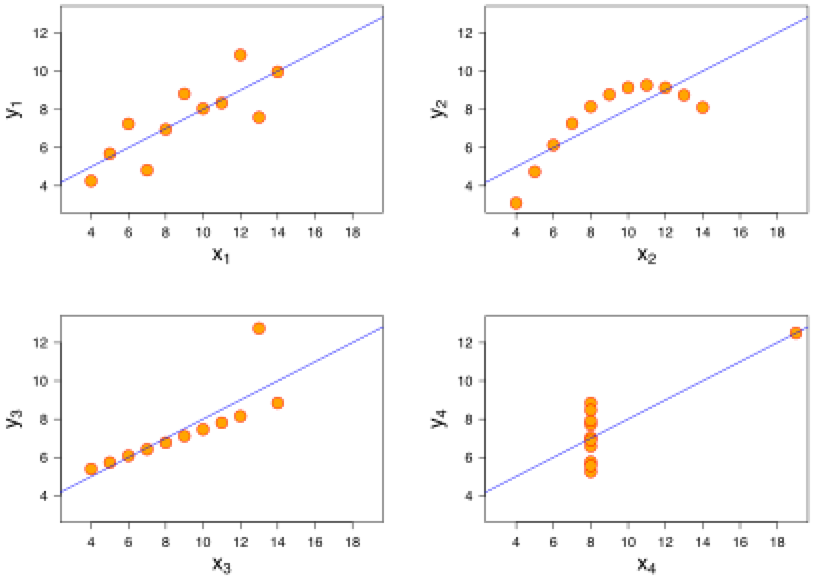
Before Francis Anscombe, it was believed that "numerical calculations are precise, but graphs are rough". But by developing these datasets he proved that plotting of variables of datasets is equally important as descriptive statistics.

Property	**Value**	**Accuracy**
Mean of *x*	9	exact
Sample variance of *x* :	11	exact
Mean of *y*	7.50	to 2 decimal places
Sample variance of *y* :	4.125	±0.003
Correlation between *x* and *y*	0.816	to 3 decimal places
Linear regression line	*y* = 3.00 + 0.500*x*	to 2 and 3 decimal places,

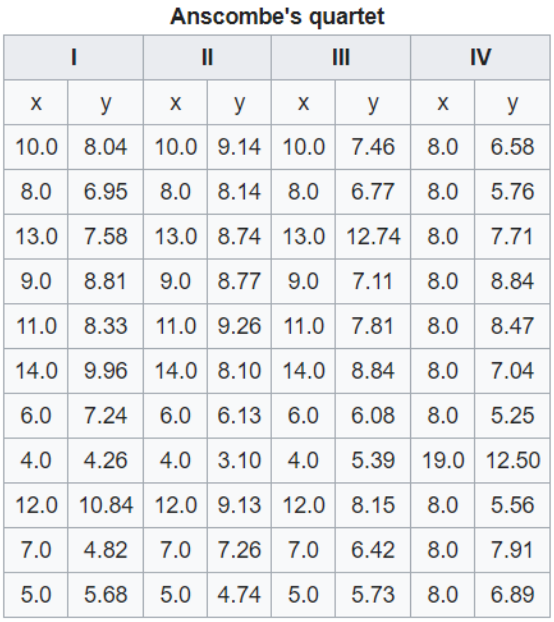
Coefficient of determination of the linear regression :	0.67	to 2 decimal places

For all four datasets:

The following plot shows the details of each set



The following table shows the details of each dataset



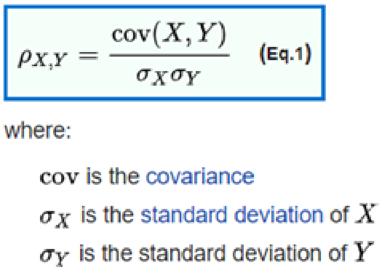
Anscombe's quartet

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

This clearly sets an example that how important it is to plot the datasets and have a feel of the datasets without going into details of statistics.

3. What is Pearson’s R?

The Pearson’s R is the ratio of covariance of the two variables to product of their standard deviations.



$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} \quad \textbf{(Eq.1)}$$

where:

cov is the covariance

σ_X is the standard deviation of X

σ_Y is the standard deviation of Y

The Pearson’s R is also commonly called as Pearson correlation coefficient. Following are the properties of Pearson’s R

- Its values lies between -1 t +1
- -1 shows the negative correlation between two variables X and Y
- 0 shows the no correlation between two variables X and Y
- +1 shows the positive correlation between two variables X and Y

This correlation coefficient when applied to a population is know as population Pearson correlation coeffient.

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

In machine learning algorithms, various features may have different scales and some these algorithms are very sensitive to the scales as the feature with more magnitude may be given more importance than with lower magnitudes. To avoid this issue, feature scaling is used. The main objective of feature scaling to bring all the features to the same scale so the relative importance of the entire feature becomes equal and it becomes more convenient for Machine learning algorithms. Normalization scaling is a max –main scaling techniques and scales all the scaled features in the range of [0,1] while the standardized scaling techniques, scales the data such its population mean is zero and standard deviation is 1.

Standardization is defines as follows:

$$x_{\text{stand}} = \frac{x - \operatorname{mean}(x)}{\text{standard deviation }(x)}$$

This technique ensures that the all transformed vectors to have a unit length.

Min = Max Scaling is defined as follows:

$$x_{\text{norm}} = \frac{x - \min(x)}{\max(x) - \min(x)}$$

This scaling technique transforms the maximum value of the column variable to 1 and corresponding minimum to 0.

Comparison between Max- main Normalization and Standardization:

For standardization scaling, the outliers are handled properly as compared to max- min normalization due to 0 population mean and 1 standard deviation while in max- min normalization, all the data is scaled to small intervals. This results in smaller standard deviation and sometimes fails to treat outliers.

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

The Quantile-Quantile plot also known as Q-Q plot is a graphical representation which measures whether a dataset is originated from the original dataset with same distribution. These distributions can be Uniform, Exponential or Normal distribution. This technique can be applied on sample sizes too.

This is a very common situation in which we receive the training and test data set separately. Deployment of Q- Q plots in such situation helps us in evaluating whether these datasets have identical theoretical distribution or not. Following features of distributions can be identified from the Q- Q plots

1. Shift in location

2. Symmetry Deviations,

3. Presence of outliers can all be spotted from this plot.

4. The presence of tail behavior.

It is used to check subsequent situations: A. If two data sets have come from populations with a common distribution. B. if two datasets have common location and scale. C. if two datasets have similar distributional shapes. D. Whether two datasets have similar tail behavior.

The Q-Q plot is drawn by following steps:

1. Data collection for creating Q- Q plot

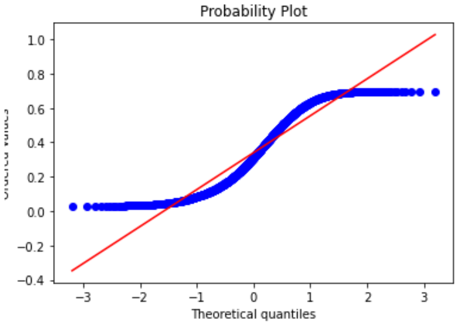
2. Data is sorted in Ascending or Descending order

3. Draw the Normal distribution curve
4. Determine the Z value or cutoff point for wach segment
5. Plot these data for each segments

Advantages of Q-Q plot

- Since Q-Q plot is like probability plot. So, while comparing two datasets the sample size need not to be equal.
- Since we need to normalize the dataset, so we don't need to care about the dimensions of values.

  Below is the sample Q- Q lot for left tailed distribution.


Probability Plot
Theoretical quantiles