Final Report

Statistical Analysis of mechanical properties of low alloy steel

By

Kumari Shubham Chaudhary (22M1840)

LIST OF CONTENTS

1. Introduction

We know that Tensile strength is the maximum amount of tensile stress a material can withstand before it fails or breaks. It is a measure of the material's ability to resist deformation under tension or stretching forces. A material's tensile strength is an important factor in determining its suitability for a particular application.

Yield strength is the maximum stress that can be applied before it begins to change shape permanently. This is an approximation of the elastic limit of the steel. If stress is added to the metal but does not reach the yield point, it will return to its original shape after the stress is removed.
As we know that various types of steel with different composition are used in industries.Here, we will be using this dataset will be used to analyse and predict the mechanical properties of steel as there are no precise theoretical methods to predict the mechanical properties of steel.

Source for the data: This dataset is taken from kaggle.com Data cleaning: We have checked the data that whether there is any null value is there but, in our data, we could not didn’t find any of these. This dataset only contains parameters relevant to our study.

2. Motivation

I have selected the data which relates the effect of various alloying elements on the mechanical properties of the material such as yield strength, Ultimate tensile strength etc. Also, this data set contains the effect of temperature on the mechanical properties for a particular alloy . To compare the temperature effects, 10 data points are taken for each alloy and to compare the alloying effect, 5 compositions namely A,B,C,D,E are taken to analyse the structure property correlation.

Description of the data*:*

- The dataset contains no missing values.

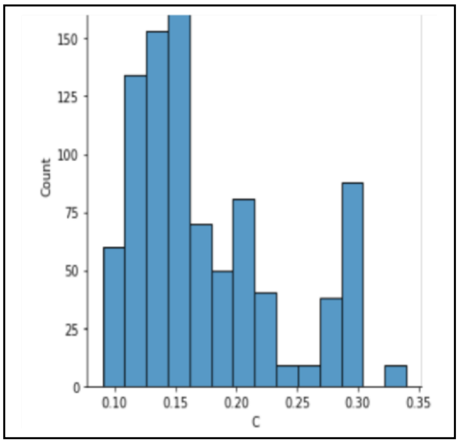
- As mentioned, the columns with percentage composition of alloying metals and impurities.
- Mechanical properties such as tensile strength and proof stress are given in columns with varying percentage composition of alloying metals and temperature.

3. EDA analysis

In EDA analysis, we have tried to understand the data by plotting various kinds of graphs in order to observe the trends the data follow and to observe if there are any outliers in the data to discard for the analysis.

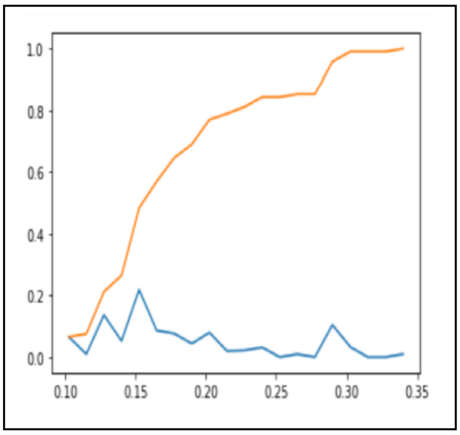
1. Tensile strength vs %C

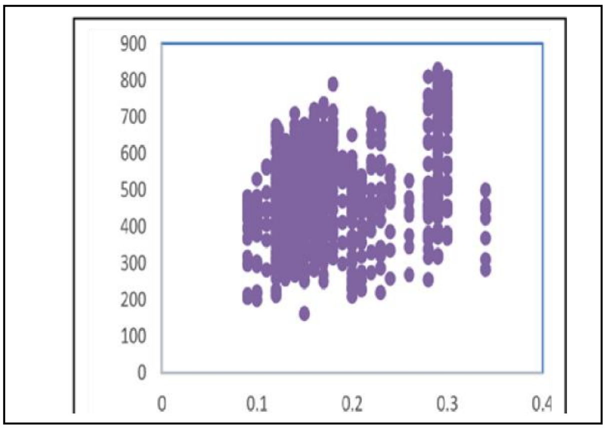
Here are the various graphical representations for each alloy content:



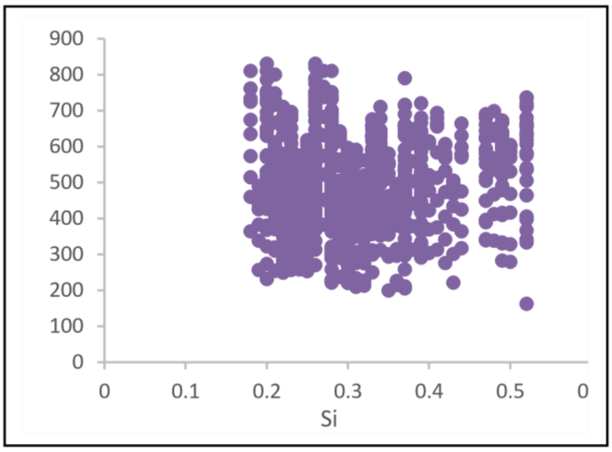
Observations:

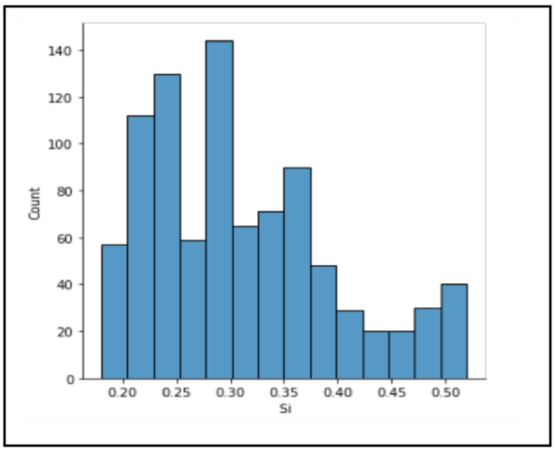
From scatter plot, it can be seen that the mostly values are scattered from 0.1 to 0.3%.
From Histogram, it can be seen that peek around 0.15. but some other %C is also there whose frequency is very less. i.e., an outlier. Most of the data lies between 0-0.3.
We have also plotted line plot to see cdf.

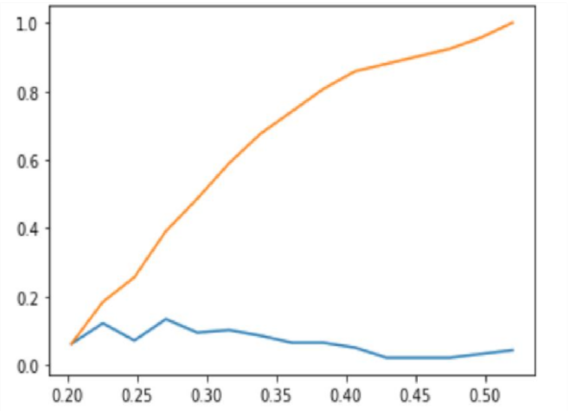




2. Tensile strength vs % si



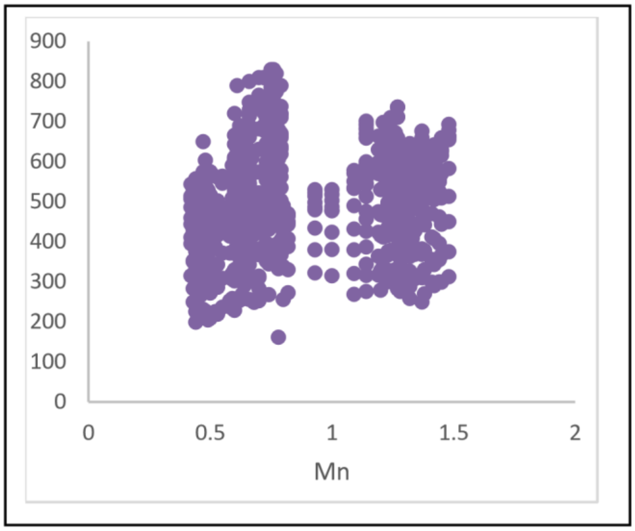


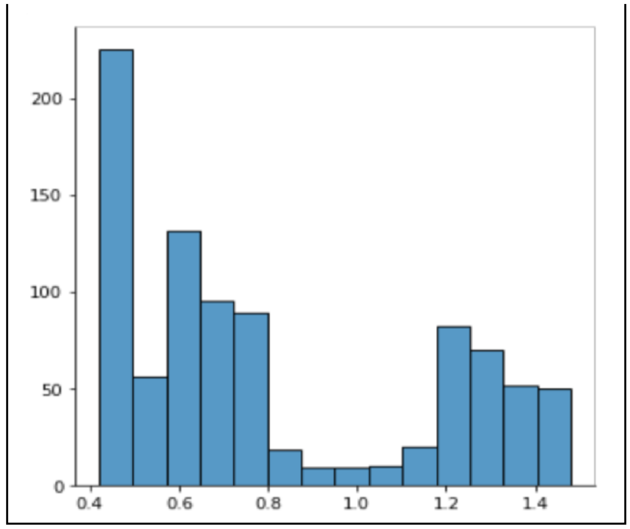


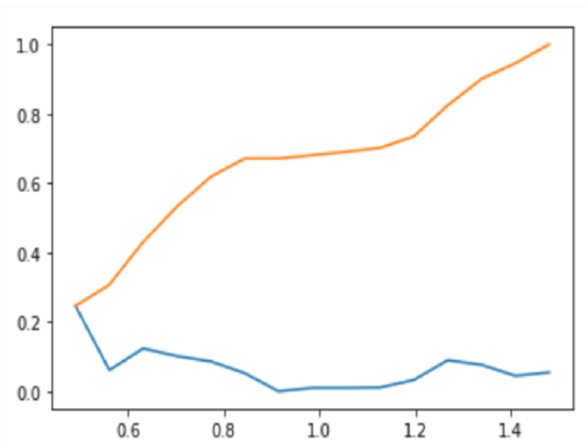
Observations:

- From scatter plot, it can be seen that the mostly values are scattered from 0.2 to 0.5%.
- From Histogram, it can be seen that peek around 0.29-0.30. Here we could not observe any outliers.
- We have also plotted line plot to see cdf.

3. **Tensile strength vs % Mn**



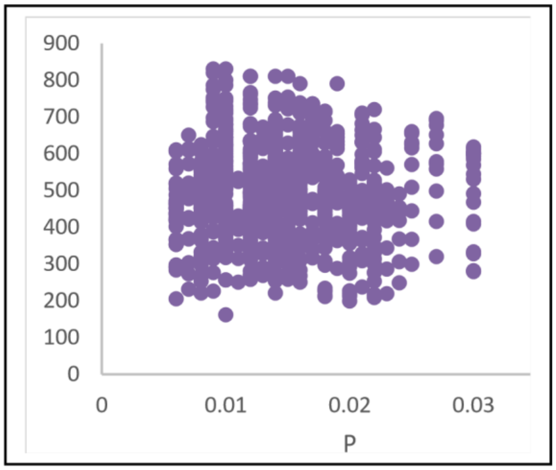


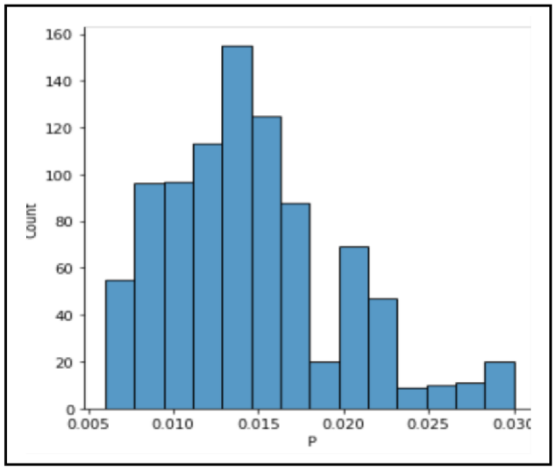


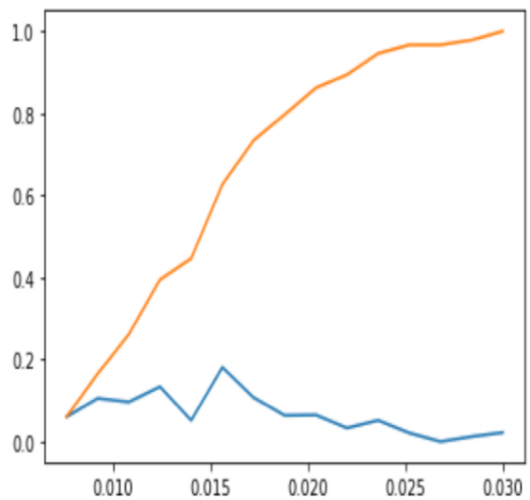
Observations:

- From scatter plot, it can be seen that the mostly values are scattered from 0.4 to 1.5%.
- From Histogram, it can be seen that peek in the range 0.4-0.5. but some other %Mn is also there whose frequency is very less. i.e., an outlier.
- We have also plotted line plot to see cdf.

4. **<u>Tensile strength vs % P</u>**



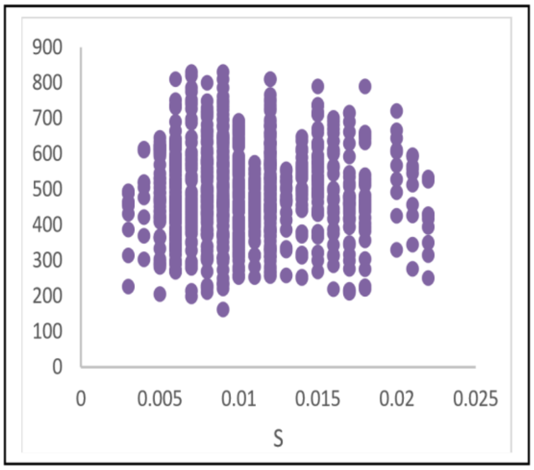


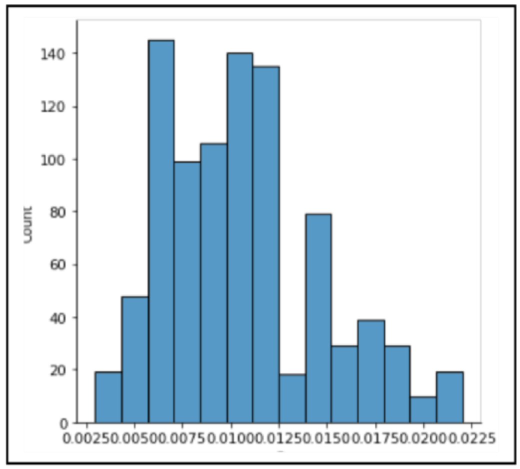


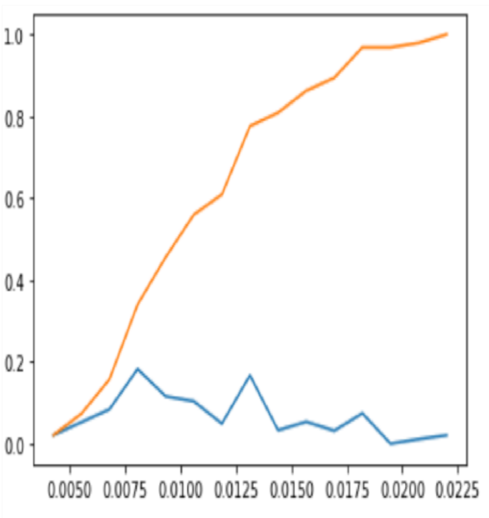
Observations:

- From scatter plot, it can be seen that the mostly values are scattered from 0.005 to 0.03%.
- From Histogram, it can be seen that peek around 0.015. but some other %P is also there whose frequency is very less. i.e., an outlier. Most of the data lies between 0.01-0.015.
- We have also plotted line plot to see cdf.

5. Tensile strength vs %S





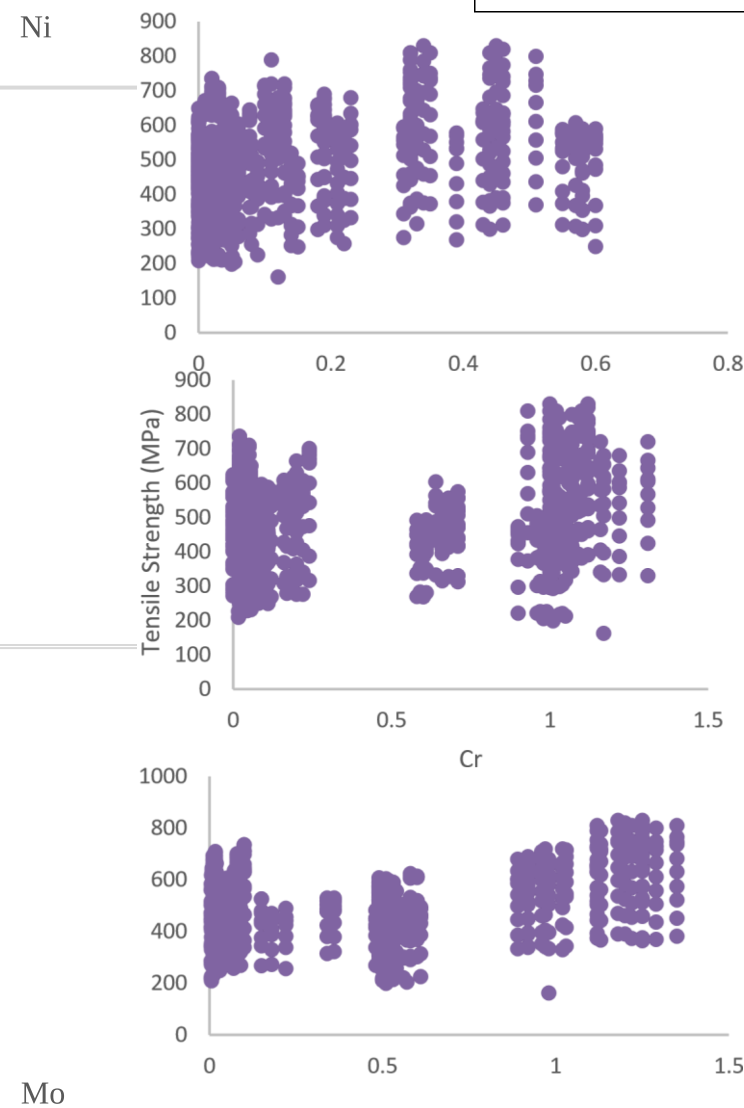


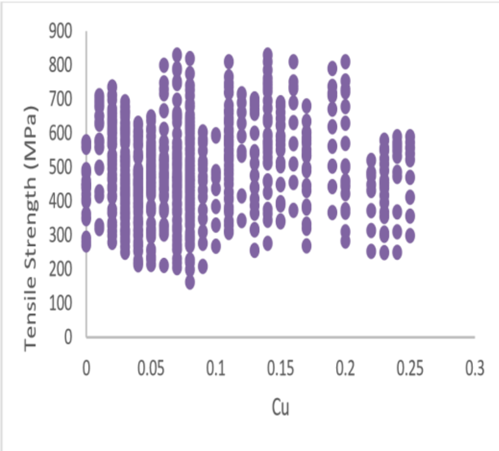
Observations:

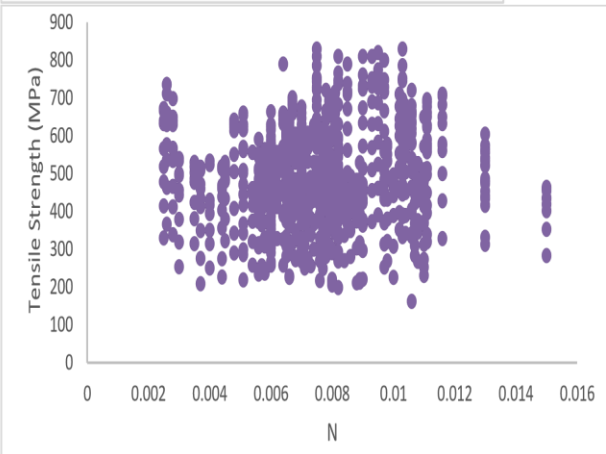
- From scatter plot, it can be seen that the mostly values are scattered from 0 to 0.02%.
- From Histogram, it can be seen that peek around 0.005. but some other %S is also there whose frequency is very less. i.e., an outlier.
- We have also plotted line plot to see cdf.

6. **Variation with other alloy contents**

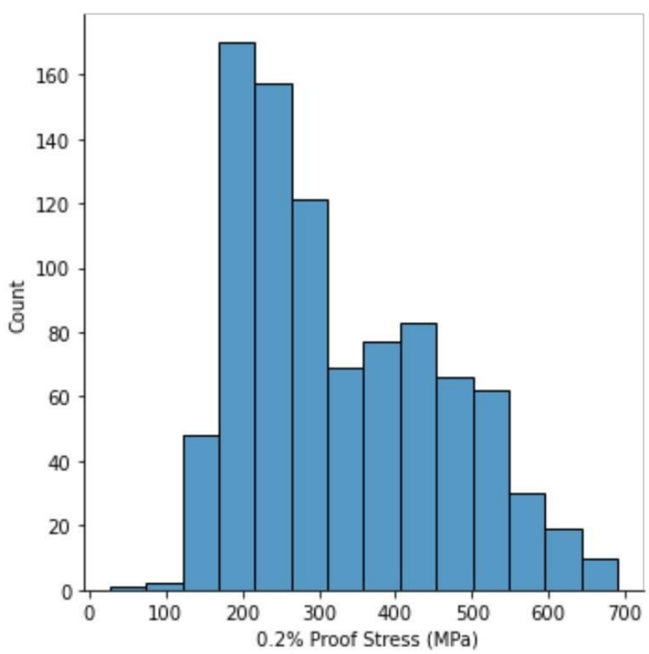
For these alloys also, we see that most values are following the general trends with tensile strength , however these are some outliers existing , which can not be explained and can be discarded in the future analysis.





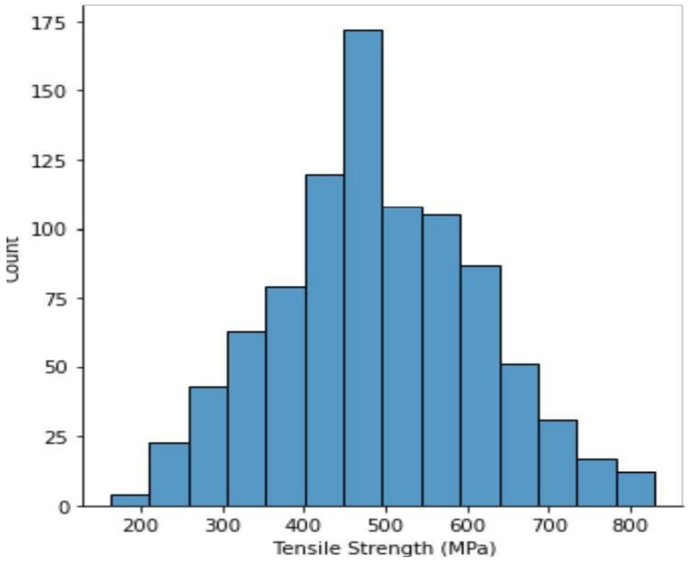


<u>Variation of 0.2% proof stress:</u>



Observations:
From Histogram, it can be seen that peek around 200 and we see that max values are in range of 150-600 MPA , but some other values are there whose frequency is very less i.e.,values less than 100 and more than 650 are very less .

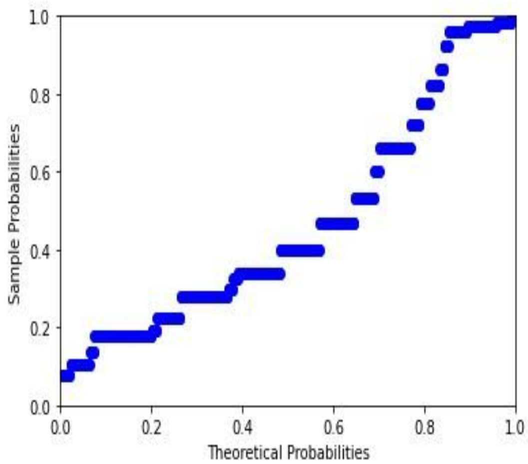
Variation of Tensile strength:

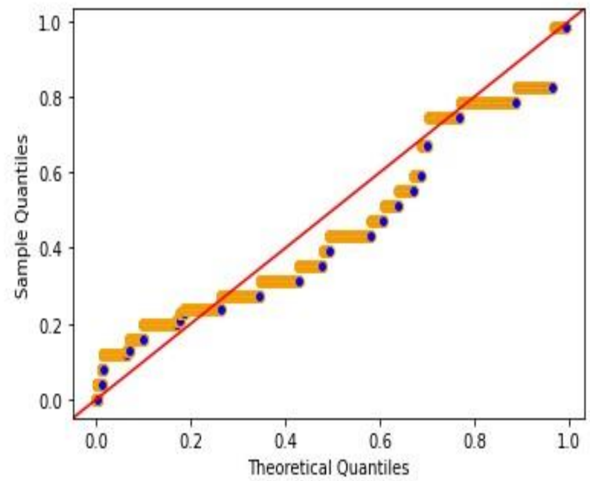


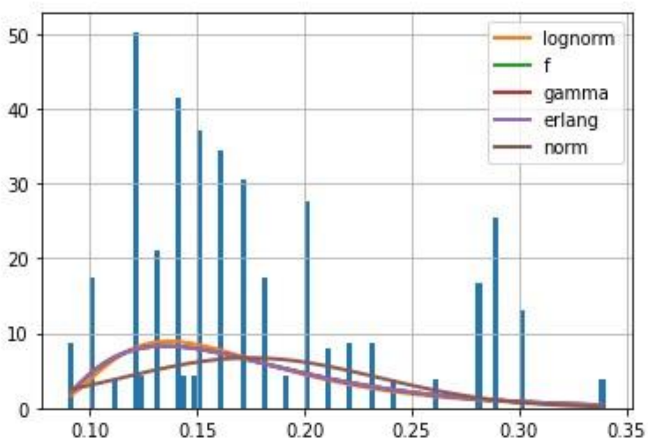
Observations:
From Histogram, it can be seen that peek around 450-500. and we see that max values are in range of 200-800 MPA , but some other values are there whose frequency is very less i.e.,values less than 200 and more than 800 are very less .

4. Distribution Analysis

1. **<u>% Carbon content</u>**





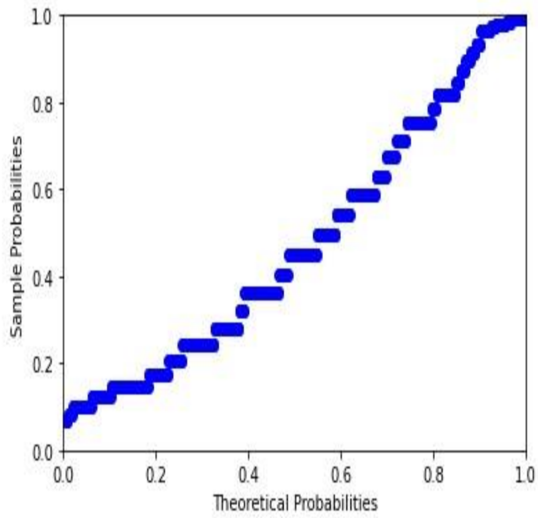


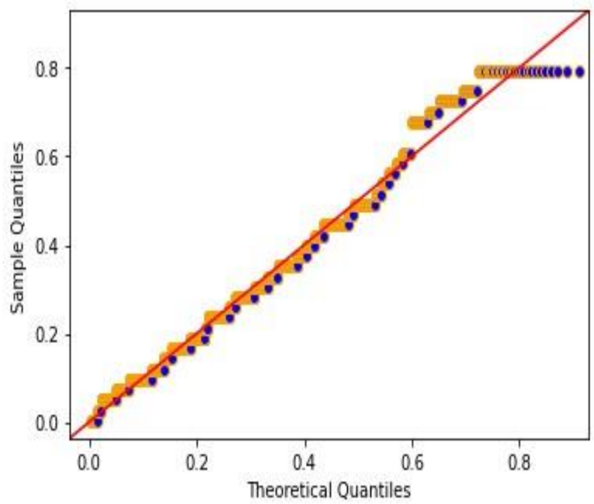
	sumsquare_error	aic	bic	kl_div
lognorm	8707.542276	-187.223345	2081.970797	inf
f	8747.610958	-193.470295	2092.990533	inf
gamma	8754.969706	-196.289342	2086.941010	inf
erlang	8754.970846	-196.290956	2086.941129	inf
norm	9121.904470	-184.545181	2117.689333	inf

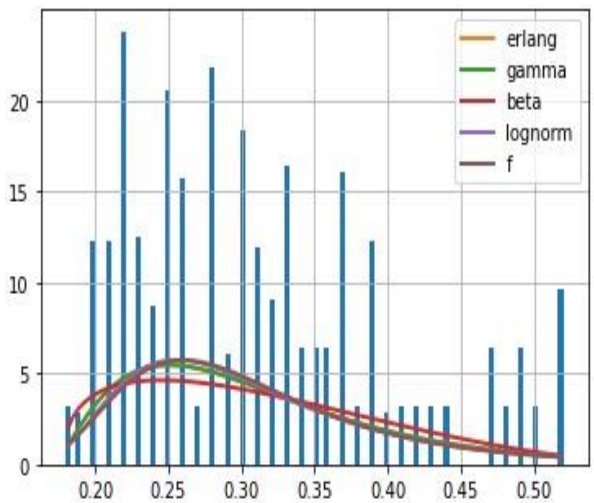
<u>Observations :</u>

From the graph we can see that the best fit is lognormal distribution for this analysis.

2. **<u>% Si content</u>**





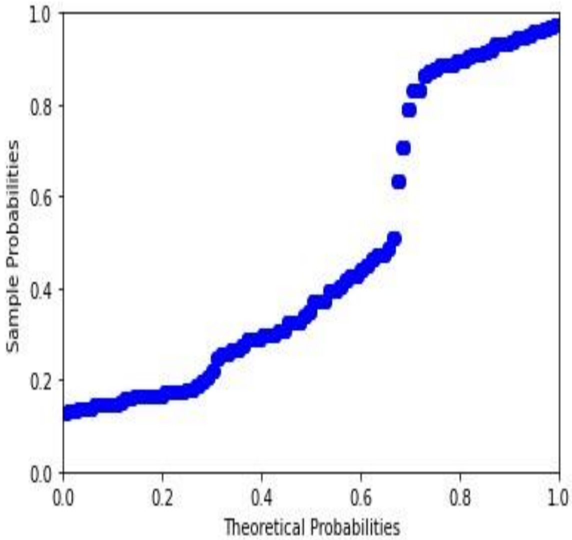


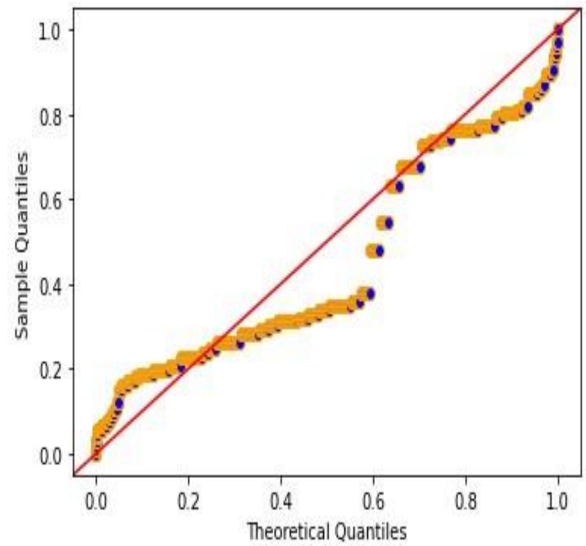
	sumsquare_error	aic	bic	kl_div
erlang	2876.659476	-153.282239	1068.553524	inf
gamma	2876.660392	-153.280546	1068.553815	inf
beta	2877.237569	-172.593422	1075.556307	inf
lognorm	2894.056349	-145.885420	1074.070408	inf
f	2900.588763	-142.250878	1082.952327	inf

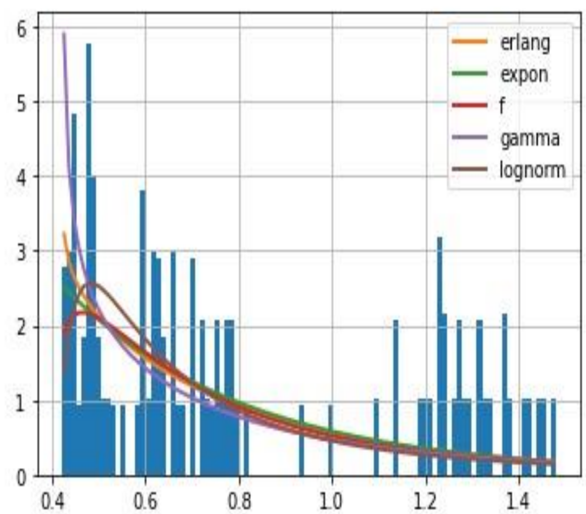
Observations :

From the graph we can see that the best fit is Erlang distribution for this analysis.

3. % Mn content





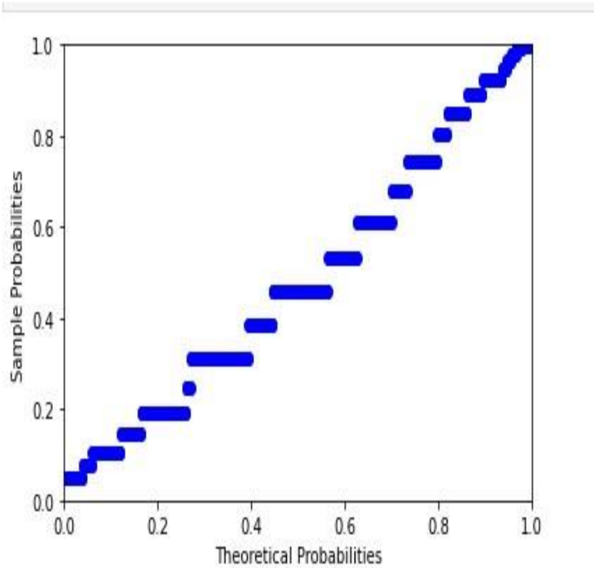


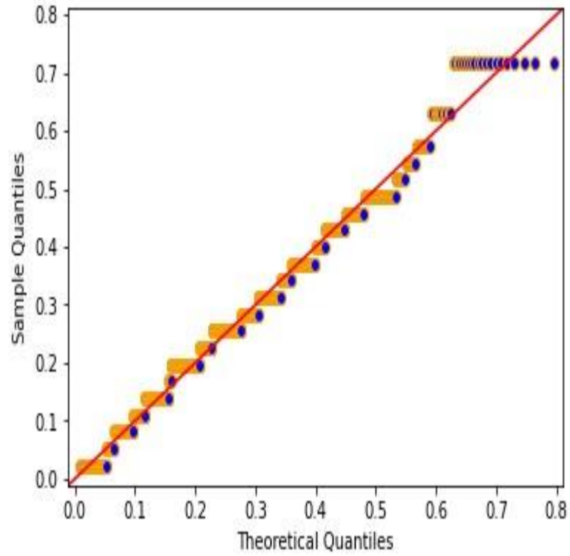
	sumsquare_error	aic	bic	kl_div
erlang	107.286873	93.902736	-1940.770479	inf
expon	109.407231	86.937756	-1929.682227	inf
f	111.819882	105.238187	-1896.086037	inf
gamma	115.497043	114.397471	-1873.299636	inf
lognorm	116.320647	116.617935	-1866.797973	inf

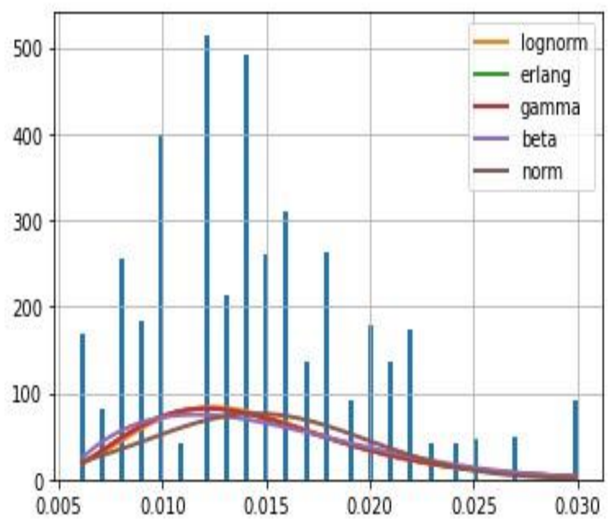
Observations :

From the graph we can see that the best fit is Erlang distribution for this analysis.

4. % P content





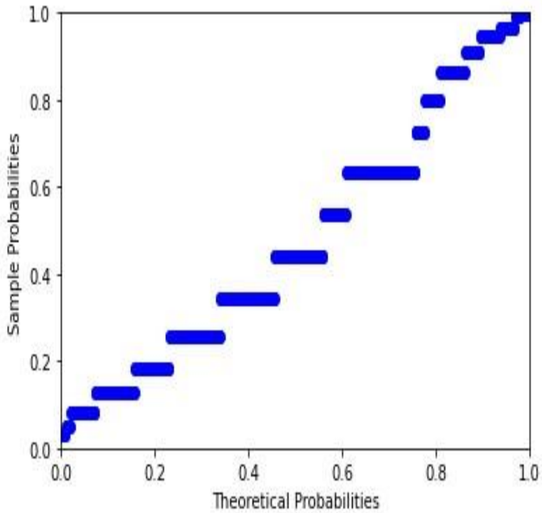


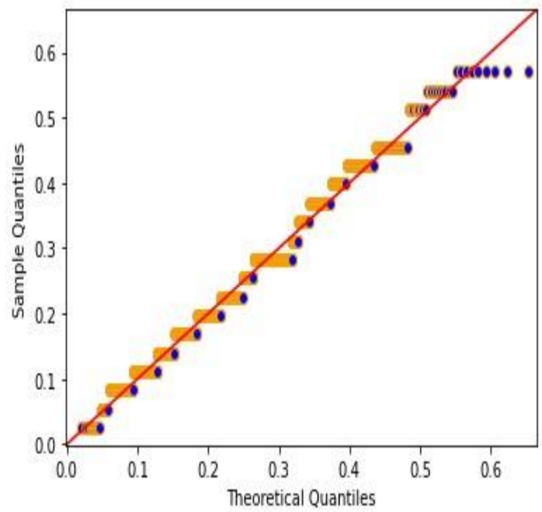
	sumsquare_error	aic	bic	kl_div
lognorm	968618.550225	-660.917801	6393.159189	inf
erlang	968994.106106	-664.782256	6393.513887	inf
gamma	968994.176346	-664.781941	6393.513953	inf
beta	970275.847677	-674.077312	6401.542332	inf
norm	972652.047016	-644.234443	6390.142574	inf

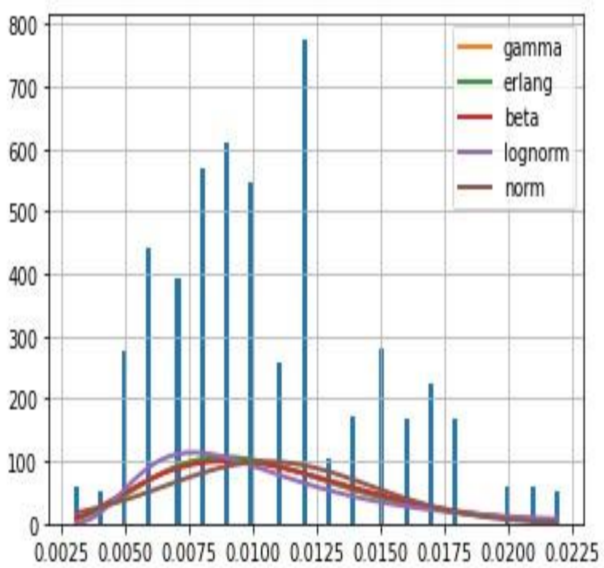
Observations :

From the graph we can see that the best fit is Lognormal distribution for this analysis.

5. % S content





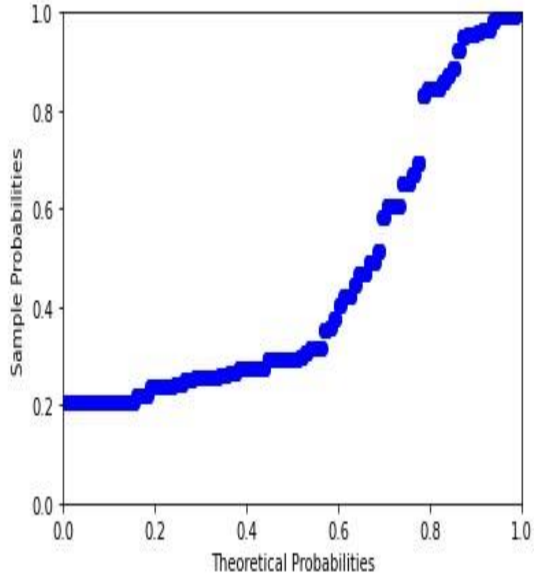


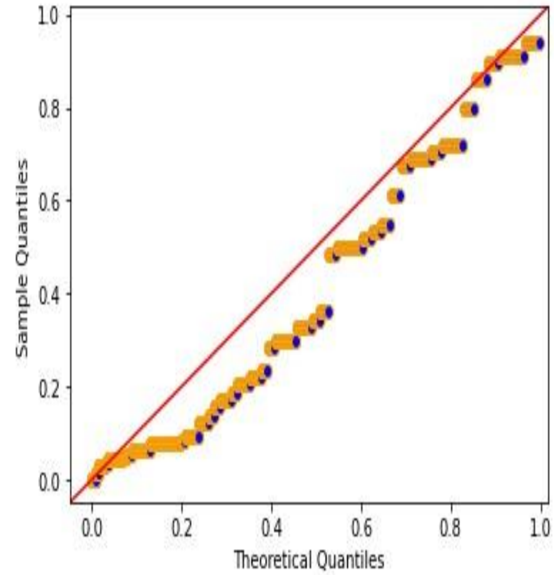
	sumsquare_error	aic	bic	kl_div
gamma	1.944633e+06	-715.655922	7030.875466	inf
erlang	1.944633e+06	-715.655586	7030.875470	inf
beta	1.945883e+06	-719.392046	7038.282282	inf
lognorm	1.951128e+06	-703.166009	7033.926278	inf
norm	1.963723e+06	-707.160159	7032.994928	inf

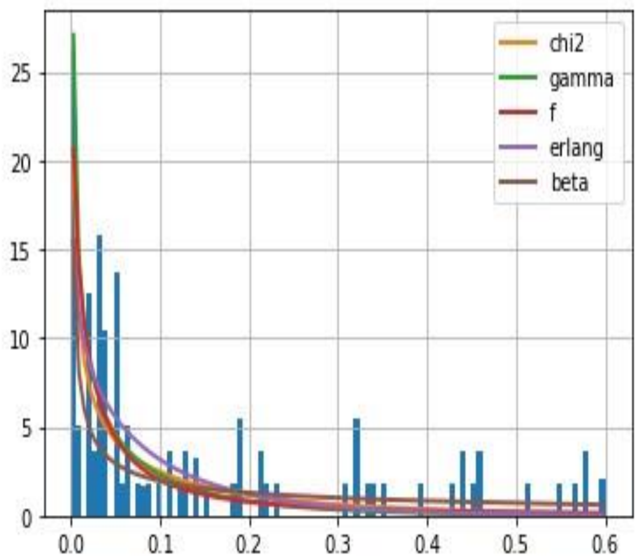
Observations :

From the graph we can see that the best fit is Gamma distribution for this analysis.

5. % Ni content





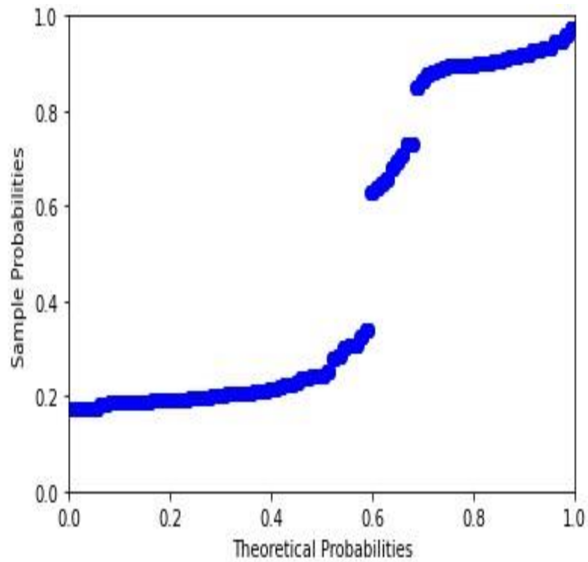


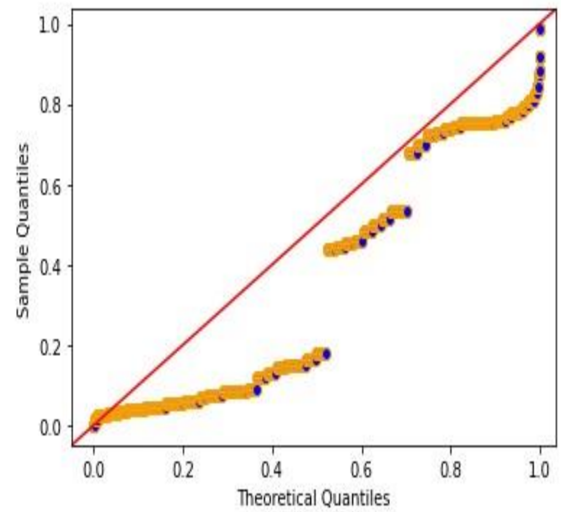
	sumsquare_error	aic	bic	kl_div
chi2	625.594346	71.206781	-327.446278	inf
gamma	652.375525	191.504713	-289.091123	inf
f	663.501892	118.139628	-266.798313	inf
erlang	706.646990	138.558116	-215.972668	inf
beta	709.809094	-25.969039	-205.068436	inf

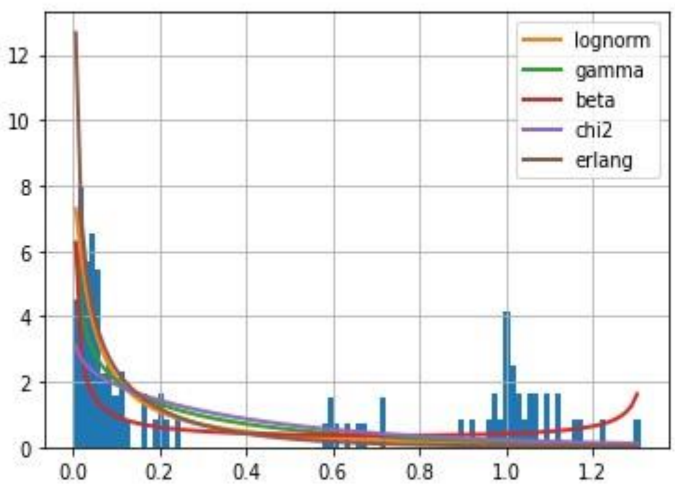
Observations :

From the graph we can see that the best fit is Chi square distribution for this analysis.

7. % Cr content





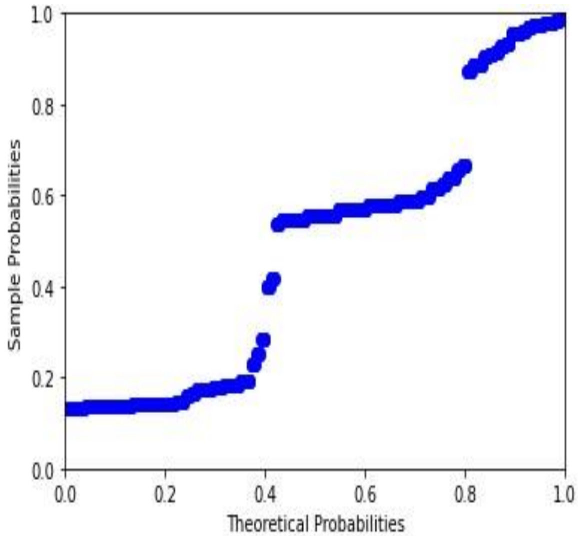


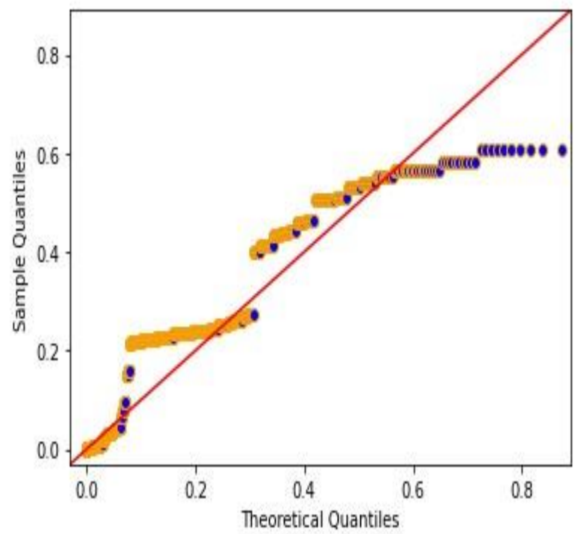
	sumsquare_error	aic	bic	kl_div
lognorm	87.997160	236.477236	-2122.125095	inf
gamma	112.009169	186.099277	-1901.357373	inf
beta	139.338303	138.605129	-1694.771921	inf
chi2	142.248357	149.964644	-1682.678062	inf
erlang	144.902217	381.510212	-1665.764623	inf

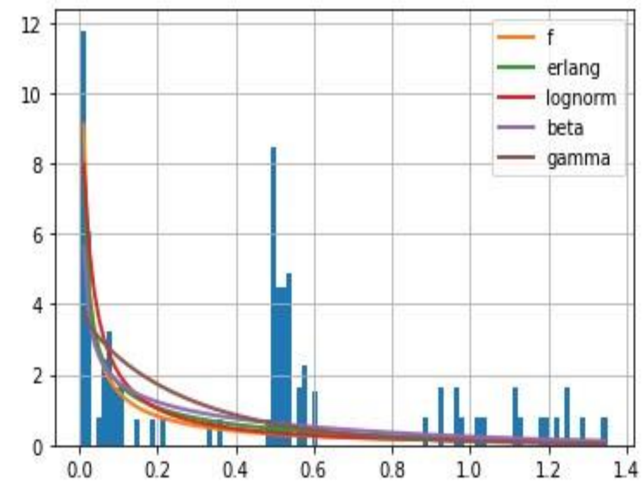
Observations :

From the graph we can see that the best fit is Lognormal distribution for this analysis.

7. %Mo content





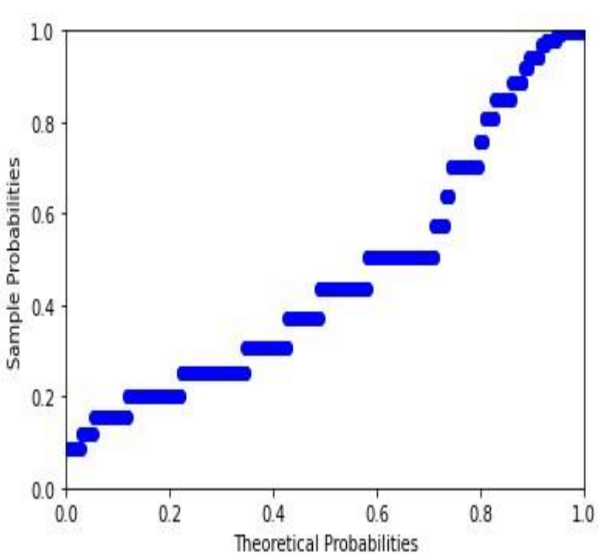


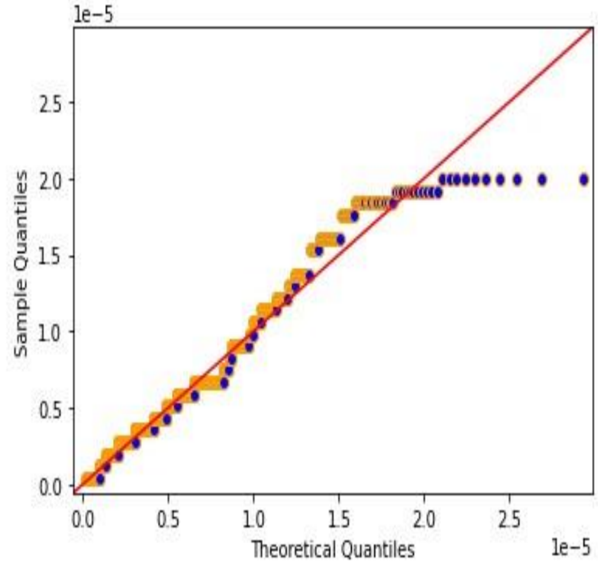
	sumsquare_error	aic	bic	kl_div
f	180.345292	264.716976	-1458.730754	inf
erlang	189.018861	199.247015	-1422.568921	inf
lognorm	202.140110	243.062188	-1361.159367	inf
beta	212.738986	158.855451	-1307.579436	inf
gamma	267.378107	226.441188	-1105.231381	inf

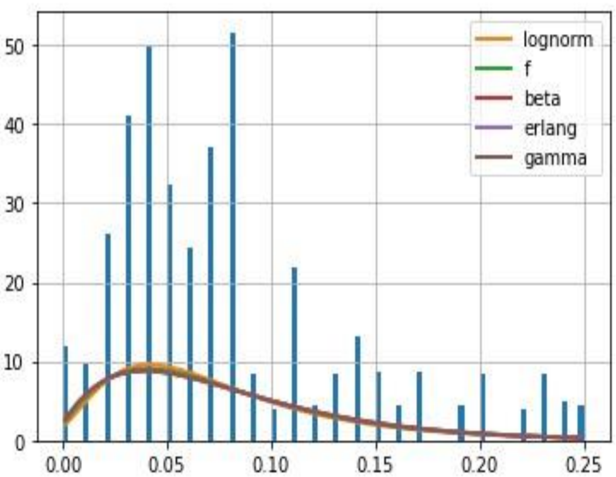
Observations :

From the graph we can see that the best fit is F distribution for this analysis.

8. % Cu content





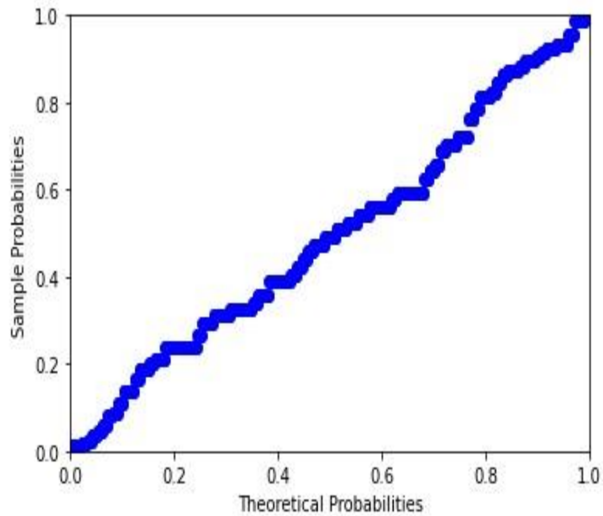


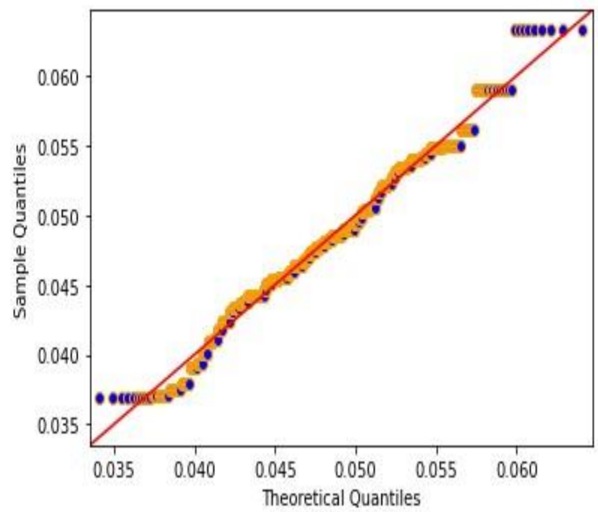
	sumsquare_error	aic	bic	kl_div
lognorm	9437.761240	-174.339099	2155.655030	inf
f	9467.862106	-178.146914	2165.387618	inf
beta	9488.862725	-181.949433	2167.414928	inf
erlang	9488.901267	-183.969533	2160.599720	inf
gamma	9488.901292	-183.967878	2160.599723	inf

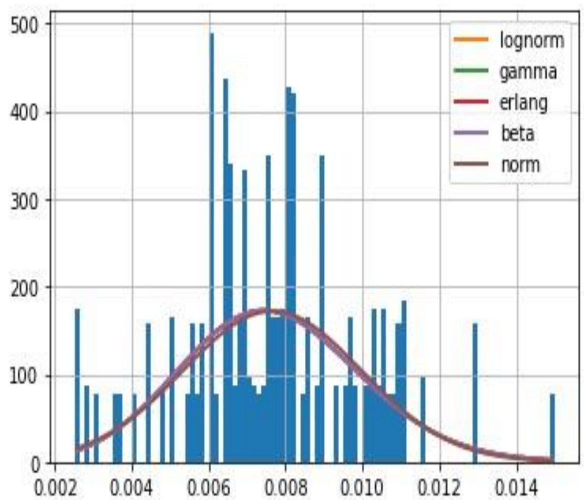
<u>**Observations :**</u>

From the graph we can see that the best fit is Lognormal distribution for this analysis.

<u>**10. % N content**</u>





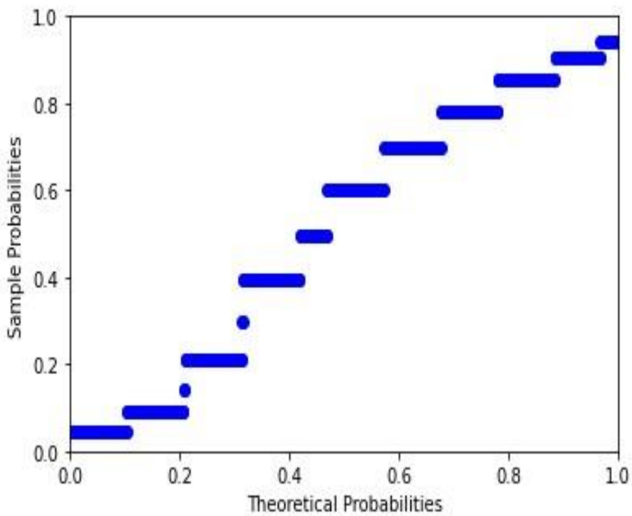


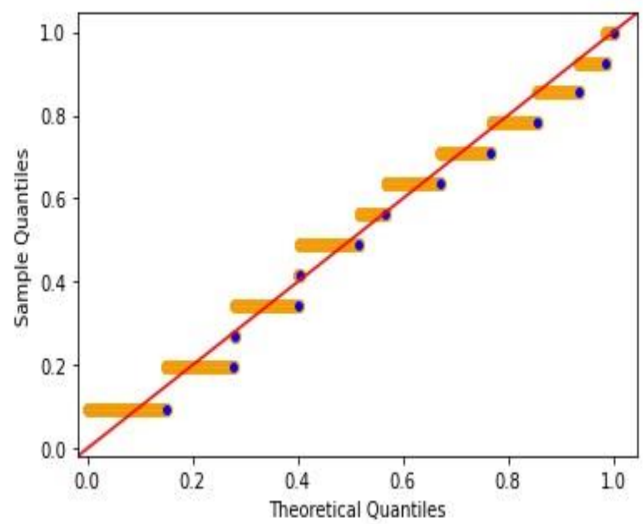
	sumsquare_error	aic	bic	kl_div
lognorm	875767.707415	-769.046162	6300.954639	inf
gamma	876056.816845	-768.859902	6301.256650	inf
erlang	876056.831544	-768.859976	6301.256666	inf
beta	876113.783515	-767.296259	6308.135072	inf
norm	880069.814512	-761.120973	6298.619542	inf

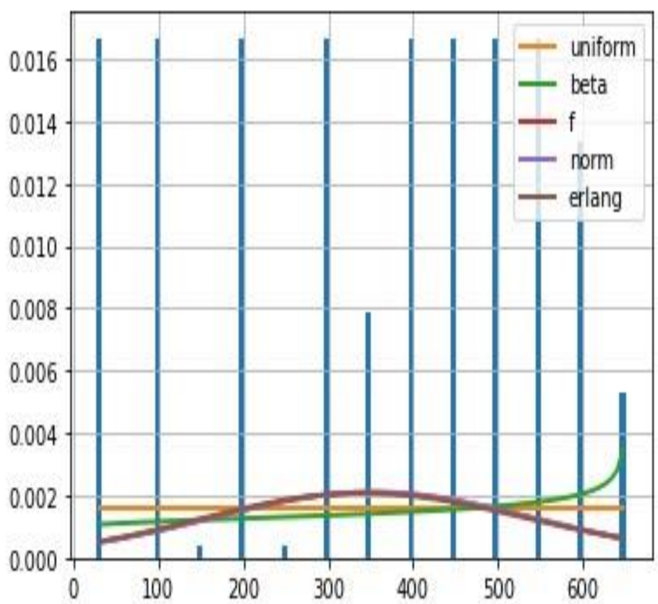
Observations :

From the graph we can see that the best fit is Lognormal distribution for this analysis.

11. Temperature distribution





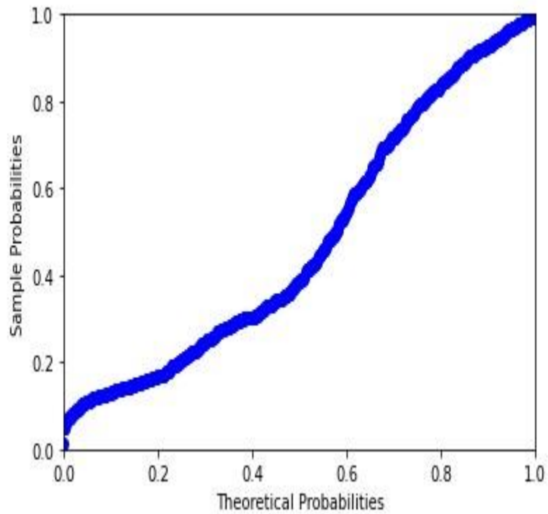


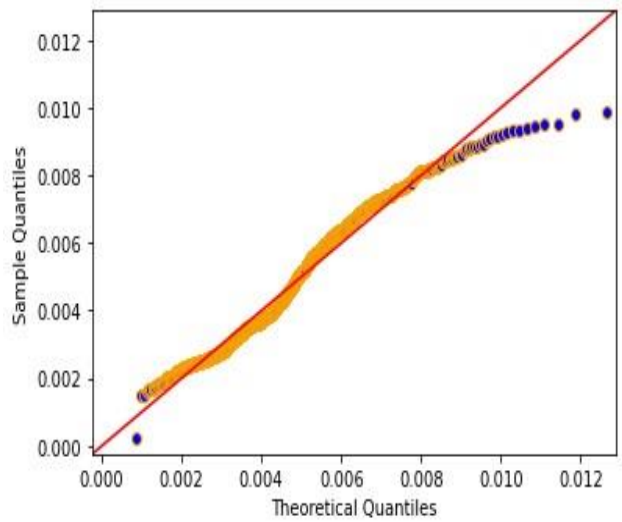
	sumsquare_error	aic	bic	kl_div
uniform	0.002232	1290.909304	-11811.525923	inf
beta	0.002240	1313.450958	-11794.769253	inf
f	0.002271	1331.333387	-11782.258873	inf
norm	0.002271	1327.214845	-11795.787271	inf
erlang	0.002273	1329.437127	-11788.254696	inf

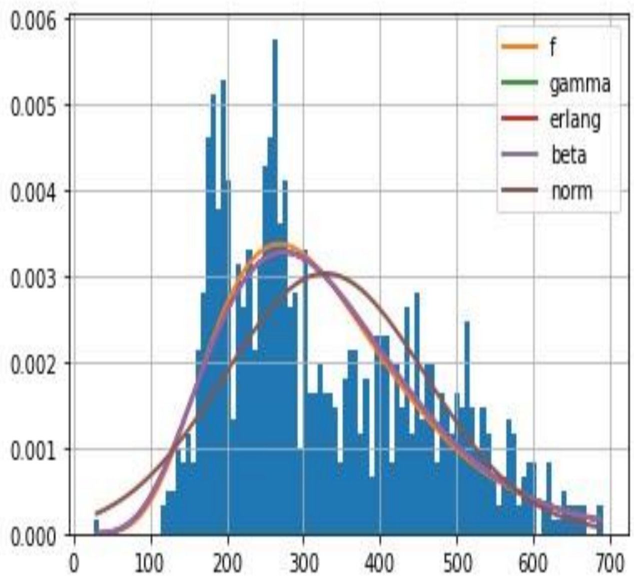
Observations :

From the graph we can see that the best fit is Uniform distribution for this analysis.

12. 0.2% Proof Stress distribution





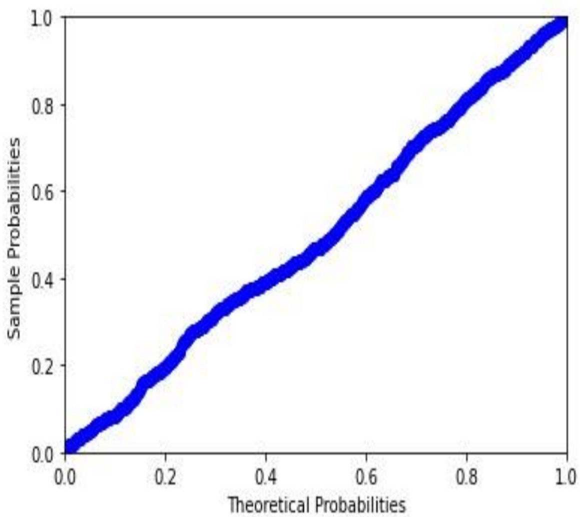


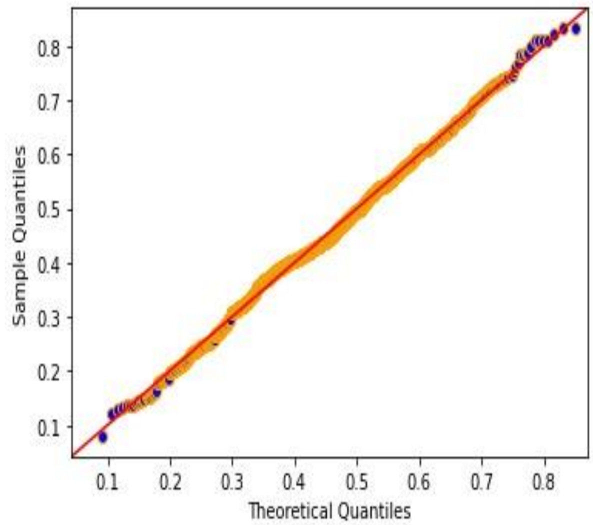
	sumsquare_error	aic	bic	kl_div
f	0.000079	1432.038210	-14851.886058	inf
gamma	0.000081	1422.841691	-14837.141097	inf
erlang	0.000081	1422.841624	-14837.141035	inf
beta	0.000081	1424.712444	-14828.483346	inf
norm	0.000113	1380.559426	-14540.799981	inf

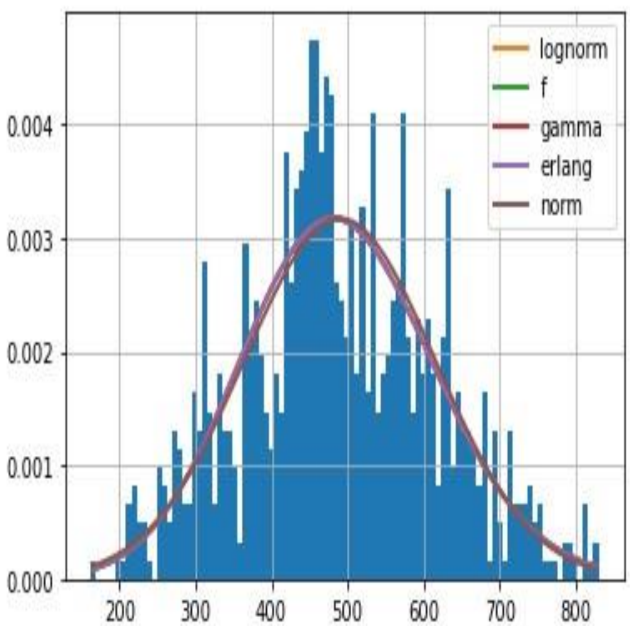
Observations :

From the graph we can see that the best fit is F distribution for this analysis.

17. Tensile strength distribution







	sumsquare_error	aic	bic	kl_div
lognorm	0.000047	1392.045297	-15331.108334	inf
f	0.000047	1394.069264	-15324.284876	inf
gamma	0.000047	1391.959922	-15331.084720	inf
erlang	0.000047	1391.953705	-15331.084361	inf
norm	0.000048	1389.711792	-15329.383187	inf

Observations :

From the graph we can see that the best fit is F distribution for this analysis.

Summary of Distribution Analysis

Sr. No	Columns	Best Fitting Distribution	Sum Square Error
1.	%C content	Lognormal	8707.542276
2.	%Si content	Erlang	2876.659476
3.	%Mn content	Erlang	107.286873
4.	%P content	Lognormal	968618.550225
5.	%S content	Gamma	1.944633e^6
6.	%Ni content	Chi square	625.594346
7.	%Cr content	Lognormal	87.997160
8.	%Mo content	F	180.345292
9.	%Cu content	Lognormal	9437.761240
10.	% V content	Erlang	18374.997726
11.	%Al content	Lognormal	180595.506666
12.	%N content	Lognormal	875767.707415
13.	%Ceq content	Gamma	18155.078274
14.	Nb+Ta content	Beta	1.278213e^9
15.	Temperature	Uniform	0.0022232
16.	0.2% Proof Stress	F	0.000079
17.	Tensile Strength	Lognormal	0.000047

6. Regression Analysis

STEPS INVOLVED:

1. First I select a feature as target (dependent variable) which is ‘0.2% Proof Stress (MPa)’.

2. Then take single-single feature as variable (independent variable).

3. Now split our dataset (variable, target) in to training and testing dataset.

4. Then fit ‘linear regression model’ on training dataset.

5. Then predict ‘0.2% Proof Stress (MPa)’ for testing data.

6. Then find out “R_SQURE” and “ADJUSTED R_SQURE” value on testing dataset which is more reliable.

7. Repeat same steps for ‘Tensile Strength’.

R_SQURE : it is a statistical measure of fit that indicates how much a dependent variable is explained by the independent variable(s) in a regression model. It’s maximum value can be +1 that tell highly dependence on that variable. ADJUSTED R_SQURE : The adjusted R-squared increases when the new term improves the model more than would be expected by chance. It decreases when a predictor improves the model by less than expected.

Linear Regression is done to see any strong dependency is observed between features and “**0.2% Proof Stress**”.

Variables	R square value	adjusted R square value
%C	0.03151415000110391	0.03161919716777428
%Si	0.07960078124100323	0.07986611717847325
%Mn	0.15301219092363727	0.1535222315600494
%P	-0.009496615503267858	-0.009528270888278678
%S	-0.015691541834740086	-0.01574384697418929
%Ni	0.24085827932807424	0.24166114025916785
%Cr	0.045212454024197934	0.045363162204278606
%Mo	0.12576784960659082	0.1261870757719461
%Cu	0.011607450908954098	0.011646142411983917
%N	-0.016259061779242634	-0.01631325865184019

Temperature	0.1212747282957316	0.12167897739005074

Conclusion: It is clearly seen that the R square values are too less for %P, %S, %Cu, %N. Among all of features 0.2% Proof stress depends most on %V content followed by %Ni and %Mn content.

Multiple Linear regression is to see if there is any combined dependency of features with "**0.2% Proof Stress**".

Variables	R square value	adjusted R square value
F1,F2	0.22324821657439187	0.22474151568191292
F1,F2,F3	0.2962364880796591	0.299218734603951
F1,F2,F3,F4	0.3411342821457769	0.34572868325211736
F1,F2,F3,F4,F5	0.35933848101931565	0.36540838779329055
F1,F2,F3,F4,F5,F6	0.42371182660316065	0.4323296942628859
F1,F2,F3,F4,F5,F6,F7	0.5099185346006612	0.5220594520911531
F1,F2,F3,F4,F5,F6,F7,F8	0.5428545560080373	0.5576765234075741
F1,F2,F3,F4,F5,F6,F7,F8,F9	0.5533859916845232	0.5704424092364434
F1,F2,F3,F4,F5,F6,F7,F8,F9,F10	0.6046331623768232	0.6254109342798069

F1,F2,F3,F4,F5,F6,F7,F8,F9,F10,F11	0.5998180862446374	0.6225698067573651
F1,F2,F3,F4,F5,F6,F7,F8,F9,F10,F11,F 12	0.6001265739637169	0.625045324439719
F1,F2,F3,F4,F5,F6,F7,F8,F9,F10,F11,F 12,F13	0.614881316115025	0.6426363755229949
F1,F2,F3,F4,F5,F6,F7,F8,F9,F10,F11,F 12,F13,F13,F14	0.6184432523187802	0.6486112158465256
F1,F2,F3,F4,F5,F6,F7,F8,F9,F10,F11,F 12,F13,F14,F15	0.8502797017564021	0.8948747910093603

Conclusion: It is clearly seen that the R square values are increasing with the combined effect of different features. And when we take input as all features then R square value is 0.85(highest) so it is good way to predict 0.2% Proof stress.

Linear Regression is done to see any strong dependency is observed between features and "**Tensile Strength**".

Variables	R square value	adjusted R square value

%C	0.045053598435921605	0.045203777097374664
%Si	0.011456825255870773	0.011495014673390314
%Mn	0.0605054449209772	0.06070712973738046
%P	-0.006531072700674434	-0.006552842943009951
%S	-0.0035681889834746627	-0.0035800829467529383
%Ni	0.09319549053379161	0.0935061421689043
%Cr	0.020448656594954476	0.020516818783604562
%Mo	0.06366629237556887	0.06387851335015404
%Cu	0.003936909491447094	0.003950032523085234
%N	-3.694154467948074e-05	-3.7064683161780465e-05
Temperature	0.3086504212364508	0.30967925597390566

Conclusion: It is clearly seen that the R square values are too less for %P, %S, %Cu, %N. Among all of features Tensile strength depends most on Temperature followed by %V, %Ni and %Mn content.

Multiple Linear regression is to see if there is any combined dependency of features with "**Tensile Strength**".

Variables	R square value	adjusted R square value

F1,F2	0.10490639605427454	0.1056081110780489
F1,F2,F3	0.1303837680537625	0.1316963563227601
F1,F2,F3,F4	0.14866384612057149	0.15066605280233003
F1,F2,F3,F4,F5	0.1533770997967281	0.15596792918518632
F1,F2,F3,F4,F5,F6	0.1663493982225972	0.16973277581356527
F1,F2,F3,F4,F5,F6,F7	0.20456928245788308	0.2094399796592612
F1,F2,F3,F4,F5,F6,F7,F8	0.21920710412564504	0.22519228103009947
F1,F2,F3,F4,F5,F6,F7,F8,F9	0.23466293395072424	0.24189569561358903
F1,F2,F3,F4,F5,F6,F7,F8,F9,F10	0.2530583038673093	0.261754465512234
F1,F2,F3,F4,F5,F6,F7,F8,F9,F10,F11	0.24683567559230646	0.25619840811477324
F1,F2,F3,F4,F5,F6,F7,F8,F9,F10,F11,F1 2	0.24707800235573107	0.257337296571194
F1,F2,F3,F4,F5,F6,F7,F8,F9,F10,F11,F1 2,F13	0.24993261341880157	0.26121429388562256
F1,F2,F3,F4,F5,F6,F7,F8,F9,F10,F11,F1 2,F13,F14	0.27198953135763004	0.2852573133750754
F1,F2,F3,F4,F5,F6,F7,F8,F9,F10,F11,F1 2,F13,F14,F15	0.6691965380824358	0.704294258611235

Conclusion: It is clearly seen that the R square values are increasing with the combined effect of different features. And when we take input as all features then R square value is 0.669(highest) so it is good way to predict Tensile strength.

7. <u>Hypothesis Testing</u>

1) Hypothesis testing for data under 0.2% proof stress Column

Step 1: Create the hypothesis

Null hypothesis: Mean= 340 Alternate hypothesis: Mean ≠ 340

Step 2 Statistical test used - Z test

To get critical Z value Consider the test as two tailed test with level of significance i.e. alpha as 0.05 then critical value of Z is +1.96 or -1.96.

Step 3 Decision making

As calculated Z value i.e. -2.196857118578904 is less than critical Z value i.e. -1.96 so we would reject the Null hypothesis.

By doing so we can conclude that the sample of 50 data we have taken from the population of 0.2% Proof stress column has significant difference from the data of population.

2) Hypothesis testing of data under Tensile Strength Column

Step 1 Create the hypothesis

Null Hypothesis: Mean = 495 Alternate Hypothesis: Mean ≠ 495

Step 2 Statistical Test used- Z test

To get critical Z value, Consider the test as two tailed test with level of significance i.e. alpha as 0.05 then critical value of Z is +1.96 or -1.96.

Step 3 Decision making

As calculated Z value i.e. -0.5268257408811524 is greater than critical Z value -1.96. So we would accept the null hypothesis. By doing so we made the Type II error in hypothesis testing.

8. **Conclusion:**

- Pre-processing of data is learnt, there is no missing or no abnormal values found. So, the data is used still for further analysis.
- Suspecting a dependency and its confirmation or rejection by applying multiple statistical tools is learnt.
- Qualitative Distributional analysis is one important learning from this work and its confirmation by least value of sum square error.
- P-P plots and Q-Q plots along with their significance in confirmation of distribution is learnt.
- Linear regression and Multiple regression applications in analysing the dependency of the variable.