Statistical Analysis of Solidification Crack Susceptibility

by

Nitesh Kumar Gouda

LIST OF CONTENTS

Introduction

Solidification defects such as porosity and hot cracking are commonly observed and existence of these drastically affects the physical and mechanical properties of the material. Hot cracking is a major problem for us so we should select our composition to minimize the hot crack susceptibility.

Austenitic stainless steels are the steels that contain no ferrite and these are susceptible to hot cracking during welding. Hot cracking means cracking that occurs during the welding, casting or hot working at the temperatures near melting point of material. The various types of the hot cracks occur in stainless steel weldments. Super solidus cracking may be manifest as solidification cracking, which occurring in presence of the liquid phase in fusion zone, or as liquation cracking in the heat-affected zone (HAZ) where it is accompanied by grain-boundary melting. Solidification cracking in the weld metal is considered the most deleterious and is more widely observed than the other types of cracking.

Solidification cracking is a major problem during welding of stainless steels. As there are many grades of steels are available and composition of all grades are different. So we should choose the composition such that it gives minimum solidification cracking susceptibility. So predicting the solidification cracking susceptibility by doing ‘longitudinal vare strain test’. Higher the value of this means higher the susceptibility of cracking. We have the small experimental dataset which contains 487 rows and 23 columns (compositions, strain, and dependent variable).

Source of the Data: The Dataset is from a Journal on “Using deep neural network with small dataset to predict material defects” published by Shuo Feng, Huiyu Zhou, Hongbiao Dong.

Motivation:

As much research experience is available on the nature of hot cracking in stainless steels and various measures which required for minimizing it, but the complete understanding of the phenomenon is still lacking. Further, new materials are continuously being developed for various applications such as power systems, nuclear, chemical and petrochemical industries, driven by requirements of higher operating temperatures. The material design criteria for these systems may vary from thermal stability, and resistance to enhanced creep under irradiation and corrosion resistance in various media. Thus, these materials may not be designed primarily to be weldable and there is a continuing need to solve welding problems in their fabrication. In this study we will do EDA, Distributional Analysis, Hypothesis testing, linear regression for analyzing and predicting the solidification crack susceptibility.

Exploratory Data Analysis:

Data cleaning:

Raw dataset with 487 rows and 23 columns was reduced to 487 rows and 22 columns after the data pre-processing step. The extra column i.e 'Sr. No.' has been removed, and we have checked the data that whether there is any duplicate entry or null value is there but in our data I didn't find any of these. This dataset only contains parameters relevant to our study.After this data cleaning, data look like as:

In [4]:
```
df.head(10)
```

Out[4]:

	C	Si	Mn	P	S	Cr	Ni	Mo	N	Nb	...	Al	Ti	V	B	Th	I	U	Ve	Strain	Longitudinal_vare_straint_test
0	0.010	0.48	1.61	0.024	0.019	17.33	10.62	2.09	0.060	0.0	...	0.02	0.0	0.0	0.0	3.18	100	12.0	4.23	4.0	1.5
1	0.011	0.58	1.06	0.032	0.013	16.95	10.50	2.15	0.078	0.0	...	0.02	0.0	0.0	0.0	3.18	100	12.0	4.23	4.0	1.1
2	0.010	0.46	1.09	0.021	0.001	17.40	11.50	2.88	0.105	0.0	...	0.02	0.0	0.0	0.0	3.18	100	12.0	4.23	4.0	0.9
3	0.010	0.51	1.60	0.021	0.001	17.55	12.95	2.76	0.113	0.0	...	0.02	0.0	0.0	0.0	3.18	100	12.0	4.23	4.0	3.7
4	0.012	0.46	1.54	0.027	0.023	16.28	10.15	2.06	0.098	0.0	...	0.02	0.0	0.0	0.0	3.18	100	12.0	4.23	4.0	1.5
5	0.018	0.51	1.37	0.021	0.009	16.95	11.25	2.23	0.097	0.0	...	0.02	0.0	0.0	0.0	3.18	100	12.0	4.23	4.0	2.4
6	0.017	0.49	1.73	0.025	0.004	16.55	11.65	2.13	0.100	0.0	...	0.02	0.0	0.0	0.0	3.18	100	12.0	4.23	4.0	3.8
7	0.020	0.48	1.39	0.025	0.015	17.27	11.51	2.20	0.083	0.0	...	0.02	0.0	0.0	0.0	3.18	100	12.0	4.23	4.0	2.4
8	0.006	0.51	1.71	0.021	0.001	17.35	13.02	2.49	0.096	0.0	...	0.02	0.0	0.0	0.0	3.18	100	12.0	4.23	4.0	3.1
9	0.010	0.52	1.75	0.019	0.001	17.25	12.90	2.56	0.101	0.0	...	0.02	0.0	0.0	0.0	3.18	100	12.0	4.23	4.0	3.0

10 rows × 22 columns

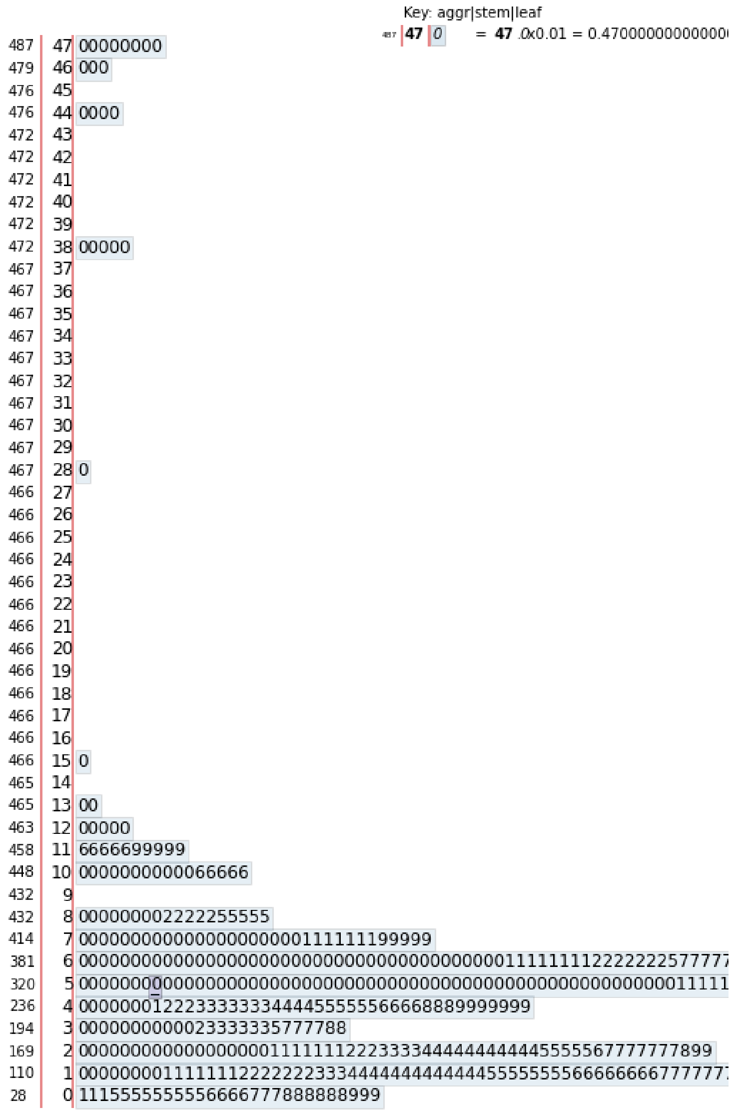
```
df.describe()
```

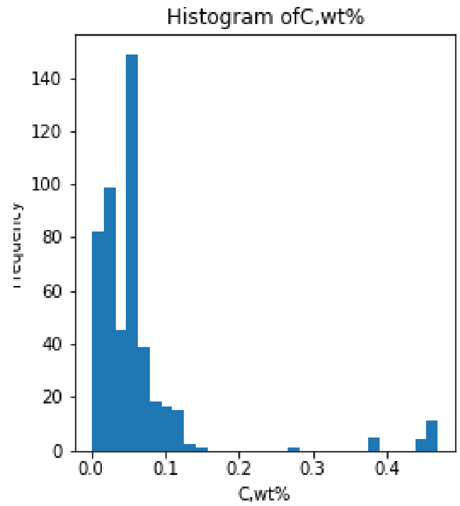
	C	Si	Mn	P	S	Cr	Ni	Mo	N	Nb	...	Al	Ti
count	487.000000	487.000000	487.000000	487.000000	487.000000	487.000000	487.000000	487.000000	487.000000	487.000000	...	487.000000	487.000000
mean	0.061951	0.627844	1.510893	0.019168	0.007691	19.112074	14.687187	0.924189	0.052364	0.082207	...	0.074885	0.113185
std	0.084147	0.662246	1.235992	0.009974	0.006582	3.276513	8.155322	1.115059	0.061406	0.181109	...	0.162976	0.193230
min	0.001000	0.010000	0.005000	0.001000	0.001000	12.630000	0.130000	0.000000	0.002000	0.000000	...	0.002000	0.000000
25%	0.020000	0.260000	0.930000	0.009500	0.003000	17.000000	9.650000	0.000000	0.019000	0.000000	...	0.020000	0.000000
50%	0.050000	0.500000	1.430000	0.021000	0.005000	18.320000	12.030000	0.190000	0.030000	0.000000	...	0.020000	0.000000
75%	0.061000	0.640000	1.700000	0.026000	0.012000	21.600000	16.870000	2.150000	0.060000	0.050000	...	0.023000	0.250000
max	0.470000	3.910000	8.360000	0.046000	0.032000	26.950000	33.950000	3.760000	0.368000	1.180000	...	1.080000	1.060000

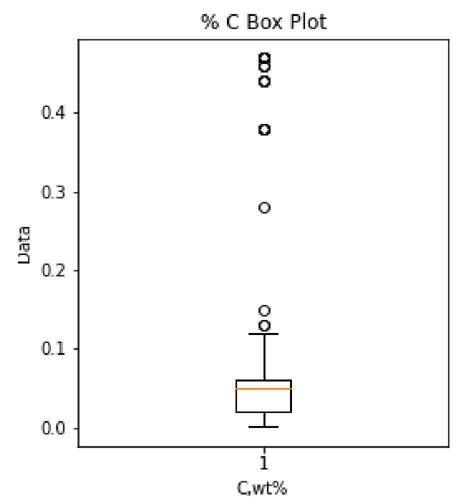
	V	B	Th	I	U	Ve	Strain	Longitudinal_vare_straint_test
count	487.000000	487.000000	487.000000	487.000000	487.000000	487.000000	487.000000	487.000000
mean	0.048747	0.000515	5.840955	147.186858	13.712526	2.669425	2.031930	3.845955
std	0.172598	0.001669	3.453002	81.701882	2.258616	1.151274	1.406767	4.188706
min	0.000000	0.000000	3.000000	50.000000	8.000000	1.250000	0.000000	0.000000
25%	0.000000	0.000000	3.180000	100.000000	12.000000	1.670000	1.000000	0.410000
50%	0.000000	0.000000	5.000000	100.000000	13.000000	2.530000	2.000000	2.500000
75%	0.000000	0.000000	6.400000	180.000000	16.000000	4.000000	3.200000	6.005000
max	0.840000	0.011400	12.700000	330.000000	17.000000	4.450000	5.000000	19.300000

Then I plotted the histograms, Box & whisker plots and for all the variables to explore the data

Stem & Leaf, Histogram, Box & Whisker Plot for wt% C



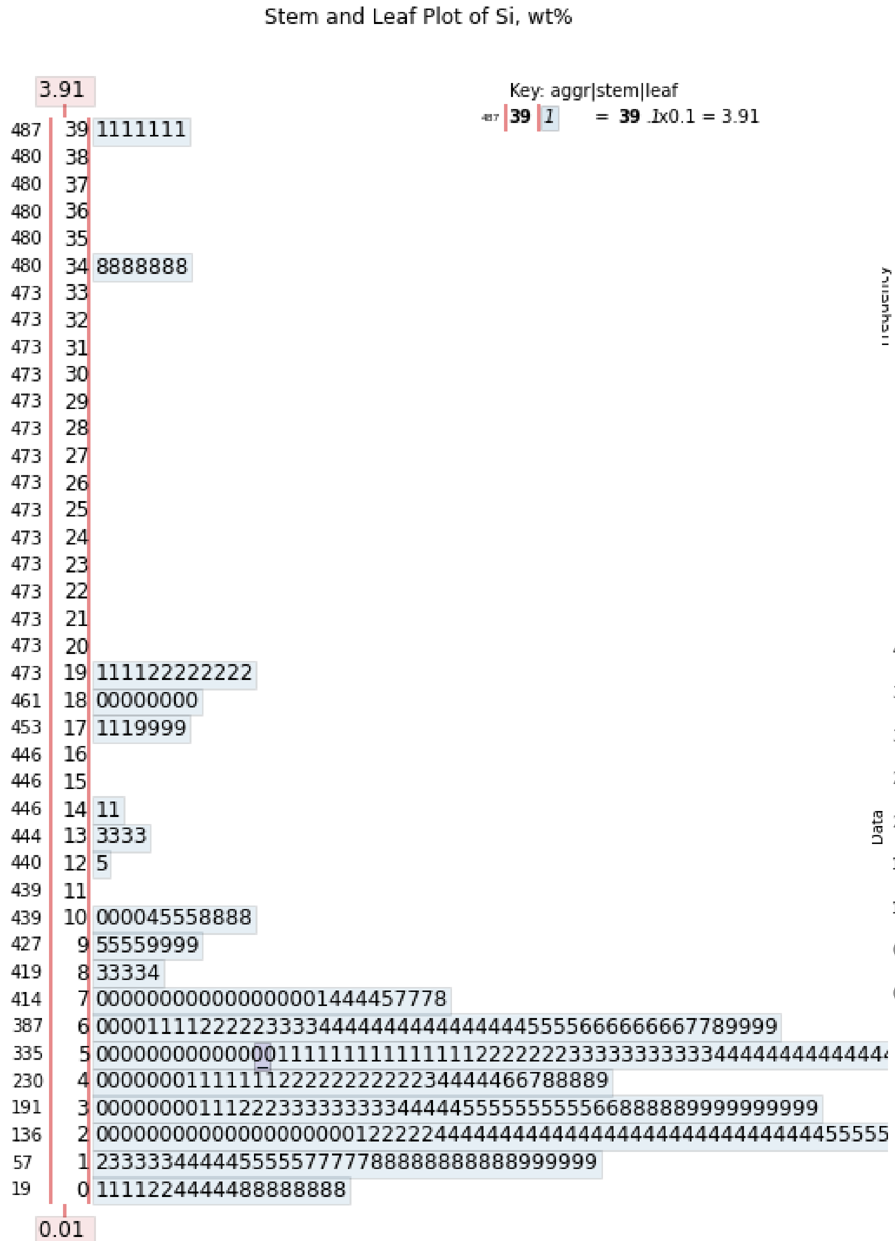


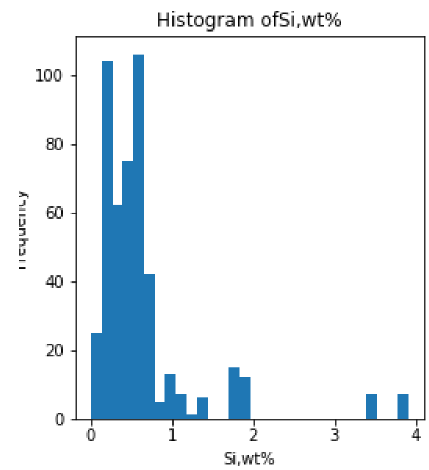


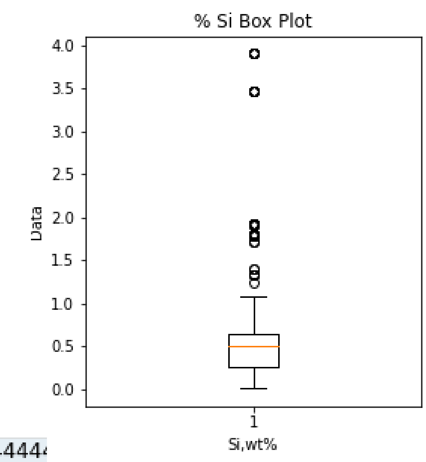
Observation:

- ❖ As from the stem & Leaf plot it can be seen that mostly values are in between 0.001 & 0.06.
- ❖ From Box & whisker plot, it can be seen that there are few outliers in the data. Q_1 = 0.02, Q_2 or median = 0.05 and Q_3= 0.061.
- ❖ From Histogram, it can be seen that mostly data lies between 0.001 & 0.15. but

Stem & Leaf, Histogram, Box & Whisker Plot for wt% Si



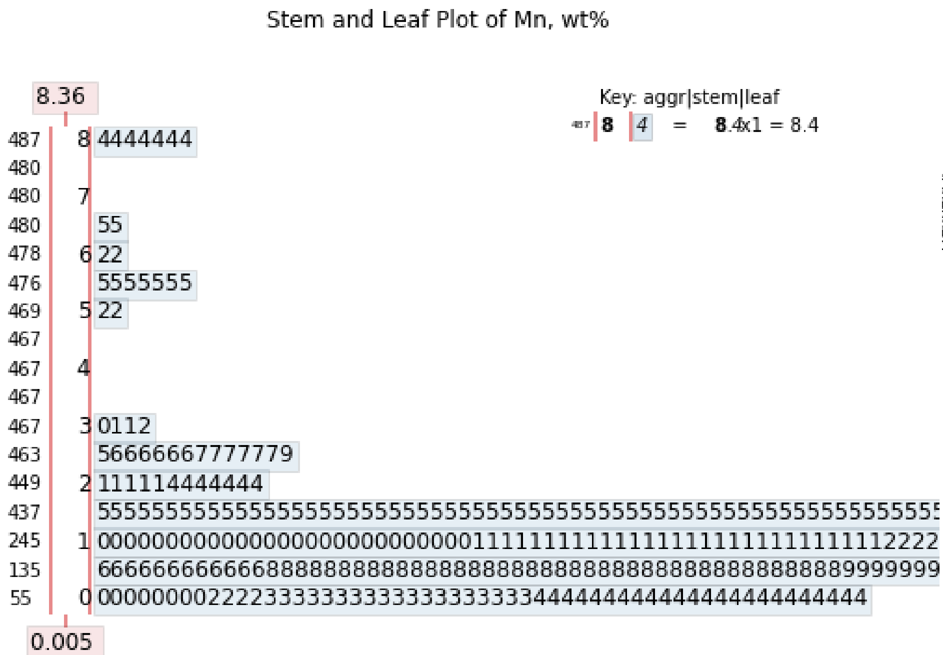


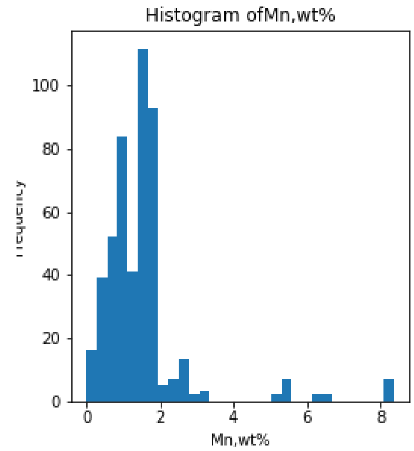


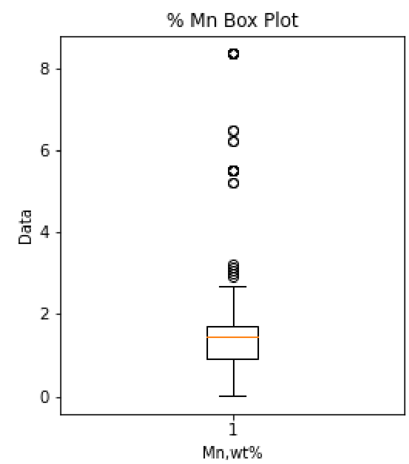
Observation:

- As from the stem & Leaf plot it can be seen that mostly values are in between 0.01 & 0.8.
- From Box & whisker plot, it can be seen that there are few outliers in the data.1st Quartile is at 0.26, 2nd Quartile or median is 0.5 and 3rd Quartile is at 0.64.
- From Histogram, it can be seen that mostly data lies between 0.01 & 2. but some other Si% is also there whose frequency is very less. i.e an outlier. Peak of the data are at about 0.6 and 0.8

Stem & Leaf, Histogram, Box & Whisker Plot for wt% Mn



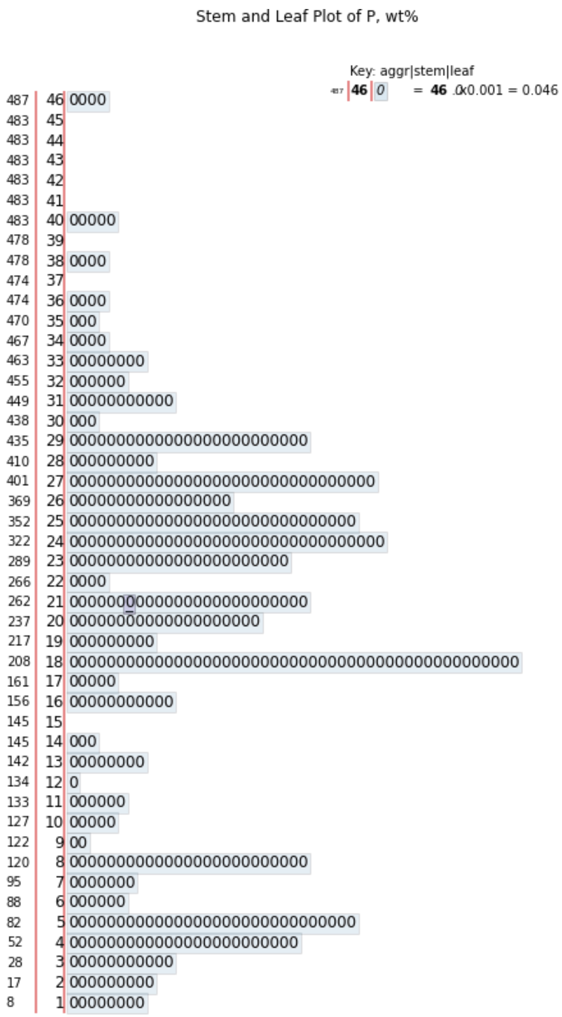


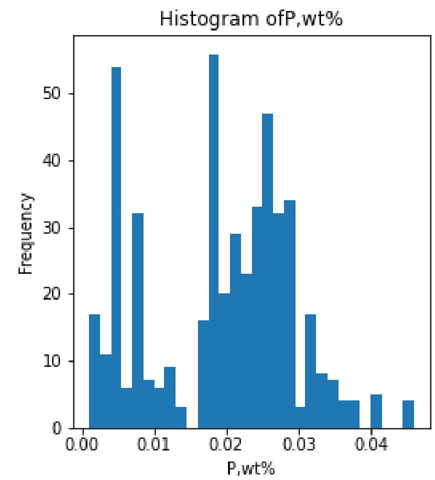


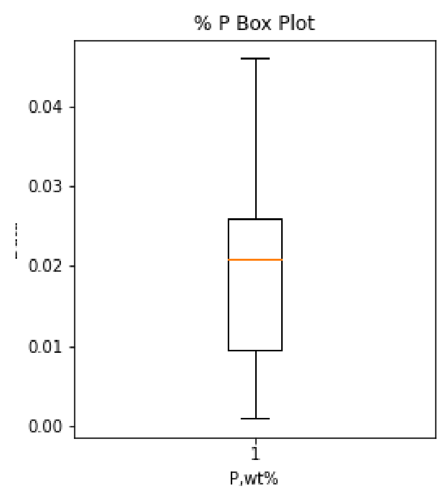
Observation:

- As from the stem & Leaf plot it can be seen that mostly values are in between 0.005 & 3.7.
- From Box & whisker plot, it can be seen that there are few outliers in the data. Q_1 = 0.93, Q_2 or median = 1.43 and Q_3= 1.7.
- From Histogram, it can be seen that mostly data lies between 0.005 & 3.7. but some other Mn % is also there whose frequency is very less. i.e an outlier. Peak of the data are at about 1.8.

Stem & Leaf, Histogram, Box & Whisker Plot for wt% P



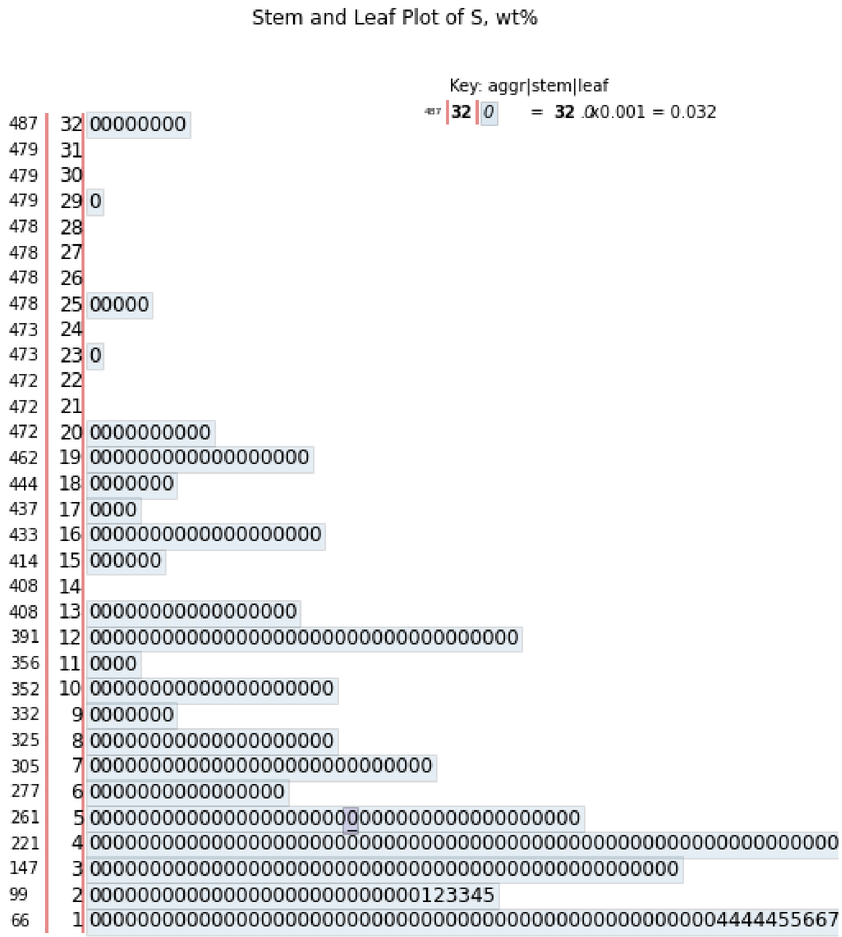


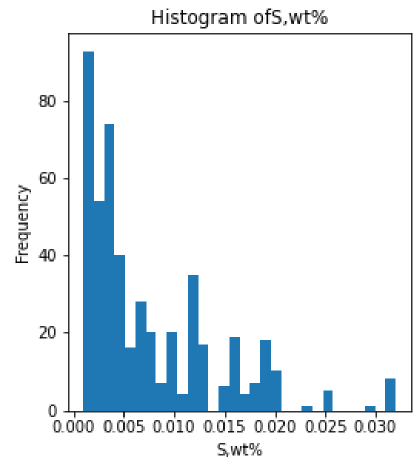


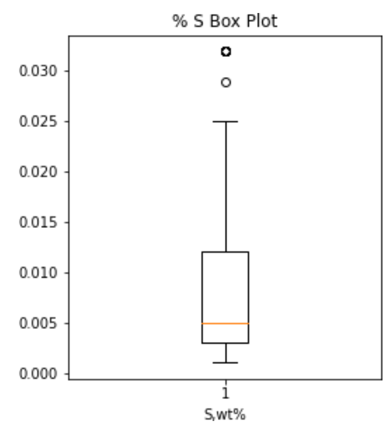
Observation:

- As from the stem & Leaf plot it can be seen that peak is at 0.018.
- From Box & whisker plot, it can be seen that there are no outliers in the data.Q_1 = 0.01, Q_2 or median = 0.021 and Q_3= 0.026, min = .001, max = 0.046.
- From Histogram, it can be seen that data lies in 2 ranges. 1st is in between .001 & .014 while 2nd is in between .016 to .036. while some other values are also there but frequency of these is very low.

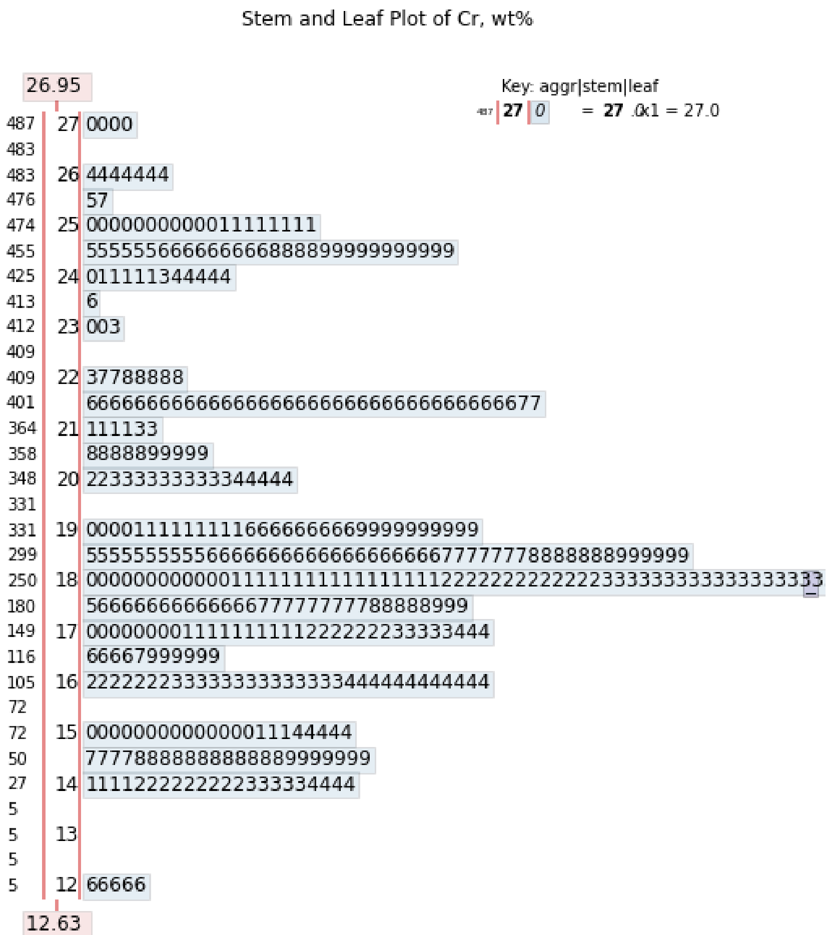
Stem & Leaf, Histogram, Box & Whisker Plot for wt% S

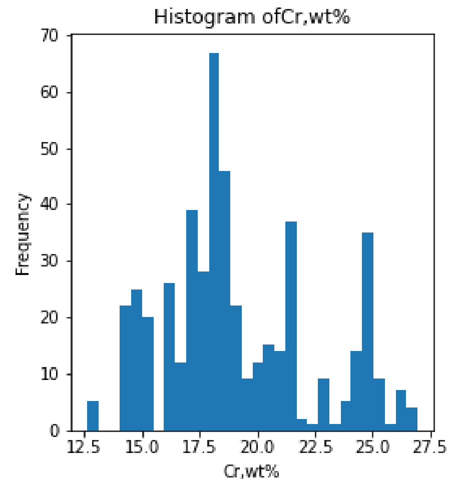


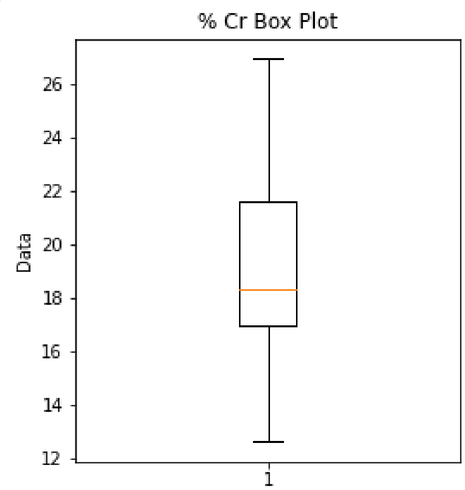




Stem & Leaf, Histogram, Box & Whisker Plot for wt% Cr



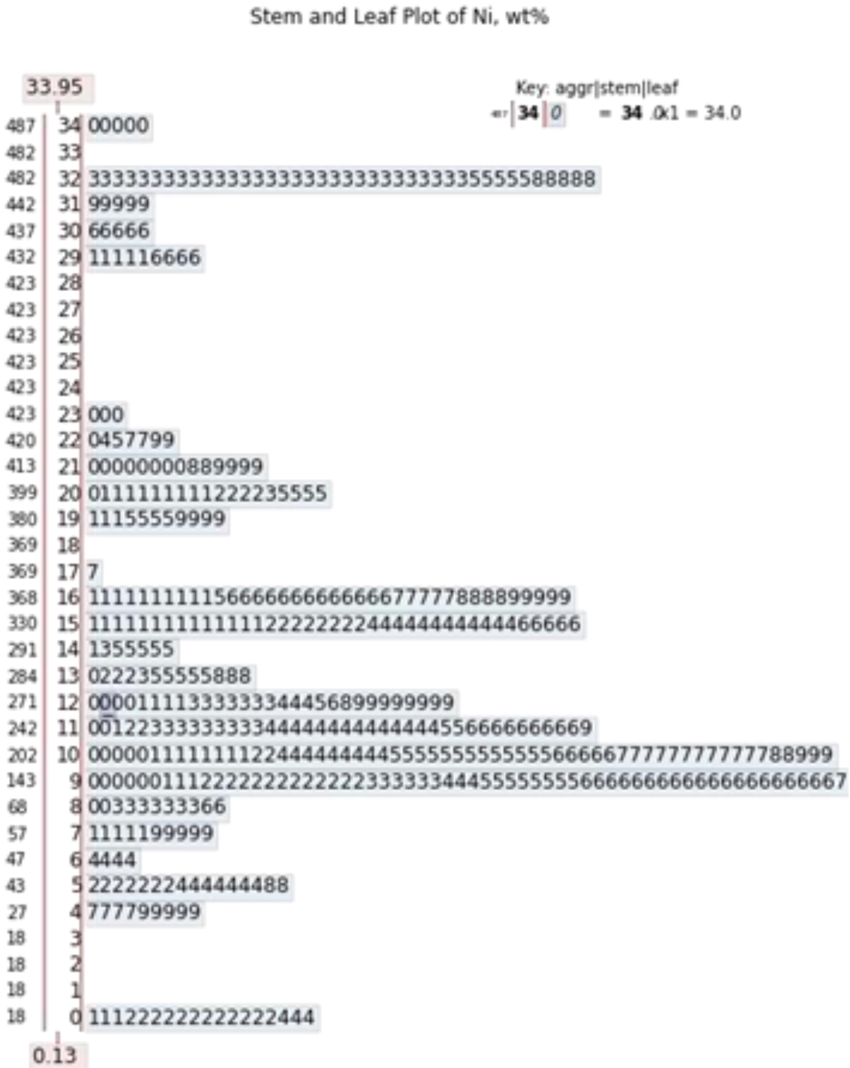


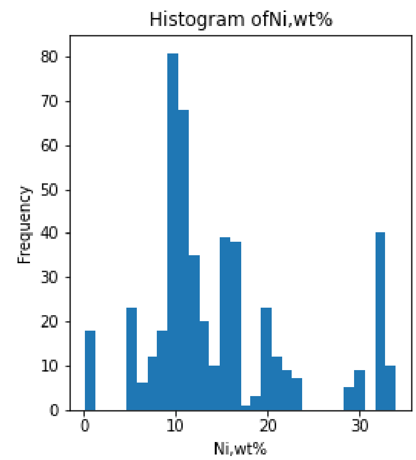


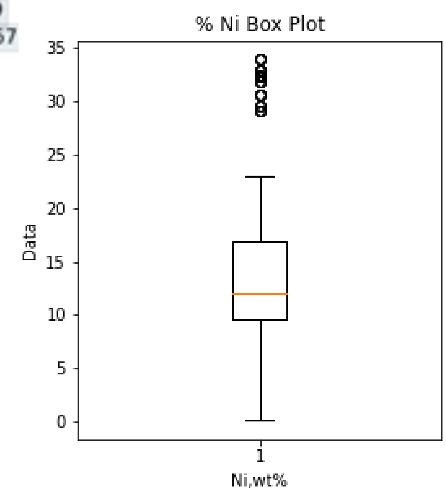
Observation:

- As from the stem & Leaf plot it can be seen that majority of values are in between 16.2 to 19.9 .
- From Box & whisker plot, it can be seen that there are no outliers in the data.Q_1 = 100, Q_2 or median = 100 and Q_3= 180, min = 50, max = 330
- From Histogram, it can be seen that peak is at about 18.3.

Stem & Leaf, Histogram, Box & Whisker Plot for wt% Ni



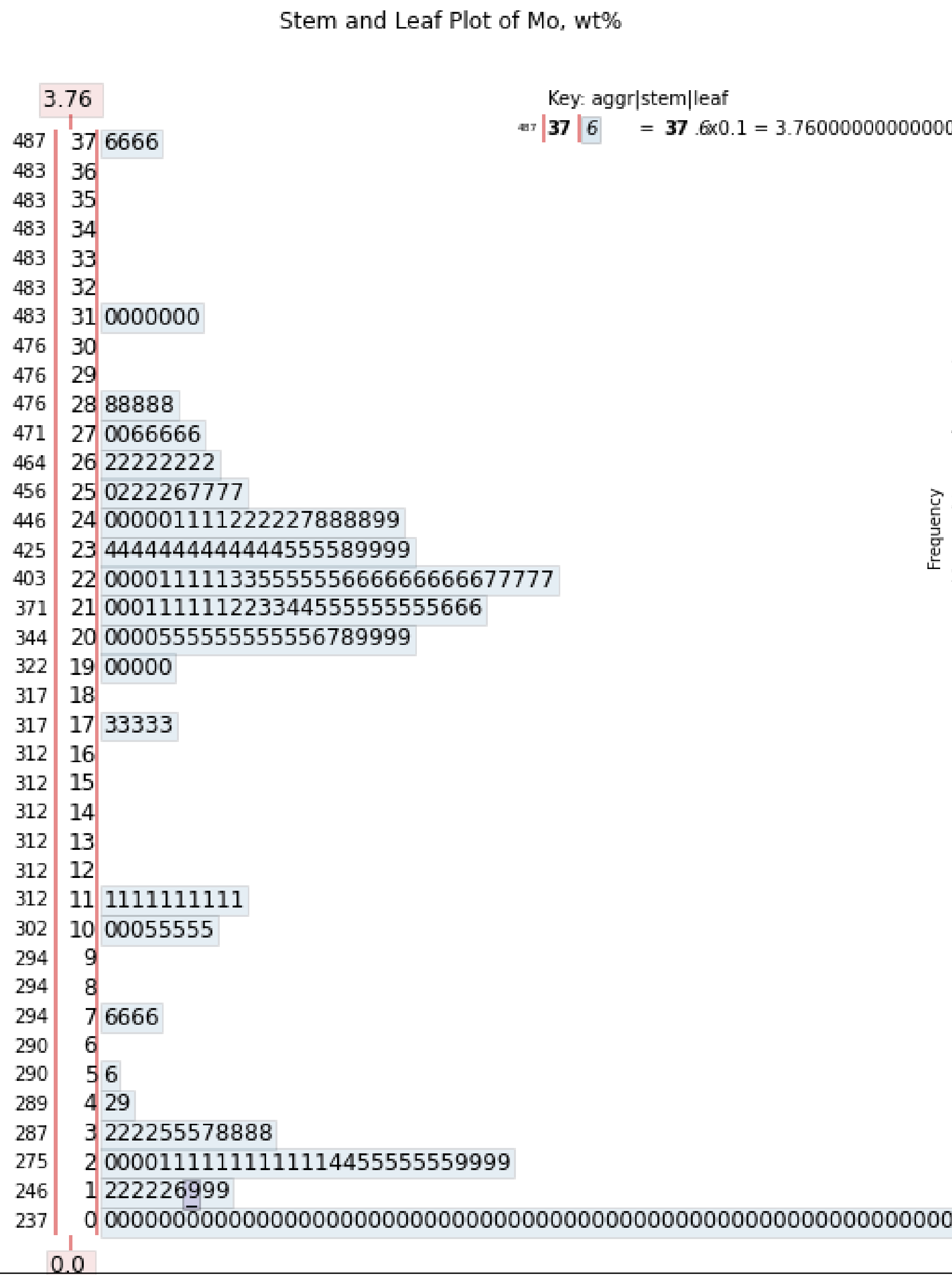


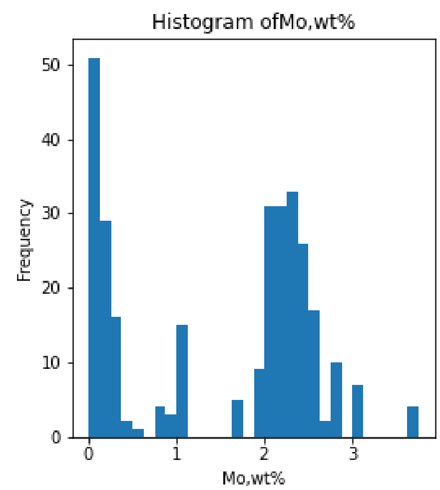


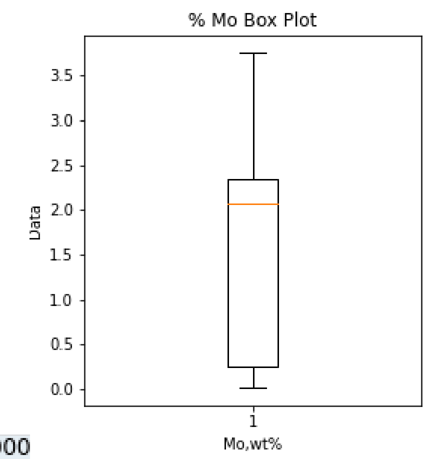
Observation:

- ❖ As from the stem & Leaf plot and histogram, it can be seen that peak is at around 10.
- ❖ As in Box and whisker plot, it can be seen that there are many outliers in the data. Data hasQ_1 = 9.65, Q_2 or median = 12.03 and Q_3= 16.87, min = .13, max= 33.95
- ❖ From Histogram, it can be seen that majority of data lies in between 4.7 & 23. While some other Ni wt % is also there in the data.

Stem & Leaf, Histogram, Box & Whisker Plot for wt% Mo



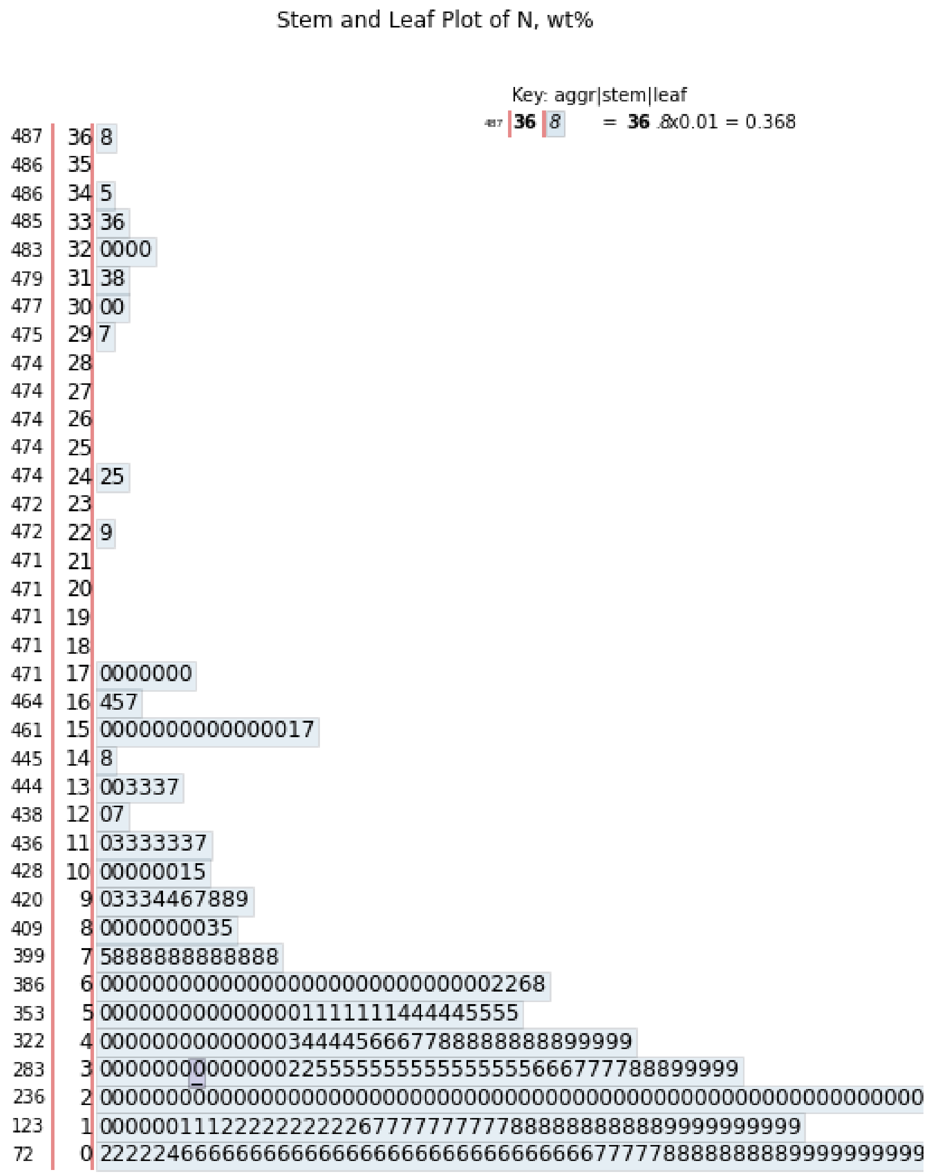


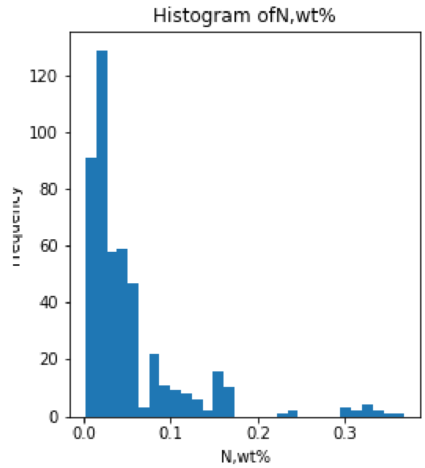


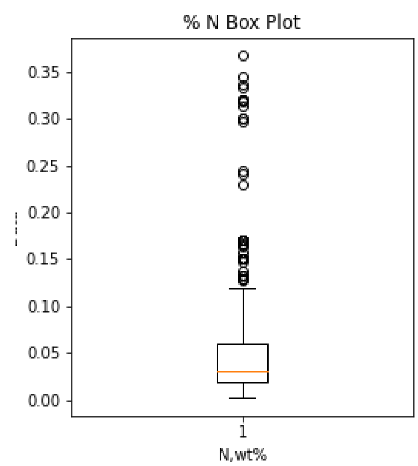
Observation:

- As from the stem & Leaf plot and histogram, it can be seen that peak is at around 0.
- As in Box and whisker plot, it can be seen that there are no outliers in the data. Data has $Q_1 = 0.25$, Q_2 or median = 2.07 and $Q_3 = 2.34$, min = 0.01.
- From Histogram, it can be seen that majority of data lies in between 1.9 & 2.88. While some other Mo wt % is also there in the data.

Stem & Leaf, Histogram, Box & Whisker Plot for wt% N



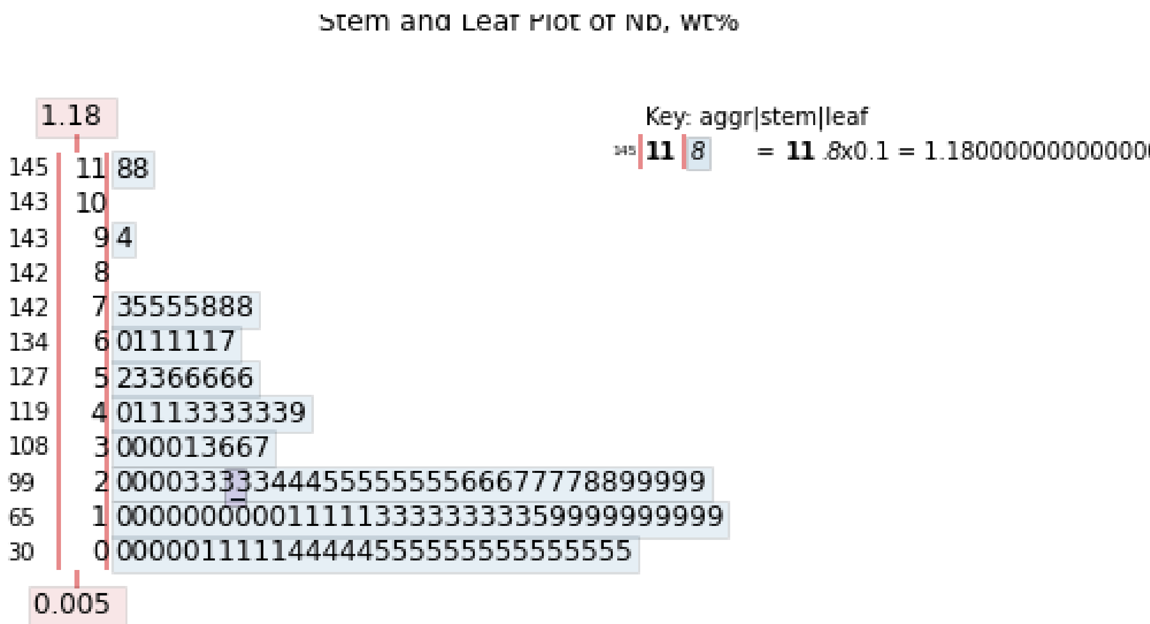


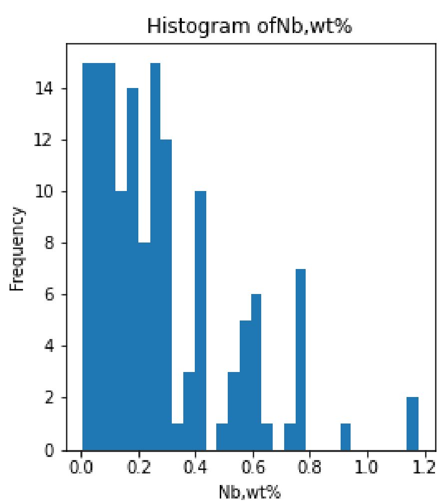


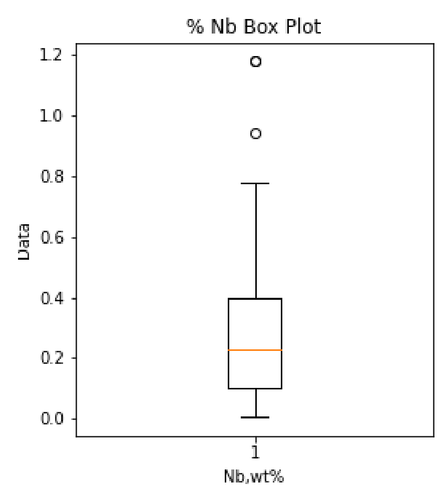
Observation:

- ❖ As from the stem & Leaf plot and histogram, it can be seen that peak is at around 0.006
- ❖ As in Box and whisker plot, it can be seen that there are many outliers in the data. Data has Q_1 = 0.019, Q_2 or median = 0.03, Q_3= .06, min = .002.
- ❖ From Histogram, it can be seen that majority of data lies in between 0.002 & 0.17. While some other N wt % is also there in the data.

Stem & Leaf, Histogram, Box & Whisker Plot for wt% Nb



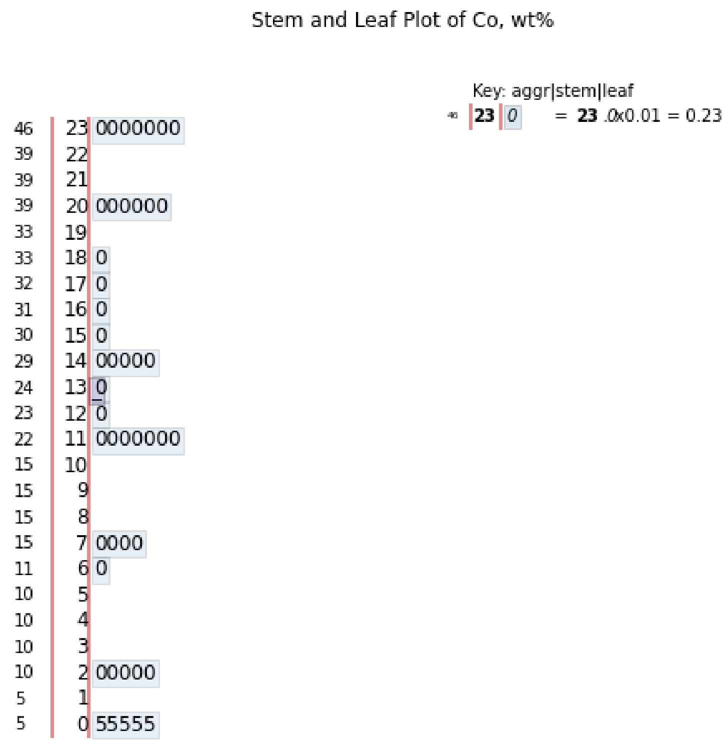


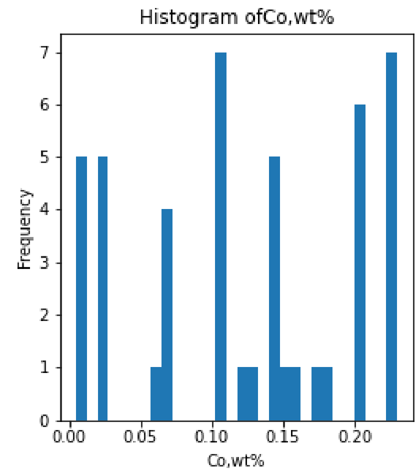


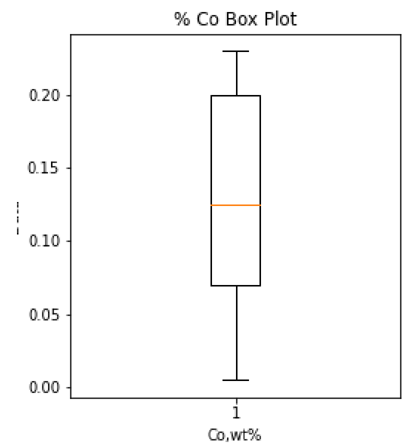
Observation:

- As in Box and whisker plot, it can be seen that there are 2 outliers in the data. Data has Q_1 = 0.1 , Q_2 or median = 0.23 , Q_3= 0.4 , min = 0.005
- From Histogram, it can be seen that majority of data lies in between 0.005 & 0.78. While some other Nb wt % is also there in the data as an outlier

Stem & Leaf, Histogram, Box & Whisker Plot for wt% Co



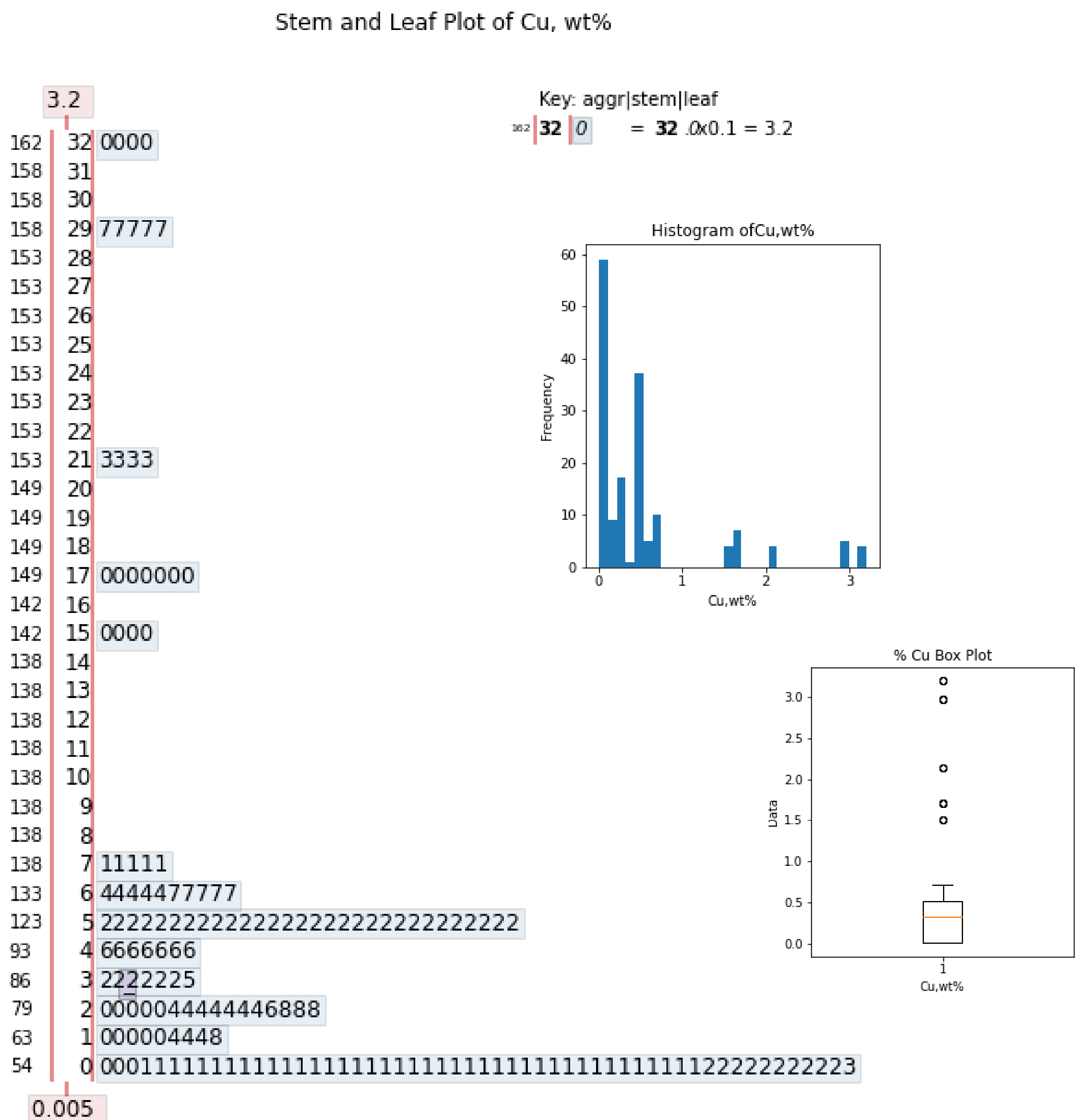




Observation:

- As in Box and whisker plot, it can be seen that there are no outliers in the data. Data has $Q_1 = 0.07$, Q_2 or median $= 0.125$, $Q_3 = 0.2$, min $= 0.005$
- From Histogram, it can be seen that there are multiple categories of data has been used for the experiment.

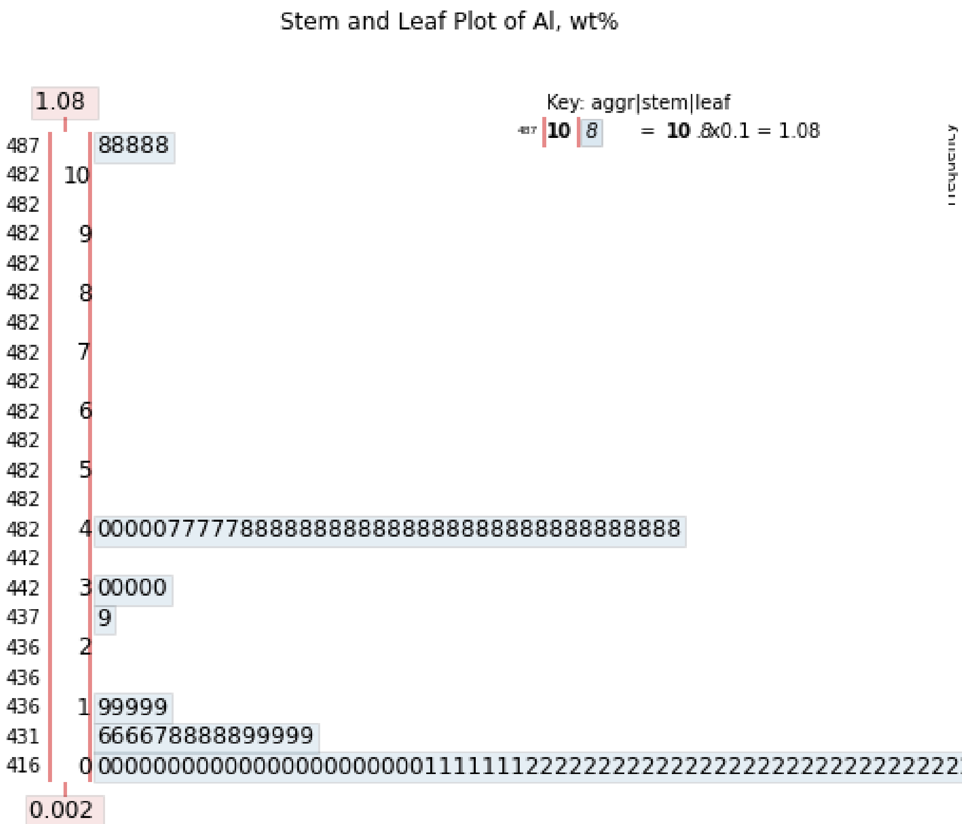
Stem & Leaf, Histogram, Box & Whisker Plot for wt% Cu

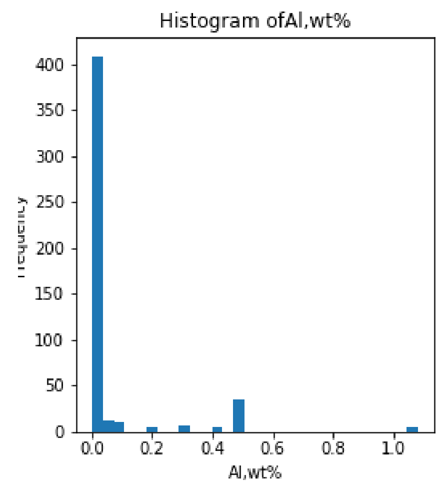


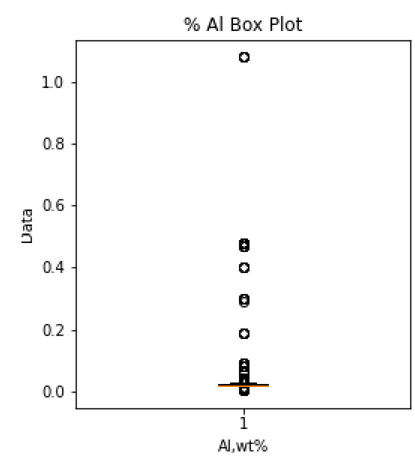
Observation:

- As in Box and whisker plot, it can be seen that there are five outliers in the data. Data has $Q_1 = 0.014$, Q_2 or median = 0.32 , $Q_3 = 0.52$, min = 0.005
- From Histogram and stem & leaf plot, it can be seen that mostly data lies between 0.005 and 0.74. while some different values also can be seen in the data.

Stem & Leaf, Histogram, Box & Whisker Plot for wt% Al



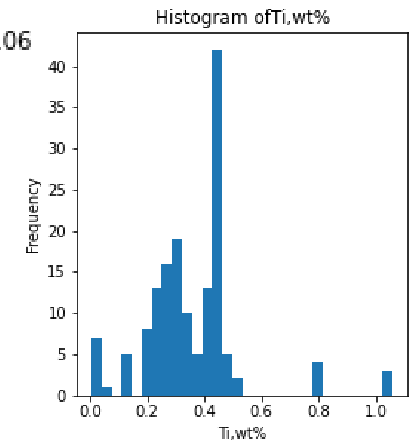


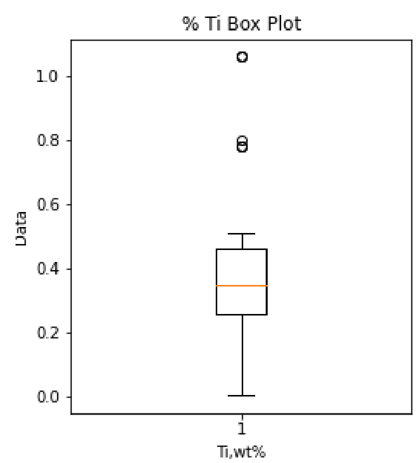


Observation:

- As in Box and whisker plot, it can be seen that there are too many outliers in the data. Data has $Q_1 = 0.02$, Q_2 or median = 0.02 , $Q_3 = 0.023$, min = 0.002
- From Histogram and stem & leaf plot, it can be seen that mostly data lies between 0.002 and 0.19.some data lies between 0.4 to 0.48 while some different values also can be seen in the data as an outlier.
- The peak can be seen at about 0.02

Stem & Leaf, Histogram, Box & Whisker Plot for wt% Ti

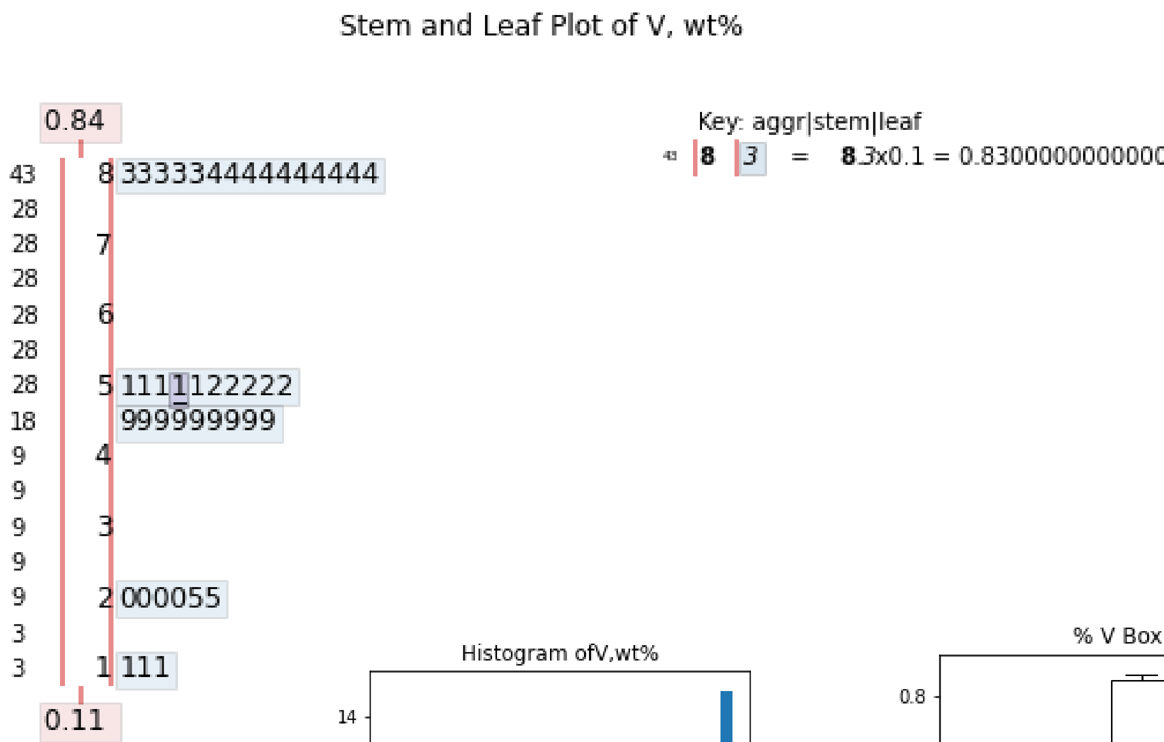


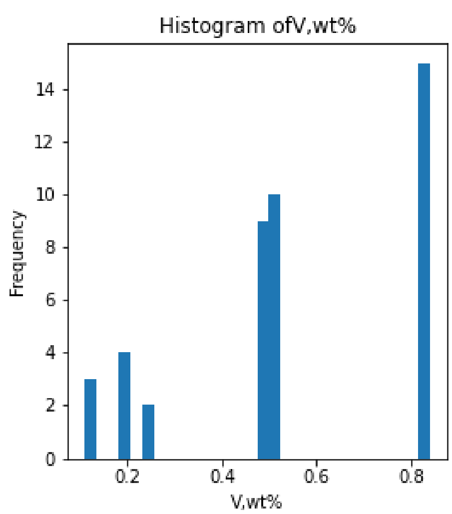


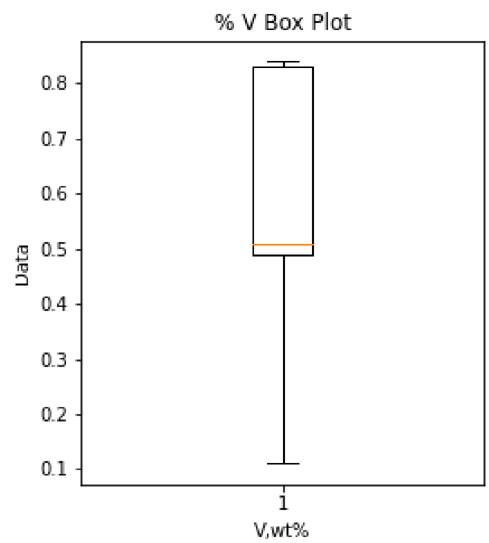
Observation:

- As in Box and whisker plot, it can be seen that there are three outliers in the data. Data has $Q_1 = 0.26$, Q_2 or median $= 0.35$, $Q_3 = 0.46$, min $= 0.005$
- From Histogram and stem & leaf plot, it can be seen that mostly data lies between 0.2 and 0.59. while some different values also can be seen in the data as an outlier.
- The peak can be seen at about 0.46

Stem & Leaf, Histogram, Box & Whisker Plot for wt% V



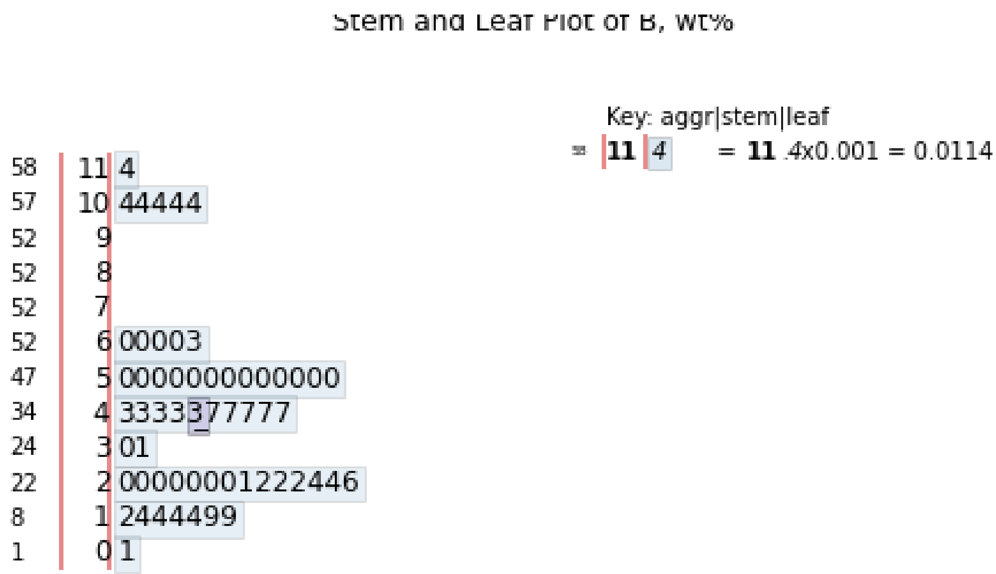




Observation:

- As in Box and whisker plot, it can be seen that there are no outliers in the data. Data has Q_1 = 0.49, Q_2 or median = 0.51 , Q_3= 0.83 , min =0.11
- From Histogram and stem & leaf plot, it can be seen that there are many kind of V wt% has been used.
- The peak can be seen at about 0.49 & 0.84

Stem & Leaf, Histogram, Box & Whisker Plot for wt% B

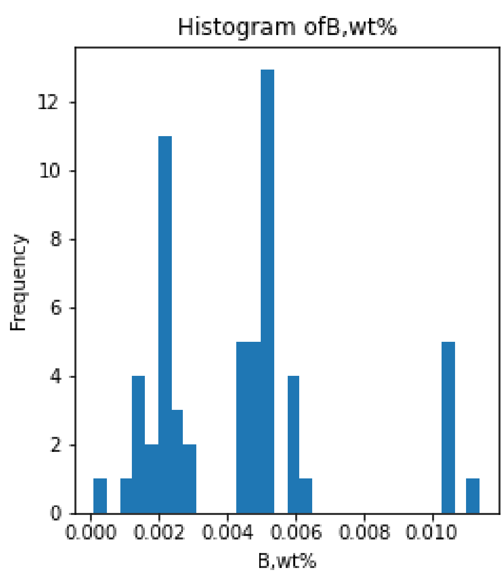


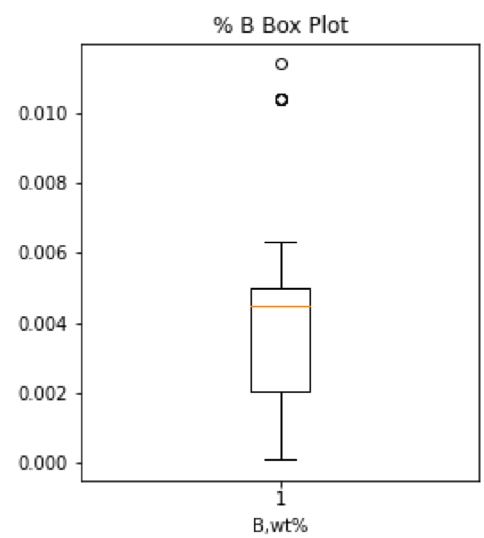
Stem and Leaf Plot of B, wt%

Key: aggr|stem|leaf

58 |11 |4 = 11.4x0.001 = 0.0114

58	11	4
57	10	44444
52	9	
52	8	
52	7	
52	6	00003
47	5	0000000000000
34	4	3333377777
24	3	01
22	2	00000001222446
8	1	2444499
1	0	1

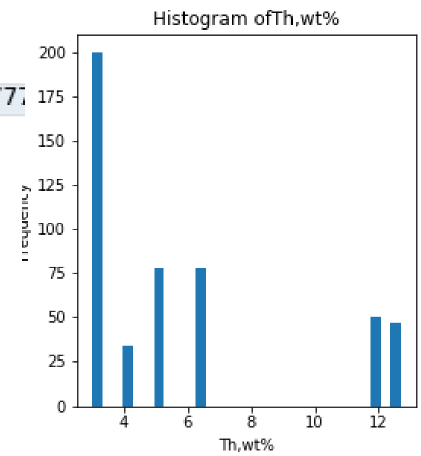


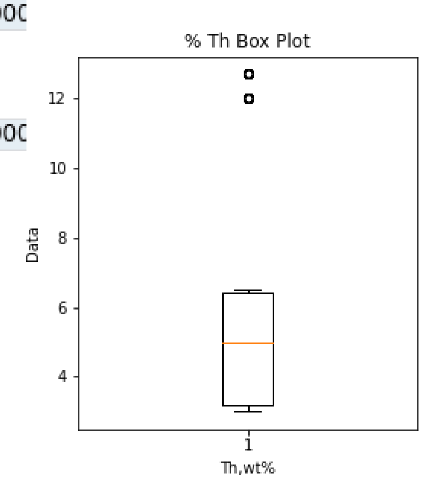


Observation:

- ❖ As in Box and whisker plot, it can be seen that there are 2 outliers in the data. Data has Q_1 = 0.002, Q_2 or median = 0.0045 , Q_3= 0.005, min =0.0001
- ❖ From Histogram and stem & leaf plot, it can be seen that data lies in between 0.0012 to 0.0026 then 0.003 to 0.0063 then 0.014 to 0.114
- ❖ peak can be seen at about 0.005

Stem & Leaf, Histogram, Box & Whisker Plot for wt% Th

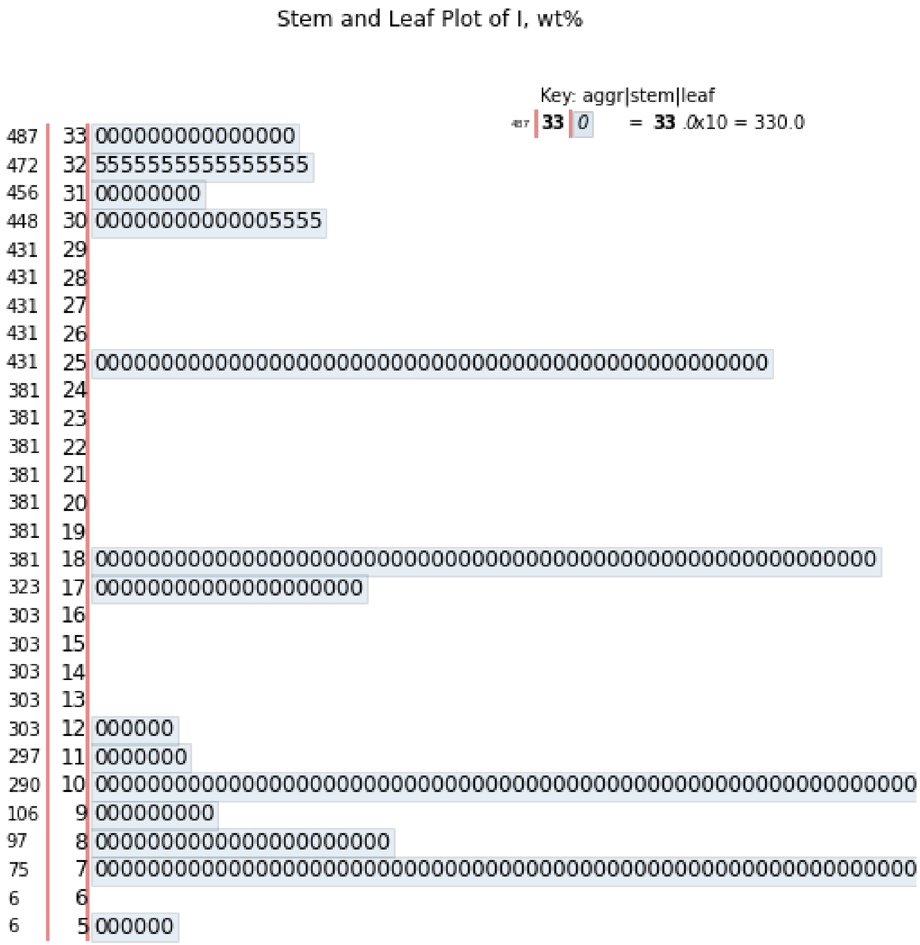


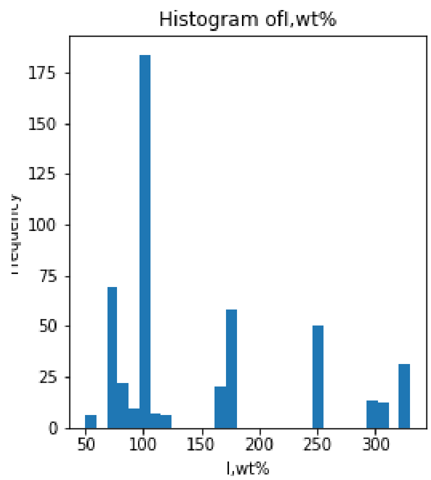


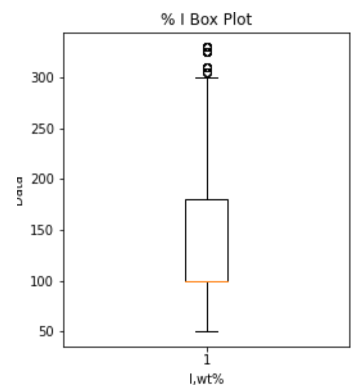
Observation:

- As in Box and whisker plot, it can be seen that there are 2 outliers in the data. Data has $Q_1 = 3.18$, Q_2 or median = 5, $Q_3 = 6.14$, min = 3
- From Histogram and stem & leaf plot, it can be seen that multiple wt% of Th has been used.
- peak can be seen at about 3

Stem & Leaf, Histogram, Box & Whisker Plot for wt% I



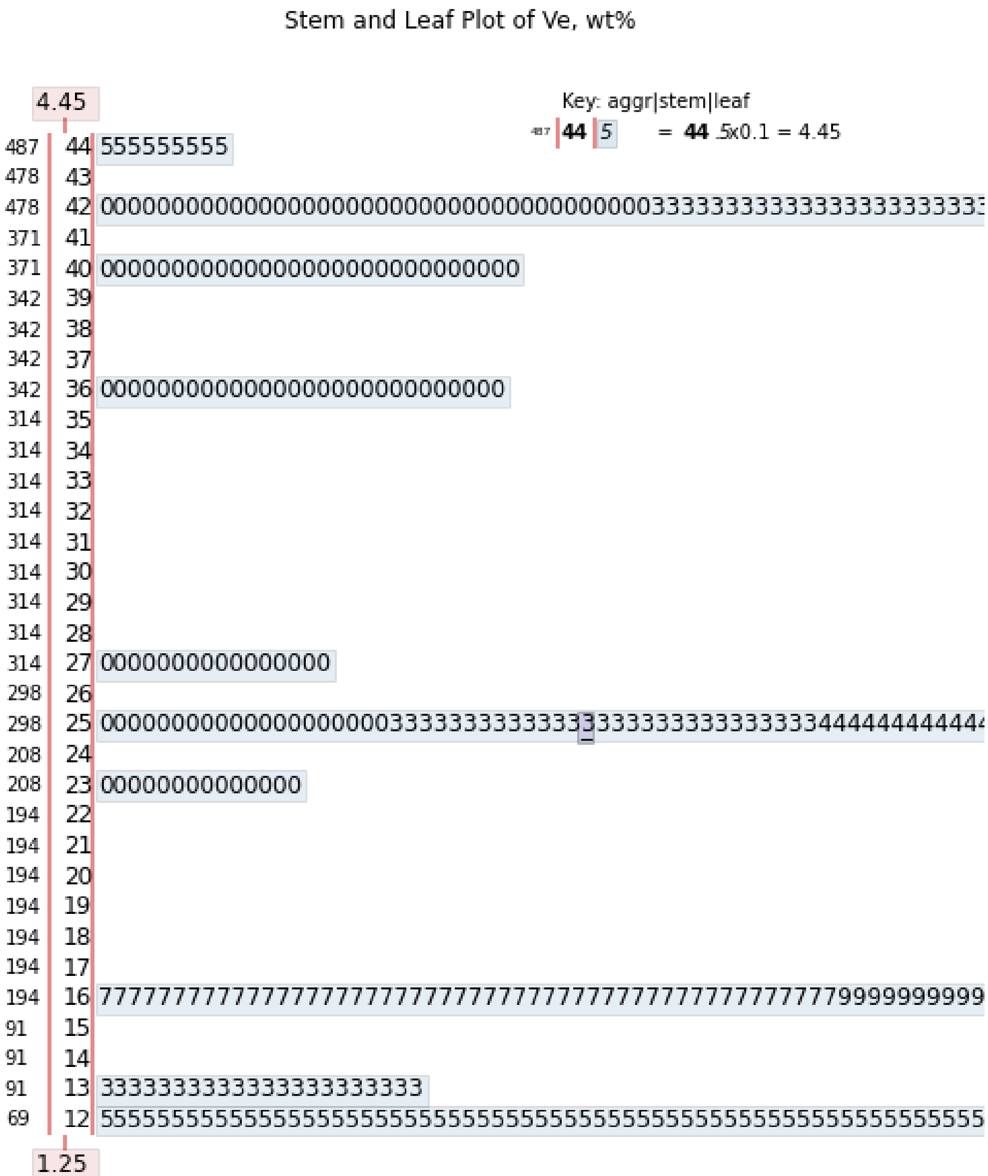


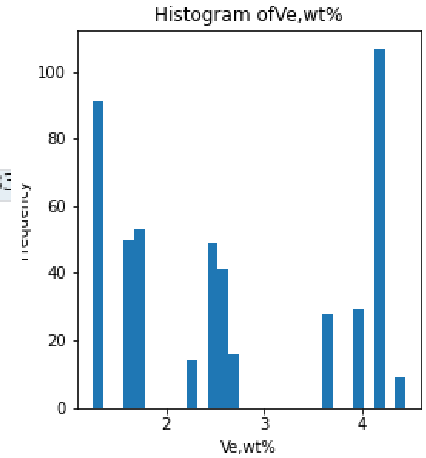


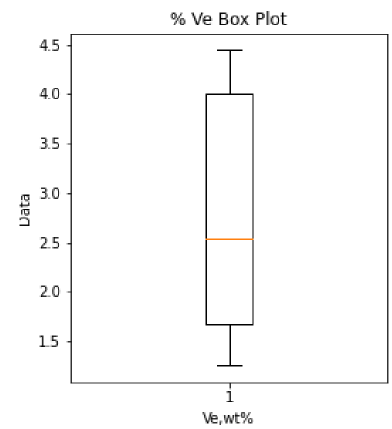
Observation:

- As in Box and whisker plot, it can be seen that there are 4 outliers in the data. Data has $Q_1 = 100$, Q_2 or median = 100, $Q_3 = 180$, min = 50, max = 330
- From Histogram and stem & leaf plot, it can be seen that multiple wt% of I has been used for the experiment.
- peak can be seen at 100

Stem & Leaf, Histogram, Box & Whisker Plot for wt% Ve



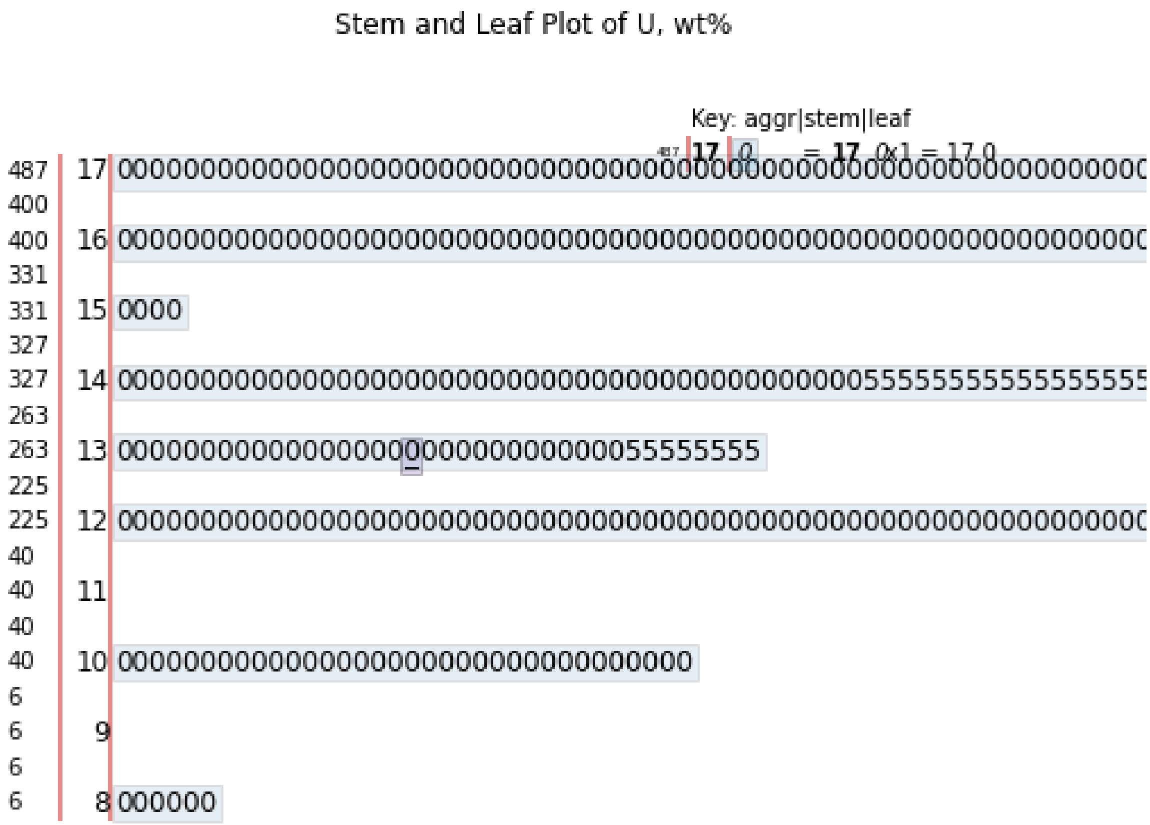


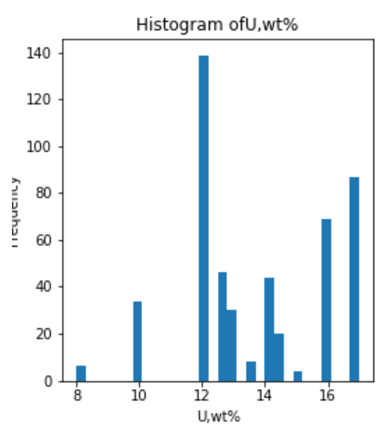


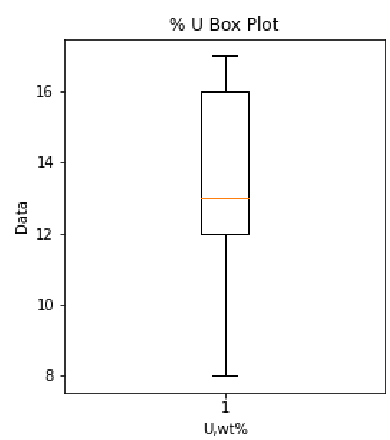
Observation:

- ❖ As in Box and whisker plot, it can be seen that there are 4 outliers in the data. Data has Q_1 = 1.67 , Q_2 or median = 2.53, Q_3= 4, min = 1.25
- ❖ From Histogram and stem & leaf plot, it can be seen that multiple wt% of Ve has been used for the experiment.
- ❖ peak can be seen at 4.2

Stem & Leaf, Histogram, Box & Whisker Plot for wt% U



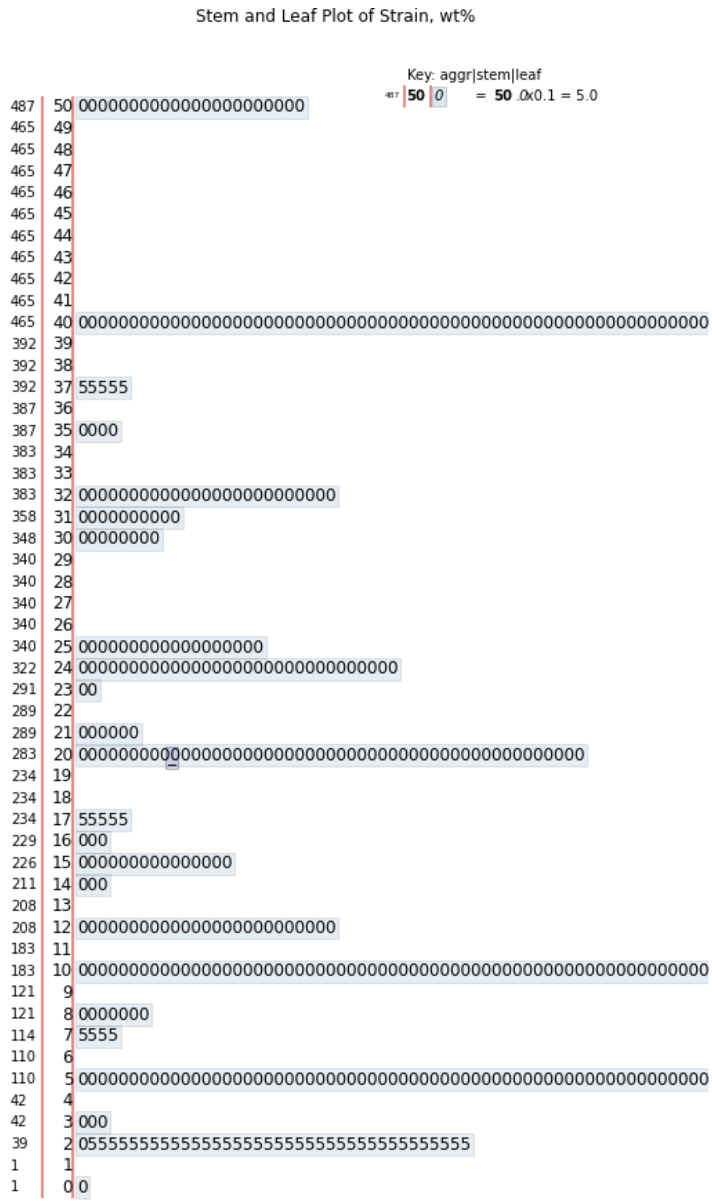


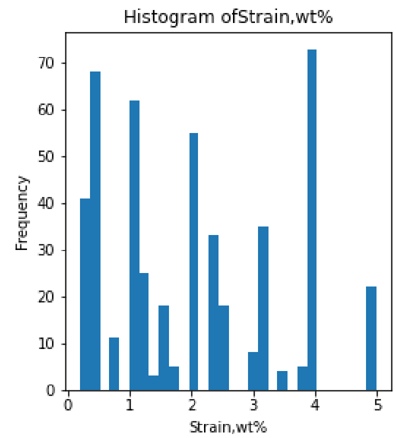


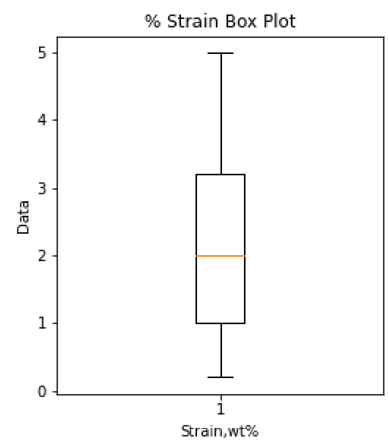
Observation:

- As in Box and whisker plot, it can be seen that there are 4 outliers in the data. Data has $Q_1 = 12$, Q_2 or median = 13, $Q_3 = 16$, min = 8
- From Histogram and stem & leaf plot, it can be seen that multiple wt% of U has been used for the experiment.
- peak can be seen at 12

Stem & Leaf, Histogram, Box & Whisker Plot for Strain







Observation:

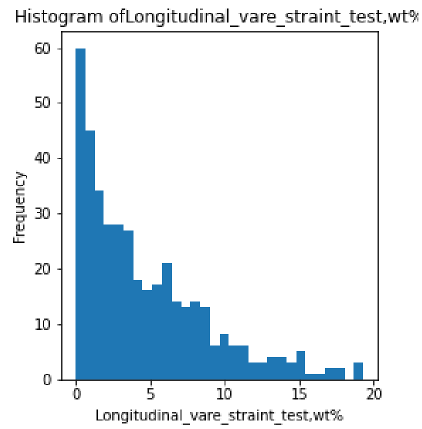
- As in Box and whisker plot, it can be seen that there are 4 outliers in the data. Data has Q_1 = 1, Q_2 or median = 2, Q_3= 3.2 , min = 0.2
- From Histogram and stem & leaf plot, it can be seen that multiple strain has been used for the experiment.
- peak can be seen at 0.5,1,2,4

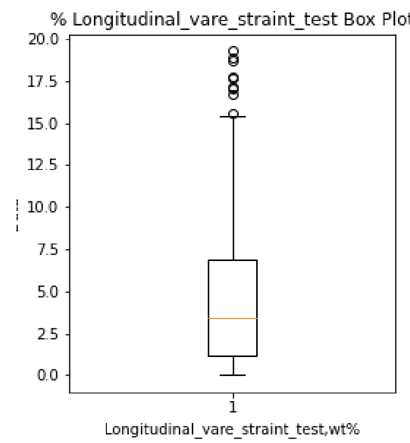
Stem & Leaf, Histogram, Box & Whisker Plot for Longitudinal_vare_starint_test

Stem and Leaf Plot of Longitudinal_vare_straint_test, wt%

Key: aggr|stem|leaf
487 |19 |3 = 19 .3x1 = 19.3

```
487  19|3
486    |78
484  18|
484    |
484  17|0168
480    |7
479  16|
479    |5
478  15|114
475    |89
473  14|234
470    |6678
466  13|013
463    |679
460  12|3
459    |5899
455  11|01223
450    |58
448  10|01123344
440    |57899
435   9|00002224
427    |6889999
420   8|00111333344444
406    |55677999
398   7|00011123344
387    |5566678999
377   6|00000001222233444
360    |556666677889999
345   5|0111112333444
332    |7788899999
322   4|0111111224444
309    |555666666667777788999999
285   3|00000011222222223344444
262    |5555555556666778888999
240   2|000000012223333444
222    |55555566666777788888888999
196   1|000000011111111111122233333444444
163    |55555555555566677777888888888889999
128   0|0000000000000000000000000000000000000000000000000000000000
```





Observation:

- ❖ As in Box and whisker plot, it can be seen that there are some outliers in the data. Data has $Q_1 = 1.2$, Q_2 or median = 3.45, Q_3= 6.92, min = 0.01
- ❖ From Histogram and stem & leaf plot, it can be seen that data is distributed between 0 to 19.3.
- ❖ peak can be seen at 0.

Backward Elimination Method:

I made a OLS regression result using python for the current scenario and this looks like as:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.463
Model:                            OLS   Adj. R-squared:                  0.445
Method:                 Least Squares   F-statistic:                     25.32
Date:                Tue, 05 Apr 2022   Prob (F-statistic):           1.49e-53
Time:                        14:28:40   Log-Likelihood:                -1236.7
No. Observations:                 487   AIC:                             2507.
Df Residuals:                     470   BIC:                             2579.
Df Model:                          16
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         -3.2536      1.389     -2.343      0.020      -5.983      -0.524
x1             1.5159      0.342      4.436      0.000       0.844       2.187
x2            -0.7511      0.202     -3.709      0.000      -1.149      -0.353
x3            74.3898     18.147      4.099      0.000      38.730     110.049
x4           -39.3578     26.743     -1.472      0.142     -91.908      13.193
x5            -0.1051      0.064     -1.632      0.103      -0.232       0.021
x6             0.2828      0.024     11.653      0.000       0.235       0.331
x7            -0.2416      0.158     -1.527      0.127      -0.552       0.069
x8             6.9841      3.416      2.044      0.041       0.271      13.697
x9            -1.1747      0.935     -1.256      0.210      -3.012       0.663
x10          -12.2488      3.657     -3.349      0.001     -19.435      -5.062
x11            0.8220      0.296      2.772      0.006       0.239       1.405
x12            1.2039      1.029      1.170      0.243      -0.818       3.225
x13            2.7564      1.155      2.387      0.017       0.487       5.026
x14          253.1904    116.647      2.171      0.030      23.977     482.404
x15            0.0047      0.002      1.883      0.060      -0.000       0.010
x16            1.4113      0.110     12.792      0.000       1.194       1.628
==============================================================================
Omnibus:                       79.719   Durbin-Watson:                   1.226
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              144.695
Skew:                           0.950   Prob(JB):                     3.80e-32
Kurtosis:                       4.876   Cond. No.                     1.40e+05
==============================================================================
```

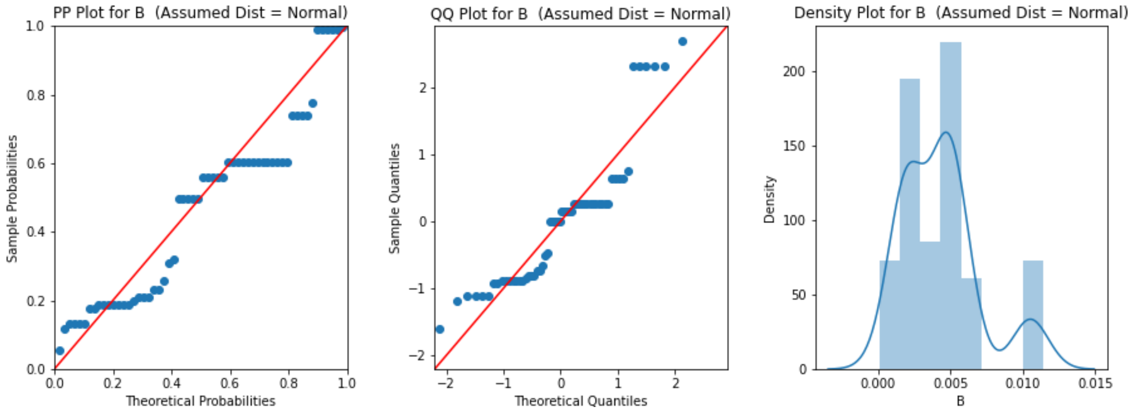
As in addition of redundant variables, R^2 is not a good measure to check because its values either remains constant or increases in case of redundant variables. So I have checked the adjusted R^2 value. If by removing the column adj. R^2 is increasing then remove that column and if it is decreasing then do not remove that column. I removed the column for which p value is maximum and then check the adjusted R2 value and repeated the same process. So summary of my OLS results is shown below.

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
Adj. R-squared: 0.441																				
C	Si	Mn	P	S	Cr	Ni	Mo	N	Nb	Co	Cu	Al	Ti	V	B	Th	I	U	Ve	Strain
Adj. R-squared: 0.442																				
C	Si	Mn	P	S	Cr	Ni	Mo	N	Nb	Co	Cu	Ti	V	B	Th	I	U	Ve	Strain	
Adj. R-squared: 0.444																				
C	Si	Mn	P	S	Cr	Ni	Mo	N	Nb	Co	Cu	Ti	V	B	Th	I	Ve	Strain		
Adj. R-squared: 0.444																				
Si	Mn	P	S	Cr	Ni	Mo	N	Nb	Co	Cu	Ti	V	B	Th	I	Ve	Strain			
Adj. R-squared: 0.445																				
Si	Mn	P	S	Cr	Ni	Mo	N	Nb	Co	Cu	Ti	V	B	Th	I	Strain				
Adj. R-squared: 0.446																				
Si	Mn	P	S	Cr	Ni	Mo	N	Nb	Co	Cu	Ti	V	B	I	Strain				Desired Columns	
Adj. R-squared: 0.445																				
Si	Mn	P	S	Cr	Ni	Mo	N	Nb	Co	Cu	V	B	I	Strain						

So this way I have reduced the 5 Variables (Al, U, C, Ve, Th) which were least important. Now I will do my further analysis on remaining 17 variables ('Si', 'Mn', 'P', 'S', 'Cr', 'Ni', 'Mo', 'N', 'Nb', 'Co', 'Cu', 'Ti', 'V', 'B', 'I', 'Strain', 'Longitudinal_vare_straint_test')

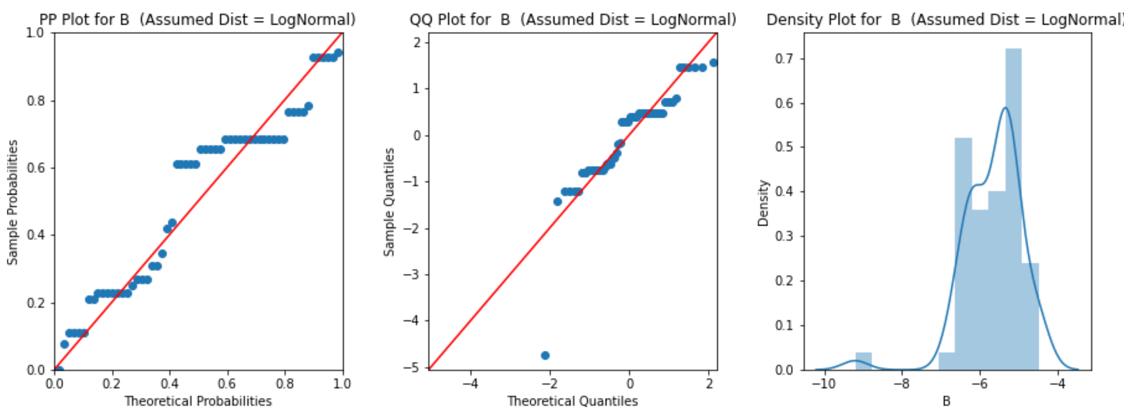
Distributional Analysis

Wt. % of B



Observation:

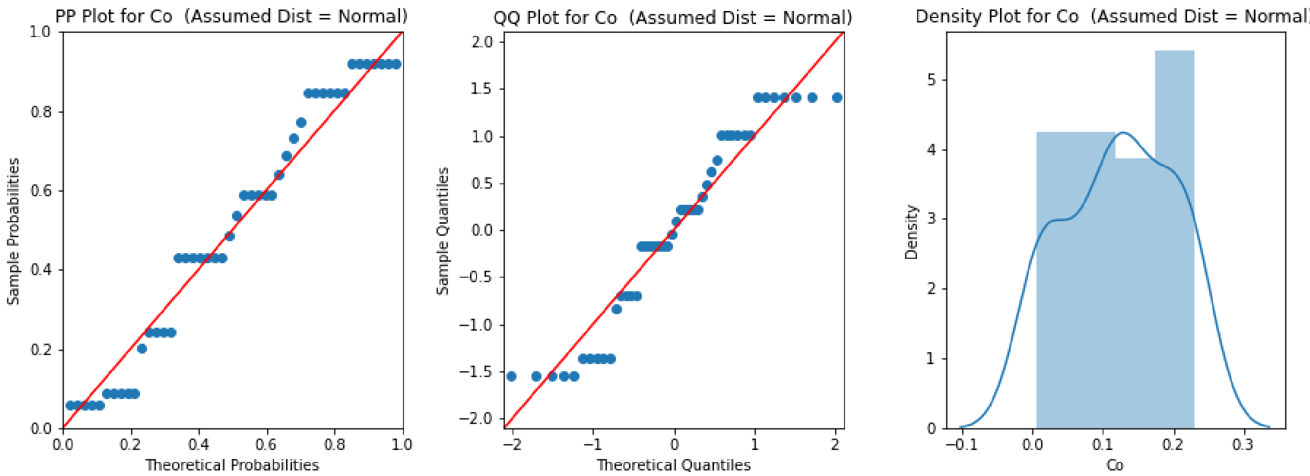
- From the density plot it could be seen that it has 2 peaks, one is in between 0 to 0.008 and second is in between 0.008 to 0.015.
- It shows that two types of Boron wt% has been used. But as from the PP and QQ plots, deviation from the standard normal distribution can be seen. So in below figure I am checking by taking assumed distribution as lognormal distribution.



Observation:

- In the QQ plot it can be seen that there is an outlier but as this is an experimental data, so might be that they have done an experiment on taking the exceptionally low or high value of wt. % of boron.
- It could also be seen from the density plot that it has 2 peaks, one is in between -10 to -8 and second is in between -8 to -4. It shows that two types of Boron wt% has been used.

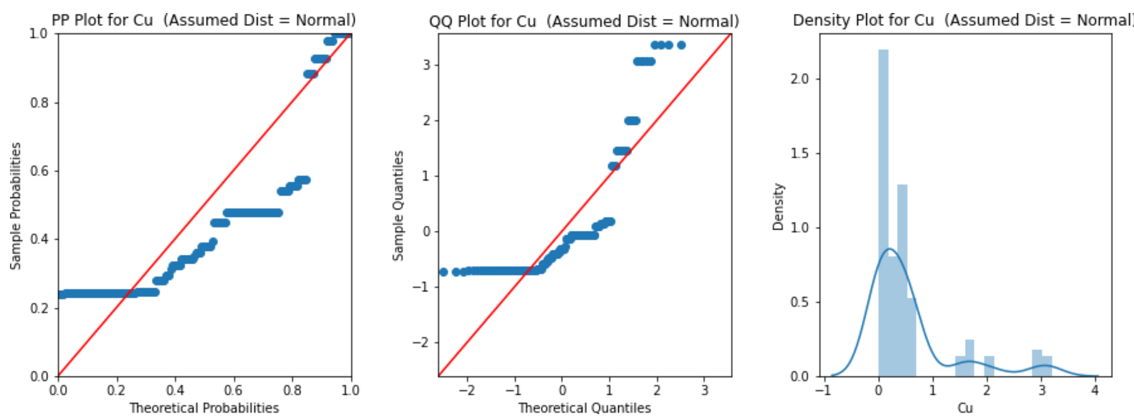
Wt. % of Co



Observation:

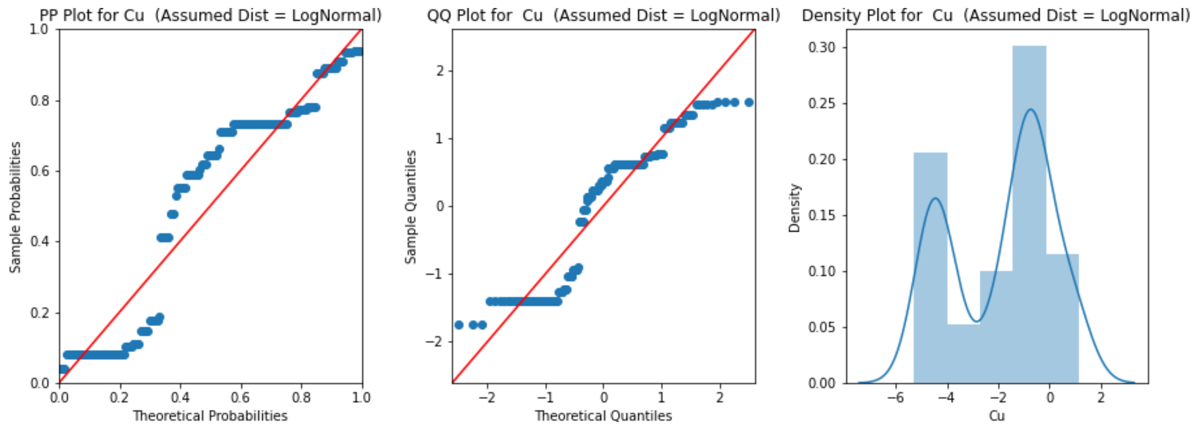
- In the density plot it can be seen that the three kind of Co wt% has been used, one is in between 0 to 0.1, second is in between 0.1 to 0.18 and thirdis in between 0.18 to 0.24.
- As from the PP and QQ plot we can assume it to be approx. normal distribution.

Wt. % of Cu



Observation:

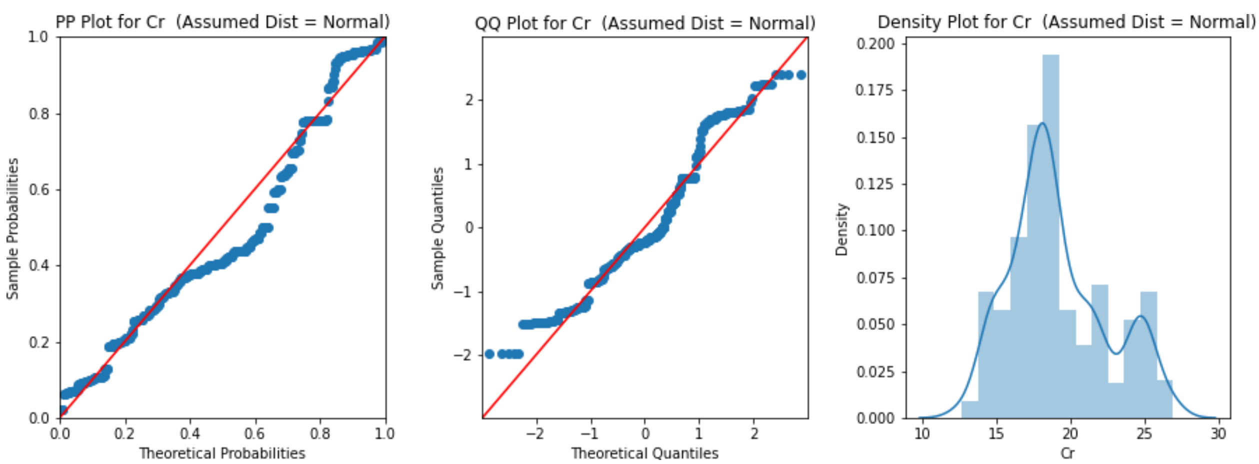
- In the density plot it can be seen that there are three peaks. One is in between 0 to1, second is in between 1 to 2.5 and third is in between 2.5 to 4. This means that three kind of Cu wt. % has been used.
- As this has a higher deviation from the Theoritical standard normal distribution so I am checking for lognormal distribution.



Observation:

- Here the theoretical distribution has better fit actual distribution as compare to previous case so I am assuming the cu wt% as lognormal distribution.

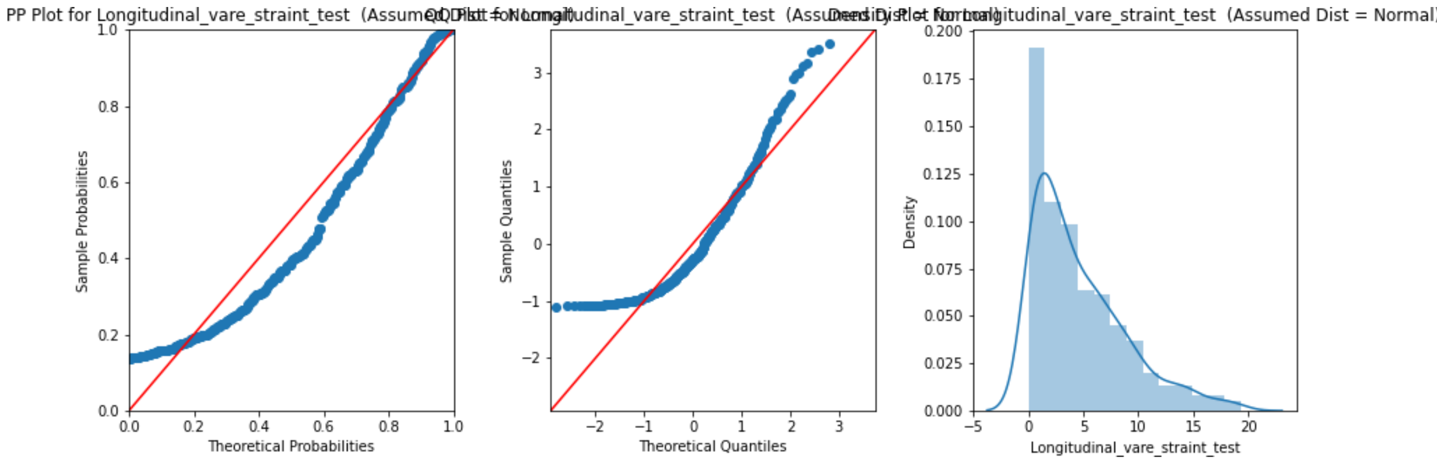
Wt. % of Cr



Observation:

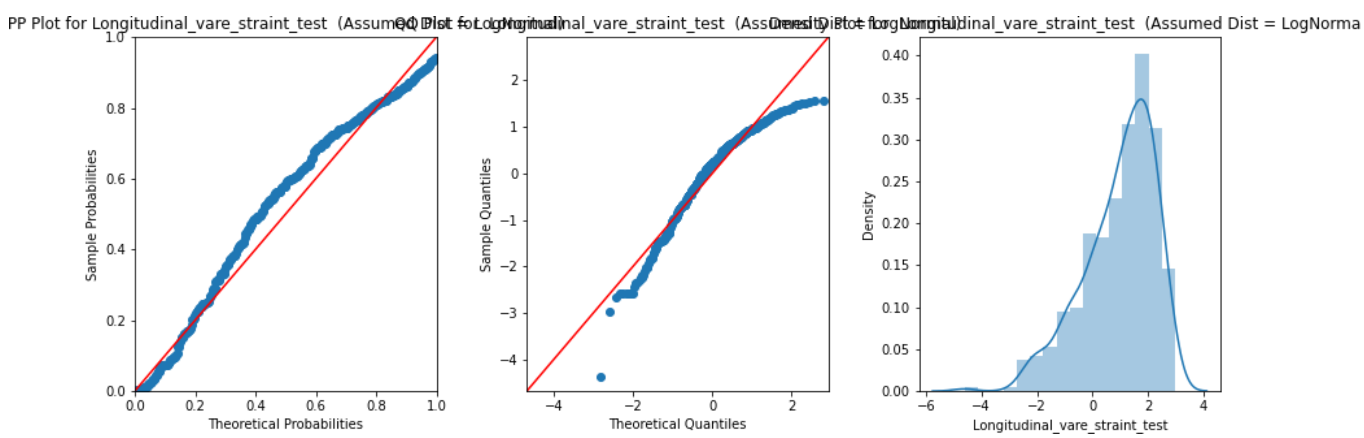
- In the density plot it can be seen that there are two peaks. One is in between 10 to 22 and second is in between 22 to 26. It means that the two types of cr wt % has been used in the experimental data.
- From the PP and QQ plot we can see that this data is almost following normal distribution so I am taking the cr wt % as normal distribution.

Longitudinal Vare straint test



Observation:

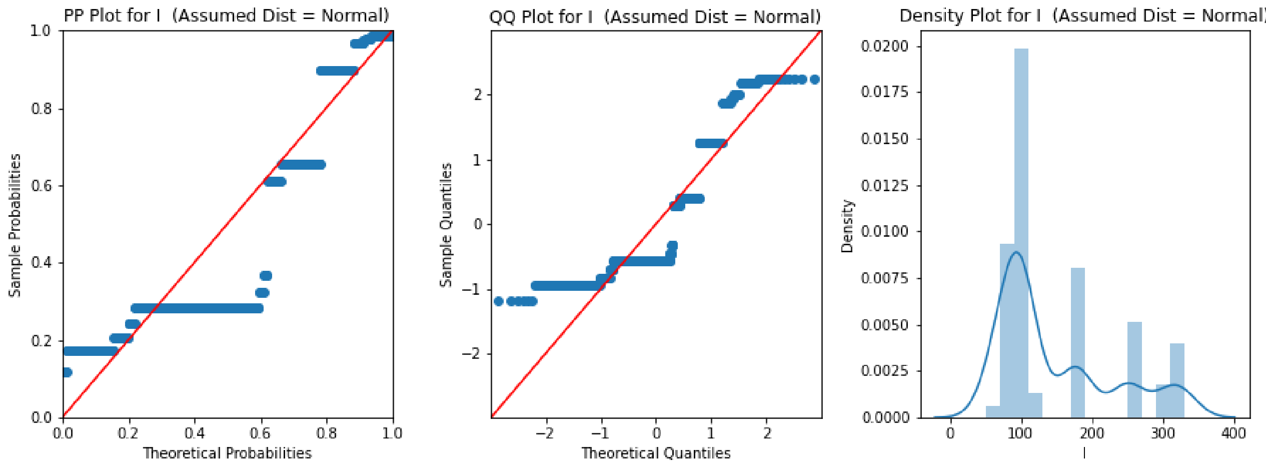
- As from the QQ plot it can be seen that it is a skewed distribution. So in below figure,I am checking the PP & QQ plot by taking my assumed distribution as lognormal distribution.



Observation:

- As we can see that this has less deviation from the previous one so I am assuming that my dependent variable is following lognormal distribution.

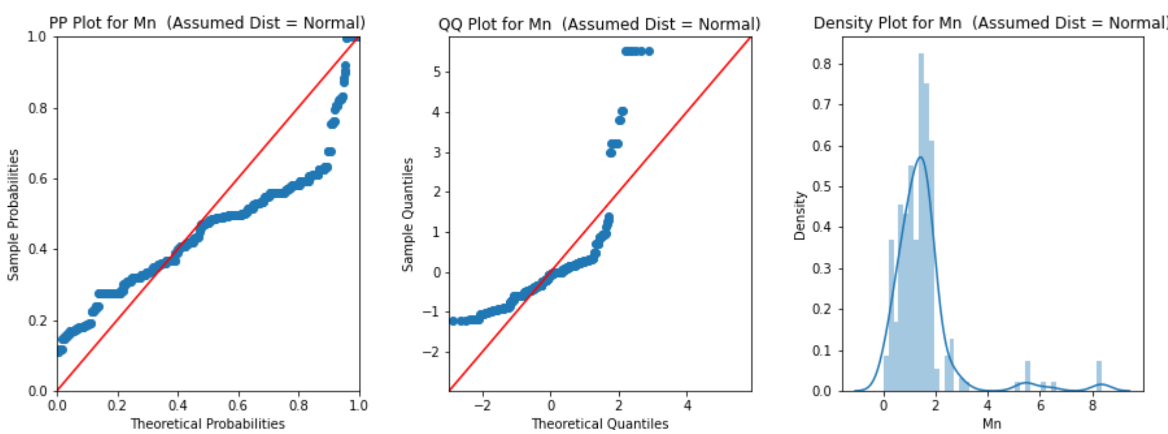
Wt % I



Observation:

- As we can see the 4 peaks in the density plot. It shows that the 4 kind of wt % I has been used in the experiment. That is in between 80-120, 160-190,270-280,290-320. As this is an experimental data so by having the different wt % of I, we might be interested to know the impact on other variables.

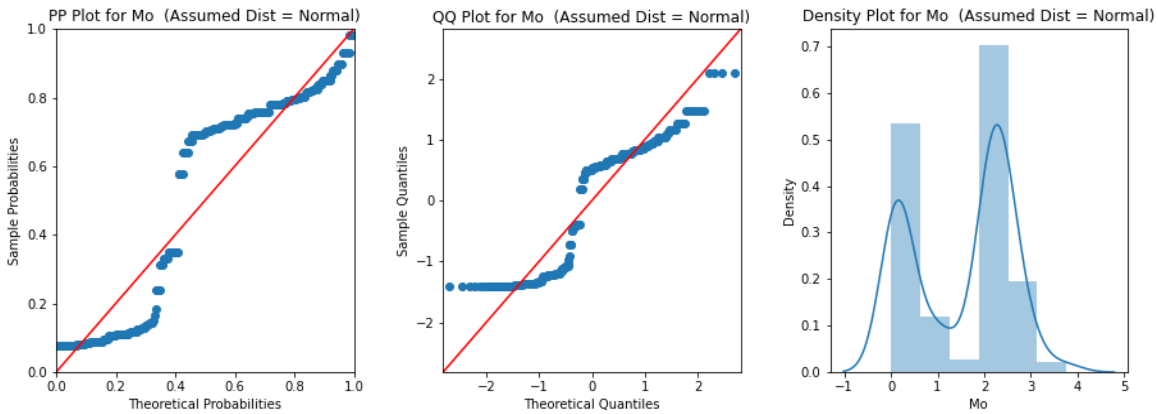
Wt % Mn



Observation:

- As we can see the 3 peaks in the density plot. It shows that the 3 kind of wt % Mn has been used in the experiment. That is in between 0-4, 4-7, 7-8. But majority of wt % lies in the range of 0-2 %. From the PP and QQ plot, I am assuming this as an approximately normal distribution

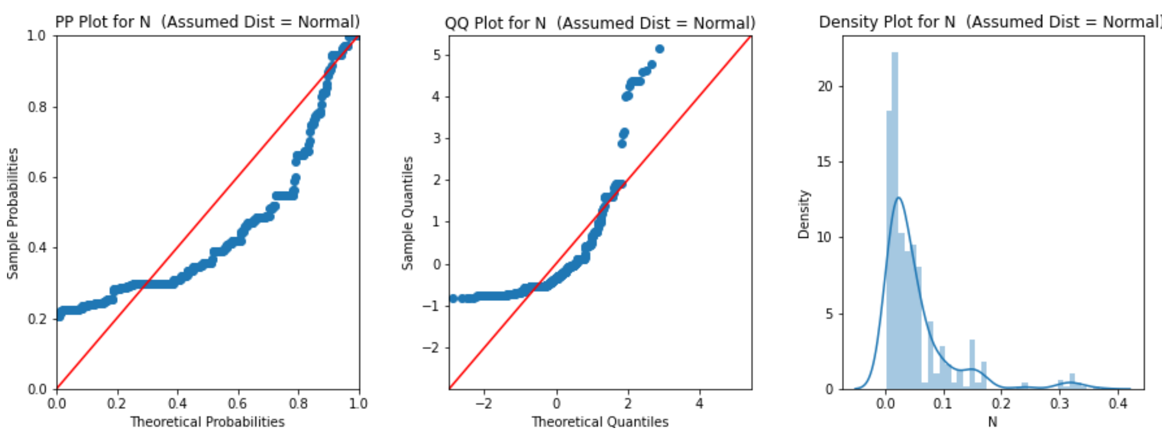
Wt % Mo



Observation:

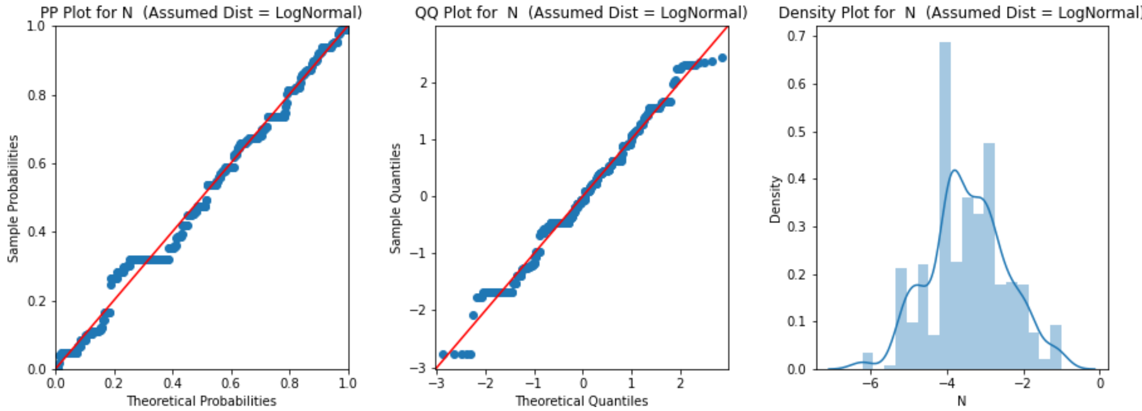
- As we can see the 2 peaks in the density plot. It shows that the 2 kind of wt % Mo has been used in the experiment. That is in between 0-1, 1-4. But majority of wt % lies in the range of 2-2.5 %. From the PP and QQ plot also, these 2 peaks can be seen.

Wt % N



Observation:

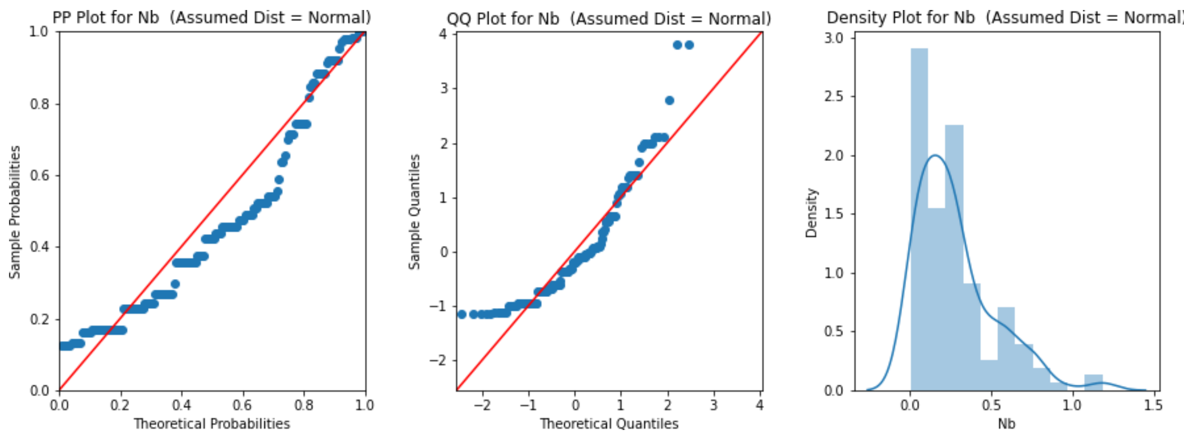
- As we can see the 2 peaks in the PP plot. It shows that the 2 kind of wt % N has been used in the experiment. That is in between 0-0.2, 0.3-0.4. But majority of wt % lies in the range of 0-0.1 %. In the below figure I am checking the distribution of my data by taking the assumed distribution as lognormal distribution.



Observation:

- As we can see that this data is almost fitting the lognormal distribution so I am assuming the distribution of N as lognormal distribution.

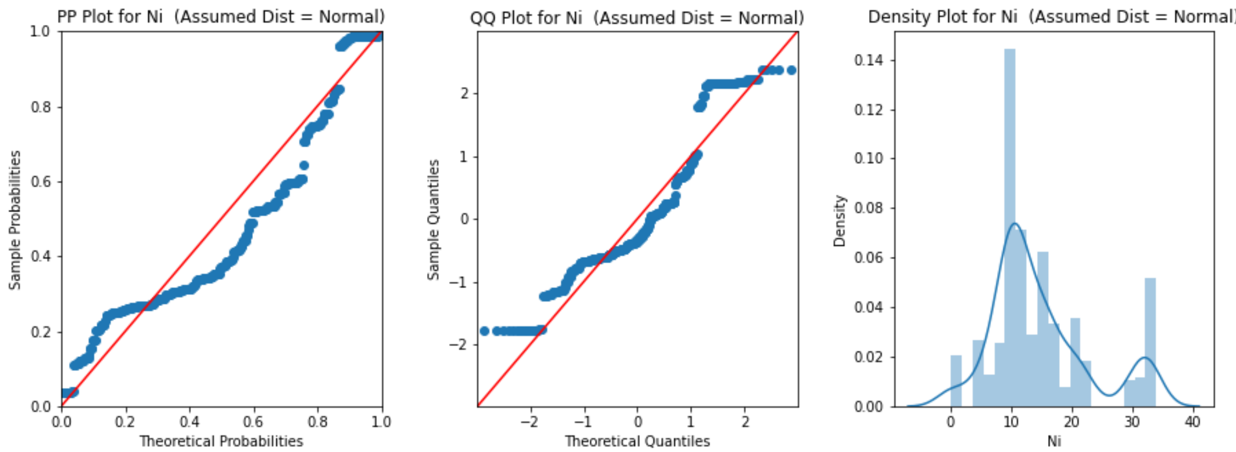
Wt % Nb



Observation:

- As we can see from the QQ plot that this data has 2 outliers. And this also can be seen in the density plot also. This data has little deviation from the normal distribution but I am assuming this to be normal distribution for further analysis.

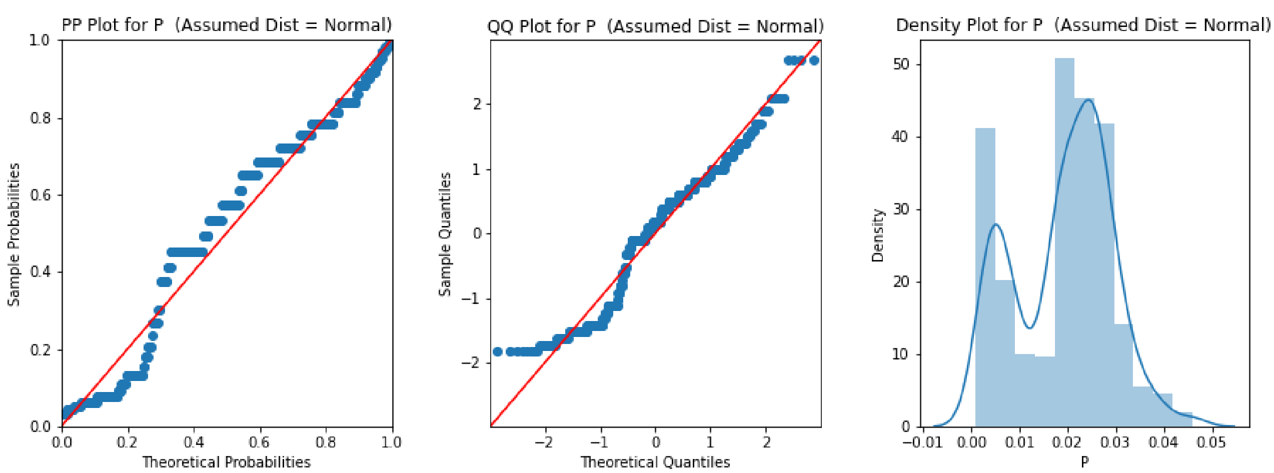
Wt % Ni



Observation:

- As we can see from the density plot that this data has 2 peaks. One is in between 0-22 while second is in between 29-34. It means that the 2 type of wt % Ni has been used. This data has little deviation from the normal distribution but I am assuming this to be normal distribution for further analysis.

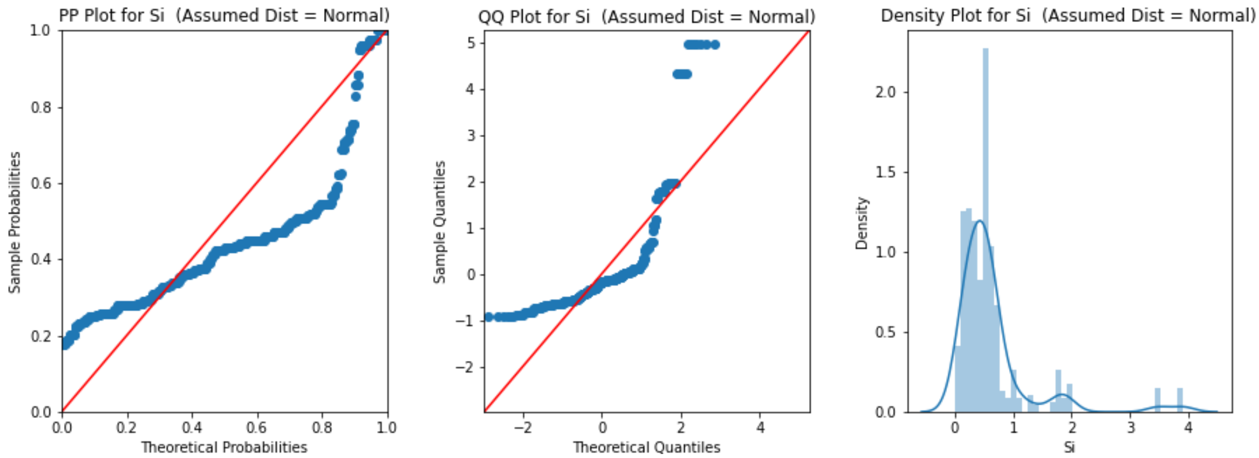
Wt % P



Observation:

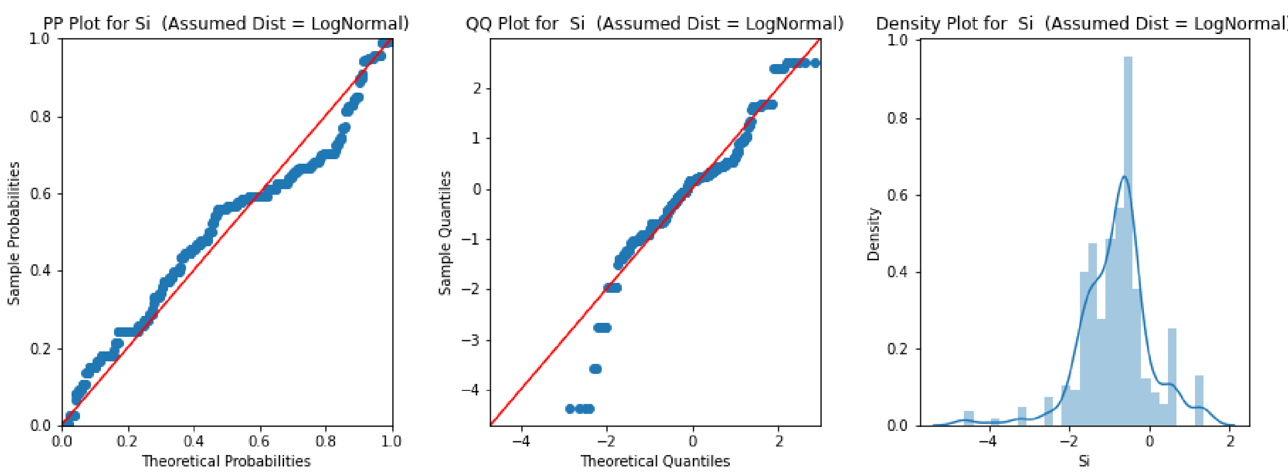
- As we can see from the PP & QQ plot that this data has 2 peaks. One is in between 0-0.01 while second is in between 0.02-0.03. It means that the 2 type of wt % P has been used. This data has little deviation from the normal distribution but I am assuming this to be normal distribution for further analysis.

Wt % Si



Observation:

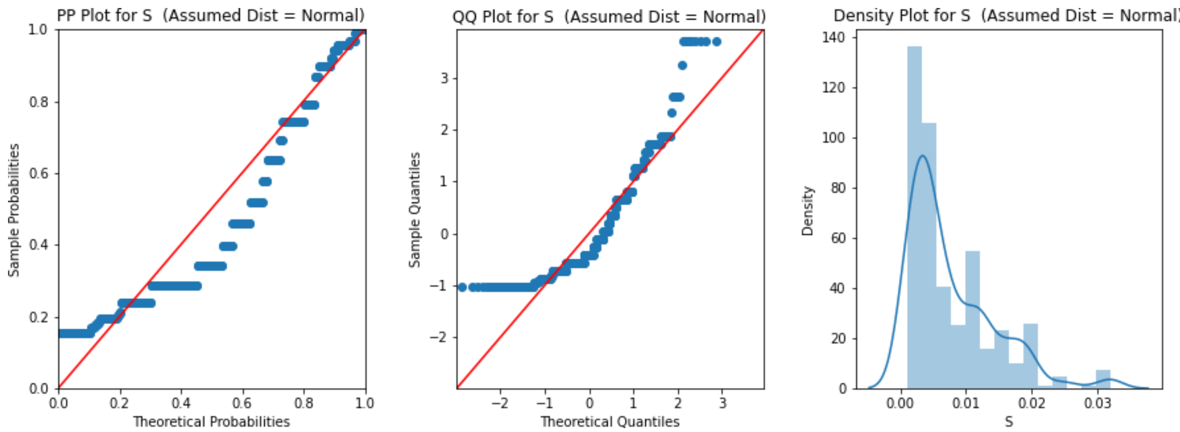
- As we can see from the PP & QQ plot that this data has 3 peaks. One is in between 0-1 while second is in between 1-2 and third is in between 3.5-4. It means that the 3 type of wt % Si has been used. This data has much deviation from the normal distribution so in below figure I am checking the distribution of wt % Si by taking the theoretical distribution as lognormal distribution



Observation:

- As we can see from the PP and QQ plot that this data is almost following the lognormal distribution. So I am assuming this to be lognormal distribution for my further analysis.

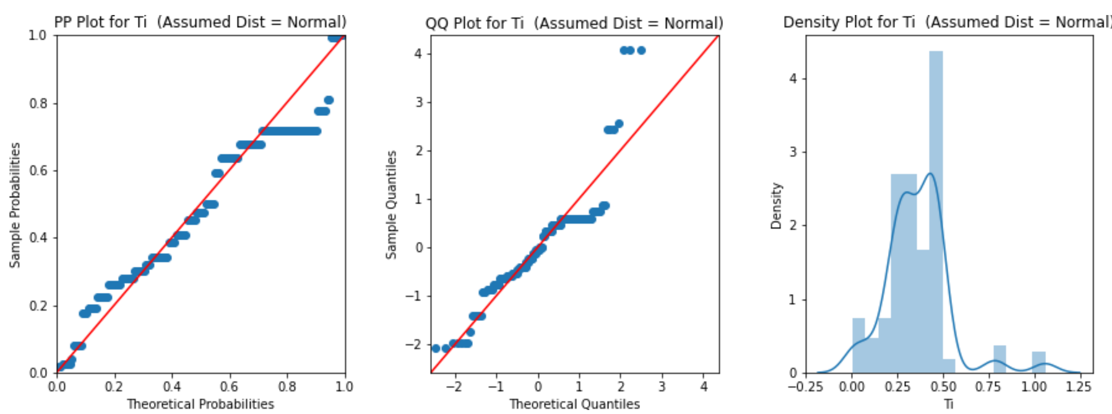
Wt % S



Observation:

- As we can see from the PP & QQ plot that this data has multiple peaks. It shows that the multiple type of wt % S has been used in the experiment. This data has deviation from the normal distribution but I am assuming for my further analysis that wt% S is normally distributed

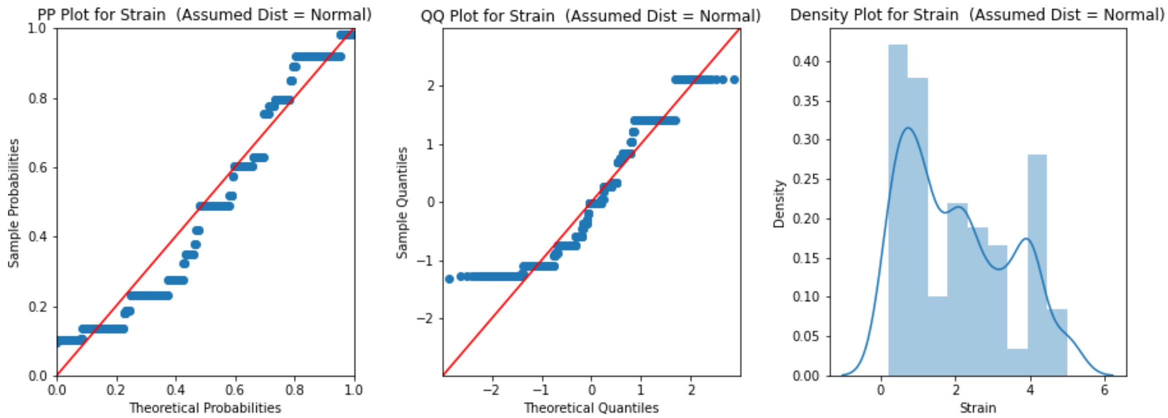
Wt % Ti



Observation:

- As we can see from the PP & QQ plot that this data has multiple peaks. It shows that the multiple type of wt % Ti has been used in the experiment.
- This data has deviation from the normal distribution but I am assuming for my further analysis that wt% Ti is normally distributed

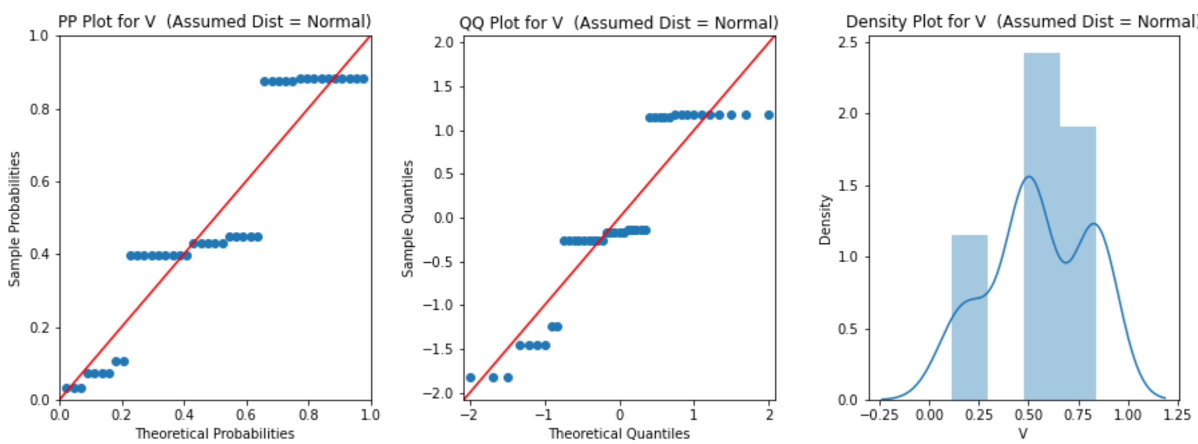
Strain



Observation:

- As we can see from the PP & QQ plot that this data has multiple peaks. It shows that the multiple type of strain has been applied in the experiment.
- This data has deviation from the normal distribution but still I am assuming for my further analysis that strain is normally distributed

Wt % V



Observation:

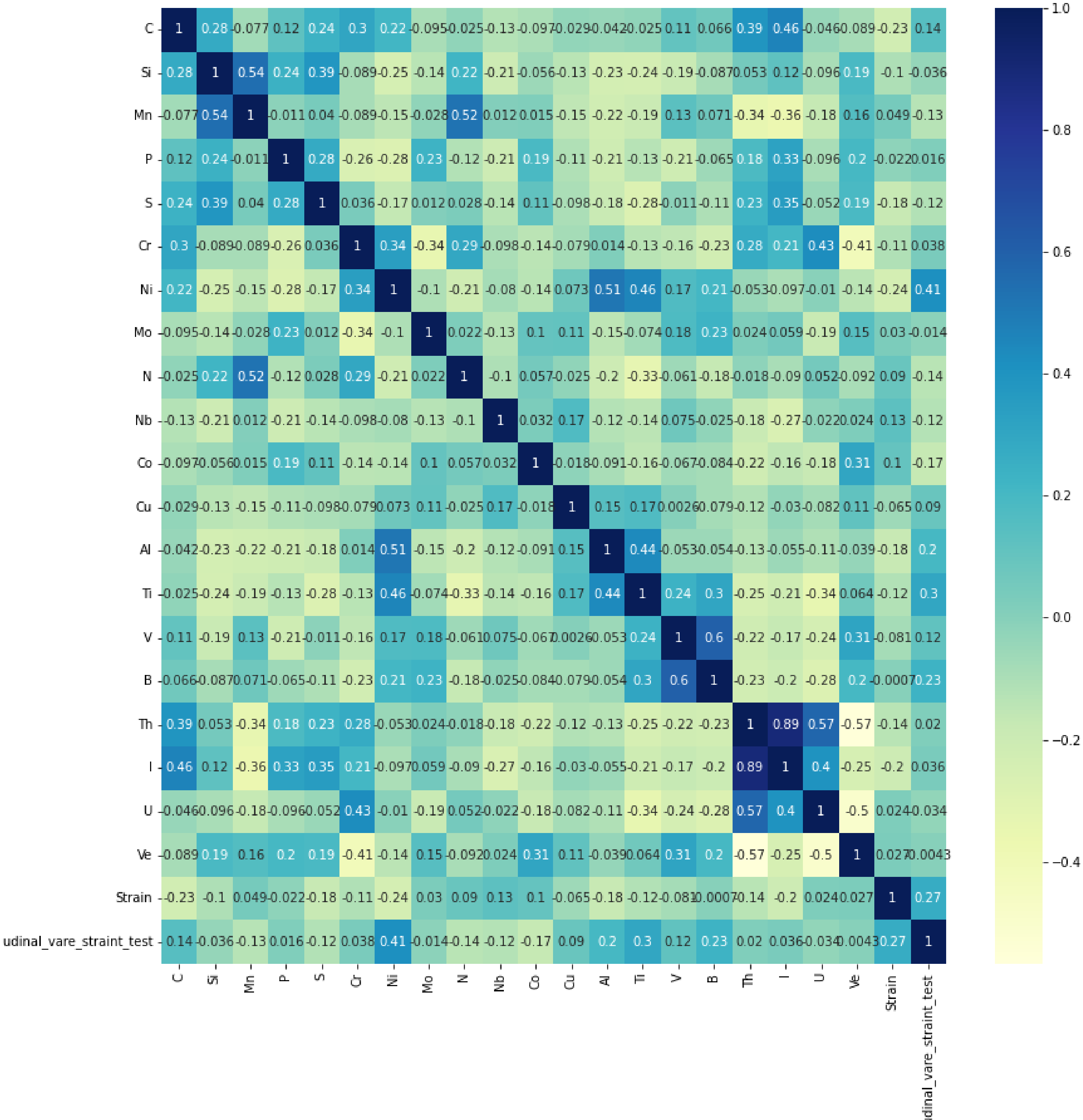
- As we can see from the PP & QQ plot that this data has 3 peaks. It shows that the 3 type of wt % V has been used in the experiment. Which are in the range of 0.15-0.25,0.5-0.6,0.6-0.77.
- This data has deviation from the standard normal distribution but I am assuming for my further analysis that wt% V is normally distributed.

Summary of Distributional Analysis:

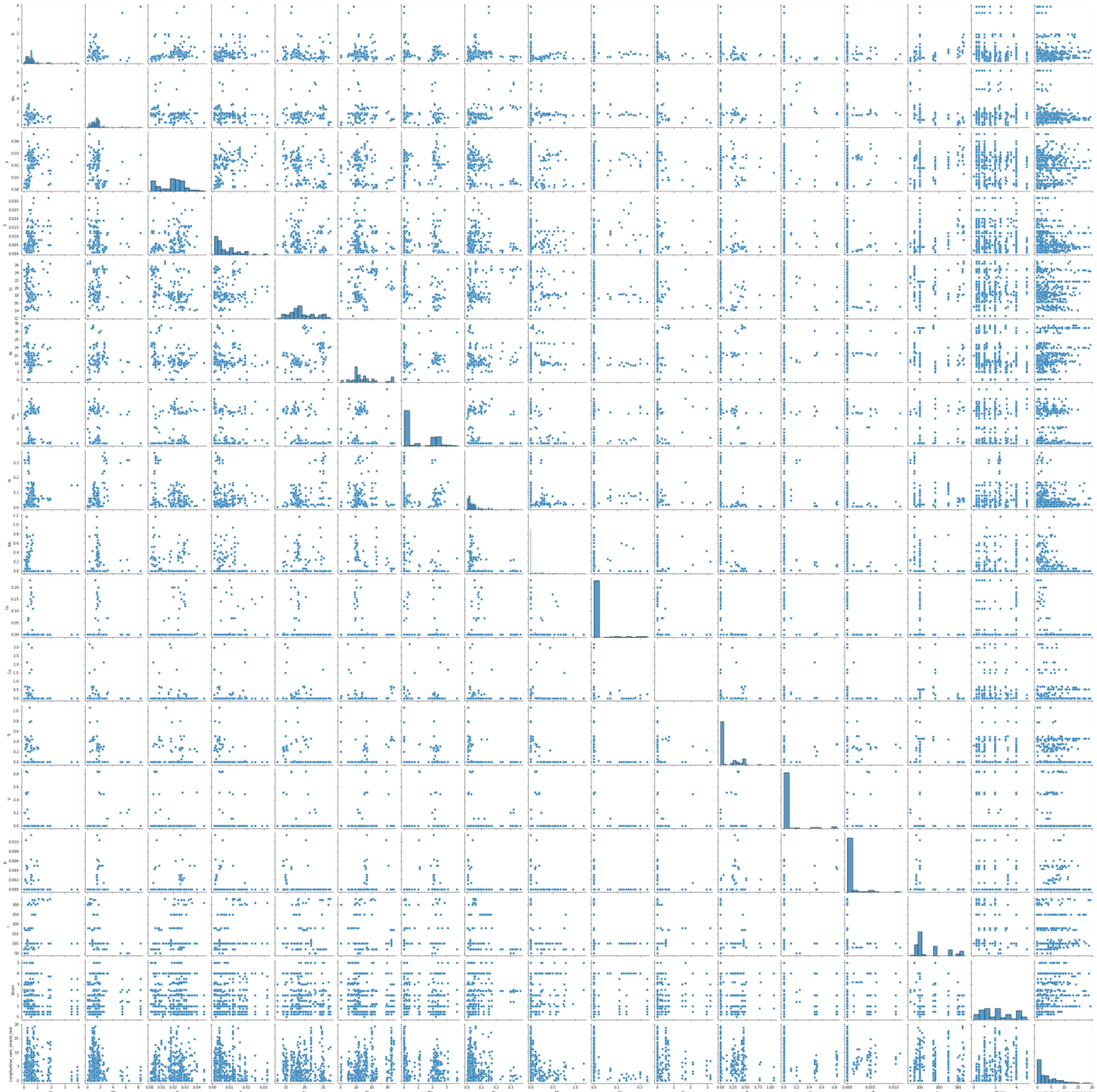
Sr. No.	Variable	Distribution
1	Si	Lognormal
2	Mn	Normal
3	P	Normal
4	S	Normal
5	Cr	Normal
6	Ni	Normal
7	Mo	Normal
8	N	Lognormal
9	Nb	Normal
10	Co	Normal
11	Cu	Lognormal
12	Ti	Normal
13	V	Normal
14	B	Lognormal
15	I	Normal
16	Strain	Normal
17	Longitudinal vare straint test (Target variable)	Lognormal

Linear Regression Analysis:

First of all I have drawn the heat map to find the correlation among all variables.

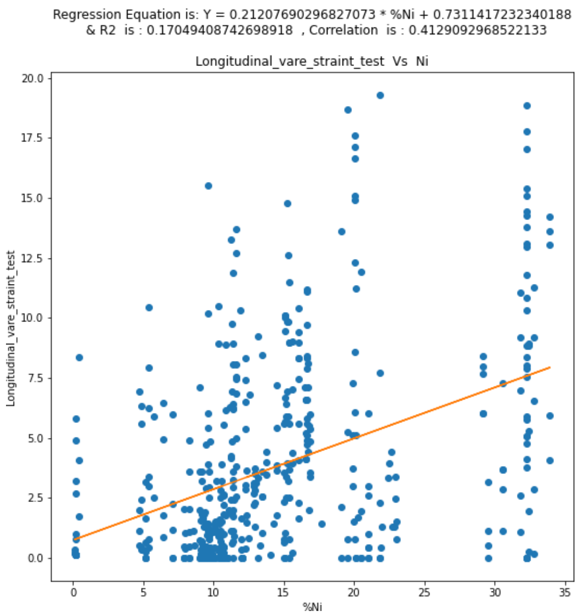


Now I have made the Pair plot to know the relation of all variables in graphical way.



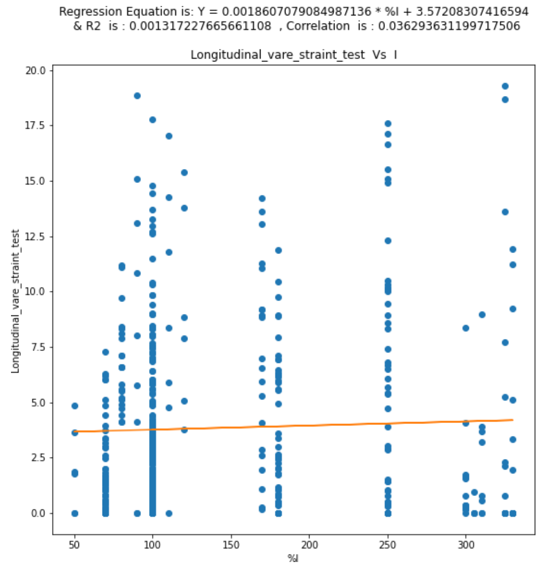
Observation: As we can see that the Ni has highest correlation. But my columns have zero values also, which should be removed. So in the linear regression plots I have removed the zero values from the columns and then I have calculated correlation coefficient and $R2$.

Now fit the linear regression equation in all independent variables and dependent variable to know the impact of each independent variable on dependent variable.



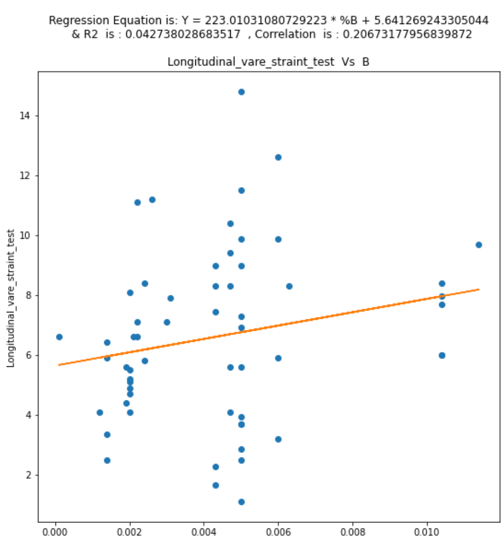
Observation: The graph between Longitudinal vare straint test vs % Ni shows that the the coefficient of correlation is 0.4129 and equation of line is:

Y = 0.212 * (% Ni) + 0.73



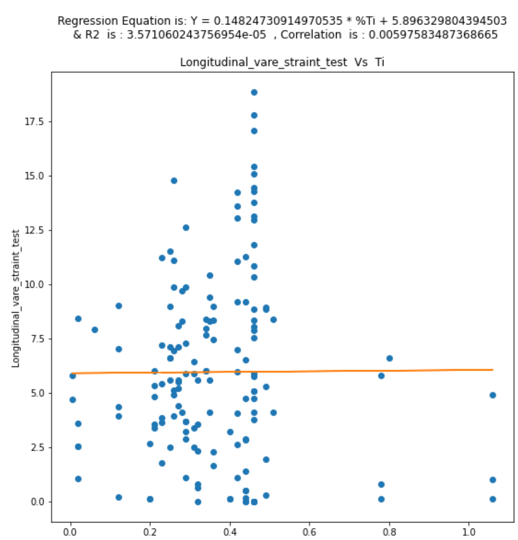
Observation: The graph between Longitudinal vare straint test vs % I shows that the the coefficient of correlation is 0.036 and equation of line is:

Y = 0.00186 * (% I) + 3.572



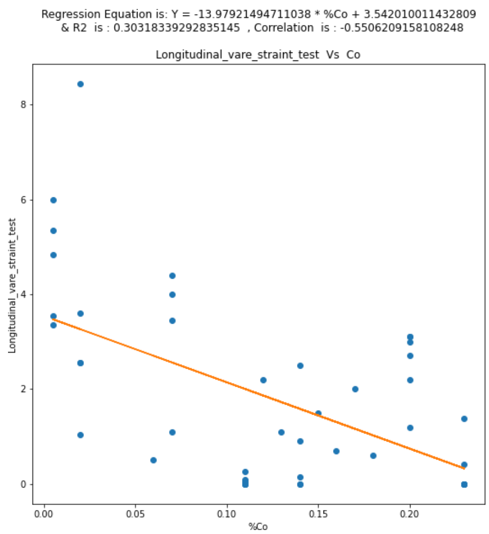
Observation: The graph between Longitudinal vare straint test vs % B shows that the the coefficient of correlation is 0.206 and equation of line is:

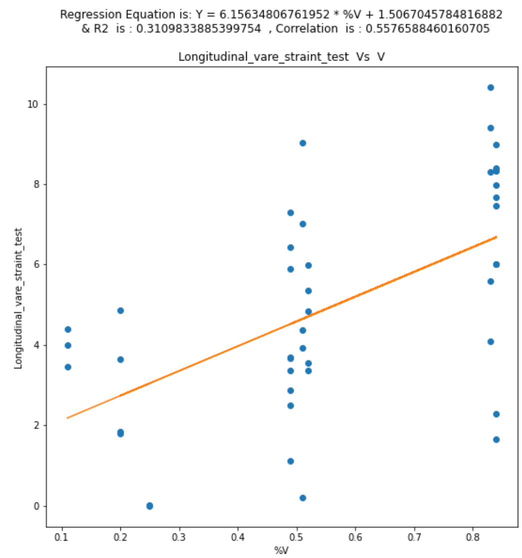
Y = 223.01 * (% B) + 5.64



Observation: The graph between Longitudinal vare straint test vs % Ti shows that the the coefficient of correlation is 0.0059 and equation of line is:

Y = 0.148 * (% Ti) + 5.89



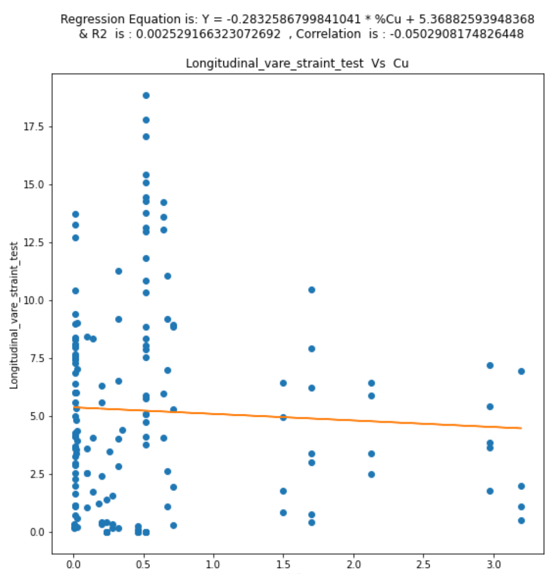


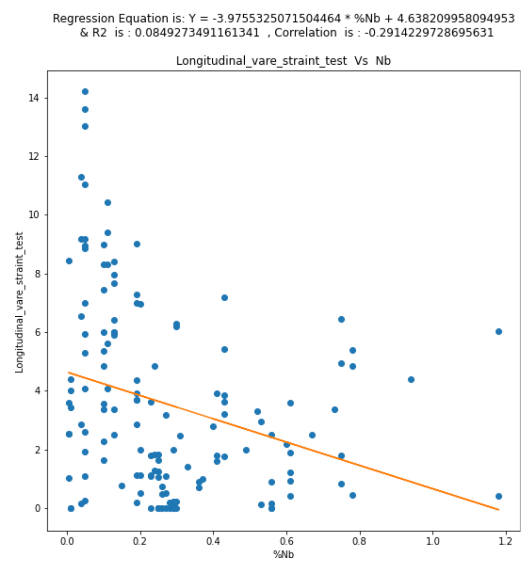
Observation: The graph between Longitudinal vare straint test vs % Co shows that the the coefficient of correlation is -0.55 and equation of line is:

Y = -13.98 * (% Co) + 3.54

Observation: The graph between Longitudinal vare straint test vs % V shows that the the coefficient of correlation is 0.56 and equation of line is:

Y = 6.156 * (% V) + 1.506



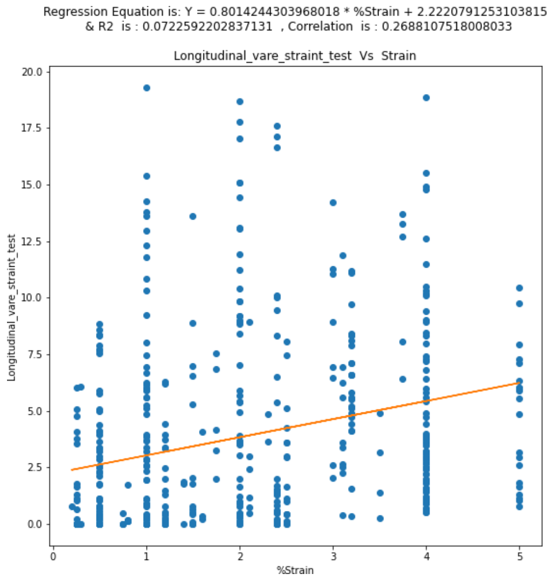


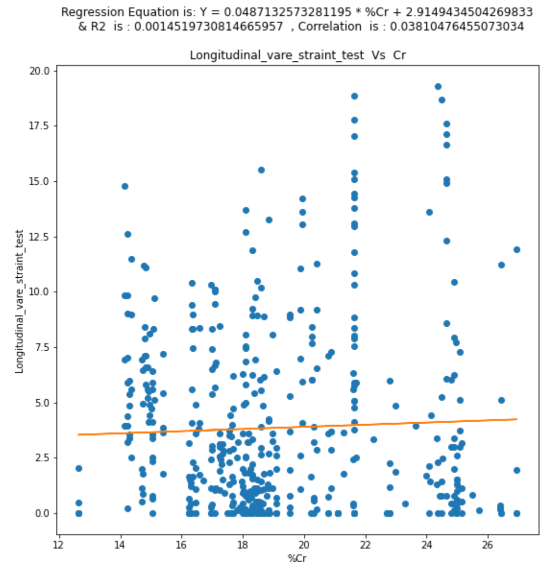
Observation: The graph between Longitudinal vare straint test vs % Cu shows that the the coefficient of correlation is -0.05 and equation of line is:

Y = -0.28 * (% Cu) + 3.54

Observation: The graph between Longitudinal vare straint test vs % Nb shows that the the coefficient of correlation is -0.29 and equation of line is:

Y = -3.97 * (% Nb) + 4.638



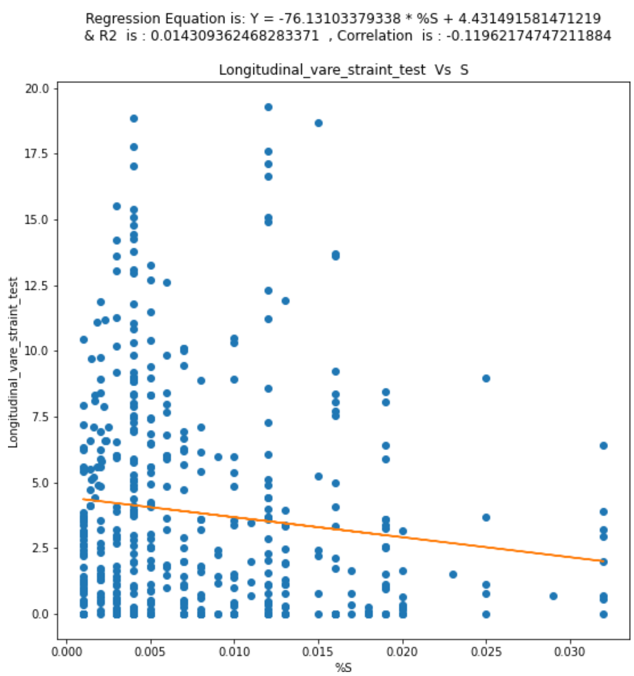


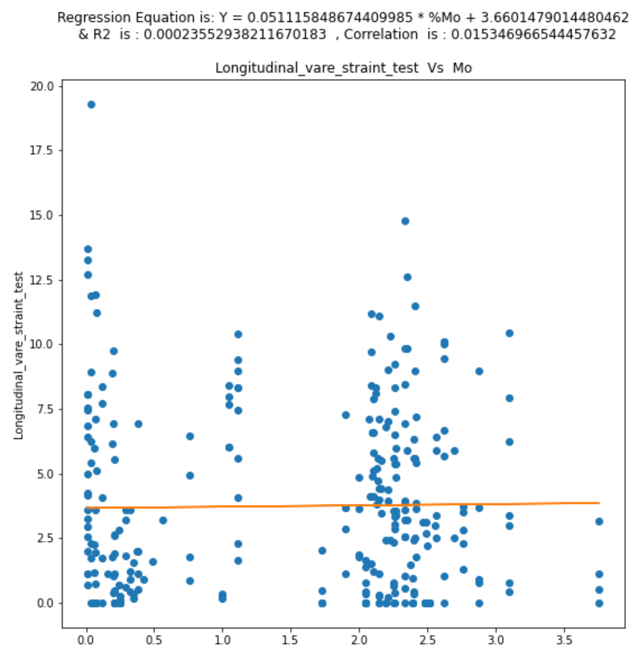
Observation: The graph between Longitudinal vare straint test vs Strain shows that the the coefficient of correlation is 0.27 and equation of line is:

Y = 0.8 * (Strain) + 2.22

Observation: The graph between Longitudinal vare straint test vs % Cr shows that the the coefficient of correlation is 0.038 and equation of line is:

Y = 0.048 * (% Cr) + 2.91



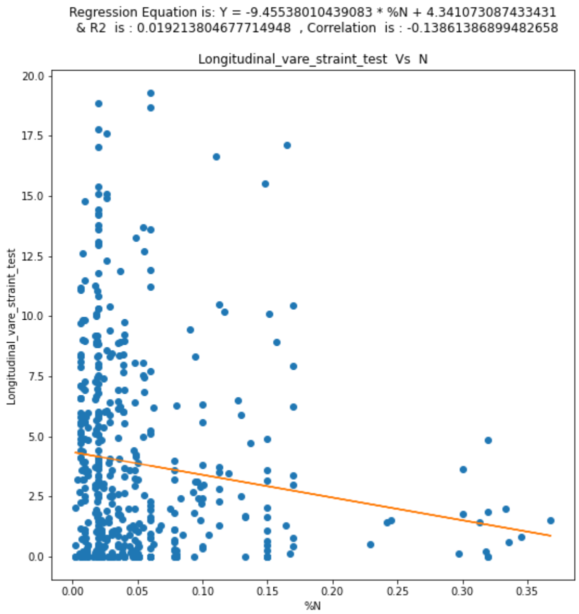


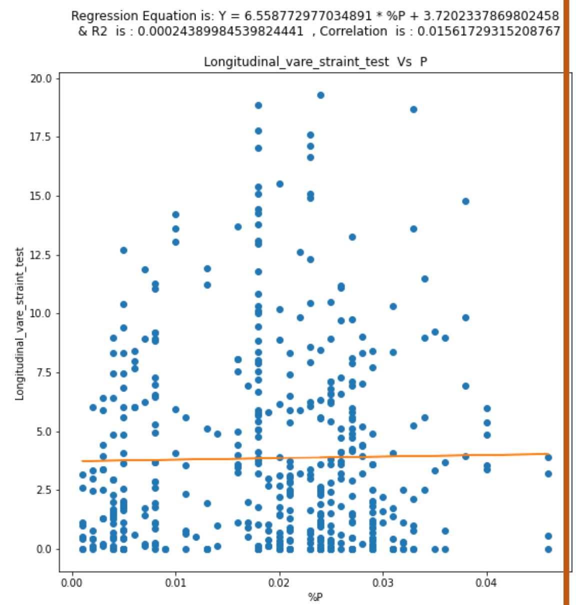
Observation: The graph between Longitudinal vare straint test vs % S shows that the the coefficient of correlation is -0.12 and equation of line is:

Y = -76.13 * (% S) + 4.43

Observation: The graph between Longitudinal vare straint test vs % Mo shows that the the coefficient of correlation is 0.015 and equation of line is:

Y = 0.05 * (% Mo) + 3.66



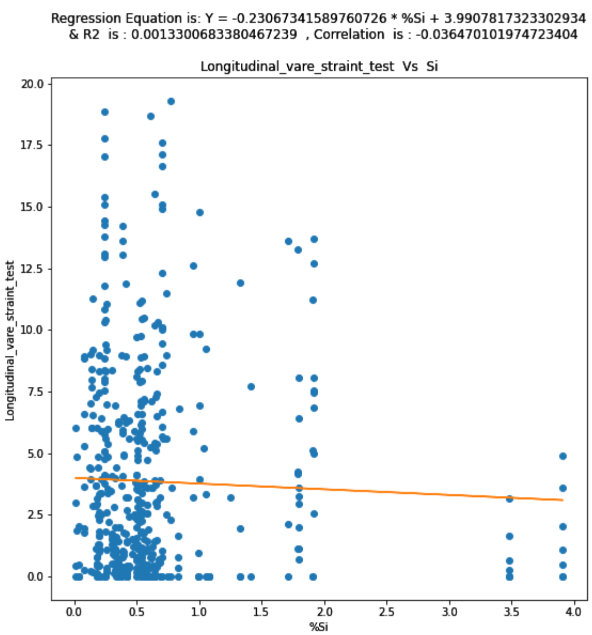


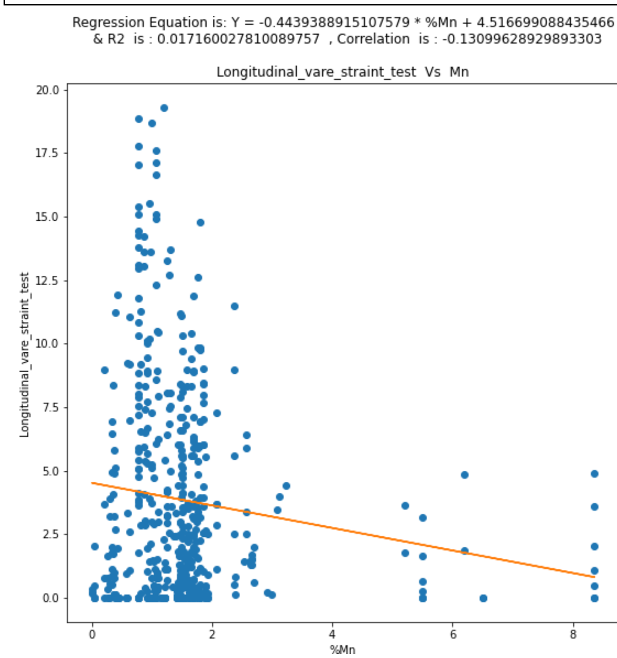
Observation: The graph between Longitudinal vare straint test vs % N shows that the the coefficient of correlation is -0.14 and equation of line is:

Y = -9.45 * (% N) + 4.34

Observation: The graph between Longitudinal vare straint test vs % P shows that the the coefficient of correlation is 0.016 and equation of line is:

Y = 6.56 * (% P) + 3.72





Observation: The graph between Longitudinal vare straint test vs % Si shows that the the coefficient of correlation is -0.036 and equation of line is:

Y = -0.23 * (% Si) + 3.99

Observation: The graph between Longitudinal vare straint test vs % Mn shows that the the coefficient of correlation is -0.13 and equation of line is:

Y = -0.44 * (% Mn) + 4.51

Hypothesis Testing

- Hypothesis testing is done when the population mean is unknown or we can say that hypothesis testing is used to determine whether a statement about the value of a population parameter should be rejected or not.
- The null hypothesis, denoted by H_o is a tentative assumption about a population parameter while the alternative hypothesis denoted by H_a is the opposite of what is stated in the null hypothesis.
- The hypothesis testing uses data from a sample to test the two competing statements indicated by H_o and $H_{a.}$ but in my data (487 rows and 22 columns) population parameters are known because this is an experimental data ,so I am performing the hypothesis testing by taking the 5 random sample of sample size = 50 and then I am taking mean of all five sample mean as sample mean.
- As in this part of submission we were supposed to do Hypothesis testing and Regression analysis. As I did my regression analysis in the EDA so here I am performing hypothesis testing for all variables.

After eliminating the less important columns in EDA, my data look like as:

	Si	Mn	P	S	Cr	Ni	Mo	N	Nb	Co	Cu	Ti	V	B	l	Strain	Longitudinal_vare_straint_test
0	0.48	1.61	0.024	0.019	17.33	10.62	2.09	0.060	0.0	0.00	0.0	0.0	0.0	0.0	100	4.0	1.5
1	0.58	1.06	0.032	0.013	16.95	10.50	2.15	0.078	0.0	0.00	0.0	0.0	0.0	0.0	100	4.0	1.1
2	0.46	1.09	0.021	0.001	17.40	11.50	2.88	0.105	0.0	0.00	0.0	0.0	0.0	0.0	100	4.0	0.9
3	0.51	1.60	0.021	0.001	17.55	12.95	2.76	0.113	0.0	0.00	0.0	0.0	0.0	0.0	100	4.0	3.7
4	0.46	1.54	0.027	0.023	16.28	10.15	2.06	0.098	0.0	0.15	0.0	0.0	0.0	0.0	100	4.0	1.5

I am performing Two-Tailed Tests for population mean and explaining the hypothesis as:

H_o: Assumed Population mean = Sample mean

$H_{a:}$ Assumed population mean is not equal to sample mean

Here for my data, actual population mean is known so I am taking actual population mean as Assumed population mean.

I have taken the level of significance (alpha) = 0.05 for my hypothesis testing.

I have calculated the Z value using the formula:

$$z = (x - \mu) / (\sigma / \sqrt{n})$$

Then I calculated the corresponding p-value and I compared with the alpha/2 and p-value/2.

If p-value/2 >alpha/2 then accept the null hypothesis and if p-value/2 <alpha/2 then reject the null hypothesis.

I have performed these all calculations for all the independent variables using python. Link of python code is attached on the top of this report.

Elements	Mu	Sigma	Xbar	Z	half_p_value	alpha/2_value
Si	0.6278	0.6616	0.6381	0.0976	0.4611	0.025
Mn	1.5109	1.2347	1.5814	0.3613	0.3589	0.025
P	0.0192	0.0100	0.0190	-0.0783	0.5312	0.025
S	0.0077	0.0066	0.0082	0.4471	0.3274	0.025
Cr	19.1121	3.2731	18.8292	-0.5467	0.7077	0.025
Ni	14.6872	8.1469	15.5180	0.6449	0.2595	0.025
Mo	1.5205	1.0652	1.4585	-0.3687	0.6438	0.025
N	0.0524	0.0613	0.0585	0.6308	0.2641	0.025
Nb	0.2761	0.2375	0.2462	-0.7970	0.7873	0.025
Co	0.1229	0.0757	0.1271	0.3458	0.3648	0.025
Cu	0.5607	0.7845	0.5725	0.0949	0.4622	0.025
Ti	0.3603	0.1720	0.3626	0.0850	0.4661	0.025
V	0.5521	0.2425	0.5629	0.2819	0.3890	0.025
B	0.0043	0.0026	0.0043	-0.0619	0.5247	0.025
I	147.1869	81.6180	145.6250	-0.1210	0.5482	0.025
Strain	2.0361	1.4037	1.9910	-0.2032	0.5805	0.025
Longitudinal_vare_straint_test	4.6246	4.1777	4.6364	0.0178	0.4929	0.025

Here it can be clearly seen that for all the variables, half p-value is greater than half alpha-value so do not reject the null hypothesis. i.e for all variables, population mean is equal to the assumed population mean.

Conclusion:

In this study we have done Exploratory Data Analysis of the solidification crack susceptibility data to understand the data by making Histograms, Box & Whisker plots and Stem and leaf plots. Then we identified the distribution of each variable and fitted Linear Regression lines to predict the solidification cracking susceptibility for any existing and new grade of steels. Then we performed the Hypothesis testing by taking level of significance as 0.05. I also found that the **V is the most** important variable which influence the solidification cracking susceptibility, and after this **Ni is the 2nd most** important variable which affects the most. . I found the multiple linear regression equation to predict the solidification crack susceptibility of any new material. The equation is :

Y = -3.2536 + 1.5159 * (% Si) -0.7511 * (% Mn) + 74.3898 * (% P) -39.36 * (% S) -0.105 * (% Cr) + 0.2828 * (% Ni) - 0.2416 * (% Mo) + 6.9841 * (% N) - 1.17 * (% Nb) - 12.24 * (% Co) + 0.82 * (% Cu) + 1.2 * (% Ti) + 2.76 * (% V) + 253.19 * (% B) + 0.0047 * (% I) + 1.41 * Strain

Drive Link for Code:
https://drive.google.com/drive/folders/1VsM086vovgkgGDCH-Q6Xfa8G7qawD1GK?usp=sharing