Springboard Data Science Career Track
Capstone Project 1
Charles Tremblay

Project Milestone Report

1. Problem

Superconductors offer extraordinary electrical conductivity with zero resistance. They are very useful in applications that require strong magnetic fields or long distance power transmission.The benefits of superconductors are clear, but are restricted by their very low critical temperature as it is far too low to be used in a real world setting. In addition, superconductivity is not a well understood phenomenon, which makes it difficult to create a superconductor with a desired critical temperature. To aid researchers study superconductivity, a model can be developed from material properties of existing superconductors with the goal of predicting its critical temperature. This can help in the study and development of new superconductors by offering insight on what properties of the material can influence its critical temperature.

2. Client

The prospective clients would be researchers who want to develop new superconducting material and better understand superconductivity. The developed model can help them analyze their data and focus their time on more promising materials. Companies developing technology that utilizes superconductors could also benefit from this model. For example, MRI machines, particle accelerators, and digital circuits. The benefit of this model could help companies develop technologies in a more affordable and scalable fashion by choosing the appropriate material.

3. Dataset

The dataset was retrieved from the UCI machine learning repository. The data set contains two files; one detailing the chemical composition of each unique superconductor, the other containing the material properties and critical temperature of the superconductor. Overall, there are 21,264 superconductors with 82 attributes in the dataset. Link: http://archive.ics.uci.edu/ml/datasets/Superconductivty+Data

4. Steps to Solve

The first step will be to perform data wrangling and EDA for exploration of data. Then I will apply statistical methods for hypothesis testing. Finally. I'll use machine learning to develop models that will be able to predict the critical temperature of a given superconductor.

5. Deliverables:

Code, report, and a slide deck for the Github repository.

Data Wrangling

The data for this project was retrieved from the UCI Machine Learning repository. It contains superconductor data such as atomic weight, thermal conductivity, density, entropy, and other properties that will be used to predict critical temperature. The data is contained in a comma separated value file that is opened by default in Excel. To access the data, Jupyter notebook was created and the pandas library was imported to read in the csv file and to store it into a pandas dataframe.

The dataframe contains 82 columns, and 21,263 rows. The cleanliness of the data was checked by using the .info() method which brought back the data type for each column, along with the number of non-null values. The majority of the data was of type float64, while a couple were of type int64. However, none of the columns contained any missing, or null values. Although there are no missing values, a duplicate value check was performed. For this, the .drop_duplicate() method was applied to the dataframe. The resulting dataframe was of the same size as the original, which means that no duplicates were found. The imported dataset was already clean in terms of missing values and duplicates, and no missing value replacement methods were needed. A screenshot of the results from the .info() method is provided below.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 21263 entries, 0 to 21262
Data columns (total 82 columns):
number_of_elements                21263 non-null int64
mean_atomic_mass                  21263 non-null float64
wtd_mean_atomic_mass              21263 non-null float64
gmean_atomic_mass                 21263 non-null float64
wtd_gmean_atomic_mass             21263 non-null float64
entropy_atomic_mass               21263 non-null float64
wtd_entropy_atomic_mass           21263 non-null float64
range_atomic_mass                 21263 non-null float64
wtd_range_atomic_mass             21263 non-null float64
std_atomic_mass                   21263 non-null float64
wtd_std_atomic_mass               21263 non-null float64
mean_fie                          21263 non-null float64
wtd_mean_fie                      21263 non-null float64
gmean_fie                         21263 non-null float64
wtd_gmean_fie                     21263 non-null float64
```

Figure 1: Screenshot of the .info() method performed on the dataframe

With the data already clean, preliminary exploratory data analysis (EDA) was performed. This was done by computing basic statistics of the data, and finally by visualizing the distribution of each variable. The first EDA step was to find statistics of the dataframe using the .describe() method. The mean, median, mode, and other statistical parameters were computed for each column. A small table containing the descriptive statistics for the first few columns is represented in Figure 2 below.

	number_of_elements	mean_atomic_mass	wtd_mean_atomic_mass	gmean_atomic_mass	wtd_gmean_atomic_mass	entropy_atomic_mass
count	21263.000000	21263.000000	21263.000000	21263.000000	21263.000000	21263.000000
mean	4.115224	87.557631	72.988310	71.290627	58.539916	1.165608
std	1.439295	29.676497	33.490406	31.030272	36.651067	0.364930
min	1.000000	6.941000	6.423452	5.320573	1.960849	0.000000
25%	3.000000	72.458076	52.143839	58.041225	35.248990	0.966676
50%	4.000000	84.922750	60.696571	66.361592	39.918385	1.199541
75%	5.000000	100.404410	86.103540	78.116681	73.113234	1.444537
max	9.000000	208.980400	208.980400	208.980400	208.980400	1.983797

Figure 2: Screenshot of descriptive statistics table of the superconductor data set.

Finally, the last step of the data wrangling project was to visualize the distribution of the variables. This was done by generating histograms for each column. The plt.hist() function was used along with the arguments set to equal to the dataframe columns and the bins argument set to 50. The resulting histograms are included in Figure 3.

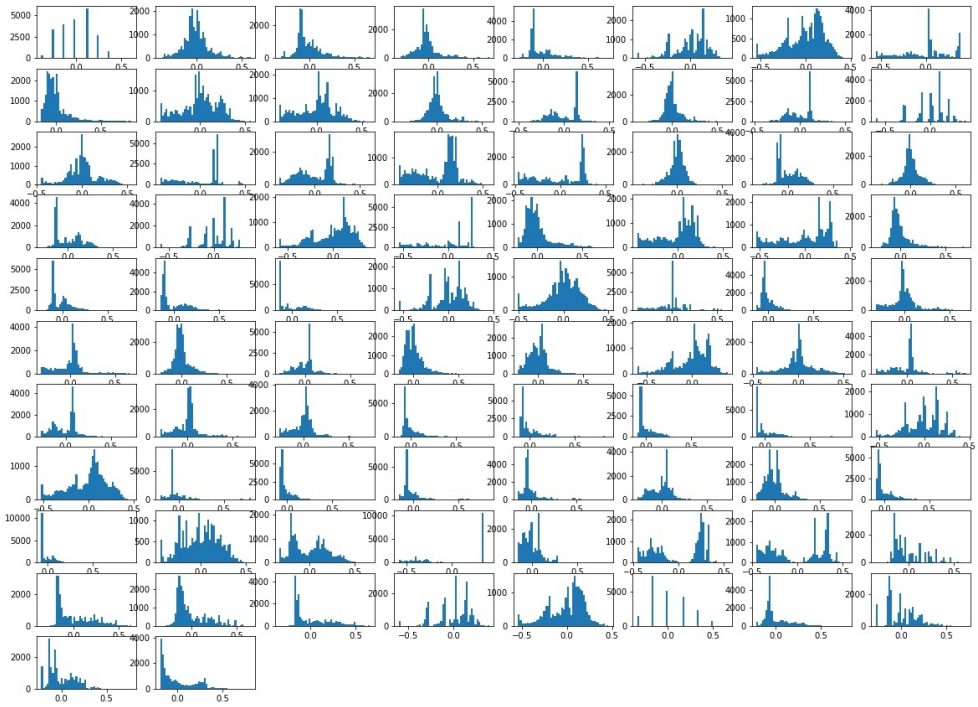

Figure 3: Visualization of each variable in the form of a histogram

The generated histograms display the distribution for each variable. From inspection, it appears that some variables follow a normal distribution, while others show skewing in either the right or left direction. In addition, we can see how clean the data is as there are no extreme discontinuity in the plots. However, It is too early in the project to make any definitive statements about the data, but this will be helpful when starting the formal exploratory data analysis project.

Exploratory Data Analysis and Statistical Inference

With the data processed, the next step is to perform exploratory data analysis. Four questions were asked to help guide the analysis.

1. What is the most common element in superconductors?
2. Is there a link between certain elements and high critical temperatures?
3. Are any variables correlated with critical temperature?
4. How are each variable distributed?

The first question asks which element is the most common in the superconductors. The data was manipulated to create a new data frame that counted the number of unique superconductors for each element.

The result is shown in Figure 4 below.

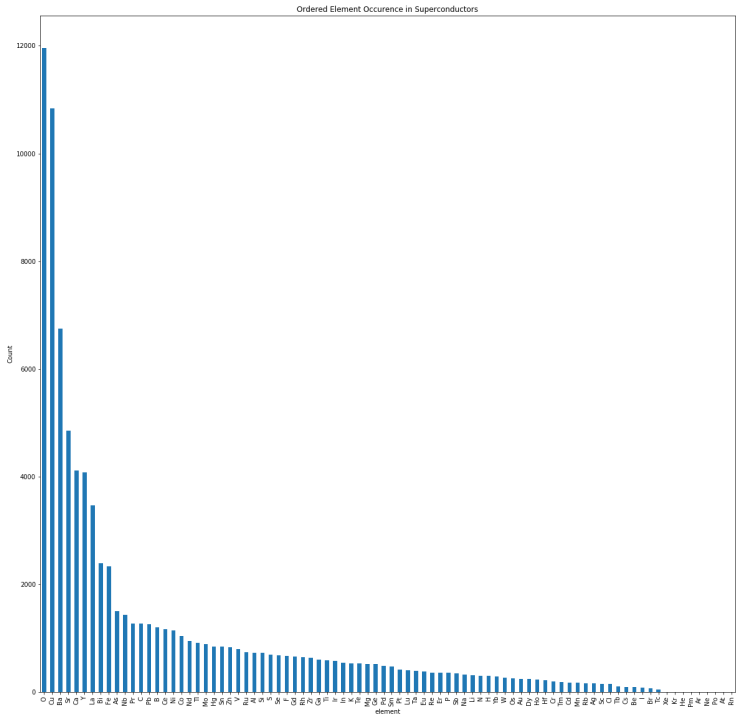

Figure 4: Bar graph of the top ten most common elements in superconductors.

From the figure, the top ten most common elements were found to be the following:

1. Oxygen
2. Copper

3. Barium
4. Strontium
5. Calcium
6. Yttrium
7. Lanthanum
8. Bismuth
9. Iron
10. Arsenic

While this result does give a good representation of the overall compositions of the superconductors, the primary objective of the project is to predict the highest critical temperature. Therefore, it is more important to determine which elements are most common in superconductors with a higher than normal critical temperature. With that said, a new figure was generated to showcase the top ten elements in superconductors that are above the 80th percentile in terms of critical temperature.

The result is shown below in Figure 5.

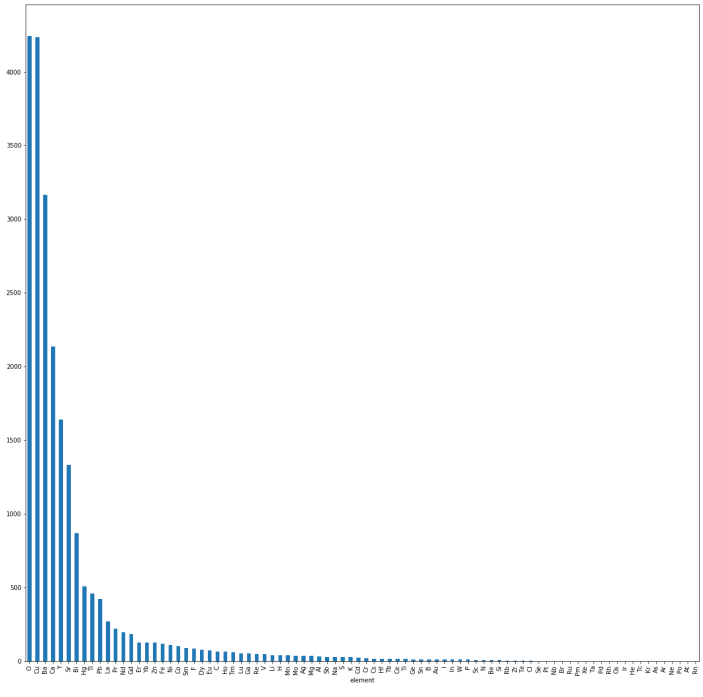

Figure 5: Bar graph of the top ten elements present in superconductors above the 80th percentile mark in terms of critical temperature

From the figure, the top ten most common elements in this class were found to be the following:

1. Oxygen
2. Copper
3. Barium
4. Calcium
5. Yttrium
6. Strontium
7. Bismuth
8. Mercury
9. Titanium
10. Lead

The last question identified ten elements that were present in superconductors that had a critical temperature above the 80 percentile. This next step will investigate further on the significance of these elements and superconductor features by applying statistical inference techniques and visualizations.

From the last question the top ten prevalent elements were as follows: Oxygen, copper, barium, calcium, yttrium, strontium, bismuth, mercury, titanium, lead.

The critical temperature distribution for each element was developed in the form of a histogram and are displayed in Figure 6 below. A table of summary statistics is also included in Table 1 showcasing the mean, standard deviation, and the count of superconductors that contained each element.

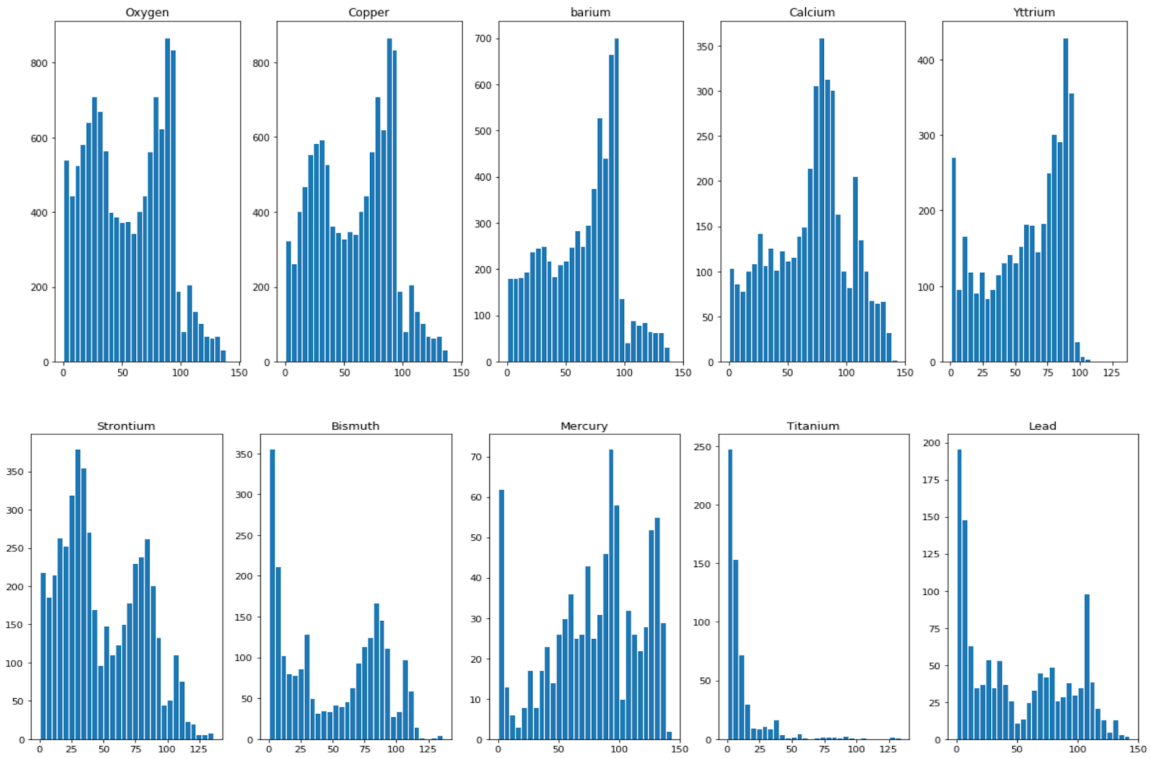

Figure 6: DIstribution of Elements and Critical Temperature

Table 1. Summary Statistics of Top Ten elements present in Superconductors with High Critical Temperature.

Element	Mean (℃)	STD	25 quartile	75 quartile	IQR	Count
Oxygen	54.67	32.90	25.5	83.3	57.80	11,964
Copper	58.30	32.06	29.7	85.0	55.33	10,838
Barium	64.31	31.48	38.0	88.8	50.75	6,751
Calcium	70.00	32.82	45.3	90.5	45.20	4,112
Yttrium	57.44	29.79	33.7	84.0	50.35	4,075
Strontium	48.91	31.16	24.0	76.7	52.70	4,852
Bismuth	49.67	37.19	10.3	84.0	73.74	2,389
Mercury	79.79	38.52	55.0	110.0	55.00	845
Titanium	11.32	17.59	2.84	11.0	8.16	589
Lead	49.77	41.00	8.01	87.3	79.31	1,255

Mercury is the most prevalent element in high temperature superconductors with a mean critical temperature of 79.79 ℃. The second highest element is calcium with a mean temperature of 70.00 ℃. Although oxygen and copper are present in the majority of the samples, their mean temperature is relatively low.

To test if the distributions followed a normal distribution, the shapiro test was performed. This was done by importing the scipy.stats module and sending each column to the function. However, the result showed that none of them were normal. This may be an issue when trying to use linear regression to develop a model. In addition, outliers were identified using the interquartile range method. The result yielded only outliers for titanium.

The second part of the project involved investigating the superconductor features forwith the goal of finding any strong correlations between each feature and the target variable (critical temperature). An initial heat map that was generated for the data story report was hard to read due to the large amount of features. This effectively diluted the matrix and made it hard to interpret the colour map. To prevent this, the feature data was cleaned by removing columns that contained the range and sample deviation of other columns. The result was a shorter dataframe with relevant features that would make physical sense when correlating it with critical temperature. A new heatmap was developed and is shown below in Figure 7.

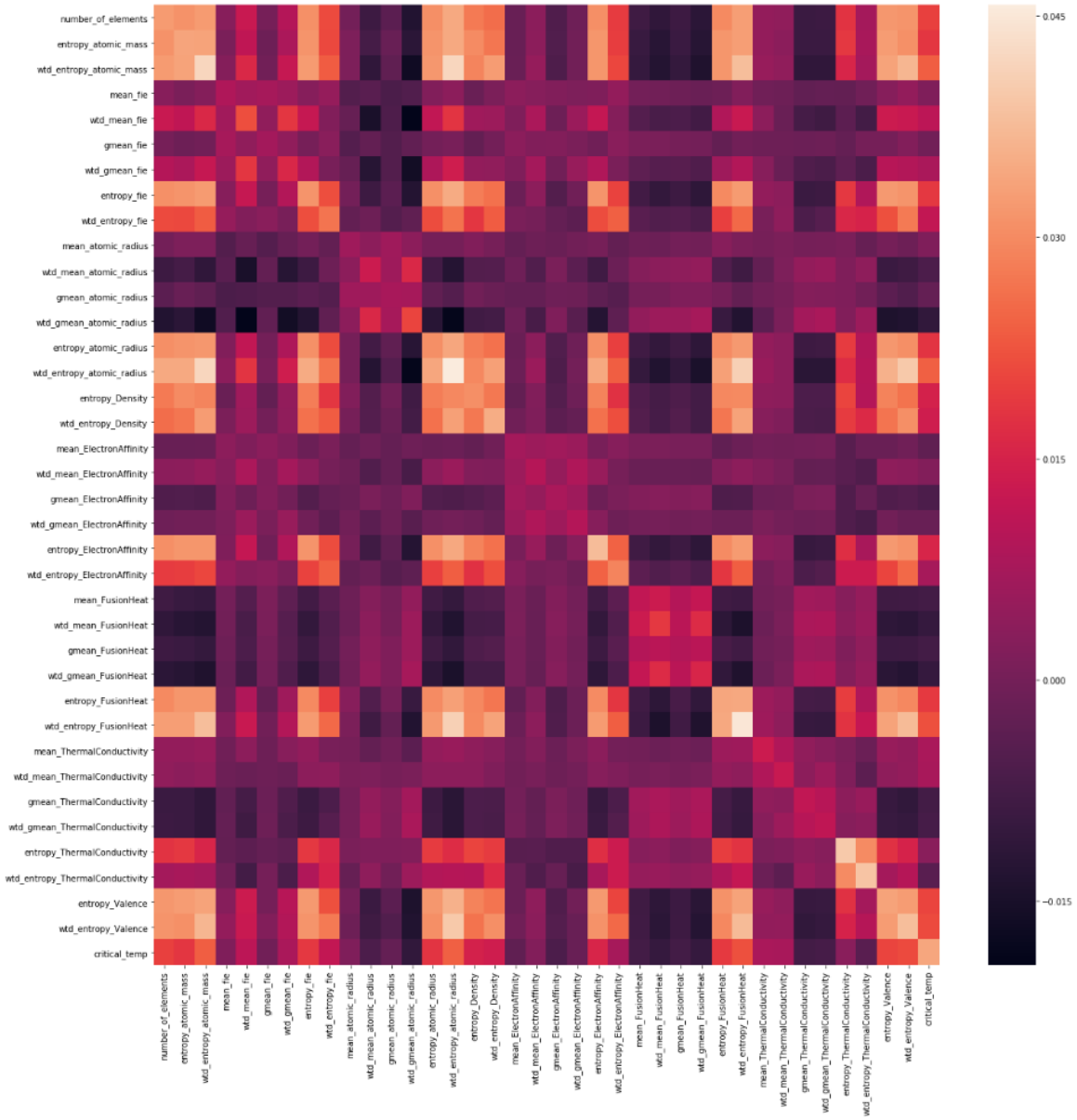

Figure 7: Heatmap of the new reduced dataframe

The new heatmap makes it easier to distinguish the correlation level in relation to the other features. In addition to the heatmap, the pearson coefficient was calculated on the current dataframe using the scipy.stats pearson function and the results are tabulated in Table 2.

Table 2. Pearson coefficients of highly correlated variables

Variable	Pearson Coefficient
number_of_elements	0.60
entropy_atomic_mass	0.54
wtd_entropy_atomic_mass	0.63
wtd_entropy_fie	0.57
wtd_entropy_atomic_radius	0.56
gmean_FusionHeat	-0.43
wtd_gmean_FusionHeat	0.55
entropy_Valence	0.60
wtd_entropy_Valence	0.59

In total, there are nine variables that showed good correlation with the critical temperature, and would probably be best to investigate further in the machine learning part of the course. To supplement these results, scatter plots of these variables were developed to visualize their relationship with the critical temperature feature. These are shown in Figure 8.

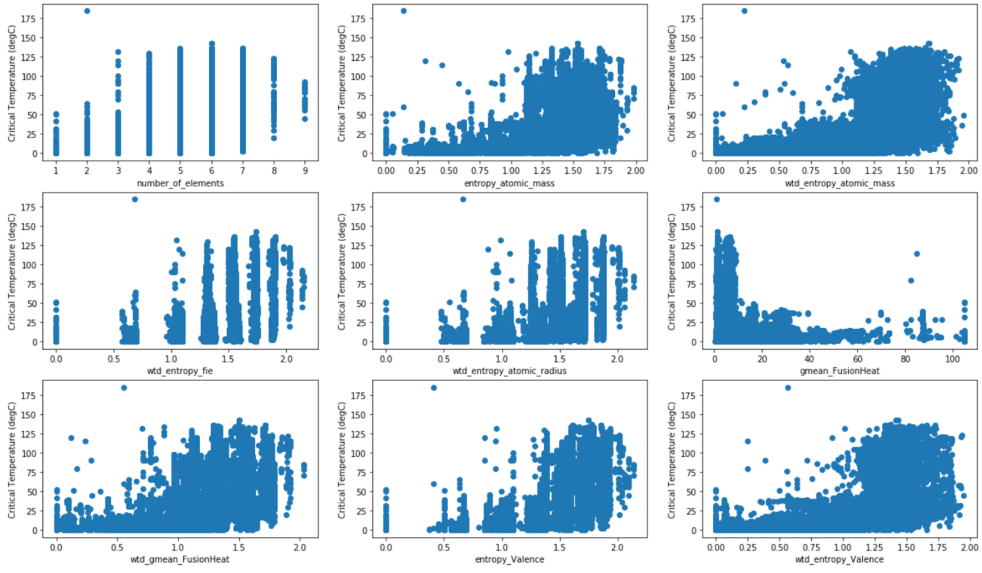

Figure 8: Scatter plots of high correlation variables with critical temperature

Boxplots of the variables were also generated to visualize the distribution of the features. This was done by utilizing the seaborn library.

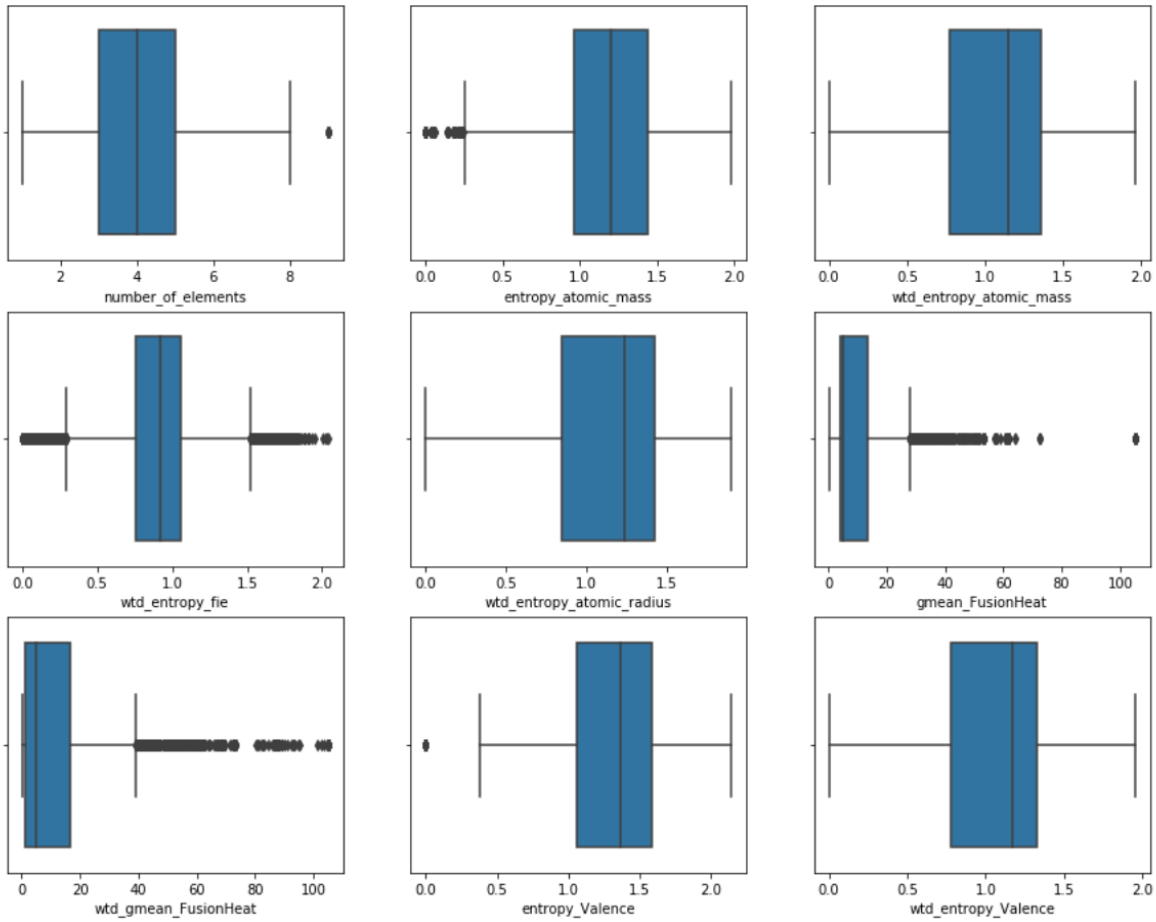

Figure 9: Boxplots of highly correlated variables with outliers generated using the seaborn library.

Certain boxplots show significant outliers. To correct this, the interquartile method was used. The removal of the outliers will help produce better results for modelling, and the resulting boxplots are shown below

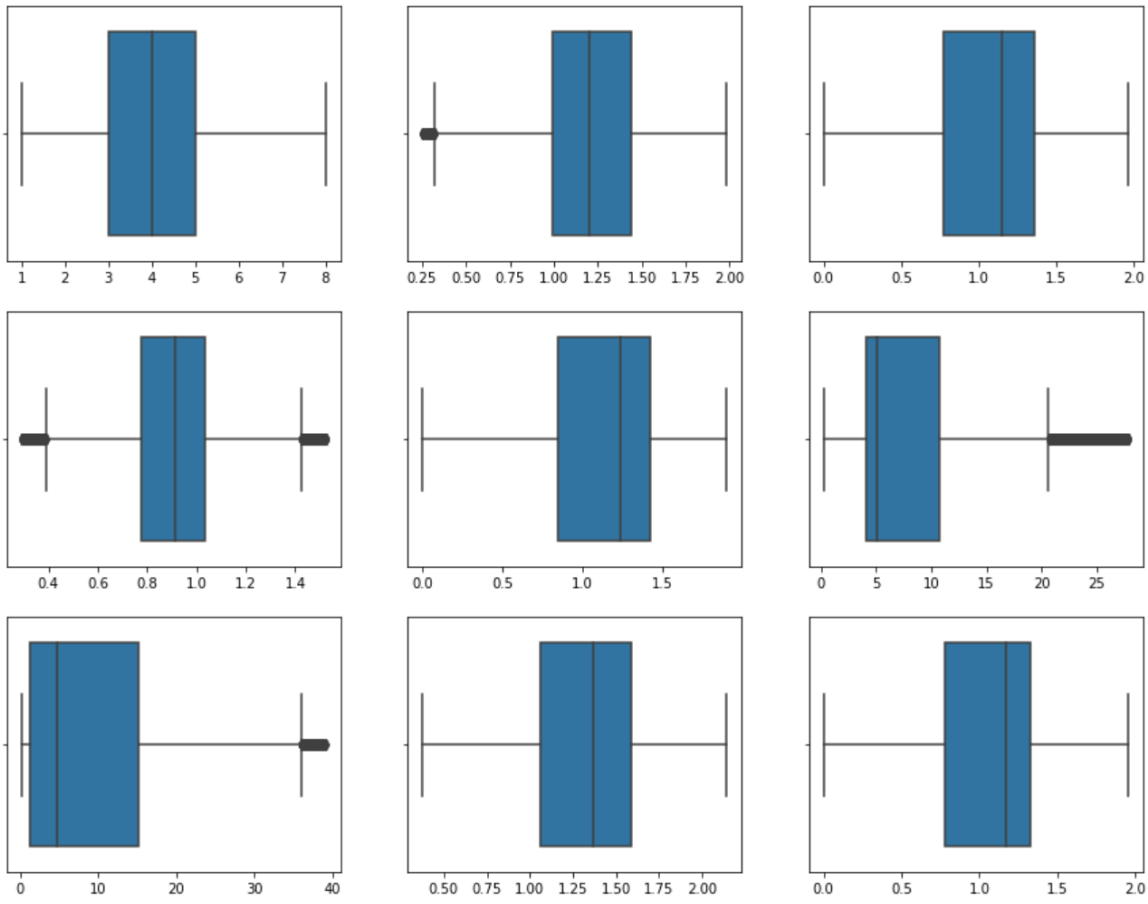

Figure 10: Boxplots of highly correlated variables without outliers generated using the seaborn library.

To conclude, the resulting statistics showed that superconductors composed with at least one element such as mercury, barium, and calcium had higher than normal critical temperatures. In addition, nine features appeared to correlate with the critical temperature variable. These results will be further investigated in the next report with the goal of using machine learning to develop a model that will be able to predict the critical temperature of a given superconductor.