Machine Learning and Analytics Material Science - Analysis of Dataset

Group Members
Danish Amin - EE17BTECH11013
Nikhil Sachan - MS17BTECH11011
Pradyumna Kabra - MS17BTECH11015
Adarsh Bujade - MS17BTECH11002

Dataset Source

We have collected this dataset from this site

https://archive.ics.uci.edu/ml/datasets/superconductivty+data

The data has been primarily collected by Kam Ham idieh (khamidieh '@' gmail.com), University of Pennsylvania, Statistics.

He also has a paper in which he has used this dataset for statistical analysis and also for Machine Learning applications.

This is the paper:
https://www.sciencedirect.com/science/article/abs/pii/S0927025618304877?via%3Dihub

Motivation

In the absence of any theory-based prediction models, the only best approach to estimate the critical temperature is to perform experiments and conclude the results for the same.

Here we have taken the statistical approach to overcome the issue as explained above. We shall be using Machine Learning approach to estimate the critical temperature of the material based on the chemical composition and physical properties.

This dataset is of superconductor materials where the type of material is Oxide and Metallic. The motivation here is to predict the critical temperature of the material based on the chemical composition and physical properties such as mean atomic weight and mean atomic radius and other parameters.

Dataset Analysis-Feature Extraction

The data we have collected has already features extracted from it. But it won't be a good analysis if we didn't show the process of feature extraction.

So first the material is picked, let's say Re_7Zr_1

We first gets it proportions, $p_1 = 6/7$ and $p_2 = 1/7$

Then we get proportion of thermal conductivity(used as an example) $t_1 = 48/71$ and $t_2 = 23/71$

And finally some intermediate value $A = p_1w_1/(p_1w_1 + p_2w_2)$ and $B = p_2w_2/(p_1w_1 + p_2w_2)$

Now we can have 10 feature for each property of material.

Dataset Analysis-Feature Extraction

Feature & description	Formula	Sample value
Mean	$= \mu = (t_1 + t_2)/2$	35.5
Weighted mean	$= \nu = (p_1 t_1) + (p_2 t_2)$	44.43
Geometric mean	$= (t_1 t_2)^{1/2}$	33.23
Weighted geometric mean	$= (t_1)^{p_1}(t_2)^{p_2}$	43.21
Entropy	$= -w_1 \ln(w_1) - w_2 \ln(w_2)$	0.63
Weighted entropy	$= -A \ln(A) - B \ln(B)$	0.26
Range	$= t_1 - t_2 \ (t_1 > t_2)$	25
Weighted range	$= p_1 t_1 - p_2 t_2$	37.86
Standard deviation	$= [(1/2)((t_1-\mu)^2 + (t_2-\mu)^2)]^{1/2}$	12.5
Weighted standard deviation	$= [p_1(t_1-\nu)^2 + p_2(t_2-\nu)^2)]^{1/2}$	8.75

Dataset Analysis-Feature Extraction

This table shows the properties of an element which are used for creating features to predict T_c.

Variable	Units	Description
Atomic Mass	Atomic mass units (AMU)	Total proton and neutron rest masses
First Ionization Energy	Kilo-Joules per mole (kJ/mol)	Energy required to remove a valence electron
Atomic Radius	Picometer (pm)	Calculated atomic radius
Density	Kilograms per meters cubed (kg/m^3)	Density at standard temperature and pressure
Electron Affinity	Kilo-Joules per mole (kJ/mol)	Energy required to add an electron to a neutral atom
Fusion Heat	Kilo-Joules per mole (kJ/mol)	Energy to change from solid to liquid without temperature change
Thermal Conductivity	Watts per meter-Kelvin (W/(m K))	Thermal conductivity coefficient κ
Valence	No units	Typical number of chemical bonds formed by the element

Dataset Analysis-Feature Extraction

After doing this on all the Properties we have 80 features and we also included one more feature that is number of atoms in the material, thus making it total of 81 features.

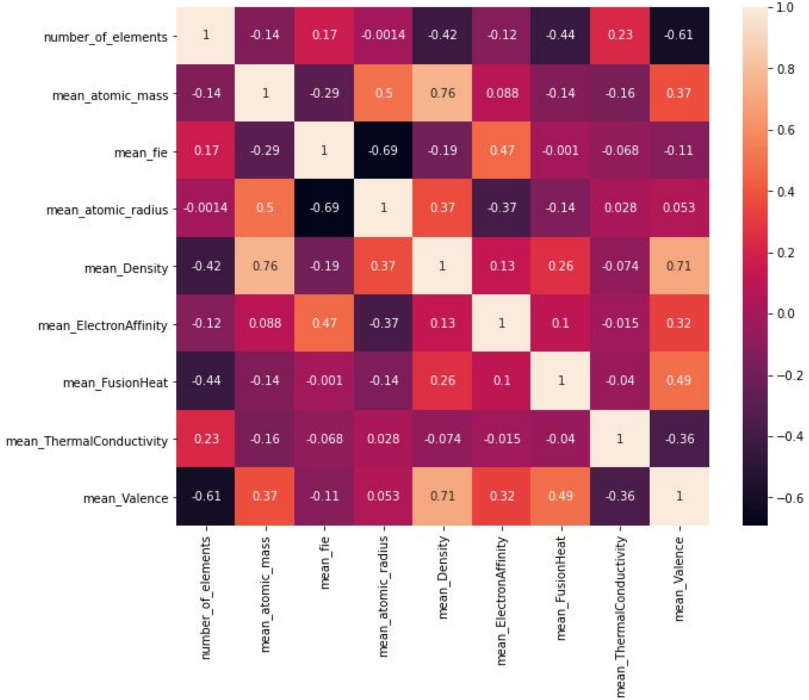
We also have a data which tells the proportion of each element a particular material has but this data is not included in features because we already have so many features and our prediction model might start to overfit if a lot of parameters are there.

Dataset Analysis-Correlation Heatmap

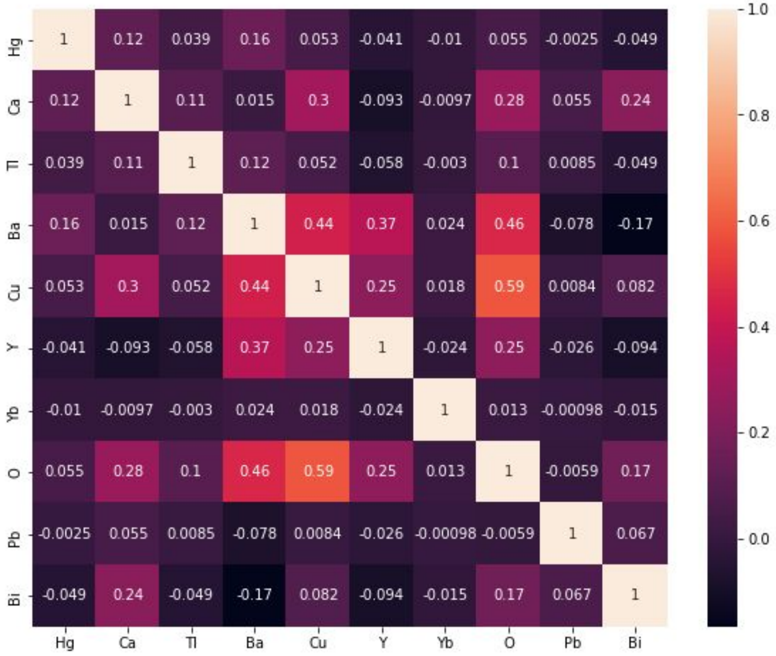
For the first Dataset containing feature below is the correlation Heatmap for only mean features of the properties.

For the second Dataset containing element proportion below is the heatmap for only 10 element having most high mean temperature.

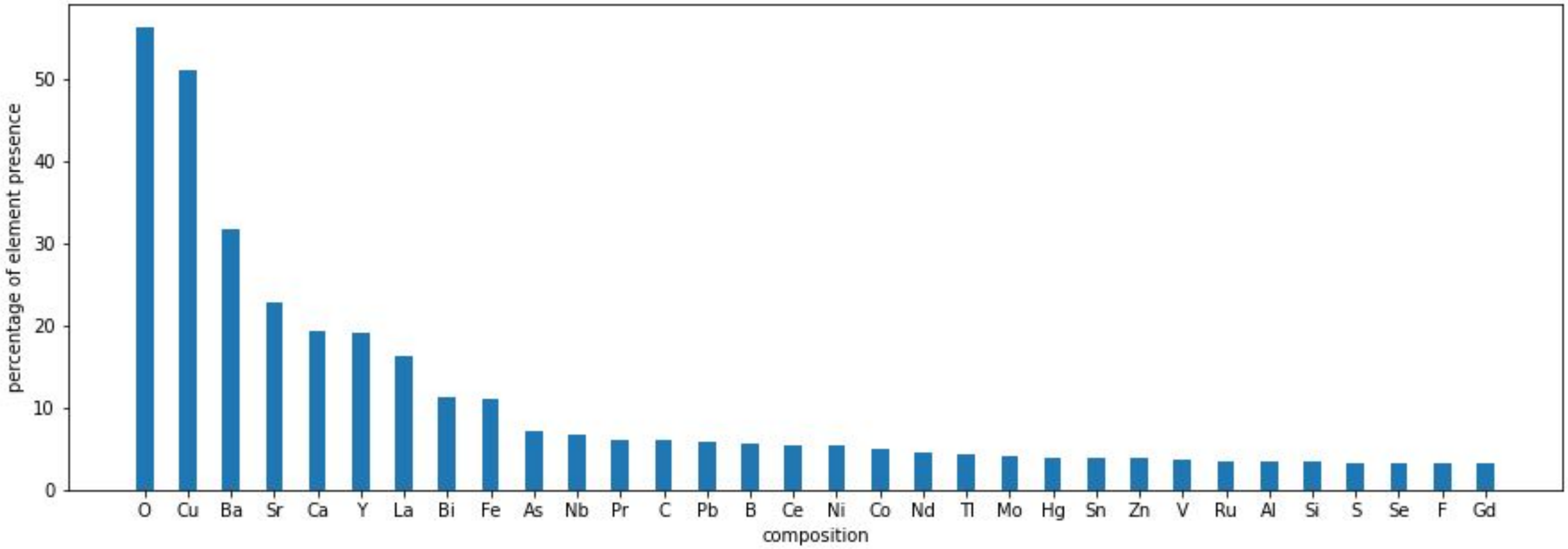
Dataset Analysis-Correlation Heatmap



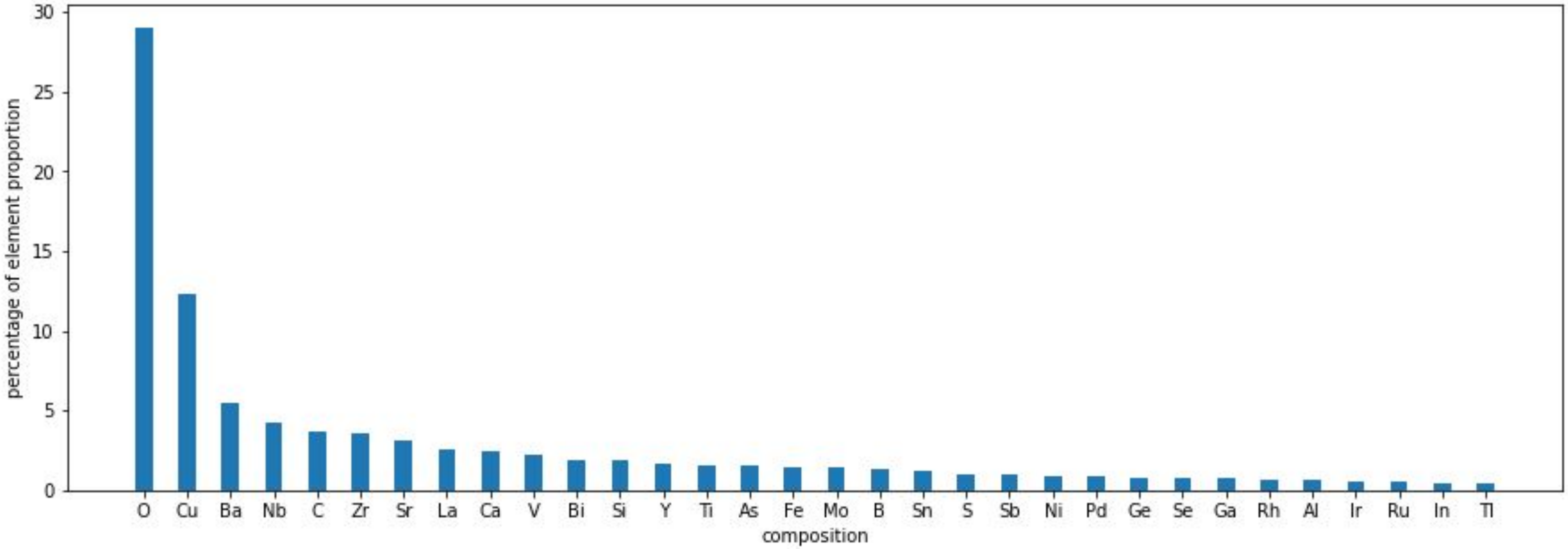
Dataset Analysis-Correlation Heatmap



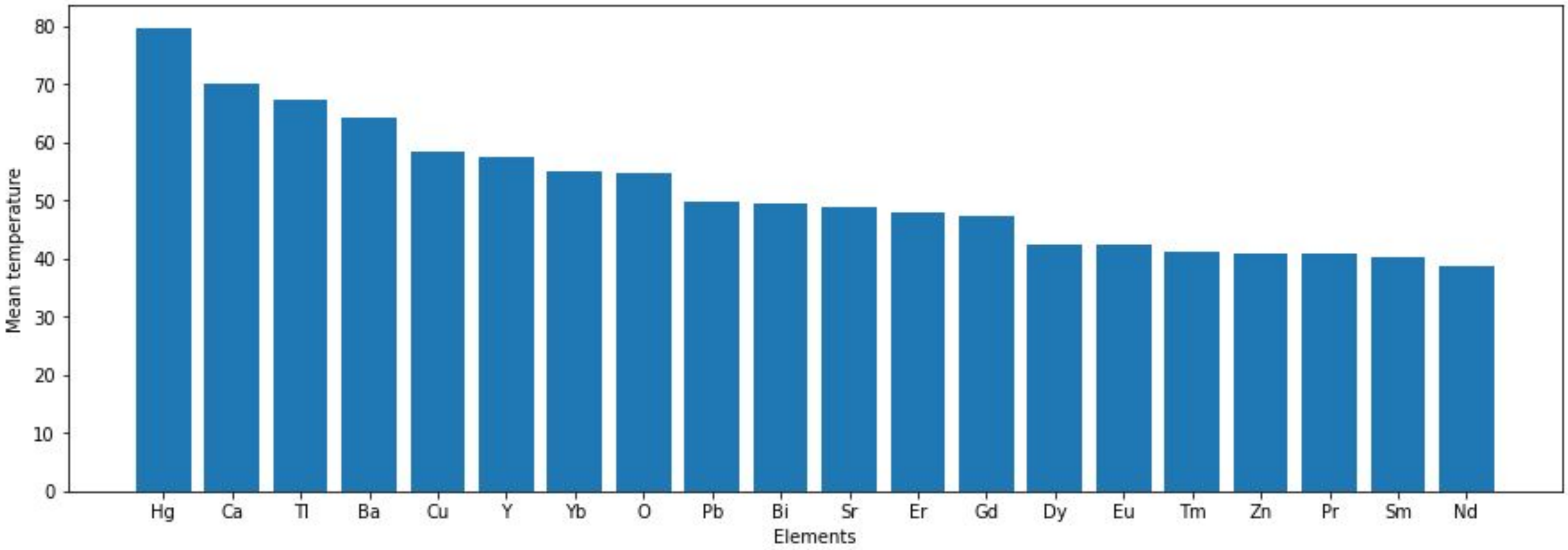
Dataset Analysis-Element Presence Bar Chart



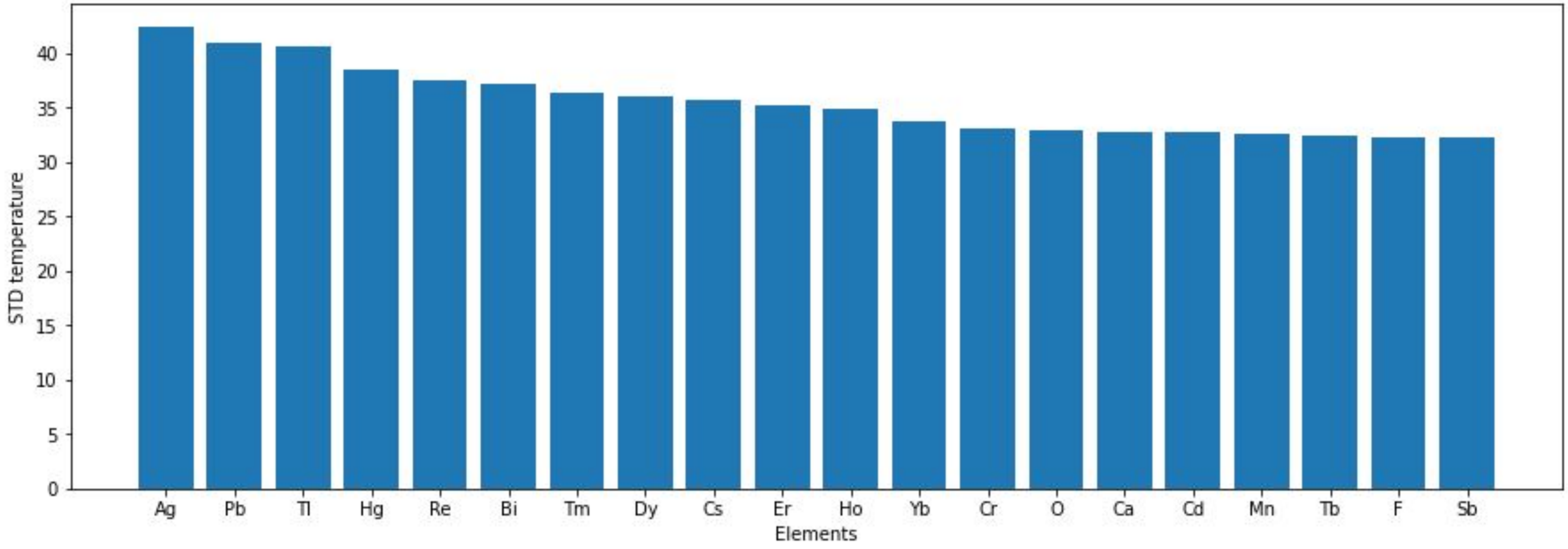
Dataset Analysis-Element Proportion Bar Chart



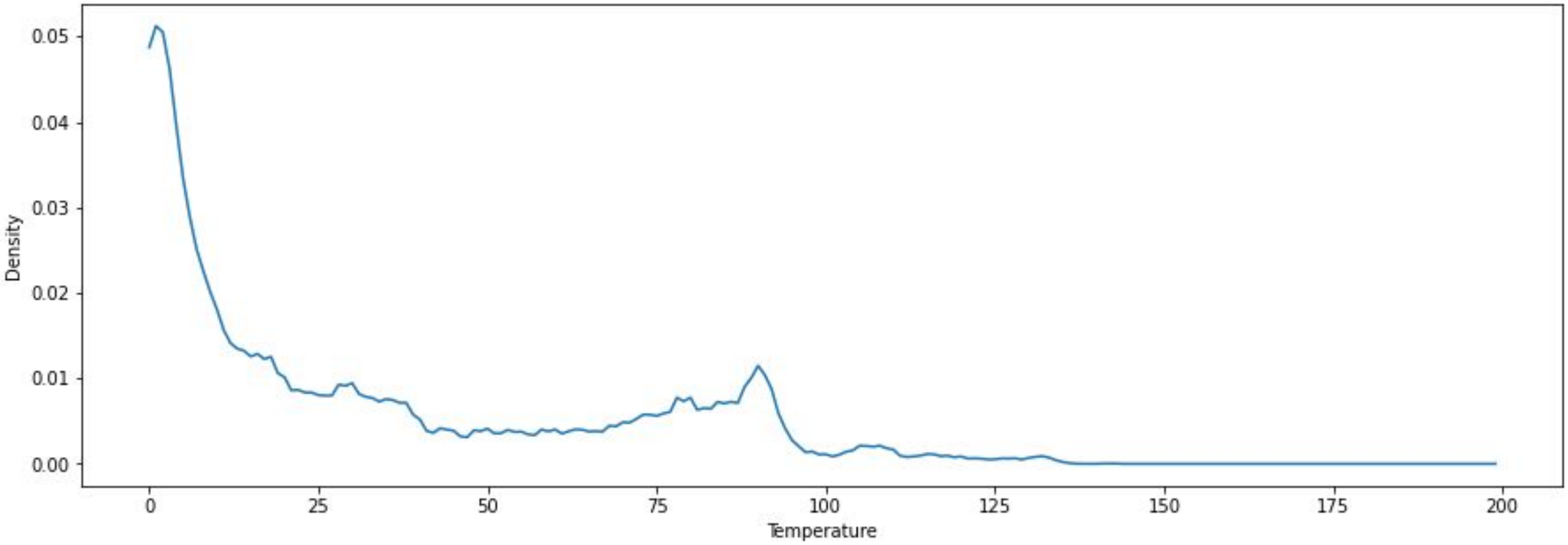
Dataset Analysis-Element Mean Temp. Bar Chart



Dataset Analysis-Element STD Temp. Bar Chart



Dataset Analysis-Temp. Density Graph



Dataset Analysis-Conclusion

- Oxygen is present in about 56% of the superconductors. Copper, barium, strontium, and calcium are the next most abundant elements.
- Iron is present in approximately 11% of the superconductors. The mean T_c of superconductors with iron is 26.9 ± 21.4 K.
- The non-iron containing superconductors' mean is 35.4 ± 35.4 K.
- Density graph is Bimodal and values are right skewed with a bump around 80 K.
- Mercury containing superconductors have the highest Tc at around 80 K on average.
- Mercury also has 4th largest standard deviation, falling behind Ag, Pb and Ti.