Data Featurizing and Data Preprocessing

# Quality of Data

###### Size of the Data

###### Number of columns

6

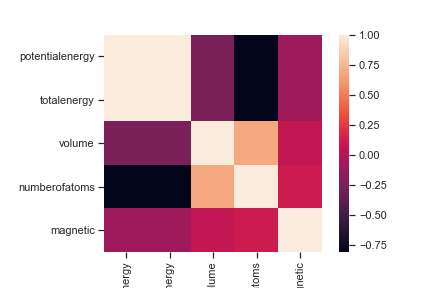
###### Number of rows

3814

###### Type of data in each of the columns

formula object  
potentialenergy float64  
totalenergy float64  
volume float64  
numberofatoms int64  
magnetic float64  
dtype: object

###### The heat matrix of all the features



###### Detailed description of the data entries

count mean std min 25% \  
potentialenergy 3814.0 -26.761224 17.175239 -111.390446 -33.255093   
totalenergy 3814.0 -26.761224 17.175239 -111.390446 -33.255093   
volume 3814.0 337.145029 201.038453 79.075438 203.596066   
numberofatoms 3814.0 5.072365 2.344566 2.000000 3.000000   
magnetic 3814.0 0.506186 1.386526 -3.998736 0.000000   
  
 50% 75% max   
potentialenergy -22.020143 -14.690745 -4.165705   
totalenergy -22.020143 -14.690745 -4.165705   
volume 264.167941 389.197282 1421.320256   
numberofatoms 4.000000 6.000000 12.000000   
magnetic 0.000000 0.000000 11.168096

###### Data Completedness: Number of incomplete data entries

0

###### Data Uniqueness: Number of duplicates

0

###### Ratio of data to errors:

0.0

# A summary of the data featurizing process

Converted formula into composition

Obtained the Valence Orbitals

Obtained the Band Center

Obtained the Stoichiometry

Obtained the Ion Property

Obtained the Element Fraction

Obtained the Oxidation Composition

Unable to obtain the Oxidation States

Obtained the Yang Solid Solution

Unable to obtain the Electron Affinity

Unable to obtain the Density Features

Unable to obtain the Dimensionality

Unable to obtain the Ewald Energy

Unable to obtain the Global Instability Index

Unable to obtain the Maximum Packing Efficiency

Unable to obtain the Coordination Number

Unable to obtain the Average Bond Length

Unable to obtain the Average Bond Angle

Unable to obtain the AGNIFingerprints

Obtained the Number of Atoms

Obtained the Number of Magnetic Atoms

# Columns after data featurizing

133

formulapotentialenergynumber\_of\_magnetic\_atomsnumber\_of\_atomstotalenergyvolumenumberofatomsmagneticcompositionavg s valence electronsavg p valence electronsavg d valence electronsavg f valence electronsfrac s valence electronsfrac p valence electronsfrac d valence electronsfrac f valence electronsband center0-norm2-norm3-norm5-norm7-norm10-normcompound possiblemax ionic charavg ionic charHHeLiBeBCNOFNeNaMgAlSiPSClArKCaScTiVCrMnFeCoNiCuZnGaGeAsSeBrKrRbSrYZrNbMoTcRuRhPdAgCdInSnSbTeIXeCsBaLaCePrNdPmSmEuGdTbDyHoErTmYbLuHfTaWReOsIrPtAuHgTlPbBiPoAtRnFrRaAcThPaUNpPuAmCmBkCfEsFmMdNoLrcomposition\_oxidYang omegaYang delta

# Data Entries

3814

# Summary of Data Preprocessing

###### The summary of the data cleaning

###### Deleting columns with no numerical significance

###### Number of columns with no numerical significance

0

###### Deleting columns with low variance

###### Performing data imputation

All missing data replaced with the mean value of the column

###### Checking for Duplicates

No duplicates present

# The summary of data transformation

###### Conducting data binning

The number of bins created:

13

###### Conducting data normalisation

Values in all the columns are now in the range of 0 to 1

# Summary of Data Reduction

###### Conducting Numerosity Reduction

Fraction of Numerosity Reduction by

1

Number of data entries reduced

0

###### Conducting Dimensionality Reduction

###### Clustering the Data

# Quality of Data

###### Size of the Data

###### Number of columns

86

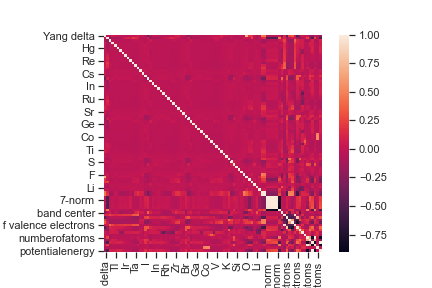
###### Number of rows

3814

###### Type of data in each of the columns

Yang delta float64  
Yang omega float64  
Bi float64  
Pb float64  
Tl float64  
 ...   
volume float64  
totalenergy float64  
number\_of\_atoms float64  
number\_of\_magnetic\_atoms float64  
potentialenergy float64  
Length: 86, dtype: object

###### The heat matrix of all the features



###### Detailed description of the data entries

count mean std min 25% 50% \  
Yang delta 3814.0 0.163668 0.121921 0.0 0.153846 0.153846   
Yang omega 3814.0 0.177060 0.139631 0.0 0.076923 0.153846   
Bi 3814.0 0.021278 0.111248 0.0 0.000000 0.000000   
Pb 3814.0 0.010447 0.074947 0.0 0.000000 0.000000   
Tl 3814.0 0.009923 0.084089 0.0 0.000000 0.000000   
... ... ... ... ... ... ...   
volume 3814.0 0.153039 0.151997 0.0 0.076923 0.076923   
totalenergy 3814.0 0.751140 0.162012 0.0 0.692308 0.769231   
number\_of\_atoms 3814.0 0.272639 0.228033 0.0 0.076923 0.153846   
number\_of\_magnetic\_atoms 3814.0 0.034851 0.118711 0.0 0.000000 0.000000   
potentialenergy 3814.0 0.751140 0.162012 0.0 0.692308 0.769231   
  
 75% max   
Yang delta 0.153846 0.923077   
Yang omega 0.230769 0.923077   
Bi 0.000000 0.923077   
Pb 0.000000 0.923077   
Tl 0.000000 0.923077   
... ... ...   
volume 0.230769 0.923077   
totalenergy 0.846154 0.923077   
number\_of\_atoms 0.384615 0.923077   
number\_of\_magnetic\_atoms 0.000000 0.923077   
potentialenergy 0.846154 0.923077   
  
[86 rows x 8 columns]

###### Data Completedness: Number of incomplete data entries

0

###### Data Uniqueness: Number of duplicates

0

###### Ratio of data to errors:

0.0

# End of Data Featurizing and Data Pre-processing

Data is now ready for input in machine learning and deep learning algorithms