Data Featurizing and Data Preprocessing

# Quality of Data

###### Size of the Data

###### Number of columns

29

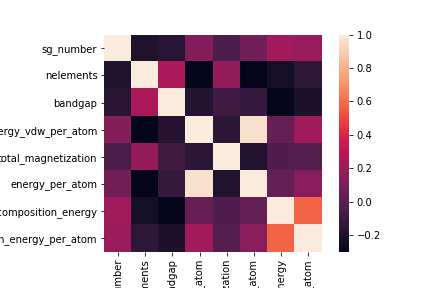
###### Number of rows

6351

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

\_id object  
sg\_number int64  
formula\_pretty object  
nelements int64  
creation\_task\_label object  
created\_at object  
chemsys object  
material\_id object  
discovery\_process object  
formula\_reduced\_abc object  
sg\_symbol object  
formula\_anonymous object  
source\_id object  
spacegroup object  
elements object  
\_tasksbuilder object  
structure object  
bandstructure object  
bandgap float64  
thermo object  
energy\_vdw\_per\_atom float64  
magnetism object  
total\_magnetization float64  
calc\_settings object  
energy\_per\_atom float64  
relative\_id object  
decomposition\_energy float64  
exfoliation\_energy\_per\_atom float64  
literature\_doi object  
dtype: object

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



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

count mean std min \  
sg\_number 6351.0 74.205952 67.610089 1.000000   
nelements 6351.0 2.410801 0.666312 1.000000   
bandgap 6351.0 1.048038 1.443890 0.000000   
energy\_vdw\_per\_atom 6351.0 -3.906346 1.749797 -9.769199   
total\_magnetization 6351.0 1.141026 3.204535 -10.014716   
energy\_per\_atom 6156.0 -4.881024 1.594898 -10.724971   
decomposition\_energy 6155.0 0.267288 0.354539 0.000000   
exfoliation\_energy\_per\_atom 4527.0 0.150947 0.204484 -1.218706   
  
 25% 50% 75% max   
sg\_number 11.000000 51.000000 145.000000 191.000000   
nelements 2.000000 2.000000 3.000000 4.000000   
bandgap 0.000000 0.221500 1.822050 9.449400   
energy\_vdw\_per\_atom -5.128683 -3.912666 -2.637456 1.782194   
total\_magnetization -0.000000 0.000000 0.245163 44.005684   
energy\_per\_atom -5.960577 -4.794826 -3.748102 -0.216389   
decomposition\_energy 0.039261 0.146934 0.359714 3.469618   
exfoliation\_energy\_per\_atom 0.058429 0.089230 0.166131 3.056498

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

8725

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

0

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

1.3737994016690285

# 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

156

\_idsg\_numbernumber\_of\_magnetic\_atomsnumber\_of\_atomsformula\_prettynelementscreation\_task\_labelcreated\_atchemsysmaterial\_iddiscovery\_processformula\_reduced\_abcsg\_symbolformula\_anonymoussource\_idspacegroupelements\_tasksbuilderstructurebandstructurebandgapthermoenergy\_vdw\_per\_atommagnetismtotal\_magnetizationcalc\_settingsenergy\_per\_atomrelative\_iddecomposition\_energyexfoliation\_energy\_per\_atomliterature\_doicompositionavg 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

6351

# 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

Deleting data entries with missing data

Number of data entries with missing data which were deleted

0

###### 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

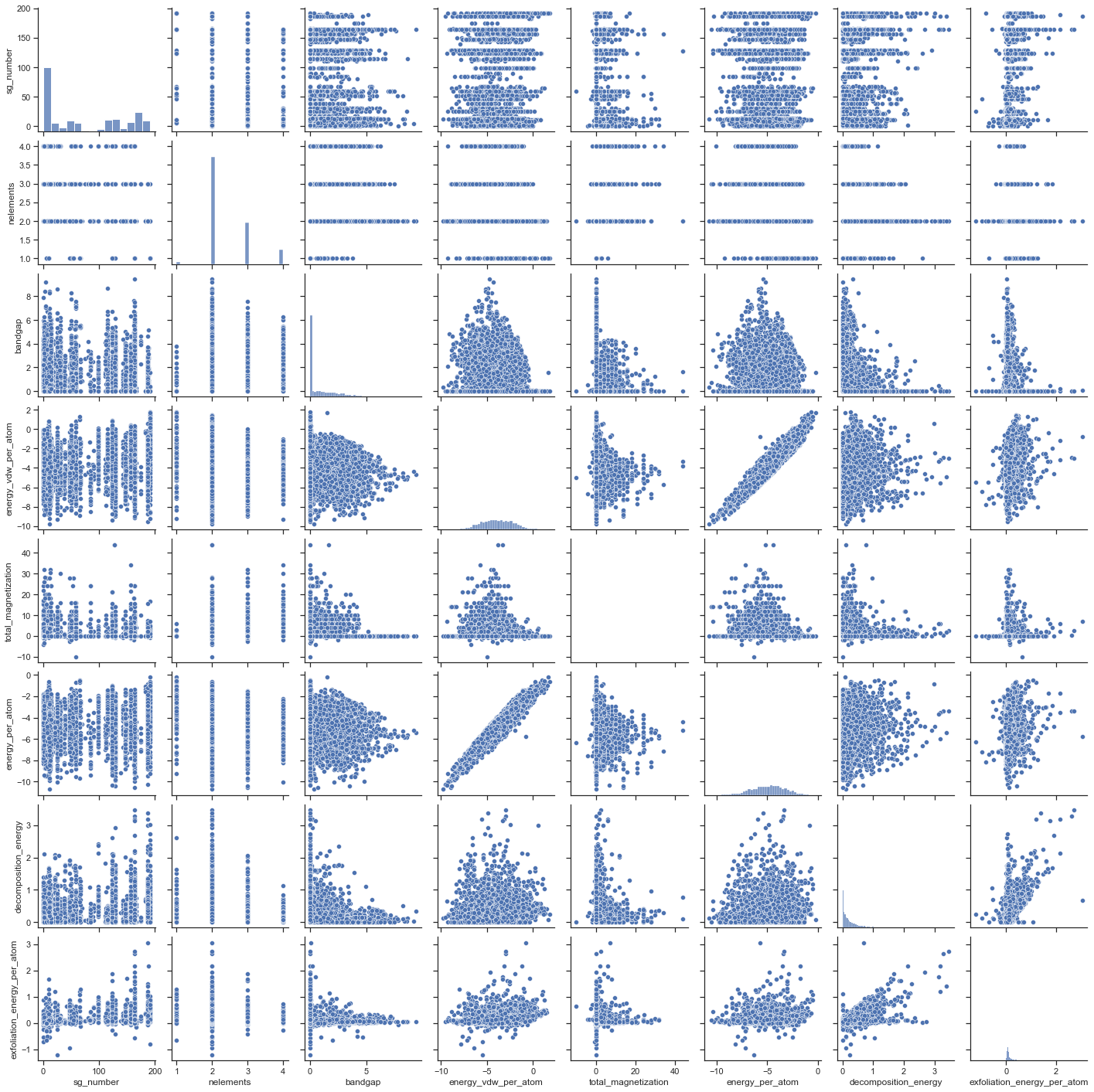
1

Number of data entries reduced

0

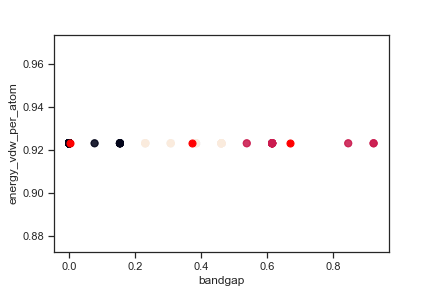
###### Conducting Dimensionality Reduction

The graph of dependence of attributes on each other



Action should be taken after a manual visual inspection of the plot

###### Clustering the Data



Created above cluster plot for following columns

Fe

Cl

# Quality of Data

###### Size of the Data

###### Number of columns

112

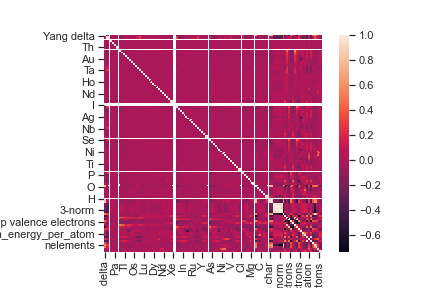
###### Number of rows

2637

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

Yang delta float64  
Yang omega float64  
Pu float64  
Np float64  
U float64  
 ...   
bandgap float64  
nelements float64  
number\_of\_atoms float64  
number\_of\_magnetic\_atoms float64  
sg\_number float64  
Length: 112, dtype: object

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



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

count mean std min 25% \  
Yang delta 2637.0 0.277763 2.344116e-01 0.000000 0.076923   
Yang omega 2637.0 0.141507 1.457631e-01 0.000000 0.000000   
Pu 2637.0 0.923077 1.110434e-16 0.923077 0.923077   
Np 2637.0 0.000700 2.541648e-02 0.000000 0.000000   
U 2637.0 0.005076 5.948857e-02 0.000000 0.000000   
... ... ... ... ... ...   
bandgap 2637.0 0.083370 1.396066e-01 0.000000 0.000000   
nelements 2637.0 0.426242 2.038102e-01 0.000000 0.307692   
number\_of\_atoms 2637.0 0.210904 2.033159e-01 0.000000 0.076923   
number\_of\_magnetic\_atoms 2637.0 0.033605 1.102892e-01 0.000000 0.000000   
sg\_number 2637.0 0.393512 3.605265e-01 0.000000 0.000000   
  
 50% 75% max   
Yang delta 0.230769 0.538462 0.923077   
Yang omega 0.153846 0.230769 0.923077   
Pu 0.923077 0.923077 0.923077   
Np 0.000000 0.000000 0.923077   
U 0.000000 0.000000 0.923077   
... ... ... ...   
bandgap 0.000000 0.153846 0.923077   
nelements 0.307692 0.615385 0.923077   
number\_of\_atoms 0.153846 0.307692 0.923077   
number\_of\_magnetic\_atoms 0.000000 0.000000 0.923077   
sg\_number 0.230769 0.769231 0.923077   
  
[112 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