**EXPLORATORY DATA ANALYSIS REPORT**

Name-Rajrishi Sarkar

Roll-203110061

# Data description

**Metallic Glass Descriptors**:

Relevant papers:

There is no paper published on this data set at the present time. The data was assembled primarily by Vanessa Nilsen under the guidance of Prof. Dane Morgan at UW Madison (ddmorgan@wisc.edu). A previous study of reduced glass transition temperature as a GFA descriptor can be found in reference [5]: https://www.sciencedirect.com/science/article/pii/S0022309300000648

X features:

The metallic glass dataset gives two columns with information about the material Composition. The first is the overall composition, and the second is the highest Composition element. The columns from four to the end are the MAGPIE features that have been generated from the material composition column and give values such as properties averaged over the material composition as well as features that are only for the majority element in each alloy [3]. The majority element features are labelled as "site1".

Y property:

The reduced glass transition temperature (Trg) has historically been used as a rough predictor for Glass Forming Ability (GFA). By making a model to predict Trg for an arbitrary alloy, it could be possible to use these values to estimate GFA directly, or as input for another model to then predict GFA.

**Checking for missing values**

1. Material.compositions=0
2. main\_element=0
3. Trg=0
4. Density\_composition\_average=0
5. IsBoron\_composition\_average=0
6. IsDBlock\_composition\_average=0
7. IsTransitionMetal\_composition\_average=0
8. NdValence\_composition\_average=0
9. NValance\_composition\_average=0
10. HeatVaporization\_max\_value=0
11. BoilingT\_difference=0
12. HeatVaporization\_difference=0
13. MeltingT\_difference=0
14. NdValence\_difference=0
15. NsUnfilled\_difference=0
16. valence\_difference=0
17. Site1\_Density=0
18. Site1\_HeatCapacityMass=0
19. Site1\_HeatFusion=0
20. Site1\_IsDBlocK=0
21. Site1\_IsTransitionMetal=0
22. Site1\_NdValence=0
23. Site1\_SpecificHeatCapacity=0

**OBSERVATION:::**

there are no missing values in any of the features including the target variable, no need to perform missing value imputation on the dataset

## **Feature statistics**

**Material.compositions main\_element Trg**

Ag10Cu35Zr55 : 1 Cu :113 Min. :0.2230

Ag10Cu50Zr40 : 1 Zr : 95 1st Qu.:0.5570

Ag10Cu55Zr35 : 1 Fe : 70 Median :0.5840

Ag20Cu40Zr40 : 1 Ca : 68 Mean :0.5774

Ag35Ca65 : 1 Mg : 49 3rd Qu.:0.6090

Al13.3Co26.7Zr60: 1 Ni : 38 Max. :0.6880

(Other) :579 (Other):152

**Density\_composition\_average IsBoron\_composition\_average**

Min. :-2.2964 Min. :-0.5856

1st Qu.:-0.6580 1st Qu.:-0.5856

Median : 0.1413 Median :-0.5856

Mean : 0.0000 Mean : 0.0000

3rd Qu.: 0.4915 3rd Qu.: 0.2045

Max. : 4.4187 Max. : 6.4790

**IsDBlock\_composition\_average IsTransitionMetal\_composition\_average**

Min. :-2.0914 Min. :-2.0914

1st Qu.:-1.1550 1st Qu.:-1.1550

Median : 0.3744 Median : 0.3744

Mean : 0.0000 Mean : 0.0000

3rd Qu.: 0.8270 3rd Qu.: 0.8270

Max. : 1.0299 Max. : 1.0299

**NdValence\_composition\_average NValance\_composition\_average**

Min. :-2.14846 Min. :-1.92286

1st Qu.:-0.84351 Qu.:-0.70925

Median : 0.04683 Median : 0.05548

Mean : 0.00000 Mean : 0.00000

3rd Qu.: 0.78511 3rd Qu.: 0.48772

Max. : 2.49636 Max. : 6.27228

**HeatVaporization\_max\_value BoilingT\_difference HeatVaporization\_diff**

Min. :-2.4367 Min. :-2.1189 Min. :-1.8446

1st Qu.:-0.7406 1st Qu.:-0.5226 1st Qu.:-0.7852

Median : 0.5639 Median : 0.1006 Median : 0.1761

Mean : 0.0000 Mean : 0.0000 Mean : 0.0000

3rd Qu.: 0.5639 3rd Qu.: 0.4384 3rd Qu.: 0.5057

Max. : 1.5330 Max. : 3.2476 Max. : 3.0847

**MeltingT\_difference NdValence\_difference NsUnfilled\_difference**

Min. :-1.4351 Min. :-1.6544 Min. :-1.1657

1st Qu.:-0.5910 1st Qu.:-0.8563 1st Qu.:-1.1657

Median :-0.2356 Median : 0.4739 Median : 0.8578

Mean : 0.0000 Mean : 0.0000 Mean : 0.0000

3rd Qu.: 0.2301 3rd Qu.: 1.0059 3rd Qu.: 0.8578

Max. : 3.8224 Max. : 1.0059 Max. : 0.8578

**valence\_difference Site1\_Density Site1\_HeatCapacityMass**

Min. :-1.7450 Min. :-1.46865 Min. :-1.2793

1st Qu.:-0.7403 1st Qu.:-0.63183 1st Qu.:-0.7280

Median : 0.2645 Median : 0.09502 Median :-0.3321

Mean : 0.0000 Mean : 0.00000 Mean : 0.0000

3rd Qu.: 0.2645 3rd Qu.: 0.64202 3rd Qu.: 0.1822

Max. : 2.2740 Max. : 4.12736 Max. : 4.9958

**Site1\_HeatFusion** **Site1\_IsDBlock** **Site1\_IsTransitionMetal**

Min. :-1.74083 Min. :-1.393 Min. :-1.393

1st Qu.:-1.06241 1st Qu.:-1.393 1st Qu.:-1.393

Median : 0.05214 Median : 0.718 Median : 0.718

Mean : 0.00000 Mean : 0.000 Mean : 0.000

3rd Qu.: 0.74378 3rd Qu.: 0.718 3rd Qu.: 0.718

Max. : 3.66452 Max. : 0.718 Max. : 0.718

**Site1\_SpecificHeatCapacity Site1\_NdValence**

Min. :-1.2808 Min. :-1.1064

1st Qu.:-0.7259 1st Qu.:-1.1064

Median :-0.3301 Median :-0.6215

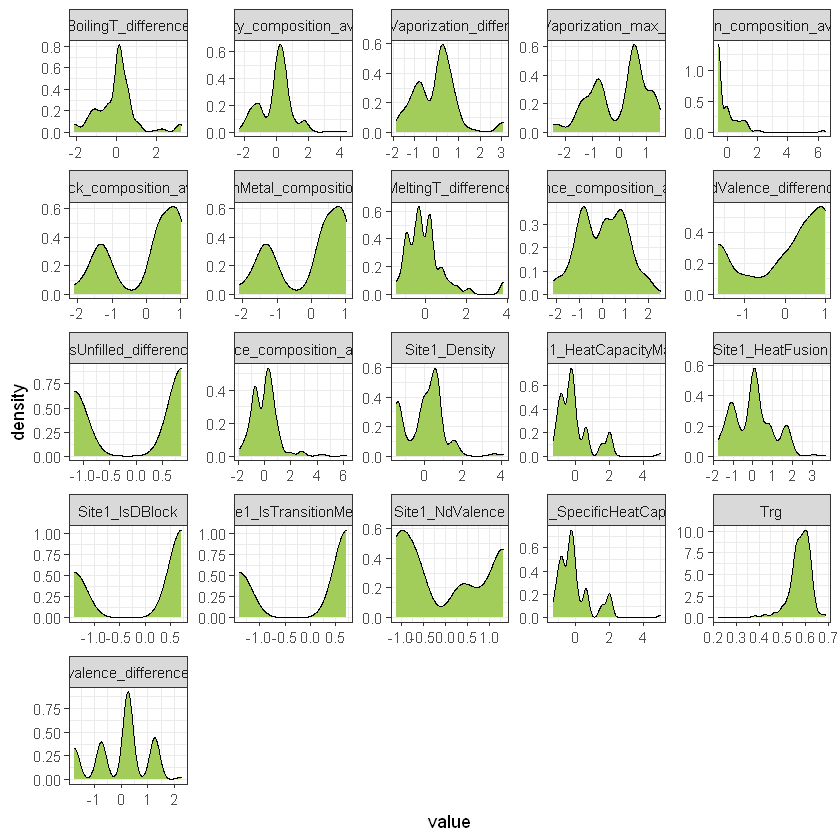
Mean : 0.0000 Mean : 0.000

3rd Qu.: 0.1804 3rd Qu.: 1.3177

Max. : 4.9969 Max. : 1.3177

**Univariate analysis**

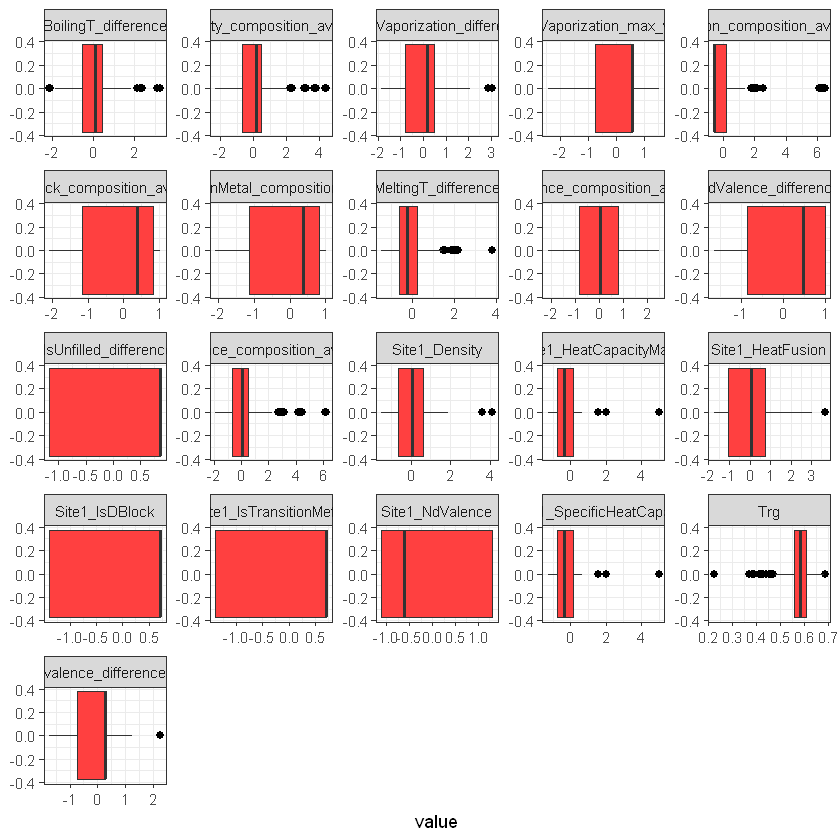
**Density plots of the numerical features**



**OBSERVATIONS:::**

1. as it can seen all the variables have different distributions.
2. The target variable "glass transition temp" is having a left skewed distribution.
3. almost all the features have 2 or more peaks in the distribution.
4. "IsBoron\_composition\_average" has an exponential distribution

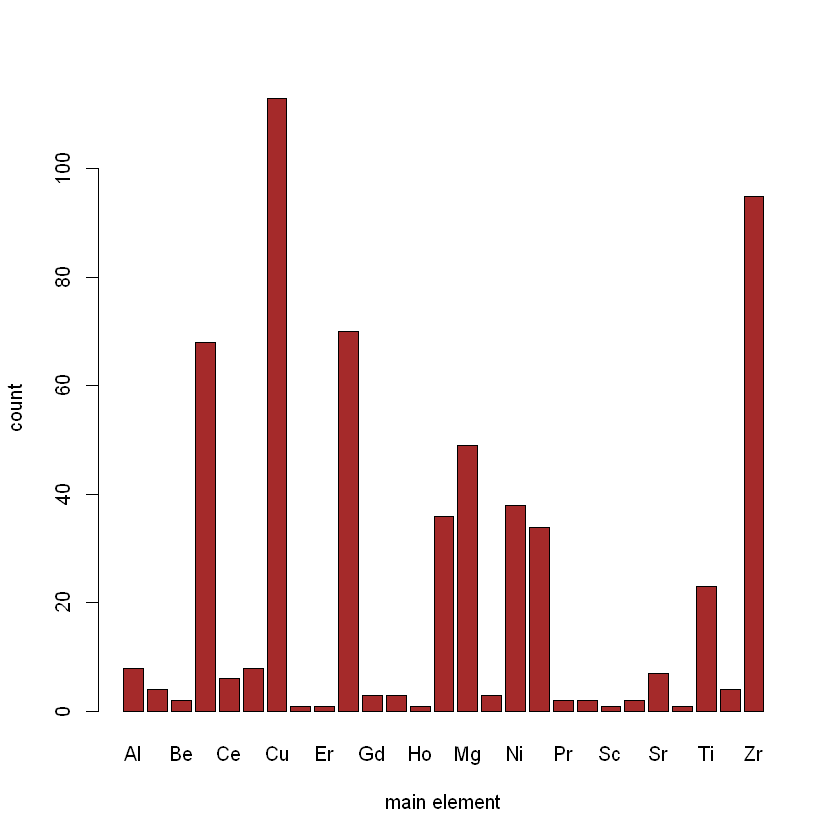
**Boxplots**



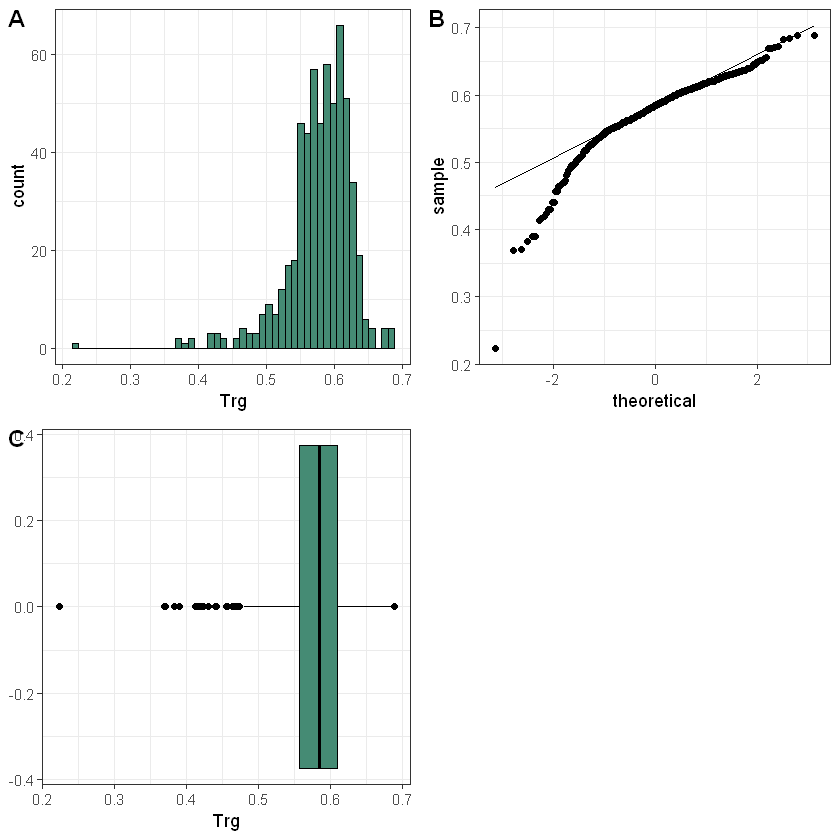
**OBSERVATIONS::**

1. "trg" has too many outliers.
2. features like "Site\_heatfusion","site1\_density","site1\_heatcapacitymass","melting\_temp\_difference","BoilingT\_difference"," IsBoron\_composition\_average" also contains outliers.

**Countplot of "main element"**



**Target feature**

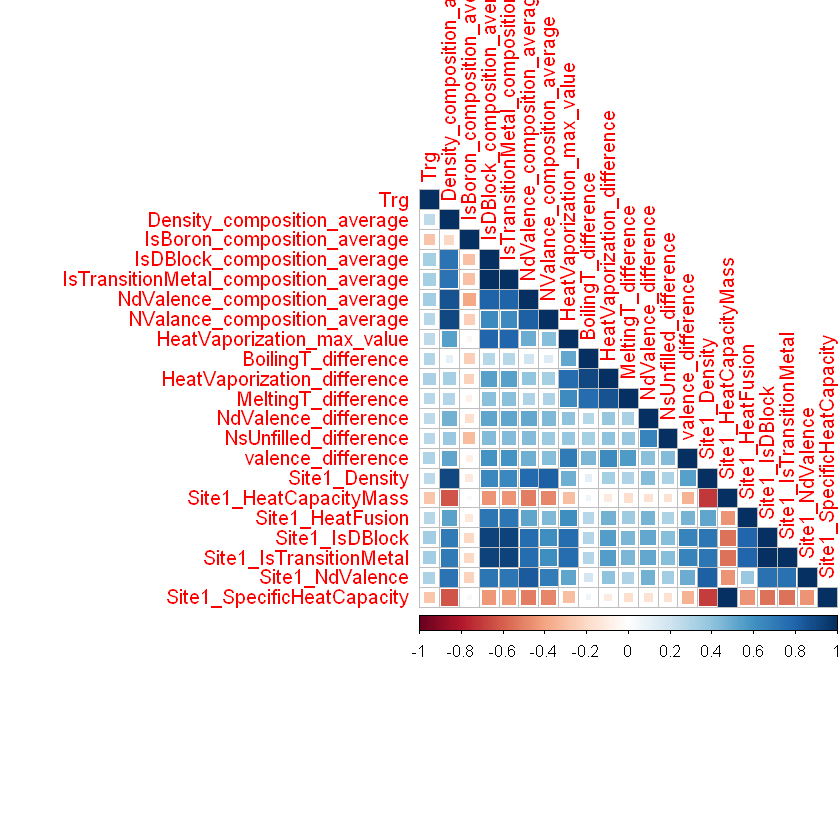


**OBSERVATIONS:::**

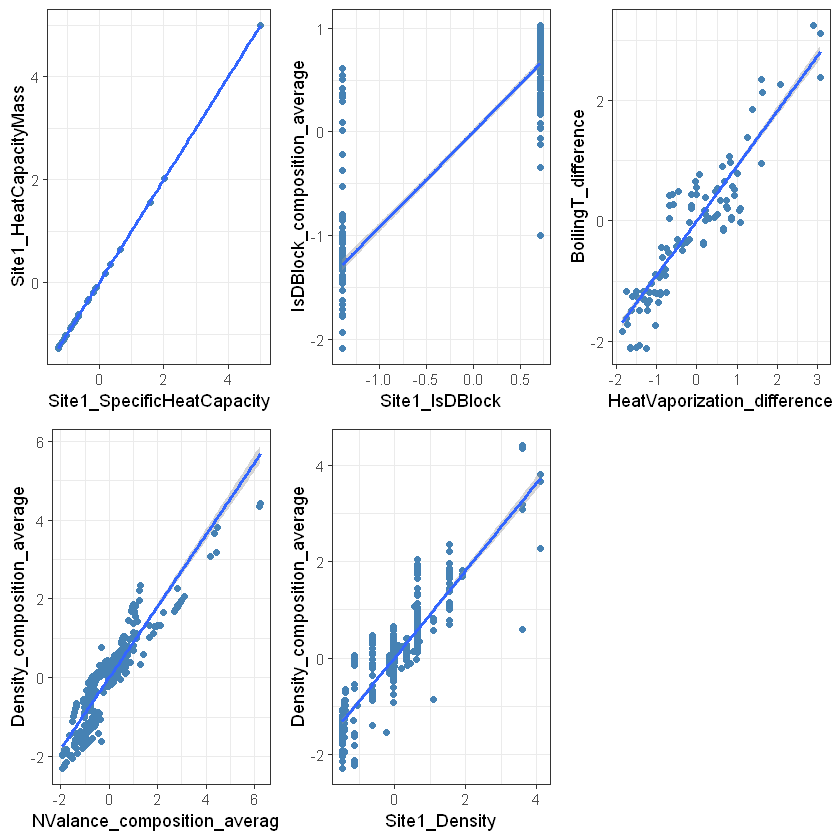
1. our target feature is 'reduced glass transition temperature' is the measure of GFA.
2. the boxplot says that it contains outliers
3. the distribution is left skewed.
4. the qq plot suggests that there is a deviation from normal distribution.

# Bivariate analysis

**Correlation between the features**



**Scatterplots of the highly correlated variables**



**OBSERVATIONS::**

The most correlated variables are

1. Site1\_SpecificHeatCapacity---Site1\_HeatCapacityMass ---- 0.9999816
2. Site1\_IsDBlock----IsDBlock\_composition\_average----0.9217491
3. HeatVaporization\_difference----BoilingT\_difference----0.9095843
4. NValance\_composition\_average----Density\_composition\_average----0.9045063
5. Site1\_Density----Density\_composition\_average----0.9030622
6. MeltingT\_difference----HeatVaporization\_difference----0.8635075
7. NdValence\_composition\_average-----Density\_composition\_average----0.8632754
8. Site1\_NdValence-----NdValence\_composition\_average---- 0.8364252
9. Site1\_Density----NValance\_composition\_average----0.8195043
10. NValance\_composition\_average----NdValence\_composition\_average---- 0.8150236
11. Site1\_NdValence----Site1\_Density----0.8148377
12. NdValence\_composition\_average-----IsDBlock\_composition\_average---- 0.8090395
13. Also no feature is highly correlated with the target variable