The glass identification dataset from the UCI Machine Learning Repository contains data on different types of glass. The dataset has 214 observations and 10 features. The features include the refractive index; an optical characteristic which tells us how fast light travels through the material, the sodium and magnesium content, and the aluminum and silicon content, among others... The target variable is the type of glass, which can be one of seven different types. The machine readability of the glass identification dataset and meta-data could be improved. The glass identification dataset is currently in a plain text file format. This file format may not be the most efficient for loading and processing large datasets. Considering a more compact file format such as CSV will be more efficient. The current column names in the glass identification dataset are not very informative. Considering giving more descriptive names to the columns, such as "refractive_index" will make the dataset easier to understand.

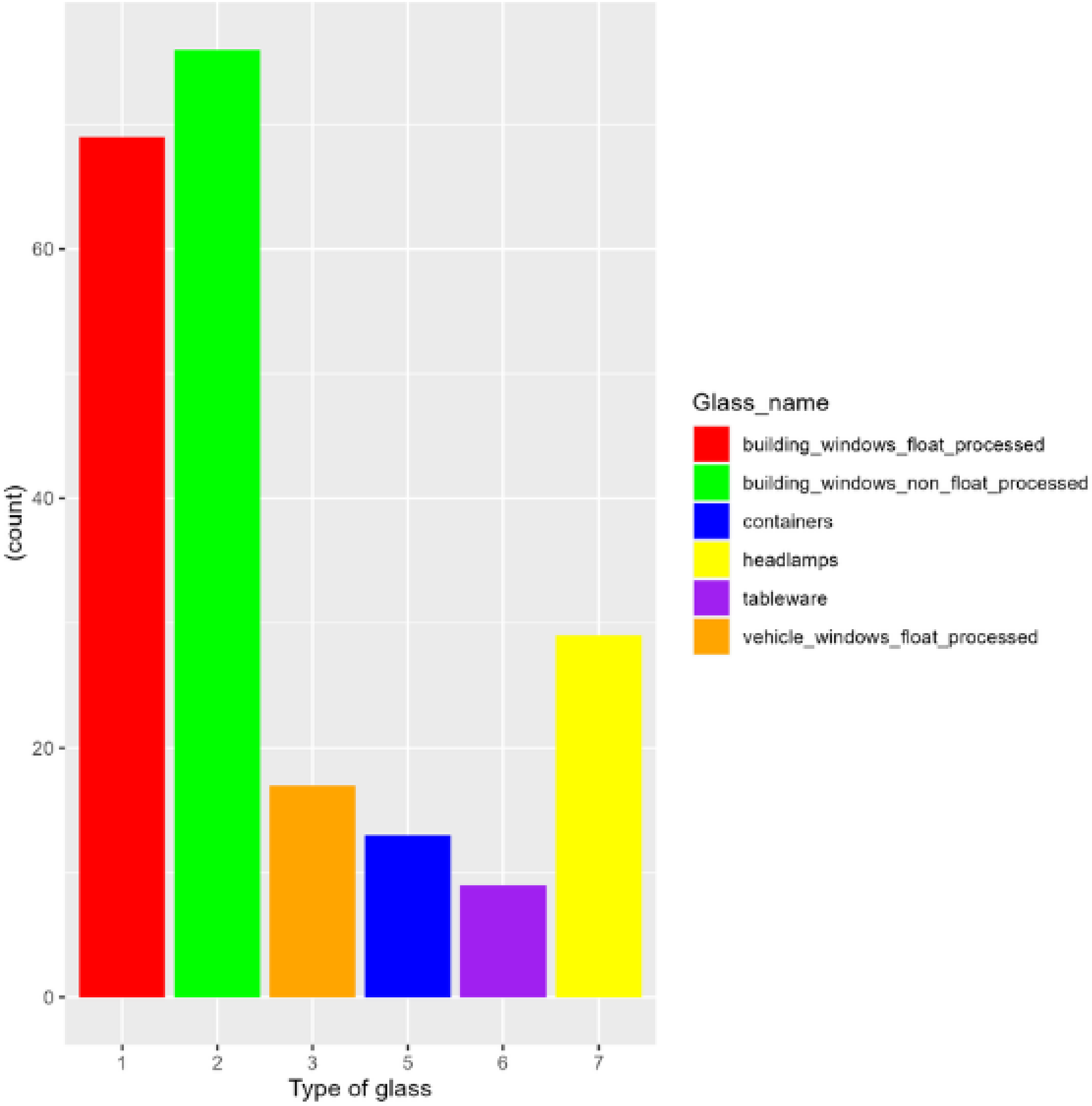
Data preparation :

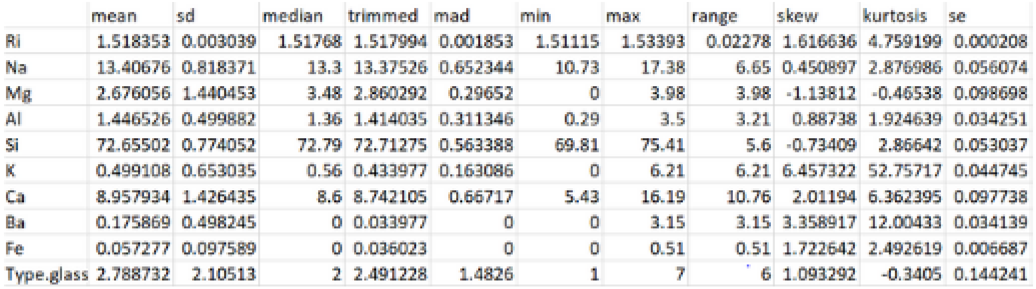
- After loading the data, we added a column to the dataset that includes the name of each type of glass corresponding to the correct row.
- The dataset does not contain any missing values.
- The Type of glass column was factorized for better management of the data.

Descriptive Statistics:

Describe function gives us some statistics about the dataset shown in the table below. *RI* varies around 1.5, from research we found that typical *RI* values for different sorts of glass lie in the range of 1.49 … 1.69. The RI values seem to be correct. Silicon is the most important oxyde of glass with a range of [68.81-75.41], followed by Natrium and Calcium.

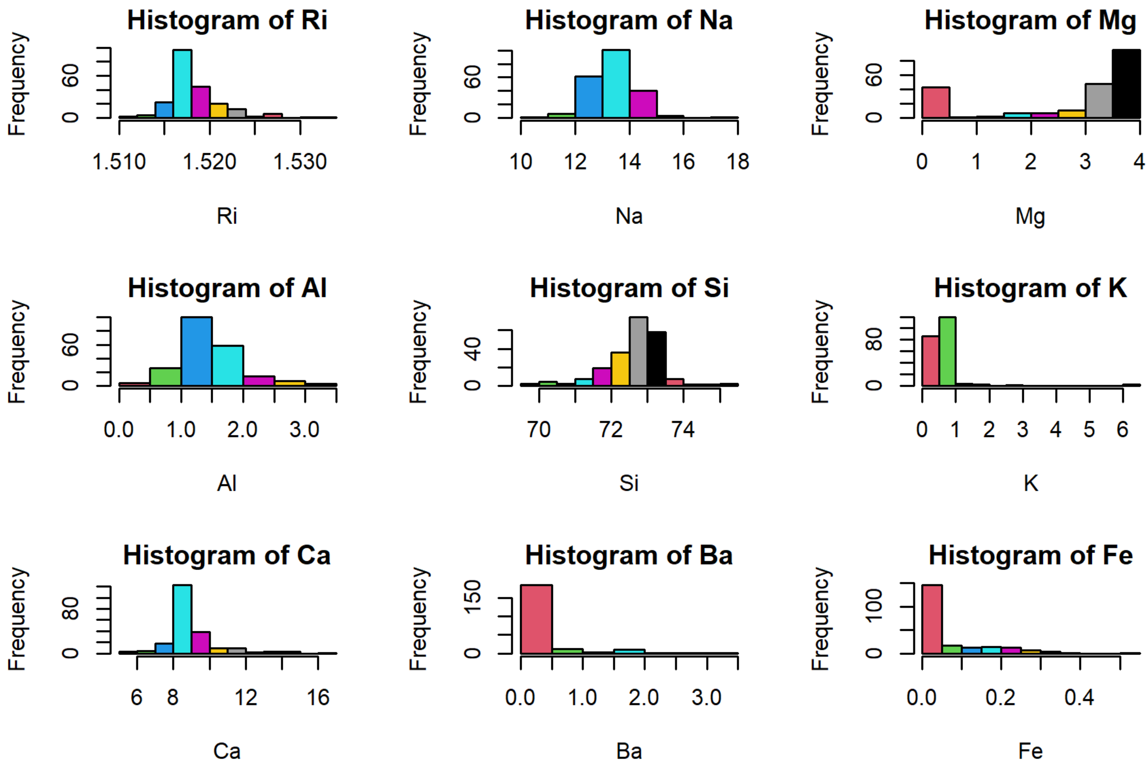
In the plot below, we see the frequency of each type of glass. Only 6 of them appear in the glass dataset. The type 4 *'Vehicle_windows_non_float_processed'* does not occur. *'building_windows_non_float_processed*' is the most common followed by *'building_windows_float_processed'*.



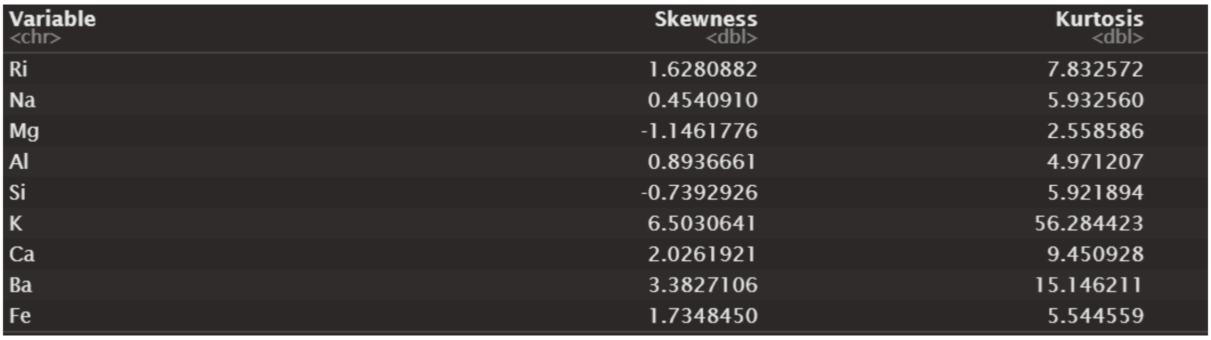


	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
Ri	1.518353	0.003039	1.51768	1.517994	0.001853	1.51115	1.53393	0.02278	1.616636	4.759199	0.000208
Na	13.40676	0.818371	13.3	13.37526	0.652344	10.73	17.38	6.65	0.450897	2.876986	0.056074
Mg	2.676056	1.440453	3.48	2.860292	0.29652	0	3.98	3.98	-1.13812	-0.46538	0.098698
Al	1.446526	0.499882	1.36	1.414035	0.311346	0.29	3.5	3.21	0.88738	1.924639	0.034251
Si	72.65502	0.774052	72.79	72.71275	0.563388	69.81	75.41	5.6	-0.73409	2.86642	0.053037
K	0.499108	0.653035	0.56	0.433977	0.163086	0	6.21	6.21	6.457322	52.75717	0.044745
Ca	8.957934	1.426435	8.6	8.742105	0.66717	5.43	16.19	10.76	2.01194	6.362395	0.097738
Ba	0.175869	0.498245	0	0.033977	0	0	3.15	3.15	3.358917	12.00433	0.034139
Fe	0.057277	0.097589	0	0.036023	0	0	0.51	0.51	1.722642	2.492619	0.006687
Type.glass	2.788732	2.10513	2	2.491228	1.4826	1	7	6	1.093292	-0.3405	0.144241

To better understand the distribution of the oxydes, we plotted a list of histograms for each as shown below :



We notice that the data has both positive and negative skeweness. Potasium (K) and Barium (Ba) are the oxydes with the higher skewness, their distributions are asymmetric and long-tailed. We notice as well that the oxydes do not follow a normal distribution. We can also presume the presence of outliers in Fe, Ba and K since we see some instances with high frequency.



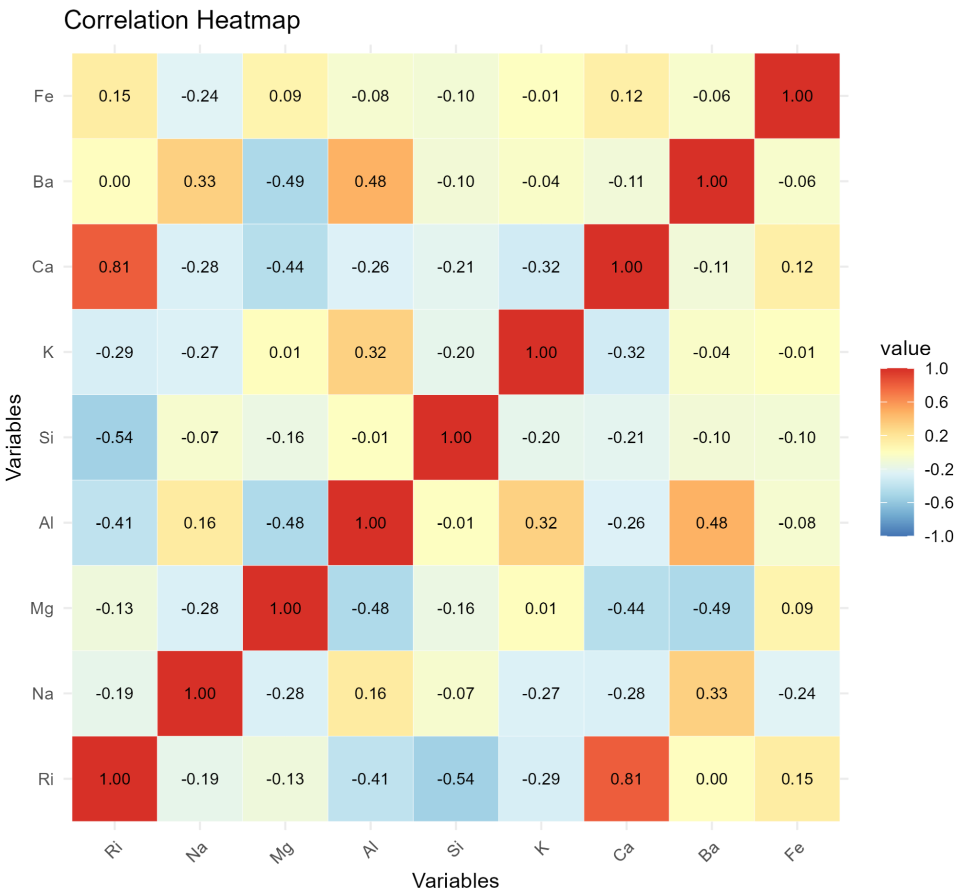
Variable <chr>	Skewness <dbl>	Kurtosis <dbl>
Ri	1.6280882	7.832572
Na	0.4540910	5.932560
Mg	-1.1461776	2.558586
Al	0.8936661	4.971207
Si	-0.7392926	5.921894
K	6.5030641	56.284423
Ca	2.0261921	9.450928
Ba	3.3827106	15.146211
Fe	1.7348450	5.544559

Correlation Matrix:

It is important to understand the correlation between numerical attributes when working with them, as certain models may not function well when attributes are strongly correlated. Therefore, we plotted the below correlation heatmap. We note that the highest correlation values are between :

Ca - RI = 0.8124949
K - Al = 0.324483684
Ba -Al = 0.478935953
Ba - Na = 0.32723299
and the lowest are between:
Si - Ri = -0.540009928
Al - Ri = -0.405670651
Al- Mg = -0.480035475
Ba - Mg = -0.492148823

From the scatterplot matrix, the most apparent relationship is that Ri increases as Ca increases.



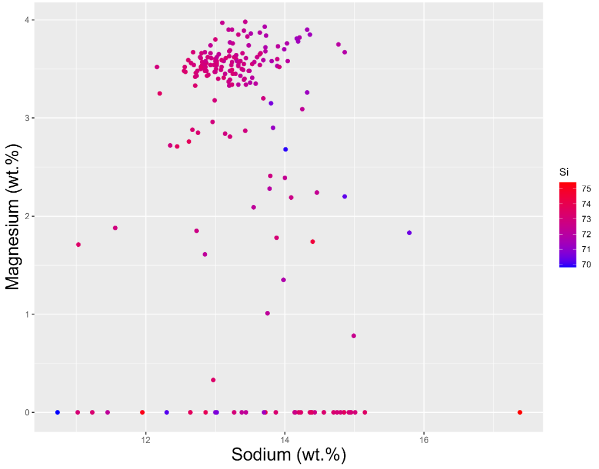
Multivariate relations:

The first scatterplot of Sodium and Magnesium colored by Silicon is important as a multivariate relationship because it allows us to see how the relationship between Sodium and Magnesium changes based on the level of Silicon. We notice that when the level of Silicon is high in the glass, the points are mostly grouped together for high values of magnesium(between 3% and 4%) and Sodium level of 12% and 15%. This plot suggest that there is a strong relationship between the levels of silicon, magnesium and sodium in the glass. It could indicate a specific manufacturing process that produces glass with consistent levels of these elements, or that the presence of high levels of silicon is dependent on specific levels of magnesium and sodium

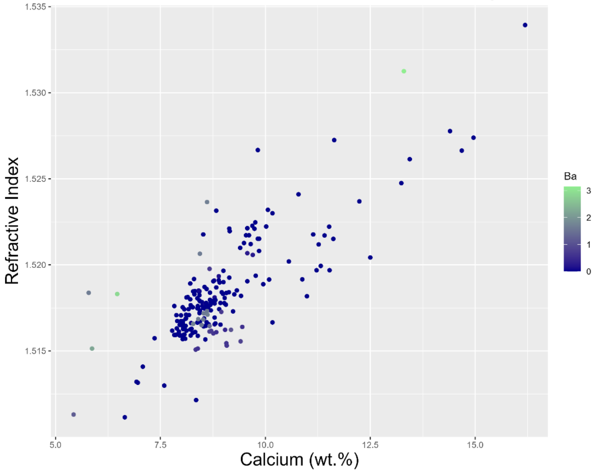
The second plot is about the distribution of calcium and the reflective index based on the changes in Barium levels. We notice that the points are mostly scattered between 7.5%-10% of calcium and 1515- 1520 of the reflective index when the barium level is between 0 and 2. This plot suggest that all the types of glasses have consistent levels of Barium. It could also suggest that the presence of a specific range of calcium and reflective index in the glass is dependent on the presence of specific levels of barium.

The third plot consists of a scatterplot of Refractive Index vs Iron with data points colored by Aluminum. This plot shows mostly low levels of aluminum for all values of Iron and refractive index, with a few exceptions when the iron level is between 0 and 0.1 on one hand and the refractive index below 1518 approximately

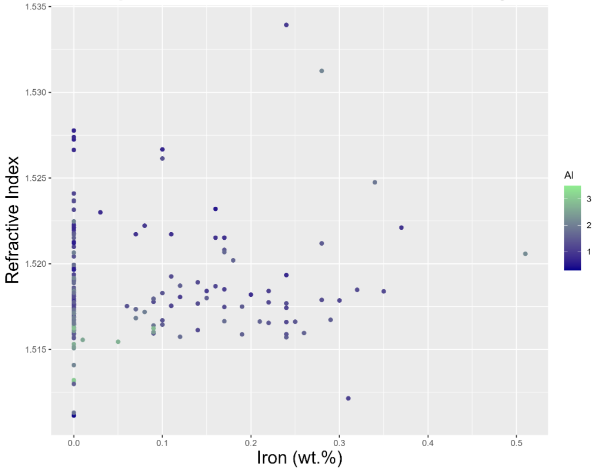
Scatterplot of Sodium and Magnesium colored by Silicon



Scatterplot of Calcium vs Refractive index colored by Barium



Scatterplot of Refractive Index vs Iron colored by Aluminum


Refractive Index
Iron (wt.%)
Al