Report 1

Dataset: ‘Glass identification’

Context: This dataset provides a set of oxide compositions for different glass types, these types are identified based on their oxide mixture and each glass has a refractive index. The data source is from USA Forensic Science Service.

Available on: (<https://archive.ics.uci.edu/dataset/42/glass+identification>).

Description: The dataset includes 11 variables; each includes 214 attributes. The data set is complete and does not have missing values. The objects are:

1. ID number: 1 to 214

2. RI: *refractive index*

3. Na: *Sodium (unit measurement: weight percent in the corresponding oxide, as are attributes 4-10)*

4. Mg: *Magnesium*

5. Al: *Aluminum*

6. Si: *Silicon*

7. K: *Potassium*

8. Ca: *Calcium*

9. Ba: *Barium*

10. Fe: *Iron*

11. Type of glass: (class attribute) range from [1, 2, 3, 5, 6, 7] each refers to a particular type of glass.

Note: In the ‘Type\_of\_glass’ factor variable, level 4 is not present in the dataset.

Interpretation of glass types:

1: building\_windows\_float\_processed.

2: building\_windows\_non\_float\_processed.

3: vehicle\_windows\_float\_processed.

5: containers.

6: tableware.

7: headlamps.

Characteristics: Multi-variate dataset

Qualitative investigation:

The following table represents a checklist the data preparation including cleaning if needed:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Checkpoints | Validity (V/NV) | Number of issues | Observation | Handling | Result |
| Readability | NV | 2 | Unstructured separation + Unassigned object names | Comma separator (,) + assigning correspondent -objects names | Structured and Determined data objects |
| Data type conversion | NV | 1 | ‘Type of glass’ object type: *Int* | As.factor(type\_of\_glass) | Factor var with 5 levels |
| Missing data | V | 0 | *-* | - | - |
| Incoherent data | V | 0 | - | - | - |
| Potential Outliers | NV | 3 | [Na, K, Ca] | Identified and saved | outliers identified |
| Duplicates | V | 0 | - | - | - |
| Inconsistent format | V | 0 | - | - | - |

Data exploration:

After preparing the data from inconsistencies highlighted in the previous table, the step forward is to explore the relationships between variables that are used for identifying glasses.

* Overview of the structure of the data set
* Str (glass)

'data.frame': 214 obs. of 11 variables:

$ id : int 1 2 3 4 5 6 7 8 9 10 ...

$ RI : num 1.52 1.52 1.52 1.52 1.52 ...

$ NA : num 13.6 13.9 13.5 13.2 13.3 ...

$ Mg : num 4.49 3.6 3.55 3.69 3.62 3.61 3.6 3.61 3.58 3.6 ...

$ AI : num 1.1 1.36 1.54 1.29 1.24 1.62 1.14 1.05 1.37 1.36 ...

$ Si : num 71.8 72.7 73 72.6 73.1 ...

$ K : num 0.06 0.48 0.39 0.57 0.55 0.64 0.58 0.57 0.56 0.57 ...

$ Ca : num 8.75 7.83 7.78 8.22 8.07 8.07 8.17 8.24 8.3 8.4 ...

$ Ba : num 0 0 0 0 0 0 0 0 0 0 ...

$ Fe : num 0 0 0 0 0 0.26 0 0 0 0.11 ...

$ Type\_of\_glass: Factor w/ 6 levels "1","2","3","5",..: 1 1 1 1 1 1 1 1 1 1 ...

* Summary (glass)

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | RI | NA | Mg | Al | Si | K | Ca | Ba | Fe | Type\_of\_glass  (Fctr) |
| Min | 1.511 | 10.73 | 0.00 | 0.290 | 69.81 | 0.00 | 5.430 | 0.00 | 0.00 | 1: 70  2: 76  3: 17  5: 13  6: 9  7: 29 |
| 1st QU | 1.517 | 12.91 | 2.115 | 1.190 | 72.28 | 0.1225 | 8.240 | 0.00 | 0.00 |
| Median | 1.518 | 13.30 | 3.480 | 1.360 | 72.79 | 0.5550 | 8.600 | 0.00 | 0.00 |
| Mean | 1.518 | 13.41 | 2.685 | 1.445 | 72.65 | 0.4971 | 8.957 | 0.175 | 0.05701 |
| 3rd QU | 1.519 | 13.82 | 3.6 | 1.630 | 73.09 | 0.61 | 9.172 | 0.00 | 0.1 |
| Max | 1.534 | 17.38 | 4.490 | 3.500 | 75.41 | 6.21 | 16.190 | 3.150 | 0.51 |
| SD | 0.003036 | 0.8166 | 1.442 | 0.4992 | 0.7745 | 0.652 | 1.4231 | 0.4972 | 0.0974 |

Observations: the key statistical measures provided in the summary of the glass dataset show that [Ba] and [Fe] have the lowest means compared to the other oxides with centralized data points spread around zero which may indicate a specificity or presence of potential outliers.

In terms of [RI], the mean equals the median which indicates a symmetrical distribution of data points with a spread that is limited between the quartiles [1.517, 1.519] in addition to a low standard deviation exhibiting fewer extreme values, this suggests that this oxide is potentially predictable since it has a stable pattern. Otherwise, [Na] and [Si] are slightly right-skewed (median< mean). In contrast, [K] is left-skewed.

The [Type\_of\_glass] factor variable shows a higher frequency for levels (2) and (1), while the lowest frequency is for the level (6).

* Barplot (numerical\_variables)

A bar chart of different types of glass

Description automatically generated

Figure . Bar Plot for the factor variable

The bar plot shows the frequency of glass types in this dataset, you can see that the types 2 and 1 represent 68.22 % of the overall frequency which is an important sample size for the analysis, while the rest of types represent 31.78 % of the glass types.

* Boxplot

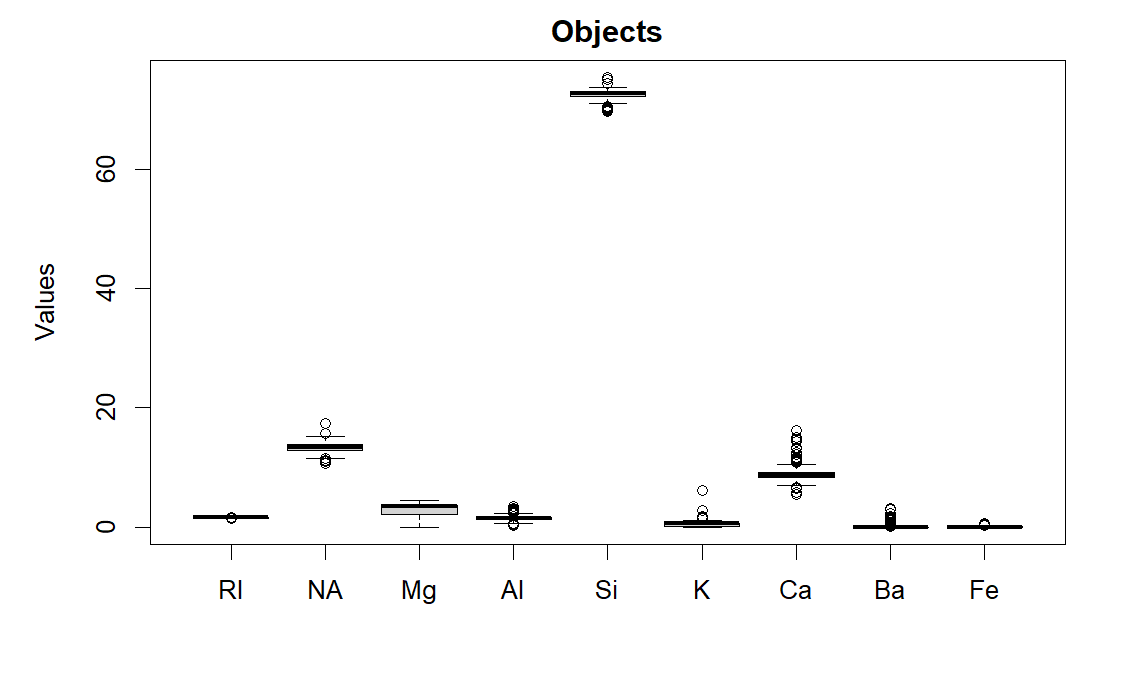


Figure . Boxplot

The upon boxplot illustrates some potential outliers to be considered such as [Ca][K][Na], also the [Mg] has a larger box which means a larger variation between the first and third quartiles.

Bivariate relations visualization

The following pairs plot visualizes the relations between numerical variables where the data points are colored based on their type of glass. The colors are identified as follows,

Types [1: Black] [2: Red] [3: Green] [5: Blue] [6: LightBlue] [7: Pink]

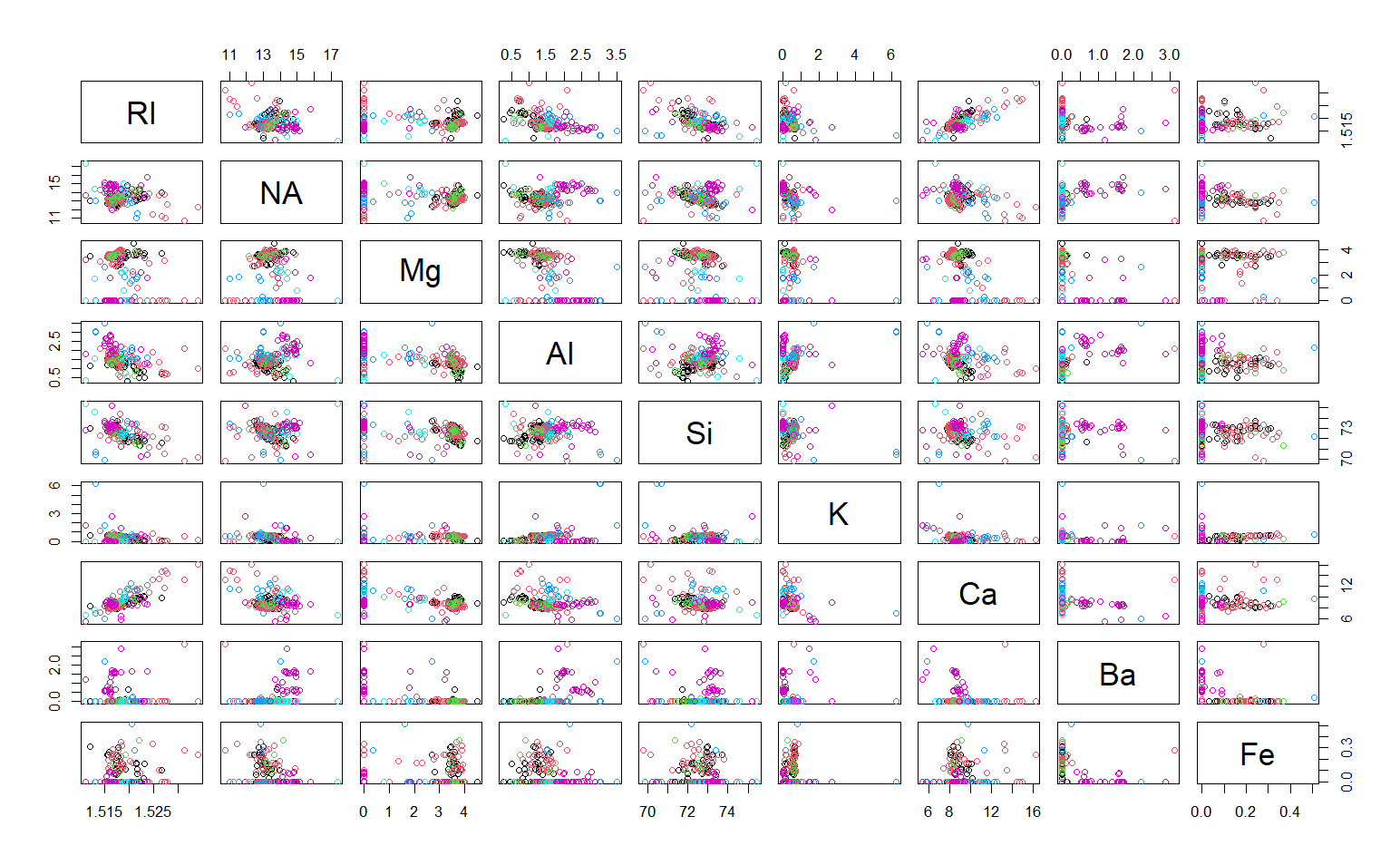




Figure . pairs(subset\_without\_id ,col=subset\_without\_id$Type\_of\_glass, pch = 1)

By Visual inspection, there are important observations that can be extracted from this multivariate plot where various patterns can be a subject of investigation. For instance, the plots indicated with a red point illustrate a clear correlation and the black points show a clustering for a specific type. Therefore, the relevant oxides to explore are [Ca], [Si] , [Al], [Ba], [Fe].

* Correlation inspection

The correlation heatmap matrix above illustrates the correlated relationship between numeric variables where the color gradient represents the strength and the direction of the correlation.

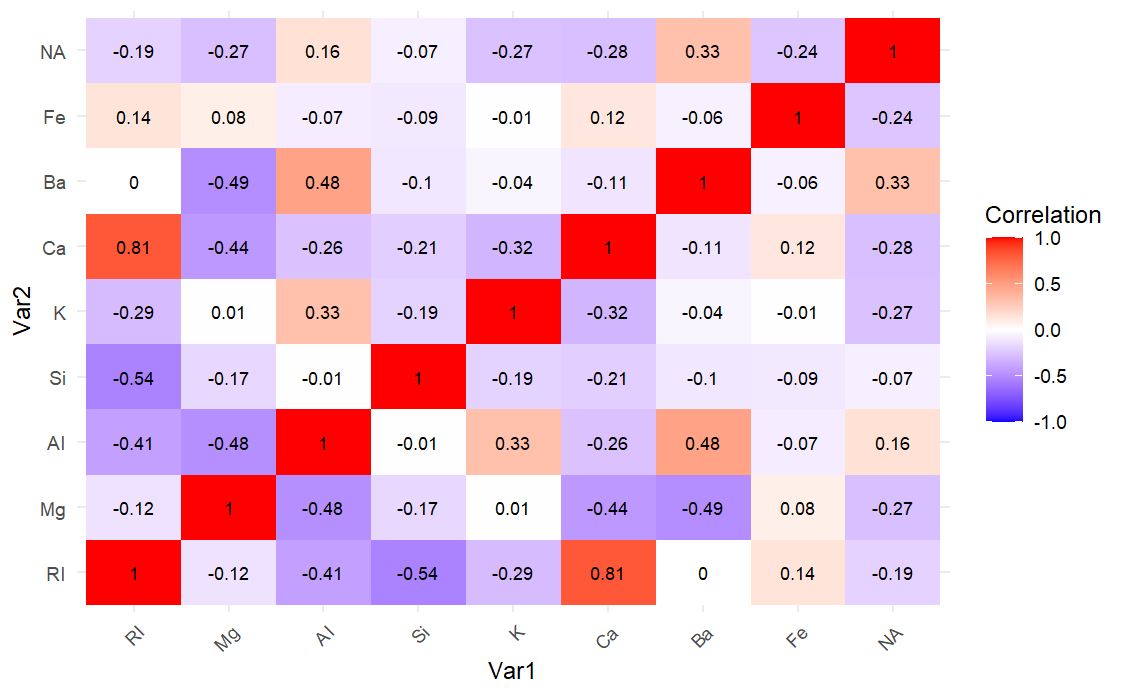


Figure . Correlation Heatmap matrix

These are the considerable correlations extracted from this figure:

Strongest positive correlation: [Ca] versus [RI].

Considerable positive correlations (0.3 <= coefficient < 0.7): [Al] vs [Ba], [Al] vs [K].

Strongest negative correlation: [Si] versus [RI].

Considerable negative correlations (-0.7 > coefficient <= -0.3): [Ca] vs [Mg], [Al] vs [RI].

* Investigation
* Determination of best predictors for [RI]:

Since we have the correlations, a linear model is fitted between the response variable [RI] and one of the potential predictors [Ca] [Si] [Al], with and without consideration for the interaction terms in order to evaluate which model provides the best balance between explanatory power and simplicity where the best predictor is the one with the lower AIC model value. The test results that the two oxides that best predict [RI] are [Ca] and [Si] with an (AIC= -2221.008).

The following plot illustrates the linear model between [RI] and [Ca] [Si] with consideration for the interaction terms between the explanatory variables.

A graph with blue dots on it

Description automatically generated

Figure . Linear model plot

* Determination of which oxides best predict the ‘Type of glass’:

According to the pairs plot, there are various features, clusters, and patterns that are a subject of investigation in which the following observation/question/answer aims to explore the relations that are needed to identify the predictors of the glass type.

Observation 1: Higher [Al] corresponds to a lower [RI], with Type 7 exhibiting the highest [Al].

Question 1 addresses whether non-float types have a lower [RI] and if a higher [Al] is exclusive to Type 7.

Answer 1, supported by a t-test, reveals that lower [RI] does not exclusively indicate non-float types. A higher [Al] potentially predicts glass types 5 or 7.

Sub-Question 1.1 identifies [Ba] as the distinguishing oxide between Types 5 and 7.

Observation 2: Types (1)(2)(3) form a cluster of [Mg] values.

Question 2 explores if a specific stable interval of [Mg] indicates Types (1)(2)(3).

Answer 2, backed by a significant difference in [Mg] between types (1) (2), suggests that Types (1)(3) share a cluster in terms of [Mg], specifically for FLOAT processed types.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 1st Qu. | Median | Mean | 3rd Qu. |
| Type (1) | 3.480 | 3.565 | 3.552 | 3.658 |
| Type (3) | 3.400 | 3.530 | 3.544 | 3.650 |

Sub-Question 2.1 acknowledges the need for further analysis and recommends external data or cross-validation for robust findings.

Observation 3: Type 7 shows an exception in containing [Ba] and lacks [Mg].

Question 3 investigates if the presence of [Ba] and the absence of [Mg] determine Type 7.

Answer 3 indicates that [Ba] is the sole distinguisher of Type 7 from Type 5.

Observation 4: Types (1)(2) contain an exception in [Fe].

Question 4 explores if the presence of [Fe] indicates a building window type.

Answer 4 reveals that the presence of [Fe] is not exclusive to building window types and identifies significant differences for Types 1\_2 with Types 6 and 7.

* key findings

The two oxides that best predict the refraction index are Calcium [Ca] and Silicon [Si].

The oxides that best predict the glass types are:

-Headlamps Type (7): [Al] and [Ba]

-Tableware type (6): [Fe]

-Containers type (5): [Al] and [Mg]

-Vehicle window float processed type (3): [Mg]

-Building window non-float processed type (2): [Mg] and [Fe]

-Building window float-processed type (1): [Mg]

Code:

