Glass identification dataset report

1. Introduction:

A criminological investigation motivated a study to classify glass on types. They can use the glass left in scene crime as evidence if it was correctly identified.

The objective of this report is to apply the methods and techniques acquired from this module for data exploration, and discover which factors are effecting the glass type and reflective index of glass.

2. Dataset description:

This database have 2140 observation in total, it contains 11 columns and 214 rows. There are one column that represents the Id number of the glass, and other 9 continuous columns that represents the chemical components of the glass, and one categorical variable that represents the glass type class:

1. Id number – Describes the 1-214 instances
2. RI – Refractive index of the glass
3. Na – Sodium
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

For the type of glass, the glass is classified on 7 classes depending on the oxides that are included in it:

- Class 1: Building windows (float processed)
- Class 2: Building windows (non-float processed)
- Class 3: Vehicle windows (float processed)
- Class 4: Vehicle windows (non-float processed)
- Class 5: Containers
- Class 6: Tableware
- Class 7: Headlamps

This database does not have any missing data.

3. Data cleaning:

First, I started by giving the columns names: Id, Ri, Na, Mg, Al, Si, K, Ca, Ba, Fe, Type. Then I deleted the column ID since it's not important in this analysis. After that, check if there is any outliers and delete them. The summary of this data indicates that data is not normally distributed; I applied a min-max transformation on it. At the end, we will have a data set of 10 columns and 204 row.

4. Data exploration:

According to the bar chart below in figure 1, it seems that building window glass have more values (1 and 2), while there is no vehicle glass in this data set.

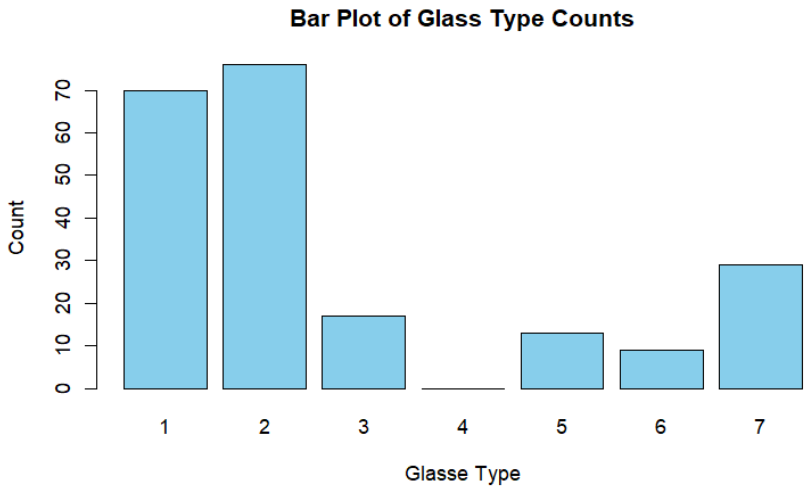

Figure 1: bar plot of glass type distribution

- **Trends and relations:**

The correlation plot in figure 2, shows the correlation between all factors in this data base. This plot shows some significant relations between the quantity of oxides and RI and the type.

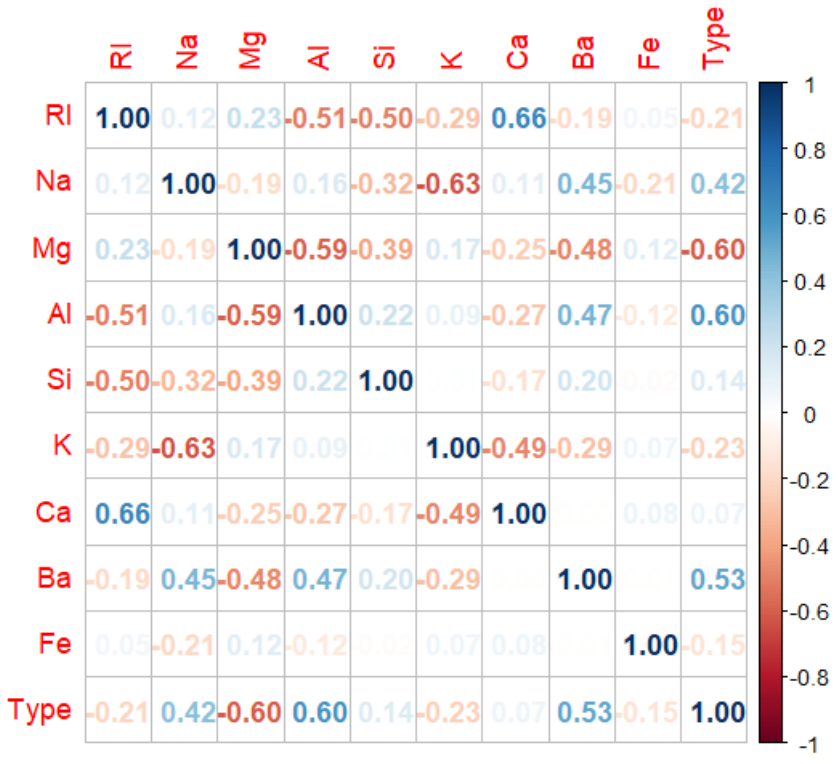

Figure 2: correlation plot

As it is clear that **the most two oxides that reflect the reflective index of glass are: calcium (Ca), aluminium (Al) and silicon (Si).** While type of glass is most correlated with **magnesium (Mg), Aluminium (Al) and barium (Ba).** There was also a remarkable correlation between components such as Na and K and Mg and AL, while there was some week relations between other variables.

RI have a strong positive correlation coefficient with Ca equals to 0.66, and negative correlation coefficient with Al and Si equals to -0.51 and -0.5 respectively. The same as with type of glass. *The type is positively correlated to Al and Ba with 0.6 and 0.53 respectively, and -0.6 with Mg.*

The plot below in figure 3 will focus on the two oxides that best predict the reflective index of glass which are Ca and Si.

As the plots demonstrate, there is a linear relation between RI and the two oxides. The higher quantity of Ca was in the glass, the higher RI goes. In the contrast, the higher quantity there was of Si the less RI will be.

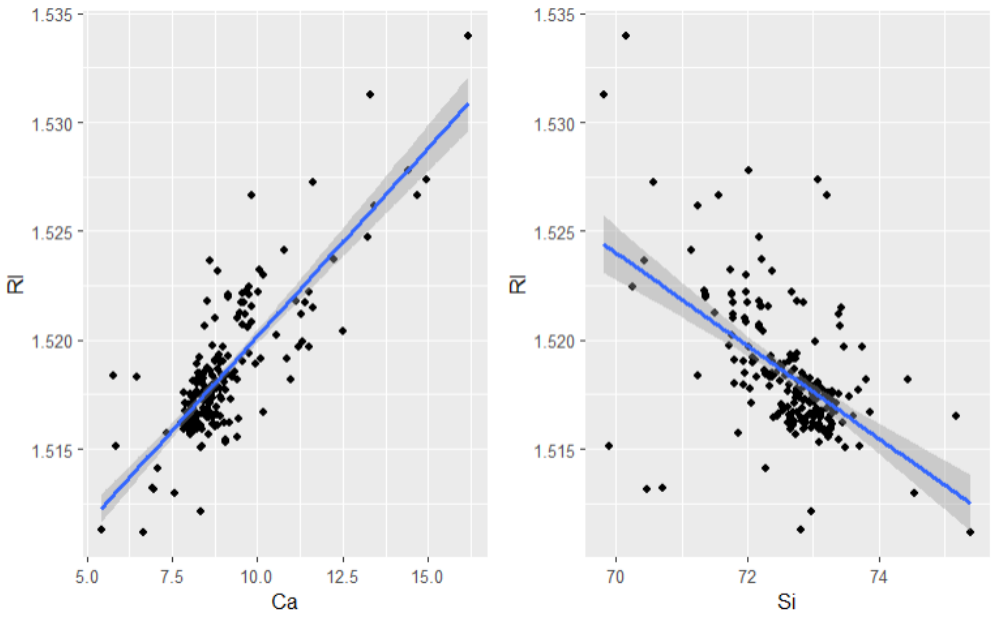

Figure 3: scatter plot for the relationship between RI and both Ca and Si

Next, we will focus on the two oxides that most predict the type of glass which are Mg and Al. To do this, a linear regression was performed.

The summary of this model (figure 4) confirms the results obtained in the correlation plot. Ba and Al are the most significant, following by Na and Ca. While Fe is the most insignificant oxide, which means: it does not predict the type of the glass. There is other interactions between oxides that predict the type, the most significant ones are: Na and Mg, K and Ca, Na and Ba respectively.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.256e+04  1.068e+04   3.049 0.002641 **
RI          -2.147e+04  7.045e+03  -3.047 0.002652 **
Na          -4.290e+02  1.196e+02  -3.587 0.000429 ***
Mg          -3.089e+02  1.004e+02  -3.077 0.002414 **
Al          -7.075e+02  1.712e+02  -4.133 5.45e-05 ***
Si          -2.941e+02  1.108e+02  -2.655 0.008644 **
K           -7.020e+02  2.254e+02  -3.115 0.002140 **
Ca          -3.673e+02  1.048e+02  -3.505 0.000575 ***
Ba          -1.609e+01  3.779e+00  -4.256 3.32e-05 ***
Fe           1.056e+02  1.114e+03   0.095 0.924641
RI:Na        2.827e+02  7.885e+01   3.585 0.000433 ***
RI:Mg        1.954e+02  6.621e+01   2.951 0.003584 **
RI:Al        4.718e+02  1.122e+02   4.206 4.07e-05 ***
RI:Si        1.940e+02  7.307e+01   2.655 0.008631 **
RI:K         5.259e+02  1.527e+02   3.445 0.000709 ***
RI:Ca        2.418e+02  6.902e+01   3.503 0.000579 ***
RI:Fe        1.018e+03  6.041e+02   1.686 0.093533 .
Na:Mg        5.995e-01  1.008e-01   5.945 1.38e-08 ***
Na:Al       -4.685e-01  2.504e-01  -1.871 0.062928 .
Na:K        -1.469e+00  4.025e-01  -3.650 0.000343 ***
Na:Ba        1.090e+00  2.628e-01   4.148 5.14e-05 ***
Na:Fe       -2.023e+01  5.268e+00  -3.840 0.000169 ***
Mg:Ca        3.564e-01  8.923e-02   3.995 9.40e-05 ***
Mg:Ba        4.517e-01  1.662e-01   2.718 0.007206 **
Mg:Fe       -1.239e+01  4.166e+00  -2.974 0.003339 **
Al:K        -2.474e+00  6.981e-01  -3.544 0.000500 ***
Al:Fe       -1.617e+01  5.907e+00  -2.737 0.006814 **
Si:K        -8.205e-01  3.133e-01  -2.619 0.009573 **
Si:Fe       -1.586e+01  4.721e+00  -3.360 0.000951 ***
K:Ca        -1.577e+00  3.381e-01  -4.666 5.95e-06 ***
K:Fe        -2.671e+01  7.990e+00  -3.343 0.001007 **
Ca:Fe       -1.803e+01  4.424e+00  -4.075 6.87e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 4: summary of the model

5. Conclusion:

In the conclusion, all the components effect the Ri and type of glass in general, but oxides of calcium and silicon are the predictors that most effect refractive index and Mg and Al are the best predictors for glass type. There is other interactions between oxides that also predict glass type such as interaction between Na and Mg, and other components that are not significant at this prediction such as Fe.

Example code:

```
# libraries
library(ggplot2)
library(GGally)
library(reshape2)
library(corrplot)
library(gridExtra)
# read data and name columns
data=read.table("C:/Users/Thinkpad/Desktop/data anal/project databases/glas
s identification/glass.data", sep=",")
names=c("Id_number","RI","Na","Mg","Al","Si","K","Ca","Ba","Fe","Type")
colnames(data)=names
# delete ID column
data=subset(data, select = c(-1) )
# get information of data
str(data)
summary(data)
# check if there is any outliers and delete them
out_ind=which(data$RI %in% c(out))
out_ind
data1= data[-which(data$RI %in% c(out)),]
# data normalization
min_max_norm = function(x) {
    (x - min(x)) / (max(x) - min(x))}
data_norm= as.data.frame(lapply(data1[1:9], min_max_norm))
data_norm$Type =data1$Type
# check the data
summary(data_norm)
nrow(data_norm)
# figure 1 : bar plot of glass type distribution
table=data.frame(type=c(1,2,3,4,5,6,7),count=c(70,76,17,0,13,9,29))
barplot(table$count, names.arg = table$type, col = 'skyblue',
        main = 'Bar Plot of Glass Type Counts', xlab = 'Glasse Type', ylab
= 'Count')
# correlation plot
correlation = cor(data_norm, method = 'spearman')
corrplot(correlation, method="number")
ggpairs(data_norm) # pair plot
# scatter plot Ca vs RI
plot1=ggplot(data,aes(x = Ca, y = RI)) + geom_point() +geom_smooth(method =
 "lm")
# scatter plot Si vs RI
plot2=ggplot(data,aes(x = Si, y = RI)) + geom_point() +geom_smooth(method =
 "lm")
grid.arrange(plot1, plot2, ncol=2)
# model to see interactions
model = lm(Type~.*., data = data)
smodel = step(model, trace = FALSE)
summary(smodel)
```