GLASS IDENTIFICATION DATA ANALYSIS

Akhil A. Naik

Contents

1 Introduction

Forensic studies frequently use glass fragments discovered at crime scenes as an evidence to correctly identify criminals thus solving the problem. In crime scenes such as house breaks where criminals can enter through shattering a window, even a small fragment of glass attached on a suspected persons cloths can solve the crime if the glass is correctly identified. Motivated by criminology investigation, *Vina Spiehler*, a personnel of *Diagnostic Products Corporation*, conducted a comparison test of her rule-based system, BEAGLE which is a product available through *VRS Consulting, Inc.*, the nearest-neighbor algorithm, and discriminant analysis (German and Vina Spiehler 1987).

Using attributes of a glass like refractive index and different types of oxides contents within it, following results were obtained to determine whether the glass was a type of *float* glass or not:

1. Windows that were float processed : 87
2. Windows that were not float processed : 76

This *Glass Identification* data set is made available in the UCI Machine Learning Repository (German and Vina Spiehler 1987). By employing this dataset, the objective is to determine by visual assessment the two oxides in glass that best predict the refractive index and the type of glass.

2 Exploratory Data Analysis

The glass identification dataset comprises of 214 observations/instances of 6 types of glasses defined in terms of their oxide content.

2.1 Attribute Information

There are total of 11 attributes present in the data set. One of the data set attributes is the Refractive Index, which is a dimensionless number that gives the indication of the light bending ability of glass (*Refractive Index*, n.d.). Different formulas affect the mechanical, electrical, chemical, optical, and thermal properties of the glasses that are produced, some of which are mentioned in (*Chemistry of Glass* 2011). Hence the other attributes comprise of oxide content (total 8) in glass which is measured as a weight percent in corresponding oxides like Na, Fe, K etc. Lastly the target variable is the type of the glass which classified in 7 different types although there are only 6 types of glass observation available in this data set. There are 163 Window glass (building windows and vehicle windows) and 51 Non-window glasses (containers, tableware, headlamps).

Attributes:

1. Id number: 1 to 214
2. RI: refractive index
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)
    - building_windows_float_processed (1)
    - building_windows_non_float_processed (2)
    - vehicle_windows_float_processed (3)
    - vehicle_windows_non_float_processed (4) (Not present in this data set)
    - containers (5)
    - tableware (6)
    - headlamps (7)

2.2 Data Manipulation And Summary Statistics

Our first step is to import the data set and prepare it for further analysis. The data is available in the *glass.data* file, with comma seperated values. Note that columns/attribute names are not present in the same file and must be imported/added from a different *glass.names* file. All the explanatory variables are continuous numerical type as expected, but note that the response variable which classifies the type of glass is provided as an floating/numerical variable hence we convert it to a categorical type factor with 7 levels with each level corresponding to the type of glass as mentioned in above section. Following table shows first few observations of the data set:

Table 1: Glass Identification Data

Id_number	RI	Na	Mg	Al	Si	K	Ca	Ba	Fe	Type_of_glass
1	1.5210	13.64	4.49	1.10	71.78	0.06	8.75	0	0	1
2	1.5176	13.89	3.60	1.36	72.73	0.48	7.83	0	0	1
3	1.5162	13.53	3.55	1.54	72.99	0.39	7.78	0	0	1

Our next step is to explore the data for missing values as missing values can significantly affect our analysis. We also check the statistical parameters like mean, median, standard deviation etc. to get an overview of the data in terms of its central tendency and spread.

Table 2: Statistical Summary Of Explainatory Variables

	RI	Na	Mg	Al	Si	K	Ca	Ba	Fe
nbr.val	214.00	214.00	214.00	214.00	214.00	214.00	214.00	214.00	214.00
nbr.null	0.00	0.00	42.00	0.00	0.00	30.00	0.00	176.00	144.00
nbr.na	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
min	1.51	10.73	0.00	0.29	69.81	0.00	5.43	0.00	0.00
max	1.53	17.38	4.49	3.50	75.41	6.21	16.19	3.15	0.51
range	0.02	6.65	4.49	3.21	5.60	6.21	10.76	3.15	0.51
sum	324.93	2869.28	574.49	309.21	15547.30	106.37	1916.79	37.46	12.20
median	1.52	13.30	3.48	1.36	72.79	0.56	8.60	0.00	0.00
mean	1.52	13.41	2.68	1.44	72.65	0.50	8.96	0.18	0.06
SE.mean	0.00	0.06	0.10	0.03	0.05	0.04	0.10	0.03	0.01
CI.mean.0.95	0.00	0.11	0.19	0.07	0.10	0.09	0.19	0.07	0.01
var	0.00	0.67	2.08	0.25	0.60	0.43	2.03	0.25	0.01
std.dev	0.00	0.82	1.44	0.50	0.77	0.65	1.42	0.50	0.10
coef.var	0.00	0.06	0.54	0.35	0.01	1.31	0.16	2.84	1.71

As we can see from above table, the features are not on the same scale. For example Fe has a mean of 0.06 whereas Si has a mean of 72.65. This can pose problem as algorithms like logistic regression require features to be on the same scale to converge smoothly. Also, the data is not evenly balanced with respect to the types of glass. As we can see from the below table, the observations for glass type 1 and type 2 constitutes more than 65% of the total data which implies unbalance data.

Table 3: Number of Instances for each glass type

Type_of_glass	num_of_observations	percent_of_TotalData
1	70	32.7
2	76	35.5
3	17	7.9
5	13	6.1
6	9	4.2
7	29	13.6

2.3 Visualisations

In this section we check if the data distributions are normal, check for outliers, explore relationships between the features and study the trends in the data via visualizations. We got a high level overview from the statistical summary which indicated that the data is not normal. The following density plots (Fig 1) confirms non-normal data for most of the features with skewness with Fe, Ba and K exhibiting high skewness. Also the boxplots for each features show certain glass types with outliers, fpr example feature Ca has plenty of outliers for glass type 2. Outliers may affect the final results but in our case we shall not remove the data/manipulate them for the time being.

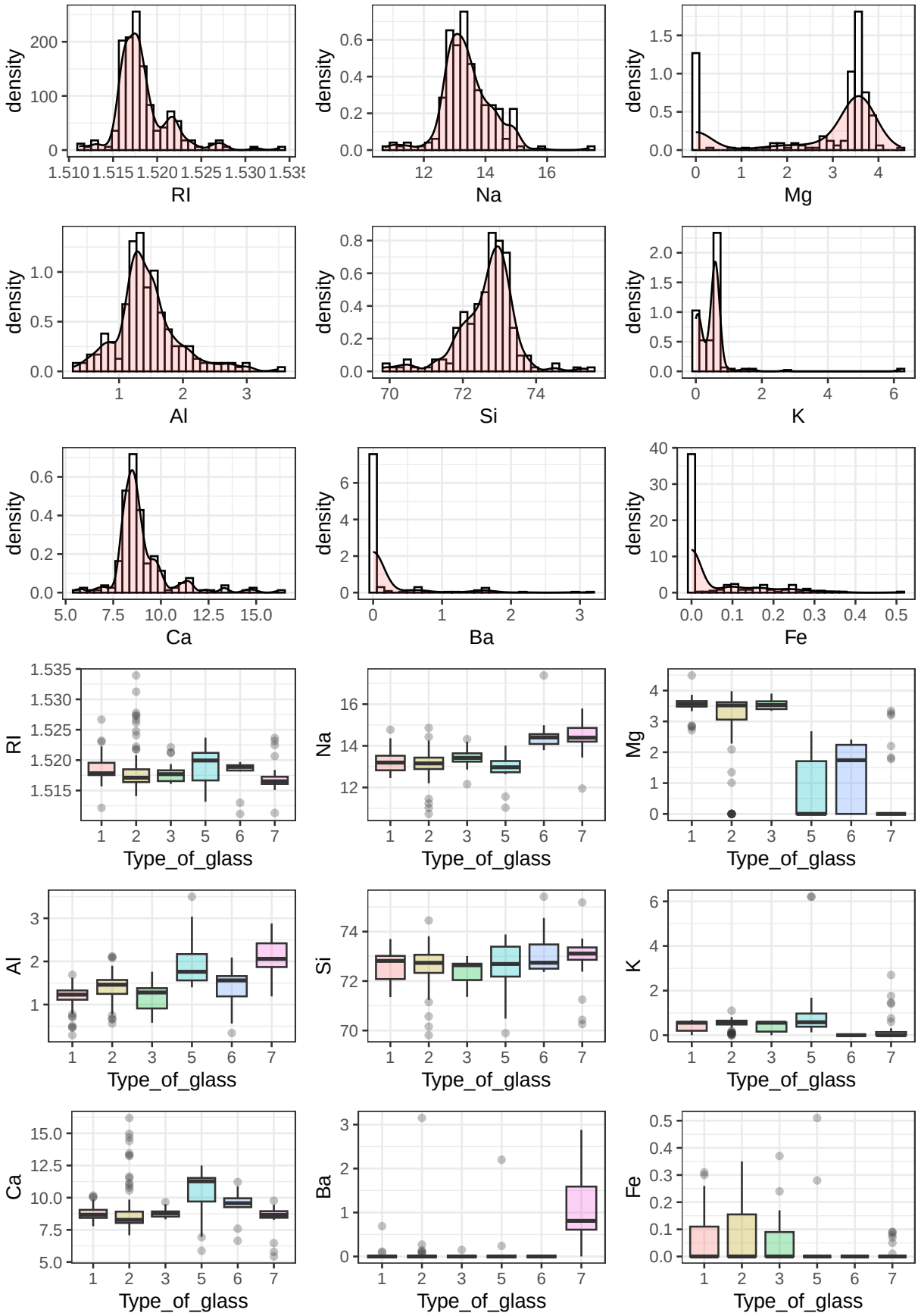

Figure 1: Density Plots and Box plots for each Glass Type

Figure 2: Pair Plots (Non transformed variables)

As we can notice from the above figure (Fig 2), due to non-normal data present and outliers, observations involving relations with features like Mg, K, Ba and Fe appear to be closely coupled or squeezed at one side of either axis. This can be avoided by normalising the features either using logarithmic scale or transforming the features along with removing the outliers if necessary. However, note that this can also be due to genuine properties of the

glass types which split them apart based on their oxide contents. We use Min-Max scaling method which allows us to scale the data values between a range of 0 to 1 and suppresses the effect of ouliers along with helping us have a smaller value of the standard deviation of the data scale (Mulani 2021; Sharkie 2020). Although the effect after scaling the data remains negligible for some features with excessive outlier, we can now check for relationship between the oxide features and refractive index along with type of glass.

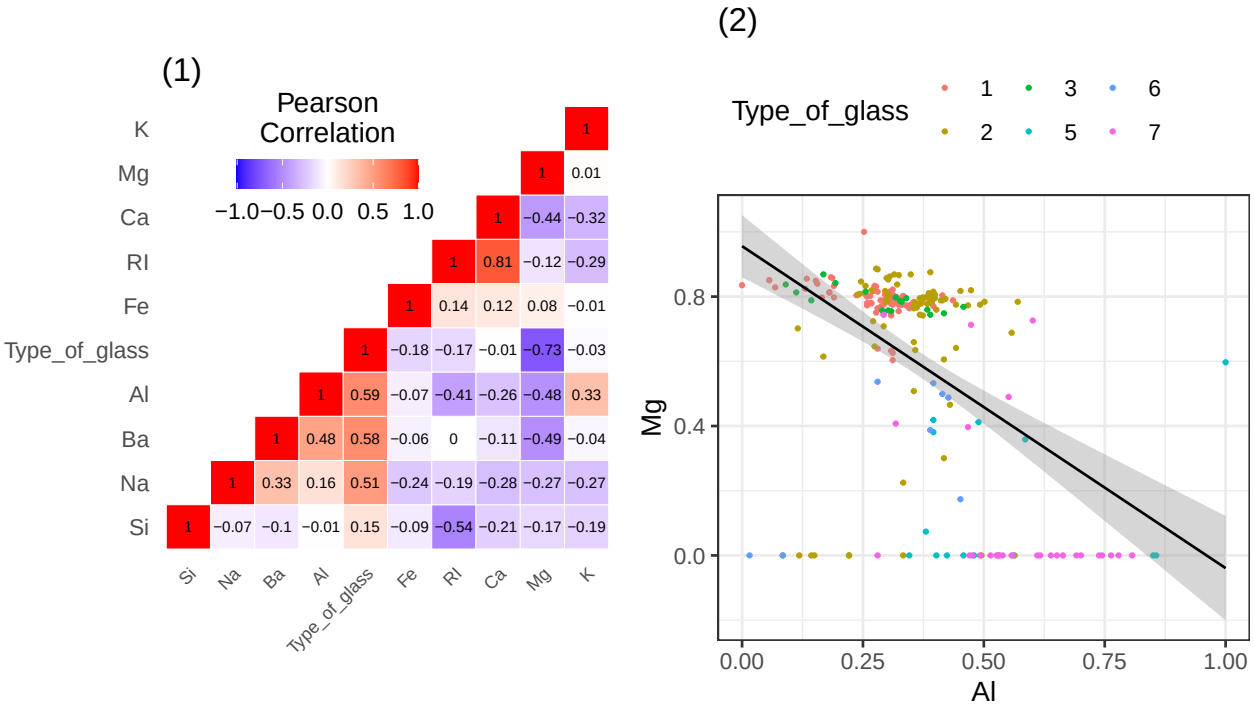

Figure 3: Correlation Heatmap (1) and Relation between two most correlated oxide contents with respect to glass type (2)

As we can notice in the pearson correlation coefficient heatmap for all the variables in the data set from the the above figure (Fig 3 (1)), Al has the strongest positive correlation (*0.59*) whereas Mg has the strongest negative correlation (*-0.73*) associated with the types of glasses. With plot between these two as seen above (Fig 3 (2)), we can primarily distinguish between the building windows (float & non-float processed) and vehicle windows float processed glass types (Type,1,2,3) which constitute majority of the points above the best fit line (high level of Mg and Al) and glass types mainly (Type,5,6,7) which lie below the line (mid-low level Mg and mid-high level Al). We also look ahead how the RI vary with the oxide contents and relate to the types of glasses in following figure. The below figure (Fig 4) shows how RI vary significantly with Si and Ca as Ri has strongest negative correlation with Si (*-0.54*) and strongest positive correlation with Ca *(0.81)*.

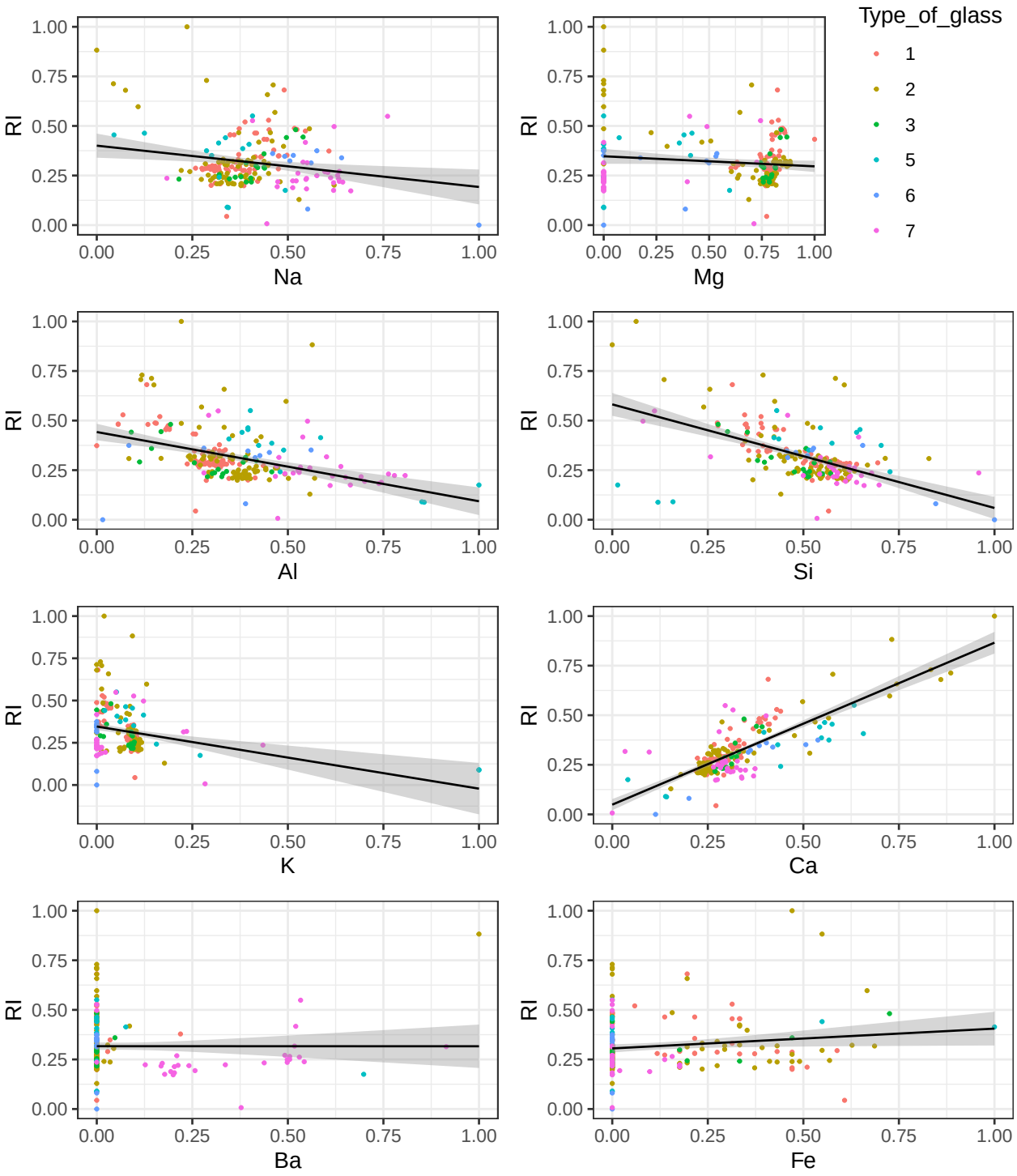

Figure 4: RI vs the oxide contents

Glasses with a lower RI and mid-high level Si are likely to be classified as non building or vehicle window glasses mainly headlamps whereas glasses with higher RI and low-mid level Si are more likely to be building windows. Similarly, best fit line between RI and and Ca also allows us to distinguish between the building and vehicle window glasses (likely to be above the line) and the remaining types which are likely to be below the line. Let us check

these best attributes for each glass type in following figure.

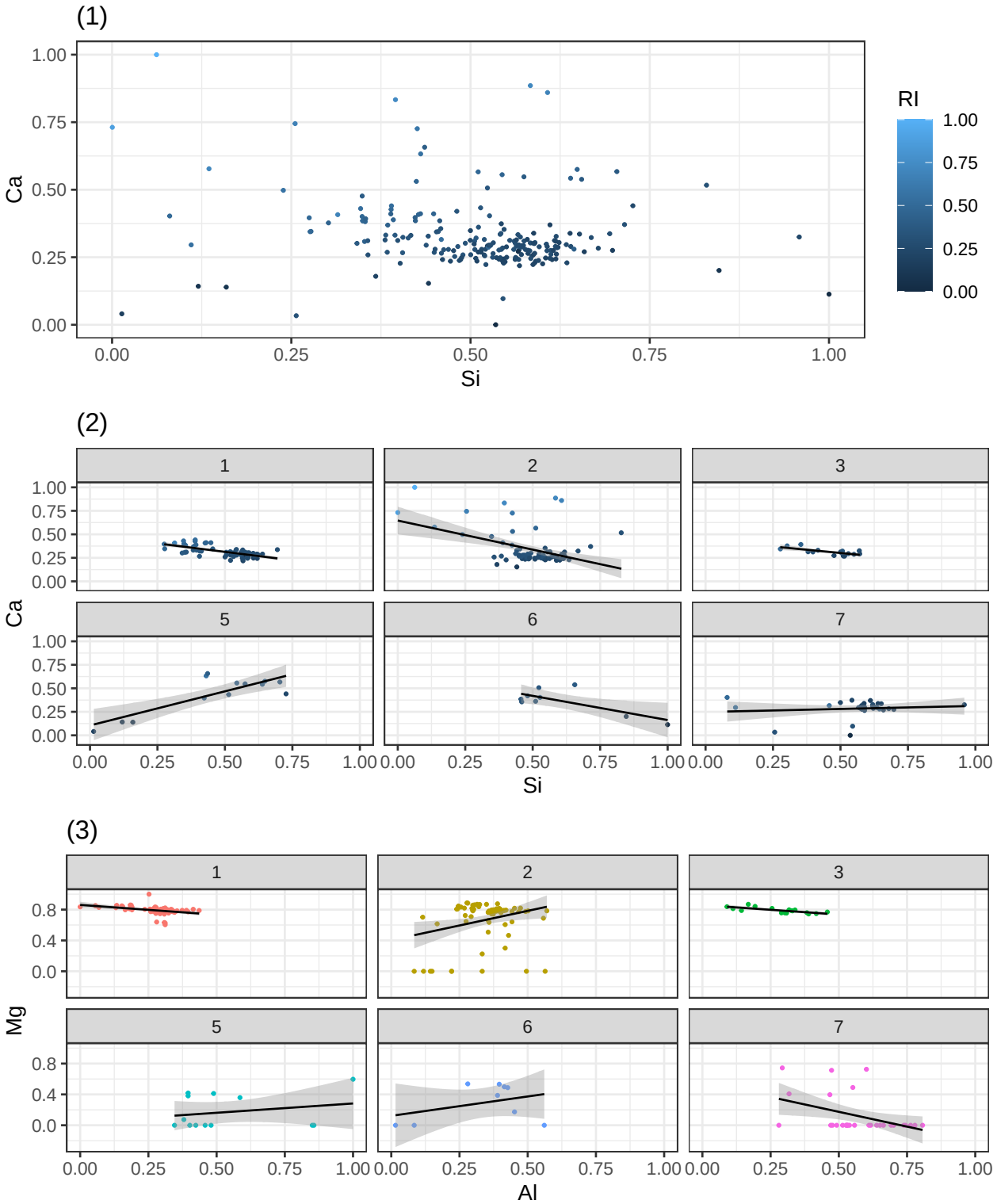

Figure 5: Relation of RI with the oxide contents (1) and its relation with respect to each glass type (2) and Mg vs Ai (3)

From above figure (Fig 5(1)) we can see how RI varies with Ca and Si. With higher the level of Si and lower the level of Ca, the refractive index is low. Whereas glasses with high level Ca and lower Si are likely to have higher refractive index. Similarly (Fig 5(3)) Mg and Al can be significant in determining the building and vehicle window glasses which have high Mg levels and low Al level and the remaining glass types which have lower Mg content with mid level Al content. We can also determine the individual glass types considering the above 3 plots, as float processed glasses (type 1 and 3) have higher Mg levels, low Si, low RI and mid to low level Si and Ca content.

3 Conclusion

We found out that Ca and Si oxide content can be significant in determining the refractive index of a glass. We also found that Mg and Al oxide content can be significant in determining the types of glasses mainly in distinguishing between, the building and vehicle window glasses which are highly associated with crime scenes vs the other daily home usage glass items. Presence of outliers and unbalanced and non normal data are likely to pose problem for better predictability. Also since the data observation are highly overlapped indicating that the types of glasses have correlation with almost every feature, visualisation with individual features alone may not bring justice for better predictions hence we may have to choose some dimensionality reduction techniques like Principal Component Analysis along with suitable machine learning algorithm for best prediction accuracy.

4 Code Snippets

```
# Importing data
columns = c('Id_number','RI','Na','Mg','Al','Si','K','Ca','Ba','Fe','Type_of_glass')
glass_data = read.table('glass.data',sep=',',col.names = columns)
# Converting type_of_glass to factor
data$Type_of_glass = as.factor(data$Type_of_glass)
```

```
#Density Plots and Box Plots
ggplot(data, aes(x=RI)) +
    geom_histogram(aes(y=..density..),# Histogram with density on y-axis
                   colour="black", fill="white") +
    geom_density(alpha=.2, fill="#FF6666") + # Overlay with transparent density plot
        theme_bw()
p10 = ggplot(data, aes(x=Type_of_glass, y=RI, fill=Type_of_glass)) +
    geom_boxplot(alpha=0.3, show.legend = FALSE)+
        theme_bw()
```

```
# Pair Plots
ggpairs(df, columns = 1:9, aes(color = Type_of_glass, alpha = 0.3),
```

```
        lower = list(continuous = wrap("points", alpha = 0.3, size=0.2)),
        upper = list(continuous = wrap("cor", size = 1.3))) +
theme( axis.line=element_blank(),
axis.text.x = element_text(face="bold",
                           size=4, angle=45),
axis.text.y = element_text(face="bold",
                           size=4),
panel.grid.major= element_blank())
```

```
# Min Max scaling
library(caret)
process = preProcess(df, method=c("range"))
norm_scale = predict(process, df) # Normalised data
```

```
# Cheking relation between the the highest correlated and
# lowest correlated oxide content with respect to glass type
p9 = ggplot(norm_scale, aes(y=Mg, x=Al, col = Type_of_glass))+
   geom_point(size=0.5) +
  ggtitle("(2)")+
  stat_smooth(method = "lm", col = "black", size=0.5) +
        theme_bw()+
  theme(legend.position='top',
        legend.justification='left',
        legend.direction='horizontal')
```

```
# Fig 5(1) - Finding relation of Refractive Index with varying Si and Ca
p11 = ggplot(norm_scale, aes(y=Ca, x=Si))+
   geom_point(size=0.5,mapping=aes(col=RI),
position='jitter') +
  ggtitle("(1)")+
  theme(legend.position='top',
        legend.justification='left',
        legend.direction='horizontal')+
        theme_bw()
```

```
# Fig 5(3) - Mg vs Al w.r.t types of glass
p13 = ggplot(norm_scale, aes(y=Mg, x=Al, col = Type_of_glass))+
   geom_point(size=0.5,show.legend = FALSE) +
  ggtitle("(3)")+ stat_smooth(method = "lm", col = "black", size=0.5) +
  facet_wrap(~ Type_of_glass)+
  theme(legend.position='top', legend.justification='left',
        legend.direction='horizontal')+theme_bw()
```

References

Chemistry of Glass. 2011. *Corning Museum Of Glass*. https://www.cmog.org/article/chemistry-glass.

German, B., and DABFT Vina Spiehler Ph.D. 1987. *Glass Identification Data Set. UCI Machine Learning Repository*. http://archive.ics.uci.edu/ml/datasets/glass+identification.

Mulani, Safa. 2021. *How to Normalize Data in r. Digital Ocean*. https://www.digitalocean.com/community/tutorials/normalize-data-in-r.

Refractive Index. n.d. *Wikipedia*. https://en.wikipedia.org/wiki/Refractive_index.

Sharkie, Data. 2020. *How to Normalize Data in r*. https://datasharkie.com/how-to-normalize-data-in-r/.