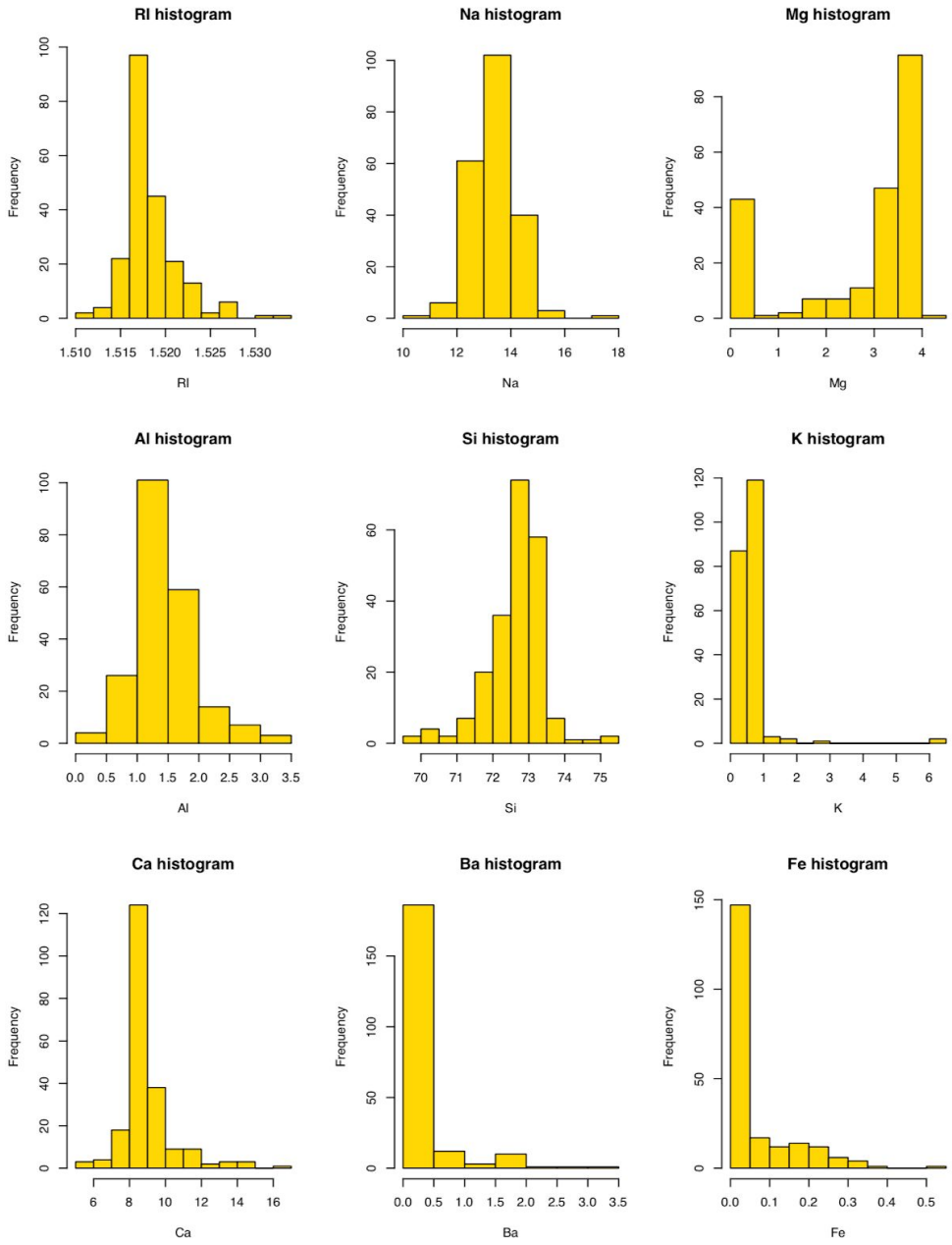
Assignment-1

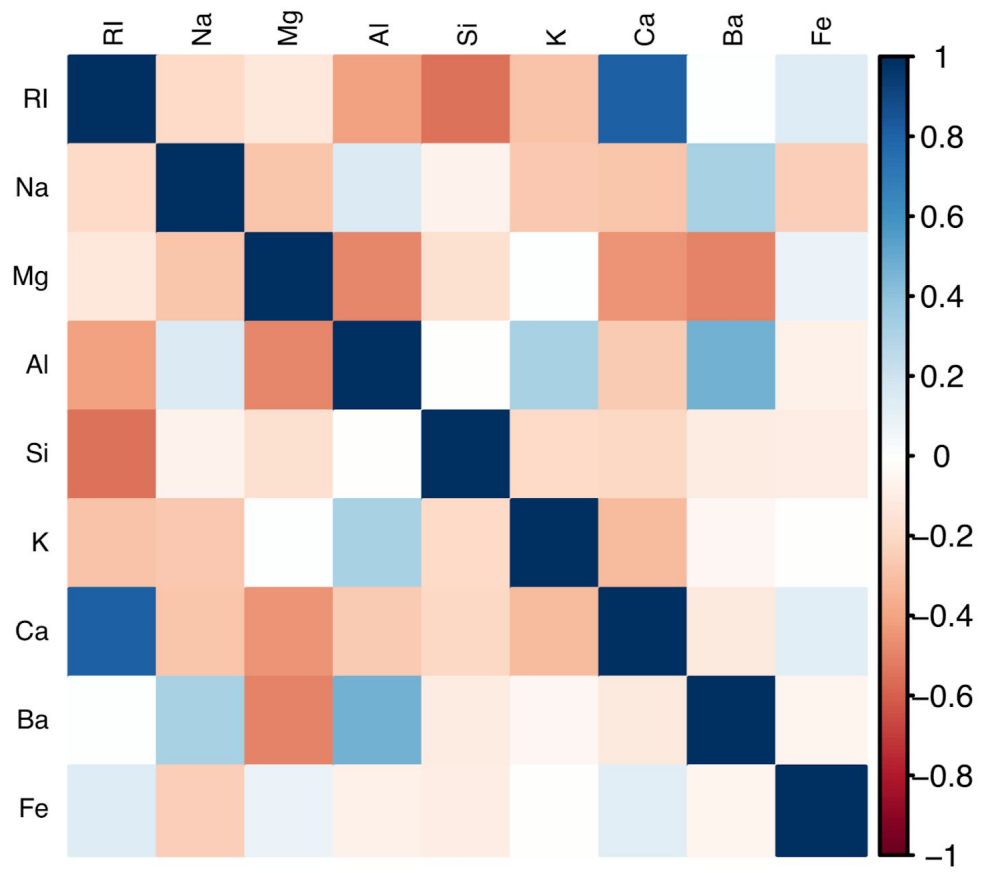
Sai Hemanth Nirujogi
G01065588

3.1. *The UC Irvine Machine Learning Repository contains a data set related to glass identification. The data consist of 214 glass samples labeled as one of seven class categories. There are nine predictors, including the refractive index and percentages of eight elements: Na, Mg, Al, Si, K, Ca, Ba, and Fe.*

(a) *Using visualizations, explore the predictor variables to understand their distributions as well as the relationships between predictors.*



As we can observe from the nine histograms, the predictors RI, Na, Al, Si have normal distributions than Mg, K, Fe, Ca, and Ba.

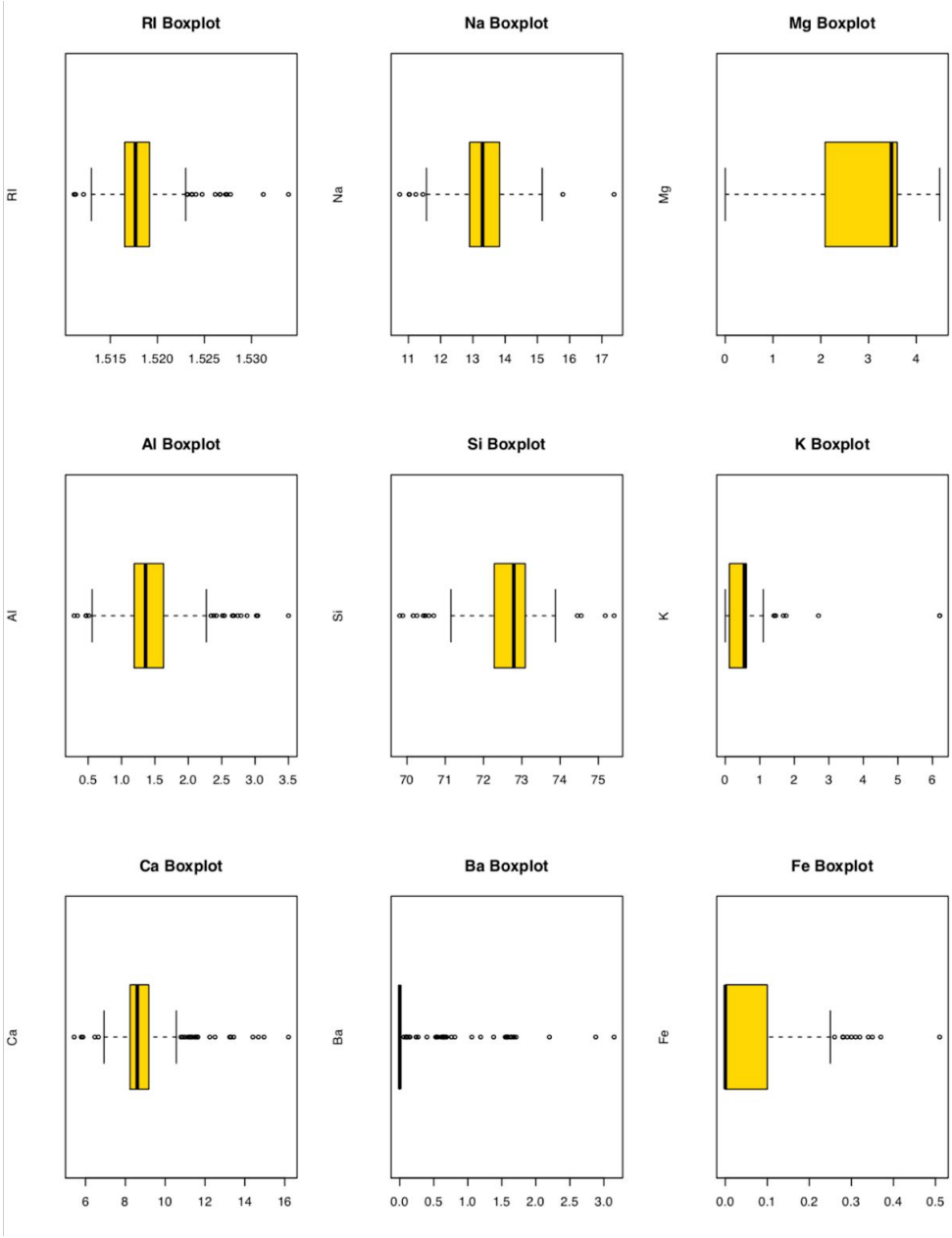


As we can observe from the Correlation plot above, RI, Ca pair shows a particularly positive relationship and the RI, Si pair shows a negative relationship.

b. *Do there appear to be any outliers in the data? Are any predictors skewed?*

- To determine any outliers in the data, Boxplots are used. The Boxplots indicate skewing in a few variables, however, does not work well for variables with extraordinary outliers.

- As we can see that the Boxplots show outliers in every plot except for Mg. From the Boxplots, it is clear that Mg is Left-Skewed and K, Ba, Fe, Ca are Right-Skewed.



C. *Are there any relevant transformations of one or more predictors that might improve the classification model?*

To improve the classification, Box-Cox Transformation is used to normalize and improve forecasting.

From the results, we can see that the Skewness is changed (reduced) after the transformation.

```
> apply( Glass[,-10], 2, skewness)
        RI         Na         Mg         Al         Si          K         Ca         Ba         Fe 
 1.6027151  0.4478343 -1.1364523  0.8946104 -0.7202392  6.4600889  2.0184463  3.3686800  1.7298107 
> apply( Glass[,-10], 2, boxcox )
        RI          Na          Mg          Al          Si           K          Ca          Ba          Fe 
 1.56566039  0.03384644 -1.13645228  0.09105899 -0.65090568  6.46008890 -0.19395573  3.36867997  1.72981071 
```

As we know from the transformation, we can see that K, Ba, Fe has high Skewness values so, we are transforming the variables by adding 0.000001.

```
> Glass$K=Glass$K+0.000001
> Glass$Ba=Glass$Ba+0.000001
> Glass$Fe=Glass$Fe+0.000001
>
> apply( Glass[,-10], 2, boxcox )
         RI          Na          Mg          Al          Si           K          Ca          Ba          Fe
 1.56566039  0.03384644 -1.13645228  0.09105899 -0.65090568 -0.78216211 -0.19395573  1.67566612  0.74424403
```

3.2. *The soybean data can also be found at the UC Irvine Machine Learning Repository. Data were collected to predict disease in 683 soybeans. The 35 predictors are mostly categorical and include information on the environmental conditions (e.g., temperature, precipitation) and plant conditions (e.g., left spots, mold growth). The outcome labels consist of 19 distinct classes.*

a. *Investigate the frequency distributions for the categorical predictors. Are any of the distributions degenerate in the ways discussed earlier in this chapter?*

By using the nearZeroVar(), we found the predictors with near zero variance are:
Leaf.mild, mycelium, and sclerotia

```
> data(Soybean)
> X=nearZeroVar( Soybean )
> colnames( Soybean )[ X ]
[1] "leaf.mild" "mycelium"  "sclerotia"
```

b. *Roughly 18 % of the data are missing. Are there particular predictors that are more likely to be missing? Is the pattern of missing data related to the classes?*

- The predictors 2-4-d-injury, cyst-nematode, diaporthe-pod-&-stem-blight, and phytophthora have missing data.
- The pattern of the missing data is related to the classes.

	FALSE	TRUE
2-4-d-injury	16	0
alternarialeaf-spot	0	91
anthracnose	0	44
bacterial-blight	0	20
bacterial-pustule	0	20
brown-spot	0	92
brown-stem-rot	0	44
charcoal-rot	0	20
cyst-nematode	14	0
diaporthe-pod-&-stem-blight	15	0
diaporthe-stem-canker	0	20
downy-mildew	0	20
frog-eye-leaf-spot	0	91
herbicide-injury	8	0
phyllosticta-leaf-spot	0	20
phytophthora-rot	68	20
powdery-mildew	0	20
purple-seed-stain	0	20
rhizoctonia-root-rot	0	20