Assignment No:1

Supriya Bachal

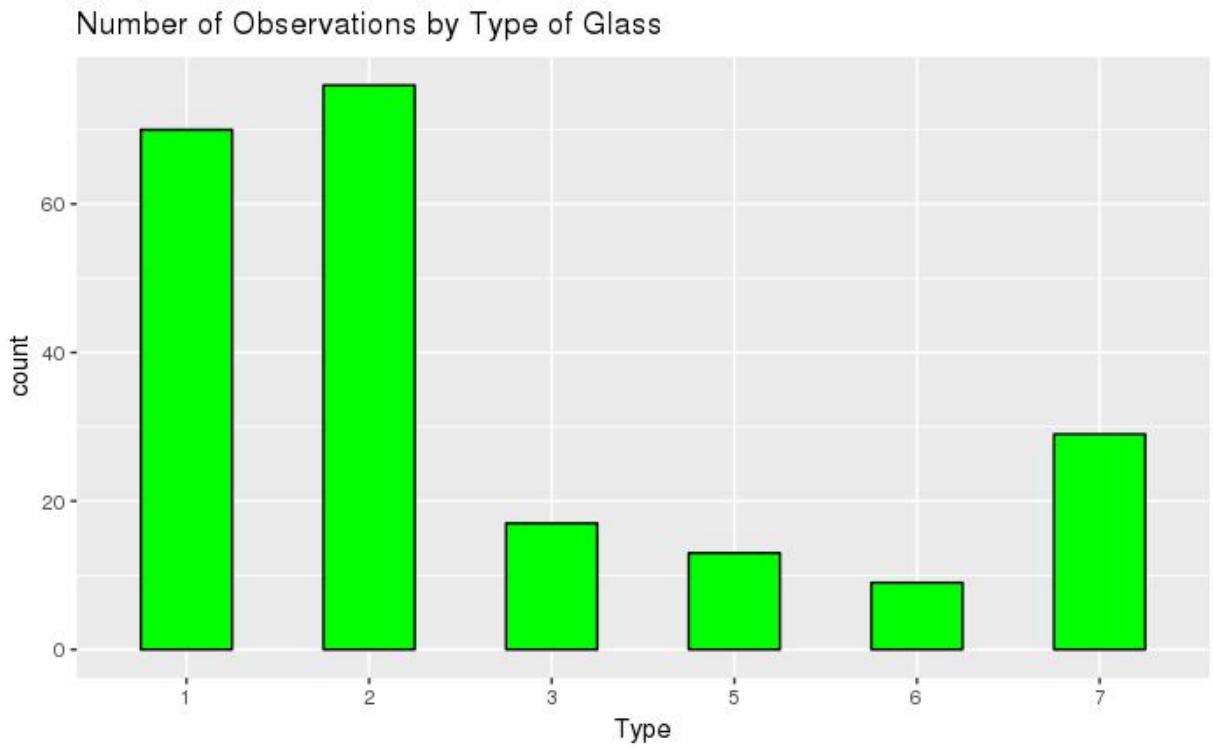
Question 3.1

(a) Using visualizations, explore the predictor variables to understand their

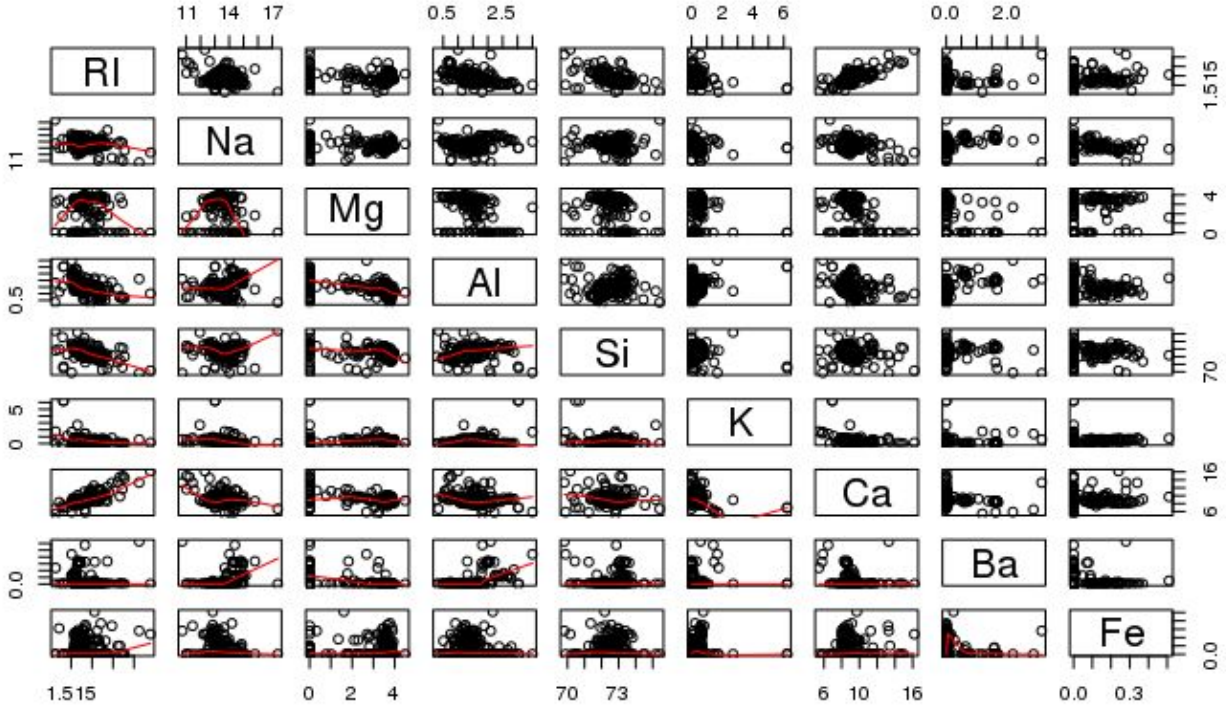
distributions as well as the relationships between predictors.

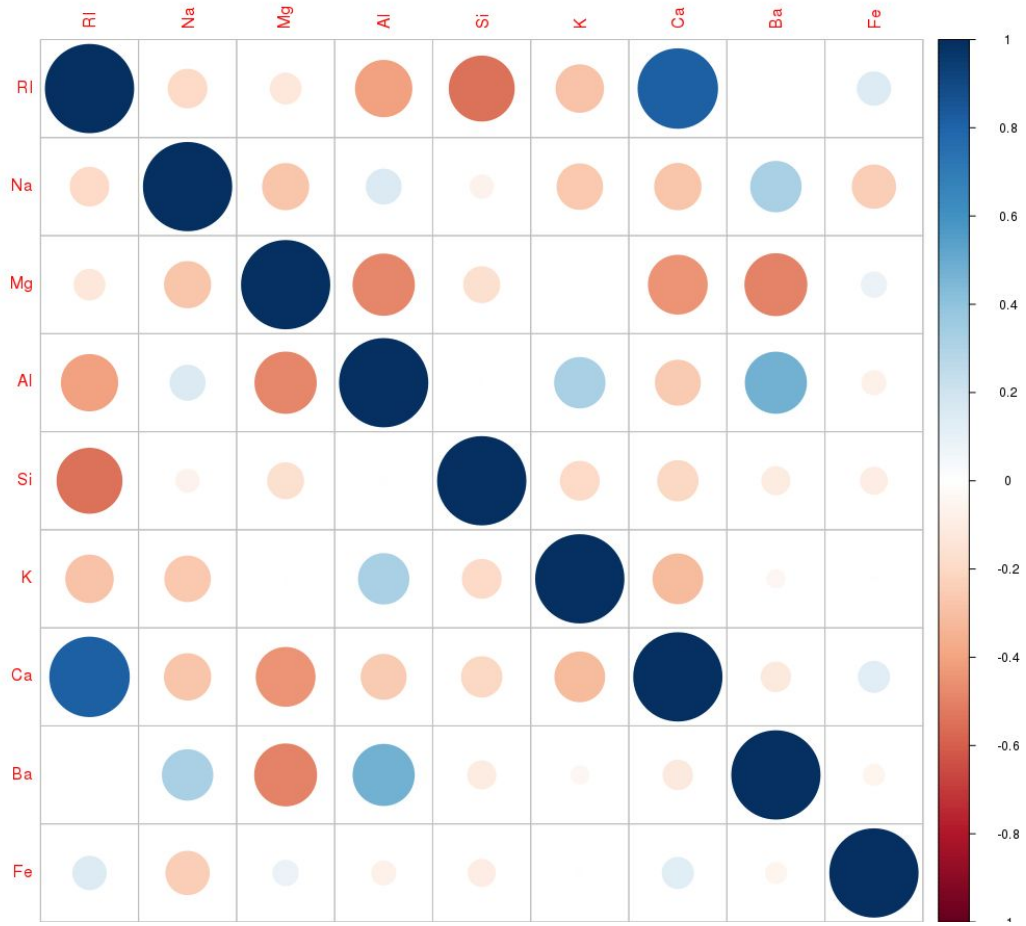
(b) Do there appear to be any outliers in the data? Are any predictors skewed?

(c) Are there any relevant transformations of one or more predictors that

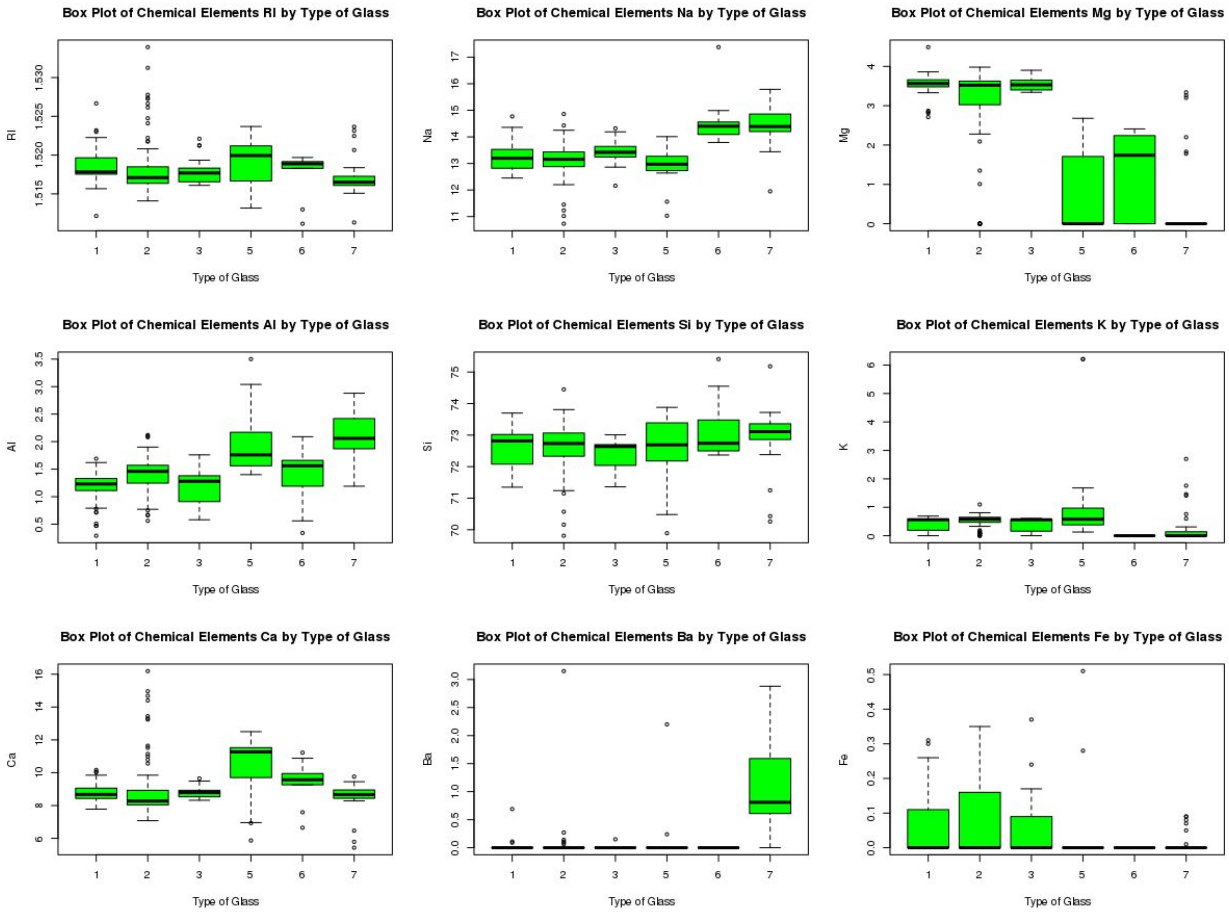


In the first step i identified the number of observations that belonged to each of the type of glass.The above histogram reveals that most of the observations are of glass Type 1 or 2.Least no of observations are of Type 6.Types 1 and 2 represent 68% of the total observations.


RI
Na
Mg
Al
Si
K
Ca
Ba
Fe

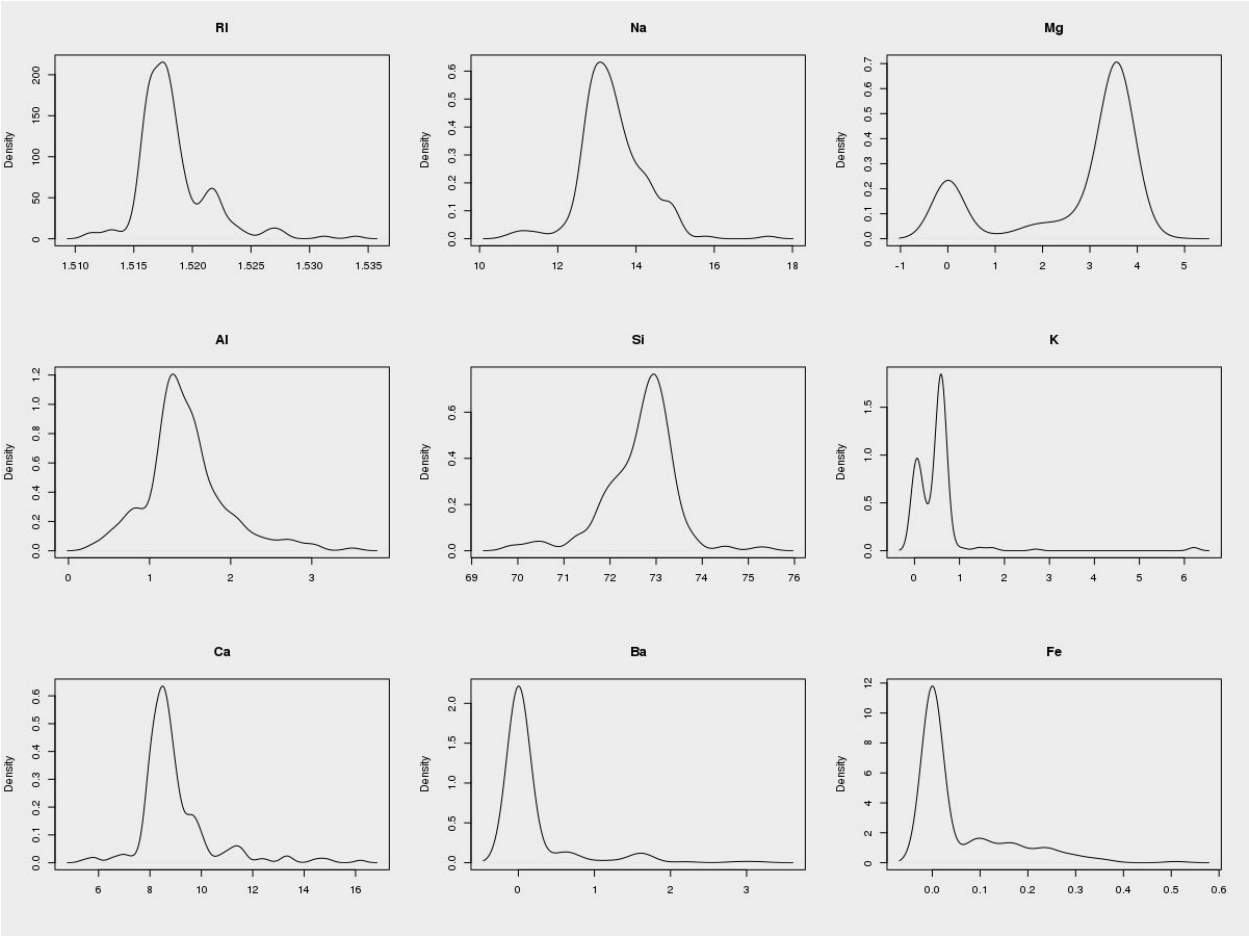


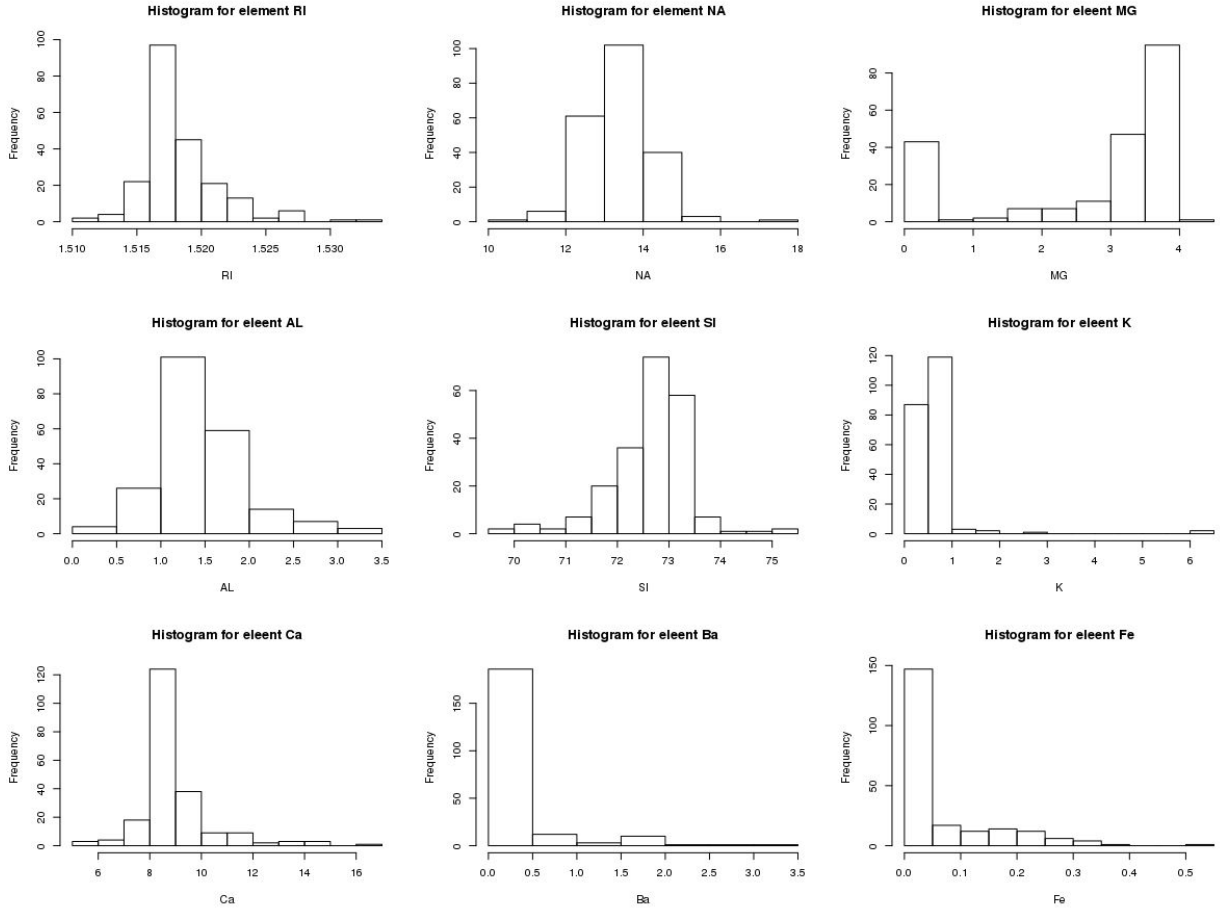
The above graphs show there is a very strong correlation between RI and Ca[0.8]. There is also a correlation between Ai and Ba as also between Ba and Na. There is also a correlation between K and Ai and so on but the strongest one is between RI and ca. The plot also shows that there is a negative correlation between Ri and Si and Ba and Mg.Seeing the high positive Correlation maybe we should drop one of these variables while building a model.



The above box plots show that there are a large no of outliers for Ca for Type 2 and for Mg for Type 2 similarly for Na for and Si. Most of the outliers are for Type 2. Additionally we also can note that for the elements Fe, Ba and k there are instances of 0 observations of a certain Types which implies that we can conclude that if a certain type of glass has traces of fe it cannot be of type 6 etc.The box plots also show that the amount of Fe and Ba in most types of glass is very low. For Mg the concentration is very high for Type 1 and Type 2 and Type 3.If the concentration of Ba is more than 0.1 there is a very high chance that the glass is of Type 7.

To further investigate this we can form histograms and density plots of the elements.


RI
Density
1.510
1.515
1.520
1.525
1.530
1.535
Na
Density
10
12
14
16
18
Mg
Density
-1
0
1
2
3
4
5
Al
Density
0
1
2
3
Si
Density
69
70
71
72
73
74
75
76
K
Density
0
1
2
3
4
5
6
Ca
Density
6
8
10
12
14
16
Ba
Density
0
1
2
3
Fe
Density
0.0
0.1
0.2
0.3
0.4
0.5
0.6



As we saw the distribution from the density plots , we see the histograms as well for the distribution of the elements.the peaks for RI is 1.515 and 1.520, for NA it is 12 to 14 , Mg it is 3 to 4 and so on. To futhur investigate the skewness of the data we use the skewness function which results in the following:

The Skewness of the various elements is as below:

RI 1.6027151	Al 0.8946104	Ca 2.0184463
Ba 3.3686800	Si -0.7202392	Na 0.4478343
Fe 1.7298107	Mg -1.1364523	K 6.4600889

We can conclude that the From the histogram that predictors RI,Mg and Fe are skewed. RI, Na and Al are slightly right skewed while Ca, Ba and Fe are very right skewed .K is double peaked and the density plots show that Mg and K are bimodal.

The most common transformations to eliminate skewness are log, square root or inverse transformations and/or outliers are spatial sign transformations. However there are many zero values in some of the elements hence we cannot use log transformations.We can try to find an appropriate transformation using Box -cox method from the Caret package. After scaling,centering and using Box cox transformations the skewness was not reduced to a great extent. However to improve results we could use the Yeo Johnson family of transformations to combat the problem of predictors having zero values.

Question 3.2:

a) Investigate the frequency distributions for the categorical predictors. Are

any of the distributions degenerate in the ways discussed earlier in this

chapter?

(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

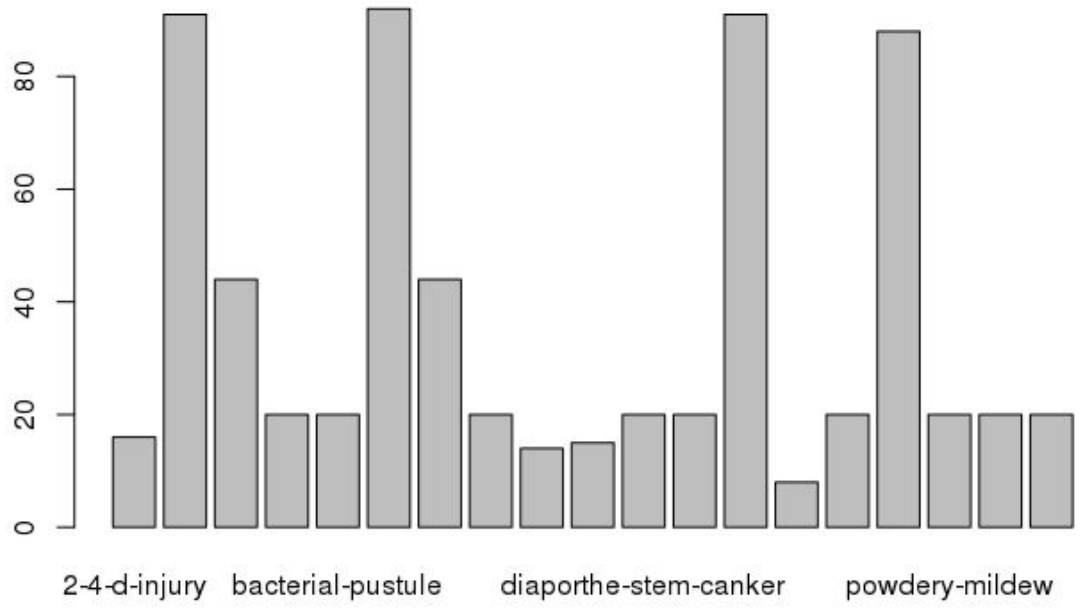
the classes?

(c) Develop a strategy for handling missing data, either by eliminating

predictors or imputation.

The first observation about this dataset was that it has missing values and the predictor as well as the predicted variables are factors.

Distribution of variables:



Degenerate distribution basically means that there is not enough variability in a predictor to be informative about the prediction.

The ordered descending table the diseases are as follows:

1 brown-spot 92

2 alternarialeaf-spot 91

3 frog-eye-leaf-spot 91

4 phytophthora-rot 88

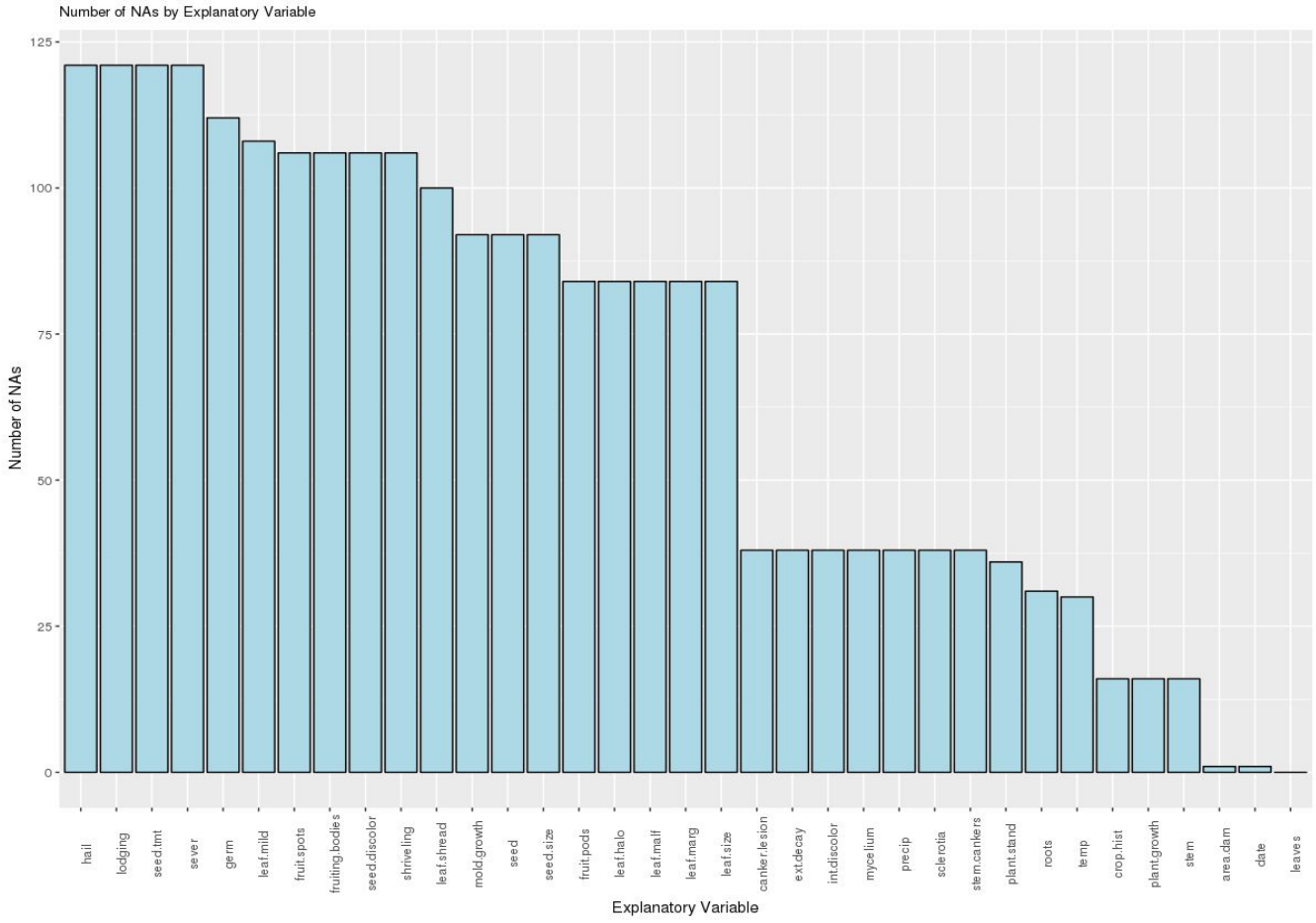
5 anthracnose 44

6 brown-stem-rot 44

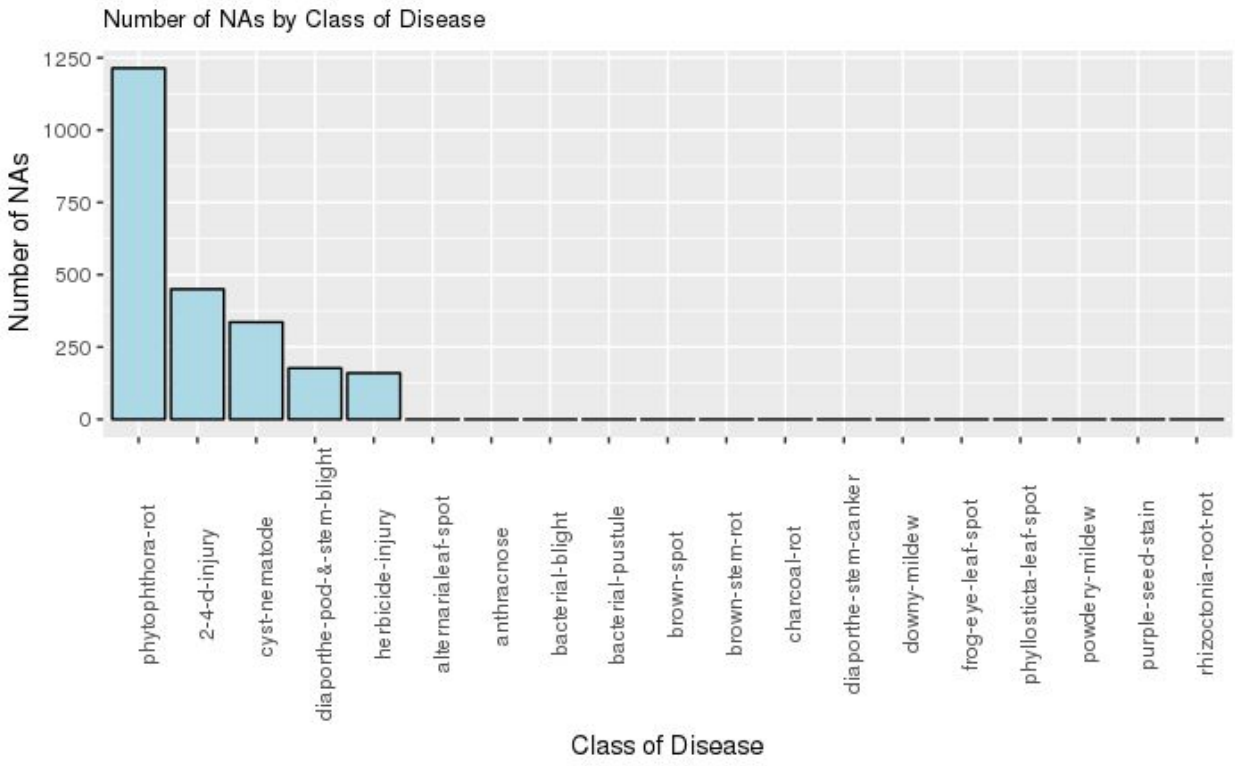
To handle the missing values

na.count	x.var
1	date
2	plant.stand
3	precip
4	temp
5	hail
6	crop.hist
7	area.dam
8	sever
9	seed.tmt
10	germ
11	plant.growth
12	leaves
13	leaf.halo
14	leaf.marg
15	leaf.size
16	leaf.shread
17	leaf.malf
18	leaf.mild
19	stem
20	lodging
21	stem.cankers
22	canker.lesion
23	fruiting.bodies

24	38	ext.decay
25	38	mycelium
26	38	int.discolor
27	38	sclerotia
28	84	fruit.pods
29	106	fruit.spots
30	92	seed
31	92	mold.growth
32	106	seed.discolor
33	92	seed.size
34	106	shriveling
35	31	roots



We find that the predictor variables have a maximum count of 120. This is about 17 percent of the data.



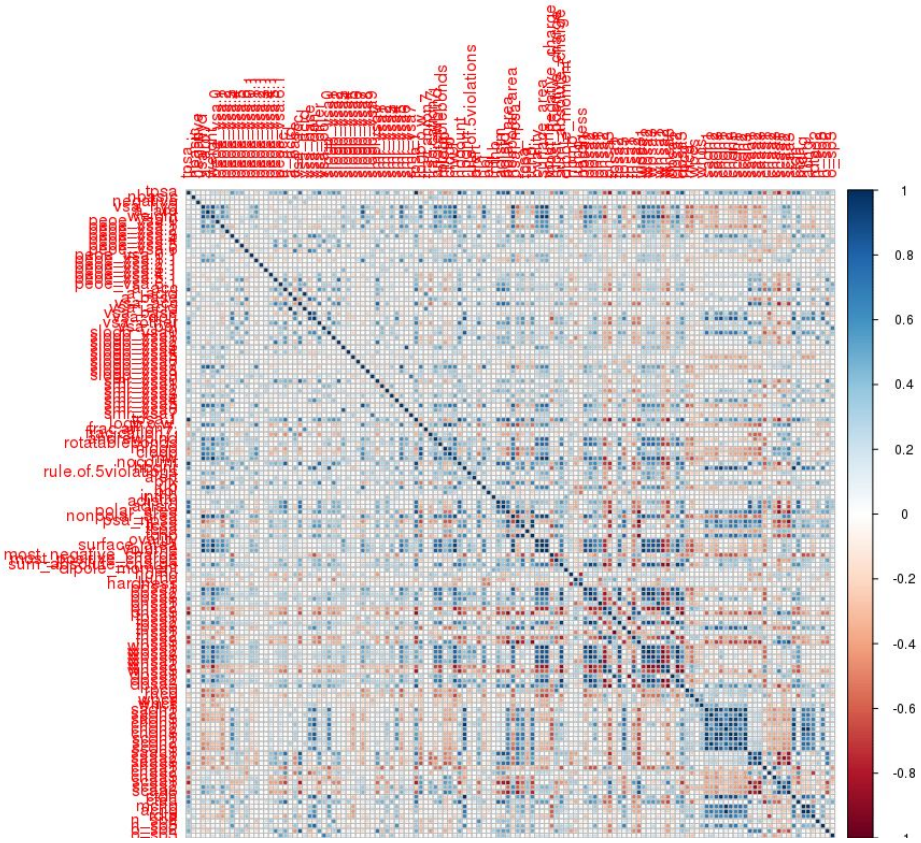
The classes that show maximum number of NA values are Phytophthora-rot 2-4-d-injury and cyst nematode.These classes are likely to have more missing values than the others.

(b) Do any of the individual predictors have degenerate distributions?

(c) Generally speaking, are there strong relationships between the predic-

tor data? If so, how could correlations in the predictor set be reduced?

Does this have a dramatic effect on the number of predictors available for



There are many positive and many negative correlations in the data, and since most of the variables are related to each other the data will be reduced to a large extent as we may drop many of the variables.

Appendix:

Question 1:

---

title: "Assignment 1"

author: "Supriya Bachal"

date: "September 30, 2017"

output: html_document

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = FALSE)
```

```{r}
library(mlbench)
data(Glass)
str(Glass)
```

```{r}
data(Glass)
summary(Glass$Type)
library(ggplot2)
ggplot(Glass,aes(x=Type))+
stat_count(width = 0.5,fill="Green",colour="black")+
ggtitle("Number of Observations by Type of Glass")
```

```{r}
par(mfrow=c(3,3))
library(fBasics)
for(i in 1:9) {
d<-density(Glass[,i])
plot(d,main=names(Glass)[i],xlab = "")
```

}

```

```{r}

par(mfrow=c(3,3))

for(i in 1:9){

  plot(Glass$Type,Glass[,i],col="Green",

       xlab="Type of Glass",ylab=names(Glass)[i],main=paste("Box Plot of Chemical Elements",names(Glass)[i],"by Type of Glass"))

}

```

```{r}

pairs(Glass[,1:9],lower.panel=panel.smooth)

m<-cor(Glass[1:9])

cor(Glass[1:9])

corrplot(m, order="hclust")

```

```{r}

library(e1071)

apply( Glass[,-10], 2, skewness )

par(mfrow=c(3,3))

hist( Glass$RI,main="Histogram for element RI",xlab="RI")

hist( Glass$Na,main="Histogram for element NA",xlab="NA")

hist( Glass$Mg,main="Histogram for eleent MG",xlab="MG")

hist( Glass$Al,main="Histogram for eleent AL",xlab="AL")

hist(Glass$Si,main="Histogram for eleent SI",xlab="SI")

```
hist(Glass$K,main="Histogram for eleent K",xlab="K")
hist(Glass$Ca,main="Histogram for eleent Ca",xlab="Ca")
hist(Glass$Ba,main="Histogram for eleent Ba",xlab="Ba")
hist(Glass$Fe,main="Histogram for eleent Fe",xlab="Fe")
```

```{r}
Sorted<-Glass[order(Glass$Type),]
library(e1071)
for(i in 1:9)
{
a[i]<-skewness(Glass[,i])

}
names(a)<-c("RI","Na","Mg","Al","Si","K","Ca","Ba","Fe")
```

```{r}
f=Glass[,-10]
trans <- preProcess(f,method = c("BoxCox"))
transformed <- predict(trans, f)

fulldata.scale = data.frame(predict(pred.scale, as.data.frame(Glass[, -10])),Type = Glass$Type)

par(mfrow=c(3,3))
for (i in names(fulldata.scale)[1:9]) {
```

```
ggplot(fulldata.scale, aes_string(x = i)) + geom_histogram(fill = "SeaGreen") + labs(y = "")
}
par(mfrow=c(3,3))
graphlist

skewness(fulldata.scale[,i])

yjTrans <- preProcess(Glass[, -10], method = "YeoJohnson")
yjData <- predict(yjTrans, newdata= Glass[, -10])
```

Q2

```
---
title: "Assignment 2"
author: "Supriya Bachal"
date: "October 1, 2017"
output: html_document
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
```

```{r}
```
```

```
d=table(Soybean$Class)
barplot(d)
which.max(d)
d[which.min(d)]
gg=data.frame(d[order(d,decreasing = T)])
head(gg)
```

```{r}
Missing values
isna.byclass<-Soybean%>%
group_by(Class)%>%
do(data.frame(sum(is.na(.))))
names(isna.byclass)<-c("Class","na.count")
isna.byclass
```

```
isna.byclass<-Soybean%>%
  group_by(Class)%>%
  do(data.frame(sum(is.na(.))))
names(isna.byclass)<-c("Class","na.count")
Isna.byclass
```

Question:3

```{r}
```

```
library(caret)

data(BloodBrain)

correlations<-cor(bbbDescr)

library(corrplot)

corrplot(correlations)
```

References:

https://rpubs.com/fraki22/122619

https://rpubs.com/chidungkt/221544

https://rpubs.com/chidungkt/221544

Chapter 3 from notes.