Data624 - Homework4

Amit Kapoor

2/28/2021

Contents

```r
library(mlbench)
library(DataExplorer)
library(GGally)
library(psych)
```

3.1

The UC Irvine Machine Learning Repository[6] 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.

The data can be accessed via:

```r
data(Glass)
str(Glass)
```

```
## 'data.frame':    214 obs. of  10 variables:
##  $ 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 ...
##  $ Al  : 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: Factor w/ 6 levels "1","2","3","5",..: 1 1 1 1 1 1 1 1 1 1 ...
```
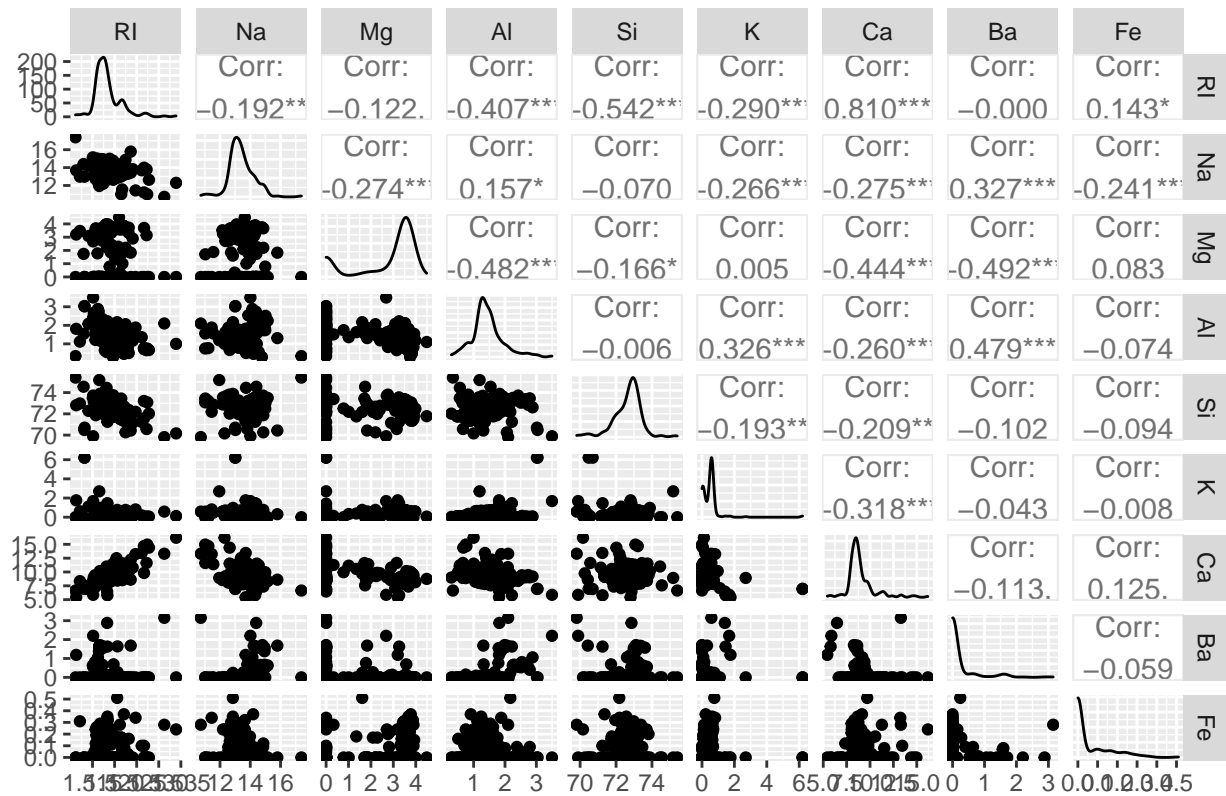
(a) Using visualizations, explore the predictor variables to understand their distributions as well as the relationships between predictors.

```r
# predictors distribution
plot_histogram(Glass,
               geom_histogram_args = list(bins = 30L),
               nrow = 3L,
               ncol = 3L)
```
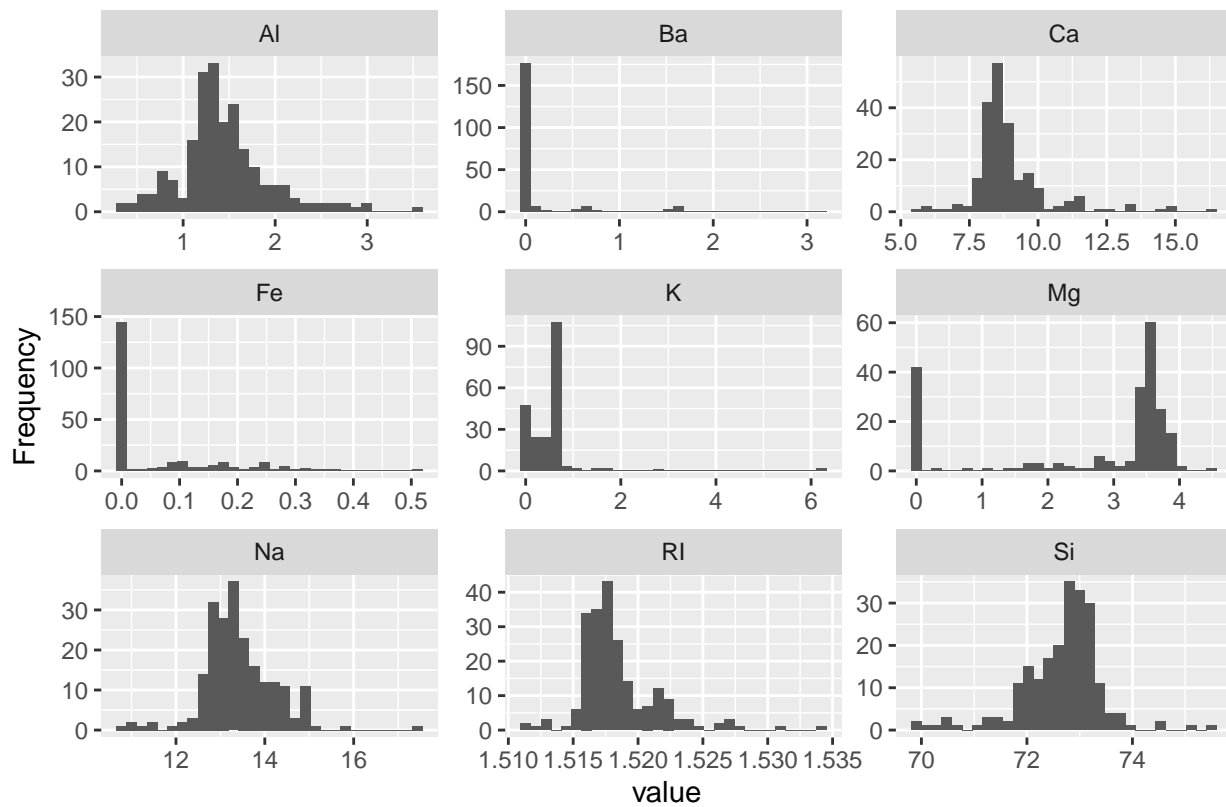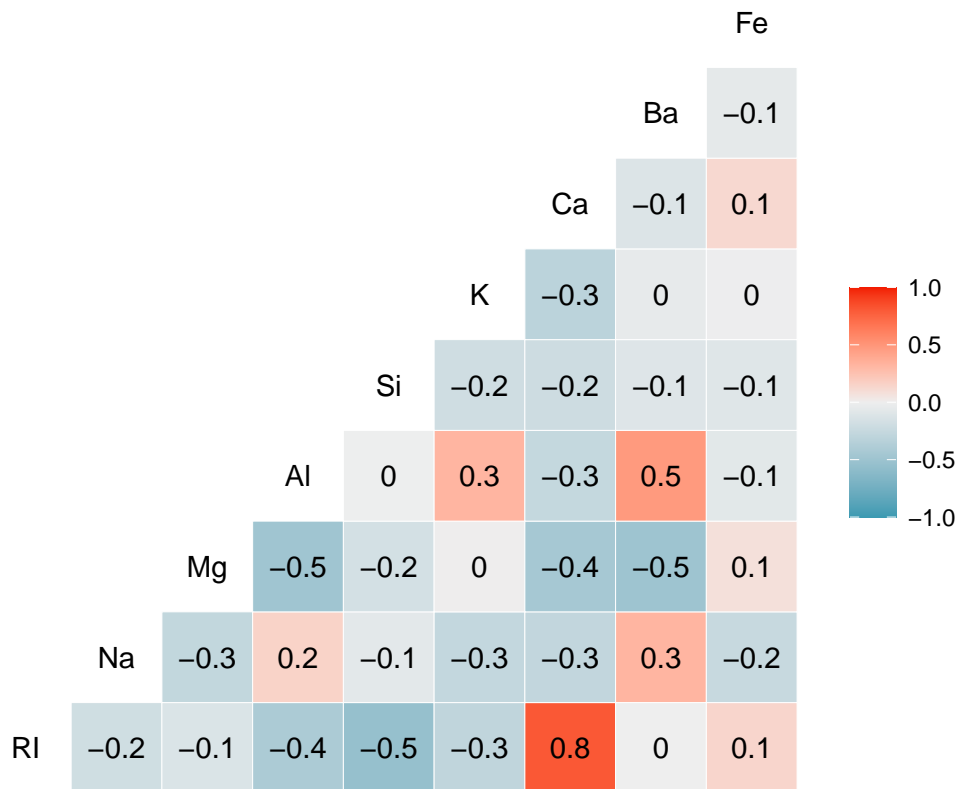
```
# scatterplot matrix
Glass %>%
  dplyr::select(-Type) %>%
  ggpairs(title = "Paiwise scatter plots") %>%
  print(progress = F)
```

Paiwise scatter plots

	RI	Na	Mg	Al	Si	K	Ca	Ba	Fe	
		Corr:	Corr:	Corr:	Corr:	Corr:	Corr:	Corr:	Corr:	RI
		−0.192**	−0.122.	-0.407**	-0.542**	−0.290**	0.810***	−0.000	0.143*	
			Corr:	Corr:	Corr:	Corr:	Corr:	Corr:	Corr:	Na
			-0.274**	0.157*	−0.070	−0.266**	−0.275**	0.327***	-0.241**	
				Corr:	Corr:	Corr:	Corr:	Corr:	Corr:	Mg
				-0.482**	−0.166*	0.005	−0.444**	−0.492**	0.083	
					Corr:	Corr:	Corr:	Corr:	Corr:	Al
					−0.006	0.326***	−0.260**	0.479***	−0.074	
						Corr:	Corr:	Corr:	Corr:	Si
						−0.193**	−0.209**	−0.102	−0.094	
							Corr:	Corr:	Corr:	K
							−0.318**	−0.043	−0.008	
								Corr:	Corr:	Ca
								−0.113.	0.125.	
									Corr:	Ba
									−0.059	
										Fe

```
# correlation
Glass %>%
  dplyr::select(-Type) %>%
  ggcorr(label = TRUE)
```

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

```
describe(Glass)
```
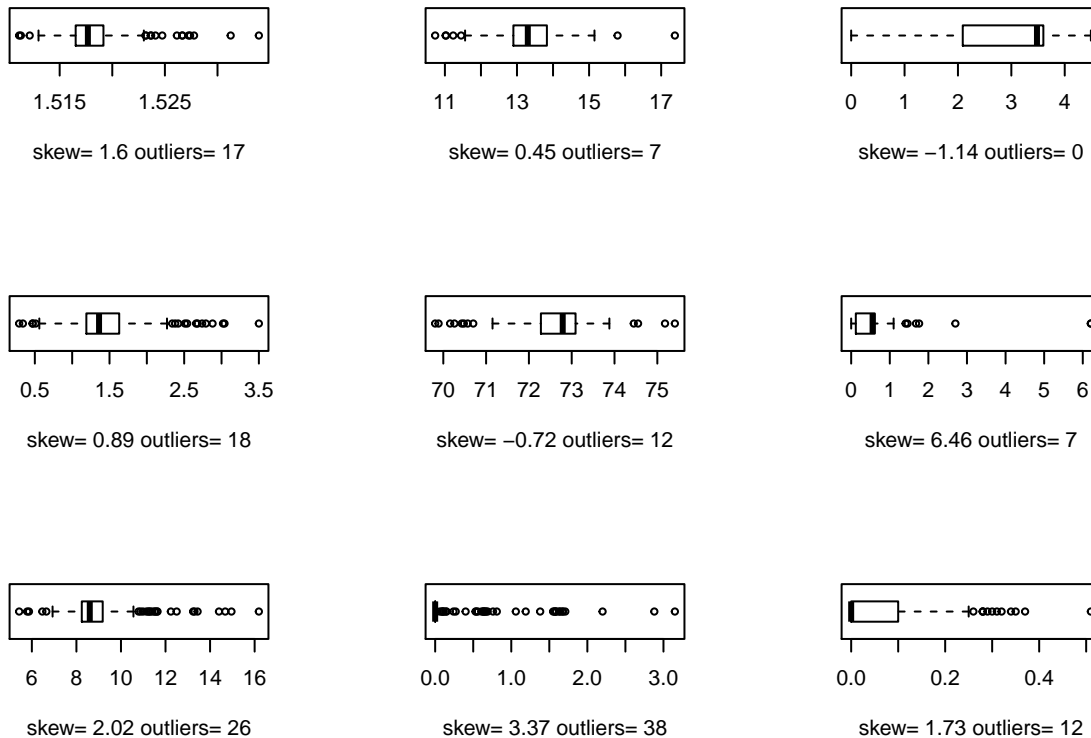
```
##           vars   n  mean    sd median trimmed  mad   min   max range  skew kurtosis
## RI          1 214  1.52 0.00   1.52    1.52 0.00  1.51  1.53  0.02  1.60     4.72
## Na          2 214 13.41 0.82  13.30   13.38 0.64 10.73 17.38  6.65  0.45     2.90
## Mg          3 214  2.68 1.44   3.48    2.87 0.30  0.00  4.49  4.49 -1.14    -0.45
## Al          4 214  1.44 0.50   1.36    1.41 0.31  0.29  3.50  3.21  0.89     1.94
## Si          5 214 72.65 0.77  72.79   72.71 0.57 69.81 75.41  5.60 -0.72     2.82
## K           6 214  0.50 0.65   0.56    0.43 0.17  0.00  6.21  6.21  6.46    52.87
## Ca          7 214  8.96 1.42   8.60    8.74 0.66  5.43 16.19 10.76  2.02     6.41
## Ba          8 214  0.18 0.50   0.00    0.03 0.00  0.00  3.15  3.15  3.37    12.08
## Fe          9 214  0.06 0.10   0.00    0.04 0.00  0.00  0.51  0.51  1.73     2.52
## Type*      10 214  2.54 1.71   2.00    2.31 1.48  1.00  6.00  5.00  1.04    -0.29
##           se
## RI      0.00
## Na      0.06
## Mg      0.10
## Al      0.03
## Si      0.05
## K       0.04
## Ca      0.10
## Ba      0.03
## Fe      0.01
## Type* 0.12
```

```
label <- function(var) {
  return( paste("skew=" , round(describe(var)$skew,2) , "outliers=" , length(boxplot(var, plot=FALSE)$ou
}
```

```
par(mfrow=c(3,3))
for (i in 1:9){
  boxplot(
    Glass[i], color='green', horizontal = T,
    xlab = label(Glass[i])
    )
}
```


skew= 1.6 outliers= 17   skew= 0.45 outliers= 7   skew= −1.14 outliers= 0

skew= 0.89 outliers= 18   skew= −0.72 outliers= 12   skew= 6.46 outliers= 7

skew= 2.02 outliers= 26   skew= 3.37 outliers= 38   skew= 1.73 outliers= 12

(c) Are there any relevant transformations of one or more predictors that might improve the classification model?

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**.

The data can be loaded via:

```
data(Soybean)
str(Soybean)
```

```
## 'data.frame':    683 obs. of  36 variables:
##  $ Class      : Factor w/ 19 levels "2-4-d-injury",..: 11 11 11 11 11 11 11 11 11 11 ...
##  $ date       : Factor w/ 7 levels "0","1","2","3",..: 7 5 4 4 7 6 6 5 7 5 ...
##  $ plant.stand: Ord.factor w/ 2 levels "0"<"1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ precip     : Ord.factor w/ 3 levels "0"<"1"<"2": 3 3 3 3 3 3 3 3 3 3 ...
##  $ temp       : Ord.factor w/ 3 levels "0"<"1"<"2": 2 2 2 2 2 2 2 2 2 2 ...
##  $ hail       : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 2 1 1 ...
```

```
##  $ crop.hist      : Factor w/ 4 levels "0","1","2","3": 2 3 2 2 3 4 3 2 4 3 ...
##  $ area.dam       : Factor w/ 4 levels "0","1","2","3": 2 1 1 1 1 1 1 1 1 1 ...
##  $ sever          : Factor w/ 3 levels "0","1","2": 2 3 3 3 2 2 2 2 2 3 ...
##  $ seed.tmt       : Factor w/ 3 levels "0","1","2": 1 2 2 1 1 1 2 1 2 1 ...
##  $ germ           : Ord.factor w/ 3 levels "0"<"1"<"2": 1 2 3 2 3 2 1 3 2 3 ...
##  $ plant.growth   : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 2 ...
##  $ leaves         : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 2 ...
##  $ leaf.halo      : Factor w/ 3 levels "0","1","2": 1 1 1 1 1 1 1 1 1 1 ...
##  $ leaf.marg      : Factor w/ 3 levels "0","1","2": 3 3 3 3 3 3 3 3 3 3 ...
##  $ leaf.size      : Ord.factor w/ 3 levels "0"<"1"<"2": 3 3 3 3 3 3 3 3 3 3 ...
##  $ leaf.shread    : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ leaf.malf      : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ leaf.mild      : Factor w/ 3 levels "0","1","2": 1 1 1 1 1 1 1 1 1 1 ...
##  $ stem           : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 2 ...
##  $ lodging        : Factor w/ 2 levels "0","1": 2 1 1 1 1 1 2 1 1 1 ...
##  $ stem.cankers   : Factor w/ 4 levels "0","1","2","3": 4 4 4 4 4 4 4 4 4 4 ...
##  $ canker.lesion  : Factor w/ 4 levels "0","1","2","3": 2 2 1 1 2 1 2 2 2 2 ...
##  $ fruiting.bodies: Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 2 ...
##  $ ext.decay      : Factor w/ 3 levels "0","1","2": 2 2 2 2 2 2 2 2 2 2 ...
##  $ mycelium       : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ int.discolor   : Factor w/ 3 levels "0","1","2": 1 1 1 1 1 1 1 1 1 1 ...
##  $ sclerotia      : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ fruit.pods     : Factor w/ 4 levels "0","1","2","3": 1 1 1 1 1 1 1 1 1 1 ...
##  $ fruit.spots    : Factor w/ 4 levels "0","1","2","4": 4 4 4 4 4 4 4 4 4 4 ...
##  $ seed           : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ mold.growth    : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ seed.discolor  : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ seed.size      : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ shriveling     : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ roots          : Factor w/ 3 levels "0","1","2": 1 1 1 1 1 1 1 1 1 1 ...
```

(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.