# Glass Data (UCI/ML)

## Basic Statistics in Exploration

```
glass <- data.table::fread("glass/glass.data")
colnames(glass) <- c("id","Ri","Na","Mg","Al","Si","K","Ca","Ba","Fe","type")

glass <- glass[,-1]
glass$type <- as.factor(glass$type)

summary(glass)
```

```
##        Ri              Na              Mg              Al
##  Min.   :1.511   Min.   :10.73   Min.   :0.000   Min.   :0.290
##  1st Qu.:1.517   1st Qu.:12.91   1st Qu.:2.115   1st Qu.:1.190
##  Median :1.518   Median :13.30   Median :3.480   Median :1.360
##  Mean   :1.518   Mean   :13.41   Mean   :2.685   Mean   :1.445
##  3rd Qu.:1.519   3rd Qu.:13.82   3rd Qu.:3.600   3rd Qu.:1.630
##  Max.   :1.534   Max.   :17.38   Max.   :4.490   Max.   :3.500
##        Si              K               Ca               Ba
##  Min.   :69.81   Min.   :0.0000   Min.   : 5.430   Min.   :0.000
##  1st Qu.:72.28   1st Qu.:0.1225   1st Qu.: 8.240   1st Qu.:0.000
##  Median :72.79   Median :0.5550   Median : 8.600   Median :0.000
##  Mean   :72.65   Mean   :0.4971   Mean   : 8.957   Mean   :0.175
##  3rd Qu.:73.09   3rd Qu.:0.6100   3rd Qu.: 9.172   3rd Qu.:0.000
##  Max.   :75.41   Max.   :6.2100   Max.   :16.190   Max.   :3.150
##        Fe          type
##  Min.   :0.00000   1:70
##  1st Qu.:0.00000   2:76
##  Median :0.00000   3:17
##  Mean   :0.05701   5:13
##  3rd Qu.:0.10000   6: 9
##  Max.   :0.51000   7:29
```

```
library(moments)
```

```
## Warning: package 'moments' was built under R version 4.0.3
```

```
# Inter-Quartile Range
apply(glass[,-"type"], 2, IQR)
```

```
##       Ri       Na       Mg       Al       Si        K       Ca       Ba
## 0.002635 0.917500 1.485000 0.440000 0.807500 0.487500 0.932500 0.000000
##       Fe
## 0.100000
```

```r
# Standard deviation
apply(glass[,-'type'], 2, sd)
```

```
##          Ri          Na          Mg          Al          Si           K
## 0.003036864 0.816603556 1.442407845 0.499269646 0.774545795 0.652191846
##          Ca          Ba          Fe
## 1.423153487 0.497219261 0.097438701
```

```r
# skewness
apply(glass[,-"type"], 2, skewness)
```

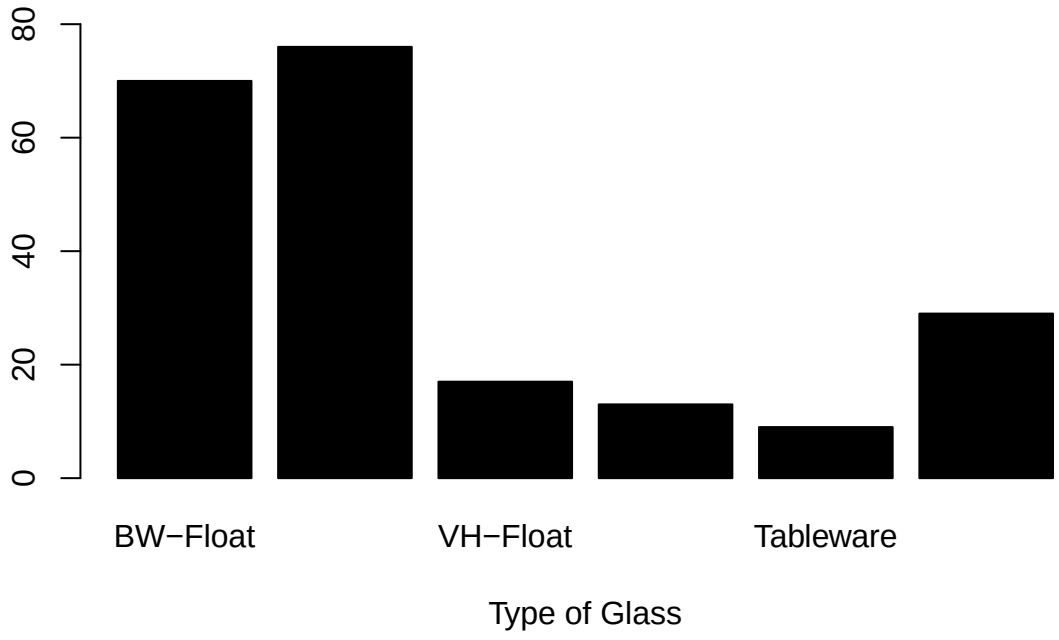
```
##         Ri         Na         Mg         Al         Si          K         Ca
##  1.6140150  0.4509917 -1.1444648  0.9009179 -0.7253173  6.5056358  2.0326774
##         Ba         Fe
##  3.3924309  1.7420068
```

```r
# kurtosis
apply(glass[,-'type'],2, kurtosis)
```

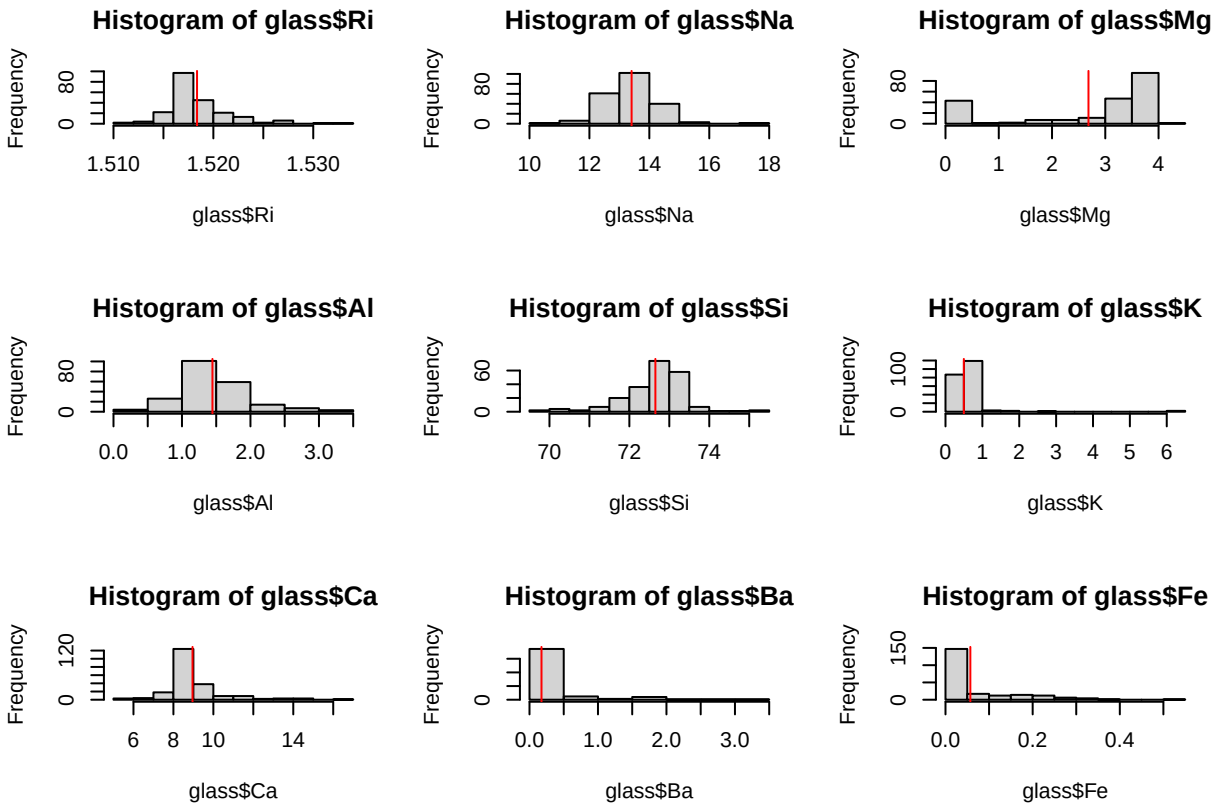
```
##        Ri        Na        Mg        Al        Si         K        Ca        Ba
##  7.789354  5.953477  2.571298  4.984832  5.871105 56.392327  9.498968 15.222071
##        Fe
##  5.572318
```

## Univariate Graphical Analysis

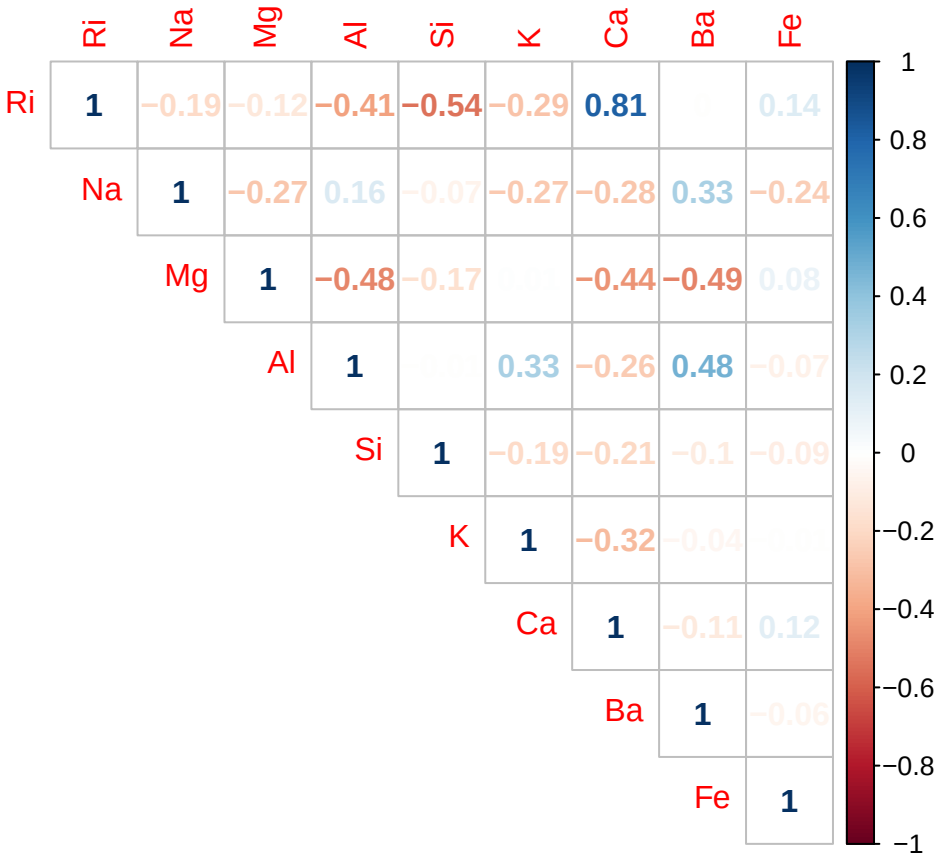
```r
### Dependent (Type Variable)
names <- c("BW-Float", "BW-Non_Float", "VH-Float",  "Container", "Tableware", "Headlamps")
plot(glass$type, col = "black",
     ylim= c(0,90),
     # yaxt = "n",
     xlab = "Type of Glass",
     names = names)
```




```
### Explanatory Variables
par(mfrow = c(3,3))
hist(glass$Ri);abline(v = mean(glass$Ri), col = "red")
hist(glass$Na);abline(v = mean(glass$Na), col = "red")
hist(glass$Mg);abline(v = mean(glass$Mg), col = "red")
hist(glass$Al);abline(v = mean(glass$Al), col = "red")
hist(glass$Si);abline(v = mean(glass$Si), col = "red")
hist(glass$K);abline(v = mean(glass$K), col = "red")
hist(glass$Ca);abline(v = mean(glass$Ca), col = "red")
hist(glass$Ba);abline(v = mean(glass$Ba), col = "red")
hist(glass$Fe);abline(v = mean(glass$Fe), col = "red")
```

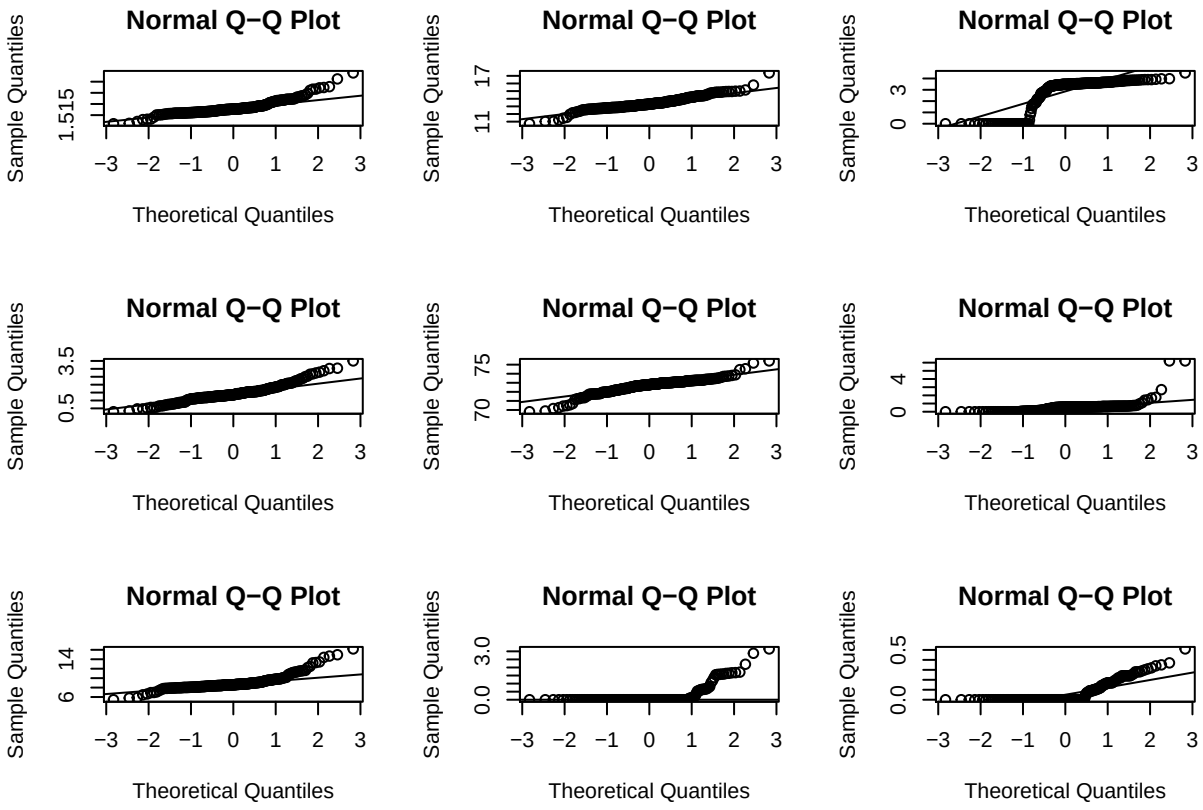



```
corrplot::corrplot(cor(glass[,-"type"]),type = "upper", method = "number")
```



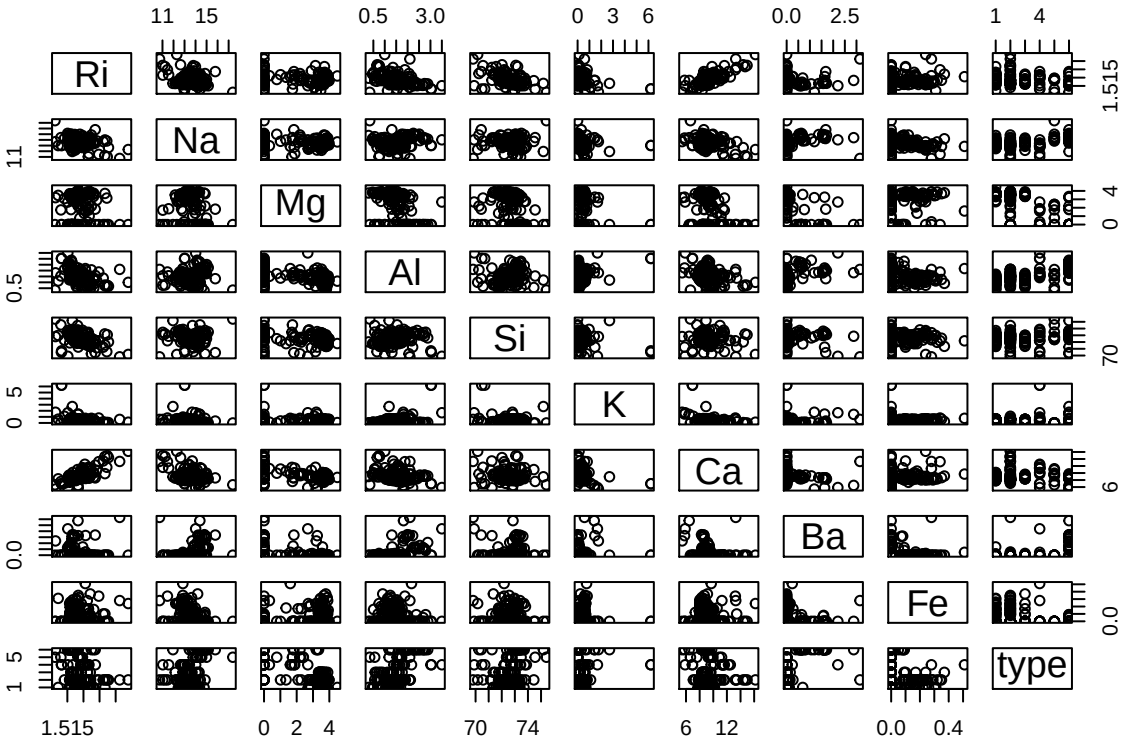

### Diagonostic

```
par(mfrow = c(3,3))
qqnorm(glass$Ri);qqline(glass$Ri)
qqnorm(glass$Na);qqline(glass$Na)
qqnorm(glass$Mg);qqline(glass$Mg)
qqnorm(glass$Al);qqline(glass$Al)
qqnorm(glass$Si);qqline(glass$Si)
qqnorm(glass$K);qqline(glass$K)
qqnorm(glass$Ca);qqline(glass$Ca)
qqnorm(glass$Ba);qqline(glass$Ba)
qqnorm(glass$Fe);qqline(glass$Fe)
```



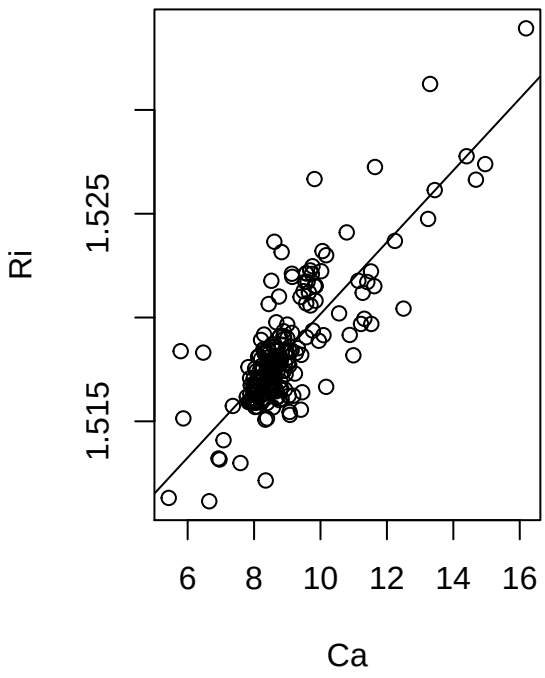

## Multivariate

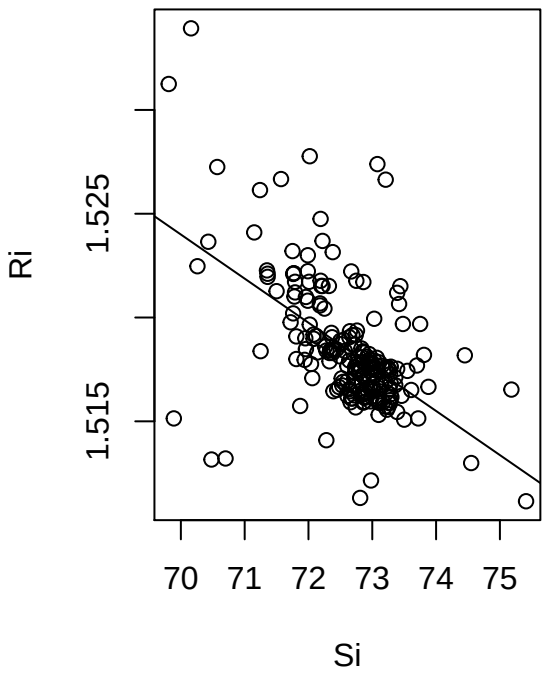
```
pairs(glass)
```




**Ri vs most correlated variables (Can be use for Modelling)**

```
par(mfrow = c(1,2))
plot(Ri~Ca, glass); abline(lm(Ri~Ca, glass))
plot(Ri~Si, glass); abline(lm(Ri~Si, glass))
```




```
## Normality Test
apply(glass[,-"type"], 2, shapiro.test)
```

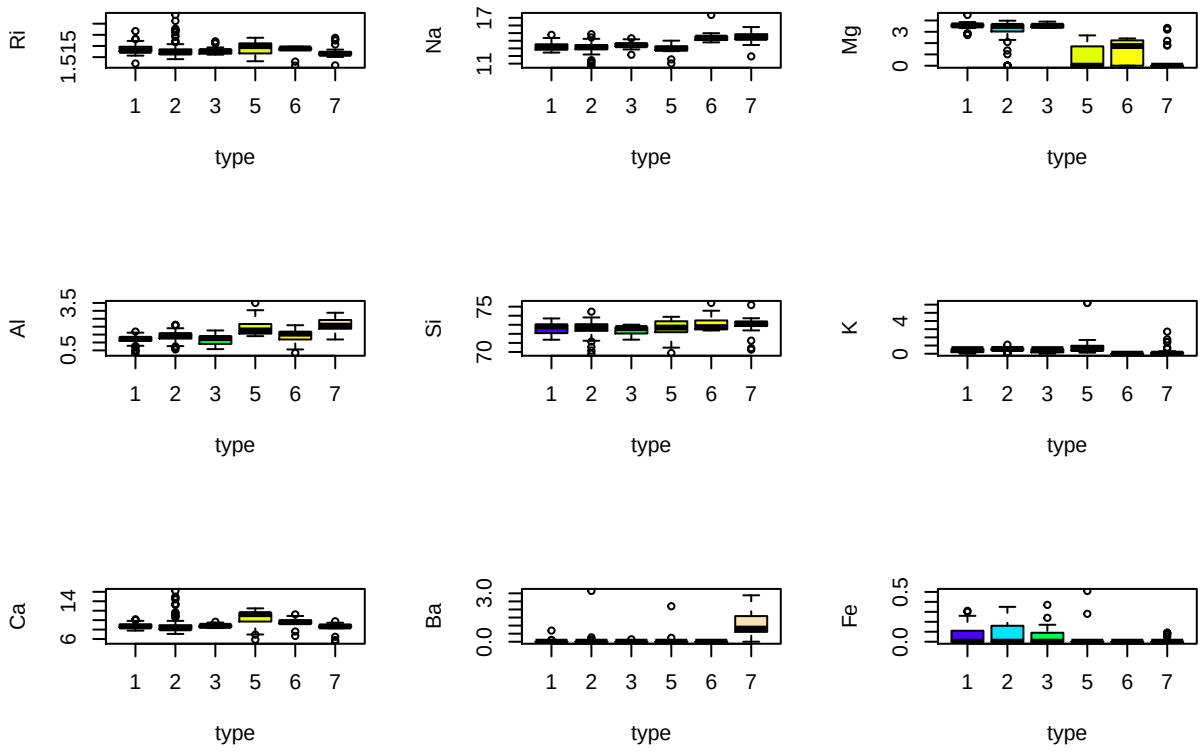
```
## $Ri
##
##  Shapiro-Wilk normality test
##
## data:  newX[, i]
## W = 0.86757, p-value = 1.077e-12
##
##
## $Na
##
##  Shapiro-Wilk normality test
##
## data:  newX[, i]
## W = 0.94576, p-value = 3.466e-07
##
##
## $Mg
##
##  Shapiro-Wilk normality test
##
## data:  newX[, i]
## W = 0.69934, p-value < 2.2e-16
##
```

```
##
## $Al
##
##  Shapiro-Wilk normality test
##
## data:  newX[, i]
## W = 0.94341, p-value = 2.083e-07
##
##
## $Si
##
##  Shapiro-Wilk normality test
##
## data:  newX[, i]
## W = 0.91966, p-value = 2.175e-09
##
##
## $K
##
##  Shapiro-Wilk normality test
##
## data:  newX[, i]
## W = 0.44162, p-value < 2.2e-16
##
##
## $Ca
##
##  Shapiro-Wilk normality test
##
## data:  newX[, i]
## W = 0.79387, p-value = 4.287e-16
##
##
## $Ba
##
##  Shapiro-Wilk normality test
##
## data:  newX[, i]
## W = 0.40857, p-value < 2.2e-16
##
##
## $Fe
##
##  Shapiro-Wilk normality test
##
## data:  newX[, i]
## W = 0.6532, p-value < 2.2e-16
```

## Type vs Independents

```
### Dependant to independent
par(mfrow = c(3,3))
```

```
boxplot(Ri~type, data = glass,
        col = topo.colors(6))

boxplot(Na~type, data = glass,
        col = topo.colors(6))

boxplot(Mg~type, data = glass,
        col = topo.colors(6))

boxplot(Al~type, data = glass,
        col = topo.colors(6))

boxplot(Si~type, data = glass,
        col = topo.colors(6))

boxplot(K~type, data = glass,
        col = topo.colors(6))

boxplot(Ca~type, data = glass,
        col = topo.colors(6))

boxplot(Ba~type, data = glass,
        col = topo.colors(6))

boxplot(Fe~type, data = glass,
        col = topo.colors(6))
```




**Description and Findings**

Purpose of this analysis is to analyze descriptively for relations, associations among variables and individual characteristics variables. Glass Identification data set has taken fromhttps://archive.ics.uci.edu/ml/datasets/glass+identification, originally came from Vina Spiehler, Ph.D., DABFT Diagnostic Products Corporations. In data set, we have total 9 variables and 214 observations to study. Moving to analysis part, here we have two objective variable which are Refractive index and Type of glass to study. Refractive index is a continuous attribute on the other hand type of glass has seven classes (class 4 is missing in data). In the univariate graphical section, we have individual graphs of each attribute. For most of attributes cluster of observations is around its mean in accompany with its mode e.g. mean of Na (Sodium) is 13.41 and the median is 13.30. Interesting characteristics is came into light for attributes K and Ba, majority of observations are extreme e.g. for K(Potassium) between 0 to 1 and for Fe (Iron) its zero. Interestingly Mg (Magnesium) has two peaks (bi-modal) at both extremes. All these attributes are sharing asymmetric characteristics so they are likely to be non-normal. For type variable, there are 6 classes present in the table where Building window floating and non floating class has dominance, on the other hand Tableware and Container are at bottom in numbers for data set (13 and 9respectively).

For association between variables, I have drawn correlation matrix which depicts variables , which are highly correlated with Iron, If iron increases by 1 unit so Ca will increase by .8 and Si (silicon) will fall for .5.Most highly correlated variables can be used for modelling Correlation does not implies causality. It only tells the direction of change in variables. Normality is seen normal for most variables but their tails are thick (except Mg, Ba, K and Fe).Since the p value is less than .05, Normality for all the attributes is rejected by Shapiro Wilkson test. For part b) I have drawn box-plot, also called box-whisker, which uses to show categorical variation of a variable against another continuous variable. Median for Ri and Fe are quite same across all the type of glasses. Magnesium has higher value for building windows of either type and vehicle window float processed and higher spread for tableware and container.