Intoduction to Machine Learning - Exercise 1

Mikko Ahro

Problem 1

Task a

Read p1.csv into dataframe and drop columns “id”, “SMILES”, “InChIKey”

```
p1data <- read.csv("data/p1.csv", header=TRUE, sep=",")
p1data <- subset(p1data, select=-c(id, SMILES, InChIKey))
```

Task b

```
p1_subset <- subset(p1data, select=c(pSat_Pa, NumOfConf, ChemPot_kJmol))
summary(p1_subset)
```

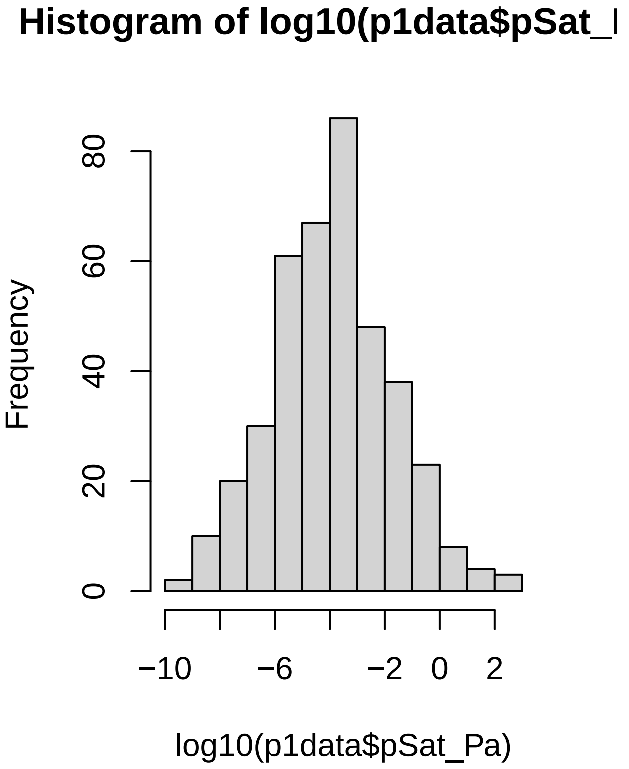
```
##     pSat_Pa            NumOfConf        ChemPot_kJmol
##  Min.   :  0.0000   Min.   :   2.00   Min.   :-3.160
##  1st Qu.:  0.0000   1st Qu.:  73.25   1st Qu.: 9.723
##  Median :  0.0001   Median : 172.50   Median :12.781
##  Mean   :  2.9620   Mean   : 223.50   Mean   :12.434
##  3rd Qu.:  0.0023   3rd Qu.: 324.25   3rd Qu.:15.659
##  Max.   :562.8970   Max.   :1058.00   Max.   :28.096
```

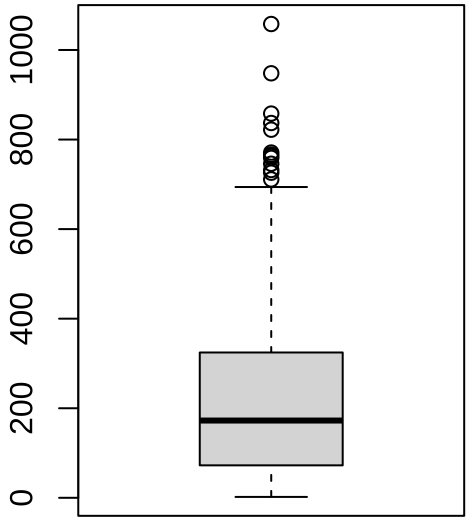
Task c

```
ChemPot_kjmol_arr <- p1data$ChemPot_kJmol
```

Task d

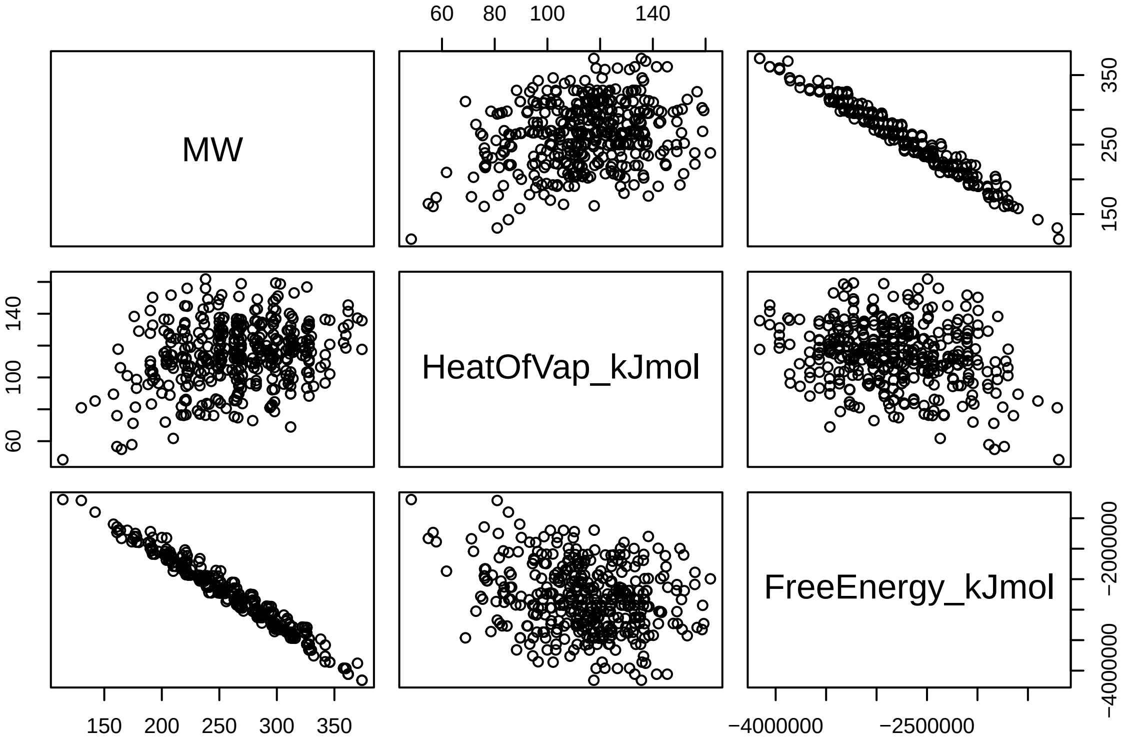
```
par(mfrow=c(1,2))
hist(log10(p1data$pSat_Pa))
boxplot(p1data$NumOfConf)
```





Task e

```
scatter_subset <- subset(p1data, select=c(MW, HeatOfVap_kJmol, FreeEnergy_kJmol))
pairs(scatter_subset)
```



```
{r eval=FALSE} # library(rmarkdown) # render("MLExercise1.Rmd")
#
```