Exercise 5

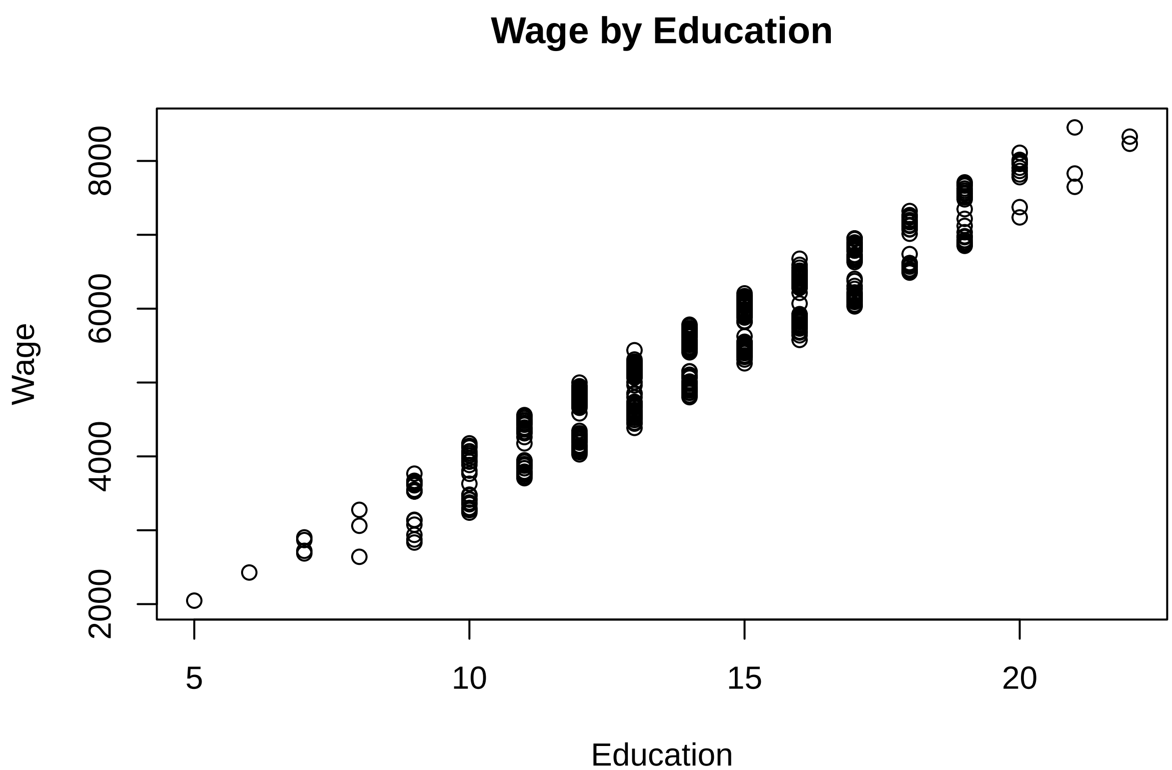
Tobias Famos

Exercise 5

Task 1

```
educationBis <- read.table("/home/tobias/unibe/statistical methods in R/Exercise 5/EducationBis.txt", he
educationBis <- subset(educationBis, select=-c(ID))
plot(educationBis$Gender, educationBis$Wage, xlab="Gender", ylab = "Wage", main="Wage by Gender")
```


```
plot(educationBis$Education, educationBis$Wage, xlab="Education", ylab = "Wage", main="Wage by Educatio
```




```r
summary(educationBis)
```

```
##    Education        Gender         Wage     
##  Min.   : 5.00   female:198   Min.   :2047  
##  1st Qu.:12.00   male  :299   1st Qu.:4679  
##  Median :14.00                Median :5520  
##  Mean   :14.26                Mean   :5463  
##  3rd Qu.:16.00                3rd Qu.:6319  
##  Max.   :22.00                Max.   :8454
```

```r
linear_model <- lm(educationBis$Wage ~ . , data=educationBis)
summary(linear_model)
```

```
## 
## Call:
## lm(formula = educationBis$Wage ~ ., data = educationBis)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -243.473  -76.073    0.354   73.126  280.275 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -569.784     23.100  -24.67   <2e-16 ***
## Education    397.975      1.543  258.00   <2e-16 ***
## Gendermale   597.904      9.173   65.18   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 100.1 on 494 degrees of freedom
## Multiple R-squared:  0.9931, Adjusted R-squared:  0.9931
```

```
## F-statistic: 3.544e+04 on 2 and 494 DF,  p-value: < 2.2e-16
```

From the output we can conclude that the model accurately describes the Wage using the variables Education and Gender. As the R^2 is 0.99 (almost 1) we can conclude that with our dataset we explain the bulk part of the variance. Additionally, we can conclude that for our dataset both gender and education have a very high significance due to the low P value.

Task 2

Read Data, remove model and ERP (model is not usable as it is unique, thus can be seen as an ID)

```
computers <- read.table("/home/tobias/unibe/statistical methods in R/Exercise 5/Computers.txt", header =
computers <- subset(computers, select = -c(model, ERP))
computers_only_numeric <- subset(computers, select = -c(vendor))
```

Take a look at the correlation of the numeric values to perform feature selection

```
cor(computers_only_numeric$PRP, computers_only_numeric)
```

```
##            MYCT      MMIN      MMAX      CACH     CGMIN     CHMAX PRP
## [1,] -0.3070994 0.7949313 0.8630041 0.6626414 0.6089033 0.6052093   1
```

Remove the features that have not sufficient correlation

```
computers <- subset(computers, select = -c( CGMIN, MYCT))
```

```
model2 <- lm(computers$PRP~., data=computers)
summary(model2)
```

```
##
## Call:
## lm(formula = computers$PRP ~ ., data = computers)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -204.838  -17.522    1.253   19.587  313.928
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -2.766e+02  7.209e+01  -3.838 0.000173 ***
## vendoramdahl          1.967e+02  8.049e+01   2.444 0.015516 *
## vendorapollo          2.795e+02  8.236e+01   3.394 0.000853 ***
## vendorbasf            2.343e+02  8.054e+01   2.909 0.004092 **
## vendorbti             2.244e+02  8.215e+01   2.732 0.006946 **
## vendorburroughs       1.700e+02  7.199e+01   2.362 0.019293 *
## vendorc.r.d           2.825e+02  7.580e+01   3.727 0.000262 ***
## vendorcambex          2.296e+02  7.807e+01   2.941 0.003715 **
## vendorcdc             2.609e+02  7.222e+01   3.612 0.000396 ***
## vendordec             2.666e+02  7.609e+01   3.503 0.000583 ***
## vendordg              2.765e+02  7.617e+01   3.630 0.000373 ***
## vendorformation       2.513e+02  7.755e+01   3.240 0.001429 **
## vendorfour-phase      2.539e+02  8.990e+01   2.824 0.005291 **
## vendorgould           2.351e+02  7.354e+01   3.197 0.001646 **
## vendorharris          2.311e+02  7.414e+01   3.117 0.002138 **
## vendorhoneywell       1.843e+02  7.278e+01   2.532 0.012219 *
## vendorhp              2.209e+02  7.411e+01   2.981 0.003283 **
## vendoribm             2.490e+02  7.421e+01   3.356 0.000970 ***
```

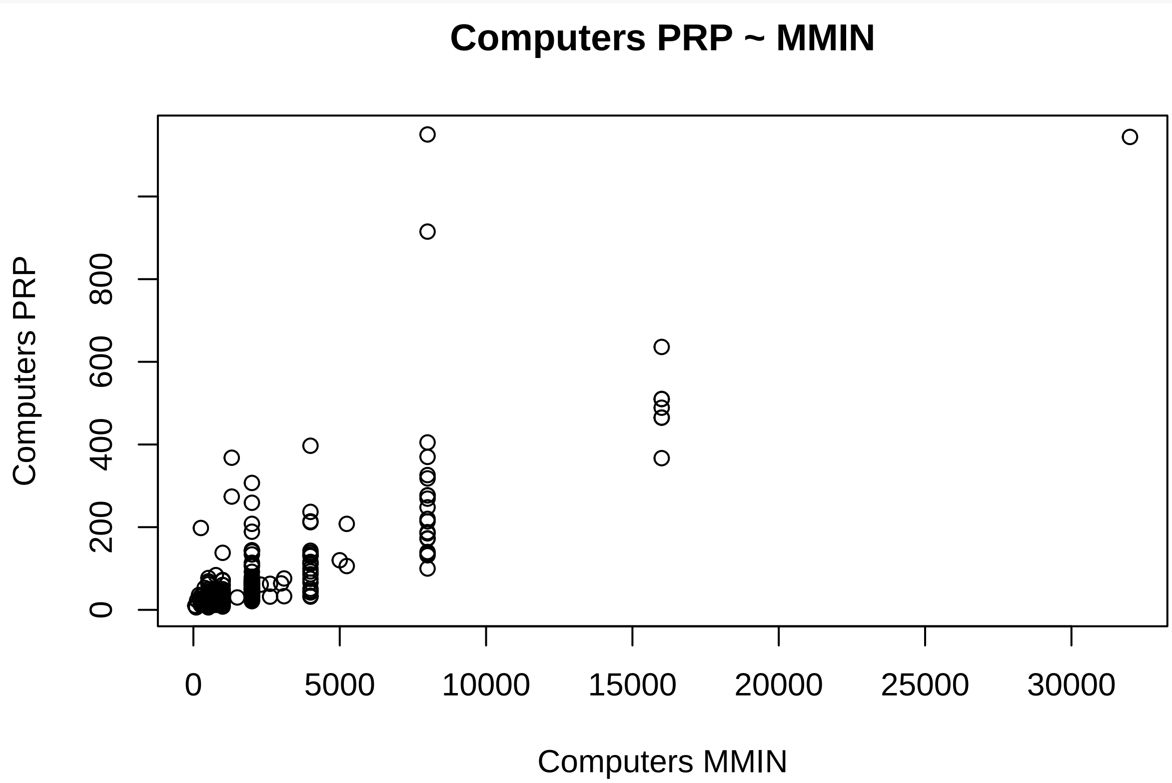
```
## vendoripl            2.222e+02  7.678e+01   2.894 0.004293 **
## vendormagnuson       2.095e+02  7.616e+01   2.751 0.006574 **
## vendormicrodata     -8.515e+00  8.884e+01  -0.096 0.923751
## vendornas            2.242e+02  7.353e+01   3.049 0.002653 **
## vendorncr            1.873e+02  7.334e+01   2.554 0.011491 *
## vendornixdorf        2.455e+02  7.838e+01   3.132 0.002037 **
## vendorperkin-elmer   2.604e+02  7.992e+01   3.259 0.001345 **
## vendorprime          2.611e+02  7.571e+01   3.449 0.000706 ***
## vendorsiemens        2.166e+02  7.421e+01   2.919 0.003974 **
## vendorsperry         2.562e+02  7.149e+01   3.583 0.000440 ***
## vendorsratus         2.307e+02  9.171e+01   2.515 0.012798 *
## vendorwang           2.873e+02  8.234e+01   3.489 0.000613 ***
## MMIN                 1.896e-02  2.066e-03   9.180  < 2e-16 ***
## MMAX                 3.525e-03  7.293e-04   4.833 2.93e-06 ***
## CACH                 6.144e-01  1.624e-01   3.783 0.000212 ***
## CHMAX                2.276e+00  2.837e-01   8.023 1.43e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 55.45 on 175 degrees of freedom
## Multiple R-squared:    0.9,  Adjusted R-squared:  0.8811
## F-statistic: 47.72 on 33 and 175 DF,  p-value: < 2.2e-16
```

By using the remaining features we can build a good model that explains 90% of the variance.

Task 3

The most important feature is MMIN

```
plot(computers$MMIN, computers$PRP , xlab = "Computers MMIN", ylab="Computers PRP", main="Computers PRP
```



There

are a few outliers. Additionally, it looks like the data is more on the lower end of MMIN and quite sparse on the higher end.