Statistical Learning. Placement Data

Elizaveta German, Roberto Russo, Arvin Rastegar

09/06/2021

Introduction

We decide to analyze what are the factors that influence success in the job market. Our task is to see what are the factors that during education, can help to be placed and have a high salary. So our questions are: what influences the placement? What influences the salary?

```
knitr::opts_chunk$set(fig.height = 9, fig.width = 12)
```

```
# Importing libraries
library('funModeling')
library('tidyverse')
library('Hmisc')
library('dummies')
library('ROCR')
library('corrplot')
library('plotly')
library("car")
library('gridExtra')
library('ISLR')
library('leaps')
library('kableExtra')
library('caret')
library('broom')
library('readr')
library('lmtest')
```

Dataset

We decided to use a dataset about academic and employability factors of the placement. It was already provided in Kaggle and contains data of 215 students from India and their characteristics, listed below:

1) Gender
2) Secondary Education percentage- 10th Grade (ssc_p)
3) Board of Education- Central/ Others (ssc_b)
4) Higher Secondary Education percentage- 12th Grade (hsc_p)
5) Board of Education- Central/ Others (hsc_b)
6) Specialization in Higher Secondary Education (hsc_s)
7) Degree Percentage (degree_p)

8) Field/type of degree education (degree_t)
9) Work Experience (workex)
10) Employability test percentage (etest_p)
11) Specialisation in MBA (specialisation)
12) MBA percentage (mba_p)
13) Status of placement (status)
14) Salary offered (salary)

Methodology

Main Goals:

1) to predict placement of the students from their academic and employability characteristics
2) to predict salary of the students

Objectives:

1) Prepare data for analysis
2) Explore data and describe the sample
3) Build hypotheses based on results of exploratory data analysis
4) Build models
   - binary logistic regression for predicting placement;
   - linear regression for predicting salary;
5) Interpret results

Preparation of data

```r
# Importing the dataset
df <- read.csv(file="Placement_Data_Full_Class.csv",
               header = TRUE,
               sep = ",")
```

```r
# Deleting index variable
df <- df[,-1]

# Replacing empty values with NA
df[df==""] <- NA

# Counting the number of NA values
na_values <- sapply(df,function(x) sum(is.na(x)))

kbl(t(na_values), booktabs = T) %>%
  kable_classic(full_width = F) %>%
  kable_styling(latex_options= c("scale_down", "hold_position", "striped"),
                font_size = 8)
```

gender	ssc_p	ssc_b	hsc_p	hsc_b	hsc_s	degree_p	degree_t	workex	etest_p	specialisation	mba_p	status	salary
0	0	0	0	0	0	0	0	0	0	0	0	0	67

```
# Substituting NA values with 0 for the salary variable
df$salary <- Recode(var = df$salary, recodes = "NA=0", as.factor = FALSE)

#Transforming scale of salary to thousands
df$salary <- df$salary / 1000

# Counting the number of unique values of each variable
unique <- sapply(df, function(x) length(unique(x)))

kbl(t(unique), booktabs = T) %>%
    kable_classic(full_width = F) %>%
    kable_styling(latex_options= c("scale_down", "hold_position", "striped"),
                  font_size = 8)
```

gender	ssc_p	ssc_b	hsc_p	hsc_b	hsc_s	degree_p	degree_t	workex	etest_p	specialisation	mba_p	status	salary
2	103	2	97	2	3	89	3	2	100	2	205	2	46

Exploratory analysis

Numerical variables

The basic exploratory data analysis is performed by plotting histograms of the numerical variables of the dataset.

```
theme_set(theme_test())
numerical <- select(df, ssc_p, hsc_p, degree_p, etest_p, mba_p, salary)
colnames(numerical) <- c(’Sec.Educ.%’, ’Higher Sec.Educ.%’, ’Degree %’,
                         ’Employab. test %’, ’MBA %’, ’Salary (thousands)’)

basic_eda <- function(numerical)
{ plot_num(numerical, bins=5, path_out = ".")
  kbl(summary(numerical), booktabs = T, caption = "Summary of numerical variables",
      col.names = c(’Sec.Educ.%’, ’Higher Sec.Educ.%’,
                    ’Degree %’, ’Employab. test %’, ’MBA %’,
                    ’Salary (thousands)’), valign = ’t’) %>%
      kable_classic(full_width = F) %>%
      kable_styling(latex_options = c("hold_position", "striped"),
                    font_size = 8)}

basic_eda(numerical)
```

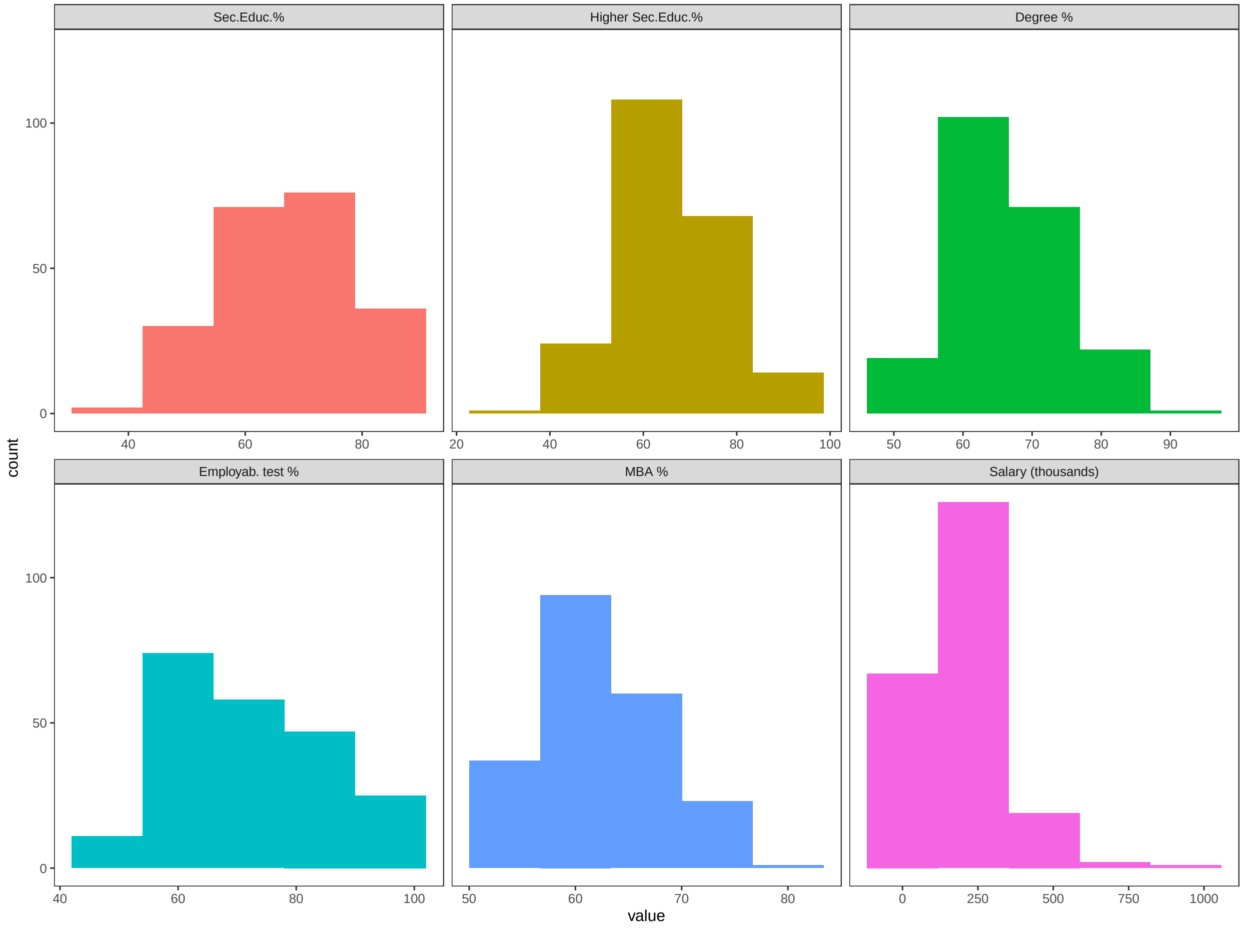


Table 1: Summary of numerical variables

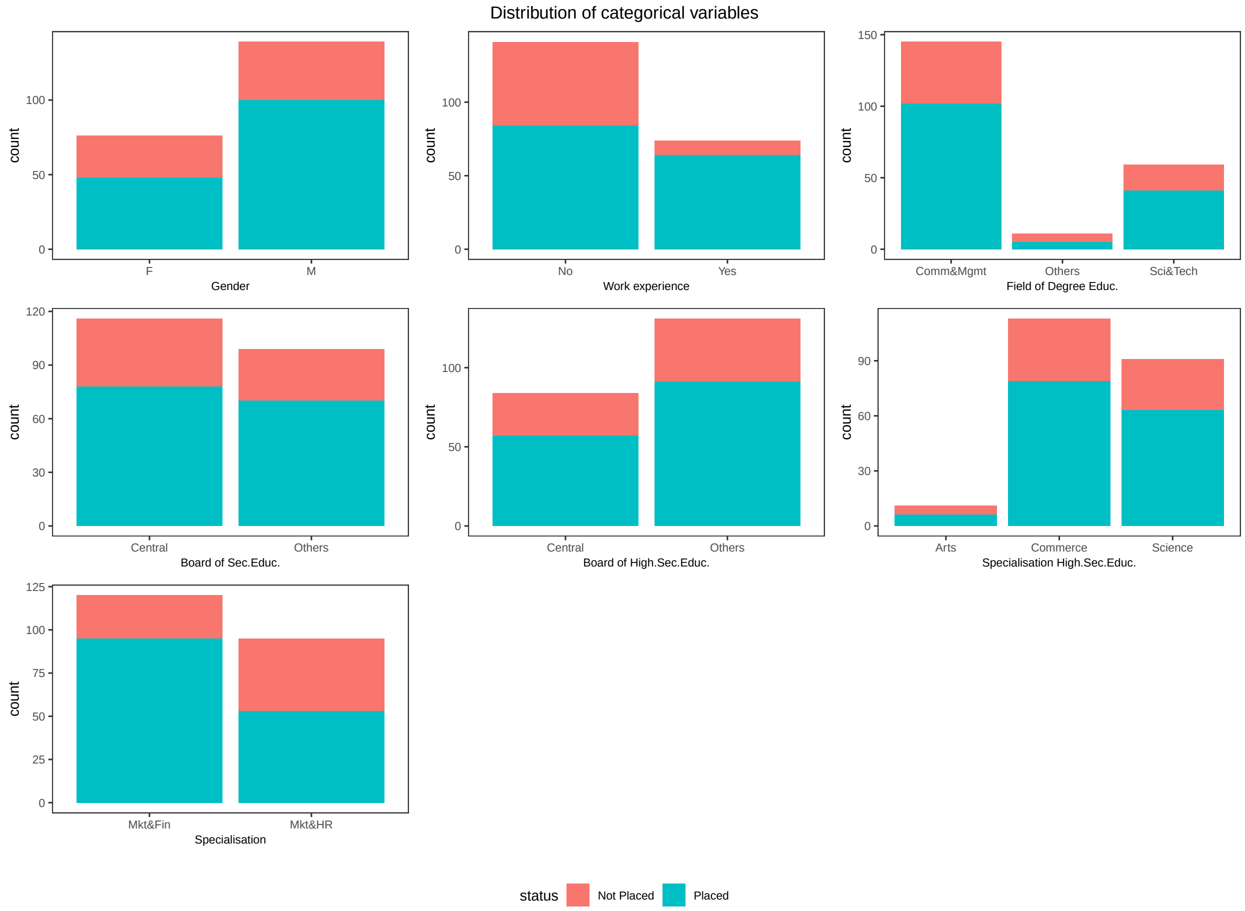
Sec.Educ.%	Higher Sec.Educ.%	Degree %	Employab. test %	MBA %	Salary (thousands)
Min. :40.89	Min. :37.00	Min. :50.00	Min. :50.0	Min. :51.21	Min. : 0.0
1st Qu.:60.60	1st Qu.:60.90	1st Qu.:61.00	1st Qu.:60.0	1st Qu.:57.95	1st Qu.: 0.0
Median :67.00	Median :65.00	Median :66.00	Median :71.0	Median :62.00	Median :240.0
Mean :67.30	Mean :66.33	Mean :66.37	Mean :72.1	Mean :62.28	Mean :198.7
3rd Qu.:75.70	3rd Qu.:73.00	3rd Qu.:72.00	3rd Qu.:83.5	3rd Qu.:66.25	3rd Qu.:282.5
Max. :89.40	Max. :97.70	Max. :91.00	Max. :98.0	Max. :77.89	Max. :940.0

Categorical variables

The graphs below show the distribution of the categorical variables splitted with respect to the variable “Status”.

```
a<-df %>% ggplot(aes(x = gender)) + geom_bar(aes(fill = status)) +
                 labs(x = "Gender") +
                 theme(text = element_text(size=10),
                 axis.title.x = element_text(size=8),
                 legend.position="bottom")
```

```
b<-df %>% ggplot(aes(x = workex)) + geom_bar(aes(fill = status),
                 show.legend = FALSE) +
                 labs(x = "Work experience") +
                 theme(text = element_text(size=10),
                 axis.title.x = element_text(size=8),
                 legend.position="bottom")

c<-df %>% ggplot(aes(x = degree_t)) + geom_bar(aes(fill = status),
                 show.legend = FALSE) +
                 labs(x = "Field of Degree Educ.") +
                 theme(text = element_text(size=10),
                 axis.title.x = element_text(size=8),
                 legend.position="bottom")

d<-df %>% ggplot(aes(x = ssc_b)) + geom_bar(aes(fill = status),
                 show.legend = FALSE) +
                 labs(x = "Board of Sec.Educ.") +
                 theme(text = element_text(size=10),
                 axis.title.x = element_text(size=8),
                 legend.position="bottom")

e<-df %>% ggplot(aes(x = hsc_b)) + geom_bar(aes(fill = status),
                 show.legend = FALSE) +
                 labs(x = "Board of High.Sec.Educ.") +
                 theme(text = element_text(size=10),
                 axis.title.x = element_text(size=8),
                 legend.position="bottom")

f<-df %>% ggplot(aes(x = hsc_s)) + geom_bar(aes(fill = status),
                 show.legend = FALSE) +
                 labs(x = "Specialisation High.Sec.Educ.") +
                 theme(text = element_text(size=10),
                 axis.title.x = element_text(size=8),
                 legend.position="bottom")

g<-df %>% ggplot(aes(x = specialisation)) + geom_bar(aes(fill = status),
                 show.legend = FALSE) +
                 labs(x = "Specialisation") +
                 theme(text = element_text(size=10),
                 axis.title.x = element_text(size=8),
                 legend.position="bottom")

g_legend<-function(a.gplot){
  tmp <- ggplot_gtable(ggplot_build(a.gplot))
  leg <- which(sapply(tmp$grobs, function(x) x$name) == "guide-box")
  legend <- tmp$grobs[[leg]]
  return(legend)}

mylegend<-g_legend(a)
graph <- grid.arrange(arrangeGrob(a + theme(legend.position = "none"),
                      b, c, d, e, f, g, nrow=3),
                      top = "Distribution of categorical variables",
                      mylegend, nrow = 2, heights=c(10, 1))
```



From the graphs above we can see for the variables secondary school board and higher secondary school board do not affect much the placement, because the percentage of placed people for each category is very similar. While the remaining variables seem to influence the percentage of the placement.The percentage of placed men is sligthly higher than the percentage of placed women. The variable Work experience seems to have a big impact on the placement, because just a little portion of the people who have work experience are not placed, while the same cannot be said for people without work experience. Also the distribution of specialization has similar behavior. The field of degree seems also to play a role, hence we observe that placed people graduated in Commerce and management and scientific fields are relatively more than people graduated in other fields, as well as secondary school specialization. Below are printed the percentage.

```
# Gender
summary(factor(filter(df, gender=='M')$status))/length(filter(df, gender=='M')$status)
```

```
## Not Placed     Placed
##  0.2805755  0.7194245
```

```
summary(factor(filter(df, gender=='F')$status))/length(filter(df, gender=='F')$status)
```

```
## Not Placed     Placed
##  0.3684211  0.6315789
```

```
# Board of Secondary School Education
summary(factor(filter(df, ssc_b==’Central’)$status))/length(filter(df,ssc_b==’Central’)$status)
```

```
## Not Placed     Placed
##  0.3275862  0.6724138
```

```
summary(factor(filter(df, ssc_b==’Others’)$status))/length(filter(df, ssc_b==’Others’)$status)
```

```
## Not Placed     Placed
##  0.2929293  0.7070707
```

```
# Board of Higher School Education
summary(factor(filter(df, hsc_b==’Central’)$status))/length(filter(df, hsc_b==’Central’)$status)
```

```
## Not Placed     Placed
##  0.3214286  0.6785714
```

```
summary(factor(filter(df, hsc_b==’Others’)$status))/length(filter(df, hsc_b==’Others’)$status)
```

```
## Not Placed     Placed
##  0.3053435  0.6946565
```

```
# Higher School specialization
summary(factor(filter(df, hsc_s==’Commerce’)$status))/length(filter(df, hsc_s==’Commerce’)$status)
```

```
## Not Placed     Placed
##   0.300885   0.699115
```

```
summary(factor(filter(df, hsc_s==’Science’)$status))/length(filter(df, hsc_s==’Science’)$status)
```

```
## Not Placed     Placed
##  0.3076923  0.6923077
```

```
summary(factor(filter(df, hsc_s==’Arts’)$status))/length(filter(df, hsc_s==’Arts’)$status)
```

```
## Not Placed     Placed
##  0.4545455  0.5454545
```

```
# Field of Degree Education
summary(factor(filter(df, degree_t==’Comm&Mgmt’)$status))/length(filter(df, degree_t==’Comm&Mgmt’)$statu
```

```
## Not Placed     Placed
##  0.2965517  0.7034483
```

```
summary(factor(filter(df, degree_t==’Sci&Tech’)$status))/length(filter(df, degree_t==’Sci&Tech’)$status
```

```
## Not Placed     Placed
##  0.3050847  0.6949153
```

```
summary(factor(filter(df, degree_t==’Others’)$status))/length(filter(df, degree_t==’Others’)$status)
```

```
## Not Placed     Placed
##  0.5454545  0.4545455
```

```
# Work experience
summary(factor(filter(df, workex==’Yes’)$status))/length(filter(df, workex==’Yes’)$status)
```

```
## Not Placed     Placed
##  0.1351351  0.8648649
```

```
summary(factor(filter(df, workex==’No’)$status))/length(filter(df, workex==’No’)$status)
```

```
## Not Placed     Placed
##  0.4042553  0.5957447
```

```
# Specialization in MBA
summary(factor(filter(df, specialisation==’Mkt&Fin’)$status))/length(filter(df, specialisation==’Mkt&Fi
```

```
## Not Placed     Placed
##  0.2083333  0.7916667
```

```
summary(factor(filter(df, specialisation==’Mkt&HR’)$status))/length(filter(df, specialisation==’Mkt&HR’)
```

```
## Not Placed     Placed
##  0.4421053  0.5578947
```

Placed vs. Not placed

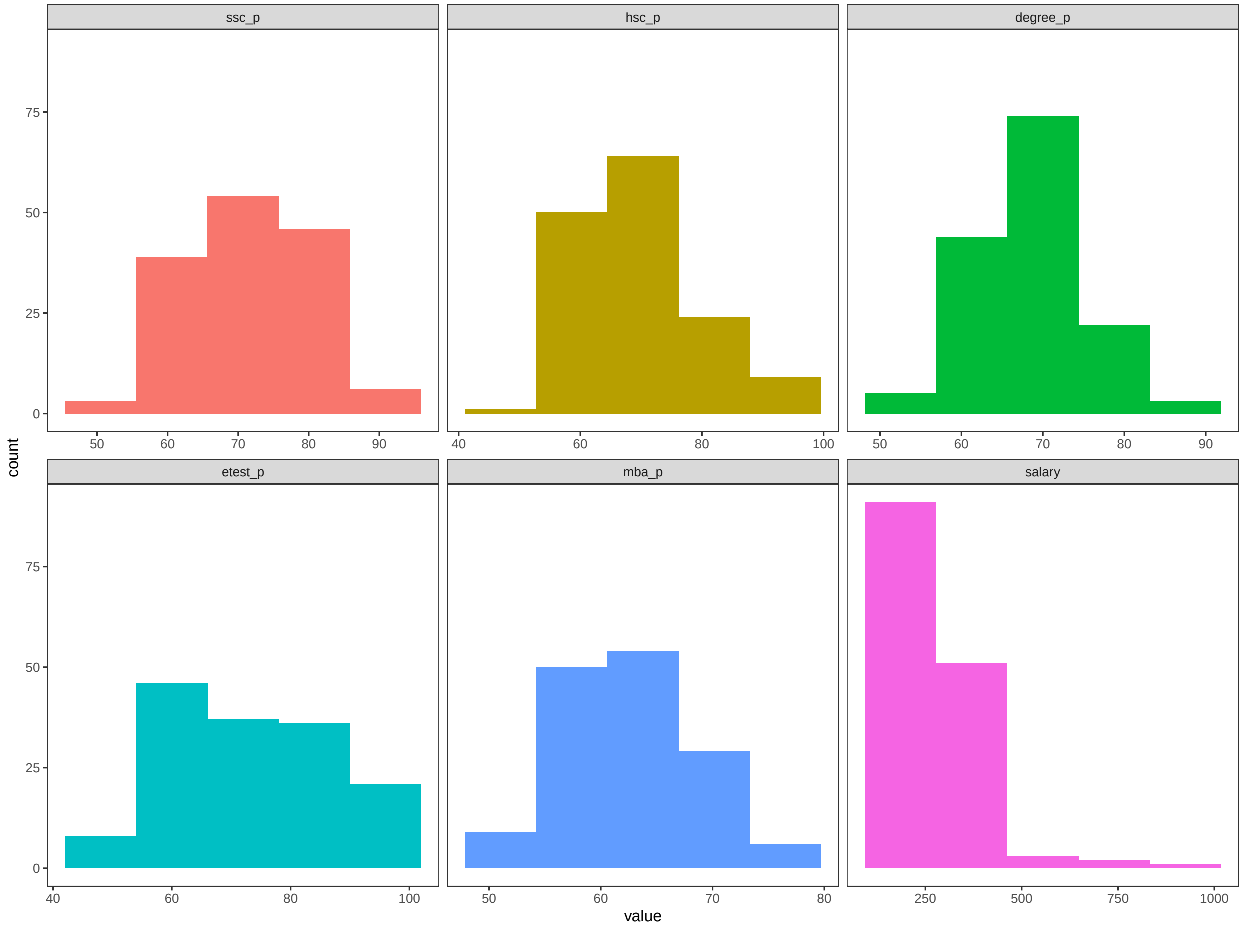
Next each numerical variable will be splitted into “Placed” and “Not placed” to see the difference in the scores that students got while studying at school and university.

```
datap<-filter(df, status==’Placed’)
datanp<-filter(df, status==’Not Placed’)
datap <- select(datap, ssc_p, hsc_p, degree_p, etest_p, mba_p, salary)
datanp<- select(datanp, ssc_p, hsc_p, degree_p, etest_p, mba_p)


kbl(summary(datap), booktabs = T, caption = "Placed",
    col.names = c(’Sec.Educ.%’, ’Higher Sec.Educ.%’,
                  ’Degree %’, ’Employab. test %’, ’MBA %’,
                  ’Salary (thousands)’), valign = ’t’) %>%
    kable_classic(full_width = F) %>%
    kable_styling(latex_options = c("scale_down", "hold_position", "striped"),
                  font_size = 8)
```

Table 2: Placed

Sec.Educ.%	Higher Sec.Educ.%	Degree %	Employab. test %	MBA %	Salary (thousands)
Min. :49.00	Min. :50.83	Min. :56.00	Min. :50.00	Min. :52.38	Min. :200.0
1st Qu.:65.00	1st Qu.:63.00	1st Qu.:65.00	1st Qu.:60.00	1st Qu.:57.77	1st Qu.:240.0
Median :72.50	Median :68.00	Median :68.00	Median :72.00	Median :62.24	Median :265.0
Mean :71.72	Mean :69.93	Mean :68.74	Mean :73.24	Mean :62.58	Mean :288.7
3rd Qu.:78.12	3rd Qu.:75.25	3rd Qu.:72.42	3rd Qu.:85.00	3rd Qu.:66.76	3rd Qu.:300.0
Max. :89.40	Max. :97.70	Max. :91.00	Max. :98.00	Max. :77.89	Max. :940.0

```r
plot_num(datap, bins=5,path_out = ".")
```

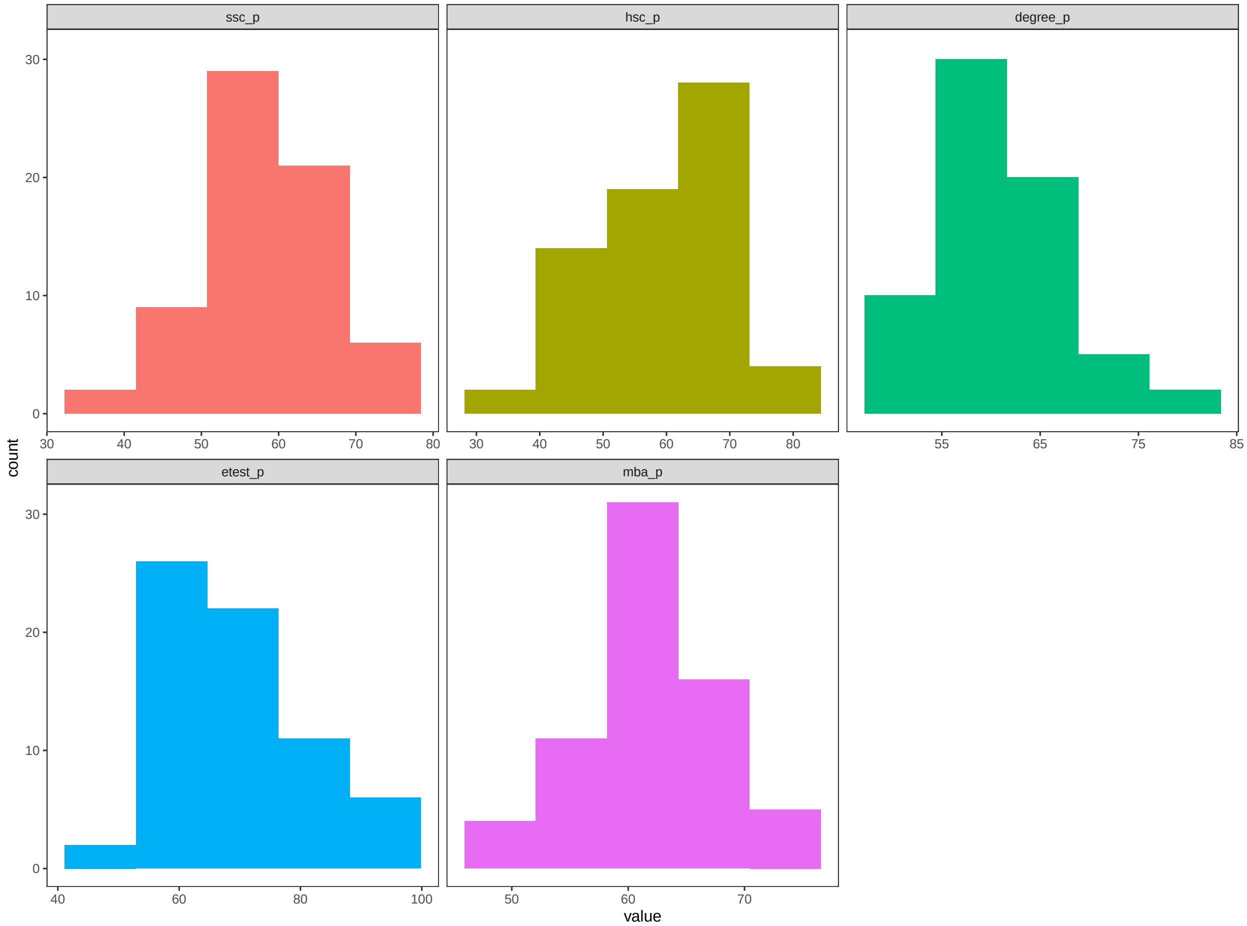


```r
kbl(summary(datanp), booktabs = T, caption = "Not Placed",
    col.names = c(’Sec.Educ.%’, ’Higher Sec.Educ.%’,
                  ’Degree %’, ’Employab. test %’, ’MBA %’), valign = ’t’) %>%
    kable_classic(full_width = F) %>%
    kable_styling(latex_options = c("scale_down","hold_position", "striped"),
                  font_size = 8)
```

Table 3: Not Placed

Sec.Educ.%	Higher Sec.Educ.%	Degree %	Employab. test %	MBA %
Min. :40.89	Min. :37.00	Min. :50.00	Min. :50.00	Min. :51.21
1st Qu.:52.00	1st Qu.:51.00	1st Qu.:57.00	1st Qu.:60.00	1st Qu.:58.48
Median :56.28	Median :60.33	Median :61.00	Median :67.00	Median :60.69
Mean :57.54	Mean :58.40	Mean :61.13	Mean :69.59	Mean :61.61
3rd Qu.:63.00	3rd Qu.:64.00	3rd Qu.:65.00	3rd Qu.:76.50	3rd Qu.:65.41
Max. :77.80	Max. :82.00	Max. :79.00	Max. :97.00	Max. :75.71

```r
plot_num(datanp, bins=5,path_out = ".")
```



The first graph shows the distribution of the numerical variables with the status='Placed', and the second graph shows the distribution of the numerical variables with the status='Not Placed'. It can be seen that there are differences for each variable, mean and median values are higher for the 'Placed' distributions. It is allowed to assume that placed students got on average better scores while they were studying.

Nevertheless, it could be difficult to predict the placement looking just at one variable because for each predictor the range of the "Placed" distribution overlaps with the range of "Not Placed" distribution. But combining the information from each of the variables can help to enhance prediction.

Salary distribution by categorical variables

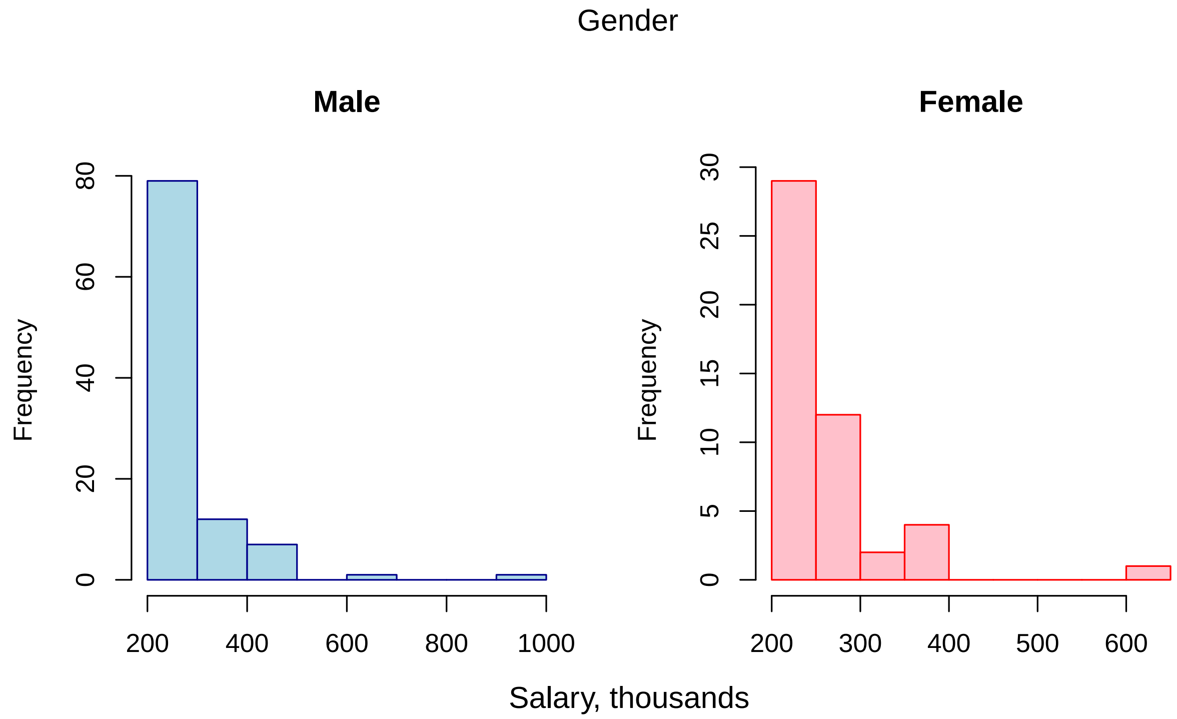
Furthermore, it is explored if any of categorical variables can affect the salary. To do this firstly salary variable should be split for each categorical variable. According to each category the histograms of the distributions are plotted and basic statistics are computed. It should be mentioned that only students, who are placed, are selected because salary data of non-placed students is not valid. That is the main limitation of the current project.

```r
datagenM<-filter(df, gender==’M’, status==’Placed’)
datagenF<-filter(df, gender==’F’, status==’Placed’)

g1 <- as.array(summary(datagenM$salary))
g2 <- as.array(summary(datagenF$salary))
t <- as.data.frame(list(g1, g2))
t <- t[,-3]
kbl(t, digits = 2, booktabs = T, caption = "Salary by Gender",
    col.names = c("Statistics", "Male", "Female"), valign=’t’) %>%
    kable_styling(latex_options = c("hold_position", "striped"))
```

Table 4: Salary by Gender

Statistics	Male	Female
Min.	200.00	200.00
1st Qu.	250.00	219.50
Median	270.00	250.00
Mean	298.91	267.29
3rd Qu.	300.00	300.00
Max.	940.00	650.00

```r
par(mfrow=c(1,2), oma = c(0, 0, 2, 0))
x <- hist(datagenM$salary, main=’Male’, xlab = NULL, col = "lightblue",
          border = "darkblue")
y <- hist(datagenF$salary, main=’Female’, xlab = NULL, col = "pink",
          border = "red")

mtext("Gender", outer = TRUE, cex = 1.2)
mtext("Salary, thousands", cex = 1.2, adj = -1.3, side=1, padj=4)
```



Gender

It can be seen that men earn on average more than women. The difference in the relative distributions is just for the interval 200k-300k that almost corresponds to the inter-quartile range of the female distribution, while for the male distribution it is smaller because the first quartile is higher, and the third is the same. We can conclude that according to the data, if the person is a male, it is less probable that he will have an annual salary smaller than 250k. But we cannot say more about the average income of men and women, because in our data this value may be affected by outliers. For that reason we inspect the outliers to decide what to keep to compute the statistics.

```
m<-length(datagenM$salary[datagenM$salary>300 & datagenM$salary<600])
f<-length(datagenF$salary[datagenF$salary>300 & datagenF$salary<600])
paste(’percentage of outliers lower than 600000 for men:’,
      m/length(datagenM$salary))
```

```
## [1] "percentage of outliers lower than 600000 for men: 0.19"
```

```
paste(’percentage of outliers lower than 600000 for women:’,
      f/length(datagenF$salary))
```

```
## [1] "percentage of outliers lower than 600000 for women: 0.125"
```

```
paste(’number of outliers higher than 600000 for men:’,
      length(datagenM$salary[datagenM$salary>=600]))
```

```
## [1] "number of outliers higher than 600000 for men: 2"
```

```
paste('number of outliers higher than 600000 for women:',
      length(datagenF$salary[datagenF$salary>=600]))
```

```
## [1] "number of outliers higher than 600000 for women: 1"
```

We see that outliers between 300k and 600k account for 19% and 12.5% for men and women respectively. So we decide to keep them. We remove the values higher than 600k because they are not relevant.

```
# Male
summary(datagenM$salary[datagenM$salary<600])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   200.0   250.0   270.0   288.4   300.0   500.0
```

```
# Female
summary(datagenF$salary[datagenF$salary<600])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   200.0   219.0   250.0   259.1   293.5   400.0
```

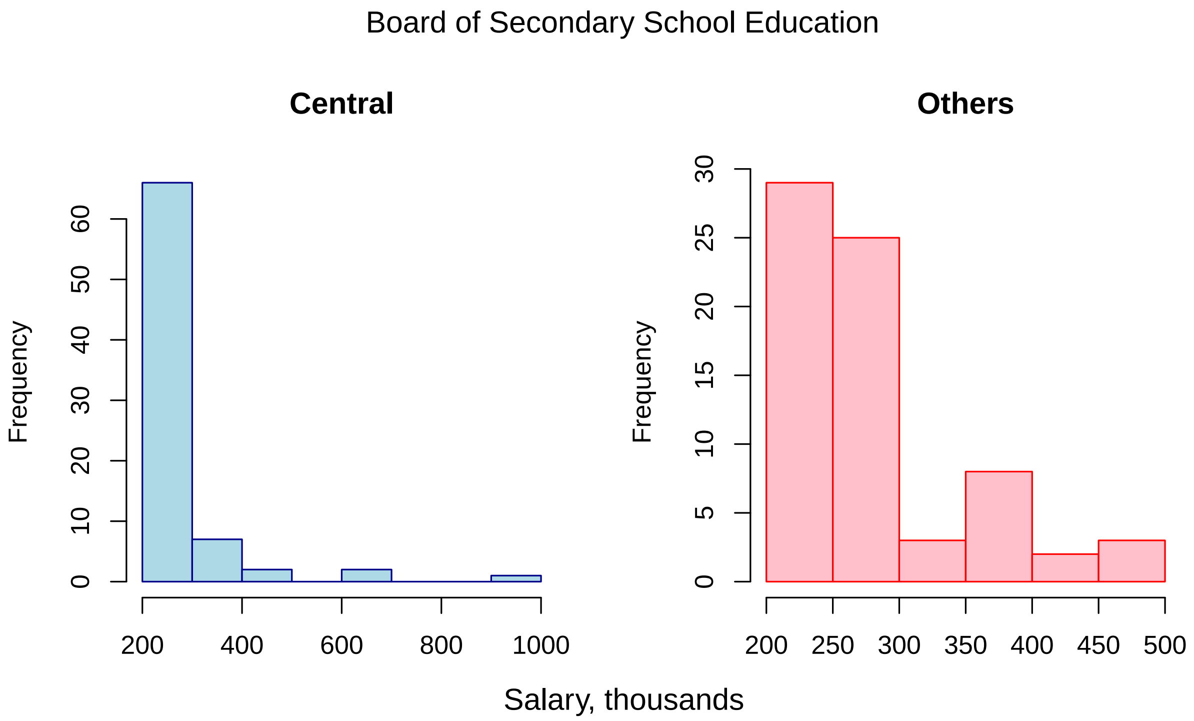
We can see that the statistics did not change much, even if we can observe that there are not women that earn more than 400k. So according to our data, the probability to have an income higher than 400k is higher if you are a man. But, except for the differences we found above, the two distributions show similar behavior and cannot be unequivocally separated.

```
datassc_bC<-filter(df, ssc_b=='Central', status=='Placed')
datassc_bO<-filter(df, ssc_b=='Others', status=='Placed')

g1 <- as.array(summary(datassc_bC$salary))
g2 <- as.array(summary(datassc_bO$salary))
t <- as.data.frame(list(g1, g2))
t <- t[,-3]
kbl(t, digits = 2, booktabs = T, caption = "Salary by Board of Secondary School Education",
    col.names = c("Statistics", "Central", "Others"), valign='t') %>%
  kable_styling(latex_options = c("hold_position", "striped"))
```

Table 5: Salary by Board of Secondary School Education

Statistics	Central	Others
Min.	200.00	200.0
1st Qu.	240.00	240.0
Median	260.00	265.0
Mean	288.17	289.2
3rd Qu.	300.00	300.0
Max.	940.00	500.0

```
par(mfrow=c(1,2), oma = c(0, 0, 2, 0))
x <- hist(datassc_bC$salary, main='Central', xlab = NULL, col = "lightblue",
          border = "darkblue")
y <- hist(datassc_bO$salary, main='Others', xlab = NULL, col = "pink",
          border = "red")
mtext("Board of Secondary School Education", outer = TRUE, cex = 1.2)
mtext("Salary, thousands", cex = 1.2, adj = -1.3, side=1, padj=4)
```



Board of Secondary School Education

Here we observe statistics that are almost equal among two distributions, except for the max value. Again we inspect the outliers.

```
m<-length(datassc_bC$salary[datassc_bC$salary>300 & datassc_bC$salary<600])
f<-length(datassc_bO$salary[datassc_bO$salary>300 & datassc_bO$salary<600])
paste('percentage of outliers lower than 600000 for Central:',
      m/length(datassc_bC$salary))
```

```
## [1] "percentage of outliers lower than 600000 for Central: 0.115384615384615"
```

```
paste('percentage of outliers lower than 600000 for Others:',
      f/length(datassc_bO$salary))
```

```
## [1] "percentage of outliers lower than 600000 for Others: 0.228571428571429"
```

```
paste(’number of outliers higher than 600000 for Central:’,
      length(datassc_bC$salary[datassc_bC$salary>=600]))
```

```
## [1] "number of outliers higher than 600000 for Central: 3"
```

```
paste(’number of outliers higher than 600000 for Others:’,
      length(datassc_bO$salary[datassc_bO$salary>=600]))
```

```
## [1] "number of outliers higher than 600000 for Others: 0"
```

Again we remove the outliers that are higher than 600k and compute the statistics.

```
summary(datassc_bC$salary[datassc_bC$salary<600])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   200.0   240.0   260.0   269.3   300.0   425.0
```

```
summary(datassc_bO$salary[datassc_bO$salary<600])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   200.0   240.0   265.0   289.2   300.0   500.0
```

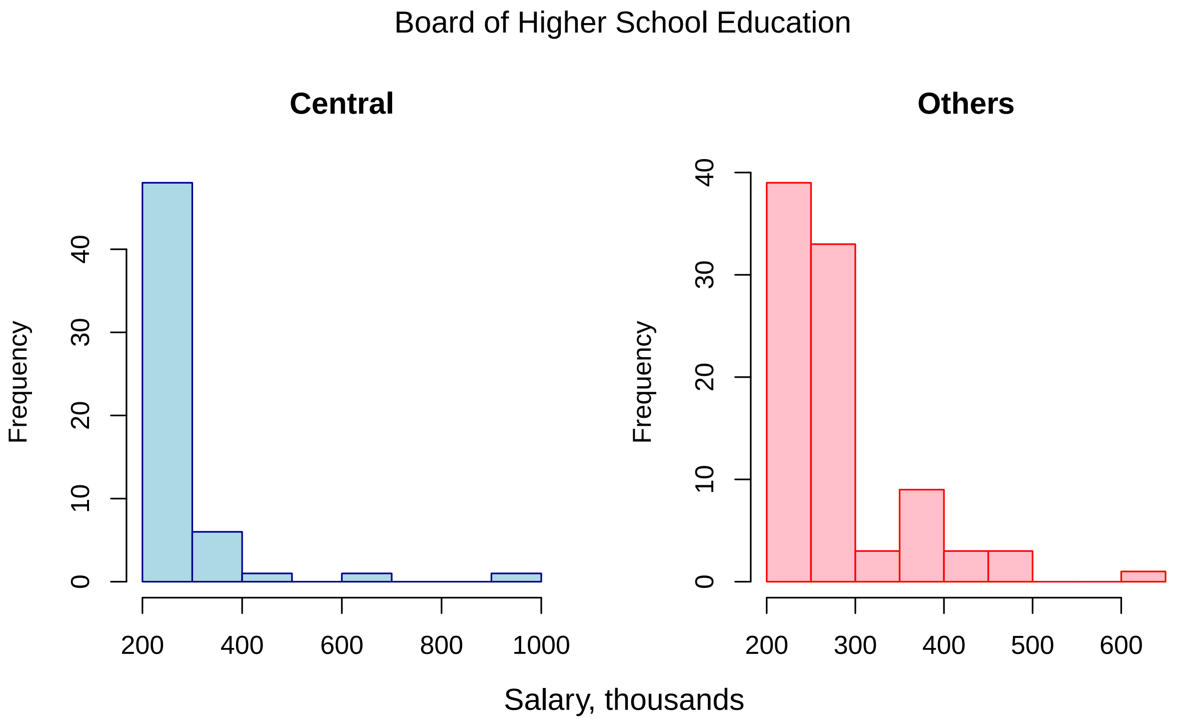
Here we can see that outliers we removed were playing a role in computing the mean of the "Central" distribution. We can say that people who studied in other secondary school board earn more on average. Also, the probability to earn more than 425k is higher if the school board is not central, but still, we cannot infer the amount of salary from the secondary school board.

```
datahsc_bC<-filter(df, hsc_b==’Central’, status==’Placed’)
datahsc_bO<-filter(df, hsc_b==’Others’, status==’Placed’)

g1 <- as.array(summary(datahsc_bC$salary))
g2 <- as.array(summary(datahsc_bO$salary))
t <- as.data.frame(list(g1, g2))
t <- t[,-3]
kbl(t, digits = 2, booktabs = T, caption = "Salary by Board of Higher School Education",
    col.names = c("Statistics", "Central", "Others"), valign=’t’) %>%
    kable_styling(latex_options = c("hold_position", "striped"))
```

Table 6: Salary by Board of Higher School Education

Statistics	Central	Others
Min.	200.00	200.0
1st Qu.	240.00	240.0
Median	260.00	265.0
Mean	289.54	288.1
3rd Qu.	300.00	300.0
Max.	940.00	650.0

```
par(mfrow=c(1,2), oma = c(0, 0, 2, 0))
x <- hist(datahsc_bC$salary, main='Central', xlab = NULL, col = "lightblue",
          border = "darkblue")
y <- hist(datahsc_bO$salary, main='Others', xlab = NULL, col = "pink",
          border = "red")
mtext("Board of Higher School Education", outer = TRUE, cex = 1.2)
mtext("Salary, thousands", cex = 1.2, adj = -1.3, side=1, padj=4)
```



Board of Higher School Education

Doing the same as before we remove outliers and compute the statistics.

```
m<-length(datahsc_bC$salary[datahsc_bC$salary>300 & datahsc_bC$salary<600])
f<-length(datahsc_bO$salary[datahsc_bO$salary>300 & datahsc_bO$salary<600])
paste('percentage of outliers lower than 600000 for Central:',
      m/length(datahsc_bC$salary))
```

```
## [1] "percentage of outliers lower than 600000 for Central: 0.12280701754386"
```

```
paste('percentage of outliers lower than 600000 for Others:',
      f/length(datahsc_bO$salary))
```

```
## [1] "percentage of outliers lower than 600000 for Others: 0.197802197802198"
```

```r
paste(’number of outliers higher than 600000 for Central:’,
      length(datahsc_bC$salary[datahsc_bC$salary>=600]))
```

```
## [1] "number of outliers higher than 600000 for Central: 2"
```

```r
paste(’number of outliers higher than 600000 for Others:’,
      length(datahsc_bO$salary[datahsc_bO$salary>=600]))
```

```
## [1] "number of outliers higher than 600000 for Others: 1"
```

```r
summary(datahsc_bC$salary[datahsc_bC$salary<600])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   200.0   240.0   260.0   270.4   300.0   425.0
```

```r
summary(datahsc_bO$salary[datahsc_bO$salary<600])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   200.0   240.0   265.0   284.1   300.0   500.0
```

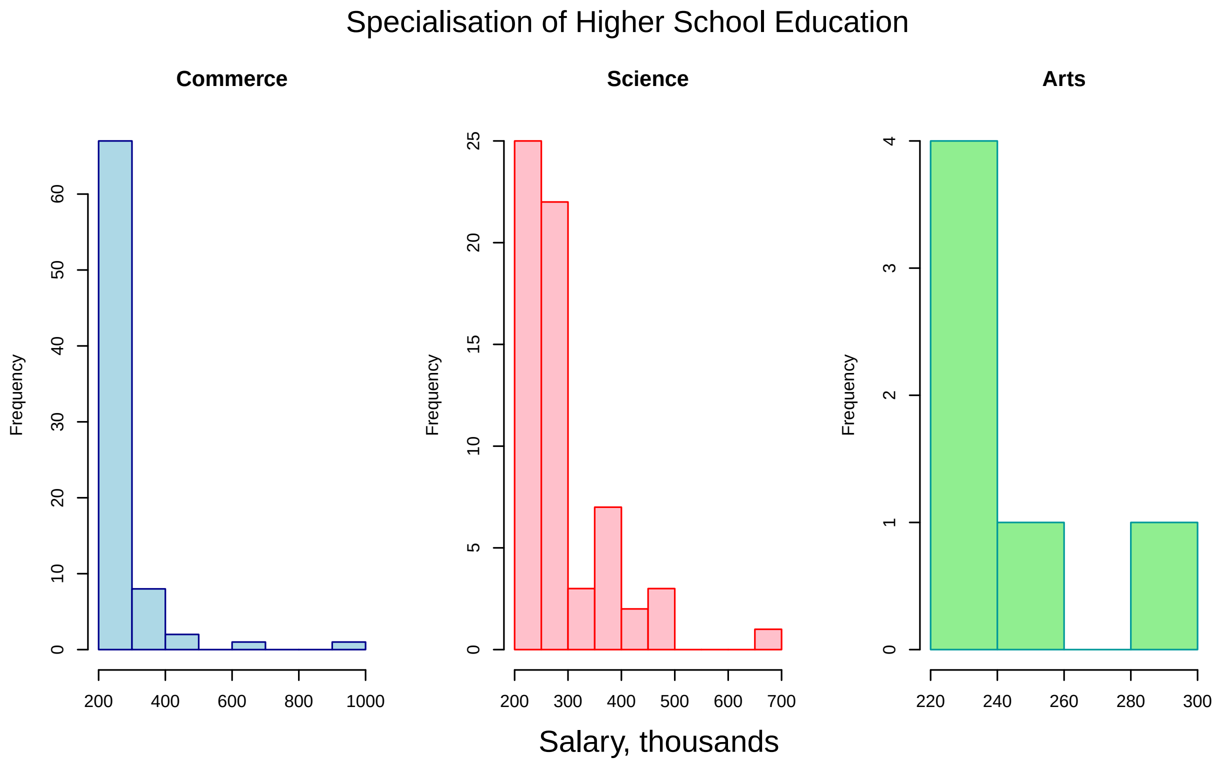
The same conclusion is applied as in the previous case. It is understandable since the different boards of education prepare students for different university career track, and consequently different working career track, we can expect the similarity in salary among people who have chosen the same board in both secondary school and higher secondary school.

```r
datahsc_sC<-filter(df, hsc_s==’Commerce’, status==’Placed’)
datahsc_sS<-filter(df, hsc_s==’Science’, status==’Placed’)
datahsc_sA<-filter(df, hsc_s==’Arts’, status==’Placed’)

g1 <- as.array(summary(datahsc_sC$salary))
g2 <- as.array(summary(datahsc_sS$salary))
g3 <- as.array(summary(datahsc_sA$salary))
t <- as.data.frame(list(g1, g2, g3))
t <- t[,-3]
t <- t[,-4]
kbl(t, digits = 2, booktabs = T,
    caption = "Salary by Specialisation of Higher School Education",
    col.names = c("Statistics", "Commerce", "Science", "Arts"), valign=’t’) %>%
    kable_styling(latex_options = c("hold_position", "striped"))
```

```r
par(mfrow=c(1,3), oma = c(0, 0, 2, 0))
x <- hist(datahsc_sC$salary, main=’Commerce’, xlab = NULL, col = "lightblue",
          border = "darkblue")
y <- hist(datahsc_sS$salary, main=’Science’, xlab = NULL, col = "pink",
          border = "red")
z <-hist(datahsc_sA$salary, main=’Arts’, xlab = NULL, col = "lightgreen",
          border = "#009999")
mtext("Specialisation of Higher School Education", outer = TRUE, cex = 1.2)
mtext("Salary, thousands", cex = 1.2, adj = -8, side=1, padj=3)
```

Table 7: Salary by Specialisation of Higher School Education

Statistics	Commerce	Science	Arts
Min.	200.00	200.00	230.00
1st Qu.	240.00	240.00	236.00
Median	265.00	260.00	238.00
Mean	287.42	294.02	248.67
3rd Qu.	300.00	310.00	247.50
Max.	940.00	690.00	300.00



Specialisation of Higher School Education

Here we have outliers just on two of the three distribution, so we remove it as before.

```
m<-length(datahsc_sC$salary[datahsc_sC$salary>300 & datahsc_sC$salary<600])
f<-length(datahsc_sS$salary[datahsc_sS$salary>300 & datahsc_sS$salary<600])
paste('percentage of outliers lower than 600000 for Commerce:',
      m/length(datahsc_sC$salary))
```

```
## [1] "percentage of outliers lower than 600000 for Commerce: 0.126582278481013"
```

```
paste('percentage of outliers lower than 600000 for Science:',
      f/length(datahsc_sS$salary))
```

```
## [1] "percentage of outliers lower than 600000 for Science: 0.238095238095238"
```

```
paste(’number of outliers higher than 600000 for Commerce:’,
      length(datahsc_sC$salary[datahsc_sC$salary>=600]))
```

```
## [1] "number of outliers higher than 600000 for Commerce: 2"
```

```
paste(’number of outliers higher than 600000 for Science:’,
      length(datahsc_sS$salary[datahsc_sS$salary>=600]))
```

```
## [1] "number of outliers higher than 600000 for Science: 1"
```

```
summary(datahsc_sC$salary[datahsc_sC$salary<600])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   200.0   240.0   265.0   274.2   300.0   425.0
```

```
summary(datahsc_sS$salary[datahsc_sS$salary<600])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   200.0   240.0   260.0   287.6   300.0   500.0
```

```
summary(datahsc_sA$salary)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   230.0   236.0   238.0   248.7   247.5   300.0
```

Also here the distributions are very similar, except for the fact that people with higher secondary school specialization in Science have a higher probability to earn more than 425k, and for the fact that people with higher secondary school specialization in Art earn generally less than the others. But the range of the “Art” distribution is comprised in the range of the other two distribution, so it cannot be distinguished.

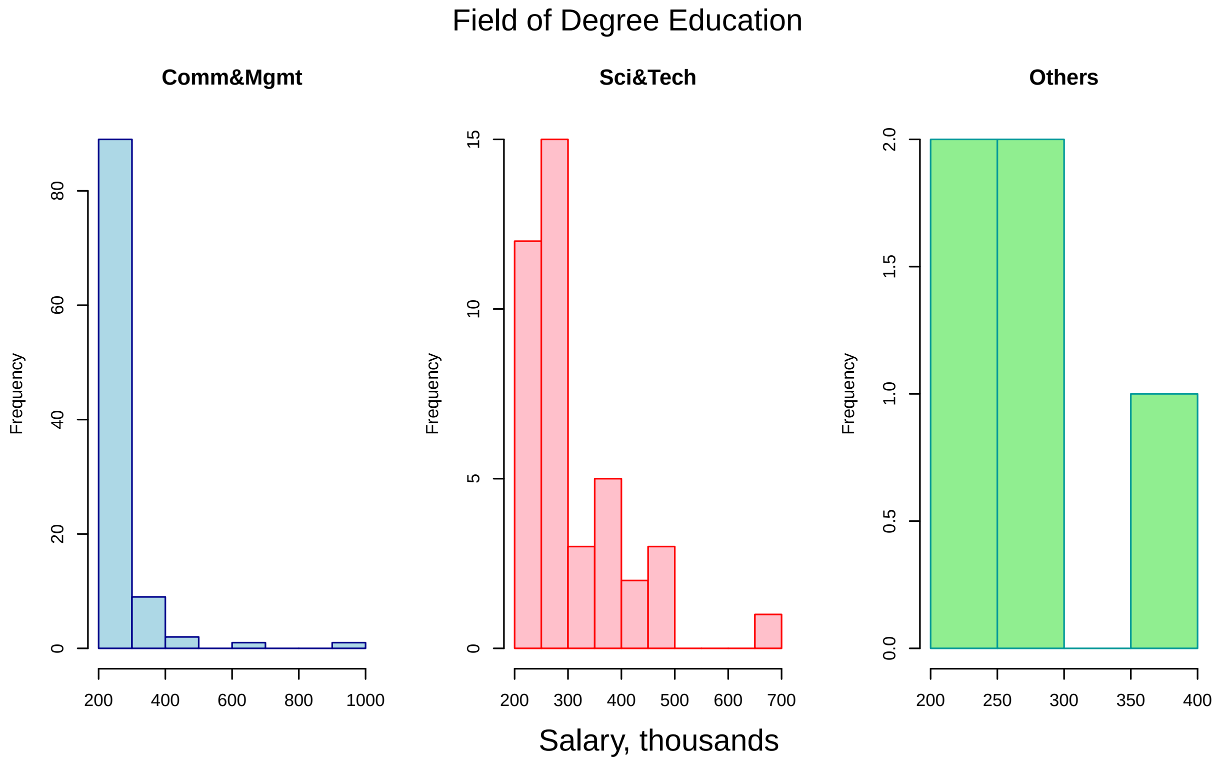
```
datadegree_tC<-filter(df, degree_t==’Comm&Mgmt’, status==’Placed’)
datadegree_tS<-filter(df, degree_t==’Sci&Tech’, status==’Placed’)
datadegree_tO<-filter(df, degree_t==’Others’, status==’Placed’)

g1 <- as.array(summary(datadegree_tC$salary))
g2 <- as.array(summary(datadegree_tS$salary))
g3 <- as.array(summary(datadegree_tO$salary))
t <- as.data.frame(list(g1, g2, g3))
t <- t[,-3]
t <- t[,-4]
kbl(t, digits = 2, booktabs = T, caption = "Salary by Field of Degree Education",
    col.names = c("Statistics", "Comm&Mgmt", "Sci&Tech", "Others"), valign=’t’) %>%
    kable_styling(latex_options = c("hold_position", "striped"))
```

```
par(mfrow=c(1,3), oma = c(0, 0, 2, 0))
x <- hist(datadegree_tC$salary, main=’Comm&Mgmt’, xlab = NULL,
          col = "lightblue", border = "darkblue")
y <- hist(datadegree_tS$salary, main=’Sci&Tech’, xlab = NULL, col = "pink",
```

Table 8: Salary by Field of Degree Education

Statistics	Comm&Mgmt	Sci&Tech	Others
Min.	200.00	200.00	240.0
1st Qu.	237.00	250.00	250.0
Median	260.00	275.00	252.0
Mean	278.63	314.61	280.4
3rd Qu.	300.00	360.00	300.0
Max.	940.00	690.00	360.0

```
         border = "red")
z <- hist(datadegree_tO$salary, main = 'Others', xlab = NULL,
         col = "lightgreen", border = "#009999")
mtext("Field of Degree Education", outer = TRUE, cex = 1.2)
mtext("Salary, thousands", cex = 1.2, adj = -8, side=1, padj=3)
```



Field of Degree Education

Doing the same thing as before, we proceed to remove the outliers, where it is needed, even if it is clear that people who graduated in the Science&Technology earn generally more than others.

```
m<-length(datadegree_tC$salary[datadegree_tC$salary>300 & datadegree_tC$salary<600])
f<-length(datadegree_tS$salary[datadegree_tS$salary>360 & datadegree_tS$salary<600])
paste('percentage of outliers lower than 600000 for Comm&Mgmt:',
      m/length(datadegree_tC$salary))
```

```
## [1] "percentage of outliers lower than 600000 for Comm&Mgmt: 0.107843137254902"
```

```
paste(’percentage of outliers lower than 600000 for Sci&Tech:’,
      f/length(datadegree_tS$salary))
```

```
## [1] "percentage of outliers lower than 600000 for Sci&Tech: 0.170731707317073"
```

```
paste(’number of outliers higher than 600000 for Comm&Mgmt:’,
      length(datadegree_tC$salary[datadegree_tC$salary>=600]))
```

```
## [1] "number of outliers higher than 600000 for Comm&Mgmt: 2"
```

```
paste(’number of outliers higher than 600000 for Sci&Tech:’,
      length(datadegree_tS$salary[datadegree_tS$salary>=600]))
```

```
## [1] "number of outliers higher than 600000 for Sci&Tech: 1"
```

```
summary(datadegree_tC$salary[datadegree_tC$salary<600])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   200.0   236.0   260.0   268.3   300.0   425.0
```

```
summary(datadegree_tS$salary[datadegree_tS$salary<600])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   200.0   250.0   275.0   305.2   352.5   500.0
```

```
summary(datadegree_tO$salary)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   240.0   250.0   252.0   280.4   300.0   360.0
```

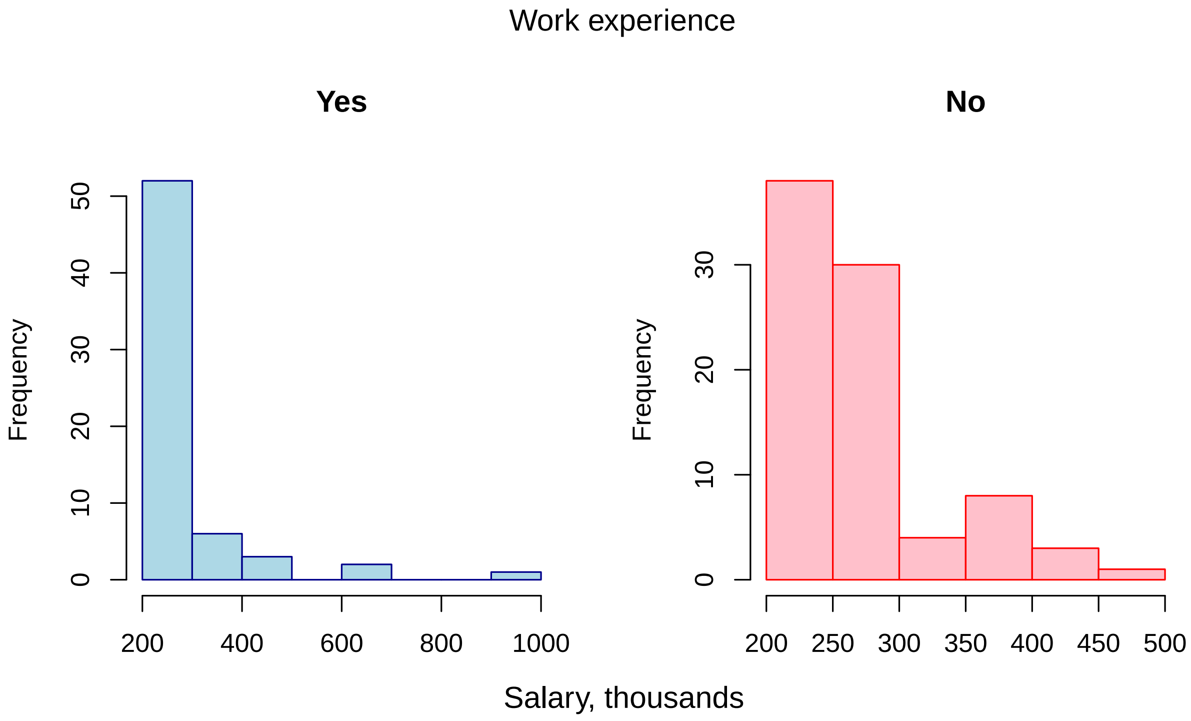
Confirming what was said before, it cannot be observed precise boundary that allows to distinguish these distributions.

```
datawexY<-filter(df, workex==’Yes’, status==’Placed’)
datawexN<-filter(df, workex==’No’, status==’Placed’)

g1 <- as.array(summary(datawexY$salary))
g2 <- as.array(summary(datawexN$salary))
t <- as.data.frame(list(g1, g2))
t <- t[,-3]
kbl(t, digits = 2, booktabs = T, caption = "Salary by Work experience",
    col.names = c("Statistics", "Yes", "No"), valign=’t’) %>%
    kable_styling(latex_options = c("hold_position", "striped"))
```

Table 9: Salary by Work experience

Statistics	Yes	No
Min.	200.00	200.00
1st Qu.	250.00	240.00
Median	267.50	262.00
Mean	303.27	277.52
3rd Qu.	300.00	300.00
Max.	940.00	500.00

```
par(mfrow=c(1,2), oma = c(0, 0, 2, 0))
x <- hist(datawexY$salary, main=’Yes’, xlab = NULL,
          col = "lightblue", border = "darkblue")
y <- hist(datawexN$salary, main=’No’, xlab = NULL, col = "pink",
          border = "red")
mtext("Work experience", outer = TRUE, cex = 1.2)
mtext("Salary, thousands", cex = 1.2, adj = -1.3, side=1, padj=4)
```



Work experience

```
m<-length(datawexY$salary[datawexY$salary>300 & datawexY$salary<600])
f<-length(datawexN$salary[datawexN$salary>300 & datawexN$salary<600])
paste(’percentage of outliers lower than 600000 for Yes:’,
      m/length(datawexY$salary))
```

```
## [1] "percentage of outliers lower than 600000 for Yes: 0.140625"
```

```
paste(’percentage of outliers lower than 600000 for No:’,
      f/length(datawexN$salary))
```

```
## [1] "percentage of outliers lower than 600000 for No: 0.19047619047619"
```

```
paste(’number of outliers higher than 600000 for Yes:’,
      length(datawexY$salary[datawexY$salary>=600]))
```

```
## [1] "number of outliers higher than 600000 for Yes: 3"
```

```
paste(’number of outliers higher than 600000 for No:’,
      length(datawexN$salary[datawexN$salary>=600]))
```

```
## [1] "number of outliers higher than 600000 for No: 0"
```

```
summary(datawexY$salary[datawexY$salary<600])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   200.0   250.0   265.0   280.8   300.0   500.0
```

```
summary(datawexN$salary[datawexN$salary<600])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   200.0   240.0   262.0   277.5   300.0   500.0
```

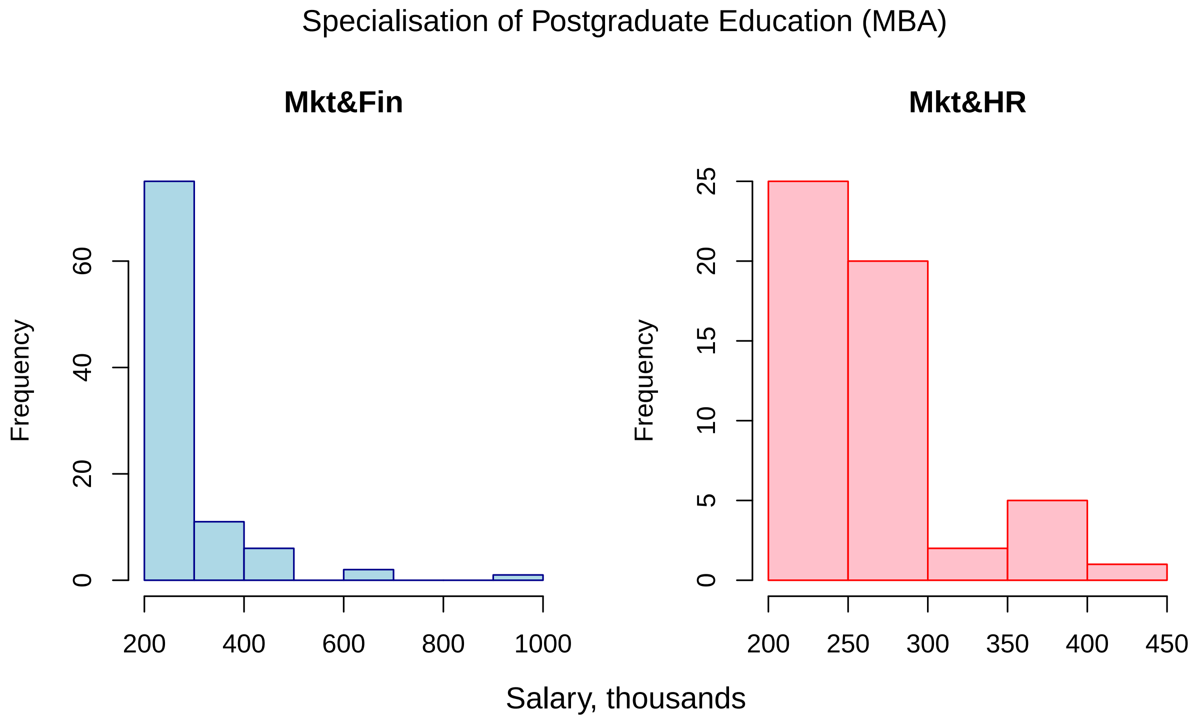
Removing the outliers shows that two distributions behave almost identically.

```
dataspeFin<-filter(df, specialisation==’Mkt&Fin’, status==’Placed’)
dataspeHR<-filter(df, specialisation==’Mkt&HR’, status==’Placed’)

g1 <- as.array(summary(dataspeFin$salary))
g2 <- as.array(summary(dataspeHR$salary))
t <- as.data.frame(list(g1, g2))
t <- t[,-3]
kbl(t, digits = 2, booktabs = T,
    caption = "Salary by Specialisation of Postgraduate Education (MBA)",
    col.names = c("Statistics", "Mkt&Fin", "Mkt&HR"), valign=’t’) %>%
    kable_styling(latex_options = c("hold_position", "striped"))
```

```
par(mfrow=c(1,2), oma = c(0, 0, 2, 0))
x <- hist(dataspeFin$salary, main=’Mkt&Fin’, xlab = NULL,
          col = "lightblue", border = "darkblue")
y <- hist(dataspeHR$salary, main=’Mkt&HR’, xlab = NULL, col = "pink",
          border = "red")
mtext("Specialisation of Postgraduate Education (MBA)",outer = TRUE, cex = 1.2)
mtext("Salary, thousands", cex = 1.2, adj = -1.3, side=1, padj=4)
```

Table 10: Salary by Specialisation of Postgraduate Education (MBA)

Statistics	Mkt&Fin	Mkt&HR
Min.	200.00	200.00
1st Qu.	240.00	240.00
Median	270.00	255.00
Mean	298.85	270.38
3rd Qu.	300.00	300.00
Max.	940.00	450.00



Specialisation of Postgraduate Education (MBA)

```
m<-length(dataspeFin$salary[dataspeFin$salary>300 & dataspeFin$salary<600])
f<-length(dataspeHR$salary[dataspeHR$salary>300 & dataspeHR$salary<600])
paste(’percentage of outliers lower than 600000 for Mkt&Fin:’,
      m/length(dataspeFin$salary))
```

```
## [1] "percentage of outliers lower than 600000 for Mkt&Fin: 0.178947368421053"
```

```
paste(’percentage of outliers lower than 600000 for Mkt&HR:’,
      f/length(dataspeHR$salary))
```

```
## [1] "percentage of outliers lower than 600000 for Mkt&HR: 0.150943396226415"
```

```
paste('number of outliers higher than 600000 for Mkt&Fin:',
      length(dataspeFin$salary[dataspeFin$salary>=600]))
```

```
## [1] "number of outliers higher than 600000 for Mkt&Fin: 3"
```

```
paste('number of outliers higher than 600000 for Mkt&HR:',
      length(dataspeHR$salary[dataspeHR$salary>=600]))
```

```
## [1] "number of outliers higher than 600000 for Mkt&HR: 0"
```

```
summary(dataspeFin$salary[dataspeFin$salary<600])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   200.0   240.0   266.5   283.8   300.0   500.0
```

```
summary(dataspeHR$salary[dataspeHR$salary<600])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   200.0   240.0   255.0   270.4   300.0   450.0
```

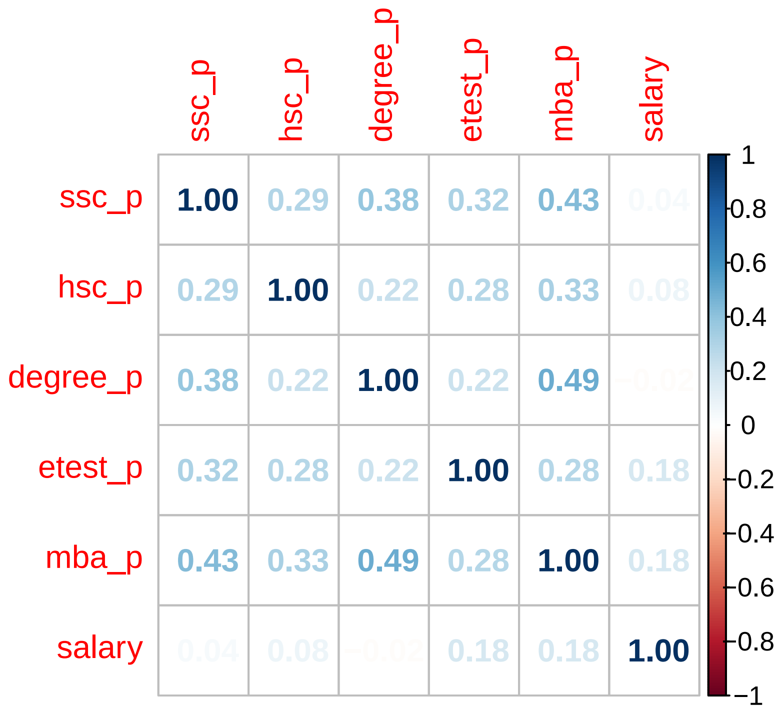
Also here removing the outliers shows that two distributions behave almost identically.

As a result of the exploratory analysis on categorical variables, there are some differences between each category but is not possible to get obvious insights on how the salary is distributed according to those variables.

Correlation matrix

Now we want to see how salary is affected by the numerical variables. Firstly, we plot the correlation matrix to see if there is any relationship between variables.

```
corrplot(cor(datap), method = 'number')
```



Since the correlation matrix does not show high correlation between salary and any other numerical variable, we decide to plot a 3D graph with the most correlated variables: mba_p and etest_p; to see if the salary is affected by the interaction of the 2 affects the salary.

```
# Interactive graph that is displayed in the html-version of this report
fig <- plot_ly(x = df$mba_p, y = df$etest_p, z = df$salary, color=df$specialisation,
               colors = c('#BF382A', '#0C4B8E'))
fig <- fig %>% add_markers()
fig <- fig %>% layout(scene = list(xaxis = list(title = 'mba_p'),
                yaxis = list(title = 'etest_p'),
                zaxis = list(title = 'salary')))
fig
```

Hypotheses

After exploring the data we can draw following hypotheses about placement:

1) The placement can be positively affected by the score percentage of school education (both secondary and higher), scores of degree education and postgraduate education (MBA): the higher is percentage, the most probable it is to get placed.
2) The placement can be affected by work experience: the chance to be placed is higher among students who already have work experience.
3) The placement can be affected by the specialization of postgraduate education (MBA): the chance to be placed is higher among students with Marketing&Finance specialisation than with Marketing&HR.

4) The placement can be affected by the field of degree education: the chance to be placed is higher among students with Commerce&Management degree than with Science&Tehnology and others.
5) Gender can also have impact on the placement: for men it is more probable to be placed than for women.

About salary:

1) We can expect the same behavior of score percentage of education on all levels from school to MBA. It should positively affect salary.
2) Salary among men is higher than among women.
3) Salary can vary between different fields of degree education. It can be higher among students with Commerce&Management degree.
4) High school specialization and MBA specialization can influence the salary.

Placement prediction

Firstly, we need to prepare the data for modeling by converting categorical variables into factors and creating a subset with all variables that can be used for placement prediction.

```
# Creating a subset with numerical variables except salary
df.new = numerical[-6]

# Converting categorical variables into factors and adding to the new dataset
df.new$gender = factor(df$gender)
df.new$ssc_b = factor(df$ssc_b)
df.new$hsc_b = factor(df$hsc_b)
df.new$hsc_s = factor(df$hsc_s)
df.new$degree_t = factor(df$degree_t)
df.new$workex = factor(df$workex)
df.new$specialisation = factor(df$specialisation)
df.new$status = factor(df$status)
```

For the prediction of placement we use binary logistic regression because the target variable has only two levels: placed and not placed. For the first attempt we are using all variables of the dataset in order to explore if any of predictors show significant importance for placement modeling.

It is crucial to mention that in this model salary is not used because of the main restriction of our dataset. Since for non-placed students there is not any information about salary (missing data), and we replaced it with 0, we are not allowed to use it for prediction, otherwise salary explains all the placement variation. For that reason we refuse to use salary as a predictor for placement.

Moreover, we are using categorical predictors, and they are included in the model as dummy variables. That means that we need to describe the control group according to which the comparisons and the interpretation of the model coefficients will be provided. Our control group consists of female students from schools related to Central Board of Education (both secondary and higher), specialisation at school - Arts, field of degree education - Commerce and Management, without work experience and specialisation of MBA - Marketing and Finance.

The First model includes all variables except salary.

```
model1 <- glm(status ~. ,family = binomial(link=’logit’), data = df.new)

summary1 <- broom::tidy(model1)
kbl(summary1, digits = 3, booktabs = T,
    caption = "Full binary logistic regression") %>%
    kable_styling(latex_options = c("hold_position", "striped"))
```

Table 11: Full binary logistic regression

term	estimate	std.error	statistic	p.value
(Intercept)	-18.372	5.324	-3.451	0.001
‘Sec.Educ.%‘	0.229	0.047	4.889	0.000
‘Higher Sec.Educ.%‘	0.107	0.038	2.838	0.005
‘Degree %‘	0.186	0.056	3.343	0.001
‘Employab. test %‘	-0.014	0.023	-0.625	0.532
‘MBA %‘	-0.214	0.059	-3.659	0.000
genderM	1.194	0.686	1.741	0.082
ssc_bOthers	0.228	0.717	0.318	0.751
hsc_bOthers	0.331	0.735	0.450	0.653
hsc_sCommerce	-1.498	1.361	-1.100	0.271
hsc_sScience	-0.911	1.457	-0.625	0.532
degree_tOthers	-1.118	1.548	-0.722	0.470
degree_tSci&Tech	-1.726	0.793	-2.177	0.029
workexYes	2.084	0.708	2.942	0.003
specialisationMkt&HR	-0.264	0.556	-0.474	0.635

From the table above we can see that six variables have a significant influence on placement on the 95% confidence level: secondary education percentage, higher education percentage, degree percentage, postgraduate MBA percentage, field of degree (only Science&Technology) and work experience. On the 90% confidence level gender of students also demonstrates statistically significant results. All other variables are not statistically significant for the prediction of placement.

Lattice analysis

For that reason as next step we can conduct the lattice analysis in order to see how quality of the model changes if we remove all non-significant variables from our model with only 7 remaining.

```
fit1 <- regsubsets(status~., data=df.new,
                   nvmax=14, method="backward")
reg_summary <- summary(fit1)

kbl(t(reg_summary$outmat), booktabs = T,
    caption = "Models with icluded variables") %>%
    kable_styling(latex_options = c("scale_down", "hold_position", "striped"))
```

```
kbl(t(reg_summary$rsq), digits = 3, booktabs = T,
    caption = "R-squared for each model") %>%
    kable_styling(latex_options = c("hold_position", "striped"))
```

Table 12: Models with icluded variables

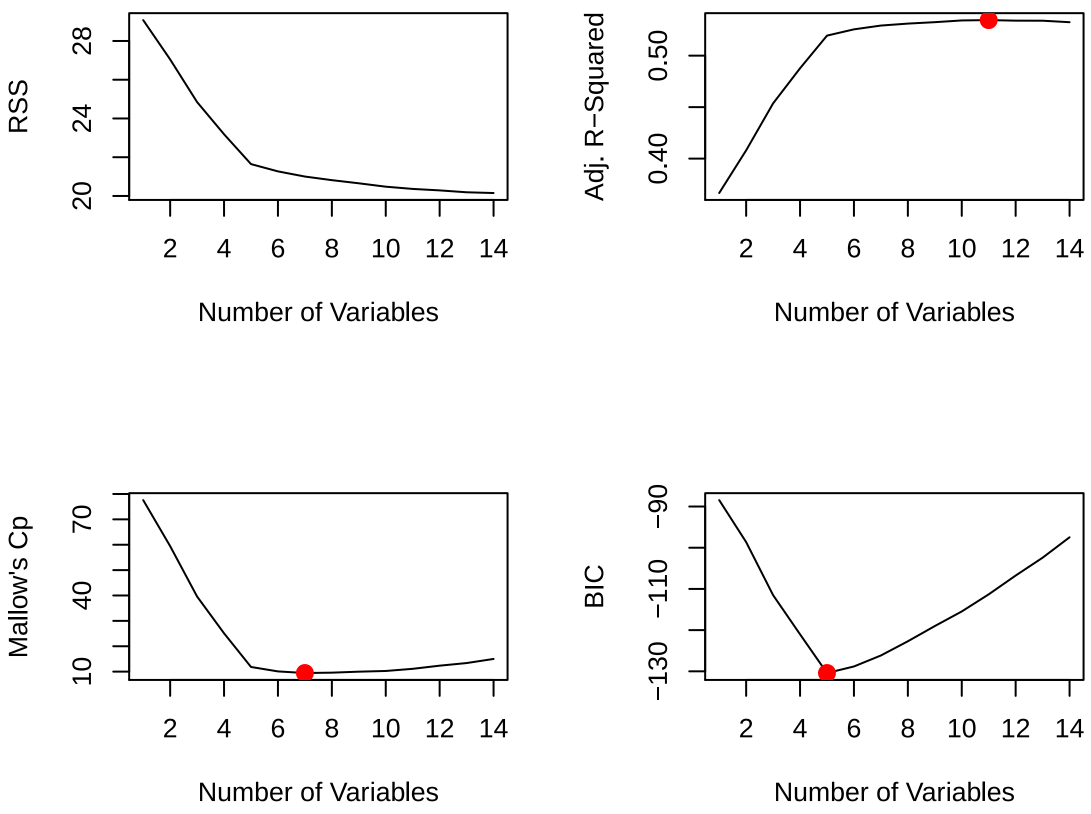
	1 (1)	2 (1)	3 (1)	4 (1)	5 (1)	6 (1)	7 (1)	8 (1)	9 (1)	10 (1)	11 (1)	12 (1)	13 (1)	14 (1)
‘Sec.Educ.%‘	*	*	*	*	*	*	*	*	*	*	*	*	*	*
‘Higher Sec.Educ.%‘		*	*	*	*	*	*	*	*	*	*	*	*	*
‘Degree %‘				*	*	*	*	*	*	*	*	*	*	*
‘Employab. test %‘								*	*	*	*	*	*	*
‘MBA %‘			*	*	*	*	*	*	*	*	*	*	*	*
genderM							*	*	*	*	*	*	*	*
ssc_bOthers												*	*	*
hsc_bOthers													*	*
hsc_sCommerce									*	*	*	*	*	*
hsc_sScience														*
degree_tOthers											*	*	*	*
degree_tSci&Tech						*	*	*	*	*	*	*	*	*
workexYes					*	*	*	*	*	*	*	*	*	*
specialisationMkt&HR										*	*	*	*	*

Table 13: R-squared for each model

0.37	0.414	0.461	0.497	0.531	0.539	0.545	0.549	0.552	0.556	0.558	0.56	0.562	0.563

As can be seen from the tables above the R-squared does not much vary between models with 7 and 14 variables (0.545 and 0.563, respectively). Therefore we can suppose that quality of the model does not increase significantly. Next, we should explore different parameters for assessment of the models - Residual Sum of Squares (RSS), adjusted R-squared, Mallow’s Cp, Bayesian information criterion (BIC).

```r
# residual sum of squares
par(mfrow=c(2,2))
plot(reg_summary$rss, xlab="Number of Variables",ylab="RSS",type="l")

# adjusted-R^2 with its largest value
rmax <- which.max(reg_summary$adjr2)
plot(reg_summary$adjr2,xlab="Number of Variables",ylab="Adj. R-Squared",type="l")
points(rmax,reg_summary$adjr2[rmax], col="red", cex=2, pch=20)

# Mallow’s Cp with its smallest value
cmin <- which.min(reg_summary$cp)
plot(reg_summary$cp,xlab="Number of Variables",ylab="Mallow’s Cp",type=’l’)
points(cmin,reg_summary$cp[cmin],col="red",cex=2,pch=20)

# BIC with its smallest value
bmin <- which.min(reg_summary$bic)
plot(reg_summary$bic,xlab="Number of Variables",ylab="BIC",type=’l’)
points(bmin,reg_summary$bic[bmin],col="red",cex=2,pch=20)
```



According to different criteria the best models are with 5, 7 and 11 variables. The maximum value of Adjusted R-Squared is achieved by model with 11 variables. But as we saw earlier, not all of predictors are statistically significant. The minimum of Cp and BIC is reached by models with 7 and 5 variables that is close to the result we obtained in the model. Relying on the results of this analysis we decided to choose model with 7 variables among others since all three models show almost equal performance, also in terms of R-Squared - it does not change a lot (0.53, 0.55 and 0.56). However, the comparison of the models with different number of variables is not correct because R-squared will obviously increase by adding new variables. Nevertheless we use it just to demonstrate that even increasing number of variables does not help crucially to enhance the quality of the model.

The Second model includes only 7 significant variables according to the previous model and lattice analysis.

```
df_sign <- df.new[, c("status","Sec.Educ.%" ,"Higher Sec.Educ.%", "Degree %",
                      "MBA %", "gender", "degree_t", "workex")]

model2 <- glm(status ~. ,family = binomial(link=’logit’), data = df_sign)

summary2 <- broom::tidy(model2)
kbl(summary2, digits = 3, booktabs = T,
    caption = "Binary logistic regression with 7 variables") %>%
    kable_styling(latex_options = c("hold_position", "striped"))
```

Table 14: Binary logistic regression with 7 variables

term	estimate	std.error	statistic	p.value
(Intercept)	-20.892	5.014	-4.166	0.000
‘Sec.Educ.%‘	0.222	0.042	5.325	0.000
‘Higher Sec.Educ.%‘	0.100	0.034	2.922	0.003
‘Degree %‘	0.187	0.054	3.480	0.001
‘MBA %‘	-0.196	0.052	-3.755	0.000
genderM	1.373	0.629	2.183	0.029
degree_tOthers	-0.477	1.161	-0.411	0.681
degree_tSci&Tech	-1.426	0.611	-2.333	0.020
workexYes	2.290	0.690	3.318	0.001

Now all variables are significant on the 95% confidence level except the degree “Others”, but we cannot remove it, because it is one of levels of dummy variable field of degree, so it remained in the model. That means that in terms of placement differences can be found only among students with Science&Technology degree and Commerce&Management degree, and not among students with Other degree and Commerce&Management degree.

Interpretation of the relationships between predictors and the target variable:

1) Secondary School Education, Higher School Education and Degree percentage are positively related to the probability of the placement. The higher is percentage the most probable is that a student will find a job.
2) Postgraduate percentage (MBA) is negatively related to the probability of placement. It is partially surprising, but explainable. During the postgraduate study the majority of students usually pay more attention to the work, and not study. If a student has a higher percentage, it probably means that this student concentrated more on study. In this case the better students study in postgraduate degree, the less attention they pay on work skills, the less probable they will be placed after study. Moreover, students who study better in postgraduate, may focus on the academic track of their career and not on the employment.
3) Probability to be placed for men is higher than for women in the control group (with Commerce&Management degree and without work experience)
4) Probability to be placed for students with Science&Technology degree is lower than for female students with Commerce&Management degree and without work experience. It can probably be explained by that it is more easy to find the job in the field related to Commerce&Management because there are more job opportunities (more places). The Science&Technology field requires precise skills and deep knowledge, and the competition can be higher because there are less places in such organisations. The field of Management has broader range of job opportunities.
5) Probability to be placed increases if a student has work experience.

In order to understand the results better we need to convert coefficients into the exponential form.

```
exp <- round(exp(model2$coefficients), digits = 3)

kbl(exp, booktabs = T,
    caption = "Exponential of coefficients", col.names = "Value") %>%
    kable_styling(latex_options = c("hold_position", "striped"))
```

Table 15: Exponential of coefficients

	Value
(Intercept)	0.000
‘Sec.Educ.%‘	1.249
‘Higher Sec.Educ.%‘	1.105
‘Degree %‘	1.206
‘MBA %‘	0.822
genderM	3.946
degree_tOthers	0.621
degree_tSci&Tech	0.240
workexYes	9.878

From the table above it can be seen, the probability of the placement will change if we increase each numerical variable by one value, and how it varies among different groups of students. For instance, having work experience increases the chance of the placement by 9.88 percent. Probability to be placed for men is almost 4 percent higher than for women.

Comparison with the null model

As next we compare the second model with the null model only with intercept to explore if the second model is statistically better and meaningful.

```
anova <- anova(model2, test="Chisq")
kbl(anova, booktabs = T,
    caption = "Comparison with the null model") %>%
    kable_styling(latex_options = c("hold_position", "striped"))
```

Table 16: Comparison with the null model

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
NULL	NA	NA	214	266.7707	NA
‘Sec.Educ.%‘	1	96.496703	213	170.2740	0.0000000
‘Higher Sec.Educ.%‘	1	15.953665	212	154.3203	0.0000649
‘Degree %‘	1	5.162563	211	149.1578	0.0230788
‘MBA %‘	1	20.748975	210	128.4088	0.0000052
gender	1	6.798158	209	121.6106	0.0091252
degree_t	2	4.672121	207	116.9385	0.0967079
workex	1	13.956736	206	102.9818	0.0001871

The results show that the model with 7 variables performs much better than the null model only with intercept. Hence, we can reject the null model. The table above illustrates the model’s behavior by adding each variable. With adding new variables the residual deviance decreases. The significant reduction in residual deviance happens when only the first variable is added. After that the gap between residual deviance gets smaller with each additional variable in the model.

Model prediction

To perform prediction analysis, firstly we split the data into train (80%) and test (20%) sets. The target variable is also recoded into numerical with values 0 and 1.

```r
# Recoding the target variable
df_sign$status <- Recode(var = df_sign$status,
                         recodes = "'Placed'=1;'Not Placed'=0; else=NA",
                         as.factor = FALSE)
# Splitting data
train <- df_sign[1:170,]
test <- df_sign[170:215,]
```

```r
# Training procedure
modelTrain <- glm(status ~. ,family=binomial(link='logit'), data=train)
fitted.results <- predict(modelTrain, newdata = test, type='response')
fitted.results <- ifelse(fitted.results > 0.5,1,0)
misClasificError <- mean(fitted.results != test$status)

kbl(round(1-misClasificError, 2), booktabs = T,
    caption = "Performance measure", col.names="Accuracy") %>%
    kable_styling(latex_options = c("hold_position", "striped"))
```

Table 17: Performance measure

Accuracy
0.91

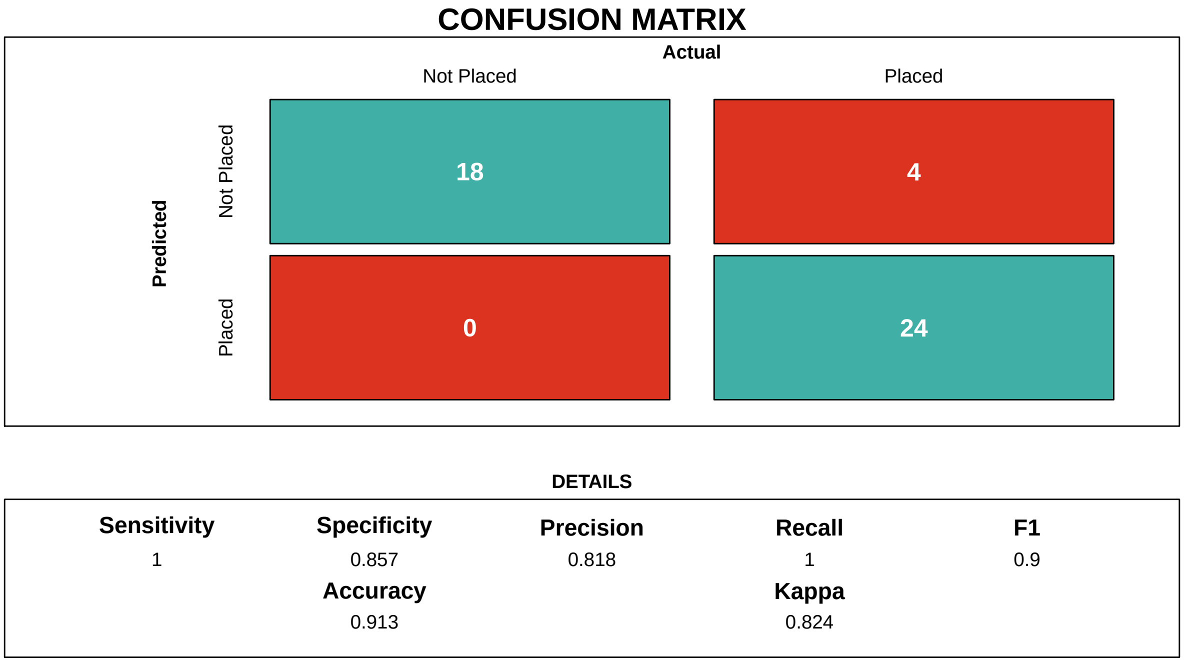
We use accuracy as the performance measure. The model reached the 91% of accuracy in the test data, what indicates that the model is well fitted, and we can use it for prediction of placement.

However the only accuracy is not a reliable score to measure performance because it does not take into account false-negative errors. In the code below, we draw the confusion matrix to see how well model predicts separate classes.

Confusion matrix

```r
# Function for drawing confusion matrix
draw_confusion_matrix <- function(cm) {
  layout(matrix(c(1,1,2)))
  par(mar=c(2,2,2,2))
  plot(c(100, 345), c(300, 450), type = "n", xlab="", ylab="",
       xaxt='n', yaxt='n')
  title('CONFUSION MATRIX', cex.main=2)
  # create the matrix
  rect(150, 430, 240, 370, col='#40B0A6')
  text(195, 440, 'Not Placed', cex=1.2)
  rect(250, 430, 340, 370, col='#DC3220')
  text(295, 440, 'Placed', cex=1.2)
  text(125, 370, 'Predicted', cex=1.3, srt=90, font=2)
  text(245, 450, ’Actual’, cex=1.3, font=2)
  rect(150, 305, 240, 365, col=’#DC3220’)
  rect(250, 305, 340, 365, col=’#40B0A6’)
  text(140, 400, ’Not Placed’, cex=1.2, srt=90)
  text(140, 335, ’Placed’, cex=1.2, srt=90)

  # add in the cm results
  res <- as.numeric(cm$table)
  text(195, 400, res[1], cex=1.6, font=2, col=’white’)
  text(195, 335, res[2], cex=1.6, font=2, col=’white’)
  text(295, 400, res[3], cex=1.6, font=2, col=’white’)
  text(295, 335, res[4], cex=1.6, font=2, col=’white’)
  # add in the specifics
  plot(c(100, 0), c(100, 0), type = "n", xlab="", ylab="", main = "DETAILS",
       xaxt=’n’, yaxt=’n’)
  text(10, 85, names(cm$byClass[1]), cex=1.5, font=2)
  text(10, 63, round(as.numeric(cm$byClass[1]), 3), cex=1.2)
  text(30, 85, names(cm$byClass[2]), cex=1.5, font=2)
  text(30, 63, round(as.numeric(cm$byClass[2]), 3), cex=1.2)
  text(50, 85, names(cm$byClass[5]), cex=1.5, font=2)
  text(50, 63, round(as.numeric(cm$byClass[5]), 3), cex=1.2)
  text(70, 85, names(cm$byClass[6]), cex=1.5, font=2)
  text(70, 63, round(as.numeric(cm$byClass[6]), 3), cex=1.2)
  text(90, 85, names(cm$byClass[7]), cex=1.5, font=2)
  text(90, 63, round(as.numeric(cm$byClass[7]), 3), cex=1.2)
  # add in the accuracy information
  text(30, 40, names(cm$overall[1]), cex=1.5, font=2)
  text(30, 20, round(as.numeric(cm$overall[1]), 3), cex=1.2)
  text(70, 40, names(cm$overall[2]), cex=1.5, font=2)
  text(70, 20, round(as.numeric(cm$overall[2]), 3), cex=1.2) }
```

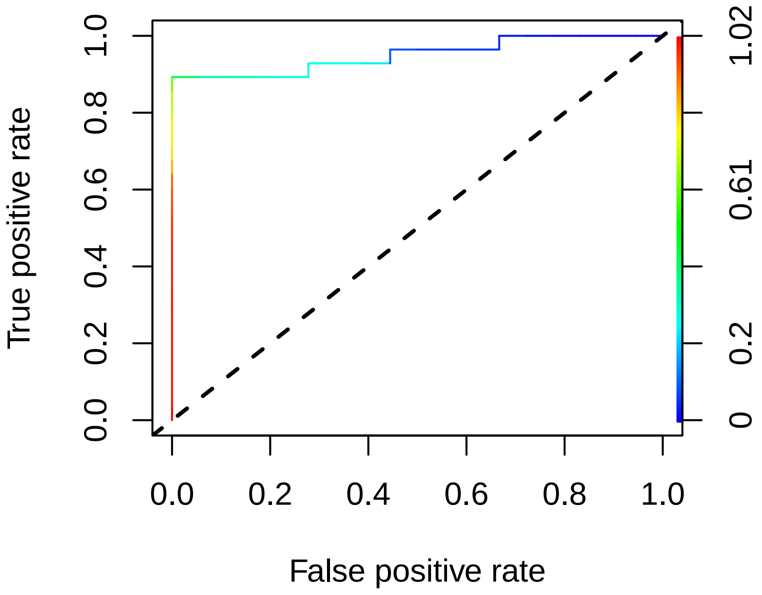
```
test$status = factor(test$status)
fitted.results = factor(fitted.results)
cm <- confusionMatrix(fitted.results, test$status)
draw_confusion_matrix(cm)
```



This plot is verifying that the built model performs very well. There is not any significant difference between two types of error in the model.

AUC graph

```
p <- predict(modelTrain,newdata = test, type='response')
pr <- prediction(p, test$status)
prf <- performance(pr, measure = "tpr", x.measure = "fpr")
plot(prf,colorize=TRUE)
abline(0,1,lwd = 2, lty = 2)
```



```r
auc <- performance(pr, measure = "auc")
auc <- auc@y.values[[1]]
round(auc, 2)
```

```
## [1] 0.95
```

The Area Under the Curve (AUC) is around 95% which shows that the model is good at identifying positive values. Overall, we can conclude that the current model with 7 variables performs very well, it is not overfitting and can be used for further research to examine if it works also for different dataset.

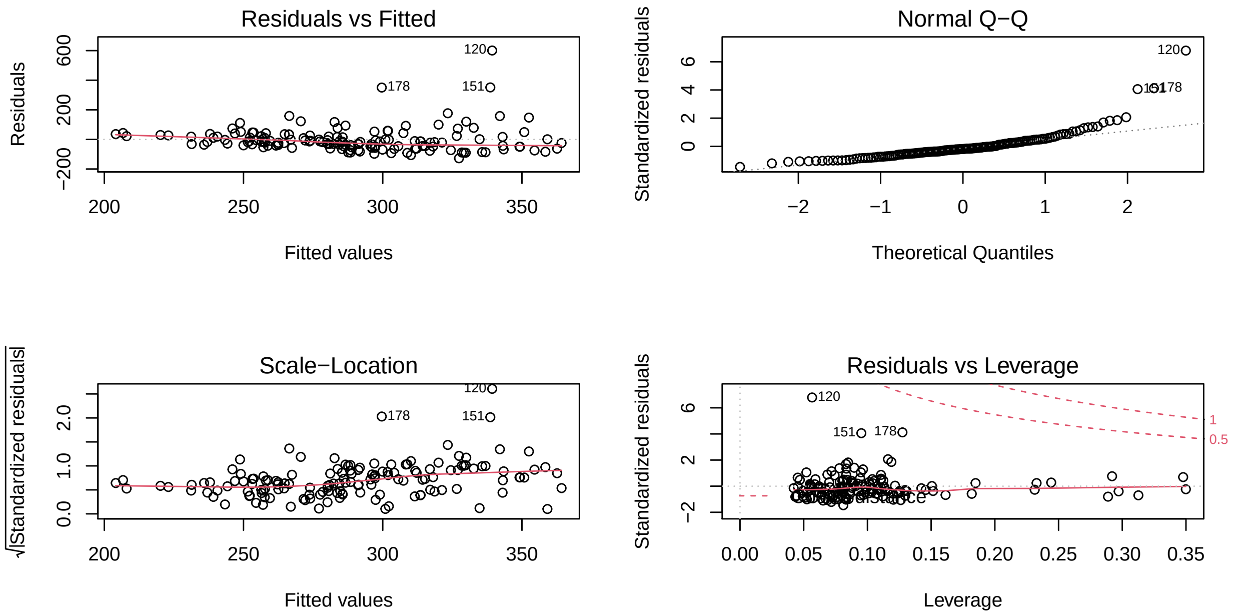
Salary prediction

Before we start modeling salary with linear regression first thing that should be done is checking for assumptions. The diagnostics will be provided to examine if we can apply this method on this data.

Checking for assumptions (diagnostics)

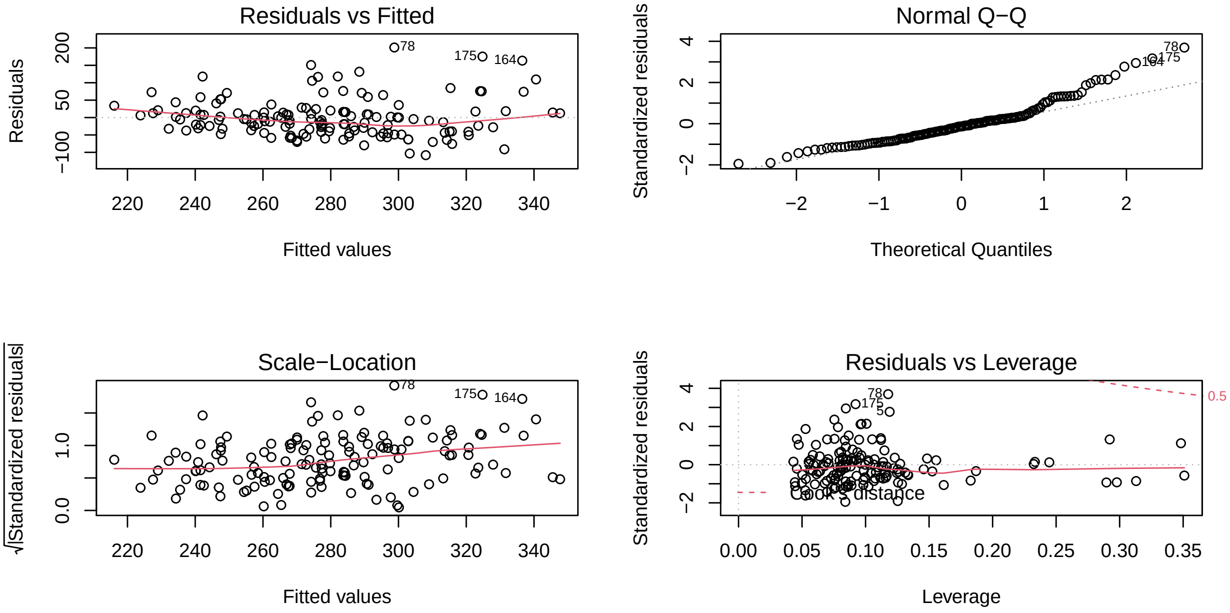
As a first step we should prepare data for linear regression. As it was mentioned above, variable “Status” will not be used in the analysis since non-placed students have 0 salary. And for that reason we cannot use status as a predictor for salary. Moreover, we should remove 0 values of salary from the analysis because otherwise the model will be biased towards 0 values, and predicting 0 salary seems to be meaningless. For that reason for predicting salary we are subsetting only students who are placed (with non-zero salary). It is another limitation of our analysis that we do not include non-placed students.

```
# Creating a subset with all variables except status
df.lin <- df.new
df.lin$salary <- df$salary
df.lin <- df.lin[-13]

# Removing 0 values of salary
df.lin <- df.lin[df.lin$salary>0,]

# Building full model with all variables for diagnostics
fit1 <- lm(salary~., data=df.lin)

# Diagnostics plots
par(mfrow=c(2,2))
plot(fit1)
```



Linearity. As can be seen on graphs above there is no strong evidence of non-linearity. The line on the residuals plot seems to be almost straight, so we can suggest that the assumption of linearity is satisfied. Normality of standardized residuals. We can see on the graph that points are close to the line, but on the top there is a little curve that is probably caused by outliers. Outliers. On the graphs some points are highlighted that means these cases can be outliers that we can remove and see if it will improve diagnostics graphs.

```
# Removing outliers
df.out <- df.lin[df.lin$salary<650,]
fit2 <- lm(salary~., data=df.out)
par(mfrow=c(2,2))
plot(fit2)
```



Removing outliers does not improve diagnostics plots and makes them even worse. It changes also the QQ-plot, and some cases are still highlighted as outliers, so probably we should return to the first version of dataset.

Heteroscedasticity. We conduct Breusch-Pagan test to identify if the variance remains the same. The null hypothesis of this test is that data are homoscedastic. The alternative hypothesis - heteroscedastic.

P-value of the test is much greater than 0.05, so there is no strong evidence to decline the null hypothesis, thus the data are homoscedastic.

```r
# Test for heteroscedasticity
het <- lmtest::bptest(salary~., data=df.lin)
het <- broom::tidy(het)
kbl(het, booktabs = T, digits = 2,
    caption = "Studentized Breusch-Pagan test") %>%
    kable_styling(latex_options = c("hold_position", "striped"))
```

Table 18: Studentized Breusch-Pagan test

statistic	p.value	parameter	method
11.15	0.67	14	studentized Breusch-Pagan test

Multicollinearity. The last assumption we should check consists of examining if the predictors are correlated with each other, because multicollinearity can raise several problems while modeling. For that reason we explore variance inflation factor (VIF) - how much variance of regression coefficients is inflated due to multicollinearity.

```r
# Test for multicollinearity
mult <- car::vif(fit1)
kbl(mult, booktabs = T, digits = 2,
    caption = "VIF test for multicollinearity") %>%
    kable_styling(latex_options = c("hold_position", "striped"))
```

Table 19: VIF test for multicollinearity

	GVIF	Df	GVIF^(1/(2*Df))
‘Sec.Educ.%‘	1.64	1	1.28
‘Higher Sec.Educ.%‘	1.57	1	1.25
‘Degree %‘	1.52	1	1.23
‘Employab. test %‘	1.37	1	1.17
‘MBA %‘	1.78	1	1.33
gender	1.26	1	1.12
ssc_b	2.08	1	1.44
hsc_b	1.90	1	1.38
hsc_s	3.31	2	1.35
degree_t	2.87	2	1.30
workex	1.19	1	1.09
specialisation	1.20	1	1.10

As graph shows VIF values of variables do not exceed value of 5. Most of values are between 1 and 2, and the biggest value is only 3.3, so the model does not have such amount of multicollinearity that can be problematic for prediction.

Overall, there is not strong evidence against using linear regression for this data since all assumptions are more or less satisfied (according to diagnostics plots and tests).

Lattice analysis

Following the model diagnostics we can try to provide variable selection with the help of lattice analysis.

```
fit1 <- regsubsets(salary~., data=df.lin,
                   nvmax=14, method="backward")
reg_summary <- summary(fit1)

kbl(t(reg_summary$outmat), booktabs = T,
    caption = "Models with icluded variables") %>%
    kable_styling(latex_options = c("scale_down", "hold_position", "striped"))
```

```
kbl(t(reg_summary$rsq), digits = 3, booktabs = T,
    caption = "R-squared for each model") %>%
    kable_styling(latex_options = c("hold_position", "striped"))
```

```
# residual sum of squares
par(mfrow=c(2,2))
plot(reg_summary$rss, xlab="Number of Variables",ylab="RSS",type="l")

# adjusted-R^2 with its largest value
rmax <- which.max(reg_summary$adjr2)
```

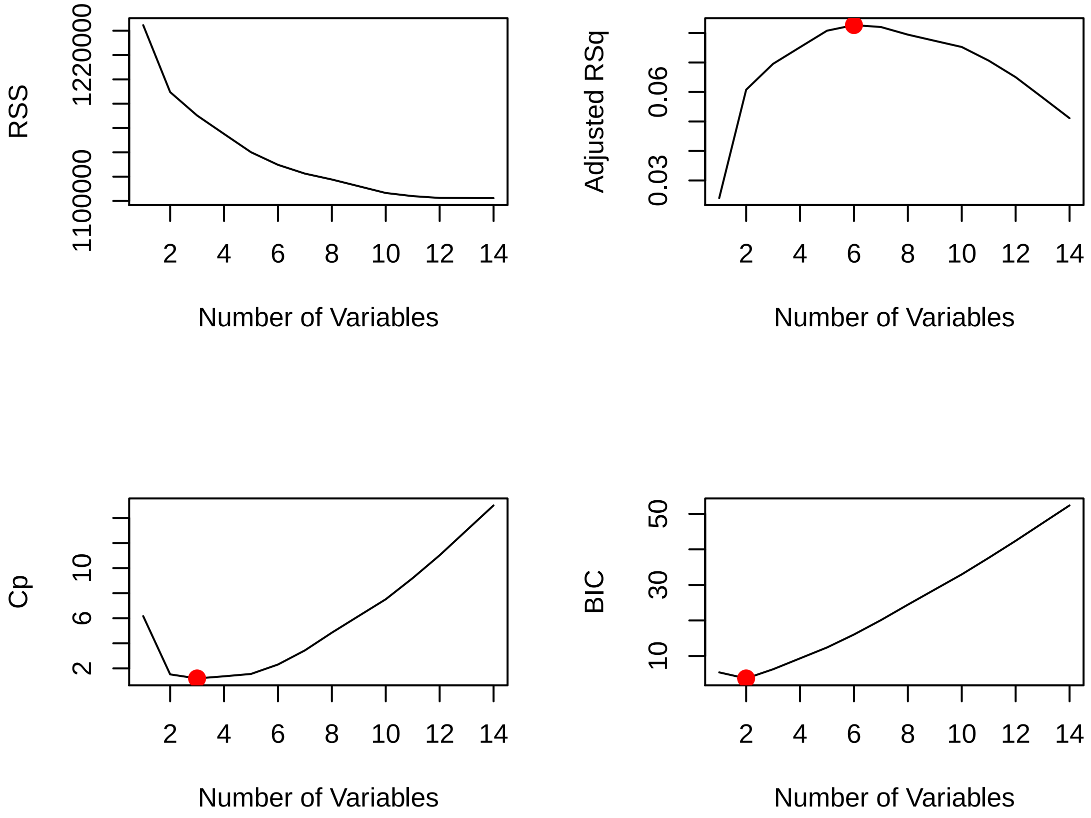
Table 20: Models with icluded variables

	1 (1)	2 (1)	3 (1)	4 (1)	5 (1)	6 (1)	7 (1)	8 (1)	9 (1)	10 (1)	11 (1)	12 (1)	13 (1)	14 (1)
‘Sec.Educ.%‘									*	*	*	*	*	*
‘Higher Sec.Educ.%‘													*	*
‘Degree %‘				*	*	*	*	*	*	*	*	*	*	*
‘Employab. test %‘					*	*	*	*	*	*	*	*	*	*
‘MBA %‘	*	*	*	*	*	*	*	*	*	*	*	*	*	*
genderM		*	*	*	*	*	*	*	*	*	*	*	*	*
ssc_bOthers														*
hsc_bOthers												*	*	*
hsc_sCommerce						*	*	*	*	*	*	*	*	*
hsc_sScience								*	*	*	*	*	*	*
degree_tOthers										*	*	*	*	*
degree_tSci&Tech			*	*	*	*	*	*	*	*	*	*	*	*
workexYes											*	*	*	*
specialisationMkt&HR							*	*	*	*	*	*	*	*

Table 21: R-squared for each model

0.031	0.074	0.089	0.1	0.112	0.12	0.126	0.13	0.134	0.138	0.14	0.141	0.141	0.141

```r
plot(reg_summary$adjr2,xlab="Number of Variables",ylab="Adjusted RSq",type="l")
points(rmax,reg_summary$adjr2[rmax], col="red", cex=2, pch=20)

# Mallow’s Cp with its smallest value
cmin <- which.min(reg_summary$cp)
plot(reg_summary$cp,xlab="Number of Variables",ylab="Cp",type=’l’)
points(cmin,reg_summary$cp[cmin],col="red",cex=2,pch=20)

# BIC with its smallest value
bmin <- which.min(reg_summary$bic)
plot(reg_summary$bic,xlab="Number of Variables",ylab="BIC",type=’l’)
points(bmin,reg_summary$bic[bmin],col="red",cex=2,pch=20)
```



As we can see on the table with R-squared even with all variables included the result is not good enough. They explain only 14% of salary variation. It means that predictors in our model do not explain the huge percentage of salary variance, and therefore, there exist other factors that have a greater impact on salary. The variables of our dataset are not suitable enough for modeling salary.

Nevertheless, we can try to build model relying on the suggestions of variable selection method. Regarding number of the variables in the model different criteria illustrate different options. The best number of variables according to Adjusted R-Squared is 6, Mallow's Cp - 3 and BIC - only 2. The last two suggestions have very low values of R-Squared. For that reason we can still experiment with 6 variables (12% of the explained variance).

Linear regression model with 6 variables based on lattice analysis

```
df_lat <- df.lin[, c("salary", "Degree %", "Employab. test %",
                      "MBA %", "gender", "degree_t", "hsc_s")]
model1 <- lm(salary~., data = df_lat)
summary1 <- broom::tidy(model1)
kbl(summary1, digits = 3, booktabs = T,
    caption = "Linear regression with 6 variables") %>%
    kable_styling(latex_options = c("hold_position", "striped"))
```

Table 22: Linear regression with 6 variables

term	estimate	std.error	statistic	p.value
(Intercept)	67.563	106.989	0.631	0.529
‘Degree %‘	-2.087	1.378	-1.514	0.132
‘Employab. test %‘	0.718	0.584	1.229	0.221
‘MBA %‘	3.799	1.567	2.424	0.017
genderM	29.533	17.406	1.697	0.092
degree_tOthers	28.822	46.295	0.623	0.535
degree_tSci&Tech	41.657	24.013	1.735	0.085
hsc_sCommerce	52.568	41.937	1.253	0.212
hsc_sScience	32.146	43.331	0.742	0.459

As we can see on the table of coefficients only 3 variables are significant on the 90% confidence level, and this model explains 12,5% of salary variance. Percentage of postgraduate education (MBA) is positively related to the salary: increasing of MBA percentage only by 1 point raises salary by 3.8 thousands. The salary among men is 29.5 thousands higher than among women. The salary among students with Science&Technology degree is 41.7 thousands higher than among students with Commerce&Management.

Exponential regression model with 6 variables

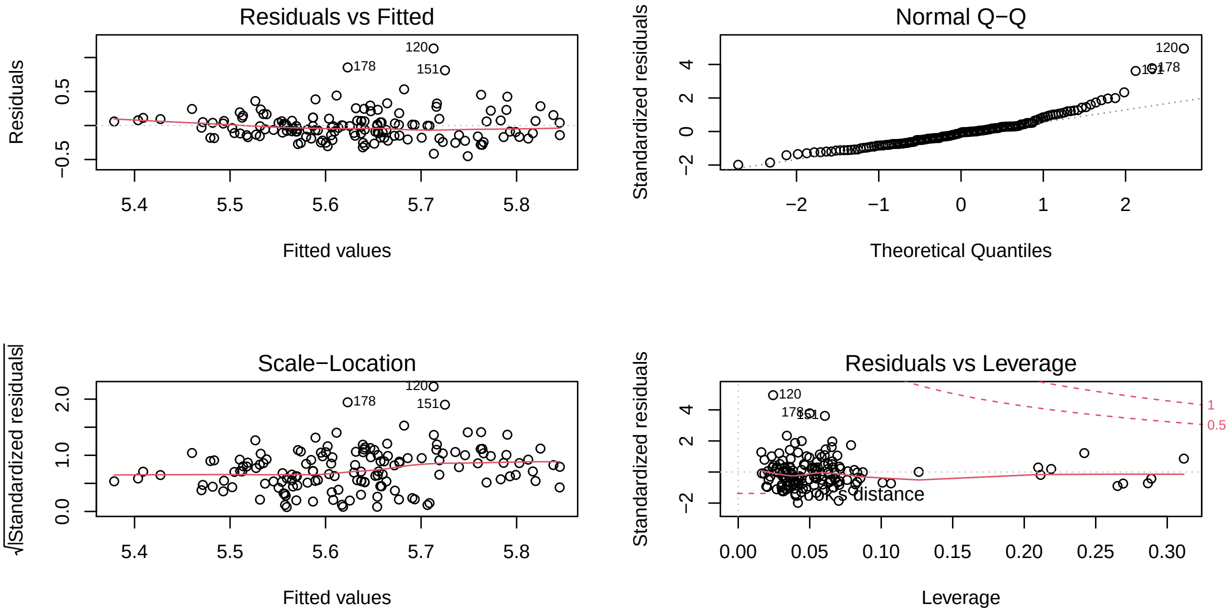
As the next step we can provide an experiment and use the exponential regression in order to see if it enhances performance by making target variable (salary) exponential in the model.

```r
model2 <- lm(log(salary)~., data = df_lat)
summary2 <- broom::tidy(model2)
kbl(summary2, digits = 3, booktabs = T,
    caption = "Exponential regression with 6 variables") %>%
    kable_styling(latex_options = c("hold_position", "striped"))
```

Table 23: Exponential regression with 6 variables

term	estimate	std.error	statistic	p.value
(Intercept)	4.932	0.276	17.876	0.000
‘Degree %‘	-0.005	0.004	-1.346	0.180
‘Employab. test %‘	0.002	0.002	1.439	0.153
‘MBA %‘	0.010	0.004	2.535	0.012
genderM	0.101	0.045	2.253	0.026
degree_tOthers	0.120	0.119	1.009	0.315
degree_tSci&Tech	0.131	0.062	2.122	0.036
hsc_sCommerce	0.151	0.108	1.392	0.166
hsc_sScience	0.088	0.112	0.791	0.430

```r
par(mfrow=c(2,2))
plot(model2)
```



The exponential function fits the data a little better than the linear model, explaining 16% of the salary variation. The same variables are significant as in the linear model. The diagnostics plots do not change much from those of linear model. Overall, we can make a conclusion that the current dataset does not perfectly fit the goal of salary prediction.

Results

In conclusion we would like to sum up results according to our predefined goals. First goal was to predict placement. For that purpose we used binary logistic regression model with 7 variables that were chosen according to the lattice analysis. We found out that Secondary School Education, Higher School Education and Degree percentage are positively related to the probability of the placement. Also having work experience increases the chance to be placed. The chance to be placed is higher for men than for women. Probability to be placed for students with Science&Technology degree is lower than for female students with Commerce&Management degree and without work experience. It is so probably because the field of Commerce&Management has broader range of job opportunities. Postgraduate percentage (MBA) is negatively related to the probability of placement that is probably because during the postgraduate study the majority of students usually pay more attention to the work skills, and not study. It results in that students who studied worse at MBA have better chances to get a job because they were more focused on it. Moreover, besides good interpretability the built model also demonstrates the good performance in terms of test accuracy. It gives 91% of accuracy and is equally good by predicting both classes - placed and not placed. Overall, we can say that it was a successful experiment with modeling placement. Almost all of our hypotheses have been confirmed.

Second goal consisted of predicting salary. In this case the results are different from those we obtained by predicting placement. The predictors in the regression model do not explain well variance of salary, and therefore, there exist other factors that have a greater impact on salary which are not included in our dataset. The full model with all variables has R-Squared equal only to 14% which is very low result. The model with fewer number of variables which were chosen according to lattice analysis explains 12,5% of salary variance. Percentage of postgraduate education (MBA) is positively related to the salary. The salary among men is higher than among women. The salary among students with Science&Technology degree is higher than among students with Commerce&Management. However, we cannot use this model for predictiion due to its not satisfiable quality. For predicting salary it will be more efficient to use different dataset with variables that can be sufficient enough for purpose of prediction.