MTHM 501 - Working with Data

Final Assessment

Which factor(s) influenced a candidate in getting placed?

Jingjie, Zhao

(700059248)

<University of Exeter>

Date: Nov.16th, 2020

Introduction

With the increasing popularity of higher education, many students have obtained a bachelor's degree or above before entering social work. This also makes the job market increasingly competitive. In order to recruit outstanding young talents, many companies choose to launch targeted campus recruitment for fresh graduates every year. Campus recruitment is a strategy that provides internships and full-time entry-level positions to attract and hire young talents. University recruitment is more suitable for large and medium-sized companies with large recruitment needs, but it can range from small positions to large-scale operations with different job requirements. So far, how to stand out in campus recruitment and how to be successfully admitted from the company of your choice before graduation has attracted public attention. Many social organizations have conducted background checks on the recruited students on campus and have reached different conclusions. The data in this report also revolves around campus recruitment. Through the discussion of a series of relationships, the question which factor(s) influenced a candidate in getting placed will be answered.

Objectives

In order to answer the final question of which factor(s) affect students' admission status in campus recruitment, the following goals will be achieved: First, I will create a neat data set. This will include the use of data wrangling to organize our raw data into a neat and effective data set. Second, use appropriate data visualization to analyze the characteristics of each column of the data. Third, establish bivariate and multivariate exploration between variables to reveal the dependencies between them. Then, fix the model based on the significant levels between each column of data and the response variable. Finally, the limitations of the optimization model should be pointed out.

Data

The data in this report comes from kaggle campus recruitment, which is the campus recruitment data of an Indian business school. The data summarizes the status of 215 students and records the information of each student after entering high school, as well as the final hiring status of the student and the salary that the company is willing to provide. In order to make the results more comprehensive, the scope of statistics includes their academic performence, degree type, board of education etc.

There are missing values in this dataset and all come from the salary column. It is determined that part of students are not employed by any company, so they have no corresponding payment. I decided to set all missing values to zero, that is, the income of these unemployed students is zero. Moreover, it was found that the names of various columns were extremely unclear. On the other hand, students' performance records are rather chaotic, which will affect the analysis of our fitted model. Therefore, it is necessary to rename each column, then created a new column and classify the results of each student.

Analysis and Results

I. Univariate Exploration

The analysis starts with the univariate exploration. Overall, there are 140 male and 75 female in the sample, each accounting for 65% and 35% of the total population. Among these students, there are 113 students belong to the Faculty of Business, 91 students belong to the Faculty of Science, and 11 students from the Faculty of Arts. In the survey of majors, 145 students are interested in business and management, and the remaining 70 students are from science and technology and other programs.

Figure 1: Univariate Exploration

The higher education percentages for all students are divided into three groups. The first group is students with scores below 60 percent, the second group is students with scores between 60 percent and 80 percent, and the last group is students with scores above 80 percent. Check the distribution of student entrance examination scores, and the results show that there is little difference in the number of students in each group. In contrast, 44% of students belong to the second group, which is slightly higher than the first group (28%) and the thrid group (27%). When investigating the students' background, one of item is specifically about whether the student have work experience before graduates. There are more students have no work experience in our sample. This is an interesting result. I look forward to use histogram to illustrate whether students' work experience will affect the results of campus recruitment in the next section. The last two sets of univariate exploration studied how many students were admitted by the company in campus recruitment and their salary distribution range. The company hired most of the students in the datasets. In fact, most companies limit the monthly payment for campus recruitment to between 200,000 and 400,000. According to the statistics, more than 63% of student salaries fall within this range.

II. Bivariate and Multivariate Exploration

In order to fully understand the interdependence between variables, it is necessary to establish a correlation diagram between variables. It is a feasible strategy to help determine which factors affect the final employment status of students through the interdependence of variables.

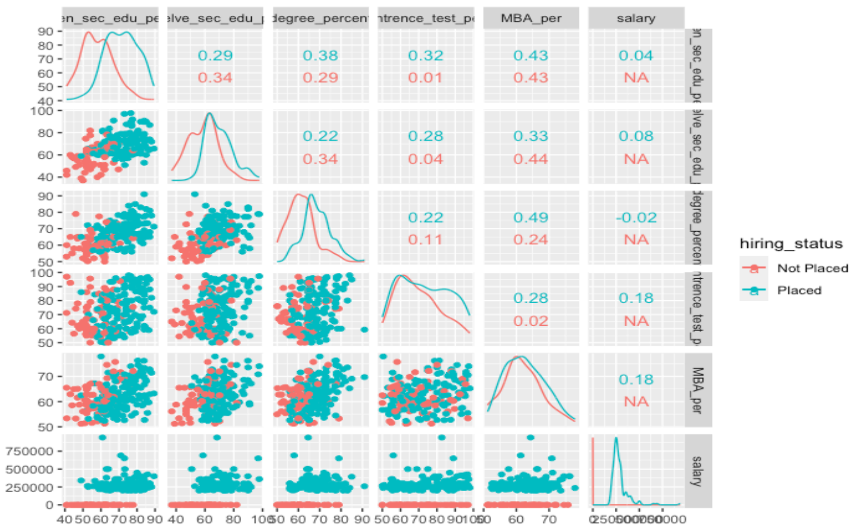


Figure 2: Correlation between variables coloured by hiring status

After establishing a rough model, it is found that all variables are positively correlated. In other words, large p-values make this sample almost no independent variables. Scatter plots show that variables are strongly correlated with each other. However, it is difficult to judge the student background represented by each point because the points are randomly distributed. As a result, Figure 2 is generated, which divides the students into two groups according to their hiring status. In general, the employed students have better performance in all aspects than the unemployed ones.

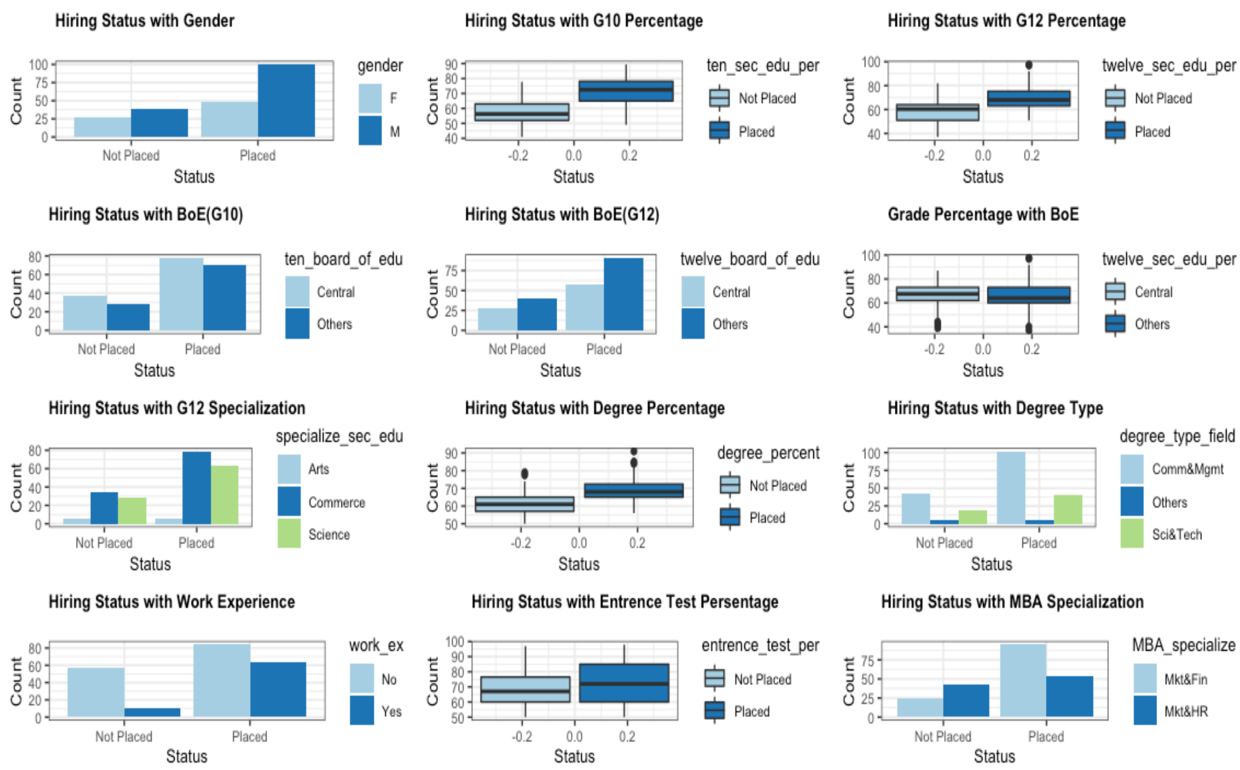


Figure 3: Bivariate Exploration

For the sake of accruacy, it is helpful to establish a binary relationship between the hiring status and each remaining variables, as shown in Figure 3. Statistics show that male acceptance rate is slightly higher than the female in campus recruitment. Students who admitted also have better performance in both tenth-grade and twelve-grade education. After delving into the situation of the student board, it is found that tenth grade students are more in the central education area, while the twelfth grade students have moved to other board areas. In comparison, the scores of students on the central board are generally higher than those on the others board. This phenomenon has caught my attention. Since the central educational environment is relatively better, why do most senior students choose to complete their final year of studies on the other board? Are they trying to avoid employment pressure after graduation or have other considerations? These details are not known from this data.

In addition, there has always been a saying that "program determines whether to be hired in the end", and research has also been carried out in this regard. The data shows that more students who graduated from the commerce have found jobs, while only a few of arts students are employed. The result may only be a reference for students of these two majors, but the field of study does limit the range of choices in the job market to some extent. It seems to be a trade-off between the field of interest of the students and the future development prospects. Problems faced by everyone in the world.

Subsequently, a detailed analysis is conducted on whether work experience can help students stand out in campus recruitment. According to the feedback, even if more students have no work experience, these students enter companies in different fields after graduates. It concludes that the focus of the headhunters is not only on the work experience, but also on the comprehensive strength of the students. This point of view is verified by following three-dimensional model. Through the relationship between G10 grade, G12 grade, degree percent and the monthly payment, it is found that students who are awarded higher salaries tend to perform well in grade 10, grade 12 and degree percentage. In the sample, there are cases where students have average grades but are awarded high salaries (i.e x = 60.8, y = 68.4, z = 64.6). A more general and convincing set of data is at the center (i.e x = 79, y = 76, z = 85). This also potentially explains the relationship between the student's overall performance and hiring status. Students with high comprehensive scores are more likely to be hired.

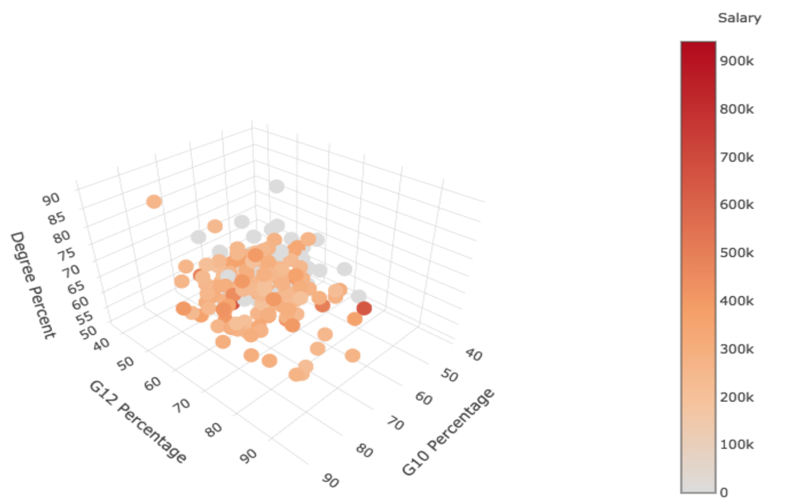


Figure 4: Multivariate Exploration with monthly payment

Model fitting

I. Hiring Status

Before presenting the final model, let's clarify what we will focus on. First of all, it is necessary to pay attention to the significant level of each variable for the response variable. Theoretically, we can remove these least important features and then reclassify our data. Generally, the least significant variables have the following characteristics: they have the highest p-value; if they are deleted, the adjusted R-squared will be reduced to the lowest value and also has least increment in residuals-sum-of-squares (RSS). The p-value represents the probability of obtaining the same result as the null hypothesis. One of the most commonly used p-value is 0.05. If the p-value is less than 0.05, the null hypothesis is considered wrong (i.e reject null hypothesis). If the p-value is greater than 0.05, the null hypothesis cannot be rejected. More than that, a small p-value indicates that we're unlikely to observe the relationship between the explanatory variable and the response variable (i.e hiring status). Therefore, we start with all features. We calculate the p-value and then eliminate the feature with the highest p-value in turn until the p-values of all features are below 0.05. Secondly, we also pay attention to the outlier. In the presence of outliers, the model selection will be affected. During the operation, choose to either remove the outlier or pay attention to its changes.

Residual plot show as in Figure 5 is the fitted model of the hiring status. The function of the residual vs. fitted model is to tell whether the residual has a linear pattern. As show in the figure, the residuals are evenly distributed on the horizontal line. In that case, the model is judged to have linear relationship. The normal-QQ plot shows whether the residuals are normally distributed. Since most of the residuals in the figure are well aligned around the dotted line, this set of data is close to a normal distribution.

Model fitting is a process of repeatedly deleting least significant features and adding it back to confirm the correctness of the deletion. In this study, after 12 time deletion processes and 10 repeated confirmations, it is found that the status of being recruited in campus recruitment have the greatest correlation with grade12 secondary education percentage, degree percent and the degree type. In particular, students with degree type in the field of science and technology are more likely to be employed, and the employment rate of students in other disciplines is not as good as that of science and technology students. In this regard, medium and large size companies do not value the practical experience of new graduates very much. They believe that students' professional performance can better reflect their own strength, and good grades also show that students have the ability to learn new skills quickly. This also reflects the company's inclusiveness.

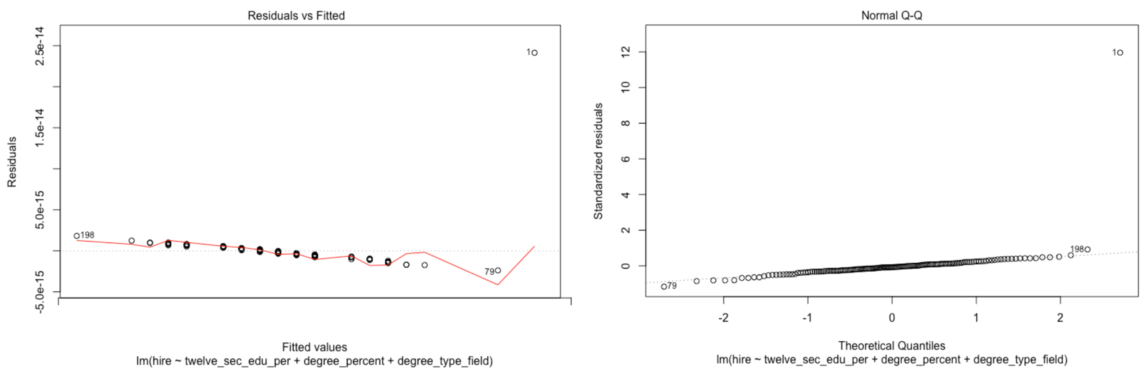


Figure 5: Model fitting with Hiring Status (model 1 vs. model 12)

II. Salary

On the basis of being hired, we also explored what kind of students will be paid a higher salary. It is observed that the salaries of new graduates are divided into two groups. Most of the data are concentrated within 400,000 and have a linear relationship show as in Figure 6. The data stacked on the right side of the residual plot are more randomly distributed and always have outliers. For these randomly distributed high-income jobs, there are no more data for more detailed research. It can also be determined from the QQ-plot that the salary is normally distributed since most of the data coincide with the dotted line.

Only employed students are selected as the research objects. Using the same method show as above, top 2 features affecting salary we identified are gender and MBA percentage. It is understood that India's economy is denominated by agriculture, handcrafts, modern industries and its supporting industries. One-fourth of the population still straggling with the food and clothe. Therefore, more powerful and intelligent workers are needed. At this point, male do have certain gender advantages. In addition, India has a large English-speaking population and is the world's most important information service industry producer, and the home country of many software engineers in the 21st century. At this time, cultivating business talents with an international background can better help the country develop technology, and our fitted model also shows that the higher the proportion of MBA students, the higher the monthly payment.

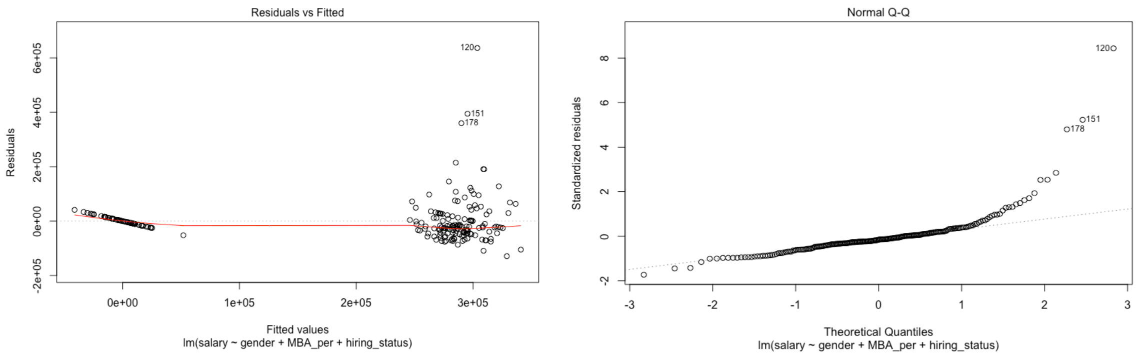


Figure 6: Model Fitting with Salary (model 1 vs. model 12)

Exploring Stationary

Since campus data is a snapshot, we do not have a time component. However, stationarity over the field the measurement is also important. Thus, we also explore the relationship between the hiring status and the degree percent, split by degree type field; and the relationship between the hiring status and the degree type field, split by degree percent, show as shown in Figure 7. Since the trend of the variable in the fitted model at different times has almost no changes, so the fitted model for hiring status has stationarity.

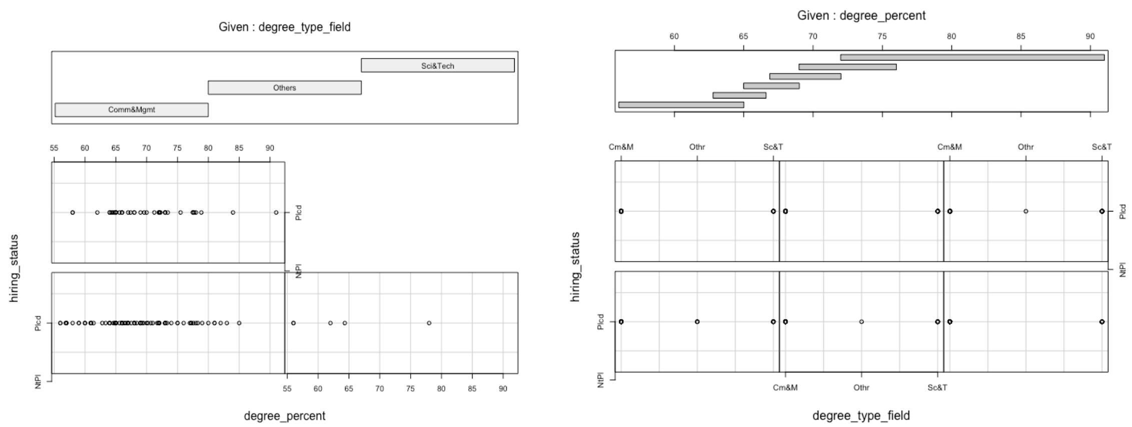


Figure 7: Stationarity (Fitted Model)

Limitations

To some extent, this data has some limitations. First of all, the data that constitutes the sample comes from an Indian school, and the collected variable data and statistical results cannot be defined as applicable to each country. After all, each country has inconsistent positioning and efforts for education and employment due to various reasons. They only mention some of these variables under the Indian social system. Since the sample size comes from one school, the correlation and pattern between variables lack universality. In general, these values should only be considered as estimated scores.

Some questions cannot be answered in this report due to data limitations. For instance, why do senior students choose to leave the central board, which has a relatively better educational environment? If we have more specific and comprehensive grasp of the data background, it may help us to improve in the future research.

Conclusion

The purpose of this project is to investigate which factor(s) influenced a candidate in getting placed in campus recruitment. To this end, data from an Indian business school was selected. The initial graphical analysis revealed some interdependencies of different strengths among the variables. During the analysis, it found that more male candidates got placed as compared to female candidates. Candidates with higher educational percentage in the grade10 and grade12 have better chance of placement. Moreover, having work experience doesn’t benefit a lot from the campus recruitment. Most of the students get hired from the field of Science and Technology, etc.

The factors that truly affect hiring status are grade12 secondary education percentage, degree percent and the degree type. In particular, students with degree type in the field of science and technology are more likely to be employed. When study the relationship between the salary and the students background, it concludes that wages are directly related to gender and MBA percentage. The results meet the national conditions of India.

Which factor(s) affect the status of students’ final employment in campus recruitment? The above report shows that the hired students tend to perform well in higher educational percentage, degree percent and the degree type.

MTHM 501: Campus Recuitrment Dataset Analysis

Jingjie Zhao

2020-11-16

Install packages

```
library(RColorBrewer)
library(ggplot2)
library(plyr)
library(dbplyr)
library(dplyr)
library(ggpubr)
library(tidyr)
library(GGally)
library(plotly)
library(gstat)
library(lattice)
library(sp)
```

Import data

```
library("rmarkdown")
setwd("/Users/jingjiezhao/Downloads")
campus_re <- read.csv("campus_recruitment.csv")
head(campus_re)
```

```
##   sl_no gender ssc_p   ssc_b hsc_p   hsc_b    hsc_s degree_p  degree_t workex
## 1     1      M 67.00  Others 91.00  Others Commerce    58.00  Sci&Tech     No
## 2     2      M 79.33 Central 78.33  Others  Science    77.48  Sci&Tech    Yes
## 3     3      M 65.00 Central 68.00 Central     Arts    64.00 Comm&Mgmt     No
## 4     4      M 56.00 Central 52.00 Central  Science    52.00  Sci&Tech     No
## 5     5      M 85.80 Central 73.60 Central Commerce    73.30 Comm&Mgmt     No
## 6     6      M 55.00  Others 49.80  Others  Science    67.25  Sci&Tech    Yes
##   etest_p specialisation mba_p     status salary
## 1    55.0         Mkt&HR 58.80     Placed 270000
## 2    86.5        Mkt&Fin 66.28     Placed 200000
## 3    75.0        Mkt&Fin 57.80     Placed 250000
## 4    66.0         Mkt&HR 59.43 Not Placed     NA
## 5    96.8        Mkt&Fin 55.50     Placed 425000
## 6    55.0        Mkt&Fin 51.58 Not Placed     NA
```

```r
str(campus_re)
```

```
## 'data.frame':    215 obs. of  15 variables:
##  $ sl_no         : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ gender        : Factor w/ 2 levels "F","M": 2 2 2 2 2 2 1 2 2 2 ...
##  $ ssc_p         : num  67 79.3 65 56 85.8 ...
##  $ ssc_b         : Factor w/ 2 levels "Central","Others": 2 1 1 1 1 2 2 1 1 1 ...
##  $ hsc_p         : num  91 78.3 68 52 73.6 ...
##  $ hsc_b         : Factor w/ 2 levels "Central","Others": 2 2 1 1 1 2 2 1 1 1 ...
##  $ hsc_s         : Factor w/ 3 levels "Arts","Commerce",..: 2 3 1 3 2 3 2 3 2 2 ...
##  $ degree_p      : num  58 77.5 64 52 73.3 ...
##  $ degree_t      : Factor w/ 3 levels "Comm&Mgmt","Others",..: 3 3 1 3 1 3 1 3 1 1 ...
##  $ workex        : Factor w/ 2 levels "No","Yes": 1 2 1 1 1 2 1 2 1 1 ...
##  $ etest_p       : num  55 86.5 75 66 96.8 ...
##  $ specialisation: Factor w/ 2 levels "Mkt&Fin","Mkt&HR": 2 1 1 2 1 1 1 1 1 1 ...
##  $ mba_p         : num  58.8 66.3 57.8 59.4 55.5 ...
##  $ status        : Factor w/ 2 levels "Not Placed","Placed": 2 2 2 1 2 1 1 2 2 1 ...
##  $ salary        : int  270000 200000 250000 NA 425000 NA NA 252000 231000 NA ...
```

```r
# Replace column names
my_column_name <- c("serial_number", "gender", "ten_sec_edu_per", "ten_board_of_edu",
                    "twelve_sec_edu_per", "twelve_board_of_edu", "specialize_sec_edu",
                    "degree_percent", "degree_type_field", "work_ex", "entrence_test_per",
                    "MBA_specialize", "MBA_per", "hiring_status", "salary")
colnames(campus_re) <- my_column_name
colnames(campus_re)
```

```
##  [1] "serial_number"      "gender"             "ten_sec_edu_per"
##  [4] "ten_board_of_edu"   "twelve_sec_edu_per" "twelve_board_of_edu"
##  [7] "specialize_sec_edu" "degree_percent"     "degree_type_field"
## [10] "work_ex"            "entrence_test_per"  "MBA_specialize"
## [13] "MBA_per"            "hiring_status"      "salary"
```

```r
# Backup
campus_re_backup <- campus_re
campus_re_backup2 <- campus_re
```

```r
# Missing values
summary(is.na(campus_re_backup))
```

```
##  serial_number     gender        ten_sec_edu_per ten_board_of_edu
##  Mode :logical   Mode :logical   Mode :logical   Mode :logical
##  FALSE:215       FALSE:215       FALSE:215       FALSE:215
##
##  twelve_sec_edu_per twelve_board_of_edu specialize_sec_edu degree_percent
##  Mode :logical      Mode :logical       Mode :logical      Mode :logical
##  FALSE:215          FALSE:215           FALSE:215          FALSE:215
##
##  degree_type_field  work_ex        entrence_test_per MBA_specialize
##  Mode :logical     Mode :logical   Mode :logical     Mode :logical
##  FALSE:215         FALSE:215       FALSE:215         FALSE:215
```

```
##
##     MBA_per          hiring_status      salary
##   Mode :logical    Mode :logical    Mode :logical
##   FALSE:215        FALSE:215        FALSE:148
##                                     TRUE :67
```

There is only salary column has missing values.Salary has null values because the students who aren't placed will not have any monthly payment. We can replace the null values with 0.

```
# Imputing Missing Values
campus_re_backup[is.na(campus_re_backup)] <- 0
summary(is.na(campus_re_backup))
```

```
##   serial_number      gender          ten_sec_edu_per ten_board_of_edu
##   Mode :logical    Mode :logical    Mode :logical    Mode :logical
##   FALSE:215        FALSE:215        FALSE:215        FALSE:215
##   twelve_sec_edu_per twelve_board_of_edu specialize_sec_edu degree_percent
##   Mode :logical      Mode :logical       Mode :logical      Mode :logical
##   FALSE:215          FALSE:215           FALSE:215          FALSE:215
##   degree_type_field  work_ex         entrence_test_per MBA_specialize
##   Mode :logical      Mode :logical   Mode :logical     Mode :logical
##   FALSE:215          FALSE:215       FALSE:215         FALSE:215
##     MBA_per          hiring_status      salary
##   Mode :logical    Mode :logical    Mode :logical
##   FALSE:215        FALSE:215        FALSE:215
```

```
campus_re_backup2[is.na(campus_re_backup2)] <- 0
summary(is.na(campus_re_backup2))
```

```
##   serial_number      gender          ten_sec_edu_per ten_board_of_edu
##   Mode :logical    Mode :logical    Mode :logical    Mode :logical
##   FALSE:215        FALSE:215        FALSE:215        FALSE:215
##   twelve_sec_edu_per twelve_board_of_edu specialize_sec_edu degree_percent
##   Mode :logical      Mode :logical       Mode :logical      Mode :logical
##   FALSE:215          FALSE:215           FALSE:215          FALSE:215
##   degree_type_field  work_ex         entrence_test_per MBA_specialize
##   Mode :logical      Mode :logical   Mode :logical     Mode :logical
##   FALSE:215          FALSE:215       FALSE:215         FALSE:215
##     MBA_per          hiring_status      salary
##   Mode :logical    Mode :logical    Mode :logical
##   FALSE:215        FALSE:215        FALSE:215
```

Data Visualization

Univariate Exploration

```
# 1. GENDER
# Check how many levels in "gender"
levels(campus_re_backup$gender)
```

```
## [1] "F" "M"
```

```r
# Count the frequency of each level in "gender"
sum(data.frame(campus_re_backup$gender) == "F")
```

```
## [1] 76
```

```r
sum(data.frame(campus_re_backup$gender) == "M")
```

```
## [1] 139
```

```r
# Percentage of each level
female_per <- (sum(data.frame(campus_re_backup$gender) == "F") /
                 length(campus_re_backup$gender)) * 100
female_per
```

```
## [1] 35.34884
```

```r
male_per <- (sum(data.frame(campus_re_backup$gender) == "M") /
               length(campus_re_backup$gender)) * 100
male_per
```

```
## [1] 64.65116
```

```r
# 2. 10th secondary Board of education
# Check how many levels in "ten_board_of_edu"
levels(campus_re_backup$ten_board_of_edu)
```

```
## [1] "Central" "Others"
```

```r
# Count the frequency of each level in "ten_board_of_edu"
sum(data.frame(campus_re_backup$ten_board_of_edu) == "Central")
```

```
## [1] 116
```

```r
sum(data.frame(campus_re_backup$ten_board_of_edu) == "Others")
```

```
## [1] 99
```

```r
# Percentage of each level
central_10_per <- (sum(data.frame(campus_re_backup$ten_board_of_edu) == "Central") /
                     length(campus_re_backup$ten_board_of_edu)) * 100
central_10_per
```

```
## [1] 53.95349
```

```r
others_10_per <- (sum(data.frame(campus_re_backup$ten_board_of_edu) == "Others") /
                    length(campus_re_backup$ten_board_of_edu)) * 100
others_10_per
```

```
## [1] 46.04651
```

```r
# 3. 12th secondary Board of education
# Check how many levels in "twelve_board_of_edu"
levels(campus_re_backup$twelve_board_of_edu)
```

```
## [1] "Central" "Others"
```

```r
# Count the frequency of each level in "twelve_board_of_edu"
sum(data.frame(campus_re_backup$twelve_board_of_edu) == "Central")
```

```
## [1] 84
```

```r
sum(data.frame(campus_re_backup$twelve_board_of_edu) == "Others")
```

```
## [1] 131
```

```r
# Percentage of each level
central_12_per <- (sum(data.frame(campus_re_backup$twelve_board_of_edu) == "Central") /
                     length(campus_re_backup$twelve_board_of_edu)) * 100
central_12_per
```

```
## [1] 39.06977
```

```r
others_12_per <- (sum(data.frame(campus_re_backup$twelve_board_of_edu) == "Others") /
                    length(campus_re_backup$twelve_board_of_edu)) * 100
others_12_per
```

```
## [1] 60.93023
```

```r
# 4. Higher Secondary Education - Specialization
# Check how many levels in "specialize_sec_edu"
levels(campus_re_backup$specialize_sec_edu)
```

```
## [1] "Arts"     "Commerce" "Science"
```

```r
# Count the frequency of each level in "specilize_sec_edu"
sum(data.frame(campus_re_backup$specialize_sec_edu) == "Arts")
```

```
## [1] 11
```

```r
sum(data.frame(campus_re_backup$specialize_sec_edu) == "Commerce")
```

```
## [1] 113
```

```r
sum(data.frame(campus_re_backup$specialize_sec_edu) == "Science")
```

```
## [1] 91
```

```r
# Percentage of each level
arts_per <- (sum(data.frame(campus_re_backup$specialize_sec_edu) == "Arts") /
               length(campus_re_backup$specialize_sec_edu)) * 100
arts_per
```

```
## [1] 5.116279
```

```r
commerce_per <- (sum(data.frame(campus_re_backup$specialize_sec_edu) == "Commerce") /
                   length(campus_re_backup$specialize_sec_edu)) * 100
commerce_per
```

```
## [1] 52.55814
```

```r
science_per <- (sum(data.frame(campus_re_backup$specialize_sec_edu) == "Science") /
                  length(campus_re_backup$specialize_sec_edu)) * 100
science_per
```

```
## [1] 42.32558
```

```r
# 5. Degree Type (Field)
# Check how many levels in "degree_type_field"
levels(campus_re_backup$degree_type_field)
```

```
## [1] "Comm&Mgmt" "Others"    "Sci&Tech"
```

```r
# Count the frequency of each level in "degree_type(field)"
sum(data.frame(campus_re_backup$degree_type_field) == "Comm&Mgmt")
```

```
## [1] 145
```

```r
sum(data.frame(campus_re_backup$degree_type_field) == "Others")
```

```
## [1] 11
```

```r
sum(data.frame(campus_re_backup$degree_type_field) == "Sci&Tech")
```

```
## [1] 59
```

```r
# Percentage of each level
c_m_per <- (sum(data.frame(campus_re_backup$degree_type_field) == "Comm&Mgmt") /
              length(campus_re_backup$degree_type_field)) * 100
c_m_per
```

```
## [1] 67.44186
```

```r
s_t_per <- (sum(data.frame(campus_re_backup$degree_type_field) == "Sci&Tech") /
              length(campus_re_backup$degree_type_field)) * 100
s_t_per
```

```
## [1] 27.44186
```

```r
others_per <- (sum(data.frame(campus_re_backup$degree_type_field) == "Others") /
                 length(campus_re_backup$degree_type_field)) * 100
others_per
```

```
## [1] 5.116279
```

```r
# 6. Work Experience
# Check how many levels in "work_ex"
levels(campus_re_backup$work_ex)
```

```
## [1] "No"  "Yes"
```

```r
# Count the frequency of each level in "work_ex"
sum(data.frame(campus_re_backup$work_ex) == "No")
```

```
## [1] 141
```

```r
sum(data.frame(campus_re_backup$work_ex) == "Yes")
```

```
## [1] 74
```

```r
# Percentage of each level
have_workex_per <- (sum(data.frame(campus_re_backup$work_ex) == "Yes")/
                      length(campus_re_backup$work_ex)) * 100
have_workex_per
```

```
## [1] 34.4186
```

```r
no_workex_per <- (sum(data.frame(campus_re_backup$work_ex) == "No") /
                    length(campus_re_backup$work_ex)) * 100
no_workex_per
```

```
## [1] 65.5814
```

```r
# 7. MBA Specialization
# Check how many levels in "MBA_specialize"
levels(campus_re_backup$MBA_specialize)
```

```
## [1] "Mkt&Fin" "Mkt&HR"
```

```r
# Count the frequency of each level in "MBA_specialize"
sum(data.frame(campus_re_backup$MBA_specialize) == "Mkt&Fin")
```

```
## [1] 120
```

```r
sum(data.frame(campus_re_backup$MBA_specialize) == "Mkt&HR")
```

```
## [1] 95
```

```r
# Percentage of each level
MF_per <- (sum(data.frame(campus_re_backup$MBA_specialize) == "Mkt&Fin") /
             length(campus_re_backup$MBA_specialize)) * 100
MF_per
```

```
## [1] 55.81395
```

```r
MH_per <- (sum(data.frame(campus_re_backup$MBA_specialize) == "Mkt&HR") /
             length(campus_re_backup$MBA_specialize)) * 100
MH_per
```

```
## [1] 44.18605
```

```r
# 8. Hiring Status
# Check how many levels in "hiring_status"
levels(campus_re_backup$hiring_status)
```

```
## [1] "Not Placed" "Placed"
```

```r
# Count the frequency of each level in "hiring_status"
sum(data.frame(campus_re_backup$hiring_status) == "Not Placed")
```

```
## [1] 67
```

```r
sum(data.frame(campus_re_backup$hiring_status) == "Placed")
```

```
## [1] 148
```

```r
# Percentage of each level
placed_per <- (sum(data.frame(campus_re_backup$hiring_status) == "Placed") /
                 length(campus_re_backup$hiring_status)) * 100
placed_per
```

```
## [1] 68.83721
```

```r
not_placed_per <- (sum(data.frame(campus_re_backup$hiring_status) == "Not Placed") /
                     length(campus_re_backup$hiring_status)) * 100
not_placed_per
```

```
## [1] 31.16279
```

```r
# 9. 10th sec edu per
# using "campus_re_backup2"
# Sectioning the 10th education percentage into 3 groups
# Note:
# Group1: 0% - 60%
# Group2: 61% - 80%
# Group3: 81% - 100%
campus_re_backup2[campus_re_backup2['ten_sec_edu_per']<=60,'ten_group']= "Group3"
campus_re_backup2[(campus_re_backup2['ten_sec_edu_per']>60) &
                    (campus_re_backup2['ten_sec_edu_per']<81),'ten_group']= "Group2"
campus_re_backup2[(campus_re_backup2['ten_sec_edu_per']>80) &
                    (campus_re_backup2['ten_sec_edu_per']<101),'ten_group']= "Group1"
# Convert chr to factor
class(campus_re_backup2$ten_group)
```

```
## [1] "character"
```

```r
campus_re_backup2$ten_group <- as.factor(campus_re_backup2$ten_group)
str(campus_re_backup2)
```

```
## 'data.frame':    215 obs. of  16 variables:
##  $ serial_number      : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ gender             : Factor w/ 2 levels "F","M": 2 2 2 2 2 2 1 2 2 2 ...
##  $ ten_sec_edu_per    : num  67 79.3 65 56 85.8 ...
##  $ ten_board_of_edu   : Factor w/ 2 levels "Central","Others": 2 1 1 1 1 2 2 1 1 1 ...
##  $ twelve_sec_edu_per : num  91 78.3 68 52 73.6 ...
##  $ twelve_board_of_edu: Factor w/ 2 levels "Central","Others": 2 2 1 1 1 2 2 1 1 1 ...
##  $ specialize_sec_edu : Factor w/ 3 levels "Arts","Commerce",..: 2 3 1 3 2 3 2 3 2 2 ...
##  $ degree_percent     : num  58 77.5 64 52 73.3 ...
##  $ degree_type_field  : Factor w/ 3 levels "Comm&Mgmt","Others",..: 3 3 1 3 1 3 1 3 1 1 ...
##  $ work_ex            : Factor w/ 2 levels "No","Yes": 1 2 1 1 1 2 1 2 1 1 ...
##  $ entrence_test_per  : num  55 86.5 75 66 96.8 ...
##  $ MBA_specialize     : Factor w/ 2 levels "Mkt&Fin","Mkt&HR": 2 1 1 2 1 1 1 1 1 1 ...
##  $ MBA_per            : num  58.8 66.3 57.8 59.4 55.5 ...
##  $ hiring_status      : Factor w/ 2 levels "Not Placed","Placed": 2 2 2 1 2 1 1 2 2 1 ...
##  $ salary             : num  270000 200000 250000 0 425000 0 0 252000 231000 0 ...
##  $ ten_group          : Factor w/ 3 levels "Group1","Group2",..: 2 2 2 3 1 3 3 1 2 3 ...
```

```r
# Check how many levels in "ten_group"
levels(campus_re_backup2$ten_group)
```

```
## [1] "Group1" "Group2" "Group3"
```

```r
# Count the frequency of each level in "ten_group"
sum(data.frame(campus_re_backup2$ten_group) == "Group1")
```

```
## [1] 29
```

```r
sum(data.frame(campus_re_backup2$ten_group) == "Group2")
```

```
## [1] 134
```

```r
sum(data.frame(campus_re_backup2$ten_group) == "Group3")
```

```
## [1] 52
```

```r
# Percentage of each level
g1_per <- (sum(data.frame(campus_re_backup2$ten_group) == "Group1") /
                 length(campus_re_backup2$ten_group)) * 100
g1_per
```

```
## [1] 13.48837
```

```r
g2_per <- (sum(data.frame(campus_re_backup2$ten_group) == "Group2") /
                     length(campus_re_backup2$ten_group)) * 100
g2_per
```

```
## [1] 62.32558
```

```r
g3_per <- (sum(data.frame(campus_re_backup2$ten_group) == "Group3") /
             length(campus_re_backup2$ten_group)) * 100
g3_per
```

```
## [1] 24.18605
```

```r
# 10. 12th sec edu per
# using "campus_re_backup2"
# Sectioning the 12th education percentage into 3 groups
# Note:
# Group1: 0% - 60%
# Group2: 61% - 80%
# Group3: 81% - 100%
campus_re_backup2[campus_re_backup2['twelve_sec_edu_per']<=60,'twelve_group']= "Group3"
campus_re_backup2[(campus_re_backup2['twelve_sec_edu_per']>60) &
                    (campus_re_backup2['twelve_sec_edu_per']<81),'twelve_group']= "Group2"
campus_re_backup2[(campus_re_backup2['twelve_sec_edu_per']>80) &
                    (campus_re_backup2['twelve_sec_edu_per']<101),'twelve_group']= "Group1"
# Convert chr to factor
class(campus_re_backup2$twelve_group)
```

```
## [1] "character"
```

```r
campus_re_backup2$twelve_group <- as.factor(campus_re_backup2$twelve_group)
str(campus_re_backup2)
```

```
## 'data.frame':    215 obs. of  17 variables:
##  $ serial_number      : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ gender             : Factor w/ 2 levels "F","M": 2 2 2 2 2 2 1 2 2 2 ...
##  $ ten_sec_edu_per    : num  67 79.3 65 56 85.8 ...
##  $ ten_board_of_edu   : Factor w/ 2 levels "Central","Others": 2 1 1 1 1 2 2 1 1 1 ...
##  $ twelve_sec_edu_per : num  91 78.3 68 52 73.6 ...
##  $ twelve_board_of_edu: Factor w/ 2 levels "Central","Others": 2 2 1 1 1 2 2 1 1 1 ...
##  $ specialize_sec_edu : Factor w/ 3 levels "Arts","Commerce",..: 2 3 1 3 2 3 2 3 2 2 ...
##  $ degree_percent     : num  58 77.5 64 52 73.3 ...
##  $ degree_type_field  : Factor w/ 3 levels "Comm&Mgmt","Others",..: 3 3 1 3 1 3 1 3 1 1 ...
##  $ work_ex            : Factor w/ 2 levels "No","Yes": 1 2 1 1 1 2 1 2 1 1 ...
##  $ entrence_test_per  : num  55 86.5 75 66 96.8 ...
##  $ MBA_specialize     : Factor w/ 2 levels "Mkt&Fin","Mkt&HR": 2 1 1 2 1 1 1 1 1 1 ...
##  $ MBA_per            : num  58.8 66.3 57.8 59.4 55.5 ...
##  $ hiring_status      : Factor w/ 2 levels "Not Placed","Placed": 2 2 2 1 2 1 1 2 2 1 ...
##  $ salary             : num  270000 200000 250000 0 425000 0 0 252000 231000 0 ...
##  $ ten_group          : Factor w/ 3 levels "Group1","Group2",..: 2 2 2 3 1 3 3 1 2 3 ...
##  $ twelve_group       : Factor w/ 3 levels "Group1","Group2",..: 1 2 2 3 2 3 3 2 2 2 ...
```

```r
# Check how many levels in "ten_group"
levels(campus_re_backup2$twelve_group)
```

```
## [1] "Group1" "Group2" "Group3"
```

```r
# Count the frequency of each level in "ten_group"
sum(data.frame(campus_re_backup2$twelve_group) == "Group1")
```

```
## [1] 18
```

```r
sum(data.frame(campus_re_backup2$twelve_group) == "Group2")
```

```
## [1] 146
```

```r
sum(data.frame(campus_re_backup2$twelve_group) == "Group3")
```

```
## [1] 51
```

```r
# Percentage of each level
g1_12_per <- (sum(data.frame(campus_re_backup2$twelve_group) == "Group1") /
             length(campus_re_backup2$twelve_group)) * 100
g1_12_per
```

```
## [1] 8.372093
```

```
g2_12_per <- (sum(data.frame(campus_re_backup2$twelve_group) == "Group2") /
              length(campus_re_backup2$twelve_group)) * 100
g2_12_per
```

```
## [1] 67.90698
```

```
g3_12_per <- (sum(data.frame(campus_re_backup2$twelve_group) == "Group3") /
              length(campus_re_backup2$twelve_group)) * 100
g3_12_per
```

```
## [1] 23.72093
```

```
# 11. Entrence test percentage
# using "campus_re_backup2"
# Sectioning the 12th education percentage into 3 groups
# Note:
# Group1: 0% - 60%
# Group2: 61% - 80%
# Group3: 81% - 100%
campus_re_backup2[campus_re_backup2['entrence_test_per']<=60,'entrence']= "Group3"
campus_re_backup2[(campus_re_backup2['entrence_test_per']>60) &
                    (campus_re_backup2['entrence_test_per']<81),'entrence']= "Group2"
campus_re_backup2[(campus_re_backup2['entrence_test_per']>80) &
                    (campus_re_backup2['entrence_test_per']<101),'entrence']= "Group1"
# Convert chr to factor
class(campus_re_backup2$entrence)
```

```
## [1] "character"
```

```
campus_re_backup2$entrence <- as.factor(campus_re_backup2$entrence)
str(campus_re_backup2)
```

```
## 'data.frame':    215 obs. of  18 variables:
##  $ serial_number      : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ gender             : Factor w/ 2 levels "F","M": 2 2 2 2 2 2 1 2 2 2 ...
##  $ ten_sec_edu_per    : num  67 79.3 65 56 85.8 ...
##  $ ten_board_of_edu   : Factor w/ 2 levels "Central","Others": 2 1 1 1 1 2 2 1 1 1 ...
##  $ twelve_sec_edu_per : num  91 78.3 68 52 73.6 ...
##  $ twelve_board_of_edu: Factor w/ 2 levels "Central","Others": 2 2 1 1 1 2 2 1 1 1 ...
##  $ specialize_sec_edu : Factor w/ 3 levels "Arts","Commerce",..: 2 3 1 3 2 3 2 3 2 2 ...
##  $ degree_percent     : num  58 77.5 64 52 73.3 ...
##  $ degree_type_field  : Factor w/ 3 levels "Comm&Mgmt","Others",..: 3 3 1 3 1 3 1 3 1 1 ...
##  $ work_ex            : Factor w/ 2 levels "No","Yes": 1 2 1 1 1 2 1 2 1 1 ...
##  $ entrence_test_per  : num  55 86.5 75 66 96.8 ...
##  $ MBA_specialize     : Factor w/ 2 levels "Mkt&Fin","Mkt&HR": 2 1 1 2 1 1 1 1 1 1 ...
##  $ MBA_per            : num  58.8 66.3 57.8 59.4 55.5 ...
##  $ hiring_status      : Factor w/ 2 levels "Not Placed","Placed": 2 2 2 1 2 1 1 2 2 1 ...
##  $ salary             : num  270000 200000 250000 0 425000 0 0 252000 231000 0 ...
##  $ ten_group          : Factor w/ 3 levels "Group1","Group2",..: 2 2 2 3 1 3 3 1 2 3 ...
##  $ twelve_group       : Factor w/ 3 levels "Group1","Group2",..: 1 2 2 3 2 3 3 2 2 2 ...
##  $ entrence           : Factor w/ 3 levels "Group1","Group2",..: 3 1 2 2 1 3 2 2 1 3 ...
```

```r
# Check how many levels in "ten_group"
levels(campus_re_backup2$entrence)
```

```
## [1] "Group1" "Group2" "Group3"
```

```r
# Count the frequency of each level in "ten_group"
sum(data.frame(campus_re_backup2$entrence) == "Group1")
```

```
## [1] 61
```

```r
sum(data.frame(campus_re_backup2$entrence) == "Group2")
```

```
## [1] 95
```

```r
sum(data.frame(campus_re_backup2$entrence) == "Group3")
```

```
## [1] 59
```

```r
# Percentage of each level
entr_1_per <- (sum(data.frame(campus_re_backup2$entrence) == "Group1") /
                length(campus_re_backup2$entrence)) * 100
entr_1_per
```

```
## [1] 28.37209
```

```r
entr_2_per <- (sum(data.frame(campus_re_backup2$entrence) == "Group2") /
                length(campus_re_backup2$entrence)) * 100
entr_2_per
```

```
## [1] 44.18605
```

```r
entr_3_per <- (sum(data.frame(campus_re_backup2$entrence) == "Group3") /
                length(campus_re_backup2$entrence)) * 100
entr_3_per
```

```
## [1] 27.44186
```

```r
# 12. salary
# using "campus_re_backup2"
# Sectioning the 12th education percentage into 8 groups
# Note:
# Group0: 0 - 199999
# Group1: 200000 - 299999
# Group2: 300000 - 399999
# Group3: 400000 - 499999
# Group4: 500000 - 599999
# Group5: 600000 - 699999
# Group6: 700000 - 799999
```

```r
# Group7: 800000 - 899999
# Group8: 900000 - 999999
campus_re_backup2[campus_re_backup2['salary'] < 200000,'salary_group']= "G0"
campus_re_backup2[(campus_re_backup2['salary'] >= 200000) &
                    (campus_re_backup2['salary'] < 300000),'salary_group']= "G1"
campus_re_backup2[(campus_re_backup2['salary'] >= 300000) &
                    (campus_re_backup2['salary'] < 400000),'salary_group']= "G2"
campus_re_backup2[(campus_re_backup2['salary'] >= 400000) &
                    (campus_re_backup2['salary'] < 500000),'salary_group']= "G3"
campus_re_backup2[(campus_re_backup2['salary'] >= 500000) &
                    (campus_re_backup2['salary'] < 600000),'salary_group']= "G4"
campus_re_backup2[(campus_re_backup2['salary'] >= 600000) &
                    (campus_re_backup2['salary'] < 700000),'salary_group']= "G5"
campus_re_backup2[(campus_re_backup2['salary'] >= 700000) &
                    (campus_re_backup2['salary'] < 800000),'salary_group']= "G6"
campus_re_backup2[(campus_re_backup2['salary'] >= 800000) &
                    (campus_re_backup2['salary'] < 900000),'salary_group']= "G7"
campus_re_backup2[(campus_re_backup2['salary'] >= 900000) &
                    (campus_re_backup2['salary'] < 1000000),'salary_group']= "G8"
# Convert chr to factor
class(campus_re_backup2$salary_group)
```

```
## [1] "character"
```

```r
campus_re_backup2$salary_group <- as.factor(campus_re_backup2$salary_group)
str(campus_re_backup2)
```

```
## 'data.frame':    215 obs. of  19 variables:
##  $ serial_number      : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ gender             : Factor w/ 2 levels "F","M": 2 2 2 2 2 2 1 2 2 2 ...
##  $ ten_sec_edu_per    : num  67 79.3 65 56 85.8 ...
##  $ ten_board_of_edu   : Factor w/ 2 levels "Central","Others": 2 1 1 1 1 2 2 1 1 1 ...
##  $ twelve_sec_edu_per : num  91 78.3 68 52 73.6 ...
##  $ twelve_board_of_edu: Factor w/ 2 levels "Central","Others": 2 2 1 1 1 2 2 1 1 1 ...
##  $ specialize_sec_edu : Factor w/ 3 levels "Arts","Commerce",..: 2 3 1 3 2 3 2 3 2 2 ...
##  $ degree_percent     : num  58 77.5 64 52 73.3 ...
##  $ degree_type_field  : Factor w/ 3 levels "Comm&Mgmt","Others",..: 3 3 1 3 1 3 1 3 1 1 ...
##  $ work_ex            : Factor w/ 2 levels "No","Yes": 1 2 1 1 1 2 1 2 1 1 ...
##  $ entrence_test_per  : num  55 86.5 75 66 96.8 ...
##  $ MBA_specialize     : Factor w/ 2 levels "Mkt&Fin","Mkt&HR": 2 1 1 2 1 1 1 1 1 1 ...
##  $ MBA_per            : num  58.8 66.3 57.8 59.4 55.5 ...
##  $ hiring_status      : Factor w/ 2 levels "Not Placed","Placed": 2 2 2 1 2 1 1 2 2 1 ...
##  $ salary             : num  270000 200000 250000 0 425000 0 0 252000 231000 0 ...
##  $ ten_group          : Factor w/ 3 levels "Group1","Group2",..: 2 2 2 3 1 3 3 1 2 3 ...
##  $ twelve_group       : Factor w/ 3 levels "Group1","Group2",..: 1 2 2 3 2 3 3 2 2 2 ...
##  $ entrence           : Factor w/ 3 levels "Group1","Group2",..: 3 1 2 2 1 3 2 2 1 3 ...
##  $ salary_group       : Factor w/ 7 levels "G0","G1","G2",..: 2 2 2 1 4 1 1 2 2 1 ...
```

```r
# Check how many levels in "ten_group"
levels(campus_re_backup2$salary_group)
```

```
## [1] "G0" "G1" "G2" "G3" "G4" "G5" "G8"
```

```r
# Count the frequency of each level in "ten_group"
sum(data.frame(campus_re_backup2$salary_group) == "G0")
```

```
## [1] 67
```

```r
sum(data.frame(campus_re_backup2$salary_group) == "G1")
```

```
## [1] 98
```

```r
sum(data.frame(campus_re_backup2$salary_group) == "G2")
```

```
## [1] 36
```

```r
sum(data.frame(campus_re_backup2$salary_group) == "G3")
```

```
## [1] 8
```

```r
sum(data.frame(campus_re_backup2$salary_group) == "G4")
```

```
## [1] 3
```

```r
sum(data.frame(campus_re_backup2$salary_group) == "G5")
```

```
## [1] 2
```

```r
sum(data.frame(campus_re_backup2$salary_group) == "G6")
```

```
## [1] 0
```

```r
sum(data.frame(campus_re_backup2$salary_group) == "G7")
```

```
## [1] 0
```

```r
sum(data.frame(campus_re_backup2$salary_group) == "G8")
```

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

```r
# Percentage of each level
sala_0_per <- (sum(data.frame(campus_re_backup2$salary_group) == "G0") /
                 length(campus_re_backup2$salary_group)) * 100
sala_0_per
```

```
## [1] 31.16279
```

```r
sala_1_per <- (sum(data.frame(campus_re_backup2$salary_group) == "G1") /
                 length(campus_re_backup2$salary_group)) * 100
sala_1_per
```

```
## [1] 45.5814
```

```r
sala_2_per <- (sum(data.frame(campus_re_backup2$salary_group) == "G2") /
                 length(campus_re_backup2$salary_group)) * 100
sala_2_per
```

```
## [1] 16.74419
```

```r
sala_3_per <- (sum(data.frame(campus_re_backup2$salary_group) == "G3") /
                 length(campus_re_backup2$salary_group)) * 100
sala_3_per
```

```
## [1] 3.72093
```

```r
sala_4_per <- (sum(data.frame(campus_re_backup2$salary_group) == "G4") /
                 length(campus_re_backup2$salary_group)) * 100
sala_4_per
```

```
## [1] 1.395349
```

```r
sala_5_per <- (sum(data.frame(campus_re_backup2$salary_group) == "G5") /
                 length(campus_re_backup2$salary_group)) * 100
sala_5_per
```

```
## [1] 0.9302326
```

```r
sala_6_per <- (sum(data.frame(campus_re_backup2$salary_group) == "G6") /
                 length(campus_re_backup2$salary_group)) * 100
sala_6_per
```

```
## [1] 0
```

```r
sala_7_per <- (sum(data.frame(campus_re_backup2$salary_group) == "G7") /
                 length(campus_re_backup2$salary_group)) * 100
sala_7_per
```

```
## [1] 0
```

```r
sala_8_per <- (sum(data.frame(campus_re_backup2$salary_group) == "G8") /
                 length(campus_re_backup2$salary_group)) * 100
sala_8_per
```

```
## [1] 0.4651163
```

Univariate Histogram

```
graph1 <- ggplot(data = campus_re_backup, aes(x = gender)) +
  geom_bar(fill = "lightblue") +
  labs(y="Count", x = "Gender") +
  theme_bw() +
  theme(text = element_text(size=10)) +
  ggtitle("Comparison of Males and Females \n") +
  theme(plot.title = element_text(size = 10, face = "bold"))

graph2 <- ggplot(data = campus_re_backup, aes(x = ten_board_of_edu)) +
  geom_bar(fill = "lightblue") +
  labs(y="Count", x = "Groups") +
  theme_bw() +
  theme(text = element_text(size=10)) +
  ggtitle("10th Grade Board of Education \n") +
  theme(plot.title = element_text(size = 10, face = "bold"))

graph3 <- ggplot(data = campus_re_backup, aes(x = twelve_board_of_edu)) +
  geom_bar(fill = "lightblue") +
  labs(y="Count", x = "Groups") +
  theme_bw() +
  theme(text = element_text(size=10)) +
  ggtitle("12th Grade Board of Education \n") +
  theme(plot.title = element_text(size = 10, face = "bold"))

graph4 <- ggplot(data = campus_re_backup, aes(x = specialize_sec_edu)) +
  geom_bar(fill = "lightblue") +
  labs(y="Count", x = "Faculty") +
  theme_bw() +
  theme(text = element_text(size=10)) +
  ggtitle("Higher Secondary Education - Specialization \n") +
  theme(plot.title = element_text(size = 10, face = "bold"))

graph5 <- ggplot(data = campus_re_backup, aes(x = degree_type_field)) +
  geom_bar(fill = "lightblue") +
  labs(y="Count", x = "Degree") +
  theme_bw() +
  theme(text = element_text(size=10)) +
  ggtitle("Higher Secondary Education Degree Type \n") +
  theme(plot.title = element_text(size = 10, face = "bold"))

graph6 <- ggplot(data = campus_re_backup, aes(x = work_ex)) +
  geom_bar(fill = "lightblue") +
  labs(y="Count", x = "Work Experience") +
  theme_bw() +
  theme(text = element_text(size=10)) +
  ggtitle("Work Experiences \n") +
  theme(plot.title = element_text(size = 10, face = "bold"))

graph7 <- ggplot(data = campus_re_backup, aes(x = MBA_specialize)) +
  geom_bar(fill = "lightblue") +
  labs(y="Count", x = "MBA Specialization") +
  theme_bw() +
```

```
  theme(text = element_text(size=10)) +
  ggtitle("MBA Specialization \n") +
  theme(plot.title = element_text(size = 10, face = "bold"))

graph8 <- ggplot(data = campus_re_backup, aes(x = hiring_status)) +
  geom_bar(fill = "lightblue") +
  labs(y="Count", x = "Status") +
  theme_bw() +
  theme(text = element_text(size=10)) +
  ggtitle("Hiring Status \n") +
  theme(plot.title = element_text(size = 10, face = "bold"))

graph9 <- ggplot(data = campus_re_backup2, aes(x = ten_group)) +
  geom_bar(fill = "lightblue") +
  labs(y="Count", x = "Percentage Groups") +
  theme_bw() +
  theme(text = element_text(size=10)) +
  ggtitle("Grade 10 Percentage Groups \n") +
  theme(plot.title = element_text(size = 10, face = "bold"))

graph10 <- ggplot(data = campus_re_backup2, aes(x = twelve_group)) +
  geom_bar(fill = "lightblue") +
  labs(y="Count", x = "Percentage Groups") +
  theme_bw() +
  theme(text = element_text(size=10)) +
  ggtitle("Grade 12 Percentage Groups \n") +
  theme(plot.title = element_text(size = 10, face = "bold"))

graph11 <- ggplot(data = campus_re_backup2, aes(x = entrence)) +
  geom_bar(fill = "lightblue") +
  labs(y="Count", x = "Percentage Groups") +
  theme_bw() +
  theme(text = element_text(size=10)) +
  ggtitle("Entrence Test Percentage \n") +
  theme(plot.title = element_text(size = 10, face = "bold"))

graph12 <- ggplot(data = campus_re_backup2, aes(x = salary_group)) +
  geom_bar(fill = "lightblue") +
  labs(y="Count", x = "Salary Groups") +
  theme_bw() +
  theme(text = element_text(size=10)) +
  ggtitle("Salary Distribution \n") +
  theme(plot.title = element_text(size = 10, face = "bold"))

ggarrange(graph1, graph2, graph3, graph4, graph5, graph6,
          graph7, graph8, graph9, graph10, graph11, graph12, ncol = 3, nrow = 4)
```

Conclusions:

- We have about 140 male (65%) and 75 female students (35%) in our sample.
- We have about 116 students (54%) from Central Board and 99 students (46%) from Other Board.
- We have about 84 students (39%) from Central Board and 131 students (61%) from Other Board.
- We have about 11 students (5%) from Arts, 113 students (53%) from Commerce and 91 students (42%) from Science.

- We have 145 students (67%) from Commerce&Management, 70 students (32%) from Science&Technology and others.
- We can see that more students have no work experience before graduation.
- There is NOT much difference in the number of our students specializing between Mkt&Fin and Mkt&HR. In comparison, slightly more students specizlize in Mkt&Fin.
- Majority of students were eventually hired by the company
- Majority of grade 10 students belong to Group2: 61% - 80%
- Approximately half of the students who belong to Group1 in grade 10 rose to Group2. The proportion of students whose grades belong to the third group (Group3) has not changed much.
- The results of the entrance examination show that there is little difference in the number of students in each group. In comparison, 44% of the students belong to the second group, which is slightly more compare to Group1 (28%) and Group3 (27%).
- Most companies limit the salary package for campus recruitment from 200,000 to 400,000. We can see that salary in this range account for 63% of the entire sample.But it does not rule out that the company will offer high wages. One student in the sample received a salary of up to 940,000.

Bivariate Exploration

```
data("iris")

# Check data type
str(campus_re_backup)

1. Restrict the dataset (NUM only)
campus_re_backup_bivariate <- campus_re_backup[,
                                               c("ten_sec_edu_per",
                                                 "twelve_sec_edu_per",
                                                 "degree_percent",
                                                 "entrence_test_per",
                                                 "MBA_per",
                                                 "salary",
                                                 "hiring_status")]

# Plot Histogram
a1<- ggpairs(campus_re_backup_bivariate,
             lower = list(continous = "smooth"),
             diag = list(continous = "barDiag"),
             axisLabels = "show")
a1
a2 <- ggscatmat(campus_re_backup_bivariate,
                columns = c("ten_sec_edu_per",
                            "twelve_sec_edu_per",
                            "degree_percent",
                            "entrence_test_per",
                            "MBA_per",
                            "salary"),
                color = "hiring_status")
a2

2. gender & hiring Status
# Bar plot
```

```
graph13 <- campus_re_backup %>%
  group_by(gender, hiring_status) %>%
  tally() %>%
  complete(hiring_status, fill = list(n = 0)) %>%
  mutate(counting = n)

graph13 <- ggplot(graph13, aes(hiring_status, counting, fill = gender)) +
  geom_bar(stat = 'identity', position = 'dodge') +
  labs(x ="Status", y = "Count") +
  theme_bw() +
  theme(text = element_text(size = 10)) +
  ggtitle("Hiring Status with Gender \n") +
  theme(plot.title = element_text(size = 10, face = "bold")) +
  scale_fill_brewer(palette="Paired")
# Hiring percentage
male_hire_per <- (100 /140) * 100
male_hire_per
female_hire_per <- ((148 - 100) / 75) * 100
female_hire_per


3. ten_sec_edu_per & hiring status
# Box Plot
graph14 <- campus_re_backup %>%
  ggplot(aes(y = ten_sec_edu_per, fill = ten_sec_edu_per)) +
  geom_boxplot(aes(fill = hiring_status)) +
  labs(x ="Status", y = "Count") +
  theme_bw() +
  theme(text = element_text(size = 10)) +
  ggtitle("Hiring Status with G10 Percentage \n") +
  theme(plot.title = element_text(size = 10, face = "bold")) +
  scale_fill_brewer(palette = "Paired")


4. twevle_sec_edu_per & hiring status
# Box Plot
graph15 <- campus_re_backup %>%
  ggplot(aes(y = twelve_sec_edu_per, fill = twelve_sec_edu_per)) +
  geom_boxplot(aes(fill = hiring_status)) +
  labs(x ="Status", y = "Count") +
  theme_bw() +
  theme(text = element_text(size = 10)) +
  ggtitle("Hiring Status with G12 Percentage \n") +
  theme(plot.title = element_text(size = 10, face = "bold")) +
  scale_fill_brewer(palette = "Paired")


5. ten_board_of_edu & hiring status
graph16 <- campus_re_backup %>%
  group_by(ten_board_of_edu, hiring_status) %>%
  tally() %>%
  complete(hiring_status, fill = list(n = 0)) %>%
  mutate(counting = n)
```

```
graph16 <- ggplot(graph16, aes(hiring_status, counting, fill = ten_board_of_edu)) +
  geom_bar(stat = 'identity', position = 'dodge') +
  labs(x ="Status", y = "Count") +
  theme_bw() +
  theme(text = element_text(size = 10)) +
  ggtitle("Hiring Status with BoE(G10) \n") +
  theme(plot.title = element_text(size = 10, face = "bold")) +
  scale_fill_brewer(palette="Paired")


6. twelve_board_of_edu & hring status
# Bar plot
graph17 <- campus_re_backup %>%
  group_by(twelve_board_of_edu, hiring_status) %>%
  tally() %>%
  complete(hiring_status, fill = list(n = 0)) %>%
  mutate(counting = n)

graph17 <- ggplot(graph17, aes(hiring_status, counting, fill = twelve_board_of_edu)) +
  geom_bar(stat = 'identity', position = 'dodge') +
  labs(x ="Status", y = "Count") +
  theme_bw() +
  theme(text = element_text(size = 10)) +
  ggtitle("Hiring Status with BoE(G12) \n") +
  theme(plot.title = element_text(size = 10, face = "bold")) +
  scale_fill_brewer(palette="Paired")


7. twelve_board_of_edu $ twelve_sed_edu_per
graph18 <- campus_re_backup %>%
  ggplot(aes(y = twelve_sec_edu_per, fill = twelve_sec_edu_per)) +
  geom_boxplot(aes(fill = twelve_board_of_edu)) +
  labs(x ="Status", y = "Count") +
  theme_bw() +
  theme(text = element_text(size = 10)) +
  ggtitle("Grade Percentage with BoE \n") +
  theme(plot.title = element_text(size = 10, face = "bold")) +
  scale_fill_brewer(palette = "Paired")


8. specialize_sec_edu & hiring status
graph19 <- campus_re_backup %>%
  group_by(specialize_sec_edu, hiring_status) %>%
  tally() %>%
  complete(hiring_status, fill = list(n = 0)) %>%
  mutate(counting = n)

graph19 <- ggplot(graph19, aes(hiring_status, counting, fill = specialize_sec_edu)) +
  geom_bar(stat = 'identity', position = 'dodge') +
  labs(x ="Status", y = "Count") +
  theme_bw() +
  theme(text = element_text(size = 10)) +
  ggtitle("Hiring Status with G12 Specialization \n") +
  theme(plot.title = element_text(size = 10, face = "bold")) +
```

```
  scale_fill_brewer(palette = "Paired")


9. degree percentage & hiring status
graph20 <- campus_re_backup %>%
  ggplot(aes(y = degree_percent, fill = degree_percent)) +
  geom_boxplot(aes(fill = hiring_status)) +
  labs(x ="Status", y = "Count") +
  theme_bw() +
  theme(text = element_text(size = 10)) +
  ggtitle("Hiring Status with Degree Percentage \n") +
  theme(plot.title = element_text(size = 10, face = "bold")) +
  scale_fill_brewer(palette = "Paired")


10. degree type field & hiring status
graph21 <- campus_re_backup %>%
  group_by(degree_type_field, hiring_status) %>%
  tally() %>%
  complete(hiring_status, fill = list(n = 0)) %>%
  mutate(counting = n)

graph21 <- ggplot(graph21, aes(hiring_status, counting, fill = degree_type_field)) +
  geom_bar(stat = 'identity', position = 'dodge') +
  labs(x ="Status", y = "Count") +
  theme_bw() +
  theme(text = element_text(size = 10)) +
  ggtitle("Hiring Status with Degree Type \n") +
  theme(plot.title = element_text(size = 10, face = "bold")) +
  scale_fill_brewer(palette = "Paired")

11. work_ex & hiring status
graph22 <- campus_re_backup %>%
  group_by(work_ex, hiring_status) %>%
  tally() %>%
  complete(hiring_status, fill = list(n = 0)) %>%
  mutate(counting = n)

graph22 <- ggplot(graph22, aes(hiring_status, counting, fill = work_ex)) +
  geom_bar(stat = 'identity', position = 'dodge') +
  labs(x ="Status", y = "Count") +
  theme_bw() +
  theme(text = element_text(size = 10)) +
  ggtitle("Hiring Status with Work Experience \n") +
  theme(plot.title = element_text(size = 10, face = "bold")) +
  scale_fill_brewer(palette = "Paired")


12. entrence_test_per & hiring status
graph23 <- campus_re_backup %>%
  ggplot(aes(y = entrence_test_per, fill = entrence_test_per)) +
  geom_boxplot(aes(fill = hiring_status)) +
  labs(x ="Status", y = "Count") +
  theme_bw() +
```

```
  theme(text = element_text(size = 10)) +
  ggtitle("Hiring Status with Entrence Test Persentage \n") +
  theme(plot.title = element_text(size = 10, face = "bold")) +
  scale_fill_brewer(palette = "Paired")


13. MBA specialize & hiring status
graph24 <- campus_re_backup %>%
  group_by(MBA_specialize, hiring_status) %>%
  tally() %>%
  complete(hiring_status, fill = list(n = 0)) %>%
  mutate(counting = n)

graph24 <- ggplot(graph24, aes(hiring_status, counting, fill = MBA_specialize)) +
  geom_bar(stat = 'identity', position = 'dodge') +
  labs(x ="Status", y = "Count") +
  theme_bw() +
  theme(text = element_text(size = 10)) +
  ggtitle("Hiring Status with MBA Specialization \n") +
  theme(plot.title = element_text(size = 10, face = "bold")) +
  scale_fill_brewer(palette = "Paired")

ggarrange(graph13, graph14, graph15,
          graph16, graph17, graph18,
          graph19, graph20, graph21,
          graph22, graph23, graph24, ncol = 3, nrow = 4)
```

Multivariate Exploration

```
# G10 grade, G12 grade, degree percent & salary (3D)
# Point colors
marker <- list(color = ~salary,
               colorscale = c('#FFE1A1', '#683531'),
               showscale = TRUE)
# Create the plot
plot1 <- plot_ly(campus_re_backup,
             x = ~ten_sec_edu_per,
             y = ~twelve_sec_edu_per,
             z = ~degree_percent, marker = marker) %>%
  add_markers() %>%
  layout(scene = list(xaxis = list(title = 'G10 Percentage'),
                      yaxis = list(title = 'G12 Percentage'),
                      zaxis = list(title = 'Degree Percent')),
         annotations = list(
           x = 1.15,
           y = 1.05,
           text = 'Salary \n',
           showarrow = FALSE
         ))
plot1
```

Model Fitting 1

```
## WHY PLACED (lm)
# Trying to find "fitted model"
# model1
str(campus_re_backup)
campus_re_backup[campus_re_backup['hiring_status'] == "Placed",'hire'] = "1"
str(campus_re_backup)
# Convert chr to num
campus_re_backup$hire <- as.numeric(campus_re_backup$hire)
# Only keep "Placed" status
campus_re_backup <- campus_re_backup[complete.cases(campus_re_backup), ]
str(campus_re_backup)
# model1
model1 <- lm(hire ~ serial_number +
               gender +
               ten_sec_edu_per +
               ten_board_of_edu +
               twelve_sec_edu_per +
               twelve_board_of_edu +
               specialize_sec_edu +
               degree_percent +
               degree_type_field +
               work_ex +
               entrence_test_per +
               MBA_specialize +
               MBA_per +
               salary, data = campus_re_backup)
summary(model1)
## significant:
# degree_type_field
# twelve_sec_edu_per
# intercept

# plot
plot(model1, which = c(1, 2))
# Remove "MBA_per" (p = 0.90023)


# model2
model2 <- lm(hire ~ serial_number +
               gender +
               ten_sec_edu_per +
               ten_board_of_edu +
               twelve_sec_edu_per +
               twelve_board_of_edu +
               specialize_sec_edu +
               degree_percent +
               degree_type_field +
               work_ex +
               entrence_test_per +
               MBA_specialize +
               salary, data = campus_re_backup)
summary(model2)
```

```
## Significant:
# degree_type_field
# twelve_sec_edu_per
# intercept
# Remove "specialize_sec_edu" (p = 0.77851)


# model3
model3 <- lm(hire ~ serial_number +
               gender +
               ten_sec_edu_per +
               ten_board_of_edu +
               twelve_sec_edu_per +
               twelve_board_of_edu +
               degree_percent +
               degree_type_field +
               work_ex +
               entrence_test_per +
               MBA_specialize +
               salary, data = campus_re_backup)
summary(model3)
## Significant:
# degree_type_field
# twelve_sec_edu_per
# intercept
# Remove "twelve_board_of_edu" (p = 0.911418)


# model4
model4 <- lm(hire ~ serial_number +
               gender +
               ten_sec_edu_per +
               ten_board_of_edu +
               twelve_sec_edu_per +
               degree_percent +
               degree_type_field +
               work_ex +
               entrence_test_per +
               MBA_specialize +
               salary, data = campus_re_backup)
summary(model4)
## Significant:
# degree_type_field
# twelve_sec_edu_per
# intercept
# Remove "work_ex" (p = 0.730570)


# model5
model5 <- lm(hire ~ serial_number +
               gender +
               ten_sec_edu_per +
               ten_board_of_edu +
               twelve_sec_edu_per +
```

```
               degree_percent +
               degree_type_field +
               entrence_test_per +
               MBA_specialize +
               salary, data = campus_re_backup)
summary(model5)
## Significant:
# degree_type_field
# twelve_sec_edu_per
# intercept
# Remove "gender" (p = 0.641972)


# model6
model6 <- lm(hire ~ serial_number +
               ten_sec_edu_per +
               ten_board_of_edu +
               twelve_sec_edu_per +
               degree_percent +
               degree_type_field +
               entrence_test_per +
               MBA_specialize +
               salary, data = campus_re_backup)
summary(model6)
## Significant:
# degree_type_field
# twelve_sec_edu_per
# intercept
# degree_percent
# Remove "salary" (p = 0.669565)


# model7
model7 <- lm(hire ~ serial_number +
               ten_sec_edu_per +
               ten_board_of_edu +
               twelve_sec_edu_per +
               degree_percent +
               degree_type_field +
               entrence_test_per +
               MBA_specialize, data = campus_re_backup)
summary(model7)
## Significant:
# degree_type_field
# twelve_sec_edu_per
# intercept
# degree_percent
# Remove "MBA_specialize" (p = 0.224075)


# model8
model8 <- lm(hire ~ serial_number +
               ten_sec_edu_per +
               ten_board_of_edu +
```

```
                twelve_sec_edu_per +
                degree_percent +
                degree_type_field +
                entrence_test_per, data = campus_re_backup)
summary(model8)
## Significant:
# degree_type_field
# twelve_sec_edu_per
# intercept
# degree_percent
# Remove "ten_sec_edu_per" (p = 0.224693)


# model9
model9 <- lm(hire ~ serial_number +
                ten_board_of_edu +
                twelve_sec_edu_per +
                degree_percent +
                degree_type_field +
                entrence_test_per, data = campus_re_backup)
summary(model9)
## Significant:
# degree_type_field
# twelve_sec_edu_per
# intercept
# degree_percent
# entrence_test_per
# Remove "serial_number" (p = 0.21963)


# model10
model10 <- lm(hire ~ ten_board_of_edu +
                 twelve_sec_edu_per +
                 degree_percent +
                 degree_type_field +
                 entrence_test_per, data = campus_re_backup)
summary(model10)
## Significant:
# degree_type_field
# twelve_sec_edu_per
# intercept
# degree_percent
# entrence_test_per
# Remove "ten_board_of_edu" (p = 0.081713)


# model11
model11 <- lm(hire ~ twelve_sec_edu_per +
                 degree_percent +
                 degree_type_field +
                 entrence_test_per, data = campus_re_backup)
summary(model11)
## Significant:
# degree_type_field
```

```
# twelve_sec_edu_per
# intercept
# degree_percent
# Remove "entrence_test_per" (p = 0.053881)


# model12
model12 <- lm(hire ~ twelve_sec_edu_per +
                degree_percent +
                degree_type_field, data = campus_re_backup)
summary(model12)
##### ALL SIGNIFICANT #####

######################## Try add it back ###################

# fit13 = fit12 + ten_board_of_edu
model13 <- lm(hire ~ twelve_sec_edu_per +
                degree_percent +
                degree_type_field +
                ten_board_of_edu, data = campus_re_backup)
summary(model13)
## p = 0.13969 DROP (NO ten_board_of_edu)


# fit14 = fit12 + serial_number
model14 <- lm(hire ~ twelve_sec_edu_per +
                degree_percent +
                degree_type_field +
                serial_number, data = campus_re_backup)
summary(model14)
## p = 0.15005 DROP (NO serial_number)


# fit15 = fit12 + ten_sec_edu_per
model15 <- lm(hire ~ twelve_sec_edu_per +
                degree_percent +
                degree_type_field +
                ten_sec_edu_per, data = campus_re_backup)
summary(model15)
## p = 0.25463 DROP (NO ten_sec_edu_per)


# fit16 = fit12 + MBA_specialize
model16 <- lm(hire ~ twelve_sec_edu_per +
                degree_percent +
                degree_type_field +
                MBA_specialize, data = campus_re_backup)
summary(model16)
## p = 0.11591 DROP (NO MBA_specialize)


# fit17 = fit12 + salary
model17 <- lm(hire ~ twelve_sec_edu_per +
                degree_percent +
```

```
                degree_type_field +
                salary, data = campus_re_backup)
summary(model17)
## p = 0.36321 DROP (NO salary)


# fit18 = fit12 + gender
model18 <- lm(hire ~ twelve_sec_edu_per +
                degree_percent +
                degree_type_field +
                gender, data = campus_re_backup)
summary(model18)
## p = 0.98047 DROP (NO gender)


# fit19 = fit12 + work_ex
model19 <- lm(hire ~ twelve_sec_edu_per +
                degree_percent +
                degree_type_field +
                work_ex, data = campus_re_backup)
summary(model19)
## p = 0.25554 DROP (NO work_ex)


# fit20 = fit12 + twelve_board_of_edu
model20 <- lm(hire ~ twelve_sec_edu_per +
                degree_percent +
                degree_type_field +
                twelve_board_of_edu, data = campus_re_backup)
summary(model20)
## p = 0.33794 DROP (NO twelve_board_of_edu)


# fit21 = fit12 + specialize_sec_edu
model21 <- lm(hire ~ twelve_sec_edu_per +
                degree_percent +
                degree_type_field +
                specialize_sec_edu, data = campus_re_backup)
summary(model21)
## p(Commerce) = 0.55925 DROP (NO specialize_sec_eduCommerce)
## p(Science) = 0.66068 DROP (NO specialize_sec_eduScience)


# fit22 = fit12 + MBA_per
model22 <- lm(hire ~ twelve_sec_edu_per +
                degree_percent +
                degree_type_field +
                MBA_per, data = campus_re_backup)
summary(model22)
## p = 0.38016 DROP (NO MBA_per)

#############################
######## FINAL MODEL ########
#############################
```

```
summary(model12)

# residual plot:
plot(model12, which = c(1, 2))
```

Model Fitting 2

```
## SALARY (lm)
# Trying to find "fitted model"
# fit1
campus_fit1 <- lm(salary ~ serial_number +
                    gender +
                    ten_sec_edu_per +
                    ten_board_of_edu +
                    twelve_sec_edu_per +
                    twelve_board_of_edu +
                    specialize_sec_edu +
                    degree_percent +
                    degree_type_field +
                    work_ex +
                    entrence_test_per +
                    MBA_specialize +
                    MBA_per +
                    hiring_status, data = campus_re_backup2)
summary(campus_fit1)
## significant:
# hiring_statusPlaced
# MBA_per
# degree_type_fieldSci&Tech
# Remove "twelve_sec_edu_per" (p = 0.9334)


# fit2
campus_fit2 <- lm(salary ~ serial_number +
                    gender +
                    ten_sec_edu_per +
                    ten_board_of_edu +
                    twelve_board_of_edu +
                    specialize_sec_edu +
                    degree_percent +
                    degree_type_field +
                    work_ex +
                    entrence_test_per +
                    MBA_specialize +
                    MBA_per +
                    hiring_status,
                  data = campus_re_backup2)
summary(campus_fit2)
## significant:
# hiring_statusPlaced
# MBA_per
# degree_type_fieldSci&Tech
# Remove "ten_board_of_edu" (p = 0.8534)
```

```
# fit3
campus_fit3 <- lm(salary ~ serial_number +
                    gender +
                    ten_sec_edu_per +
                    twelve_board_of_edu +
                    specialize_sec_edu +
                    degree_percent +
                    degree_type_field +
                    work_ex +
                    entrence_test_per +
                    MBA_specialize +
                    MBA_per +
                    hiring_status,
                  data = campus_re_backup2)
summary(campus_fit3)
## significant:
# hiring_statusPlaced
# MBA_per
# degree_type_fieldSci&Tech
# Remove "serial_number" (p = 0.8079)


# fit4
campus_fit4 <- lm(salary ~ gender +
                    ten_sec_edu_per +
                    twelve_board_of_edu +
                    specialize_sec_edu +
                    degree_percent +
                    degree_type_field +
                    work_ex +
                    entrence_test_per +
                    MBA_specialize +
                    MBA_per +
                    hiring_status,
                  data = campus_re_backup2)
summary(campus_fit4)
## significant:
# hiring_statusPlaced
# MBA_per
# degree_type_fieldSci&Tech
# Remove "twelve_board_of_edu" (p = 0.5799)


# fit5
campus_fit5 <- lm(salary ~ gender +
                    ten_sec_edu_per +
                    specialize_sec_edu +
                    degree_percent +
                    degree_type_field +
                    work_ex +
                    entrence_test_per +
                    MBA_specialize +
```

```
                    MBA_per +
                    hiring_status,
                  data = campus_re_backup2)
summary(campus_fit5)
## significant:
# hiring_statusPlaced
# MBA_per
# degree_type_fieldSci&Tech
# Remove "specialize_sec_edu" (p = 0.4361)


# fit6
campus_fit6 <- lm(salary ~ gender +
                    ten_sec_edu_per +
                    degree_percent +
                    degree_type_field +
                    work_ex +
                    entrence_test_per +
                    MBA_specialize +
                    MBA_per +
                    hiring_status,
                  data = campus_re_backup2)
summary(campus_fit6)
## significant:
# hiring_statusPlaced
# MBA_per
# Remove "degree_type_field" (p = 0.8535)


# fit7
campus_fit7 <- lm(salary ~ gender +
                    ten_sec_edu_per +
                    degree_percent +
                    work_ex +
                    entrence_test_per +
                    MBA_specialize +
                    MBA_per +
                    hiring_status,
                  data = campus_re_backup2)
summary(campus_fit7)
## significant:
# hiring_statusPlaced
# MBA_per
# gender
# Remove "ten_sec_edu_per" (p = 0.6641)


# fit8
campus_fit8 <- lm(salary ~ gender +
                    degree_percent +
                    work_ex +
                    entrence_test_per +
                    MBA_specialize +
                    MBA_per +
```

```
                    hiring_status,
                  data = campus_re_backup2)
summary(campus_fit8)
## significant:
# hiring_statusPlaced
# MBA_per
# gender
# intercept
# Remove "MBA_specialize" (p = 0.3143)


# fit9
campus_fit9 <- lm(salary ~ gender +
                    degree_percent +
                    work_ex +
                    entrence_test_per +
                    MBA_per +
                    hiring_status,
                  data = campus_re_backup2)
summary(campus_fit9)
## significant:
# hiring_statusPlaced
# MBA_per
# gender
# intercept
# Remove "degree_percent" (p = 0.2994)


# fit10
campus_fit10 <- lm(salary ~ gender +
                    work_ex +
                    entrence_test_per +
                    MBA_per +
                    hiring_status,
                  data = campus_re_backup2)
summary(campus_fit10)
## significant:
# hiring_statusPlaced
# MBA_per
# gender
# intercept
# Remove "work_ex" (p = 0.21086)


# fit11
campus_fit11 <- lm(salary ~ gender +
                     entrence_test_per +
                     MBA_per +
                     hiring_status,
                   data = campus_re_backup2)
summary(campus_fit11)
## significant:
# hiring_statusPlaced
# MBA_per
```

```
# gender
# intercept
#Remove "entrence_test_per" (p = 0.1369)


# fit12
campus_fit12 <- lm(salary ~ gender +
                     MBA_per +
                     hiring_status,
                   data = campus_re_backup2)
summary(campus_fit12)

######################## Try add it back ###################

# fit13 = fit12 + work_ex
campus_fit13 <- lm(salary ~ gender +
                     MBA_per +
                     hiring_status +
                     work_ex,
                   data = campus_re_backup2)
summary(campus_fit13)
## p = 0.22826 DROP (NO work_ex)


# fit14 = fit12 + degree_percent
campus_fit14 <- lm(salary ~ gender +
                     MBA_per +
                     hiring_status +
                     degree_percent,
                   data = campus_re_backup2)
summary(campus_fit14)
## p = 0.367 DROP (NO degree_percent)


# fit15 = fit12 + MBA_specialize
campus_fit15 <- lm(salary ~ gender +
                     MBA_per +
                     hiring_status +
                     MBA_specialize,
                   data = campus_re_backup2)
summary(campus_fit15)
## p = 0.19425 DROP (NO MBA_specialize)


# fit16 = fit12 + ten_sec_edu_per
campus_fit16 <- lm(salary ~ gender +
                     MBA_per +
                     hiring_status +
                     ten_sec_edu_per,
                   data = campus_re_backup2)
summary(campus_fit16)
## p = 0.62998 DROP (NO ten_sec_edu_per)
```

```
# fit17 = fit12 + degree_type_field
campus_fit17 <- lm(salary ~ gender +
                     MBA_per +
                     hiring_status +
                     degree_type_field,
                   data = campus_re_backup2)
summary(campus_fit17)
## p(Others) = 0.71860
## p(Sci&Tech) = 0.11596 DROP (NO degree_type_field)


# fit18 = fit12 + specialize_sec_edu
campus_fit18 <- lm(salary ~ gender +
                     MBA_per +
                     hiring_status +
                     specialize_sec_edu,
                   data = campus_re_backup2)
summary(campus_fit18)
## p(Commerce) = 0.48079
## p(Science) = 0.43263 DROP (NO specialize_sec_edu)


# fit19 = fit12 + twelve_board_of_edu
campus_fit19 <- lm(salary ~ gender +
                     MBA_per +
                     hiring_status +
                     twelve_board_of_edu,
                   data = campus_re_backup2)
summary(campus_fit19)
## p = 0.57598 DROP (NO twelve_board_of_edu)


# fit20 = fit12 + serial_number
campus_fit20 <- lm(salary ~ gender +
                     MBA_per +
                     hiring_status +
                     serial_number,
                   data = campus_re_backup2)
summary(campus_fit20)
## p = 0.63659 DROP (NO serial_number)


# fit21 = fit12 + ten_board_of_edu
campus_fit21 <- lm(salary ~ gender +
                     MBA_per +
                     hiring_status +
                     ten_board_of_edu,
                   data = campus_re_backup2)
summary(campus_fit21)
## p = 0.82053 DROP (NO ten_board_of_edu)


# fit22 = fit12 + twelve_sec_edu_per
campus_fit22 <- lm(salary ~ gender +
```

```
                      MBA_per +
                      hiring_status +
                      twelve_sec_edu_per,
                    data = campus_re_backup2)
summary(campus_fit22)
## p = 0.97686 DROP (NO twelve_sec_edu_per)


#############################
######## FINAL MODEL ########
#############################
summary(campus_fit12)
plot(campus_fit1, c(1, 2))
plot(campus_fit12, c(1, 2))
```

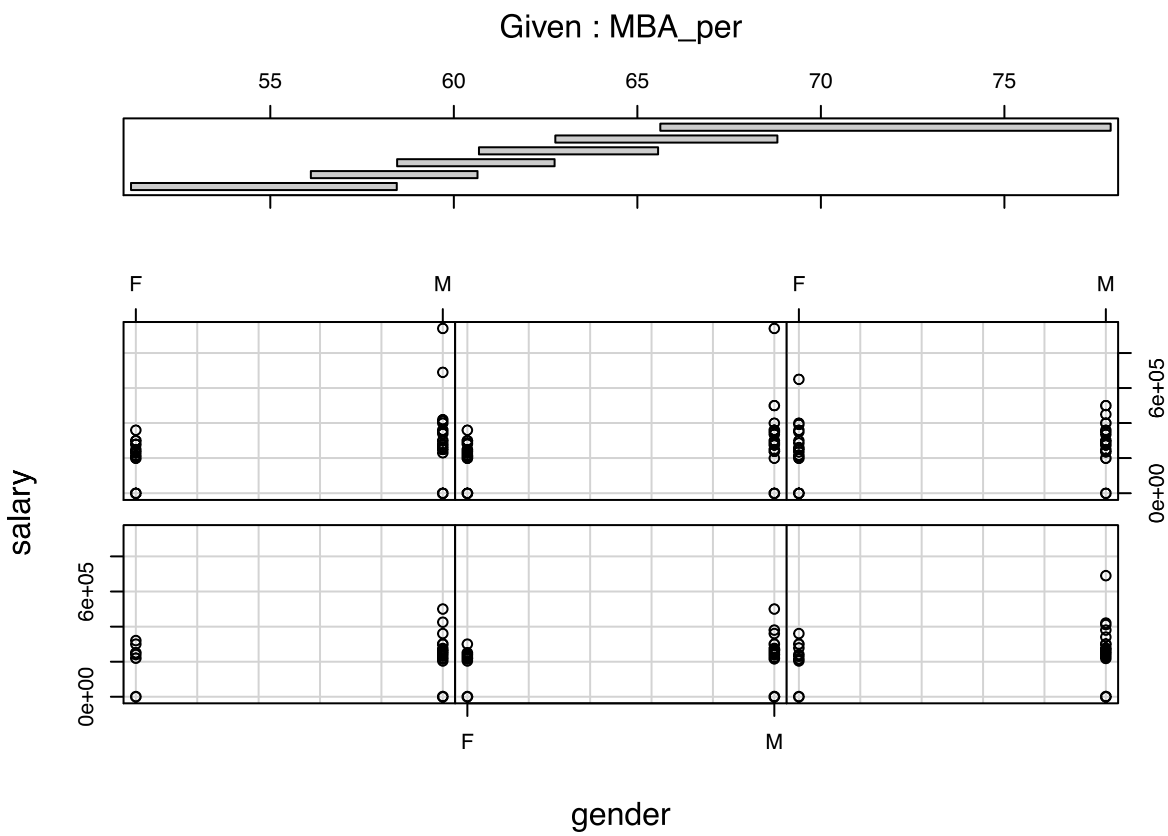
Spatial Modelling

```
# Exploring Stationarity
# Based on model 12
# hire ~ twelve_sec_edu_per + degree_percent + degree_type_field
# NOW,
# A.
# hiring_status & degree_percent, GIVEN degree_type_field
coplot(hiring_status ~ degree_percent | degree_type_field,
             data = campus_re_backup)
# hiring_status & degree_type_field, GIVEN degree_percent
coplot(hiring_status ~ degree_type_field | degree_percent,
             data = campus_re_backup)

# B.
# hiring_status & twelve_sec_edu_per, GIVEN degree_type_field
coplot(hiring_status ~ twelve_sec_edu_per | degree_type_field,
             data = campus_re_backup)
# hiring_status & degree_type_field, GIVEN twelve_sec_edu_per
coplot(hiring_status ~ degree_type_field | twelve_sec_edu_per,
             data = campus_re_backup)

# C.
# hiring_status & twelve_sec_edu_per, GIVEN degree_percent
coplot(hiring_status ~ twelve_sec_edu_per | degree_percent,
             data = campus_re_backup)
# hiring_status & degree_percent, GIVEN twelve_sec_edu_per
coplot(hiring_status ~ degree_percent | twelve_sec_edu_per,
             data = campus_re_backup)
```

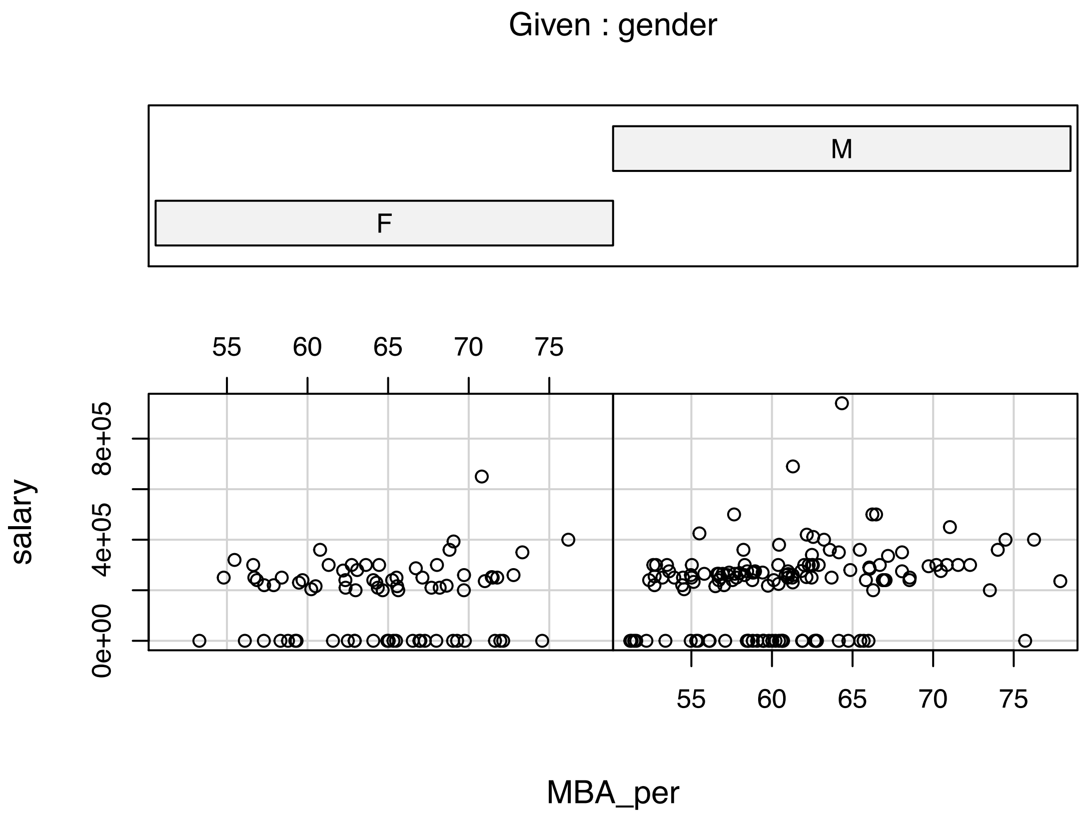
Exploring Stationarity

```
# Based on campus_fit12
# salary ~ gender + MBA_per + hiring_status
# NOW,
# A.
# salary & gender, GIVEN MBA_per
```

```r
coplot(salary ~ gender | MBA_per,
       data = campus_re_backup)
```

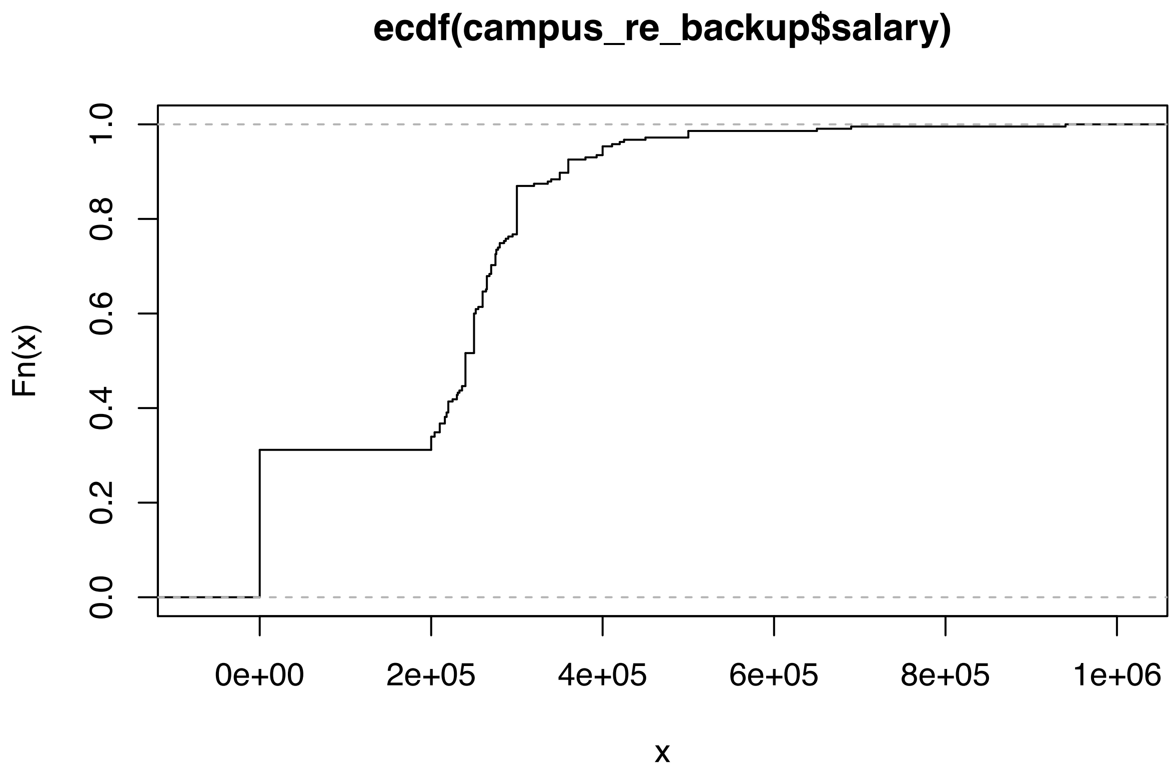



```r
# B.
# salary & MBA_per, GIVEN gender
coplot(salary ~ MBA_per | gender,
       data = campus_re_backup)
```



Cumulative distribution function of salary

```
cdf_hiring_status <- ecdf(campus_re_backup$salary)
plot(cdf_hiring_status, verticals = T, do.points = F)
```


ecdf(campus_re_backup$salary)
Fn(x)
0.0
0.2
0.4
0.6
0.8
1.0
0e+00
2e+05
4e+05
6e+05
8e+05
1e+06
x