UNIVERSITAT POLITÈCNICA
DE CATALUNYA
BARCELONATECH

Final Project

Ettore Falde, Samoussa Fofana, Federico Basaglia

11/29/2021

Contents

1 Introduction

In this report we consider that the new CEO of a specific IT company has contacted us because she wants us to **analyze the current Human Resources status** of the company. She has just sent a data set with all available employee information. As we can see, the **company has two locations**: the first one in **London**, and the second one in **Barcelona**.

The new CEO is concerned about several issues. She truly believes in gender equality in organizations as it implies a signal to society. On the other hand, she is concerned that the offices in Barcelona do not follow a similar structure to the one in London. **In her opinion, the structure of the Barcelona offices should tend towards the London structure.** In her meeting with us, she also told us that she would like to know the attitudes (e.g., satisfaction) of the employees across the different departments and if anything could be done to improve them. Finally, she commented that she is very concerned about the company's succession strategy and in particular some positions in certain departments.

Let's consider that the new CEO of a specific IT company has contacted us because she wants us to analyze the current Human Resources status of the company. She has just sent a data set with all available employee information. This information is in the attached data set. As we can see, the company has two locations: the first one in London, and the second one in Barcelona.

The new CEO is concerned about several issues. She truly believes in gender equality in organizations as it implies a signal to society. On the other hand, she is concerned that the offices in Barcelona do not follow a similar structure to the one in London. In her opinion, the structure of the Barcelona offices should tend towards the London structure. In her meeting with us, she also told us that she would like to know the attitudes (e.g., satisfaction) of the employees across the different departments and if anything could be done to improve them. Finally, she commented that she is very concerned about the company's succession strategy and in particular some positions in certain departments.

Based on this information, we need to carry out an exploratory data analysis and prepare a technical report (with Rmarkdown) and a technical presentation (5-10 minutes).

Note: It is highly recommended to seek external sources of information (either in dataset or report formats) for the analysis and the reporting.

Based on this information, we will to carry out an exploratory data analysis and prepare a technical report (with Rmarkdown) and a technical presentation (5-10 minutes).

2 Setup the software

The software used for the development of the study and the writing of the report is R[1]. The first step is to define the work directory and to load the libraries needed:

```
library(tidyverse)
library(ggplot2)
library(GGally)
library(gridExtra)
library(yardstick)
library(broom)
library(janitor)
library(caTools)
library(ROCR)
library(corrplot)
library(tidytext)
library(glue)
library(scales)
```

```
library(plotly)
library(patchwork)
library(skimr)
library(RColorBrewer)
```

3 Importing Data

The first step is to load the dataset in the system, and check the names of the variables.

```
source("functions_script.R")
```

```
mydb <- read.csv2("dataset.csv")
# web_db <- read.csv("WA_Fn-UseC_-HR-Employee-Attrition.csv")
# names(web_db)
```

3.1 Variables analysis

We got the dataset from the website of Atenea, it is composed by 1506 observations of 36 variables. The variables selected for this dataset are:

1. **Age**: Variable that represent the age of the employee
2. **Attrition**: variable that represent the departure of employees from the organization for any reason
3. **BusinessTravel**: Represent how often an employee travel for work purpose
4. **DailyRate**: The amount of money the employees are paid per day
5. **Department**: Department of the company at which the employee belong
6. **DistanceFromHome**: Employee home distance from the workplace
7. **Education**: Educational level of the employee (1=Below College, 2=College, 3=Bachelor, 4=Master,5 = Doctor)
8. **EducationField**: Education field of employee (Human Resources, Life Sciencies, Marketing, Medical, Technical Degree, Other)
9. **EmployeeCount**: Coolumn all equal to 1 to count the total number of employee in the data set
10. **EmployeeNumber**: unique number to identify the employee
11. **EnvironmentSatisfaction**: level of environment satisfaction (1=Low, 2=Medium, 3=High, 4=Very High)
12. **Gender**: Gender of the employee (Male, Female)
13. **HourlyRate**: The amount of money the employees are paid per hour
14. **JobInvolvement**: Level of involvement of the employee (1=Low, 2=Medium, 3=High, 4=Very High)
15. **JobLevel**: Is a category of authority in the company (1=low, 5=High)
16. **JobRole**: Represent the role cover by the employee (Sales Executive, Research Scientist, Laboratory Technician, Manufacturing Director, Healthcare Representative, Manager, Sales Representative, Research Director, Human Resources)
17. **JobSatisfaction**: Level of satisfaction of the employee (1=Low, 2=Medium, 3=High, 4=Very High)
18. **MaritalStatus**: Marital status of the employee (Divorced, Married, Single)
19. **MonthlyIncome**: Monthly income of the employee
20. **MonthlyRate**: Monthly rate of employee
21. **NumCompaniesWorked**: Number of companies for ehich the employee worked
22. **Over18**: If the age of the employee is higher than 18 (Y = yes, N = no)
23. **OverTime**: If the employee perform over time (Yes, No)
24. **PercentSalaryHike**: Represent the percentage inrease of a salary
25. **PerformanceRating**: Performance rating of the employee (1=Low, 2=Good,3=Excellent,4=Outstanding)

26. **RelationshipSatisfaction**: Relationship satisfaction of the employee (1=Low, 2=Medium, 3=High, 4=Very High)
27. **StandardHours**: Standard working hour per **week?** (80 for everyone)
28. **StockOptionLevel**: Stock option level
29. **TotalWorkingYears**: Total years of working
30. **TrainingTimesLastYear**: Training hours of the last year
31. **WorkLifeBalance**: the amount of time you spend doing your job compared with the amount of time you spend with your family and doing things you enjoy (1=Bad,2=Good, 3=Better, 4=Best)
32. **YearsAtCompany**: Total years of working at the company
33. **YearsInCurrentRole**: Total years spent in the current position
34. **YearsSinceLastPromotion**: How many year ago the employee had the last promotion
35. **YearsWithCurrManager**: How many years the employee is with the actual manager
36. **City**: where the employee works (London, Barcelona)

4 Cleaning Data

4.1 Names

In this sub-point we are going to change the names of the variables in order to have all he names of the variables with the same layout.

```
mydb <- mydb %>% clean_names(., "snake")
```

4.2 Dimensions

First of all, we are going to check the actual dimension of our dataset. Hence, from the following code we can understand that there are 36 variables in total and

```
mydb %>% dim()
```

```
## [1] 1506   36
```

```
mydb %>% nrow()
```

```
## [1] 1506
```

```
mydb %>% ncol()
```

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

4.3 Head and Tail

Here, we are going to check the first 10 elements at the beginning and at the end of the dataset. Consecutively, we are going to check the top and the bottom values of the main relevant variables to catch some errors.

```
mydb %>% head(10)
mydb %>% tail(10)
mydb<-rename(mydb, age =i_age)
mydb %>% arrange(desc(age)) %>% top_n(10, age)
mydb %>% arrange(age) %>% top_n(-10, age)
```

4.4 Removing

In this part of the data cleaning we are going to remove all the blank rows, the duplicates and strange values that may affect our analysis.

```
# Remove blank rows and columsn
mydb <- mydb %>% remove_empty(c("rows", "cols"))

# Removing entries with too high and too low age
mydb <- mydb %>% filter(age <= 80 & age >= 16)
mydb <- mydb %>% filter(job_involvement <= 4)
mydb <- mydb %>% filter(num_companies_worked >= 0)
```

Therefore, as we can see, this line of code did not affected our dataset. So, this mean that there are no rows or columns that are empty.

Now, we are going to pass to the study of duplicates, by the *employee_number* variable that we suggest it is the key.

```
# Duplicates removal
mydb %>% get_dupes(employee_number)
mydb <- mydb %>% distinct(employee_number, .keep_all= TRUE)
```

4.5 Checking n's

Hence, now it is time to check the n's.

4.6 Droping na

To conclude, the cleaning of the dataset, we are going to remove every line with at least one empy gap.

```
mydb <- mydb %>% drop_na()
```

From know on we can easiliy proceed with our analysis.

5 Analysis

Our analysis consists in different parts.

1. First of all, we want to describe the gender equality inside the company and understand if there are discrepancies and how we can solve those problems.
2. Secondly, we want to understand the attrition factor and have a clear comprehension of what is the structure of the officies between Barcelona and London. This would be very helpful in otder to decrease the percentace of people who want to leave and how the company can improve the level of comfort of its employees.
3. Thirdly, we think is very important to describe the attitude of our employees and what it their satisfaction level. So, combining this solution with the previous question we can improve our decision and suggest to the company where it can adopt new HR methodologies.

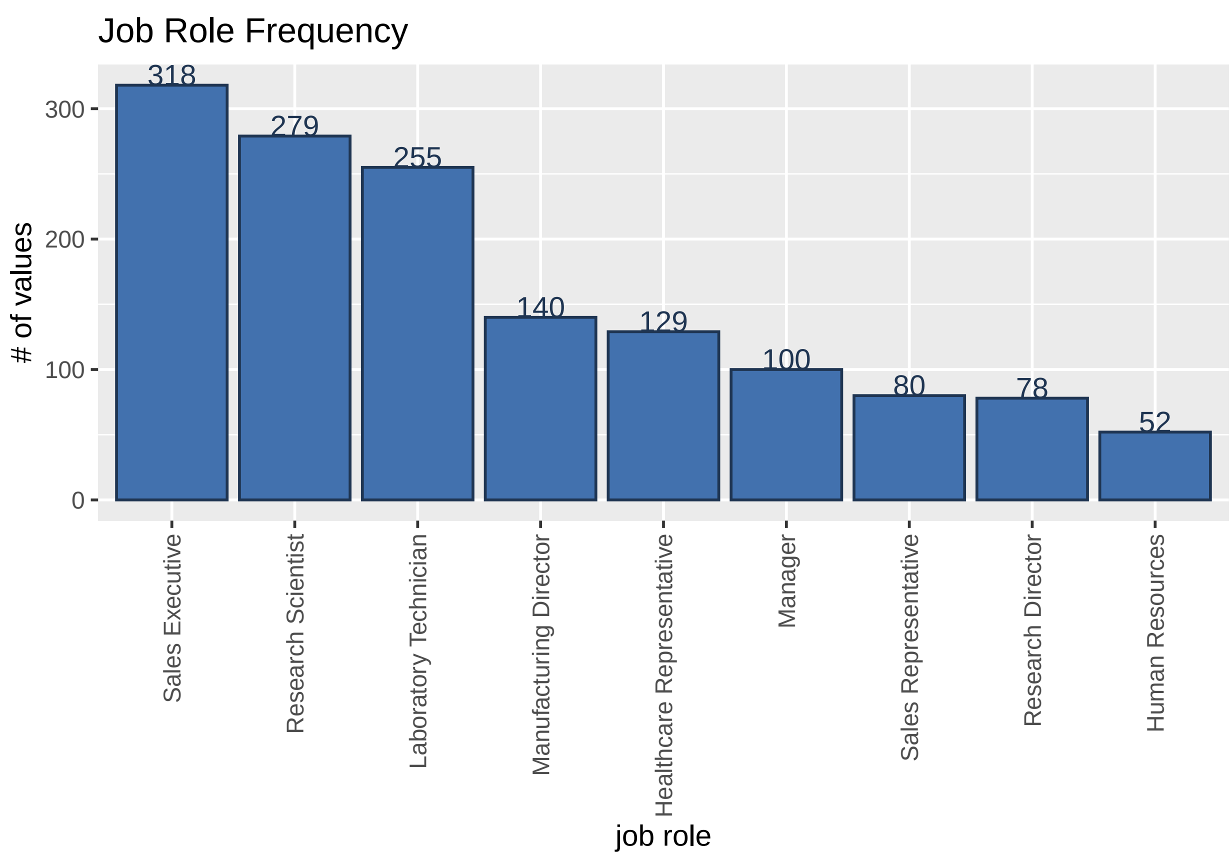
4. Fourthly, we want to understand how the company can find its successor and what could be the best solution.

5. To conclude, there is also the good practice to import in our project new datasets that could improve the decions that we can take.

5.1 Introduction Analysis

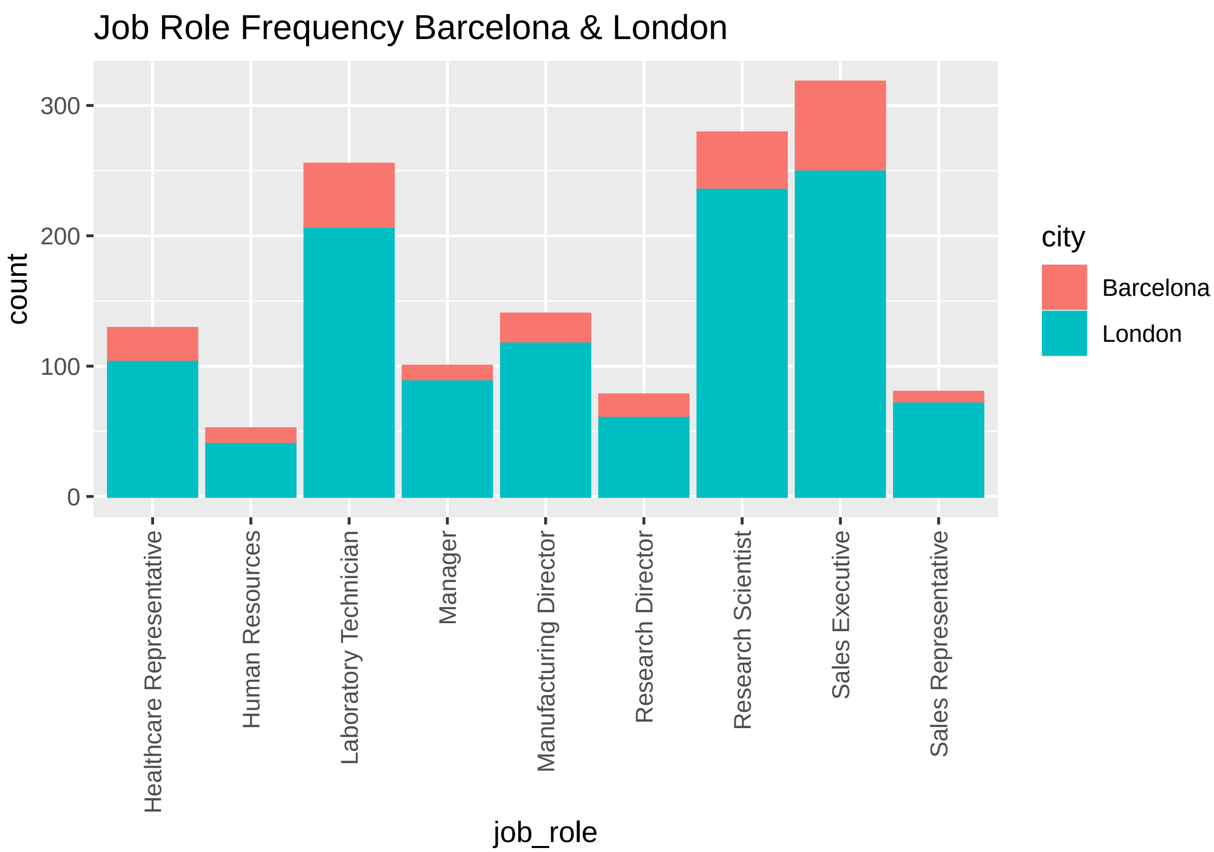
5.1.1 Plot job role frequency

```
# This would help to analyse the number of types of jobs and compare with types
k <- mydb %>% tabyl(job_role)

ggplot(k, aes(x = reorder(job_role, -n), y = n)) +
  geom_bar(stat = "identity", fill = "#4271AE", colour = "#1F3552") +
  theme(axis.text.x=element_text(angle=90,hjust=1,vjust=0.5)) +
  geom_text(aes(label = n), vjust = 0, colour = "#1F3552")+
  labs(y= "# of values", x = "job role") +
  ggtitle("Job Role Frequency")
```




```
rm("k")
```

```
ggplot(mydb, aes(x = job_role, color = city, fill = city)) +
  geom_bar(stat = "count") +
  theme(axis.text.x=element_text(angle=90,hjust=1,vjust=0.5)) +
  ggtitle("Job Role Frequency Barcelona & London")
```



5.2 Gender Analysis

First of all, we want to understand the salary that sex has. This would give us a better overview on how the salary is distributed inside the company. Moreover, we will also take into account the possibile differences that we have in the Barcelona and London headwarters.

5.2.1 Monthly Income

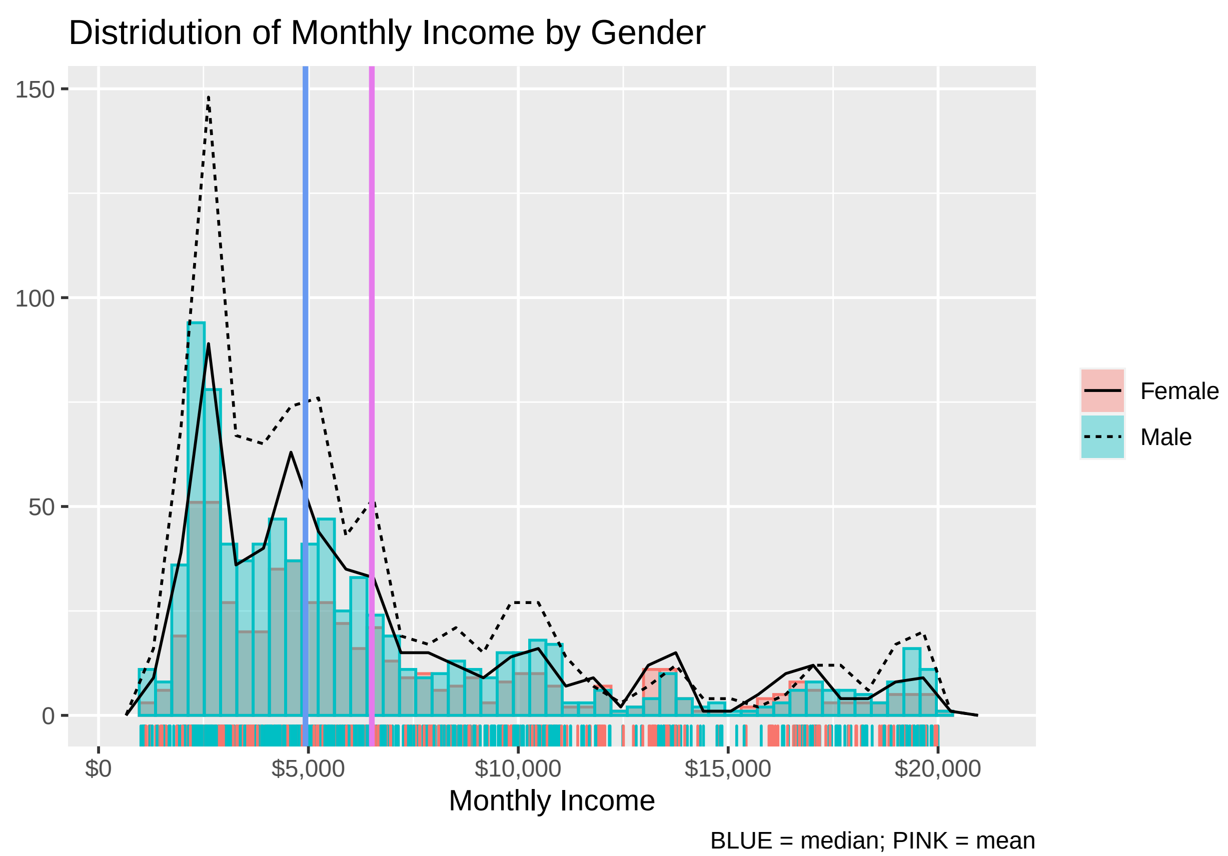
It is pretty clear that the monthly income could be one of the most important variable that will determine the differences inside a compnay taking into accout the gender analysis.

Introduction

```
summary(mydb$monthly_income)
```

```
##     Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     1009    2909    4930    6510    8386   19999
```

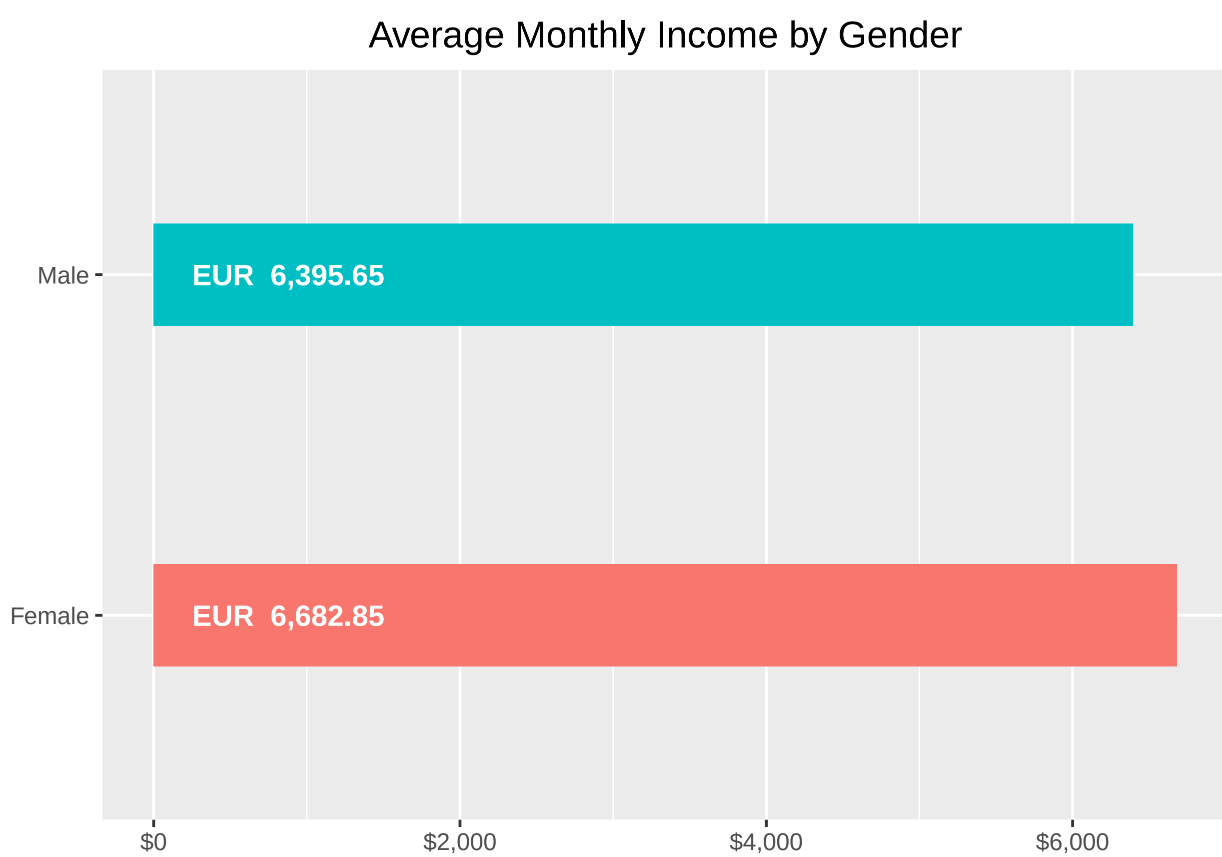
```
mydb %>% select(monthly_income, gender) %>%
  ggplot(aes(monthly_income)) +
  geom_histogram(
    aes(monthly_income, color = gender, fill = gender),
    alpha = 0.4,
    position = "identity",
    bins = 50,
  ) +
  geom_freqpoly(aes(linetype = gender), bins = 30) +
  geom_rug(aes(color = gender)) +
  scale_x_continuous(labels = label_dollar()) +
  geom_vline(aes(xintercept = mean(monthly_income)), color = "#E67AEC", size = 1) +
  geom_vline(aes(xintercept = median(monthly_income)), color = "#6899F1", size = 1) +
  guides(color = "none") +
  labs(title = "Distridution of Monthly Income by Gender",
       caption = "BLUE = median; PINK = mean",
       x = "Monthly Income",
       y = NULL,
       fill = NULL,
       linetype = NULL)
```



Hence, from this graph we can see that most of the employee have a salary lower than the average. In ddition, we can also see that generally we have the dotted male line almost always above the female line, but this can be considered acceptable, due to the fact that in our dataset we have more men and women. To conclude, we can also see that the major discrepancies are given from the lower range of salary, while the higher is the salary, the lower are the differences between the female and male monthly income.

Now, we will take a closer look to the monthly income between male an female . So, we can observe that the male have a worse results. While, the female has in medium an higher salray than man.

```
mydb %>%
  select(gender, monthly_income) %>%
  group_by(gender) %>%
  summarise(avg_income = round(mean(monthly_income), 2), .groups = "drop") %>%
  ggplot(aes(x = gender, y = avg_income)) +
  geom_col(aes(fill = gender), width = 0.3, show.legend = FALSE) +
  geom_text(
    aes(
      x = gender,
      y = 0.01,
      label = dollar(avg_income, prefix = "EUR  ")
    ),
    hjust = -0.2,
    size = 4,
    colour = "white",
    fontface = "bold"
  ) +
  coord_flip() +
  scale_y_continuous(labels = label_dollar()) +
  theme(plot.title = element_text(size = 14, hjust = 0.5)) +
  labs(title = "Average Monthly Income by Gender",
       x = NULL,
       y = NULL)
```

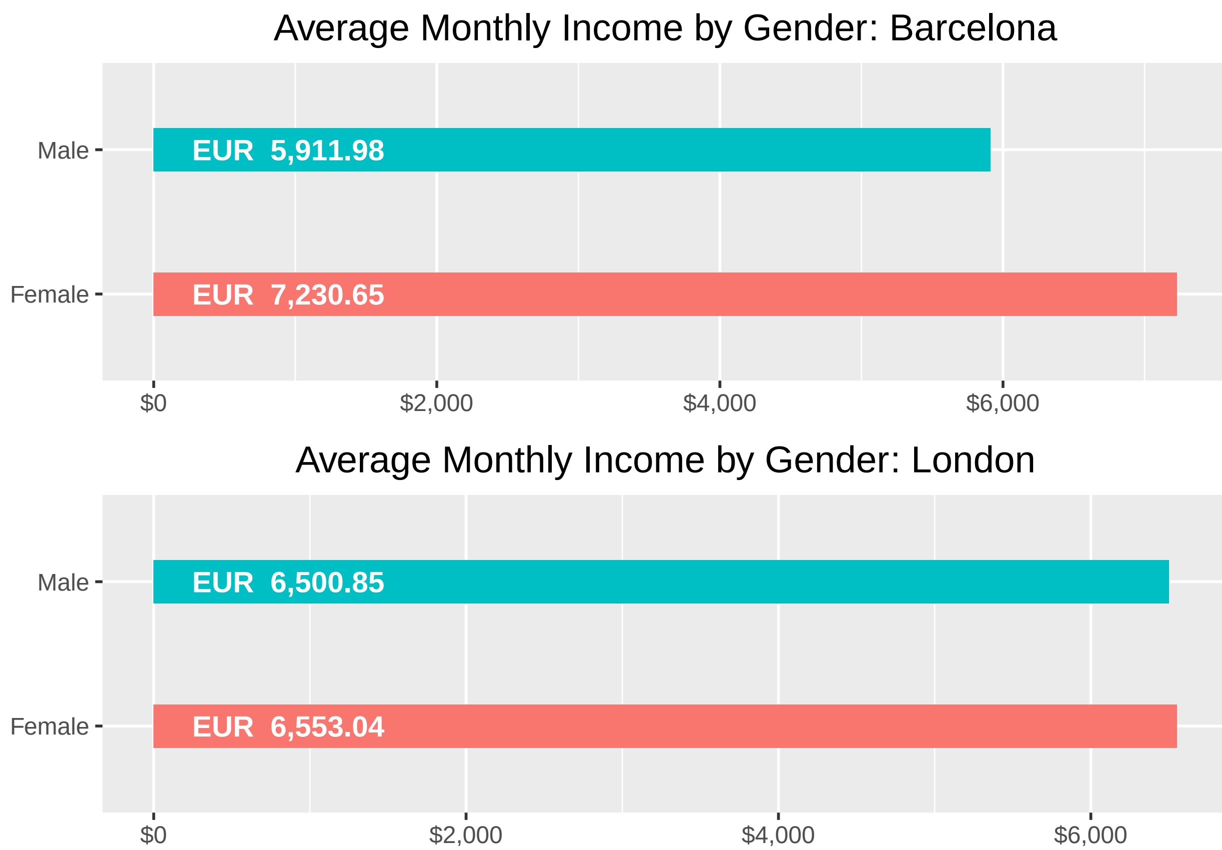


Clearly, this difference is so minimal that we need to go further in detail to clearly understand if we can adopt any strategy to balance the gender inside a company or not.

So, due to we also need to understan where is located this difference, we will filter by country.

```
# Barcelona
g1 <- mydb %>% filter(city == "Barcelona") %>%
  select(gender, monthly_income) %>%
  group_by(gender) %>%
  summarise(avg_income = round(mean(monthly_income), 2), .groups = "drop") %>%
  ggplot(aes(x = gender, y = avg_income)) +
  geom_col(aes(fill = gender), width = 0.3, show.legend = FALSE) +
  geom_text(
    aes(
      x = gender,
      y = 0.01,
      label = dollar(avg_income, prefix = "EUR  ")
    ),
    hjust = -0.2,
    size = 4,
    colour = "white",
    fontface = "bold"
  ) +
  coord_flip() +
  scale_y_continuous(labels = label_dollar()) +
  theme(plot.title = element_text(size = 14, hjust = 0.5)) +
  labs(title = "Average Monthly Income by Gender: Barcelona",
       x = NULL,
       y = NULL)


# London
g2 <- mydb %>% filter(city == "London") %>%
  select(gender, monthly_income) %>%
  group_by(gender) %>%
  summarise(avg_income = round(mean(monthly_income), 2), .groups = "drop") %>%
  ggplot(aes(x = gender, y = avg_income)) +
  geom_col(aes(fill = gender), width = 0.3, show.legend = FALSE) +
  geom_text(
    aes(
      x = gender,
      y = 0.01,
      label = dollar(avg_income, prefix = "EUR  ")
    ),
    hjust = -0.2,
    size = 4,
    colour = "white",
    fontface = "bold"
  ) +
  coord_flip() +
  scale_y_continuous(labels = label_dollar()) +
  theme(plot.title = element_text(size = 14, hjust = 0.5)) +
  labs(title = "Average Monthly Income by Gender: London",
       x = NULL,
       y = NULL)
```

```
grid.arrange(g1, g2, nrow = 2)
```



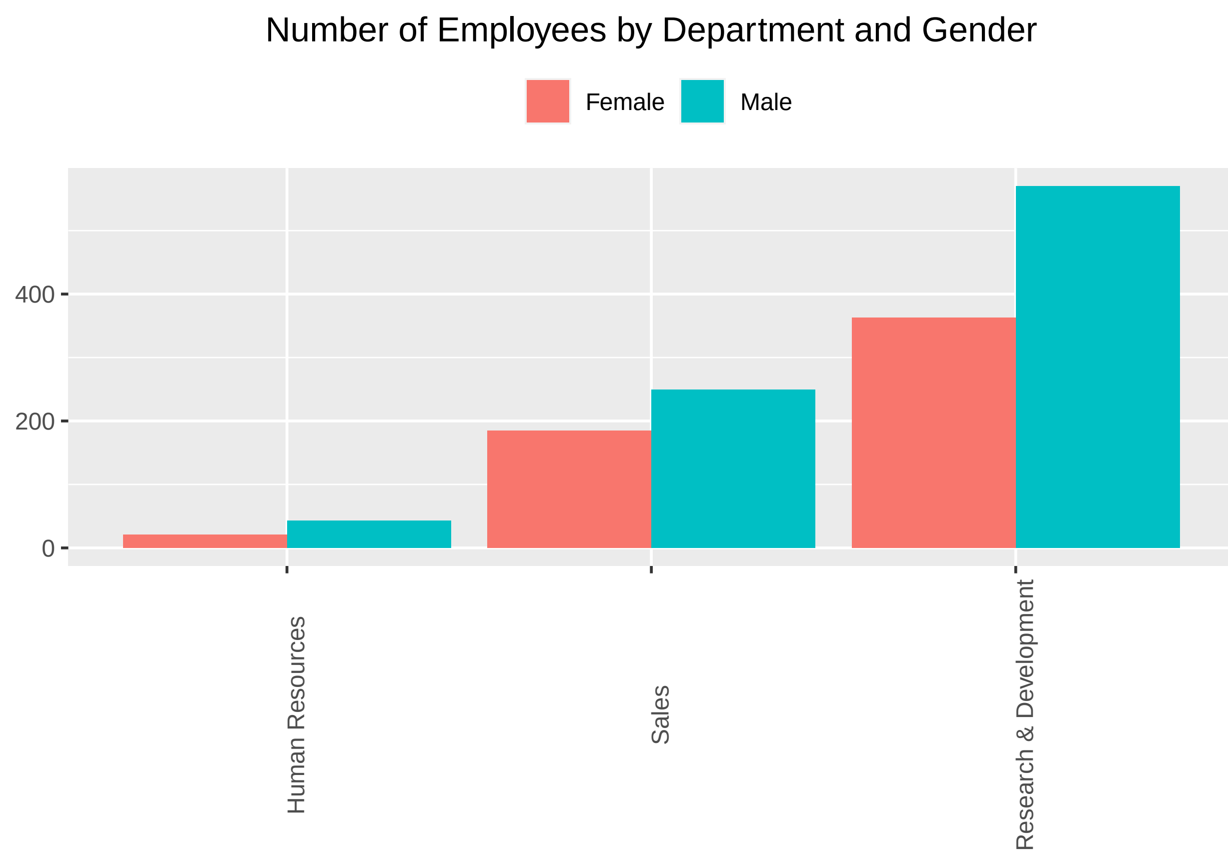

```
rm(g1, g2)
```

In this differentiation is definetely more clear how in Barcelona the monthly salary seems to privilege the female gender despite the male one. While, in London even if women has slightly higher average monthly salary, this is so small that can be omitted. In conclusion, we can focus to Barcelona and try to identify here the causes of this differences.

5.2.2 Department

```
mydb %>%
  group_by(department, gender) %>%
  summarise(amount = n(), .groups = "drop") %>%
  ggplot(aes(
    x = fct_reorder(department, amount),
    y = amount,
    fill = gender
  )) +
  geom_col(position = "dodge") +
  theme(axis.text.x = element_text(angle = 90),
        plot.title = element_text(hjust = 0.5),
```

```
        legend.position = "top") +
  labs(title = "Number of Employees by Department and Gender",
x = NULL,
y = NULL,
fill = NULL)
```

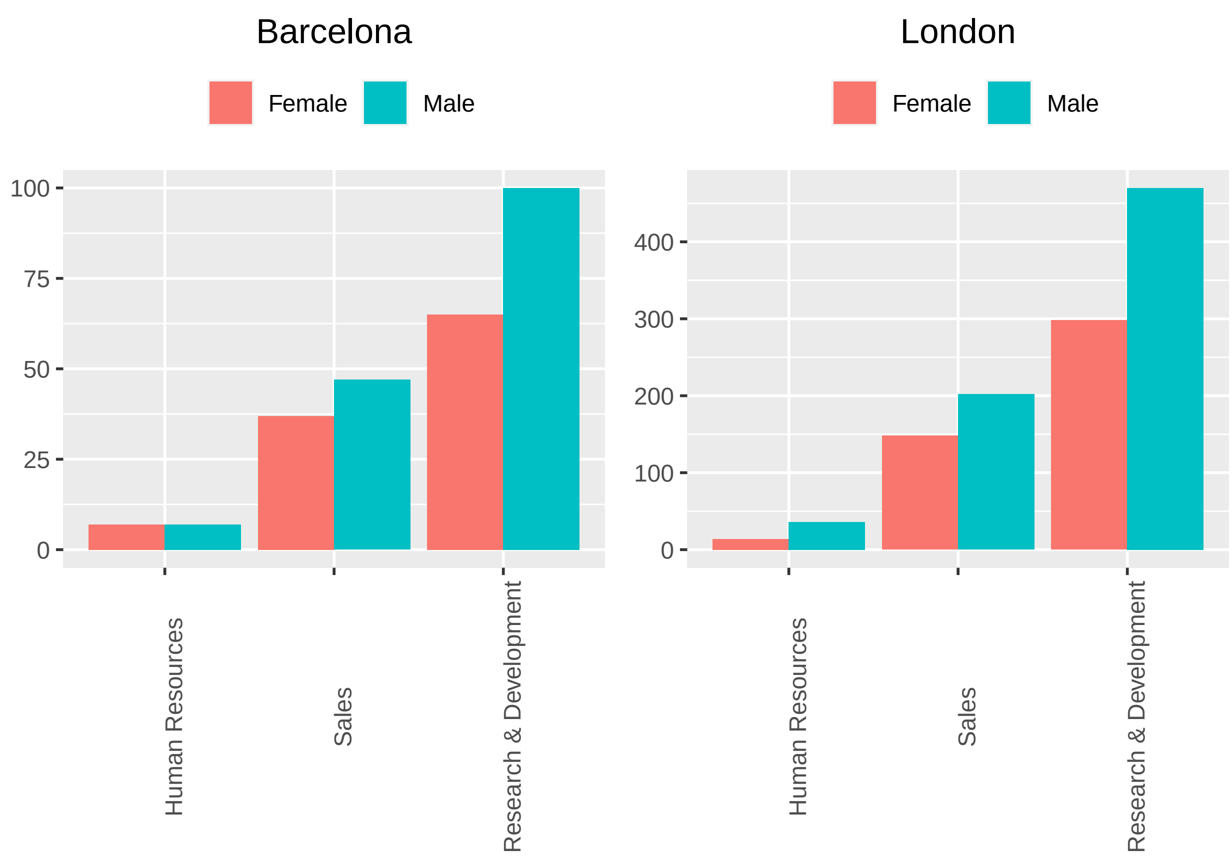


Therefore, in this graph we can see how inside the company we have the majority of employees in the field of research and development. Moreover, also here we can see that in general the male sex has an higher frequency in each department compared to to female one.

Also here we found pretty important to stat that the differences between Barcelona and London to decide ultimately where to take decisions.

```
# Barcelona
g1 <- mydb %>% filter(city == "Barcelona") %>%
  group_by(department, gender) %>%
  summarise(amount = n(), .groups = "drop") %>%
  ggplot(aes(
    x = fct_reorder(department, amount),
    y = amount,
    fill = gender
  )) +
  geom_col(position = "dodge") +
  theme(axis.text.x = element_text(angle = 90),
        plot.title = element_text(hjust = 0.5),
        legend.position = "top") +
  labs(title = "Barcelona",
```

```
x = NULL,
y = NULL,
fill = NULL)

# London
g2 <- mydb %>% filter(city == "London") %>%
  group_by(department, gender) %>%
  summarise(amount = n(), .groups = "drop") %>%
  ggplot(aes(
    x = fct_reorder(department, amount),
    y = amount,
    fill = gender
  )) +
  geom_col(position = "dodge") +
  theme(axis.text.x = element_text(angle = 90),
        plot.title = element_text(hjust = 0.5),
        legend.position = "top") +
  labs(title = "London",
x = NULL,
y = NULL,
fill = NULL)

grid.arrange(g1, g2, nrow = 1)
```



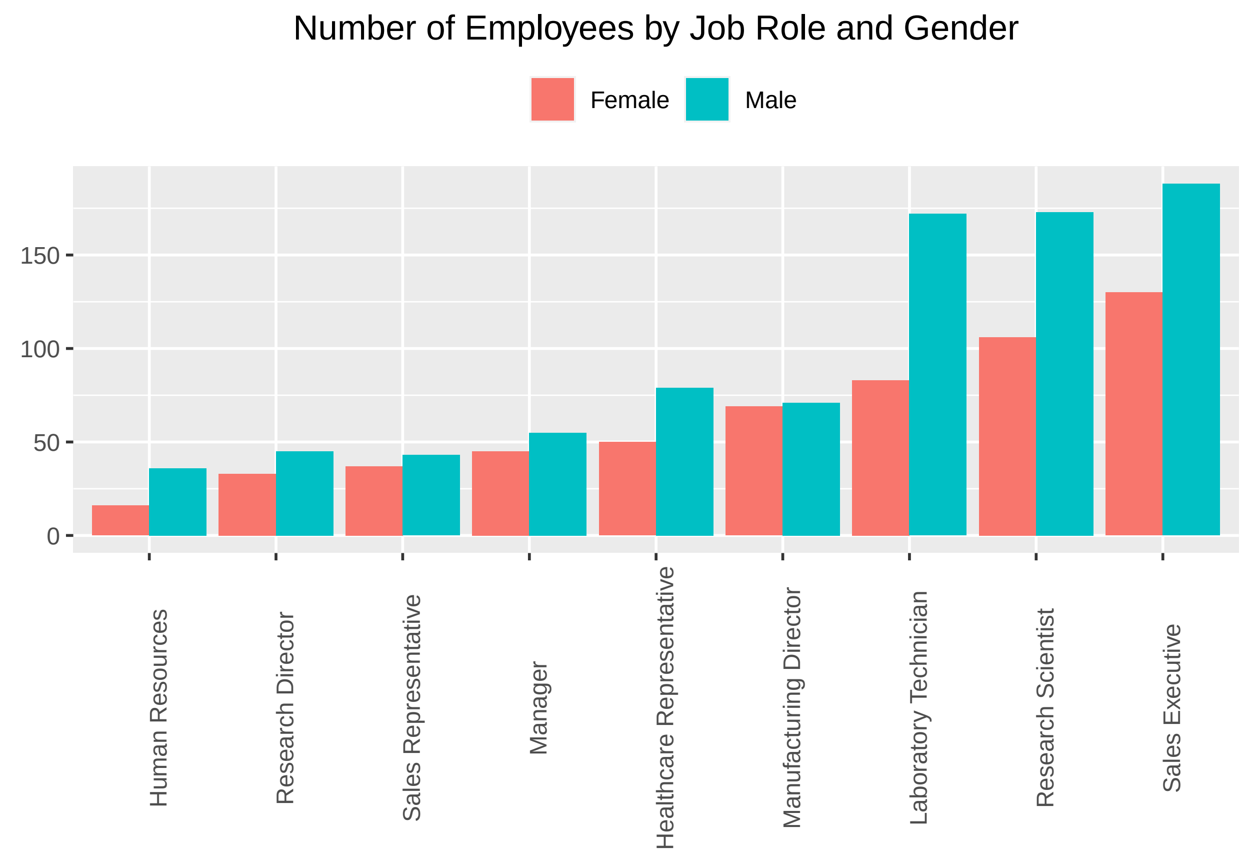
```
rm(g1, g2)
```

Now, we can start seeing that the major differences about the number of employees and their distribution is mainly in London, while in Barcelona for instance, In the HR department we have almost the same number of employee per gender. This balance inside the company is a good sign of gender equality, even if there are some department, such as sales and R&D where the male sex is predominant In both London and Barcelona.

5.2.3 Job Role

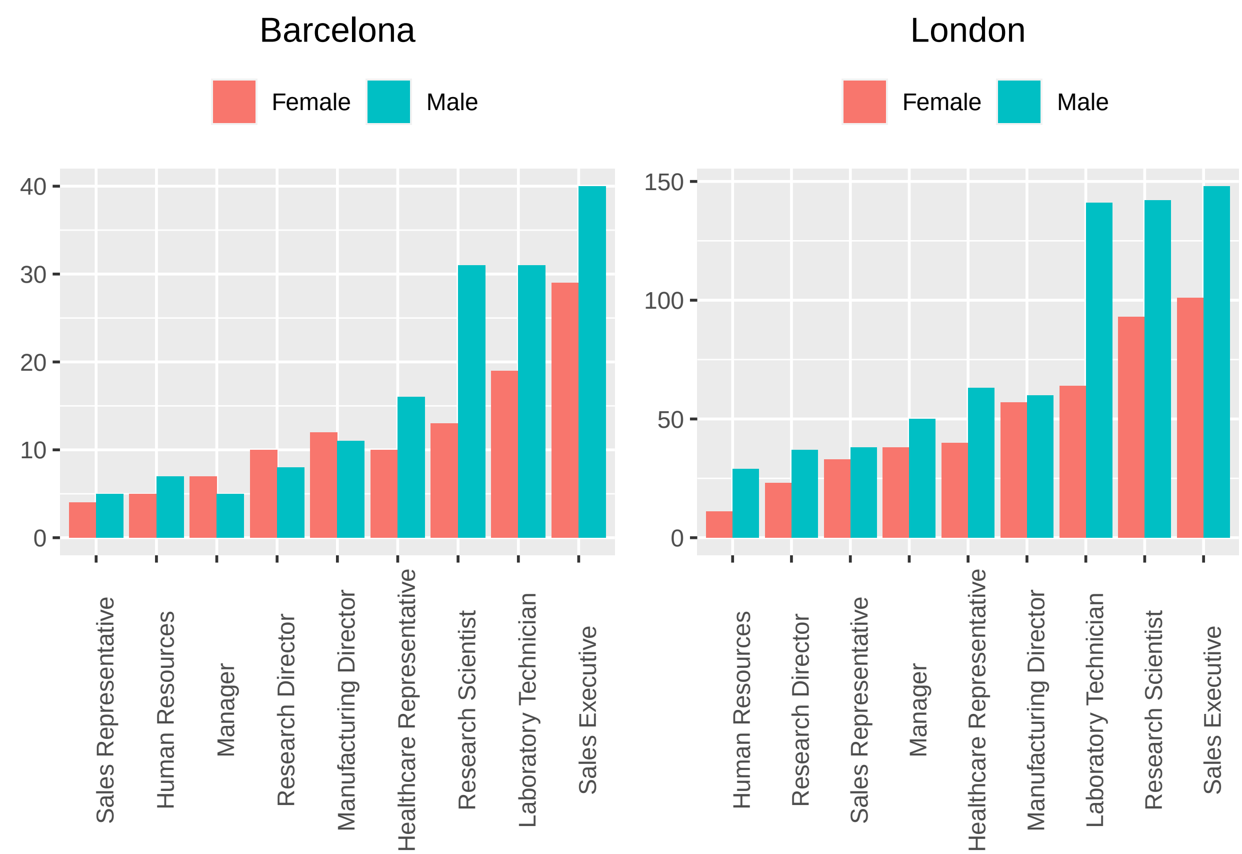
The same analysis as above is represented below taking into account the *job_role* variable.

```
mydb %>%
  group_by(job_role, gender) %>%
  summarise(amount = n(), .groups = "drop") %>%
  ggplot(aes(
    x = fct_reorder(job_role, amount),
    y = amount,
    fill = gender
  )) +
  geom_col(position = "dodge") +
  theme(axis.text.x = element_text(angle = 90),
        plot.title = element_text(hjust = 0.5),
        legend.position = "top") +
  labs(title = "Number of Employees by Job Role and Gender",
x = NULL,
y = NULL,
fill = NULL)
```




```
# Barcelona
g1 <- mydb %>% filter(city == "Barcelona") %>%
  group_by(job_role, gender) %>%
  summarise(amount = n(), .groups = "drop") %>%
  ggplot(aes(
    x = fct_reorder(job_role, amount),
    y = amount,
    fill = gender
  )) +
  geom_col(position = "dodge") +
  theme(axis.text.x = element_text(angle = 90),
        plot.title = element_text(hjust = 0.5),
        legend.position = "top") +
  labs(title = "Barcelona",
x = NULL,
y = NULL,
fill = NULL)

# London
g2 <- mydb %>% filter(city == "London") %>%
  group_by(job_role, gender) %>%
  summarise(amount = n(), .groups = "drop") %>%
  ggplot(aes(
    x = fct_reorder(job_role, amount),
    y = amount,
    fill = gender
```

```
  )) +
  geom_col(position = "dodge") +
  theme(axis.text.x = element_text(angle = 90),
        plot.title = element_text(hjust = 0.5),
        legend.position = "top") +
  labs(title = "London",
x = NULL,
y = NULL,
fill = NULL)

grid.arrange(g1, g2, nrow = 1)
```




```
rm(g1, g2)
```

This two graphs seems to be relevant in some job roles. In fact, as we can see in Barcelona, women play an essential role in the two Director job roles.

5.2.4 Seniority

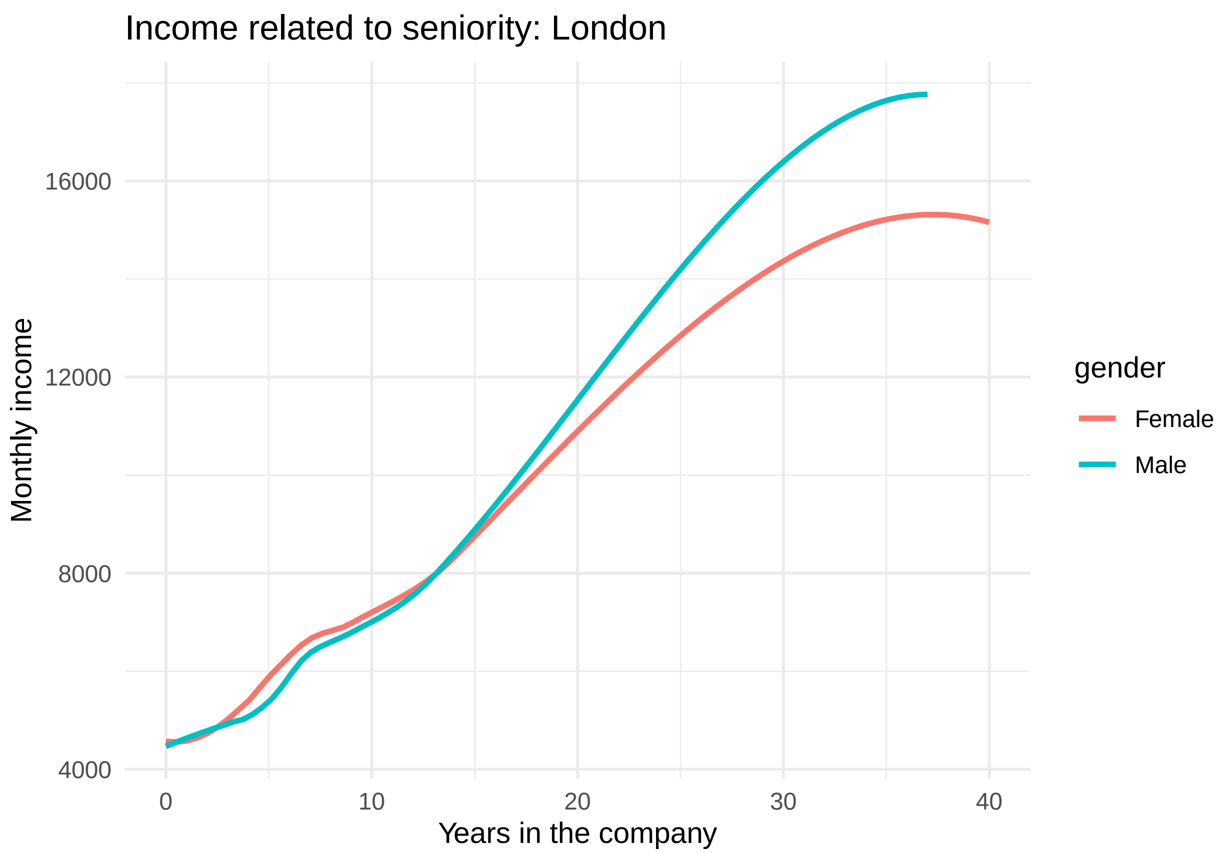
Taking in consideration both company location In this graphs in order to show that up in a better way, we decided to remove the confidence interval and the points.

```
ggplot(mydb, aes(x=years_at_company, y=monthly_income, colour=gender))+
  geom_smooth(method = 'loess', formula = y ~ x, se=F) +
  labs(title = "Income related to seniority to all company location",
       x = "Years in the company",
       y = "Monthly income") +
  theme_minimal()
```

So, as we can see here, generally speaking the female gender has an higher salary compare to the male gender. This is true till approximately the 20th year of seniority inside the company where the male gender will overcome the female monthly income in a dramatic way.

Taking in consideration London

```
ggplot(mydb%>%filter(city=="London"),
       aes(x=years_at_company, y=monthly_income, colour=gender)) +
  geom_smooth(method = 'loess', formula = y ~ x, se=F) + labs(
      title = "Income related to seniority: London",
       x = "Years in the company",
       y = "Monthly income"
    ) +
  theme_minimal()
```



In the specific case of London we can stat that generally speaking the monthly income is well balanced between male and female work positions, but around the 15th year of seniority the male gender will have a boost in therm of monthly income.

Taking in consideration Barcelona

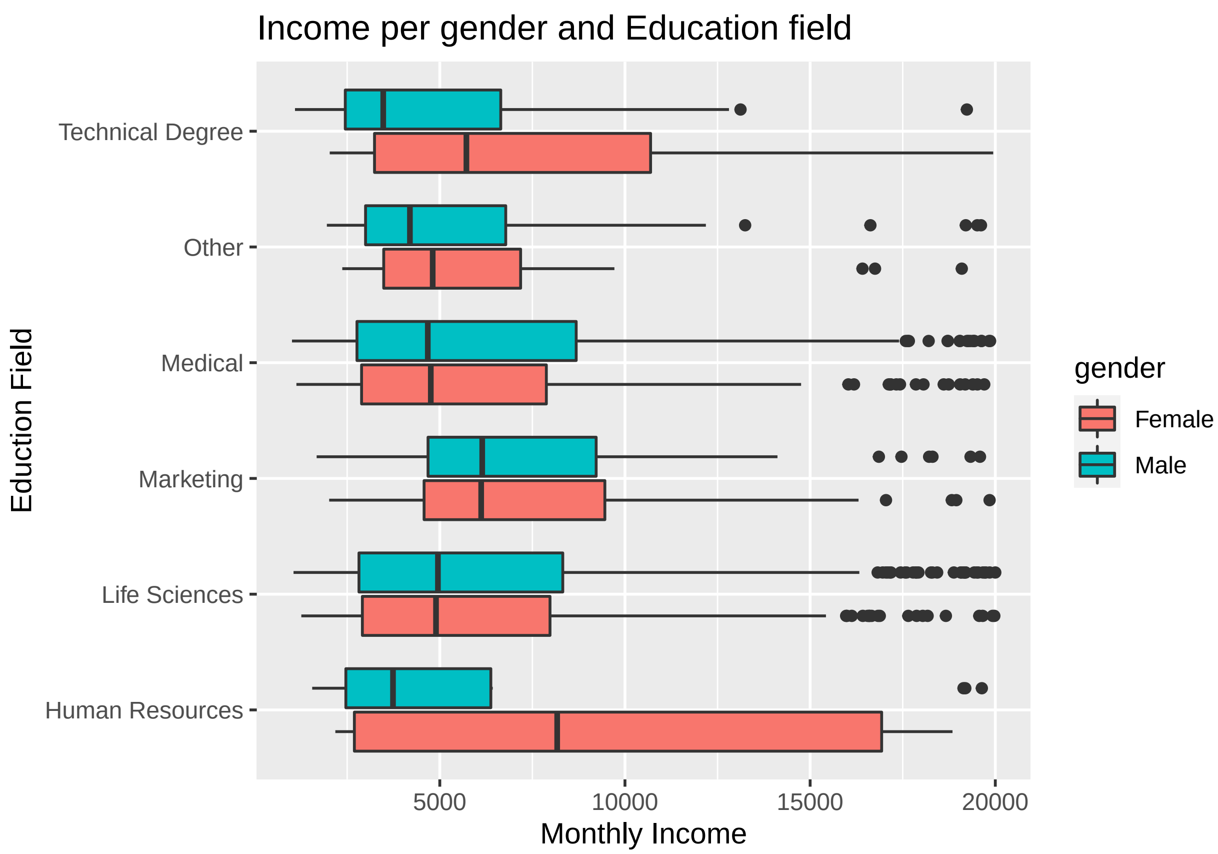
```
ggplot(mydb%>%filter(city=="Barcelona"),
       aes(x=years_at_company, y=monthly_income, colour=gender)) +
  geom_smooth(method = 'loess', formula = y ~ x, se=F) + labs(
      title = "Income related to seniority: Barcelona",
       x = "Years in the company",
       y = "Monthly income"
    ) +
  theme_minimal()
```

Now, in Barcelona we can see hat the resoult we obtained is pretty different than before. In fact, the female gender has generally an higher incoem compared to the male gender.

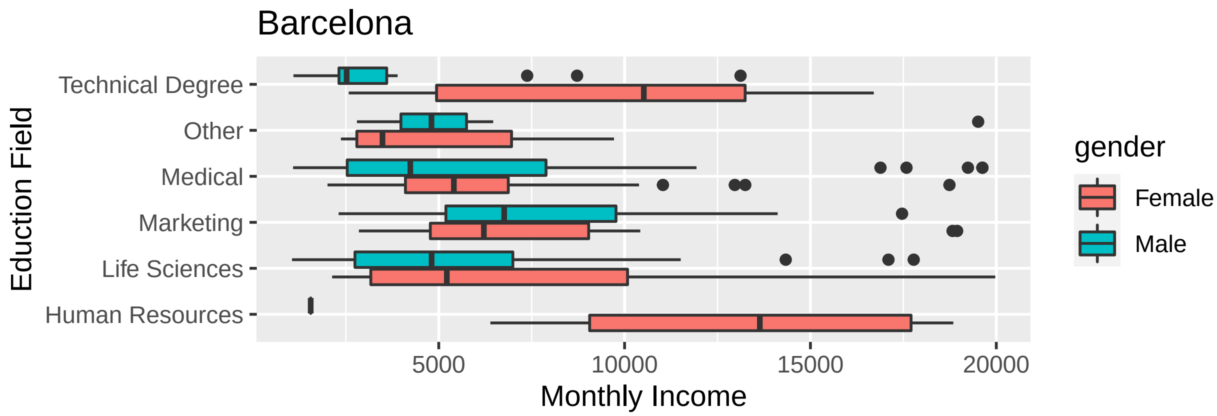
5.2.5 Education field

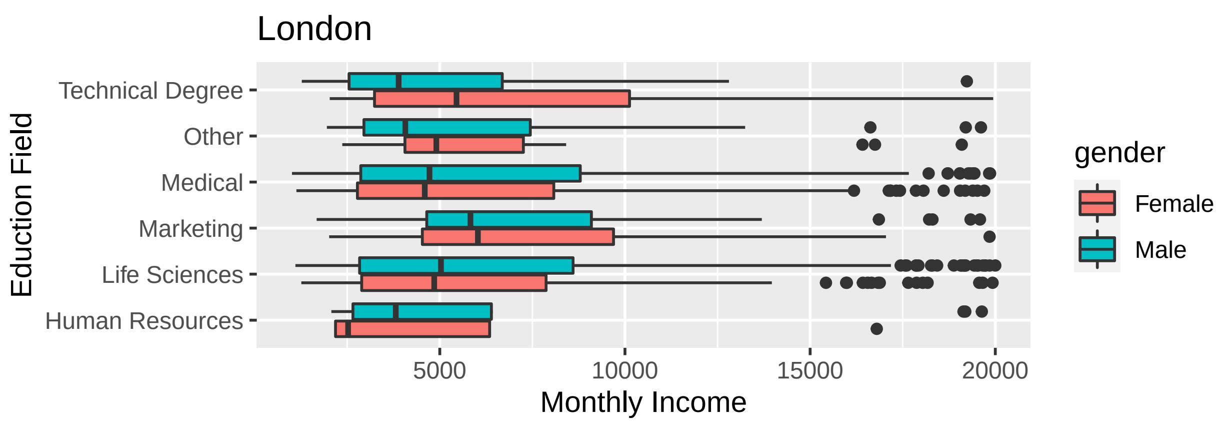
Performance rating education field.

```
ggplot(mydb, aes(y= education_field, x = monthly_income, fill = gender))+
  labs(y = "Eduction Field", x = "Monthly Income", title = "Income per gender and Education field") +
  geom_boxplot()
```




```
# Barcelona
g1 <- ggplot(mydb%>%filter(city == "Barcelona"), aes(y= education_field, x = monthly_income, fill = gend
  labs(y = "Eduction Field", x = "Monthly Income", title = "Barcelona") +
  geom_boxplot()

# London
g2 <- ggplot(mydb%>%filter(city == "London"), aes(y= education_field, x = monthly_income, fill = gender)
  labs(y = "Eduction Field", x = "Monthly Income", title = "London") +
  geom_boxplot()

grid.arrange(g1, g2, nrow = 2)
```






```
rm(g1, g2)
```

So, in the first graph we can see that there are some differences. Hence we decided to locate them and understood that the major problems are in Barcelona. In fact, in the HR sector, women have an higher salary than men, also for technical degree we can see how the female gender seems to have a clear advantage on that.

Clearly, per education field we can see also that in some cases also the male gender seems to have a slightly higer salary, but the main focus are those from which the discrepancy are very high, such as HR in Barcelona and Technical Degree, always in Barcelona.

5.2.6 Regression

To conclude, we wanted also to see if the gender is a relevan variable that affect *monthly_income* or other variables relavant for our study.

```
mod1 <- lm(monthly_income ~ ., data = mydb)
summary(mod1)
```

```
##
## Call:
## lm(formula = monthly_income ~ ., data = mydb)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -3908.5  -708.4    -2.6   671.1  4386.4
##
## Coefficients: (2 not defined because of singularities)
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      3.470e+02  7.669e+02   0.453 0.650964
## age                             -5.641e+00  4.709e+00  -1.198 0.231111
## attritionYes                     3.134e+01  9.423e+01   0.333 0.739467
## business_travelTravel_Frequently  1.522e+02  1.179e+02   1.291 0.196892
## business_travelTravel_Rarely     1.809e+02  1.007e+02   1.797 0.072541 .
## daily_rate                       8.238e-02  7.538e-02   1.093 0.274661
## departmentResearch & Development  3.289e+02  3.880e+02   0.848 0.396800
## departmentSales                  1.243e+02  4.096e+02   0.303 0.761617
## distance_from_home              -4.345e+00  3.731e+00  -1.165 0.244375
## education                       -1.657e+01  3.035e+01  -0.546 0.585303
## education_fieldLife Sciences    -1.192e+02  2.915e+02  -0.409 0.682715
## education_fieldMarketing        -2.094e+01  3.114e+02  -0.067 0.946408
## education_fieldMedical          -1.418e+02  2.928e+02  -0.484 0.628357
## education_fieldOther            -2.207e+02  3.142e+02  -0.702 0.482566
## education_fieldTechnical Degree -4.758e+01  3.056e+02  -0.156 0.876305
## employee_count                          NA         NA      NA       NA
## employee_number                  7.576e-02  5.037e-02   1.504 0.132803
## environment_satisfaction        -3.513e+00  2.804e+01  -0.125 0.900322
## genderMale                       1.033e+02  6.209e+01   1.664 0.096398 .
## hourly_rate                      1.065e+00  1.492e+00   0.714 0.475601
## job_involvement                 -9.031e+01  4.298e+01  -2.101 0.035832 *
## job_level                        2.758e+03  6.892e+01  40.018  < 2e-16 ***
## job_roleHuman Resources         -4.542e+01  4.133e+02  -0.110 0.912508
## job_roleLaboratory Technician   -5.865e+02  1.413e+02  -4.150 3.53e-05 ***
## job_roleManager                  4.247e+03  2.105e+02  20.177  < 2e-16 ***
## job_roleManufacturing Director  -4.501e+01  1.394e+02  -0.323 0.746829
## job_roleResearch Director        4.072e+03  1.846e+02  22.064  < 2e-16 ***
## job_roleResearch Scientist      -5.202e+02  1.399e+02  -3.718 0.000209 ***
## job_roleSales Executive          1.204e+02  2.746e+02   0.439 0.661048
## job_roleSales Representative    -5.310e+02  3.059e+02  -1.736 0.082827 .
## job_satisfaction                 8.567e-01  2.776e+01   0.031 0.975385
## marital_statusMarried            1.722e+01  8.113e+01   0.212 0.831897
## marital_statusSingle            -3.919e+01  1.119e+02  -0.350 0.726291
## monthly_rate                    -4.583e-03  4.246e-03  -1.079 0.280650
## num_companies_worked             1.128e+01  1.361e+01   0.829 0.407411
## over18Y                         -2.533e+01  4.363e+02  -0.058 0.953715
## over_timeYes                     7.032e+01  7.004e+01   1.004 0.315615
## percent_salary_hike              1.746e+01  1.306e+01   1.337 0.181600
## performance_rating              -1.627e+02  1.316e+02  -1.237 0.216350
## relationship_satisfaction        2.005e+01  2.817e+01   0.712 0.476750
## standard_hours                          NA         NA      NA       NA
## stock_option_level              -4.319e+01  4.853e+01  -0.890 0.373648
## total_working_years              4.752e+01  8.490e+00   5.598 2.62e-08 ***
## training_times_last_year        -1.715e+01  2.369e+01  -0.724 0.469296
## work_life_balance               -1.562e+01  4.295e+01  -0.364 0.716218
## years_at_company                 3.395e+00  1.057e+01   0.321 0.748023
## years_in_current_role            5.206e+00  1.356e+01   0.384 0.701107
## years_since_last_promotion       2.366e+01  1.219e+01   1.940 0.052544 .
## years_with_curr_manager         -3.138e+01  1.416e+01  -2.216 0.026842 *
## cityLondon                      -1.143e+01  7.823e+01  -0.146 0.883837
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1129 on 1383 degrees of freedom
## Multiple R-squared:  0.9445, Adjusted R-squared:  0.9426
## F-statistic: 500.8 on 47 and 1383 DF,  p-value: < 2.2e-16
```

```
mod2 <- lm(monthly_income ~ gender, data = mydb)
summary(mod2)
```

```
##
## Call:
## lm(formula = monthly_income ~ gender, data = mydb)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -5554  -3586  -1555   1924  13603
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   6682.8      197.5  33.837   <2e-16 ***
## genderMale    -287.2      254.5  -1.129    0.259
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4711 on 1429 degrees of freedom
## Multiple R-squared:  0.0008906,  Adjusted R-squared:  0.0001914
## F-statistic: 1.274 on 1 and 1429 DF,  p-value: 0.2593
```

```
names(mydb)
```

```
##  [1] "age"                        "attrition"
##  [3] "business_travel"            "daily_rate"
##  [5] "department"                 "distance_from_home"
##  [7] "education"                  "education_field"
##  [9] "employee_count"             "employee_number"
## [11] "environment_satisfaction"   "gender"
## [13] "hourly_rate"                "job_involvement"
## [15] "job_level"                  "job_role"
## [17] "job_satisfaction"           "marital_status"
## [19] "monthly_income"             "monthly_rate"
## [21] "num_companies_worked"       "over18"
## [23] "over_time"                  "percent_salary_hike"
## [25] "performance_rating"         "relationship_satisfaction"
## [27] "standard_hours"             "stock_option_level"
## [29] "total_working_years"        "training_times_last_year"
## [31] "work_life_balance"          "years_at_company"
## [33] "years_in_current_role"      "years_since_last_promotion"
## [35] "years_with_curr_manager"    "city"
```

```
mod3 <- lm(monthly_income ~ age + gender + business_travel + education_field + education + distance_fro
summary(mod3)
```

```
##
## Call:
## lm(formula = monthly_income ~ age + gender + business_travel +
##     education_field + education + distance_from_home + job_level +
##     job_role + marital_status + total_working_years + performance_rating +
##     years_at_company, data = mydb)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3891.7  -705.5     2.1   660.6  4258.4
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       193.964    450.615   0.430 0.666940
## age                                -4.646      4.638  -1.002 0.316737
## genderMale                        104.622     61.651   1.697 0.089917 .
## business_travelTravel_Frequently  140.283    116.457   1.205 0.228563
## business_travelTravel_Rarely      166.238    100.023   1.662 0.096735 .
## education_fieldLife Sciences      -44.626    268.716  -0.166 0.868124
## education_fieldMarketing           30.145    287.329   0.105 0.916458
## education_fieldMedical            -57.443    269.850  -0.213 0.831460
## education_fieldOther             -132.605    291.814  -0.454 0.649599
## education_fieldTechnical Degree    31.189    283.180   0.110 0.912314
## education                         -17.829     30.166  -0.591 0.554597
## distance_from_home                 -4.104      3.703  -1.108 0.268013
## job_level                        2772.607     68.512  40.469  < 2e-16 ***
## job_roleHuman Resources          -328.629    221.961  -1.481 0.138945
## job_roleLaboratory Technician    -608.474    140.212  -4.340 1.53e-05 ***
## job_roleManager                  4114.167    182.436  22.551  < 2e-16 ***
## job_roleManufacturing Director    -91.199    138.511  -0.658 0.510374
## job_roleResearch Director        4006.173    183.527  21.829  < 2e-16 ***
## job_roleResearch Scientist       -528.001    139.542  -3.784 0.000161 ***
## job_roleSales Executive          -105.359    126.509  -0.833 0.405089
## job_roleSales Representative     -701.603    182.652  -3.841 0.000128 ***
## marital_statusMarried              48.096     77.150   0.623 0.533118
## marital_statusSingle               23.255     83.002   0.280 0.779380
## total_working_years                49.009      8.232   5.953 3.32e-09 ***
## performance_rating                -27.222     82.507  -0.330 0.741496
## years_at_company                   -4.357      6.479  -0.672 0.501379
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1130 on 1405 degrees of freedom
## Multiple R-squared:  0.9435, Adjusted R-squared:  0.9425
## F-statistic: 937.9 on 25 and 1405 DF,  p-value: < 2.2e-16
```

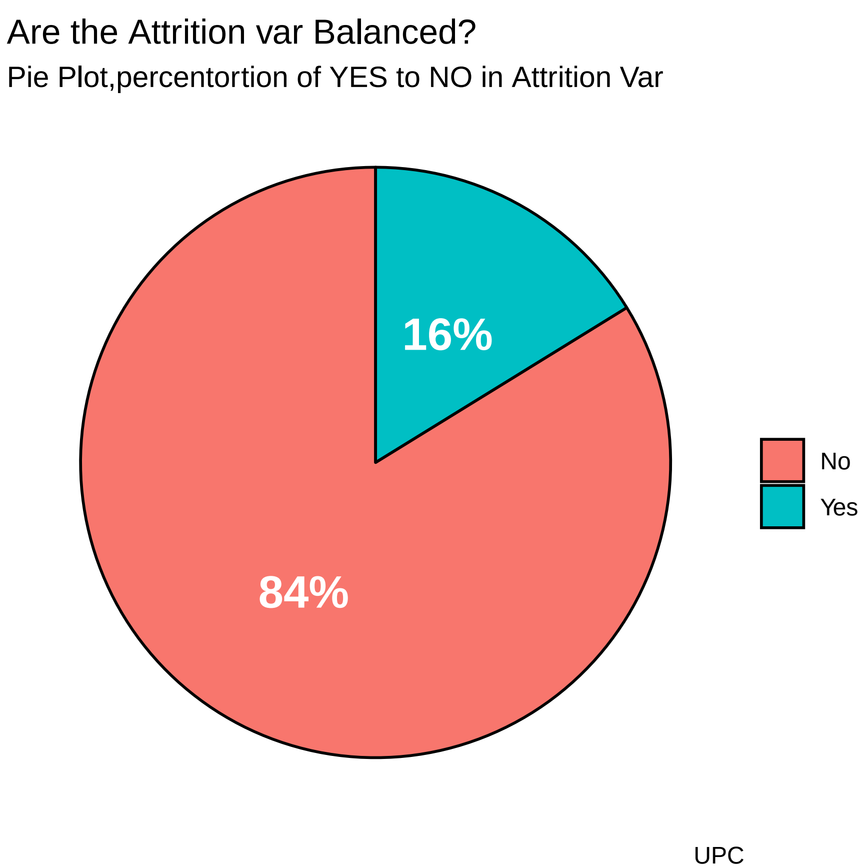
```
rm(mod1, mod2, mod3)
```

Hence, after this tree linear regression models we finally understood that actually the gender variable isn't significative to determine the monthly income. Anyway, this doesn't means that there are no discrepancy between gender, but simply that we need to develop an internal company analysis understand how in Barcelona for some kind of roles women and men have higher salaries compared to the opposite sex.

5.3 Attrition Analysis

First of all, we want to understand how the percentage of attrition is distributed inside the company. Then, we will differenciate per country.

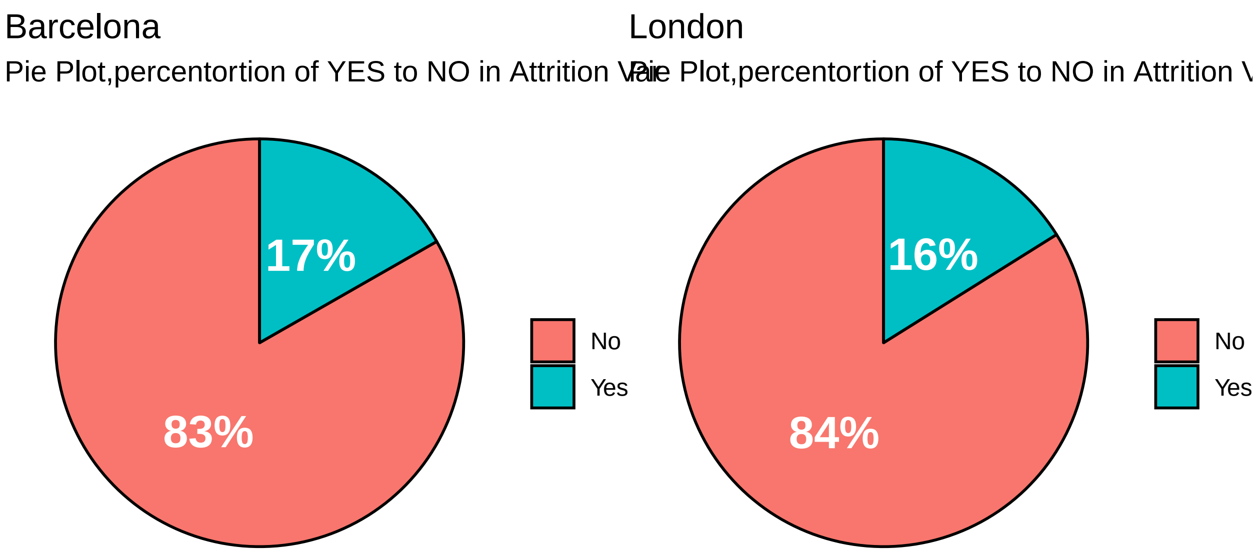
```
# Create a temporary dataset to get the percentage of employees that would leave the company
temp <- mydb %>%
  group_by(attrition) %>%
  summarize(counts = n()) %>%
  mutate(percent = percentage_func(counts)) %>%
  arrange(desc(percent))

# Create a pie chart to see the percentage of people that will leave the company
pie_chart_func(dataset = temp,
               counts_var = temp$counts,
               var_interest = temp$attrition,
               title = "Are the Attrition var Balanced?",
               subtitle = "Pie Plot,percentortion of YES to NO in Attrition Var",
               caption = "UPC")
```




```
# Removing the un-used variables
rm("temp")
```

```
# Barcelona
# Create a temporary dataset to get the percentage of employees that would leave the company
```

```
t1 <- mydb %>% filter(city == "Barcelona") %>%
  group_by(attrition) %>%
  summarize(counts = n()) %>%
  mutate(percent = percentage_func(counts)) %>%
  arrange(desc(percent))

g1 <- pie_chart_func(dataset = t1,
               counts_var = t1$counts,
               var_interest = t1$attrition,
               title = "Barcelona",
               subtitle = "Pie Plot,percentortion of YES to NO in Attrition Var",
               caption = "")

# London
t2 <- mydb %>% filter(city == "London") %>%
  group_by(attrition) %>%
  summarize(counts = n()) %>%
  mutate(percent = percentage_func(counts)) %>%
  arrange(desc(percent))

# Create a pie chart to see the percentage of people that will leave the company
g2 <- pie_chart_func(dataset = t2,
               counts_var = t2$counts,
               var_interest = t2$attrition,
               title = "London",
               subtitle = "Pie Plot,percentortion of YES to NO in Attrition Var",
               caption = "")


grid.arrange(g1, g2, nrow = 1)
```

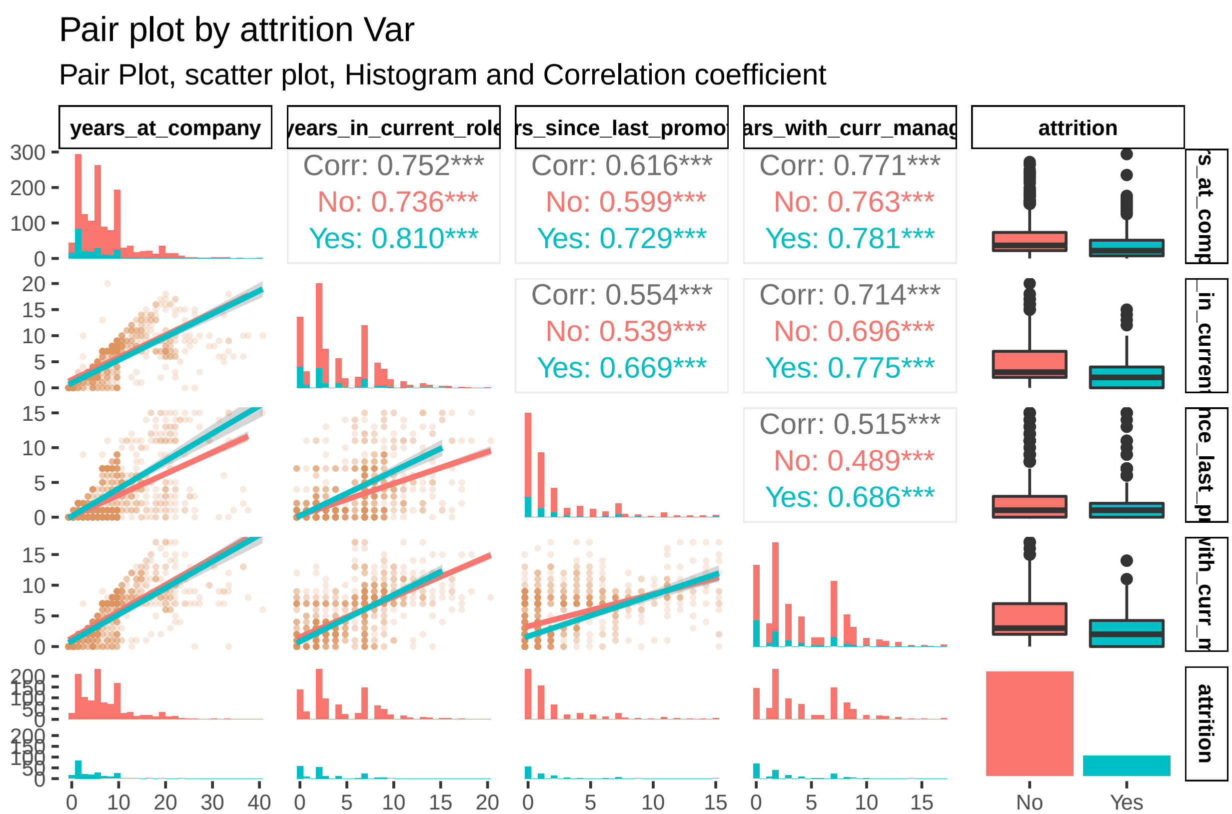



```
# Removing the un-used variables
rm(g1, g2, t1, t2)
```

Fortunately, as we can see the percentage of attrition between Barcleona and London is almost the same, so we can proceed with a generalized analysis.

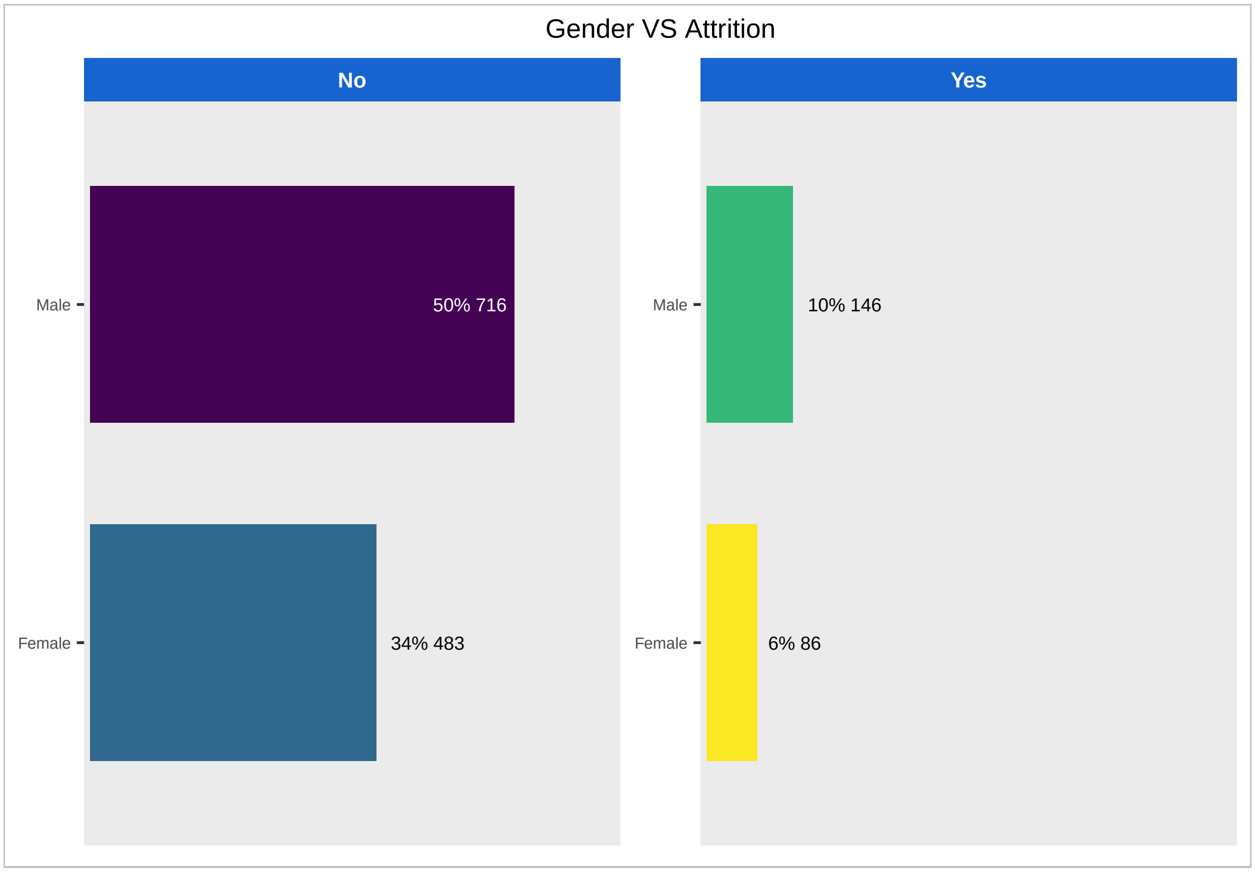
```
mydb %>% select(starts_with("years"), attrition) %>%

ggpairs(
  aes(color = attrition),
  lower = list(continuous = wrap(
    "smooth",
    alpha = 0.2,
    size = 0.5,
    color = "#DE945E"
  )),
  diag = list(continuous = "barDiag"),
  upper = list(continuous = wrap("cor", size = 4))
) +
  theme(
    axis.text = element_text(size = 8),
    panel.background = element_rect(fill = "white"),
    strip.background = element_rect(fill = "white"),
    strip.background.x = element_rect(colour = "black"),
    strip.background.y = element_rect(colour = "black"),
```

```
  strip.text = element_text(color = "black", face = "bold", size = 8)
) +
labs(
  title = "Pair plot by attrition Var",
  subtitle = "Pair Plot, scatter plot, Histogram and Correlation coefficient",
  caption = "",
  x = NULL,
  y = NULL
)
```



5.3.1 Gender vs Attrition

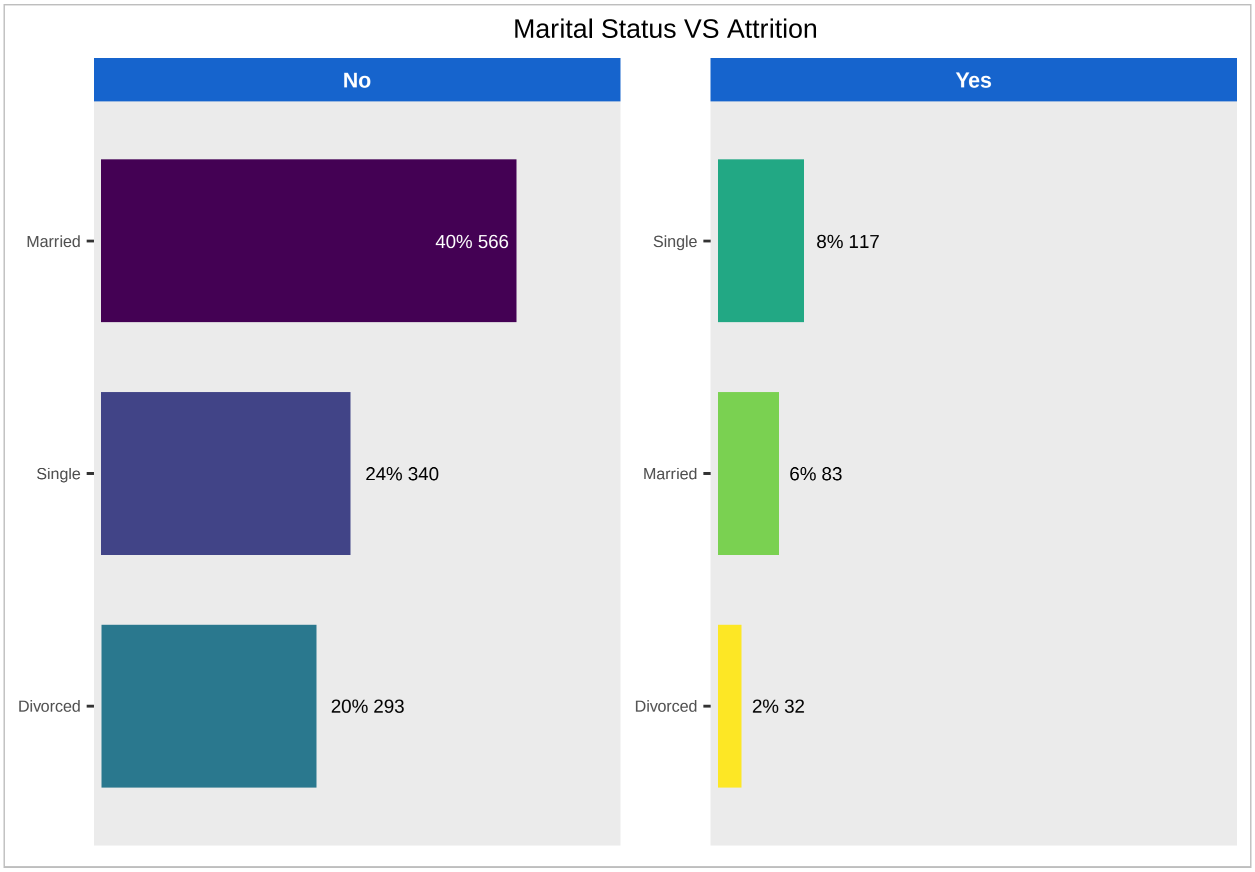
```
bar_plot_proportions(gender, attrition)
```



So, as we can see male tend to leave the comapny more the women does, but we also have to take into account the fact that there are more men in the company. Hence, the relation seems to be pretty balanced.

5.3.2 Marital status vs Attrition

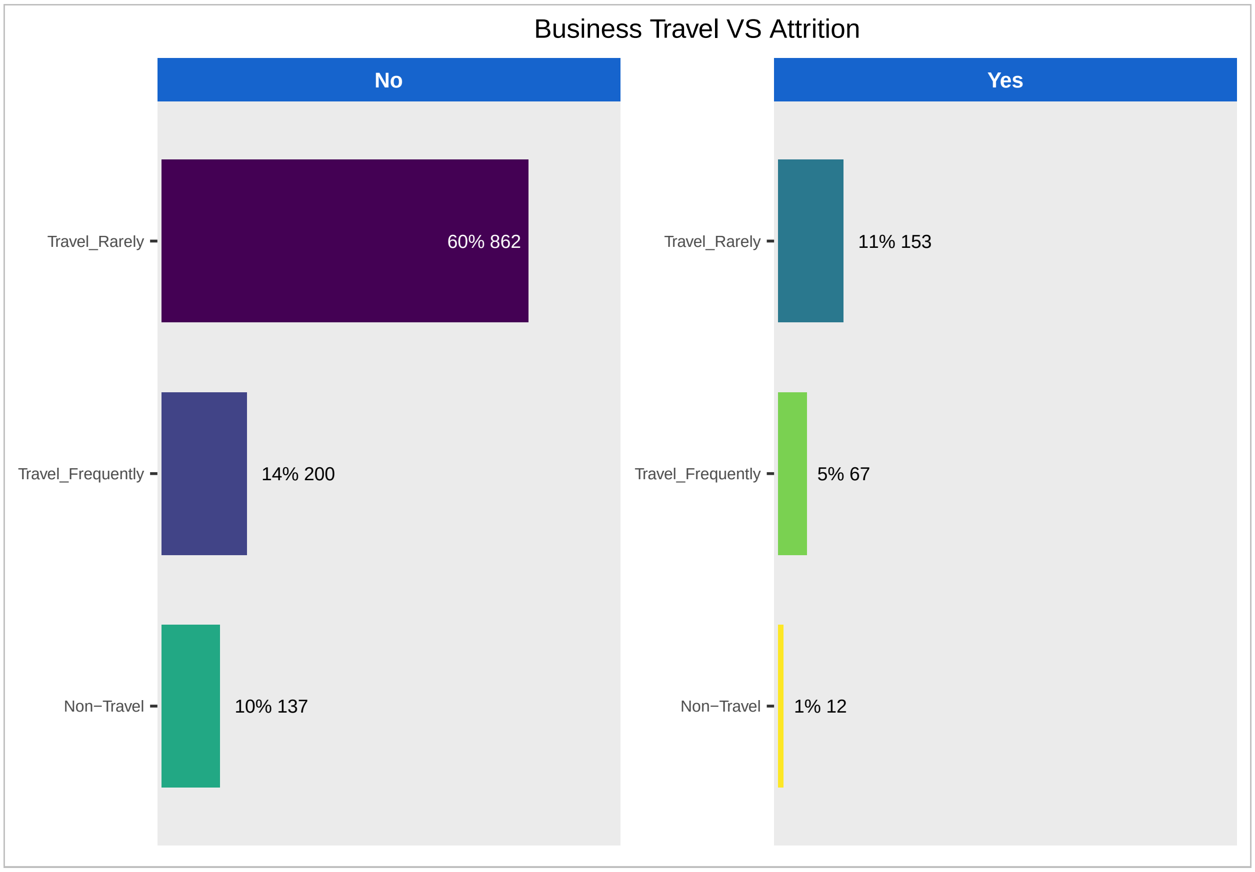
```
bar_plot_proportions(marital_status, attrition)
```



Here, we can see how single people tend to leave the company more, compared to married and divorced.

The *business_travel* variable could be a very important variable that will tell us if the travels affect the attrition variable.

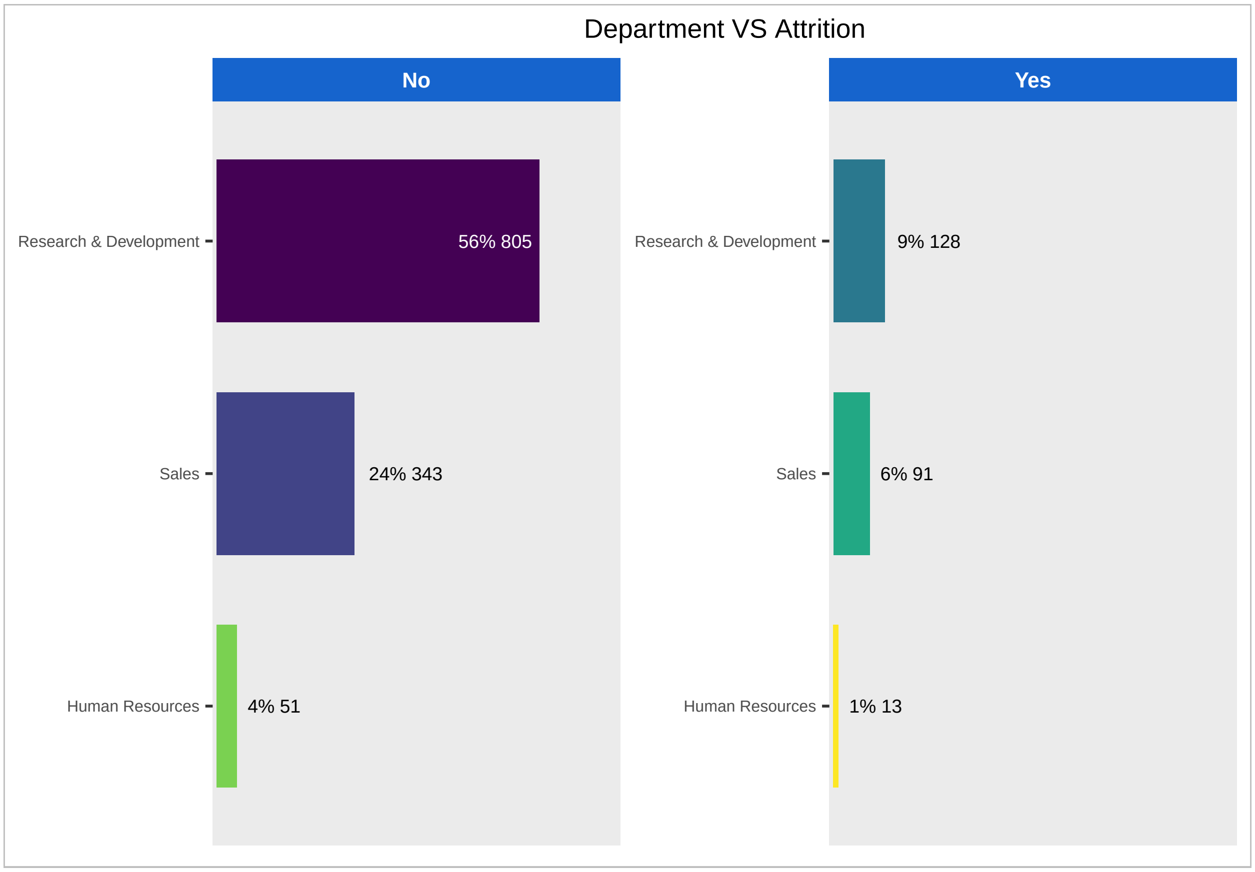
```
bar_plot_proportions(business_travel, attrition)
```



At the end, we have that the main obs that affect the attrition is the travel_rarely, which could that the non routine and the change of plans for the single employee could cause an higher attrition. Hence, we suggest to clarify with higher advance if that employee have to travel or not, and finally take the new results and compute a further analysis.

The *department* variable could be a very important variable that will tell us if the department affect the attrition variable.

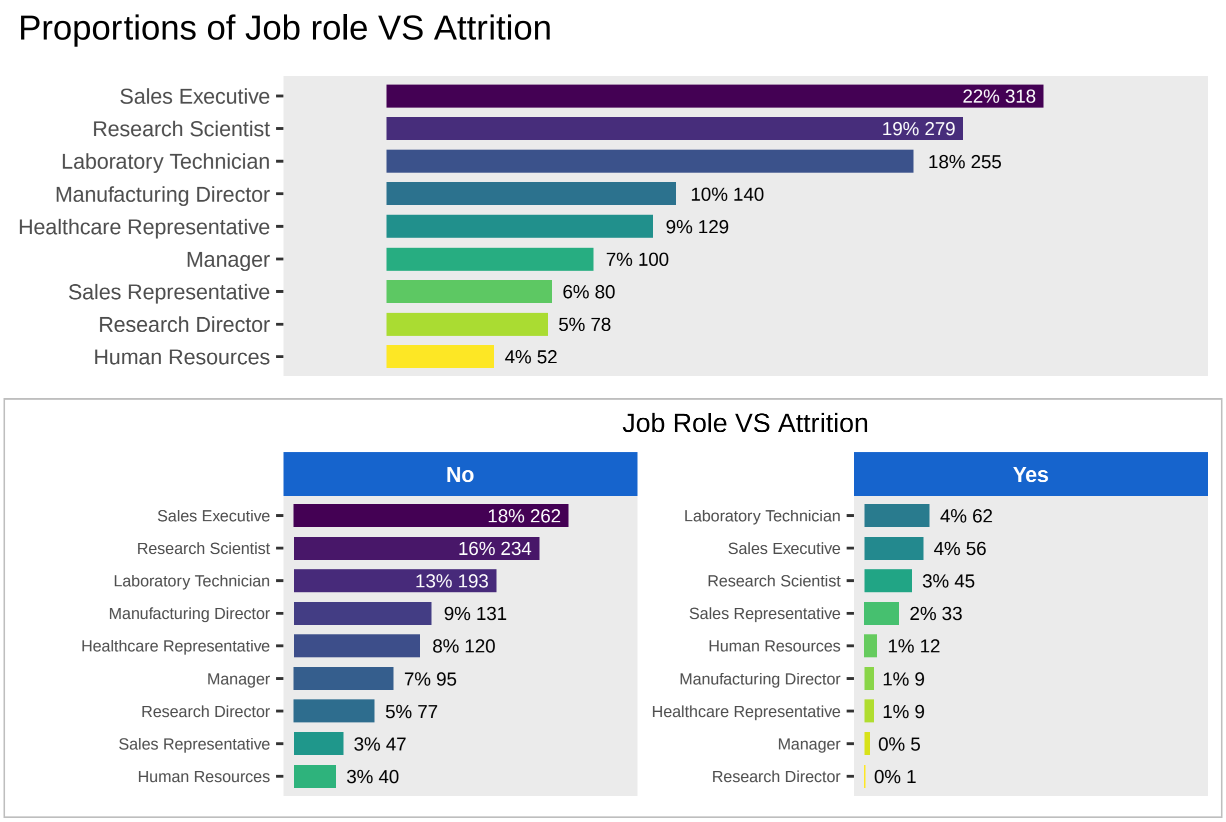
```
bar_plot_proportions(department, attrition)
```



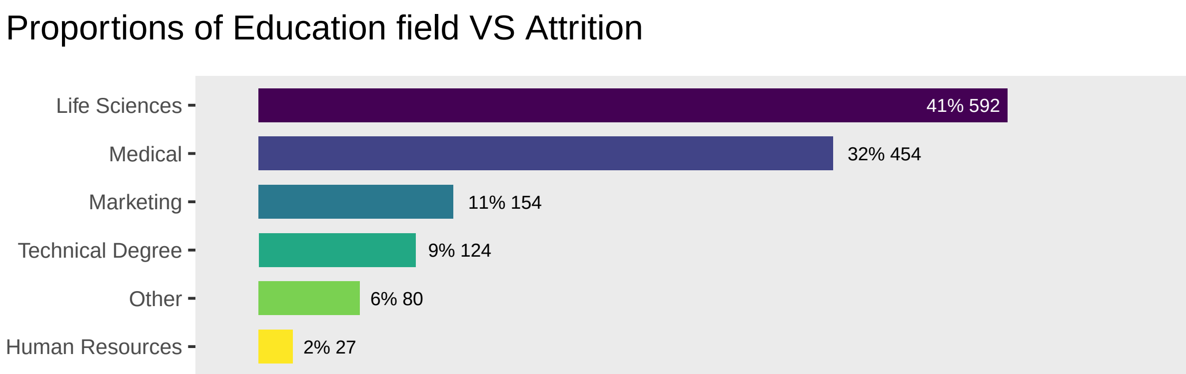
Here instead, we can see how the R&D department seems the one with higher attrition. Anyway, we have also have to take into consideration that the majority of the people inside the company work in this department.

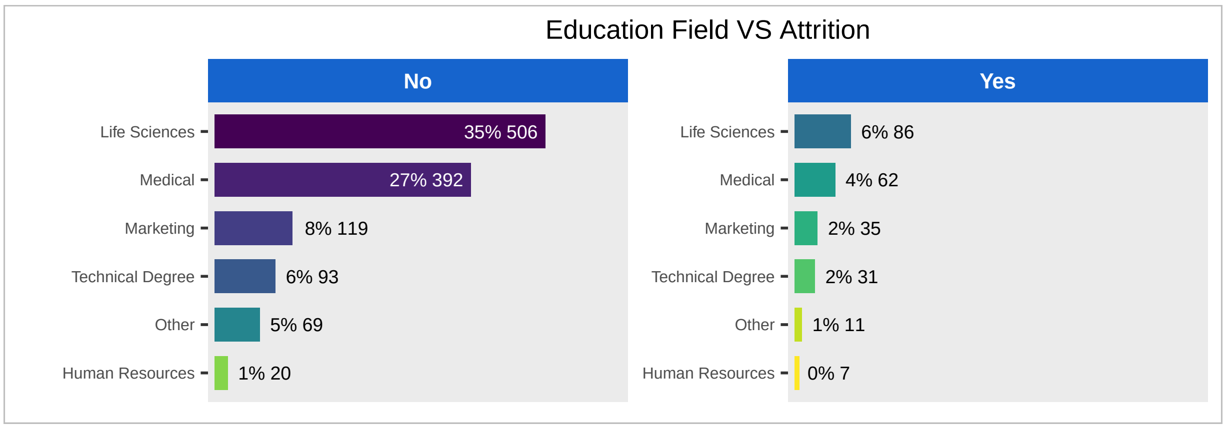
Now, we want to study also the

```
plt_job_role <- bar_plot_proportions(job_role)
plt_job_role_att <- bar_plot_proportions(job_role, attrition)
(plt_job_role /
   plt_job_role_att) +
  plot_annotation(
  title = "Proportions of Job role VS Attrition",
  caption = ""
) &
  theme(plot.caption = element_text(color = "#969696", size = 7))
```




```r
plt_education_field <- bar_plot_proportions(education_field)
plt_education_field_att <- bar_plot_proportions(education_field, attrition)

(plt_education_field /
   plt_education_field_att) +
  plot_annotation(
  title = "Proportions of Education field VS Attrition",
  caption = "Data Source: Kaggle IBM HR Employee Attrition"
) &
  theme(plot.caption = element_text(color = "#969696", size = 7))
```





Data Source: Kaggle IBM HR Employee Attrition

5.3.3 Regression

```
mydb$attrition <-ifelse(mydb$attrition=="Yes",1,0)
mod1 <- glm(attrition ~ ., data = mydb)
summary(mod1)
```

```
##
## Call:
## glm(formula = attrition ~ ., data = mydb)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.56380  -0.20926  -0.08298   0.08646   1.13122
##
## Coefficients: (2 not defined because of singularities)
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       4.916e-01  2.185e-01   2.250 0.024588 *
## age                              -3.542e-03  1.341e-03  -2.641 0.008358 **
## business_travelTravel_Frequently  1.517e-01  3.341e-02   4.542 6.06e-06 ***
## business_travelTravel_Rarely      6.205e-02  2.872e-02   2.161 0.030875 *
## daily_rate                       -2.586e-05  2.151e-05  -1.202 0.229386
## departmentResearch & Development -8.383e-03  1.107e-01  -0.076 0.939666
## departmentSales                  -7.603e-03  1.169e-01  -0.065 0.948150
## distance_from_home                3.733e-03  1.060e-03   3.521 0.000445 ***
```

```
## education                          4.978e-04  8.662e-03   0.057 0.954184
## education_fieldLife Sciences      -8.436e-02  8.315e-02  -1.015 0.310491
## education_fieldMarketing          -3.531e-02  8.887e-02  -0.397 0.691157
## education_fieldMedical            -9.266e-02  8.353e-02  -1.109 0.267465
## education_fieldOther              -1.043e-01  8.962e-02  -1.164 0.244614
## education_fieldTechnical Degree    8.256e-03  8.722e-02   0.095 0.924596
## employee_count                            NA         NA      NA       NA
## employee_number                   -1.202e-05  1.438e-05  -0.836 0.403552
## environment_satisfaction          -4.232e-02  7.920e-03  -5.343 1.07e-07 ***
## genderMale                         3.227e-02  1.772e-02   1.821 0.068761 .
## hourly_rate                       -1.807e-04  4.258e-04  -0.424 0.671339
## job_involvement                   -5.786e-02  1.219e-02  -4.748 2.27e-06 ***
## job_level                         -5.677e-03  2.889e-02  -0.197 0.844246
## job_roleHuman Resources            9.030e-02  1.179e-01   0.766 0.443924
## job_roleLaboratory Technician      1.374e-01  4.041e-02   3.400 0.000693 ***
## job_roleManager                    3.070e-02  6.833e-02   0.449 0.653228
## job_roleManufacturing Director     5.641e-04  3.978e-02   0.014 0.988689
## job_roleResearch Director         -1.384e-02  6.124e-02  -0.226 0.821273
## job_roleResearch Scientist         3.917e-02  4.011e-02   0.977 0.328942
## job_roleSales Executive            7.343e-02  7.833e-02   0.937 0.348721
## job_roleSales Representative       2.369e-01  8.716e-02   2.718 0.006640 **
## job_satisfaction                  -3.807e-02  7.855e-03  -4.847 1.39e-06 ***
## marital_statusMarried              1.582e-02  2.315e-02   0.684 0.494339
## marital_statusSingle               1.034e-01  3.182e-02   3.249 0.001186 **
## monthly_income                     2.552e-06  7.673e-06   0.333 0.739467
## monthly_rate                       4.119e-07  1.212e-06   0.340 0.734037
## num_companies_worked               1.743e-02  3.856e-03   4.521 6.69e-06 ***
## over18Y                            2.014e-01  1.244e-01   1.619 0.105642
## over_timeYes                       2.057e-01  1.921e-02  10.705  < 2e-16 ***
## percent_salary_hike               -2.603e-03  3.729e-03  -0.698 0.485281
## performance_rating                 1.798e-02  3.756e-02   0.479 0.632330
## relationship_satisfaction         -2.225e-02  8.018e-03  -2.775 0.005596 **
## standard_hours                            NA         NA      NA       NA
## stock_option_level                -2.024e-02  1.384e-02  -1.462 0.143914
## total_working_years               -4.342e-03  2.447e-03  -1.774 0.076240 .
## training_times_last_year          -1.410e-02  6.750e-03  -2.089 0.036913 *
## work_life_balance                 -3.240e-02  1.223e-02  -2.650 0.008153 **
## years_at_company                   2.993e-03  3.014e-03   0.993 0.320805
## years_in_current_role             -7.562e-03  3.864e-03  -1.957 0.050563 .
## years_since_last_promotion         1.128e-02  3.471e-03   3.251 0.001179 **
## years_with_curr_manager           -7.716e-03  4.043e-03  -1.908 0.056548 .
## cityLondon                         1.143e-02  2.232e-02   0.512 0.608768
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.1037191)
##
##     Null deviance: 194.39  on 1430  degrees of freedom
## Residual deviance: 143.44  on 1383  degrees of freedom
## AIC: 867.43
##
## Number of Fisher Scoring iterations: 2
```

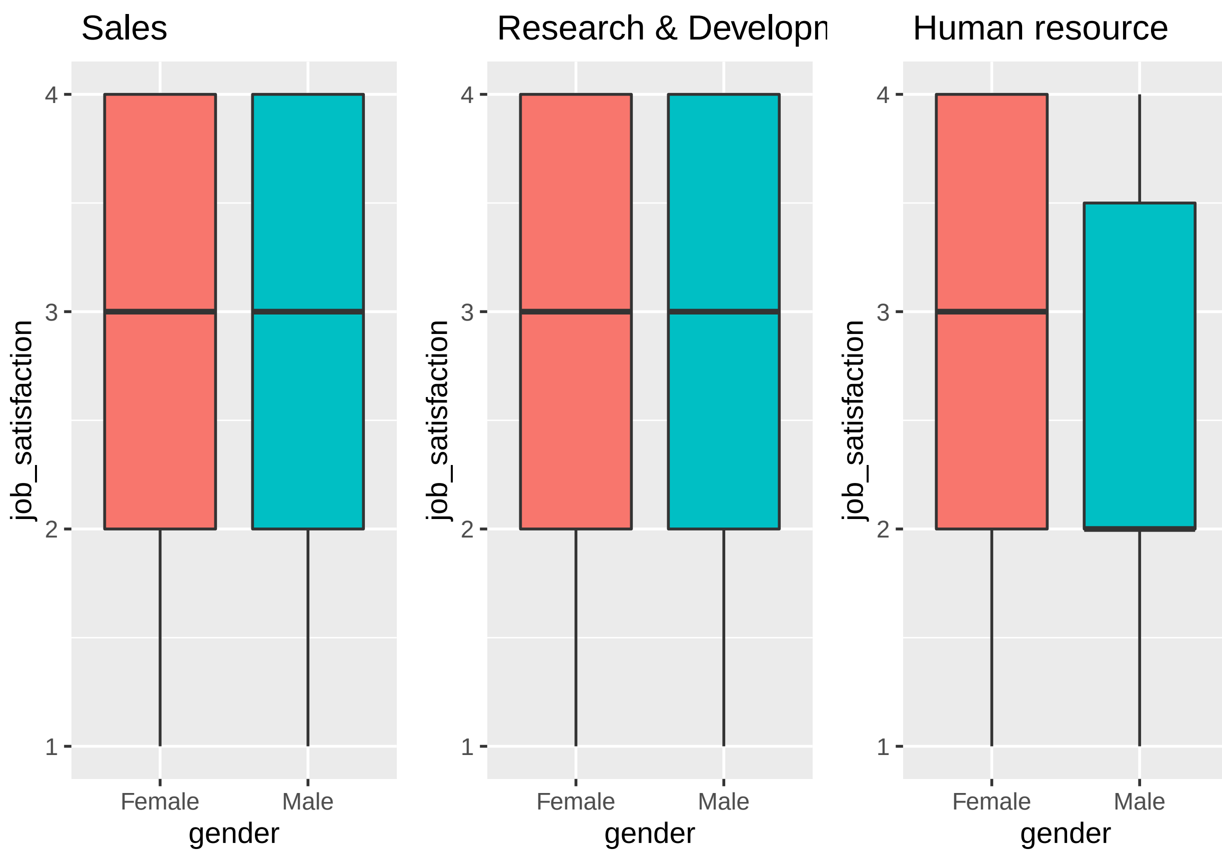
```
# mod2 <- lm(monthly_income ~ gender, data = mydb)
# summary(mod2)
#
# names(mydb)
# mod3 <- lm(monthly_income ~ age + gender + business_travel + education_field + education + distance_f
# summary(mod3)

rm(mod1, mod2, mod3)
```

```
## Warning in rm(mod1, mod2, mod3): oggetto 'mod2' non trovato
```

```
## Warning in rm(mod1, mod2, mod3): oggetto 'mod3' non trovato
```

satisfaction of male and female in each department considering the whole data set

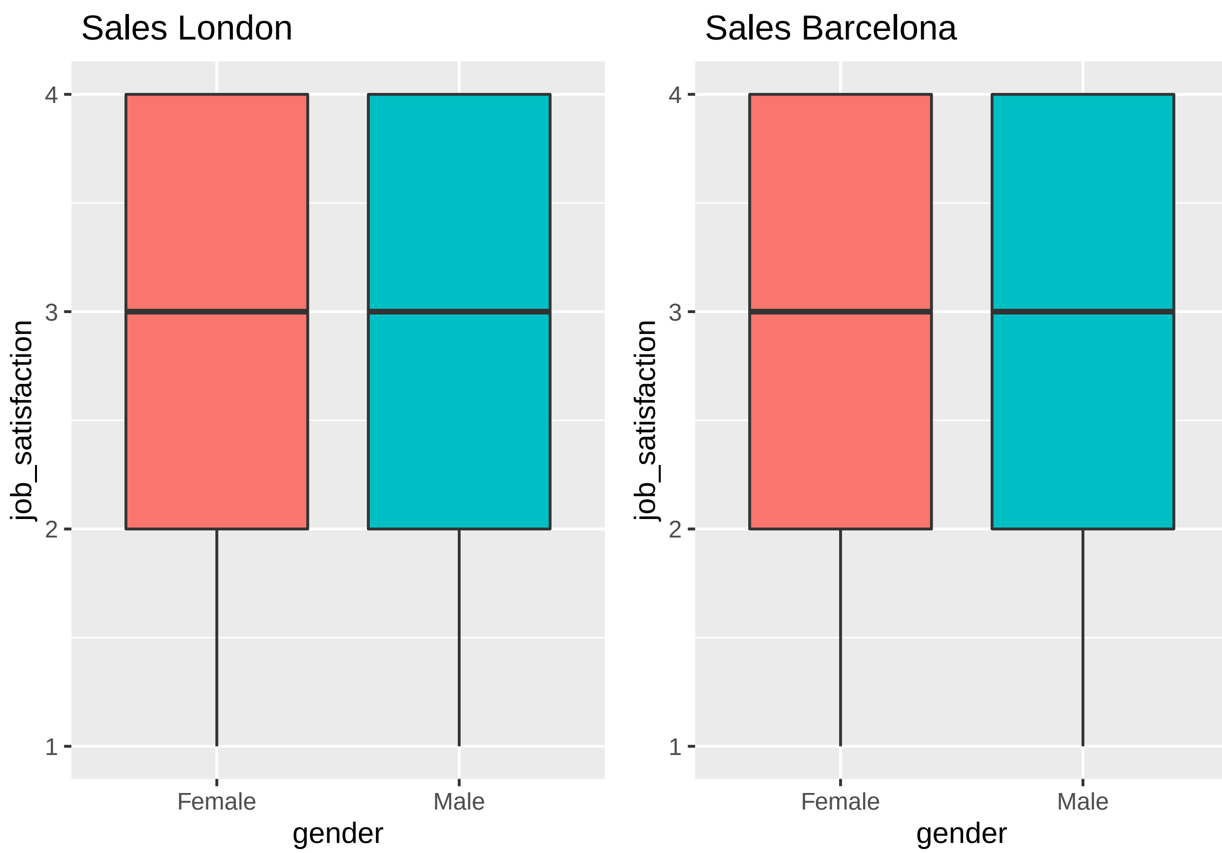
```
##Sales
g1 <- ggplot(mydb%>%filter(department =="Sales"),
             aes(x= gender, y=job_satisfaction, fill= gender))+
  geom_boxplot(show.legend = FALSE)+
  labs(title = " Sales")

##Research & Development
g2 <- ggplot(mydb%>%filter(department =="Research & Development"),
             aes(x= gender, y=job_satisfaction, fill= gender))+
  geom_boxplot(show.legend = FALSE)+
  labs(title = " Research & Development ")

##Human Resource
g3 <- ggplot(mydb%>%filter(department =="Human Resources"),
             aes(x= gender, y=job_satisfaction, fill= gender))+
  geom_boxplot(show.legend = FALSE)+
  labs(title = " Human resource ")

grid.arrange(g1, g2, g3, ncol = 3)
```



5.4 Differences between the two cities in term of job satisfaction

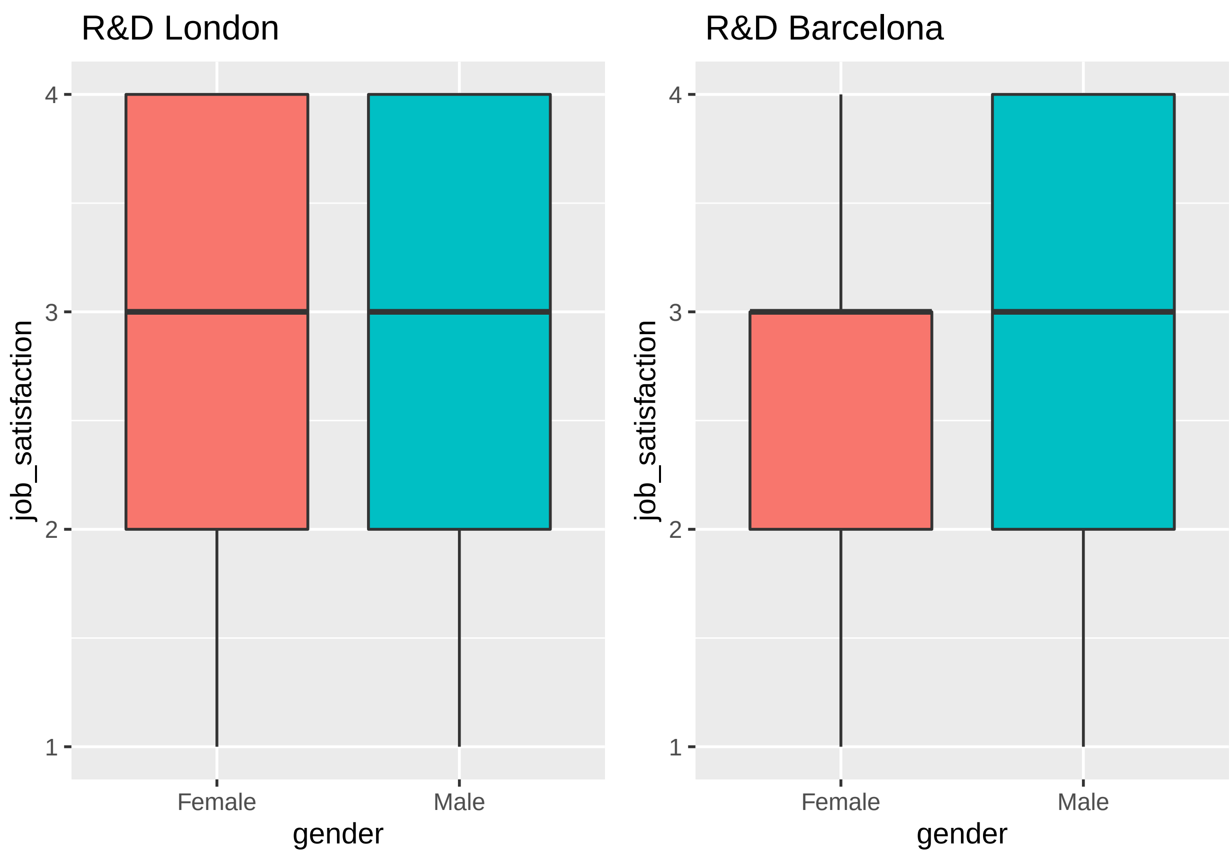
5.4.1 Sales, Gender, Job satisfaction

```
##Sales  London
g4 <- ggplot(mydb%>%filter(department =="Sales" & city == "London"),
             aes(x= gender, y= job_satisfaction, fill= gender))+
  geom_boxplot(show.legend = FALSE)+
  labs(title = " Sales London ")

##Sales  Barcelona
g5 <- ggplot(mydb%>%
               filter(department =="Sales" & city == "Barcelona"),
             aes(x= gender, y=job_satisfaction, fill= gender))+
  geom_boxplot(show.legend = FALSE)+
  labs(title = " Sales Barcelona ")
grid.arrange(g4,g5, ncol = 2)
```



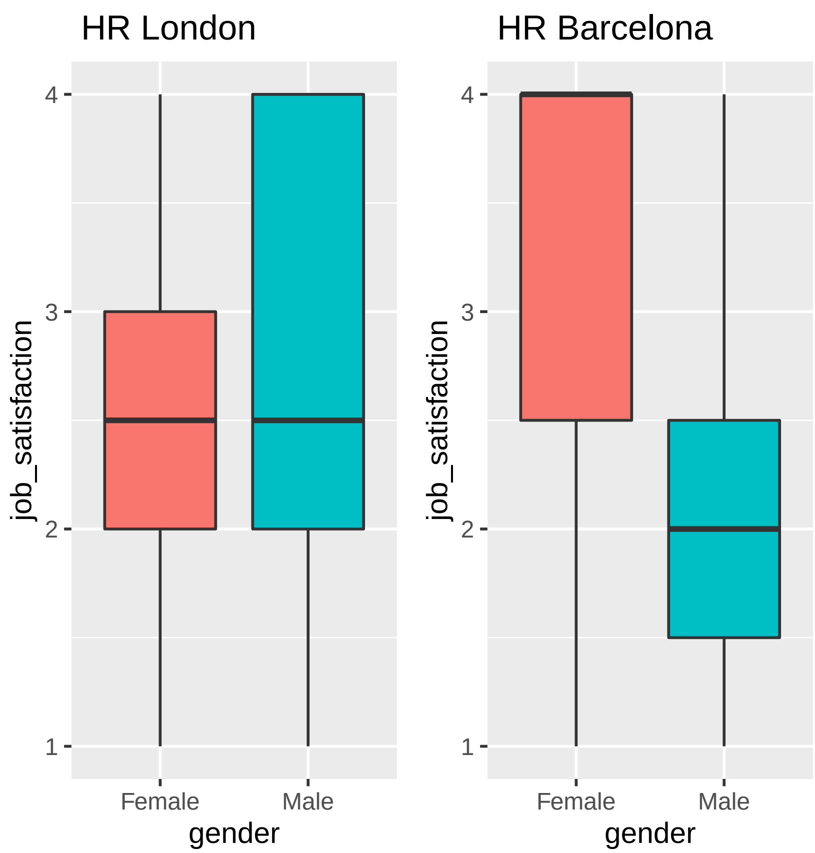
5.4.2 R&D, Gender, Job satisfaction

```
##Research & Development London
g6 <- ggplot(mydb%>%
               filter(department =="Research & Development" & city=="London"),
             aes(x= gender, y=job_satisfaction, fill= gender))+
  geom_boxplot(show.legend = FALSE)+
  labs(title = " R&D London")

##Research & Development Barcelona
g7 <- ggplot(mydb%>%
               filter(department =="Research & Development" & city=="Barcelona"),
             aes(x= gender, y=job_satisfaction, fill= gender))+
  geom_boxplot(show.legend = FALSE)+
  labs(title = " R&D Barcelona")

grid.arrange(g6,g7, ncol = 2)
```

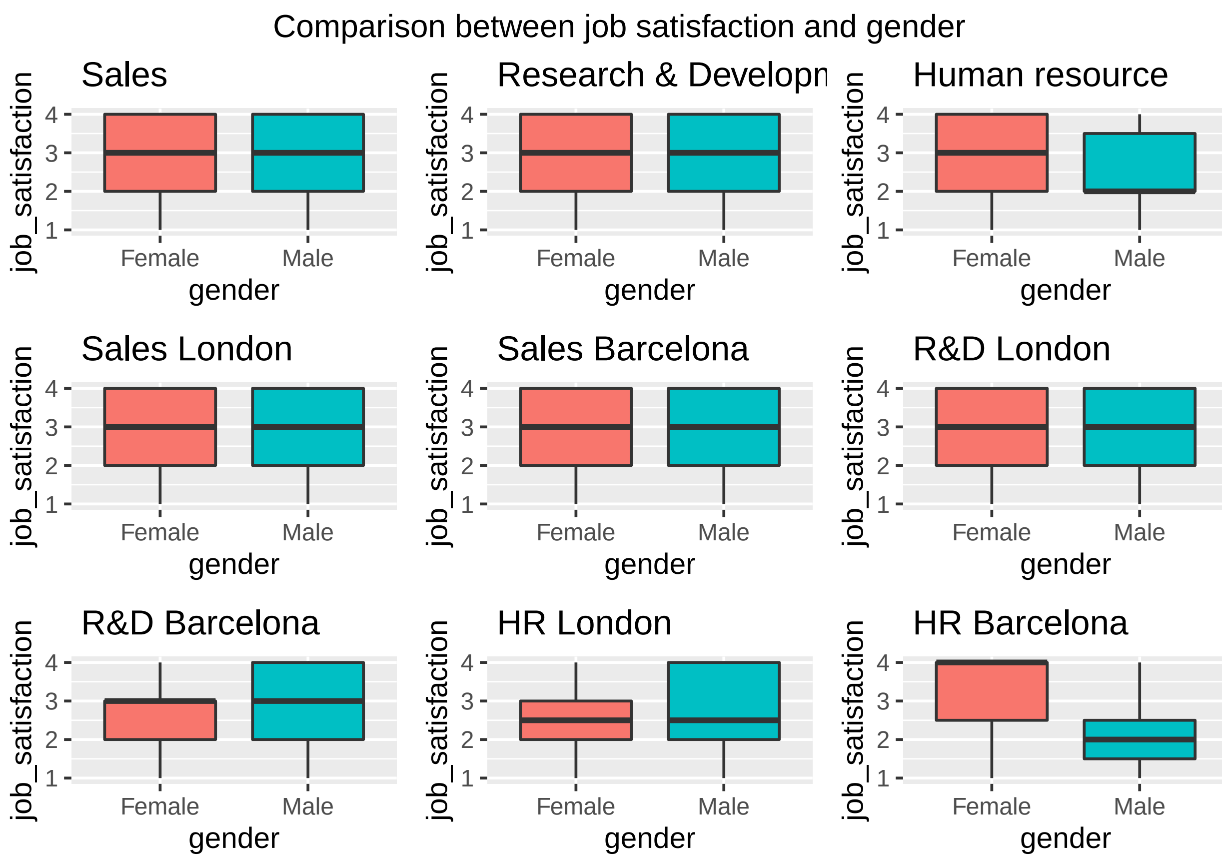


5.4.3 HR, Gender, Job satisfaction

```
##Human Resource London

g8 <- ggplot(mydb%>%
               filter(department =="Human Resources" & city=="London"),
             aes(x= gender, y=job_satisfaction, fill= gender))+
  geom_boxplot(show.legend = FALSE)+
  labs(title = " HR London")

##Human Resource BArcelona
g9 <- ggplot(mydb%>%
               filter(department =="Human Resources" & city=="Barcelona"),
             aes(x= gender, y=job_satisfaction, fill= gender))+
  geom_boxplot(show.legend = FALSE)+
    labs(title = " HR Barcelona")

grid.arrange(g8,g9, ncol= 3)
```



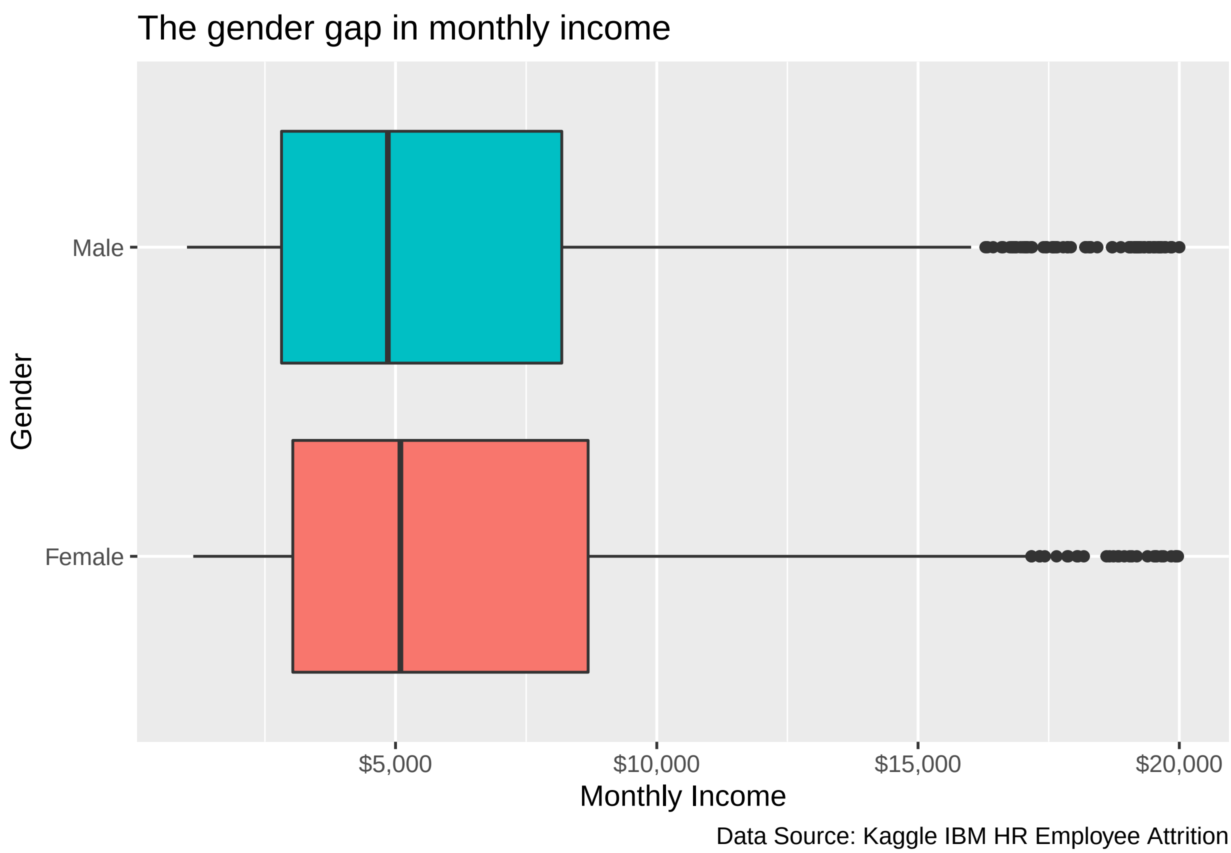

```
grid.arrange(g1, g2, g3, g4, g5, g6, g7, g8, g9, ncol = 3, top = "Comparison between job satisfaction ar
```




```
# clean variables
rm(g1, g2, g3, g4, g5, g6, g7, g8, g9)
```

```
mean_mi <- round(mean(mydb$monthly_income),2)
median_mi <- round(median(mydb$monthly_income),2)
```

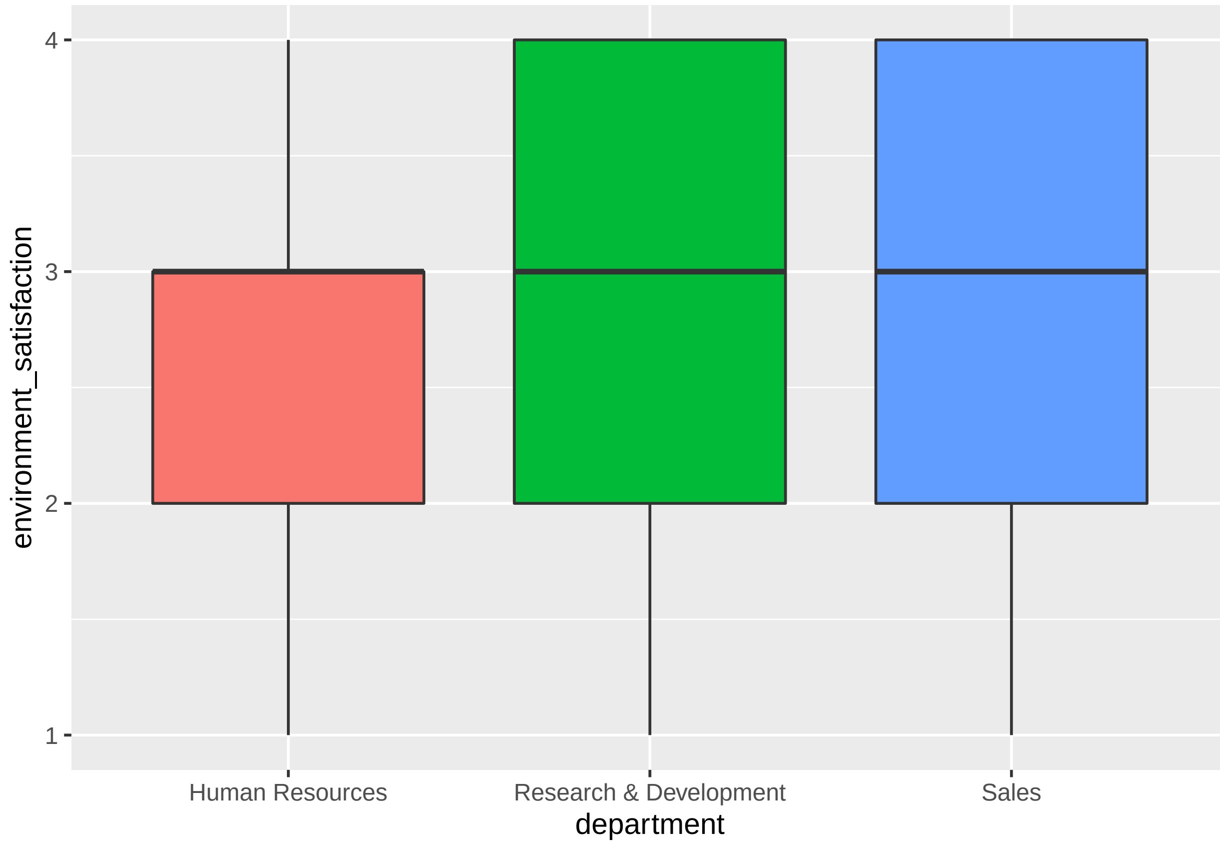
```
ggplot(mydb,
       aes(x = gender,
           y = monthly_income,
           fill = gender)) +
  geom_boxplot(show.legend = FALSE) +
  coord_flip() +
   scale_y_continuous(labels = label_dollar()) +
  labs(title = "The gender gap in monthly income",
       caption = "Data Source: Kaggle IBM HR Employee Attrition",
       x = "Gender",
       y = "Monthly Income")
```



5.5 Environment Satisfaction per department

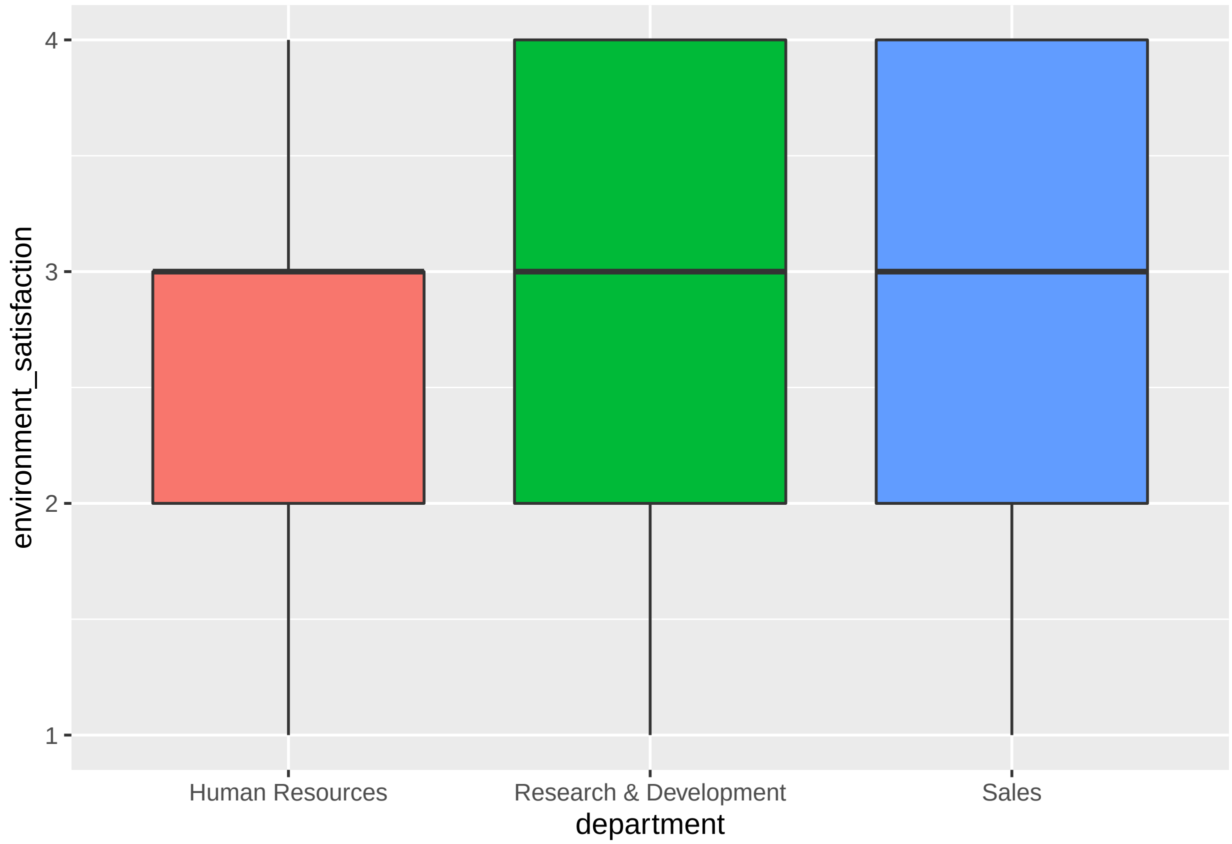
5.5.1

```
ggplot(mydb, aes(x= department, y= environment_satisfaction, fill = department))+
  geom_boxplot(show.legend = FALSE)
```



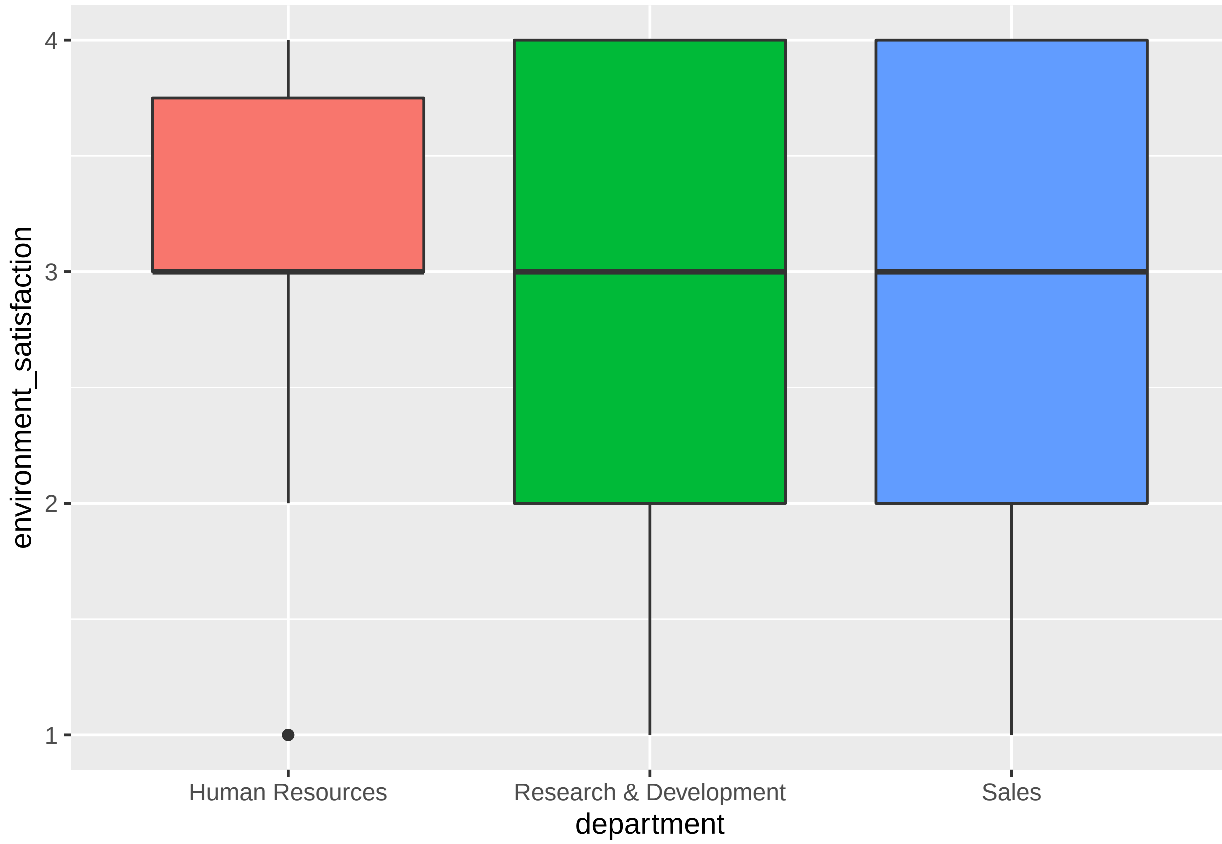
London male

```
ggplot(mydb%>%filter(city == "London", gender== "Male"), aes(x= department, y= environment_satisfaction
  geom_boxplot(show.legend = FALSE)
```



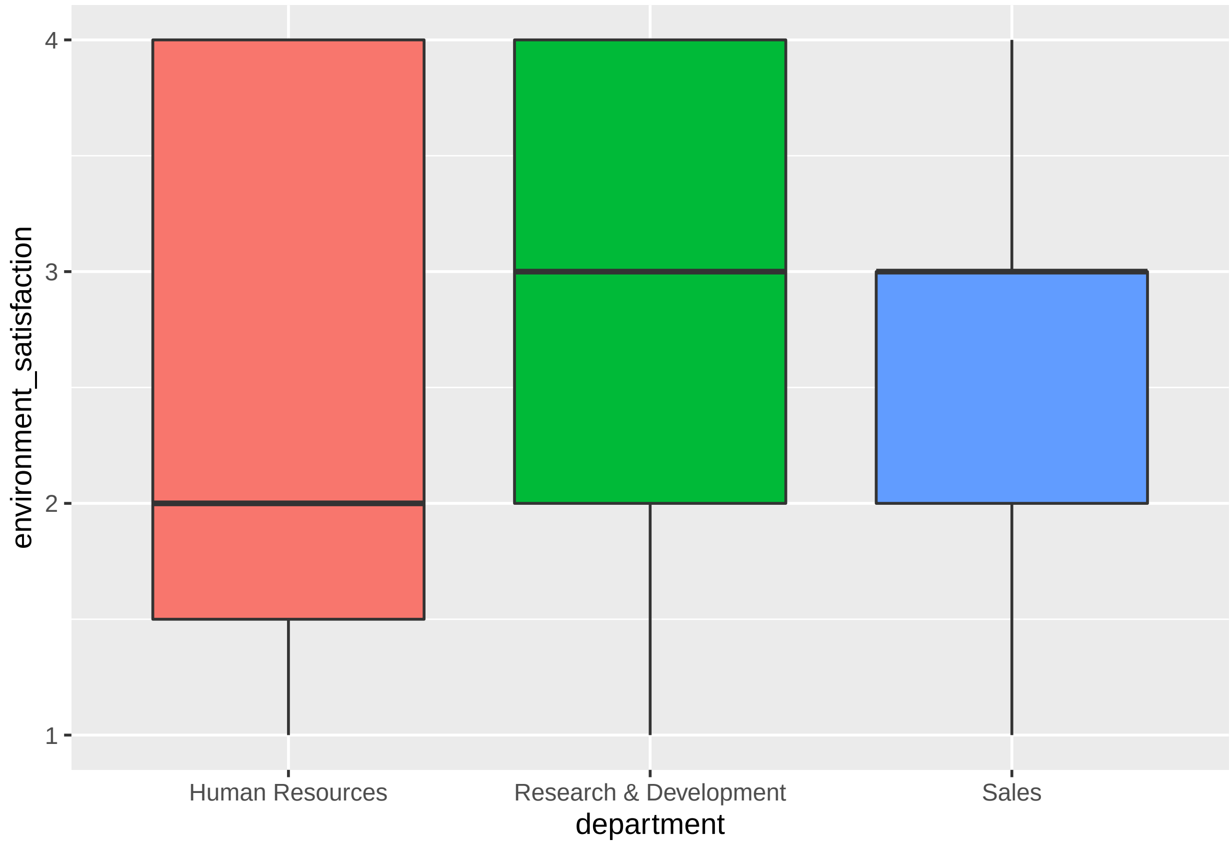
London female

```
ggplot(mydb%>%filter(city == "London", gender== "Female"), aes(x= department, y= environment_satisfacti
  geom_boxplot(show.legend = FALSE)
```



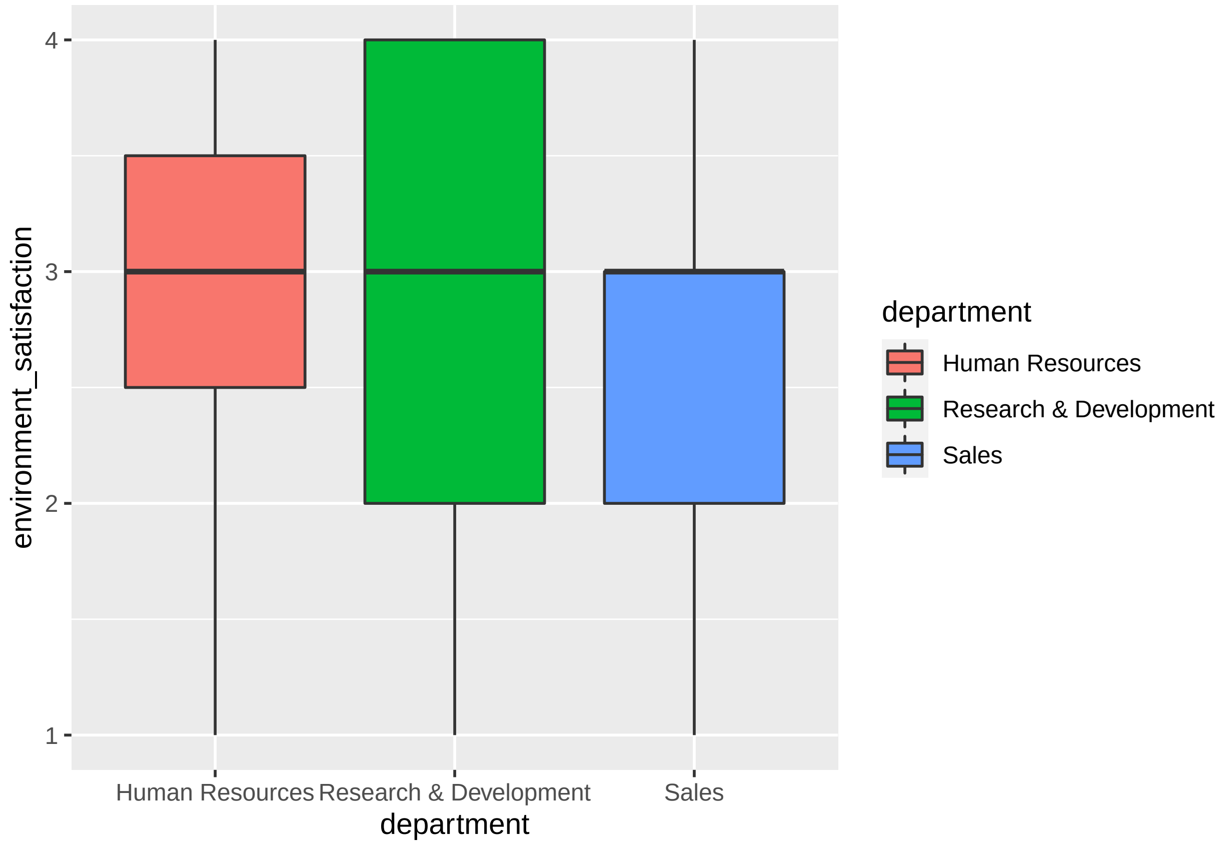
Barcelona malemale

```
ggplot(mydb%>%filter(city == "Barcelona", gender== "Male"), aes(x= department, y= environment_satisfact
  geom_boxplot(show.legend = FALSE)
```



Barcelona Female

```
ggplot(mydb%>%filter(city == "Barcelona", gender== "Female"), aes(x= department, y= environment_satisfac
  geom_boxplot()+ labs(show.legend = FALSE)
```



linear regression to understand which variable influence monthly income

```
 #m1<-lm(city ~ ., mydb%>%drop_na()%>% select(-employee_count,-employee_number,-standard_hours))
#summary(m1)
```

- performance rating e percent salary hike,

```
# Fa togliere
# ggplot(mydb%>%filter(city == "Barcelona"), aes(y= percent_salary_hike, x = performance_rating))+
#   geom_point()
#
# ggplot(mydb%>%filter(city == "London"), aes(y= percent_salary_hike, x = performance_rating))+
#   geom_point()
#
# ggplot(mydb, aes(x= percent_salary_hike, y = monthly_income)) + geom_point()
```

We can

-
- total working year e age, years at company, relationship satisfaction, numb of companies worked, monthly income, job role, job level
- environment satisfaction e over time (Yes), over 18, hourly rate, attrition yes
- job involvement job role, business travel and attrition

- job satisfaction attrition, years with current manager, overtime, marital status, hourly rate

-year since last promotion and, years in current role,years at company, job roles, attrition

6 Adding dataset

```
london_data<- read.csv("1st Dataset.csv")
london_data<- london_data%>%filter(Geography == "London", Sex == c("Female", "Male"), WorkingPattern ==
```

7 Conclusions

- Gender equality seems to be well reached in London, while in Barcelona we can suggest to concentrate more in defining same salaries per

8 Referemces

1. *M., L. (2004). Moneyball: The art of winning an unfair game. New york: John Wiley and sons.*
2. Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686
3. H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.
4. David Robinson, Alex Hayes and Simon Couch (2021). broom: Convert Statistical Objects into Tidy Tibbles. R package version 0.7.9. https://CRAN.R-project.org/package=broom