Final Project

Ettore Falde, Samoussa Fofana, Federico Basaglia

11/29/2021

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).

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)
```

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)
```

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)

Cleaning Data

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")
```

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
```

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 =age)
mydb %>% arrange(desc(age)) %>% top_n(10, age)
mydb %>% arrange(age) %>% top_n(-10, age)
```

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)
```

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

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.

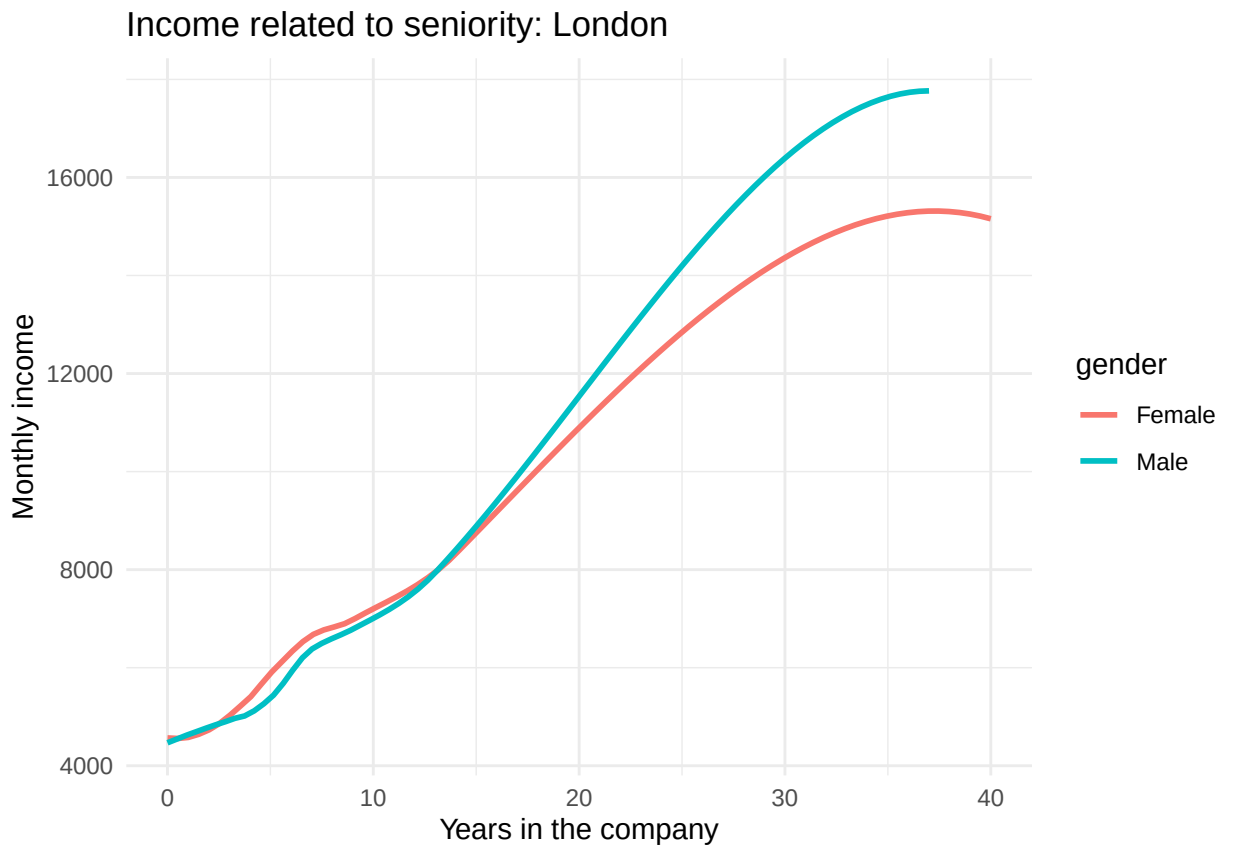
Analysis

scatter plot monthly income with respect to age of wich you work in the company differentiate by gender 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()
```

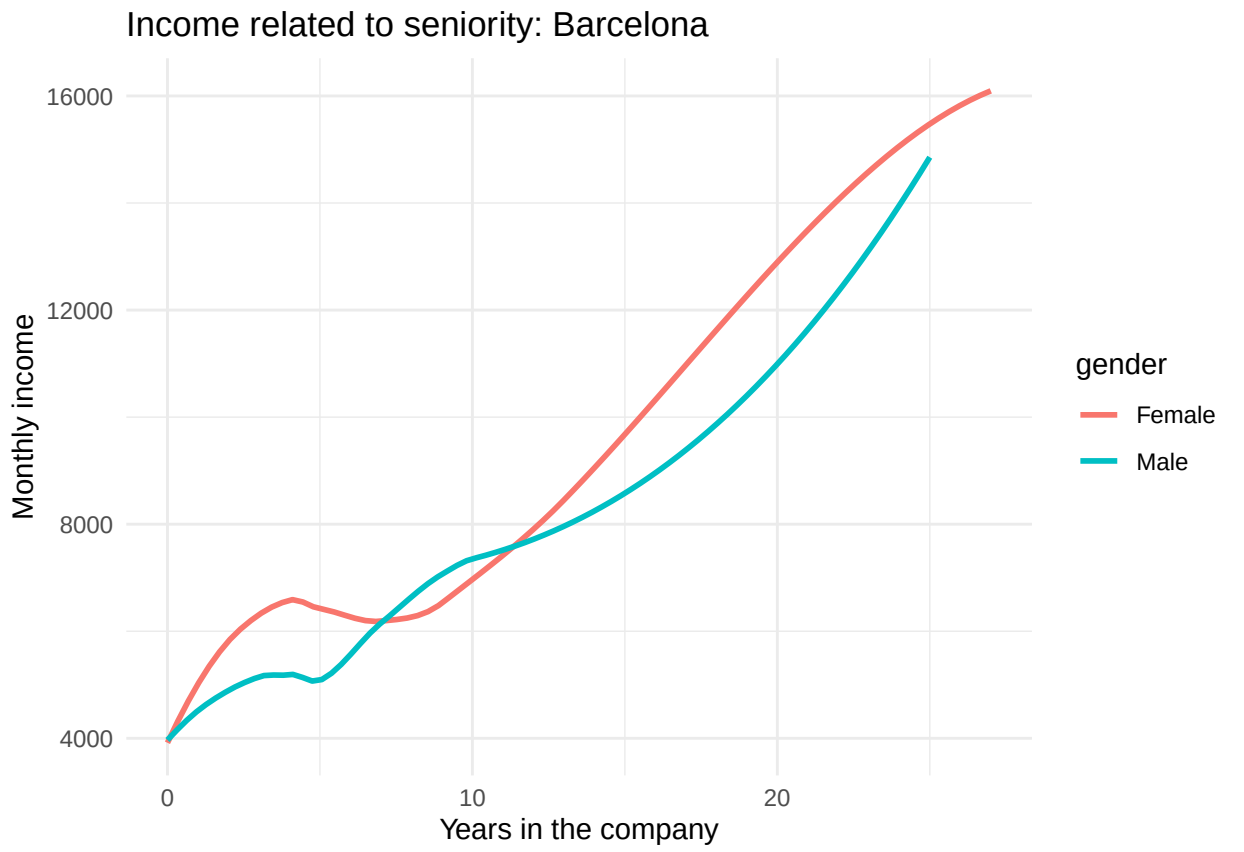
Taking in consideration London

```
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()
```



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()
```




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

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

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

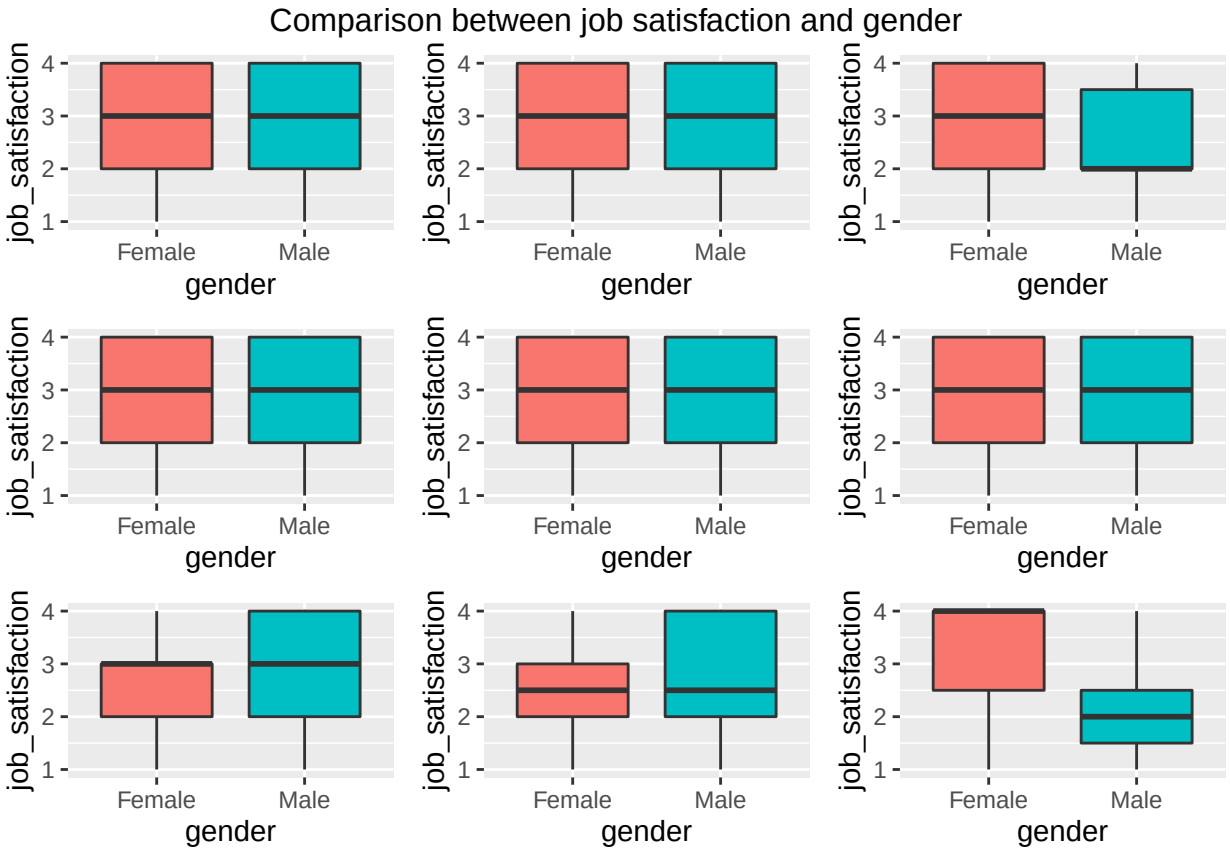
satisfaction of male and female in each department in the different city

```
##Sales  London
g4 <- ggplot(mydb%>%filter(department =="Sales" & city == "London"),
             aes(x= gender, y= job_satisfaction, fill= gender))+
```

```
  geom_boxplot(show.legend = FALSE)

##Sales  Barcelona
g5 <- ggplot(mydb%>%
               filter(department =="Sales" & city == "Barcelona"),
             aes(x= gender, y=job_satisfaction, fill= gender))+
  geom_boxplot(show.legend = FALSE)

##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)

##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)

##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)

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

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

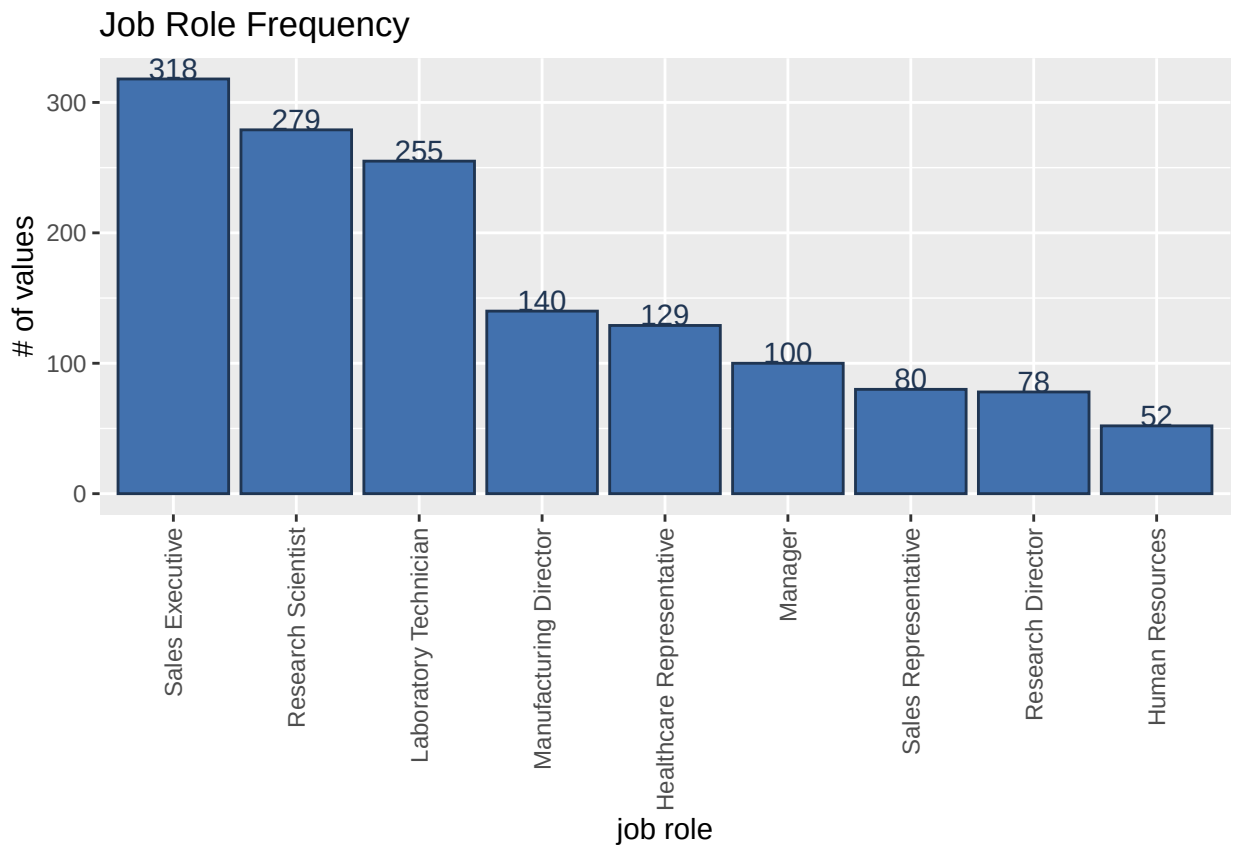
Differences between the two cities in term of job satisfaction male and female




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

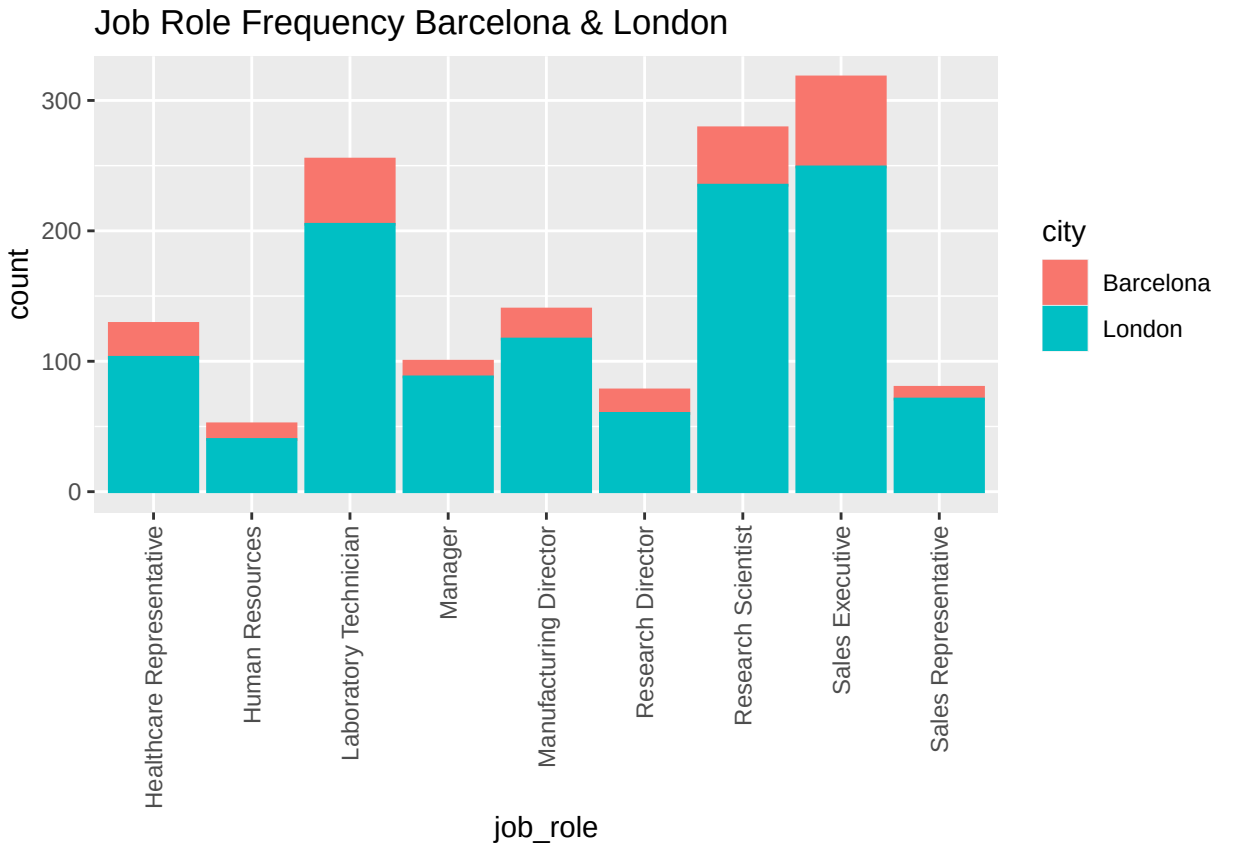
```
# This would help to analyse the number of types of jobs and compare with types
k <- mydb %>% tabyl(job_role)

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")
```

Plot job role frequency

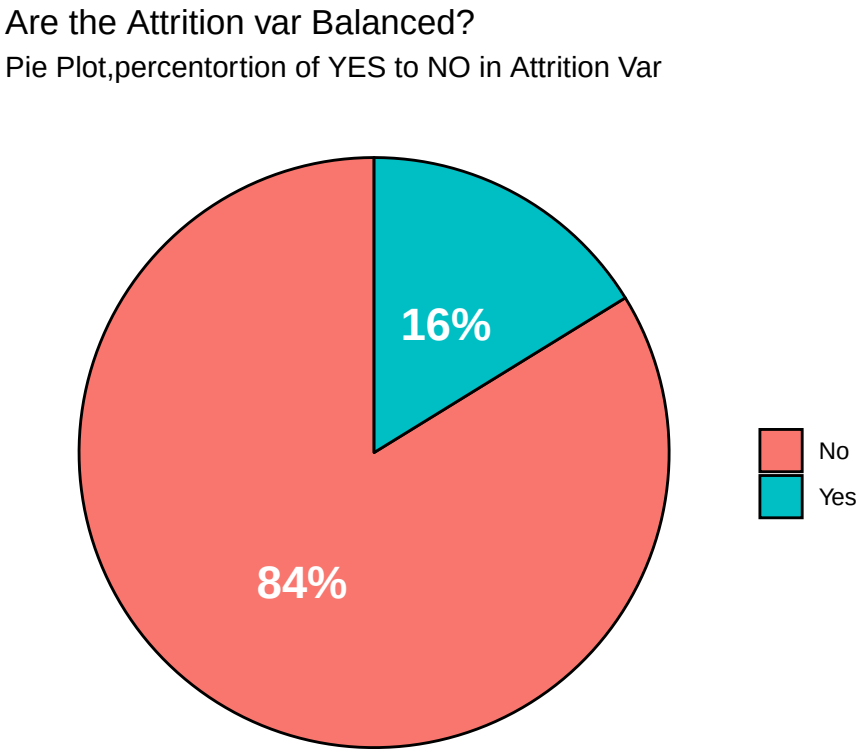



```
rm("k")
```

```
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")
```



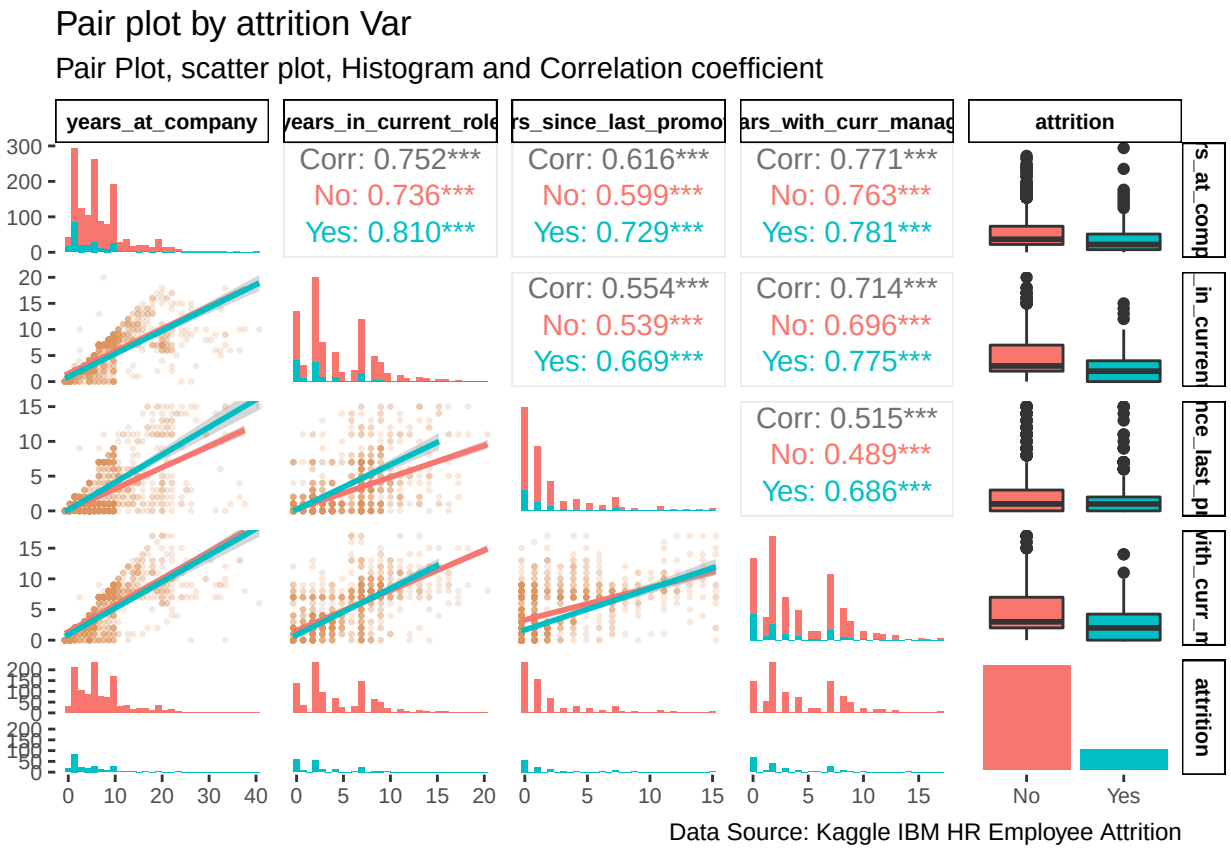

```
# Create a temporary dataset to get the percentage of employees that would leave the company
temp <- mydb %>%
  group_by(attrition) %>%
  summarize(counts = n()) %>%
  mutate(percent = percentage_func(counts)) %>%
  arrange(desc(percent))

# 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")
```

```
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 = "Data Source: Kaggle IBM HR Employee Attrition",
    x = NULL,
    y = NULL
  )
```



Conclusions

...

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