Midterm Project

An Huynh

June 11th, 2023

## R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

# 1. Introduction

**Problem**: Before the pandemic, the company experienced an increasing rate of employee departures. Employee attrition is a challenge to a tech company as certain skills are needed to ensure quality products and smooth operations. Hence, we are looking at multiple factors from the HR dataset and attempting to find the correlation between those factors and the leaving company status to discover which ones contribute to employees’ quitting.

**Summary**: This dataset has 13 variables and 1100 observations. The variables consist of leaving status, salary, commuting distance, job levels, departments, weekly hours, frequency of business travel, years at the company, years since promotion, job satisfaction, performance ratings, and marital status.

**Analyses summary**: In this analysis report, we examine the relationship between leaving and six other factors - salary, commuting distance, job satisfaction, performance ratings, weekly hours, and frequency of business travel. Among the six factors, salary, performance ratings, and working hours per week are strong indicators for leaving. Interestingly, the data also show that sales and product development departments have the highest turnover rates.

**Methods**: This report comprise of summary tables and data visualizations. Main packages used in this project include: ‘readxl’ to import .xlsx file into R; ‘dplyr’ for pipe operating; ‘ggplot2’ for data visualization; and ‘skimr’ to skim through data.

# 2. Data and Model

library(readxl)  
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(ggplot2)  
library(skimr)  
  
employees <- read\_xlsx('/Users/AnhHuynh/Documents/SUMMER 2023/MIS 431/Mid-term project/Employee Data Formatted - MIS 431 Summer 2023 (1100 Records) f.xlsx')  
  
employees <- employees %>%  
 mutate\_if(is.character, as.factor)  
str(employees)

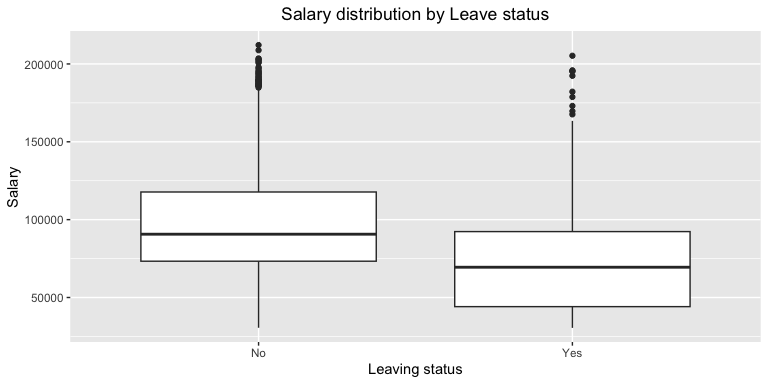
## tibble [1,100 × 13] (S3: tbl\_df/tbl/data.frame)  
## $ left\_company : Factor w/ 2 levels "No","Yes": 2 1 2 1 1 1 1 1 1 1 ...  
## $ department : Factor w/ 6 levels "Finance and Operations",..: 6 6 4 2 6 3 3 6 1 4 ...  
## $ job\_level : Factor w/ 5 levels "Associate","Director",..: 2 4 1 2 1 4 4 2 4 2 ...  
## $ salary : num [1:1100] 118681 85576 46236 117227 36635 ...  
## $ weekly\_hours : num [1:1100] 56 42 56 50 46 48 44 47 50 51 ...  
## $ business\_travel : Factor w/ 3 levels "Frequently","None",..: 3 1 3 1 3 1 3 3 1 3 ...  
## $ yrs\_at\_company : num [1:1100] 6 10 0 8 2 7 1 1 9 7 ...  
## $ yrs\_since\_promotion: num [1:1100] 0 1 0 3 2 3 0 0 1 7 ...  
## $ previous\_companies : num [1:1100] 5 5 6 1 2 1 3 3 3 3 ...  
## $ job\_satisfaction : Factor w/ 4 levels "High","Low","Medium",..: 4 3 1 1 3 4 2 1 1 1 ...  
## $ performance\_rating : Factor w/ 5 levels "Exceeds Expectations",..: 3 1 4 2 1 3 3 2 3 1 ...  
## $ marital\_status : Factor w/ 3 levels "Divorced","Married",..: 3 2 3 2 2 3 2 1 3 2 ...  
## $ miles\_from\_home : num [1:1100] 1 8 2 3 2 2 3 24 23 27 ...

# Analysis 1

## Question: Is there a correlation between leaving and salary?

**Answer**: Yes, both summary tables and data visualization show that median salary for those who leave is lower than those who do not. In addition, the boxplot shows that salary of 25% of those who leave are less than $50000 while that figure is roughly $75000 for 25% of those who stay. Seventy-five percent of those who don’t leave earn nearly $125000 compared to roughly $100000 of 75% of those who leave.

## # A tibble: 2 × 6  
## left\_company employee\_counts min\_salary median\_salary max\_salary less\_60  
## <fct> <int> <dbl> <dbl> <dbl> <int>  
## 1 No 915 30559. 90649. 212135. 91  
## 2 Yes 185 30488. 69437. 205267. 63

 ***Summary of results***: Statistics and data visualizations indicate that salary is a reason for leaving. The 25th, 50th and 75th salary percentiles are lower for those who leave compared to those who don’t. Hence, it is recommended that managers review payment structure to ensure fairness and equality among emmployees.

# Analysis 2

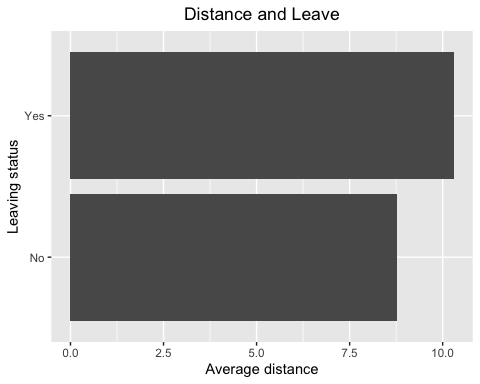
## Question: Is commutting distance a reason for leaving?

**Answer**: Distance is not an indicator of leaving. The table shows that the minimum and maximum commuting distance are the same among those who leave and do not leave. The average distance is about 1 mile different between the two groups.

# Summary table  
factor2 <- employees %>%  
 group\_by(left\_company) %>%  
 summarize (average\_dist = mean(miles\_from\_home),  
 min\_dist = min(miles\_from\_home),  
 max\_dist = max(miles\_from\_home))  
factor2

## # A tibble: 2 × 4  
## left\_company average\_dist min\_dist max\_dist  
## <fct> <dbl> <dbl> <dbl>  
## 1 No 8.78 1 29  
## 2 Yes 10.3 1 29

#Data visualization  
ggplot(factor2, aes(x=left\_company,y=average\_dist)) +   
 geom\_col() +  
 coord\_flip() +  
 labs (title = "Distance and Leave", x="Leaving status", y = "Average distance") +  
 theme(plot.title = element\_text(hjust=0.5))

 ***Summary of result***: Since the average distance between the two groups is just a mile difference, we can conclude that commuting distance is not a reason for leaving. Therefore, managers can keep hiring people living far from office as long as they are willing to commute.

# Analysis 3

## Question: Do people leave because of low job satisfaction?

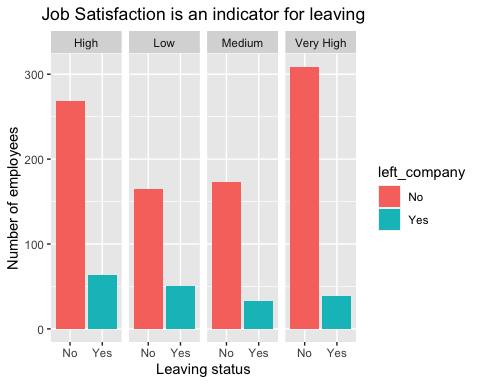
**Answer**: Yes, low job satisfaction is a reason for leaving. In particular, for those who leave, low job satisfaction accounts for 27%, while this figure is just 18% for those who stay. In addition, the difference in percentage for ‘very high’ job satisfacion is 12.73% between employees who leave and those who don’t.

#Summary table  
employees %>%  
 group\_by(left\_company, job\_satisfaction) %>%  
 summarize(n=n()) %>%  
 arrange(job\_satisfaction)

## `summarise()` has grouped output by 'left\_company'. You can override using the  
## `.groups` argument.

## # A tibble: 8 × 3  
## # Groups: left\_company [2]  
## left\_company job\_satisfaction n  
## <fct> <fct> <int>  
## 1 No High 268  
## 2 Yes High 63  
## 3 No Low 165  
## 4 Yes Low 50  
## 5 No Medium 173  
## 6 Yes Medium 33  
## 7 No Very High 309  
## 8 Yes Very High 39

#Data visuaization  
ggplot(employees, aes(x=left\_company, fill=left\_company)) +  
 geom\_bar()+  
 facet\_grid(.~job\_satisfaction) +   
 labs(title = "Job Satisfaction is an indicator for leaving",  
 x="Leaving status", y = "Number of employees") +  
 theme(plot.title = element\_text(hjust=0.5))

 ***Summary of result***: There is not much difference in the middle level, ‘Medium’ job satisfaction, but there are extremes on the two ends of this ordinal category. We can conclude that employees who have ‘Very high’ job satisfaction tend to stay and those who have ‘Low’ job satisfaction tend to leave. However, since the data set does not include factors related to job satisfaction, we cannot make any further conclusion. More research needs to be done to find out which factors impact job satisfaction. Is it working environment, benefits, management, coworkers or anything else.

# Analysis 4

## Question: Which department has the highest rate of leaving?

**Answers**: Sales and Product Development departments have the highest turnover rate - 29%. Research department has the lowest turnover rate, only 4%.

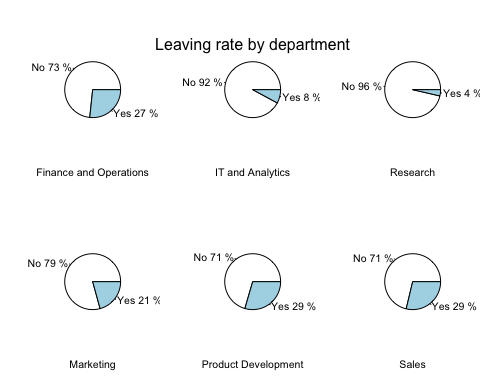
factor4 <-employees%>%  
 group\_by(department,left\_company) %>%  
 summarize(count = n()) %>%  
 mutate(percent = round(count/sum(count),2)\*100, "%") %>%  
 arrange(-percent)

## `summarise()` has grouped output by 'department'. You can override using the  
## `.groups` argument.

factor4

## # A tibble: 12 × 5  
## # Groups: department [6]  
## department left\_company count percent `"%"`  
## <fct> <fct> <int> <dbl> <chr>  
## 1 Research No 214 96 %   
## 2 IT and Analytics No 269 92 %   
## 3 Marketing No 134 79 %   
## 4 Finance and Operations No 66 73 %   
## 5 Product Development No 98 71 %   
## 6 Sales No 134 71 %   
## 7 Product Development Yes 41 29 %   
## 8 Sales Yes 54 29 %   
## 9 Finance and Operations Yes 24 27 %   
## 10 Marketing Yes 35 21 %   
## 11 IT and Analytics Yes 23 8 %   
## 12 Research Yes 8 4 %

attach(employees)  
par(mfrow=c(2,3))  
count1 <- table(employees$left\_company[department=="Finance and Operations"])  
lb1<- paste(c('No', 'Yes'),round(count1/sum(count1),2)\*100,'%')  
pie(count1, labels = lb1,xlab="Finance and Operations")  
count2 <-table(employees$left\_company[department == "IT and Analytics"])  
lb2 <- paste(c('No','Yes'),round(count2/sum(count2),2)\*100, '%')  
pie(count2, labels = lb2, xlab="IT and Analytics")  
mtext(side = 3, text = "Leaving rate by department")  
count3 <- table(employees$left\_company[department=="Research"])  
lb3 <- paste(c('No','Yes'), round(count3/sum(count3),2)\*100, '%')  
pie(count3, labels = lb3, xlab="Research")  
count4 <- table(employees$left\_company[department == "Marketing"])  
lb4 <- paste(c('No','Yes'), round(count4/sum(count4),2)\*100, '%')  
pie(count4, labels = lb4, xlab="Marketing")  
count5 <- table(employees$left\_company[department == "Product Development"])  
lb5 <- paste(c('No','Yes'), round(count5/sum(count5),2)\*100, '%')  
pie(count5, labels = lb5, xlab="Product Development")  
count6 <- table(employees$left\_company[department == "Sales"])  
lb6 <- paste(c('No','Yes'), round(count6/sum(count6),2)\*100, '%')  
pie(count6, labels = lb6, xlab="Sales")

 ***Summary of results***: Pie charts show the percentage of leaving for each department. Sales and Product development departments have the highest leaving rate. One possible explanation is that commission or compensation is insufficient compared to workload. There may be other reasons requiring further investigation from department managers and top leaders to reduce attrition rates from these two departments.

# Analysis 5

## Question: What is the relationship between performance rating and leaving status?

**Answer**: Employees who leave tend to have low to medium performance, and those who stay have medium to very high performance.

# Summary table  
factor5\_1 <- employees %>%  
 filter(left\_company=="Yes") %>%  
 group\_by(performance\_rating) %>%  
 summarize(n\_employee = n()) %>%  
 arrange(performance\_rating)  
factor5\_1

## # A tibble: 5 × 2  
## performance\_rating n\_employee  
## <fct> <int>  
## 1 Exceeds Expectations 29  
## 2 Exceptional 2  
## 3 Meets Expectations 79  
## 4 Minimally Effective 62  
## 5 Not Effective 13

factor5\_2 <- employees %>%  
 filter(left\_company=="No") %>%  
 group\_by(performance\_rating) %>%  
 summarize(n\_employee = n()) %>%  
 arrange(performance\_rating)  
factor5\_2

## # A tibble: 5 × 2  
## performance\_rating n\_employee  
## <fct> <int>  
## 1 Exceeds Expectations 337  
## 2 Exceptional 202  
## 3 Meets Expectations 317  
## 4 Minimally Effective 48  
## 5 Not Effective 11

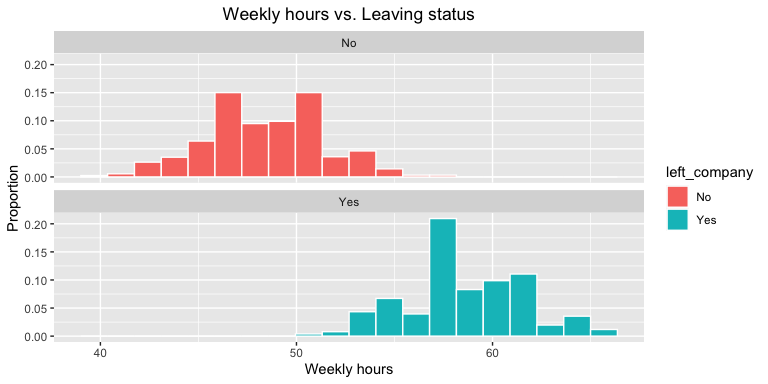
***Summary of results***: Comparing the stats between two summary tables, the percentage of employees whose performance is beyond expectations and exceptional is two and twenty times higher for those who stay compared to those who leave. Besides, only 1.2% of those who don’t leave perform poorly compared to 4.2% of those who leave. Retaining too many low-performance workers can impede the growth of the company in general. However, more investigation needs to be executed to identify why employees perform not so well, and what managers can do to help. Some may need more training and coaching to excel in their works. Identifying current employees’ problems and help them advance will save the company money compared to recruiting and training new employees from scratch.

# Analysis 6

## Question : Is number of hours working a reason for leaving?

**Answer**: Yes, long weekly hours is a strong indicator for leaving.

## # A tibble: 2 × 4  
## left\_company avg\_hr min\_hr max\_hr  
## <fct> <dbl> <dbl> <dbl>  
## 1 No 48.4 40 58  
## 2 Yes 58.7 51 66

 ***Summary of result***: Statistics and the histograms strongly indicate that long working hours contribute to the reason for leaving. Average, minimum, and maximum working hours of those who leave are approximately 10 hours more than those who don’t. Combined this result with the salary analysis, we can say that the employees who leave work longer hours for lower salary. These combined factors can negatively impact their job performance. To this point, managers can see there are many red flags within the business operations.

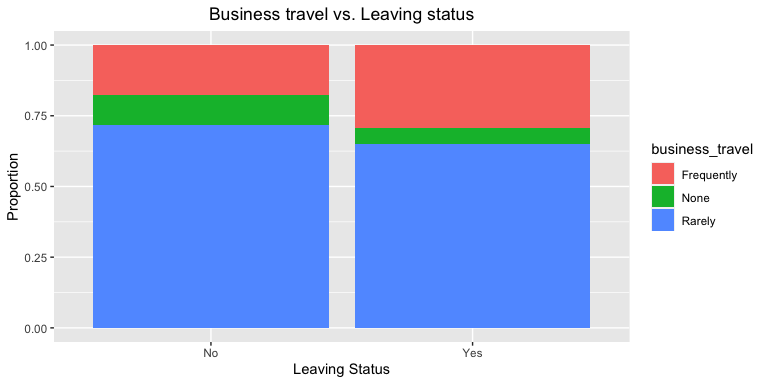
# Analysis 7

## Question: Is business travel frequency contribute to leaving?

**Answer**: This statistics alone may not reveal much about reason for leaving. However, when combined with other factors analyzed above, frequency of business travel do contribute to reasons for leaving.

## `summarise()` has grouped output by 'left\_company'. You can override using the  
## `.groups` argument.

## # A tibble: 6 × 3  
## # Groups: left\_company [2]  
## left\_company business\_travel n\_employee  
## <fct> <fct> <int>  
## 1 No Frequently 160  
## 2 Yes Frequently 54  
## 3 No None 98  
## 4 Yes None 11  
## 5 No Rarely 657  
## 6 Yes Rarely 120

 ***Summary of result***: Statistics and stacked bar chart point out that employees who leave traveled more frequently than those who didn’t. Frequent business travel may impact the work-life balance in a negative way. Remember, the people who quit are those who work longer hours for lower salary, and it turns out they have more business trips than those who stay.

# Analysis 8

## Question: What is the distribution of salary among departments?

## # A tibble: 6 × 4  
## department avg\_salary lowest\_salary highest\_salary  
## <fct> <dbl> <dbl> <dbl>  
## 1 Finance and Operations 96138. 37115. 212135.  
## 2 IT and Analytics 93782. 33541. 208804.  
## 3 Marketing 91809. 30559. 203186.  
## 4 Product Development 91146. 32444. 195345.  
## 5 Research 98980. 33277. 203529.  
## 6 Sales 89920. 30488. 197965.

 ***Summary of result***: On average, Sales department has the lowest salary, followed by Product Development department. Sales employees also earn the lowest among the minimum salary group and Product Development employees earn the lowest among the maximum salary group. These salary statistics match the leaving rates for these two departments analyzed above.

# 3. Conclusion

* In conclusion, the combination of salary, job satisfaction, performance ratings, weekly hours and frequency of business travel contribute to reasons for leaving.
* A pattern detected from this analysis is that employees who left worked longer hours, traveled more frequently, but had lower salary. Those who leave work 10 hours more weekly and travel more frequenty while earning about $30000 less than those who don’t. The combination of these three factors may explain for low job satisfaction and low performance rating. It’s no surprise that those who stuck in this cycle leave the company eventually.
* The two departments that witness high turnover rates are Sales and Product Development. Employees from these two departments also earn the lowest among the lowest earning group and the highest earning group, respectively. Reasons can’t be detected from the dataset, but one assumption is that the commission or compensation is insufficient.
* Managers at all level need to conduct a thorough internal investigation to find out the loopholes in the business structure and have specific strategies to fix them to reduce attrition rates.

# 4. Appendix

knitr::opts\_chunk$set(echo = TRUE)  
library(readxl)  
library(dplyr)  
library(ggplot2)  
library(skimr)  
  
employees <- read\_xlsx('/Users/AnhHuynh/Documents/SUMMER 2023/MIS 431/Mid-term project/Employee Data Formatted - MIS 431 Summer 2023 (1100 Records) f.xlsx')  
  
employees <- employees %>%  
 mutate\_if(is.character, as.factor)  
str(employees)  
# Summary table  
employees %>%  
 group\_by(left\_company) %>%  
 summarize(employee\_counts = n(),  
 min\_salary = min(salary),  
 median\_salary = median(salary),  
 max\_salary = max(salary),  
 less\_60= sum(salary <= 60000))  
   
   
# Data visualization  
ggplot(employees, aes(x=left\_company,y=salary)) +  
 geom\_boxplot() +  
 labs(title = "Salary distribution by Leave status",  
 x = "Leaving status", y="Salary") +  
 theme(plot.title = element\_text(hjust = 0.5))  
# Summary table  
factor2 <- employees %>%  
 group\_by(left\_company) %>%  
 summarize (average\_dist = mean(miles\_from\_home),  
 min\_dist = min(miles\_from\_home),  
 max\_dist = max(miles\_from\_home))  
factor2  
#Data visualization  
ggplot(factor2, aes(x=left\_company,y=average\_dist)) +   
 geom\_col() +  
 coord\_flip() +  
 labs (title = "Distance and Leave", x="Leaving status", y = "Average distance") +  
 theme(plot.title = element\_text(hjust=0.5))  
#Summary table  
employees %>%  
 group\_by(left\_company, job\_satisfaction) %>%  
 summarize(n=n()) %>%  
 arrange(job\_satisfaction)  
  
#Data visuaization  
ggplot(employees, aes(x=left\_company, fill=left\_company)) +  
 geom\_bar()+  
 facet\_grid(.~job\_satisfaction) +   
 labs(title = "Job Satisfaction is an indicator for leaving",  
 x="Leaving status", y = "Number of employees") +  
 theme(plot.title = element\_text(hjust=0.5))  
factor4 <-employees%>%  
 group\_by(department,left\_company) %>%  
 summarize(count = n()) %>%  
 mutate(percent = round(count/sum(count),2)\*100, "%") %>%  
 arrange(-percent)  
factor4  
attach(employees)  
par(mfrow=c(2,3))  
count1 <- table(employees$left\_company[department=="Finance and Operations"])  
lb1<- paste(c('No', 'Yes'),round(count1/sum(count1),2)\*100,'%')  
pie(count1, labels = lb1,xlab="Finance and Operations")  
count2 <-table(employees$left\_company[department == "IT and Analytics"])  
lb2 <- paste(c('No','Yes'),round(count2/sum(count2),2)\*100, '%')  
pie(count2, labels = lb2, xlab="IT and Analytics")  
mtext(side = 3, text = "Leaving rate by department")  
count3 <- table(employees$left\_company[department=="Research"])  
lb3 <- paste(c('No','Yes'), round(count3/sum(count3),2)\*100, '%')  
pie(count3, labels = lb3, xlab="Research")  
count4 <- table(employees$left\_company[department == "Marketing"])  
lb4 <- paste(c('No','Yes'), round(count4/sum(count4),2)\*100, '%')  
pie(count4, labels = lb4, xlab="Marketing")  
count5 <- table(employees$left\_company[department == "Product Development"])  
lb5 <- paste(c('No','Yes'), round(count5/sum(count5),2)\*100, '%')  
pie(count5, labels = lb5, xlab="Product Development")  
count6 <- table(employees$left\_company[department == "Sales"])  
lb6 <- paste(c('No','Yes'), round(count6/sum(count6),2)\*100, '%')  
pie(count6, labels = lb6, xlab="Sales")  
# Summary table  
factor5\_1 <- employees %>%  
 filter(left\_company=="Yes") %>%  
 group\_by(performance\_rating) %>%  
 summarize(n\_employee = n()) %>%  
 arrange(performance\_rating)  
factor5\_1  
  
factor5\_2 <- employees %>%  
 filter(left\_company=="No") %>%  
 group\_by(performance\_rating) %>%  
 summarize(n\_employee = n()) %>%  
 arrange(performance\_rating)  
factor5\_2  
# Summary table  
factor6 <- employees %>%  
 group\_by(left\_company) %>%  
 summarize(avg\_hr = mean(weekly\_hours),  
 min\_hr = min(weekly\_hours),  
 max\_hr = max(weekly\_hours))  
factor6  
  
# Data visualization  
ggplot(employees, aes(x=weekly\_hours, fill=left\_company)) +  
 geom\_histogram(aes(y=after\_stat(density)),color="white",bins = 20) +  
 facet\_wrap(~left\_company, nrow = 2) +  
 labs(x="Weekly hours", y="Proportion", title = "Weekly hours vs. Leaving status") +  
 theme(plot.title = element\_text(hjust=0.5))  
#Summary table  
factor6 <- employees %>%  
 group\_by(left\_company,business\_travel) %>%  
 summarize(n\_employee = n()) %>%  
 arrange(business\_travel)  
factor6  
  
# Data visualization  
ggplot(factor6, aes(x=left\_company,y=n\_employee, fill=business\_travel)) +  
 geom\_bar(position = "fill", stat = "identity") +   
 labs(title = "Business travel vs. Leaving status",  
 x = "Leaving Status",  
 y = "Proportion") +  
 theme(plot.title = element\_text(hjust = 0.5))  
#Summary table  
factor7 <- employees %>%  
 group\_by(department) %>%  
 summarize(avg\_salary = mean(salary),  
 lowest\_salary = min(salary),  
 highest\_salary = max(salary))  
factor7  
  
#Data visualization  
ggplot(employees, aes(x=salary, fill=left\_company)) +  
 geom\_histogram(bins = 50) +  
 facet\_grid(.~department) +  
 labs(title = "Salary distribution",  
 x = "Salary") +  
 theme(plot.title = element\_text(hjust = 0.5))