Project_1: Which Factor Affects Satisfaction Level The Most?

Vi Le

2024-02-10

Project and Data Introduction

Importing data and necessary libraries.

Data set downloaded from Kaggle: https://www.kaggle.com/datasets/redpen12/employees-satisfaction-analysis
The data used in this project contains employee ID, satisfaction level, last performance evaluation, numbers of project total, average monthly hour, year with company, promotion within last 5 years, work accident, department(10 departments total), and salary type (low, medium, high).

A few libraries needed for this project are:

- Cleaning and analyzing data: dplyr, reshape, reshape2, tidyverse, modelr.
- Visualizing data: ggplot2, RColorBrewer, ggpubr.

Goal and process of project:

- Goal: to determine which factor affects the satisfaction level the most, and the reason why.
- Process:

– Step 1: Compare satisfaction level with last performance evaluation.

– Step 2: Detect outliers (if any).

– Step 3: Show current satisfaction level associates with each factor (department, salary type, work accident, promotion, average monthly hour, time spend with company, number of project) to see which factor has more uneven pattern.

– Step 4: When factor(s) is/are determined from above, show any possible relationship between factors.

– Step 5: Perform deeper analysis from here, based on results in step 4.

```
library(dplyr)
library(ggplot2)
library(reshape2)
library(reshape)
library(RColorBrewer)
library(tidyverse)
library(modelr)
library(ggpubr)
df = read.csv("Employee_Attrition.csv", sep=",", header=TRUE)
```

Summary of data

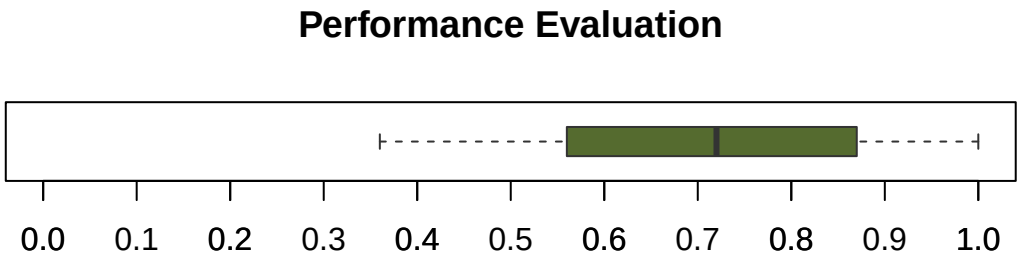
```
##      Emp.ID      satisfaction_level last_evaluation  number_project
##  Min.   :    1   Min.   :0.0900     Min.   :0.3600   Min.   :2.000
##  1st Qu.: 3750   1st Qu.:0.4400     1st Qu.:0.5600   1st Qu.:3.000
##  Median : 7500   Median :0.6400     Median :0.7200   Median :4.000
##  Mean   : 7500   Mean   :0.6128     Mean   :0.7161   Mean   :3.803
##  3rd Qu.:11250   3rd Qu.:0.8200     3rd Qu.:0.8700   3rd Qu.:5.000
##  Max.   :14999   Max.   :1.0000     Max.   :1.0000   Max.   :7.000
##  NA's   :788     NA's   :788        NA's   :788      NA's   :788
##  average_montly_hours time_spend_company Work_accident    promotion_last_5years
##  Min.   : 96.0        Min.   : 2.000     Min.   :0.0000   Min.   :0.0000
##  1st Qu.:156.0        1st Qu.: 3.000     1st Qu.:0.0000   1st Qu.:0.0000
##  Median :200.0        Median : 3.000     Median :0.0000   Median :0.0000
##  Mean   :201.1        Mean   : 3.498     Mean   :0.1446   Mean   :0.0213
##  3rd Qu.:245.0        3rd Qu.: 4.000     3rd Qu.:0.0000   3rd Qu.:0.0000
##  Max.   :310.0        Max.   :10.000     Max.   :1.0000   Max.   :1.0000
##  NA's   :788          NA's   :788        NA's   :788      NA's   :788
##      dept              salary
##  Length:15787       Length:15787
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##
##
##
##
```

```
## [1] 6304
```

```
## [1] 39.93159
```

Step 1: Compare satisfaction level and last evaluation.

```
current_level = df %>%
  select(satisfaction_level, last_evaluation)

attach(current_level)
par(mfrow=c(2,1))
boxplot(current_level$satisfaction_level, horizontal = T,
        main = "Satisfaction Level", col = 'darkolivegreen',
        border='gray20', ylim=c(0,1))
axis(side = 1, at=seq(0,1,by=0.1))

boxplot(current_level$last_evaluation, horizontal = T,
        main = "Performance Evaluation", col = 'darkolivegreen',
        border='gray20', ylim=c(0,1))
axis(side = 1, at=seq(0,1,by=0.1))
```



Step 2: Detect outliers (if any)

```
## numeric(0)
```

```
## numeric(0)
```

```
## # A tibble: 5 x 3
##   Measures Current  Last
##   <chr>      <dbl> <dbl>
## 1 Min        0.09  0.36
## 2 Q1 (25%)   0.44  0.56
## 3 Q3 (75%)   0.82  0.87
## 4 Sd         0.249 0.171
## 5 IQR        0.38  0.31
```

As in 2 box plots and the table shown above, a few things can be concludes:

- There is no outliers in satisfaction level or last performance evaluation.
- The last evaluation result had a better performance than the satisfaction level (higher Min, Q1, Q3).

-> Possible interpretation: Some employees have good performance but are not satisfied with job.

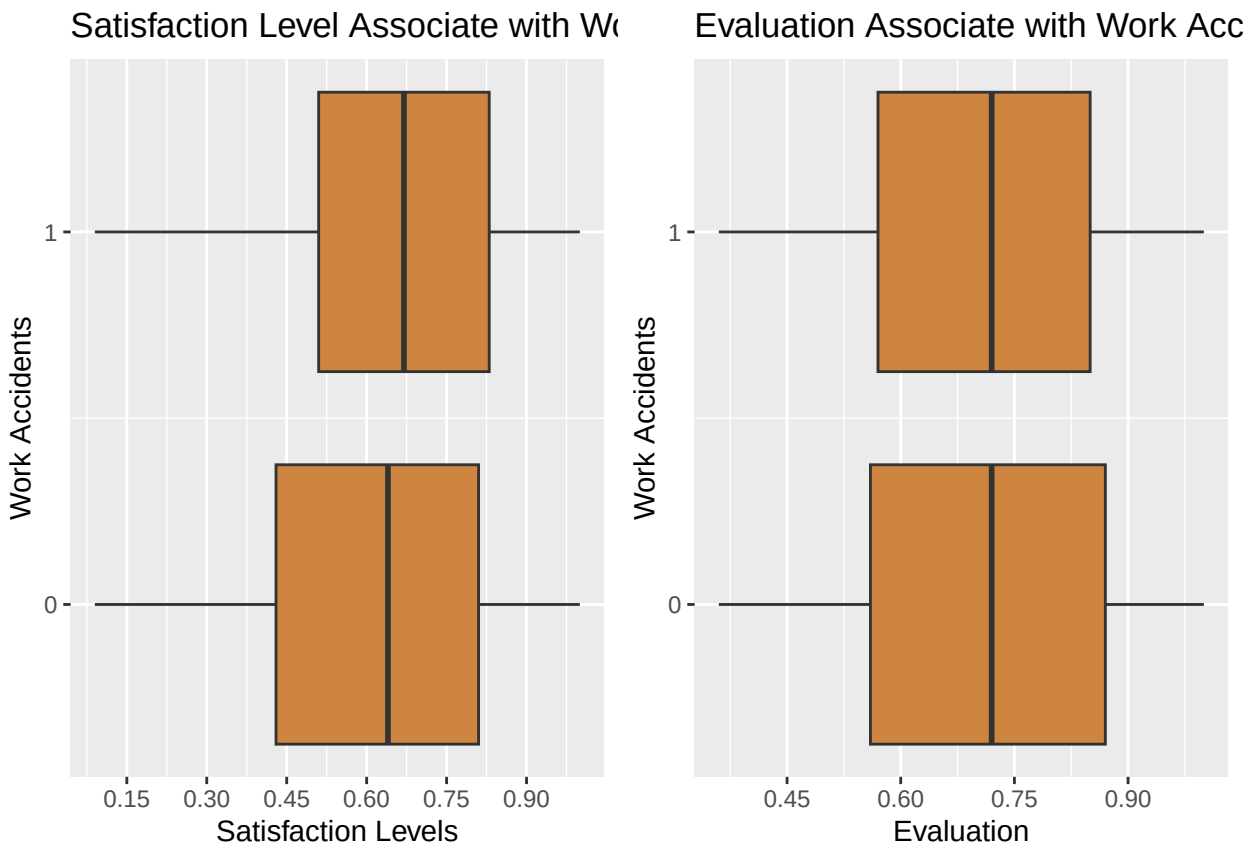
Next step, let's take a look of what could possibly be the reason the current satisfaction level is not match with last evaluation.

Step 3: Show current satisfaction level and last evaluation associate with each factor.

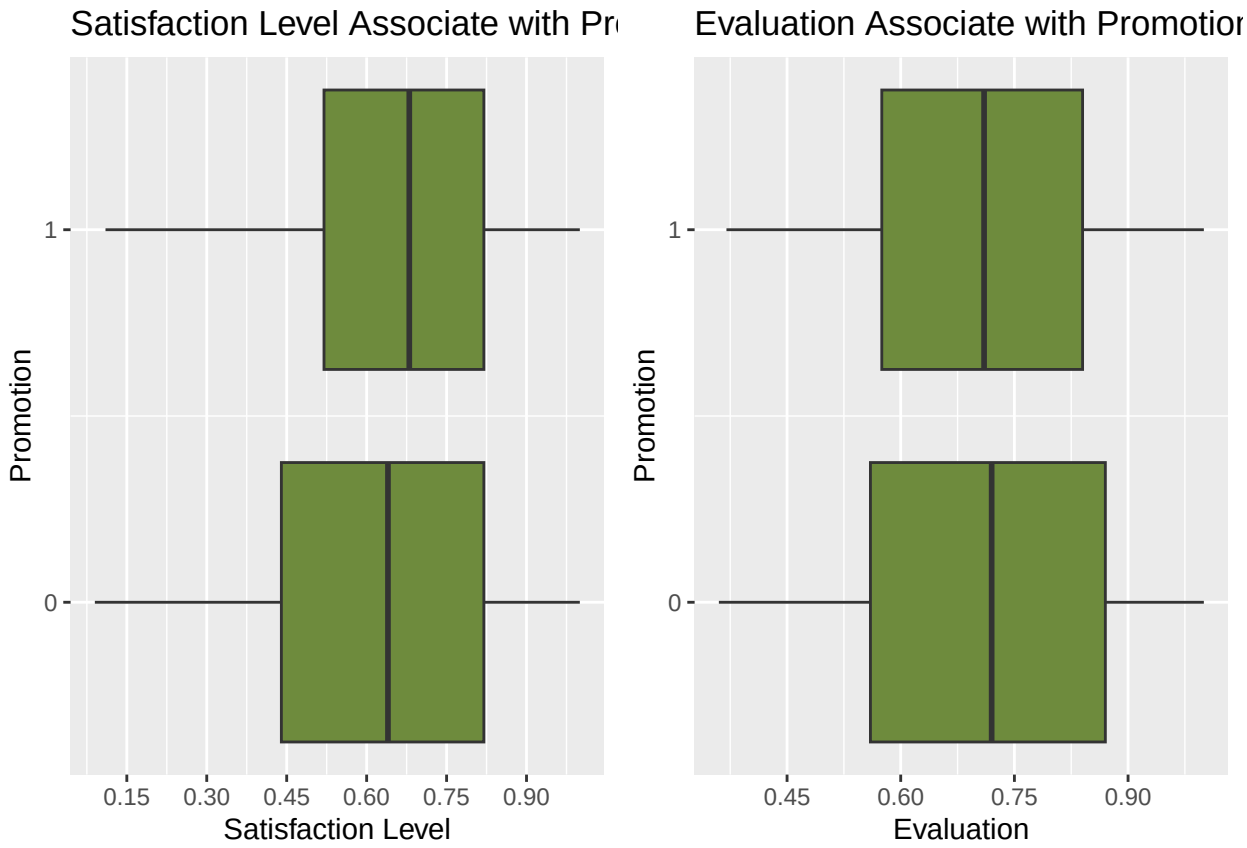
```
## $x
## [1] "Average Hours"
##
## $y
## [1] "Satisfaction Level"
##
## $title
## [1] "Satisfaction Level Associate with Monthly Hours"
##
## attr(,"class")
## [1] "labels"
```

```
## $x
## [1] "Average Hours"
##
## $y
## [1] "Evaluation"
##
## $title
## [1] "Evaluation Associate with Monthly Hours"
##
## attr(,"class")
## [1] "labels"
```

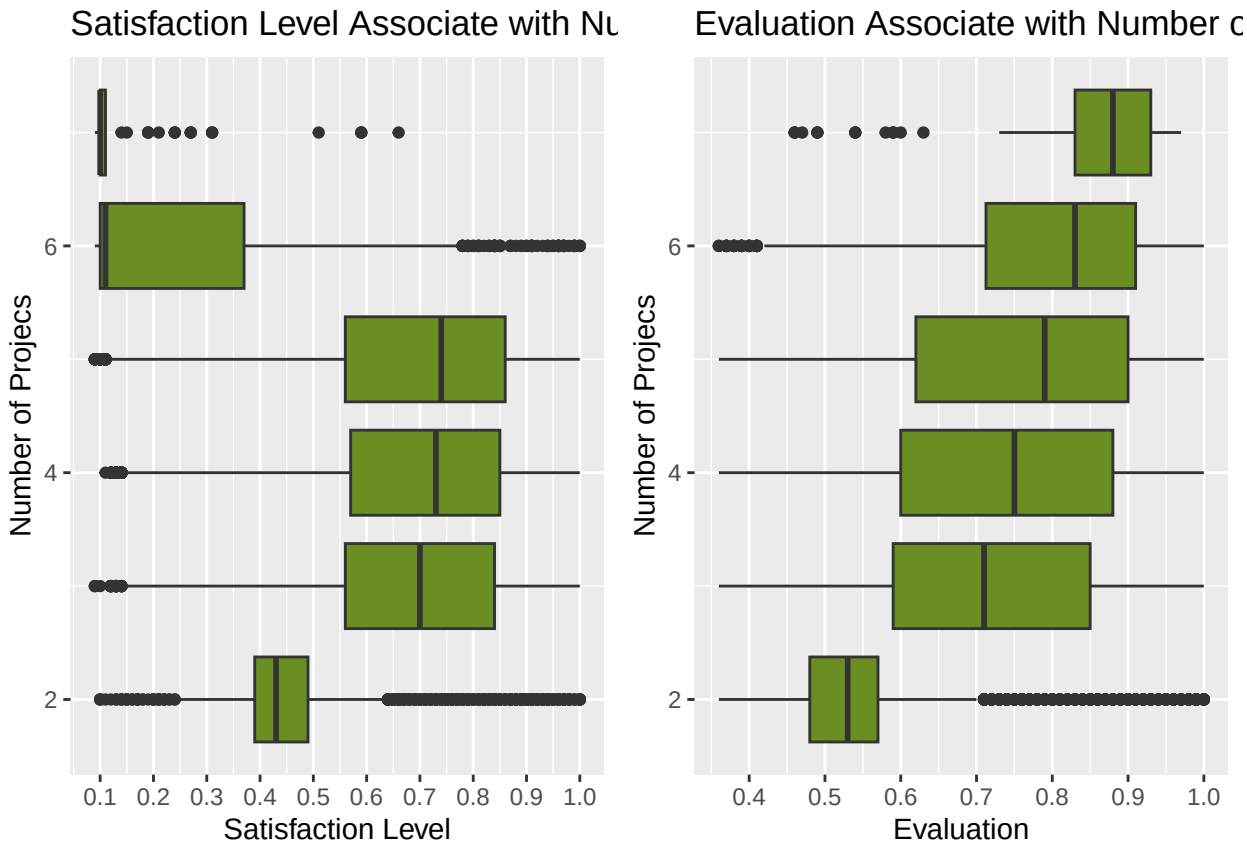
```
ggarrange(p1,p11,ncol=2)
```



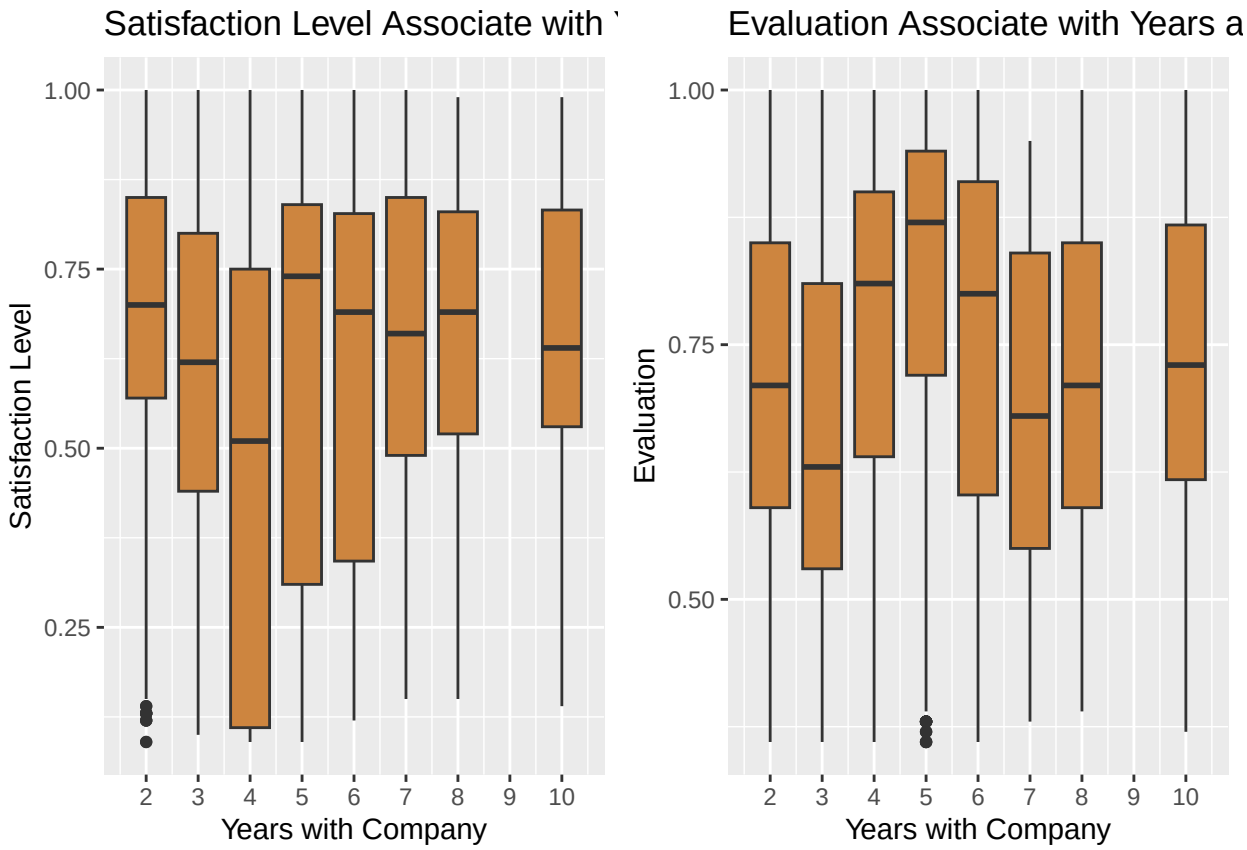

```
ggarrange(p2,p22,ncol=2)
```



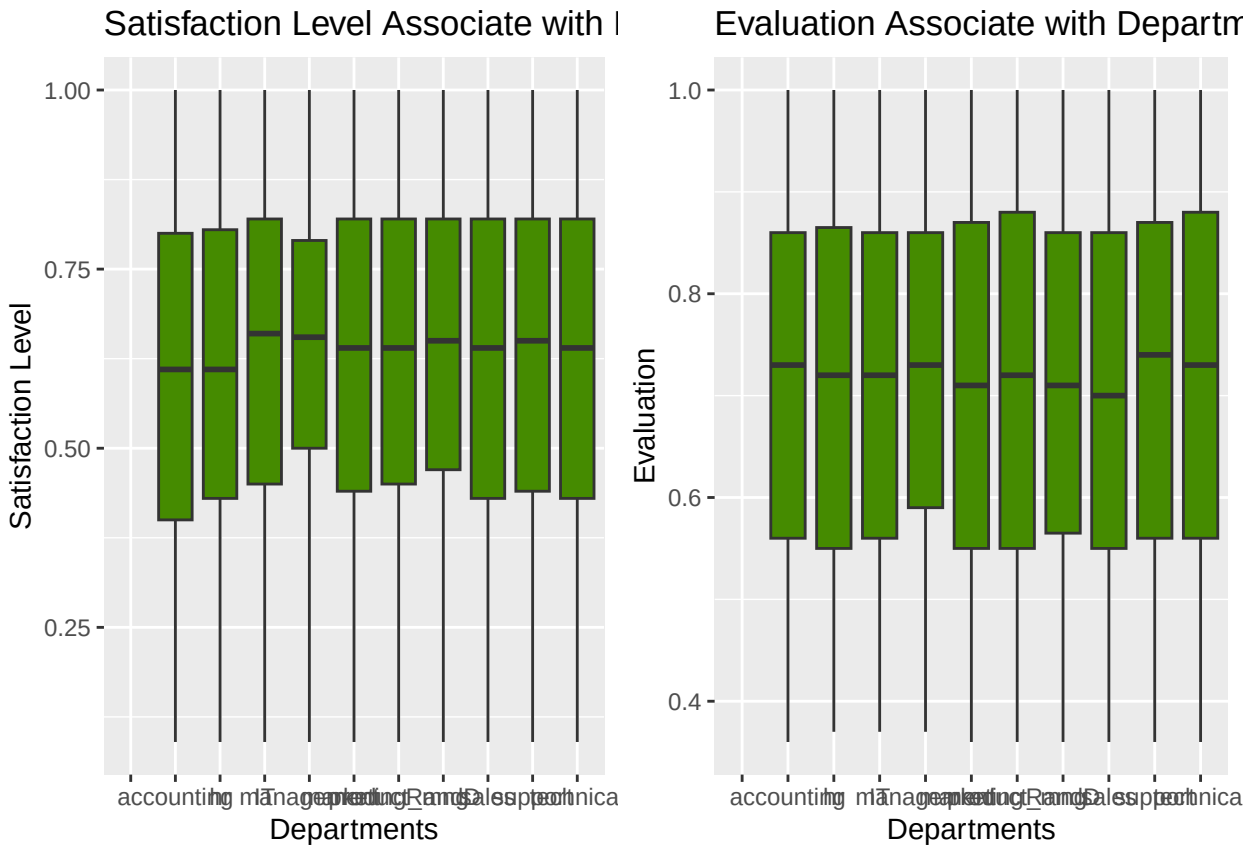

```
ggarrange(p3,p33,ncol=2)
```




```
ggarrange(p4,p44,ncol=2)
```

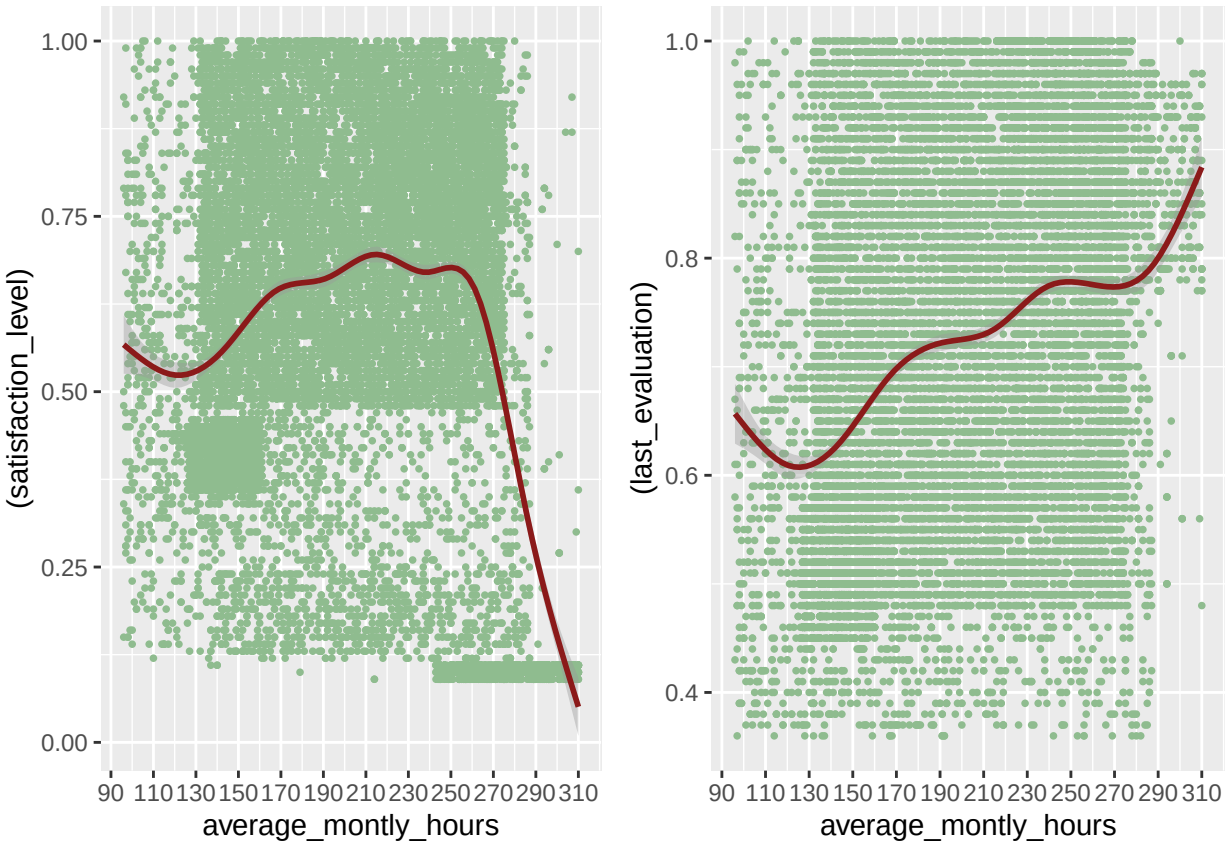



```
ggarrange(p5,p55,ncol=2)
```




```
ggarrange(p6,p66,ncol=2)
```


```
ggarrange(p7,p77,ncol=2)
```



Observation:

– Satisfaction level and last evaluation in the plots of Number of Projects, Years with Company, and Average Monthly Hours has big difference between each groups than compared to other plots.

- In **Number of Project plot, Years with Company plot, and Average Monthly Hours plot**, groups with lowest satisfaction level have significantly higher performance evaluation, in some cases even, have highest evaluation. **WHY?**
- To understand deeper why there is such a contrary between lowest and highest groups of each plot associate with satisfaction level and performance evaluation, let's take a deeper look into mentioned categories.

Step 4: Analysis on Number of Projects, Years with Company, and Average Monthly Hours

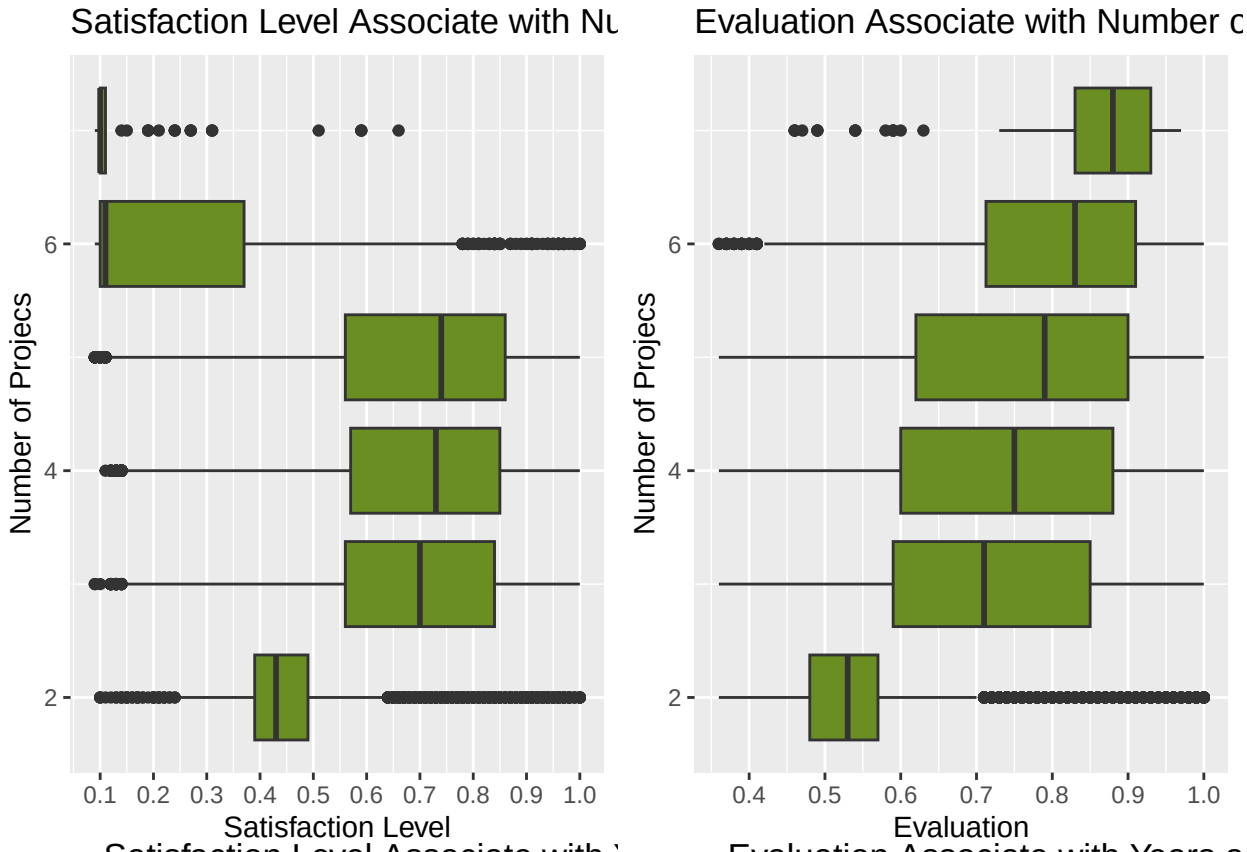
Compute counts of employees associates with number of projects and years with company.

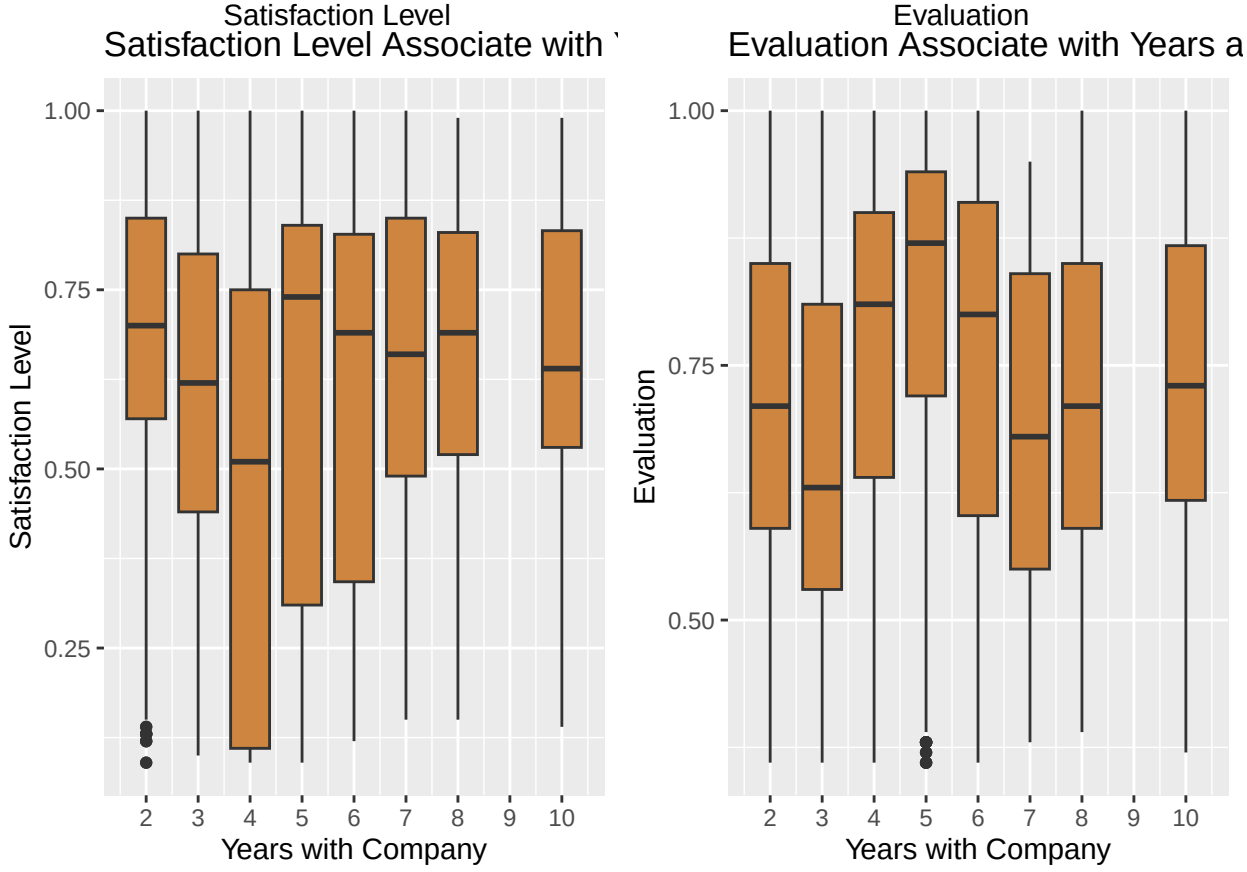
Below is the plot tile plot represents the values of counts of employees associates with number of projects and years with company. The darker the tile, the higher the counts of employees.

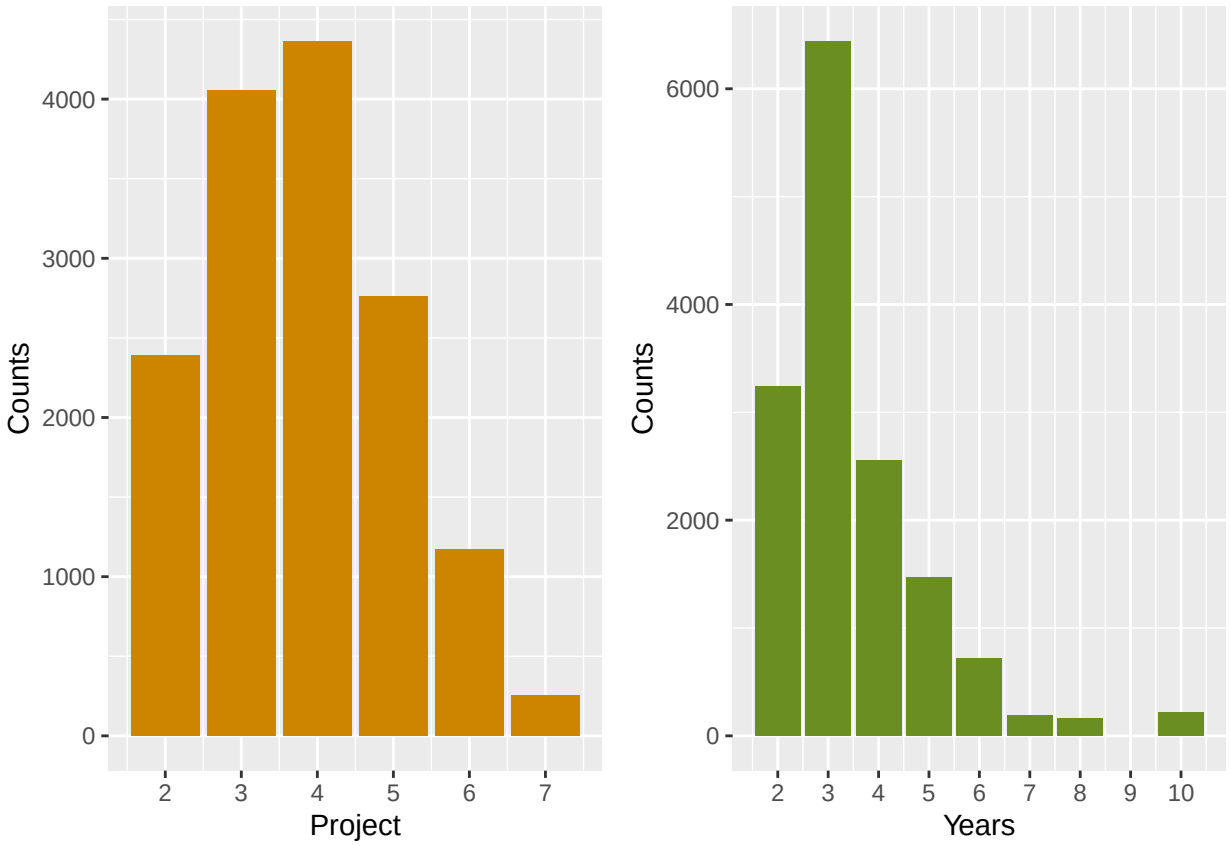
```
green <- brewer.pal(9, "Greens")
green_range <- colorRampPalette(green)

ypr = df%>%
  select(time_spend_company, number_project) %>%
  group_by(number_project) %>%
  count(number_project, time_spend_company)

yprplot = ggplot(data = ypr) +
  geom_tile(mapping = aes(x=number_project, y=time_spend_company, fill=factor(n)),
            show.legend = FALSE, color=NA)+
  scale_fill_manual(values = green_range(40)) +
  labs(title = "Counts of Employees Associates With Years at Company and Number of Projects",
       x="Projects", y="Years") +
  scale_x_continuous(breaks=seq(0,10,1)) +
  scale_y_continuous(breaks=seq(2,12,1))
```

```
yprplot
```


```
##   time_spend_company    2    3    4   5   6   7  NA
## 1                  2  224 1255 1144 554  66   1  NA
## 2                  3 1854 1782 1798 866 136   7  NA
## 3                  4  136  530  577 431 673 210  NA
## 4                  5   83  135  445 592 180  38  NA
## 5                  6   53  139  215 224  87  NA  NA
## 6                  7   16   58   64  38  12  NA  NA
## 7                  8   12   62   46  34   8  NA  NA
## 8                 10   10   94   76  22  12  NA  NA
## 9                 NA   NA   NA   NA  NA  NA  NA 788
```


Satisfaction Level Associate with Nu
Number of Projecs
6
4
2
0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
Satisfaction Level
Evaluation Associate with Number o
Number of Projecs
6
4
2
0.4 0.5 0.6 0.7 0.8 0.9 1.0
Evaluation


Satisfaction Level Associate with
Satisfaction Level
1.00
0.75
0.50
0.25
2 3 4 5 6 7 8 9 10
Years with Company
Evaluation Associate with Years a
Evaluation
1.00
0.75
0.50
2 3 4 5 6 7 8 9 10
Years with Company



Based on the plots:

- Back on the "Count of Employees Associates with Years at Company and Number of Project" plot, even though 4-year group does NOT have the highest count of employees associates with number of projects, but it DOES have the highest employee counts in 6 and 7 projects specifically (6 projects is 673, and 7 projects is 210), compared to the other year group associates with 6 or 7 projects.
- People did 6 and 7 projects have the lowest satisfaction level, but very high performance evaluation.
- People spend 4 years with company have the lowest performance in satisfaction level, but have the second best performance in the last evaluation.

-> Things are getting clearer now. Let's take a closer look into these data.

Step 5: A deeper look

```
sixseven = df %>%
  select(number_project) %>%
  filter(number_project == 6 | number_project == 7) %>%
  count(number_project)

sixseven
```

```
##   number_project    n
## 1              6 1174
## 2              7  256
```

```
(ypr_stat[3, 6]/sixseven[1,2])*100
```

```
## [1] 57.32538
```

```
(ypr_stat[3, 7]/sixseven[2,2])*100
```

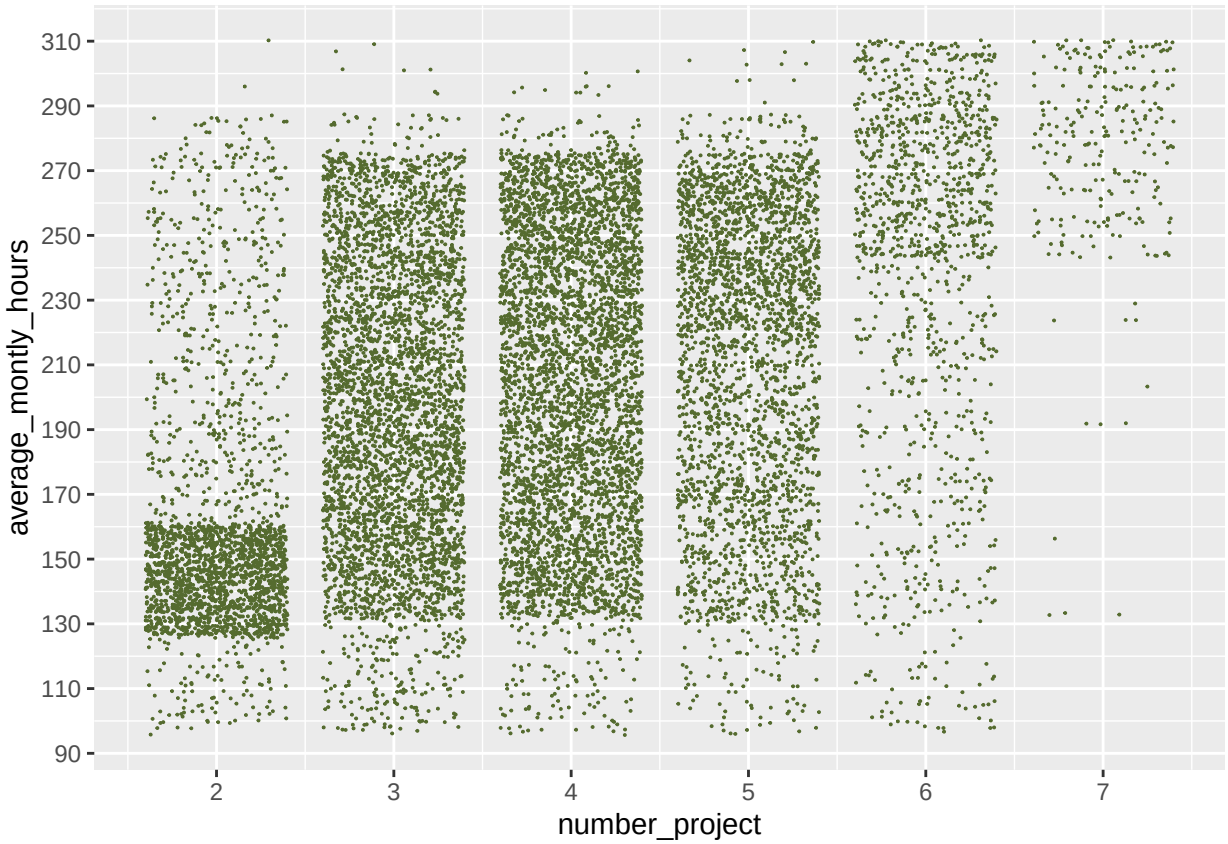
```
## [1] 82.03125
```

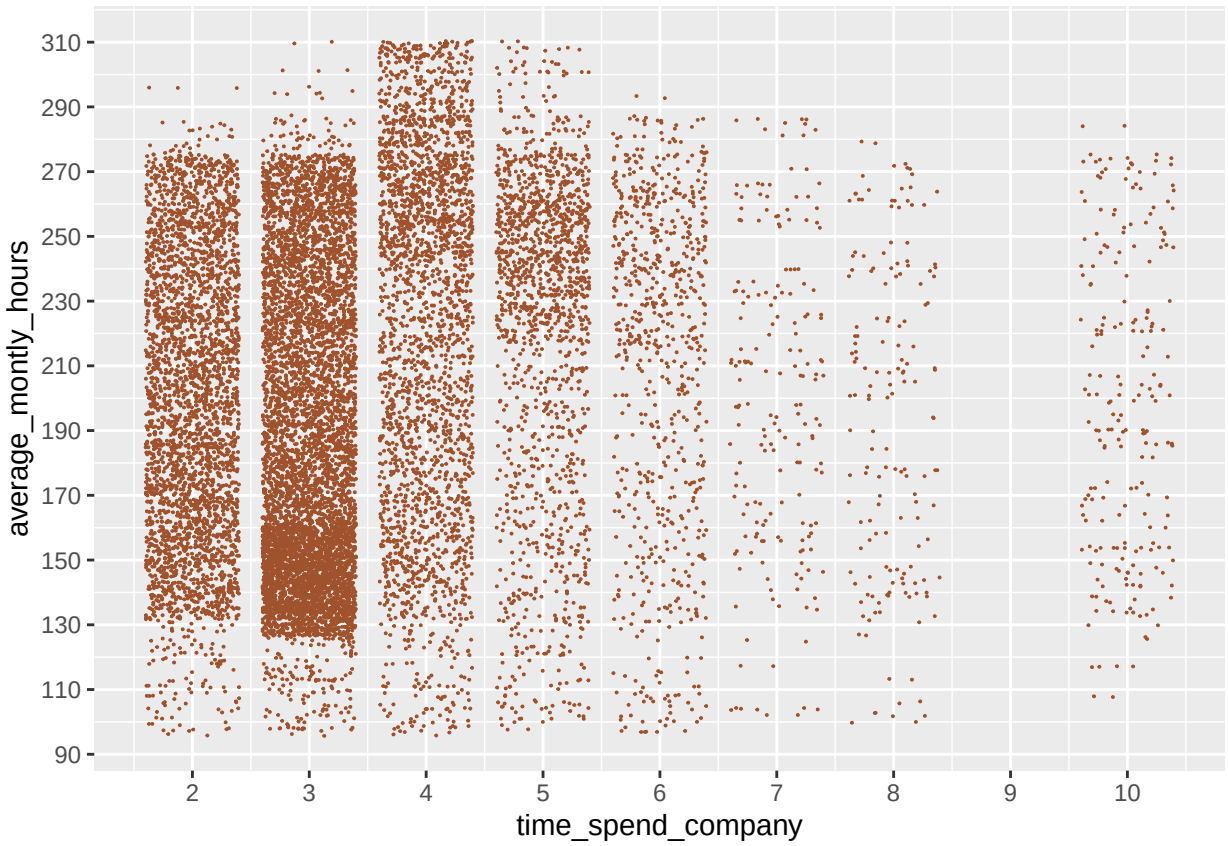
```
## # A tibble: 1,041 x 3
## # Groups:   number_project [7]
##    number_project average_montly_hours     n
##             <int>                 <int> <int>
##  1              2                    96     1
##  2              2                    98     3
##  3              2                    99     2
##  4              2                   100     6
##  5              2                   101     3
##  6              2                   102     6
##  7              2                   103     6
##  8              2                   104     5
##  9              2                   105     3
## 10              2                   106     3
## # i 1,031 more rows
```

```
## $x
## [1] "Projects"
##
## $y
## [1] "Hours"
##
## $title
## [1] "Number of Projects Group by Monthly Hours"
##
## attr(,"class")
## [1] "labels"
```

```
## $x
## [1] "Years"
##
## $y
## [1] "Hours"
##
## $title
## [1] "Years at Company Group by Monthly Hours"
```

```
##
## attr(,"class")
## [1] "labels"
```





- Project 6 and 7 have the two lowest employee count (not many employees did 6 or 7 projects).
- However, employees in 4-year group take about 57% in 6-project group and about 82% in 7-project group. This means those differences mentioned above are heavily affected by employees did 6 or 7 projects.
- As expected, 6-project and 7-project groups have the most employees who work the most average hours a month, which are also the groups have the lowest satisfaction level. Additionally, 4-year group also have the most employees wh have the most average monthly hours.

From all the results and interpretation above, we can assume a few things:

- More than half of employees in 4-year group did 6 or 7 projects, and those employees heavily affect the satisfaction level of 4-year and 6 or 7-project groups.

- Naturally, more projects means more working hours a month. In this case, highest hours a month falls into 6 or 7-project groups, and a lot falls into 4-year group.

4 years at company is enough to have certain experience, but does not seem enough to handle more than average number of projects. Therefore, projects done by employees in this group might take a little longer compared to more experience group, which could possibly lead to more problem and stress.