University of Science VNUHCM

REPORT FOR STAT452

Applied Statistics for Engineers and Scientists

Human Resources Analytics

Pham Hoai Phu Thinh

18125044

Supervior: Assoc. Prof. Dinh Ngoc Thanh

MSc. Nguyen Huu Toan

Contents

Chapter 1

Dataset

1.1 Introduction

The Human Resources Analytics is a simulated dataset from Kaggle and the focus is to understand why the best and most experienced employees is leaving the company. The dataset can be downloaded via the link here.
The dataset contains 1188 rows and originally 10 attributes, but we just consider 6 of them, namely satisfaction, evalution, hours, left, promotion, and salary.

1.2 Format and structure of the dataset

No.	**Attribute**	**Meaning**
1	satisfation	The Satisfaction Level of employees (%)
2	evalution	The last evalution of employees (%)
3	hours	Avarage hours which employees spent on in the company monthly
4	left	Whether the employee has left (0 = FALSE, 1 = TRUE)
5	promotion	Whether had a promotion in the last 5 years (0 = FALSE, 1 = TRUE)
6	salary	How they feel about their salary (high, low or medium)

Chapter 2

Descriptive Statistics

In this chapter we will analyse and describe the information shown in the data.

2.1 Set up working environment and load data

First, we need to set the working environment in the directory which contains the .csv file. Next, read and store data in the variable called *data* by using *attach* funtion.
In order to count the number of rows and columns in data, we take advantage of *nrow* and *ncol* functions in R.
Code

```
#set working environment
setwd("Documents/HCMUS/STAT452/Project")
#read data
data<-read.csv("human_resource.csv")
#store data
save(data, file = 'data.rda')
load('data.rda')
attach(data)
#count the number of rows and columns
nrow(data)
ncol(data)
```

Listing 2.1: Setting the environment and store data

Result

```
> nrow(data)
[1] 14999
> ncol(data)
[1] 7
```

In R, it provides the function called *summary* that can be used for calculating some basic statistic: min, first quartile, median, third quartile and max for quantitative attributes; and the frequency for qualitative attrubutes.
Code

```
#name the column
names(data)<-c("satisfaction", "evaluation", "hours", "left", "promotion",
    "salary")
#left
data$left<-factor(data$left)
levels(data$left)<-c("No", "Yes")
#promotion
data$promotion<-factor(data$promotion)
levels(data$promotion)<-c("No", "Yes")

summary(data)
```

Listing 2.2: Summary

And we receive the result:

```
> summary(data)
  satisfaction      evaluation        hours        left     promotion     salary
 Min.   :0.1200   Min.   :0.3700   Min.   : 98.0   No :758   No :1160   high  :111
 1st Qu.:0.4500   1st Qu.:0.5400   1st Qu.:144.0   Yes:430   Yes:  28   low   :622
 Median :0.6500   Median :0.6650   Median :180.5                        medium:455
 Mean   :0.6351   Mean   :0.6898   Mean   :185.6
 3rd Qu.:0.8100   3rd Qu.:0.8400   3rd Qu.:224.0
 Max.   :1.0000   Max.   :1.0000   Max.   :285.0
>
```

2.2 Basic statistics for attributes

2.2.1 General

There are a number of steps repeated through the report; therefore, they will be presented to avoid the recurrence.

- Use *pie* function to create pie charts.
- Use *barplot* fucntion to create the bar charts with frequency table and stacked bar graphs thanks to *table* function.
- Use *hist* function to create histogram.
- The box drawn by using the values of the three quartiles is called boxplot. Boxplots are created by calling *boxplot* function.
- With two quantitative variables, we use scatter plots, which are drawn by using *plot* function.

2.2.2 *salary* attribute

We draw pie chart representing the proportion of 3 levels of salary.
Code

```
#create table
psal<-table(data$salary)

#draw pie chart
pie(psal, main = "Pie chart for salary" ,radius = 1.0, clockwise = TRUE,
    col = rainbow(3))
```

Listing 2.3: Pie chart for salary

Result

Comments

- As shown in the figure, the percentage of employees doing self-evaluation about their salary occupies the least with high salary, whereas the data of the low one accounts for the most, which is approximate six times as much as that of high salary.
- In addition, the proportion of workers feeling that their salary is medium is about 45%.

2.2.3 *left* attribute

We draw pie chart to illustrate the proportion of employees who have left the job.
Code

```
#create table left
pleft<-table(data$left)

#draw pie chart for left
pie(pleft, main = "Pie chart for left", radius = 1.0, clockwise = TRUE,
    col = c("springgreen", "yellow2"))
```

Listing 2.4: Pie chart for left

Result

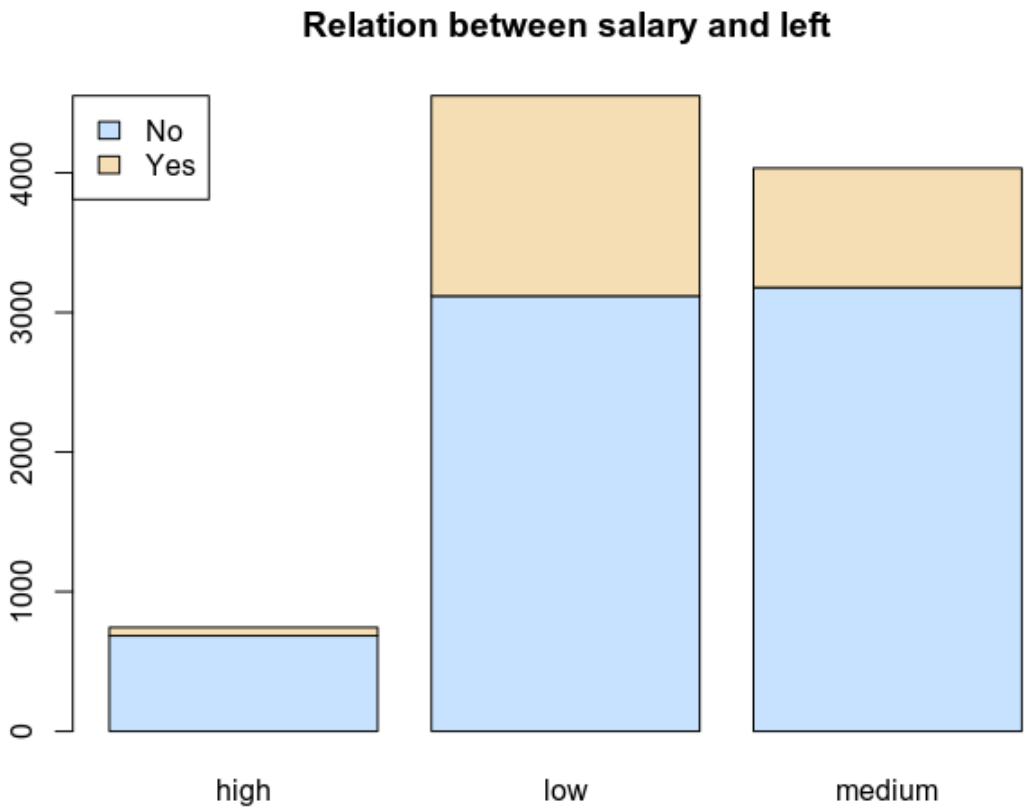
Comments
The percentage of employees who have not left the job is significantly higher than that of those who have. The figure for *no* is at about three-fifth, approximate three times as high as the data of *yes*.

Next, we want to know the relation between *salary* and *left*, answering the question of whether the lower salary is, the more percentage employees decide to leave.
We are going to draw stacked bar graph between *salary* and *left*.
Code

```
#create table salaray and left
saleft<-table(left, salary)

#draw the graph
barplot(saleft, main = "Relation between salary and left", col = c("
    slategray1", "wheat"))
legend("topleft", c("No", "Yes"), fill = c("slategray1", "wheat"))
```

Listing 2.5: Draw stacked bar graph

Result



Comments

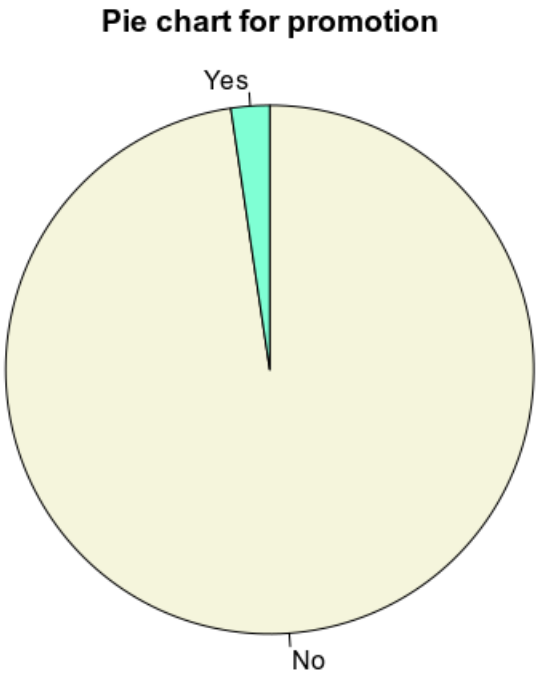
- The more salary they receive, the less proportion they leave the company.
- Almost no employees who feel their salary is high left the job, which means the proportion of them leaving is less than about 5%.

2.2.4 *promotion* attribute

First, we draw pie chart for promotion which shows data on the percentage of employees promoted in the last 5 years.
Code

```
#create table
promo<-table(data$promotion)
#draw pie chart
pie(promo, radius = 1.0, main = "Pie chart for promotion",clockwise = TRUE
    , col = c("beige", "aquamarine"))
```

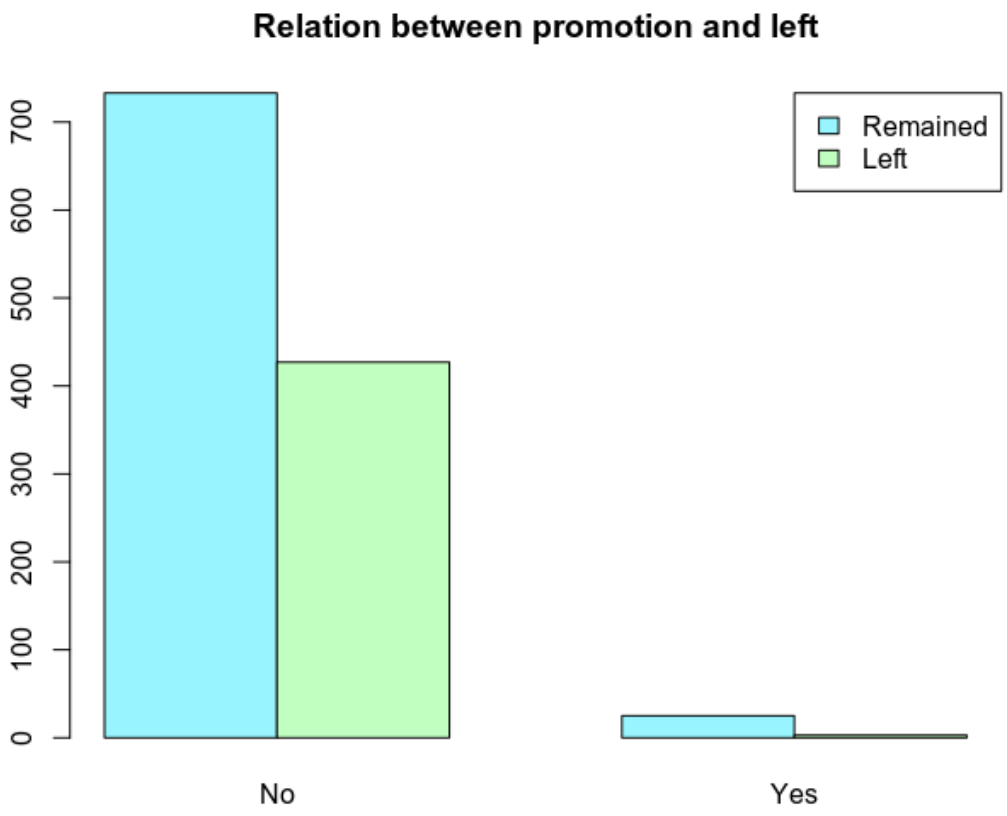
Result



Comments
The proportion of employees promoted in the last 5 years is small.
Next, we are going to find the relation between promotion in salary and whether they left or not by drawing bar graphs.
Code

```
#create table of left and promotion
promoleft<-table(left, promotion)

#draw bar graphs
barplot(promoleft, main = "Relation between promotion and left", beside =
    TRUE, col = c("cadetblue1", "darkseagreen1"))
legend("topright", c("Remained", "Left"), fill = c("cadetblue1", "
    darkseagreen1"))
```

Result



Comments
Almost no employees who were promoted in the last 5 years left the company, whereas the firgure for those who were not is approximate one-half.

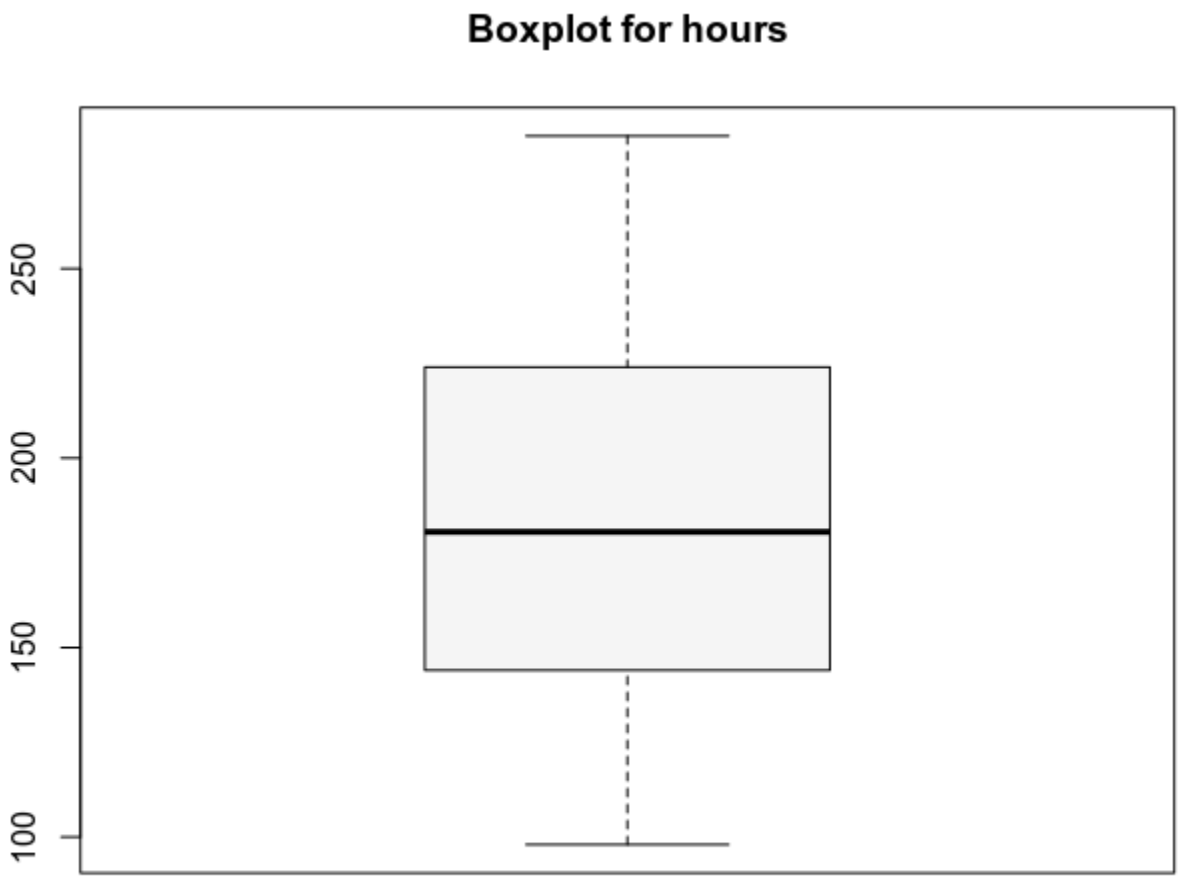
2.2.5 *hours* attribute

We create a boxplot to analyze this attribute.
Code

```
#crate boxplot
boxplot(hours, main = "Boxplot for hours" ,col = c("gray96"))
```

Listing 2.6: Create boxplot for hours attribute

Result



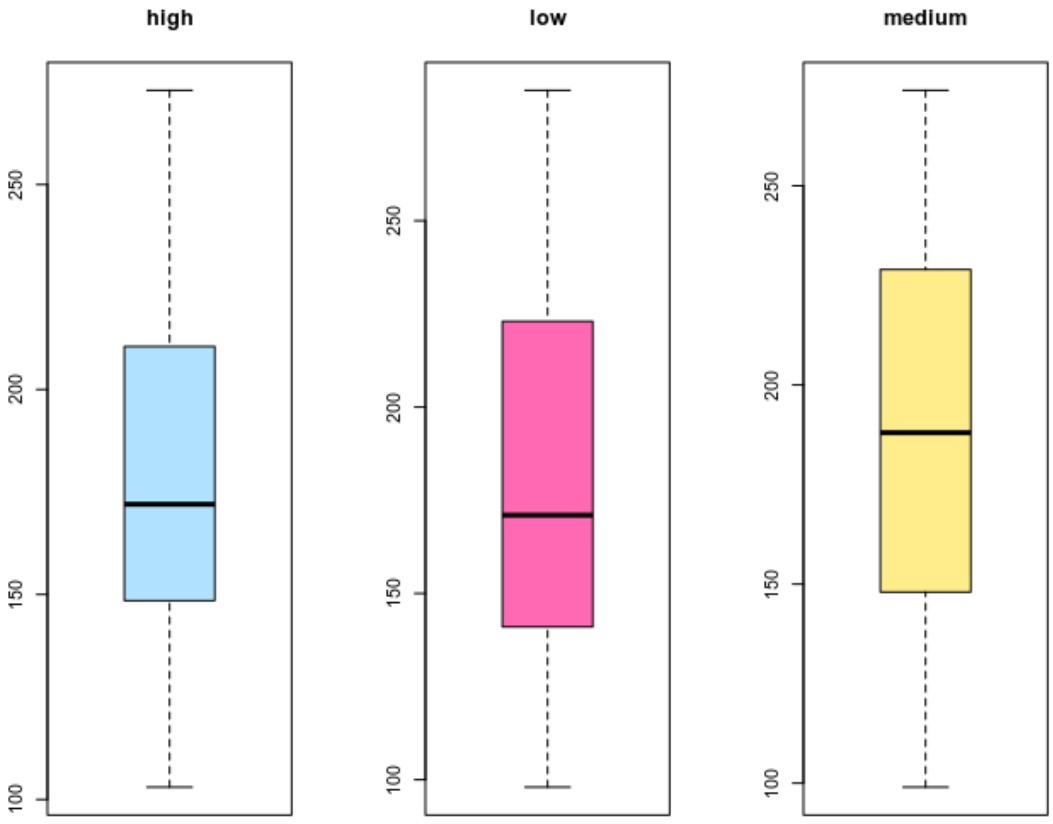
Comments
As shown in the figure, the number of hours employees spent monthly in the company varies from 100 to approximate 300.
Overall, most of them spent from 140 to 230 hours. Moreover, there are a half of them spending more than 180 hours per month.

Next, we draw the boxplot for hours devided into 3 types of salary in order to find whether the salary could affect the amount of time spent on company or not.
Code

```
#devide column
par(mfrow = c(1,3))
#draw boxplot
boxplot(subset(hours, salary == 'high'), main = 'high', col = c("
    lightskyblue1"))
boxplot(subset(hours, salary == 'low'), main = 'low', col = c("hotpink"))
boxplot(subset(hours, salary == 'medium'), main = 'medium', col = c("
    lightgoldenrod1"))
```

Listing 2.7: Boxplot for hours by salary

Result



Comments
There is no significant differences in the numbers of hours spent on the company in high and low salary at 170, whereas the average hours of the medium are higher, at about 190.

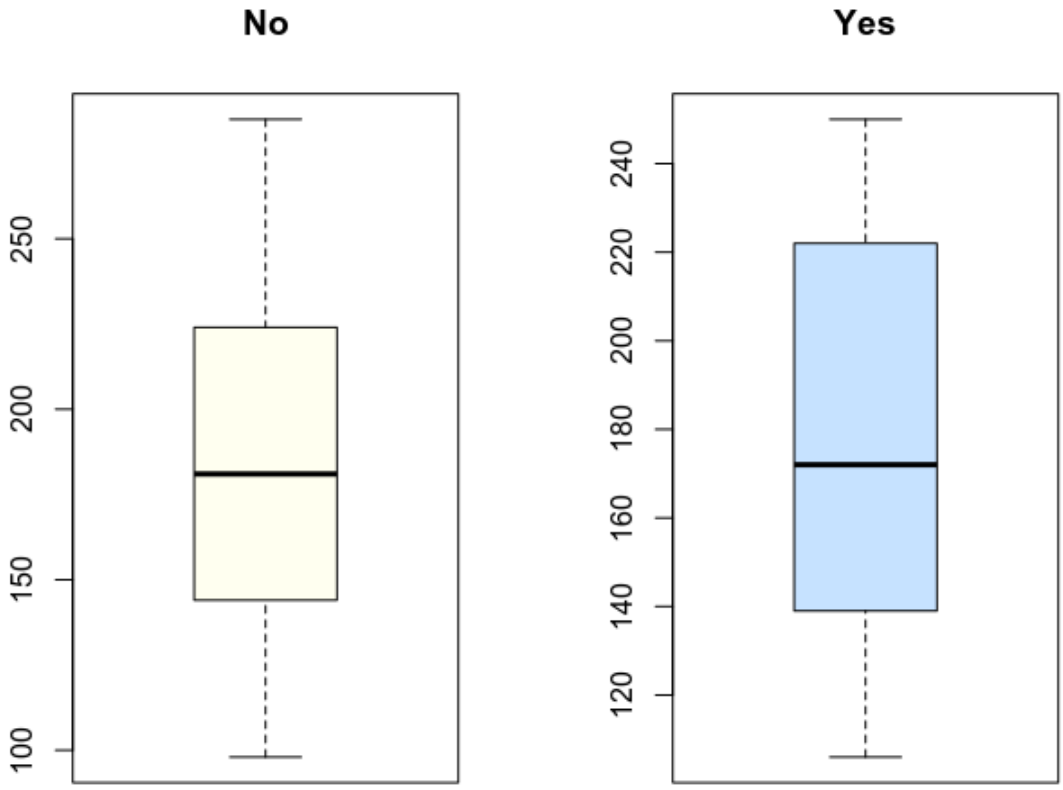
Then, we consider the relation between hours and promotion attributes. Again, boxplots for 2 categories are being drawn.

Code

```
#devide column
par(mfrow = c(1,2))

#Draw boxplots
boxplot(subset(hours, promotion == "No"), main = "No", col = c("ivory"))
boxplot(subset(hours, promotion == "Yes"), main = "Yes", col = c("
    slategray1"))
```

Listing 2.8: Boxplots for hours by promotion

Result



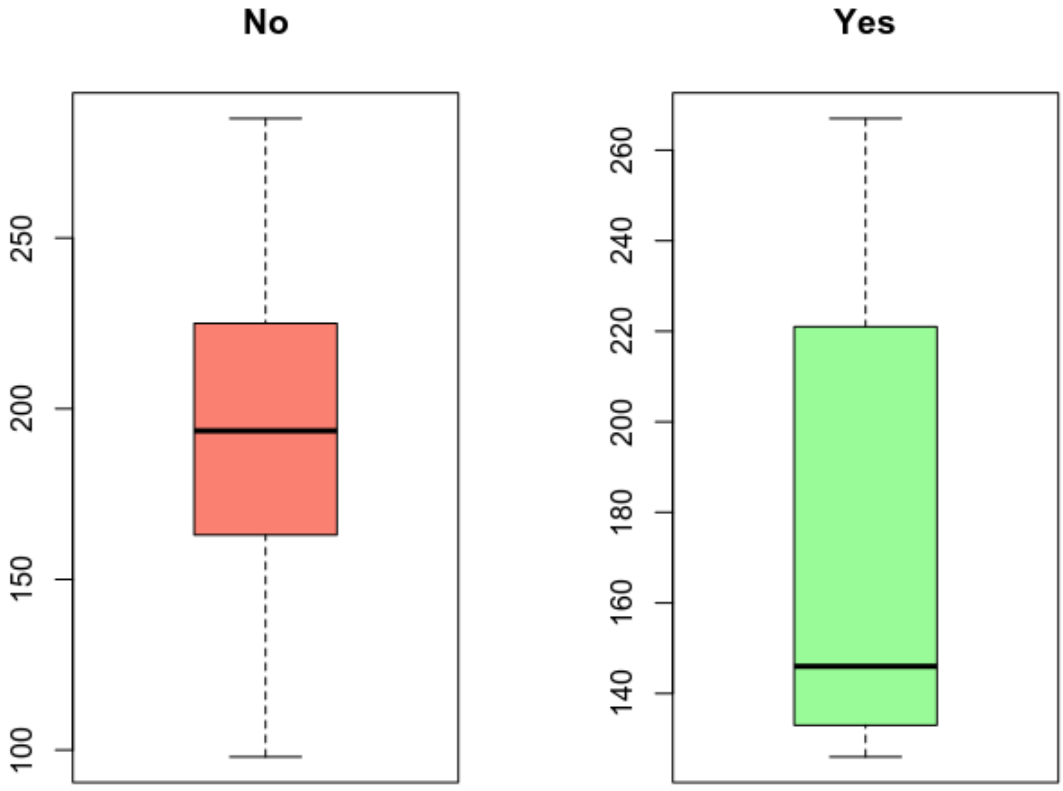
Comments

There is no significant differences in the relation between the amount of time they spent on the company and the probability they got promoted.

We want to find out whether the decision of employees quitting the job has any relation to the amount of working time or not. To compare, we draw boxplots.
Code

```
#Devide column
par(mfrow = c(1,2))

#Draw
boxplot(subset(hours, left == "No"), main = "No", col = c("salmon"))
boxplot(subset(hours, left == "Yes"), main = "Yes", col = c("palegreen"))
```

Listing 2.9: Boxplots for hours by left



Comments
The number of hours of employees leaving the job, at average 190, is much lower than that of they did not, just under 150.

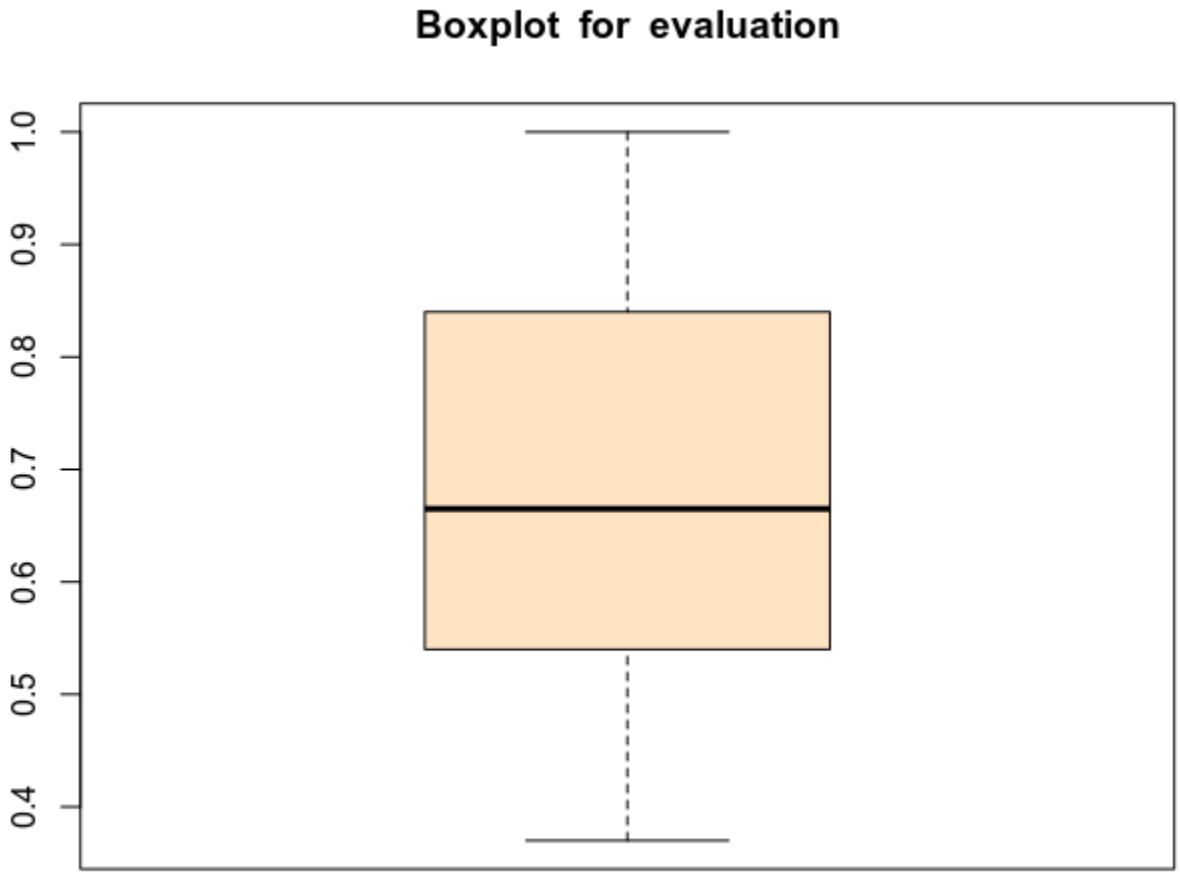
2.2.6 *evaluation* attribute

First, we will create a boxplot to illustrate overall evaluation in the dataset.
Code

```
#create boxplot
boxplot(evaluation, main = "Boxplot for evaluation", col = c("bisque"))
```

Listing 2.10: Boxplot for evaluation

Result



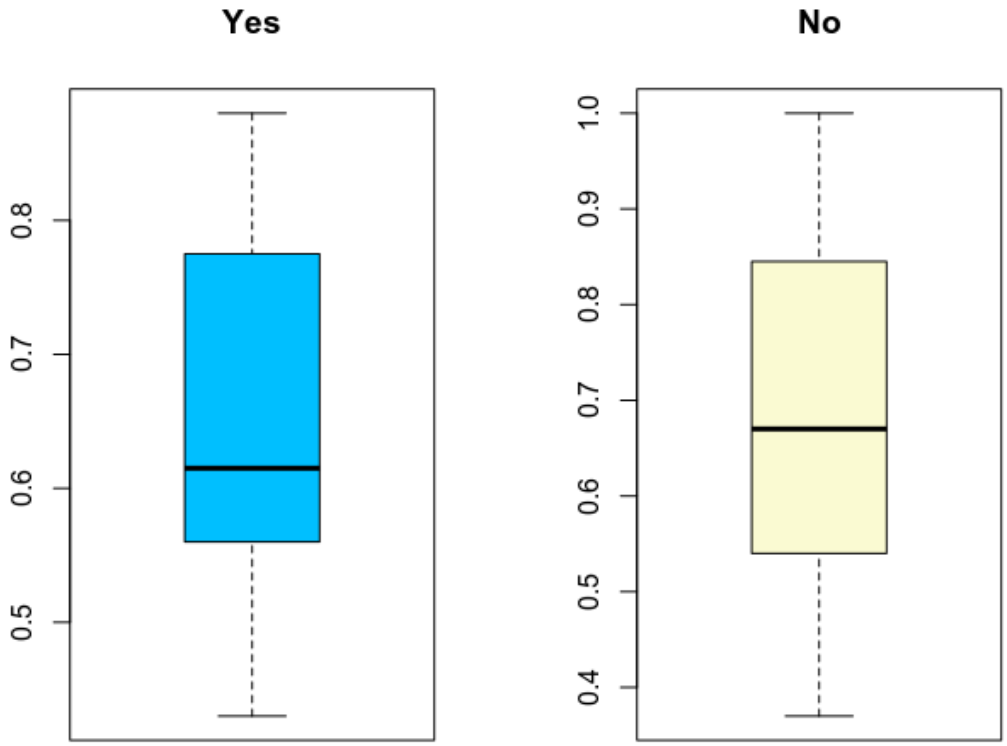
Comments
The performance of employees has been qualified from about 0.38 to 1.0. There are a half of them having the evaluation rate larger than 0.68. In addition, 50% of them having this rate range from 0.58 to 0.86.

Now, we consider about the relation between promotion and evaluation, whether those who got promotion work better than those who did not?
Code

```
#devide column
par(mfrow = c(1,2))
#draw
boxplot(evaluation[promotion == "Yes"], main = "Yes", col = c('deepskyblue'))
boxplot(evaluation[promotion == "No"], main = "No", col = c('lightgoldenrodyellow'))
```

Listing 2.11: Boxplot for evaluation by promotion

Result



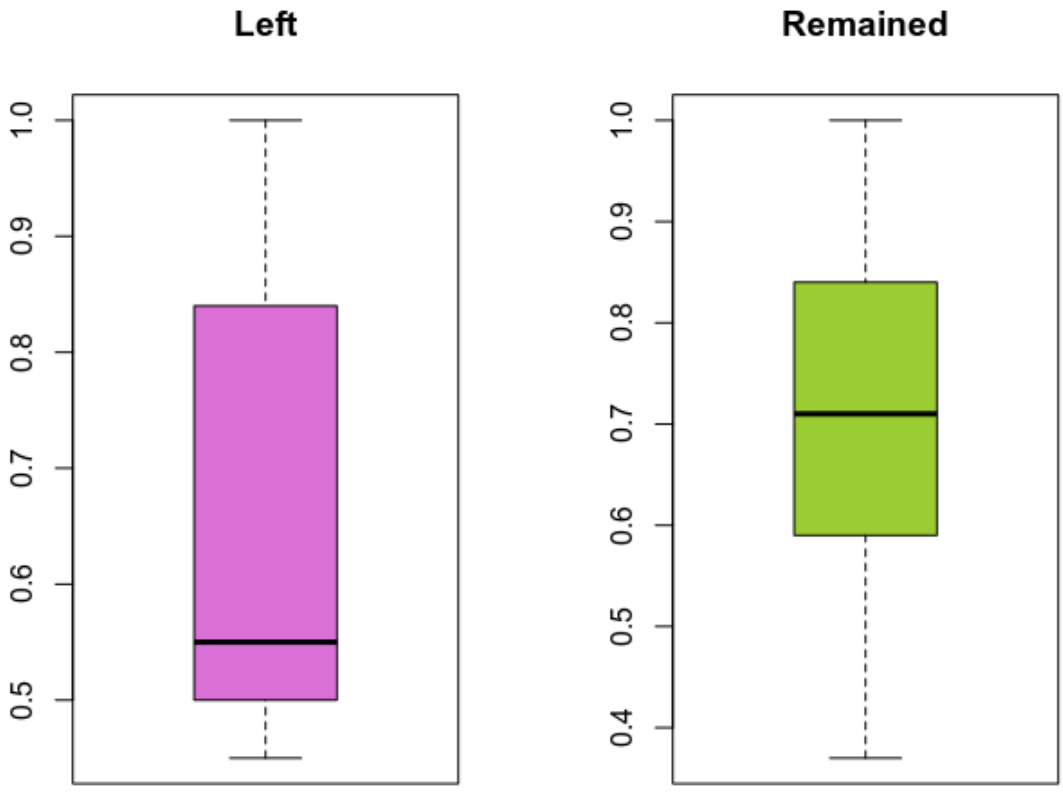
Comments
The fact that those get promoted does not relate to the performance of them in the work. Whereas the average evaluation of them who did not got promoted was at 0.68, the figure for them who did was lower, at 0.62.

Next, we are going to draw boxplots by left attribute to find out the link between the decision of leaving the job and the performance in the work. *Code*

```
#devide column
par(mfrow = c(1,2))

#draw boxplots
boxplot(evaluation[left == 'Yes'], main = 'Left', col = c('orchid'))
boxplot(evaluation[left == 'No'], main = 'Remained', col = c('yellowgreen'
    ))
```

Listing 2.12: Boxplots for evaluation by left

Result



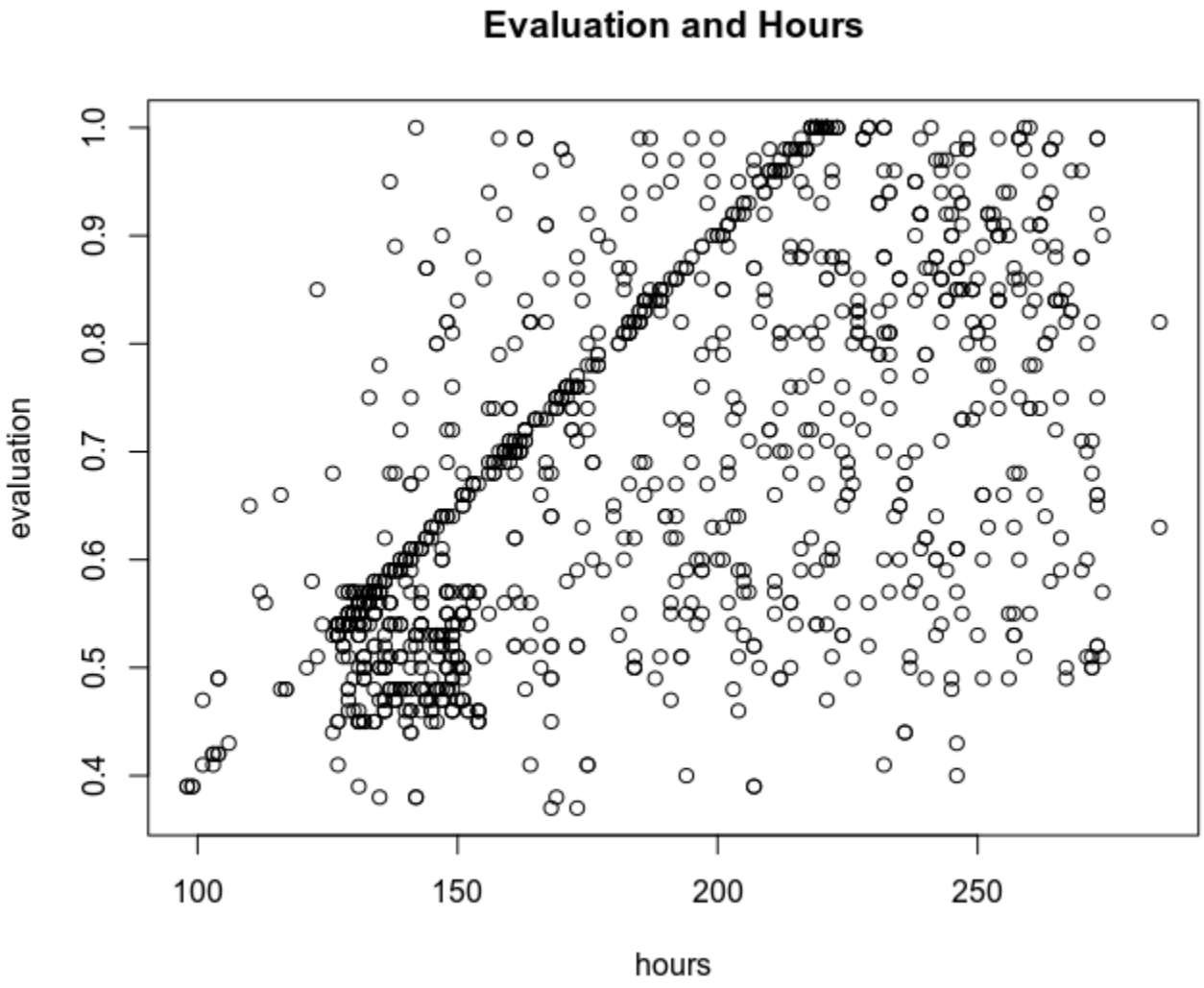
Comments
As shown in the figures, the performance of those who left was at 0.55, while the data of them remaining was significant higher, at 0.71.

Next, we draw the plot to illustrate the relation between evaluation and hours attributes.

Code

```
#draw plot

plot(hours, evaluation, main = 'Evaluation and Hours')
```

Result



Comments

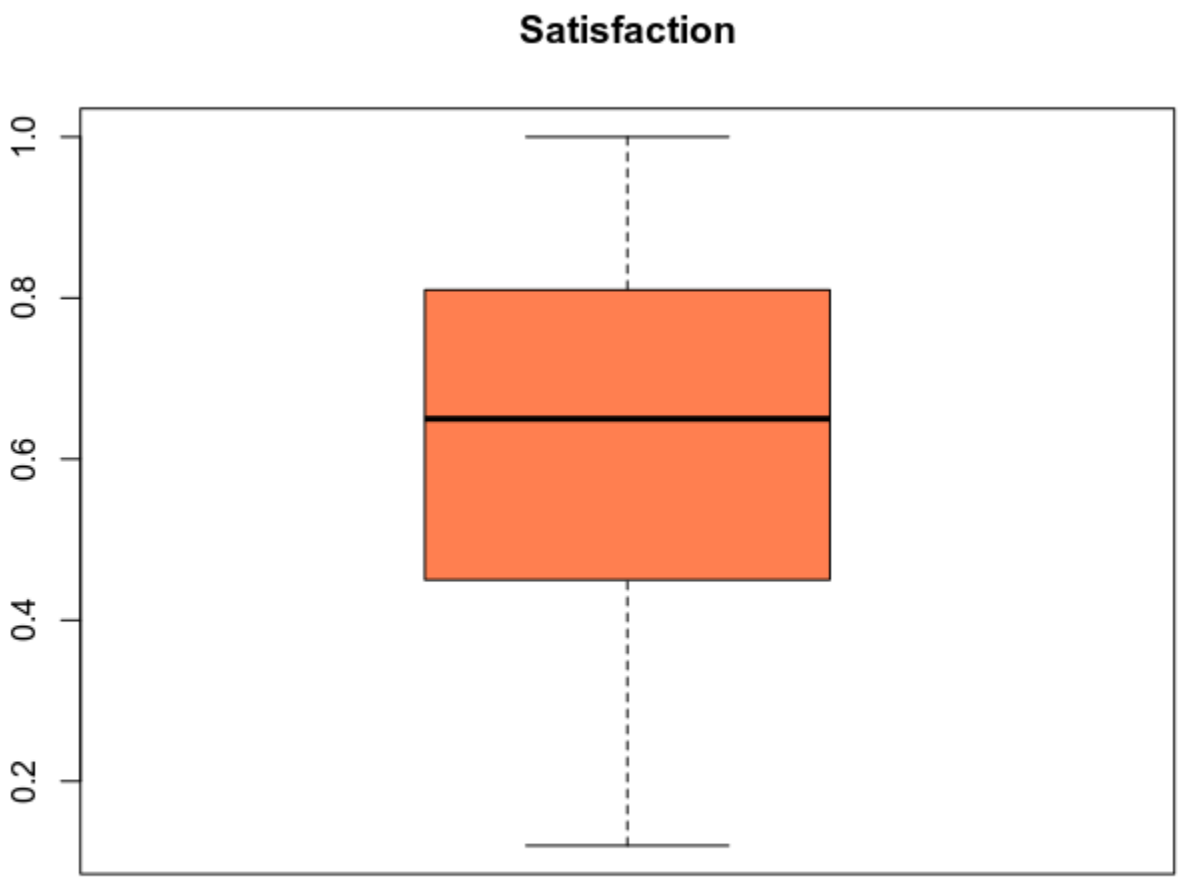
The relation between two attributes is not really clear, but we can find that from the figure, it is possible that the more hours they spent on working, the higher evaluation they get.

2.2.7 *satisfaction* attribute

First, we draw a boxplot to show the overall satisfaction of all employees.
Code

```
#draw boxplot

boxplot(satisfaction, main = 'Satisfaction', col = c('coral'))
```

Result



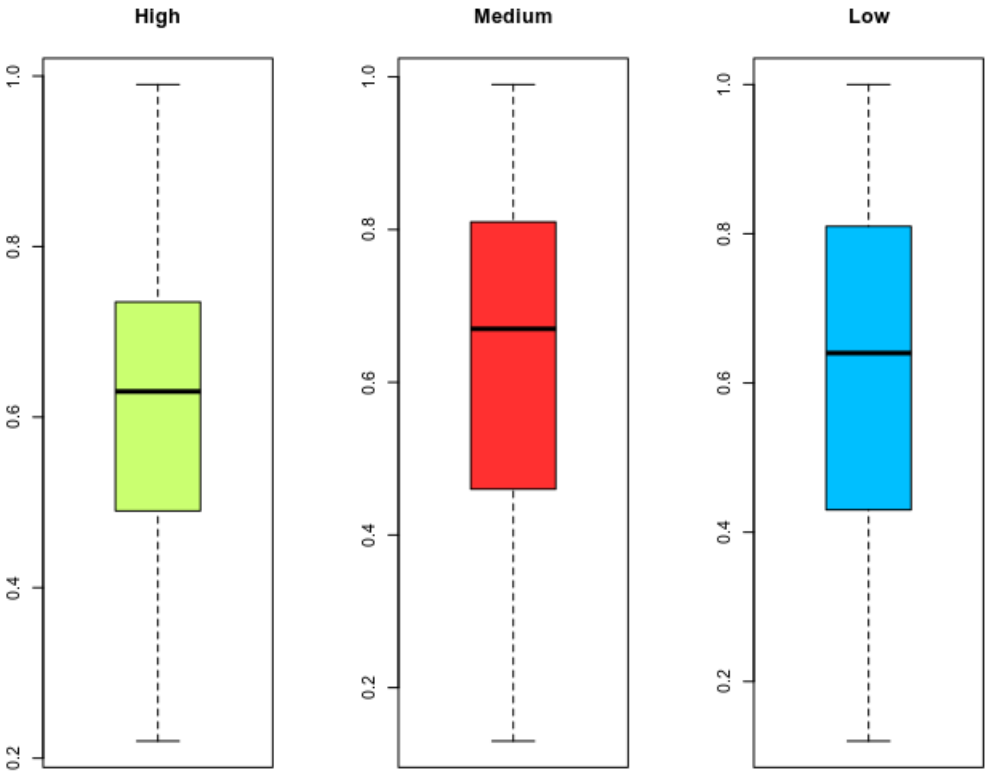
Comments
The average satisfaction rate is at about 0.64, and a half of the workers have this rate range from 0.48 to 0.8.

Then, we split the whole dataset into 3 sets by salary attribute and we draw the boxplots to compare.

Code

```
#devide into 3 columns
par(mfrow = c(1,3))

#draw boxplots
boxplot(satisfaction[salary == 'high'], main = 'High', col = c('
    darkolivegreen1'))
boxplot(satisfaction[salary == 'medium'], main = 'Medium', col = c('
    firebrick1'))
boxplot(satisfaction[salary == 'low'], main = 'Low', col = c('deepskyblue1
    '))
```

Result



Comments

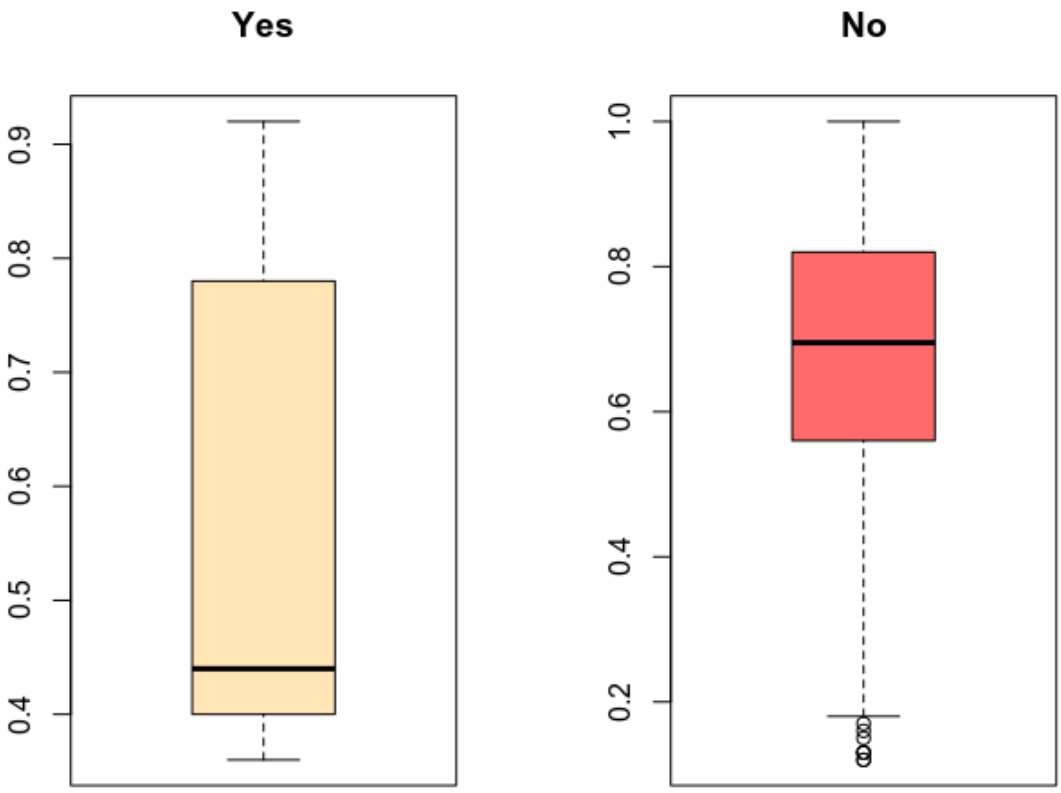
As shown in the figures, there are no significant differences between salary and satisfaction. However, the rate of medium is the highest among 3 categories.

Next, we consider satisfaction and left attribute, and draw the boxplots.

Code

```
#deivde into 2 columns
par(mfrow = c(1,2))

#draw boxplots
boxplot(satisfaction[left == 'Yes'], main = 'Yes', col = c('moccasin'))
boxplot(satisfaction[left == 'No'], main = 'No', col = c('indianred1'))
```

Result



Comments

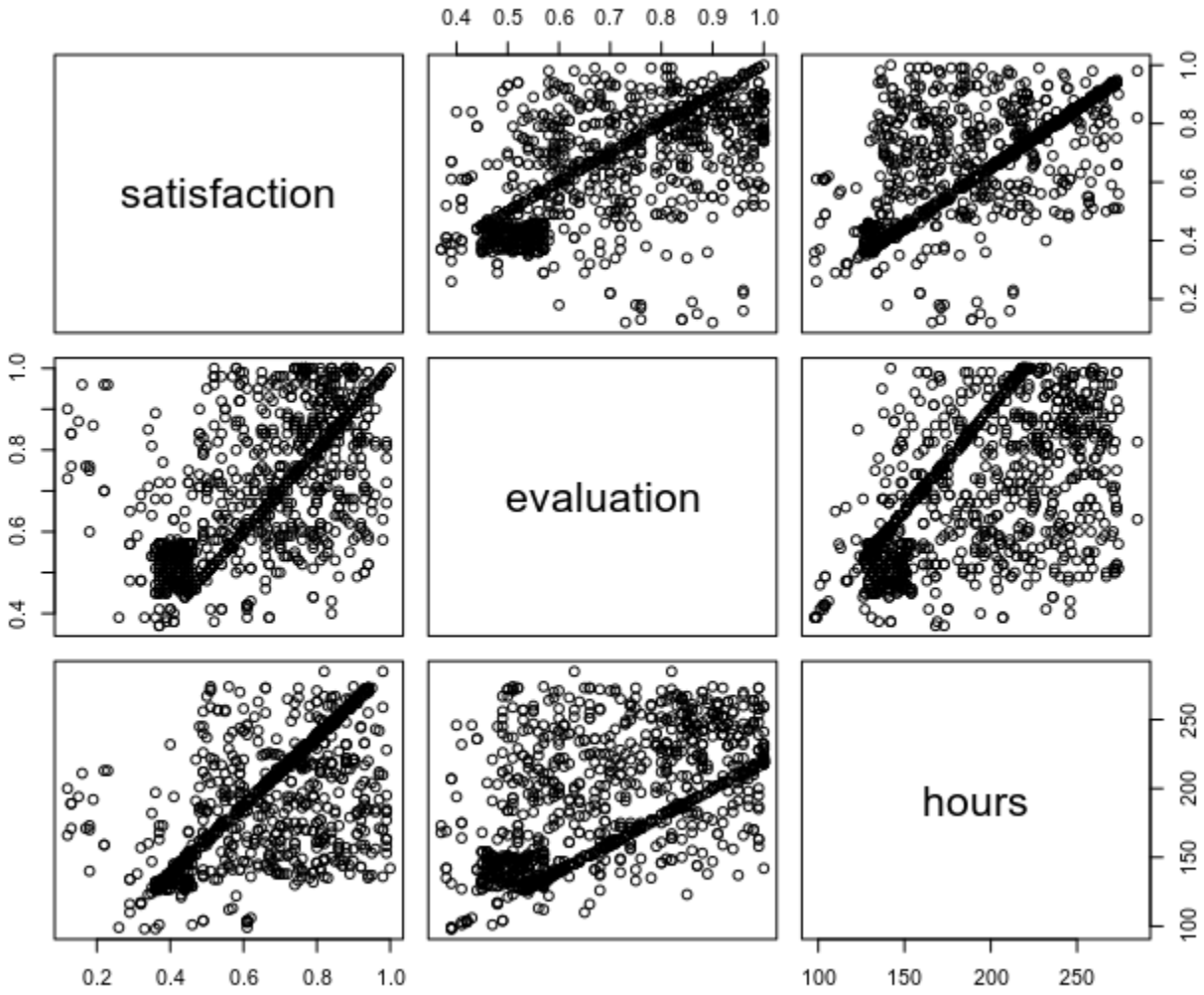
It is clear that the average satisfaction of those who left the job is much lower than the figure for those who do not, at 0.45 and 0.65 respectively. It also shows that there are many outliers in the figure for No.

Next, to illustrate the relation between satisfaction and evaluation as well as hours, we draw the matrix of scatterplot.

Code

```
#draw matrix of scatterplot
pairs(~satisfaction + evaluation + hours)
```

Result



Comments

Although the relation among them is not clear, it is possible that there exists an equation expressing this relation.

Chapter 3

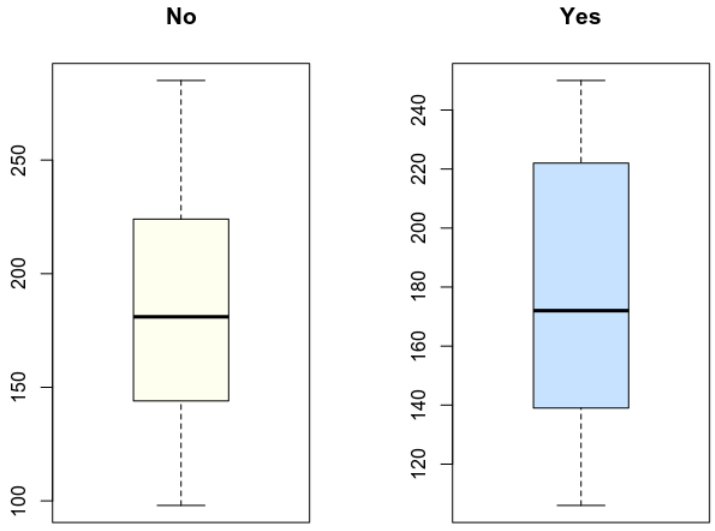
Inferential Statistics

In this chapter, based on the comments in the Descriptive Statistics, we hypothesis and test some problems related to satisfaction, evaluation and hours attributes.

3.1 *hours* attribute

***Problem 1**: True means of hours by promotion are the same.*

Based on the figure about hours by promotion in the previous chapter, we can assume that the time they spend on the company is the same in 2 different categories of promotion.



Hours by promotion

To test this hypothesis, we utilise *t.test* function in R.

$$H_0 : hours_{yes} = hours_{no}$$

$$H_a : hours_{yes} \neq hours_{no}$$

Code

```
#use t test
test<-t.test(hours ~ promotion)
#print the result
test
```

Result

```
> test<-t.test(hours ~ promotion)
> test

        Welch Two Sample t-test

data:  hours by promotion
t = 0.75878, df = 28.345, p-value = 0.4542
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -11.17901  24.34526
sample estimates:
 mean in group No mean in group Yes
         185.7974          179.2143

>
```

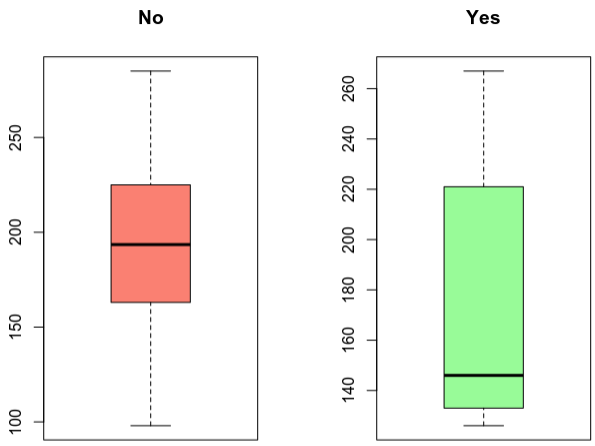
As shown in the figure, the p-value is 0.4542 at 95% confidence interval. Compared to the level of significance $\alpha = 0.05$, the p-value is much higher, so we do not have enough evidence to reject the null hypothesis.
Conclusion
Therefore, we accept the hypothesis that the true means of hours are the same by promotion.

***Problem 2**: True mean of hours of employees who has left is lower than that of them who has not.*

Based on the figure about hours by left attribute in the previous chapter, we can see that the true mean of hours of those who left is much lower than that of those who did not. So, we are going to test this hypothesis.

$$H_0 : hours_{left} = hours_{remained}$$

$$H_a : hours_{left} < hours_{remained}$$



Hours by left

To test the validity of this hypothesis, we use *t.test* function.
Code

```
#test the hypothesis
test <- t.test(hours[left == 'Yes'], y = hours[left == 'No'], alternative
    = "less")
#show the result
test
```

Result

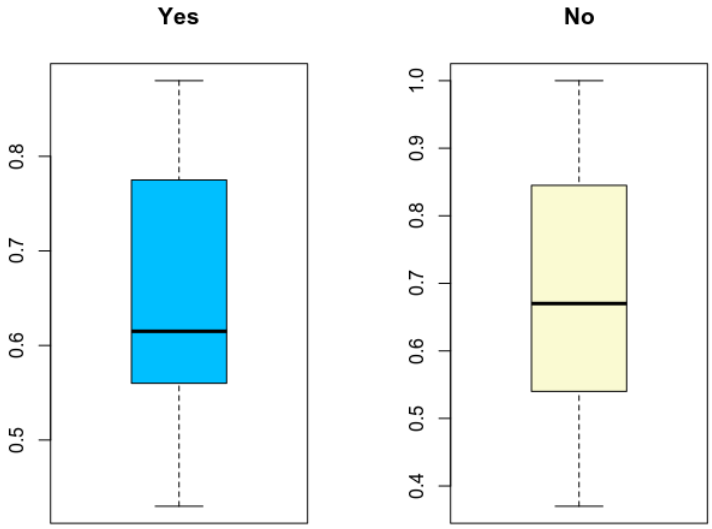
```
> test <- t.test(hours[left == 'Yes'], y = hours[left == 'No'], alternative = "less")
> test

        Welch Two Sample t-test

data:  hours[left == "Yes"] and hours[left == "No"]
t = -9.4285, df = 803.05, p-value < 2.2e-16
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
      -Inf -21.42524
sample estimates:
mean of x mean of y
 169.0791  195.0383

>
```

As shown in the figure, the p-value is very small at 95% confidence interval. Compared to $\alpha = 0.05$ level of significance, the p-value is much lower, so we have enough evidence to reject the null hypothesis.
Conclusion
Therefore, we accept the alternative hypothesis that the mean of hours of workers who left the company is lower than that of them remaining.

3.2 *evaluation* attribute

***Problem 3**: True means of evaluation by promotion are the same.*

Based on the figure about evaluation by promotion in the previous chapter, we can assume that the time they spend on the company is the same in 2 different categories of promotion.



Evaluation by promotion

To test this hypothesis, we utilise *t.test* function in R.

$$H_0 : evaluation_{yes} = evaluation_{no}$$

$$H_a : evaluation_{yes} \neq evaluation_{no}$$

Code

```
#use t test
test<-t.test(evaluation ~ promotion)

#print the result
test
```

Result

```
> test <- t.test(evaluation ~ promotion)
> test

        Welch Two Sample t-test

data:  evaluation by promotion
t = 1.5173, df = 29.014, p-value = 0.14
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.01442215  0.09733348
sample estimates:
 mean in group No mean in group Yes
        0.6907414         0.6492857

>
```

As shown in the figure, the p-value is 0.14 at 95% confidence interval. Compared to the level of significance $\alpha = 0.05$, the p-value is much higher, so we do not have enough evidence to reject the null hypothesis.

Conclusion

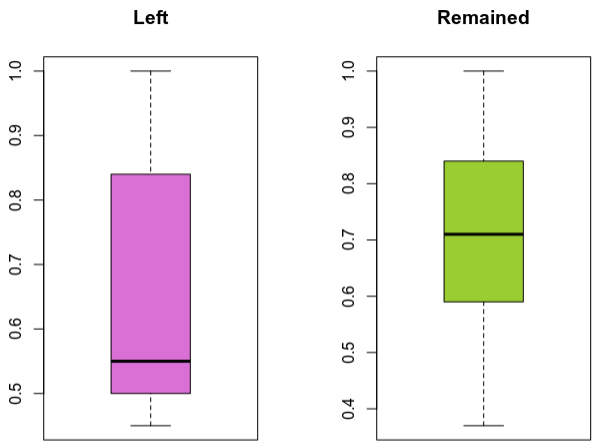
Therefore, we accept the hypothesis that the true means of evaluation are the same by promotion.

***Problem 4**: True mean of evaluation of employees who has left is lower than that of them who has not.*

Based on the figure about evaluation by left attribute in the previous chapter, we can see that the true mean of evaluation of those who left is much lower than that of those who did not. So, we are going to test this hypothesis.

$$H_0 : evaluation_{remained} = evaluation_{left}$$

$$H_a : evaluation_{remained} > evaluation_{left}$$



Evaluation by left

To test the validity of this hypothesis, we use *t.test* function.
Code

```
#test the hypothesis
test <- t.test(evaluation[left == 'No'], y = evaluation[left == 'Yes'],
    alternative = "greater")
#show the result
test
```

Result

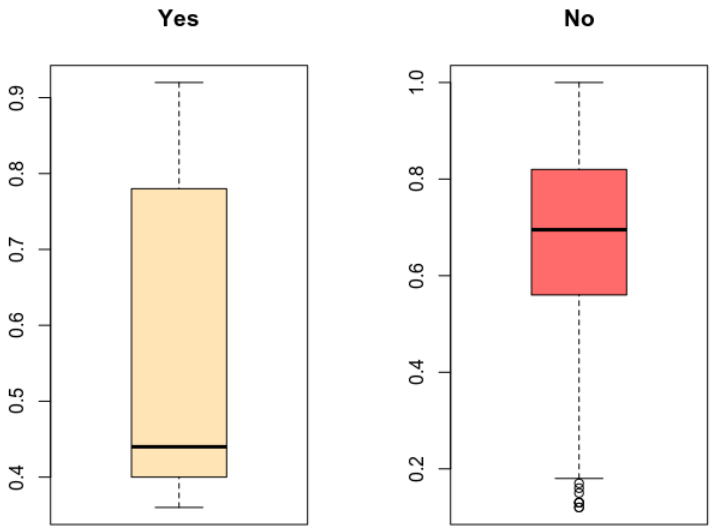
```
> test <- t.test(evaluation[left == 'No'], y = evaluation[left == 'Yes'], alternative = "greater")
> test

        Welch Two Sample t-test

data:  evaluation[left == "No"] and evaluation[left == "Yes"]
t = 7.1868, df = 765.51, p-value = 7.887e-13
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 0.05989754        Inf
sample estimates:
mean of x mean of y
0.7178892 0.6401860

>
```

As shown in the figure, the p-value is very small, at $7.887 * 10^{-13}$, at 95% confidence interval. Compared to $\alpha = 0.05$ level of significance, the p-value is much lower, so we have enough evidence to reject the null hypothesis.
Conclusion
Therefore, we accept the alternative hypothesis that the mean of evaluation of workers who left the company is lower than that of them remaining.

3.3 *satisfaction* attribute

***Problem 5**: True mean of satisfaction of those who left the company is lower than that of those who did not.*

Based on the figure about satisfaction by left attribute in the previous chapter, we can see that the true mean of satisfaction of those who left is much lower than that of those who did not. So, we are going to test this hypothesis.



Satisfaction by left

To test the validity of this hypothesis, we use *t.test* function.
Code

```
#test the hypothesis
test <- t.test(satisfaction[left == 'Yes'], y = satisfaction[left == 'No'
    ], alternative = "less")
#show the result
test
```

Result

```
> test <- t.test(satisfaction[left == 'Yes'], y = satisfaction[left == 'No'], alternative = "less")
> test

        Welch Two Sample t-test

data:  satisfaction[left == "Yes"] and satisfaction[left == "No"]
t = -12.095, df = 819.55, p-value < 2.2e-16
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
       -Inf -0.1199518
sample estimates:
mean of x mean of y
0.5465116 0.6853694

>
```

As shown in the result, the p-value is very small, at $2.2 * 10^{-16}$, at 95% confidence interval. Compared to $\alpha = 0.05$ level of significance, the p-value is much lower, so we have enough evidence to reject the null hypothesis.

Conclusion

Therefore, we accept the alternative hypothesis that the mean of satisfaction of those who left the company is lower than that of them who still remain.

Chapter 4

Regression

In this chapter, we will build various models based on the dataset.

4.1 Simple Regression

4.1.1 *evaluation* by *hours*

We are going to build the regression equation:

$$evaluation = \beta_0 + \beta_1 * hours + \epsilon$$

To calculate the regression coefficients, we use *lm* function in R.
Code

```
#build the model
model1 <- lm(evaluation ~ hours)
#show the result
model1
```

Result

```
> model1 <- lm(evaluation ~ hours)
> model1

Call:
lm(formula = evaluation ~ hours)

Coefficients:
(Intercept)        hours
   0.280961     0.002202

>
```

The regression equation now becomes:

$$evaluation = 0.280961 + 0.002202 * hours + \epsilon$$

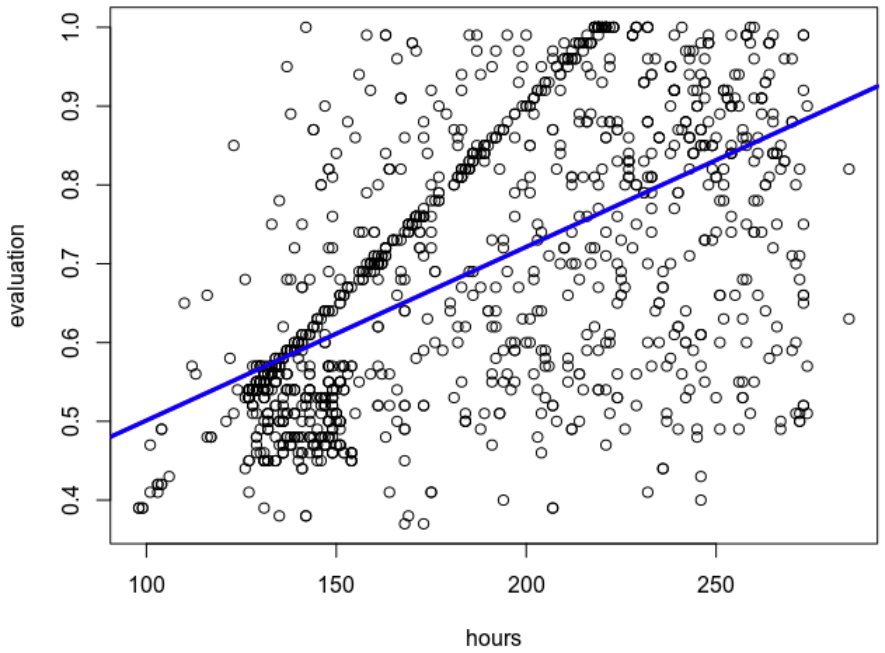
which means:

- $\beta_0 = 0.280961$, meaning that if hours is equal to 0, the average mean of evaluation will be 0.280961
- $\beta_1 = 0.002202$, meaning that if hours increases by 1, evaluation will increase by 0.002202

We draw the regression line on the plot.
Code

```
#plot the graph
plot(evaluation ~ hours)
#add the line
abline(model1, col = 'blue', lwd = 3)
```

Result



To check the goodness of the equation, we check it by using *summary* function in R. *Code*

```
summary(model1)
```

Result

```
> summary(model1)

Call:
lm(formula = evaluation ~ hours)

Residuals:
     Min       1Q   Median       3Q      Max
-0.42268 -0.09391 -0.00604  0.09826  0.40634

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.810e-01  1.724e-02   16.29   <2e-16 ***
hours       2.202e-03  9.019e-05   24.42   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1423 on 1186 degrees of freedom
Multiple R-squared:  0.3345,    Adjusted R-squared:  0.334
F-statistic: 596.2 on 1 and 1186 DF,  p-value: < 2.2e-16
```

From the result, we can see that:

- Call: $evaluation = \beta_0 + \beta_1 * hours + \epsilon$
- Residuals: descriptive analysis of the residuals $\hat{\epsilon} = \hat{y} - y$
- Coefficients:
  - Estimate: $\beta_0 = 0.281, \beta_1 = 0.002202$
  - Standard Error: $\beta_0 = 0.01724, \beta_1 = 9.019 * 10^{-5}$
  - t-value: $\beta_0 = 16.29, \beta_1 = 24.42$
- Residual standard error $\sigma = 0.1423$
- $R^2 = 0.334$

To estimate the confident interval of the coefficients, we use *confint* function in R.
Code

```
confint(model1)
```

Result

```
> confint(model1)
                    2.5 %      97.5 %
(Intercept) 0.247129378 0.314793282
hours       0.002025155 0.002379047
>
```

At 95% confidence inverval, the estimates for coefficients are $\beta_0 \in (0.247129378, 0.314792182)$ and $\beta_1 \in (0.002025155, 0.002379047)$

4.1.2 *satisfaction* by *evaluation*

We are going to build the regression equation:

$$satisfaction = \beta_0 + \beta_1 * evaluation + \epsilon$$

To calculate the regression coefficients, we use *lm* function in R.
Code

```
#build the model
model2 <- lm(satisfaction ~ evaluation)
#show the result
model2
```

Result

```
> model2 <- lm(satisfaction ~ evaluation)
> model2

Call:
lm(formula = satisfaction ~ evaluation)

Coefficients:
(Intercept)   evaluation
     0.1247       0.7400
```

The regression equation now becomes:

$$satisfaction = 0.1247 + 0.7400 * evaluation + \epsilon$$

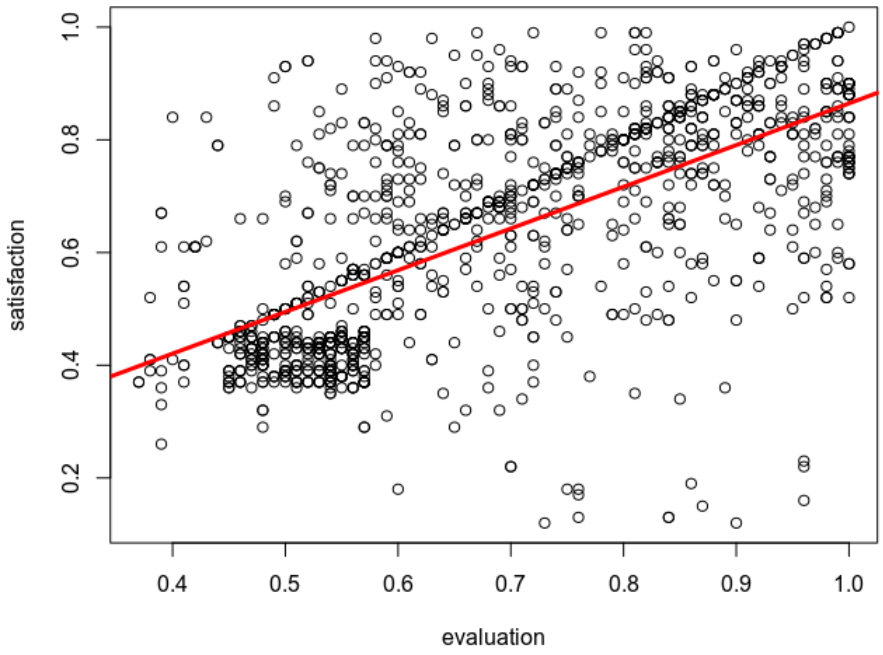
which means:

- $\beta_0 = 0.1247$, meaning that if evaluation is equal to 0, the average mean of satisfaction will be 0.1247
- $\beta_1 = 0.7400$, meaning that if evaluation increases by 1, satisfaction will increase by 0.7400

We draw the regression line on the plot.
Code

```
#plot the graph
plot(satisfaction ~ evaluation)
#add the line
abline(model2, col = 'red', lwd = 3)
```

Result



To check the goodness of the equation, we check it by using *summary* function in R.
Code

```
summary(model2)
```

Result

```
> summary(model2)

Call:
lm(formula = satisfaction ~ evaluation)

Residuals:
     Min       1Q   Median       3Q      Max
-0.67510 -0.08990  0.00522  0.08731  0.43532

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.12465    0.01758    7.09  2.3e-12 ***
evaluation   0.74005    0.02471   29.95  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1484 on 1186 degrees of freedom
Multiple R-squared:  0.4305,    Adjusted R-squared:  0.4301
F-statistic: 896.7 on 1 and 1186 DF,  p-value: < 2.2e-16

>
```

From the result, we can see that:

- Call: $satisfaction = \beta_0 + \beta_1 * evaluation + \epsilon$
- Residuals: descriptive analysis of the residuals $\hat{\epsilon} = \hat{y} - y$
- Coefficients:
  - Estimate: $\beta_0 = 0.12465, \beta_1 = 0.74005$
  - Standard Error: $\beta_0 = 0.01758, \beta_1 = 0.02471$
  - t-value: $\beta_0 = 7.09, \beta_1 = 29.95$
- Residual standard error $\sigma = 0.1484$
- $R^2 = 0.4301$

To estimate the confident interval of the coefficients, we use *confint* function in R.
Code

```
confint(model2)
```

Result

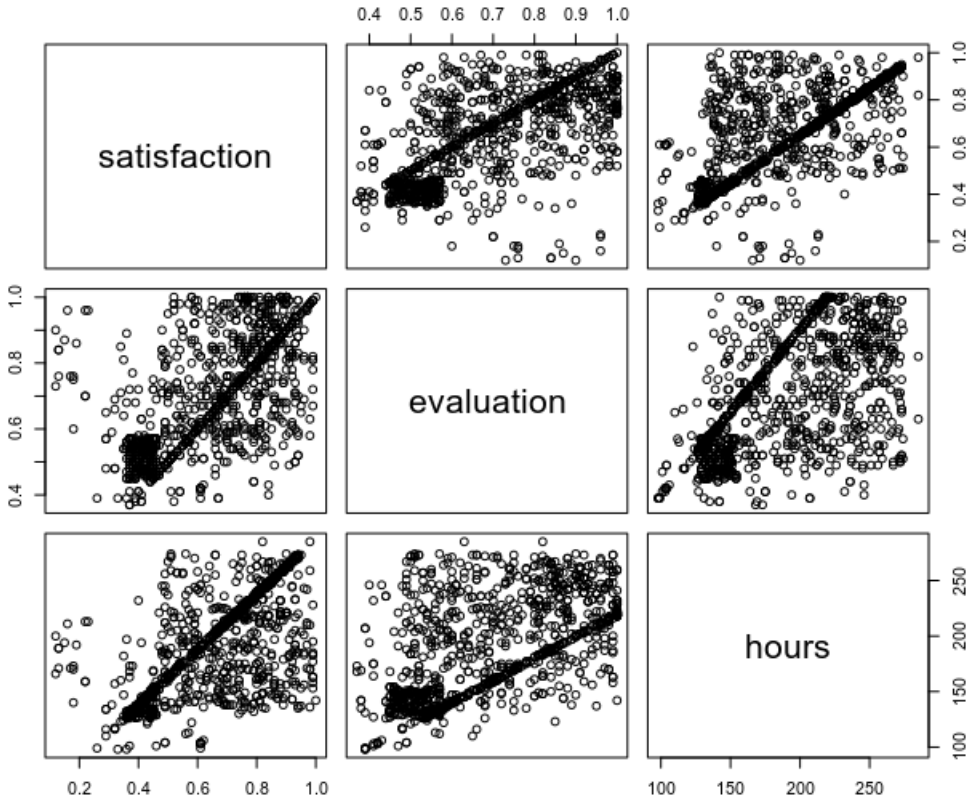
```
> confint(model2)
                 2.5 %    97.5 %
(Intercept) 0.09015683 0.1591475
evaluation  0.69155847 0.7885333
>
```

At 95% confidence inverval, the estimates for coefficients are $\beta_0 \in (0.09015683, 0.01591475)$ and $\beta_1 \in (0.69155847, 0.7885333)$

4.2 Multiple Regression

Based on the figure showing the relation among *satisfaction*, *evaluation* and *hours* attributes, we are going to build the regression equation:

$$satisfaction = \beta_0 + \beta_1 * evaluation + \beta_2 * hours + \epsilon$$



Code

```
#build model
model3 <- lm(satisfaction ~ evaluation + hours)
#show model
model3
```

Result

```
> model3 <- lm(satisfaction ~ evaluation + hours)
> model3

Call:
lm(formula = satisfaction ~ evaluation + hours)

Coefficients:
(Intercept)   evaluation        hours
  -0.025644     0.457686     0.001859

>
```

The regression equation now becomes:

$$satisfaction = -0.025644 + 0.457686 * evaluation + 0.001859 * hours + \epsilon$$

which means:

- β_0 = -0.025644, meaning that if hours is equal to 0 and evaluation is equal to 0, the average mean of evaluation will be -0.025644
- β_1 = 0.457686, meaning that if evaluation increases by 1, satisfaction will increase by 0.457686
- β_2 = 0.001859, meaning that if hours increases by 1, satisfaction will increase by 0.001859

To check the goodness of the equation, we check it by using *summary* function in R.
Code

```
summary(model3)
```

Result

```
> summary(model3)

Call:
lm(formula = satisfaction ~ evaluation + hours)

Residuals:
     Min       1Q   Median       3Q      Max
-0.64593 -0.06599 -0.01656  0.06102  0.48740

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.025645   0.017597  -1.457    0.145
evaluation   0.457686   0.026785  17.087   <2e-16 ***
hours        0.001859   0.000102  18.226   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1312 on 1185 degrees of freedom
Multiple R-squared:  0.5552,    Adjusted R-squared:  0.5545
F-statistic: 739.6 on 2 and 1185 DF,  p-value: < 2.2e-16

>
```

From the result, we can see that:

- Call: $satisfaction = \beta_0 + \beta_1 * evaluation + \beta_2 * hours + \epsilon$
- Residuals: descriptive analysis of the residuals $\hat{\epsilon} = \hat{y} - y$
- Coefficients:
  - Estimate: $\beta_0 = -0.025645, \beta_1 = 0.0457686, \beta_2 = 0.001859$
  - Standard Error: $\beta_0 = 0.017597, \beta_1 = 0.026785, \beta_2 = 0.000102$
  - t-value: $\beta_0 = -1.457, \beta_1 = 17.087, \beta_2 = 18.226$
- Residual standard error $\sigma = 0.1312$
- $R^2 = 0.5545$

However, the p-value for β_0 is 0.145, which is much high compared to $\alpha = 0.05$. So, we have enough evidence to remove β_0 from the equation. The equation now becomes:

$$satisfaction = \beta_1 * evaluation + \beta_2 * hours + \epsilon$$

Repeat the steps above,
Code

```
#build
fit <- lm(satisfaction ~ 0 + evaluation + hours)
fit
#test the reduced equation
anova(model3, fit)
#analysis
summary(fit)
```

Result

```
> fit <- lm(satisfaction ~ 0 + evaluation + hours)
> fit

Call:
lm(formula = satisfaction ~ 0 + evaluation + hours)

Coefficients:
evaluation       hours
  0.440992    0.001789

> anova(model3, fit)
Analysis of Variance Table

Model 1: satisfaction ~ evaluation + hours
Model 2: satisfaction ~ 0 + evaluation + hours
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1   1185 20.403
2   1186 20.439 -1 -0.036568 2.1238 0.1453
> summary(fit)
```

With Pr(>F) = 0.145, we have enough evidence to accept the reduced equation. The equation becomes:

$$satisfaction = 0.440992 * evaluation + 0.001789 * hours + \epsilon$$

```
> summary(fit)

Call:
lm(formula = satisfaction ~ 0 + evaluation + hours)

Residuals:
     Min       1Q   Median       3Q      Max
-0.64085 -0.07118 -0.01880  0.06367  0.48091

Coefficients:
           Estimate Std. Error t value Pr(>|t|)
evaluation 4.410e-01  2.422e-02   18.20   <2e-16 ***
hours      1.789e-03  9.013e-05   19.85   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1313 on 1186 degrees of freedom
Multiple R-squared:  0.9611,    Adjusted R-squared:  0.961
F-statistic: 1.464e+04 on 2 and 1186 DF,  p-value: < 2.2e-16

>
```

From the result, we can see that:

- Call: $satisfaction = \beta_1 * evaluation + \beta_2 * hours + \epsilon$
- Residuals: descriptive analysis of the residuals $\hat{\epsilon} = \hat{y} - y$
- Coefficients:
  - Estimate: $\beta_1 = 0.440992, \beta_2 = 0.0.001789$
  - Standard Error: $\beta_1 = 2.422 * 10^{-2}, \beta_2 = 9.013 * 10^{-5}$
  - t-value: $\beta_1 = 18.20, \beta_2 = 19.85$
- Residual standard error $\sigma = 0.1313$
- $R^2 = 0.961$

4.3 Conclusion

Based on the analysis above, we can conclude that the last model is the best fit:

$$satisfaction = 0.440992 * evaluation + 0.001789 * hours + \epsilon$$

which has the best $R^2 = 0.961$.