Separation analysis

I. Data sources

https://www.kaggle.com/analystanand/employee-attrition

II. Research Questions

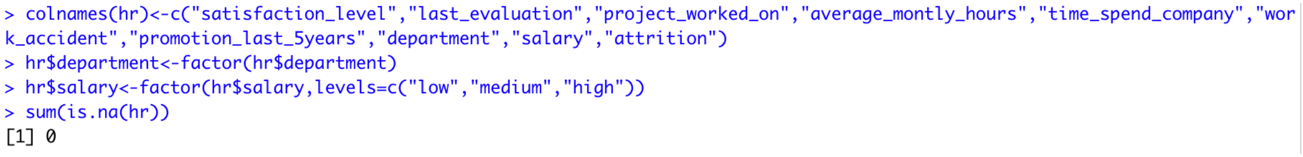
1. main descriptive statistics
2. Relationship between factors and separation
2.1. satisfaction 2.2. employee evaluation 2.3. number of projects
2.4. average monthly working hours 2.5. years of work 2.6. whether to be promoted 2.7. company department recession 2.8. salary
3. Separation modeling 3.1. correlation coefficient diagram
3.2. Decision tree 3.3. ROC/AUC curve

III. Data Understanding

This data set has 15000 rows of data, which contains 10 fields with the following meanings:

- satisfaction_level: satisfaction
- last_evaluation: employee evaluation
- projects_worked_on: number of completed projects
- average_montly_hours: average
- time_spend_company: number of years working in the company
- work_accident: workplace injury
- Attrition:whether_separation
- promotion_last_5years: whether promotion in the past five years
- department: company division
- salary: salary

IV. Data cleaning

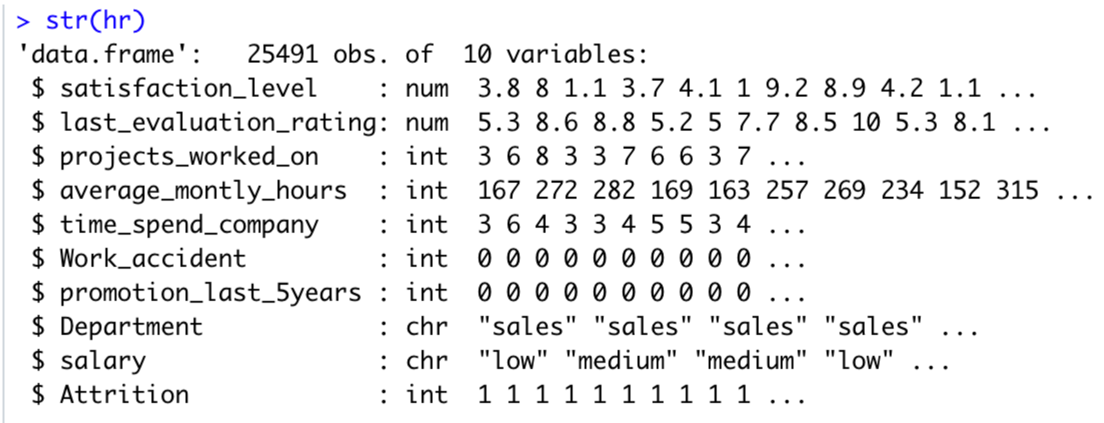


```
> colnames(hr)<-c("satisfaction_level","last_evaluation","project_worked_on","average_montly_hours","time_spend_company","wor
k_accident","promotion_last_5years","department","salary","attrition")
> hr$department<-factor(hr$department)
> hr$salary<-factor(hr$salary,levels=c("low","medium","high"))
> sum(is.na(hr))
[1] 0
```

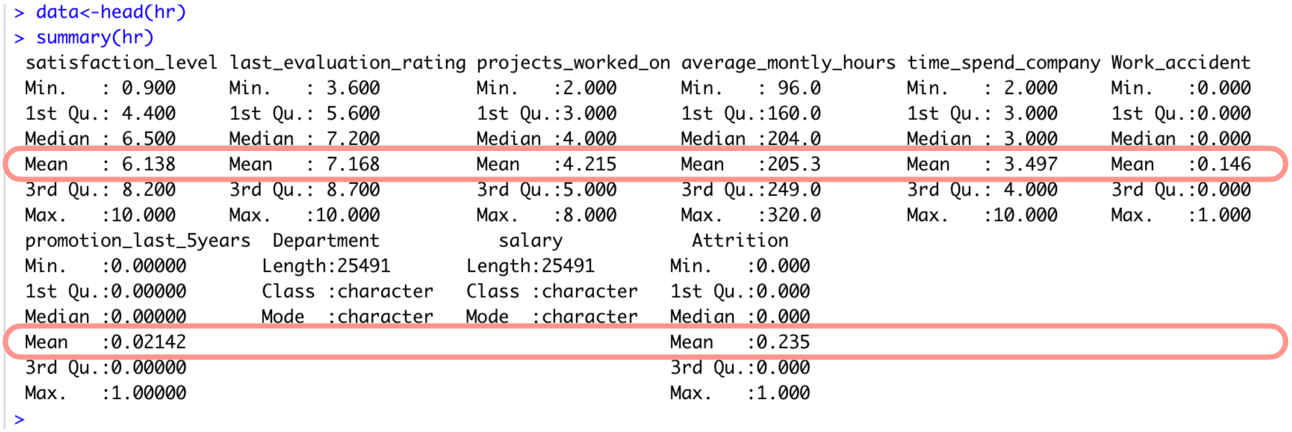
After variable renaming and factorization, there is no missing values are found in this dataset.

V. Data Analysis

1. Main descriptive statistics of each variable



```
> str(hr)
'data.frame':	25491 obs. of  10 variables:
 $ satisfaction_level     : num  3.8 8 1.1 3.7 4.1 1 9.2 8.9 4.2 1.1 ...
 $ last_evaluation_rating: num  5.3 8.6 8.8 5.2 5 7.7 8.5 10 5.3 8.1 ...
 $ projects_worked_on     : int  3 6 8 3 3 7 6 6 3 7 ...
 $ average_montly_hours   : int  167 272 282 169 163 257 269 234 152 315 ...
 $ time_spend_company     : int  3 6 4 3 3 4 5 5 3 4 ...
 $ Work_accident          : int  0 0 0 0 0 0 0 0 0 0 ...
 $ promotion_last_5years  : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Department             : chr  "sales" "sales" "sales" "sales" ...
 $ salary                 : chr  "low" "medium" "medium" "low" ...
 $ Attrition              : int  1 1 1 1 1 1 1 1 1 1 ...
```



```
> data<-head(hr)
> summary(hr)
 satisfaction_level last_evaluation_rating projects_worked_on average_montly_hours time_spend_company Work_accident
 Min.   : 0.900     Min.   : 3.600         Min.   :2.000      Min.   : 96.0        Min.   : 2.000     Min.   :0.000
 1st Qu.: 4.400     1st Qu.: 5.600         1st Qu.:3.000      1st Qu.:160.0        1st Qu.: 3.000     1st Qu.:0.000
 Median : 6.500     Median : 7.200         Median :4.000      Median :204.0        Median : 3.000     Median :0.000
 Mean   : 6.138     Mean   : 7.168         Mean   :4.215      Mean   :205.3        Mean   : 3.497     Mean   :0.146
 3rd Qu.: 8.200     3rd Qu.: 8.700         3rd Qu.:5.000      3rd Qu.:249.0        3rd Qu.: 4.000     3rd Qu.:0.000
 Max.   :10.000     Max.   :10.000         Max.   :8.000      Max.   :320.0        Max.   :10.000     Max.   :1.000
 promotion_last_5years  Department           salary            Attrition
 Min.   :0.00000       Length:25491       Length:25491       Min.   :0.000
 1st Qu.:0.00000       Class :character   Class :character   1st Qu.:0.000
 Median :0.00000       Mode  :character   Mode  :character   Median :0.000
 Mean   :0.02142                                             Mean   :0.235
 3rd Qu.:0.00000                                             3rd Qu.:0.000
 Max.   :1.00000                                             Max.   :1.000
>
```

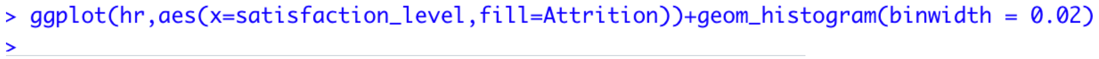
There are 25491 rows of data in this dataset, and from the statistical results,

- Employee satisfaction with the company: level6
- Employee rating: 7.2
- Average number of projects per employee: 4
- Average number of hours worked by employees: 205h/month
- Average number of years worked by employees: 3
- Accident rate: 15%
- Promotion rate in the last 5 years: 2%
- Turnover rate: 23.5%

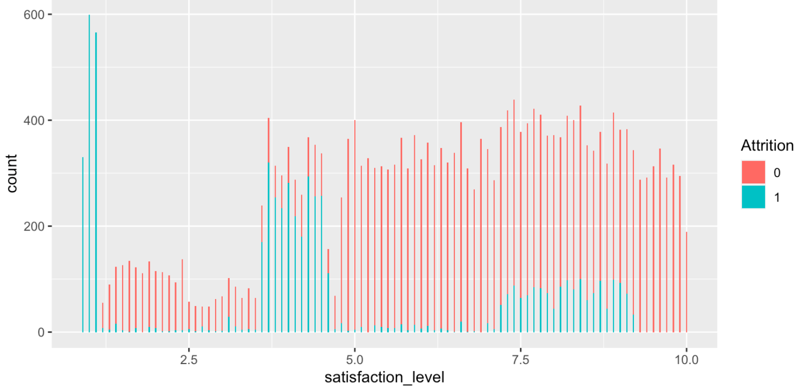
(all the numerical values are approximate values)

2. Relationship between variables and turnover

2.1 The relationship between employee satisfaction with the company and separation

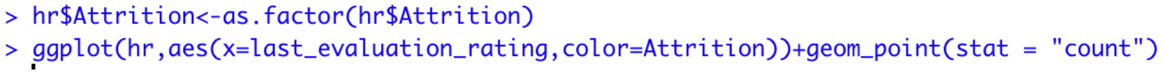


```
> ggplot(hr,aes(x=satisfaction_level,fill=Attrition))+geom_histogram(binwidth = 0.02)
>
```

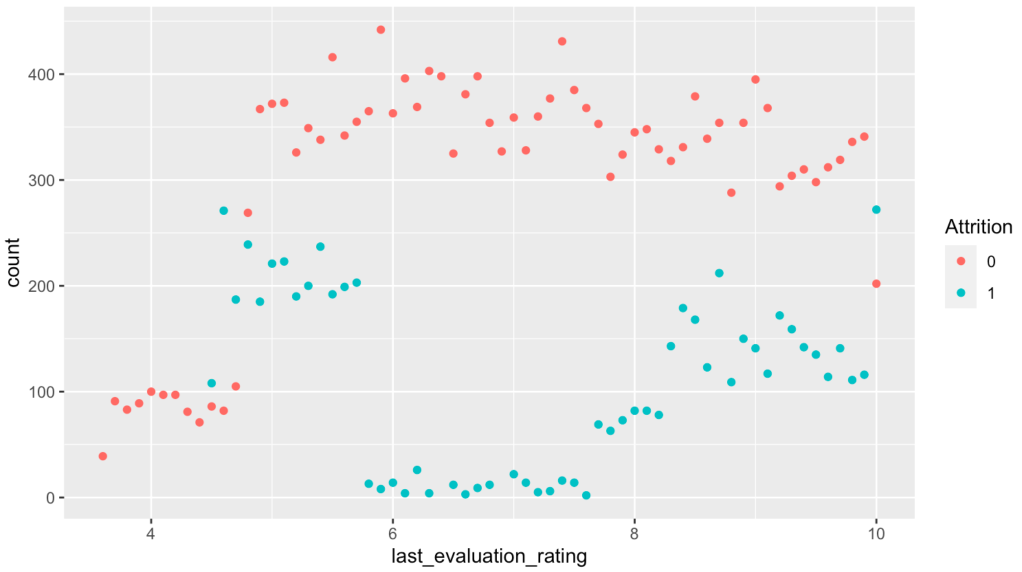


From the figure, it can be seen that most of the staff who have left the company are satisfied below level1, followed by level3-5, and the least number of leavers are satisfied between level7-9.

2.2 The relationship between staff evaluation and separation

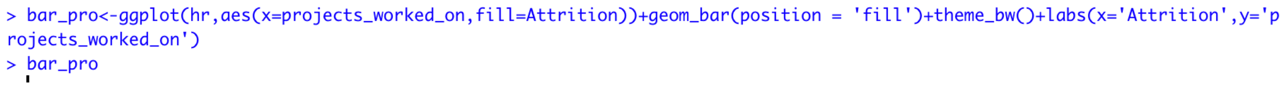


```
> hr$Attrition<-as.factor(hr$Attrition)
> ggplot(hr,aes(x=last_evaluation_rating,color=Attrition))+geom_point(stat = "count")
```

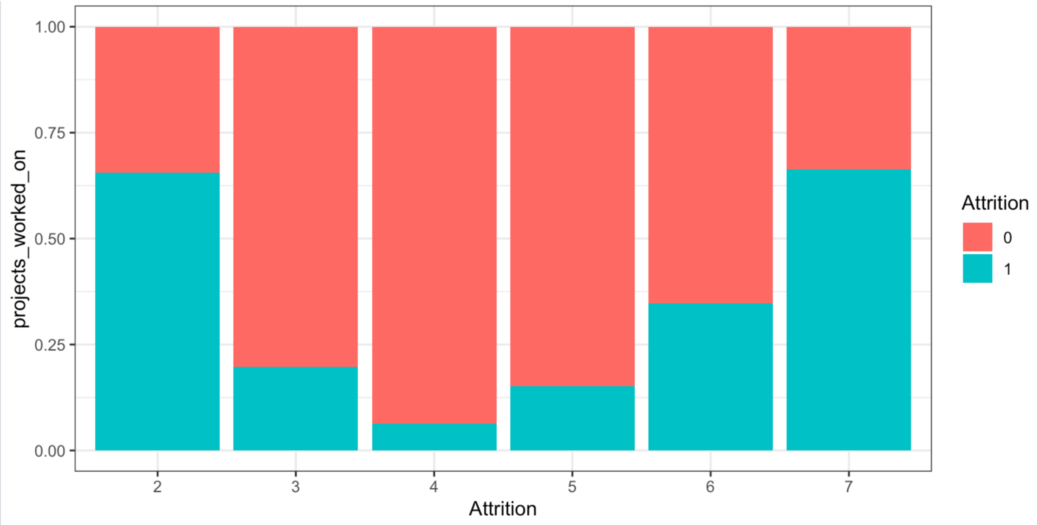


The statistical results show that the dispersion of staff evaluation is high, and most of the evaluations of separated staff and retained staff are concentrated in the range of 5-10, and even the separated staff have given high scores.

2.3 The relationship between the number of employees participating in projects and leaving

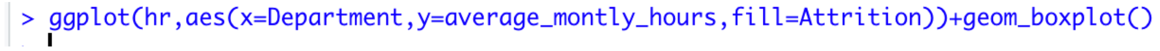


```
> bar_pro<-ggplot(hr,aes(x=projects_worked_on,fill=Attrition))+geom_bar(position = 'fill')+theme_bw()+labs(x='Attrition',y='p
rojects_worked_on')
> bar_pro
```

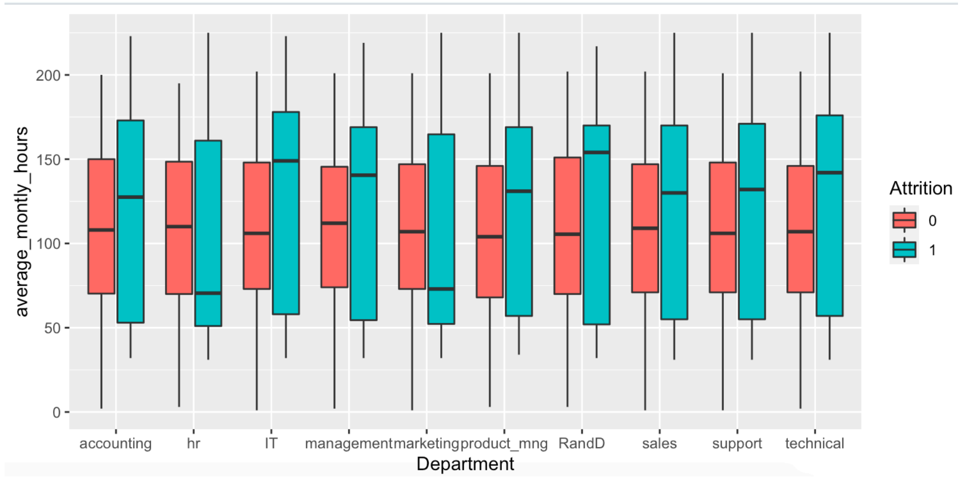


As shown in the figure, the interval of having participated in 2 and 7 projects brings together the largest number of employees who have left the company, and the separation rate of employees who have participated in more than 4 projects is getting higher.

2.4 Relationship between average monthly working hours and turnover in different departments

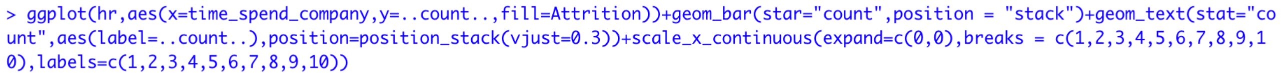


```
> ggplot(hr,aes(x=Department,y=average_montly_hours,fill=Attrition))+geom_boxplot()
```

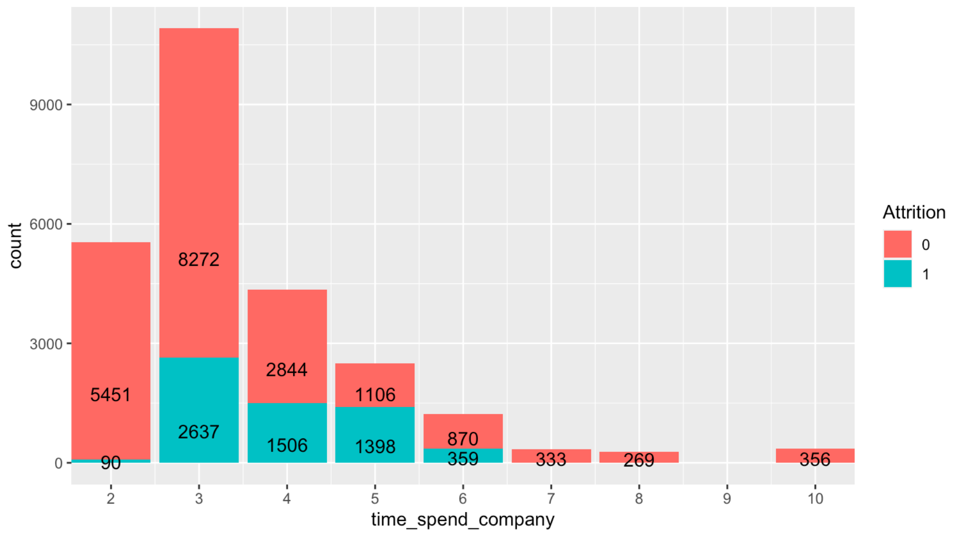


The statistical results show that the average monthly working hours of retained employees are relatively stable regardless of the department, while the average monthly working hours of leavers are polarized, implying that the average monthly working hours of employees in different departments are more stable.
The average monthly working hours of employees who leave the company are polarized, which means that long or short working hours are related to employees leaving the company.

2.5 Relationship between the number of years of service and employee turnover

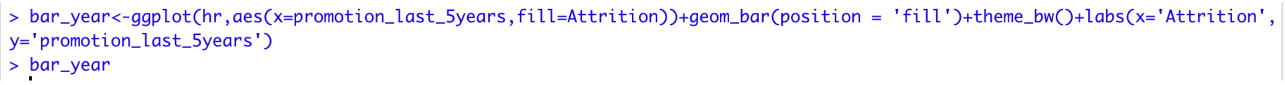


```
> ggplot(hr,aes(x=time_spend_company,y=..count..,fill=Attrition))+geom_bar(star="count",position = "stack")+geom_text(stat="co
unt",aes(label=..count..),position=position_stack(vjust=0.3))+scale_x_continuous(expand=c(0,0),breaks = c(1,2,3,4,5,6,7,8,9,1
0),labels=c(1,2,3,4,5,6,7,8,9,10))
```

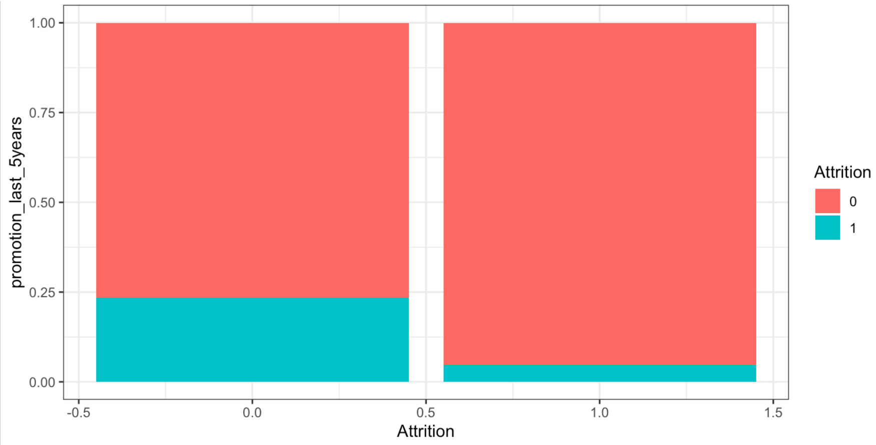


As can be seen from the graph, the largest number of employees left the company between 3 and 5 years of service, with a decreasing trend, and no employees left the company after 7 years of service.

2.6 The relationship between the presence of promotion and separation within 5 years

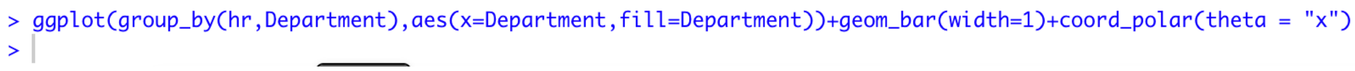


```
> bar_year<-ggplot(hr,aes(x=promotion_last_5years,fill=Attrition))+geom_bar(position = 'fill')+theme_bw()+labs(x='Attrition',
y='promotion_last_5years')
> bar_year
```

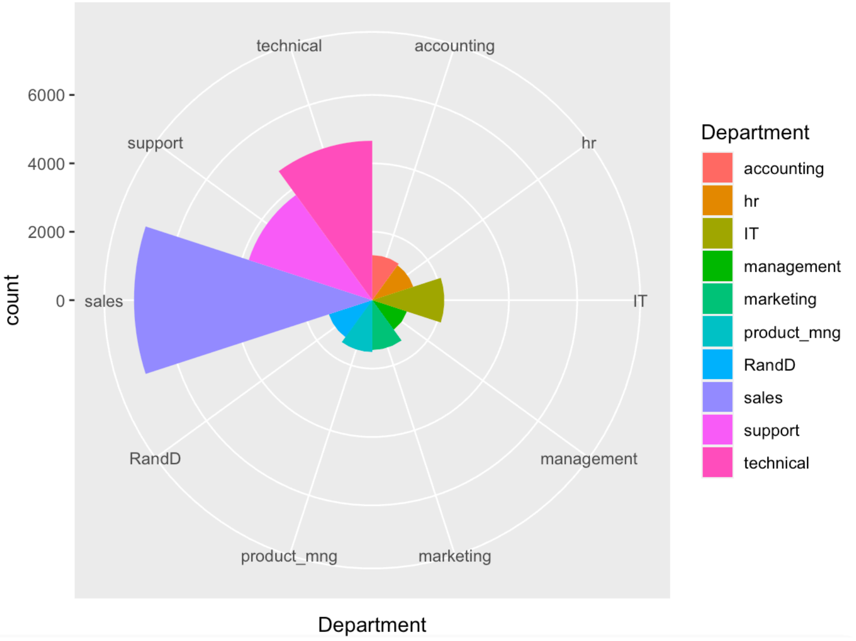


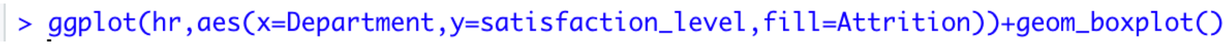
It can be seen that the separation rate of employees who have not been promoted in the past 5 years is much higher than the separation rate of employees who have been promoted.

2.7 Relationship between Departments and Separation

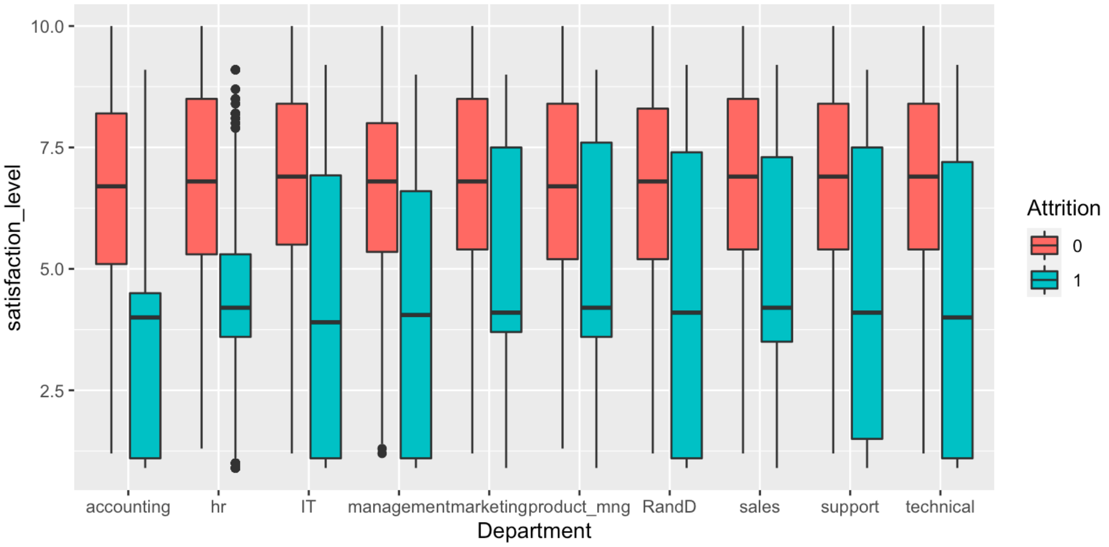


```
> ggplot(group_by(hr,Department),aes(x=Department,fill=Department))+geom_bar(width=1)+coord_polar(theta = "x")
>
```



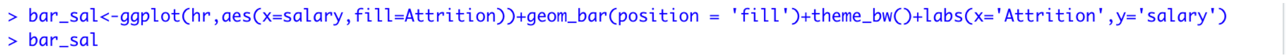


```
> ggplot(hr,aes(x=Department,y=satisfaction_level,fill=Attrition))+geom_boxplot()
```

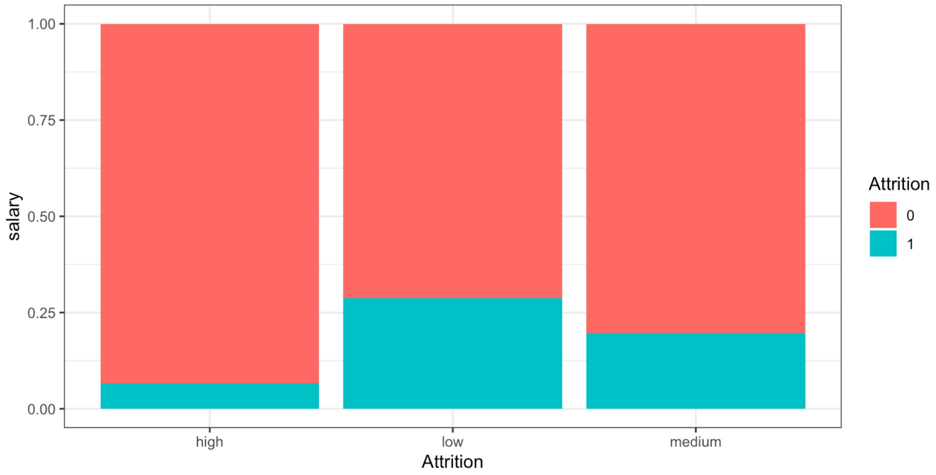


The graph clearly shows that the satisfaction rate of retained employees in each department is more evenly distributed, while the satisfaction rate of separated employees is very different from that of retained employees.

2.8 Relationship between employee salary and turnover

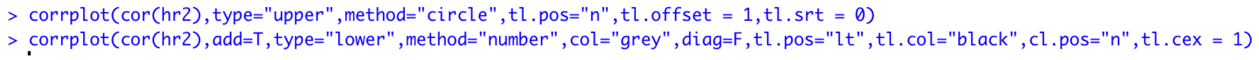


```
> bar_sal<-ggplot(hr,aes(x=salary,fill=Attrition))+geom_bar(position = 'fill')+theme_bw()+labs(x='Attrition',y='salary')
> bar_sal
```

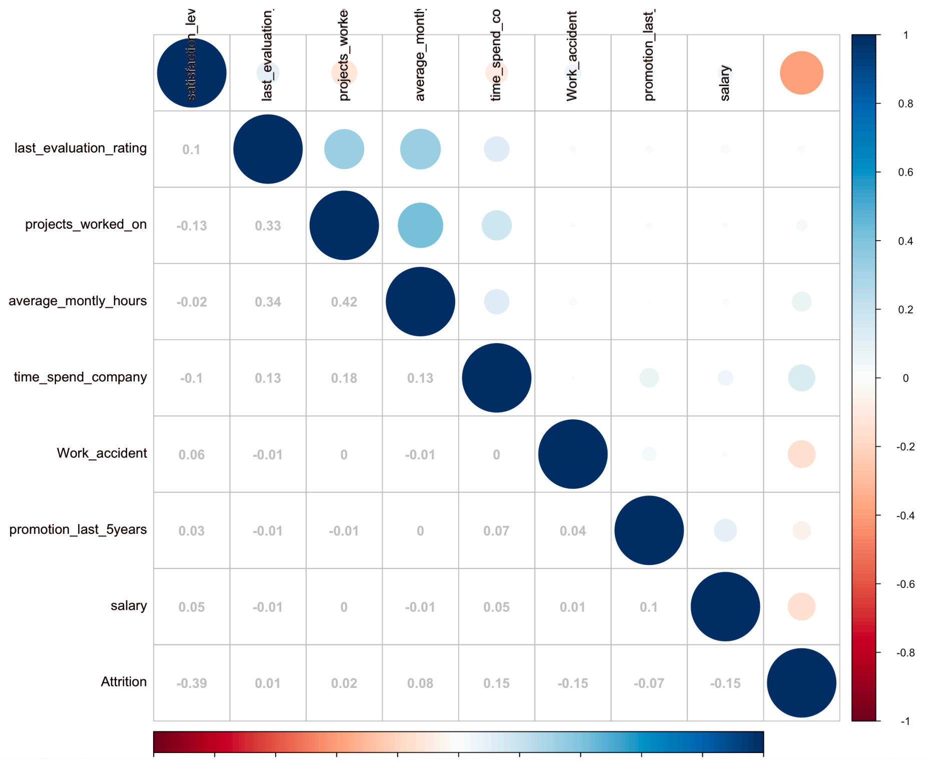


It is easy to see that the higher the salary, the lower the turnover rate.

3. 1 Correlation coefficient plot

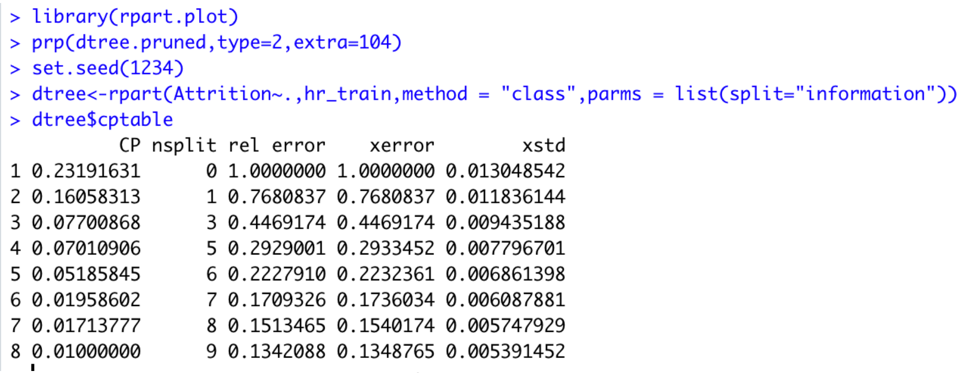


```
> corrplot(cor(hr2),type="upper",method="circle",tl.pos="n",tl.offset = 1,tl.srt = 0)
> corrplot(cor(hr2),add=T,type="lower",method="number",col="grey",diag=F,tl.pos="lt",tl.col="black",cl.pos="n",tl.cex = 1)
```

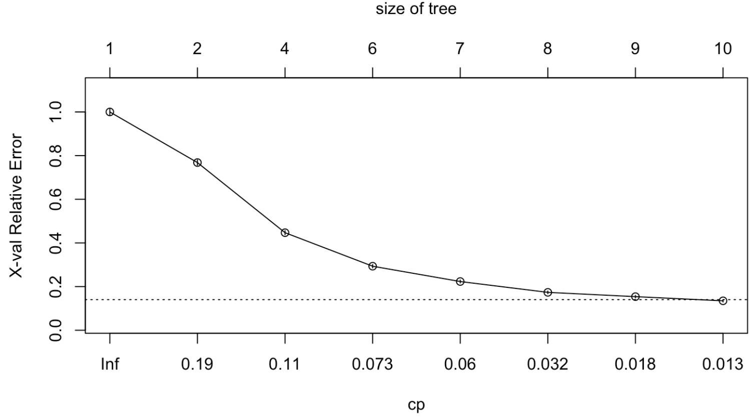


The color bar on the right side of the correlation coefficient graph represents the correlation, blue is positive correlation and red is negative correlation, the closer to the two ends the greater the correlation coefficient. From the graph, the factor with the highest correlation with turnover is satisfaction (-0.39), which has a high correlation, followed by salary (-0.15), years of service (0.15), and job errors (-0.15).

3.2 Decision tree

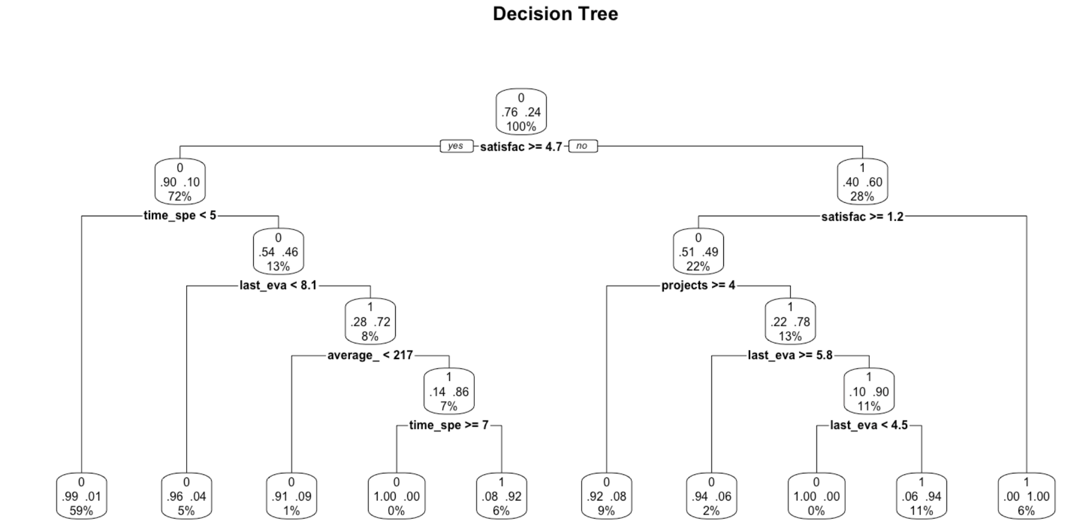


```
> library(rpart.plot)
> prp(dtree.pruned,type=2,extra=104)
> set.seed(1234)
> dtree<-rpart(Attrition~.,hr_train,method = "class",parms = list(split="information"))
> dtree$cptable
          CP nsplit rel error    xerror        xstd
1 0.23191631      0 1.0000000 1.0000000 0.013048542
2 0.16058313      1 0.7680837 0.7680837 0.011836144
3 0.07700868      3 0.4469174 0.4469174 0.009435188
4 0.07010906      5 0.2929001 0.2933452 0.007796701
5 0.05185845      6 0.2227910 0.2232361 0.006861398
6 0.01958602      7 0.1709326 0.1736034 0.006087881
7 0.01713777      8 0.1513465 0.1540174 0.005747929
8 0.01000000      9 0.1342088 0.1348765 0.005391452
```



The dashed line is based on an upper limit obtained by a standard deviation criterion. We select the tree corresponding to the leftmost cp value under the dashed line and prune the decision tree.

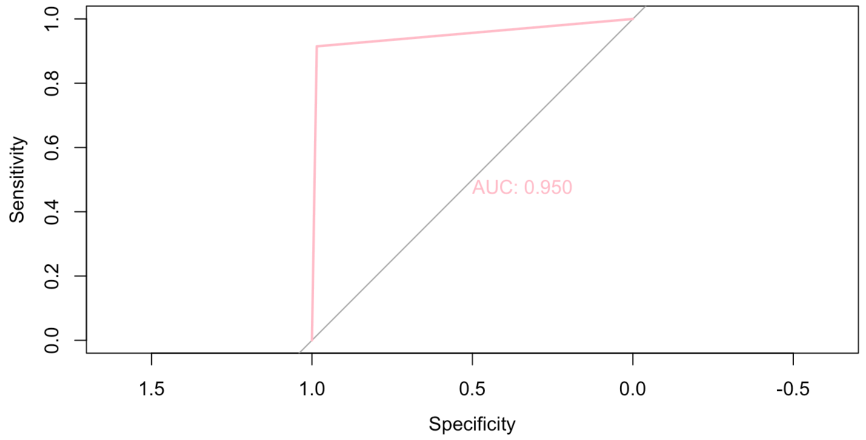
```
> dtree.pruned<-prune(dtree,cp=0.01)
> library(rpart.plot)
> prp(dtree.pruned,type = 2,extra = 104,fallen.leaves = TRUE,main="Decision Tree")
```



The top of the decision tree is the satisfaction level, and the satisfaction level is greater than or equal to 4.7. If the satisfaction level holds, the tree goes down from the left branch, otherwise it goes down from the right branch. The classification is completed when the observation reaches the end node.

```
> dtree.pruned<-prune(dtree,cp=0.01)
> dtree.pruned.pred<-predict(dtree.pruned,hr_good_train,type="class")
> roc(as.numeric(hr_good_train$Attrition),as.numeric(dtree.pruned.pred),plot=TRUE,print.thres=TRUE,print.auc=TRUE,col="pink")
```

3.3 ROC/AUC curve



AUC is the area under the ROC curve, and the closer the AUC is to 1, the better the prediction model is. The AUC = 0.95, which means the prediction model is very effective.