Human Resources Dataset Analysis

MATH 257 GROUP PROJECT

Zhengxia Yi, Miyetani Chauke, Ray Chen, Qiaoqiao Jiang, and Junyang Liu

5/25/2017

Project Background

A successful company is normally created by a group of really talented people. Ideally, as company grows, more and more talented people will join the company. However, we also observe the common fact that as the company grows, more and more people leave the company for various reasons. In this project, we are interested in exploring the reasons and identifying the characteristics of the employee that left the company. Hopefully, we can use what we have discovered in this project to help companies better assess their employees and make sure really talented people grow with the company.

Data

Our dataset is collected from Kaggle. The url of the dataset is provided here:

(https://www.kaggle.com/ludobenistant/hr-analytics)

The description of the dataset is down below. Dataset description:

- *satisfaction_level* : Employee satisfaction level for the company
- *last_evaluation* : Most recent employer satisfaction level for employee
- *number_project* : Number of projects completed at work
- *average_montly_hours* : Average hours at workplace in a month
- *time_spend_company* : Number of years spent in the company
- *Work_acciden* : Whether they have had a work accident
- *left* : Whether the employee left the workplace or not
- *promotion_last_5years* : Whether they have had a promotion in the last 5 years
- *sales* : Department
- *Salary* : level of salary

Goal

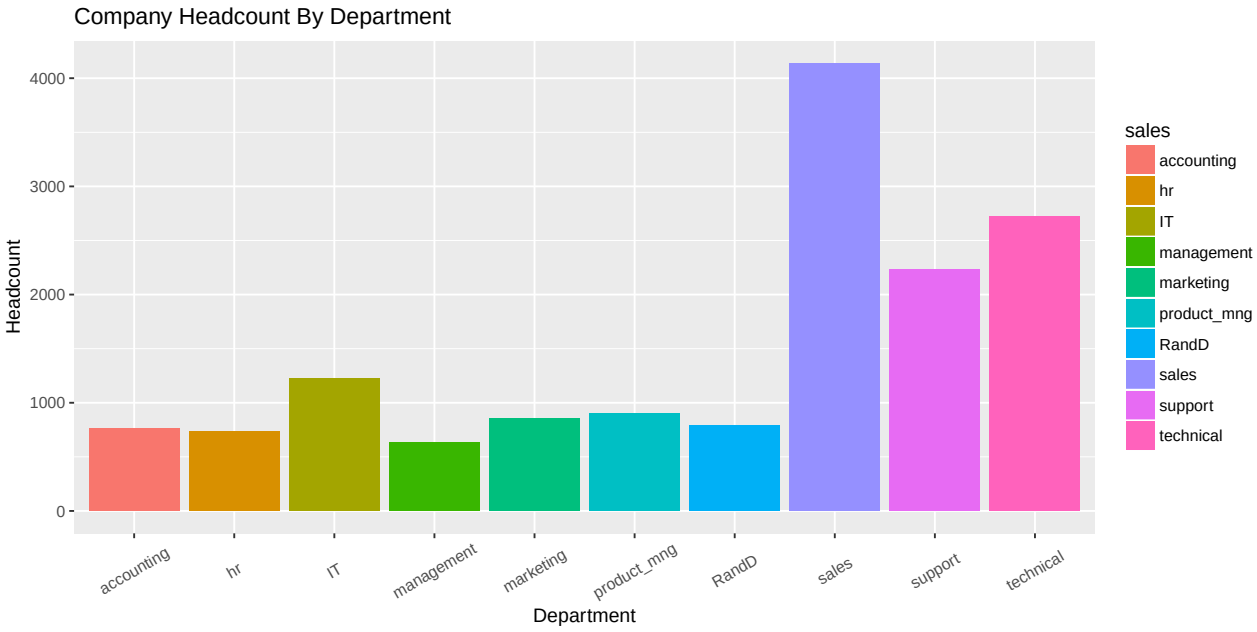
The Goal of this project is to discover the patterns of people who left the company. What is influencing people leaving the company? we also want to focus on why good people leave, and verify our finding by applying the finding we obtain for prediction analysis and see how accurate our conclusion is. We believe our finding has important applications in the business world in terms of HR analytics, but is also applicable to customer churn and student retention. Hopefully, we can have better understanding of assessing our employee and keep good employees stay with company.

Analysis

Preliminary Data analysis

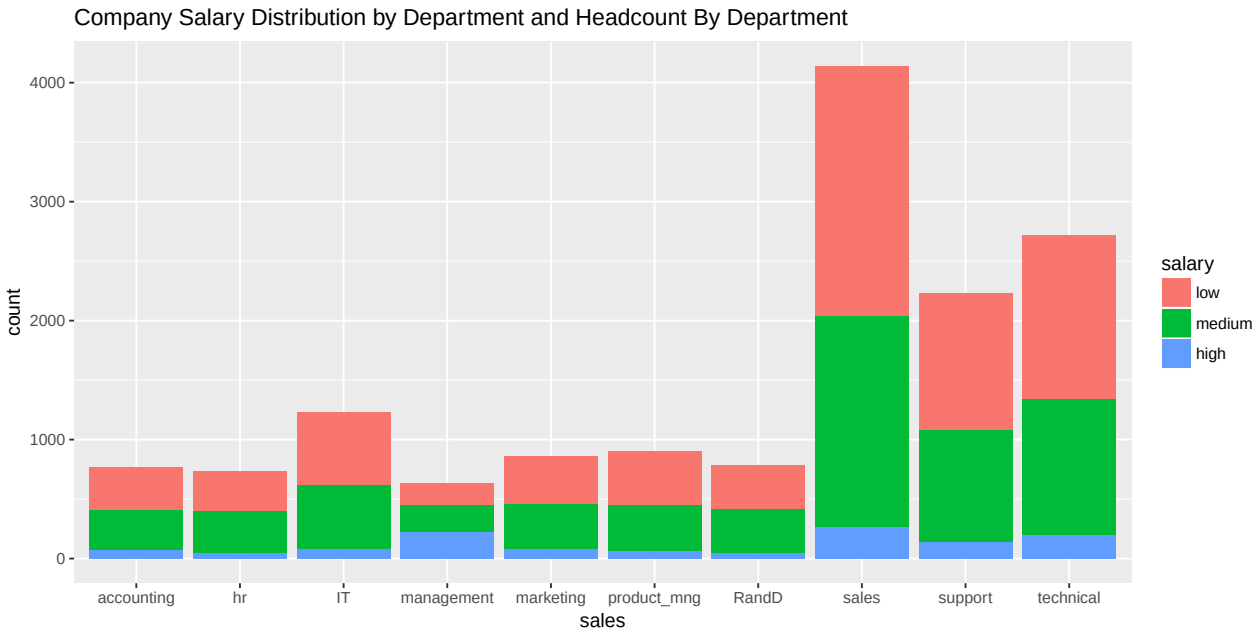
Before we dive into any specific analysis,let us get to know more about this dataset.The dataset tells us a lot of stories if we look closely. let's first take a look at the employee structure of the company.

```
## Loading required package: lattice
```

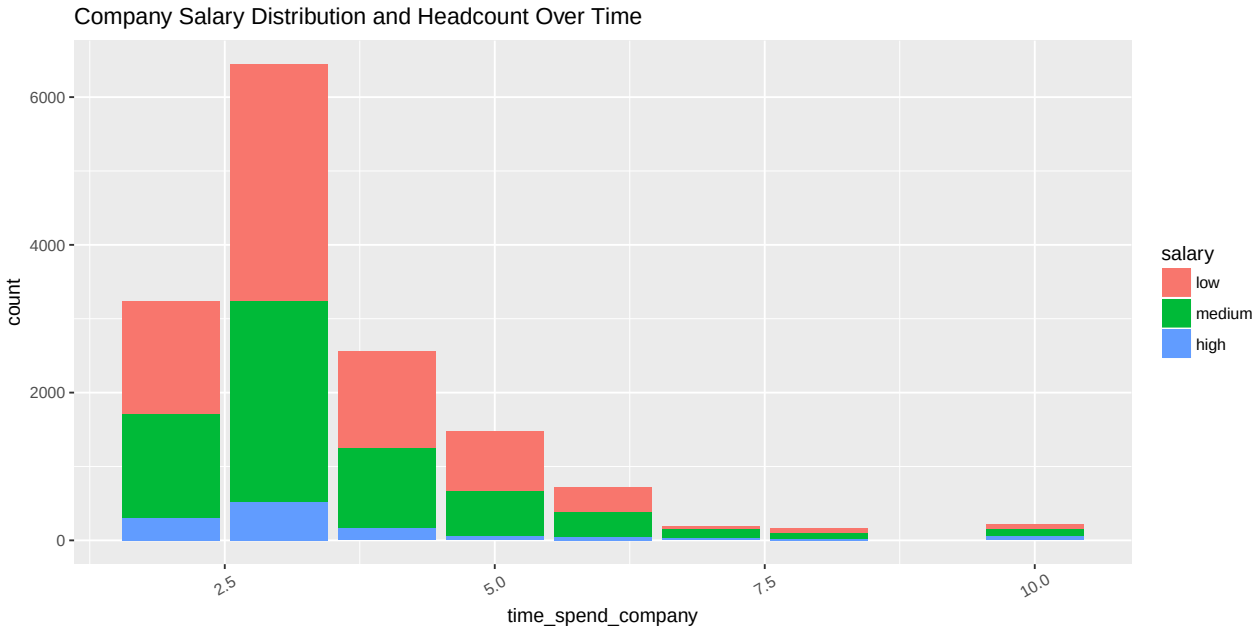


From the Headcount histogram we can tell that *sales*, *technical*, and *support* department are the departments with most employees.

Now let's take a look the salary structure of employees in the company.



From the Stacked chart above, we can see that *sales*, *support*, and *technical* department have lots of people earning low level salary. It might raise concerns when most employees of the company's top 3 biggest departments are earning low salary.



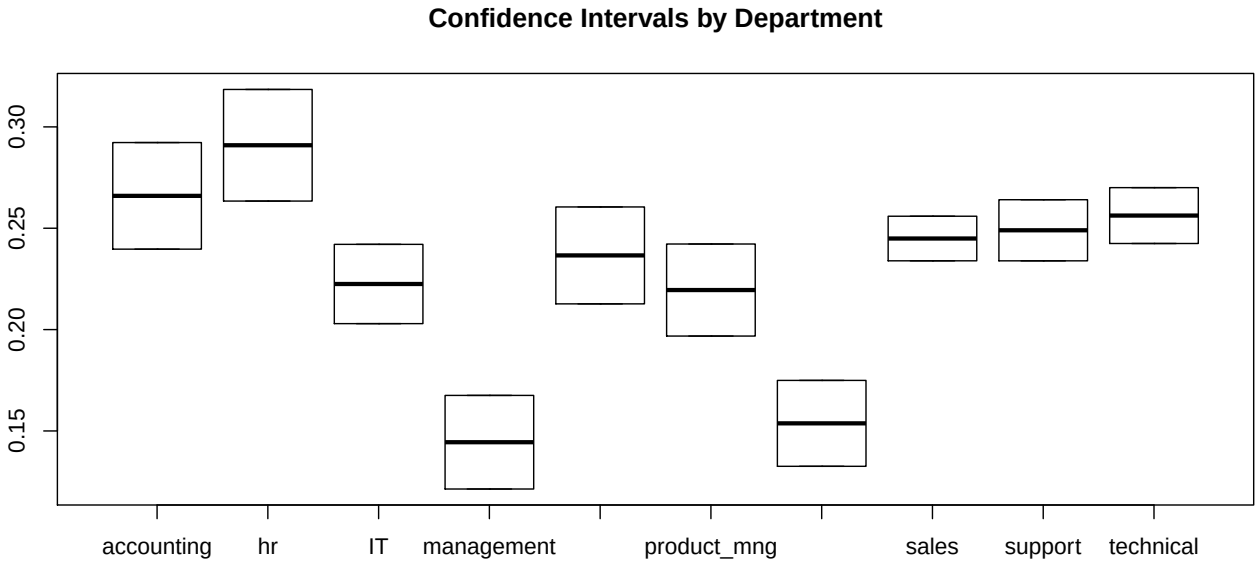
This stacked barchart shows the distribution and headcounts of the employee salary over time.we can see that most of employes are at their 3rd and 4th year of work. we are a little concered about employees with 3-4 years of experience who still earn low or medium salary. It means that perhaps they are not appreciated in the company or maybe too shy or scared to ask for a rise.

Now, let's take a look at the employees who left the company. we want to know what is the characteristics of those people?let's find the confidence intervals for all the department proportions of those who left.

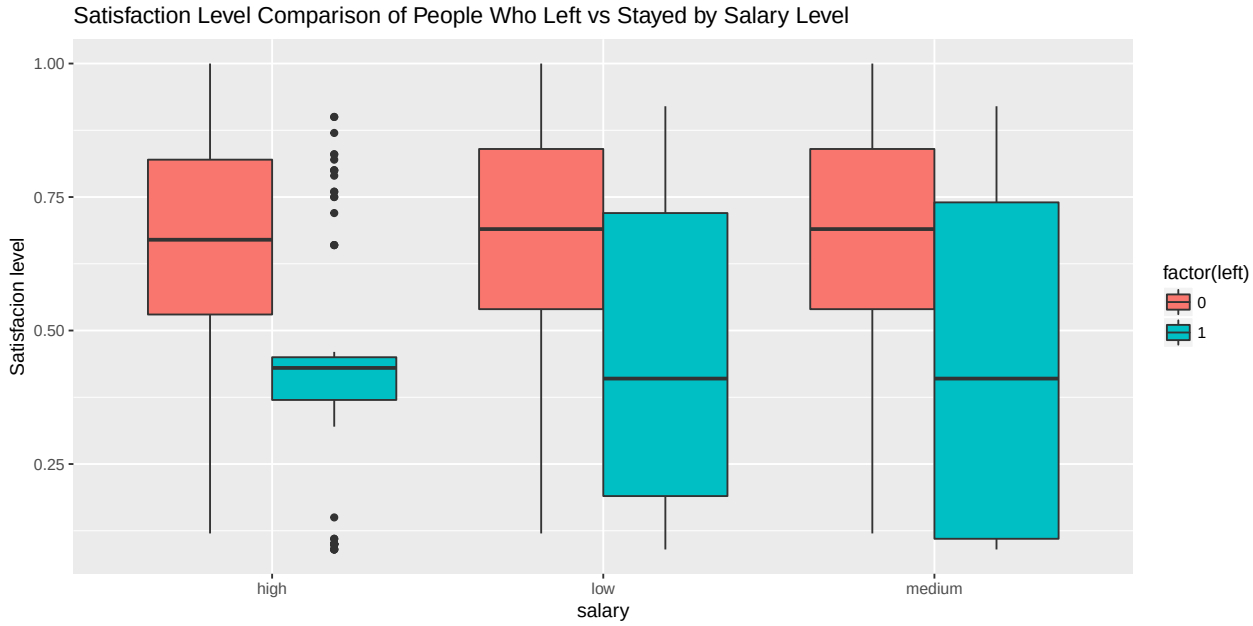
A 95% CI's for department proportion of those who left is performed:

```
##                lower     upper      mean
## accounting  0.2396800 0.2922626 0.2659713
## hr          0.2633987 0.3184687 0.2909337
## IT          0.2029405 0.2420472 0.2224939
## management  0.1213548 0.1675341 0.1444444
## marketing   0.2126918 0.2605017 0.2365967
## product_mng 0.1968071 0.2422173 0.2195122
## RandD       0.1325607 0.1749361 0.1537484
## sales       0.2339301 0.2559249 0.2449275
## support     0.2339153 0.2640658 0.2489906
## technical   0.2424742 0.2700258 0.2562500
```

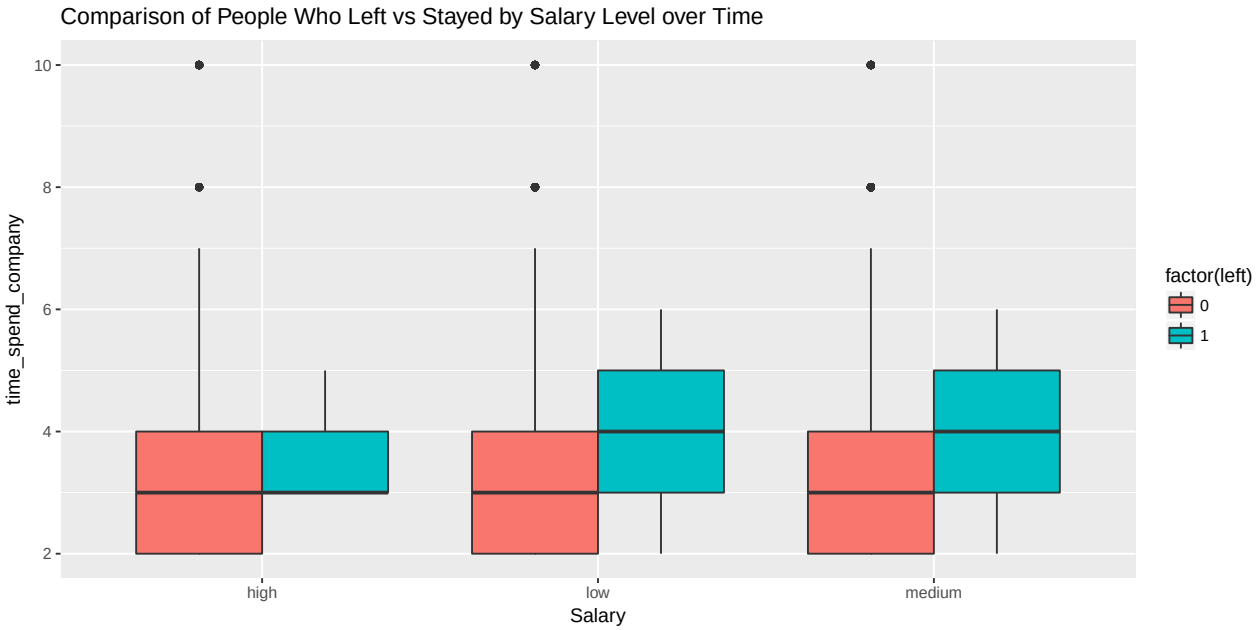
From the confidence intervals above we can see that although the ratio varies across departments, it stays very stay across different departments.



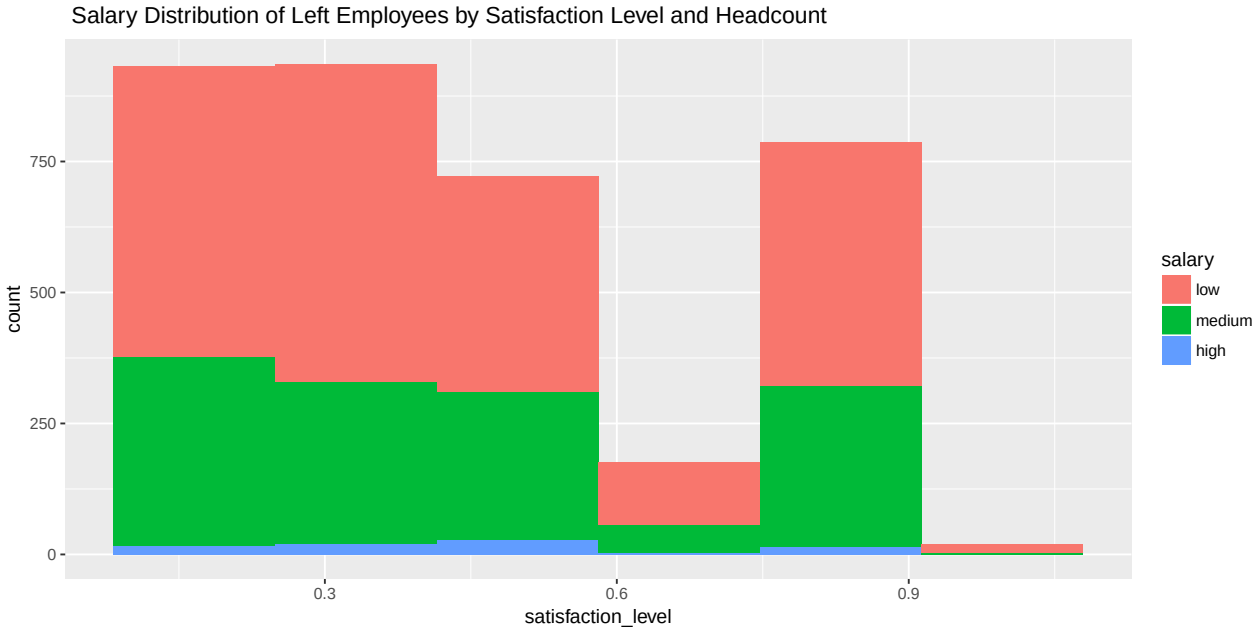
This graph provides the confidence intervals for the proportion of employees leaving for each department. Looking at the box plot and the data frame of confidence intervals, we see that the CI for department hr overlaps only with accounting, technical, and support. This suggests that hr is statistically different from the rest of the departments at a 95% significance level. We also see that management and RandD departments overlap only with each other, meaning that those two departments are statistically different from the rest of the departments at a 95% significance level.



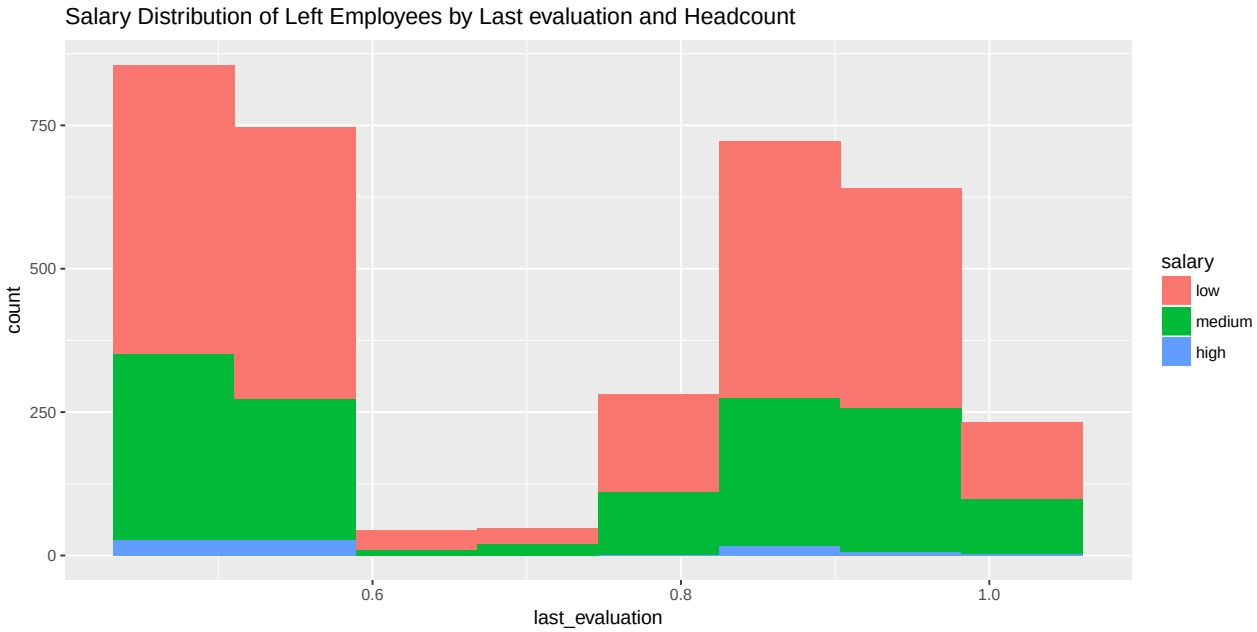
From the Boxplot we can conclude that the people who left company have lower satisfaction level than the people who stayed in the company in general.



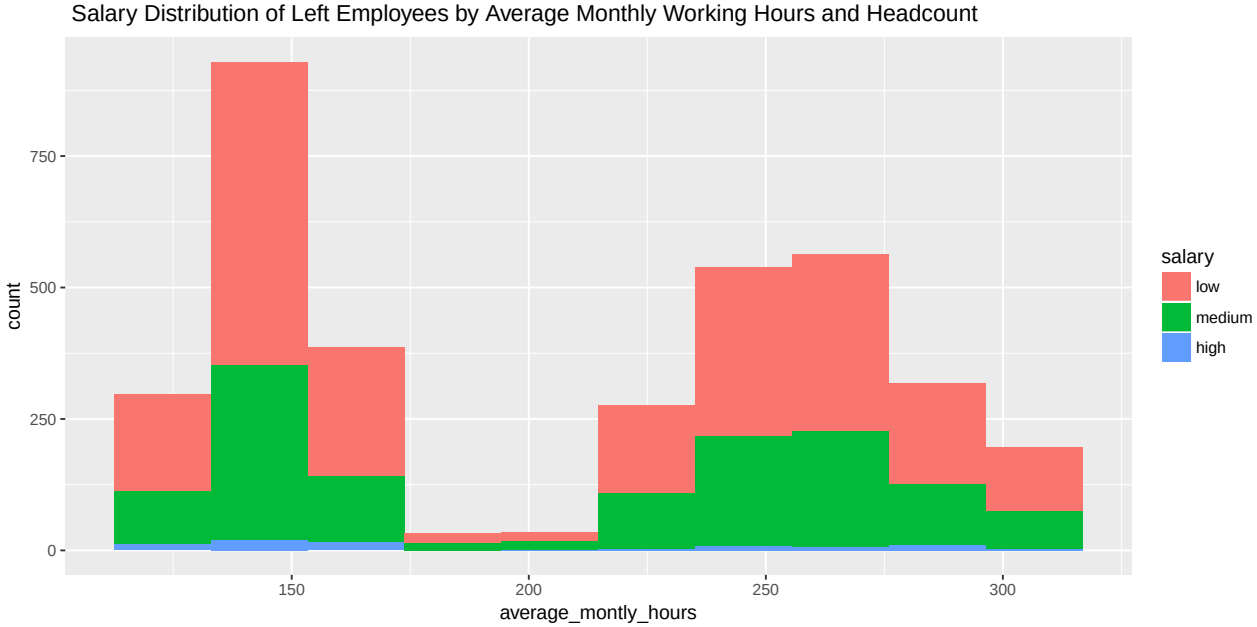
From the boxplot we can conclude that people who left the company tend to spend more time with company. It is probably due to overtime.



From the staced barchart above, we can see a lot of people who left the company when they are paid with low or midium salary. Furthermore, it shows that a lot of emmployees left they company when their satisfaction level was low. It also shows that employees rarely left the company if they are very satisfied with the comapany or have high salary.



From the stacked barchart above, we can see many people who did perform very well and earned medium or low salary left the company. However, we see there are also a lot of top performers that earned medium or low salary left the company too. It is worth to note that salary might be a driver for their departure.



From the stacked bar chart above, we can see the mixed pattern.A lot of low salary employees who worked either more than average or less than average left company. It can imply that the one who falls into the first group worked over time and didn't get compensation properly.

We know that companies don't want to retain everybody. Some people don't work well as we can see from their evaluation, but clearly there are also many good workers left the company. These are the people the company should have retained.

We define that the people who received an evaluation above average, or spent at least four years in the company, or worked on more than 5 projects are "good employees". We will apply this definition in later analysis.

Now, let's dive deeper into the data.

Data setup

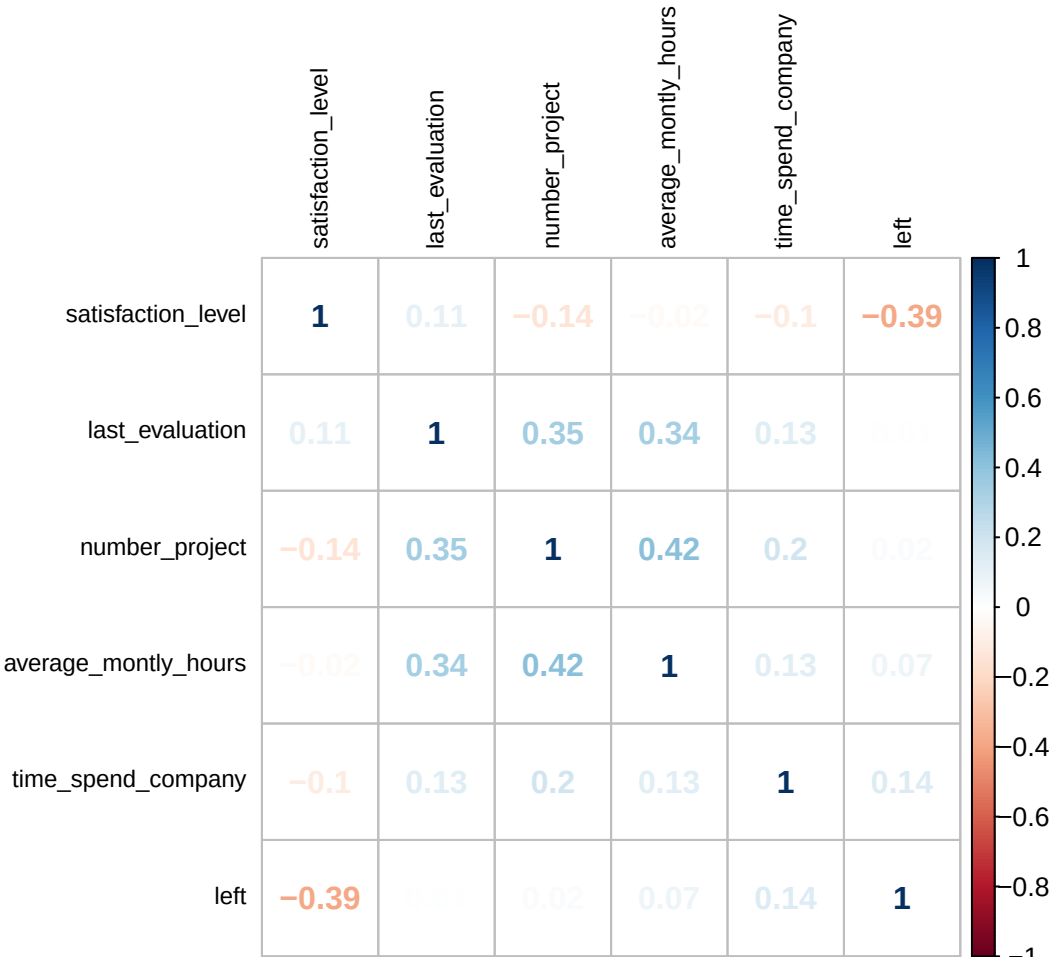
```
str(HR)
```

```
## 'data.frame':    14999 obs. of  10 variables:
##  $ satisfaction_level   : num  0.38 0.8 0.11 0.72 0.37 0.41 0.1 0.92 0.89 0.42 ...
##  $ last_evaluation      : num  0.53 0.86 0.88 0.87 0.52 0.5 0.77 0.85 1 0.53 ...
##  $ number_project       : int  2 5 7 5 2 2 6 5 5 2 ...
##  $ average_montly_hours : int  157 262 272 223 159 153 247 259 224 142 ...
##  $ time_spend_company   : int  3 6 4 5 3 3 4 5 5 3 ...
##  $ Work_accident        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ left                 : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ promotion_last_5years: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ sales                : Factor w/ 10 levels "accounting","hr",..: 8 8 8 8 8 8 8 8 8 ...
##  $ salary               : Ord.factor w/ 3 levels "low"<"medium"<..: 1 2 2 1 1 1 1 1 1 ...
```

```
summary(HR)
```

```
##  satisfaction_level last_evaluation  number_project  average_montly_hours
##  Min.   :0.0900     Min.   :0.3600   Min.   :2.000   Min.   : 96.0
##  1st Qu.:0.4400     1st Qu.:0.5600   1st Qu.:3.000   1st Qu.:156.0
##  Median :0.6400     Median :0.7200   Median :4.000   Median :200.0
##  Mean   :0.6128     Mean   :0.7161   Mean   :3.803   Mean   :201.1
##  3rd Qu.:0.8200     3rd Qu.:0.8700   3rd Qu.:5.000   3rd Qu.:245.0
##  Max.   :1.0000     Max.   :1.0000   Max.   :7.000   Max.   :310.0
##
##  time_spend_company Work_accident         left
##  Min.   : 2.000     Min.   :0.0000   Min.   :0.0000
##  1st Qu.: 3.000     1st Qu.:0.0000   1st Qu.:0.0000
##  Median : 3.000     Median :0.0000   Median :0.0000
##  Mean   : 3.498     Mean   :0.1446   Mean   :0.2381
##  3rd Qu.: 4.000     3rd Qu.:0.0000   3rd Qu.:0.0000
##  Max.   :10.000     Max.   :1.0000   Max.   :1.0000
##
##  promotion_last_5years         sales         salary
##  Min.   :0.00000       sales      :4140   low   :7316
##  1st Qu.:0.00000       technical  :2720   medium:6446
##  Median :0.00000       support    :2229   high  :1237
##  Mean   :0.02127       IT         :1227
##  3rd Qu.:0.00000       product_mng: 902
##  Max.   :1.00000       marketing  : 858
##                        (Other)    :2923
```

From the structure and summary of our dataset, we can find that *Work_accident*, *left promotion_last_5years*, *salary*,and *sales* should be factors variables. Therefore, we converted them into corrected form and eliminated them in variable correlation plot.



From the correlation plot, we can see that there is no strong correlation between any of two variables.

Principle Component Analysis

Now we want to know what the key components of our numeric variables are. These components may lead us to the deeper layer of our data.

```
## Standard deviations:
## [1] 1.8154384 1.1424933 0.4145560 0.3614769 0.3104356
##
## Rotation:
##                                PC1         PC2         PC3          PC4
## satisfaction_level     0.06494749 -0.85317465  0.35295820 -0.147687169
## last_evaluation        0.52235326 -0.06326494  0.43058817  0.587001544
## number_project         0.49370026  0.31829886 -0.01564635  0.194516383
## average_montly_hours   0.51009467  0.19468709  0.25453360 -0.771867256
## time_spend_company     0.46796622 -0.35898933 -0.79055814  0.001415325
##                                PC5
## satisfaction_level     -0.3485453
## last_evaluation         0.4395084
## number_project         -0.7854072
## average_montly_hours    0.2033075
## time_spend_company      0.1647729
```

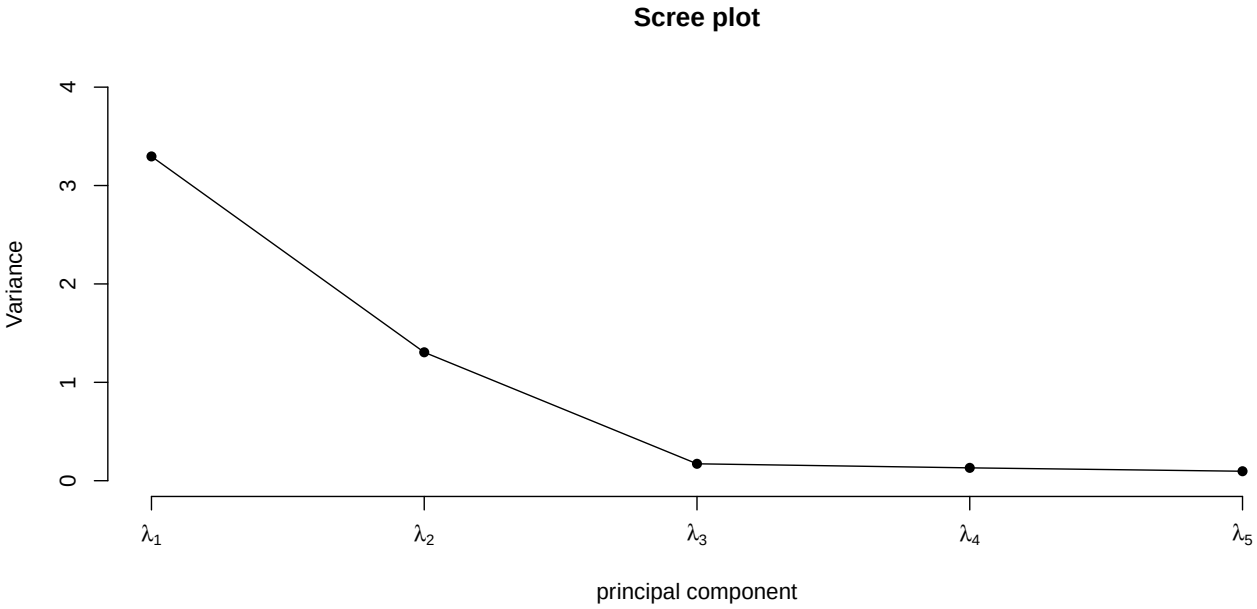
The 1st PC is combination of *satisfaction_level*, *last_evaluation*, *number_project*, *average_montly_hours*, and *time_spend_company*.Employees who are above average on *satisfaction_level*, *last_evaluation*, *number_project*, *average_montly_hours*, and *time_spend_company* tend to have high score on PC1. Employees who are below average on *satisfaction_level*, *last_evaluation*, *number_project*, *average_montly_hours*, and *time_spend_company* tend to have low score on PC1. We can imply that PC1 is overall fitness of the employee to the company. if the employee fits the company well and very is valuable to the company, they tend to have higher score.

The 2nd PC is the contrast between *satisfaction_level*, *last_evaluation*, *time_spend_company*, and *number_project*, *average_montly_hours*. Employees with high PC2 score do below average on *number_project*, *average_montly_hours* but do above average on *satisfaction_level*, *last_evaluation*, *time_spend_company*. Employees with low PC2 score do above average on *number_project*, *average_montly_hours* but do below average on *satisfaction_level*, *last_evaluation*, *time_spend_company*. We can imply that PC2 is related to work-life balance of the employee. Working long hours and doing more projects reduce the satisfaction level of the employee.

The cumulative percentage of variance explained by one more lambda is:

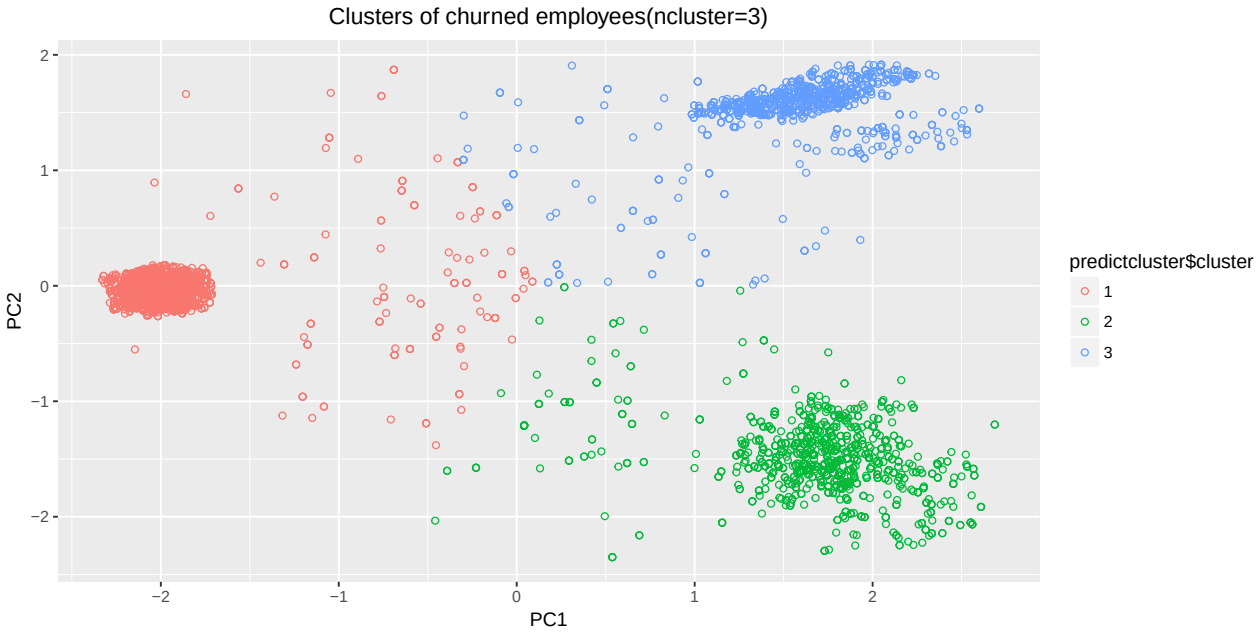
```
## [1] 0.6591633 0.9202215 0.9545928 0.9807260 1.0000000
```

From the cumulative sum of variance explained by each component we can conclude that in order to achieve the goal of "> 90% variance explained by PC" we need to keep 2 out of 5 PCs.It implies that there are 2 major components drive the dataset.



We normally keep the number of PCs that has variance greater than 1. There are 2 components have a variance value greater than 1. Moreover, we also see an obvious elbow point at the 2nd PC on the scree plot. It confirms the conclusion we draw previously which 2PCs are significant for the dataset.

cluster analysis on first two PC



There are 3 Groups of people:1st group has very high PC1 score and low PC2 score. 2nd group has very high PC2 score and low PC1 score. 3rd group has high PC1 score and high PC2 score. Therefore, we can imply that the 1st group is the group of people who is very valuable to the company because they sacrified their work-life balance and devoted their time to the company, the 2nd group is the group of people who are new to the company, the 3rd group of people is someone who really fit the company. They enjoyed their work and performed well.

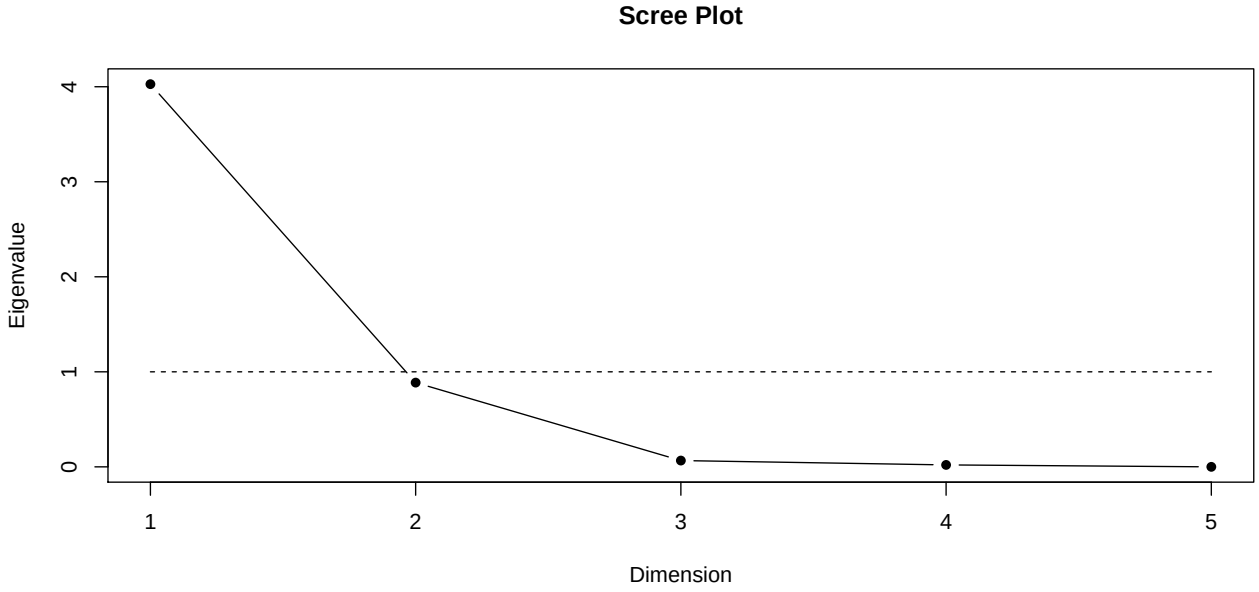
Factor Analysis

Here we try to perform factor analysis to see if we can discover similar patterns that we have right now.

```
##
## Call:
## factanal(x = hr[, c(1:5)], factors = 2, rotation = "varimax")
##
## Uniquenesses:
##   satisfaction_level      last_evaluation       number_project
##                0.040                0.127                0.089
## average_montly_hours   time_spend_company
##                0.140                0.210
##
## Loadings:
##                      Factor1 Factor2
## satisfaction_level            0.980
## last_evaluation       0.910   0.212
## number_project        0.932  -0.205
## average_montly_hours  0.926
## time_spend_company    0.750   0.477
##
##                Factor1 Factor2
## SS loadings      3.117   1.278
```

```
## Proportion Var   0.623   0.256
## Cumulative Var   0.623   0.879
##
## Test of the hypothesis that 2 factors are sufficient.
## The chi square statistic is 0.71 on 1 degree of freedom.
## The p-value is 0.399
```

We can see that first factor is dominated by *last_evaluation*, *number_project*, *average_montly_hours*, and *time_spend_company*, which implies that these employees are very valuable to the company. The second factor is donminated by *satisfaction_level*. Here we can suspect that the most important driver of employement is employee satisfaction vs employees' value to the company.



We here confirm again that 2 hiden factors is driving the employement.

logistic regression is also useful for factor analysis, let's do it on which factors affect whether or not an employee leaves the most.

```
## Warning: glm.fit: algorithm did not converge
```

```
##
## Call:
## glm(formula = left ~ ., family = binomial(link = "logit"), data = hr)
##
## Deviance Residuals:
##        Min          1Q      Median          3Q         Max
## -2.409e-06  -2.409e-06  -2.409e-06  -2.409e-06  -2.409e-06
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -2.657e+01  3.743e+04  -0.001    0.999
## satisfaction_level       -1.003e-13  3.730e+04   0.000    1.000
## last_evaluation          -2.617e-14  7.300e+04   0.000    1.000
## number_project           -3.950e-14  8.602e+03   0.000    1.000
## average_montly_hours      4.690e-16  2.285e+02   0.000    1.000
## time_spend_company       -5.667e-16  1.237e+04   0.000    1.000
## Work_accident1           -4.704e-14  2.817e+04   0.000    1.000
## promotion_last_5years1 -9.360e-14  8.248e+04   0.000    1.000
## saleshr                  -1.052e-15  3.484e+04   0.000    1.000
```

```
## salesIT               -8.273e-15  3.305e+04   0.000    1.000
## salesmanagement        8.389e-15  4.512e+04   0.000    1.000
## salesmarketing        -1.023e-14  3.540e+04   0.000    1.000
## salesproduct_mng       4.173e-15  3.559e+04   0.000    1.000
## salesRandD             1.633e-14  4.095e+04   0.000    1.000
## salessales             1.635e-13  2.750e+04   0.000    1.000
## salessupport          -1.282e-14  2.936e+04   0.000    1.000
## salestechnical         1.115e-14  2.840e+04   0.000    1.000
## salary2               -6.568e-14  1.265e+04   0.000    1.000
## salary3               -7.482e-14  4.032e+04   0.000    1.000
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 0.0000e+00  on 3570  degrees of freedom
## Residual deviance: 2.0717e-08  on 3552  degrees of freedom
## AIC: 38
##
## Number of Fisher Scoring iterations: 25
```

Number of Fisher Scoring iterations: 5

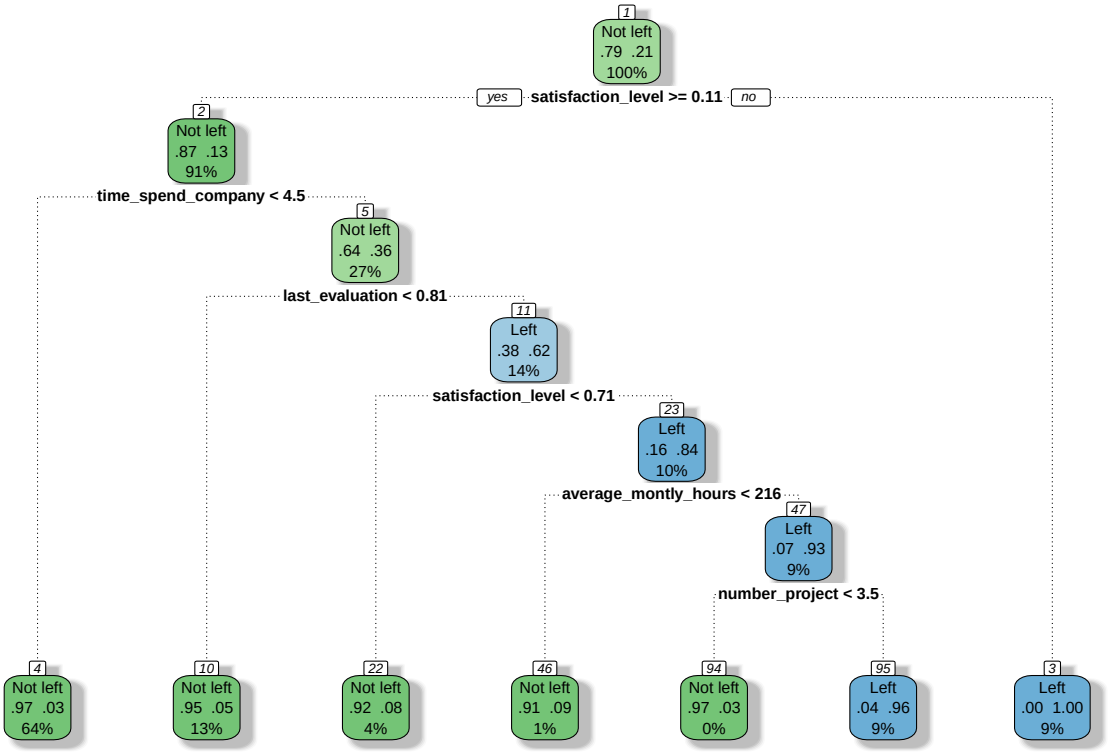
The only variables with p-values less than .05 come from factor levels of the department. Of the factor levels of department, only departments from management and research and development have estimates that are significant on a 95% level. Because these two departments have means that differ the most from the overall mean, it makes sense that only these two departments seem to have a significant effect on whether or not an employee leaves.

Classification And Regression Tree (CART) and Random forest model (RFM)

Now lets use our knowledges that we gain from previous analysis to perform predictions.

CART

First, a single decision tree using rpart; this can give an idea of the splits in the prediction of attrition from the company.This analysis looks only at the high performers(good people), which is defined as last_evaluation >= 0.70 or time_spend_company >= 4 or number_project > 5, and CART runs on a subset of the dataset using only employees whose last evaluation was >= 0.70, time spent at the company was >= 4, and number of projects > 5.



The decision tree shows that low satisfaction level is a good predictor of a high performer leaving the company. We can see that those high performers with satisfaction levels below 0.11 and last evaluation less than 0.81 were highly likely to leave. High performers that have time spent less than 4.5 years in the company are unlikely to be at risk of leaving. For some high performers who already had high satisfaction levels (>0.71), risk of leaving the company was related to having worked a high average number of monthly hours (>216), and a large number of projects (>3.5).

RFM

Using the full dataset, this analysis will provide the factors most influencing why people leave this company (all performance levels are analyzed)

Model

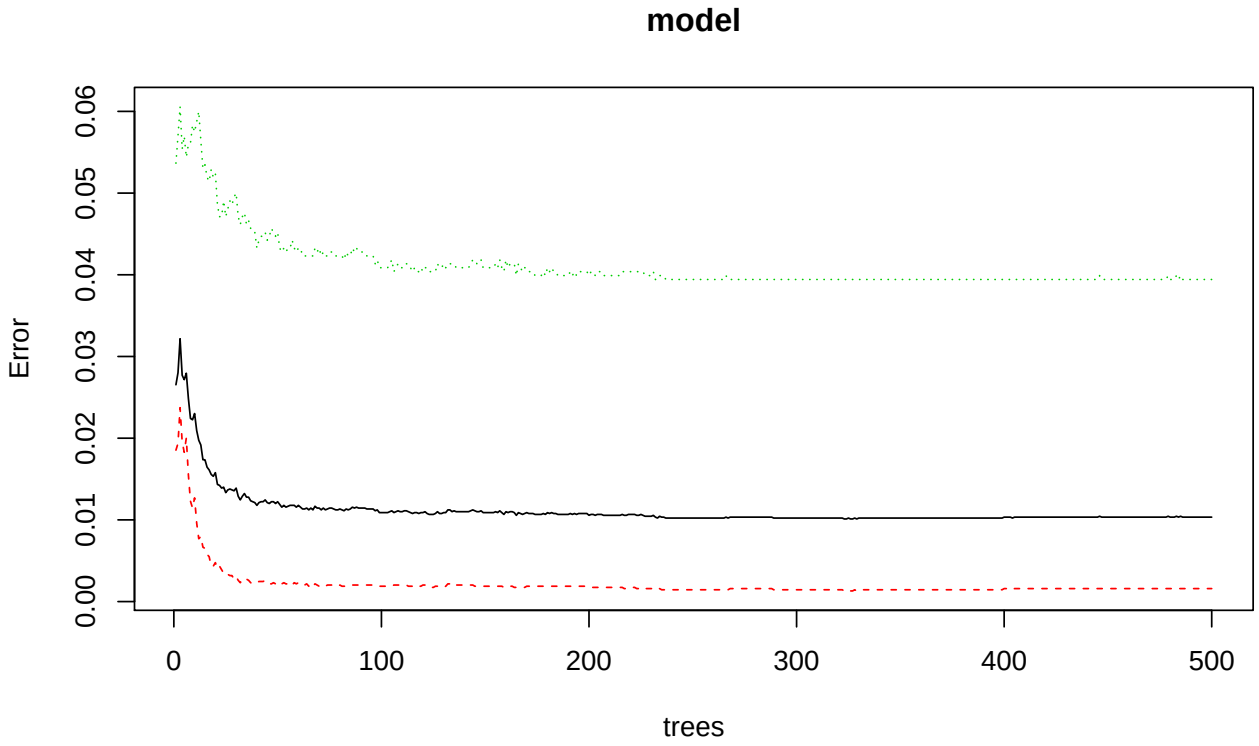
```
## [1] 14999
```

```
##
## Call:
##  randomForest(formula = left ~ ., data = training)
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 3
##
##         OOB estimate of  error rate: 1.03%
## Confusion matrix:
##          Not left Left class.error
## Not left     6908   11 0.001589825
## Left           82 1998 0.039423077
```

```
tbl<-importance(model);
tbl
```

```
##                         MeanDecreaseGini
## satisfaction_level           1059.765872
## last_evaluation               389.784209
## number_project                618.734550
## average_montly_hours          455.308427
## time_spend_company            564.173993
## Work_accident                  14.515630
## promotion_last_5years           2.143771
## sales                          49.534503
## salary                         25.233775
```

Satisfaction level is the strongest predictor of who will leave the company. This is corroborated by a decision tree (not pictured) showing one of the most important criteria to be classified as leaving the company is satisfaction levels below 0.46. In addition, those who had more than 2.5 projects were more likely to leave the company. Lastly, time spent in the company less than 4.5 was a strong predictor of not leaving.



The plot of the model above shows that the large number of trees used in the RFM helped to lower the prediction error rate for "left" the company (green), out-of-bag (OOB) error (black), and "not left" the company (red). The most important variables output by the RFM have provided a description of which factors influence classification as "left" the most. Clearly satisfaction level is important.

Predictions

```
# Confusion Matrix
pred<-predict(model, testing, type="response")
conf.mat<-print(table(testing$left,pred),1)
```

```
##           pred
##            Not left Left
```

```
##   Not left     3591   12
##   Left           48  485
```

From the confusion matrix above, we can conclude that this RFM has high accuracy in predicting who is going to leave the company.

Conclusion

For all employees, including high performers, a low satisfaction level is a strong predictor of quitting, as is having spent less time in the company. High performers are also noticeably influenced by having high average monthly hours and a high number of projects in leaving the company.

- The satisfaction level is the major parameter to determine if an employee stay with the company.
- Salary is a big influence factor in determing employee satisfaction level and whether they left the company.
- Time is a significant impact. Employee want to stay with company, not doing excessive high pressure overtimes. Employees with 3-4 projects assigned tend to stay. It shows that they are providing value to the company but by not doing too many project to ruin their work-life balance.

why do good people leave? For the employees that the company may want to retain, there are a few red flags that they might leave:

- They are not satisfied with their job, which is affected by a few things like:
  - They work too much or too little.
  - They are not working on diversified projects (less than 3 projects)
- They have been with company about 3 or 4 years, and is looking for a change for either personal reason or they just get tired of the company.
- There seems to be a relationship between an employee's performance and their happiness. A constant review and updating for the process of evaluation will likely benefit the company in retaining its employees.

So at the end, from all the analysis we have above, it may seem straightforward for the company to predict who is leaving. if the company can pay more salary to the high performers, have them do some interesting projects (not too many), while let them keep a good work-life balance, those high performers will mostlikely stay and grow with company.However, since our data is simulated by nature, in the real world, we still need to aquire more information and have diversified methods to keep good employees happy.