Toolbox: Human Resource Data Analysis

Xiaoxu Zhang

April 30, 2017

1. Introduction

2. Problem Statement

2.1 Problem

Employee attrition is one of the biggest challenges that the company has to face. There are many different reasons and possible factors for employees leaving. Retaining valued employees is the final purpose and needs targeted strategies. But are there reliables ways to figure out if and why the best and most experienced employees are leaving prematurely? Human resourece department is looking forward to analysis based on data dealing with this problem. Several steps could be accomplished: 1. The existing situation of employees leaving in the current company. 2. The possible reason why employees leave. 3. Predicting who will be the next to leave.

2.2 Data Overview

In order to solve the above problem, a related data set is necessary. Here is a data set found on Kaggle (www.kaggle.com/ludobenistant/hr-analytics). After reading in the data set, a quick overview is represeted as following.

```
# Reading the csv data set called "HR_comma_sep" and looking at the overall structure of data.
hr_data<-read.csv("HR_comma_sep.csv",header=T,sep=",")
str(hr_data)
```

```
## '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 ...
##  $ department           : Factor w/ 10 levels "accounting","hr",..: 8 8 8 8 8 8 8 8 8 8 ...
##  $ salary               : Factor w/ 3 levels "high","low","medium": 2 3 3 2 2 2 2 2 2 2 ...
```

In order to predict which employee will leave next, understanding variables in detailed comes to the first. There are 10 variables in this data set, as well as 14999 rows. Each row represents one sepcific employee in the company. Following is a table of variable name and its corresponding description.

The data set does highly relate to the problem to be solved, as it includes one variable of whether the employee has left, and various variables which can help to figure out the possible factors could cause the leaving.

Variable name	Description
satisfaction_level	How the employee statisfies the company. Highest being 1 and lowest is 0.09.
last_evaluation	How the company evaluates the employee. It is the last evaluation.
number_project	There are employees who are assigned up to 7 projects and as least as 2 projects.
average_montly_hours	On an monthly average, how many hours the employee spend in office.
time_spend_company	The company has employees whose stay varied from 2 to 2 years.
Work_accident	Whether the employee has a work accident.
left	Whether the employee has left. Totally 3571 (out of 14999) employees left.
promotion_last_5years	Only 319 (out of 14999) employees are promoted in the last 5 years.
department	There are totally 10 departments in the company.
salary	Classified into high/medium/low salary level.

3. Data Exploration

3.1 Data Processing

Missing Value

At the beginging of exploring the data set, it is necessary to check whether missing values or other invalid values exist. If so, it comes to complete missing values with proper strategies and methods. If not, continuing following analysis. The number of missing value in the data set is actually 0. Now, a completed data set is ready for following analysis.

Correlation

Calculating the correlations between all different combinations of data allows us to get first hints on why people leave. However, correlation requires that the type of variable is numeric so that changing the class of variables from factor to numeric.

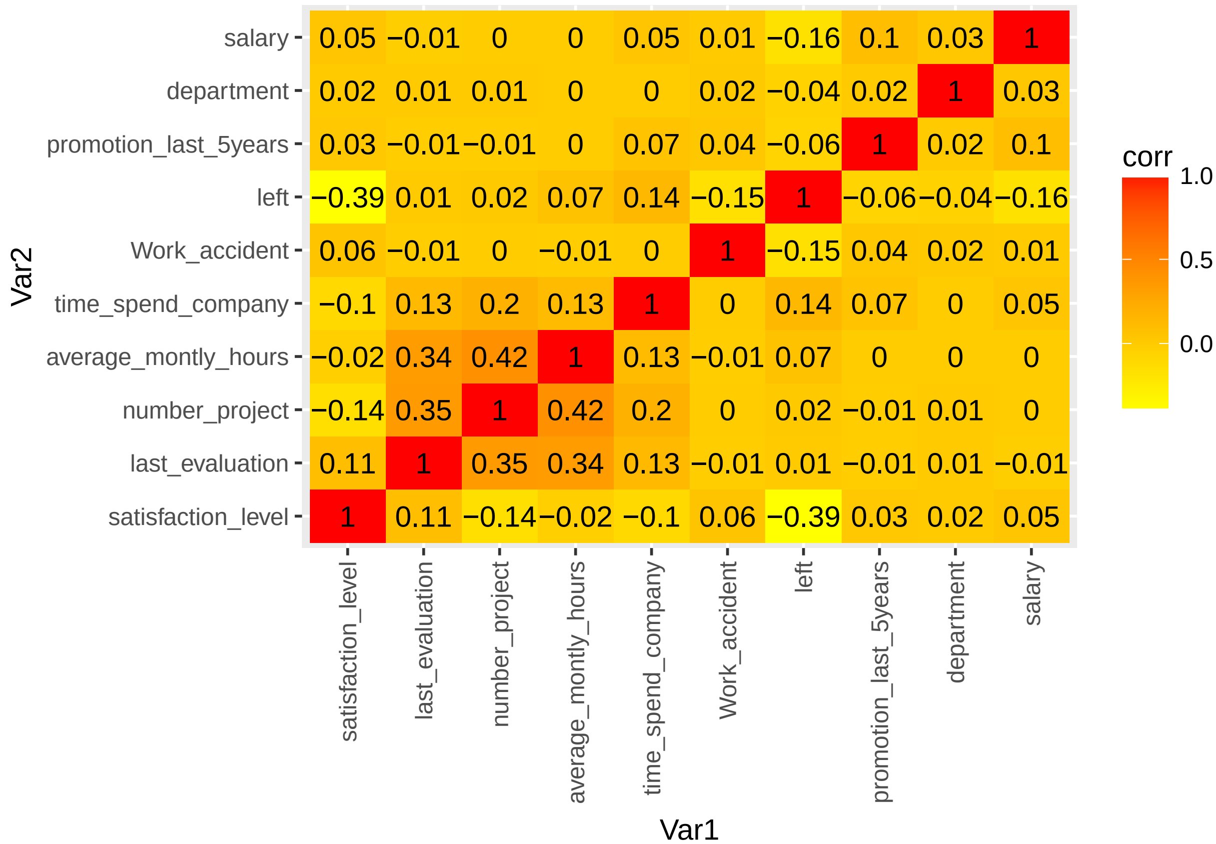
```
# Taking a look at the class of all variables
sapply(hr_data,class)
```

```
##    satisfaction_level       last_evaluation        number_project
##             "numeric"             "numeric"             "integer"
##  average_montly_hours    time_spend_company         Work_accident
##             "integer"             "integer"             "integer"
##                  left promotion_last_5years            department
##             "integer"             "integer"              "factor"
##                salary
##              "factor"
```

Here we can see that "department" and "salary" are factor.

```
# Changing to numeric type
hrdata<-hr_data
hrdata$department<-as.numeric(1:10)[match(hrdata$department,unique(hrdata$department))]
hrdata$salary<-as.numeric(1:3)[match(hrdata$salary,c("low","medium","high"))]
# Caculating correlation between each pair of varialbes
corr<-melt(cor(hrdata))
names(corr)<-c("Var1","Var2","corr")
```

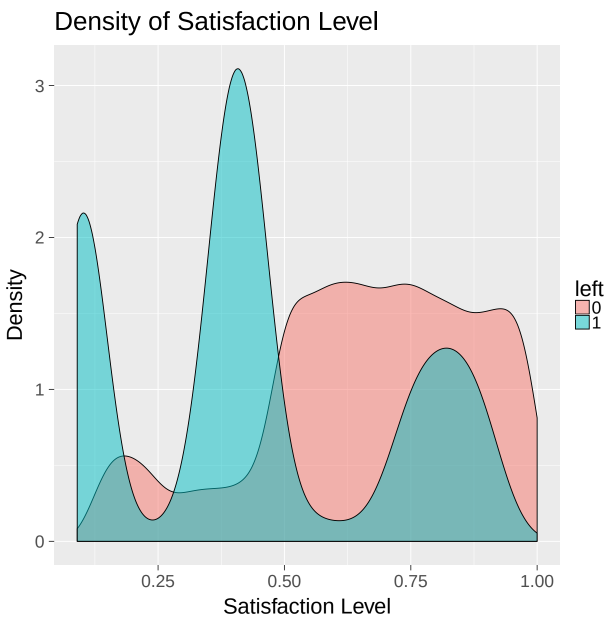
```r
# Making correlation visualized
ggplot(corr, aes(Var1, Var2, fill = corr)) + geom_tile() +
    scale_fill_gradient(low = "yellow",  high = "red") +
    geom_text(aes(label = round(corr, 2))) +
        theme(axis.text.x=element_text(angle=90,hjust=1,vjust=0.5))
```

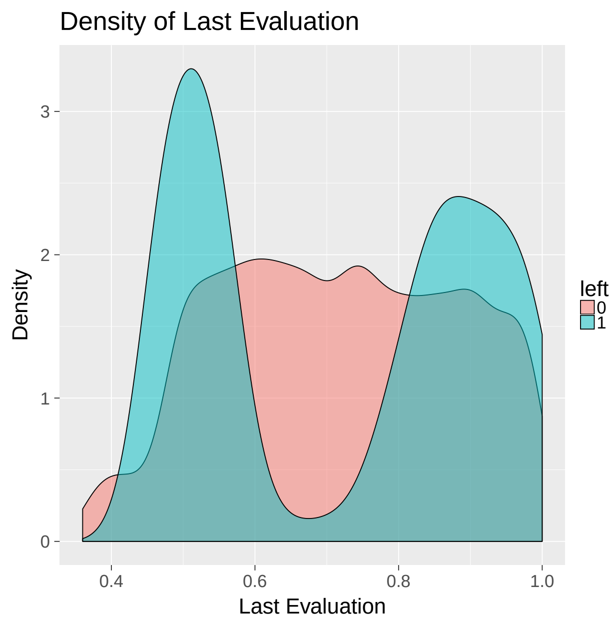


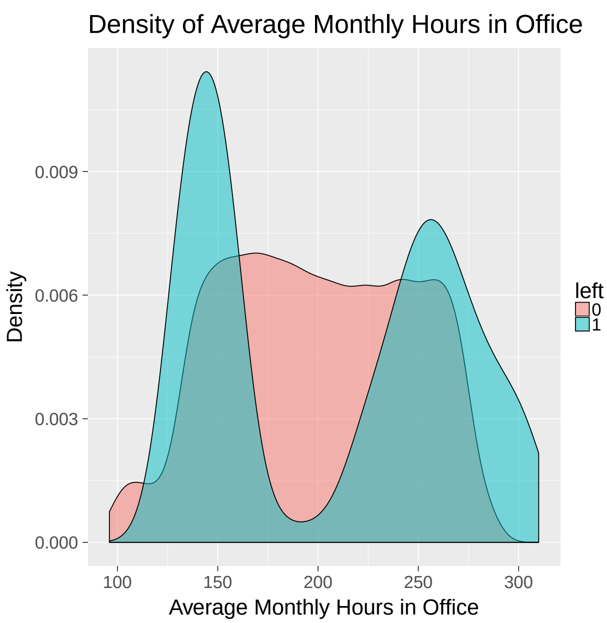
As we can see from the above graph, the top four factors that are relatively high corelated with "left" are "satisfication level", "salary", "work accident", and "time spend at company". To be specific, the most correlated factor is the level how employees statify the company, and the higher satisfaction level, the less possbility to leave.

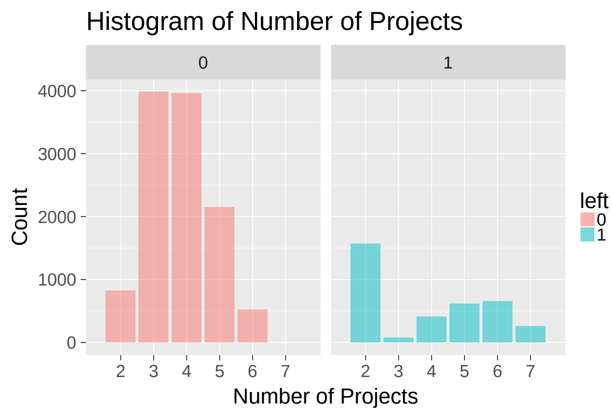
3.2 Data Analysis

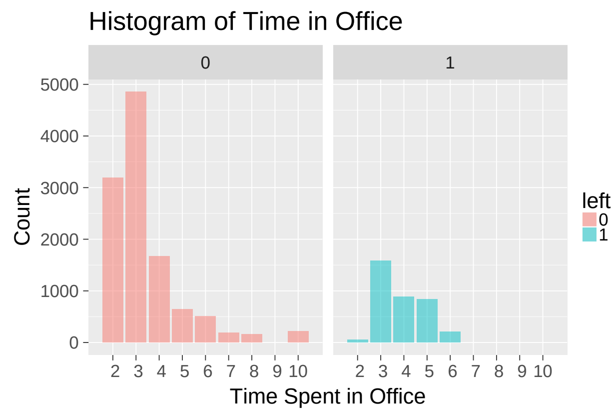
Although people who left the company have their own reasons as an individual, comparing those who left and did not leave would give more perspectives. So dividing the entire into two groups and comparing them of each variable. There are three continuous variables: "satisfaction level","last evaluation","average monthly hours". Meanwhile, others could be treated as categorical variables. Different plots would be selected to figure out their trend for above two types of variables, repectively, density and histogram. Eventually each plot will use different color to represent whether employees left or not.

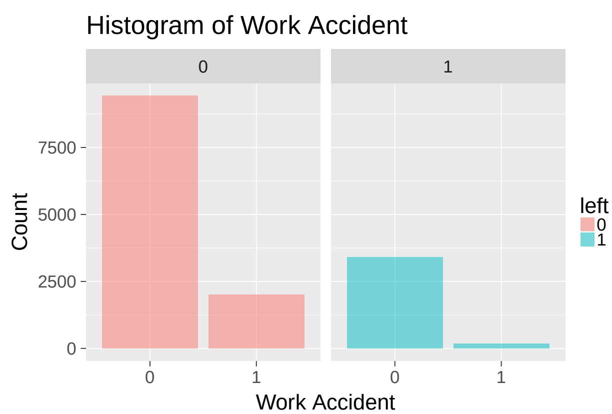

Density of Satisfaction Level
Density
3
2
1
0
0.25
0.50
0.75
1.00
Satisfaction Level
left
0
1

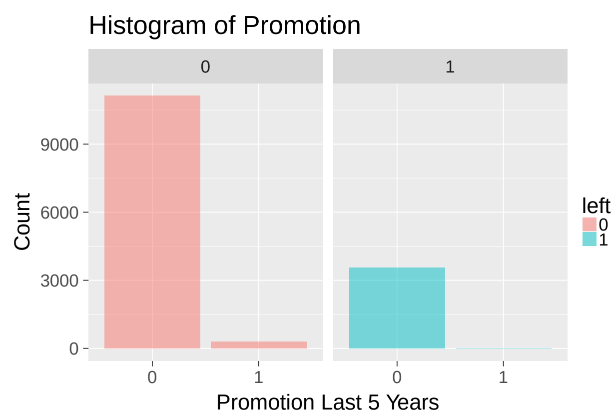

Density of Last Evaluation
Density
3
2
1
0
0.4
0.6
0.8
1.0
Last Evaluation
left
0
1

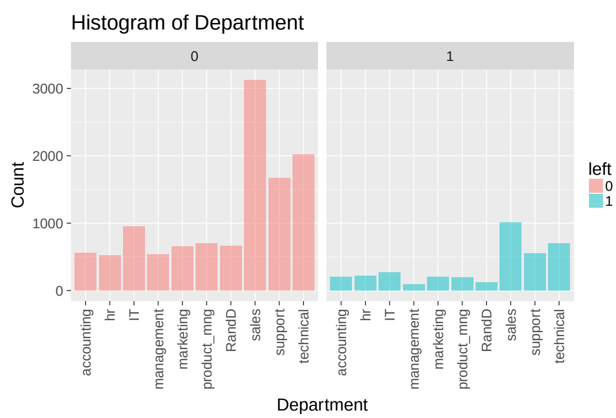

Density of Average Monthly Hours in Office
Density
0.009
0.006
0.003
0.000
100
150
200
250
300
Average Monthly Hours in Office
left
0
1

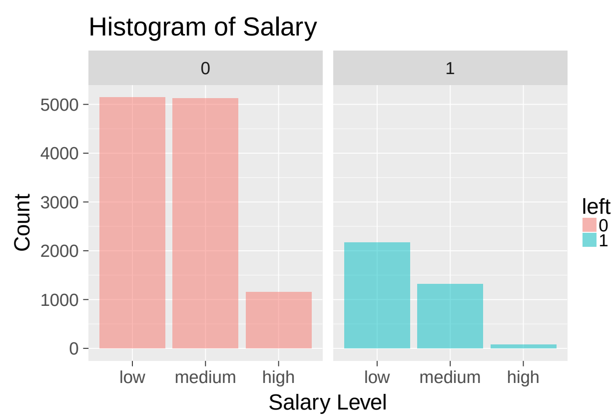












Obersavertion

- **Satisfaction Level and Last Evaluation**

The mean satisfaction level of those who have left is apparently lower than those who did not. As for people who left, there are three peaks of satisfaction level instead of two. Is there any possible classification of these group? Going back to the correlation graph, other two variables are highly correlated with satisfaction level. They are “last evaluation” and “number of projects”. Now, turning to analyze the relationship between them.

As shown above, employees who left are gathering in three parts so that also can be divided into three subgroups:

1. Best Match: people who possess both high satisfaction and high evaluation. They are content with the company, and the company is also content with them. They seem to be the best match with the company, but they decide to leave. The reason behind this group might be more individual rather than caused by the company.

2. Over Qualified: people who possess low satisfaction but high evaluation. They are too excellent to be content by the company. They decide to leave probably because they are pursuing better platform instead of standing at the same point.

3. Worst Match: people who possess both low satisfaction and low evaluation. They are the opposite side of the best match. Their bad performance might also leads to their leaving, In other words, they might be fired by the company.

- **Average Monthly Hours and Number of Projects**

For employees who left, there are two peaks in the density of average monthly hours. That means they are much more probably to leave if they spend too much or too little time in office. Meanwhile, employees who left have either too many or too few projects. Actually, to some degree, the number of projects you are assigned lead to the amount of time you spend in office.

- **Work Accident and Promotion**

Comparing with employees who left or not, similar results are shown on these two variables. In terms of the percentage of having work accident, employees who left are lower than those who did not, as the same as the percentage of being promoted.