**# Project Theme: Human Resources Analytics**

A data mining project about why employees leave their work.

We used a dataset from kaggle.com with the website: https://www.kaggle.com/ludobenistant/hr-analytics  
And our result in the github: https://github.com/DataMining2017NJU/HR\_Analytics/tree/develop

Team Information:

|  |  |  |  |
| --- | --- | --- | --- |
| Name | Role | Student\_ID | Work |
| 李雪菲 | Leader | MF1372074 | Grasp the progress of project,preprocessing,and decision tree |
| 孔敏 | Member | MF1732063 | project,preprocessing,and decision tree |
| 安磊 | Member | MF1732001 |  |
| 户胜浩 | Member | MF1732004 |  |

1. **Introduction:**  
   **Task:**

We want to **clustering** dataset to see the class of the data and use **decision tree** to conclude and predict why people leave their job and what kind of people who is possible to leave.  
**About Dataset:**   
This dataset is simulated. In the dataset, fields, explaination and type are listed below:

Table 1 fields in the dataset

|  |  |  |
| --- | --- | --- |
| satisfaction\_level | Level of satisfaction (0-1) | Numeric |
| last\_evaluation | Time since last performance evaluation (in Years) | Numeric |
| number\_project | Number of projects completed while at work | Numeric |
| average\_montly\_hours | Average monthly hours at workplace | Numeric |
| time\_spend\_company | Number of years spent in the company | Numeric |
| Work\_accident | Whether the employee left the workplace or not (1 or 0) Factor | Numeric |
| left | Whether the employee was promoted in the last five years | Numeric |
| promotion\_last\_5years | Whether the employee was promoted in the last five years | Numeric |
| sales | Department in which they work for | String |
| salary | Relative level of salary (high) | String |

Data exploration and preprocessing.

Typically, cleaning the data requires a lot of work and can be a very tedious procedure. This dataset from Kaggle is super clean and contains no missing values. But still, I will have to examine the dataset to make sure that everything else is readable and that the observation values match the feature names appropriately.

· Does the dataset need preprocessing?

· What methods do you use for preprocessing. § Data Mining:

· What algorithms you use for your task?

· What evaluation methods, and metrics to evaluate

We use weka to draw a histograph of each attributes. We found attribute number\_project has a shape that similar to normal distribution.



We also use weka to calculate the statistics data of dataset, and get the result below:

Table 2 statistics of fields

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Min | Max | Std | Mean | Mode | Median |
| satisfaction\_level | 0.09 | 1 | 0.249 | 0.613 | 0.1 | 0.64 |
| last\_evaluation | 0.36 | 1 | 0.171 | 0.716 | 0.55 | 0.72 |
| number\_project | 2 | 7 | 1.233 | 3.803 | 4 | 4 |
| average\_montly\_hours | 96 | 310 | 49.943 | 201.037 | 0 135/ 1 156 | 200 |
| time\_spend\_company | 2 | 10 | 1.46 | 3.498 | 3 | 3 |
| Work\_accident | 0 | 1 | 0.352 | 0.145 | 0 | 0 |
| left | 0 | 1 | 0.426 | 0.238 | 0 | 0 |
| promotion\_last\_5years | 0 | 1 | 0.144 | 0.021 | 0 | 0 |

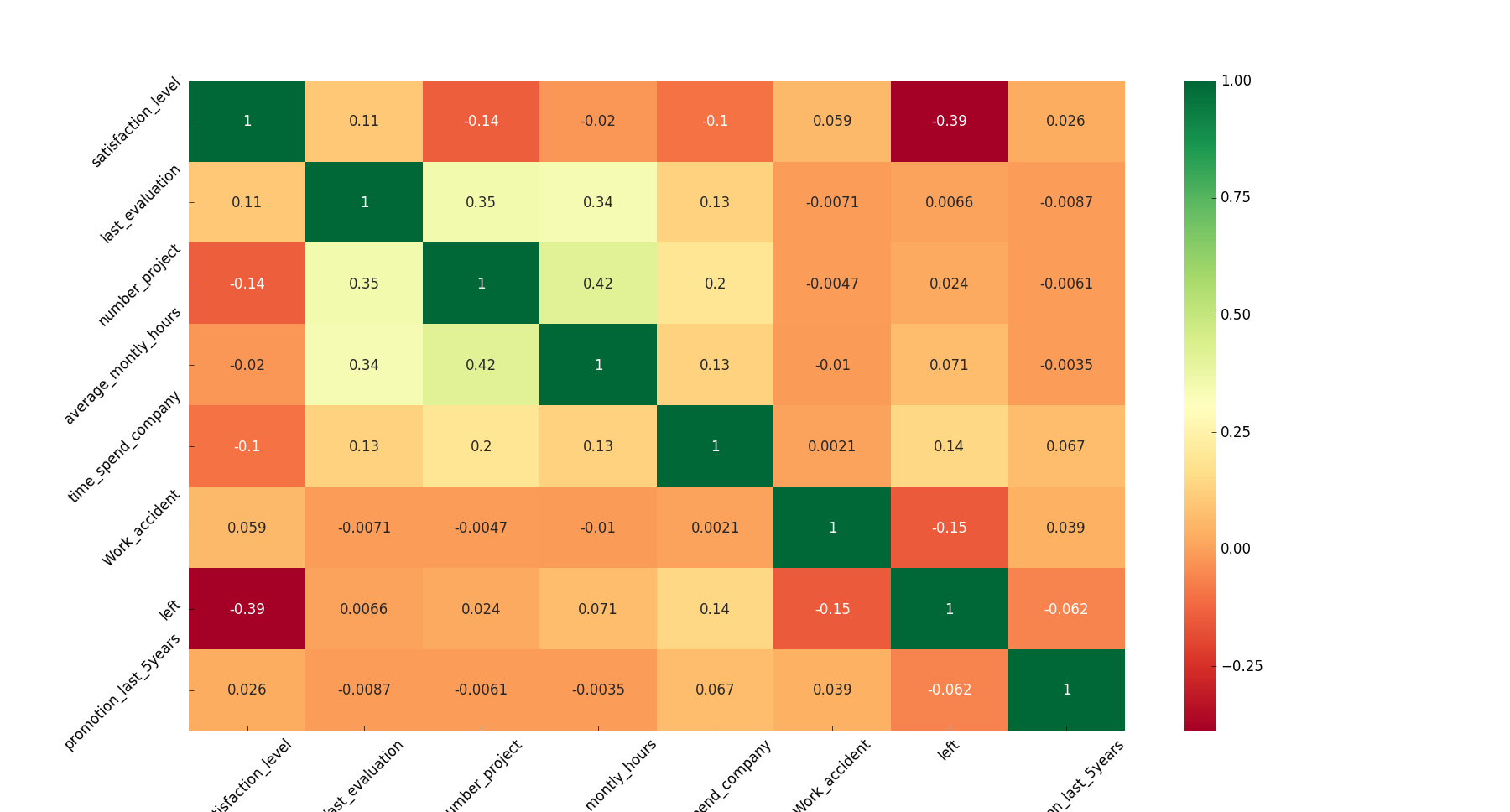
From table 2, we found that the mean/mode/median of number\_project are familiar with each other.So it`s possible number\_project fit the normal distribution.

Table 3 sales count

|  |  |  |
| --- | --- | --- |
| sales | label | Count |
| 1 | sales | 4140 |
| 2 | accounting | 767 |
| 3 | hr | 739 |
| 4 | technical | 2720 |
| 5 | support | 2229 |
| 6 | management | 630 |
| 7 | IT | 1227 |
| 8 | product\_mng | 902 |
| 9 | marketing | 858 |

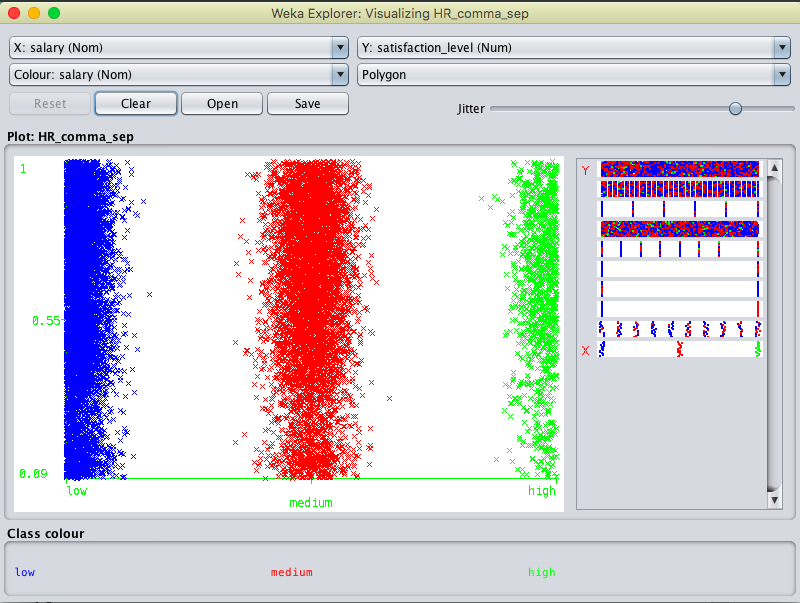
|  |  |  |
| --- | --- | --- |
| left | Stay | Left\_Rate |
| 3517 | 11429 | 23.808% |

Figure 1 Correlation of attributes

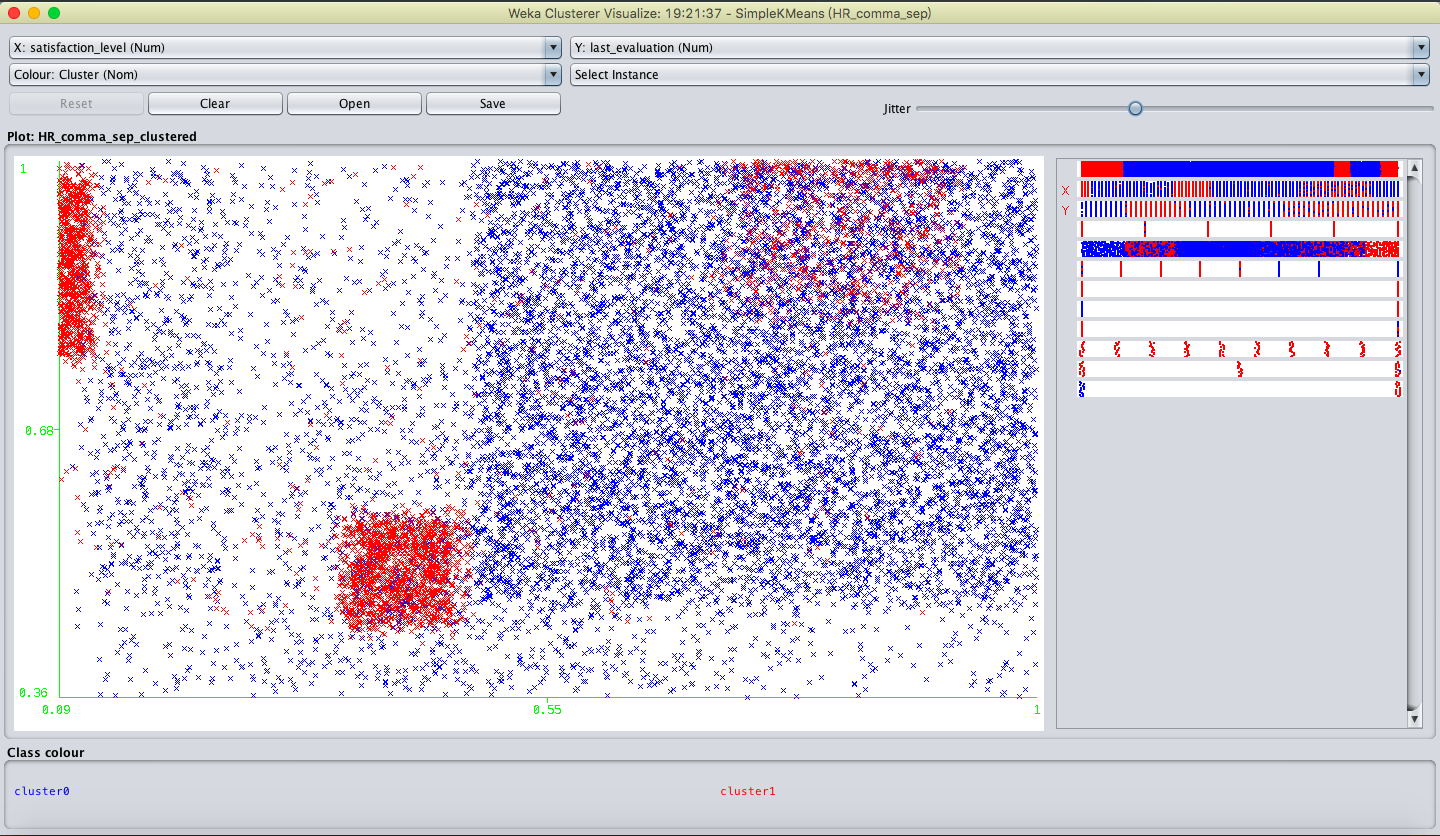


In Figure1,

Figure 2



In figure2 ,which have x\_axle: salary, and y\_axle: satisfaction\_level. In our opinion , the most possible reason maybe the salary.But from this graph we`ll know, there are many people who earned less but still have high satisfaction, and earned a lot but has low satisfiction.



the results?

· Report the results

· Write discussion about the results

§ Conclusion.