Big Data Project Report

Submitted By Team 5:

Members:	**SEC**	**B.N**
Ramy said	1	21
Shehab Alaa	1	31
Ammar AL-Sayed	2	5
Loai Ali	2	11

Submitted To: Eng. Hussien Fadl

1. Table of contents

2. Brief problem description

A large company named XYZ, employs, at any given point of time, around 4000 employees. However, every year, around 15% of its employees leave the company and need to be replaced with the talent pool available in the job market. The management believes that this level of attrition (employees leaving, either on their own or because they got fired) is bad for the company, because of the following reasons

1. The former employees' projects get delayed, which makes it difficult to meet timelines, resulting in a reputation loss among consumers and partners
2. A sizeable department has to be maintained, for the purposes of recruiting new talent
3. More often than not, the new employees have to be trained for the job and/or given time to acclimatize themselves to the company

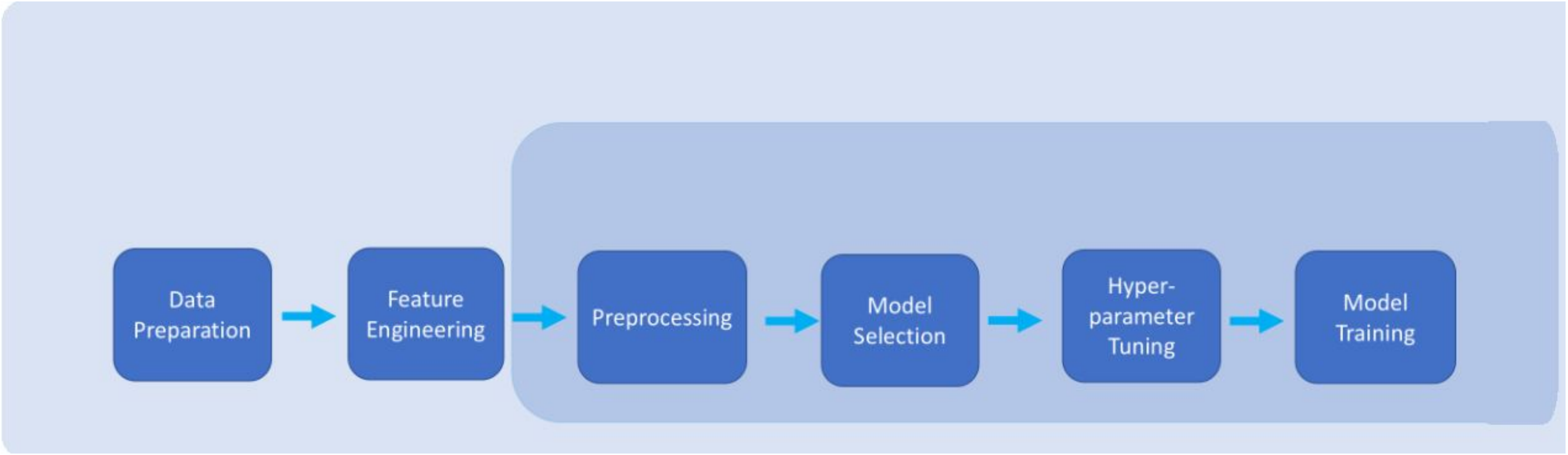
Hence, the management has contracted an HR analytics firm to understand what factors they should focus on, in order to curb attrition. In other words, they want to know what changes they should make to their workplace, in order to get most of their employees to stay. Also, they want to know which of these variables is most important and needs to be addressed right away.

The goal of the case study

We are required to understand why this high level of attrition happens. The results thus obtained will be used by the management to understand what changes they should make to their workplace, in order to get most of their employees to stay.

Also, build a model to predict the probability of attrition of a given employee.

3. Project pipeline



4. Dataset description

Data set contain info of 4410 employees and this info can be divided into 5 categories:

1. General: general data about the employee as age, education, attrition, gender, marital status,etc..
2. Manager Survey: employees feedback survey about their managers (job involvement, rating)
3. Employee Survey: employees feedback about the work in the company (environment satisfaction, job satisfaction, work-life balance)
4. Log in time: login times of employees from 1/1/2015 to 19/5/2019.
5. Log out time: logout times of employees from 1/1/2015 to 19/5/2019.

<u>Note</u>: a complete description of the data set can be found in the appendix.
And data can be found in HR Analytics Case Study on Kaggle

5. Analysis and solution of the problem

A. Data preprocessing.

a. Data was collected in different CSV files so we have to group the data by employee id

b. Remove rows that are missing some values (removed data of 80 employees so it wasn't a big deal).

c. Creating new columns

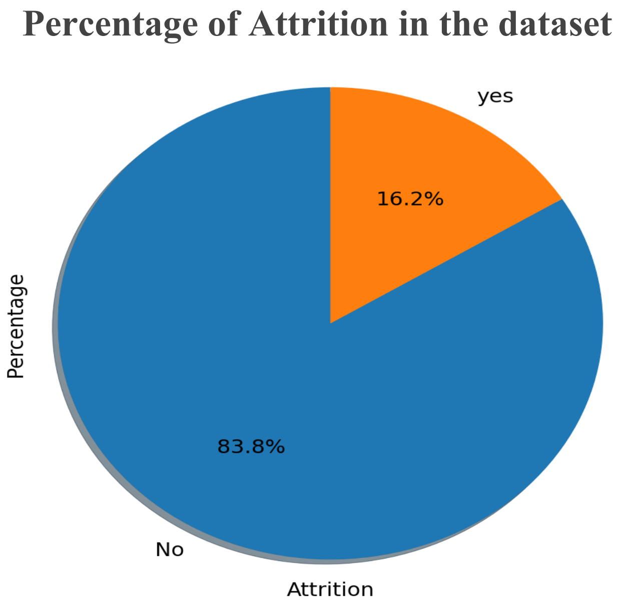
- Mean working hours: from the login and logout times we have created a new variable called mean working time of each employee

d. Remove useless columns

- Employee id (this set as index)
- Over 18 (binary column): all employees in the data set are above 18 so the variable is not needed.
- Standard hours: because all employees in the data set have standard hours of working of 8 hours in the contract.

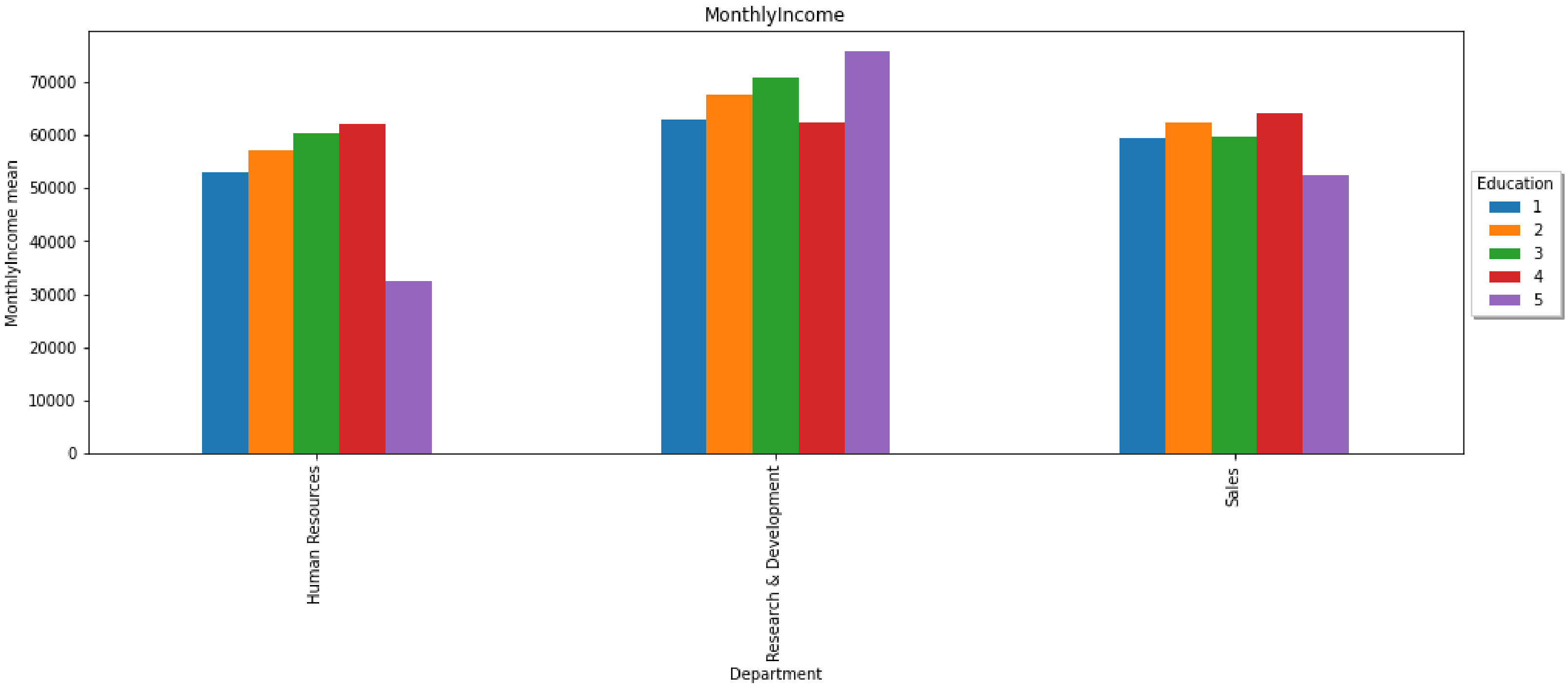
B. Data visualization & data insights

Note: there will be simple comments on each graph and business insights and conclusions will be in the conclusions section.



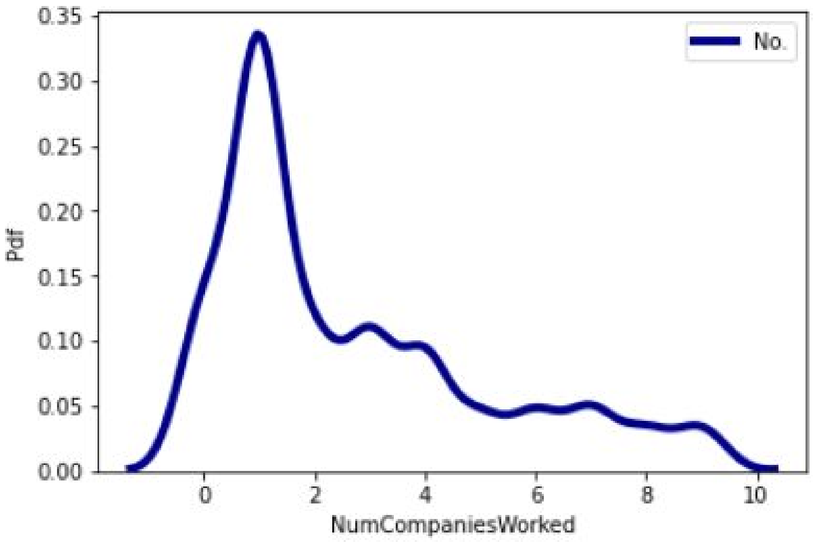
Comment: Data set is **imbalanced** so will not depend on accuracy but we will report better performance indications like FI score and ROC curve

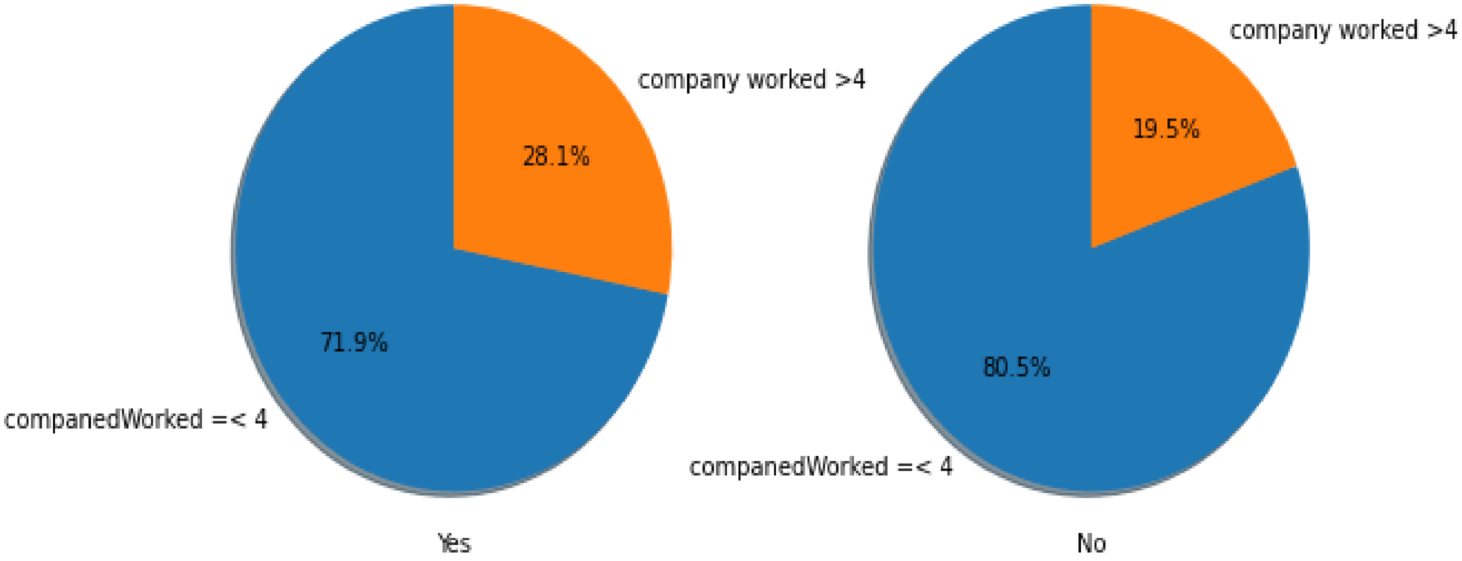
Group employees by department and education level against monthly income mean



Comment: Human Resources has the lowest income

Distribution of num of companies each employee has worked for

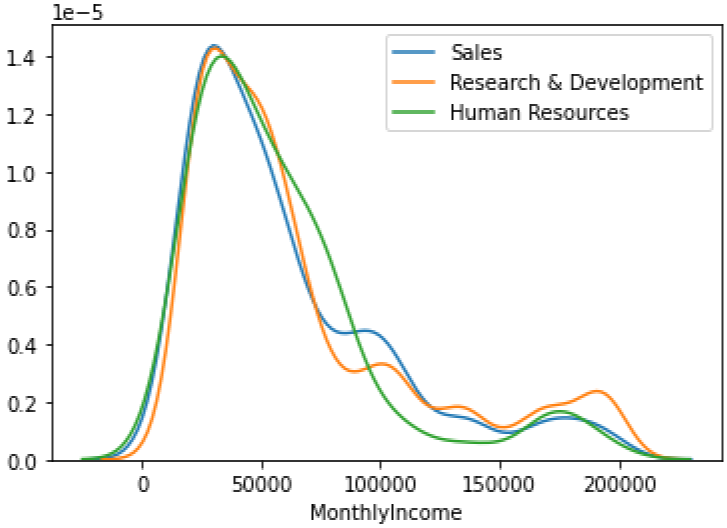




Comment: Most of the company employees have worked on <= 4 companies before and comparing ratios of people worked at more companies than 4 in case of stayed in the company or left we can see that ratio of people worked in more companies is bigger in case of attrition than staying at the company

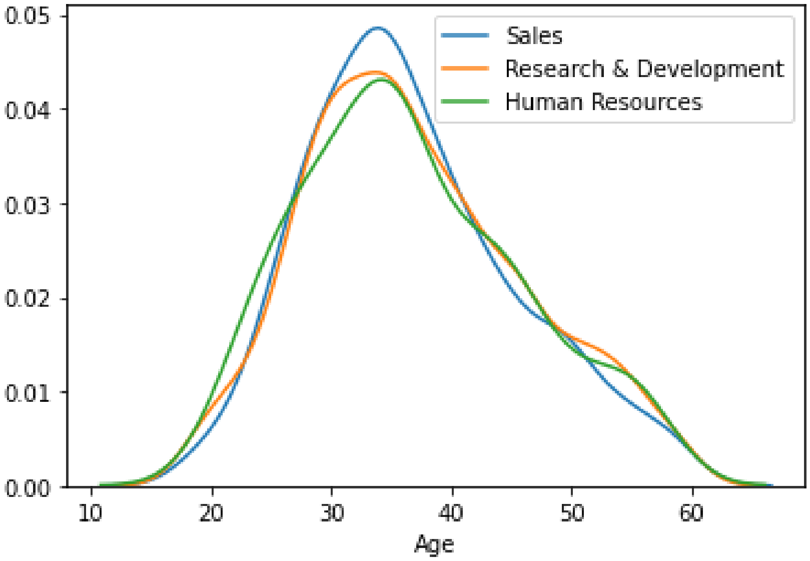
Explore monthly income

Distribution with department



Comment: The monthly income is almost the same pattern for all departments except

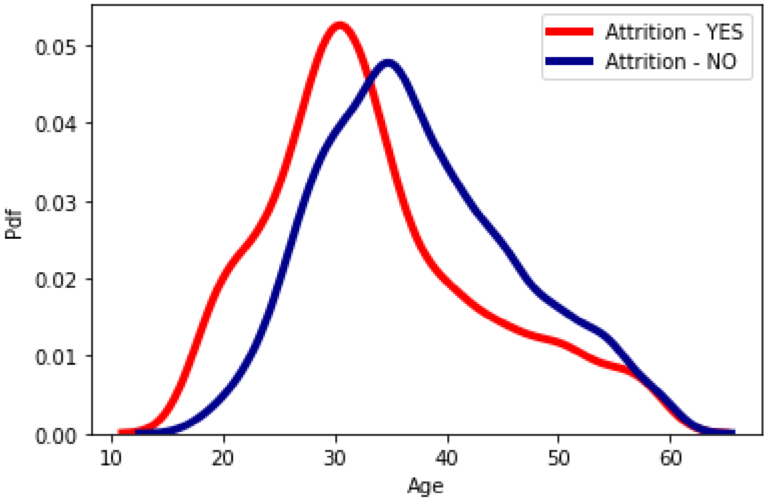
Distribution with Age



Comment: The monthly income is almost the same pattern over all ages in different departments

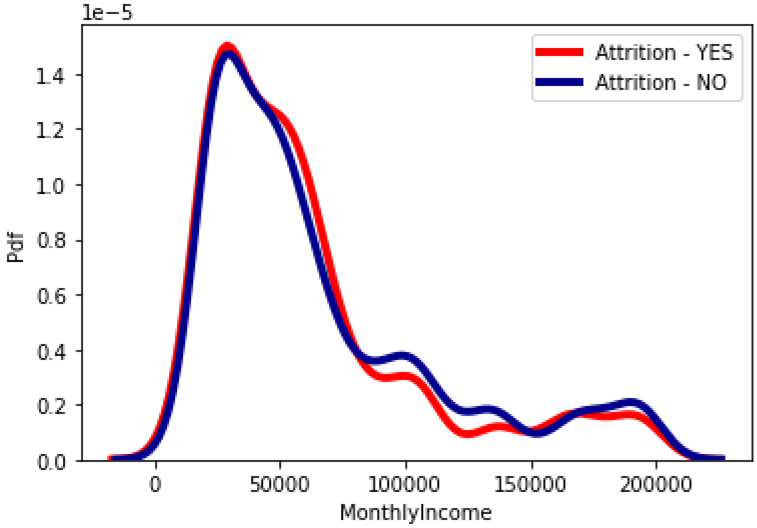
Visualization of the relation between attrition and other factors.

Age



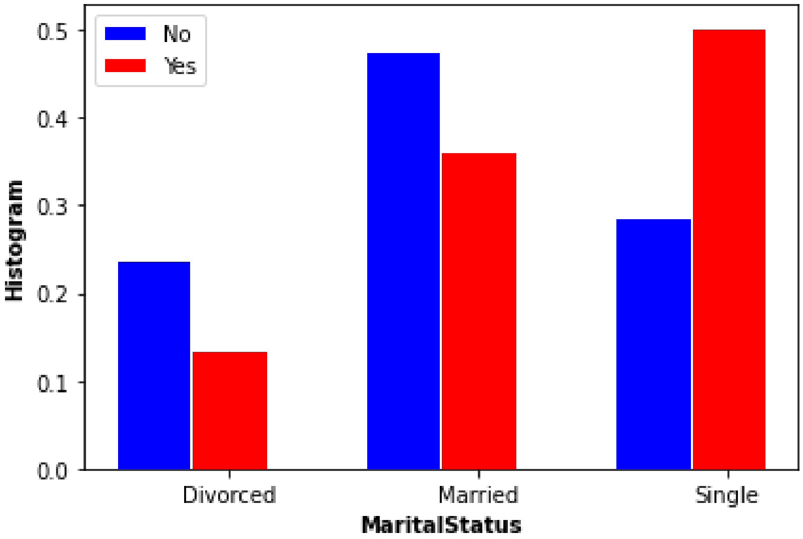
Comment: Mean age of leaved employees is smaller than the mean of stayed

Monthly Income distribution for leaved and stayed employees



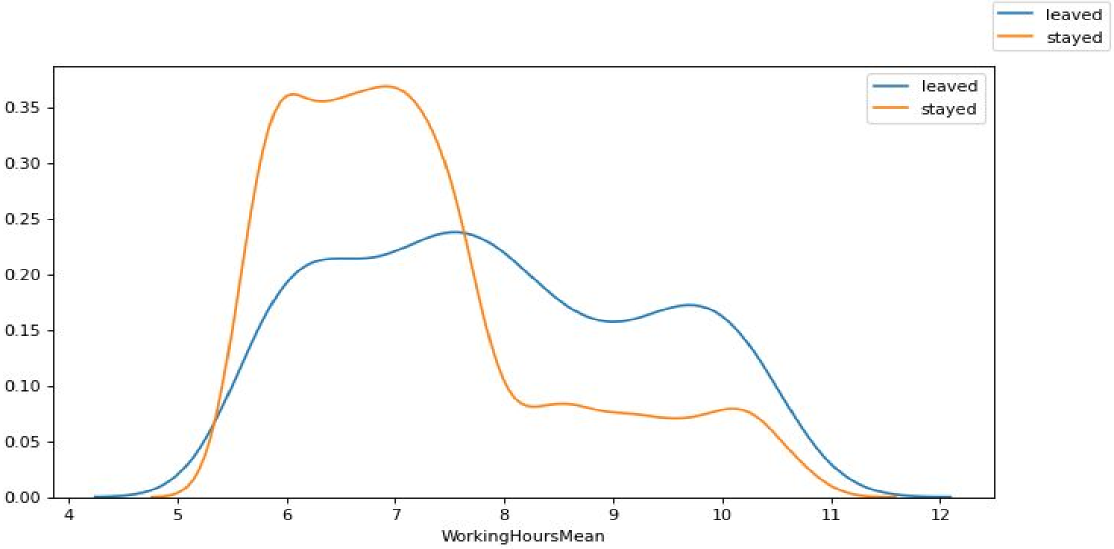
Comment: both left and stayed employees have almost the same distribution

Marital status

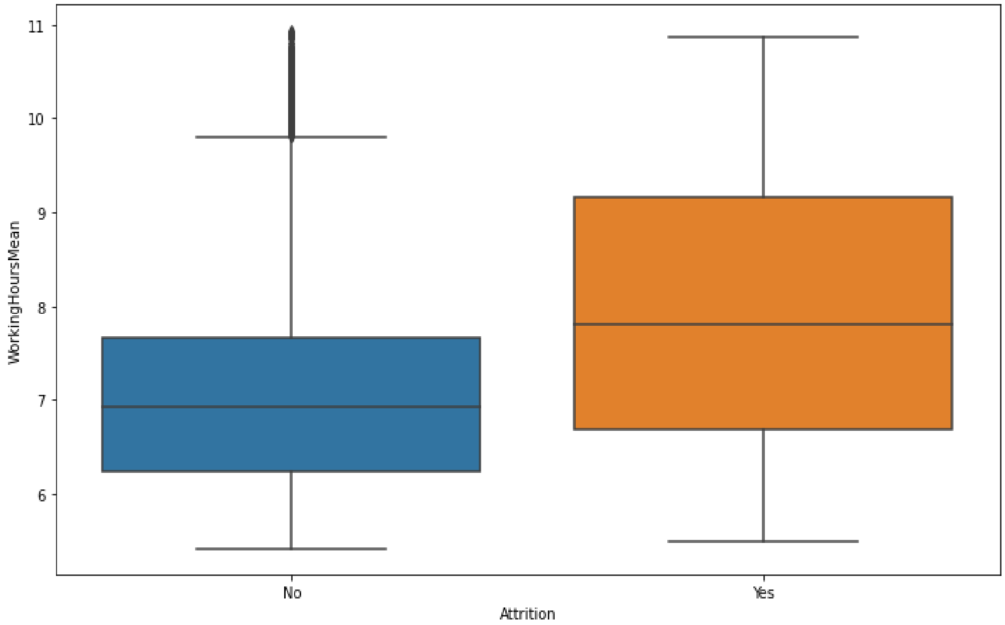


Comment: Most of the left employees are single

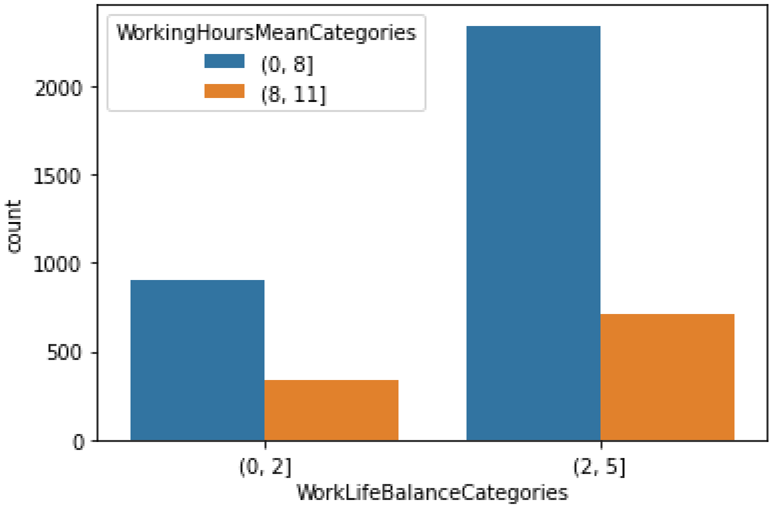
Working hours



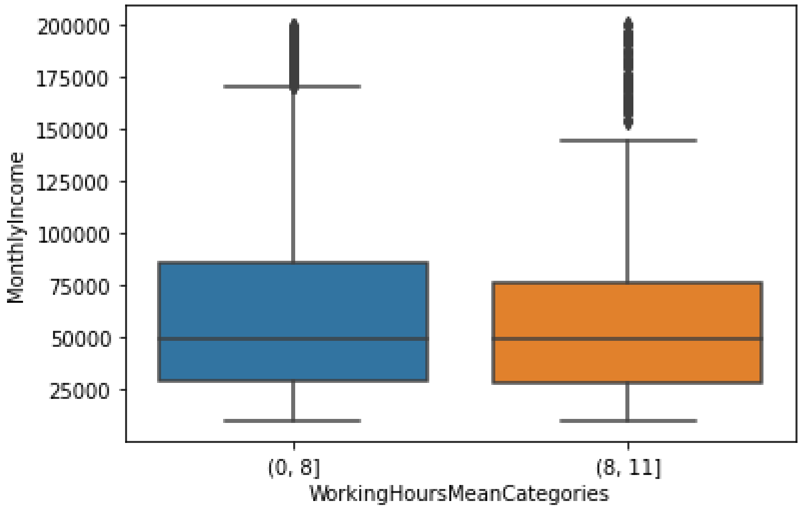
Comment: The mean of stayed employees is shifted towards smaller working hours mean and they have smaller deviation.



Comment: this is emphasized using a box plot which is more clear

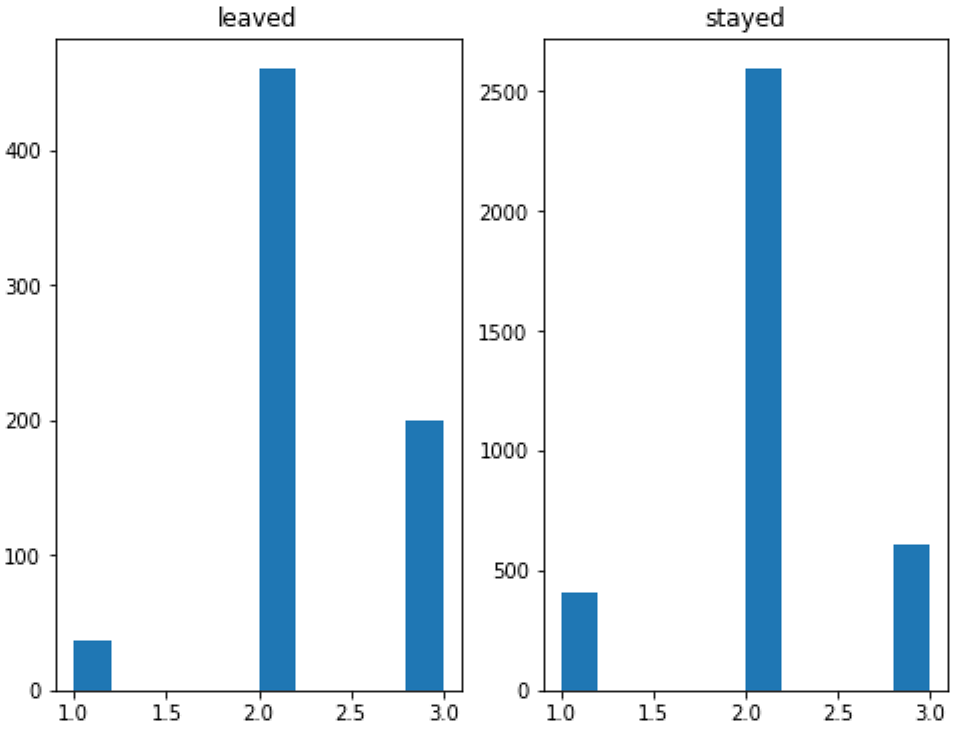


Comment: Ration between employees that has [0 -> 8] hours work to employees works for more hours affects there survey about work-life balance



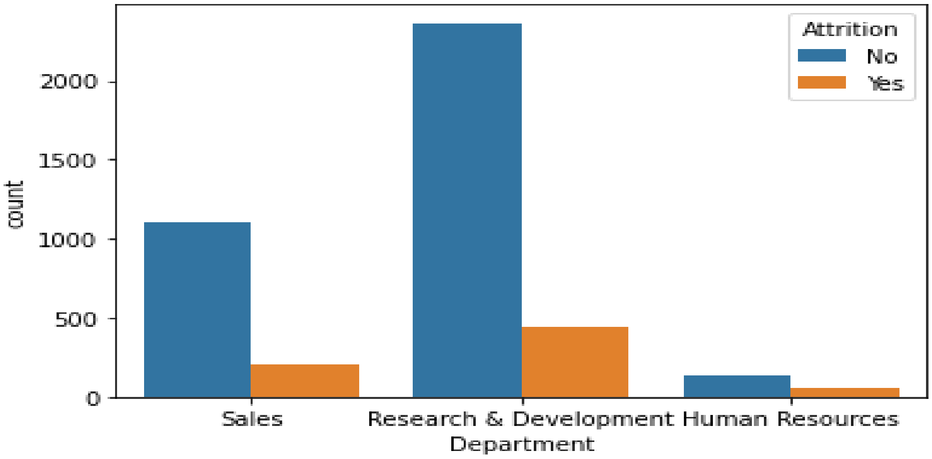
Comment: Employees who work more have almost equal monthly income mean and like 25 of them take between [50,000 -> 75,000] (and much outliers) on the other hand people who work less has a higher range of salary (and this is weird that company doesn't give them a bonus for working more hours so why they work for more hours ?!!)

Business Travel

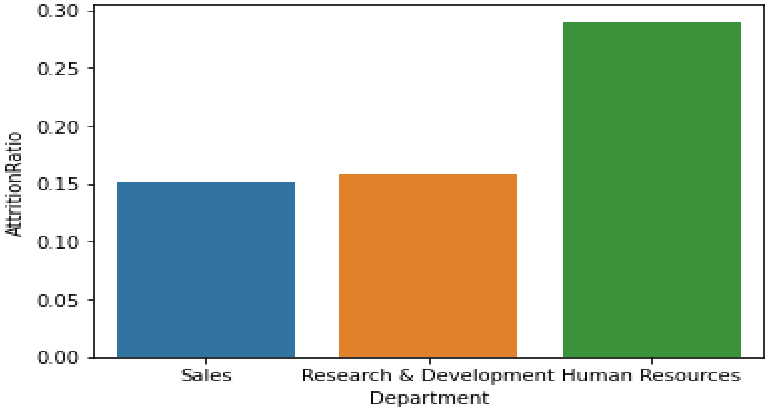


Comment: We can see that almost left and stayed employees had the same business travel opportunities

Department



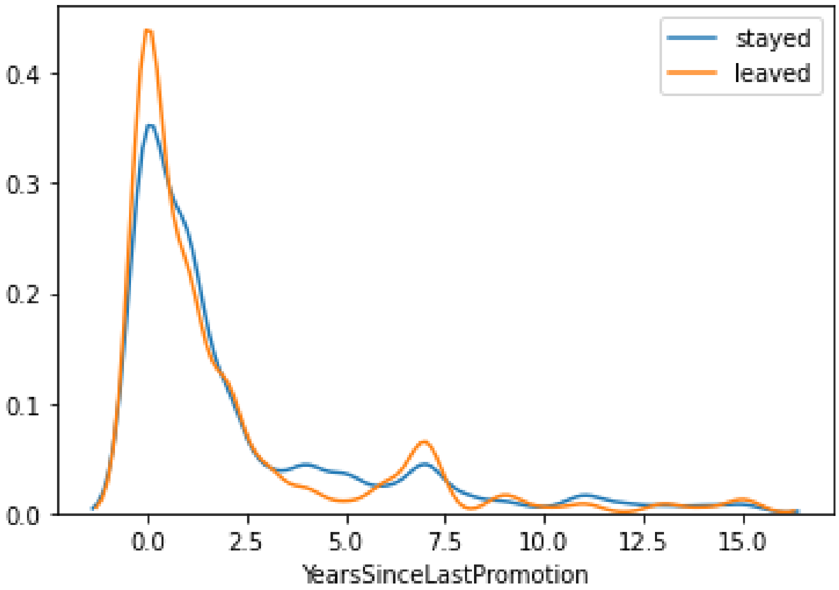
Comment: Seems most people left were from the R&D department. but this not useful as this department has the largest number of employees and also there are a lot stayed so let us examine something more useful like the ratio between left and total at each department



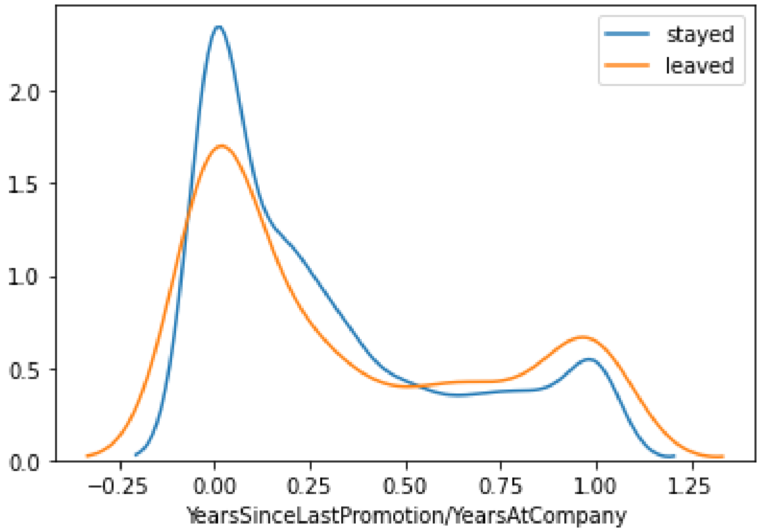
Comment: We can see that the biggest ration in HR department 30 % of them left

Employee satisfaction

Years last promotion

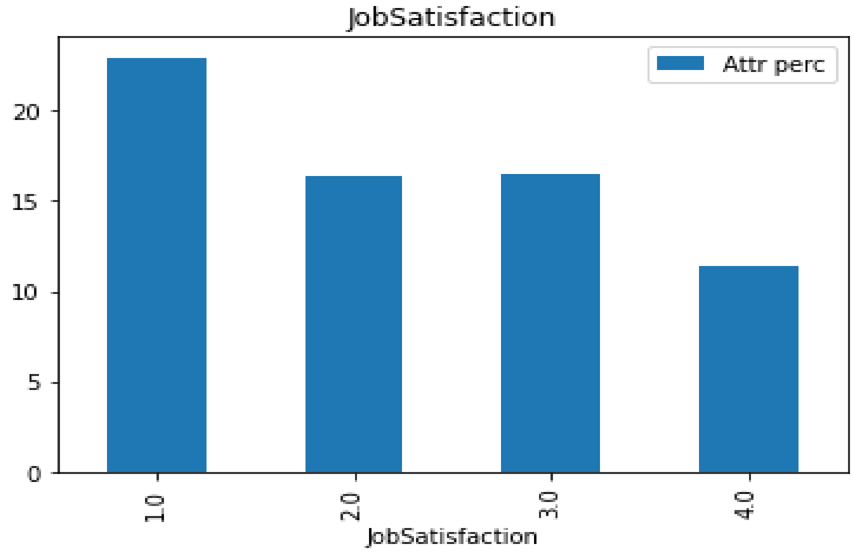


Comment: these distributions show that no effect of years since last promotion in attrition, as the two dist are almost the same.



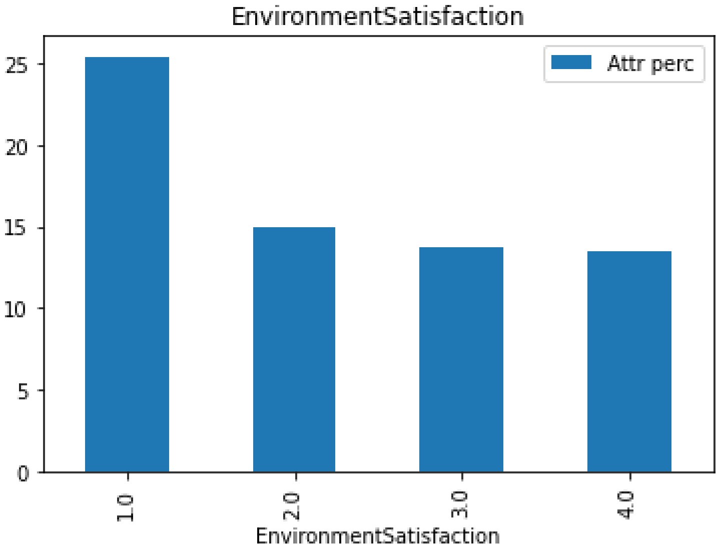
Comment: But when we divide it to the number of years stayed at the company, we can find a light effect of this ratio, and in after certain value of this ratio (about 0.51) the prob that the employee leave becomes higher

Job satisfaction



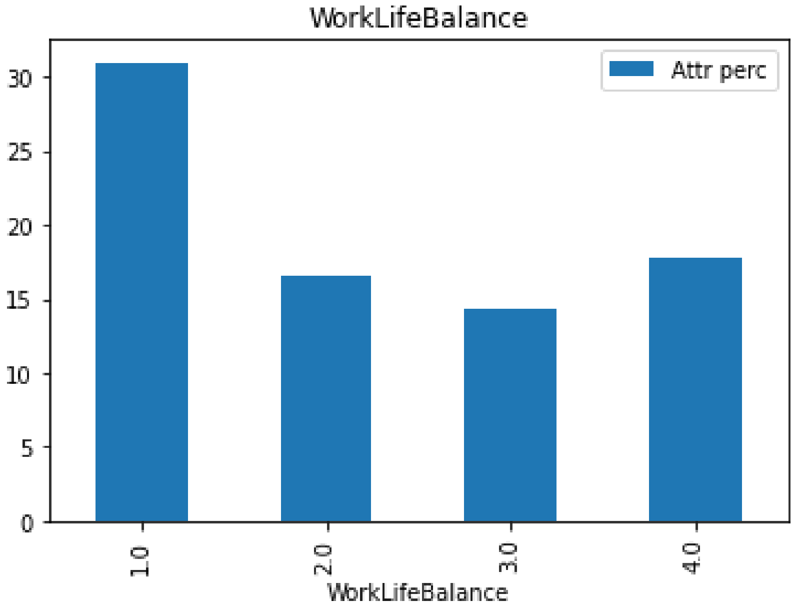
Comment: Of course more job satisfaction leads to less attrition percentage

Environment satisfaction



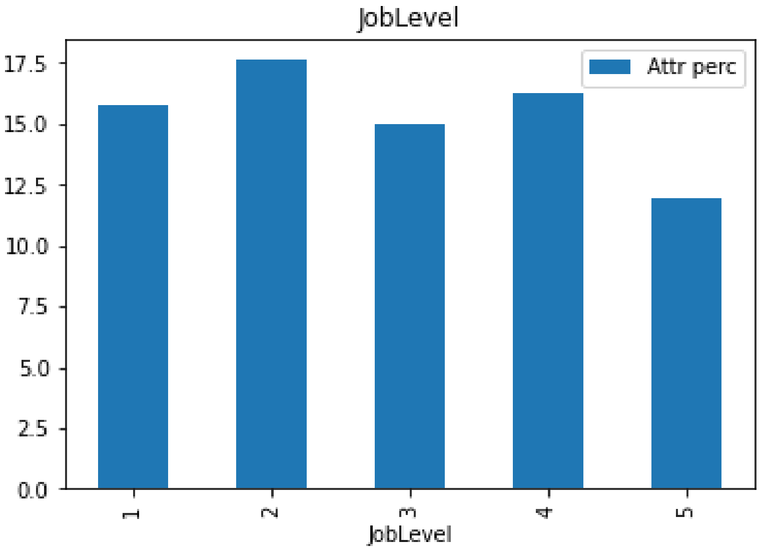
Comment: Also more environment satisfaction leads to less attrition percentage

Work-life balance



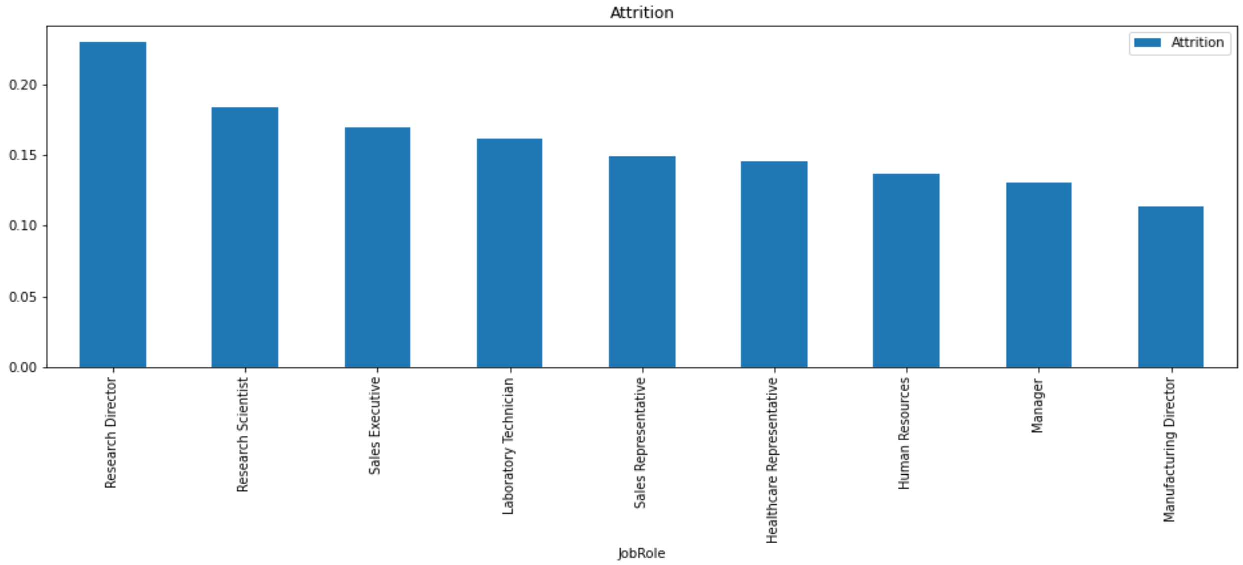
Comment: Also better work-life balance lead to less attrition percentage

Job level



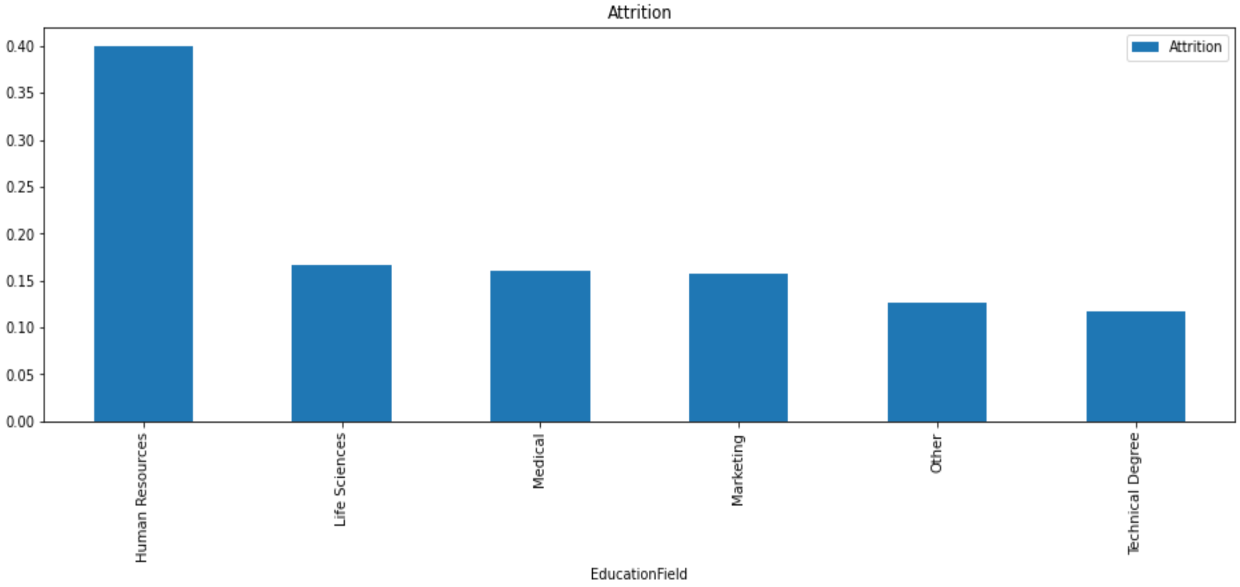
Comment: Job level does not contribute much to attrition

Job Role Effect



Comment: Most employees left pursued research roles because most employees work in R&D dep.

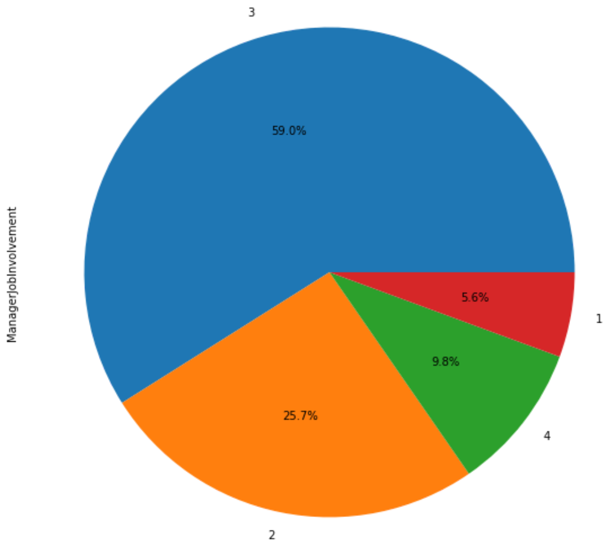
Education Field

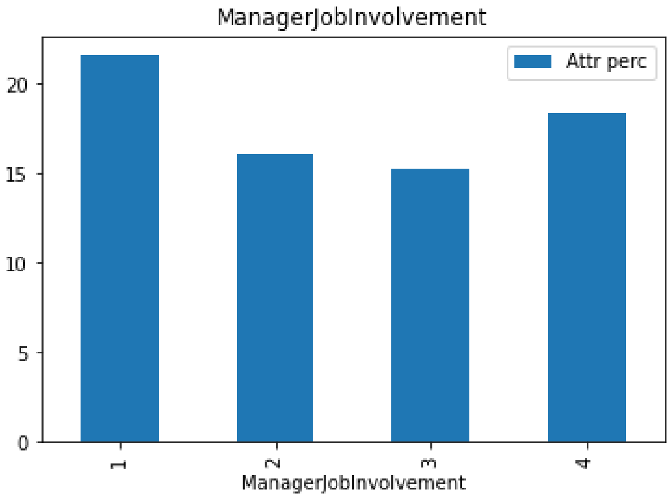


Comment: 40 % of employees left had human resources education.

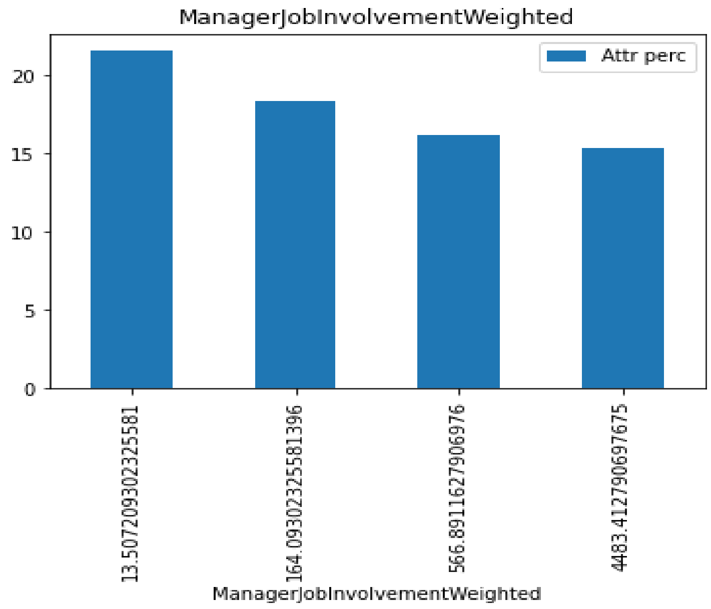
Manager survey effect on Attrition

ManagerJob Involvement





Comment: we can see there is something not clear here why attrition perc decreases and then increases that's because there are very small portion who rated as 4 as ex if only 5 rated 4 and 2 left we now have 40% attrition rate so maybe a better solution is sum these rates of each category and reweighting them with the percentage of the number of people rated for that rate we get that plot.



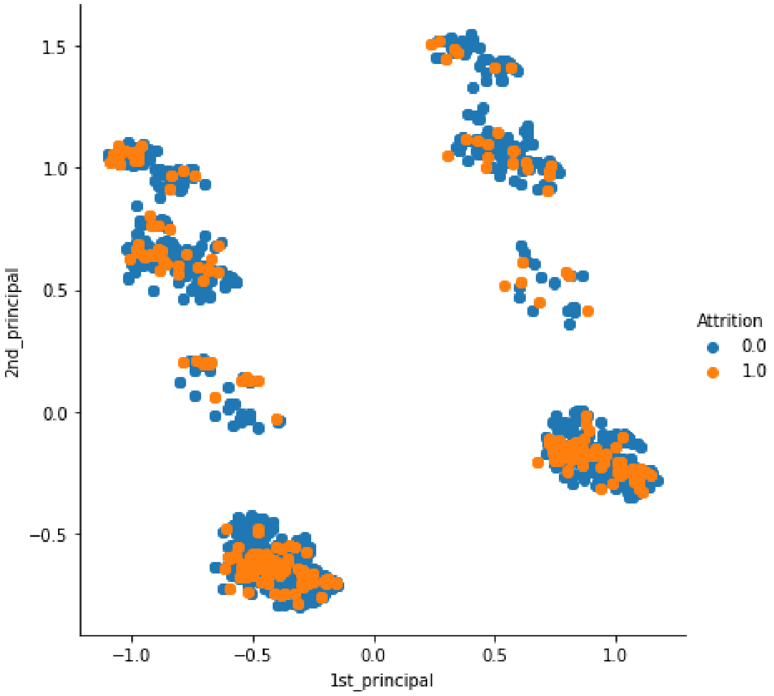
c. Data preparation & Feature engineering

- One hot encoding: one-hot encoding of categorical features like(education Field, gender, departments, etc..)
- Data normalization: min-max normalization of numeric features like (age, working hours mean, monthly income,etc..)
- Creating new columns.
    - Mean working hours: from the login and logout times we have created a new variable called mean working time of each employee as all employees almost have exact login time.

d. Models

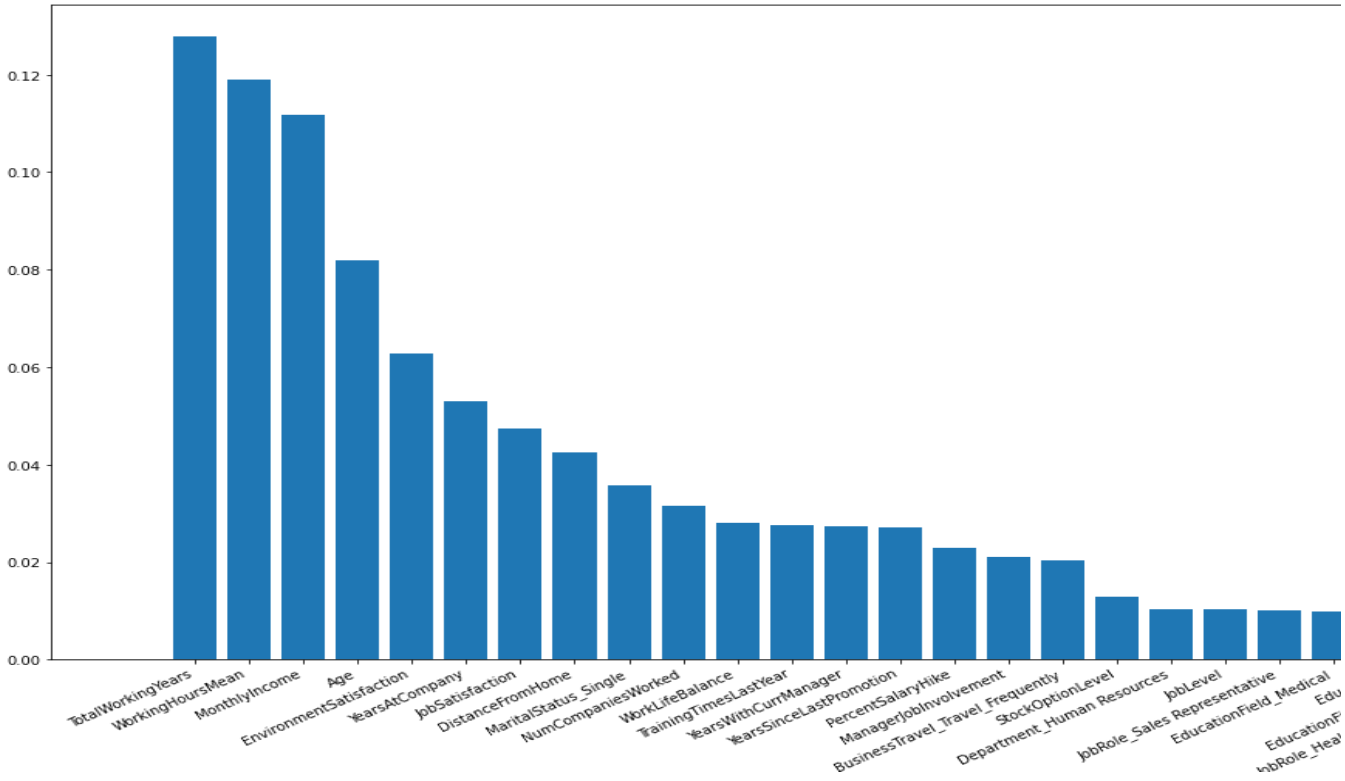
- Data firstly prepared as mentioned in c.
- Three models are built.

1. Logistic regression
    - It performs so bad on test set 80% accuracy and & .5 f1 score
    - The following graph explains that data is not linearly separable (the two classes we have employees that have left the company and employees that have stayed (Attrition)) and this is proved by doing PCA and visualization (of course the plot is not so accurate because the loss is 70% due to PCA dimensionality reduction but we can get from it some sense that linear classifiers (as logistic regression) is not suitable to this problem because it is not linearly separable)



○

2. [Selected Model] Neural Network
   - Neural Network with
     1. 3 hidden layers (100,50,25).
     2. Tanh activation functions. (to introduce nonlinearity)
     3. Sigmoid at the output layer.
   - The model performed very well on the test set 98% accuracy & .94 f1_score
   - This model is selected without doing any tuning to hyperparameters; we didn’t even need a validation.
3. Gradient Boost Classifier
   - This model performed nearly same NN (some random runs got better results than NN) but we stick to NN
   - This model mainly built for examining feature importance (as logistic didn't achieve good performance) and gain further insights about features and to support our claims and conclusions we made from visualizations of data



A cropped version from feature importance

e. Results and Evaluation.

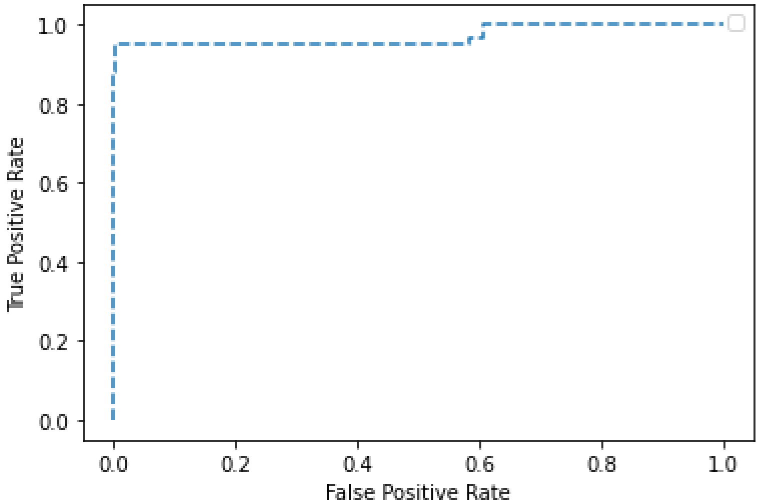
Model accuracy on train and test data.

Model	Training set	Test set
Neural Network	Accuracy: 100% F1 Score: 100%	Accuracy: 98%2 F1 Score: 94%

Confusion Matrix

Actual / Predicted	Positive	Negative
Positive	TP = 363	FN = 3
Negative	FP = 3	TN = 56

ROC curve



AUC (area under curve of ROC curve) = 0.969

f. Model Trials

- There is one unsuccessful trial that's by building logistic regression to predict the attrition probability and gives poor results but this is explained in the model building section.
- In the context of trials we have tried a neural network trained only on personal features and gets accuracy much better than logistic regression on all features (85 % f1 score compared to 50%), which tells us two things.
  a. Personal features have a great impact on attrition and must also be considered by the company when choosing employees.
  b. Logistic Regression is bad in this problem when the decision boundary is not linear

g. Any Enhancements and future work

- Maybe try to understand more things like why the hr department 30% of them leaves the company that will require more data mining about that department, the environment of work and etc.
- Try different models to get a better f1 score or use ensemble methods.

h. Conclusions and business insights

There are factors which will result in an employee to stay or to leave. Factors which highly affect attrition:

1. Total Working Years
   - people have more working years tends to stay in the company
   - That's also because of aging there is a clear relation between working in more companies and getting old this affects the probability of attrition as being young has less risk to leave the job
2. Department
   - There is an obvious problem in the hr department as about 30% of the company tends to leave
   - We have also seen that they have the lowest monthly income comparing it to R&D and Sales departments
3. Environment Satisfaction
   - Of course, we have seen that better environment satisfaction lead to a lower probability of attrition
   - So the company can provide more suitable, attracting and stress-free environment
4. Job Satisfaction

- Also, job satisfaction has an impact on attrition the more employee is satisfied with the job the less he will tend to leave
- So the company can hire better hr employees to solve employees problems with job satisfaction

5. Personal features (Marital status & Age & number of companies employees worked on).
   - Seems that being single and young gives you more feasibility to risk leaving the job and go for another one and vice versa
   - Also in the model trials, we have built a model based only on personal features (Marital status, Age ….) and we found that it's accuracy much better than logistic regression on all features (85 % f1 score compared to 50%), that means, Personal features have a great impact on the attrition and must also be considered by the company when choosing employees.

6. Working hours mean
   - People who have smaller working hours tend to stay at the company because they have work-life balance and no job stress.
   - On the other hand, people who work much have poor work-life balance and there is no return from the company hence they tend to leave the company.
   - So the company has two solutions one is to ensure work-life balance by setting strict rules on working hours.
   - Or they provide any kind of return for people that tend to work more

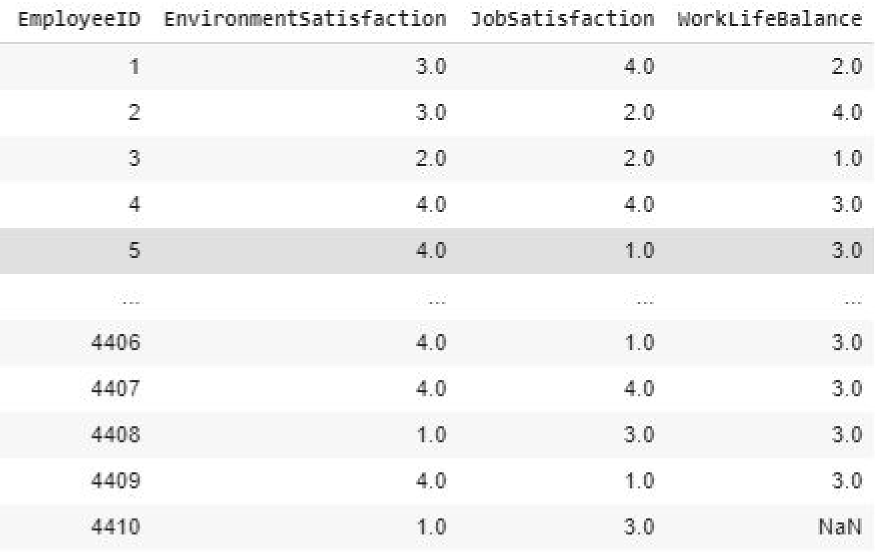
7. Work-life balance
   - Employees who rate for higher work-life balance has a lower attrition rate
   - So as mention in Working hours mean the company has to deal with that

8. Managers Effect on Employees attrition
   - We can see that the more employees feel that managers more involved in their jobs they have less probability of leaving their jobs
   - So the company must choose their managers carefully and must choose managers that only try to work the employees and not only give orders.

i. Appendix

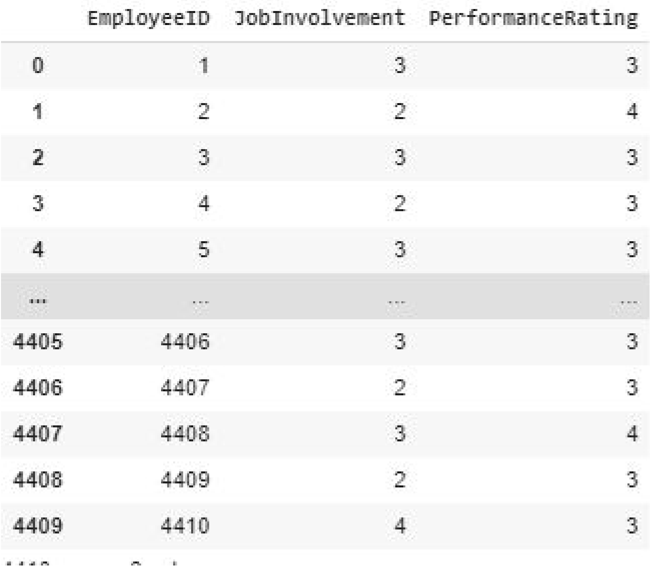
Dataset Sample

Employee Survey



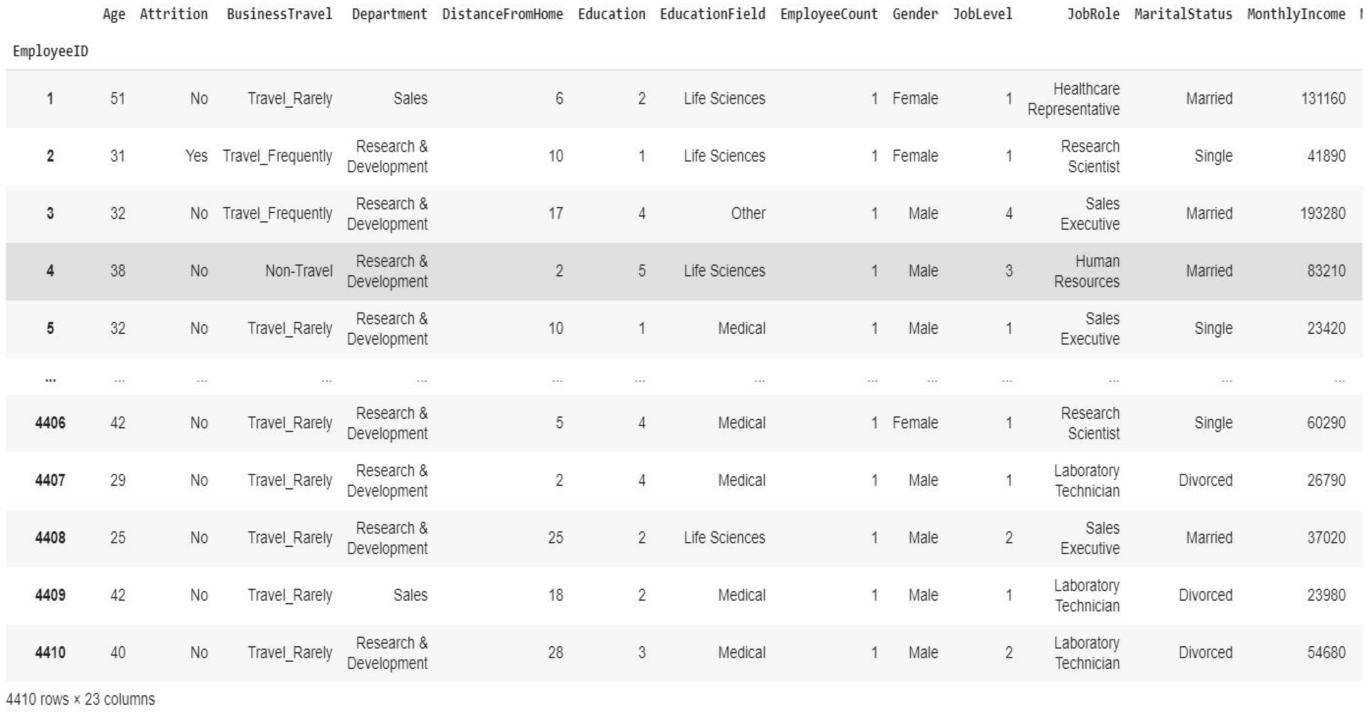
EmployeeID	EnvironmentSatisfaction	JobSatisfaction	WorkLifeBalance
1	3.0	4.0	2.0
2	3.0	2.0	4.0
3	2.0	2.0	1.0
4	4.0	4.0	3.0
5	4.0	1.0	3.0
...	...	...	...
4406	4.0	1.0	3.0
4407	4.0	4.0	3.0
4408	1.0	3.0	3.0
4409	4.0	1.0	3.0
4410	1.0	3.0	NaN

Manager Survey



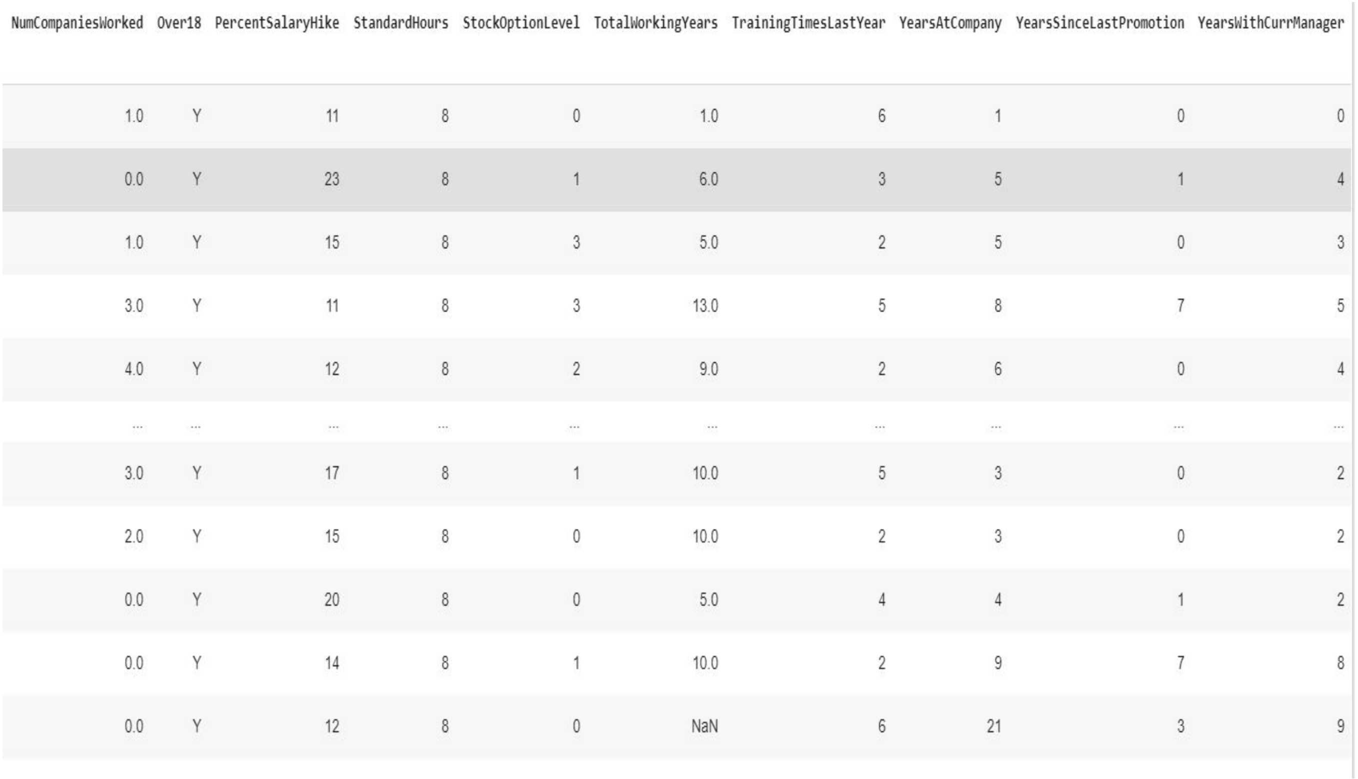
	EmployeeID	JobInvolvement	PerformanceRating
0	1	3	3
1	2	2	4
2	3	3	3
3	4	2	3
4	5	3	3
...	...	...	...
4405	4406	3	3
4406	4407	2	3
4407	4408	3	4
4408	4409	2	3
4409	4410	4	3

General



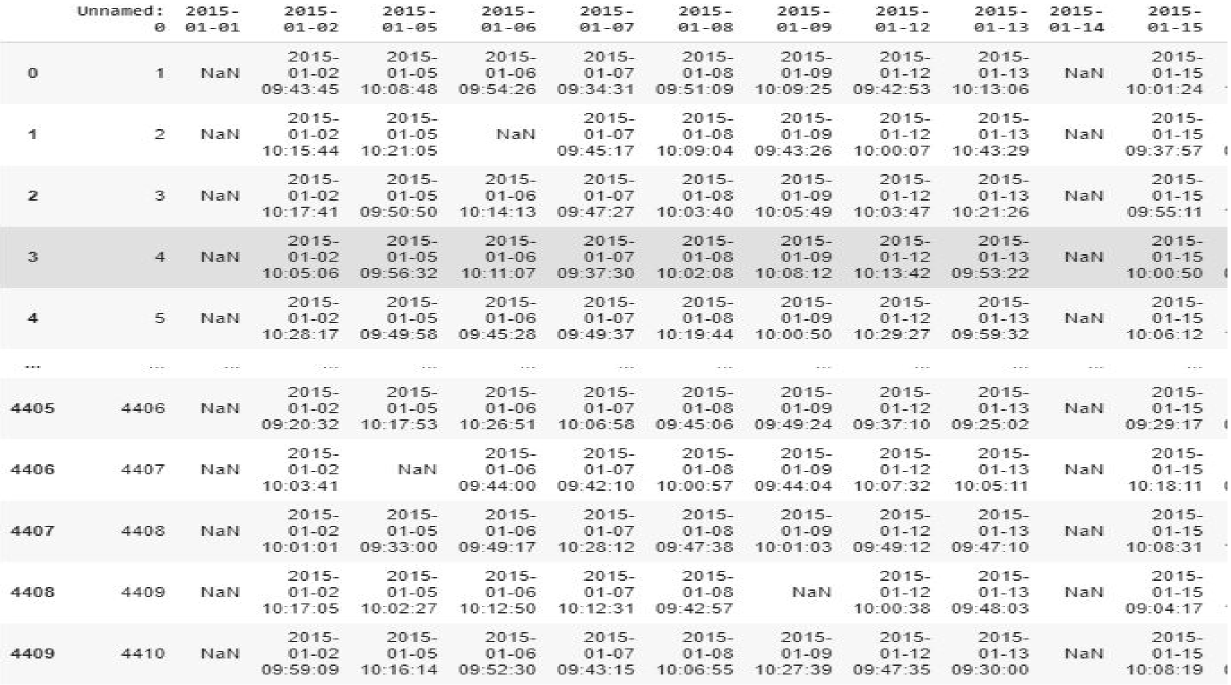
	Age	Attrition	BusinessTravel	Department	DistanceFromHome	Education	EducationField	EmployeeCount	Gender	JobLevel	JobRole	MaritalStatus	MonthlyIncome
EmployeeID													
1	51	No	Travel_Rarely	Sales	6	2	Life Sciences	1	Female	1	Healthcare Representative	Married	131160
2	31	Yes	Travel_Frequently	Research & Development	10	1	Life Sciences	1	Female	1	Research Scientist	Single	41890
3	32	No	Travel_Frequently	Research & Development	17	4	Other	1	Male	4	Sales Executive	Married	193280
4	38	No	Non-Travel	Research & Development	2	5	Life Sciences	1	Male	3	Human Resources	Married	83210
5	32	No	Travel_Rarely	Research & Development	10	1	Medical	1	Male	1	Sales Executive	Single	23420
...	...	...	...	...	...	...	...	...	...	...	...	...	...
4406	42	No	Travel_Rarely	Research & Development	5	4	Medical	1	Female	1	Research Scientist	Single	60290
4407	29	No	Travel_Rarely	Research & Development	2	4	Medical	1	Male	1	Laboratory Technician	Divorced	26790
4408	25	No	Travel_Rarely	Research & Development	25	2	Life Sciences	1	Male	2	Sales Executive	Married	37020
4409	42	No	Travel_Rarely	Sales	18	2	Medical	1	Male	1	Laboratory Technician	Divorced	23980
4410	40	No	Travel_Rarely	Research & Development	28	3	Medical	1	Male	2	Laboratory Technician	Divorced	54680

4410 rows × 23 columns



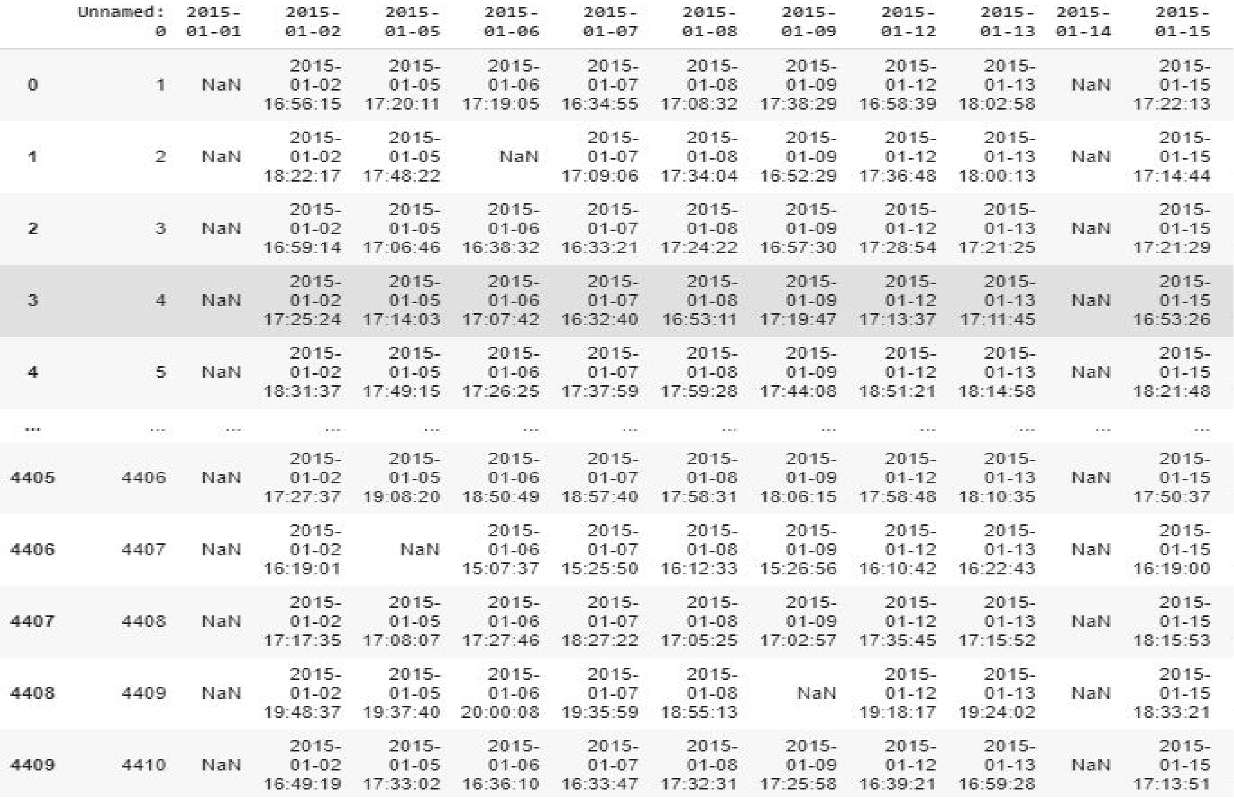
NumCompaniesWorked	Over18	PercentSalaryHike	StandardHours	StockOptionLevel	TotalWorkingYears	TrainingTimesLastYear	YearsAtCompany	YearsSinceLastPromotion	YearsWithCurrManager
1.0	Y	11	8	0	1.0	6	1	0	0
0.0	Y	23	8	1	6.0	3	5	1	4
1.0	Y	15	8	3	5.0	2	5	0	3
3.0	Y	11	8	3	13.0	5	8	7	5
4.0	Y	12	8	2	9.0	2	6	0	4
...	...	...	...	...	...	...	...	...	...
3.0	Y	17	8	1	10.0	5	3	0	2
2.0	Y	15	8	0	10.0	2	3	0	2
0.0	Y	20	8	0	5.0	4	4	1	2
0.0	Y	14	8	1	10.0	2	9	7	8
0.0	Y	12	8	0	NaN	6	21	3	9

Log in time (cropped till 2015-01-15 to fit in the page)



	Unnamed: 0	2015-01-01	2015-01-02	2015-01-05	2015-01-06	2015-01-07	2015-01-08	2015-01-09	2015-01-12	2015-01-13	2015-01-14	2015-01-15
0	1	NaN	2015-01-02 09:43:45	2015-01-05 10:08:48	2015-01-06 09:54:26	2015-01-07 09:34:31	2015-01-08 09:51:09	2015-01-09 10:09:25	2015-01-12 09:42:53	2015-01-13 10:13:06	NaN	2015-01-15 10:01:24
1	2	NaN	2015-01-02 10:15:44	2015-01-05 10:21:05	NaN	2015-01-07 09:45:17	2015-01-08 10:09:04	2015-01-09 09:43:26	2015-01-12 10:00:07	2015-01-13 10:43:29	NaN	2015-01-15 09:37:57
2	3	NaN	2015-01-02 10:17:41	2015-01-05 09:50:50	2015-01-06 10:14:13	2015-01-07 09:47:27	2015-01-08 10:03:40	2015-01-09 10:05:49	2015-01-12 10:03:47	2015-01-13 10:21:26	NaN	2015-01-15 09:55:11
3	4	NaN	2015-01-02 10:05:06	2015-01-05 09:56:32	2015-01-06 10:11:07	2015-01-07 09:37:30	2015-01-08 10:02:08	2015-01-09 10:08:12	2015-01-12 10:13:42	2015-01-13 09:53:22	NaN	2015-01-15 10:00:50
4	5	NaN	2015-01-02 10:28:17	2015-01-05 09:49:58	2015-01-06 09:45:28	2015-01-07 09:49:37	2015-01-08 10:19:44	2015-01-09 10:00:50	2015-01-12 10:29:27	2015-01-13 09:59:32	NaN	2015-01-15 10:06:12
...	...	...	...	...	...	...	...	...	...	...	...	...
4405	4406	NaN	2015-01-02 09:20:32	2015-01-05 10:17:53	2015-01-06 10:26:51	2015-01-07 10:06:58	2015-01-08 09:45:06	2015-01-09 09:49:24	2015-01-12 09:37:10	2015-01-13 09:25:02	NaN	2015-01-15 09:29:17
4406	4407	NaN	2015-01-02 10:03:41	NaN	2015-01-06 09:44:00	2015-01-07 09:42:10	2015-01-08 10:00:57	2015-01-09 09:44:04	2015-01-12 10:07:32	2015-01-13 10:05:11	NaN	2015-01-15 10:18:11
4407	4408	NaN	2015-01-02 10:01:01	2015-01-05 09:33:00	2015-01-06 09:49:17	2015-01-07 10:28:12	2015-01-08 09:47:38	2015-01-09 10:01:03	2015-01-12 09:49:12	2015-01-13 09:47:10	NaN	2015-01-15 10:08:31
4408	4409	NaN	2015-01-02 10:17:05	2015-01-05 10:02:27	2015-01-06 10:12:50	2015-01-07 10:12:31	2015-01-08 09:42:57	NaN	2015-01-12 10:00:38	2015-01-13 09:48:03	NaN	2015-01-15 09:04:17
4409	4410	NaN	2015-01-02 09:59:09	2015-01-05 10:16:14	2015-01-06 09:52:30	2015-01-07 09:43:15	2015-01-08 10:06:55	2015-01-09 10:27:39	2015-01-12 09:47:35	2015-01-13 09:30:00	NaN	2015-01-15 10:08:19

Log Out time (cropped till 2015-01-15 to fit in the page)



	Unnamed: 0	2015-01-01	2015-01-02	2015-01-05	2015-01-06	2015-01-07	2015-01-08	2015-01-09	2015-01-12	2015-01-13	2015-01-14	2015-01-15
0	1	NaN	2015-01-02 16:56:15	2015-01-05 17:20:11	2015-01-06 17:19:05	2015-01-07 16:34:55	2015-01-08 17:08:32	2015-01-09 17:38:29	2015-01-12 16:58:39	2015-01-13 18:02:58	NaN	2015-01-15 17:22:13
1	2	NaN	2015-01-02 18:22:17	2015-01-05 17:48:22	NaN	2015-01-07 17:09:06	2015-01-08 17:34:04	2015-01-09 16:52:29	2015-01-12 17:36:48	2015-01-13 18:00:13	NaN	2015-01-15 17:14:44
2	3	NaN	2015-01-02 16:59:14	2015-01-05 17:06:46	2015-01-06 16:38:32	2015-01-07 16:33:21	2015-01-08 17:24:22	2015-01-09 16:57:30	2015-01-12 17:28:54	2015-01-13 17:21:25	NaN	2015-01-15 17:21:29
3	4	NaN	2015-01-02 17:25:24	2015-01-05 17:14:03	2015-01-06 17:07:42	2015-01-07 16:32:40	2015-01-08 16:53:11	2015-01-09 17:19:47	2015-01-12 17:13:37	2015-01-13 17:11:45	NaN	2015-01-15 16:53:26
4	5	NaN	2015-01-02 18:31:37	2015-01-05 17:49:15	2015-01-06 17:26:25	2015-01-07 17:37:59	2015-01-08 17:59:28	2015-01-09 17:44:08	2015-01-12 18:51:21	2015-01-13 18:14:58	NaN	2015-01-15 18:21:48
...	...	...	...	...	...	...	...	...	...	...	...	...
4405	4406	NaN	2015-01-02 17:27:37	2015-01-05 19:08:20	2015-01-06 18:50:49	2015-01-07 18:57:40	2015-01-08 17:58:31	2015-01-09 18:06:15	2015-01-12 17:58:48	2015-01-13 18:10:35	NaN	2015-01-15 17:50:37
4406	4407	NaN	2015-01-02 16:19:01	NaN	2015-01-06 15:07:37	2015-01-07 15:25:50	2015-01-08 16:12:33	2015-01-09 15:26:56	2015-01-12 16:10:42	2015-01-13 16:22:43	NaN	2015-01-15 16:19:00
4407	4408	NaN	2015-01-02 17:17:35	2015-01-05 17:08:07	2015-01-06 17:27:46	2015-01-07 18:27:22	2015-01-08 17:05:25	2015-01-09 17:02:57	2015-01-12 17:35:45	2015-01-13 17:15:52	NaN	2015-01-15 18:15:53
4408	4409	NaN	2015-01-02 19:48:37	2015-01-05 19:37:40	2015-01-06 20:00:08	2015-01-07 19:35:59	2015-01-08 18:55:13	NaN	2015-01-12 19:18:17	2015-01-13 19:24:02	NaN	2015-01-15 18:33:21
4409	4410	NaN	2015-01-02 16:49:19	2015-01-05 17:33:02	2015-01-06 16:36:10	2015-01-07 16:33:47	2015-01-08 17:32:31	2015-01-09 17:25:58	2015-01-12 16:39:21	2015-01-13 16:59:28	NaN	2015-01-15 17:13:51

Dataset Description Table

Variable	Meaning	Value
Age	Age of the employee	
Attrition	Whether the employee left in the previous year or not	
Business travel	How frequently the employees traveled for business purposes in the last year	
Department	Department in the company	
DistanceFromHome	Distance from home in km	
Education	Education Level	1 'Below College'

		2 'College' 3 'Bachelor' 4 'Master' 5 'Doctor'
EducationField	Field of education	
Employee Count	Employee count	
Employee number	Employee number/id	
Environment Satisfaction	Work Environment Satisfaction Level	1 'Low' 2 'Medium' 3 'High' 4 'Very High'
Gender	Gender of employee	
JobInvolvement	Job Involvement Level	1 'Low' 2 'Medium' 3 'High' 4 'Very High'
JobLevel	Job level at the company on a scale of 1 to 5	
Job role	Name of the job role in the company	
Job satisfaction	Job Satisfaction Level	1 'Low' 2 'Medium' 3 'High' 4 'Very High'
Marital status	Marital status of the employee	
Monthly income	Monthly income in rupees per month	
NumCompaniesWorked	Total number of companies the employee has worked for	
Over18	Whether the employee is above 18 years of age or not	

PercentSalaryHike	Percent salary hike for last year	
Performance rating	Performance rating for last year	1 'Low' 2 'Good' 3 'Excellent' 4 'Outstanding'
Relationship satisfaction	Relationship satisfaction level	1 'Low' 2 'Medium' 3 'High' 4 'Very High'
StandardHours	Standard hours of work for the employee	
StockOptionLevel	Stock options level of the employee	
TotalWorkingYears	Total number of years the employee has worked so far	
TrainingTimesLastYear	Number of times training was conducted for this employee last year	
Work-life balance	Work-life balance level	1 'Bad' 2 'Good' 3 'Better' 4 'Best'
YearsAtCompany	Total number of years spent at the company by the employee	
YearsSinceLastPromotion	Number of years since the last promotion	
YearsWithCurrManager	Number of years under current manager	