UC San Diego

GROUP 14 PRESENTS

EMPLOYEE ATTRITION
WHY DO WORKERS QUIT?

University of California, San Diego
Department of Electrical and Computer Engineering
Team Members : Nikhil Mohan, Xuezhu Hong, Qi Ma, Changhan Ge
03.13.2019

CONTENTS

INTRODUCTION

- Attrition is basically the turnover rate of employees within an organization. Our goal is to analyze the factors which determine attrition and suggest possible remedial measures.
- Variety of factors can determine attrition, and these are, but not limited to
    1. Hostile environment
    2. Long work hours with less pay
    3. Employees looking for better opportunities.
    4. Poor management

CONTENTS

DATASET

- **IBM Attrition Dataset**
  - A US Science and Technology Company
  - Source: Kaggle

    https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset
- **BRC Attrition Dataset**
  - A Chinese real estate & tourism investment group
  - Source: Provided by Ms. Jane Liu who served as a HR at BRC
- **Employee Review Dataset**
  - Company reviews written by employees of several silicon valley companies
  - Source: Kaggle

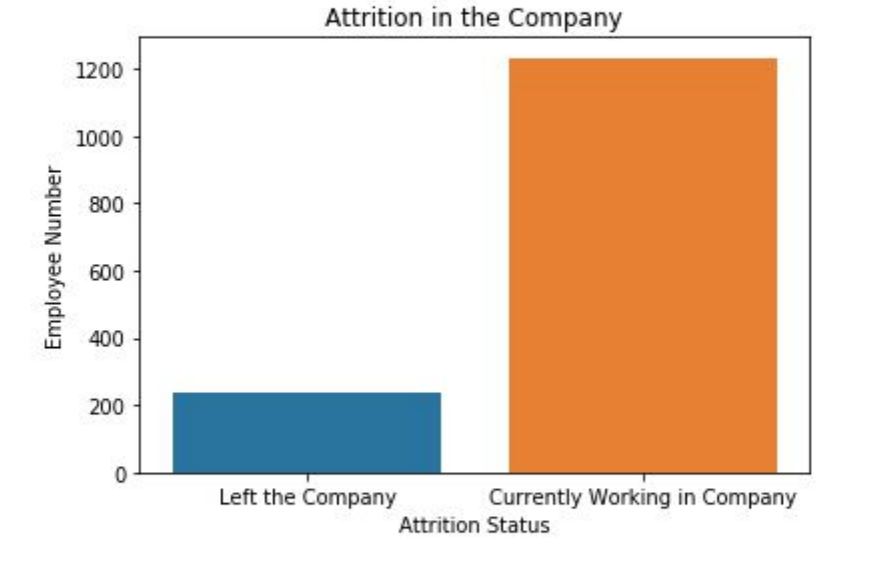
    https://www.kaggle.com/petersunga/google-amazon-facebook-employee-reviews
- **Basics about dataset**

	Total Sample	Left the Company	Still in the Company	Attrition Rate
IBM	1470	237	1233	16%
BRC	838	752	86	90%
Review	67000		N/A	

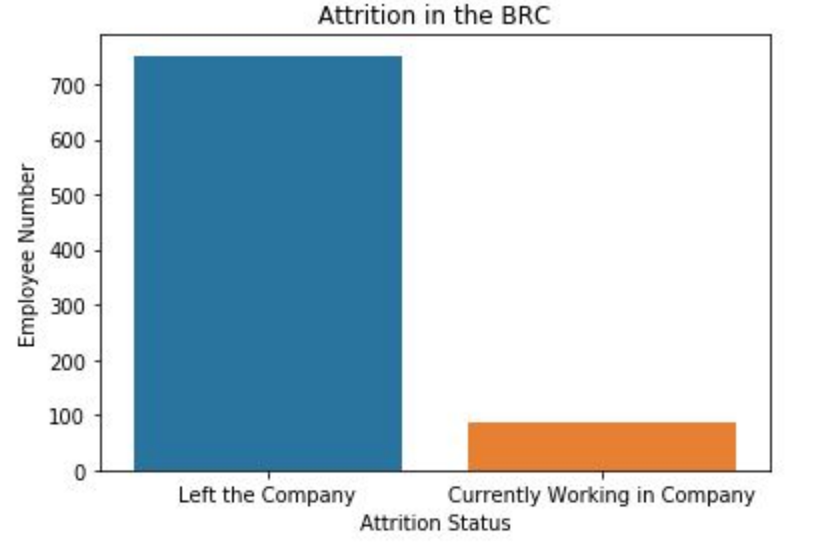
CONTENTS

Distribution of features - Attrition

IBM DATASET

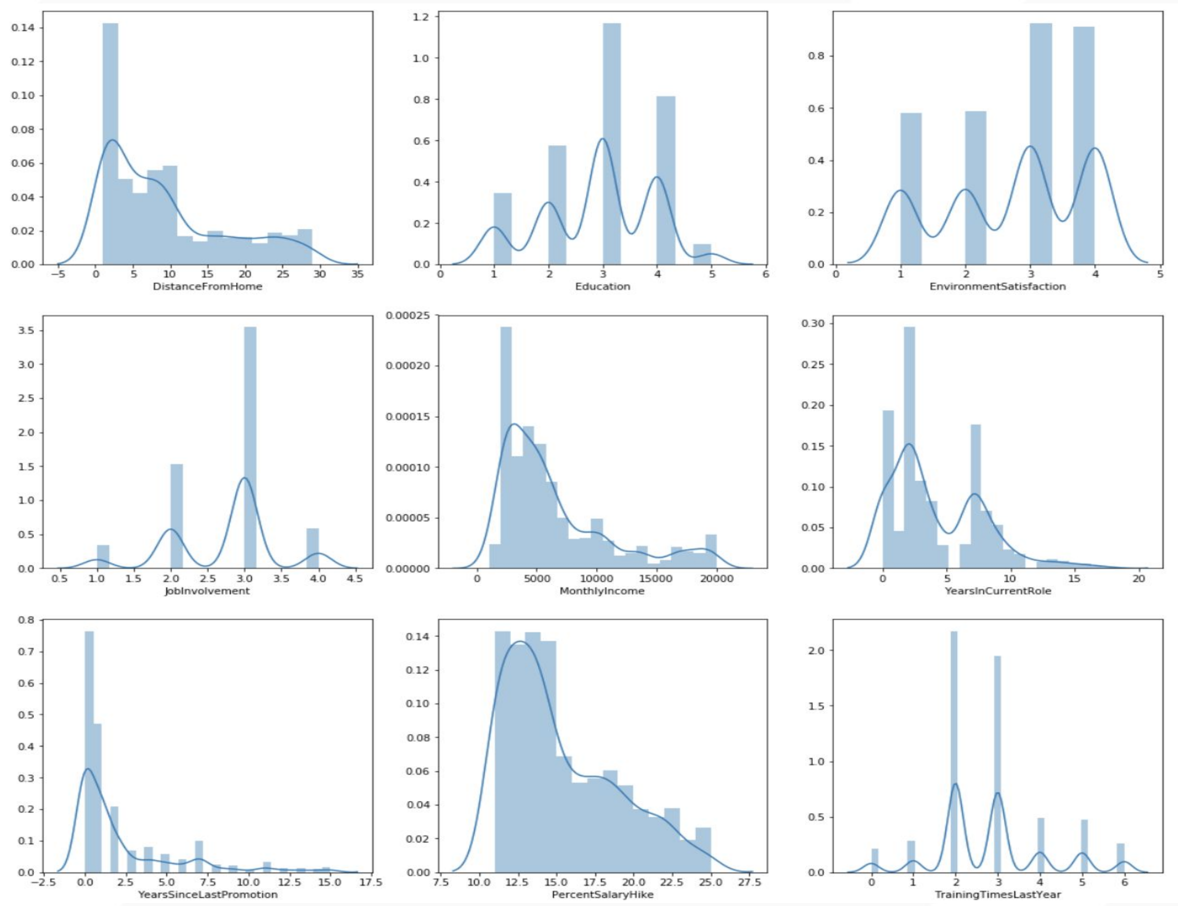


BRC DATASET



- **An important aspect to consider is the fact that we are dealing with an imbalanced dataset in both the cases. This is crucial when we consider the accuracy of our prediction models.**
- **Due to the contrasting nature of the two datasets, we will try to draw suitable conclusions from both datasets wherever possible.**

Distribution of auxiliary features - IBM Dataset



- **A good practice is to look at the spread of features within a dataset.**
- **We use the IBM dataset as the data is more clean and thorough.**
- **The curves are more or less a mixture of gaussians and this is a good indicator that our dataset consists of mainly independent samples and is therefore unbiased.**

Age & Sex

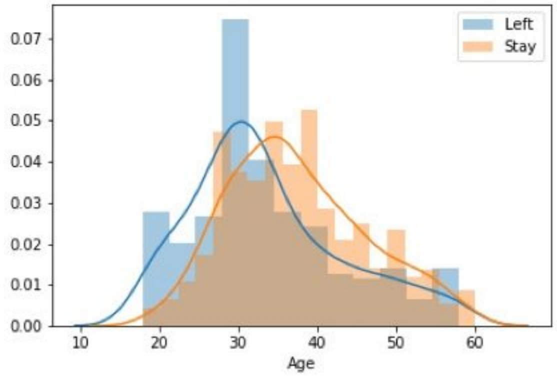
❖ **Age**

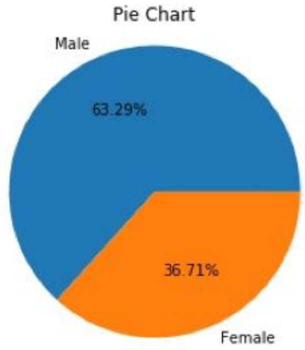
- **Elder employee intends a stable life**
- **Young employees intend to seeking better opportunities**

❖ **Gender**

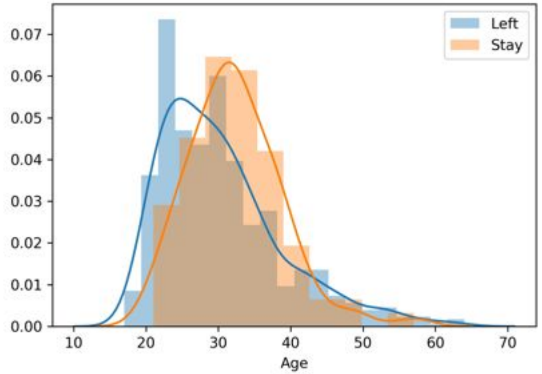
- **Male is more likely to quit the job than female**

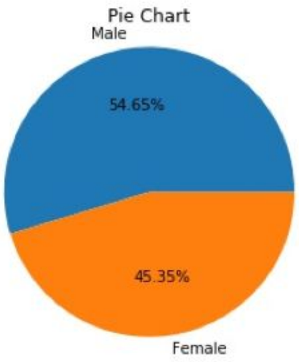
IBM DATASET



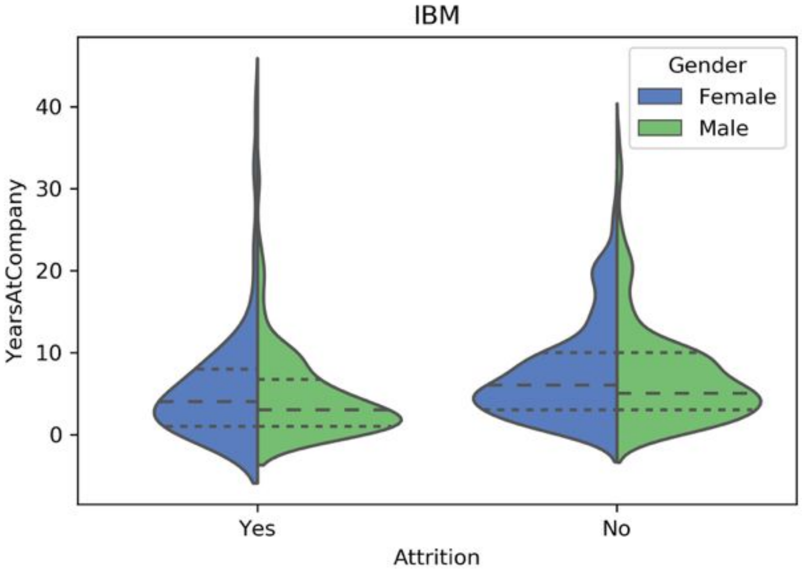


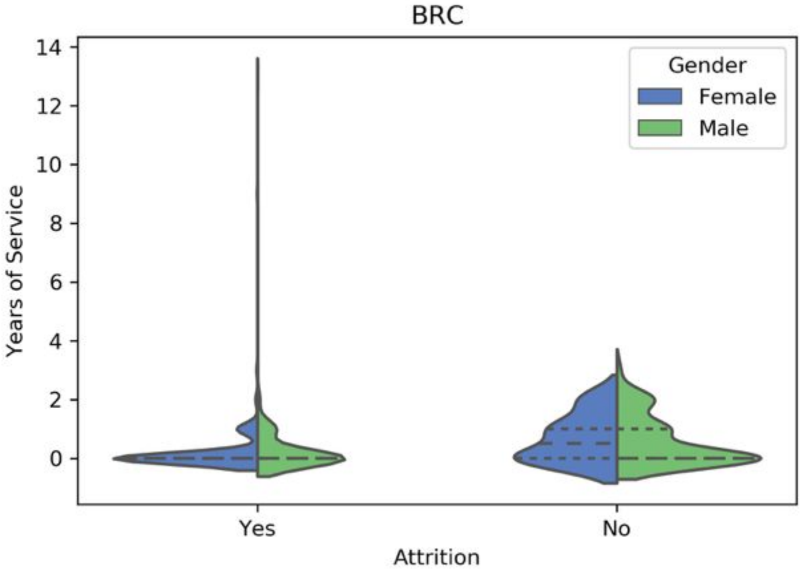
BRC DATASET





Year of Service

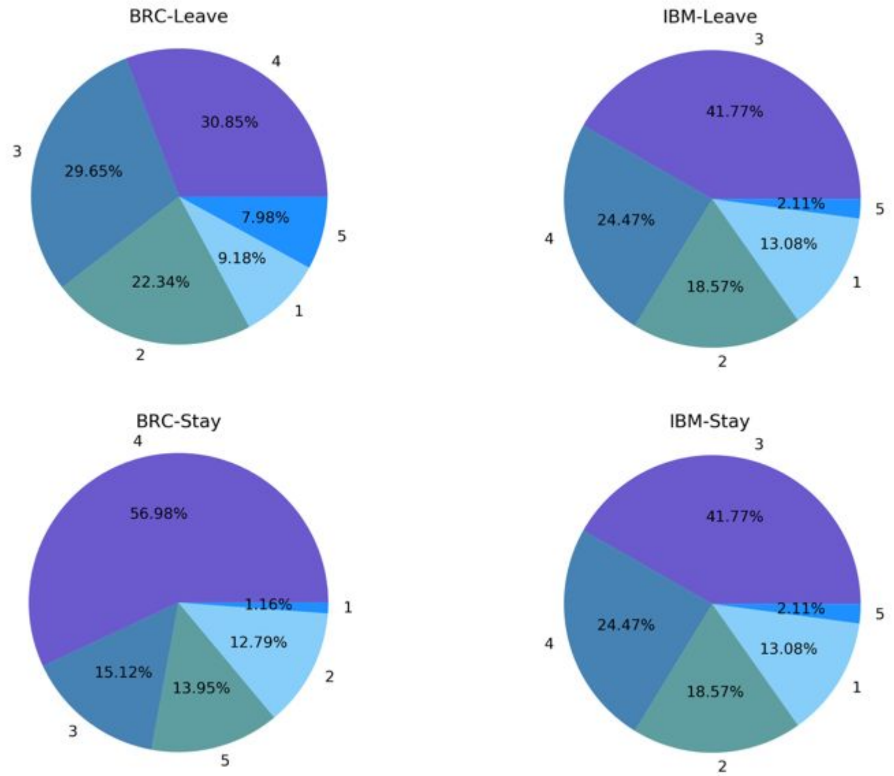




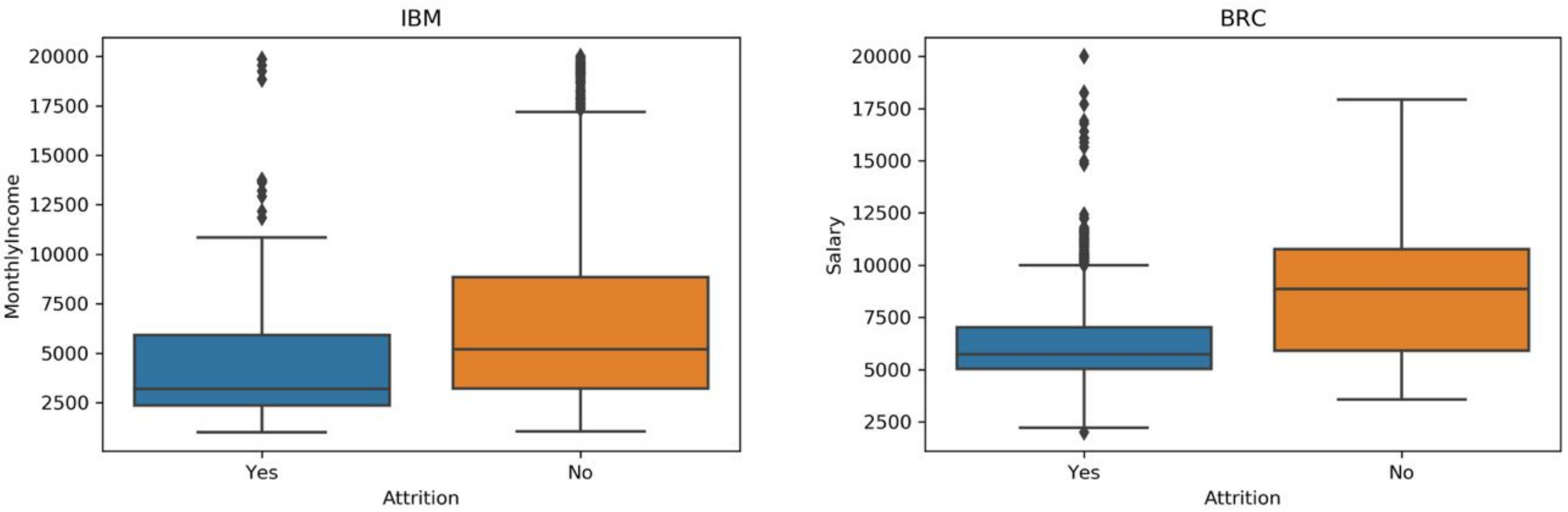
- ❖ **The year of service doesn’t differ very much on gender**
- ❖ **The shorter the year of service, the higher the possibility to quit**

Education

- ❖ **Education Level (1-5 Low to High)**
- **The education level of people who leaves company and people who stays in company share the same distribution.**
- **There is no apparent indication that education level has direct relationship with attrition**

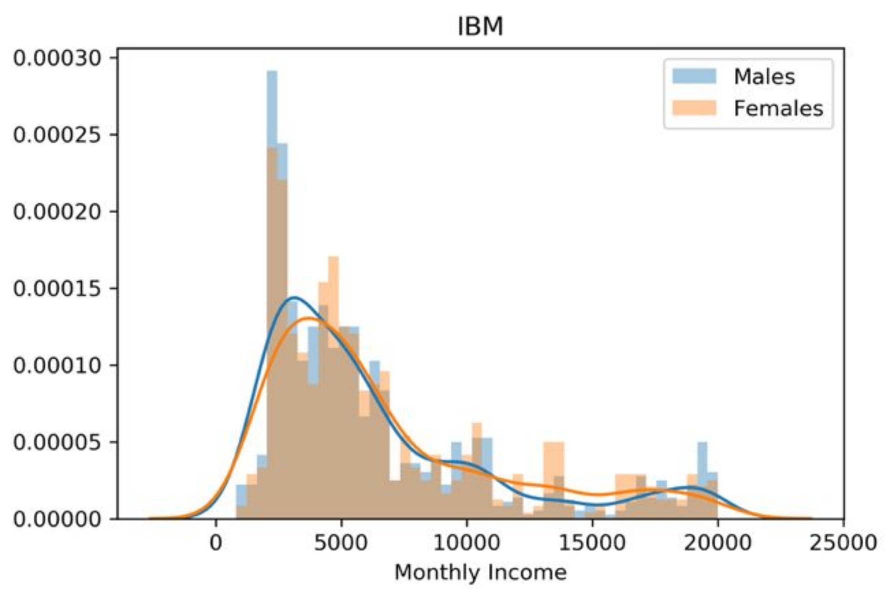


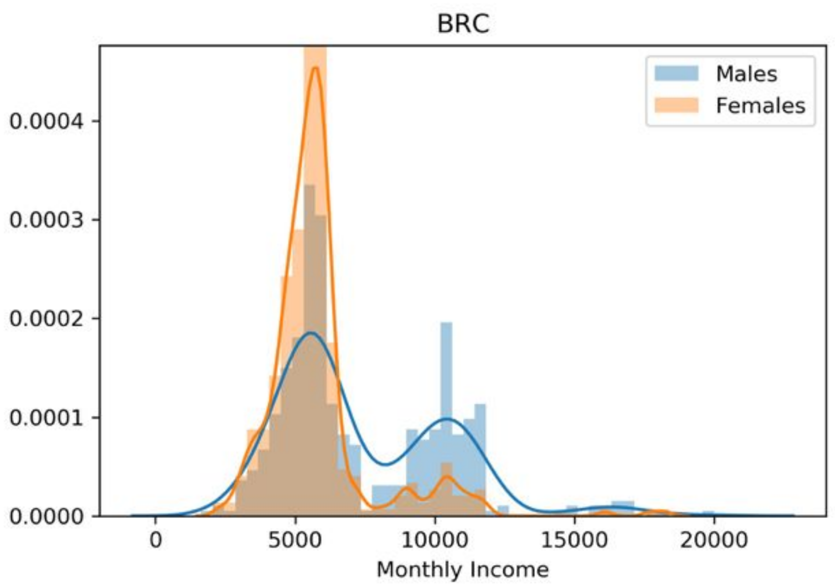
Monthly Income



- ❖ **Employees who stay has a higher salary than employees who leave**
- ❖ **Salary is a important factor on people’s leaving**

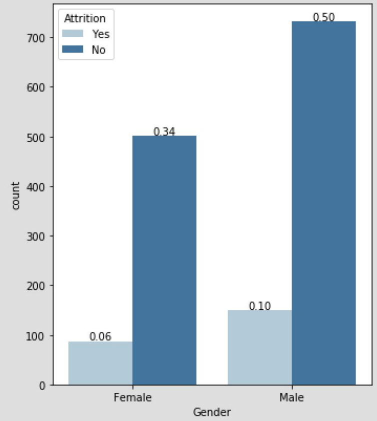
Monthly Income & Gender

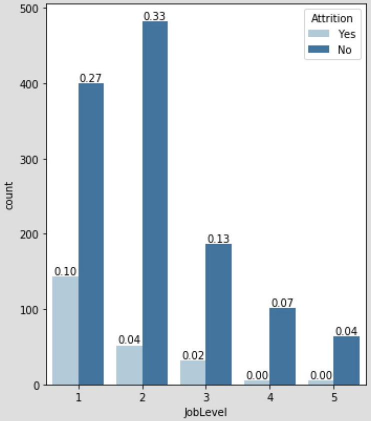


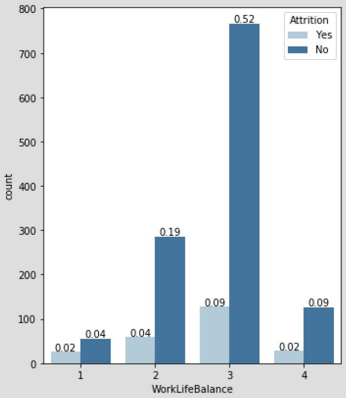


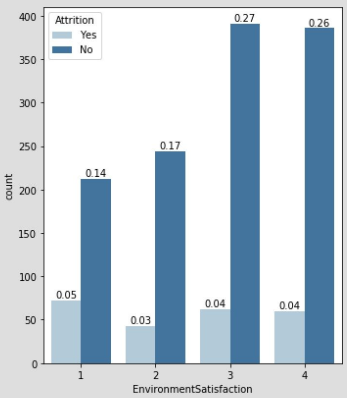
- ❖ **In IBM, Males and Females has little difference on monthly income**
- ❖ **In BRC, Male gets more salary than female**

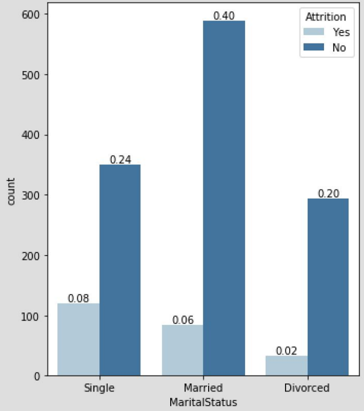
Attrition comparison

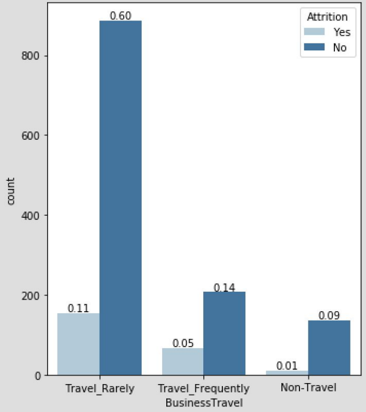






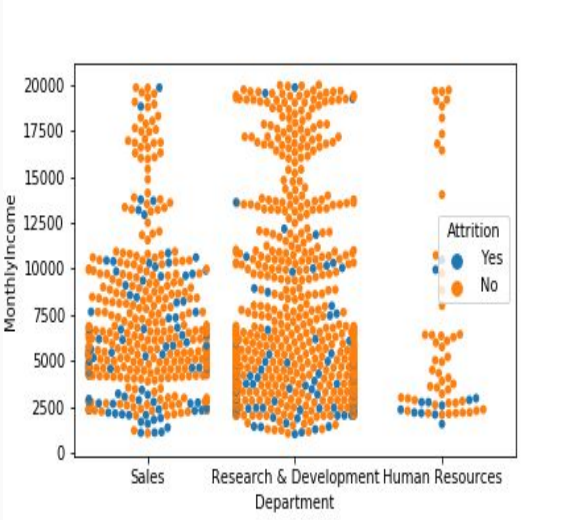


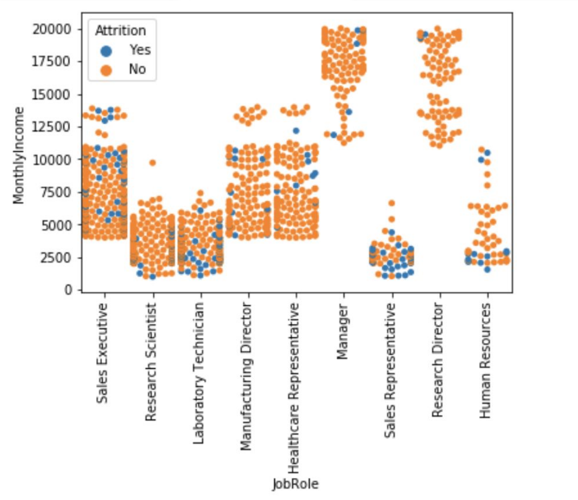


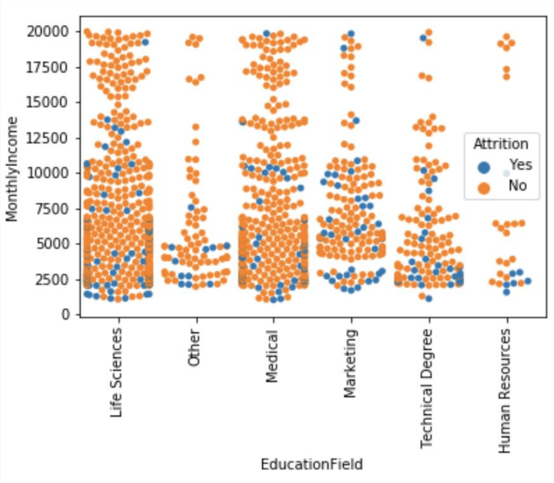


- **Attrition ratio: female < male**
- **People at job level 2 might be more willing to stay, but those at level 1 tend to quit**
- **People with WorkLifeBalance 3 tend to stay**
- **More satisfied to environment, more likely to stay**
- **Single people hop more**
- **If the position with less travel, people will like the job more**

SwarmPlot - factors influence income

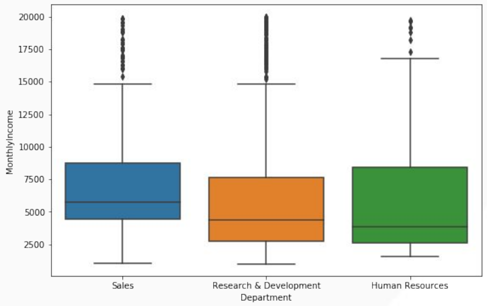


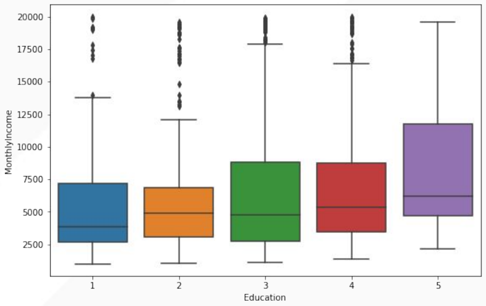


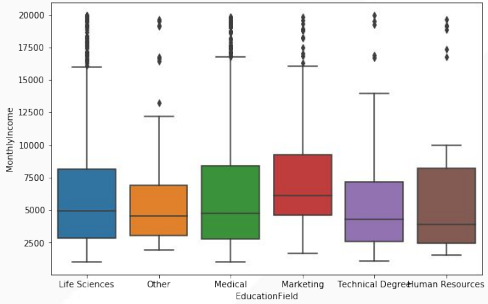


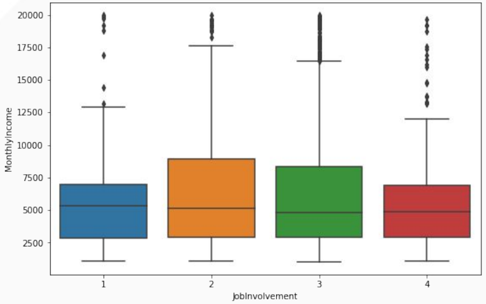
- With SwarmPlot, we can get location of every single point, and the overall distribution
- e.g, the income distribution of people from Sales and R & D are relatively evenly, but not for HRs
- Manager and research director earn much more
- Quite a bit samples are from Life Sciences and Medical background individuals

Monthly Income and other features-IBM Dataset









- **Even though R & D is hard, the income is lower than Sales and HRs**
- **Education level (1-5, low-high), higher degree, higher income**
- **People from marketing field tend to get higher salary**
- **Income for people with medium job involvement ranges a lot**

Monthly Income By Age - IBM dataset

- Younger employees are typically paid lesser than the more experienced employees.
- This can be accounted for, by observing that more experienced employees tend to hold more prestigious job roles.

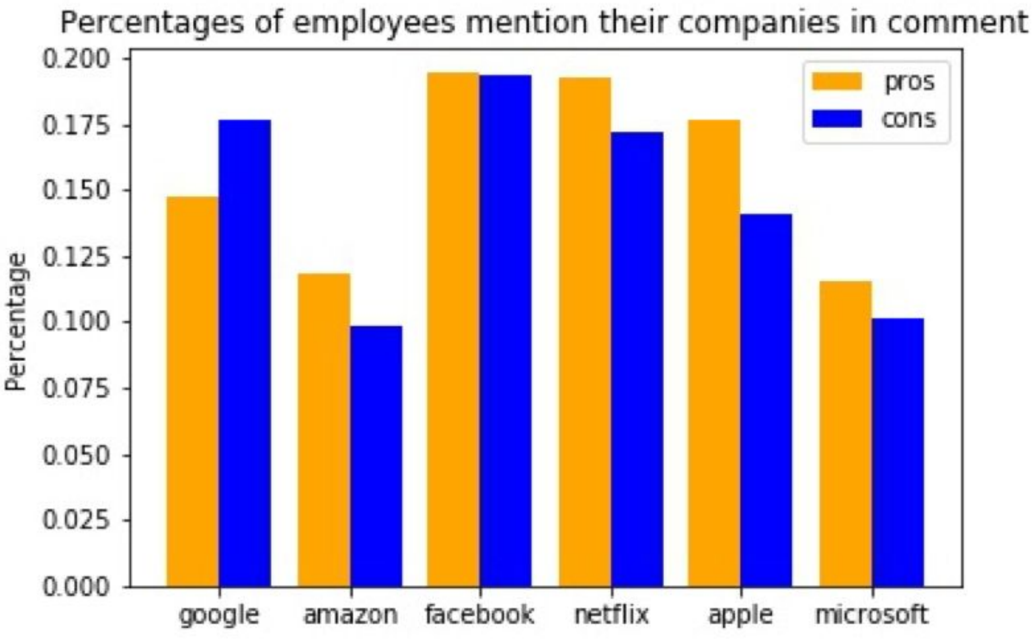
Monthly Income By Age - BRC dataset

- **The similar trend is also observed in the BRC dataset.**
- **BRC's salary distribution is more hierarchical**

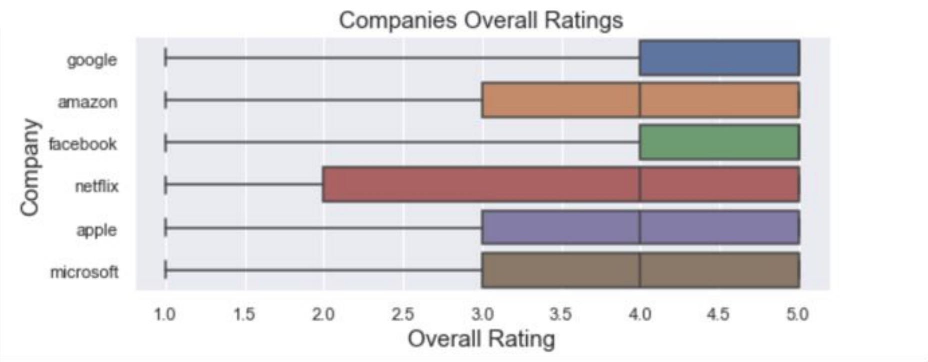
Review Dataset - How do People Mention Their own Companies

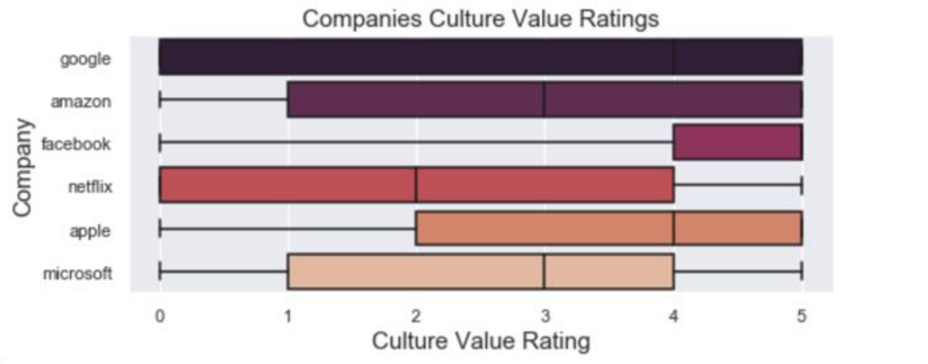
- Employees of Amazon and Microsoft tends not mentioning their companies in reviews

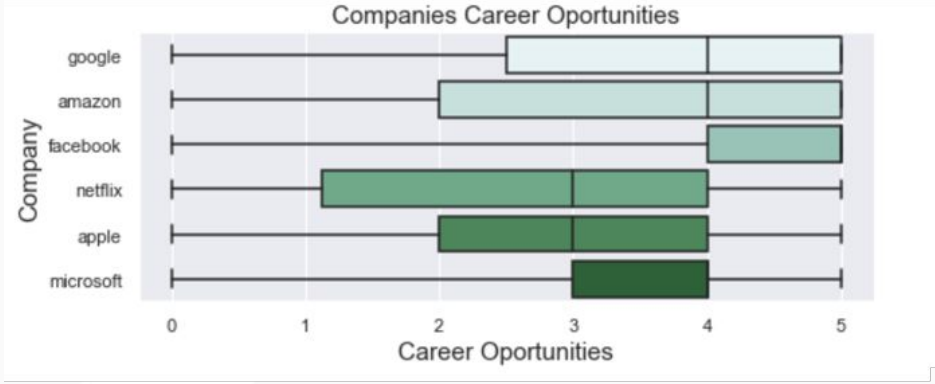
- Different from others, Google' employees are more willing to mention their company when talking about disadvantages



Review Dataset-Ratings on different features







❖ **From most rating range, Facebook might be the best one!**

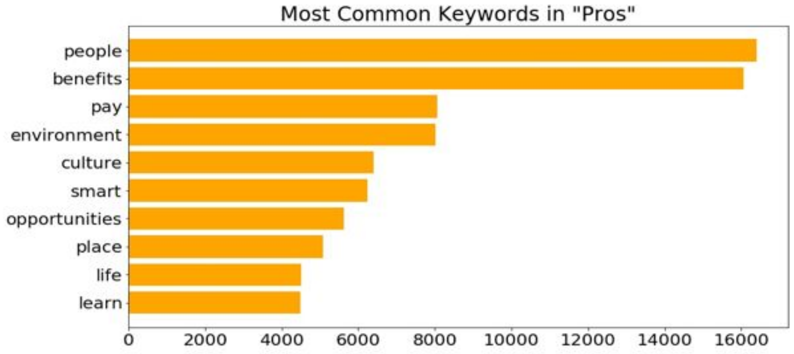
Summary of People’s comments

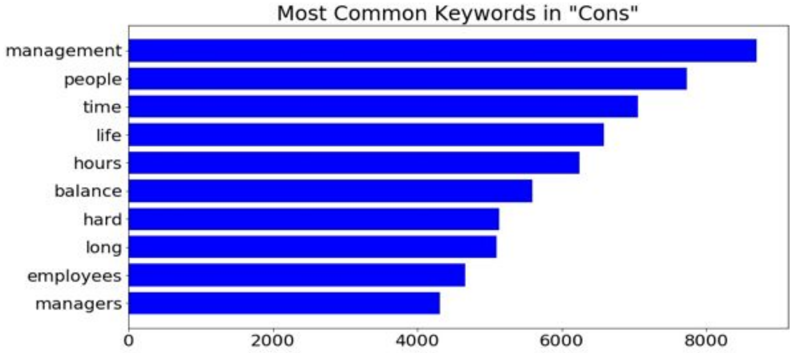
- **Common ‘good’ features:**

Colleague, Benefits, Salary, Environment, Culture, Opportunities, Place, Learn, Life

- **Common ‘bad’ features:**

Management, Colleague, Working Time, Life, Balance, Hard



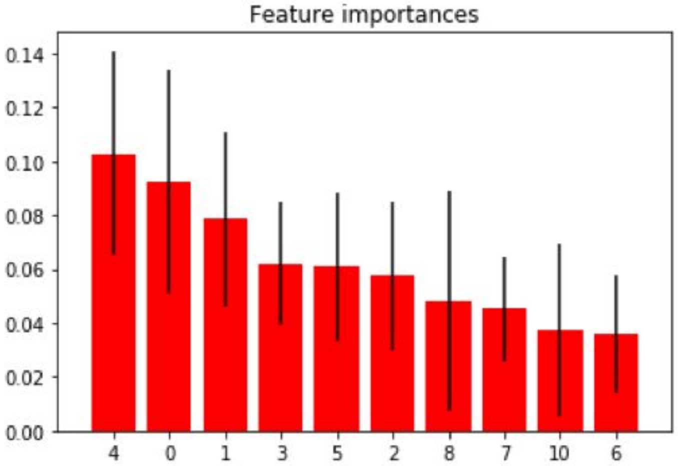


CONTENTS

- Introduction
- Dataset
- Analysis
- **Prediction**
- Conclusion

Feature Importance via Extra Decision Tree Classifier

```
Feature ranking:
1. feature 4 - MonthlyIncome (0.102904)
2. feature 0 - Age (0.092525)
3. feature 1 - DailyRate (0.078458)
4. feature 3 - HourlyRate (0.062046)
5. feature 5 - MonthlyRate (0.060752)
6. feature 2 - DistanceFromHome (0.057355)
7. feature 8 - TotalWorkingYears (0.048040)
8. feature 7 - PercentSalaryHike (0.045227)
9. feature 10 - YearsAtCompany (0.037490)
10. feature 6 - NumCompaniesWorked (0.035893)
```



- Performance analysis on the IBM dataset using extremely randomized trees.
- Features ranked on the basis of Gini importance which counts the number of times a feature is used to split a node weighted by the number of samples it splits

CONTENTS

CONCLUSION

- Contribution 1 — Major reasons to quit
- Contribution 1 — Results given by machine learning model
- Contribution 1 — Measures to be taken to prevent attrition

THANK YOU!

Q & A