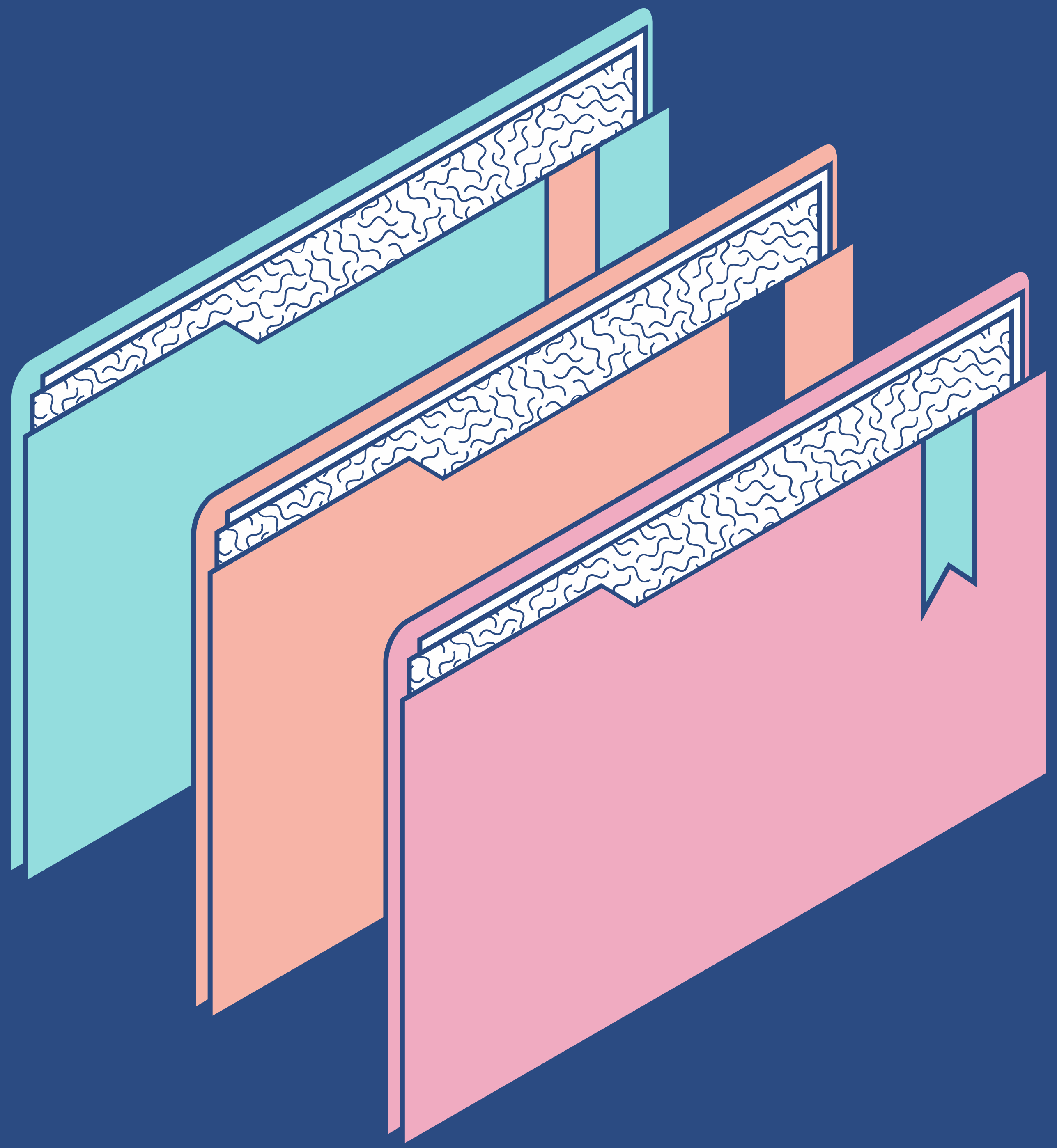
Tackling Talent Retention at IBM

HR Analytics using Supervised and Unsupervised Machine Learning in R

Marcus Loke, Jisu Baek, Yuzhe Sun, Julia Ju



Agenda

KEY TOPICS DISCUSSED IN THIS PRESENTATION

- Problem Statement + Aim of Project
- Research Questions
- Methodology + Results
- Discussion
- Recommendations + Limitations

Problem Statement

WHY DO SO MANY TECH EMPLOYEES LEAVE?

In 2018, turnover in tech industry was the highest at 13.2%

As compared to other industries like Government/Education (11.2%) and Financial Services (10.8%)

From 2012 to 2020, IBM had a reduction of 20% in its workforce

This does not bode well when talent retention is key to driving revenue growth

Employees leave for a myriad of reasons

Job fit, pay satisfaction, career development, etc.

Repercussions of Attrition

Slow the business and productivity losses

If a software developer leaves, it takes 43 days on average to hire a new one (approx. 1.5 months of productivity loss)

Loss of intellectual capital

Creates bottlenecks

Revenue loss

Costs around US$33K for each employee that leaves

Impact on workplace culture

Reduces morale of the team

Aim of our project

REDUCE ATTRITION IN IBM BY:

1. Using ML to predict attrition
2. Uncovering key factors that lead to attrition
3. Characterizing "high-risk" employees for targeted retention strategies
4. Make recommendations that are amenable to experimentation

Research Questions

1

What are the key driving factors influencing attrition the most at IBM?

Having such insights would allow us to create watch-areas in IBM

2

Who is likely to leave IBM?

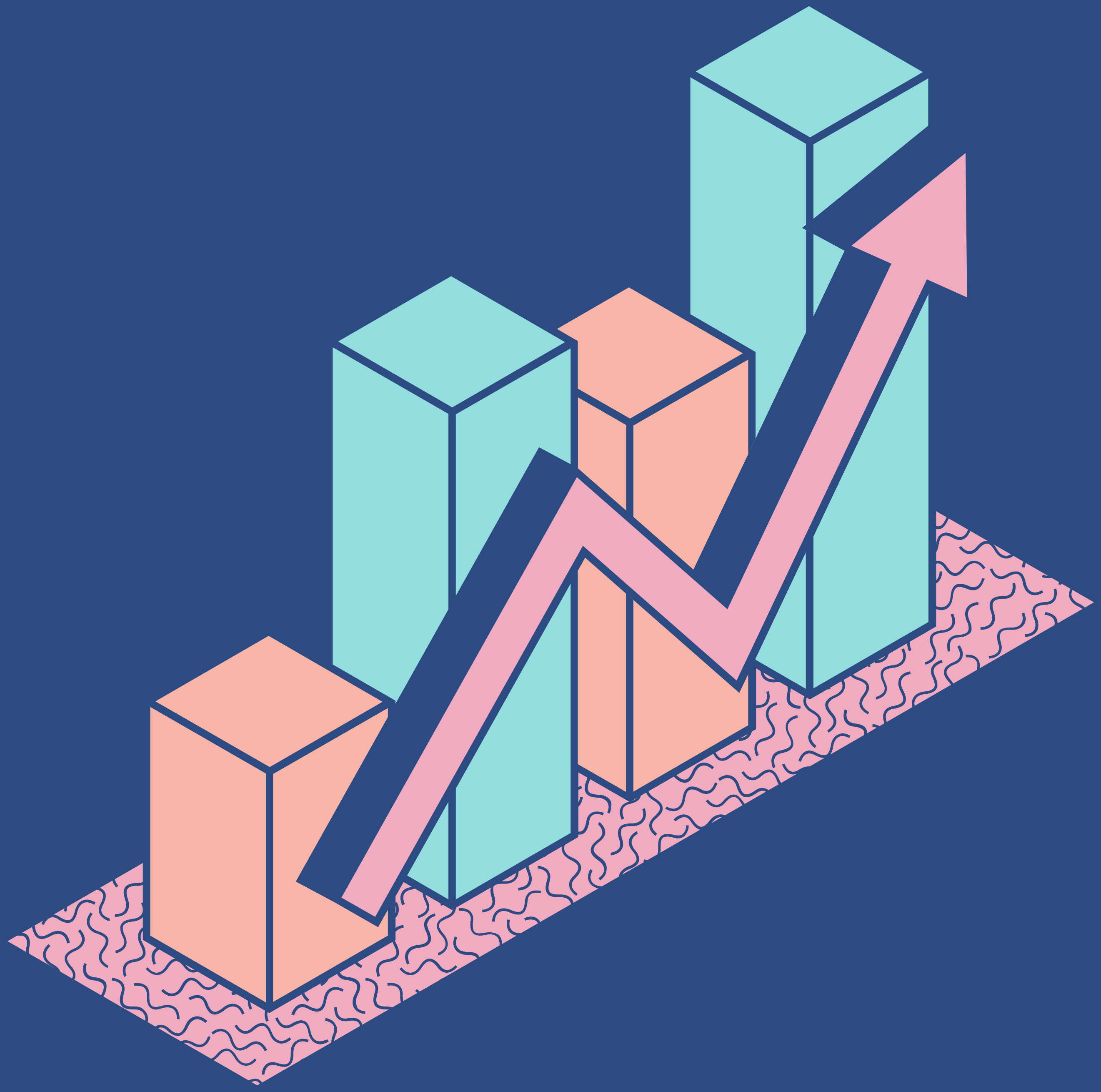
This prediction problem would allow us to identify talents who at at risk of leaving

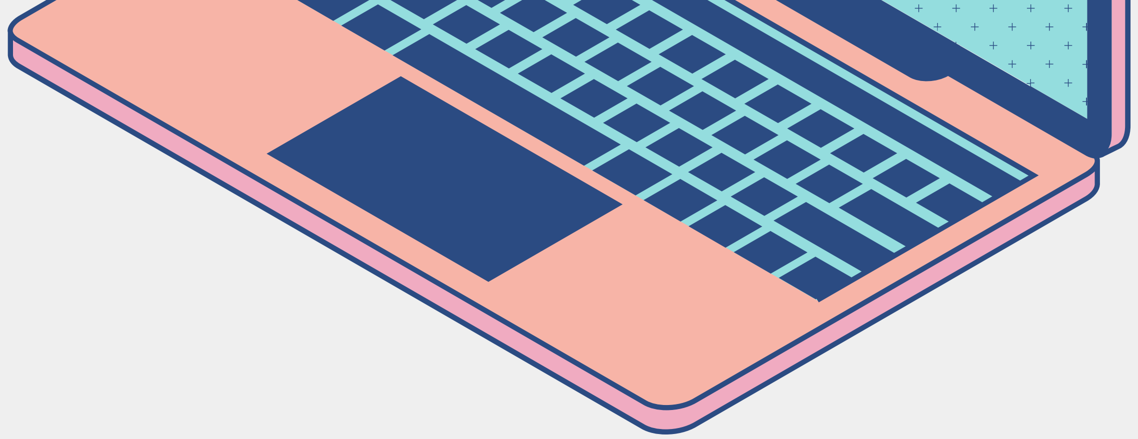
3

What is the employee type that has the highest tendency to leave IBM?

Characterize and personify these "high-risk" individuals to allow better understanding

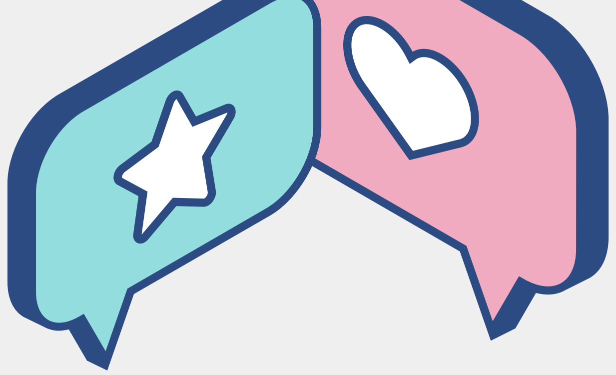
Methodology + Results





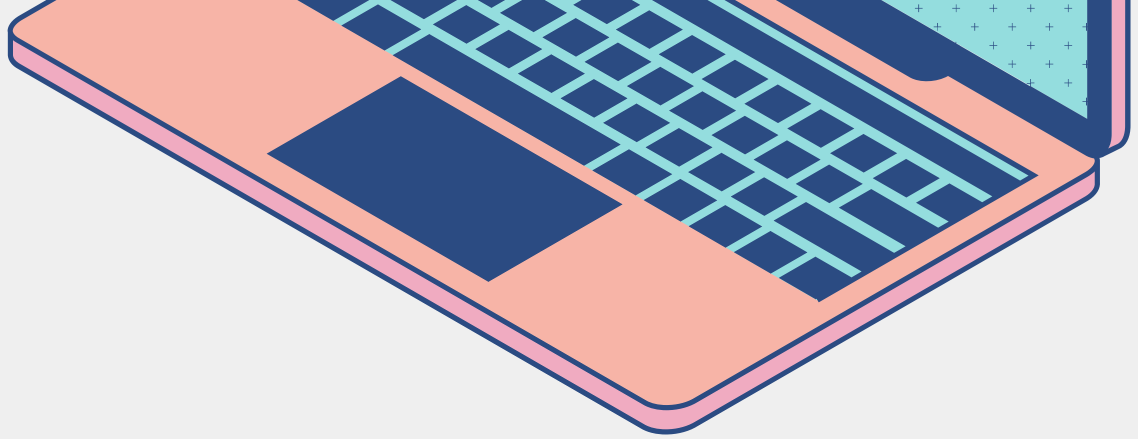
IBM Internal HR Data

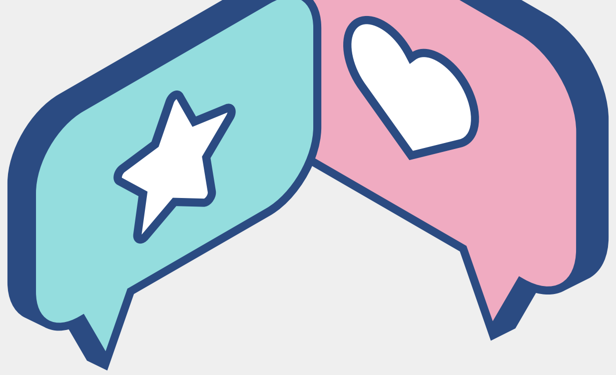
- Contains employee information such as gender, monthly salary, department, attrition status, etc.
- 32 variables
- Outcome variable: Attrition
- We are able to perform prediction modeling using this dataset



Glassdoor Text Reviews

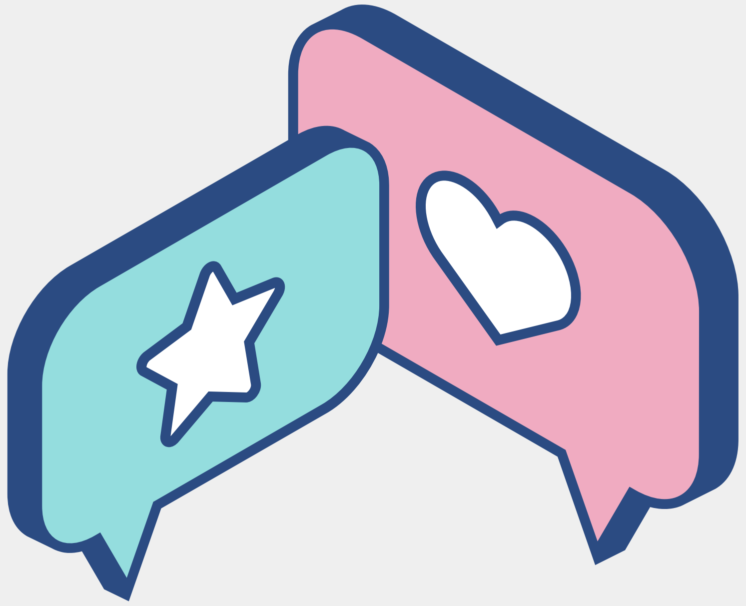
- Contains text reviews from past and present employees of IBM, their roles, etc.
- 8 variables
- How can we make use of the text reviews to augment our prediction modeling?





IBM Internal HR Data + Glassdoor Text Reviews

- The main idea is to use sentiment scores in the text reviews as a predictor in the model
- Compute sentiment scores for each role in the reviews
- Join both datasets based on roles
- We also performed clustering on the IBM dataset to see if it improves the model accuracy



So many different types of roles!

```
> unique(dat_gd$Role)
 [1] "19 Feb 2021 - Executive"
 [2] "26 Aug 2014 - Advisory Engineer"
 [3] "4 Jun 2020 - Bid Proposal Manager"
 [4] "21 May 2021 - Applications Developer"
 [5] "2 May 2021 - Technical Writer"
 [6] "18 May 2021 - Project Manager"
 [7] "26 May 2021 - Graphics Manager"
 [8] "3 Mar 2021 - Content Director"
 [9] "30 May 2021 - Software Developer"
[10] "30 May 2021 - CBD Consultant"
[11] "28 Apr 2021 - User Experience Designer"
[12] "30 May 2021 - Systems Engineer"
[13] "30 May 2021 - Administrative"
[14] "28 May 2021 - Software Development Manager"
[15] "24 May 2021 - VP-HR"
[16] "23 May 2021 - Computer Programmer"
[17] "18 May 2021 - User Experience Design Lead"
[18] "19 Apr 2021 - Partner"
[19] "30 May 2021 - Country Manager"
[20] "28 May 2021 - Data Center Technician III"
[21] "24 Feb 2021 - Client Technical Specialist"
[22] "5 Apr 2021 - CyberSecurity Engineer"
```

And many more...

Before we conduct the sentiment analysis...

- We will categorize the roles into 6 different role categories:
  - AESP (Assistant Engineering & Scientific Personnel)
  - Corporate
  - Director
  - ESP (Engineering & Scientific Personnel)
  - Manager
  - Sales

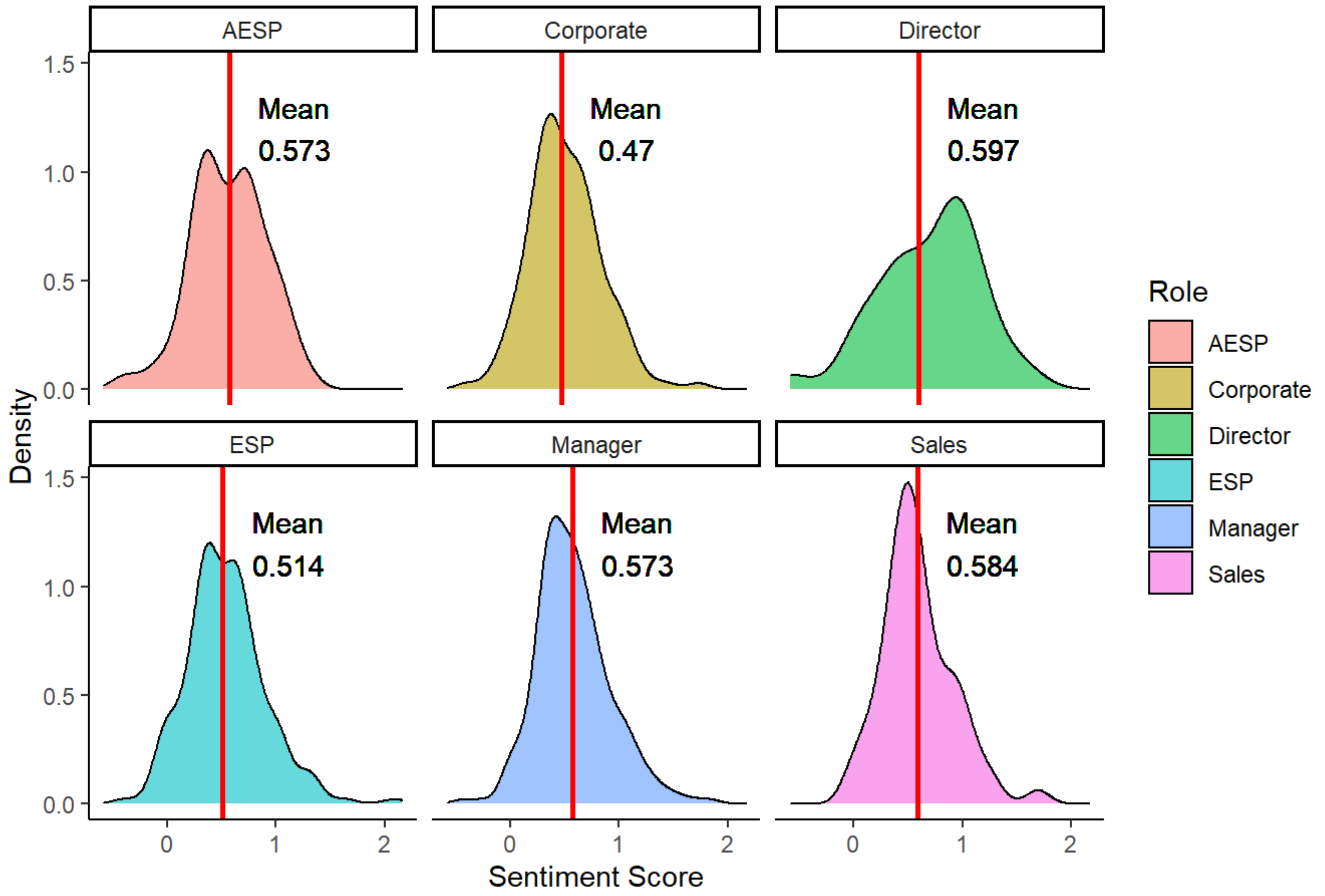
- The goal is to have an aggregated sentiment score for each role

Sentiment Analysis

Pros

```
##          Role word_count        sd ave_sentiment
## 1:       AESP        871 0.3842475     0.5730397
## 2:  Corporate       2186 0.3597134     0.4696195
## 3:   Director        642 0.5245235     0.5968319
## 4:        ESP       6274 0.3822417     0.5135917
## 5:    Manager       3470 0.3515144     0.5727328
## 6:      Sales        740 0.3391197     0.5839903
```

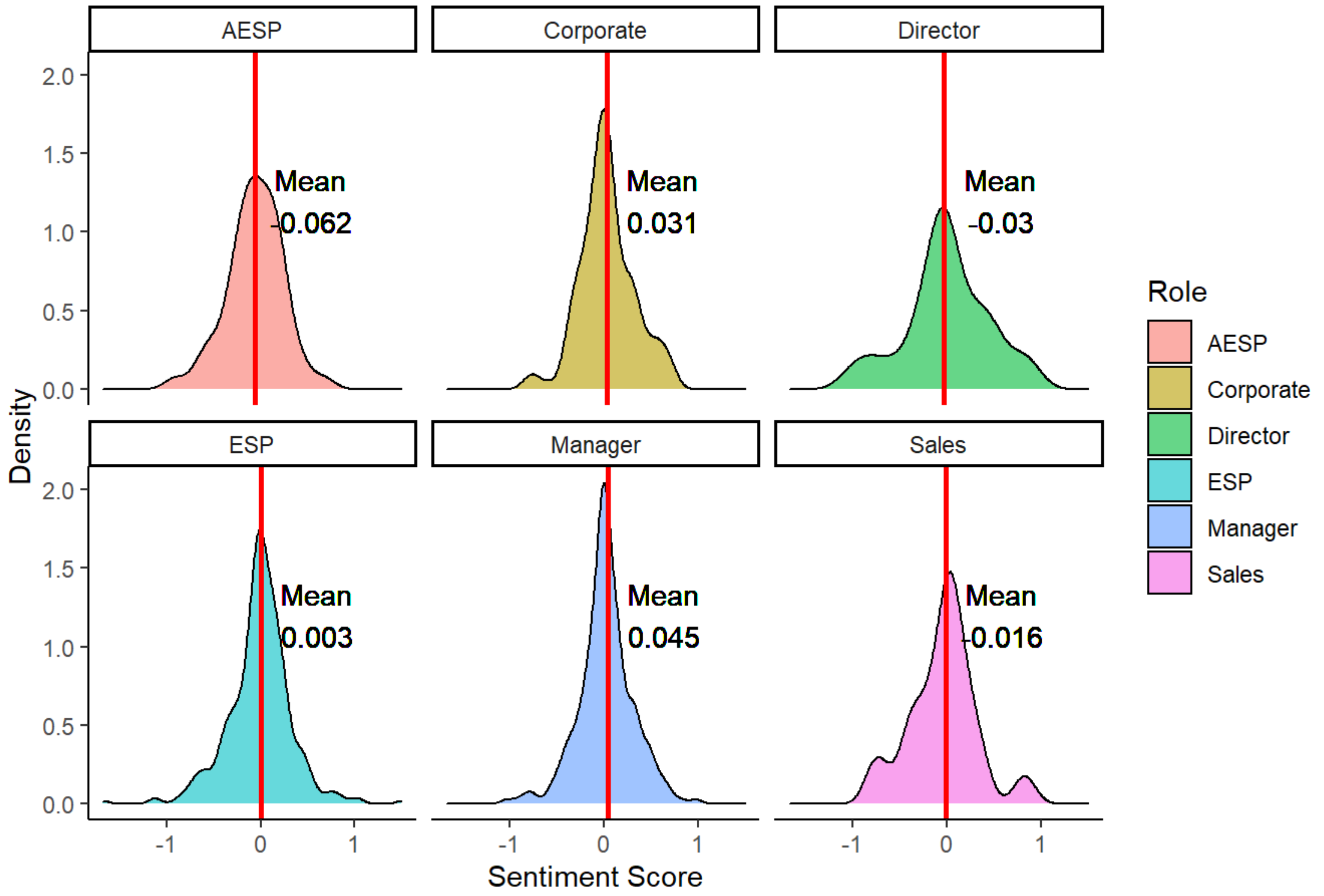
Distribution of Sentiment Scores for Reviews in Pros



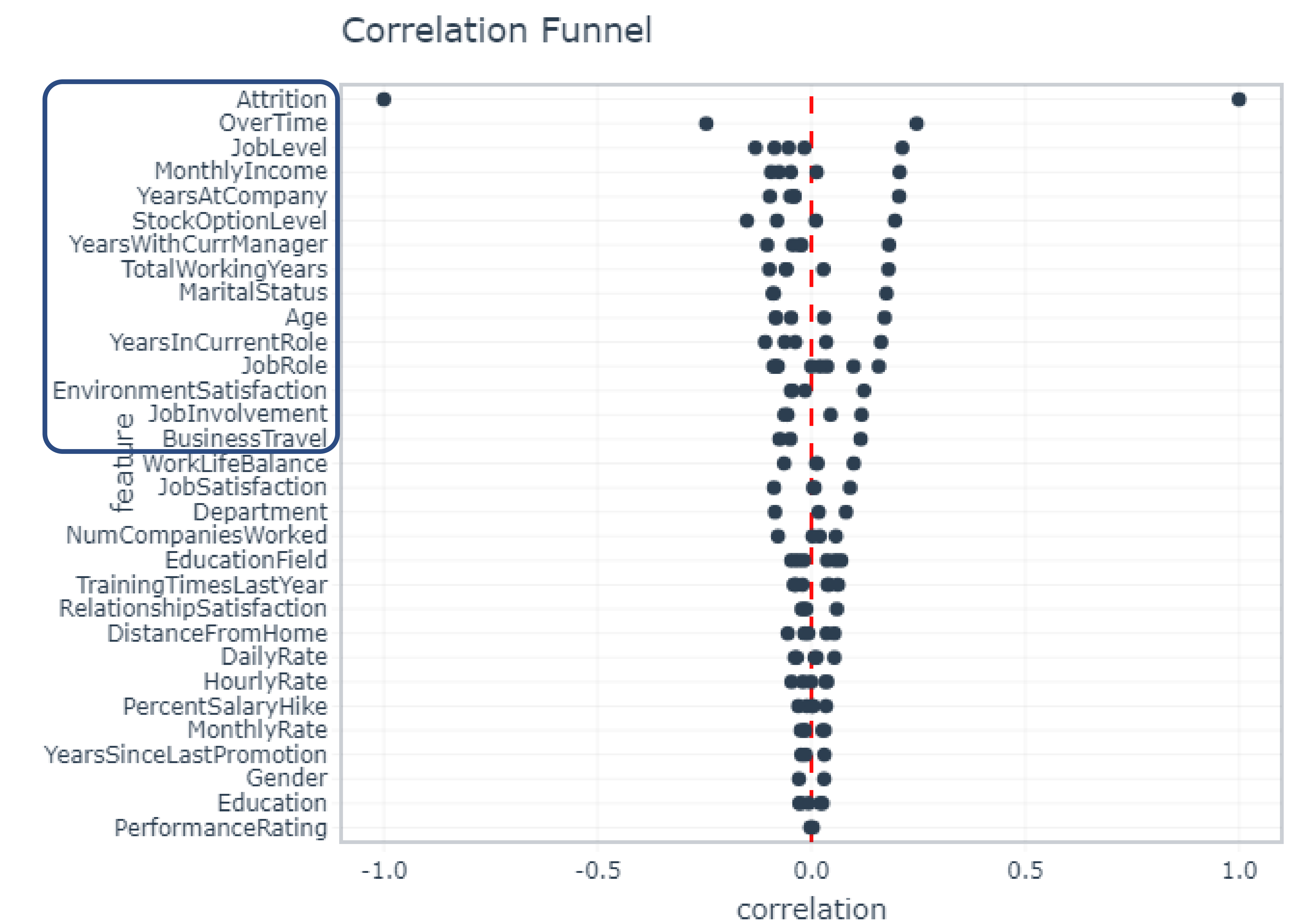
Cons

```
##          Role word_count        sd ave_sentiment
## 1:       AESP        960 0.3226684  -0.061601871
## 2:  Corporate       2392 0.3291854   0.031341807
## 3:   Director        664 0.3920271  -0.030371233
## 4:        ESP       7116 0.3493464   0.002896978
## 5:    Manager       4079 0.3316412   0.044559058
## 6:      Sales       1310 0.3581594  -0.015564249
```

Distribution of Sentiment Scores for Reviews in Cons

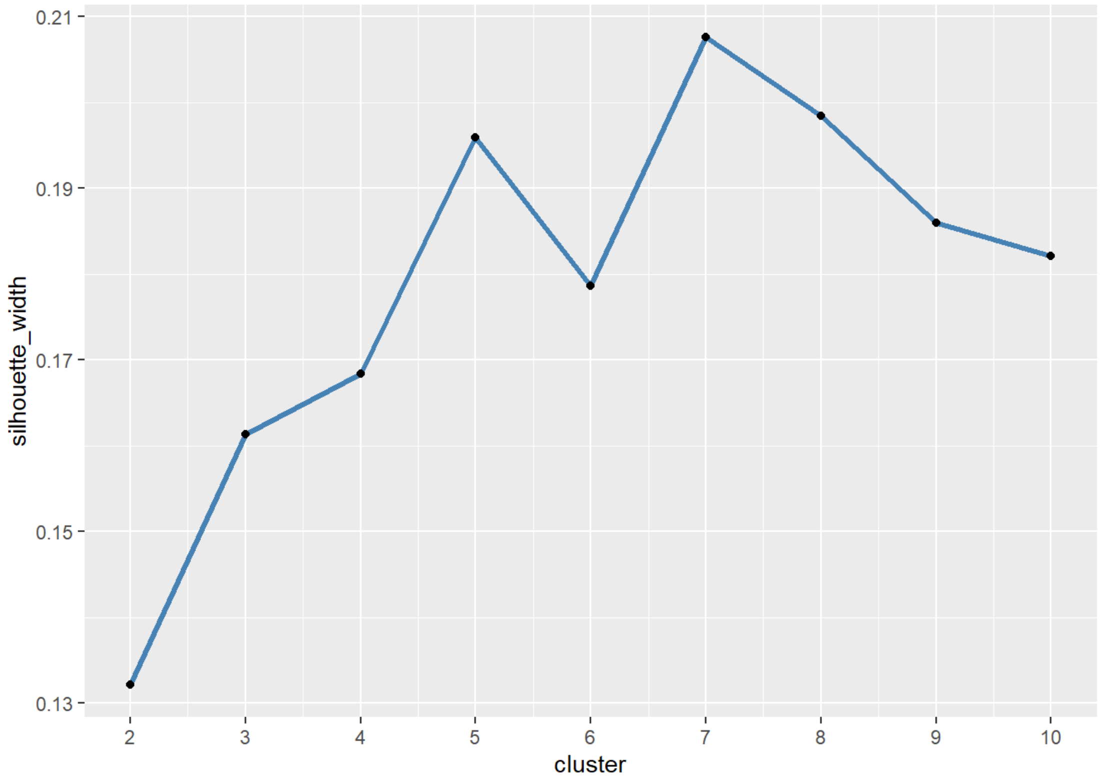


Clustering on IBM Dataset



Methodology

- Cluster the dataset based on variables that are highly correlated with Attrition
- We decided to select variables that had >0.1 in correlation for the clustering analysis (ended up with 14 variables)
- We used the Gower Distance for the distance matrix as the variables had both continuous and ordinal data types



7-Cluster Solution

- Silhouette plot suggests a 7-cluster solution
- The medoids show the "exemplary" employee for each cluster
- Employee in cluster 3 is risky

```
##      EmployeeNumber Attrition OverTime JobLevel MonthlyIncome YearsAtCompany
## 463             621        No       No        2          5337             10
## 1381           1945        No       No        2          5561              5
## 710             991       Yes      Yes        1          2321              3
## 69               88        No       No        1          2194              3
## 1002           1411        No       No        1          3629              3
## 118             154        No       No        3          9738              9
## 700             976        No       No        4         17099              9
##      StockOptionLevel YearsWithCurrManager TotalWorkingYears MaritalStatus Age
## 463                 0                    7                10        Single  34
## 1381                1                    4                 6       Married  35
## 710                 0                    2                 4        Single  31
## 69                  1                    2                 5       Married  35
## 1002                0                    2                 8        Single  37
## 118                 1                    8                10       Married  36
## 700                 1                    8                26       Married  52
##      YearsInCurrentRole              JobRole EnvironmentSatisfaction
## 463                   7      Sales Executive                       4
## 1381                  3      Sales Executive                       2
## 710                   2   Research Scientist                       3
## 69                    2   Research Scientist                       2
## 1002                  2 Laboratory Technician                      1
## 118                   7      Sales Executive                       2
## 700                   8              Manager                       4
##      JobInvolvement    BusinessTravel cluster
## 463               4     Travel_Rarely       1
## 1381              3     Travel_Rarely       2
## 710               2        Non-Travel       3
## 69                3 Travel_Frequently       4
## 1002              3     Travel_Rarely       5
## 118               3 Travel_Frequently       6
## 700               3     Travel_Rarely       7
```

Turnover Rate by Clusters

- Approximately 88% of employees in cluster 3 left IBM
- That represents about 52% of attrition in the entire IBM population

```
## # A tibble: 7 x 5
##   cluster Cluster_Turnover_Ra~ Turnover_Count Cluster_Size Population_Turnover_~
##     <int>                <dbl>          <dbl>        <int>                 <dbl>
## 1       1                10.9              27          247                 11.4
## 2       2                 6.37             20          314                  8.44
## 3       3                87.9             123          140                 51.9
## 4       4                 7.39             19          257                  8.02
## 5       5                11.3              24          213                 10.1
## 6       6                12.6              19          151                  8.02
## 7       7                 3.38              5          148                  2.11
```

Prediction Modeling

1

STEP

Logistic Regression with Backward Selection

Baseline model with selected variables from IBM dataset

2

STEP

Logistic Regression with Sentiment Scores (Pros)

Compare with baseline model

3

STEP

Logistic Regression with Sentiment Scores (Cons)

Compare with baseline model

4

STEP

Logistic Regression with Clustering

Compare with baseline model

5

STEP

Logistic Regression with Both Sentiment Scores & Clustering

Spoiler alert: this was the best model!

What is the metric for our model?

- FP: Predicting that an employee <u>would leave</u> but he/she <u>did not</u>

- FN: Predicting that an employee <u>would not leave</u> but he/she <u>did</u>

FN are more detrimental to the organization.

$$Sensitivity = TP/(TP + FN)$$

$$Specificity = TN/(TN + FP)$$

$$Accuracy = (TN + TP)/(TN + TP + FN + FP)$$

Model 5 resulted in the best model because it had the best sensitivity, accuracy and AUC.

##	model	description	auc	accuracy	specificity	sensitivity
## 1	1	logmod with bw select	0.8940799	0.8204545	0.8102981	0.8732394
## 2	2	logmod with senti (pros)	0.8982404	0.8590909	0.8563686	0.8732394
## 3	3	logmod with senti (cons)	0.9003015	0.8613636	0.8590786	0.8732394
## 4	4	logmod with clust	0.8963319	0.8386364	0.8319783	0.8732394
## 5	5	logmod with senti & clust	0.9015611	0.8704545	0.8699187	0.8732394
## 6	6	trees with 10-fold cv	0.6325814	0.8500000	0.9674797	0.2394366

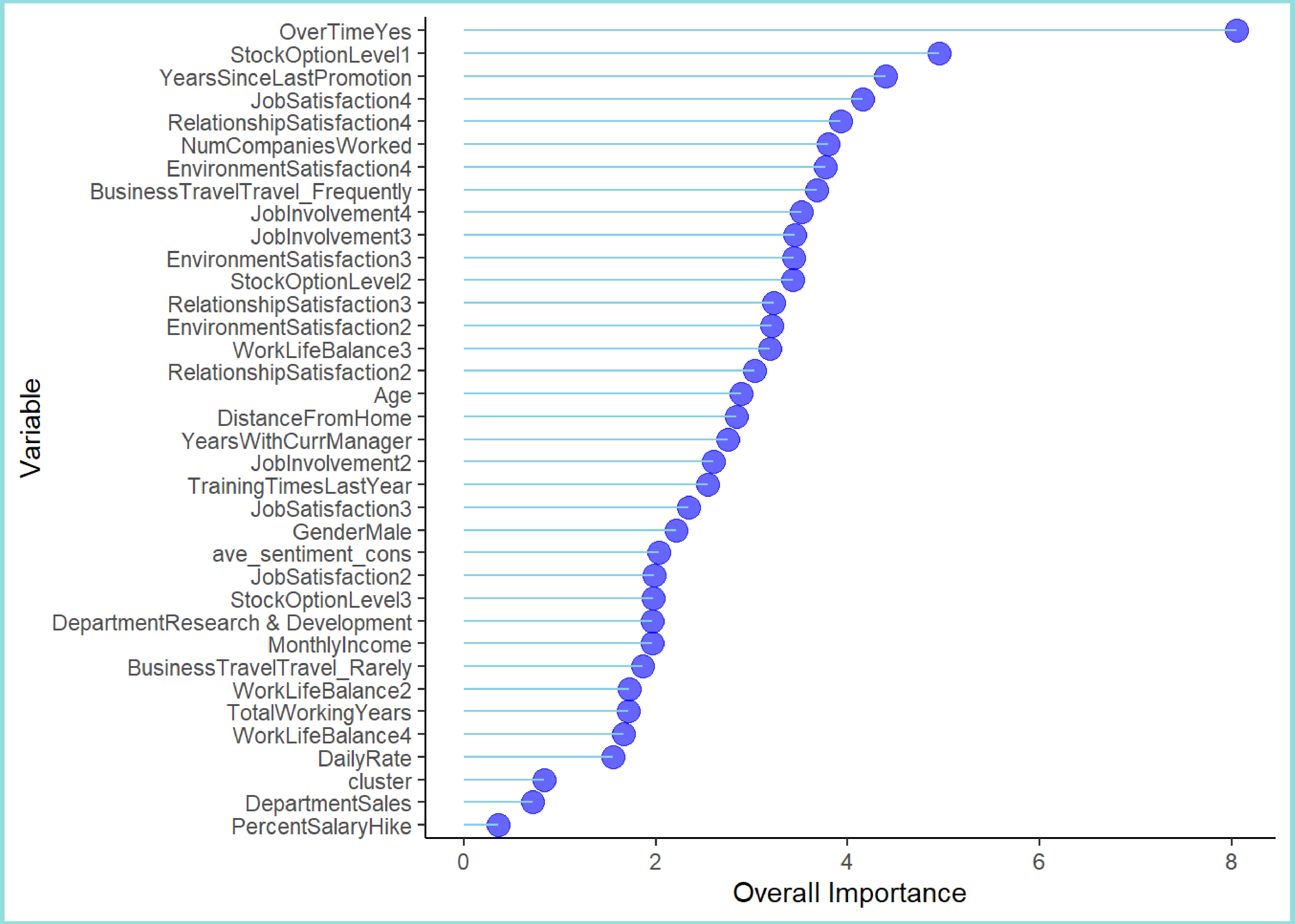
Sentiment scores and clustering were able to improve the prediction accuracy of the baseline model

Discussion

1 ***What are the key driving factors influencing attrition the most at IBM?***

By analyzing the coefficients of the regression model:

- Working overtime would increase the likelihood of attrition by about 6 times
- Having stock options would reduce the likelihood of attrition by 0.3 times
- For each year that an employee is not promoted, there is a 1.2 times higher likelihood of leaving IBM



Discussion

2 ***Who is likely to leave IBM?***

We have created a prediction model that is able to achieve the following on the test set:

- Sensitivity = 87.3%
- Accuracy = 90.2%

```
##   model                  description       auc  accuracy specificity sensitivity
## 1     5 logmod with senti & clust 0.9015611 0.8704545   0.8699187   0.8732394
```

Discussion

3 ***What is the employee type that has the highest tendency to leave IBM?***

- Employees that have the highest risk of leaving are in cluster 3
- Their personas are shown in the table
- We will propose some recommendations to improve the retention rate of this type of employees

```
##      EmployeeNumber Attrition OverTime JobLevel MonthlyIncome YearsAtCompany
## 463             621        No       No        2          5337             10
## 1381           1945        No       No        2          5561              5
## 710             991       Yes      Yes        1          2321              3
## 69               88        No       No        1          2194              3
## 1002           1411        No       No        1          3629              3
## 118             154        No       No        3          9738              9
## 700             976        No       No        4         17099              9
##      StockOptionLevel YearsWithCurrManager TotalWorkingYears MaritalStatus Age
## 463                 0                    7                10        Single  34
## 1381                1                    4                 6       Married  35
## 710                 0                    2                 4        Single  31
## 69                  1                    2                 5       Married  35
## 1002                0                    2                 8        Single  37
## 118                 1                    8                10       Married  36
## 700                 1                    8                26       Married  52
##      YearsInCurrentRole              JobRole EnvironmentSatisfaction
## 463                   7      Sales Executive                       4
## 1381                  3      Sales Executive                       2
## 710                   2   Research Scientist                       3
## 69                    2   Research Scientist                       2
## 1002                  2 Laboratory Technician                      1
## 118                   7      Sales Executive                       2
## 700                   8              Manager                       4
##      JobInvolvement    BusinessTravel cluster
## 463               4     Travel_Rarely       1
## 1381              3     Travel_Rarely       2
## 710               2        Non-Travel       3
## 69                3 Travel_Frequently       4
## 1002              3     Travel_Rarely       5
## 118               3 Travel_Frequently       6
## 700               3     Travel_Rarely       7
```

Recommendations

VARIABLES FROM MODEL #5 (BEST MODEL)

Unable to control

Gender DistanceFromHome Age NumCompaniesWorked TotalWorkingYears

Indirect variables

EnvironmentSatisfaction WorkLifeBalance RelationshipSatisfaction JobSatisfaction

Costly

MonthlyIncome PercentSalaryHike DailyRate YearsSinceLastPromotion

YearsWithCurrManager Department **Minor variables**

What we can control

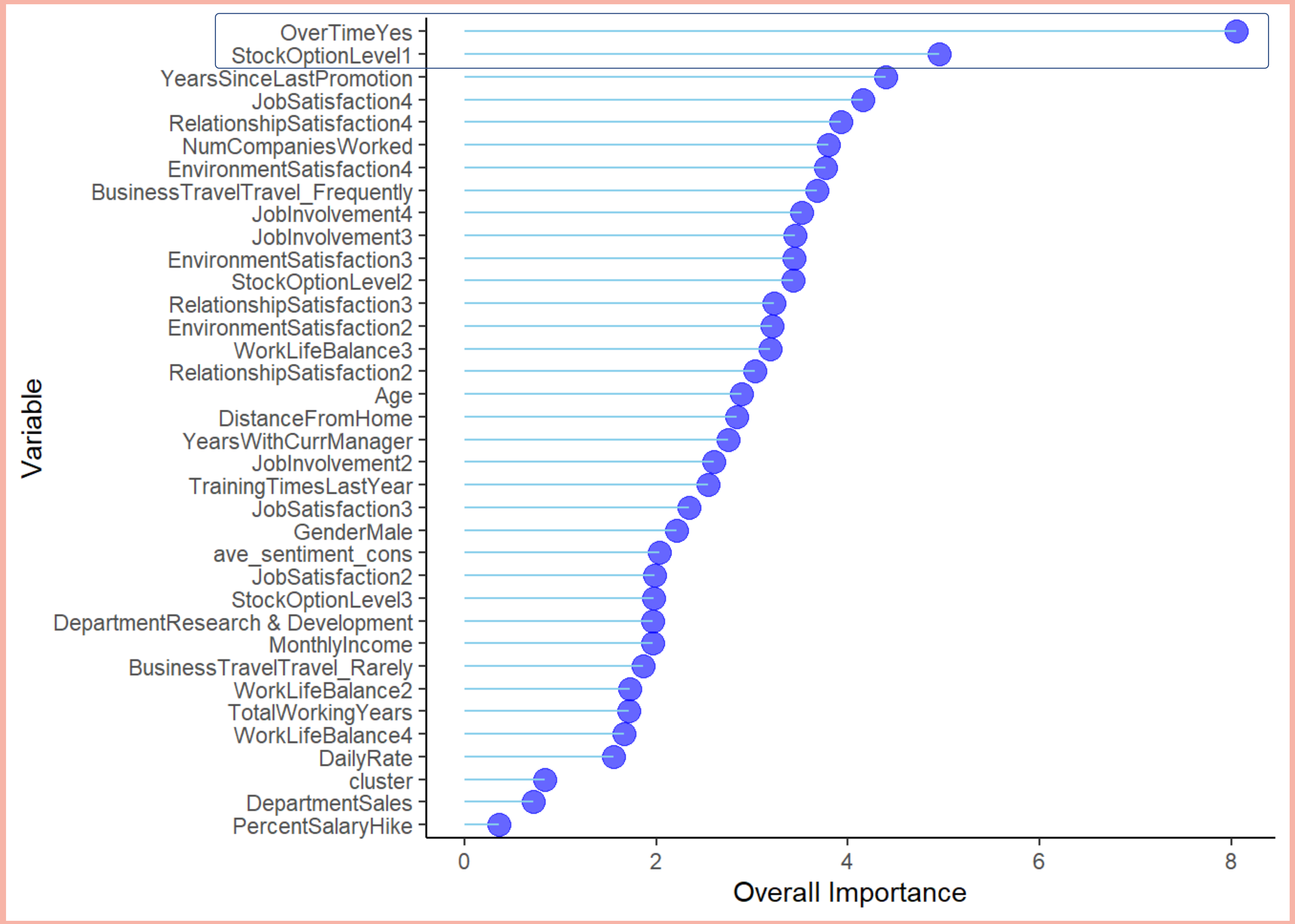
OverTime JobInvolvement TrainingTimesLastYear BusinessTravel StockOptionLevel

Recommendations

VARIABLES FROM MODEL #5 (BEST MODEL)

What we can control

OverTime | **JobInvolvement** | **TrainingTimesLastYear** | **BusinessTravel** | **StockOptionLevel**



Overtime

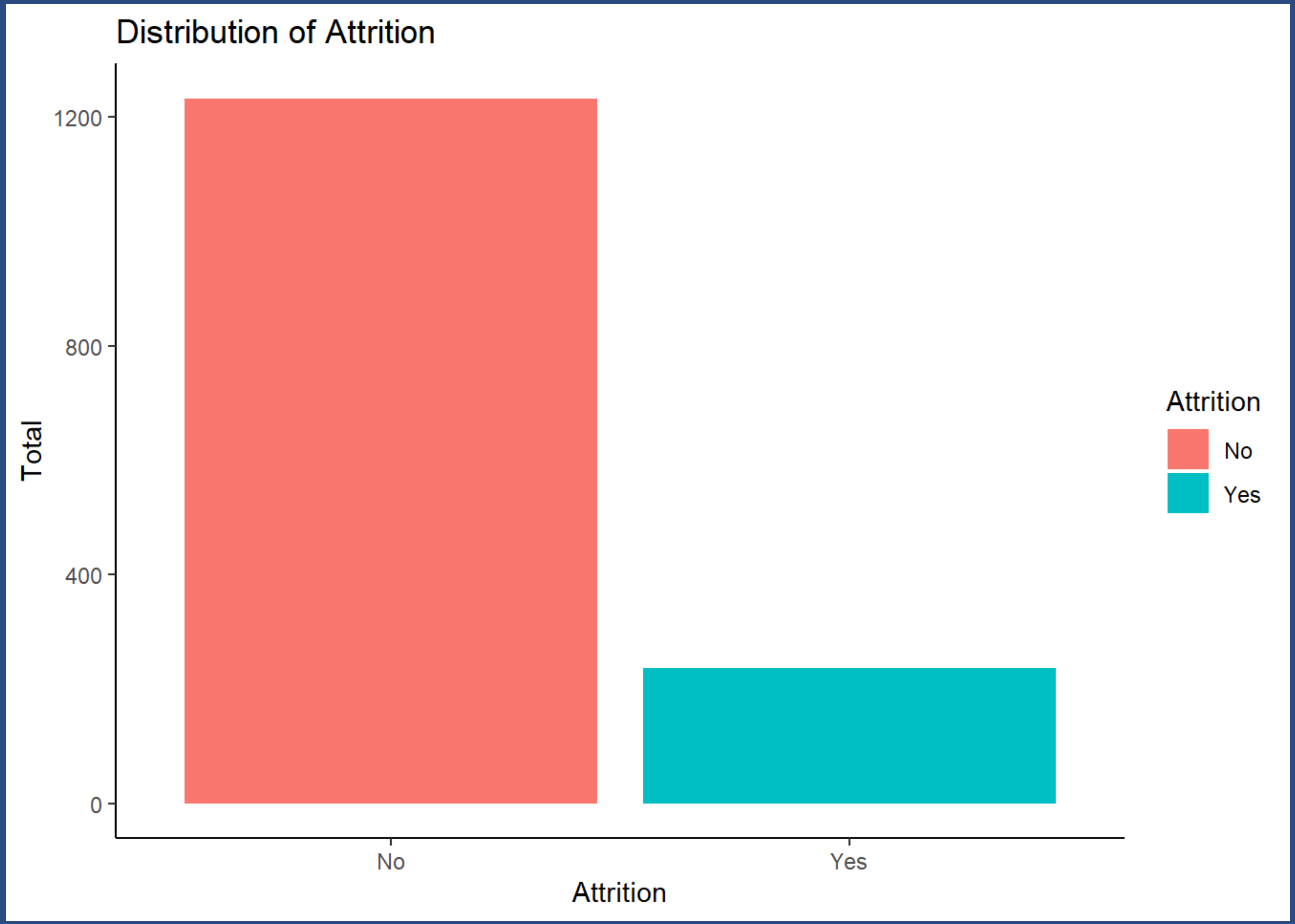
A rescission of overtime culture can have the potential to <u>reduce the likelihood of attrition by 6 times</u> (while holding other variables constant)

Stock Options

Giving stock options to employees can have the potential to <u>reduce the likelihood of attrition by 69%</u> (while holding other variables constant)

Limitations

IMBALANCE IN ATTRITION STATUSES



More stayed than left

Such an imbalance in our train set would result in poorer prediction accuracy in our models

Future works to treat imbalance

- Try to collect more observations on employees who left IBM
- Explore upsampling techniques

Do you have any questions?

Send it to us!

Thank you for listening!