HR data analysis and contract termination prediction

Gabriel Scozzarro

1.0 Introduction

2.0 Preparation and reading data

Data was imported from the set contained inside the file HRDataset_v14.csv

ï..Employee_Name	EmpID	MarriedID	MaritalStatusID	GenderID	EmpStatusID	DeptID	PerfScoreID	FromDiversityJobFairID
Length:311	Min. :10001	Min. :0.0000	Min. :0.0000	Min. :0.0000	Min. :1.000	Min. :1.000	Min. :1.000	Min. :0.00000
Class :character	1st Qu.:10078	1st Qu.:0.0000	1st Qu.:0.0000	1st Qu.:0.0000	1st Qu.:1.000	1st Qu.:5.000	1st Qu.:3.000	1st Qu.:0.00000
Mode :character	Median :10156	Median :0.0000	Median :1.0000	Median :0.0000	Median :1.000	Median :5.000	Median :3.000	Median :0.00000
NA	Mean :10156	Mean :0.3987	Mean :0.8103	Mean :0.4341	Mean :2.392	Mean :4.611	Mean :2.977	Mean :0.09325
NA	3rd Qu.:10234	3rd Qu.:1.0000	3rd Qu.:1.0000	3rd Qu.:1.0000	3rd Qu.:5.000	3rd Qu.:5.000	3rd Qu.:3.000	3rd Qu.:0.00000
NA	Max. :10311	Max. :1.0000	Max. :4.0000	Max. :1.0000	Max. :5.000	Max. :6.000	Max. :4.000	Max. :1.00000
NA	NA	NA	NA	NA	NA	NA	NA	NA

Salary	Termd	PositionID	Position	State	Zip	DOB	Sex	MaritalDesc
Min. : 45046	Min. :0.0000	Min. : 1.00	Length:311	Length:311	Min. : 1013	Length:311	Length:311	Length:311
1st Qu.: 55502	1st Qu.:0.0000	1st Qu.:18.00	Class :character	Class :character	1st Qu.: 1902	Class :character	Class :character	Class :character
Median : 62810	Median :0.0000	Median :19.00	Mode :character	Mode :character	Median : 2132	Mode :character	Mode :character	Mode :character
Mean : 69021	Mean :0.3344	Mean :16.85	NA	NA	Mean : 6555	NA	NA	NA
3rd Qu.: 72036	3rd Qu.:1.0000	3rd Qu.:20.00	NA	NA	3rd Qu.: 2355	NA	NA	NA
Max. :250000	Max. :1.0000	Max. :30.00	NA	NA	Max. :98052	NA	NA	NA
NA	NA	NA	NA	NA	NA	NA	NA	NA

CitizenDesc	HispanicLatino	RaceDesc	DateofHire	DateofTermination	TermReason	EmploymentStatus	Department	ManagerName
Length:311	Length:311	Length:311	Length:311	Length:311	Length:311	Length:311	Length:311	Length:311
Class :character	Class :character	Class :character	Class :character	Class :character	Class :character	Class :character	Class :character	Class :character
Mode :character	Mode :character	Mode :character	Mode :character	Mode :character	Mode :character	Mode :character	Mode :character	Mode :character
NA	NA	NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA	NA	NA	NA

ManagerID	RecruitmentSource	PerformanceScore	EngagementSurvey	EmpSatisfaction	SpecialProjectsCount	LastPerformanceReview_Date	DaysLateLast30	Absences
Min. : 1.00	Length:311	Length:311	Min. :1.12	Min. :1.000	Min. :0.000	Length:311	Min. :0.0000	Min. : 1.00
1st Qu.:10.00	Class :character	Class :character	1st Qu.:3.69	1st Qu.:3.000	1st Qu.:0.000	Class :character	1st Qu.:0.0000	1st Qu.: 5.00
Median :15.00	Mode :character	Mode :character	Median :4.28	Median :4.000	Median :0.000	Mode :character	Median :0.0000	Median :10.00
Mean :14.57	NA	NA	Mean :4.11	Mean :3.891	Mean :1.219	NA	Mean :0.4148	Mean :10.24
3rd Qu.:19.00	NA	NA	3rd Qu.:4.70	3rd Qu.:5.000	3rd Qu.:0.000	NA	3rd Qu.:0.0000	3rd Qu.:15.00
Max. :39.00	NA	NA	Max. :5.00	Max. :5.000	Max. :8.000	NA	Max. :6.0000	Max. :20.00
NA's :8	NA	NA	NA	NA	NA	NA	NA	NA

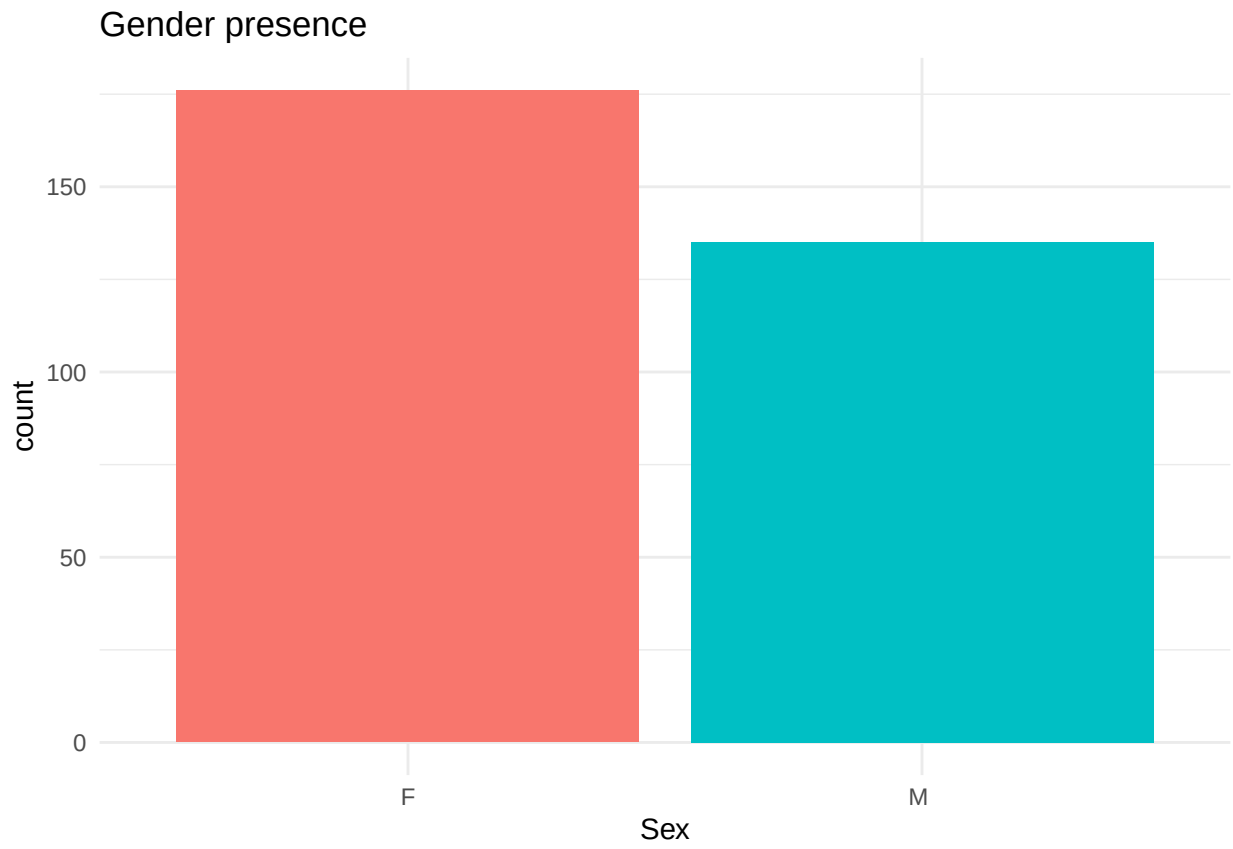
The total amount of employees are 311 and 36 features was collected for each of them.

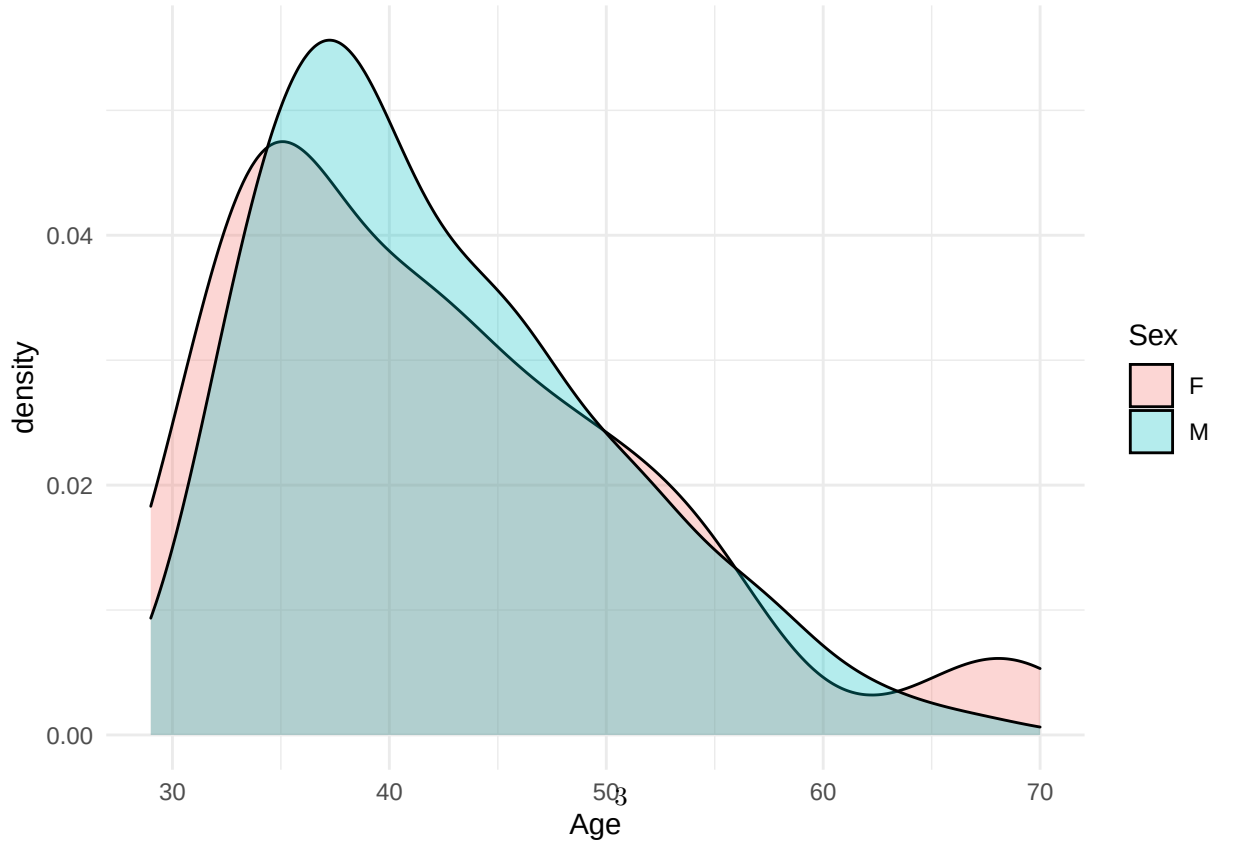
Table 1: Data summary table

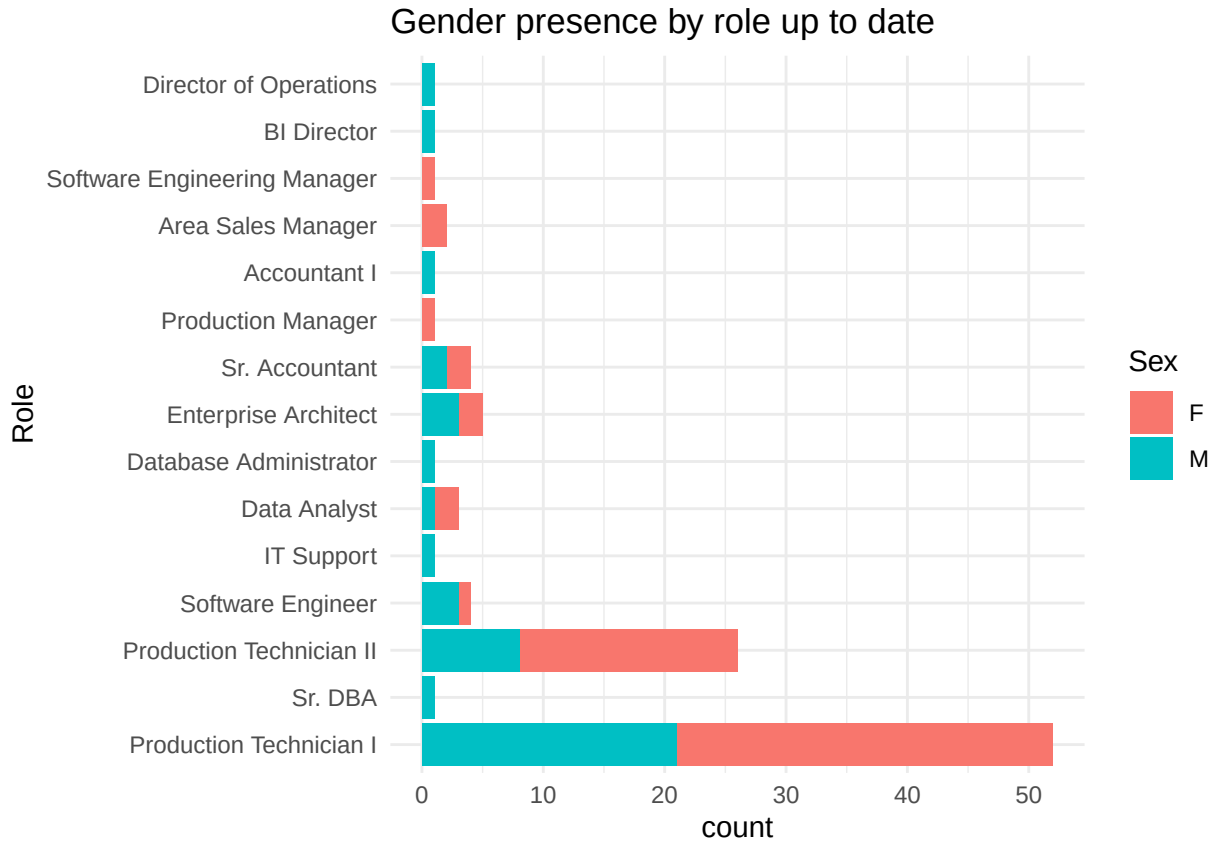
Feature	Description	Type
ï..Employee_Name	Employee's full name	Text
EmpID	Employee ID is unique to each employee	Text
MarriedID	Is the person married (1 or 0 for yes or no)	Binary
MaritalStatusID	Marital status code that matches the text field MaritalDesc	Integer
GenderID	Gender ID that mathces Sex	Binary
EmpStatusID	Employment status code that matches text field EmploymentStatus	Integer
DeptID	Department ID code that matches the department the employee works in	Integer
PerfScoreID	Performance Score code that matches the employee's most recent performance score	Integer
FromDiversityJobFairID	Was the employee sourced from the Diversity job fair? 1 or 0 for yes or no	Binary
Salary	The person's annual pay rate	Float
Termd	Has this employee been terminated - 1 or 0	Binary
PositionID	An integer indicating the person's position	Integer
Position	The text name/title of the position the person has	Text
State	The state that the person lives in	Text
Zip	The zip code for the employee	Text
DOB	Date of Birth for the employee	Date
Sex	Sex - M or F	Text
MaritalDesc	The marital status of the person (divorced, single, widowed, separated, etc)	Text
CitizenDesc	Label for whether the person is a Citizen or Eligible NonCitizen	Text
HispanicLatino	Yes or No field for whether the employee is Hispanic/Latino	Text
RaceDesc	Description/text of the race the person identifies with	Text
DateofHire	Date the person was hired	Date
DateofTermination	Date the person was terminated, only populated if, in fact, Termd = 1	Date
TermReason	A text reason / description for why the person was terminated	Text
EmploymentStatus	A description/category of the person's employment status. Anyone currently working full time = Active	Text
Department	Name of the department that the person works in	Text
ManagerName	The name of the person's immediate manager	Text
ManagerID	A unique identifier for each manager	Integer
RecruitmentSource	The name of the recruitment source where the employee was recruited from	Text
PerformanceScore	Performance Score text/category (Fully Meets, Partially Meets, PIP, Exceeds)	Text
EngagementSurvey	Results from the last engagement survey, managed by our external partner	Float
EmpSatisfaction	A basic satisfaction score between 1 and 5, as reported on a recent employee satisfaction survey	Integer
SpecialProjectsCount	The number of special projects that the employee worked on during the last 6 months	Integer
LastPerformanceReview_Date	The most recent date of the person's last performance review	Date
DaysLateLast30	The number of times that the employee was late to work during the last 30 days	Integer
Absences	The number of times the employee was absent from work	Integer

3.0 EDA

3.1 Gender analysis

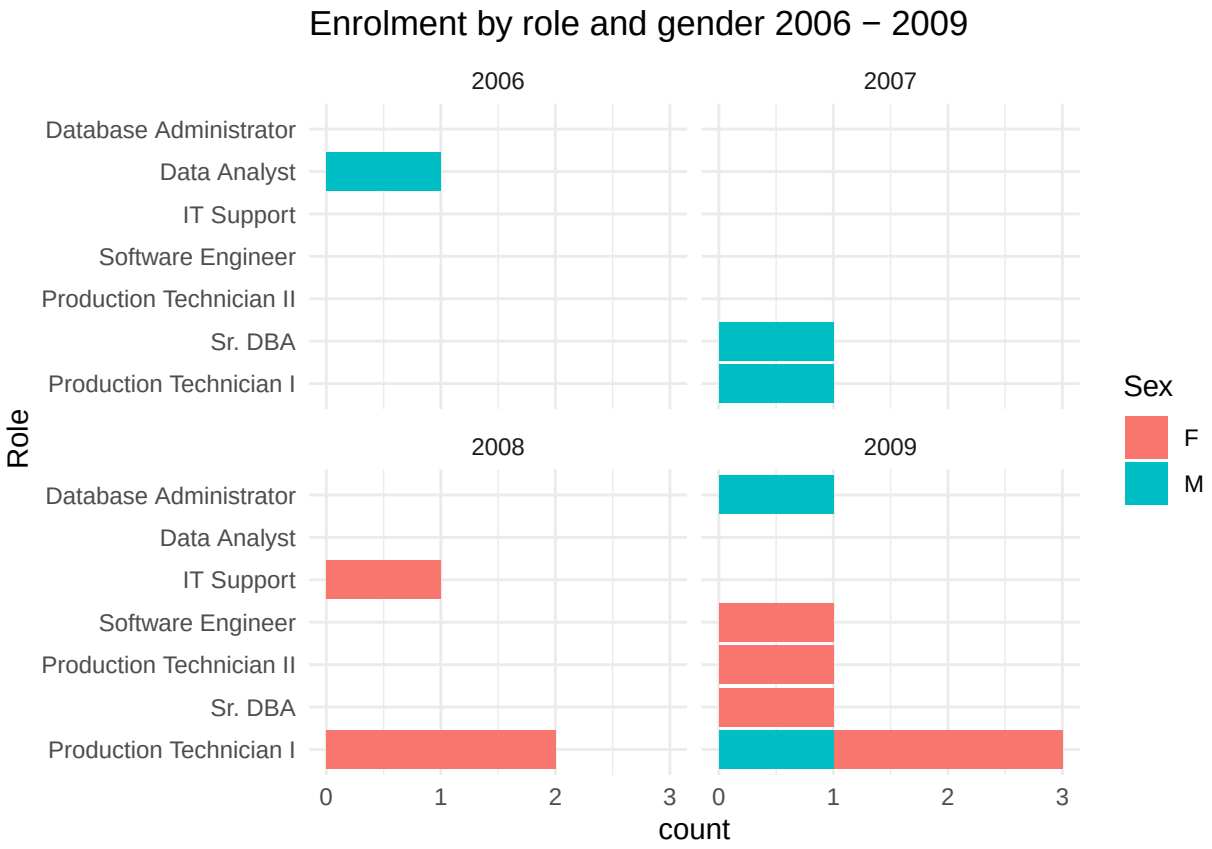


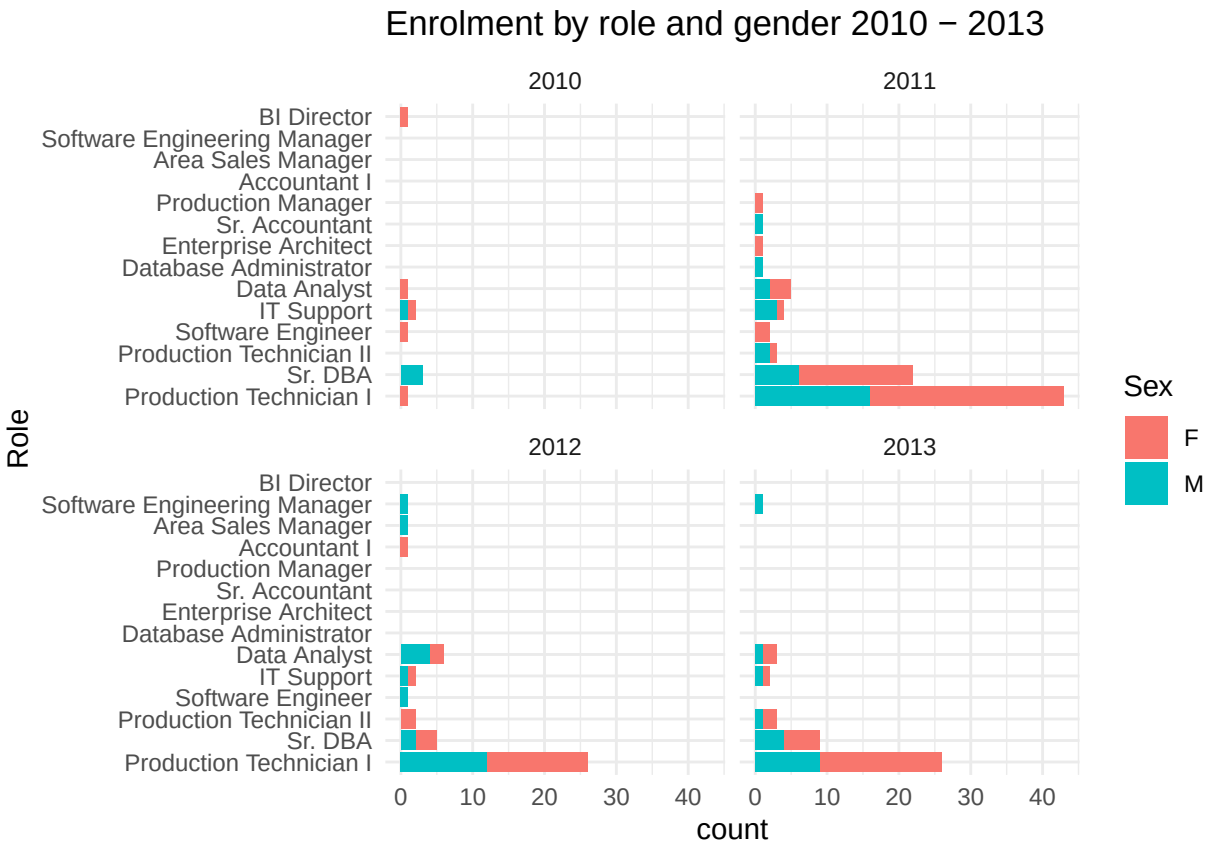


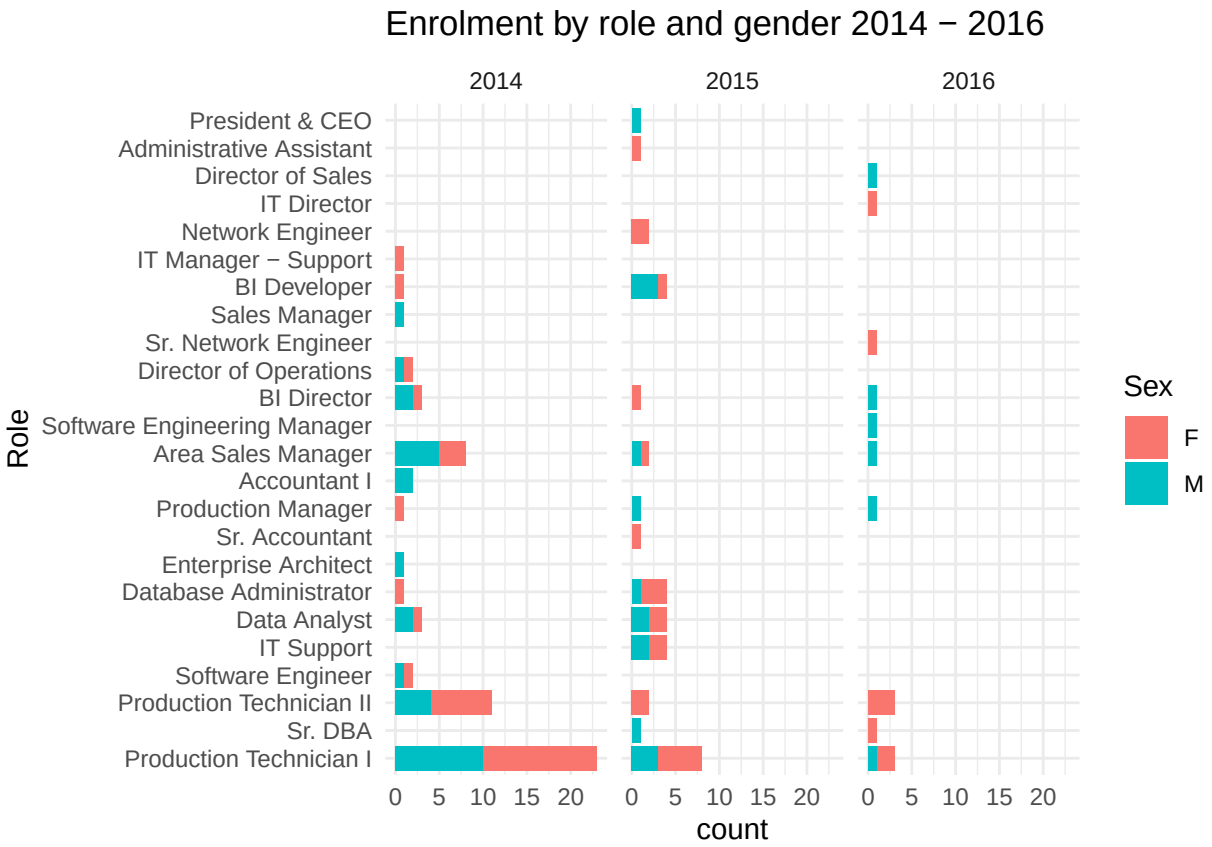


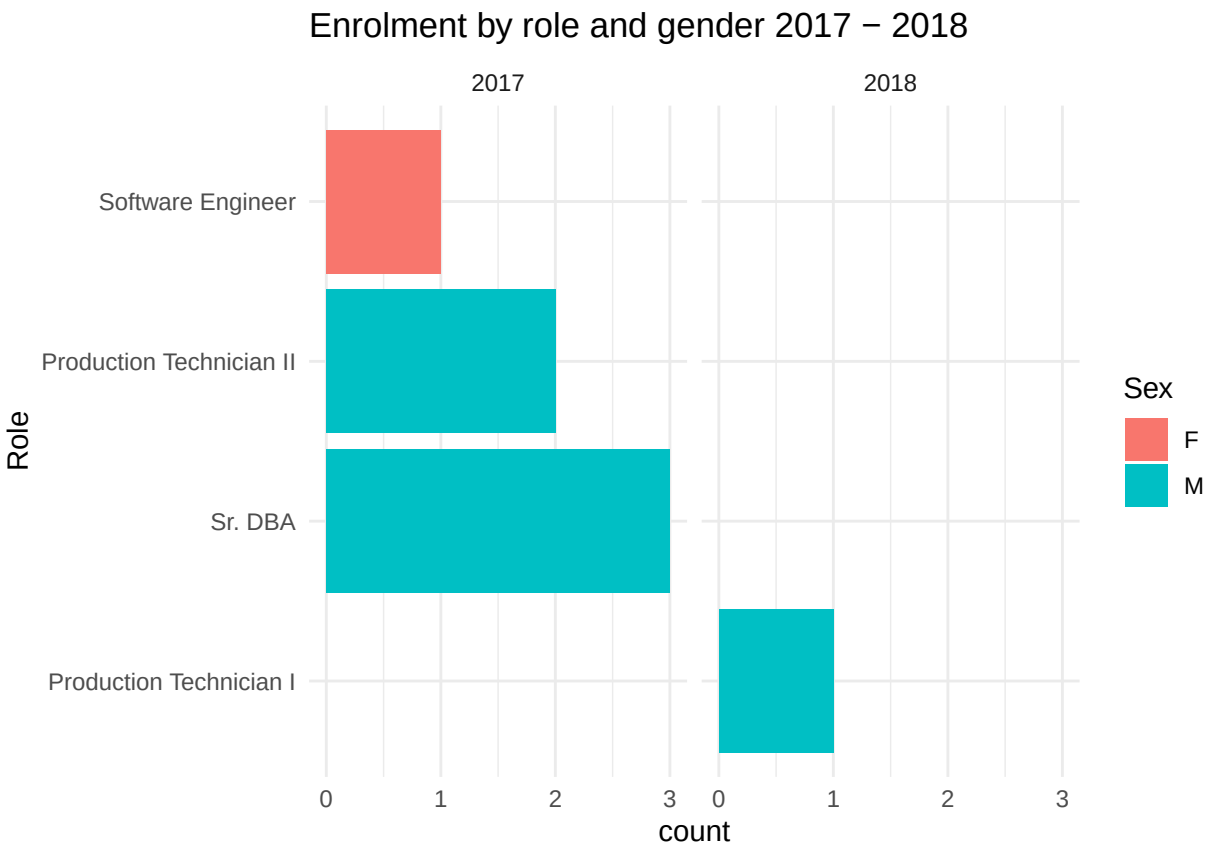
The company has a majority presence of women (176) compare to men (135) with a very close average age of respectively 42.5 for women and 42.3 for men. Regardless the average age we can see that the age distribution is shift to the left for both men and women meaning that the the majority of the employees are younger compare to the average age.

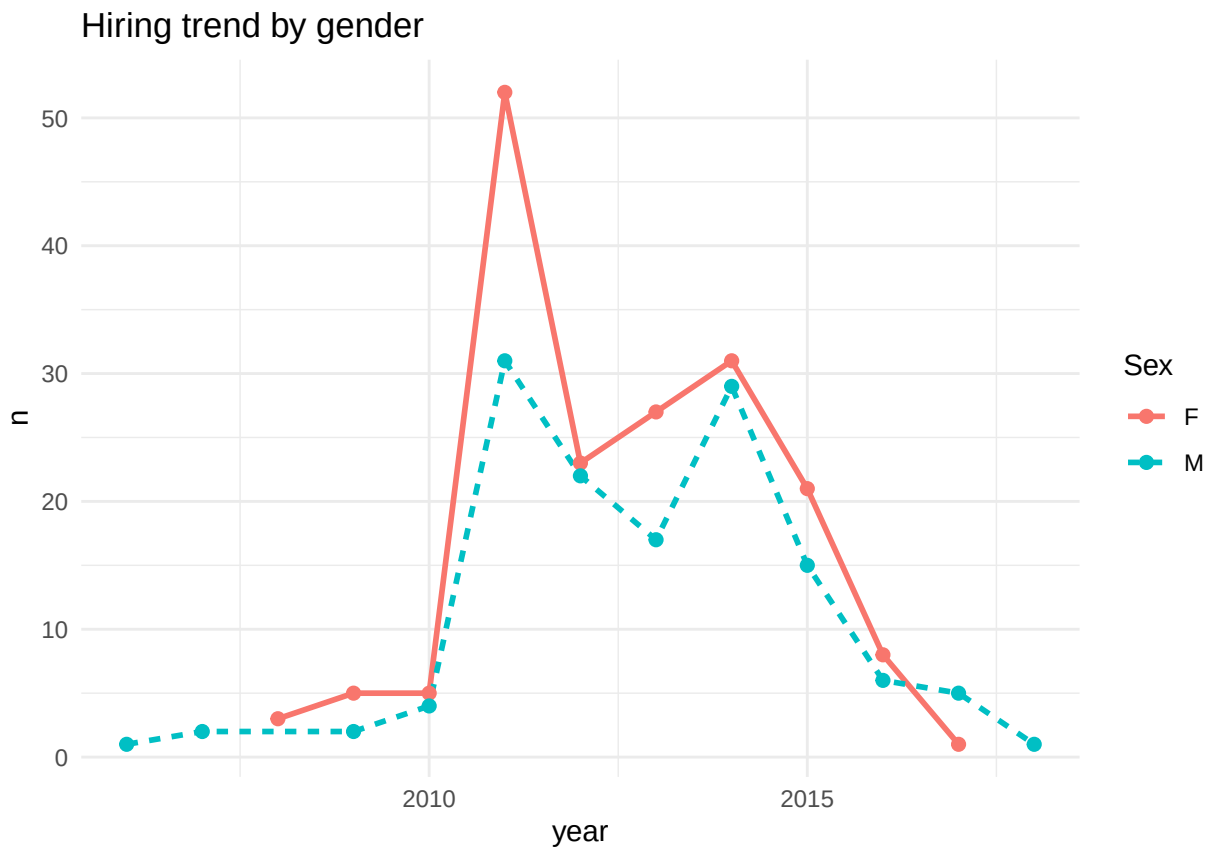
The company has a large an prevalent presence of women employed in the production as Product Technician I and Product Technician II. Managerial roles also has a 100% prevalence of women guiding fundamental aspects such as software engineering, sales and production.

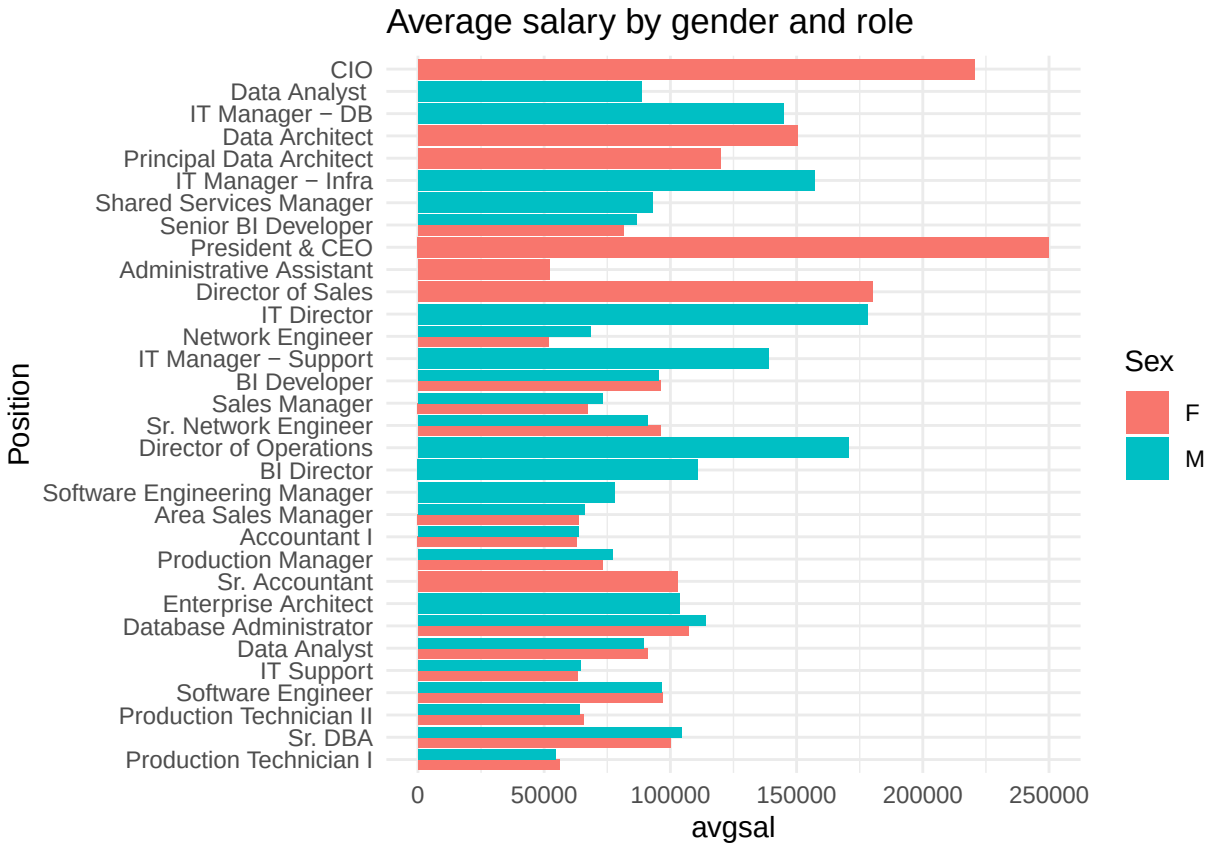

Enrolment by role and gender 2006 − 2009
2006
2007
2008
2009
Database Administrator
Data Analyst
IT Support
Software Engineer
Production Technician II
Sr. DBA
Production Technician I
Role
0
1
2
3
count
Sex
F
M

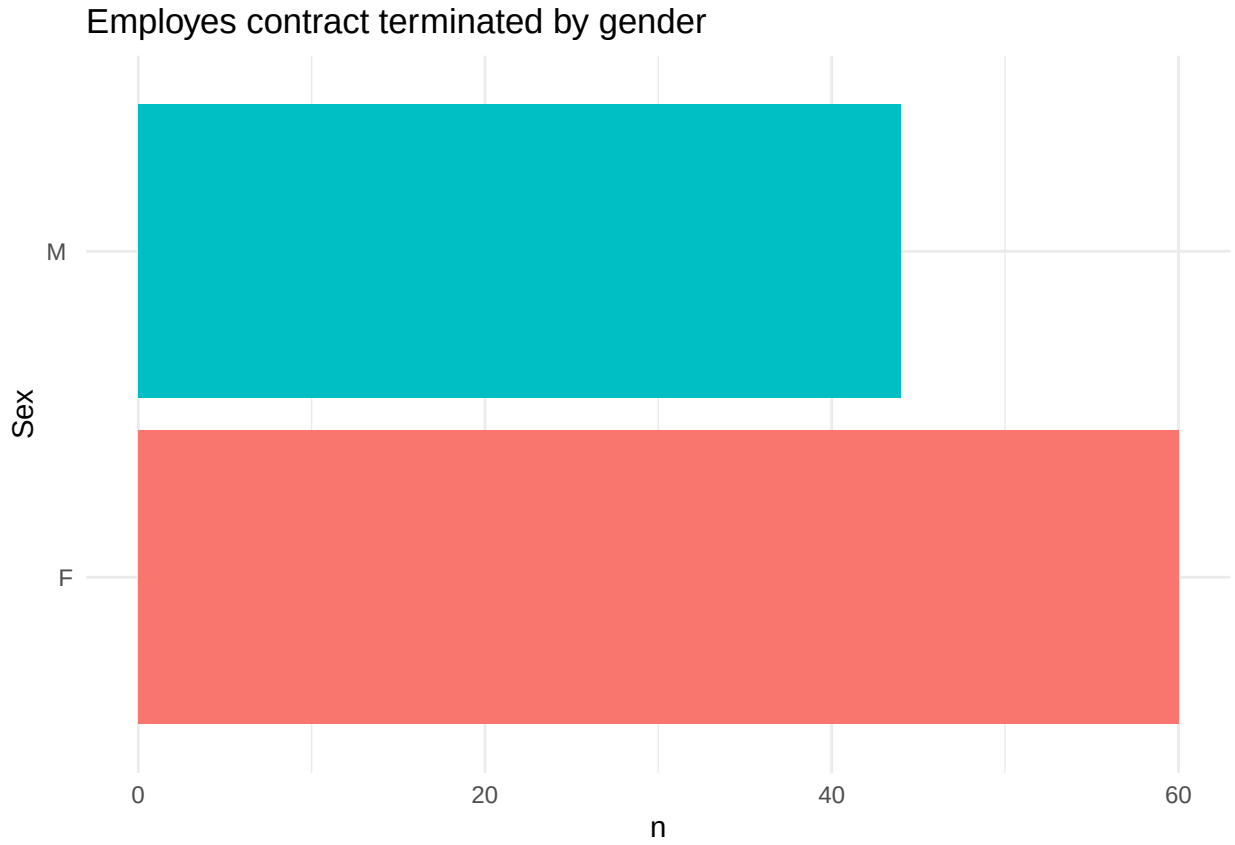

Enrolment by role and gender 2010 − 2013
2010
2011
2012
2013
BI Director
Software Engineering Manager
Area Sales Manager
Accountant I
Production Manager
Sr. Accountant
Enterprise Architect
Database Administrator
Data Analyst
IT Support
Software Engineer
Production Technician II
Sr. DBA
Production Technician I
Role
0
10
20
30
40
count
Sex
F
M

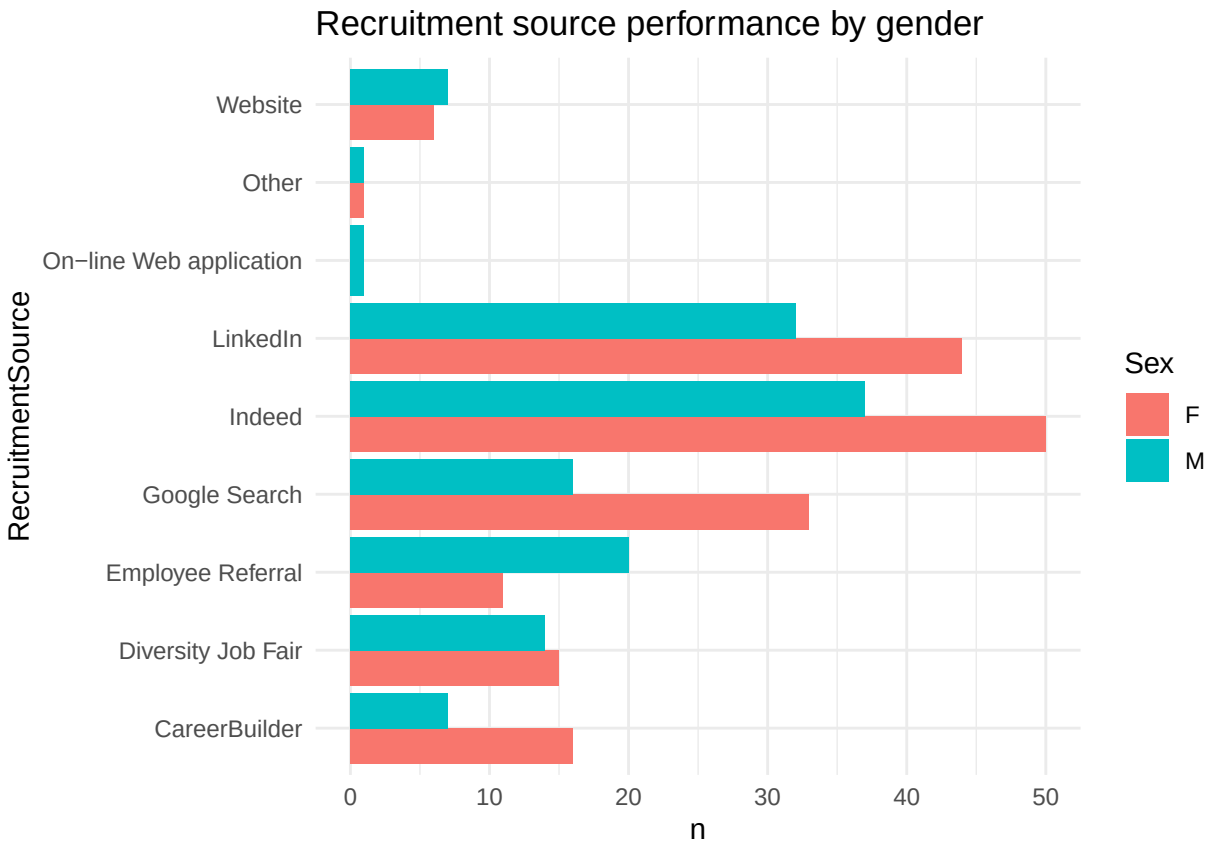

Enrolment by role and gender 2014 − 2016
2014
2015
2016
President & CEO
Administrative Assistant
Director of Sales
IT Director
Network Engineer
IT Manager − Support
BI Developer
Sales Manager
Sr. Network Engineer
Director of Operations
BI Director
Software Engineering Manager
Area Sales Manager
Accountant I
Production Manager
Sr. Accountant
Enterprise Architect
Database Administrator
Data Analyst
IT Support
Software Engineer
Production Technician II
Sr. DBA
Production Technician I
Role
0 5 10 15 20
count
Sex
F
M

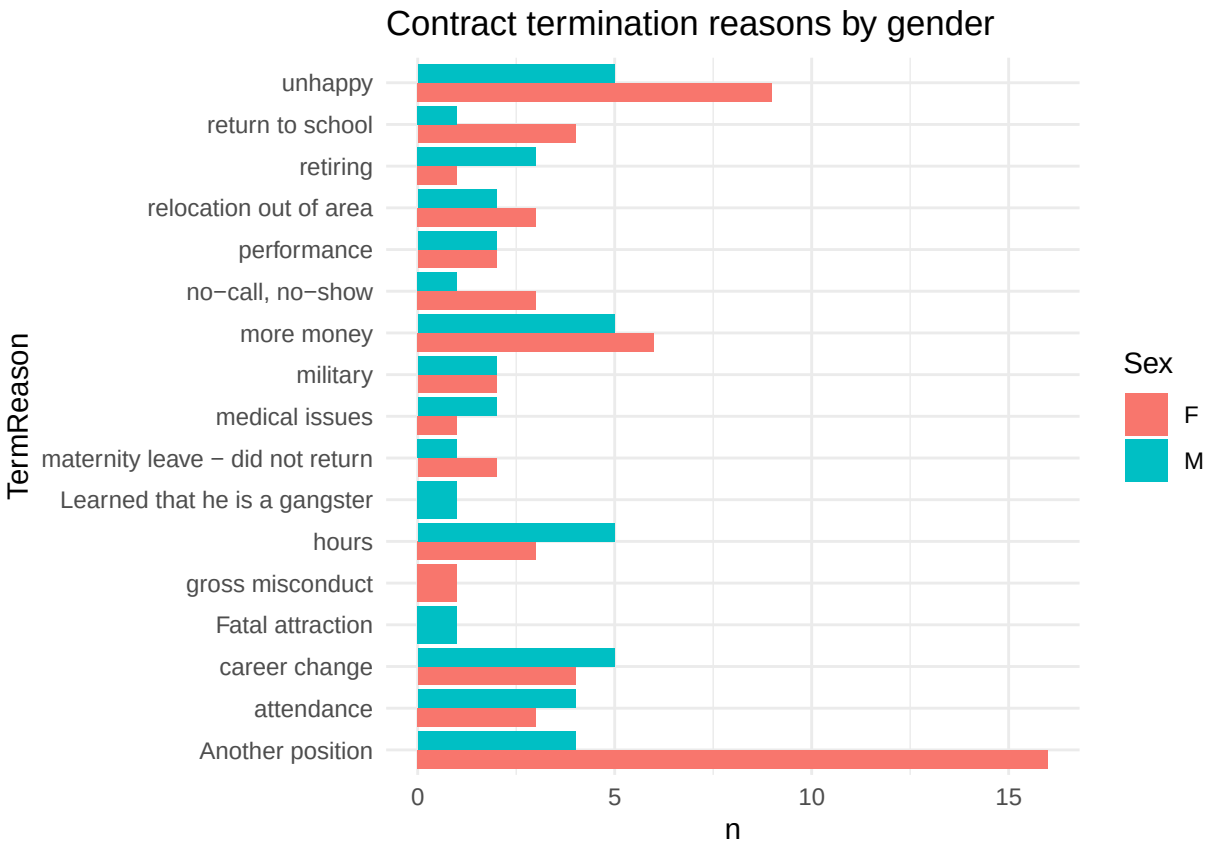

Enrolment by role and gender 2017 − 2018
2017
2018
Software Engineer
Production Technician II
Sr. DBA
Production Technician I
Role
0
1
2
3
count
Sex
F
M

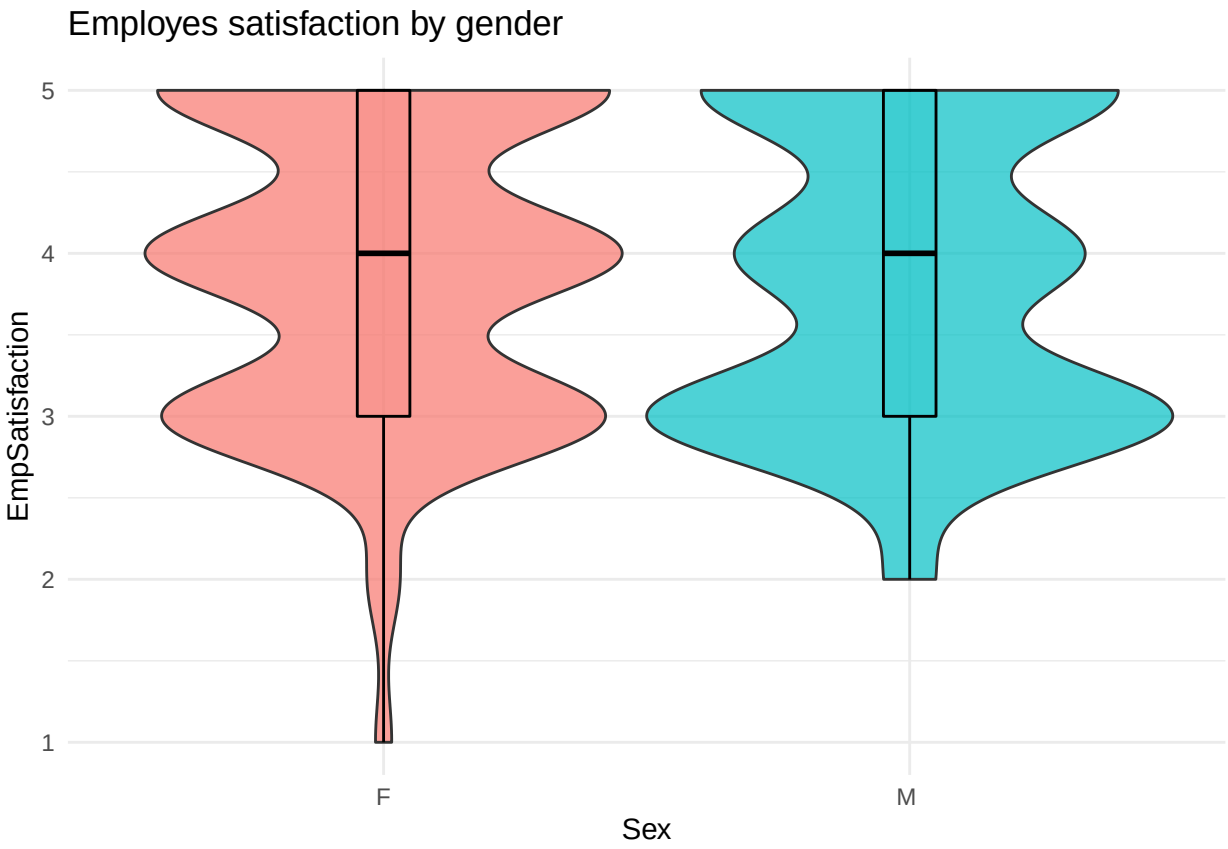

Hiring trend by gender
50
40
30
20
10
0
n
2010
2015
year
Sex
F
M

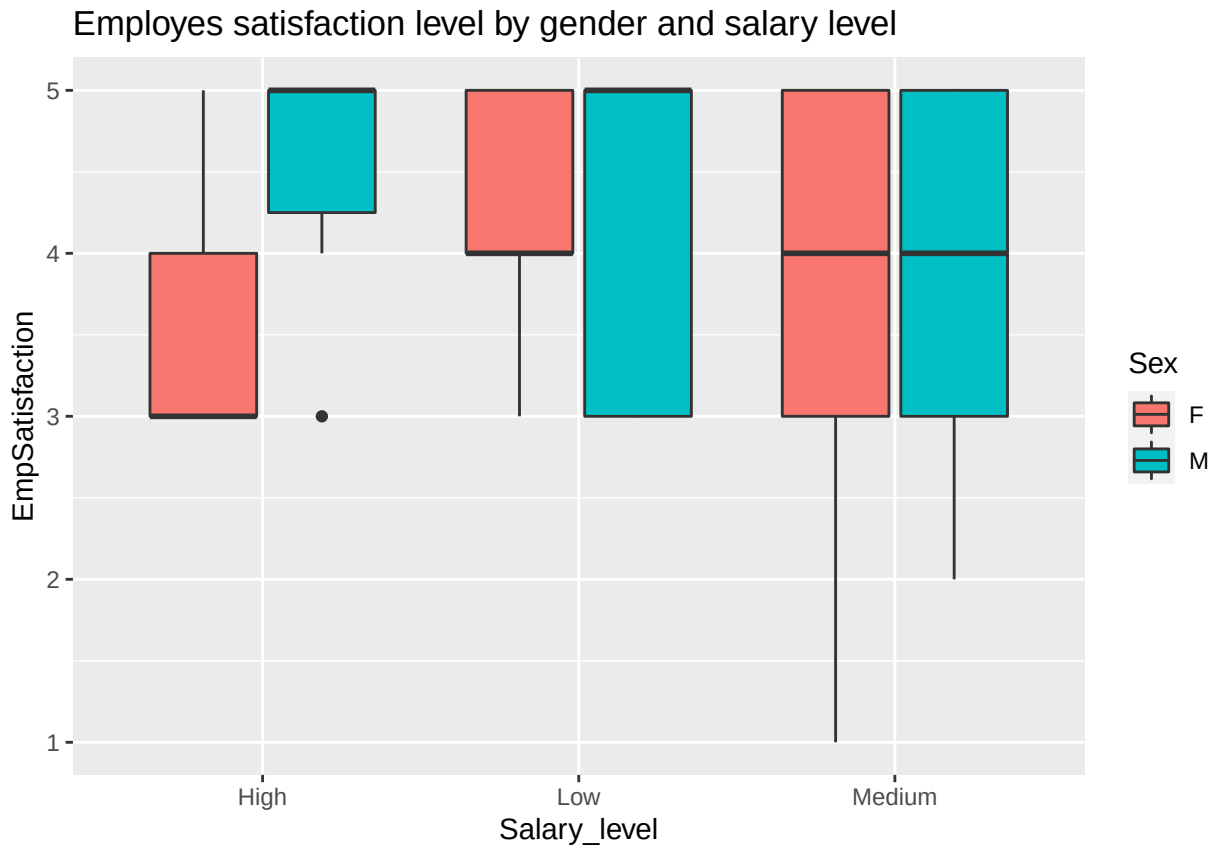

Average salary by gender and role
Position
CIO
Data Analyst
IT Manager − DB
Data Architect
Principal Data Architect
IT Manager − Infra
Shared Services Manager
Senior BI Developer
President & CEO
Administrative Assistant
Director of Sales
IT Director
Network Engineer
IT Manager − Support
BI Developer
Sales Manager
Sr. Network Engineer
Director of Operations
BI Director
Software Engineering Manager
Area Sales Manager
Accountant I
Production Manager
Sr. Accountant
Enterprise Architect
Database Administrator
Data Analyst
IT Support
Software Engineer
Production Technician II
Sr. DBA
Production Technician I
0
50000
100000
150000
200000
250000
avgsal
Sex
F
M


Employes contract terminated by gender
Sex
M
F
0
20
40
60
n


Recruitment source performance by gender
RecruitmentSource
Website
Other
On−line Web application
LinkedIn
Indeed
Google Search
Employee Referral
Diversity Job Fair
CareerBuilder
0
10
20
30
40
50
n
Sex
F
M


Contract termination reasons by gender
TermReason
unhappy
return to school
retiring
relocation out of area
performance
no−call, no−show
more money
military
medical issues
maternity leave − did not return
Learned that he is a gangster
hours
gross misconduct
Fatal attraction
career change
attendance
Another position
0
5
10
15
n
Sex
F
M


Employes satisfaction by gender
EmpSatisfaction
5
4
3
2
1
F
M
Sex

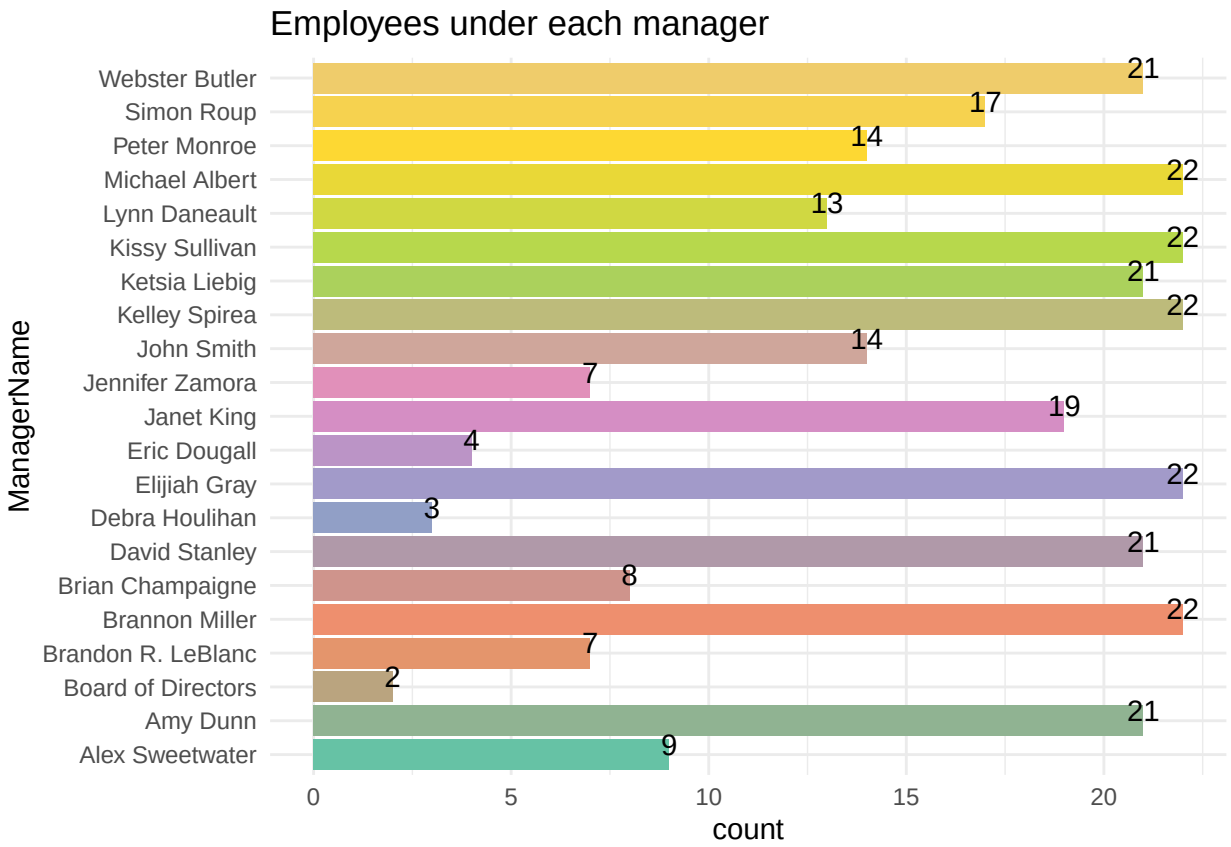


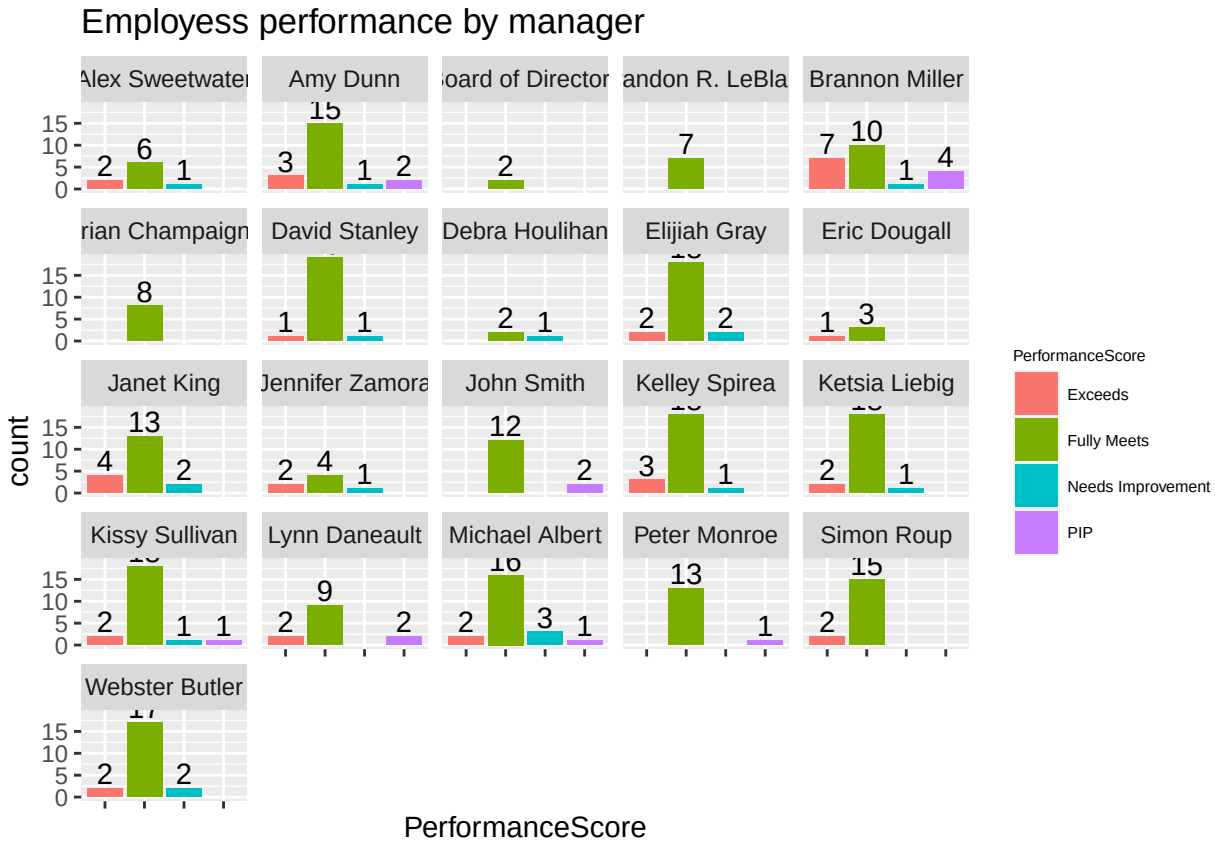
3.2 Manager and performance analysis

In the next part managers and relative performance was analyzed. Company has 21 different managers listed as follow:

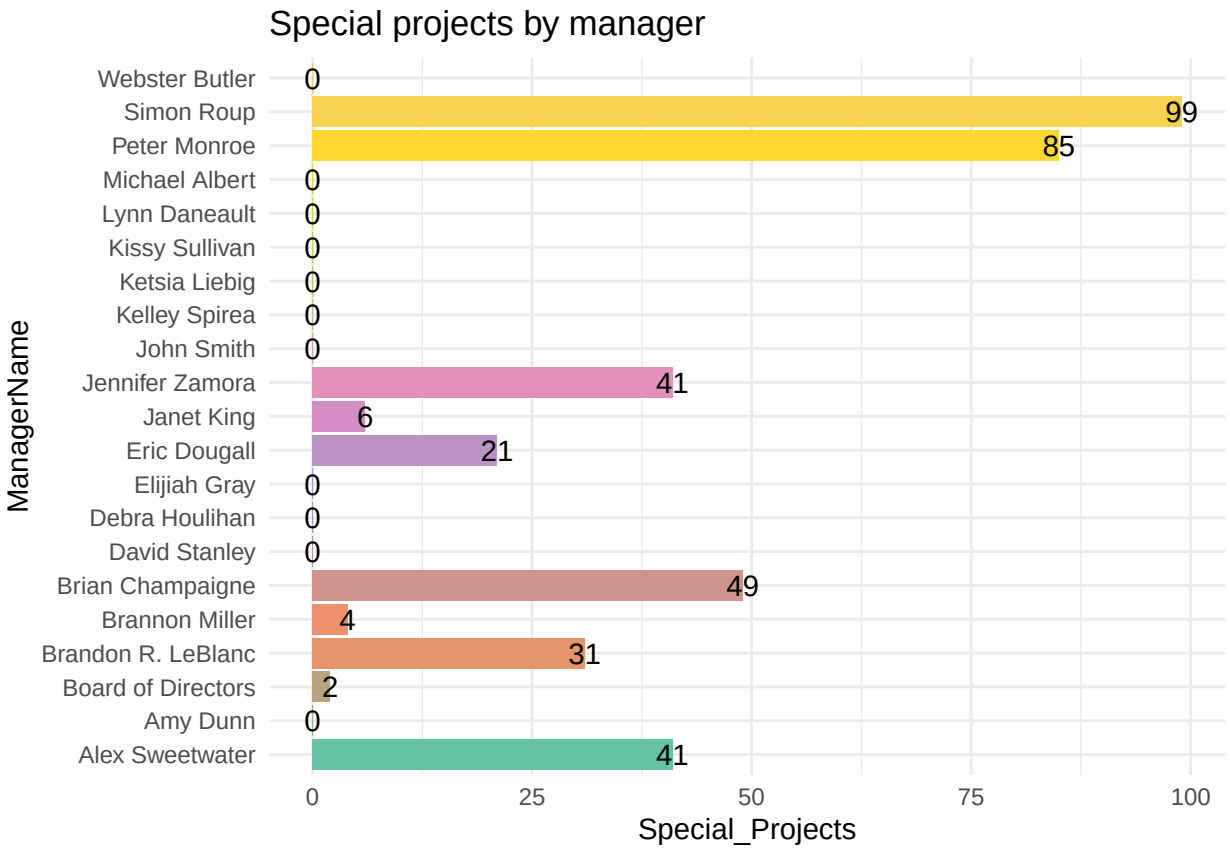
Table 2: Manager list

x
Michael Albert
Simon Roup
Kissy Sullivan
Elijiah Gray
Webster Butler
Amy Dunn
Alex Sweetwater
Ketsia Liebig
Brannon Miller
Peter Monroe
David Stanley
Kelley Spirea
Brandon R. LeBlanc
Janet King
John Smith
Jennifer Zamora
Lynn Daneault
Eric Dougall
Debra Houlihan
Brian Champaigne
Board of Directors


Employees under each manager
ManagerName
Webster Butler
Simon Roup
Peter Monroe
Michael Albert
Lynn Daneault
Kissy Sullivan
Ketsia Liebig
Kelley Spirea
John Smith
Jennifer Zamora
Janet King
Eric Dougall
Elijiah Gray
Debra Houlihan
David Stanley
Brian Champaigne
Brannon Miller
Brandon R. LeBlanc
Board of Directors
Amy Dunn
Alex Sweetwater
21
17
14
22
13
22
21
22
14
7
19
4
22
3
21
8
22
7
2
21
9
0
5
10
15
20
count


Employess performance by manager
Alex Sweetwater
Amy Dunn
Board of Director
Brandon R. LeBla
Brannon Miller
Brian Champaign
David Stanley
Debra Houlihan
Elijiah Gray
Eric Dougall
Janet King
Jennifer Zamora
John Smith
Kelley Spirea
Ketsia Liebig
Kissy Sullivan
Lynn Daneault
Michael Albert
Peter Monroe
Simon Roup
Webster Butler
count
PerformanceScore
PerformanceScore
Exceeds
Fully Meets
Needs Improvement
PIP

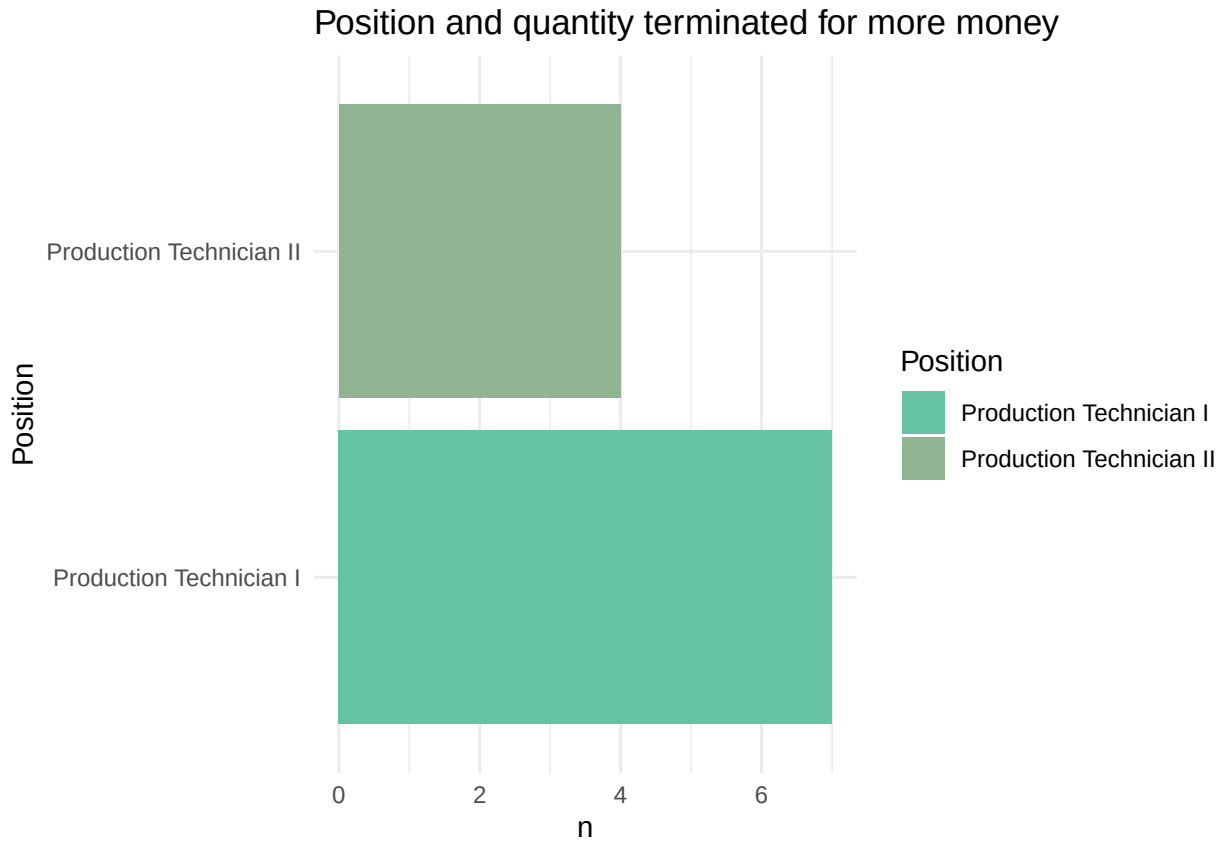

Termination reason by manager
lex Sweetwate
Amy Dunn
ndon R. LeBla
Brannon Miller
David Stanley
Debra Houlihar
Elijiah Gray
Janet King
ennifer Zamora
John Smith
Kelley Spirea
Ketsia Liebig
Kissy Sullivan
Lynn Daneault
Michael Albert
Peter Monroe
Simon Roup
Webster Butler
count
TermReason
TermReason
Another position
attendance
career change
Fatal attraction
gross misconduct
hours
Learned that he is a gangster
maternity leave − did not return
medical issues
military
more money
no−call, no−show
performance
relocation out of area
retiring
return to school
unhappy

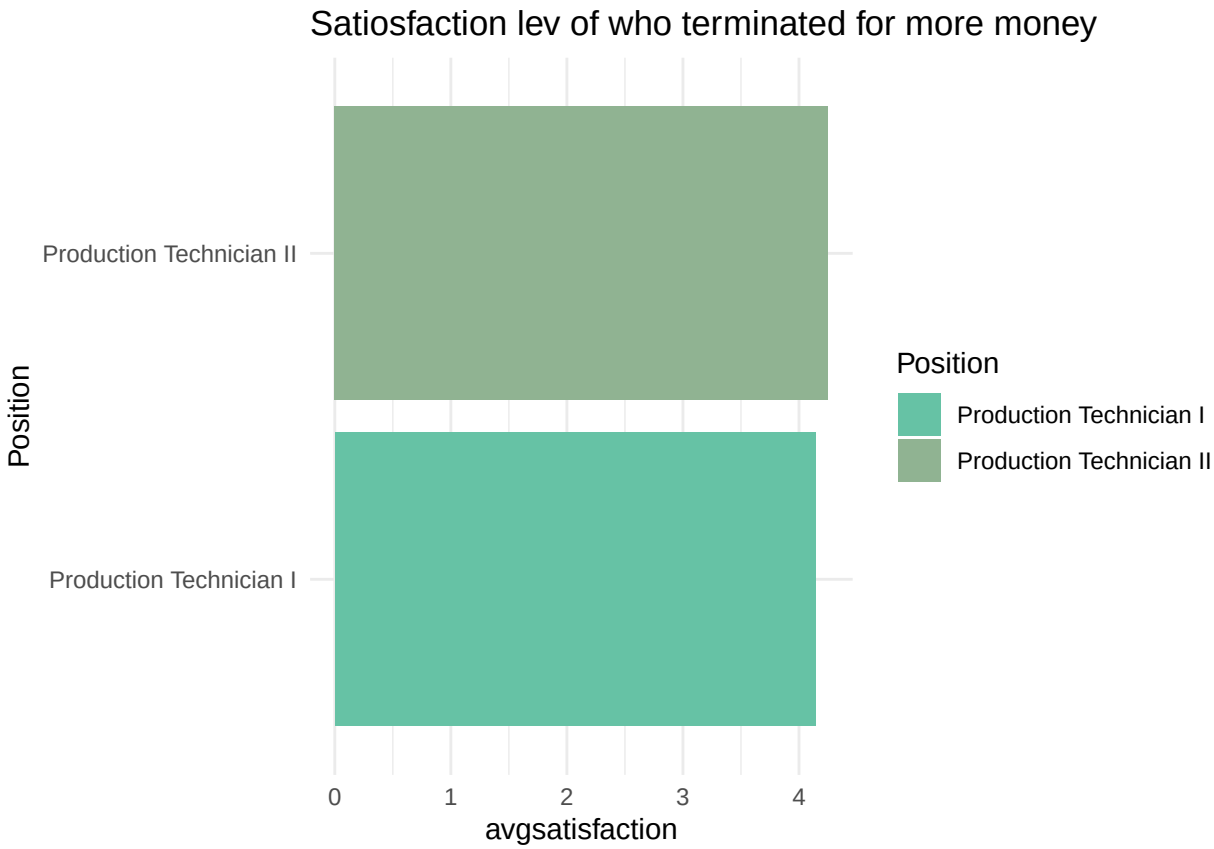

Special projects by manager
ManagerName
Webster Butler
Simon Roup
Peter Monroe
Michael Albert
Lynn Daneault
Kissy Sullivan
Ketsia Liebig
Kelley Spirea
John Smith
Jennifer Zamora
Janet King
Eric Dougall
Elijiah Gray
Debra Houlihan
David Stanley
Brian Champaigne
Brannon Miller
Brandon R. LeBlanc
Board of Directors
Amy Dunn
Alex Sweetwater
0
99
85
0
0
0
0
0
0
41
6
21
0
0
0
49
4
31
2
0
41
0
25
50
75
100
Special_Projects

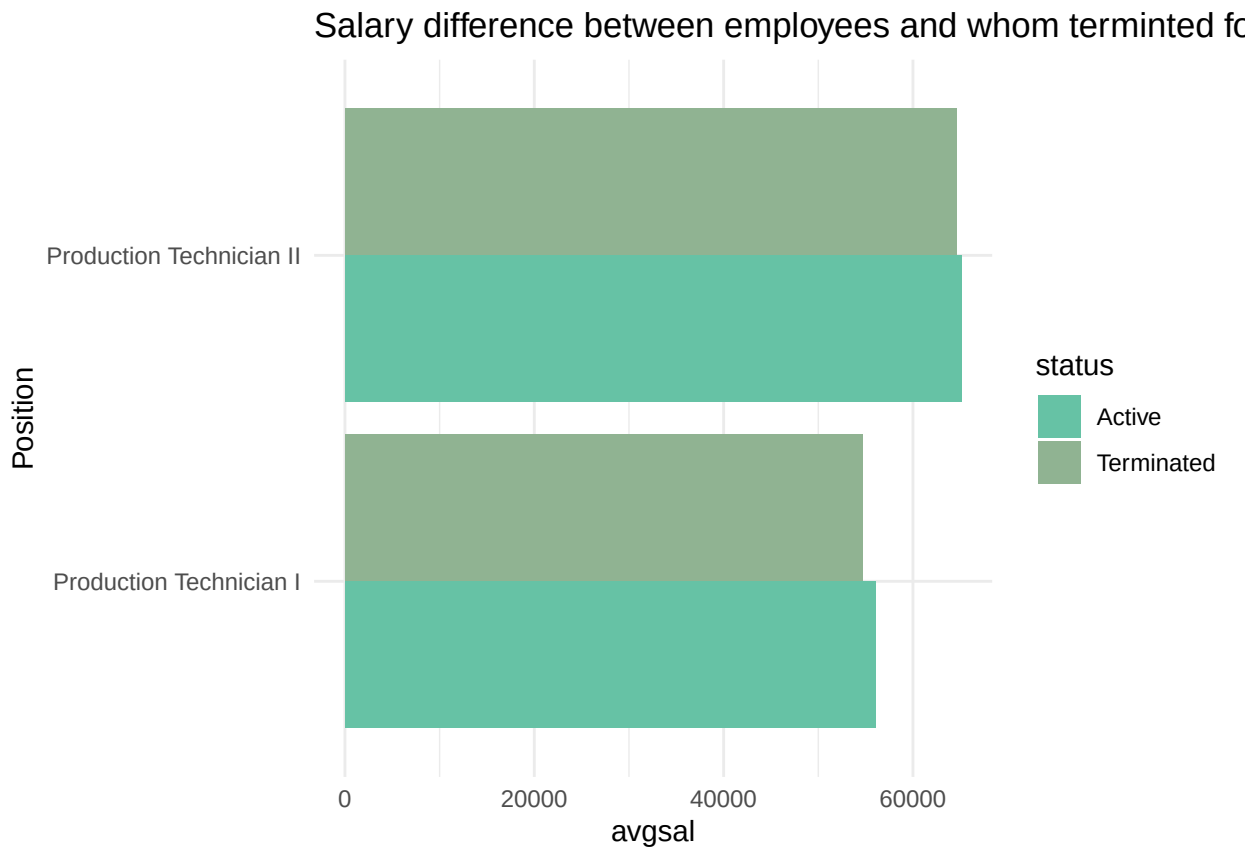

Average employees satisfaction by manager
Webster Butler
Simon Roup
Peter Monroe
Michael Albert
Lynn Daneault
Kissy Sullivan
Ketsia Liebig
Kelley Spirea
John Smith
Jennifer Zamora
Janet King
Eric Dougall
Elijiah Gray
Debra Houlihan
David Stanley
Brian Champaigne
Brannon Miller
Brandon R. LeBlanc
Board of Directors
Amy Dunn
Alex Sweetwater
ManagerName
0
1
2
3
4
Average Satisfaction

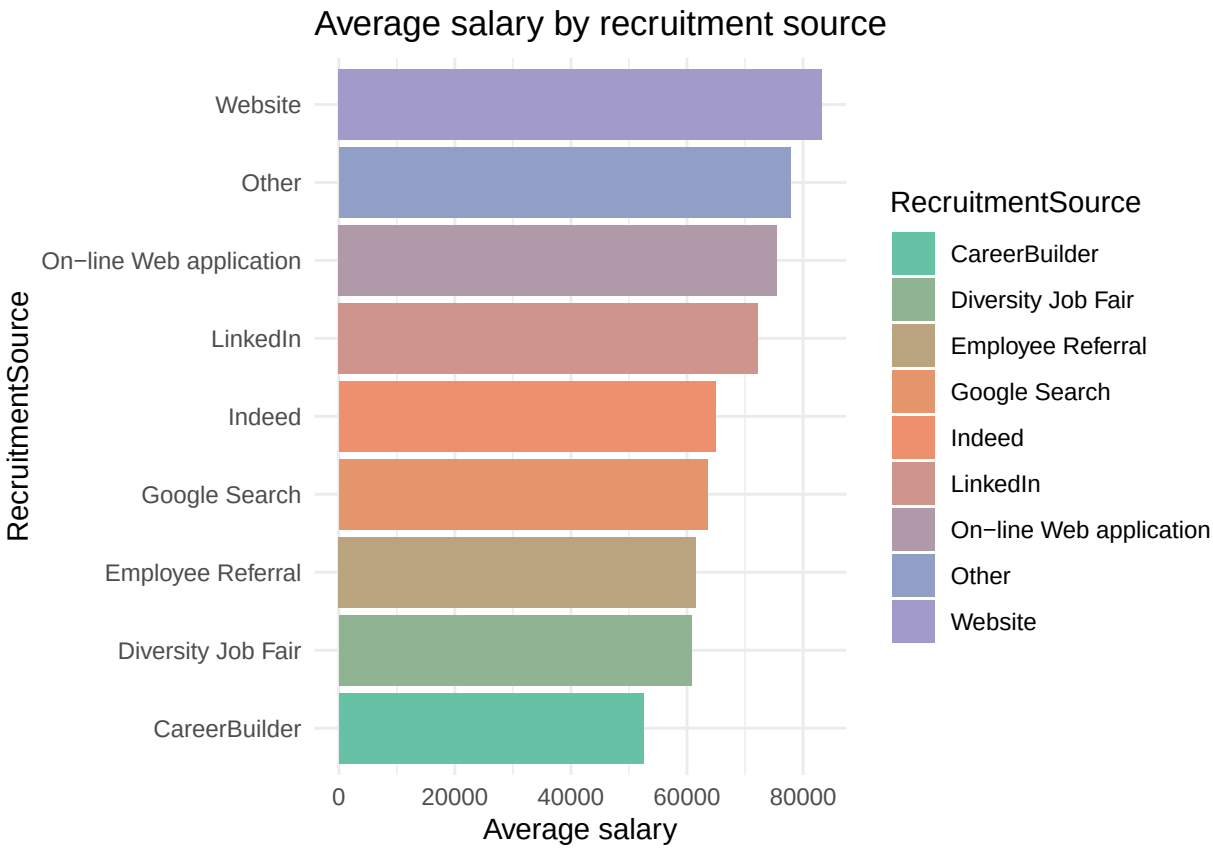
3.3 Termination for salary analysis

In the next part the correlation between salary and employees whom terminated the contract for more money reasons.

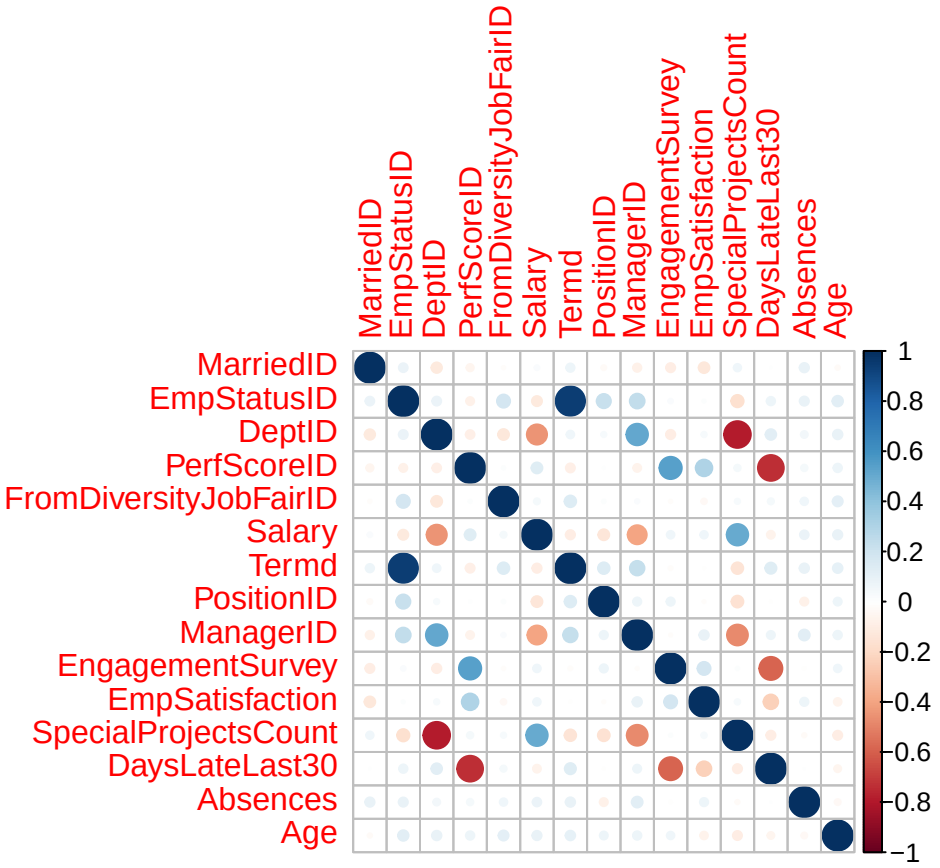



Satiosfaction lev of who terminated for more money
Production Technician II
Production Technician I
Position
0
1
2
3
4
avgsatisfaction
Position
Production Technician I
Production Technician II

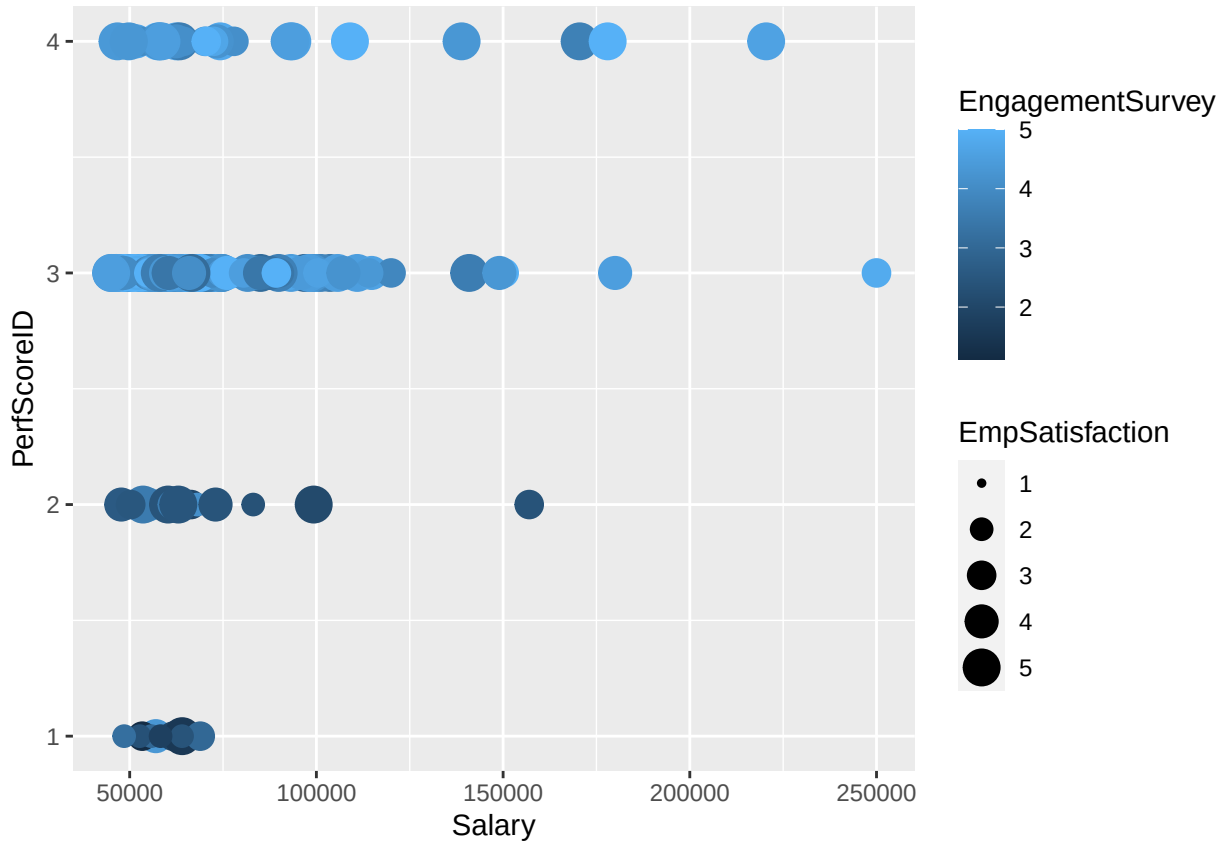

Salary difference between employees and whom terminted fc
Position
Production Technician II
Production Technician I
0
20000
40000
60000
avgsal
status
Active
Terminated


Average salary by recruitment source
RecruitmentSource
Website
Other
On−line Web application
LinkedIn
Indeed
Google Search
Employee Referral
Diversity Job Fair
CareerBuilder
0
20000
40000
60000
80000
Average salary
RecruitmentSource
CareerBuilder
Diversity Job Fair
Employee Referral
Google Search
Indeed
LinkedIn
On−line Web application
Other
Website

4.0 Correlation



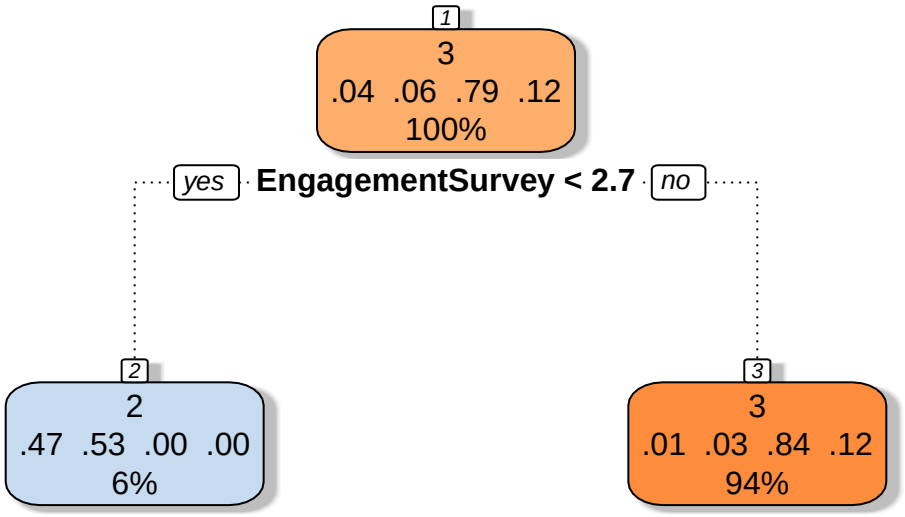
It is possible to appreciate an interesting correlation between the Performance Score (PerfScoreID in the graph) and other variables in the data set: Salary, Manager, Engagement, Employee satisfaction, Absences and Age. This can suggest that the employees performance can be subjected to variation according to those other variables. Lets explore this possible insight deeper.



In the plot is possible to appreciate a slightly pattern. It's worth to try build a prediction model that takes as input Salary, Manager, Engagement, Employee satisfaction, Absences, Age and maybe other variables to predict how the employees performance will be.

5.0 Prediction model

```
## CART
##
## 247 samples
##   6 predictor
##   4 classes: '1', '2', '3', '4'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 222, 223, 222, 223, 221, 222, ...
## Resampling results across tuning parameters:
##
##   cp          Accuracy   Kappa
##   0.0000000  0.7897436  0.19254666
##   0.0754717  0.8014359  0.21329276
##   0.1509434  0.7855897  0.03849572
##
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was cp = 0.0754717.
```

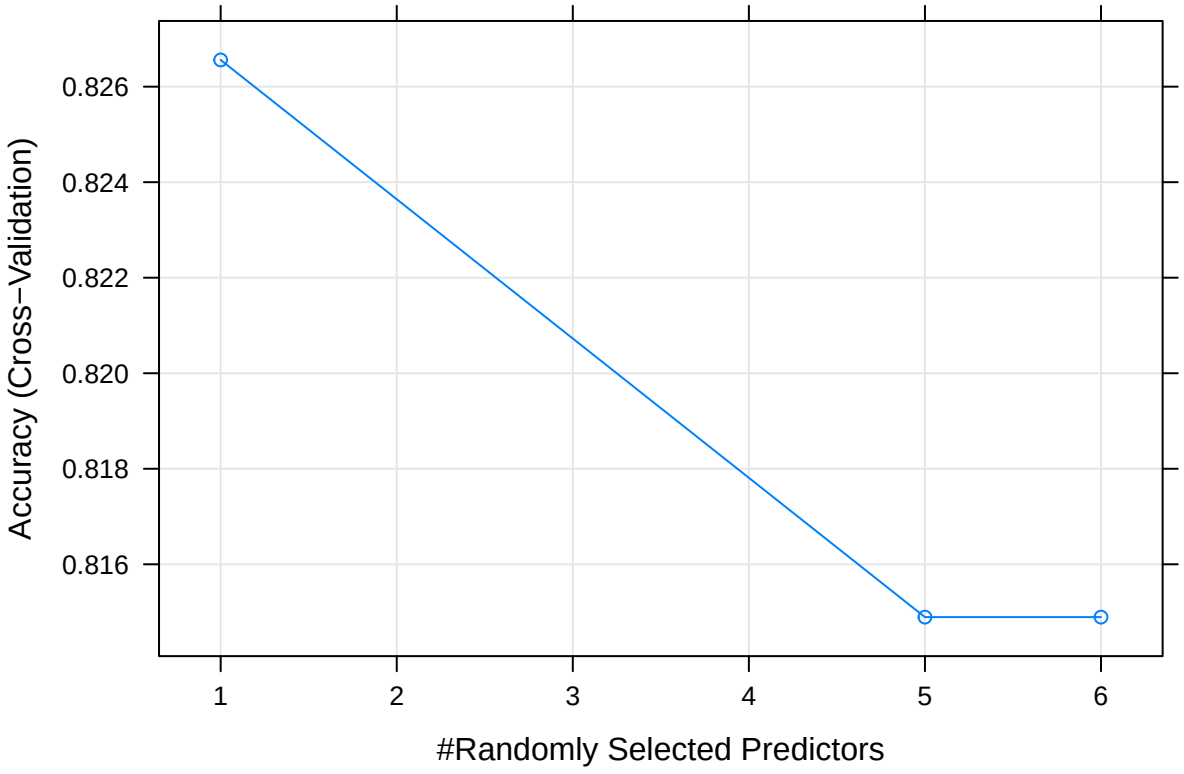



```
## rpart variable importance
##
##                  Overall
## EngagementSurvey 100.000
## EmpSatisfaction   49.962
## Salary            12.818
## Absences           4.828
## Age                4.658
## ManagerID          0.000
```

I start trying to fit a simple CART model, which generated the decision tree in the plot. It also estimated the importance of each variable in the model. Overall the model has an estimated accuracy of 0.8014359 and an error of 0.2.

Lets try to fit a more complex model to raise our accuracy.

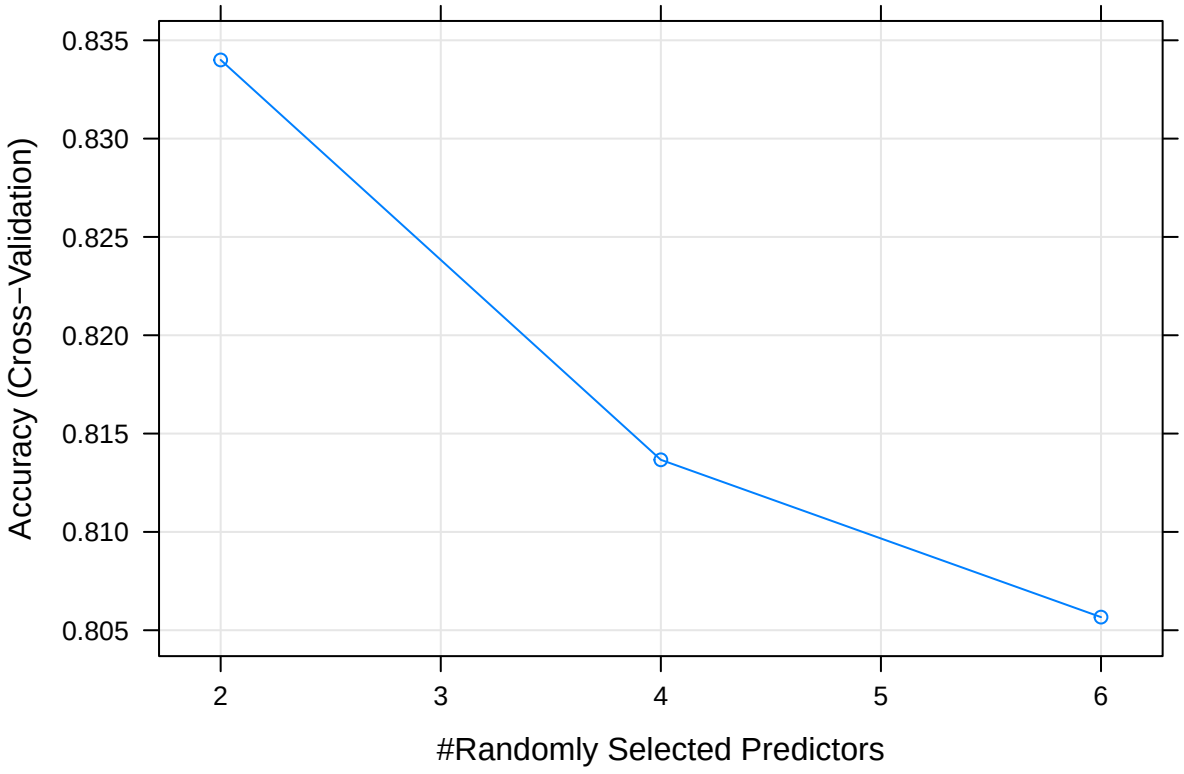
```
## Random Forest
##
## 247 samples
##   6 predictor
##   4 classes: '1', '2', '3', '4'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 222, 224, 221, 222, 223, 223, ...
## Resampling results across tuning parameters:
##
##   mtry  Accuracy   Kappa
##   1     0.8265591  0.3027629
##   5     0.8148941  0.3162735
##   6     0.8148941  0.3162735
##
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was mtry = 1.
```

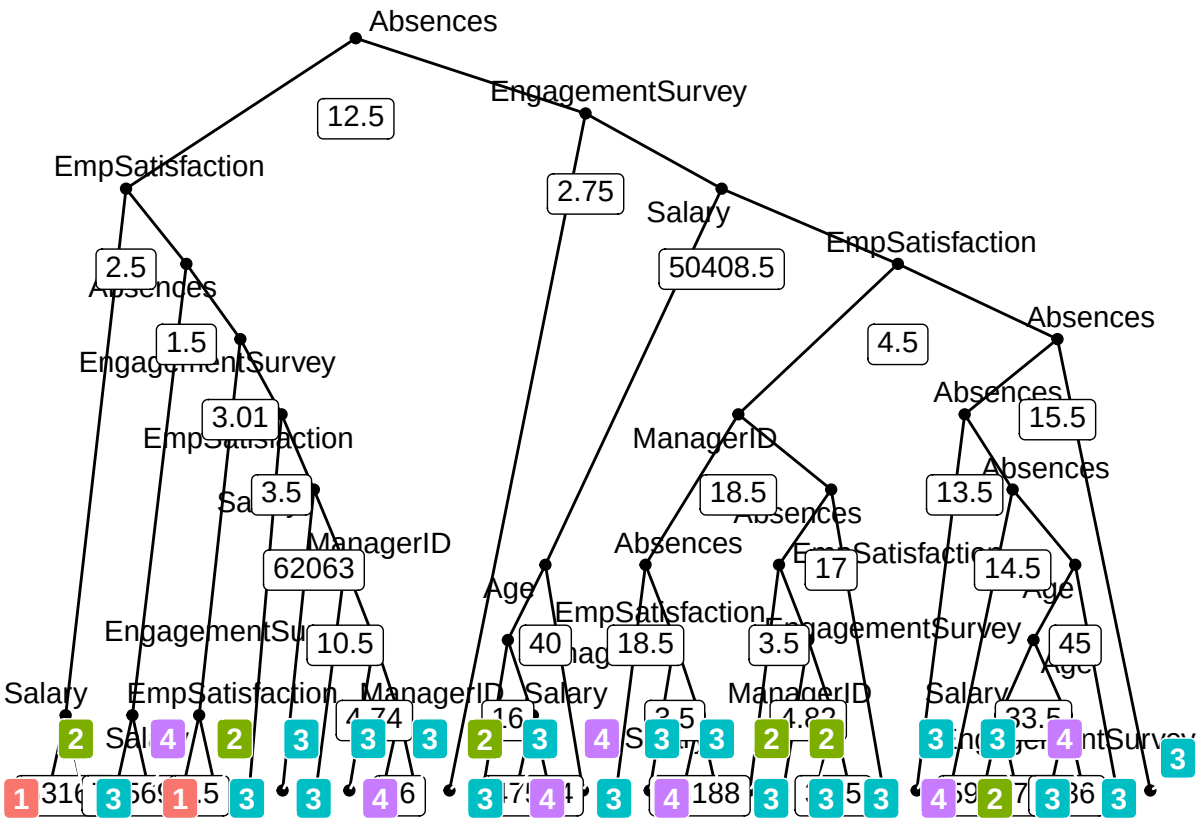


Using a default model we can see that we have a result of 0.8265591, witch is slightly better than the previous one.

Let's see if the model can be tune to reach a better result.

```
## Random Forest
##
## 247 samples
##   6 predictor
##   4 classes: '1', '2', '3', '4'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 223, 223, 222, 223, 222, 222, ...
## Resampling results across tuning parameters:
##
##   mtry  Accuracy   Kappa
##   2     0.8340000  0.3675143
##   4     0.8136667  0.3182035
##   6     0.8056667  0.2955213
##
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was mtry = 2.
```





We further increase the accuracy at 0.834 with an error of 0.17.