Case Study 2

Nicholas Mueller

Nmueller@smu.edu

Objective

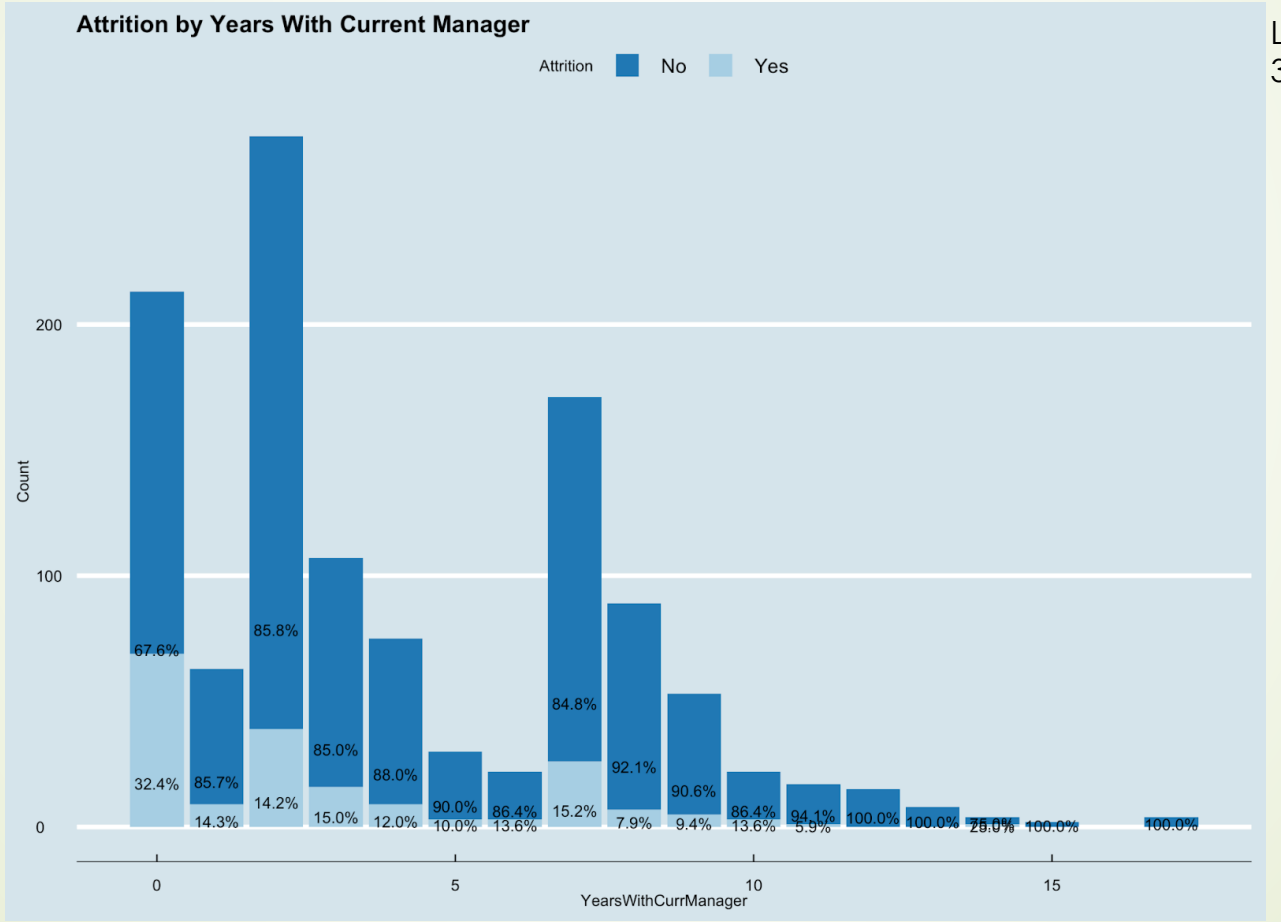
- Identifying the top 3 factors that lead to attrition
- Discover job specific trends
- Provide any other interesting trends and observations
- Build a model that can predict attrition and monthly Incomes

Data Description

- 3 data sets:
  - Data set 1:
    - 36 variables/columns
    - 870 Observations
    - No Missing values
  - Data Set 2:
    - 35 Variables/ Columns
    - 300 Observations
    - Missing Attrition Column
  - Data Set 3:
    - 35 Variables/Columns
    - 300 Observations
    - Missing Monthly Income Column

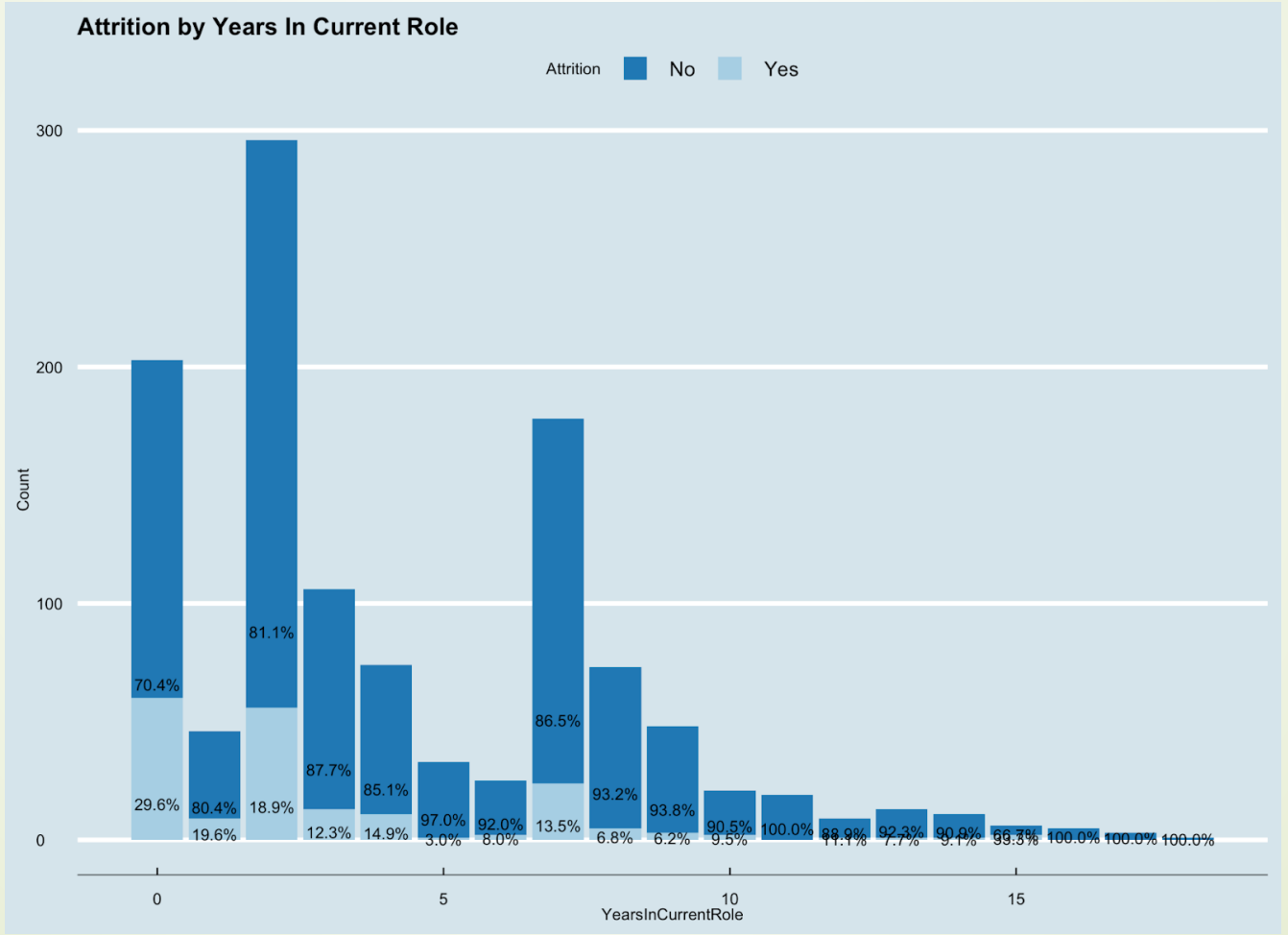
Top 3 Factors Leading To Attrition

Number 3: Early Stages In Career



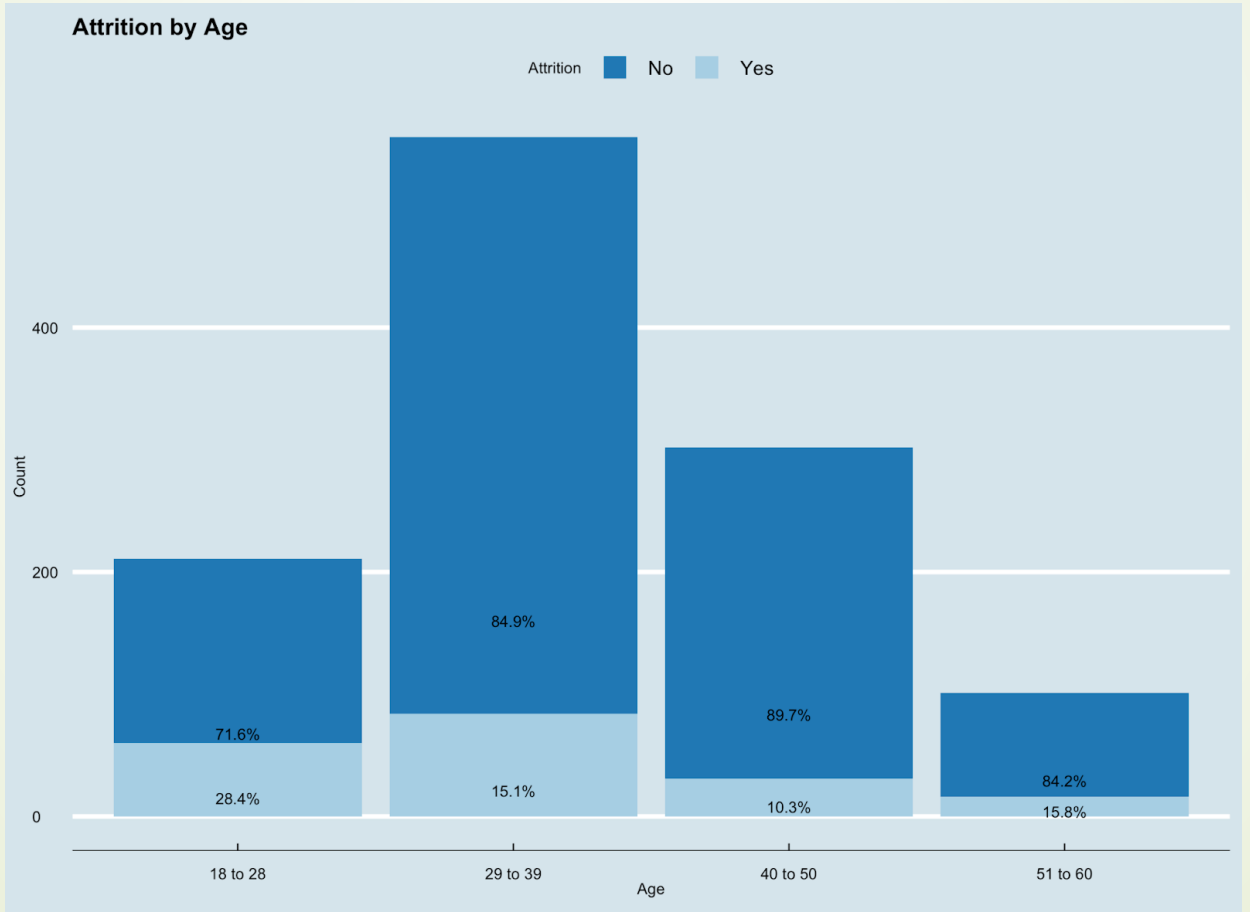
Less than 1 Years With Current Manager
32.4% Attrition rate

Number 3: Early Stages In Career



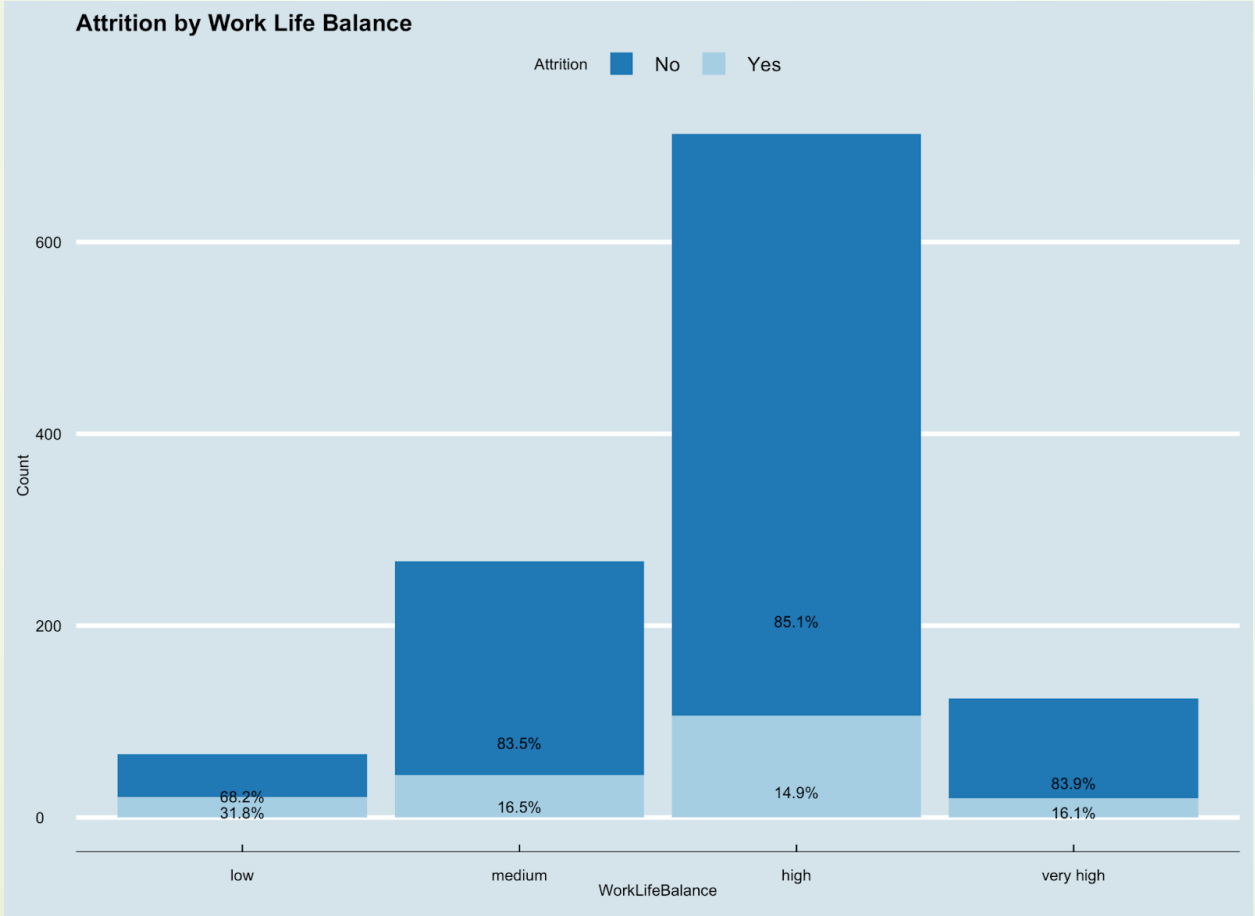
Less Than 1 Years In Current Role
29.6% Attrition rate

Number 3: Early Stages In Career



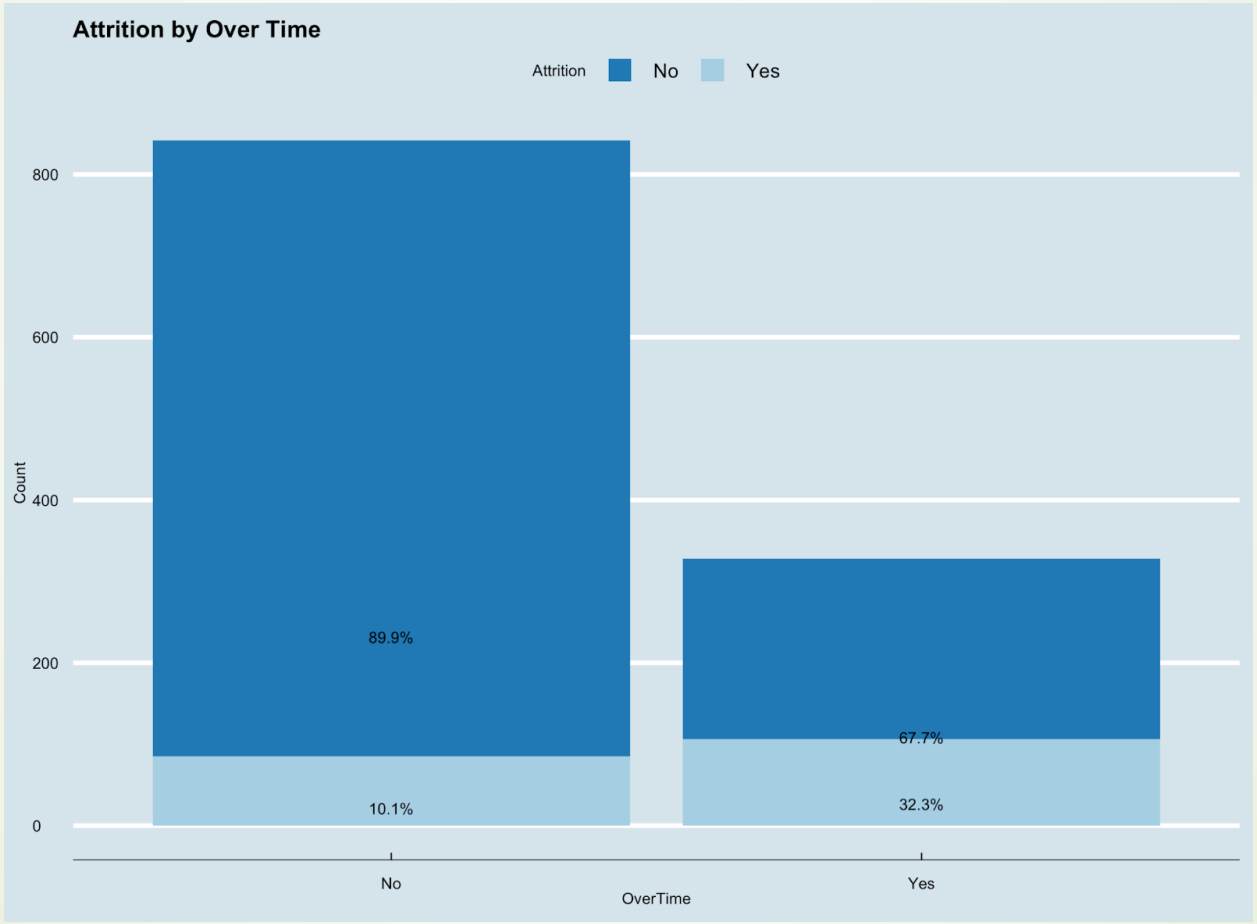
Ages 18 to 28
28.4% Attrition rate

Number 2: Work Life Balance



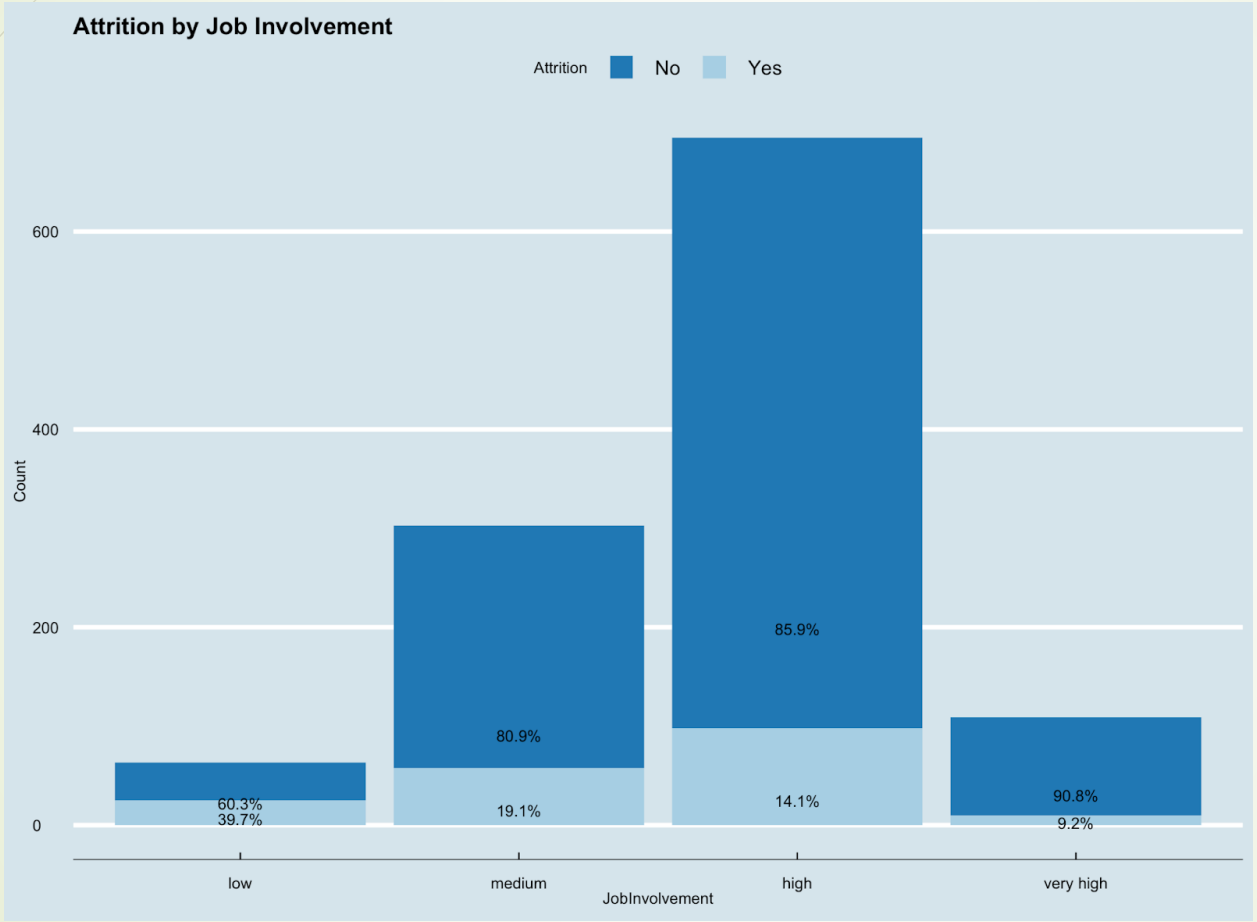
Low Work Life Balance
31.8% Attrition rate

Number 2: Work Life Balance



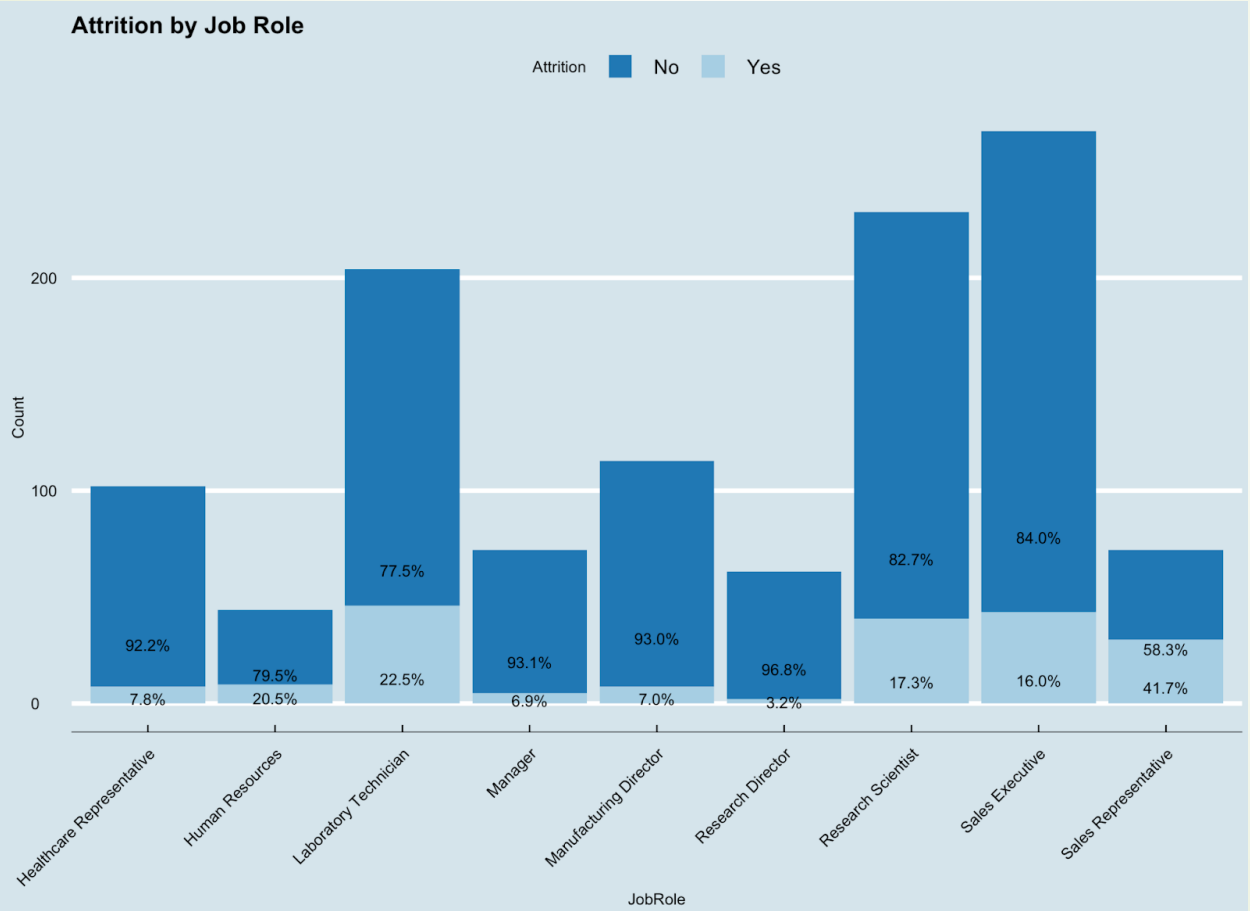
Yes To Over Time
32.3% Attrition rate

Number 2: Work Life Balance



Low Job Involvement
39.7% Attrition rate

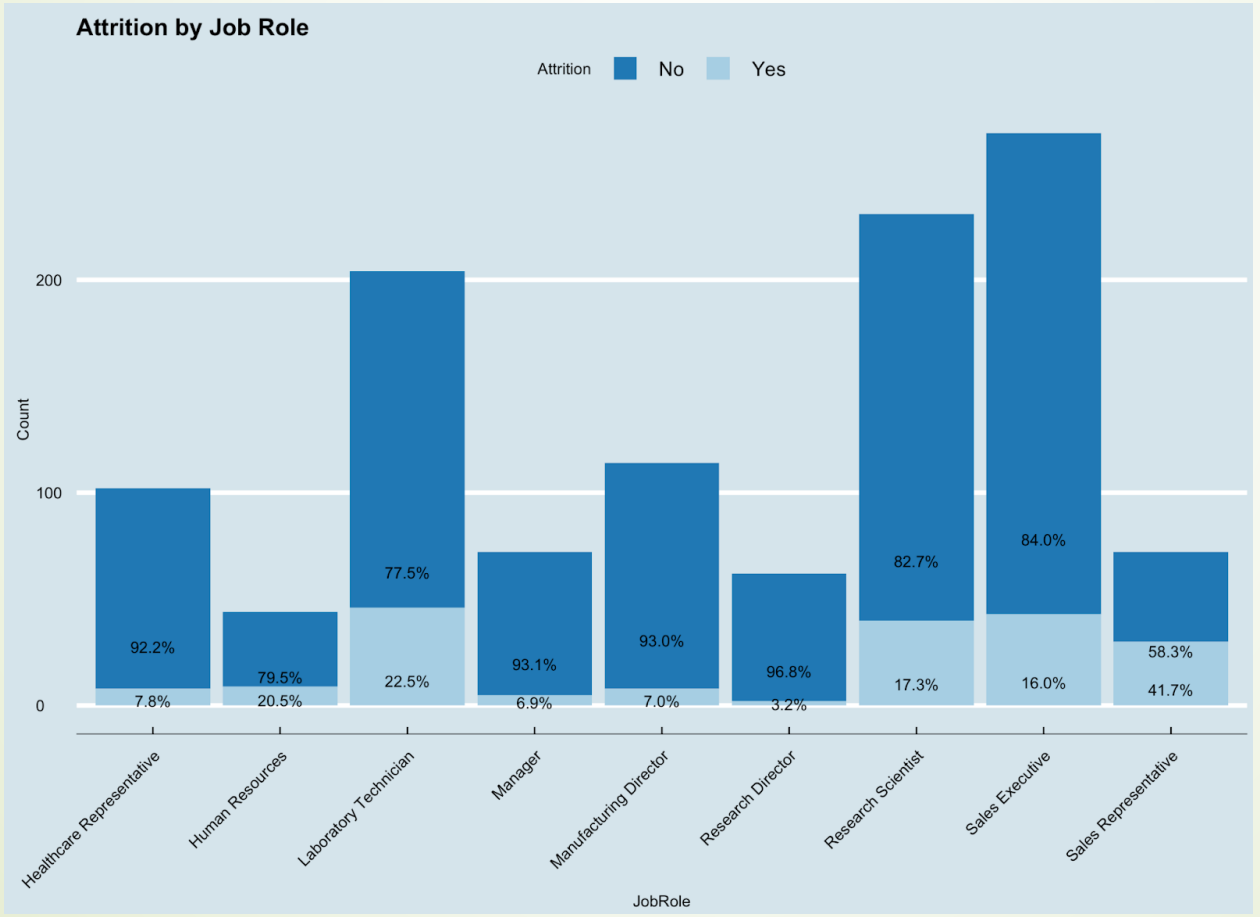
Number 1: Job Role



Sales Representative
41.7% Attrition rate

Job Specific Trends

Top 3 Lowest Attrition Rates

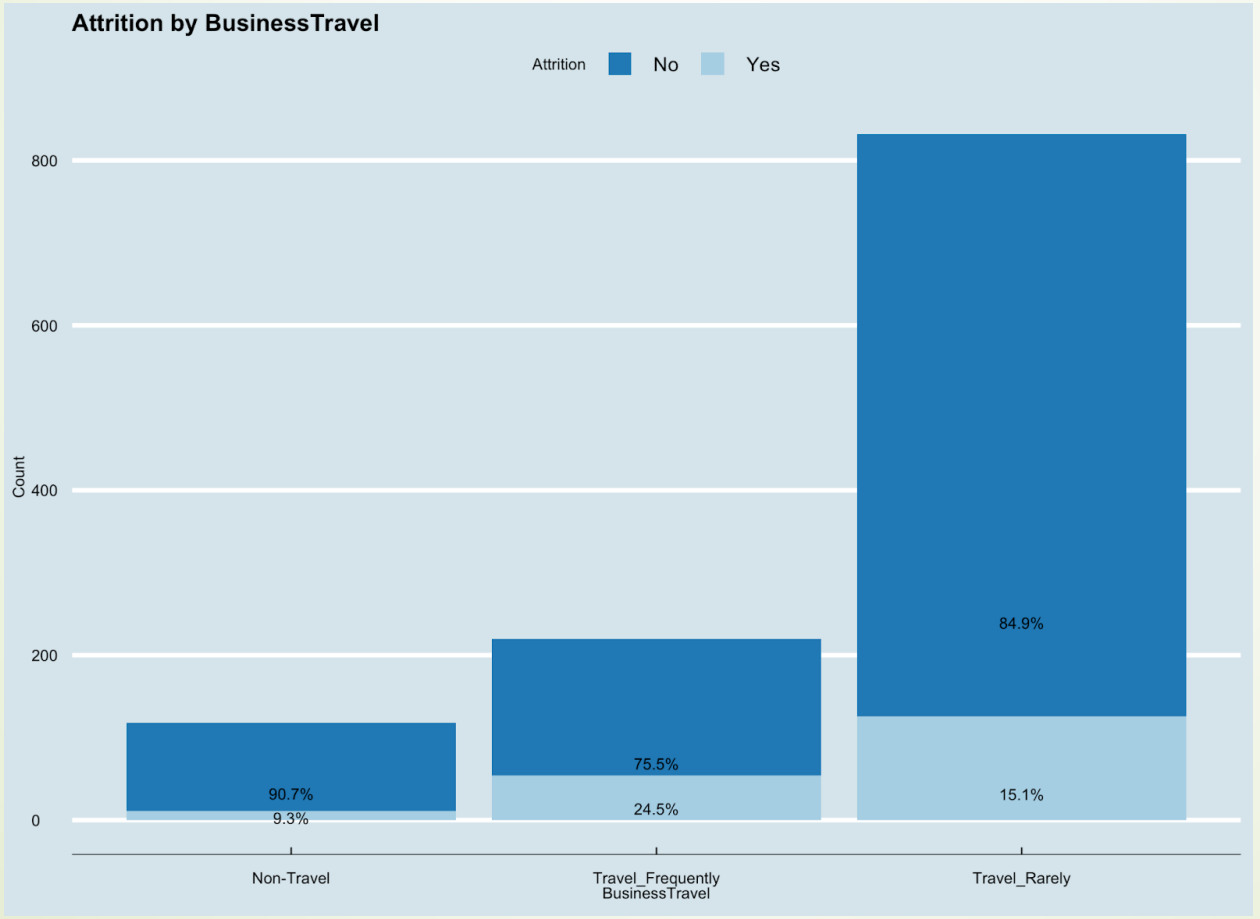


Research Director
3.2%

Manager
6.9%

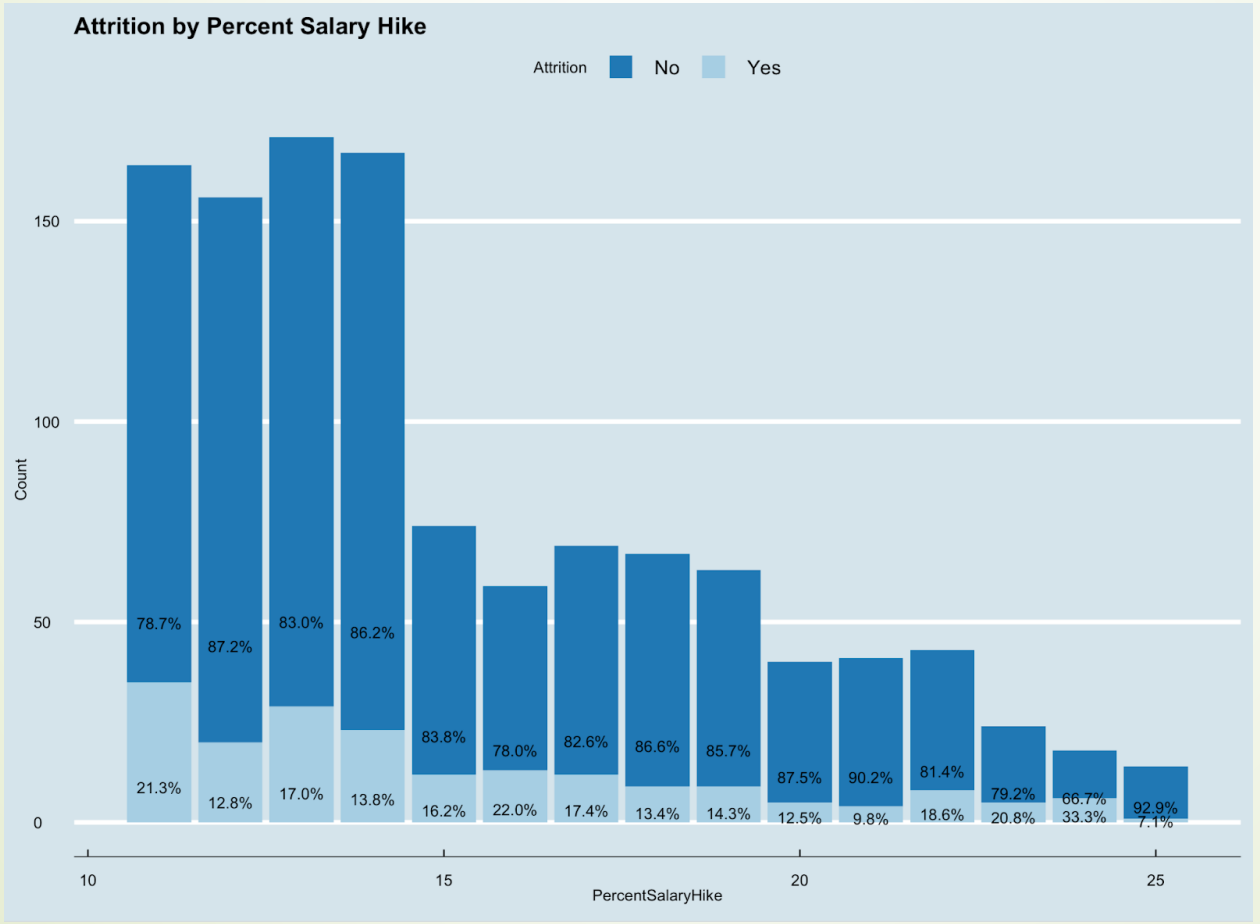
Manufacturing Director
7.0%

Travel For Work

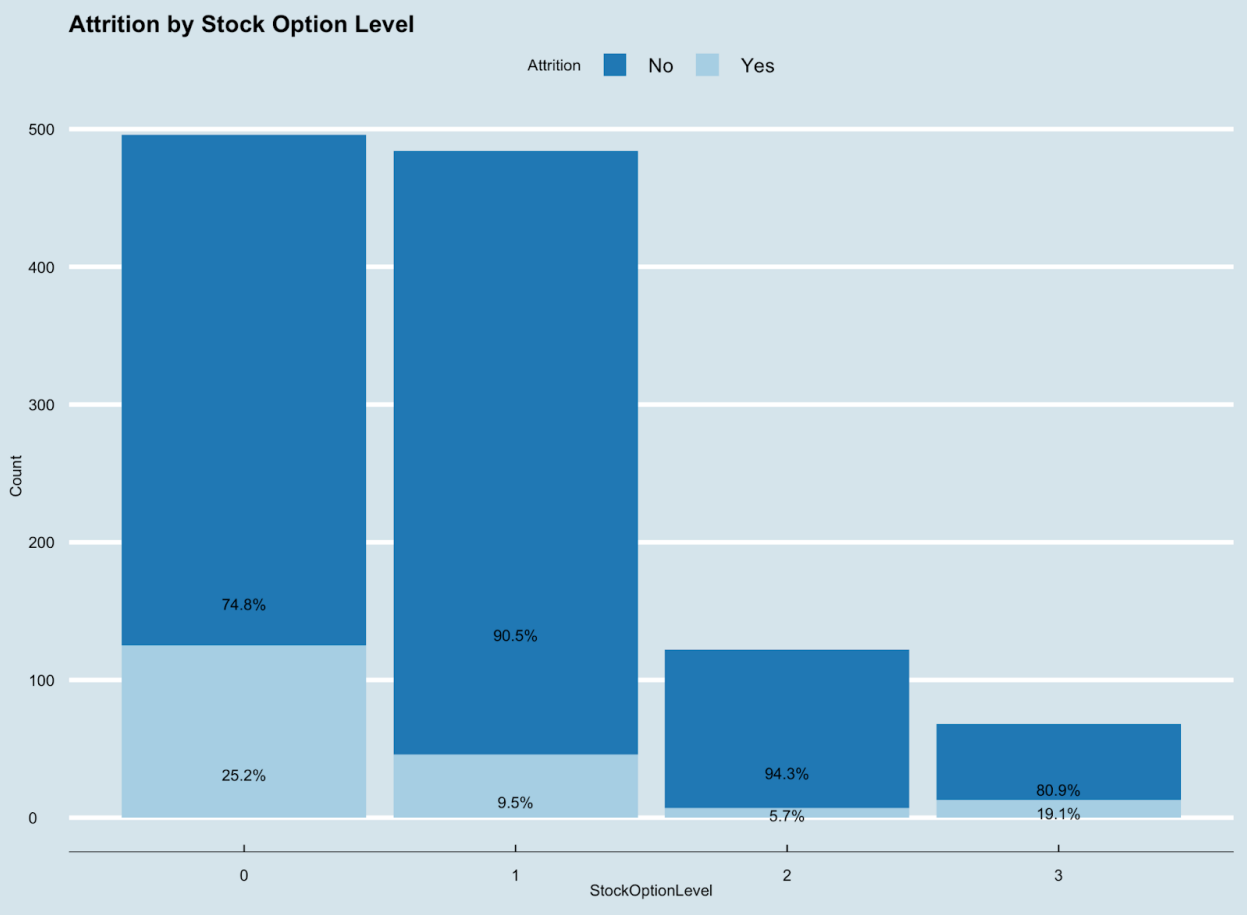


Interesting Trends

Salary Hike


Attrition by Percent Salary Hike
Attrition
No
Yes
Count
150
100
50
0
78.7%
87.2%
83.0%
86.2%
83.8%
78.0%
82.6%
86.6%
85.7%
87.5%
90.2%
81.4%
79.2%
66.7%
92.9%
21.3%
12.8%
17.0%
13.8%
16.2%
22.0%
17.4%
13.4%
14.3%
12.5%
9.8%
18.6%
20.8%
33.3%
7.1%
10
15
20
25
PercentSalaryHike

Stock Options



5.7% Attrition rate with level 2 stock option!!!

Predicting Attrition And Monthly Income

Naïve Bayes Process

- A simple but powerful machine learning algorithm that can be used for classification.
- Consists of:
  - Complete data set with all variables
  - A data set with similar variables except the one we would like to predict
- We will then eliminate unnecessary variables (refer to RMD or knit file for process)
- Preform several runs of the model
- Down sample and up sample unbalanced data(ex: Attrition has 730 “No” and 140 “Yes”)
- Compare
- PREDICT!

Compare Model Outputs

	No Change In Sample	Up Sampling	Down Sampling
Accuracy	.83	.70	.68
Sensitivity	.91	.63	.62
Specificity	.46	.78	.74

Linear Regression

- We are able to detect linearity among several variables
- Consists of:
  - Complete data set with all variables
  - A data set with similar variables except the one we would like to predict
- Use statistical evidence like a p-value to determine what factors to keep.
- Run Several Models until we achieve our desired output (<3000 RMSE)
- Compare
- PREDICT!

Compare Model Output

	Model 1	Model 2
MSPE	2,022,394	2,014,688
RMSE	1,422	1,419

Refer to RMD or Knit File For Information

Variables include:

DailyRate
HourlyRate
JobInvolvment
JobLevel
MonthlyRate
OverTime
YearsAtCompany
YearsInCurrentRole
PercentSalaryHike
YearsSinceLastPromotion
YearsWithCurrManager

The End!

Nicholas Mueller

nmueller@smu.edu

Questions?