DDSAnalyticsReport

Quentin,Sita,Olha,Tosin

4/4/2018

Contents

Abstract

DDS Analytics is a analytics company that specializes in talent management solutions for Fortune 1000 companies. Talent management is defined as the iterative process of developing and retaining employees. It may include workforce planning, employee training programs, identifying high-potential employees and reducing/preventing voluntary employee turnover (attrition). To gain a competitive edge over its competition, DDS Analytics is planning to leverage data science for talent management. The executive leadership has identified predicting employee turnover as its first application of data science for talent management. This report is the full data analysis of our findings after exploring the many facets of the data.

Executive Summary

DDS Analytics' primary task is investigation of the drivers of employee attrition using a data-driven analytics approach. From our descriptive analysis, the top drivers of employee attrition were identified. Besides drivers of attrition, the employee database provided for the analysis reveals trends that could be harnessed by the Human Resources Department for talent retention. This analysis is based on 35 factors that Human Resources track annually for all the company's 1470 employees. Some of these factors include personal information, satisfaction indices and typical human resources employee records. Insights gathered from statistical analysis of the data suggest that the top drivers of attrition are business travel, overtime, employee commute distance, employee tendency for job-hopping as well as job and relationship satisfaction. These factors are discussed in details in this report and recommendations provided for managing these causes of attrition.

I. Introduction

Talent is a precious commodity, especially in the corporate sector. Every organization no matter how big or small seeks not only to hire, but also to retain the best talent possible. Employee retention however, is much easier said than done these days, especially when the average amount of years an employee is likely to remain at a specific place of employment is 4.5 years as of 2014. What are the factors that lead to employee attrition? This is an age old question that most companies continue to try to answer. Can employee attrition be slowed, or avoided all together? Is there any data to support that job satisfaction is the main cause for an employee remaining at a job? All these questions and more will be explored in this study. Our goal is to find

sound answers to these questions rooted in strong statistical analysis of a dataset we have gathered from some of the most influential companies in the world.

II. Background

As a Talent Management organization we set out to leverage the strength of Data Science in order to maintain the best employees at our organization DDS Analytics. In order for the organization to move forward in this endeavor it was necessary to study the key factors in maintaining the best employees. As a result this study was launched by the management team in order to gain more insight into this arena.

III. Methodology

Data Cleaning & Preparation

In order to make sure our data was ready for proper analysis we had to go through a series of steps in order to prepare the data. These steps include the following...

1. Data Import
2. Data Type Conversion for quantitative variables
3. Factorization For Categorical Variables
4. Handling of Missing or Inaccurate Values

We will discuss the process of each step below.

Data Import

We have obtained a dataset which includes 35 variables with 1470 observations. The dataset is will be referred to as `talentMgmtData`. In order to better understand our data we needed to proceed with cleaning it appropriately.

Data Type Conversion for quantitative variables

Analysis could not be done properly if our variables are not in the correct type. Several of the numerical based variables showed up as doubles which was not necessarily appropriate for research and analysis. For example, the years of experience for specific employees needed to be converted to integer types, and Standard Working hours had no reason to show as a double, so it to was converted to an integer. This process was followed for each one of the 35 variables on order to ensure each columns assigned data type made sense for the context of the study.

Factorization For Catagorical Variables

Within our `talentMgmtData` we noticed that there were several columns that were improperly listed as integers when they were simply categories. In order for us to be able to do factor analysis we picked out the columns that would be better suited to be categories instead of integers so that we could get a better understanding of how our talent was spread out over different situations. For example, Columns like Department, BusinessTravel, Over18, & Job satisfaction to name a few, are all better suited to be treated as categorical variables for our research purposes. As a result, we have converted the appropriate columns that explained how our data points were separated by making them factors instead of numerical values.

Handling of Missing or Inaccurate Values

In order for our mathematical calculations to work we had to convert many of our data points into numerical values so that our regression analysis would be more accurate. For example, DailyRate, Total Working Years, Years with Current Manager, are all examples of numerical based data points that can have mathematical operations performed on them. As a result we located variables like these and made sure the data type was appropriate giving us more flexibility during analysis.

DATA CLEANING

```
## [1] 1470   35
```

```
##  [1] "Age"                       "Attrition"
##  [3] "BusinessTravel"            "DailyRate"
##  [5] "Department"                "DistanceFromHome"
##  [7] "Education"                 "EducationField"
##  [9] "EmployeeCount"             "EmployeeNumber"
## [11] "EnvironmentSatisfaction"   "Gender"
## [13] "HourlyRate"                "JobInvolvement"
## [15] "JobLevel"                  "JobRole"
## [17] "JobSatisfaction"           "MaritalStatus"
## [19] "MonthlyIncome"             "MonthlyRate"
## [21] "NumCompaniesWorked"        "Over18"
## [23] "OverTime"                  "PercentSalaryHike"
## [25] "PerformanceRating"         "RelationshipSatisfaction"
## [27] "StandardHours"             "StockOptionLevel"
## [29] "TotalWorkingYears"         "TrainingTimesLastYear"
## [31] "WorkLifeBalance"           "YearsAtCompany"
## [33] "YearsInCurrentRole"        "YearsSinceLastPromotion"
## [35] "YearsWithCurrManager"
```

```
##  [1] "Age"          "Attrition"    "BusinessTrvl" "DailyRate"
##  [5] "Department"   "DistFromHome" "YrsOfEdu"     "EduField"
##  [9] "EmployeeCnt"  "EmployeeNum"  "EnvSatfctn"   "Gender"
## [13] "HourlyRate"   "JobInvolmnt"  "JobLevel"     "JobRole"
## [17] "JobSatfctn"   "MaritalStat"  "MonthlyIncm"  "MonthlyRate"
## [21] "NumCmpWorked" "Over18"       "OverTime"     "PrcntSalHike"
## [25] "PerfRating"   "RlnSatfctn"   "StandardHrs"  "StockOptLvl"
## [29] "TtlWrkngYrs"  "TrngTmsLstYr" "WrkLifeBal"   "YrsAtCompany"
## [33] "YrsInCrntRl"  "YrsSncLstPrn" "YrsWthCurMgr"
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
## Observations: 1,470
## Variables: 31
## $ Age          <dbl> 41, 49, 37, 33, 27, 32, 59, 30, 38, 36, 35, 29, 3...
## $ Attrition    <chr> "Yes", "No", "Yes", "No", "No", "No", "No", "No",...
## $ BusinessTrvl <chr> "Travel_Rarely", "Travel_Frequently", "Travel_Rar...
```

```
## $ DailyRate    <dbl> 1102, 279, 1373, 1392, 591, 1005, 1324, 1358, 216...
## $ Department   <chr> "Sales", "Research & Development", "Research & De...
## $ DistFromHome <dbl> 1, 8, 2, 3, 2, 2, 3, 24, 23, 27, 16, 15, 26, 19, ...
## $ YrsOfEdu     <dbl> 2, 1, 2, 4, 1, 2, 3, 1, 3, 3, 3, 2, 1, 2, 3, 4, 2...
## $ EduField     <chr> "Life Sciences", "Life Sciences", "Other", "Life ...
## $ EnvSatfctn   <dbl> 2, 3, 4, 4, 1, 4, 3, 4, 4, 3, 1, 4, 1, 2, 3, 2, 1...
## $ Gender       <chr> "Female", "Male", "Male", "Female", "Male", "Male...
## $ HourlyRate   <dbl> 94, 61, 92, 56, 40, 79, 81, 67, 44, 94, 84, 49, 3...
## $ JobInvolmnt  <dbl> 3, 2, 2, 3, 3, 3, 4, 3, 2, 3, 4, 2, 3, 3, 2, 4, 4...
## $ JobLevel     <dbl> 2, 2, 1, 1, 1, 1, 1, 1, 3, 2, 1, 2, 1, 1, 1, 3, 1...
## $ JobRole      <chr> "Sales Executive", "Research Scientist", "Laborat...
## $ JobSatfctn   <dbl> 4, 2, 3, 3, 2, 4, 1, 3, 3, 3, 2, 3, 3, 4, 3, 1, 2...
## $ MaritalStat  <chr> "Single", "Married", "Single", "Married", "Marrie...
## $ MonthlyIncm  <dbl> 5993, 5130, 2090, 2909, 3468, 3068, 2670, 2693, 9...
## $ MonthlyRate  <dbl> 19479, 24907, 2396, 23159, 16632, 11864, 9964, 13...
## $ NumCmpWorked <dbl> 8, 1, 6, 1, 9, 0, 4, 1, 0, 6, 0, 0, 1, 0, 5, 1, 0...
## $ OverTime     <chr> "Yes", "No", "Yes", "Yes", "No", "No", "Yes", "No...
## $ PrcntSalHike <dbl> 11, 23, 15, 11, 12, 13, 20, 22, 21, 13, 13, 12, 1...
## $ PerfRating   <dbl> 3, 4, 3, 3, 3, 3, 4, 4, 4, 3, 3, 3, 3, 3, 3, 3, 3...
## $ RlnSatfctn   <dbl> 1, 4, 2, 3, 4, 3, 1, 2, 2, 2, 3, 4, 4, 3, 2, 3, 4...
## $ StockOptLvl  <dbl> 0, 1, 0, 0, 1, 0, 3, 1, 0, 2, 1, 0, 1, 1, 0, 1, 2...
## $ TtlWrkngYrs  <dbl> 8, 10, 7, 8, 6, 8, 12, 1, 10, 17, 6, 10, 5, 3, 6,...
## $ TrngTmsLstYr <dbl> 0, 3, 3, 3, 3, 2, 3, 2, 2, 3, 5, 3, 1, 2, 4, 1, 5...
## $ WrkLifeBal   <dbl> 1, 3, 3, 3, 3, 2, 2, 3, 3, 2, 3, 3, 2, 3, 3, 3, 2...
## $ YrsAtCompany <dbl> 6, 10, 0, 8, 2, 7, 1, 1, 9, 7, 5, 9, 5, 2, 4, 10,...
## $ YrsInCrntRl  <dbl> 4, 7, 0, 7, 2, 7, 0, 0, 7, 7, 4, 5, 2, 2, 2, 9, 2...
## $ YrsSncLstPrn <dbl> 0, 1, 0, 3, 2, 3, 0, 0, 1, 7, 0, 0, 4, 1, 0, 8, 0...
## $ YrsWthCurMgr <dbl> 5, 7, 0, 0, 2, 6, 0, 0, 8, 7, 3, 8, 3, 2, 3, 8, 5...

## [1] FALSE

## [1] "Travel_Rarely"     "Travel_Frequently" "Non-Travel"

## [1] "Sales Executive"           "Research Scientist"
## [3] "Laboratory Technician"     "Manufacturing Director"
## [5] "Healthcare Representative" "Manager"
## [7] "Sales Representative"      "Research Director"
## [9] "Human Resources"

## [1] "Yes" "No"

## [1] "Female" "Male"

## [1] "Single"   "Married"  "Divorced"

## [1] "Yes" "No"

## [1] "Sales"                  "Research & Development"
## [3] "Human Resources"

## [1] "Life Sciences"    "Other"            "Medical"
## [4] "Marketing"        "Technical Degree" "Human Resources"

## Observations: 1,470
## Variables: 31
## $ Age          <int> 41, 49, 37, 33, 27, 32, 59, 30, 38, 36, 35, 29, 3...
## $ Attrition    <fctr> Yes, No, Yes, No, No, No, No, No, No, No, No, No...
## $ BusinessTrvl <fctr> Travel_Rarely, Travel_Frequently, Travel_Rarely,...
## $ DailyRate    <dbl> 1102, 279, 1373, 1392, 591, 1005, 1324, 1358, 216...
```

```
## $ Department   <fctr> Sales, Research & Development, Research & Develo...
## $ DistFromHome <int> 1, 8, 2, 3, 2, 2, 3, 24, 23, 27, 16, 15, 26, 19, ...
## $ YrsOfEdu     <int> 2, 1, 2, 4, 1, 2, 3, 1, 3, 3, 3, 2, 1, 2, 3, 4, 2...
## $ EduField     <fctr> Life Sciences, Life Sciences, Other, Life Scienc...
## $ EnvSatfctn   <fctr> 2, 3, 4, 4, 1, 4, 3, 4, 4, 3, 1, 4, 1, 2, 3, 2, ...
## $ Gender       <fctr> Female, Male, Male, Female, Male, Male, Female, ...
## $ HourlyRate   <dbl> 94, 61, 92, 56, 40, 79, 81, 67, 44, 94, 84, 49, 3...
## $ JobInvolmnt  <fctr> 3, 2, 2, 3, 3, 3, 4, 3, 2, 3, 4, 2, 3, 3, 2, 4, ...
## $ JobLevel     <fctr> 2, 2, 1, 1, 1, 1, 1, 1, 3, 2, 1, 2, 1, 1, 1, 3, ...
## $ JobRole      <fctr> Sales Executive, Research Scientist, Laboratory ...
## $ JobSatfctn   <fctr> 4, 2, 3, 3, 2, 4, 1, 3, 3, 3, 2, 3, 3, 4, 3, 1, ...
## $ MaritalStat  <fctr> Single, Married, Single, Married, Married, Singl...
## $ MonthlyIncm  <dbl> 5993, 5130, 2090, 2909, 3468, 3068, 2670, 2693, 9...
## $ MonthlyRate  <dbl> 19479, 24907, 2396, 23159, 16632, 11864, 9964, 13...
## $ NumCmpWorked <int> 8, 1, 6, 1, 9, 0, 4, 1, 0, 6, 0, 0, 1, 0, 5, 1, 0...
## $ OverTime     <fctr> Yes, No, Yes, Yes, No, No, Yes, No, No, No, No, ...
## $ PrcntSalHike <dbl> 11, 23, 15, 11, 12, 13, 20, 22, 21, 13, 13, 12, 1...
## $ PerfRating   <fctr> 3, 4, 3, 3, 3, 3, 4, 4, 4, 3, 3, 3, 3, 3, 3, 3, ...
## $ RlnSatfctn   <fctr> 1, 4, 2, 3, 4, 3, 1, 2, 2, 2, 3, 4, 4, 3, 2, 3, ...
## $ StockOptLvl  <fctr> 0, 1, 0, 0, 1, 0, 3, 1, 0, 2, 1, 0, 1, 1, 0, 1, ...
## $ TtlWrkngYrs  <int> 8, 10, 7, 8, 6, 8, 12, 1, 10, 17, 6, 10, 5, 3, 6,...
## $ TrngTmsLstYr <int> 0, 3, 3, 3, 3, 2, 3, 2, 2, 3, 5, 3, 1, 2, 4, 1, 5...
## $ WrkLifeBal   <fctr> 1, 3, 3, 3, 3, 2, 2, 3, 3, 2, 3, 3, 2, 3, 3, 3, ...
## $ YrsAtCompany <int> 6, 10, 0, 8, 2, 7, 1, 1, 9, 7, 5, 9, 5, 2, 4, 10,...
## $ YrsInCrntRl  <int> 4, 7, 0, 7, 2, 7, 0, 0, 7, 7, 4, 5, 2, 2, 2, 9, 2...
## $ YrsSncLstPrn <int> 0, 1, 0, 3, 2, 3, 0, 0, 1, 7, 0, 0, 4, 1, 0, 8, 0...
## $ YrsWthCurMgr <int> 5, 7, 0, 0, 2, 6, 0, 0, 8, 7, 3, 8, 3, 2, 3, 8, 5...
```

Preliminary Analysis

3.A

```
## [1] 19 60
```

As we can see above, only Ages 19 to 60 exists within our dataset.

Now that the data has been prepared and formatted accordingly it is necessary to explore it to its depths.

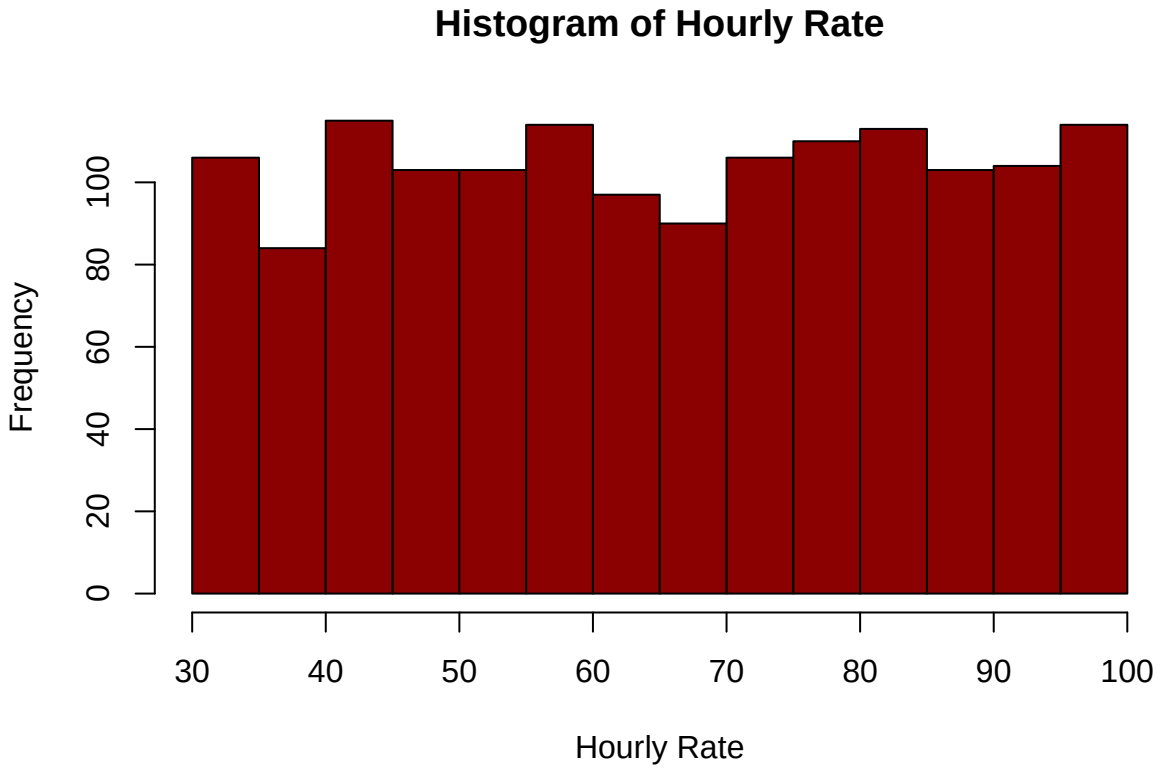
3.B Descriptive statistics

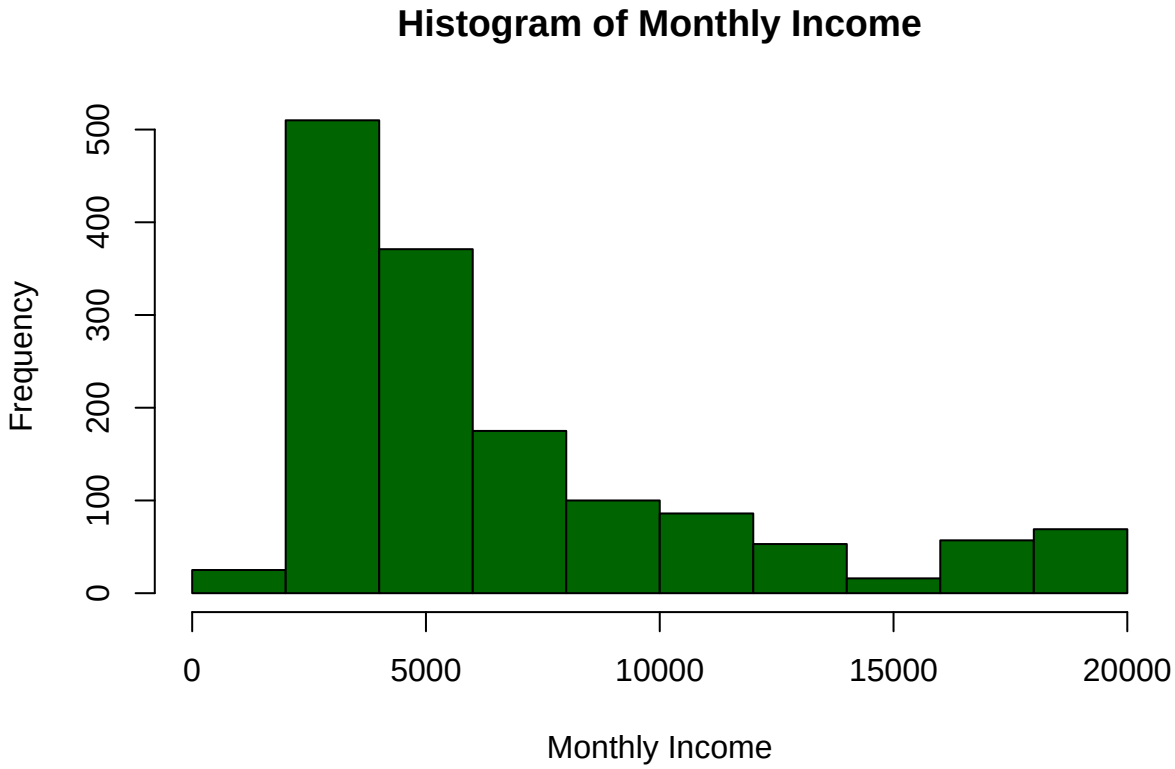
Lets check descriptive statistics for some of the data we have

DescriptiveStatistics	HourlyRate	MonthlyRate	MonthlyIncome	TotalWorkYears	YearsAtCompany	Age	YrsOfEdu
min	30.00000	2094.000	1009.000	0.000000	0.000000	19.000000	1.000000
max	100.00000	26999.000	19999.000	40.000000	40.000000	60.000000	5.000000
mean	65.87893	14312.212	6530.207	11.341313	7.046512	37.027360	2.915185
sd	20.34338	7124.669	4706.273	7.757034	6.121228	9.052093	1.025173

The average Hourly rate is $65.88/hour, the average Monthly Rate is $14,312.21/month, the average Monthly Income is $6,530.21/month, average Total Worked Years is 11.34 years, average Worked years at the company is 7.04 years, average Age of the employees is 37 years and average Years of Education of the employees is 2.9 years ('Bachelor' degree).

Lets check histograms of Hourly Rate and Monthly income.





On the histograms we can see almost equal spread of hourly rates within the company, but we can not say the same about monthly income, it means that employees work different amount of hours (some of them are part time, and some of them work with overtime (more then 40hours), we do not have information if any

bonuses were paid in the company, so it does not make sense to continue analyze working hours). Histogram of income shows a right skewed distribution for our Monthly Income in our dataset. It is also clear that the majority of the population in this dataset makes between $1000 and $6000 per month. The higher we go out in income the more the distribution becomes narrower.

3.C

Understanding Gender, Education, & Occupations

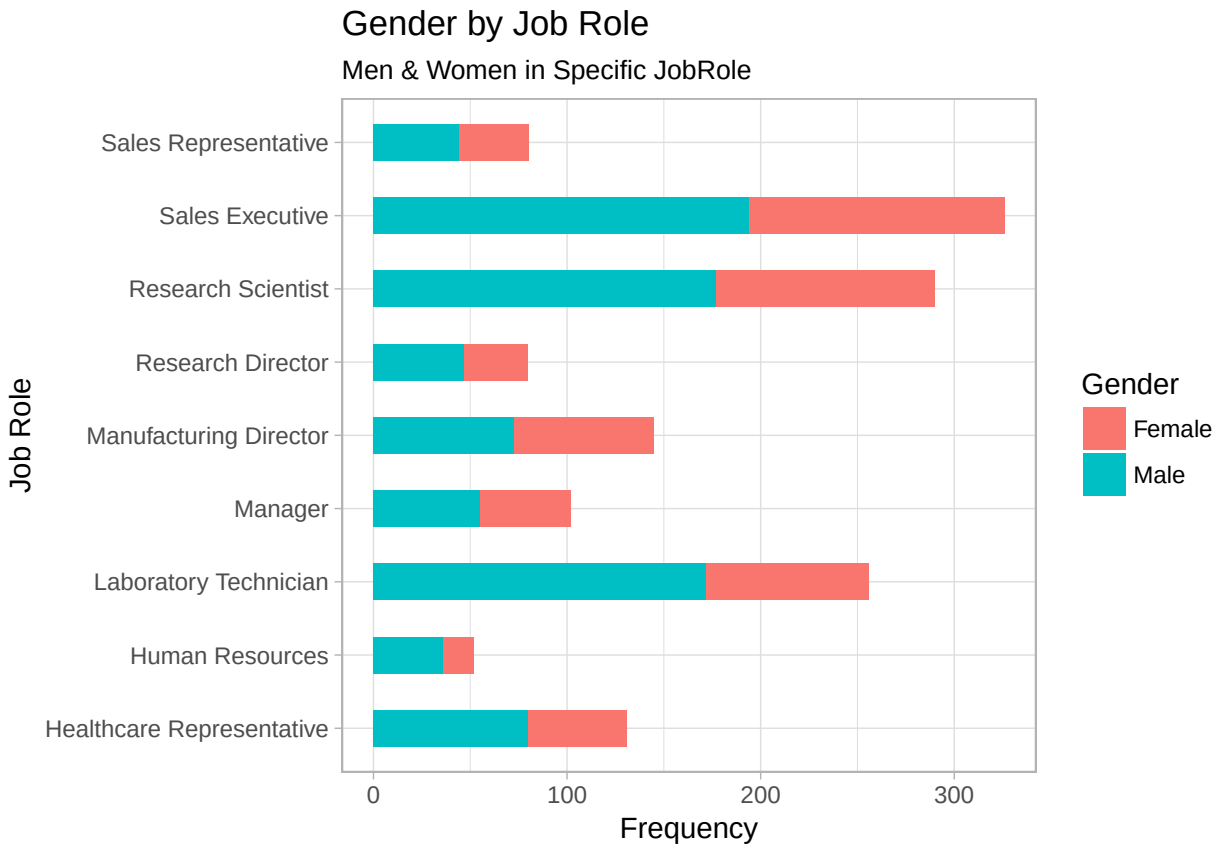
Next we explore how Gender, Education and Job Role is broken down within our dataset. Below are frequency tables of our findings for these three categories...

Gender	Frequency
Female	584
Male	878

Education	Frequency
Human Resources	27
Life Sciences	603
Marketing	158
Medical	460
Other	82
Technical Degree	132

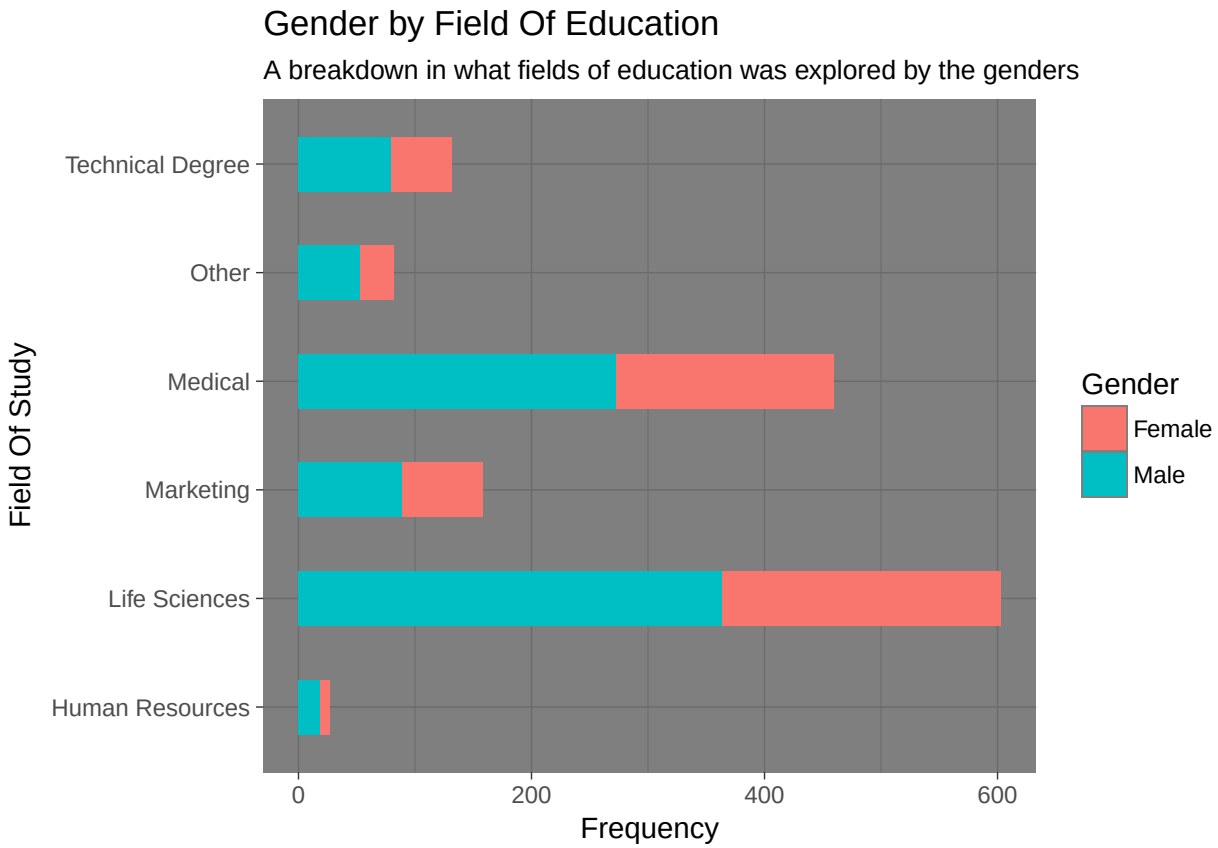
Job Role	Frequency
Healthcare Representative	131
Human Resources	52
Laboratory Technician	256
Manager	102
Manufacturing Director	145
Research Director	80
Research Scientist	290
Sales Executive	326
Sales Representative	80

Next, we examine the frequency for gender and how it is distributed across job roles visually.



Our chart shows a bimodal distribution of our job roles. What is interesting here is we can see that sales executive is the most commonly occurring job in our dataset with more than 300 people represented in that category. It is also interesting to note that the spread between male and female in that job category looks almost equally represented. Our Lowest category is Human Resources Which has partitioned of mostly males even though there are just under 50 people in this category as a whole.

What about Education? How is education represented across the genders. We take a look at that distribution next.



We can see from this chart that Human Resources has the lowest participation in terms of education which actually makes sense given that it is our lowest filled job role. Life Science seems to be the most popular field of study between all the listed education choices with 600 different people in our dataset who studied in this field. This does not necessarily match with our discovery regarding our popular job role. Life Science skills can translate into making great Sales Executives, but this educational field of study does not seem to be directly related to the Sales Executive.

3.D

Finally, we take a look at a frequency Table for all the different management positions.

ManagementOnly	Frequency
Manager	102
Manufacturing Director	145
Research Director	80

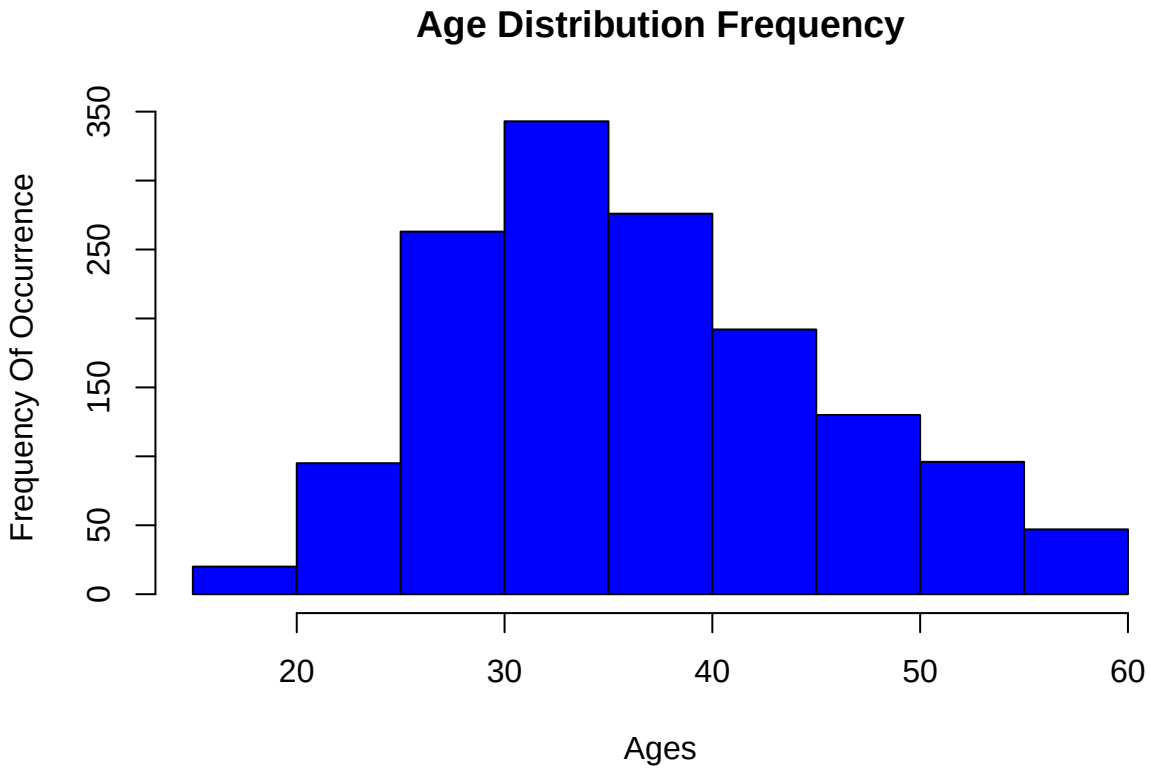
Now that we have an overall view of our data now we can begin to try to answer our questions of interest regarding our dataset. Our questions of interest are simply the following...

1. What are the factors that lead to employee attrition?
2. Can employee attrition be slowed, or avoided all together?

Our next section will use all of our recent discoveries about the `talentData` to answer these inquiries.

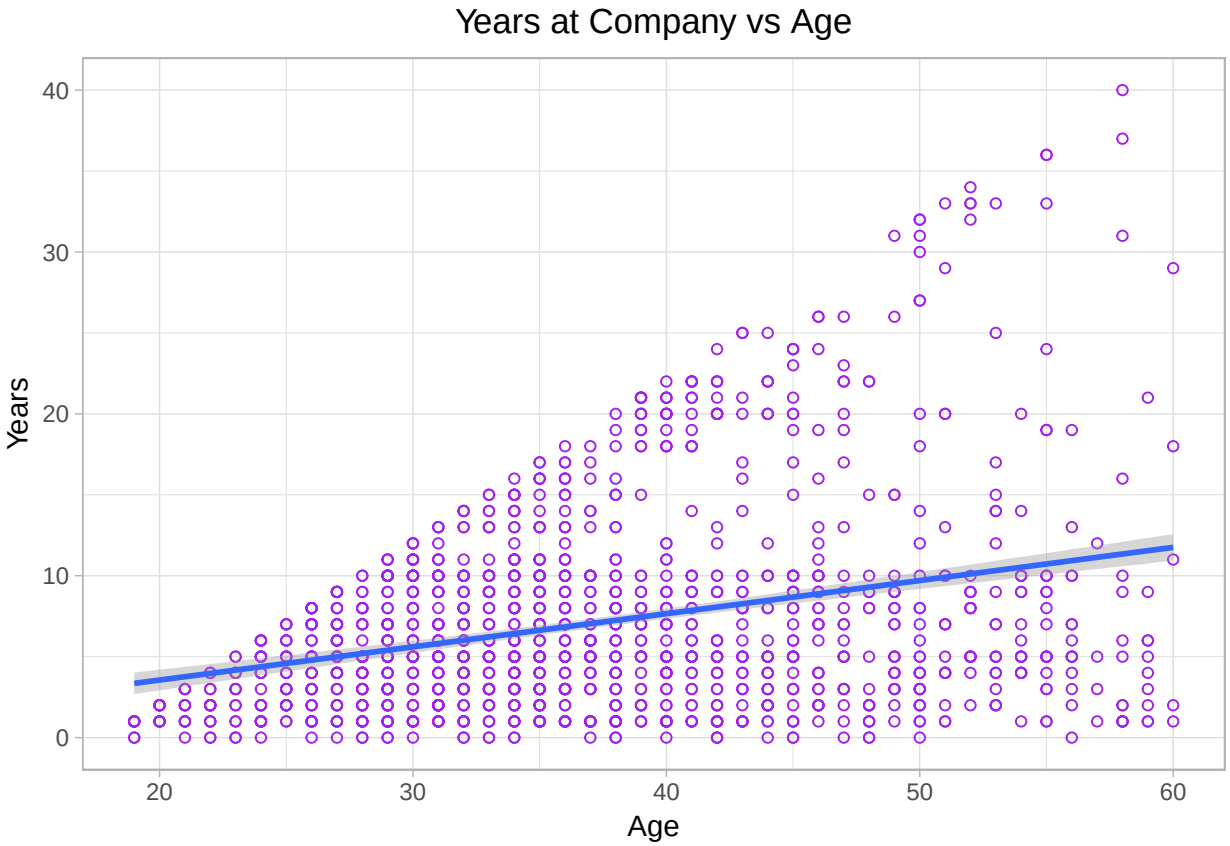
IV. Deeper Analysis and Visualization

When it comes to jobs the first thing that we want to look at is our age distribution. This is an important step in our EDA process as we would like to get an idea of how old or young our these individuals in our entire dataset as it might give us a good place to start. We are only interested in exploring individuals in the workforce that are older than 18, so all of the forward analysis will take this constraint into consideration.



Based on the chart above we can see that our age range is between 20 and 60 with a large portion of ages being between 25 and 45. This is insightful as it might help us with interpretation of our findings moving forward. It is also worth noting that between 30 and 35 is the most occurring age of all the age groups with over 350 recorded ages within this group!

Another thing that we would like to explore is whether or not there is a relationship between our Ages and the number of years they have spent at a specific company. If there is a relationship, we would like to understand whether or not the relationship is positive or negatively correlated.



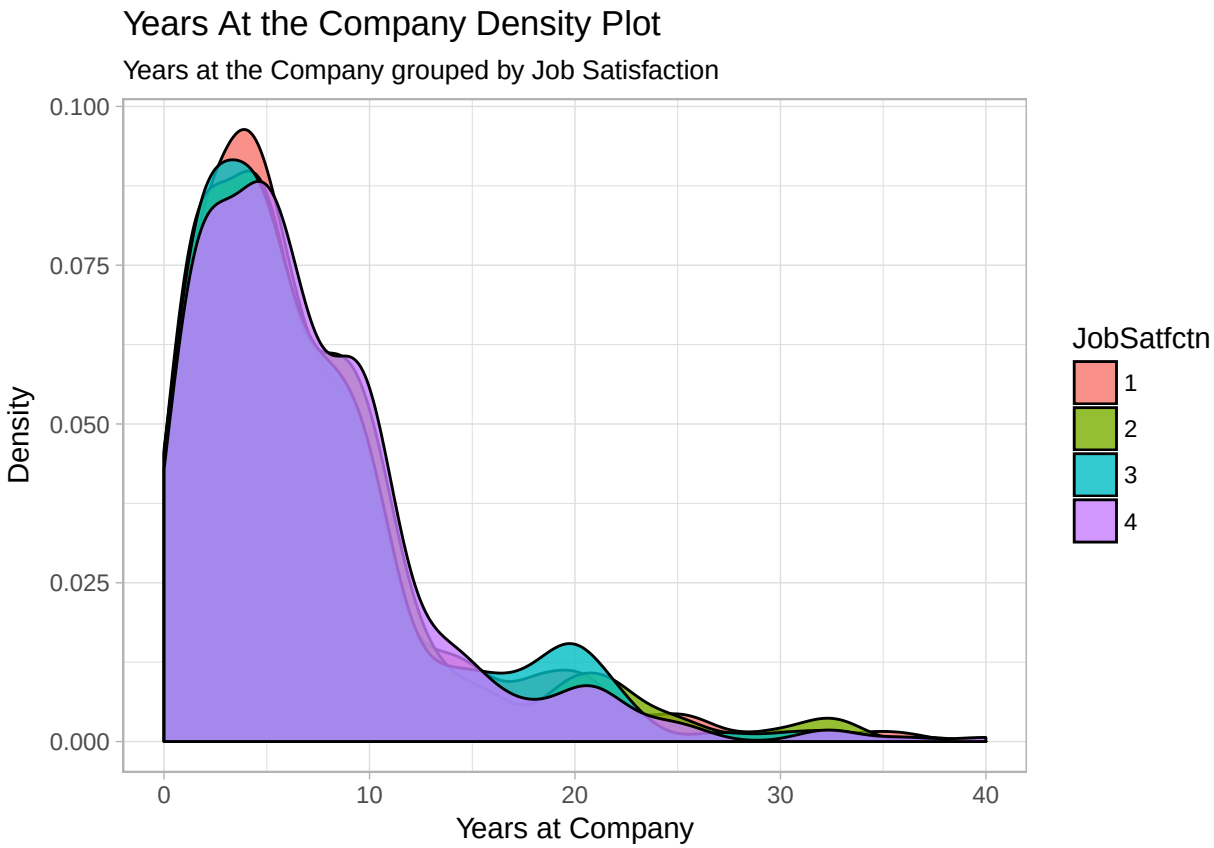
Our findings here are not surprising. We notice that visually there might be evidence of a positive linear relationship between age and years at a specific company. This finding suggests that the older you are the more likely you are to have a greater number of years at a specific company. To know for sure we examine Pearson correlation between `YrsAtCompany` and `Age`.

```
##
##  Pearson's product-moment correlation
##
## data:  talentData$Age and talentData$YrsAtCompany
## t = 12.148, df = 1460, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.2556936 0.3488375
## sample estimates:
##      cor
## 0.302989
```

Based on the results of our correlation test we have a Pearsons correlation value of `0.302989` (95% CI: 0.25 to 0.34) which is more evidence of a positive linear relationship between Age and Years at a specific company. It is important for us to keep this relationship in mind moving forward for the rest of the study.

Exploring the Relationship between Years At a the company and Satisfaction

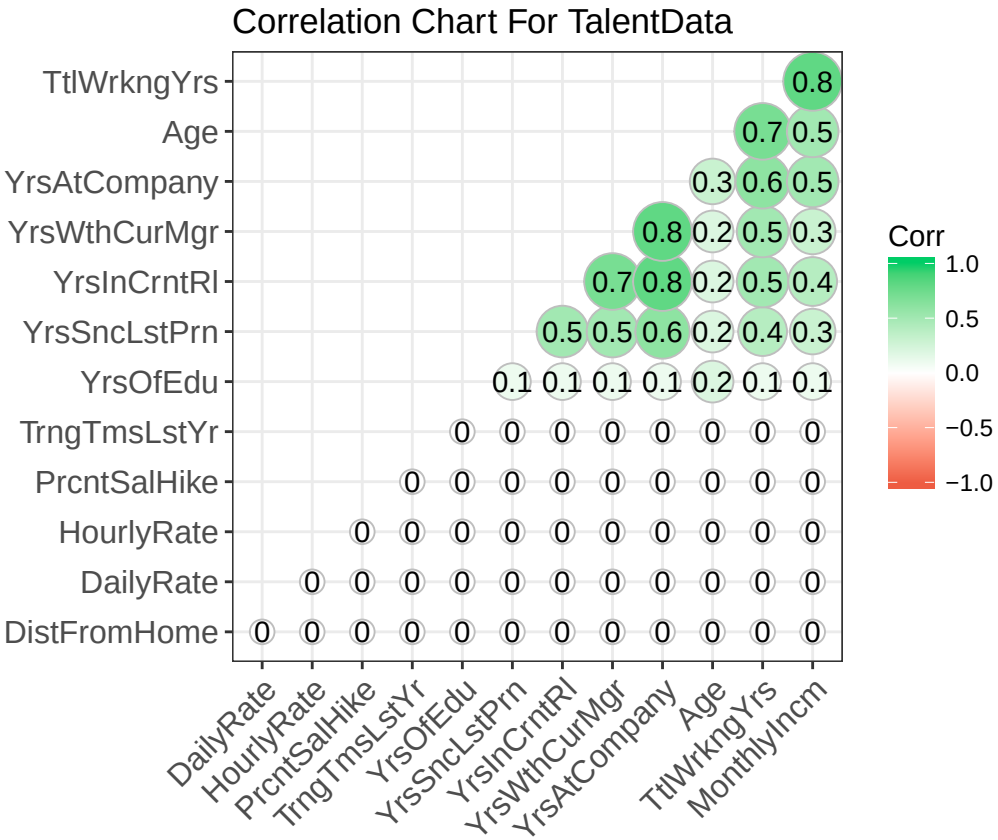
Now that we know that Age, and Income have a positive impact on employee retention we would like to visually confirm our assumption that Job Satisfaction also contributes in a major way to someone remaining at a company. We can do this by examining our different levels of employee satisfaction and their Years they have remained at a particular company.



The Chart above is very telling. We take a look at our probability distribution across different Job Satisfaction levels. 1 indicates that there is low employee job satisfaction, and 4 represents that there is really high employee job satisfaction. If we examine the probability of each within the context of the years an individual stays at a company it becomes clear that lower job satisfaction indicates that this category has the lowest number of years spent at a company. This is no surprise, but it does give us more information regarding negative factors to employee attrition.

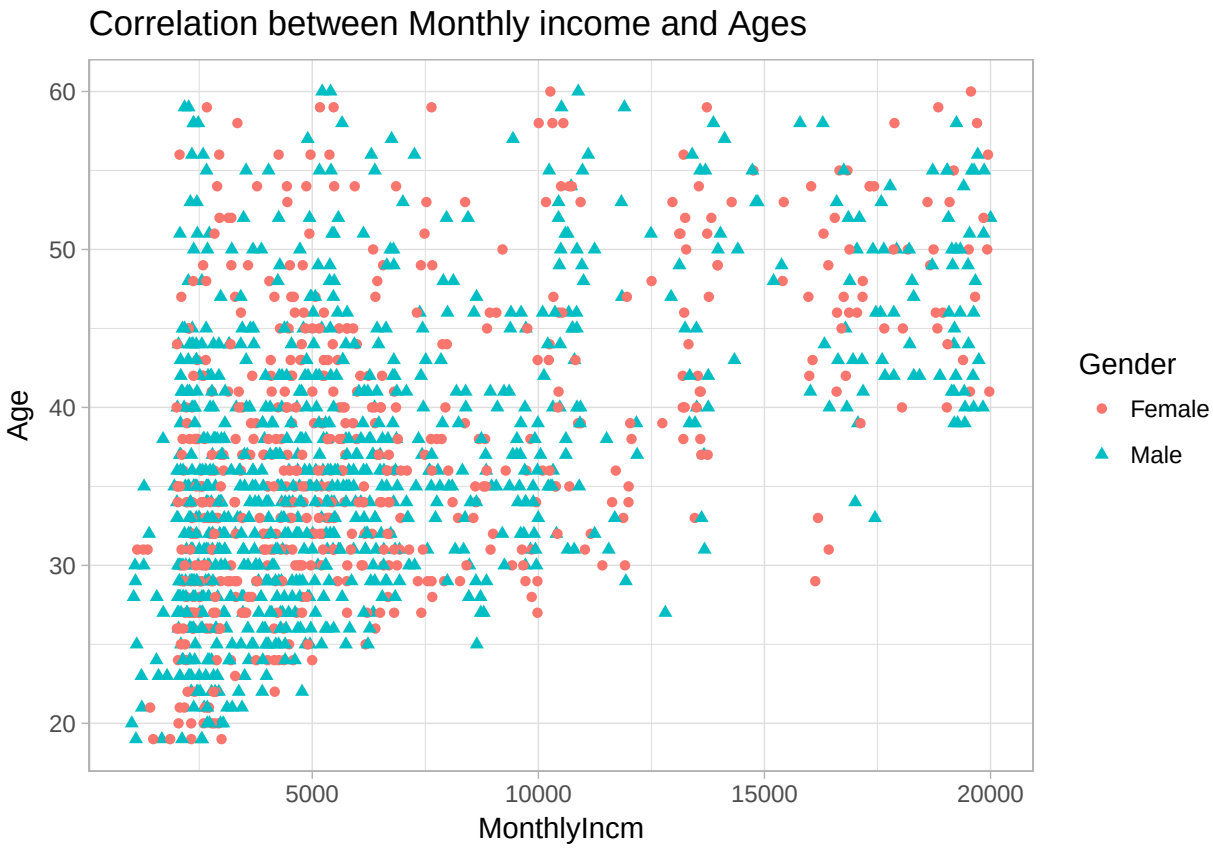
Correlation Plot of TalentData

In our data set there are a large number of numerical based values. We would like to get an idea of the relationship of those variables relate to one another in one glance. Below is a comprised chart of all the correlations regarding our talent data.



The chart above shows us that there are no negative linear relationships in our data between our variables. We can also see that there are some positive relationships between variables that are linearly correlated and we have now been able to narrow these down so they can be examined in depth.

4.C Let's check if there is any relationship between Age and Income. Does Gender makes any effect on the Monthly income?




```
##
## Call:
## lm(formula = MonthlyIncm ~ Age + Gender, data = talentData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9927.2 -2623.9  -711.1  1817.3 12595.1
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2872.67     475.14  -6.046 1.88e-09 ***
## Age           256.12      11.85  21.616  < 2e-16 ***
## GenderMale   -134.31     218.91  -0.614     0.54
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4096 on 1459 degrees of freedom
## Multiple R-squared:  0.2434, Adjusted R-squared:  0.2424
## F-statistic: 234.7 on 2 and 1459 DF,  p-value: < 2.2e-16
```

From regression analysis above we can say that Gender does not make significant change in the Monthly employee income. But Age is indeed significant variable ($p<0.0001$), it can explain 24% of monthly income change.

4D. What Factors Cause Employee Turnover?

Our goal is to determine which indicators might lead to employee attrition. The best way for us to find those is to create a regression model for Attrition prediction.

Let's use stepwise selection method to come up with the model which has only significant variables.

```
##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select
```

Step	Df	Deviance	Resid. Df	Resid. Dev	AIC
	NA	NA	1402	767.0627	887.0627
- PerfRating	1	0.0828371	1403	767.1455	885.1455
- PrcntSalHike	1	0.1933362	1404	767.3389	883.3389
- MaritalStat	2	2.2988539	1406	769.6377	881.6377
- Department	2	2.1596799	1408	771.7974	879.7974
- MonthlyRate	1	0.3098227	1409	772.1072	878.1072
- YrsOfEdu	1	0.4950893	1410	772.6023	876.6023
- HourlyRate	1	0.6259302	1411	773.2282	875.2282
- MonthlyIncm	1	1.7260800	1412	774.9543	874.9543

```
##
## Call:
## glm(formula = Attrition ~ Age + BusinessTrvl + DailyRate + DistFromHome +
##     EduField + EnvSatfctn + Gender + JobInvolmnt + JobLevel +
##     JobRole + JobSatfctn + NumCmpWorked + OverTime + RlnSatfctn +
##     StockOptLvl + TtlWrkngYrs + TrngTmsLstYr + WrkLifeBal + YrsAtCompany +
##     YrsInCrntRl + YrsSncLstPrn + YrsWthCurMgr, family = binomial,
##     data = talentData)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7445  -0.4451  -0.2018  -0.0622   3.4579
##
## Coefficients:
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept)                   3.6181563  1.2706264   2.848 0.004406 **
## Age                          -0.0260159  0.0141398  -1.840 0.065781 .
## BusinessTrvlTravel_Frequently  2.2155197  0.4504153   4.919 8.71e-07 ***
## BusinessTrvlTravel_Rarely      1.1947786  0.4150545   2.879 0.003994 **
## DailyRate                     -0.0004036  0.0002330  -1.733 0.083178 .
## DistFromHome                   0.0565932  0.0115241   4.911 9.07e-07 ***
## EduFieldLife Sciences         -0.7714993  0.8046792  -0.959 0.337676
## EduFieldMarketing             -0.2933875  0.8518262  -0.344 0.730529
## EduFieldMedical               -0.7990737  0.8063506  -0.991 0.321698
## EduFieldOther                 -0.7016476  0.8824394  -0.795 0.426542
## EduFieldTechnical Degree       0.3523726  0.8278938   0.426 0.670381
## EnvSatfctn2                   -1.1018306  0.2921475  -3.771 0.000162 ***
## EnvSatfctn3                   -1.2434005  0.2695765  -4.612 3.98e-06 ***
## EnvSatfctn4                   -1.3715407  0.2689086  -5.100 3.39e-07 ***
## GenderMale                     0.3754770  0.1951634   1.924 0.054366 .
```

```
## JobInvolmnt2                  -1.1922220  0.3705098  -3.218 0.001292 **
## JobInvolmnt3                  -1.5416306  0.3499986  -4.405 1.06e-05 ***
## JobInvolmnt4                  -2.1188441  0.4823939  -4.392 1.12e-05 ***
## JobLevel2                     -1.6674659  0.4475994  -3.725 0.000195 ***
## JobLevel3                     -0.4012516  0.5658105  -0.709 0.478224
## JobLevel4                     -1.6470379  0.9884982  -1.666 0.095673 .
## JobLevel5                      0.7434310  1.2848912   0.579 0.562863
## JobRoleHuman Resources         0.5335634  0.7531066   0.708 0.478645
## JobRoleLaboratory Technician   0.7308250  0.5930632   1.232 0.217841
## JobRoleManager                -0.9794199  1.0253470  -0.955 0.339472
## JobRoleManufacturing Director  0.4472787  0.5610328   0.797 0.425311
## JobRoleResearch Director      -2.1787101  1.0751884  -2.026 0.042729 *
## JobRoleResearch Scientist     -0.3235446  0.6129802  -0.528 0.597623
## JobRoleSales Executive         1.3251283  0.4753929   2.787 0.005313 **
## JobRoleSales Representative    1.3392198  0.6642588   2.016 0.043788 *
## JobSatfctn2                   -0.6838494  0.2862798  -2.389 0.016906 *
## JobSatfctn3                   -0.6777859  0.2538438  -2.670 0.007583 **
## JobSatfctn4                   -1.2901946  0.2699270  -4.780 1.75e-06 ***
## NumCmpWorked                   0.2101087  0.0410569   5.117 3.10e-07 ***
## OverTimeYes                    2.1658772  0.2095253  10.337  < 2e-16 ***
## RlnSatfctn2                   -0.9841531  0.2974752  -3.308 0.000938 ***
## RlnSatfctn3                   -1.0310791  0.2678812  -3.849 0.000119 ***
## RlnSatfctn4                   -1.0356753  0.2682320  -3.861 0.000113 ***
## StockOptLvl1                  -1.4613095  0.2185146  -6.687 2.27e-11 ***
## StockOptLvl2                  -1.4527612  0.3754485  -3.869 0.000109 ***
## StockOptLvl3                  -0.7644229  0.4069912  -1.878 0.060350 .
## TtlWrkngYrs                   -0.0688767  0.0306138  -2.250 0.024458 *
## TrngTmsLstYr                  -0.1707231  0.0762054  -2.240 0.025071 *
## WrkLifeBal2                   -0.9765924  0.3870782  -2.523 0.011636 *
## WrkLifeBal3                   -1.5026213  0.3646337  -4.121 3.77e-05 ***
## WrkLifeBal4                   -1.0953274  0.4412842  -2.482 0.013060 *
## YrsAtCompany                   0.0988123  0.0427345   2.312 0.020765 *
## YrsInCrntRl                   -0.1460829  0.0511317  -2.857 0.004277 **
## YrsSncLstPrn                   0.1664042  0.0454302   3.663 0.000249 ***
## YrsWthCurMgr                  -0.1413055  0.0504870  -2.799 0.005128 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1282.54  on 1461  degrees of freedom
## Residual deviance:  774.95  on 1412  degrees of freedom
## AIC: 874.95
##
## Number of Fisher Scoring iterations: 7

##
## Call:
## glm(formula = Attrition ~ BusinessTrvl + DistFromHome + EnvSatfctn +
##     JobInvolmnt + JobSatfctn + NumCmpWorked + OverTime + RlnSatfctn +
##     TtlWrkngYrs + WrkLifeBal + YrsAtCompany + YrsInCrntRl + YrsSncLstPrn +
##     YrsWthCurMgr, family = binomial, data = talentData)
##
## Deviance Residuals:
```

```
##     Min       1Q   Median       3Q      Max
## -1.9744  -0.5218  -0.3133  -0.1402   3.5586
##
## Coefficients:
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    1.372745   0.609779   2.251 0.024372 *
## BusinessTrvlTravel_Frequently  1.947783   0.402503   4.839 1.30e-06 ***
## BusinessTrvlTravel_Rarely      1.139439   0.376812   3.024 0.002495 **
## DistFromHome                   0.039682   0.009983   3.975 7.03e-05 ***
## EnvSatfctn2                   -1.069100   0.260644  -4.102 4.10e-05 ***
## EnvSatfctn3                   -1.121141   0.232287  -4.827 1.39e-06 ***
## EnvSatfctn4                   -1.214059   0.234053  -5.187 2.14e-07 ***
## JobInvolmnt2                  -1.085428   0.327927  -3.310 0.000933 ***
## JobInvolmnt3                  -1.532624   0.311002  -4.928 8.31e-07 ***
## JobInvolmnt4                  -2.140413   0.437132  -4.896 9.76e-07 ***
## JobSatfctn2                   -0.590929   0.253570  -2.330 0.019783 *
## JobSatfctn3                   -0.677072   0.224069  -3.022 0.002514 **
## JobSatfctn4                   -1.293745   0.243731  -5.308 1.11e-07 ***
## NumCmpWorked                   0.157428   0.035645   4.416 1.00e-05 ***
## OverTimeYes                    1.791276   0.177999  10.063  < 2e-16 ***
## RlnSatfctn2                   -0.725094   0.259866  -2.790 0.005267 **
## RlnSatfctn3                   -0.688730   0.233335  -2.952 0.003161 **
## RlnSatfctn4                   -0.888801   0.240851  -3.690 0.000224 ***
## TtlWrkngYrs                   -0.129640   0.021205  -6.114 9.74e-10 ***
## WrkLifeBal2                   -0.938501   0.331982  -2.827 0.004699 **
## WrkLifeBal3                   -1.188394   0.307358  -3.866 0.000110 ***
## WrkLifeBal4                   -0.791590   0.378381  -2.092 0.036435 *
## YrsAtCompany                   0.079836   0.036210   2.205 0.027467 *
## YrsInCrntRl                   -0.123868   0.043254  -2.864 0.004186 **
## YrsSncLstPrn                   0.153451   0.039009   3.934 8.36e-05 ***
## YrsWthCurMgr                  -0.113428   0.044172  -2.568 0.010233 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1282.5  on 1461  degrees of freedom
## Residual deviance:  944.9  on 1436  degrees of freedom
## AIC: 996.9
##
## Number of Fisher Scoring iterations: 6
```

AIC of the StepwiseModel is 874.95 and R is 1 - (Residual Deviance/Null Deviance)= 1 - 774.95/1285.54 = 0.4. AIC of the CustomModel is 996.9 and R is 1 - (Residual Deviance/Null Deviance)= 1 - 944.9/1285.54 = 0.26.

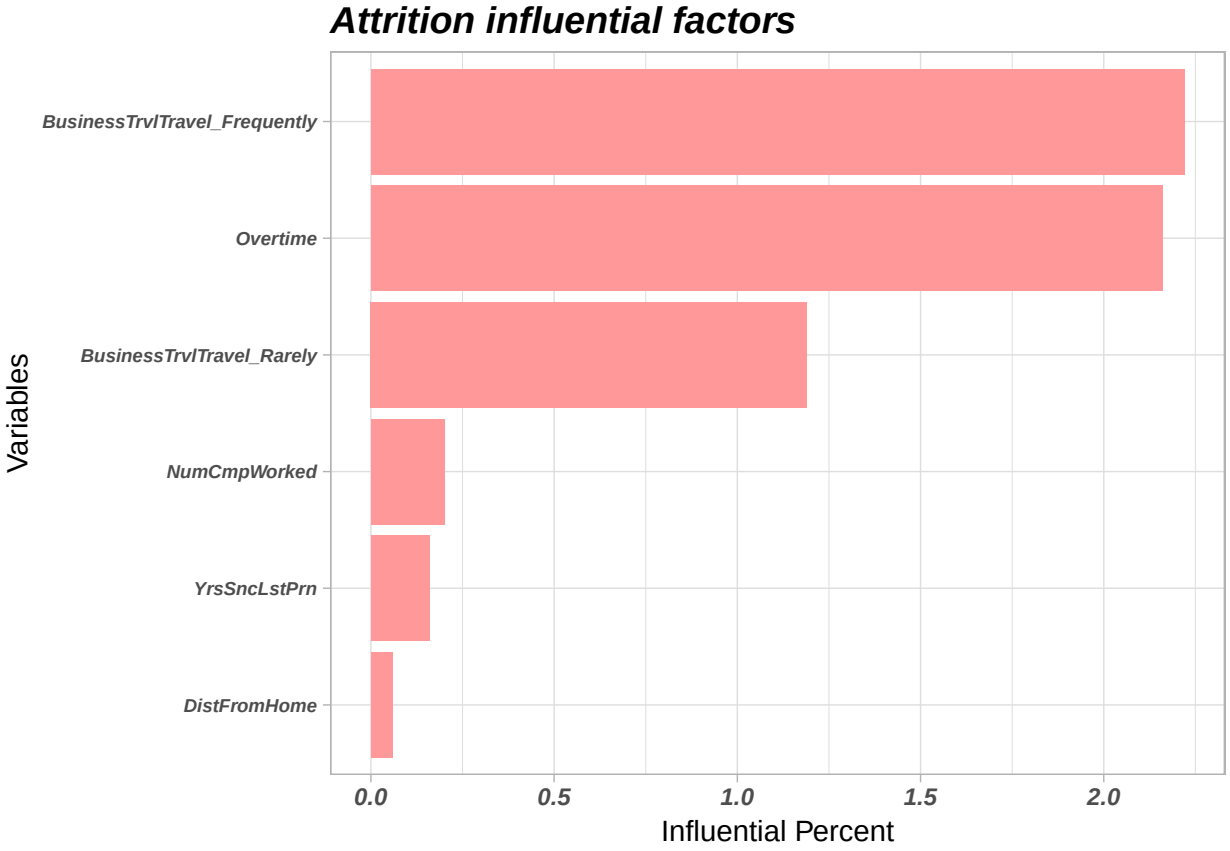
Model assumptions: Logistic regression does not make many of the key assumptions of linear regression and general linear models that are based on ordinary least squares algorithms particularly regarding linearity, normality, homoscedasticity, and measurement level. First, binary logistic regression requires the dependent variable to be binary - Assumption is met. Second, logistic regression requires the observations to be independent of each other. In other words, the observations should not come from repeated measurements or matched data - Assumption is met. Third, logistic regression requires there to be little or no multicollinearity among the independent variables. This means that the independent variables should not be too highly correlated with each other - Assumption is met. Fourth, logistic regression assumes linearity of independent variables - Assumption is met. Finally, logistic regression typically requires a large sample size - Assumption

is met (1470 observations).

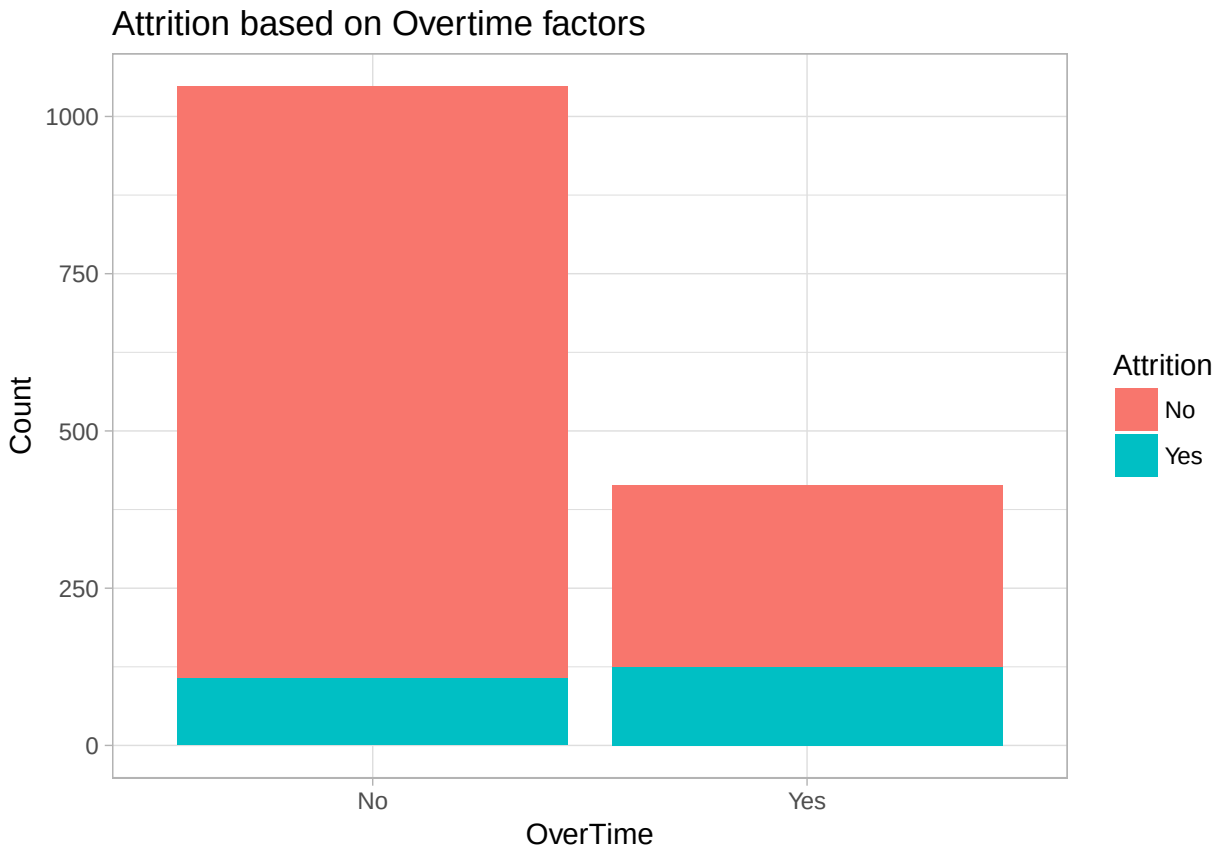
Interpretation of Stepwise Model

We definitely see that Stepwise Model is more predictive. Let's pick the most significant variables that may effect on attrition (positive slope will indicate increasing of attrition chance additively):

- Increasing of Overtime by 1 hour may predict that average attrition possibility will increase by 2.16
- BusinessTrvlTravel_Frequently may predict that average attrition possibility will increase by 2.22
- BusinessTrvlTravel_Rarely may predict that average attrition possibility will increase by 1.19
- Increasing of DistFromHome for 1 mile (assuming that distance were given in miles) may predict that average attrition possibility will increase by 0.06
- If employee has Job role Sales Executive we may predict that average attrition possibility will increase by 1.32
- If employee has Job role Sales Representative we may predict that average attrition possibility will increase by 1.34
- If number of companies where an employee worked increases by 1, we may predict that average attrition possibility will increase by 0.2
- Increasing of Years after last Promotion by 1 year may predict that average attrition possibility will increase by 0.16

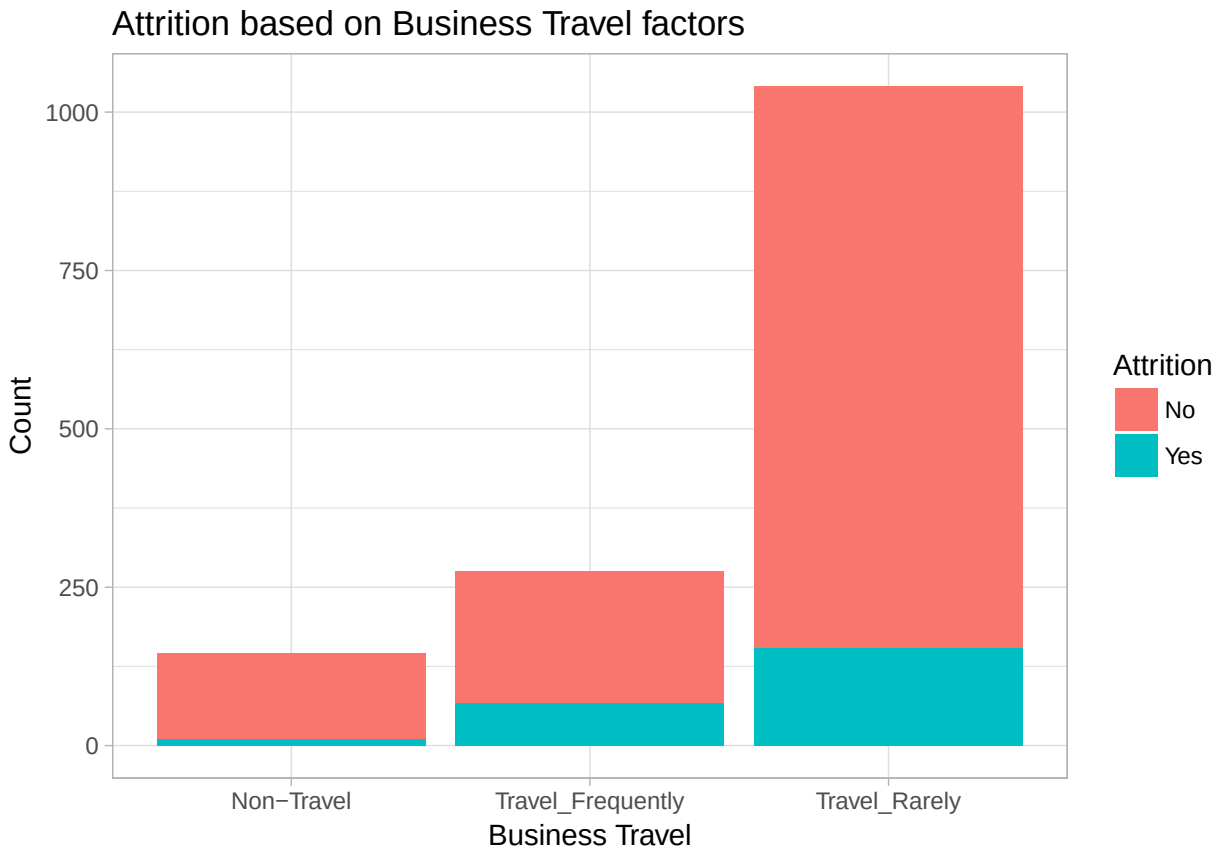


As we can see from the histogram above, the most influential factors for Attrition are Business Travels and Overtime. Let's see Overtime and Business Travels by Attrition factors on histograms and confirm this conclusion with actual percentage.



OverTime	Attrition	Freq	Attrition_Percent
No	No	940	0.00
Yes	No	289	0.00
No	Yes	108	10.31
Yes	**Yes**	**125**	**30.19**

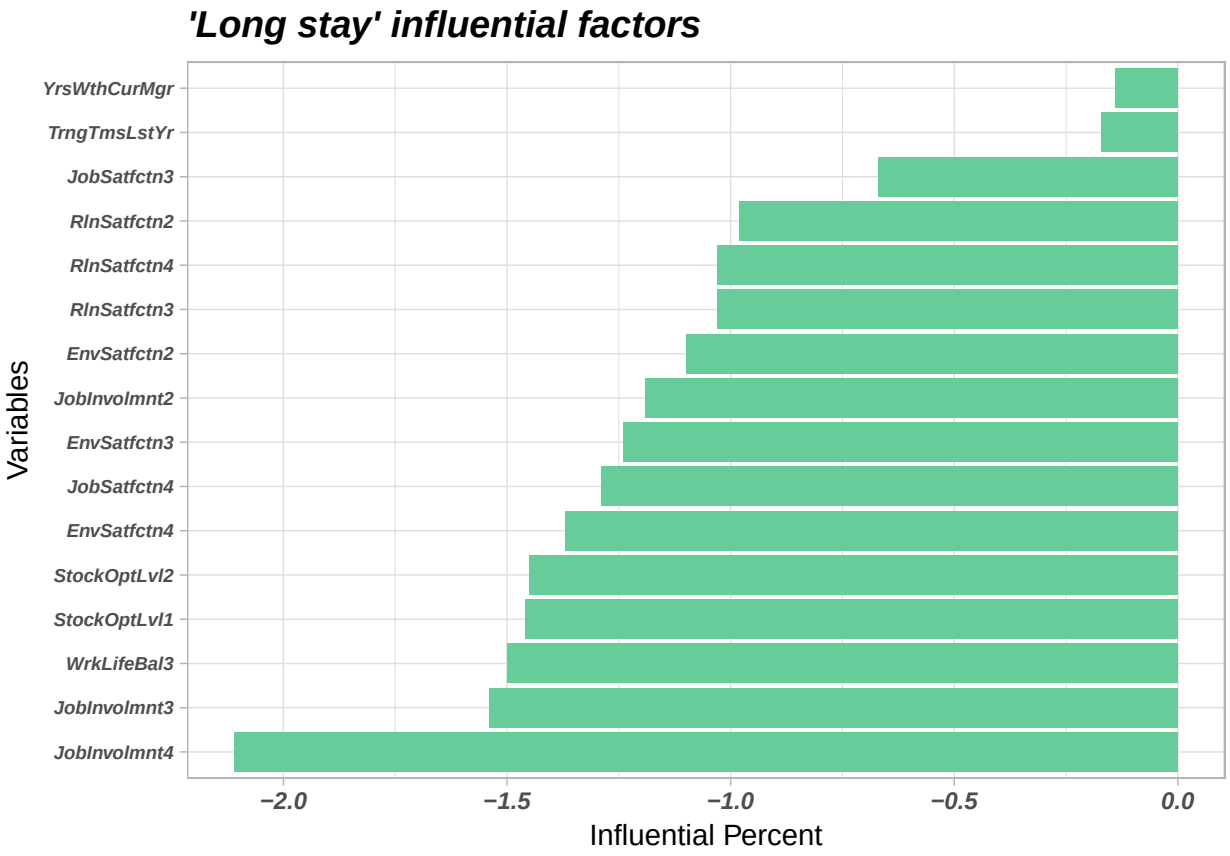
Attrition rate within employees with OverTime is 43.25%. It is 276% higher then attrition rate within employees without Overtime working hours.



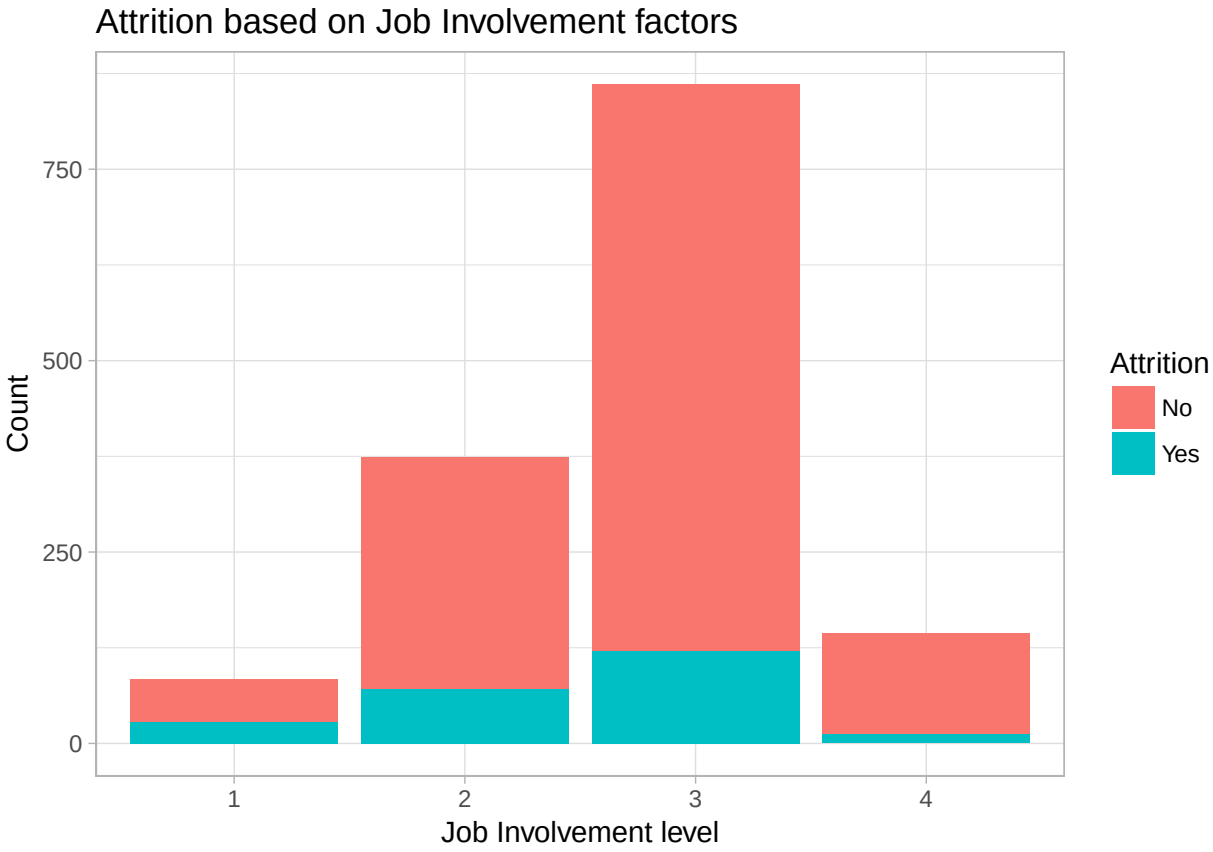
BusinessTrvl	Attrition	Freq	Attrition_Percent
Non-Travel	No	135	0.00
Travel_Frequently	No	208	0.00
Travel_Rarely	No	886	0.00
Non-Travel	Yes	11	7.53
Travel_Frequently	**Yes**	**67**	**24.36**
Travel_Rarely	**Yes**	**155**	**14.89**

Attrition rate within employees who has Frequent Business Travels is 32.21%, and 17.49% for those who has Rarely Business Travels. It is 295% and 114% higher then the attrition rate within NON travel employees, in respect to Frequent and Rarely Business Travels.

Please see below a histogram with variables that may have a good affect for “Long stay” employees, those who are satisfied with their working position and do not want to quit.



Lets confirm our findings with actual numbers.



JobInvolmnt	Attrition	Freq
1	No	55
2	No	303
3	No	740
4	No	131
1	Yes	28
2	Yes	71
3	Yes	121
4	Yes	13

JobInvolmnt	Attrition	Freq	Attrition_Percent
1	No	55	0.00
2	No	303	0.00
3	No	740	0.00
4	No	131	0.00
1	Yes	28	33.73
2	Yes	71	18.98
3	**Yes**	**121**	**14.05**
4	**Yes**	**13**	**9.03**

Average Attrition rate for those who has Job Involvement level “high” and “very high” is smaller by 50% comparing with those who has Job Involvement level “low” and “medium”.

StockOptLvl	Attrition	Freq	Attrition_Percent
0	No	473	0.00
1	No	540	0.00
2	No	146	0.00
3	No	70	0.00
0	Yes	150	24.08
1	**Yes**	**56**	**9.40**
2	**Yes**	**12**	**7.59**
3	Yes	15	17.65

It is very interesting that employees prefer Stock Opt Level 1 and 2 verses Stock Opt Level 0 and 3. Average attrition rate is smaller by 67.5% for those who has Stock Opt Level 1 and 2.

Prescriptive Analytics

The implications of this analysis for the company are crucial to the company's growth. Average staff replacement cost is about $15,000, comprising primarily recruitment and training cost of new hires. A retention plan of $7,000 per employee, addressing stock options especially for low level staff, improvement of job and relationship satisfaction, management of job-related travels and other perks will save the company an average of $8,000 per person in attrition cost in the long run. To reduce attrition resulting from reasons highlighted above, the must revamp its current pay structure and introduce benefits that motivate employee loyalty and morale . A few recommendations that can be implemented at a low cost are discussed below:

Limit Business Travel

We found that Frequent business travel was a major contributor to churn in the dataset. If a company wants to reduce the the chances of someone wanting to leave a specific job they should make sure that worker has a reduced responsibility to travel as it could make all the difference in deciding whether to change companies.

Understand How workers feel about Overtime

We found that Not having overtime was very significant in determining employee attrition. This might indicate that management should limit the amount of times they ask employees to pick up extra hours, or only consider asking those that need the extra hours for personal reasons. This can go a long way into contributing to employee attrition.

Consider Remote Work

Based on studying the significance of the distance traveled from home to get to work i.e. commuting we found that this is an extremely significant factor. Based on a 95% confidence Interval the optimal distance is somewhere between 2 miles to 6 miles. If an employee has to travel more than that one should consider offering a remote option in order to limit attrition due to the Distance from Home.

Consider Hiring people that worked for less number of Companies

After studying the data we found that the ideal number of companies worked for those that did experience attrition was 1 to 3 companies on a 95% confidence interval. This indicates that people that work for multiple companies are more likely to contribute to the Attrition rate. This is an extreme solution, but one should consider the amount of companies a potential employee worked before hiring them, as it might be a key predictor as to whether or not that individual will stay or leave.

Satisfaction Indicators

We found that Job Involvement Level 4 ("very high") and 3 ("high"), Environment Satisfaction, Job Satisfaction Level 4 ("very high"), Stock Option Levels 1 and 2 are all contributed to employee churn in a major way. This indicates that management should consider doing things to make the work environment more comfortable so that the employees feel great about going to work every day. Employees also need to feel heavily involved in their job in as well as be satisfied doing it. It is important to conduct reviews and surveys to find out how employees feel in these areas so that an overall general pulse can be gathered on how the company is performing here.

Work relationships must be maintained

Employees consider work relationships to be an extremely important factor and it also contributes to employee "long stay" position. It makes sense to have relationship building events that improves ties within employee to employee relationships. People generally have to like who they work with, and management can improve this by tracking how their employees are relating to each other on the job. Management should consider improving situations that cause bad relationships between workers.

Conclusions

After extensive statistical analysis we have been able to conclude that there are indeed several factors that lead to employee attrition. Based on what these factors are, there are things the business can do to make it much harder for someone to decide if they want to leave the company. We have compiled a list of recommendations that Human Resources can implement to help reduce Attrition. It should be stated that the above stated factors will not guarantee employee attrition will stop, but it will increase the chances of slowing the pace of attrition.

References

Forbes “True Or False Employees Today only stay one or Two Years : David Sturt and Todd Nordstrom”