CaseStudy2

Alexy Morris

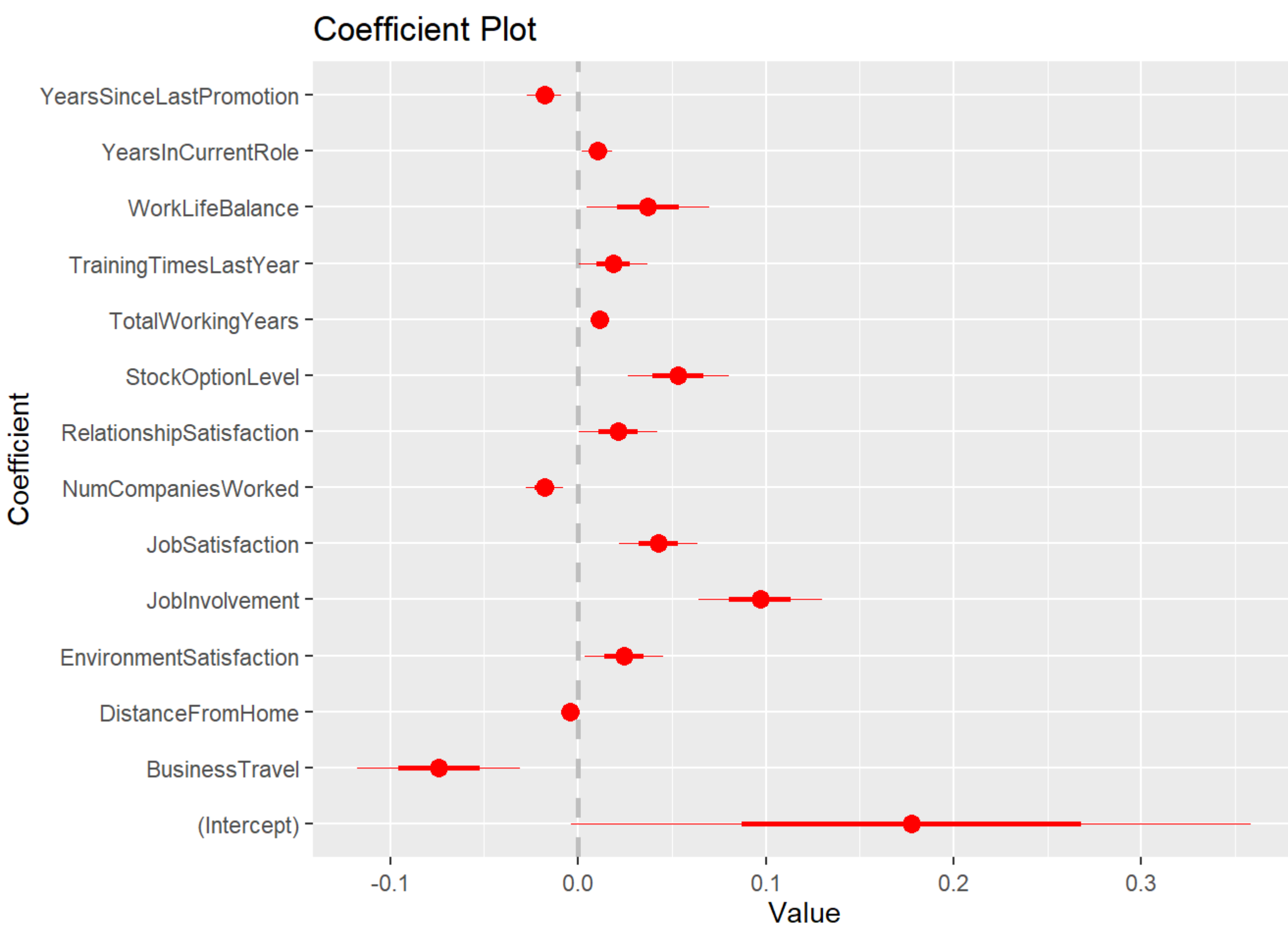
12/4/2021

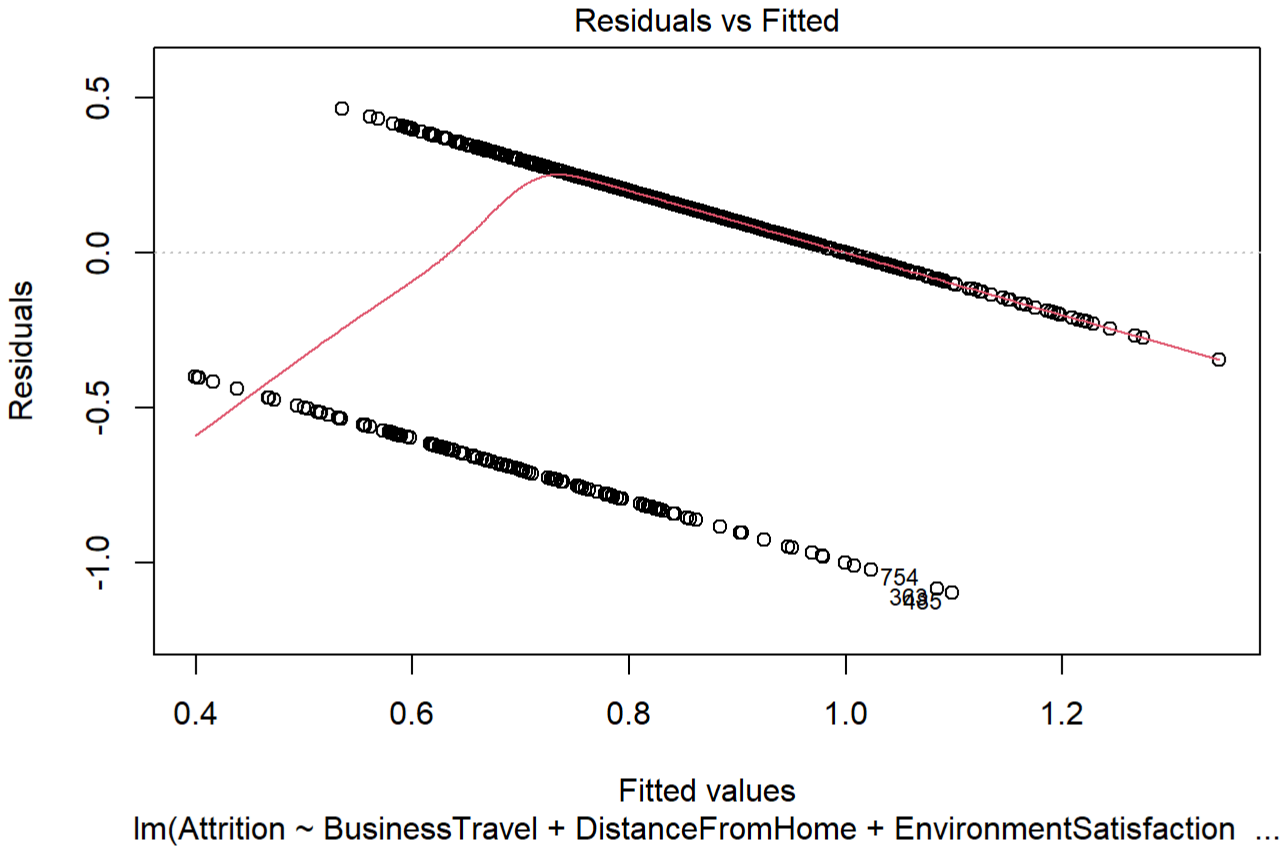
Introduction

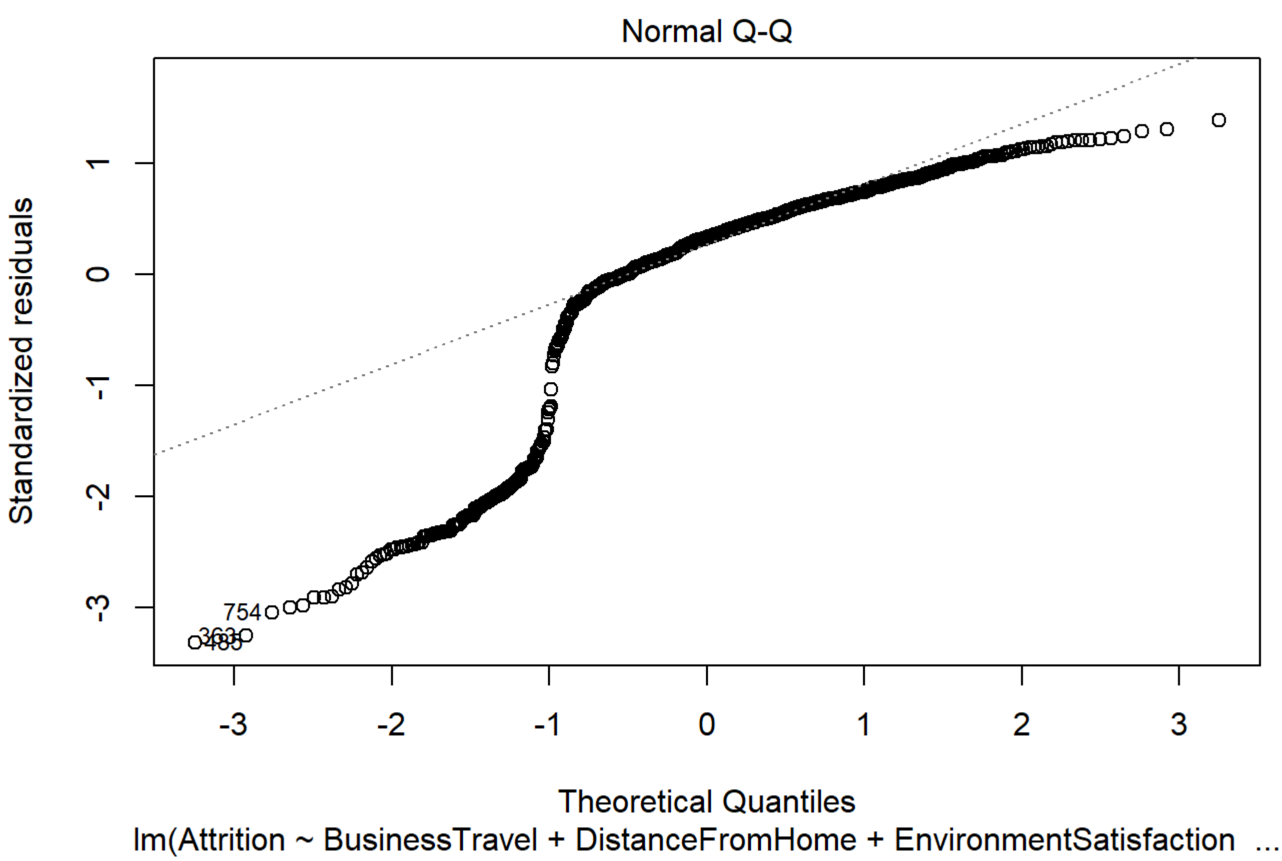
The collection of analyzations and plots below is the work of our firm, DDSAnalytics, in identifying trends in the data received from Frito Lays in regards to both attrition, salary and other notable variables.

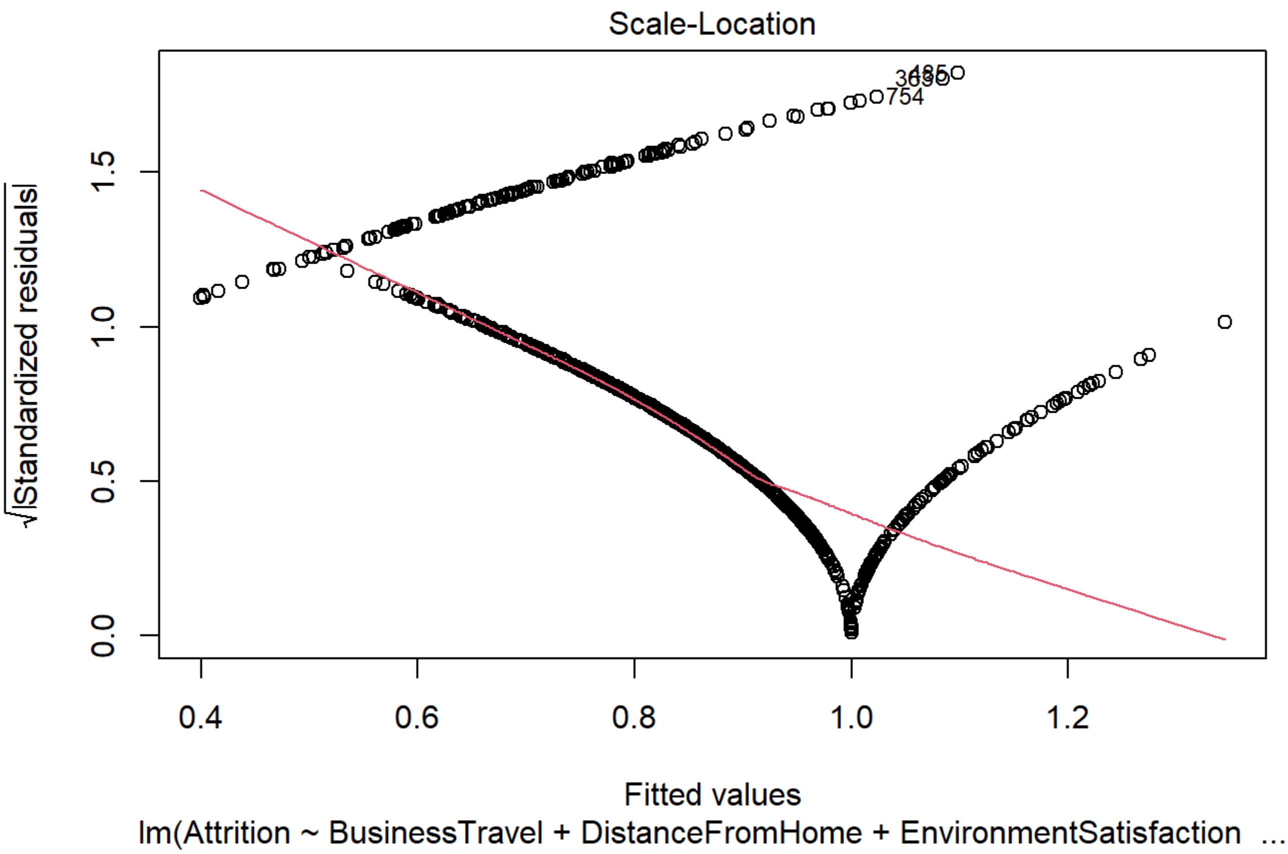
Our teams first steps in regards to the data set was to first look through the data and tidy it into a usable state. Our next step was to subset the data into a new set that was appropriate for a attrition based model. Here appropriate was used to denote a dataset devoid of identifiers and any variables associated with a discriminated class. We immediately tossed said variables because if they were found to hold a significant affect on attrition it would open the company up to potential discrimination based lawsuits.

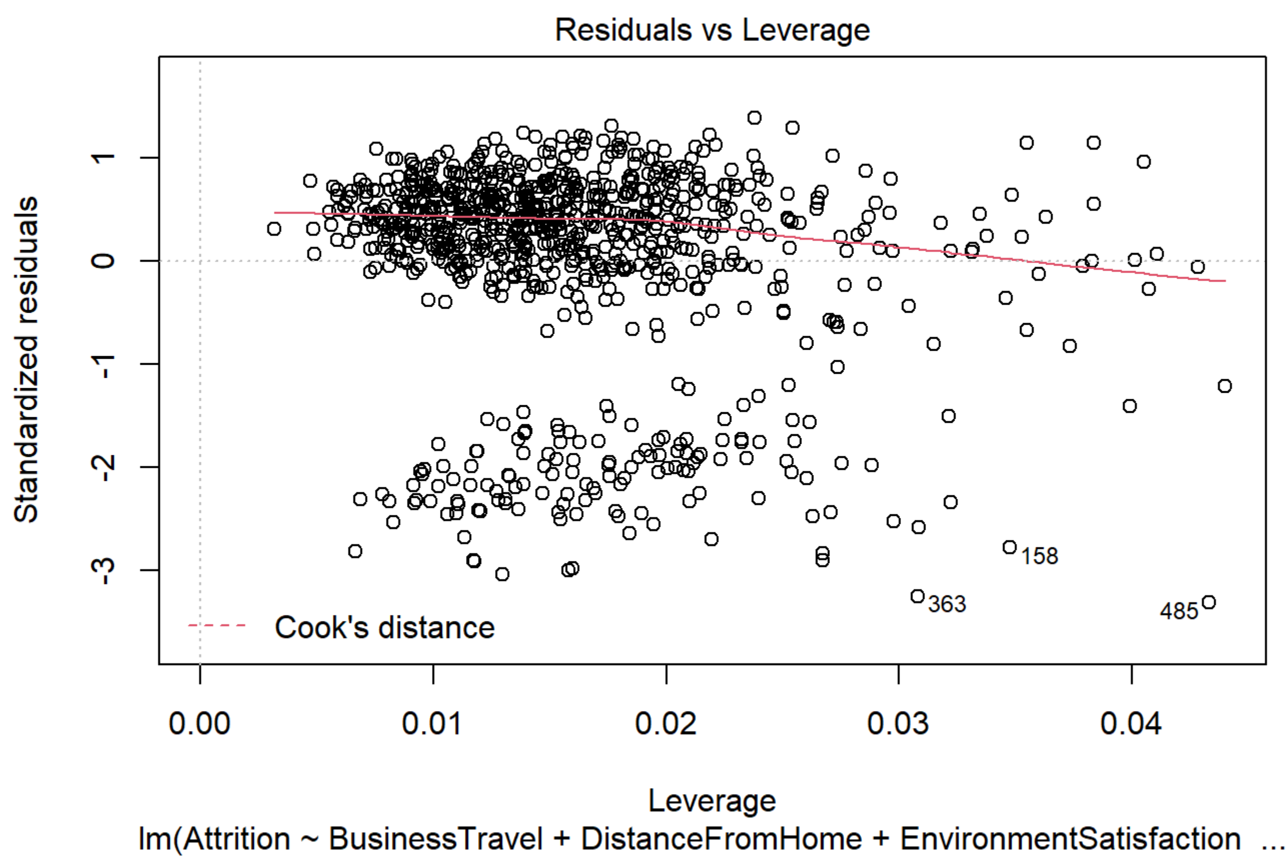
From there we created our attrition model using the backwards selection method and produced a the following coefficients which we plotted for easy comparison. In our model we have found that job involvement, stock option level, and job satisfaction play the biggest role in determining Attrition.



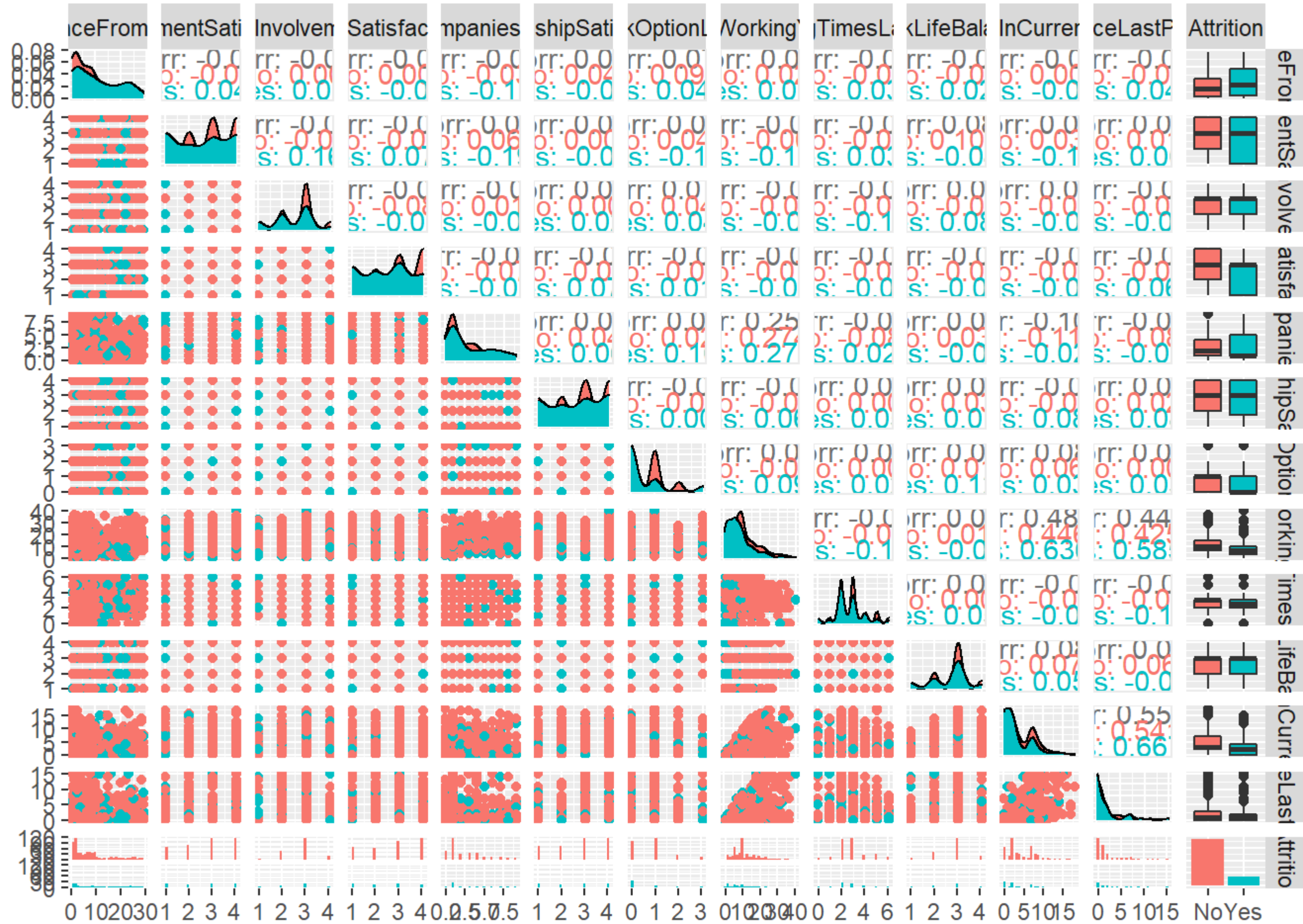

Residuals vs Fitted
Residuals
0.5
0.0
-0.5
-1.0
754
363
485
0.4
0.6
0.8
1.0
1.2
Fitted values
lm(Attrition ~ BusinessTravel + DistanceFromHome + EnvironmentSatisfaction ...


Normal Q-Q
Standardized residuals
1
0
-1
-2
-3
754
363
485
-3
-2
-1
0
1
2
3
Theoretical Quantiles
lm(Attrition ~ BusinessTravel + DistanceFromHome + EnvironmentSatisfaction ...


Scale-Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
0.4
0.6
0.8
1.0
1.2
754
Fitted values
lm(Attrition ~ BusinessTravel + DistanceFromHome + EnvironmentSatisfaction ...


Residuals vs Leverage
Standardized residuals
-3
-2
-1
0
1
0.00
0.01
0.02
0.03
0.04
158
363
485
Cook's distance
Leverage
lm(Attrition ~ BusinessTravel + DistanceFromHome + EnvironmentSatisfaction ...

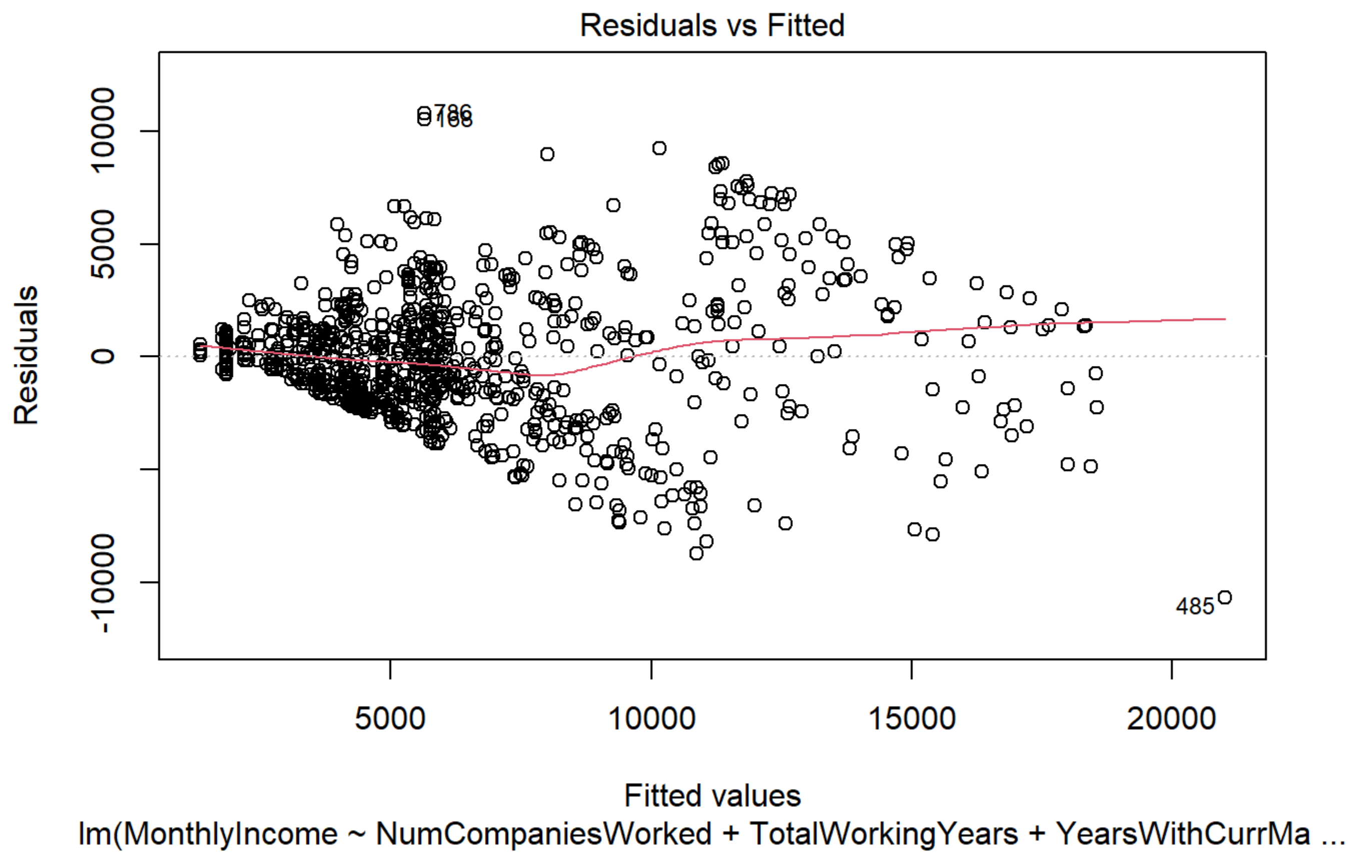
To check our model we did examine the correlations to get a better understanding our our variables in relation to each other.

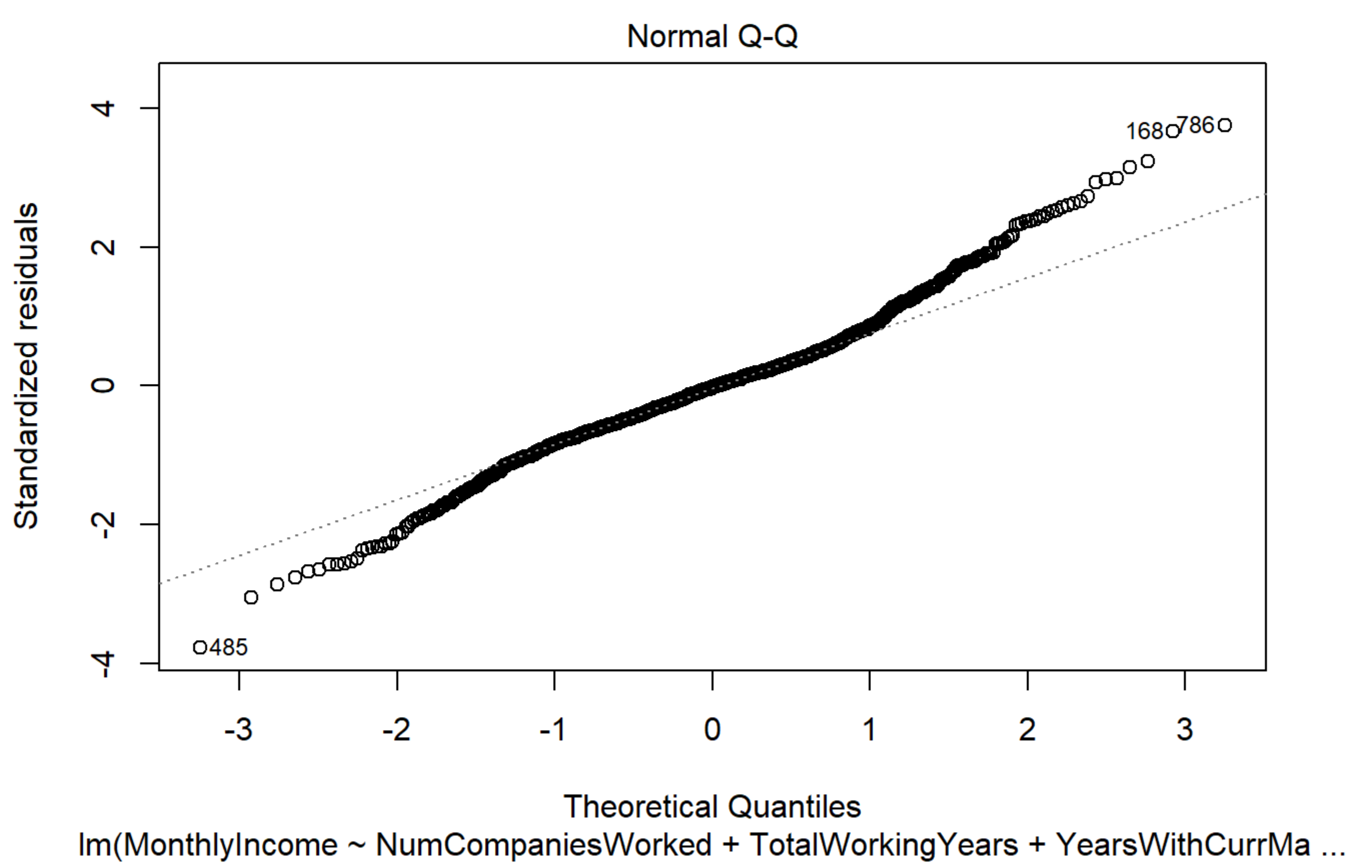


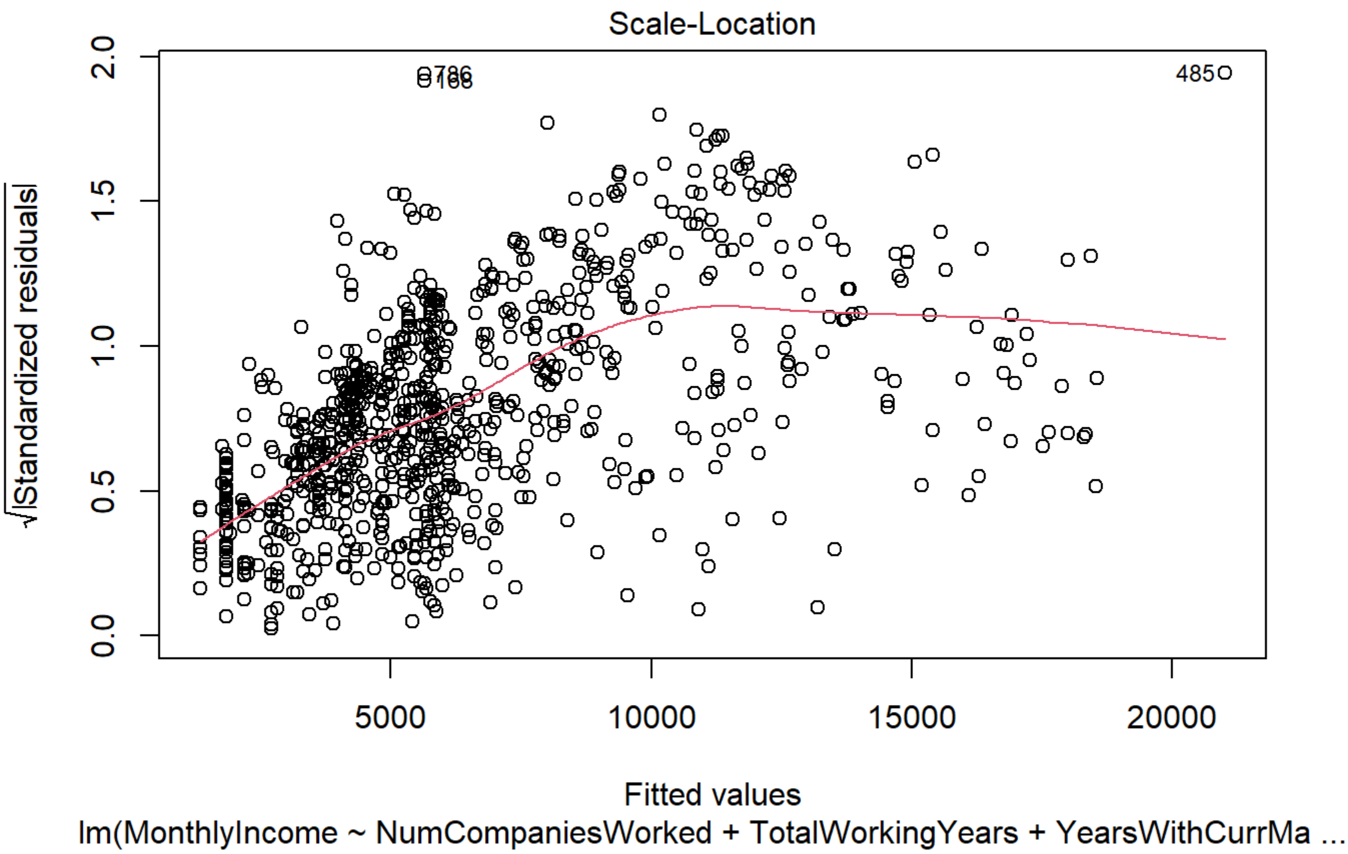
In addition, our team has also created a knn model to predict attrition based upon our specified variables and we have included a new data set in our package to show our predictions.

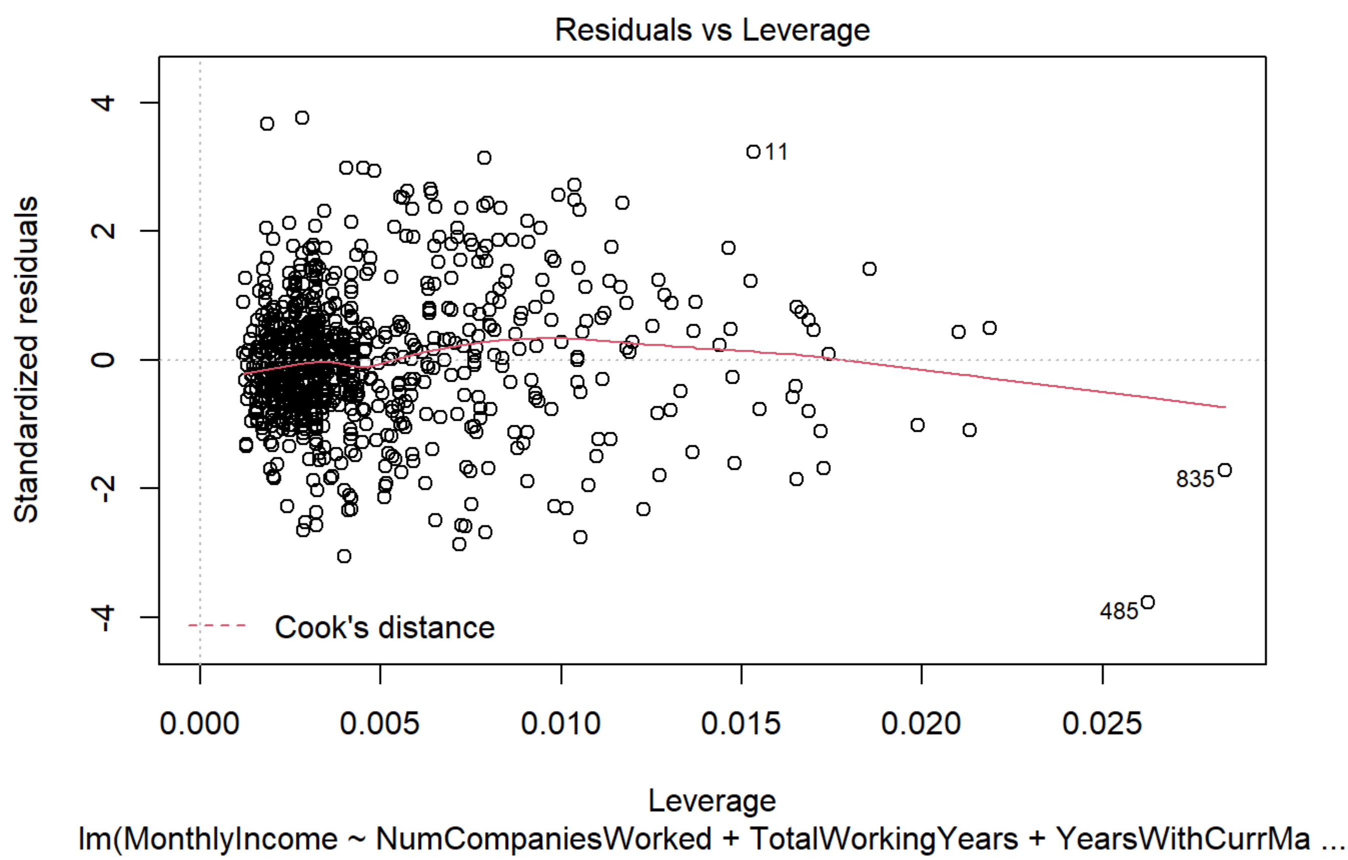
```
## # A tibble: 6 x 2
##      ID AttritionPrediction
##   <dbl> <fct>
## 1  1171 No
## 2  1172 No
## 3  1173 Yes
## 4  1174 No
## 5  1175 No
## 6  1176 No
```

Now switching gears to the employees monthly income, salary, we have again used the backward selection method to create a linear model to a select few of our original variables.


Residuals vs Fitted
Residuals
Fitted values
lm(MonthlyIncome ~ NumCompaniesWorked + TotalWorkingYears + YearsWithCurrMa ...
786
168
485
10000
5000
0
-10000
5000
10000
15000
20000


Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(MonthlyIncome ~ NumCompaniesWorked + TotalWorkingYears + YearsWithCurrMa ...
168
786
485
4
2
0
-2
-4
-3
-2
-1
0
1
2
3


Scale-Location
√|Standardized residuals|
Fitted values
lm(MonthlyIncome ~ NumCompaniesWorked + TotalWorkingYears + YearsWithCurrMa ...
485

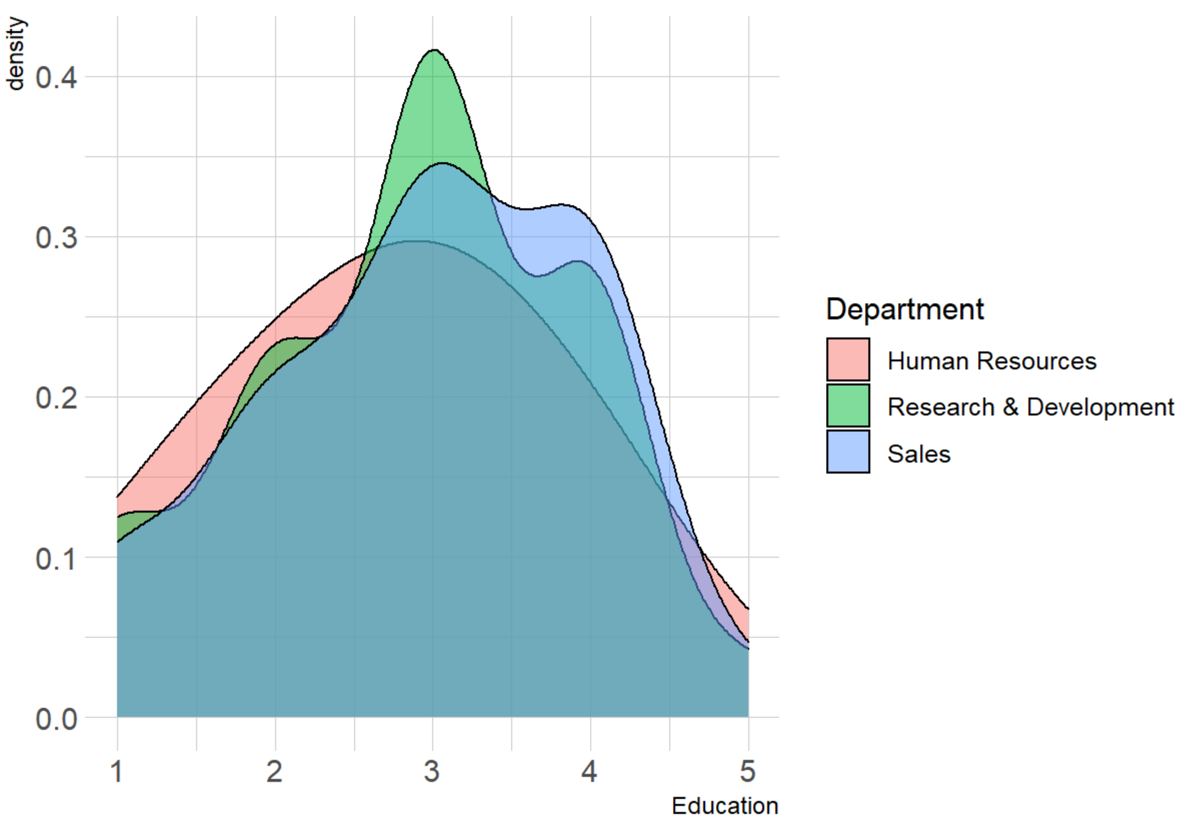

Residuals vs Leverage
Standardized residuals
Leverage
lm(MonthlyIncome ~ NumCompaniesWorked + TotalWorkingYears + YearsWithCurrMa ...
Cook's distance
11
835
485

The predictions made from said model are shown below and our firm believes that this model will be applicable to other data sets and as such can be used as a good jumping off point for future employees salary.

```
## # A tibble: 6 x 2
##      ID ModeledSalary
##   <dbl>         <dbl>
## 1   871         4336.
## 2   872         3476.
## 3   873        12232.
## 4   874         3707.
## 5   875         4336.
## 6   876        10668.
```

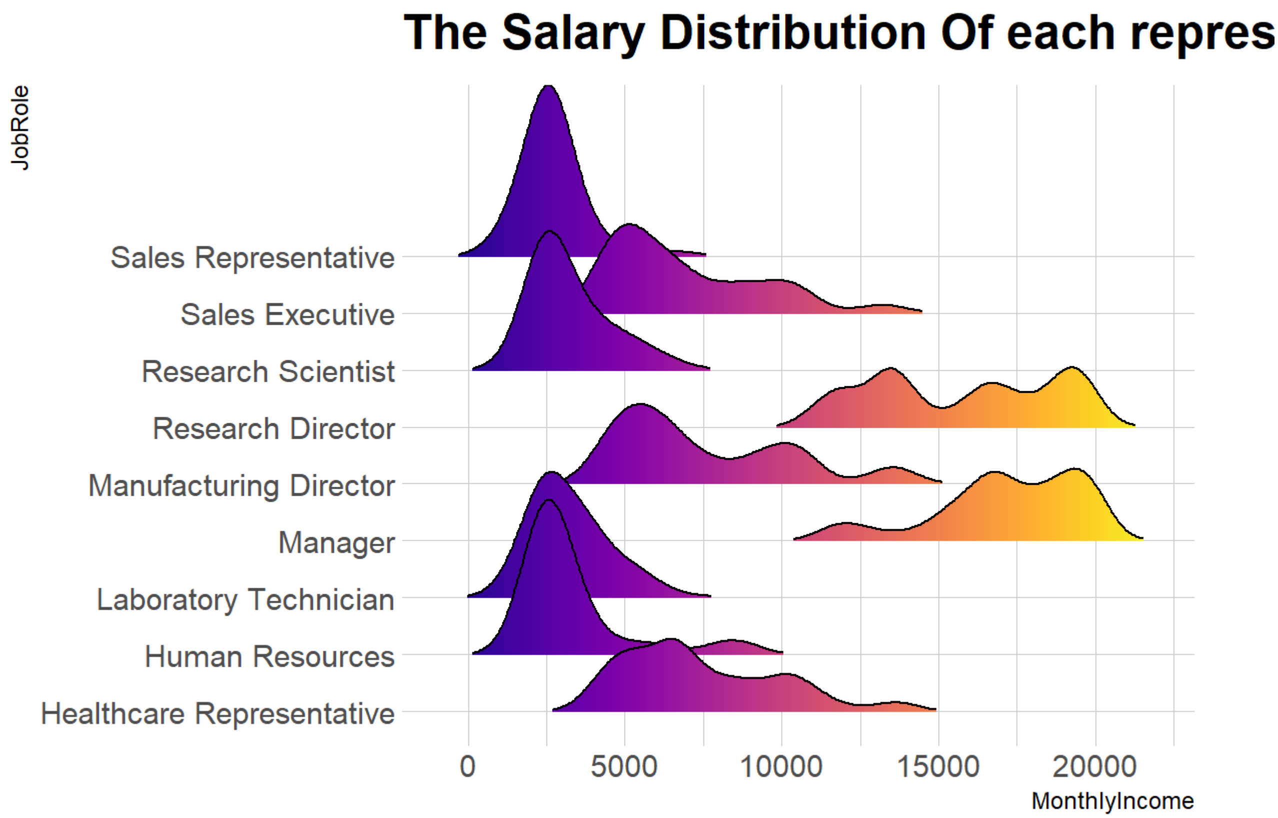
From here our firm has decided to explore additional trends in the data set, the additions includes have been marked useful on our part and if any follow up is desired we would be happy to dig.

In this plot relating the education distributions by department, we saw a mostly expected plot however at the end we noticed that the human resources department easily tops the other departments. It may be worth inquiring over why such a high level is needed in said department, or vice versa, why the other department are falling behind in terms of education.

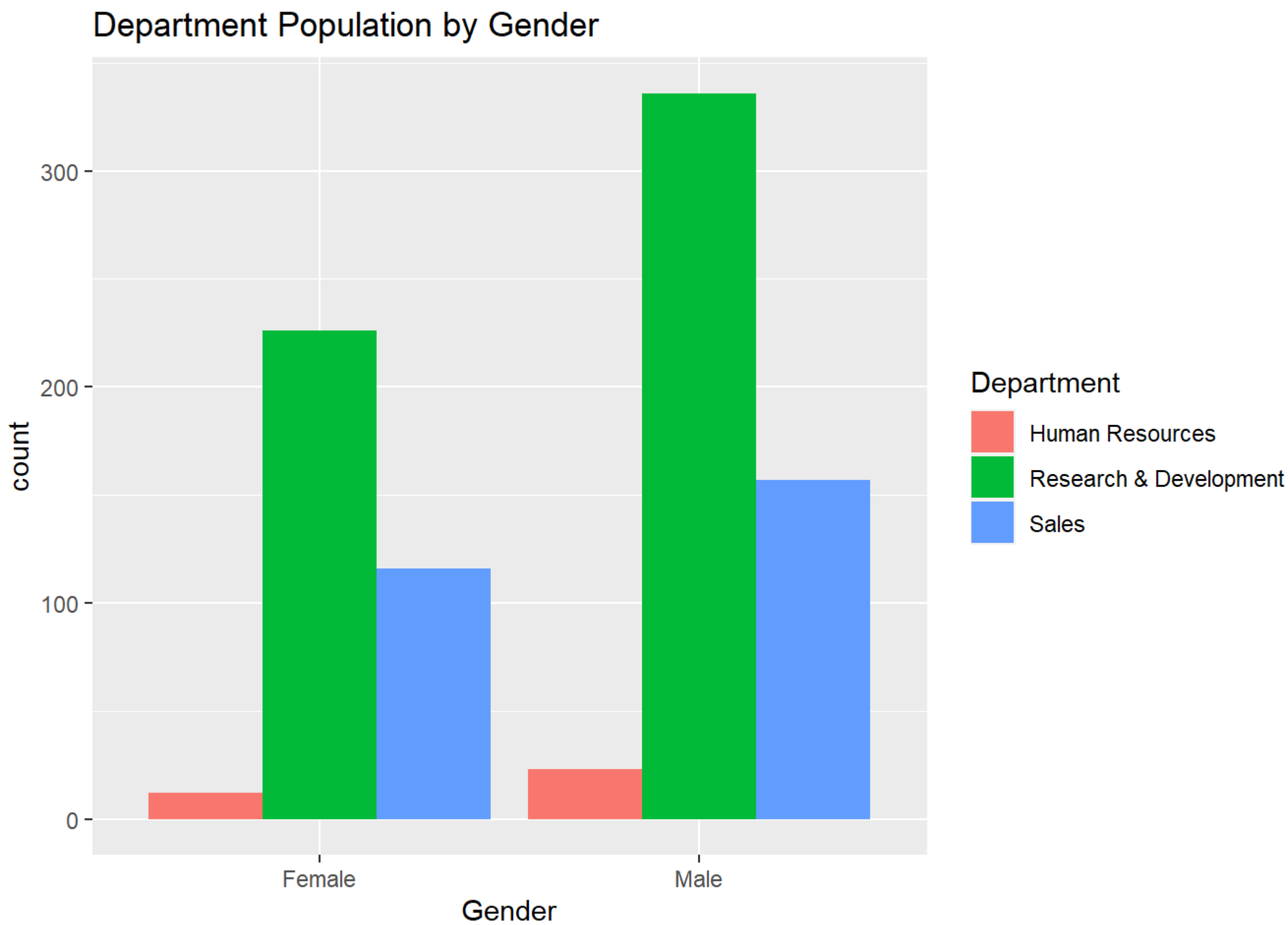


The next plot is in regards to monthly income by job role and here we see a very large array of distributions, almost none of which are normal. This plot is telling us that while the above model is well suited to the overall population, a specific role would most likely value vastly different attributes that would play a big role in determining a role specific salary.

```
## Picking joint bandwidth of 671
```



We also did a quick check on the gender make up of the different department and we noticed the gape between male and female employees grew wider with each department. This may be cause for concern as it might escalate to a potential lawsuit in the future.



Our final note is regarding the difference in correlations between job level and gender in regards to job involvement. Although the below is not a massive different between men and women, the fact that a noticeable difference exists in our sample does raise some concern for a potential bigger difference in the company as a whole.

In Summary, we have created a number of models to predict the variabes of attrition and salary; we have also identified job involvement, stock option level, and job satisfaction as our top causes of attrition. Finally we have identified a number of potentioal red flags in regards to lawsuits.