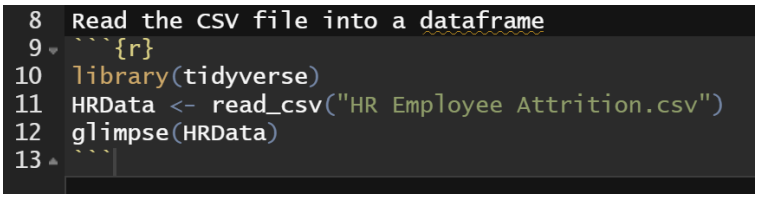
Employee Analysis | Attrition Report

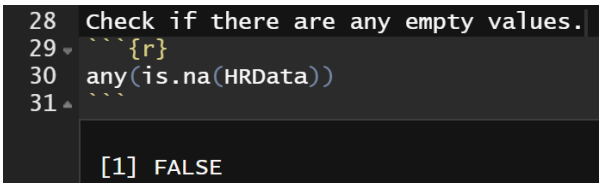
Nathaniel Bowen, Elizabeth Curinga, Sean Deery, Mackenzie Houser, Binosh Padman
IST 687 Team 4

Introduction

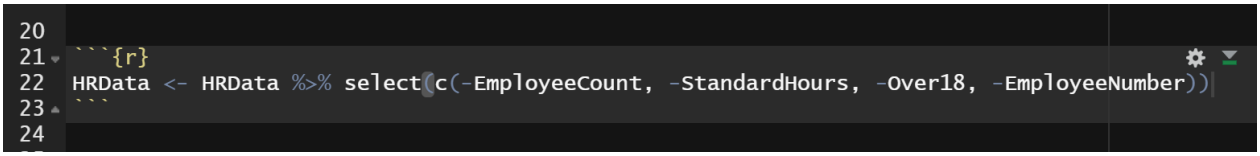
The data set that we chose is from the Employee Analysis Attrition Report from Kaggle (https://www.kaggle.com/datasets/whenamancodes/hr-employee-attrition?select=HR+Employee+Attrition.csv). This is a fictional data set created by IBM data scientists.

The dataset contains 1,470 rows and 35 columns with no missing data. Attrition ("Yes" or "No") will be our dependent variable. "Employee Attrition is defined as the natural process by which employees leave the workforce – for example, through resignation for personal reasons or retirement – and are not immediately replaced" (spiceworks.com).





We can get rid of Employee Number because there is no useful meaning behind it. We can also remove Employee Count , Over 18 and Standard Hours because they are the same for each employee.



The rest of the variables can be used as independent variables, and can be roughly categorized as Personal Details, Work Life and Income variables:

Personal Details

- Age
- Education
- Education Field
- Gender
- Marital Status
- Num Companies Worked
- Relationship Satisfaction

Work Life

- Business Travel
- Department
- Distance From Home
- Environmental Satisfaction
- Job Involvement
- Job Level
- Job Role
- Job Satisfaction
- Over Time
- Performance Rating
- Total Working Years
- Training Times Last Year
- Work Life Balance
- Years At Company
- Years In CurrentRole
- Years Since Last Promotion
- Years With Current Manager

Income

- Daily Rate
- Hourly Rate
- Monthly Income
- Monthly Rate
- Percent Salary Hike
- Stock Option Level

Attrition can be a significant factor for any business to consider, so we came up with a list of six questions to answer with our data analysis.

1. What is the impact of attrition rate on company reputation?
2. What are the most significant factors in employee attrition?
3. How does pay and time worked at the company affect attrition?
4. What are the differences in attrition between departments?
5. How does distance from home affect attrition? Would remote work efforts help with attrition?
6. Is overtime and work life balance correlated and how do they affect attrition?

Data Exploration

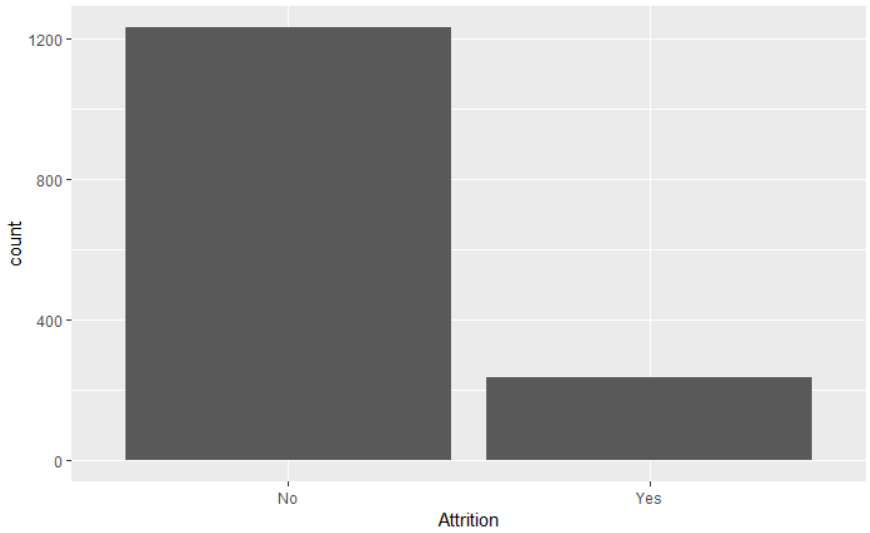
The Dependent Variable - Attrition

Our dependent variable, Attrition, is a column of “Yes” and “No” values, indicating whether employees have left the company. We can see that there are 237 “Yes” values, 16% of the total, and 1233 “No” values, 84% of the total. Since the values are not split 50/50, this indicates that our models will likely be more accurate in predicting employees who do not leave and less accurate in predicting those who leave.

Furthermore, we can take this Attrition rate and compare it to the market. A recent article by Sr. Product Marketing Manager for the Oracle Netsuite Global Business Unit states that “On average, every year, a company will experience 18% turnover in its workforce.” (netsuite.com)

At 16%, our company is below that average which is a good sign. Other factors that should be considered when it comes to Attrition is whether it is voluntary or involuntary, how it affects key roles within the company, and how it affects any specific demographics.

```
Check out Attrition.
```{r}
summary(HRData$Attrition)
table(HRData$Attrition)
nrow(HRData[HRData$Attrition=="No",])/nrow(HRData)
nrow(HRData[HRData$Attrition=="Yes",])/nrow(HRData)
```
```

```
   Length     Class      Mode
     1470 character character

  No  Yes
1233  237
[1] 0.8387755
[1] 0.1612245
```



Functions to Explore Independent Variables

We have 30 variables that we can look at, so to make things easier, we created some functions to show some initial descriptive statistics and visualizations.

```
Function to explore numerical columns
```{r}
exploreNumVariable <- function(num_variable) {
  # Get a summary of the data
  variable.summary <- summary(HRData[,num_variable])
  # Create the Group By
  variable.groupby <- HRData %>%
    group_by(Attrition) %>%
    summarize(var_mean=mean(.data[[num_variable]]))
  # Create the Histogram
  variable.hist <- ggplot(HRData) +
    aes(x=.data[[num_variable]], fill=Attrition) +
    geom_histogram()
  # Create the Box Plot
  variable.box <- ggplot(HRData) +
    aes(x=Attrition, y=.data[[num_variable]]) +
    geom_boxplot()
  return(list(variable.summary, variable.groupby, variable.hist, variable.box))
}
```
```

```
Function to explore categorical columns
```{r}
exploreCatVariable <- function(cat_variable) {
  # Create a table showing frequency distribution
  variable.table <- table(HRData[, cat_variable])
  # Create a bar chart showing frequency distribution by Attrition
  variable.bar <- ggplot(HRData) +
    aes(x=.data[[cat_variable]], fill=Attrition) +
    geom_bar(position="dodge") +
    theme(axis.text.x=element_text(angle=90, hjust=1))
  return(list(variable.table, variable.bar))
}

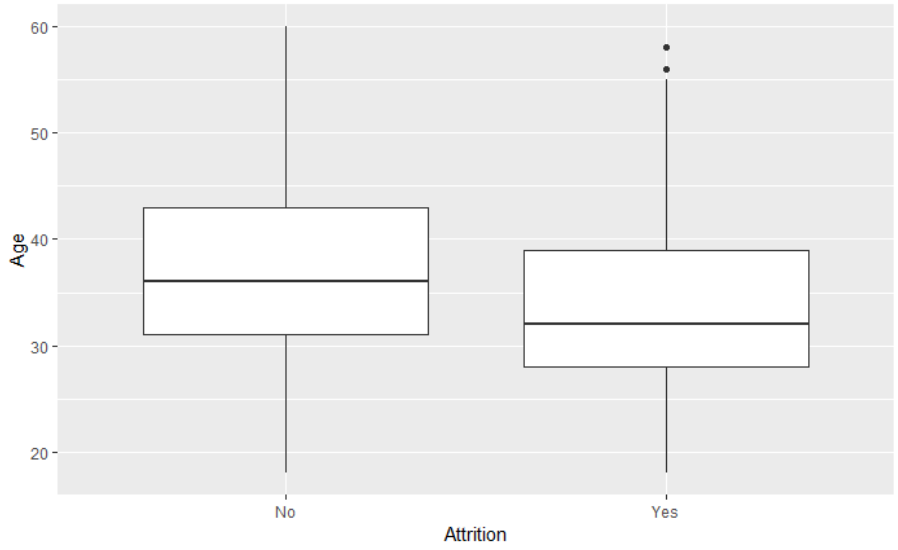
```
```

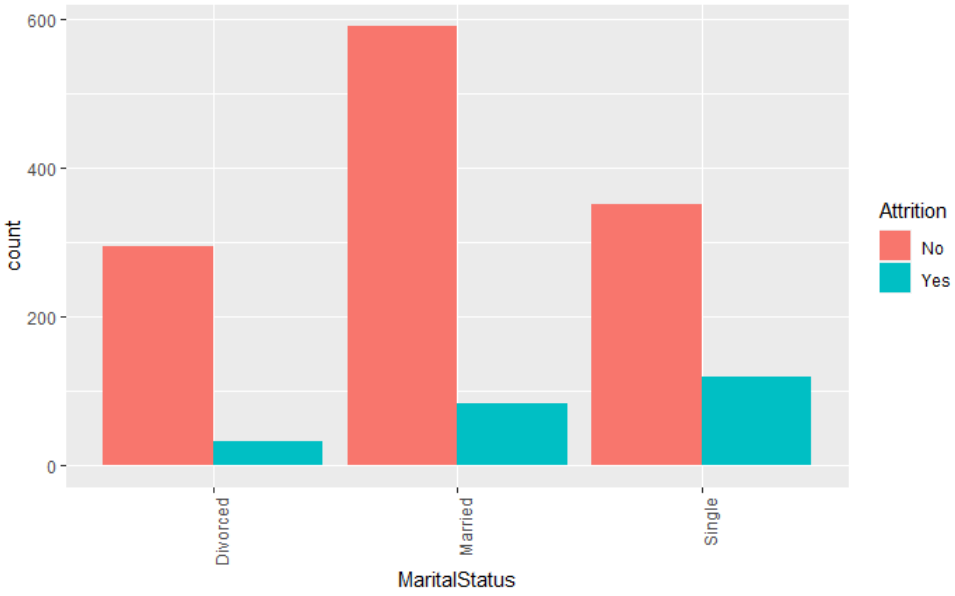
Personal Details

Overall, it looks the Age and Marital Status of employees will be significant factors in determining Attrition. It seems like younger single employees have higher rates of attrition.

The average Age of employees who leave is 33, whereas the average Age of employees who have stayed is 37. The percentage of single employees who leave is 26%, whereas the percentage for married employees is 12% and the percentage for divorced employees is 10%.

This could be because younger people are still looking for the career they want, and single people might have more flexibility to make changes.

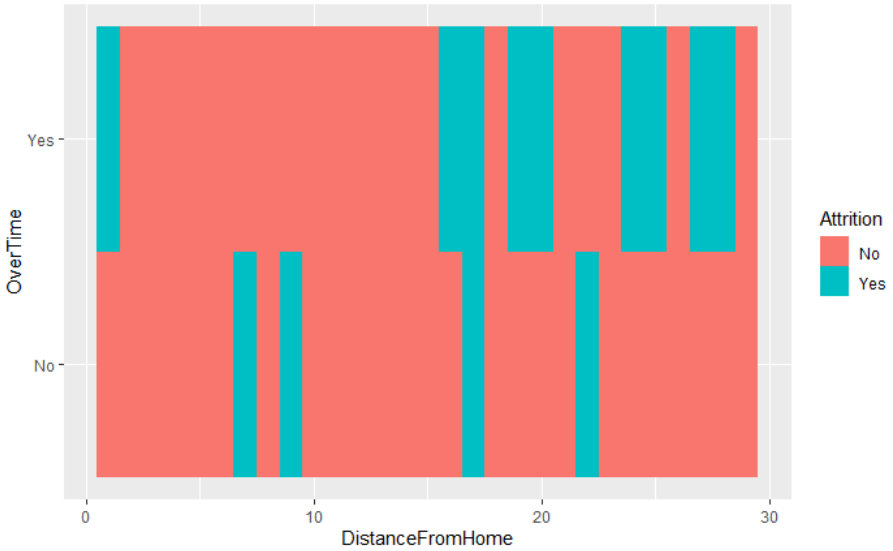




Work Life

Overall, it looks like many of the Work Life variables can be significant in our analysis. Sales and Technicians seem to have a high attrition rate. Also, people who work overtime and live farther away, or those who have another reason for low satisfaction, have a higher attrition rate.

The visual below is a heatmap using the geom_tile() function in ggplot2 to represent the higher frequency of attrition for people working overtime and having a longer commute to work. The correlation of Attrition and Distance From Home says that the longer the distance, the more likely someone is to leave. The p-value of the correlation says that it is statistically significant at a 95% confidence level.




```r
cor.test(HRData$Attrition, HRData$DistanceFromHome, method = "pearson")
```

```
        Pearson's product-moment correlation

data:  HRData$Attrition and HRData$DistanceFromHome
t = 2.9947, df = 1468, p-value = 0.002793
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.02690331 0.12853894
sample estimates:
       cor
0.07792358
```

The average attrition rate can be found numerous ways. Below is code that has multiple datasets split up by department (HR, R&D, and Sales) found using a manual equation to calculate the average. From our companies attrition rate- we can also conclude that our employee retention rate by department is: Human Resources 80.95%, Research & Development 86.16%, and Sales 79.37%.

average attrition by department

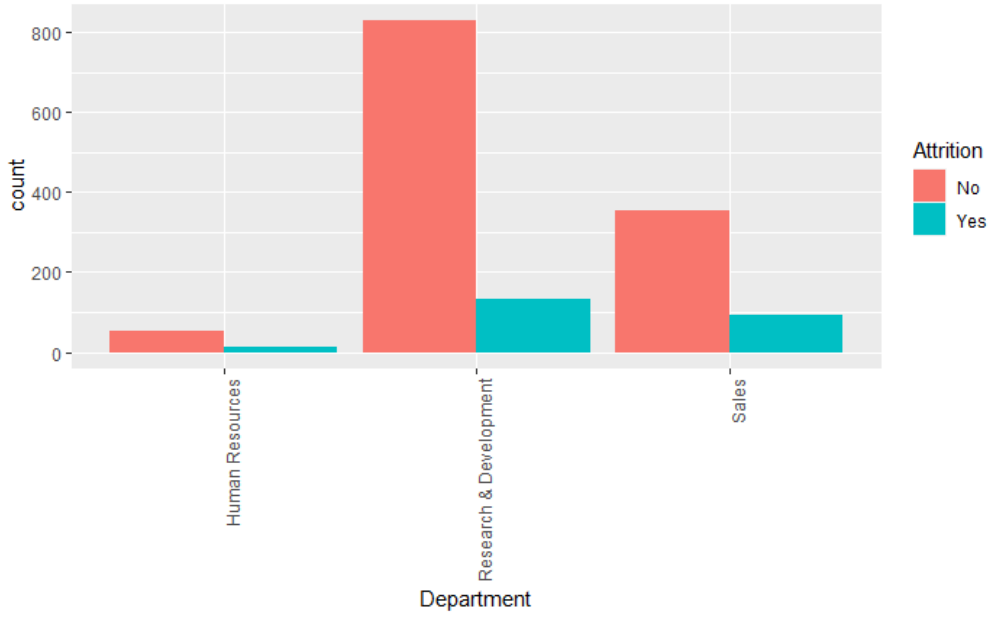
```r
sum(hr_hr$Attrition == 'Yes', na.rm = TRUE)/ nrow(hr_hr)

sum(rd_hr$Attrition == 'Yes', na.rm = TRUE)/ nrow(rd_hr)

sum(sales_hr$Attrition == 'Yes', na.rm = TRUE)/ nrow(sales_hr)

##Sales has the highest attrition Rate
```

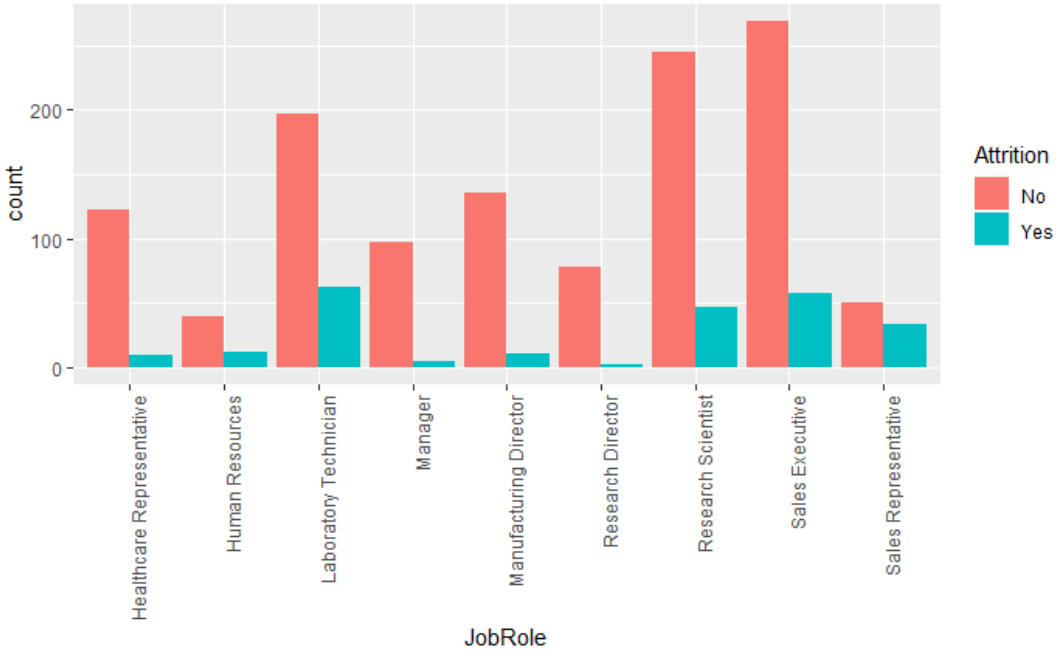
```
[1] 0.1904762
[1] 0.1383975
[1] 0.206278
```

Looking at the histogram of the Attrition frequencies by department- at first glance, Research & Development appears to have the highest attrition rate. However, there are 961 employees in the R&D department and only 63 in Human Resources and 446 in Sales. The percentages show that Sales has a greater percentage of people who leave (20.63%) compared to Human Resources (19.05%) and Research and Development (13.84%). So, the Research & Development department has the highest count of attrition, but Sales still has the highest rate of attrition. To expand on this, the Job Role column shows that Sales Representatives and Laboratory Technicians have a high percentage of employees who leave compared to the other job roles.

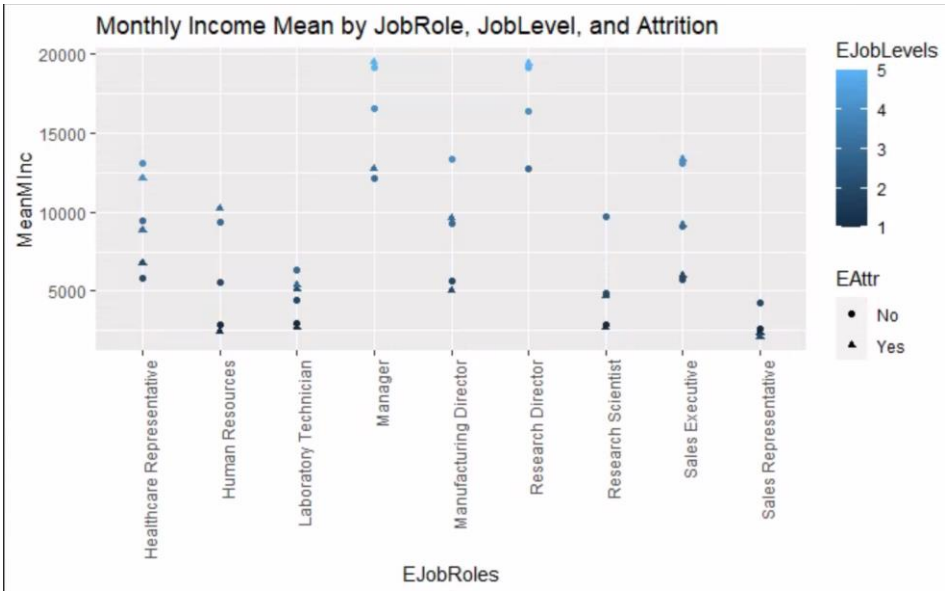
```{r}
nrow(hr_hr)
nrow(rd_hr)
nrow(sales_hr)
```

```
[1] 63
[1] 961
[1] 446
```





Also, when comparing Monthly Income, Job Role, and Job Income, it was the Job Role that seemed to have more of a determining factor on Attrition.



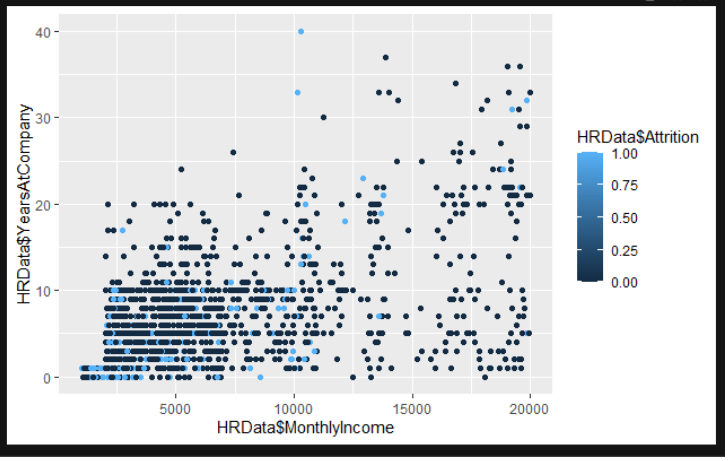
Income

Shiny App link for data exploration on Income:

https://lcavonattic.shinyapps.io/DataExplorationEmployeeAttritionRateIncomeExploration/?_ga=2.167337149.2031634688.1671493398-349767424.1671493398

Overall, it looks like the Daily Rate and the Monthly Income will be significant in determining attrition. The average Daily Rate for employees who leave is 750 where the average Daily Rate for employees who stay is 813. The average Monthly Income for people who leave is 4787 where the average for people who stay is 6832. This makes sense as people who are getting paid more will be less likely to leave the company.

```
ggplot(HRData, aes(x =HRData$MonthlyIncome, y =HRData$YearsAtCompany, color =
HRData$Attrition)) +
  geom_point()
```



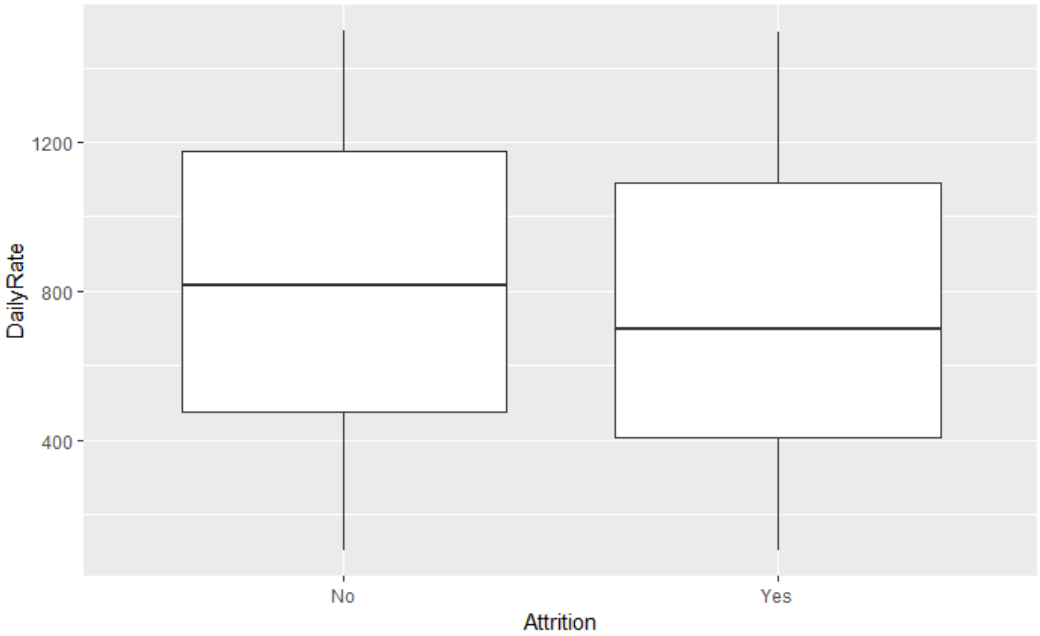
Does salary hike have an impact on attrition?

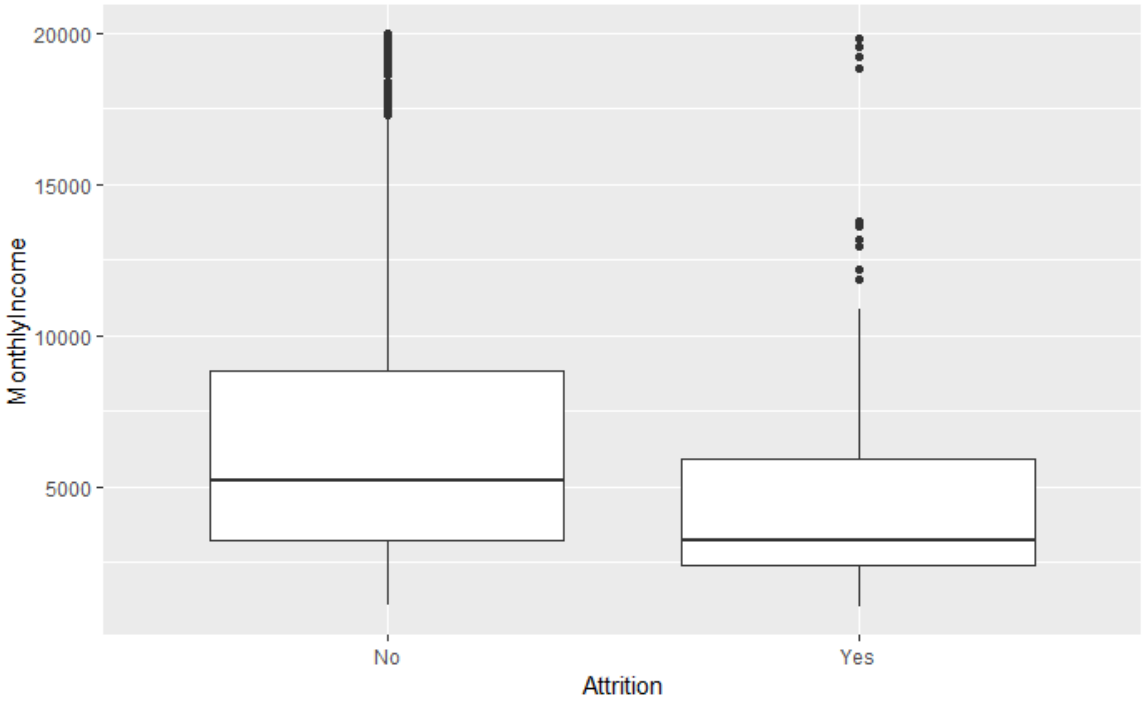
```r
cor.test(HRData$Attrition, HRData$PercentSalaryHike, method = "pearson")
quit <- HRData[HRData$Attrition == 1,]
stayed <- HRData[HRData$Attrition == 0,]
mean(quit$PercentSalaryHike)
mean(stayed$PercentSalaryHike)
mean(HRData$PercentSalaryHike)
##correlation means attrition is lower with higher salary spike
```

```
        Pearson's product-moment correlation

data:  HRData$Attrition and HRData$PercentSalaryHike
t = -0.51646, df = 1468, p-value = 0.6056
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.06456117  0.03767522
sample estimates:
       cor
-0.0134782

[1] 15.09705
[1] 15.23114
[1] 15.20952
```



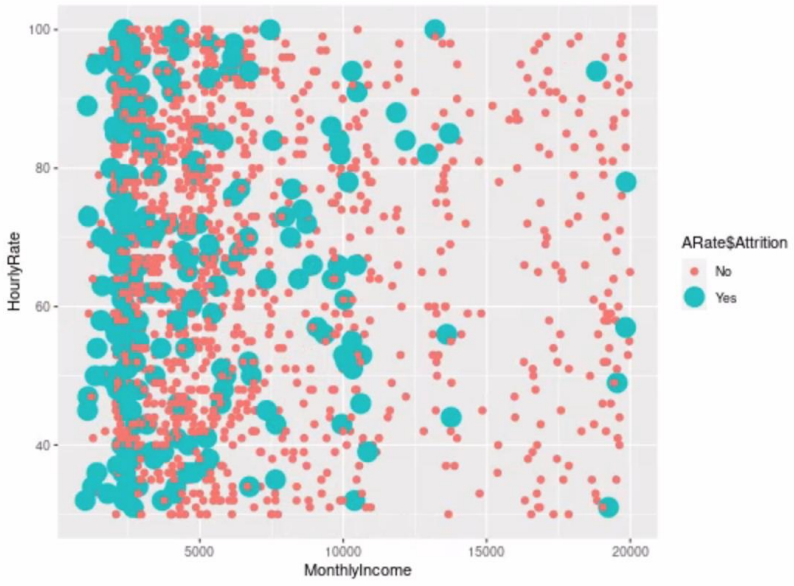


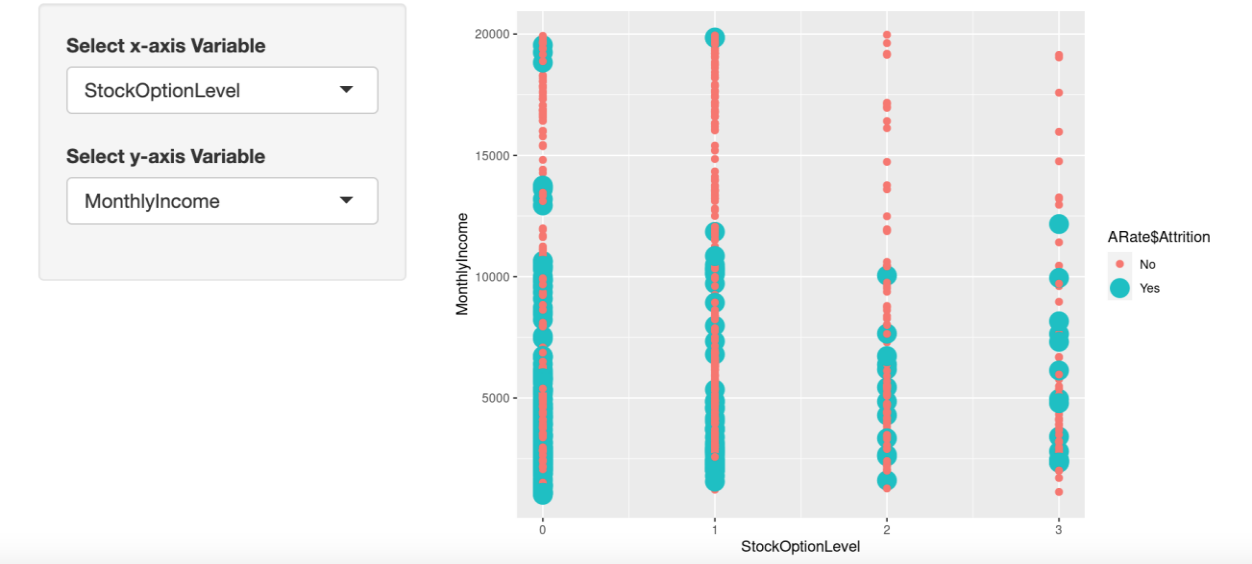
Select x-axis Variable

MonthlyIncome

Select y-axis Variable

HourlyRate





Statistical Modeling

Splitting the Data

We split the data into training and testing data sets so we could review the accuracy of the models. We chose to split 70% for training data and the remaining 30% for testing data.

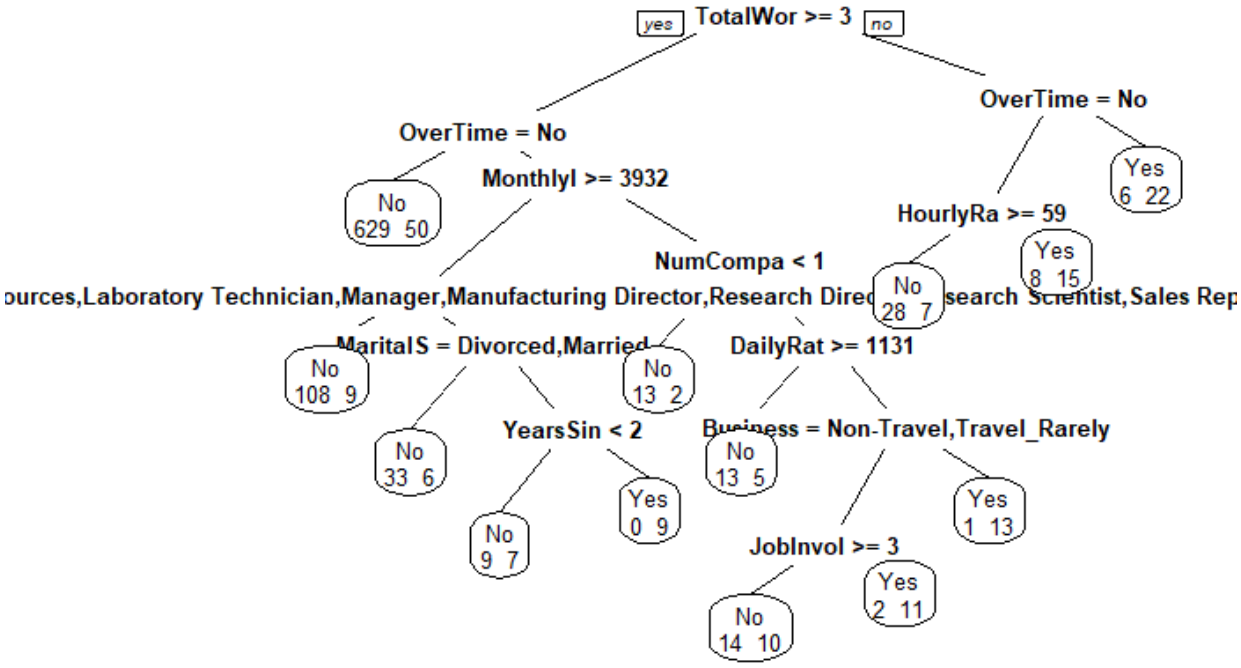
```
Split the data into train and test data
```{r}
trainList <- createDataPartition(y=HRData$Attrition, p=.70, list=FALSE)
training <- HRData[trainList,]
testing <- HRData[-trainList,]
```
```

Decision Tree

The decision tree model was able to predict the test data with 83% accuracy. The most significant variables in order were Monthly Income, Over time, Stock Option Level, Job Level, Job Role, Marital Status, Total Working Years, Years At Company, Business Travel, and Hourly Rate.

```r
library(rpart)
HRTree <- rpart(Attrition ~ ., data=training)
```

```r
library(rpart.plot)
prp(HRTree, faclen=0, cex=0.8, extra=1)
```



```r
Create a Confusion Matrix from the predictions on the test data
```{r}
treePred <- predict(HRTree, newdata=testing, type="class")
actualAttrition <- as.factor(testing$Attrition=="Yes")
confMatrix <- table(treePred, actualAttrition)
confMatrix
```
```

```
           actualAttrition
treePred FALSE TRUE
     No    354   57
     Yes    15   14
```

```r
```{r}
accuracy <- 1 - (sum(confMatrix) - sum(diag(confMatrix))) / sum(confMatrix)
accuracy
```
```

```
[1] 0.8363636
```

```r
```{r}
varImp(HRTree) %>% arrange(desc(Overall))
```
```

Description: df [30 × 1]

	Overall <dbl>
MonthlyIncome	44.132218
OverTime	42.067268
StockOptionLevel	38.192535
JobLevel	34.567684
JobRole	26.047592
MaritalStatus	25.834820
TotalWorkingYears	23.050358
YearsAtCompany	14.733740
BusinessTravel	10.916244
HourlyRate	9.085804

1-10 of 30 rows

Logit

The logit analysis showed that certain roles were significant towards job attrition. Lab techs and Sales Representatives appear more likely to leave than other roles. The graphs we have further uphold this observation. Overtime also appears to be significant and it shows in the other part of our analysis that it increases the odds of employees leaving. Single employees are significant as well. It makes sense that the variables that were higher on our variable importance are also some of the variables that were significant in the logit analysis.

```
$:
$formula = Attrition ~ MonthlyIncome + OverTime + StockOptionLevel +
$JobLevel + JobRole + MaritalStatus + TotalWorkingYears +
$YearsAtCompany, family = binomial(logit), data = Dataset)

$ance Residuals:
$Min       1Q   Median       3Q      Max
$387  -0.5916  -0.3826  -0.2128   3.1563

$ficients:
$                                Estimate  Std. Error z value    Pr(>|z|)
$ercept)                      -3.10962454  0.59001487  -5.270 0.000000136 ***
$hlyIncome                     0.00007247  0.00007102   1.020    0.307522
$Time[T.Yes]                   1.54351029  0.16178148   9.541     < 2e-16 ***
$kOptionLevel                 -0.21819699  0.13978260  -1.561    0.118530
$evel                         -0.11492222  0.26974783  -0.426    0.670082
$ole[T.Human Resources]        1.55292541  0.52244979   2.972    0.002955 **
$ole[T.Laboratory Technician]  1.49637940  0.44209973   3.385    0.000713 ***
$ole[T.Manager]               -0.37604887  0.72403451  -0.519    0.603496
$ole[T.Manufacturing Director] -0.04858960  0.49190527  -0.099    0.921314
$ole[T.Research Director]      -1.16938226  0.89255106  -1.310    0.190143
$ole[T.Research Scientist]      0.71685492  0.44947933   1.595    0.110744
$ole[T.Sales Executive]         0.98785531  0.39138149   2.524    0.011602 *
$ole[T.Sales Representative]    2.02787350  0.49034283   4.136 0.000035399 ***
$talStatus[T.Married]           0.19379770  0.24141111   0.803    0.422107
$talStatus[T.Single]            0.84487993  0.30595221   2.761    0.005754 **
$lWorkingYears                 -0.03371357  0.02087729  -1.615    0.106344
$sAtCompany                    -0.02235463  0.02089865  -1.070    0.284769

$if. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

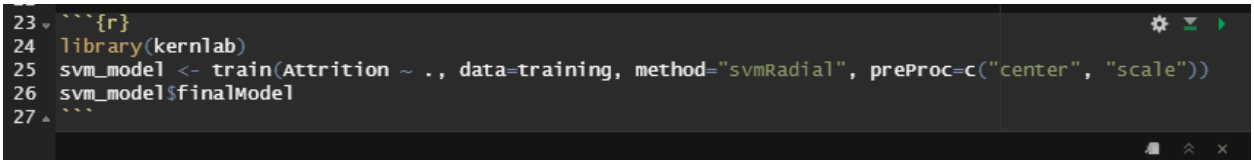
$persion parameter for binomial family taken to be 1)

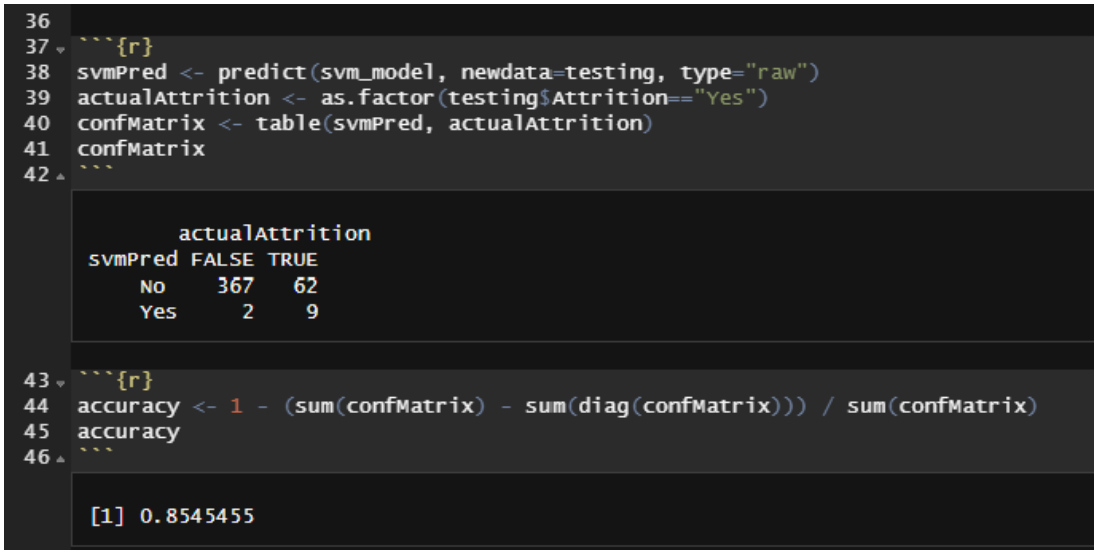
$Null deviance: 1298.6  on 1469  degrees of freedom
$dual deviance: 1069.0  on 1453  degrees of freedom
$ 1103

$er of Fisher Scoring iterations: 6
```

SVM

The SVM model, which was also trained with 70% of the data, was able to predict the test data with 85% accuracy. While the model was more accurate, there is no way to look at which factors were important to the predictions.





```r
svmPred <- predict(svm_model, newdata=testing, type="raw")
actualAttrition <- as.factor(testing$Attrition=="Yes")
confMatrix <- table(svmPred, actualAttrition)
confMatrix
```

```
       actualAttrition
svmPred FALSE TRUE
    No    367   62
    Yes     2    9
```

```r
accuracy <- 1 - (sum(confMatrix) - sum(diag(confMatrix))) / sum(confMatrix)
accuracy
```

```
[1] 0.8545455
```

Business Questions

What is the impact of attrition rate on company reputation?

Attrition can have a large impact on the company's reputation, but the type of Attrition plays a large part in it. Beyond the attrition rate, the impact to the company's reputation relies on the percentage of Attrition that is voluntary vs involuntary and if the Attrition is demographic specific.

At 16%, our company is below that average which is a good sign. Another good sign is that Marital Status was the only demographic variable that the decision tree found important in creating the model. This indicates people of certain demographics are not getting preferential treatment.

Other data that would be useful to have is whether the Attrition was voluntary or involuntary. We could also use exit interview responses, where we could use Text Mining and Sentiment Analysis to analyze employee's perceptions of the company.

What are the most significant factors in employee attrition?

The top ten most significant variables for the decision tree in order were Monthly Income, Overtime, Stock Option Level, Job Level, Job Role, Marital Status, Tota lWorking Years, Years At Company, Business Travel, and Hourly Rate.

If we look at the variables that we categorized as Personal Details, we can see MaritalStatus is the only one in the top ten. From our exploratory analysis we can see that employees who are single have a much higher attrition rate than married or divorced employees, which makes sense as those who are single probably have more flexibility to move jobs.

Looking at the variables that relate to Work Life, we can see Overtime, Job Level, Job Role, Total Working Years, Years At Company, and Business Travel are all in the top ten. These variables make sense as travel and working overtime can lead to burnout. It also makes sense that Job Role and Job Level play a part in Attrition as some jobs, like Lab Technician, may be seen as more stepping stone roles for people looking to advance their career.

Lastly for the variables that relate to Income, we can see Monthly Income, Stock Option Level, and Hourly Rate made the top ten. Monthly Income was the most significant variable in the decision tree, which makes sense as we saw that people who left the company had an average monthly pay well below the average monthly pay of those who stayed. These variables taken together indicate that changes to employee compensation can affect the company's attrition rate.

How does pay and time worked at the company affect attrition?

Overall, it looks like the Daily Rate and the Monthly Income will be significant in determining Attrition. This makes sense as people who are getting paid more will be less likely to leave the company.

Our graph shows that people with less pay and years at the company are more likely to leave. Another conclusion we reached to explain the outliers was that those who had been at the company for years and were well paid may have left due to retirement.

The average Daily Rate for employees who leave is 750 where the average Daily Rate for employees who stay is 813. The average Monthly Income for people who leave is 4,787 where the average for people who stay is 6,832.

The data set did not provide the units for Daily Rate or Monthly Income, so one limitation is not knowing exactly what they are referring to.

What are the differences in attrition between departments?

The Sales Department has the largest percentage of people who leave compared to Human Resources and Research and Development. The percentages show that Sales has an attrition rate of 20.63% compared to Human Resources 19.05% and Research and Development 13.84%. There are 961 employees in the R&D department and only 446 in Sales and 63 in Human Resources. The Research & Development department has the highest count of attrition, but Sales still has the highest rate of attrition. Within the job role, Sales Representatives and Laboratory Technicians have a high percentage of employees who leave compared to the other job roles.

How does distance from home affect attrition? Would remote work efforts help with attrition?

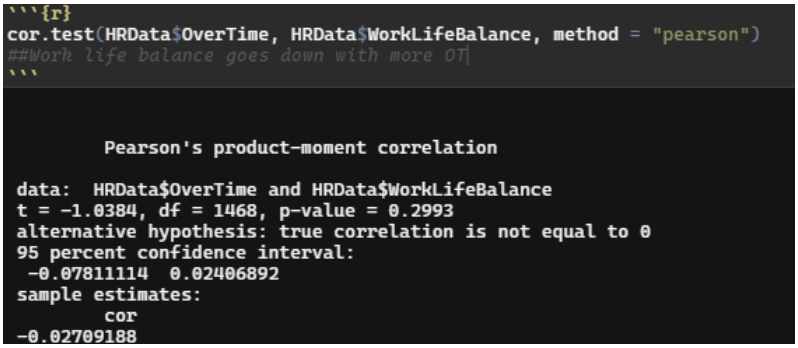
The average Distance From Home for employees who leave is 10.6 where the average for people who stay is 8.9 and the overall average is 9.19. We do not have a unit of measurement but we can compare values across employees. The correlation of Attrition and Distance from Home says that the longer the distance, the more likely someone is to leave.

A remote work option could benefit the company employee attrition rate, at least for employees whose Distance From Home is greater than average. Ethically, it would be important to offer the same options to all employees. However, not all job roles have the ability to be done from home.

It would be helpful to know if the employees already receive a remote work option. That way we could further analyze the Distance From Home variable and its impact on employee attrition or other variables like: Job Satisfaction, Overtime, Environment Satisfaction, etc.

Is overtime and work life balance correlated and how do they affect attrition?

Both Work Life Balance and Overtime are correlated with Attrition, which makes sense as people will look for other job opportunities when they are not enjoying life or working too many hours. What is interesting is that we found Work Life Balance and Overtime did not have a significant correlation with a p-value of 0.2993. Given these findings, employee's views about Overtime may be subjective, perhaps by differences in generations. We could say however, the company could better control Attrition by paying attention to employee's Work Life Balance and how much Overtime they are working.



```r
cor.test(HRData$OverTime, HRData$WorkLifeBalance, method = "pearson")
##Work life balance goes down with more OT
```

```
        Pearson's product-moment correlation

data:  HRData$OverTime and HRData$WorkLifeBalance
t = -1.0384, df = 1468, p-value = 0.2993
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.07811114  0.02406892
sample estimates:
        cor
-0.02709188
```

```{r}
cor.test(HRData$Attrition, HRData$WorkLifeBalance, method = "pearson")
##Attrition goes down with overtime
```

```
        Pearson's product-moment correlation

data:  HRData$Attrition and HRData$WorkLifeBalance
t = -2.4548, df = 1468, p-value = 0.01421
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.11469157 -0.01285361
sample estimates:
        cor
-0.06393905
```

```{r}
cor.test(HRData$Attrition, HRData$OverTime, method = "pearson")
##Attrition goes up with overtime
```

```
        Pearson's product-moment correlation

data:  HRData$Attrition and HRData$OverTime
t = 9.7292, df = 1468, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.1974754 0.2935516
sample estimates:
     cor
0.246118
```

Works Cited

BasuMallick, Chiradeep. “What Is Employee Attrition? Definition, Attrition Rate, Factors, and Reduction Best Practices.” *Spiceworks*, 11 Mar. 2021, https://www.spiceworks.com/hr/engagement-retention/articles/what-is-attrition-complete-guide/.

Chauhan, Aman. “Employee Analysis: Attrition Report.” *Kaggle*, 12 Sept. 2022, https://www.kaggle.com/datasets/whenamancodes/hr-employee-attrition?select=HR%2BEmployee%2BAttrition.csv.

NetSuite.com. “Why Employees Quit & How to Keep Them.” *Oracle NetSuite*, https://www.netsuite.com/portal/resource/articles/human-resources/employee-turnover-statistics.shtml#:~:text=What%20is%20a%20good%20employee,is%20known%20as%20involuntary%20turnover.