CaseStudy2

Kari Theobald, Shane Weinstock, Jeremey Simpson, Gabe Gonzales

4/3/2019

# EXECUTIVE SUMMARY

DDSAnalytics provided employee profiles on the current workforce of ChemicalRepo. In this dataset we were requested to review for the 3 contributing reasons of attrition in the ChemicalRepo. As Data Scientist, we there is a request from DDSAnalytics management to determine other interesting facts within this data on ChemicalRepo. We conclude the contributing factors for attrition is age of employment and those within the sales representative department. We recommend to either reduce the sales team to those below the age of 25 or increase it beyond 35 years of age. We also recommend on hiring to not hold back from higher employees with at least 8 jobs before they are 45. We conclude this is normal and not can be contributed to jobs held while in high school and college. Overall there is a larger section of men holding employment at ChemicalRepo by 60% of the workforce. We also determine those who live greater than 10 miles from their work are less likely to attrition. In conclusion ChemicalRepo could increase the number of females in their workforce, increase the hiring age beyond 35 years old for new employees and look for employees with a commute greater than 10 miles from the office. Based off the information given, attrition will with reduce by concentration in this area of adjustments.

# HYPOTHESIS

We hypothesize, the sales departments, age and overall commute distance contributes to attrition within a company.

# INTRODUCTION

DDSAnalytics a Fortune 1000 Company, is requesting their Data Scientist to review the data on ChemicalRepo. The management at DDSAnalytics instructs to the Data Scientist, all results from the data must carry the exclusion of all employees under the age of 18. DDSAnalytics also request three factors associated to attrition in ChemicalRepo. DDSAnalytics would like to suggest to ChemicalRepo workforce planning and identify high-potential employees who attribute to turnover. The management in DDSAnalytics will be using this data research as leverage in their talent management business. The Data Scientist will meet the request of reviewing attrition, they will also examine specific job role trends and report the findings.

#install.packages("readxl")  
#install.packages("ggplot2")  
#Establish the working directory.  
library(readxl)

## Warning: package 'readxl' was built under R version 3.5.2

setwd("~/Desktop/Kari/SMU /Doing Data Science/Homework/CaseStudy2/CaseStudy2")

### 2.a ) Read the CSV file (we have xlsx) into the dataset call it Data. Output how many rows and columns

New.Data <- read\_xlsx("~/Desktop/Kari/SMU /Doing Data Science/Homework/CaseStudy2/CaseStudy2/CaseStudy2-data.xlsx")

## readxl works best with a newer version of the tibble package.  
## You currently have tibble v1.4.2.  
## Falling back to column name repair from tibble <= v1.4.2.  
## Message displays once per session.

Data <- data.frame(New.Data)  
head(Data) # original import of all data

## Age Attrition BusinessTravel DailyRate Department  
## 1 41 Yes Travel\_Rarely 1102 Sales  
## 2 49 No Travel\_Frequently 279 Research & Development  
## 3 37 Yes Travel\_Rarely 1373 Research & Development  
## 4 33 No Travel\_Frequently 1392 Research & Development  
## 5 27 No Travel\_Rarely 591 Research & Development  
## 6 32 No Travel\_Frequently 1005 Research & Development  
## DistanceFromHome Education EducationField EmployeeCount EmployeeNumber  
## 1 1 2 Life Sciences 1 1  
## 2 8 1 Life Sciences 1 2  
## 3 2 2 Other 1 4  
## 4 3 4 Life Sciences 1 5  
## 5 2 1 Medical 1 7  
## 6 2 2 Life Sciences 1 8  
## EnvironmentSatisfaction Gender HourlyRate JobInvolvement JobLevel  
## 1 2 Female 94 3 2  
## 2 3 Male 61 2 2  
## 3 4 Male 92 2 1  
## 4 4 Female 56 3 1  
## 5 1 Male 40 3 1  
## 6 4 Male 79 3 1  
## JobRole JobSatisfaction MaritalStatus MonthlyIncome  
## 1 Sales Executive 4 Single 5993  
## 2 Research Scientist 2 Married 5130  
## 3 Laboratory Technician 3 Single 2090  
## 4 Research Scientist 3 Married 2909  
## 5 Laboratory Technician 2 Married 3468  
## 6 Laboratory Technician 4 Single 3068  
## MonthlyRate NumCompaniesWorked Over18 OverTime PercentSalaryHike  
## 1 19479 8 Y Yes 11  
## 2 24907 1 Y No 23  
## 3 2396 6 Y Yes 15  
## 4 23159 1 Y Yes 11  
## 5 16632 9 Y No 12  
## 6 11864 0 Y No 13  
## PerformanceRating RelationshipSatisfaction StandardHours  
## 1 3 1 80  
## 2 4 4 80  
## 3 3 2 80  
## 4 3 3 80  
## 5 3 4 80  
## 6 3 3 80  
## StockOptionLevel TotalWorkingYears TrainingTimesLastYear WorkLifeBalance  
## 1 0 8 0 1  
## 2 1 10 3 3  
## 3 0 7 3 3  
## 4 0 8 3 3  
## 5 1 6 3 3  
## 6 0 8 2 2  
## YearsAtCompany YearsInCurrentRole YearsSinceLastPromotion  
## 1 6 4 0  
## 2 10 7 1  
## 3 0 0 0  
## 4 8 7 3  
## 5 2 2 2  
## 6 7 7 3  
## YearsWithCurrManager  
## 1 5  
## 2 7  
## 3 0  
## 4 0  
## 5 2  
## 6 6

dim(Data) # how many rows and columns in the original data set.

## [1] 1470 35

### 2.b ) Change col names to less than 12 characters in the dataframe.

names <- c( "Age", "Attrition", "Feq.Travel", "Daily.Rate", "Department", "mlg.Home", "Education", "Ed.Field", "Emp.Count", "Emp.Number", "Env.Sat", "Gender", "HourlyRate", "Job.Invol", "Job.Level", "Title", "Satisfied", "M.S.D", "Income.mos","Rate.mos", "Jobs.Worked", "Over18", "OT", "Percent.Inc" , "Performance", "Sat.Relation", "Hours", "Stock", "Yrs.Wrkd", "Training", "Work.Life", "YOS", "YCRole","YbtwnPromo", "YwMgmt")  
  
Data <- setNames(Data, names)  
  
head(Data) #check new names

## Age Attrition Feq.Travel Daily.Rate Department  
## 1 41 Yes Travel\_Rarely 1102 Sales  
## 2 49 No Travel\_Frequently 279 Research & Development  
## 3 37 Yes Travel\_Rarely 1373 Research & Development  
## 4 33 No Travel\_Frequently 1392 Research & Development  
## 5 27 No Travel\_Rarely 591 Research & Development  
## 6 32 No Travel\_Frequently 1005 Research & Development  
## mlg.Home Education Ed.Field Emp.Count Emp.Number Env.Sat Gender  
## 1 1 2 Life Sciences 1 1 2 Female  
## 2 8 1 Life Sciences 1 2 3 Male  
## 3 2 2 Other 1 4 4 Male  
## 4 3 4 Life Sciences 1 5 4 Female  
## 5 2 1 Medical 1 7 1 Male  
## 6 2 2 Life Sciences 1 8 4 Male  
## HourlyRate Job.Invol Job.Level Title Satisfied M.S.D  
## 1 94 3 2 Sales Executive 4 Single  
## 2 61 2 2 Research Scientist 2 Married  
## 3 92 2 1 Laboratory Technician 3 Single  
## 4 56 3 1 Research Scientist 3 Married  
## 5 40 3 1 Laboratory Technician 2 Married  
## 6 79 3 1 Laboratory Technician 4 Single  
## Income.mos Rate.mos Jobs.Worked Over18 OT Percent.Inc Performance  
## 1 5993 19479 8 Y Yes 11 3  
## 2 5130 24907 1 Y No 23 4  
## 3 2090 2396 6 Y Yes 15 3  
## 4 2909 23159 1 Y Yes 11 3  
## 5 3468 16632 9 Y No 12 3  
## 6 3068 11864 0 Y No 13 3  
## Sat.Relation Hours Stock Yrs.Wrkd Training Work.Life YOS YCRole  
## 1 1 80 0 8 0 1 6 4  
## 2 4 80 1 10 3 3 10 7  
## 3 2 80 0 7 3 3 0 0  
## 4 3 80 0 8 3 3 8 7  
## 5 4 80 1 6 3 3 2 2  
## 6 3 80 0 8 2 2 7 7  
## YbtwnPromo YwMgmt  
## 1 0 5  
## 2 1 7  
## 3 0 0  
## 4 3 0  
## 5 2 2  
## 6 3 6

summary(Data)

## Age Attrition Feq.Travel Daily.Rate   
## Min. :18.00 Length:1470 Length:1470 Min. : 102.0   
## 1st Qu.:30.00 Class :character Class :character 1st Qu.: 465.0   
## Median :36.00 Mode :character Mode :character Median : 802.0   
## Mean :36.92 Mean : 802.5   
## 3rd Qu.:43.00 3rd Qu.:1157.0   
## Max. :60.00 Max. :1499.0   
## Department mlg.Home Education Ed.Field   
## Length:1470 Min. : 1.000 Min. :1.000 Length:1470   
## Class :character 1st Qu.: 2.000 1st Qu.:2.000 Class :character   
## Mode :character Median : 7.000 Median :3.000 Mode :character   
## Mean : 9.193 Mean :2.913   
## 3rd Qu.:14.000 3rd Qu.:4.000   
## Max. :29.000 Max. :5.000   
## Emp.Count Emp.Number Env.Sat Gender   
## Min. :1 Min. : 1.0 Min. :1.000 Length:1470   
## 1st Qu.:1 1st Qu.: 491.2 1st Qu.:2.000 Class :character   
## Median :1 Median :1020.5 Median :3.000 Mode :character   
## Mean :1 Mean :1024.9 Mean :2.722   
## 3rd Qu.:1 3rd Qu.:1555.8 3rd Qu.:4.000   
## Max. :1 Max. :2068.0 Max. :4.000   
## HourlyRate Job.Invol Job.Level Title   
## Min. : 30.00 Min. :1.00 Min. :1.000 Length:1470   
## 1st Qu.: 48.00 1st Qu.:2.00 1st Qu.:1.000 Class :character   
## Median : 66.00 Median :3.00 Median :2.000 Mode :character   
## Mean : 65.89 Mean :2.73 Mean :2.064   
## 3rd Qu.: 83.75 3rd Qu.:3.00 3rd Qu.:3.000   
## Max. :100.00 Max. :4.00 Max. :5.000   
## Satisfied M.S.D Income.mos Rate.mos   
## Min. :1.000 Length:1470 Min. : 1009 Min. : 2094   
## 1st Qu.:2.000 Class :character 1st Qu.: 2911 1st Qu.: 8047   
## Median :3.000 Mode :character Median : 4919 Median :14236   
## Mean :2.729 Mean : 6503 Mean :14313   
## 3rd Qu.:4.000 3rd Qu.: 8379 3rd Qu.:20462   
## Max. :4.000 Max. :19999 Max. :26999   
## Jobs.Worked Over18 OT Percent.Inc   
## Min. :0.000 Length:1470 Length:1470 Min. :11.00   
## 1st Qu.:1.000 Class :character Class :character 1st Qu.:12.00   
## Median :2.000 Mode :character Mode :character Median :14.00   
## Mean :2.693 Mean :15.21   
## 3rd Qu.:4.000 3rd Qu.:18.00   
## Max. :9.000 Max. :25.00   
## Performance Sat.Relation Hours Stock   
## Min. :3.000 Min. :1.000 Min. :80 Min. :0.0000   
## 1st Qu.:3.000 1st Qu.:2.000 1st Qu.:80 1st Qu.:0.0000   
## Median :3.000 Median :3.000 Median :80 Median :1.0000   
## Mean :3.154 Mean :2.712 Mean :80 Mean :0.7939   
## 3rd Qu.:3.000 3rd Qu.:4.000 3rd Qu.:80 3rd Qu.:1.0000   
## Max. :4.000 Max. :4.000 Max. :80 Max. :3.0000   
## Yrs.Wrkd Training Work.Life YOS   
## Min. : 0.00 Min. :0.000 Min. :1.000 Min. : 0.000   
## 1st Qu.: 6.00 1st Qu.:2.000 1st Qu.:2.000 1st Qu.: 3.000   
## Median :10.00 Median :3.000 Median :3.000 Median : 5.000   
## Mean :11.28 Mean :2.799 Mean :2.761 Mean : 7.008   
## 3rd Qu.:15.00 3rd Qu.:3.000 3rd Qu.:3.000 3rd Qu.: 9.000   
## Max. :40.00 Max. :6.000 Max. :4.000 Max. :40.000   
## YCRole YbtwnPromo YwMgmt   
## Min. : 0.000 Min. : 0.000 Min. : 0.000   
## 1st Qu.: 2.000 1st Qu.: 0.000 1st Qu.: 2.000   
## Median : 3.000 Median : 1.000 Median : 3.000   
## Mean : 4.229 Mean : 2.188 Mean : 4.123   
## 3rd Qu.: 7.000 3rd Qu.: 3.000 3rd Qu.: 7.000   
## Max. :18.000 Max. :15.000 Max. :17.000

lapply(Data, class) # review the class of the variables in the dataframe.

## $Age  
## [1] "numeric"  
##   
## $Attrition  
## [1] "character"  
##   
## $Feq.Travel  
## [1] "character"  
##   
## $Daily.Rate  
## [1] "numeric"  
##   
## $Department  
## [1] "character"  
##   
## $mlg.Home  
## [1] "numeric"  
##   
## $Education  
## [1] "numeric"  
##   
## $Ed.Field  
## [1] "character"  
##   
## $Emp.Count  
## [1] "numeric"  
##   
## $Emp.Number  
## [1] "numeric"  
##   
## $Env.Sat  
## [1] "numeric"  
##   
## $Gender  
## [1] "character"  
##   
## $HourlyRate  
## [1] "numeric"  
##   
## $Job.Invol  
## [1] "numeric"  
##   
## $Job.Level  
## [1] "numeric"  
##   
## $Title  
## [1] "character"  
##   
## $Satisfied  
## [1] "numeric"  
##   
## $M.S.D  
## [1] "character"  
##   
## $Income.mos  
## [1] "numeric"  
##   
## $Rate.mos  
## [1] "numeric"  
##   
## $Jobs.Worked  
## [1] "numeric"  
##   
## $Over18  
## [1] "character"  
##   
## $OT  
## [1] "character"  
##   
## $Percent.Inc  
## [1] "numeric"  
##   
## $Performance  
## [1] "numeric"  
##   
## $Sat.Relation  
## [1] "numeric"  
##   
## $Hours  
## [1] "numeric"  
##   
## $Stock  
## [1] "numeric"  
##   
## $Yrs.Wrkd  
## [1] "numeric"  
##   
## $Training  
## [1] "numeric"  
##   
## $Work.Life  
## [1] "numeric"  
##   
## $YOS  
## [1] "numeric"  
##   
## $YCRole  
## [1] "numeric"  
##   
## $YbtwnPromo  
## [1] "numeric"  
##   
## $YwMgmt  
## [1] "numeric"

# Results indicate we are operating with character and numeric classes.   
  
# Tidy the data to manageable dataframe.   
# Reduce the number of col in the dataframe and call the new dataframe Small.  
Small<- data.frame(Data$Attrition, Data$Age, Data$Gender, Data$M.S.D, Data$YOS, Data$Education, Data$Ed.Field, Data$Title, Data$Rate.mos, Data$mlg.Home)   
  
#Change col names in the smaller dataframe.   
names <- c( "Attrition", "Age", "Gender", "M.S.D", "YOS", "Ed.Lvl", "Education", "Title", "Income.mos", "mlg.Home")  
Small <- setNames (Small, names)  
head(Small)

## Attrition Age Gender M.S.D YOS Ed.Lvl Education  
## 1 Yes 41 Female Single 6 2 Life Sciences  
## 2 No 49 Male Married 10 1 Life Sciences  
## 3 Yes 37 Male Single 0 2 Other  
## 4 No 33 Female Married 8 4 Life Sciences  
## 5 No 27 Male Married 2 1 Medical  
## 6 No 32 Male Single 7 2 Life Sciences  
## Title Income.mos mlg.Home  
## 1 Sales Executive 19479 1  
## 2 Research Scientist 24907 8  
## 3 Laboratory Technician 2396 2  
## 4 Research Scientist 23159 3  
## 5 Laboratory Technician 16632 2  
## 6 Laboratory Technician 11864 2

### 3.a ) Is anyone under 18 participating in the study?

Over18 <- grep("N", Small$Over18)  
Over18

## integer(0)

Response is 0. So, all are considered over 18 by their response. There are 8 people who state they are 18 but we understand this response to be 18, plus a few days or months. There is no exclusion of data because of age in the study.

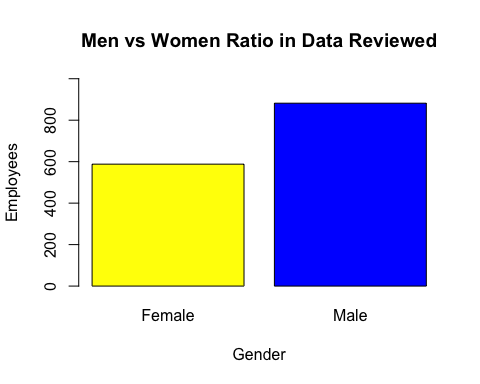
### 3.b) Provide descriptive Statistics on at least 7 variables.

summary(Small)

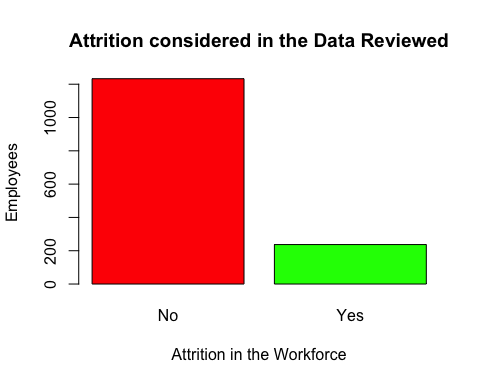
## Attrition Age Gender M.S.D YOS   
## No :1233 Min. :18.00 Female:588 Divorced:327 Min. : 0.000   
## Yes: 237 1st Qu.:30.00 Male :882 Married :673 1st Qu.: 3.000   
## Median :36.00 Single :470 Median : 5.000   
## Mean :36.92 Mean : 7.008   
## 3rd Qu.:43.00 3rd Qu.: 9.000   
## Max. :60.00 Max. :40.000   
##   
## Ed.Lvl Education Title   
## Min. :1.000 Human Resources : 27 Sales Executive :326   
## 1st Qu.:2.000 Life Sciences :606 Research Scientist :292   
## Median :3.000 Marketing :159 Laboratory Technician :259   
## Mean :2.913 Medical :464 Manufacturing Director :145   
## 3rd Qu.:4.000 Other : 82 Healthcare Representative:131   
## Max. :5.000 Technical Degree:132 Manager :102   
## (Other) :215   
## Income.mos mlg.Home   
## Min. : 2094 Min. : 1.000   
## 1st Qu.: 8047 1st Qu.: 2.000   
## Median :14236 Median : 7.000   
## Mean :14313 Mean : 9.193   
## 3rd Qu.:20462 3rd Qu.:14.000   
## Max. :26999 Max. :29.000   
##

### 3.c) Provide a histogram for 2 of the variables.

# Plot Gender vs Number of Employees  
plot(Small$Gender, xlab = "Gender", ylab= "Employees", main="Men vs Women Ratio in Data Reviewed", ylim=c(0,1000), col=c("Yellow", "Blue"))



#Plot Number of Employee Attrition vs Total Number of Employees.  
plot(Small$Attrition , xlab = "Attrition in the Workforce", ylab= "Employees", main="Attrition considered in the Data Reviewed", col= c("Red", "Green"))



# 3.d) Give Frequency for Gender, Education and Occupation. How many are in management positions?

summary(Small$Gender) #Gender Summary Statistics.

## Female Male   
## 588 882

# Result : Female Male   
# 588 882   
summary(Small$Education) #Education Field Summary Statistics.

## Human Resources Life Sciences Marketing Medical   
## 27 606 159 464   
## Other Technical Degree   
## 82 132

# Result : Human Resources Life Sciences Marketing Medical Other Technical Degree   
# 27 606 159 464 82 132   
summary(Small$Ed.Lvl) #Education Level Summary Statistics.

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.000 2.000 3.000 2.913 4.000 5.000

# Education  
# 1 'Below College'  
# 2 'College'  
# 3 'Bachelor'  
# 4 'Master'  
# 5 'Doctor'  
  
summary(Small$Title) #Department titles and summary statistics.

## Healthcare Representative Human Resources   
## 131 52   
## Laboratory Technician Manager   
## 259 102   
## Manufacturing Director Research Director   
## 145 80   
## Research Scientist Sales Executive   
## 292 326   
## Sales Representative   
## 83

#Healthcare Representative Human Resources Laboratory Technician Manager   
# 131 52 259 102   
# Manufacturing Director Research Director Research Scientist Sales Executive   
# 145 80 292 326   
# Sales Representative   
# 83   
#

Result : Number of management positions (Mfg and Research Directors, Sales Exec and Managers) Total (145 + 80 + 326 + 102) = 653

### Using only the data where Attrition is True.

setwd("~/Desktop/Kari/SMU /Doing Data Science/Homework/CaseStudy2/CaseStudy2")  
New.Data <- read\_xlsx("~/Desktop/Kari/SMU /Doing Data Science/Homework/CaseStudy2/CaseStudy2/CaseStudy2-data.xlsx")  
  
Data <- data.frame(New.Data)  
  
names <- c( "Age", "Attrition", "Feq.Travel", "Daily.Rate", "Department", "mlg.Home", "Education", "Ed.Field", "Emp.Count", "Emp.Number", "Env.Sat", "Gender", "HourlyRate", "Job.Invol", "Job.Level", "Title", "Satisfied", "M.S.D", "Income.mos","Rate.mos", "Jobs.Worked", "Over18", "OT", "Percent.Inc" , "Performance", "Sat.Relation", "Hours", "Stock", "Yrs.Wrkd", "Training", "Work.Life", "YOS", "YCRole","YbtwnPromo", "YwMgmt")  
  
Data <- setNames(Data, names)  
  
Small <- data.frame(Data$Attrition, Data$Age, Data$Gender, Data$M.S.D, Data$YOS, Data$Ed.Field, Data$Title, Data$Rate.mos, Data$mlg.Home)   
  
#Change col names in the smaller dataframe.   
names <- c( "Attrition", "Age", "Gender", "M.S.D", "YOS", "Education", "Title", "Income.mos", "mlg.Home")  
Small <- setNames (Small, names)  
  
library(plyr)  
Small$Attrition <- revalue(Small$Attrition, c("Yes"=1, "No" = 0)) # DF Small, change in the Attrition column all Yes answers to 1 and all No answers to 0.  
head(Small$Attrition) # show the dataframe Small was revalued with 1's and 0's in the Attrition column.

## [1] 1 0 1 0 0 0  
## Levels: 0 1

True.Attrition<- subset(Small, Small$Attrition==1) #create a dataframe called True.Attrition out of the dataframe Small. Pull out only the rows in Small for when the Attrition column is = 1.  
head(True.Attrition) #show only the variables names the dataframe True.Attrition.

## Attrition Age Gender M.S.D YOS Education Title  
## 1 1 41 Female Single 6 Life Sciences Sales Executive  
## 3 1 37 Male Single 0 Other Laboratory Technician  
## 15 1 28 Male Single 4 Life Sciences Laboratory Technician  
## 22 1 36 Male Single 5 Life Sciences Sales Representative  
## 25 1 34 Male Single 4 Medical Research Scientist  
## 27 1 32 Female Single 10 Life Sciences Research Scientist  
## Income.mos mlg.Home  
## 1 19479 1  
## 3 2396 2  
## 15 12947 24  
## 22 6986 9  
## 25 17102 6  
## 27 4681 16

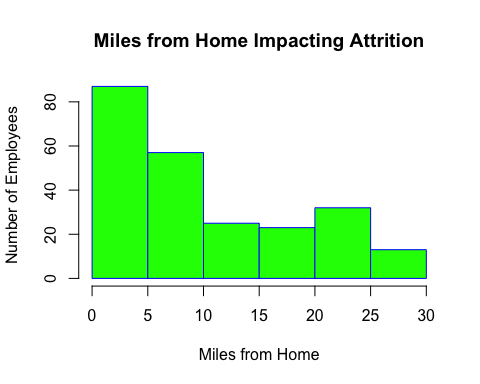
summary(True.Attrition) #summary statistics on variables impacted by attrition.

## Attrition Age Gender M.S.D YOS   
## 0: 0 Min. :18.00 Female: 87 Divorced: 33 Min. : 0.000   
## 1:237 1st Qu.:28.00 Male :150 Married : 84 1st Qu.: 1.000   
## Median :32.00 Single :120 Median : 3.000   
## Mean :33.61 Mean : 5.131   
## 3rd Qu.:39.00 3rd Qu.: 7.000   
## Max. :58.00 Max. :40.000   
##   
## Education Title Income.mos   
## Human Resources : 7 Laboratory Technician :62 Min. : 2326   
## Life Sciences :89 Sales Executive :57 1st Qu.: 8870   
## Marketing :35 Research Scientist :47 Median :14618   
## Medical :63 Sales Representative :33 Mean :14559   
## Other :11 Human Resources :12 3rd Qu.:21081   
## Technical Degree:32 Manufacturing Director:10 Max. :26999   
## (Other) :16   
## mlg.Home   
## Min. : 1.00   
## 1st Qu.: 3.00   
## Median : 9.00   
## Mean :10.63   
## 3rd Qu.:17.00   
## Max. :29.00   
##

When Years of Service is 0, the employee is indicating they have worked less than 1 year.

### Graph a histogram of Attrition employees compared to commute distance.

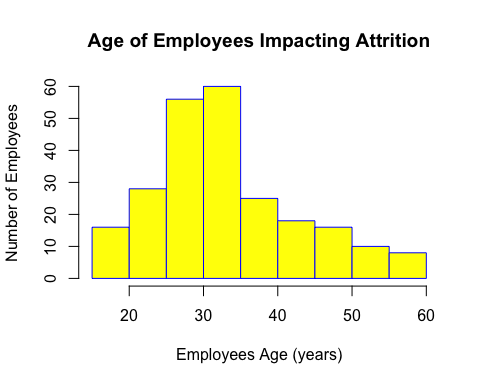
hist(True.Attrition$mlg.Home,   
 main="Miles from Home Impacting Attrition",   
 xlab="Miles from Home",  
 ylab="Number of Employees ",   
 border="blue",   
 col="green")



Results conclude distance from home appears to impact attrition. There is right skew to the data. Indicating those living close to their place of employment will attrition out of the company.

### Graph a histogram of Attrition employees compared to age of employee.

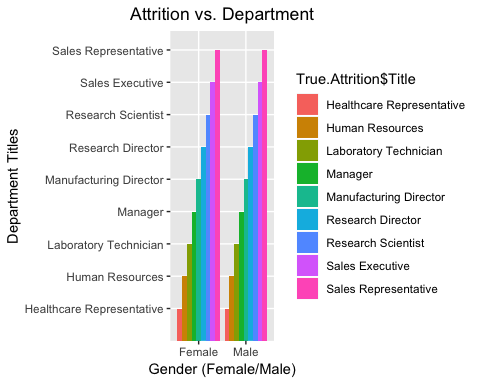
hist(True.Attrition$Age,  
 main="Age of Employees Impacting Attrition",   
 xlab="Employees Age (years)",  
 ylab="Number of Employees",  
 border = "blue",  
 col="yellow")



The graphical results show a right skew. The older the age of the employee, it is less likely for the company to experience attrition. We show a spike of attrition between the age of 25 to 35. Outside of the age bracket the attrition goes down.

### Other components impacting attrition. Review Department Titles compared to Gender.

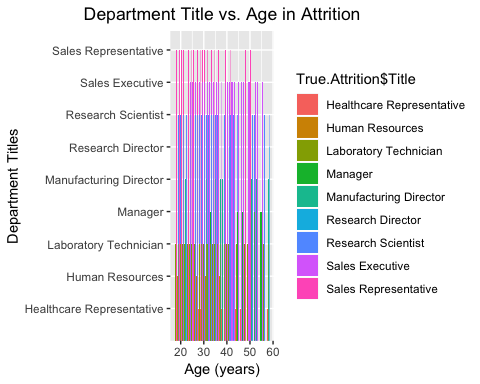
library (ggplot2)  
par(las=1)  
Attrition.Gender <- ggplot(True.Attrition, aes(x=True.Attrition$Gender, y=True.Attrition$Title, fill = True.Attrition$Title)) + geom\_bar(stat="identity", position="dodge") + xlab("Gender (Female/Male)") + ylab("Department Titles") + ggtitle("Attrition vs. Department") + theme(plot.title = element\_text(hjust = .5))  
  
Attrition.Gender



Results indicate attrition is dominant in the Sales Executives and Sales Representative field for both men and women. This could largely be due to travel required.

### Determine what is the age bracket is for each department that is showing attrition.

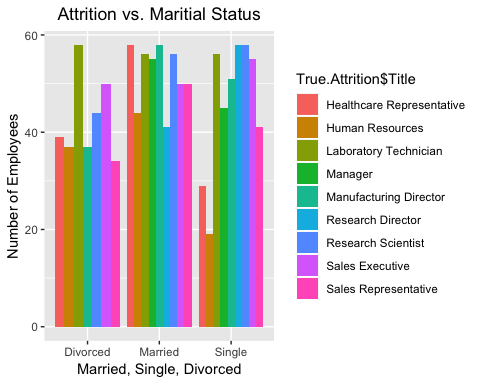
par(las=1)  
SalesvsAge <- ggplot(True.Attrition, aes(x=True.Attrition$Age, y=True.Attrition$Title, fill = True.Attrition$Title)) + geom\_bar(stat="identity", position="dodge") + xlab("Age (years)") + ylab("Department Titles") + ggtitle("Department Title vs. Age in Attrition") + theme(plot.title = element\_text(hjust = .5))  
  
SalesvsAge



Results indicate the sales representative are dominantly 18-40, which coincides with the overall average age for attrition which approximately 25-35 year old. We recommend to keep a younger sales team below the ### age of 25 or over the age of 35 to assist in reducing attrition. The workforce does have sales representative over 45 and in their mid 50’s and the attrition rate of these age groups is significantly ### less.

### Other influences on attrition? Does being married, single or divorced have any impact on Sales Representatives. We show married people are less likely to cause attrition than single and divorced employees. What does this company have in attrition and marital status?

par(las=1)  
SalesvsMSD <- ggplot(True.Attrition, aes(x=True.Attrition$M.S.D, y=True.Attrition$Age, fill = True.Attrition$Title)) + geom\_bar(stat="identity", position="dodge") + xlab("Married, Single, Divorced") + ylab("Number of Employees") + ggtitle("Attrition vs. Maritial Status") + theme(plot.title = element\_text(hjust = .5))  
SalesvsMSD



Results indicate there is dominating feature in marital status for Sales Representatives in the company. In fact, we have a significant number who are married and with married people showing less probability to attrition we do conclude marital status to be a role in attrition with this company in this department/title.

### 4.c Is there a relationship between age and income, color each point based off of gender?

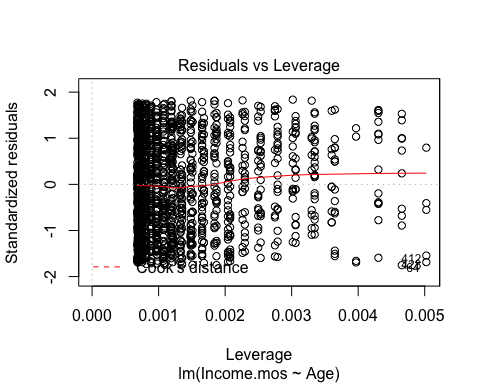
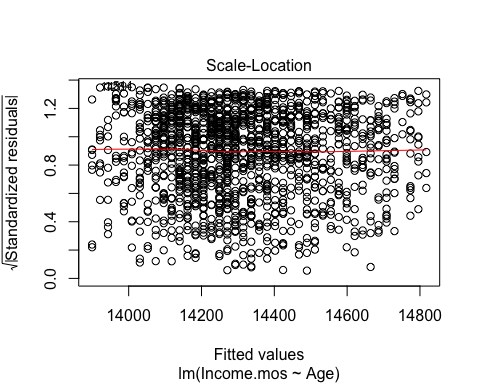
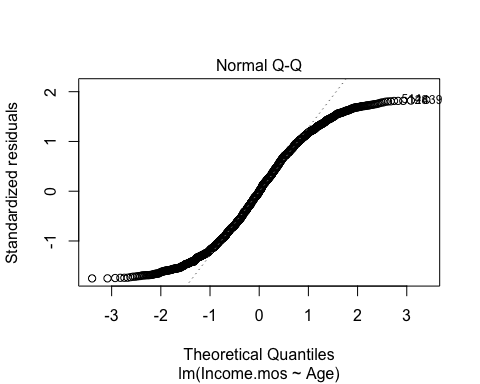
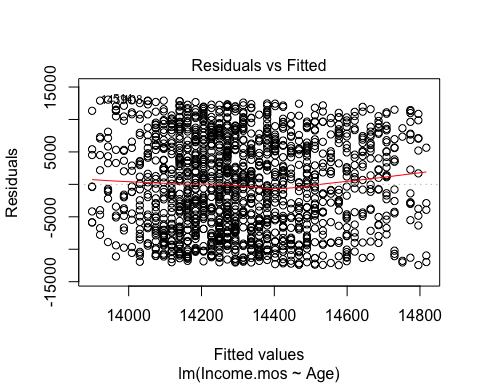
library(ggplot2)  
par(las=2)  
  
## Overall test of data when comparing women to men in the workforce at ChemicalRepo.  
Age2Income<-ggplot(Small, aes(x=Age, y=Income.mos, group=Gender, color =Gender)) + geom\_point(aes(color = Gender)) + stat\_smooth(method="lm") + xlab("Age (years) of Employee") + ylab("Monthly Income") + ggtitle("Workforce Income compared to Age") + theme(plot.title = element\_text(hjust = 0.5))  
Age2Income



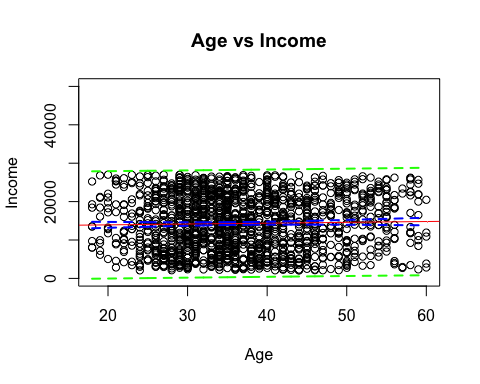
# Overall test when comparing both men and women in the workforce as one.   
# Original Linear Regression to discover correlation between Age and Income  
ageincome.lm <- lm(Income.mos ~ Age, data = Small)  
ageincome.res <- resid(ageincome.lm)  
summary(ageincome.lm)

##   
## Call:  
## lm(formula = Income.mos ~ Age, data = Small)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -12452 -6193 -45 6111 13056   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 13506.10 773.19 17.468 <2e-16 \*\*\*  
## Age 21.86 20.33 1.075 0.282   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 7117 on 1468 degrees of freedom  
## Multiple R-squared: 0.0007869, Adjusted R-squared: 0.0001062   
## F-statistic: 1.156 on 1 and 1468 DF, p-value: 0.2825

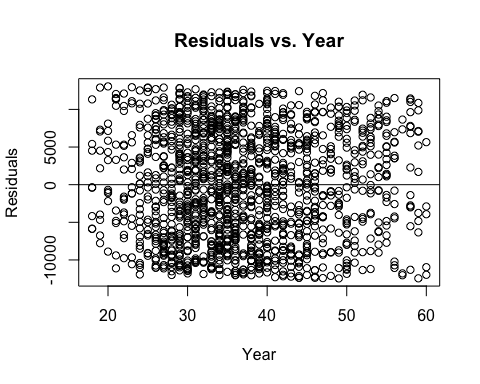
# Simple Regression Model  
plot(ageincome.lm)



plot(x = Small$Age, y = Small$Income.mos, ylim=c(0,50000), xlab = "Age", ylab = "Income", main = "Age vs Income")  
displayDF <- data.frame(Small)  
  
# Regression Model  
## Add the regression line to the existing scatterplot  
abline(ageincome.lm, col = "red")  
## Create "new" data to make confidence and prediction intervals  
newx <- displayDF$Age  
newx <- sort(newx)  
## Confidence Internal  
conf <- predict(ageincome.lm, newdata = data.frame(Age = newx), interval = c("confidence"),   
 type = c("response"), level = .95)  
## Prediction Interval  
pred <- predict(ageincome.lm, newdata = data.frame(Age = newx), interval = c("predict"),   
 type = c("response"), level = .95)  
## Add prediction and confidence intervals to the scatterplot  
lines(newx, conf[,2], col = "blue", lty = 2, lwd = 2)  
lines(newx, conf[,3], col = "blue", lty = 2, lwd = 2)  
lines(newx, pred[,2], col = "green", lty = 2, lwd = 2)  
lines(newx, pred[,3], col = "green", lty = 2, lwd = 2)



## Residual Plot  
plot(Small$Age, ageincome.res, ylab = "Residuals", xlab = "Year", main = "Residuals vs. Year")  
abline(0,0)



# Requires Transformation - log or log/log  
Small$logAge <- log(Small$Age)  
Small$logIncome <- log(Small$Income.mos)  
# Checking to see if Log Income  
logageincome.lm <- lm(logIncome ~ Age, data = Small)  
logageincome.res <- resid(logageincome.lm)  
summary(logageincome.lm)

##   
## Call:  
## lm(formula = logIncome ~ Age, data = Small)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.7662 -0.3959 0.1635 0.5231 0.8267   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 9.346636 0.068824 135.806 <2e-16 \*\*\*  
## Age 0.001508 0.001809 0.834 0.405   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.6335 on 1468 degrees of freedom  
## Multiple R-squared: 0.0004732, Adjusted R-squared: -0.0002077   
## F-statistic: 0.6949 on 1 and 1468 DF, p-value: 0.4046

logageincome2.lm <- lm(logIncome ~ logAge, data = Small)  
summary(logageincome2.lm)

##   
## Call:  
## lm(formula = logIncome ~ logAge, data = Small)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.7620 -0.4008 0.1645 0.5225 0.8196   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 9.28922 0.23700 39.196 <2e-16 \*\*\*  
## logAge 0.03161 0.06608 0.478 0.632   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.6336 on 1468 degrees of freedom  
## Multiple R-squared: 0.0001559, Adjusted R-squared: -0.0005252   
## F-statistic: 0.2289 on 1 and 1468 DF, p-value: 0.6324

With an extremely low R-Squared and no transformation narrowing our results for comparison, it is safe to say that there is not a correlation between Age and Income across these many variables. This would likely change given we select a smaller data set or a particular job, there is simply too much variation across all 1,400 data points.

When we review the relationship between age to income in a regression split between men and women there is a small variance. Women have no significant change between age and income while men have a small increase in pay as they become older. Overall the variance is negligible at Chemicalrepo.

### 4.d What about Life Satisfaction? Is there a discernable relationship there to what? (Trends and Observations) In this study we margined out the Research Scientist. We reviewed the overall satisfaction levels of Research Scientist relative to the distance they travel to work, their Age, their monthly Income, Marital Status and Years between Promotions.

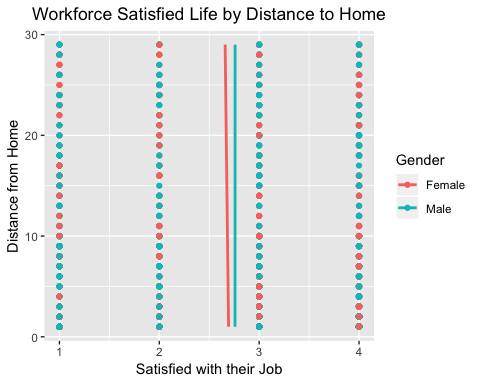
library(ggplot2)  
par(las=2)  
  
library(plyr)  
DSJob <- subset(Data, Title=="Research Scientist") #subset out from the large dataframe Data only those who have the title Research Scientist.

### Other comparisons.

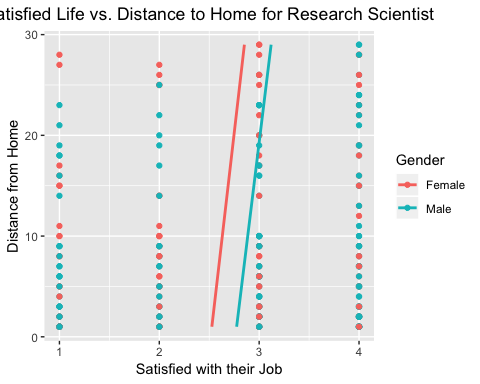
### Compare Satisfaction in the overall workforce in ChemicalRepo by commute distance. (miles from home).

### Compare Satisfaction by Research Scientist in ChemicalRepo by their commute.’

# Total Workforce comparison of Satisfaction vs miles from home.  
LifeSat2Home<-ggplot(Data, aes(x=mlg.Home, y=Satisfied, group=Gender, color=Gender)) + geom\_point(aes(color = Gender)) + stat\_smooth(method=lm, se=FALSE) + xlab("Distance from Home") + ylab("Satisfied with their Job") + ggtitle("Workforce Satisfied Life by Distance to Home") + theme(plot.title = element\_text(hjust = 0.5)) + coord\_flip()  
LifeSat2Home



# Research Scientist Satisfied vs miles from home.  
LifeSat2HomeDS<-ggplot(DSJob, aes(x=mlg.Home, y=Satisfied, group=Gender, color=Gender)) + geom\_point(aes(color = Gender)) + stat\_smooth(method=lm, se=FALSE) + xlab("Distance from Home") + ylab("Satisfied with their Job") + ggtitle("Satisfied Life vs. Distance to Home for Research Scientist") + theme(plot.title = element\_text(hjust = 0.5)) + coord\_flip()  
LifeSat2HomeDS

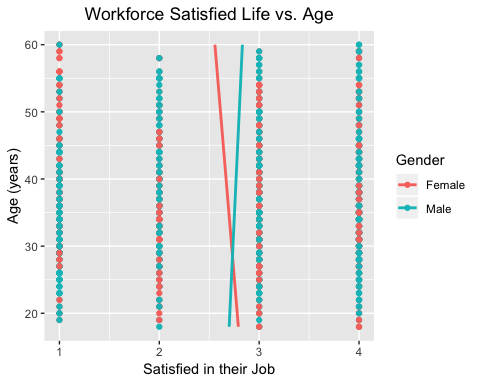


In a scale of 1 to 4 with 4 being Very Highly in satisfaction, we conclude the employees as a whole in ChemicalRepo report a Medium to High satisfaction with the distance in commute. Men appear to be more satisfied than women in the workforce by their commute distance. When we review Research Scientist, they too have a positive relationship to the company as men are more satisfied than women in how far they commute to work. Interesting factor, both men and women who work as Research Scientist find a higher satisfaction with the company the more miles they drive to work. The longer the Research Scientist commutes the happier they are within the company.

### Compare Satisfaction in the overall workforce in ChemicalRepo by Age of employee.

### Compare Satisfaction by Research Scientist in ChemicalRepo Age of employee.

# Total Workforce comparison of Satisfaction vs Age of employee.  
Workforce.Age<-ggplot(Data, aes(x=Age, y=Satisfied, group=Gender, color=Gender)) + geom\_point(aes(color = Gender)) + stat\_smooth(method=lm, se=FALSE) + xlab("Age (years)") + ylab("Satisfied in their Job") + ggtitle("Workforce Satisfied Life vs. Age") + theme(plot.title = element\_text(hjust = 0.5)) + coord\_flip()  
Workforce.Age



# Research Scientist Satisfied vs Age of employee.  
DS.Age<-ggplot(DSJob, aes(x=Age, y=Satisfied, group=Gender, color=Gender)) + geom\_point(aes(color = Gender)) + stat\_smooth(method="lm", se=FALSE) + xlab("Age (years) of Research Scientist") + ylab("Satisfied in their Job") + ggtitle("The Age of a Research Scientist Satisfied in their Job") + theme(plot.title = element\_text(hjust = 0.5))  
DS.Age

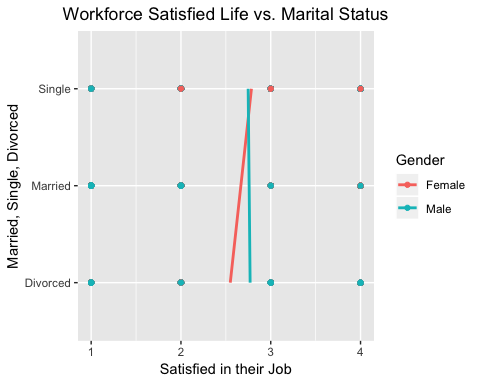


In a scale of 1 to 4 with 4 being Very Highly in satisfaction, we conclude the employees as a whole in ChemicalRepo with men having a higher job satisfaction the older men become and older women become less satisfied within the company. When reviewing the Research Scientist department, we find a very similar comparison for men and women. Women decline in job satisfaction at ChemicalRepo at an older age. While men increase within a small margin in job satisfaction at an older age in this job title at ChemicalRepo.

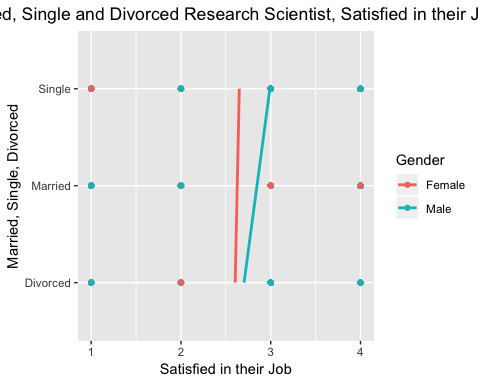
### Compare Satisfaction in the overall workforce in ChemicalRepo by Marital Status.

### Compare Satisfaction by Research Scientist in ChemicalRepo by Marital Status.

# Total Workforce comparison of Satisfaction vs Marital Status.  
Workforce.Alter<-ggplot(Data, aes(x=M.S.D, y=Satisfied, group=Gender, color=Gender)) + geom\_point(aes(color = Gender)) + stat\_smooth(method=lm, se=FALSE) + xlab("Married, Single, Divorced") + ylab("Satisfied in their Job") + ggtitle("Workforce Satisfied Life vs. Marital Status") + theme(plot.title = element\_text(hjust = 0.5)) + coord\_flip()  
Workforce.Alter



# Research Scientist Satisfied vs Marital Status.  
DS.Alter<-ggplot(DSJob, aes(x=M.S.D, y=Satisfied, group=Gender, color=Gender)) + geom\_point(aes(color = Gender)) + stat\_smooth(method="lm", se=FALSE) + xlab("Married, Single, Divorced") + ylab("Satisfied in their Job") + ggtitle("Married, Single and Divorced Research Scientist, Satisfied in their Job") + theme(plot.title = element\_text(hjust = 0.5)) + coord\_flip()  
DS.Alter

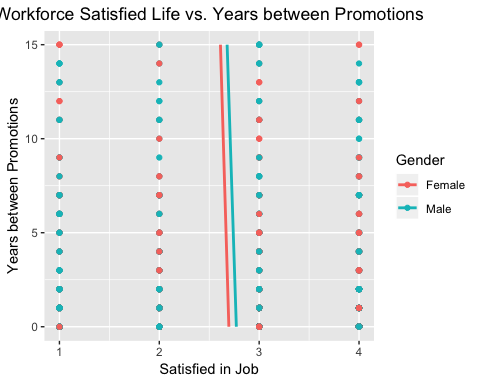


In a scale of 1 to 4 with 4 being Very Highly in satisfaction, we conclude the employees as a whole in ChemicalRepo being male, are Medium to Highly satisfied. We conclude women who are divorced are less satisfied as single women. Yet married women are demonstrating the highest satisfaction rating between the 3 types of marital status. Married women are almost Highly satisfied at ChemicalRepo. When reviewing Research Scientist women are consistent in job satisfaction at a Medium to Highly satisfied ranking. While men who are married are closer to highly satisfied compared to the respectively lower, Single and Divorced Research Scientist.

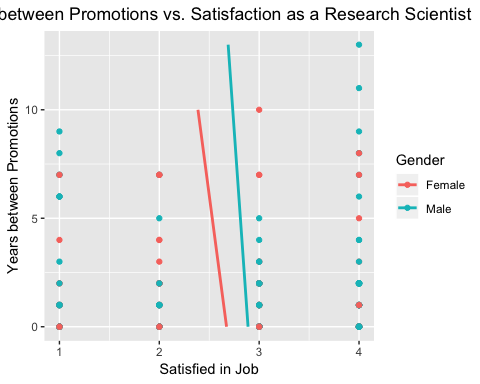
### Compare Satisfaction in the overall workforce in ChemicalRepo by Yrs between promotion.

### Compare Satisfaction by Research Scientist in ChemicalRepo by Yrs between promotion.

# Total Workforce comparison of Satisfaction vs Years between Promotions.  
Workforce.Promo<-ggplot(Data, aes(x=YbtwnPromo, y=Satisfied, group=Gender, color=Gender)) + geom\_point(aes(color = Gender)) + stat\_smooth(method=lm, se=FALSE) + xlab("Years between Promotions") + ylab("Satisfied in Job") + ggtitle("Workforce Satisfied Life vs. Years between Promotions") + theme(plot.title = element\_text(hjust = 0.5)) + coord\_flip()  
Workforce.Promo



# Research Scientist Satisfied vs Years between Promotions.   
DS.Promo <- ggplot(DSJob, aes(x=YbtwnPromo, y=Satisfied, group=Gender, color=Gender)) + geom\_point(aes(color = Gender)) + stat\_smooth(method="lm", se=FALSE) + xlab("Years between Promotions") + ylab("Satisfied in Job") + ggtitle("Years between Promotions vs. Satisfaction as a Research Scientist") + theme(plot.title = element\_text(hjust = 0.5)) + coord\_flip()  
DS.Promo

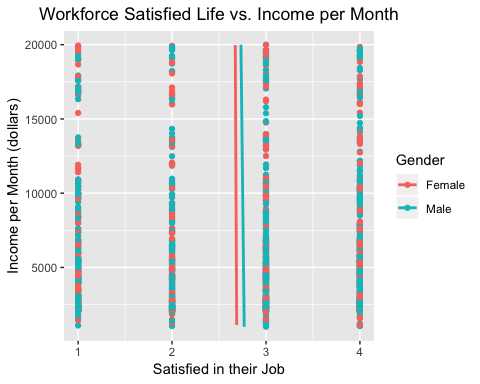


In a scale of 1 to 4 with 4 being Very Highly in satisfaction, we conclude the employees as a whole in ChemicalRepo are between Medium to Highly Satisfied. Both women and men become slightly less satisfied in their Jobs the longer they wait between promotions. In review of employees titled as Research Scientist, women become less satisfied at a faster rate than men in years between promotions. Women max out at 10 years between promotion but men continue to wait longer as Research Scientist. This could be due to years of service within the company rather than the candidate being ignored for other reasons.

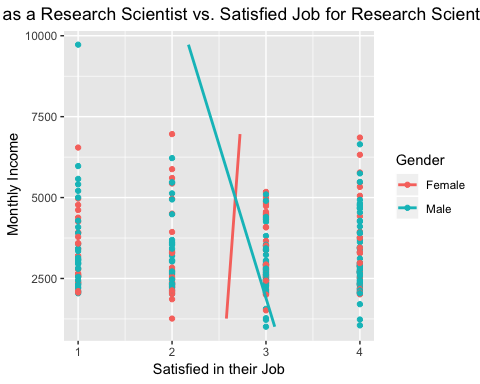
### Compare Satisfaction in the overall workforce in ChemicalRepo by Income earned per month.

### Compare Satisfaction by Research Scientist in ChemicalRepo by Income earned per month.

# Total Workforce comparison of Satisfaction vs Income per month.  
Workforce.Income<-ggplot(Data, aes(x=Income.mos, y=Satisfied, group=Gender, color=Gender)) + geom\_point(aes(color = Gender)) + stat\_smooth(method=lm, se=FALSE) + xlab("Income per Month (dollars)") + ylab("Satisfied in their Job") + ggtitle("Workforce Satisfied Life vs. Income per Month") + theme(plot.title = element\_text(hjust = 0.5)) + coord\_flip()  
Workforce.Income



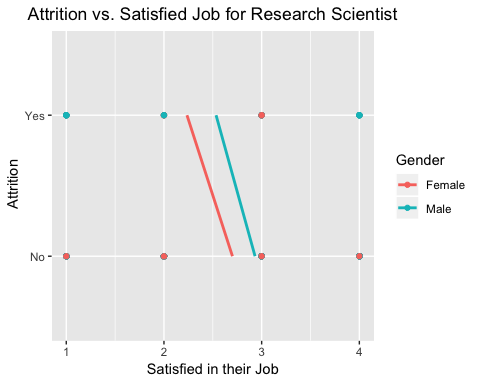
# Compare income compared to job satisfaction (male vs female).  
DS.Income <- ggplot(DSJob, aes(x=Income.mos, y=Satisfied, group=Gender, color=Gender)) + geom\_point(aes(color = Gender)) + stat\_smooth(method="lm", se=FALSE) + xlab(" Monthly Income") + ylab("Satisfied in their Job") + ggtitle("Income as a Research Scientist vs. Satisfied Job for Research Scientist") + theme(plot.title = element\_text(hjust = 0.5)) + coord\_flip()  
DS.Income



In a scale of 1 to 4 with 4 being Very Highly in satisfaction, we conclude the employees as a whole in ChemicalRepo are between Medium to Highly Satisfied. While men have a slightly higher approval rating in job satisfaction compared to women, the difference is marginal. The results indicate as a Research Scientist, men are paid a higher monthly income overall. The more men are paid the less satisfied they are with work. They start with ranking their satisfaction beyond Highly Satisfied but as they increase past $5k a month, men go below the job satisfaction of women. Men who make almost $10k a month rank a Medium Satisfaction with the job. Which is in reverse to most would consider as a normal response to an increase in pay. We can contribute this to more demands on the male employees who are making more income. However, the more women are paid the more satisfied they are with working.

### How satisfied are Research Scientist in their job at ChemicalRepo.

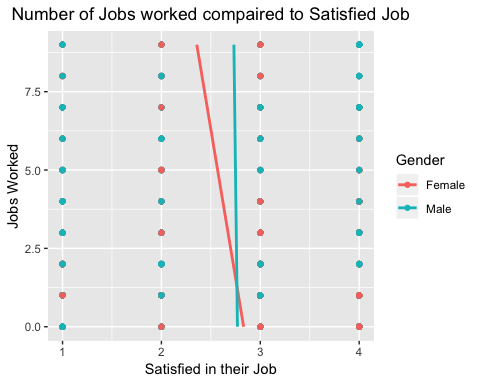
DS.Attrition <- ggplot(DSJob, aes(x=Attrition, y=Satisfied, group=Gender, color=Gender)) + geom\_point(aes(color = Gender)) + stat\_smooth(method="lm", se=FALSE) + xlab(" Attrition") + ylab("Satisfied in their Job") + ggtitle("Attrition vs. Satisfied Job for Research Scientist") + theme(plot.title = element\_text(hjust = 0.5)) + coord\_flip()  
  
DS.Attrition



In a scale of 1 to 4 with 4 being Very Highly in satisfaction, we conclude Research Scientist leaving the company are still above medium rankings for being satisfied with their job. Women leaving are slightly less pleased in the job than men. The same is true for those currently working within the company. We find Women are less satisfied in their job as Research Scientist than men. However, they both rank the job satisfaction between a Medium to High satisfaction.

### 4.d Looking only at life satisfaction data only for all of the employees in ChemicalRepo.

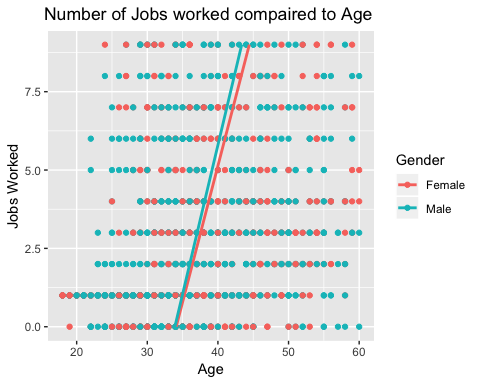
library(ggplot2)  
par(las=2)  
  
Turnover <- ggplot(Data, aes(x=Jobs.Worked, y=Satisfied, group=Gender, color=Gender)) + geom\_point(aes(color = Gender)) + stat\_smooth(method="lm", se=FALSE) + xlab("Jobs Worked") + ylab("Satisfied in their Job") + ggtitle("Number of Jobs worked compaired to Satisfied Job") + theme(plot.title = element\_text(hjust = 0.5)) + coord\_flip()  
Turnover



Results indicate women are happier when they start in the workforce. Women who frequently change employment demonstrate they have a less satisfied work experience. Men however can change the jobs equal to women and maintain the same satisfaction as they did when they took their 1st job. There is no impact on men in being satisfied with work compared to the number of times men #change #employment. Overall women and men share equal satisfaction working when they have changed employer less than 2.5 times.

### How many former jobs is considered normal in ChemicalRepo?

Age.Jobs <- ggplot(Data, aes(x=Jobs.Worked, y=Age, group=Gender, color=Gender)) + geom\_point(aes(color = Gender)) + stat\_smooth(method="lm", se=FALSE) + xlab("Jobs Worked") + ylab("Age") + ggtitle("Number of Jobs worked compaired to Age ") + theme(plot.title = element\_text(hjust = 0.5)) + coord\_flip()  
  
Age.Jobs



Results also indicate on average both women and men between the approximate age of 33 to 45 change jobs at almost the same rate. On average both men and women by the age of 40 have changed jobs almost 5 times. When compared to the maximum attrition age between 25-35 we conclude both men and women in their 30’s could be working at their first job after college. We find the attrition relative to the employee taking their first job in the workforce. It is relative to note the number of jobs at a young age could be summer high school jobs and as a student working. The stronger age bracket for less attrition are those over 35, be aware, those who work between 1 to almost 8 jobs between the age of 33-45 is considered average. There is no apparent risk by this age bracket and this many number of jobs worked.

# CONCLUSION

We hypothesized sales representatives, age and overall commute distance contributes to attrition within a company. We find there is evidence to conclude the sales representatives, age and commute influence attrition. In addition, we find there no guarantee the longer someone works for a company the more income they will earn. The variance of any increase in pay base off of age is negligible compared to years within a company and the multiple departments in ChemicalRepo. Overall our hypothesis did prove to be correct on attrition and our hypothesis is correct. We do add ChemicalRepo could increae the income in Female employees and increase the job satisfaction. We also find commuting distance from work to home is a positive relationship. Those who work farther from home are most satisfied in the job as a Research Scientist.

sessionInfo()

## R version 3.5.1 (2018-07-02)  
## Platform: x86\_64-apple-darwin15.6.0 (64-bit)  
## Running under: macOS Sierra 10.12.6  
##   
## Matrix products: default  
## BLAS: /Library/Frameworks/R.framework/Versions/3.5/Resources/lib/libRblas.0.dylib  
## LAPACK: /Library/Frameworks/R.framework/Versions/3.5/Resources/lib/libRlapack.dylib  
##   
## locale:  
## [1] en\_US.UTF-8/en\_US.UTF-8/en\_US.UTF-8/C/en\_US.UTF-8/en\_US.UTF-8  
##   
## attached base packages:  
## [1] stats graphics grDevices utils datasets methods base   
##   
## other attached packages:  
## [1] ggplot2\_3.1.1 plyr\_1.8.4 readxl\_1.3.1   
##   
## loaded via a namespace (and not attached):  
## [1] Rcpp\_1.0.0 bindr\_0.1.1 knitr\_1.21 magrittr\_1.5   
## [5] tidyselect\_0.2.5 munsell\_0.5.0 colorspace\_1.3-2 R6\_2.3.0   
## [9] rlang\_0.3.1 stringr\_1.3.1 dplyr\_0.7.8 tools\_3.5.1   
## [13] grid\_3.5.1 gtable\_0.2.0 xfun\_0.4 withr\_2.1.2   
## [17] htmltools\_0.3.6 assertthat\_0.2.0 yaml\_2.2.0 lazyeval\_0.2.1   
## [21] digest\_0.6.18 tibble\_1.4.2 crayon\_1.3.4 bindrcpp\_0.2.2   
## [25] purrr\_0.3.0 codetools\_0.2-15 glue\_1.3.0 evaluate\_0.12   
## [29] rmarkdown\_1.11 labeling\_0.3 stringi\_1.2.4 compiler\_3.5.1   
## [33] pillar\_1.3.1 cellranger\_1.1.0 scales\_1.0.0 pkgconfig\_2.0.2