HR Data Report

Jakob Orel

11/24/2020

Analyzing human resources data is extremely important for the upper level managers and C-suite executives to understand the influences on their employees. It is important to analyze what makes employees more effective and productive. This can lead to making decisions on if an organization should hire or fire employees, conduct employee training, or promote individuals. It could also lead to better strategies for recruiting new hires or retaining current employees. All of this information is greatly important to managers of a company.

Table of Contents

Introduction to the Data

```
str(data)
```

```
## 'data.frame':    311 obs. of  38 variables:
##  $ Employee_Name             : chr  "Adinolfi, Wilson  K" "Ait Sidi, Karthikeyan   " "Akinkuolie, Sa
##  $ EmpID                     : int  10026 10084 10196 10088 10069 10002 10194 10062 10114 10250 ...
##  $ MarriedID                 : int  0 1 1 1 0 0 0 0 0 0 ...
##  $ MaritalStatusID           : int  0 1 1 1 2 0 0 4 0 2 ...
##  $ GenderID                  : int  1 1 0 0 0 0 0 1 0 1 ...
##  $ EmpStatusID               : int  1 5 5 1 5 1 1 1 3 1 ...
##  $ DeptID                    : int  5 3 5 5 5 5 4 5 5 3 ...
##  $ PerfScoreID               : int  4 3 3 3 3 4 3 3 3 3 ...
##  $ FromDiversityJobFairID    : int  0 0 0 0 0 0 0 0 1 0 ...
##  $ Salary                    : int  62506 104437 64955 64991 50825 57568 95660 59365 47837 50178 ...
##  $ Termd                     : int  0 1 1 0 1 0 0 0 0 0 ...
##  $ PositionID                : int  19 27 20 19 19 19 24 19 19 14 ...
##  $ Position                  : chr  "Production Technician I" "Sr. DBA" "Production Technician II" "
##  $ State                     : chr  "MA" "MA" "MA" "MA" ...
##  $ Zip                       : int  1960 2148 1810 1886 2169 1844 2110 2199 1902 1886 ...
##  $ DOB                       : chr  "07/10/83" "05/05/75" "09/19/88" "09/27/88" ...
##  $ Age                       : int  37 45 32 32 31 43 41 37 50 32 ...
```

```
##  $ Sex                        : chr  "M" "M" "F" "F" ...
##  $ MaritalDesc                : chr  "Single" "Married" "Married" "Married" ...
##  $ CitizenDesc                : chr  "US Citizen" "US Citizen" "US Citizen" "US Citizen" ...
##  $ HispanicLatino             : chr  "No" "No" "No" "No" ...
##  $ RaceDesc                   : chr  "White" "White" "White" "White" ...
##  $ DateofHire                 : chr  "07/05/11" "03/30/15" "07/05/11" "01/07/08" ...
##  $ DateofTermination          : chr  "" "06/16/16" "09/24/12" "" ...
##  $ TermReason                 : chr  "N/A-StillEmployed" "career change" "hours" "N/A-StillEmployed"
##  $ NumYearsWorked             : int  9 1 1 12 5 8 6 7 11 5 ...
##  $ EmploymentStatus           : chr  "Active" "Voluntarily Terminated" "Voluntarily Terminated" "Acti
##  $ Department                 : chr  "Production       " "IT/IS" "Production       " "Production
##  $ ManagerName                : chr  "Michael Albert" "Simon Roup" "Kissy Sullivan" "Elijiah Gray" ..
##  $ ManagerID                  : int  22 4 20 16 39 11 10 19 12 7 ...
##  $ RecruitmentSource          : chr  "LinkedIn" "Indeed" "LinkedIn" "Indeed" ...
##  $ PerformanceScore           : chr  "Exceeds" "Fully Meets" "Fully Meets" "Fully Meets" ...
##  $ EngagementSurvey           : num  4.6 4.96 3.02 4.84 5 5 3.04 5 4.46 5 ...
##  $ EmpSatisfaction            : int  5 3 3 5 4 5 3 4 3 5 ...
##  $ SpecialProjectsCount       : int  0 6 0 0 0 0 4 0 0 6 ...
##  $ LastPerformanceReview_Date: chr  "01/17/19" "02/24/16" "05/15/12" "01/03/19" ...
##  $ DaysLateLast30             : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Absences                   : int  1 17 3 15 2 15 19 19 4 16 ...
```

This dataset contains employee information for 311 employees of a synthetically created organization for the use in courses at the New England College of Business. Dr. Carla Patalona used this dataset to teach Tableau visualization. In Excel, I added several columns including the Age of the employee (using DOB) and NumYearsWorked (using DateofHire and DateofTerm). I also had to do some data cleaning by formatting the date columns to be easily readable in R. There were also several erroneous values with spaces and capitalization issues that needed correction.

Questions of an HR Manager

There are hundreds of important questions that an HR manager may have about their employees. These examples may be useful questions to analyze the data further.

1. Do certain managers affect the performance score of employees?

- Some managers may score their employees or treat their employees worse. This may affect productivity and should be noted for professional development.

2. Is the organization diverse in sex and race?

- Diversity is extremely important in a company to ensure there is inclusion of all employees and to expand the perspectives and ideas of team members. The percentage of employees should align relatively closely to the percentage of the population in race and gender.

3. How are we recruiting for minority groups in gender or race? How can we hire more effectively to become more diverse?

- Recruiting can be an effective way to improve the diversity of the organization. I would predict that there is little evidence of a difference in recruiting for gender, but race may be different. For example, employees who are white may be more likely to be hired from an online application where as minority groups may be hired from in-person interactions or referrals.

4. Is there bias for the salaries of employees based on gender or race?

- This is extremely important. I would assume that there is no significant difference in the average salaries of these groups, but it is important to make sure. This may be affected by other factors including experience and position.

5. How do we keep employees longer? What variables lead to a higher likelihood of employees quitting?

- This may be very difficult to predict but I assume that Age, NumYearsWorked, EmpSatisfaction, and the EngagementSurvey variables will be the most important features in determining if an employee is terminated.

Exploring Manager's Effect on Performance

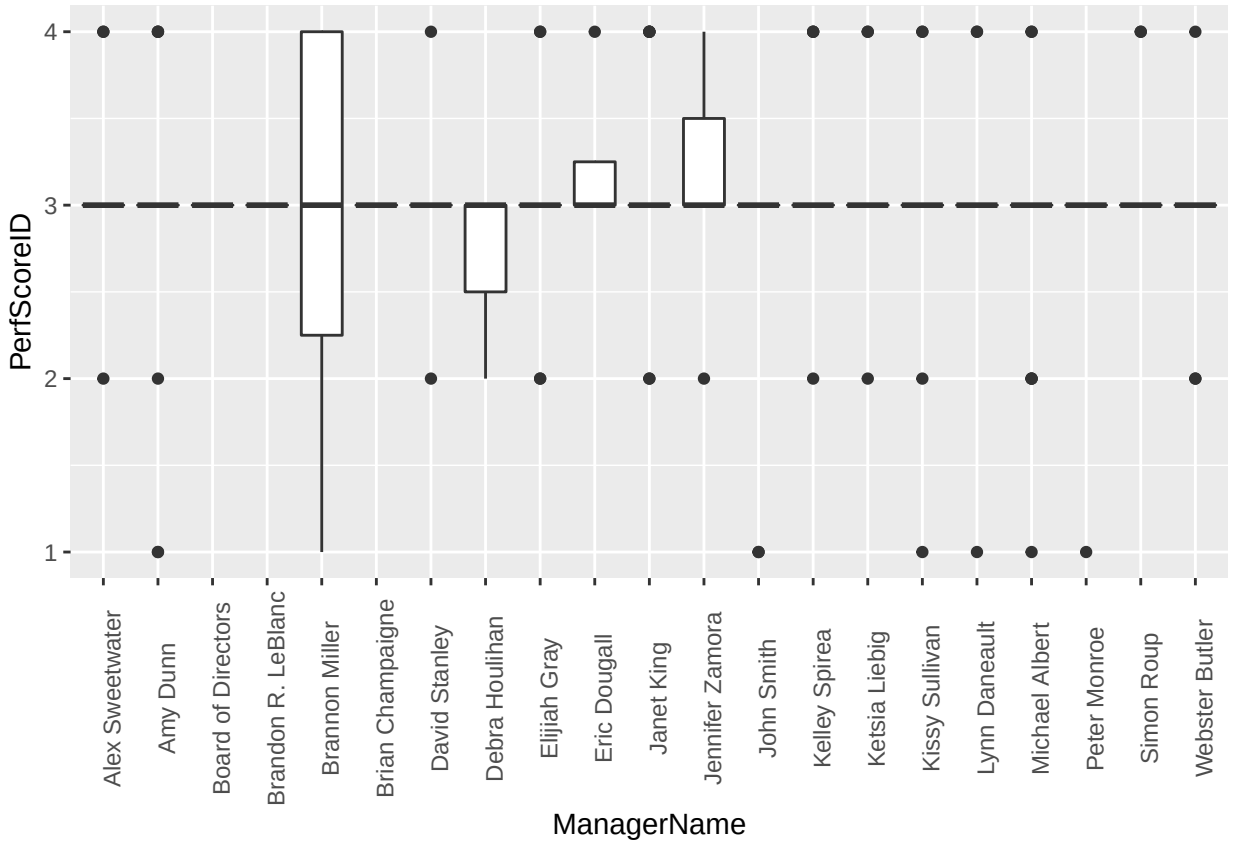
```
anyNA(data$ManagerID)
```

```
## [1] TRUE
```

```
# get rid of employees where Manager is NA
managers <- data[which(!is.na(data$ManagerID)),]

table(managers$ManagerID)
```

```
##
##  1  2  3  4  5  6  7  9 10 11 12 13 14 15 16 17 18 19 20 21 22 30 39
##  6 19  1 17  7  4 14  2  9 21 22  8 21  3 22 14 22 21 22 13 21  1 13
```

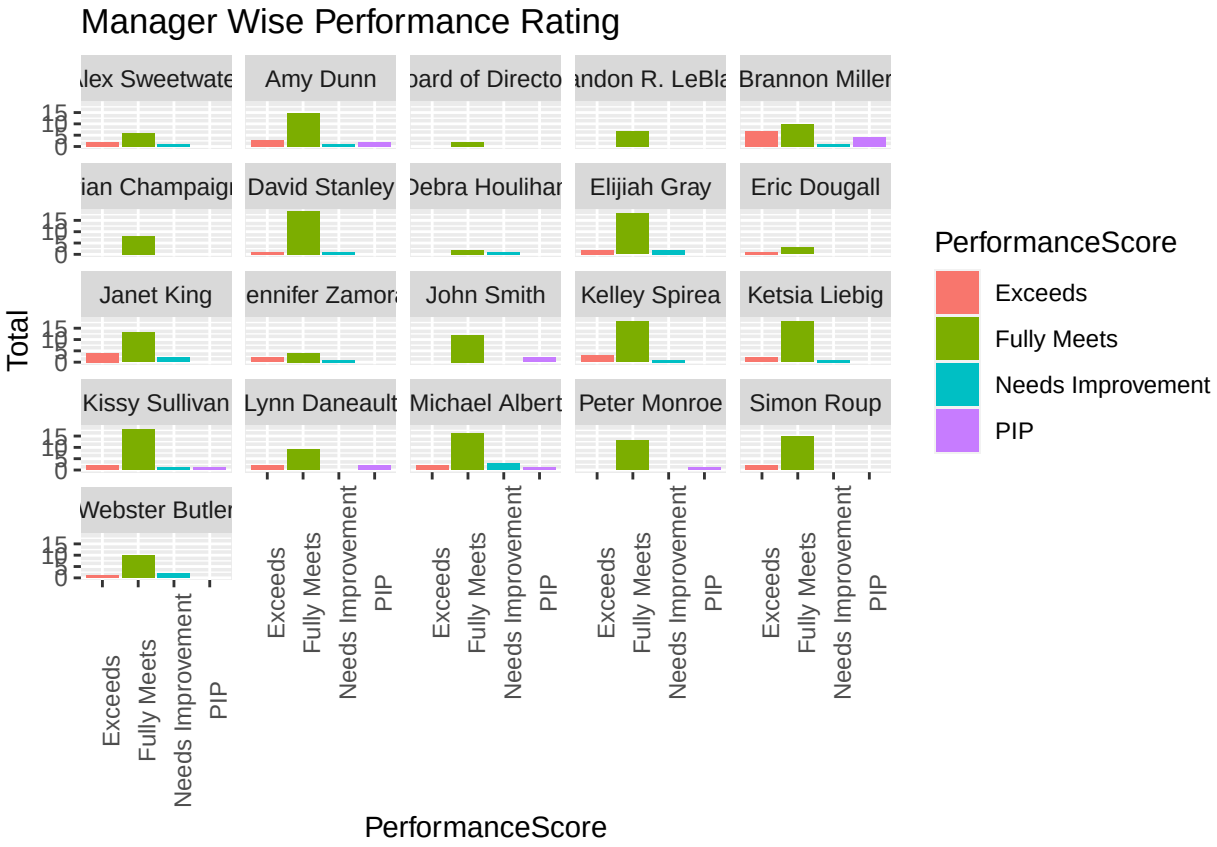
```
# Is there a correlation between performance scores and Manager?
ggplot(managers, aes(x=ManagerName, y=PerfScoreID)) +
        geom_boxplot() +
  theme(axis.text.x = element_text(angle=90))
```




```
# There are too many managers to look into this and see valuable reasons for better performance.

manager_scores <-managers%>%group_by(ManagerName,PerformanceScore)%>%summarise(Total=n())
```

```
## `summarise()` regrouping output by 'ManagerName' (override with `.groups` argument)
```

```
ggplot(manager_scores,aes(x=PerformanceScore,y=Total,fill=PerformanceScore)) +
 geom_bar(stat="identity",position="dodge") +
 facet_wrap(~ManagerName) +
 theme(axis.text.x = element_text(angle=90)) +
 ggtitle("Manager Wise Performance Rating")
```



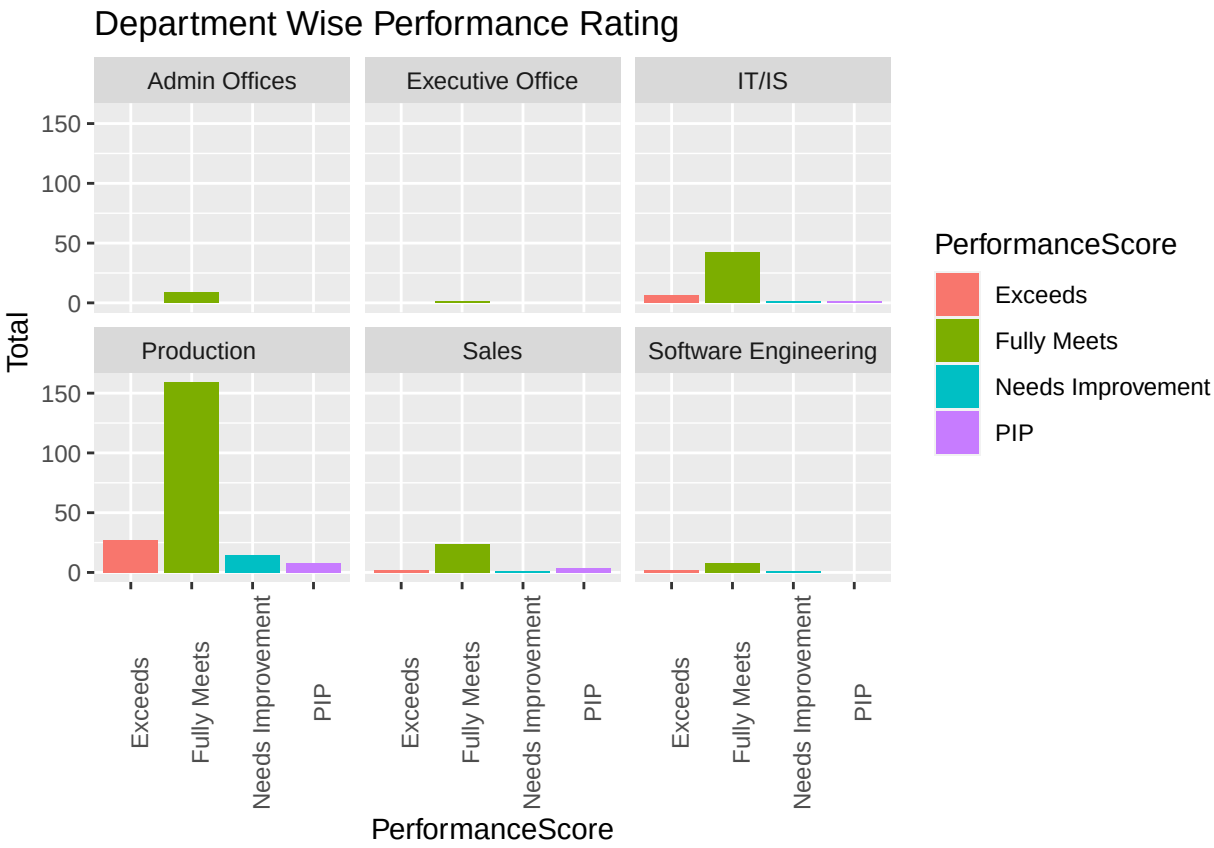
It appears that David Stanley, Janet King, and Simon Roup have some of the highest performance Brannon Miller appears to give the most PIPs and has a higher spread in Performance Scores. This is also seen in the box plots. This may be useful to understand how employees are performing under certain pressures. Maybe a professional development meeting with Brannon Miller may be useful to understand his reports and their performance. Maybe this also has to do with the department that Brannon oversees. It could be that that department may not perform as well.

Exploring Department's Performance

```
department_scores <-data%>%group_by(Department,PerformanceScore)%>%summarise(Total=n())
```

```
## `summarise()` regrouping output by 'Department' (override with `.groups` argument)
```

```
ggplot(department_scores,aes(x=PerformanceScore,y=Total,fill=PerformanceScore)) +
 geom_bar(stat="identity",position="dodge") +
 facet_wrap(~Department) +
 theme(axis.text.x = element_text(angle=90)) +
 ggtitle("Department Wise Performance Rating")
```



The production department clearly has the most employees with employees that exceed performance and also need improvement. There does not appear to be a large trend but Sales has some poor performance.

Exploring Racial Diversity in the Organization

```
# Lets look into the racial diversity of the organization.
employed <- data[which(!data$Termd), ]
anyNA(data$RaceDesc)
```

```
## [1] FALSE
```

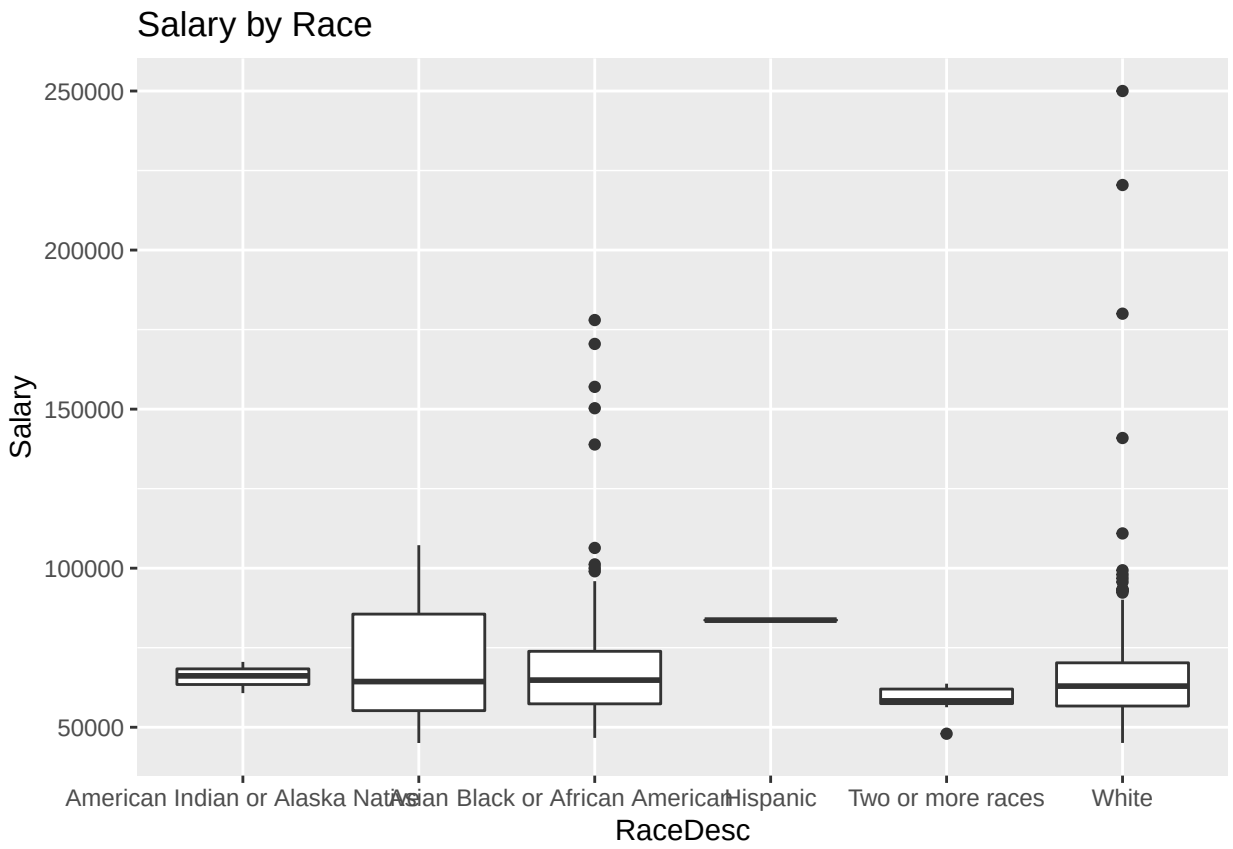
```
PercTable(employed$RaceDesc)
```

```
##
##                                  freq   perc
##
## American Indian or Alaska Native    3   1.4%
## Asian                              20   9.7%
## Black or African American          51  24.6%
## Hispanic                            1   0.5%
## Two or more races                   8   3.9%
## White                             124  59.9%
```

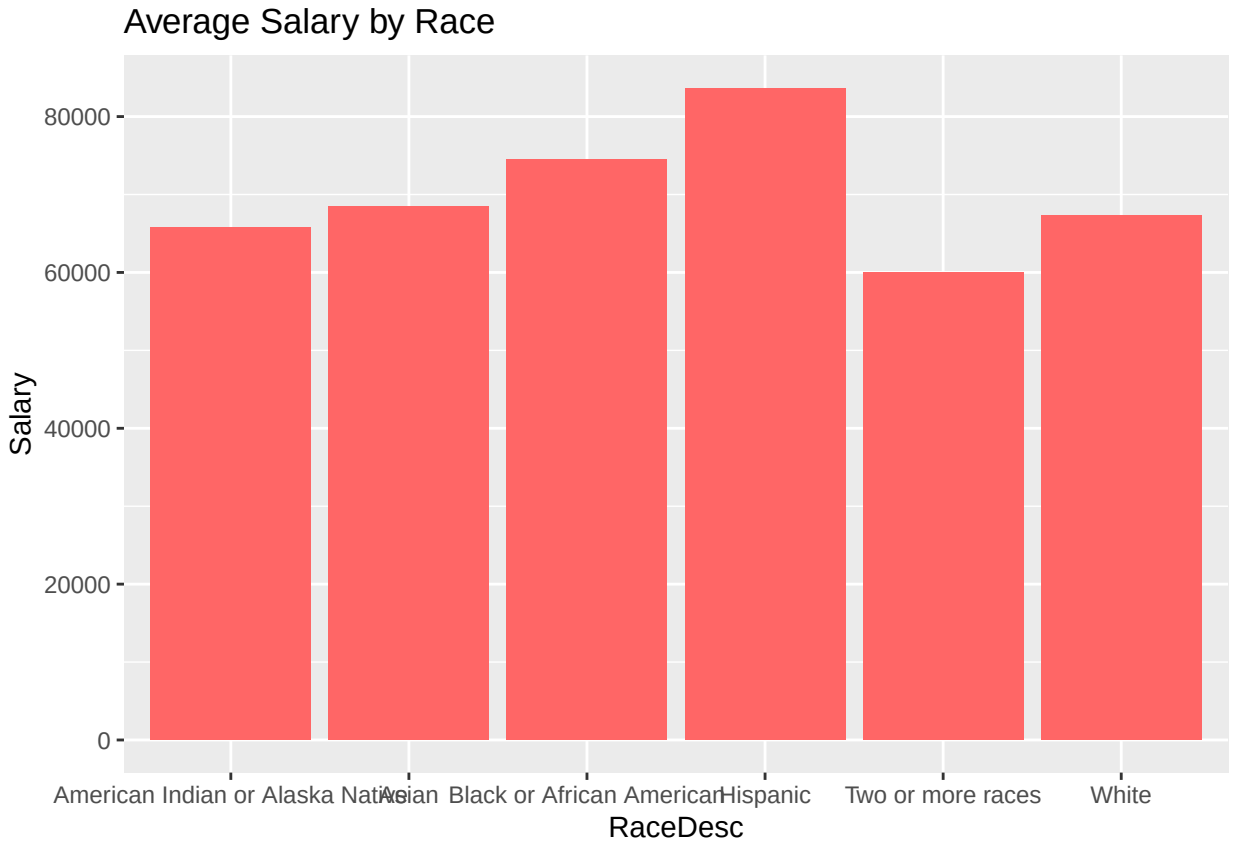
```
# This organization appears to be relatively diverse currently and is relatively proportionate to the U
# Is there a difference in the race of people who were terminated?

terminated <- data[which(data$Termd==1), ]

PercTable(terminated$RaceDesc)
```

```
##
##                              freq   perc
##
## Asian                           9   8.7%
## Black or African American      29  27.9%
## Two or more races               3   2.9%
## White                          63  60.6%
```

```
# Takeaways:
# 1. It appears the company is currently racially diverse
# 2. It appears there are not any large disparities in the races of those terminated.

# Now that I believe the proportions are ok. Is pay rate affected by race?

ggplot(employed, aes(x=RaceDesc, y=Salary)) +
  geom_boxplot() +
  labs(title = "Salary by Race")
```




```
# There does not appear to be much bias in salary based on race. There is only 1 Hispanic employee. The
ggplot(data, aes(x=RaceDesc, y=Salary)) +
  geom_bar(stat="summary", fun=mean, fill="#FF6666") +
  ggtitle("Average Salary by Race")
```

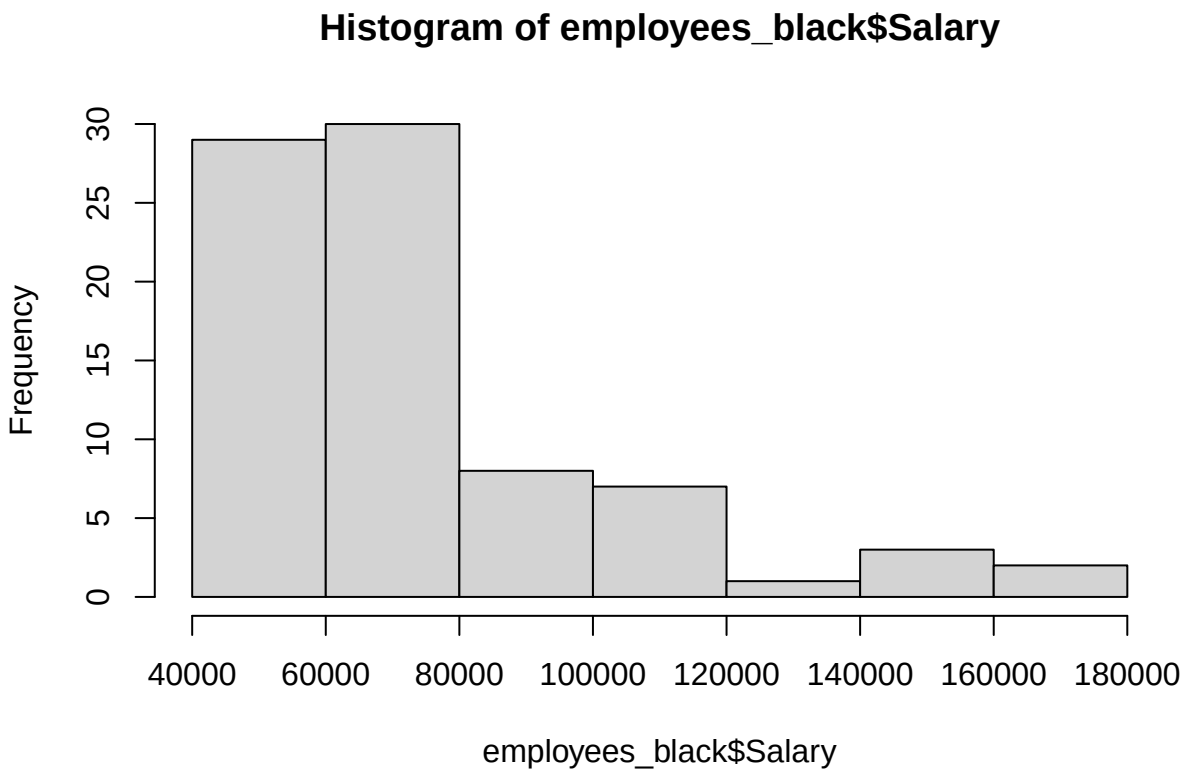



```
# The Hispanic average salary appears high because there is only one employee.
```

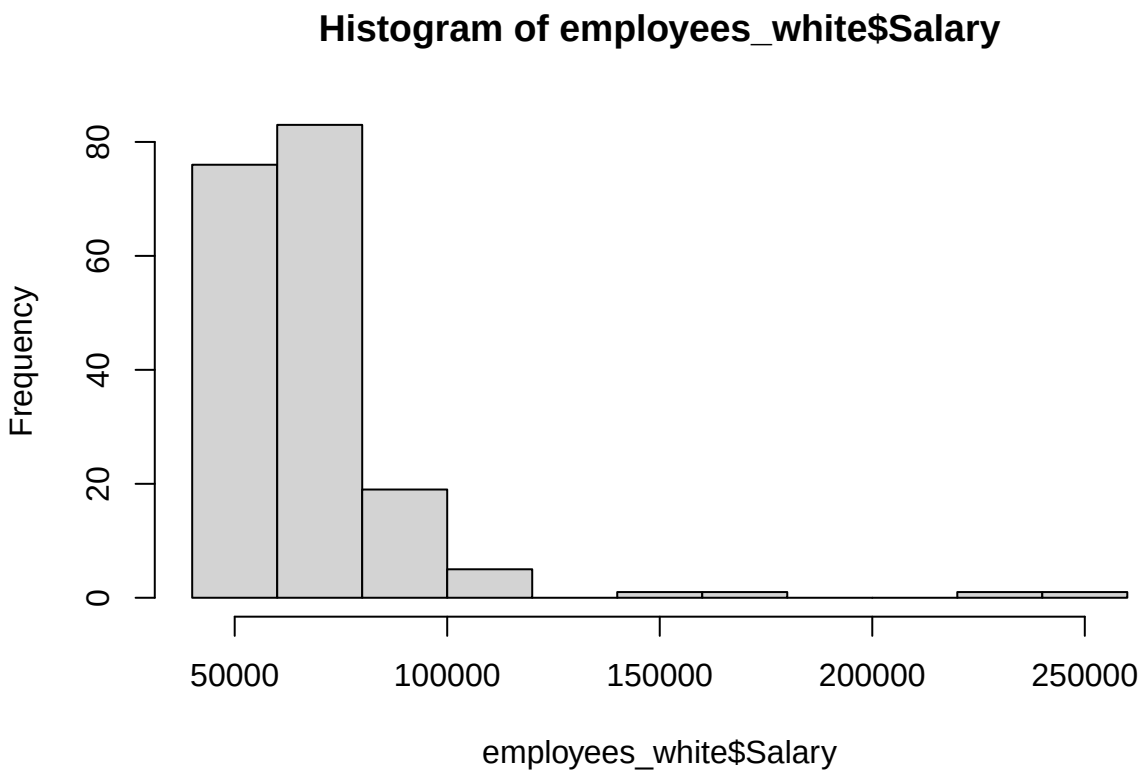
The percentages of employees by race are relative to the percentages of the population provided by the US Census Bureau for 2018 (White (76.3%), Black (13.4%), Asian (5.8%), American Native (1.3%), Two or more (2.8%), Hispanic (18.5%)). It appears that the average salary for Black/African American employees (M = 74431.02) is higher than White employees (M = 67287.55).

Test Average Salaries by Race

```
# Is the average salary for Black/African American employees significantly different from the average s
employees_black <- data[which(data$RaceDesc=="Black or African American"),]
employees_white <- data[which(data$RaceDesc=="White"),]
hist(employees_black$Salary)
```




```
hist(employees_white$Salary)
```




```
shapiro.test(employees_black$Salary)
```

```
##
```

```
##  Shapiro-Wilk normality test
##
## data:  employees_black$Salary
## W = 0.77076, p-value = 7.782e-10
```

```
shapiro.test(employees_white$Salary)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  employees_white$Salary
## W = 0.61039, p-value < 2.2e-16
```

```
# Salary is not normally distributed. Tiny p-value in Shapiro test and skewed histograms.
# Samples are not paired and not normal so we will use the Mann-Whitney test.
wilcox.test(employees_black$Salary, employees_white$Salary)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  employees_black$Salary and employees_white$Salary
## W = 8496.5, p-value = 0.07879
## alternative hypothesis: true location shift is not equal to 0
```

```
# p-value is high for Mann-Whitney test meaning that the salary is approximately the same. They come fr
# The test does not indicate any significant difference.
# Lets run a t.test just or fun even though it is not normally distributed
t.test(employees_black$Salary, employees_white$Salary)
```

```
##
##  Welch Two Sample t-test
##
## data:  employees_black$Salary and employees_white$Salary
## t = 1.9109, df = 130.3, p-value = 0.05821
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   -252.0894 14539.0484
## sample estimates:
## mean of x mean of y
##  74431.02  67287.55
```

```
# This p-value is also high meaning there is no significant difference in Salary between employees who
# This is good news for the organization!
```

The results of the Mann-Whitney test lead us to believe that there is no statistically significant difference between the average salary of Black and White employees in this organization, but this is relatively misleading. For example, the results of this test would be appropriate for a random sample of employees from all Target stores in Iowa, but it is not appropriate for this dataset because it is only the employees from one certain organization. The results for the Target stores could be used to generalize the population of all Target store employees instead of just one. The results of the test could be replicated and would be exactly the same (given no employees were terminated or hired) in this dataset This is an example of how statistics can be used inappropriately. In this case, the average salary for Black employees is actually truly greater than the average salary of White employees because this is not a random sample from a population. This could be for a number of reasons and should be explored further (perhaps experience and position will affect salary).

Exploring Diversity in Sex

```
# Lets look to see if there is gender diversity in the organization.
PercTable(employed$Sex)
```

```
##
##     freq   perc
##
## F    116  56.0%
## M     91  44.0%
```

```
PercTable(terminated$Sex)
```

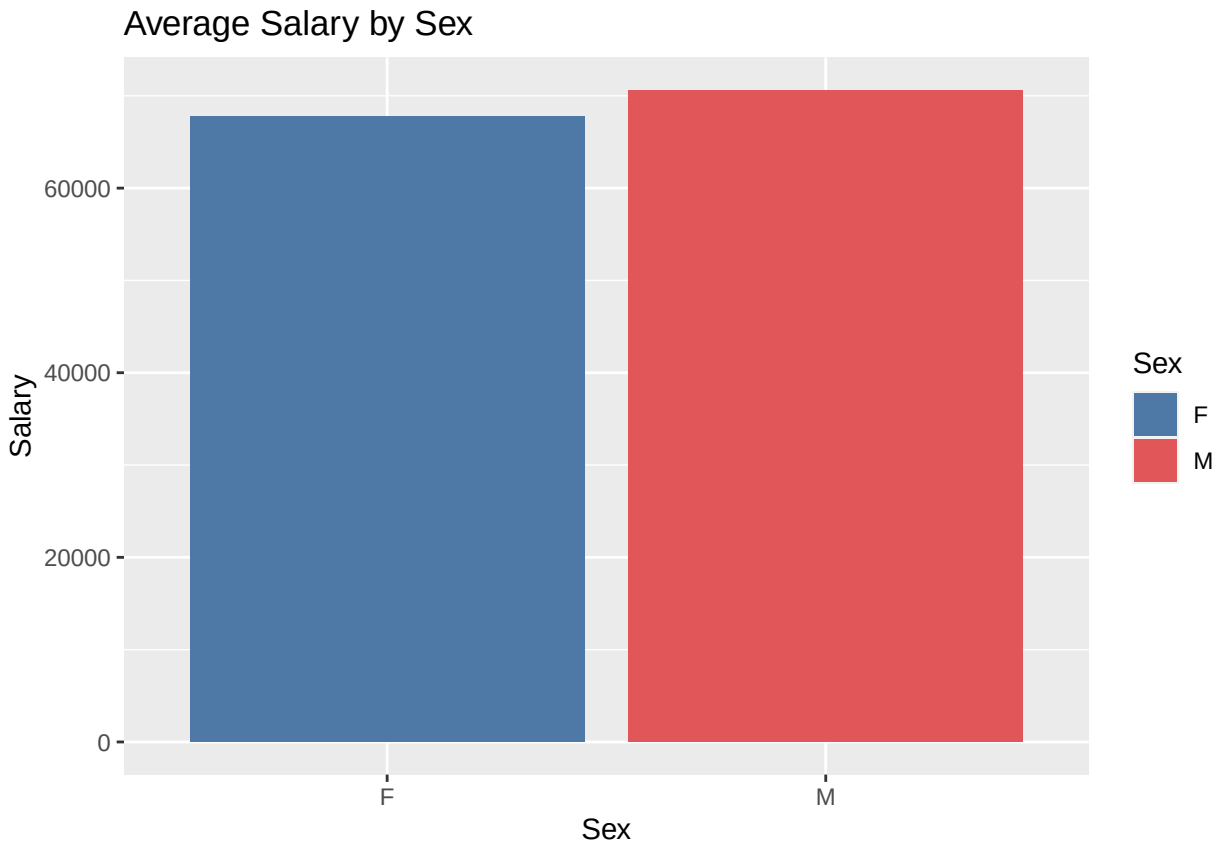
```
##
##     freq   perc
##
## F     60  57.7%
## M     44  42.3%
```

```
# This organization has a higher percentage of females than males.
# The gender of employees terminated is proportionate to those still employed.

employed_gender <- employed %>% group_by(Sex,Position) %>% summarise(Total=n())
```

```
## `summarise()` regrouping output by 'Sex' (override with `.groups` argument)
```

```
# A note, Highcharter is useful in HTML as an interactive widget.
hchart(employed_gender,type="column",hcaes(x=Position,y=Total,group=Sex)) %>%
  hc_add_theme(hc_theme_google())
```

```
# Besides the IT department it looks like there are quite a few more women in this organization
# especially in the Production department.

# Is there bias for sex in salaries?
ggplot(data, aes(x=Sex, y=Salary, fill=Sex)) +
  geom_bar(stat="summary", fun=mean) +
  scale_fill_manual("Sex", values= c("M"="#e15759", "F"="#4e79a7")) +
  ggtitle("Average Salary by Sex")
```




```
# Looks like Males have slightly higher average salary. Lets test this.
```

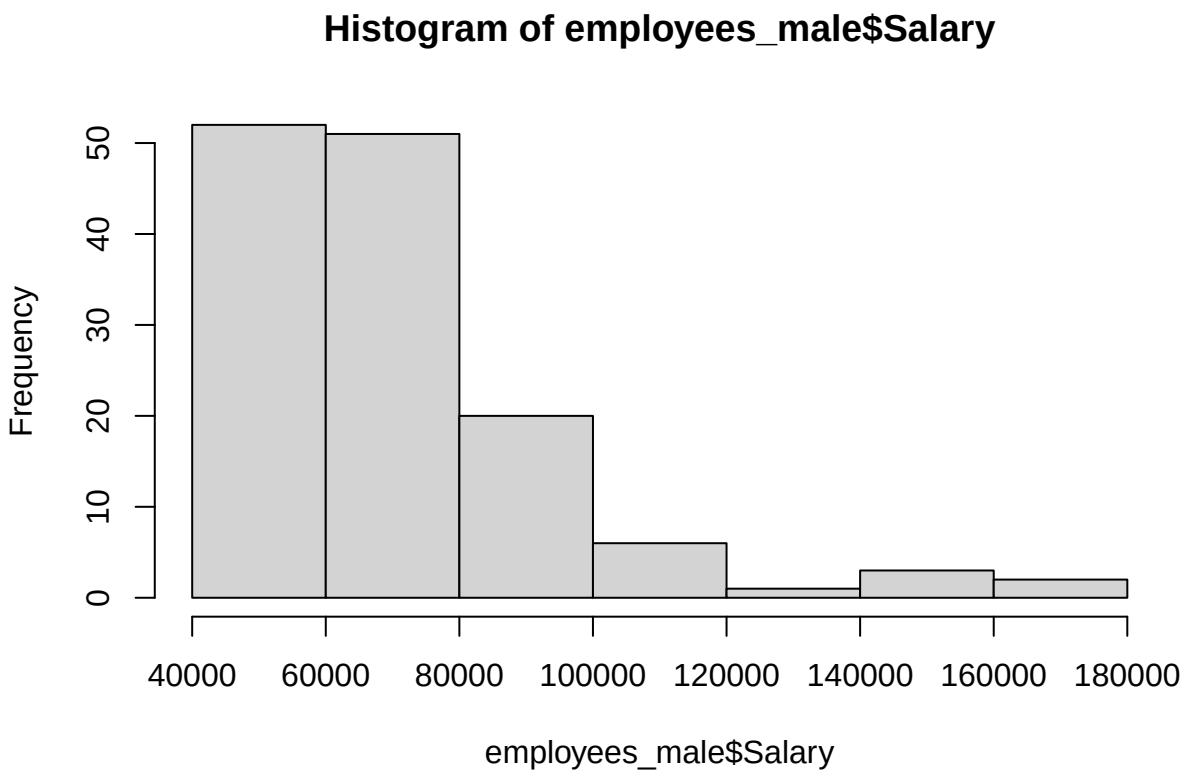
Test Average Salaries by Sex

```
# Lets test to see if the average salary for males is significantly different from females
employees_female <- data[which(data$Sex=="F"),]
employees_male <- data[which(data$Sex=="M"),]
table(data$Sex)
```

```
##
##   F   M
## 176 135
```

```
hist(employees_female$Salary)
```


```
hist(employees_male$Salary)
```




```
shapiro.test(employees_female$Salary)
```

```
##
```

```
##  Shapiro-Wilk normality test
##
## data:  employees_female$Salary
## W = 0.60527, p-value < 2.2e-16
```

```
shapiro.test(employees_male$Salary)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  employees_male$Salary
## W = 0.78057, p-value = 6.17e-13
```

```
# Salary is not normally distributed. Tiny p-value in Shapiro test and skewed histograms.
# Samples are not paired and not normal
wilcox.test(employees_female$Salary, employees_male$Salary)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  employees_female$Salary and employees_male$Salary
## W = 10870, p-value = 0.1988
## alternative hypothesis: true location shift is not equal to 0
```

```
# p-value is high for Mann-Whitney test meaning that the salary is approximately the same. They come fr
# The test does not indicate any significant difference.
# Lets run a t.test just or fun even though it is not normally distributed
t.test(employees_female$Salary, employees_male$Salary)
```

```
##
##  Welch Two Sample t-test
##
## data:  employees_female$Salary and employees_male$Salary
## t = -0.9956, df = 296.39, p-value = 0.3203
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -8461.792  2776.447
## sample estimates:
## mean of x mean of y
##  67786.73  70629.40
```

```
# This p-value is also high meaning there is no significant difference in Salary between employees who
# This is good news for the organization!
```

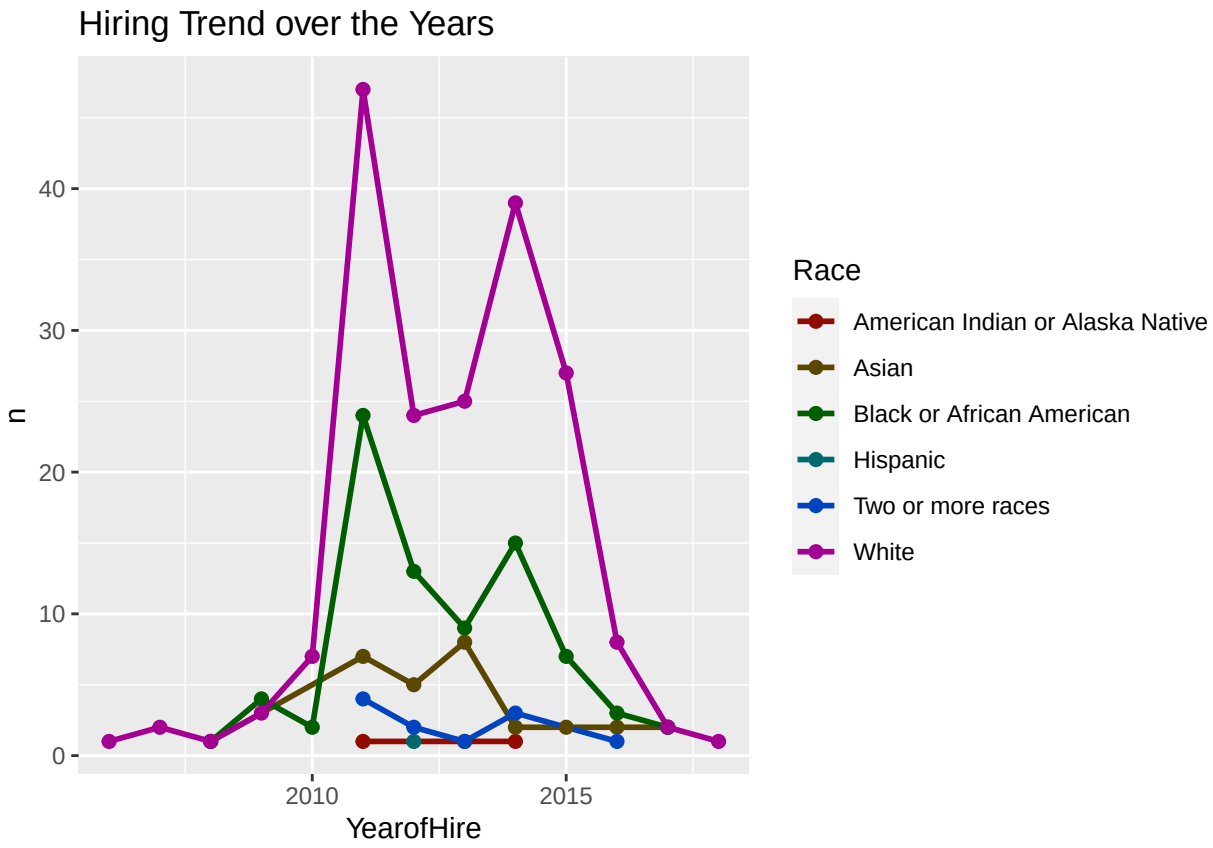
Similar to the statistical test done on the difference in average salary for race, this test is not that relevant. It would only be appropriate if we were looking at a random sample out of a population, but instead we are looking at a defined sample. If we wanted to explore why males had higher average salaries than females it would be worth looking into the positions and experience of each employee. It may also be worth noting that median may be a more accurate statistic because an employee who is paid much higher such as the CEO will cause a bias in the average.

Exploring Recruitment by Race

Let's explore how the organization is hiring new employees. Perhaps we can find advantages to different forms of recruiting.

```
# Format date of hire
data$DateofHire <- as.Date(data$DateofHire, "%m/%d/%y")
```

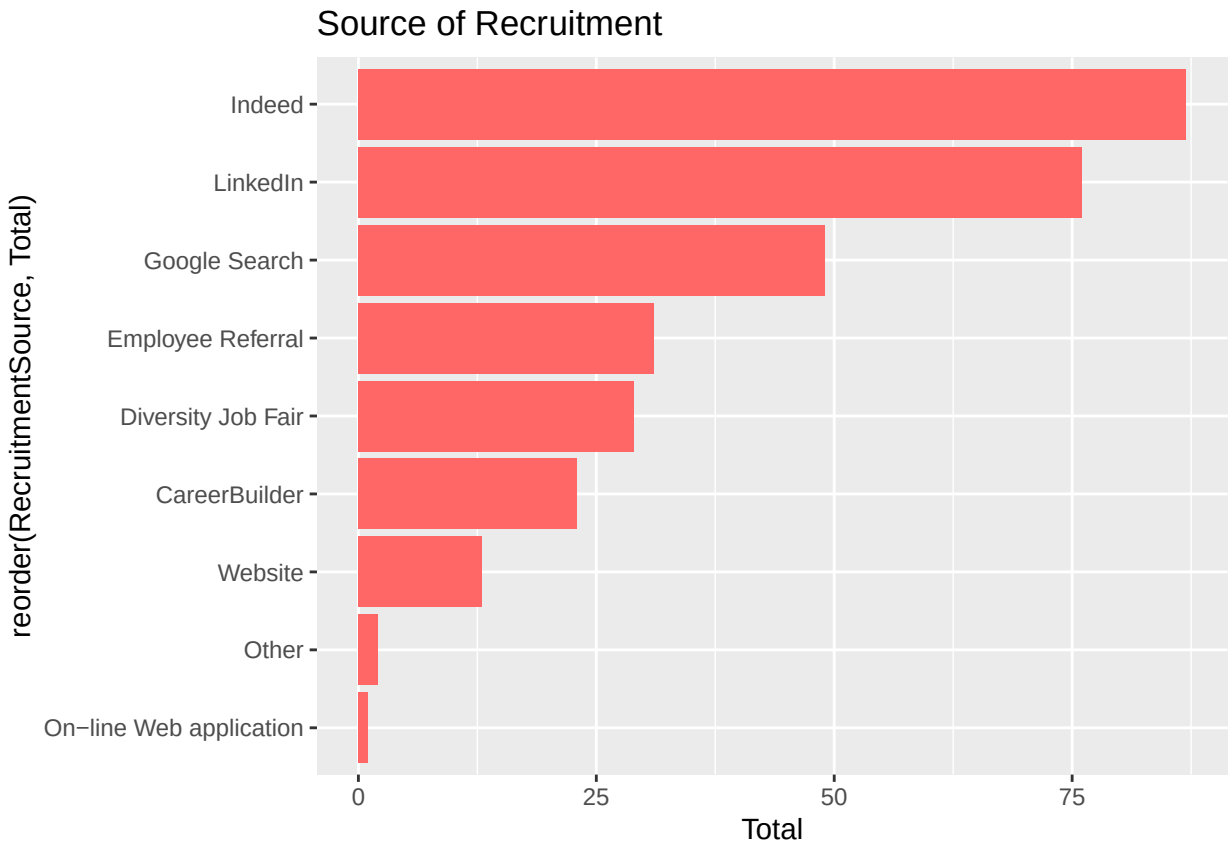
```r
data$YearofHire <- year(data$DateofHire)

# count the number of hires by race over the years
hiring_pattern <- data %>% count(YearofHire,RaceDesc)
# graph this in aline graph
ggplot(hiring_pattern,aes(YearofHire,y=n,group=RaceDesc,color=RaceDesc)) +
  geom_line(size=1) +
  geom_point(size=2)  +
  scale_color_hue(name="Race",l=30) +
  ggtitle("Hiring Trend over the Years")
```




```r
# There was an increase in hiring (especially  people who are Black/African American) but the trend has

# Lets look into how these people were hired
recruitment_source <-data %>% group_by(RecruitmentSource) %>%summarise(Total=n()) %>%arrange(desc(Total
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```r
ggplot(recruitment_source,aes(reorder(RecruitmentSource, Total),y=Total)) +
  geom_bar(stat="identity", fill= "#FF6666") +
  coord_flip() +
  ggtitle("Source of Recruitment")
```




```
# Is there bias in recruitment for race?
data%>%filter(RecruitmentSource=="Employee Referral") %>%count(RaceDesc)
```

```
##                    RaceDesc  n
## 1                     Asian  1
## 2 Black or African American  5
## 3                     White 25
```

```
data%>%filter(RecruitmentSource=="Diversity Job Fair") %>%count(RaceDesc)
```

```
##                    RaceDesc  n
## 1 Black or African American 29
```

```
data%>%filter(RecruitmentSource=="Indeed") %>%count(RaceDesc)
```

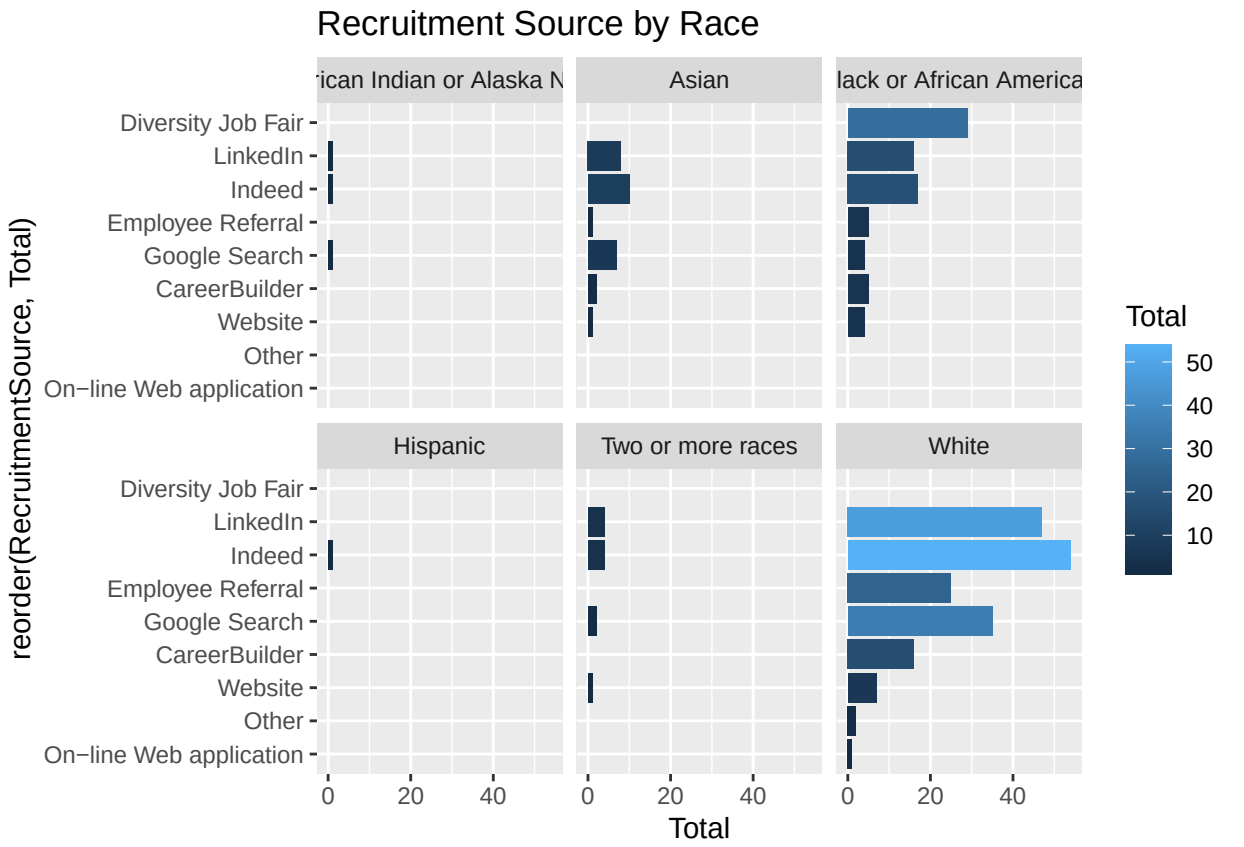
```
##                           RaceDesc  n
## 1 American Indian or Alaska Native  1
## 2                             Asian 10
## 3        Black or African American 17
## 4                          Hispanic  1
## 5                Two or more races  4
## 6                             White 54
```

```
recruitment_race <-data%>%group_by(RecruitmentSource,RaceDesc) %>%summarise(Total=n())
```

```
## `summarise()` regrouping output by 'RecruitmentSource' (override with `.groups` argument)
```

```
ggplot(recruitment_race,aes(reorder(RecruitmentSource,Total),Total,fill=Total)) +
  geom_bar(stat="identity") +
  coord_flip() +
```

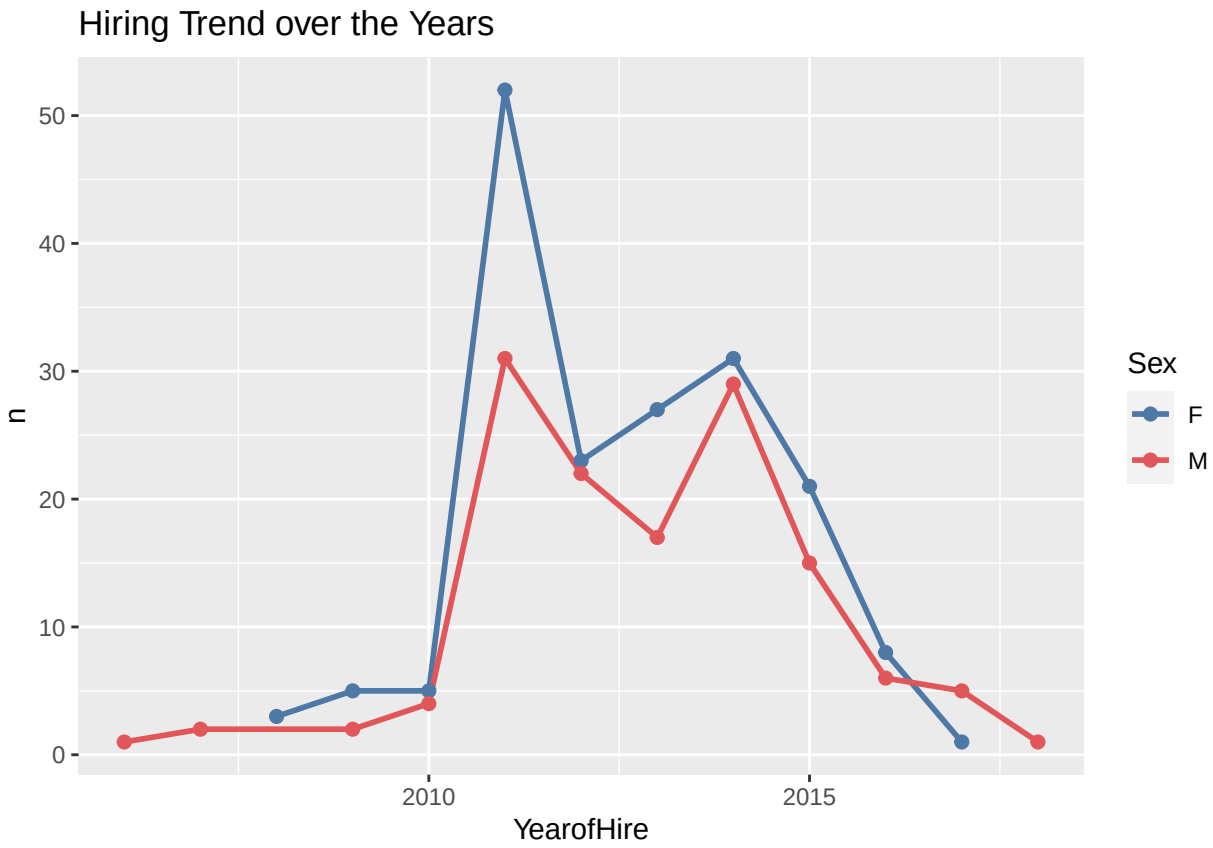
```
  facet_wrap(~RaceDesc) +
  ggtitle("Recruitment Source by Race")
```



It appears most White people are hired from Job Sites while Black employees are primarily hired from Diversity Job Fairs. This is a good indicator that these intentional diversity job fairs are effective and useful for the organization to have diverse ideas.

Exploring Recruitment by Sex

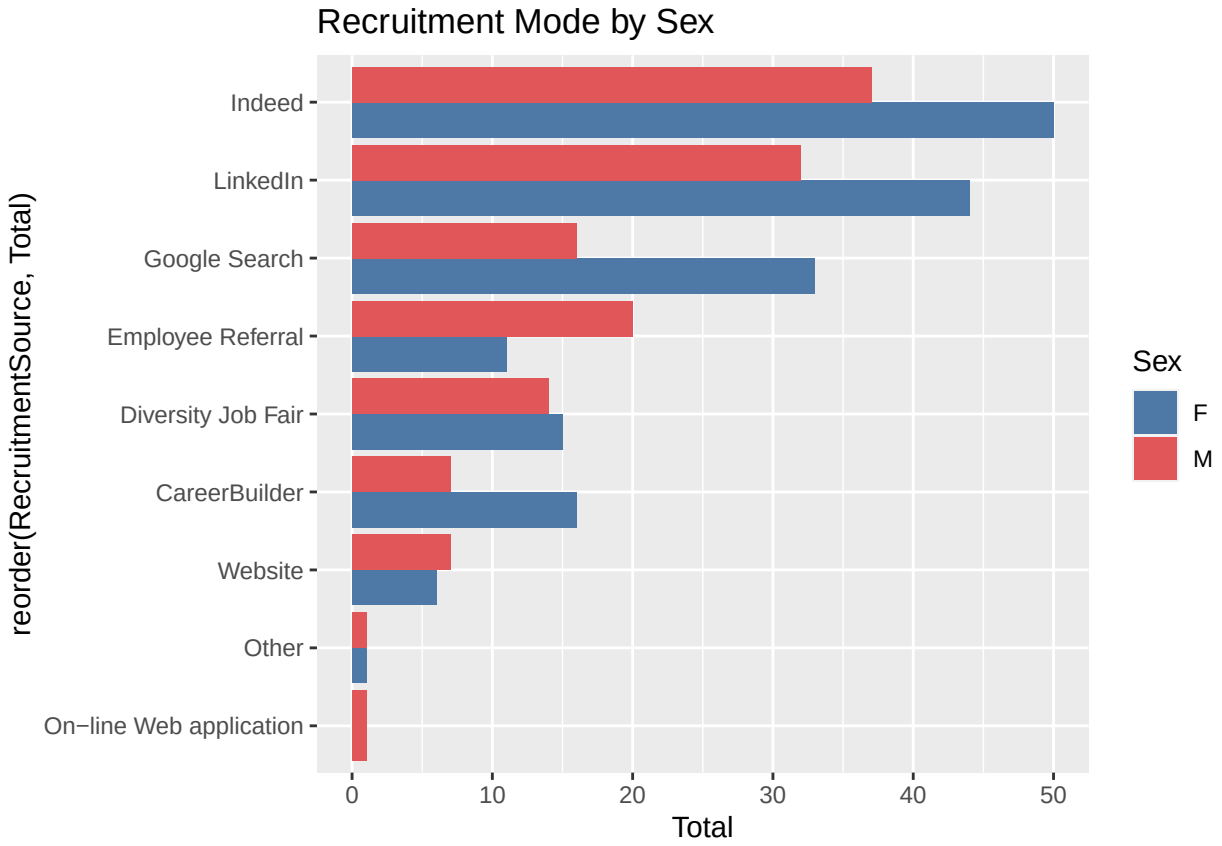
```
# count the number of hires by race over the years
hiring_pattern_gender <- data %>% count(YearofHire,Sex)
# graph this in aline graph
ggplot(hiring_pattern_gender,aes(YearofHire,y=n,group=Sex,color=Sex)) +
  geom_line(size=1) +
  geom_point(size=2)  +
  scale_color_manual(name="Sex",values= c("M"="#e15759", "F"="#4e79a7")) +
  ggtitle("Hiring Trend over the Years")
```




```
# It appears that they have had a trend of hiring more females than males but in recent years it has fl
# How are they being recruited?

recruitment_gender <-data%>%group_by(RecruitmentSource,Sex) %>%summarise(Total=n())
```

```
## `summarise()` regrouping output by 'RecruitmentSource' (override with `.groups` argument)
```

```
ggplot(recruitment_gender,aes(reorder(RecruitmentSource,Total),Total,fill=Sex)) +
  geom_bar(stat="identity", position="dodge") +
  scale_fill_manual("Sex", values= c("M"="#e15759", "F"="#4e79a7")) +
  coord_flip() +
  ggtitle("Recruitment Mode by Sex")
```



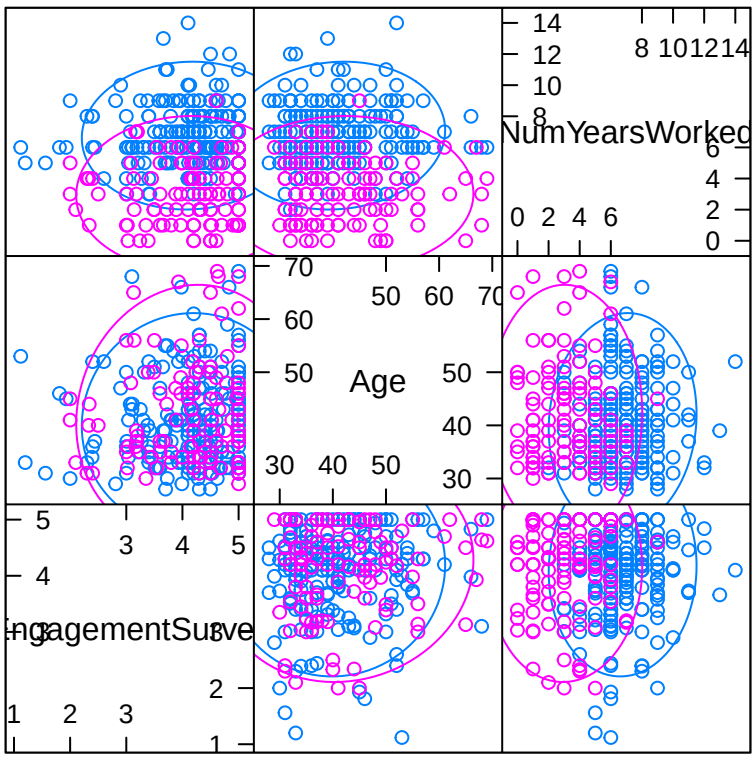
It appears females are hired more from Job Sites and Google searches and males are hired more by employee referral. It may be a good idea for the organization to see how many applications they receive from males on job sites if they would like to hire more men. These averages may be proven with statistical tests, but they would not be that useful to the organization.

Exploring the Factors of Terminated Employees

Now it is time to explore the big question: What features lead to a higher likelihood of someone quitting? Some things that I think are logically related:

1. EmpSatisfaction
2. Age
3. EngagementSurvey
4. Absences
5. NumYearsWorked

```
features <- data[,c(33, 17, 26)]

# Termd includes those who were fired, but it is an easy variable and there are not many who were fired
featurePlot(x=features, y=as.factor(data$Termd), plot='ellipse')
```



Scatter Plot Matrix

```
# This feature plot shows the NumYearsWorked is a relatively strong factor because ellipses have clearly

# Maybe we should look at some other features?
#features$MaritalStatusID <- data[,4]
#features$Salary <- data[,10]
#features$PerfScoreID <- data[,8]
# Turns out these have little effect at all.

# Lets just give some regression models a try.

summary(lm(Termd ~ Age + EmpSatisfaction, data))
```

```
##
## Call:
## lm(formula = Termd ~ Age + EmpSatisfaction, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.4843 -0.3371 -0.2948  0.6294  0.7333
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     0.106541   0.177388   0.601   0.5485
## Age             0.005415   0.003027   1.789   0.0746 .
## EmpSatisfaction 0.001038   0.029525   0.035   0.9720
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4716 on 308 degrees of freedom
```

```
## Multiple R-squared:  0.01031,    Adjusted R-squared:  0.00388
## F-statistic: 1.604 on 2 and 308 DF,  p-value: 0.2028
```

```
summary(lm(Termd ~ EngagementSurvey + Absences, data))
```

```
##
## Call:
## lm(formula = Termd ~ EngagementSurvey + Absences, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.4288 -0.3553 -0.2887  0.6207  0.7477
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       0.294306   0.149864   1.964   0.0504 .
## EngagementSurvey -0.009979   0.033917  -0.294   0.7688
## Absences          0.007923   0.004578   1.731   0.0845 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4717 on 308 degrees of freedom
## Multiple R-squared:  0.009936,   Adjusted R-squared:  0.003507
## F-statistic: 1.546 on 2 and 308 DF,  p-value: 0.2149
```

```
summary(lm(Termd ~ Salary + PerfScoreID, data))
```

```
##
## Call:
## lm(formula = Termd ~ Salary + PerfScoreID, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.4836 -0.3452 -0.3011  0.6389  0.7929
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.302e-01  1.491e-01   4.227 3.13e-05 ***
## Salary      -1.574e-06  1.072e-06  -1.468    0.143
## PerfScoreID -6.286e-02  4.592e-02  -1.369    0.172
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4705 on 308 degrees of freedom
## Multiple R-squared:  0.01483,    Adjusted R-squared:  0.008432
## F-statistic: 2.318 on 2 and 308 DF,  p-value: 0.1002
```

```
summary(lm(Termd ~ NumYearsWorked + EmpSatisfaction + Age, data))
```

```
##
## Call:
## lm(formula = Termd ~ NumYearsWorked + EmpSatisfaction + Age,
##     data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -0.7155 -0.2706 -0.0364  0.2249  1.0754
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      0.864771   0.141598   6.107 3.06e-09 ***
## NumYearsWorked  -0.119047   0.007661 -15.540  < 2e-16 ***
## EmpSatisfaction -0.010103   0.022136  -0.456   0.6484
## Age              0.004039   0.002270   1.779   0.0762 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3534 on 307 degrees of freedom
## Multiple R-squared:  0.4461, Adjusted R-squared:  0.4407
## F-statistic: 82.41 on 3 and 307 DF,  p-value: < 2.2e-16
```

```
# The only strong feature in these regressions is NumYearsWorked with a p-value < 2e-16. For every year

# These are weak regression models.
```

This is a very tough question. It may require some more advanced techniques like using decision trees to determine the most important features. There are many reasons why someone may quit and it may just be that we do not have variables that have strong correlations for determining if someone will quit. For example, they may quit because of the Salary, Manager, diversity of the organization, school, retirement or simply because they just want a new job. More advanced machine learning models would be able to accomadate for the many different variables that may affect if an employee is terminated.