HR Data EDA Project

Larissa Woods

2024-06-11

HR Data Exploratory Analysis

This is an exploratory analysis of Kaggle's Human Resources Data Set that can be found at this link: https://www.kaggle.com/datasets/rhuebner/human-resources-data-set?resource=download

```
install.packages("tidyverse")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.4'
## (as 'lib' is unspecified)
```

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become error
```

```
library(lubridate)
library(datasets)
library(tidyr)
library(ggplot2)
hrdata <- read_csv("HRDataset_v14.csv")
```

```
## Rows: 311 Columns: 36
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (18): Employee_Name, Position, State, DOB, Sex, MaritalDesc, CitizenDesc...
## dbl (18): EmpID, MarriedID, MaritalStatusID, GenderID, EmpStatusID, DeptID, ...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Data Cleaning

Create a label for the data in case any filtering is done throughout the analysis. Check for missing values and duplicates.

```
#Data Cleaning
sum(duplicated(hrdata))  # Check for duplicates
```

```
## [1] 0
```

Add Age Column

Adding an age column that calculates the age of each employee based off of their date of birth. This will help with determining if age is a significant factor in my analyses.

```
# Convert DOB column to Date format
hrdata$DOB <- as.Date(hrdata$DOB, format = "%m/%d/%Y")

# Check for NA values after conversion
sum(is.na(hrdata$DOB))
```

```
## [1] 0
```

```
# Calculate age in years
hrdata$Age <- floor(as.numeric(difftime(Sys.Date(), hrdata$DOB, units = "days")) / 365.25)

# Check for NA values after calculating age
sum(is.na(hrdata$Age))
```

```
## [1] 0
```

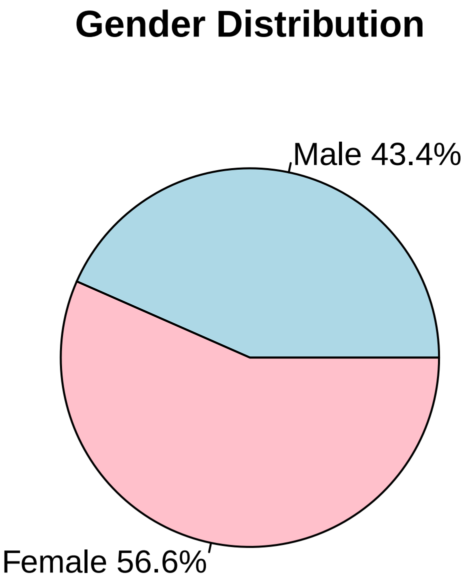
```
all_employees <- hrdata
active_employees <- hrdata[hrdata$EmploymentStatus == "Active", ]
termed_employees <- hrdata[hrdata$EmploymentStatus != "Active", ]
```

DEI Data

Gender

First, let's take a look at the gender distribution across all employees.

```
# Gender Distribution Pie Chart

# Count the number of males and females
male_count <- sum(all_employees$GenderID == 1)
female_count <- sum(all_employees$GenderID == 0)

# Calculate percentages
total_count <- male_count + female_count
male_percentage <- (male_count / total_count) * 100
female_percentage <- (female_count / total_count) * 100

# Create a vector of counts and percentages
gender_counts <- c(Male = male_count, Female = female_count)
gender_percentages <- c(Male = sprintf("%.1f%%", male_percentage), Female = sprintf("%.1f%%", female_pe

# Create colors for the pie chart
colors <- c("lightblue", "pink")

# Create pie chart
pie(gender_counts, labels = paste(names(gender_counts), gender_percentages), col = colors, main = "Gend
```

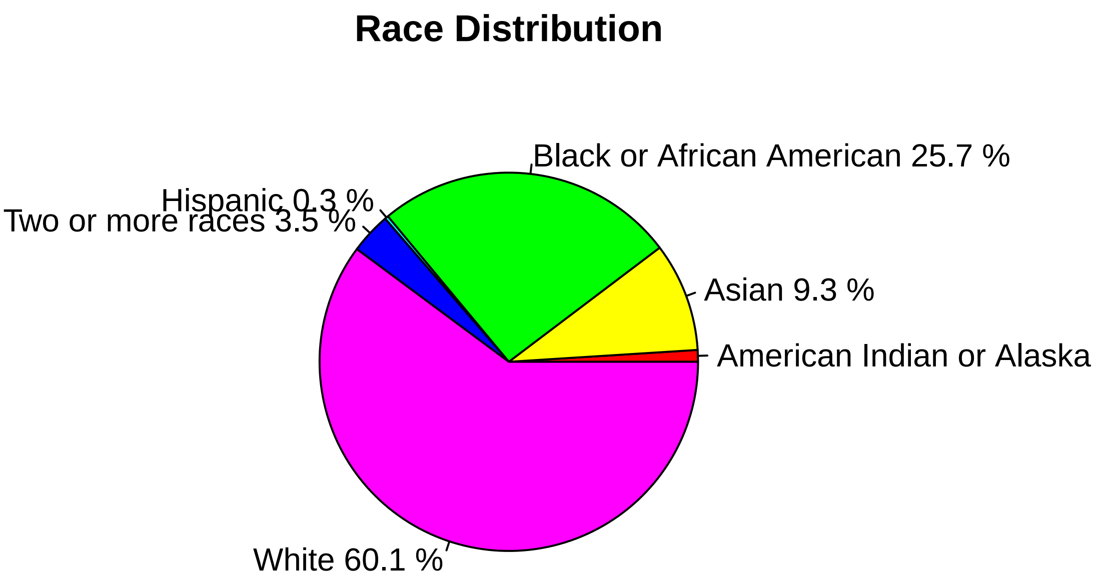


According to Qualtrics.com, in 2022 the percentage of women in the workforce was 39.49 percent worldwide and 46.38 percent in the United States. This data shows that women make up 56.6 percent of the company's workforce, which is above both stats.

Race

Now, let's take a look at the Race Distribution across all employees.

```
# Count the number of employees for each race
race_counts <- table(all_employees$RaceDesc)

# Calculate percentages
total_count <- sum(race_counts)
race_percentages <- prop.table(race_counts) * 100

# Round percentages to one decimal place
race_percentages <- round(race_percentages, 1)

# Create colors for the pie chart
colors <- rainbow(length(race_counts))

# Create pie chart
pie(race_percentages, labels = paste(names(race_percentages), race_percentages, "%"), col = colors, mai
```

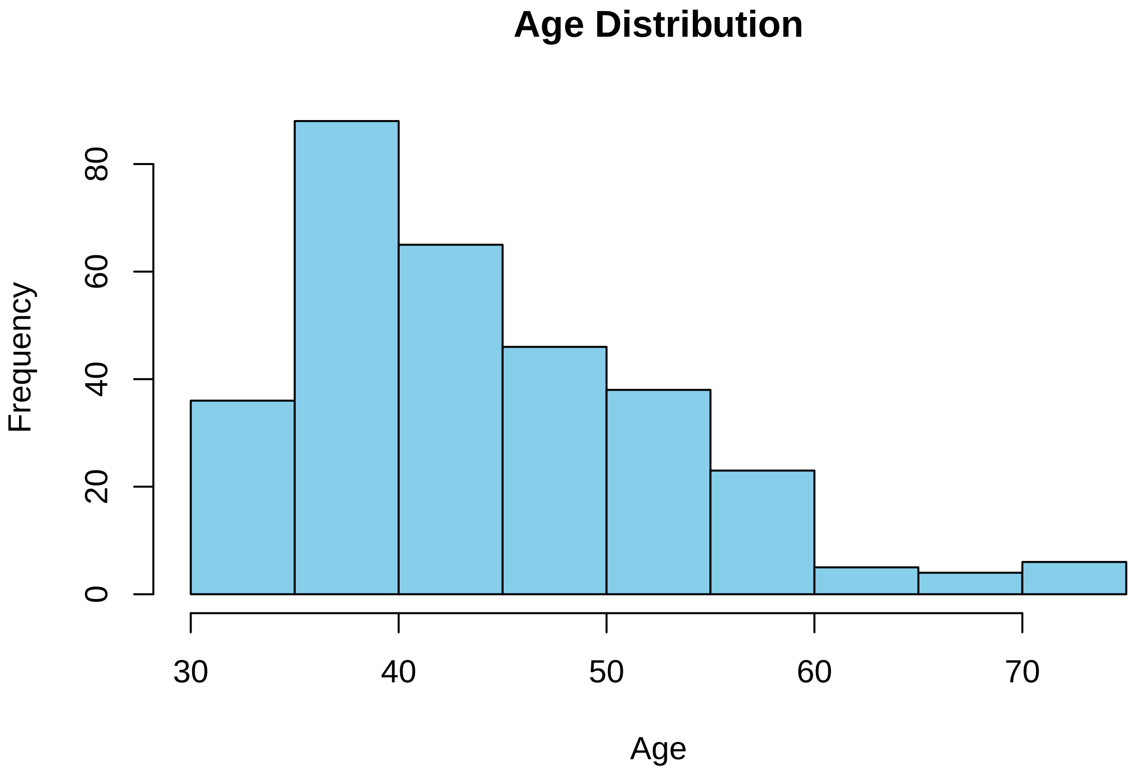


According to the United States Bureau of Labor Statististics, in 2021, 77 pecent of the labor force was White, 13 percent was black, 7 percent was Asian, 1 percent was American Indians and Alaska Natives, and less than half a percent were Native Hawaiians and other Pacific Islanders. Two or more races was 2 percent. By looking at the Race Distribution chart for this data, there seems to be a diverse staff as it compares to the makeup of the US Labor force.

Age

Finally, let's take a look at the age distribution of the employees.

```
# Create a histogram of age distribution
hist(all_employees$Age, main = "Age Distribution", xlab = "Age", ylab = "Frequency", col = "skyblue", bo
```

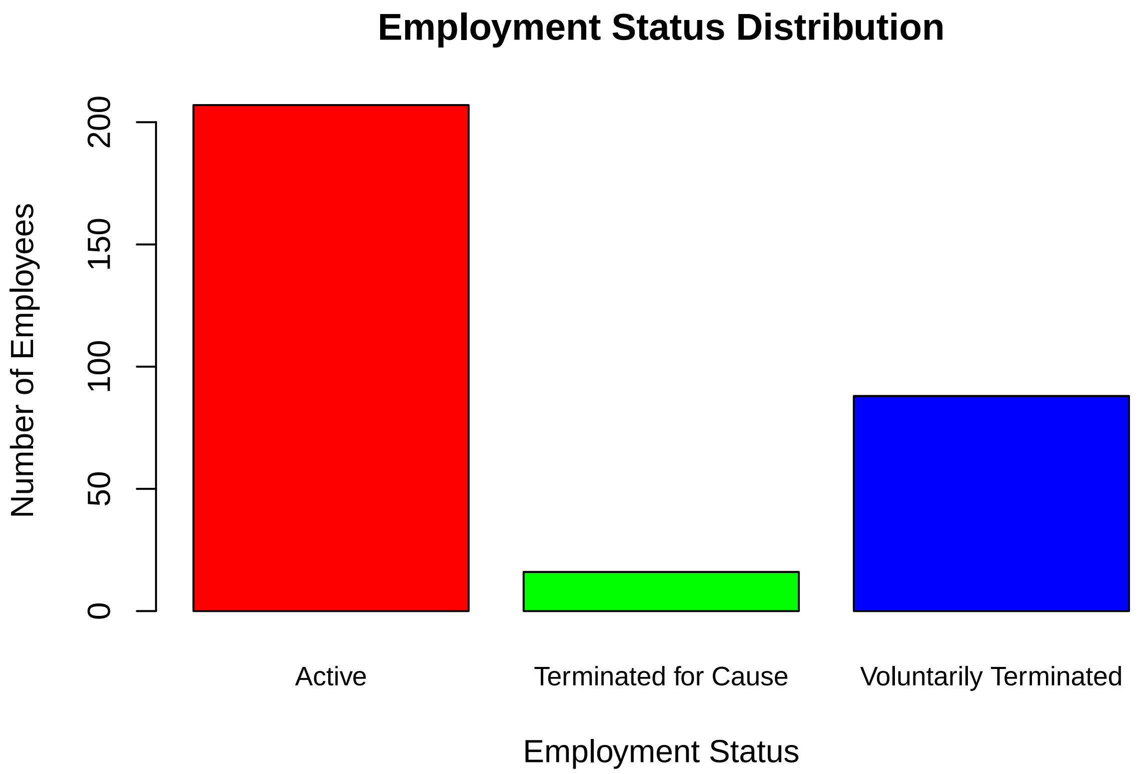


According to the data, it appears that there is no one below the age of 30 years that works for this company. The highest number of employees fall within the 35-40 year age range.

Additional Information

Employment Status

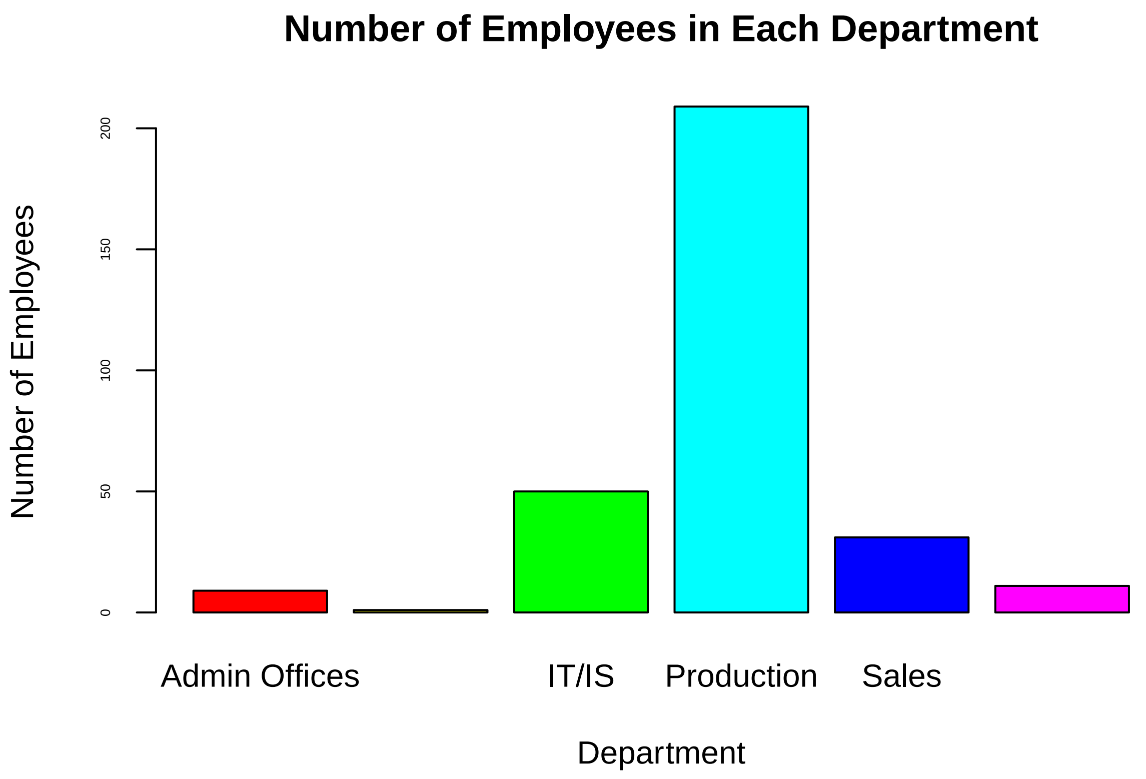
```r
# Count the number of employees in each employment status
employment_counts <- table(all_employees$EmploymentStatus)

# Create bar plot
barplot(employment_counts, main = "Employment Status Distribution",
        xlab = "Employment Status", ylab = "Number of Employees",
        col = rainbow(length(employment_counts)), cex.names = 0.8)
```



Departments

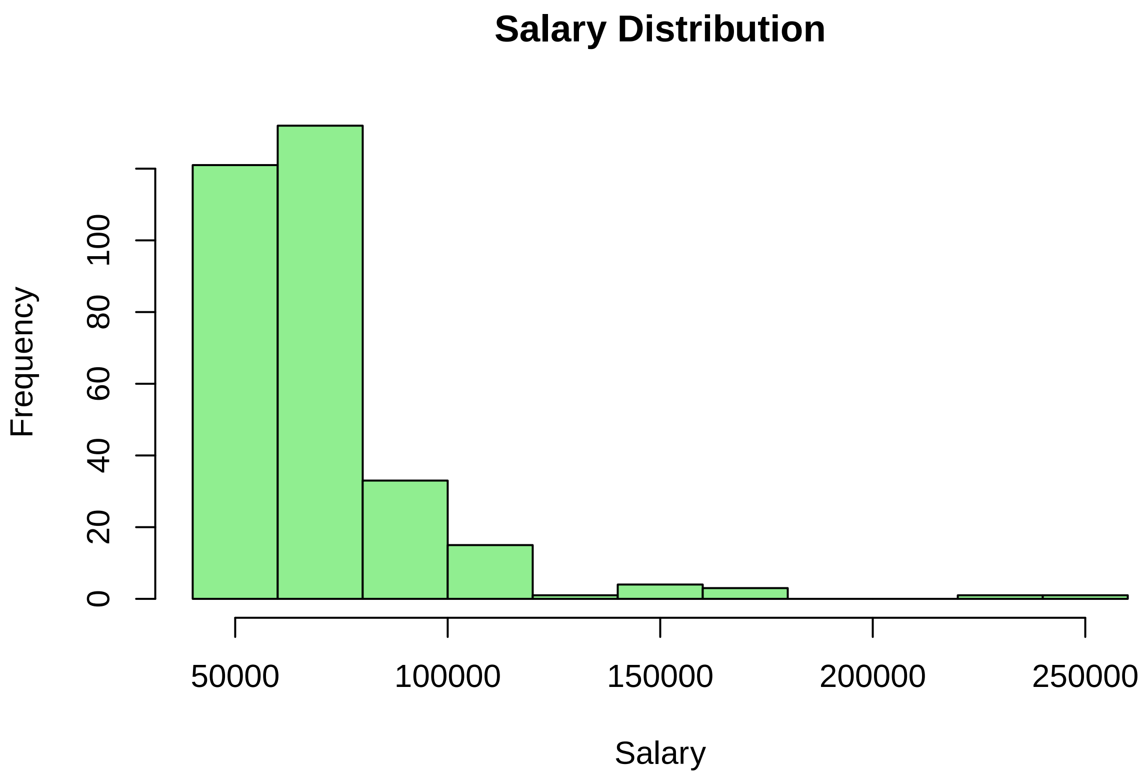
Let's take a look at the number of employees in each department.

```r
# Count the number of employees in each department
department_counts <- table(all_employees$Department)

# Create bar plot
barplot(department_counts, main = "Number of Employees in Each Department",
        xlab = "Department", ylab = "Number of Employees", col = rainbow(length(department_counts)), ce
```



Salary

```
# Create a histogram of salary distribution
hist(all_employees$Salary, main = "Salary Distribution", xlab = "Salary", ylab = "Frequency", col = "li
```

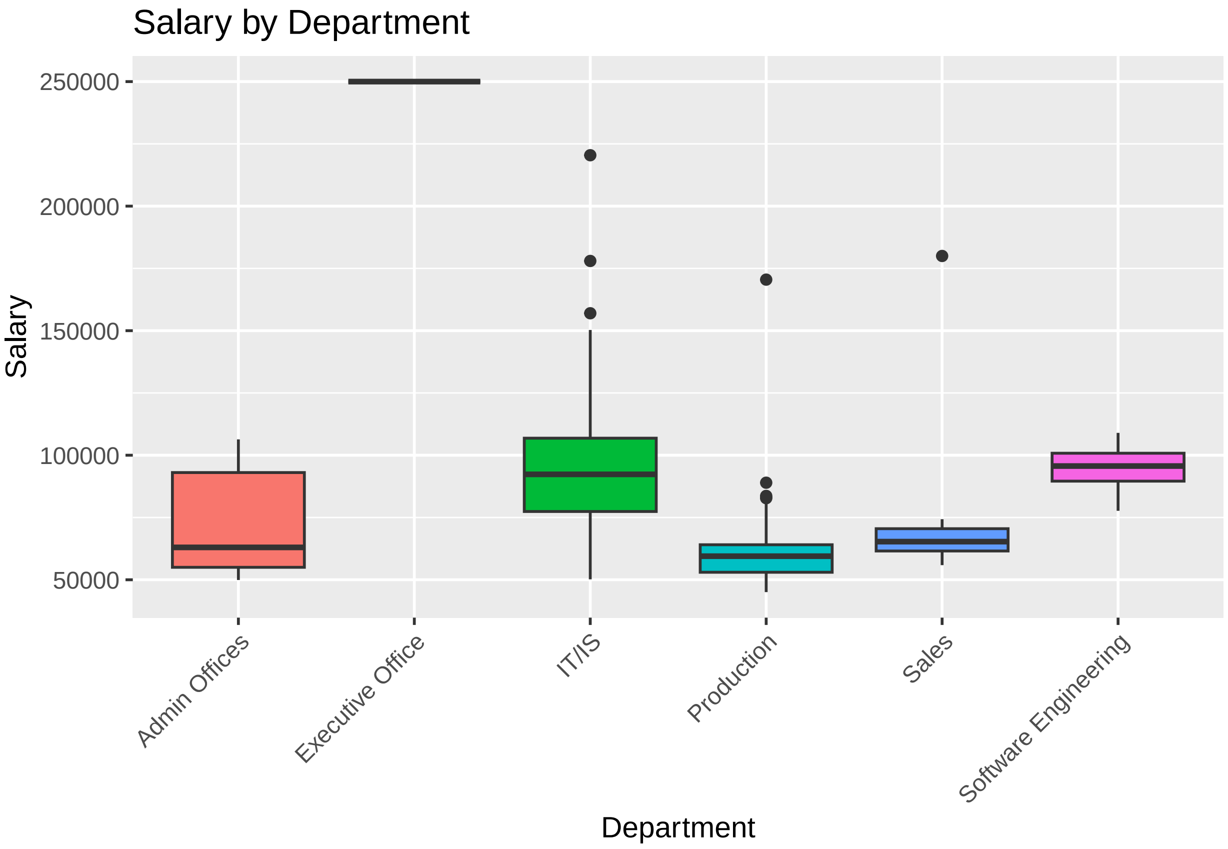


Salary by Department

Let's take a look at the salaries of the employees across departments.

```
# Let's see if Department has anything to do with Salary

# Create a box plot of Salary by Department
ggplot(hrdata, aes(x = Department, y = Salary, fill = Department)) +
  geom_boxplot() +
  labs(x = "Department", y = "Salary", title = "Salary by Department") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))+
  guides(fill = FALSE)
```

```
## Warning: The `<scale>` argument of `guides()` cannot be `FALSE`. Use "none" instead as
## of ggplot2 3.3.4.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



It looks like some departments make more than others, but are there any significant differences? Let's perform an ANOVA to get a better idea.

```
# Perform ANOVA to test for differences in Salary across departments
anova_result <- aov(Salary ~ Department, data = hrdata)

# Summary of ANOVA results
summary(anova_result)
```

```
##                 Df    Sum Sq   Mean Sq F value Pr(>F)
## Department       5 9.675e+10 1.935e+10   59.35 <2e-16 ***
## Residuals   305 9.944e+10 3.260e+08
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Perform Tukey's HSD test for pairwise comparisons
tukey_result <- TukeyHSD(anova_result)

# Summary of Tukey's HSD test results
print(tukey_result)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Salary ~ Department, data = hrdata)
##
## $Department
##                                              diff          lwr         upr
## Executive Office-Admin Offices         178208.111  123622.80451  232793.418
## IT/IS-Admin Offices                     25272.751    6522.08253   44023.420
## Production-Admin Offices               -11838.343  -29467.47252    5790.786
## Sales-Admin Offices                     -2730.631  -22338.25178   16876.990
## Software Engineering-Admin Offices      23197.566     -77.68728   46472.819
## IT/IS-Executive Office                -152935.360 -205234.80674 -100635.913
## Production-Executive Office           -190046.455 -241954.36097 -138138.548
## Sales-Executive Office                -180938.742 -233551.50994 -128325.974
## Software Engineering-Executive Office -155010.545 -209097.34566 -100923.745
## Production-IT/IS                       -37111.095  -45263.55584  -28958.633
## Sales-IT/IS                            -28003.382  -39841.25474  -16165.509
## Software Engineering-IT/IS              -2075.185  -19320.87571   15170.505
## Sales-Production                         9107.713    -858.92436   19074.350
## Software Engineering-Production         35035.909   19016.78139   51055.037
## Software Engineering-Sales              25928.196    7754.45004   44101.943
##                                           p adj
## Executive Office-Admin Offices        0.0000000
## IT/IS-Admin Offices                   0.0018693
## Production-Admin Offices              0.3883291
## Sales-Admin Offices                   0.9986828
## Software Engineering-Admin Offices    0.0513349
## IT/IS-Executive Office                0.0000000
## Production-Executive Office           0.0000000
## Sales-Executive Office                0.0000000
## Software Engineering-Executive Office 0.0000000
## Production-IT/IS                      0.0000000
## Sales-IT/IS                           0.0000000
## Software Engineering-IT/IS            0.9993511
## Sales-Production                      0.0953706
## Software Engineering-Production       0.0000000
## Software Engineering-Sales            0.0007738
```

According to the ANOVA there is a statistically significant difference in salary across departments. We can use Tukey's HSD test to see which specific departments have significantly different mean salaries.

According to Tukey's HSD test, there are statistically significant differences in salary between the departments below: * Executive Office and all other departments * IT/IS and Admin Offices, Production, Sales * Software Engineering and Production, Sales

Overall, the executive office makes the most compared to all other departments.

Also, IT/IS and Software Engineering departments make more than Production and Sales departments.

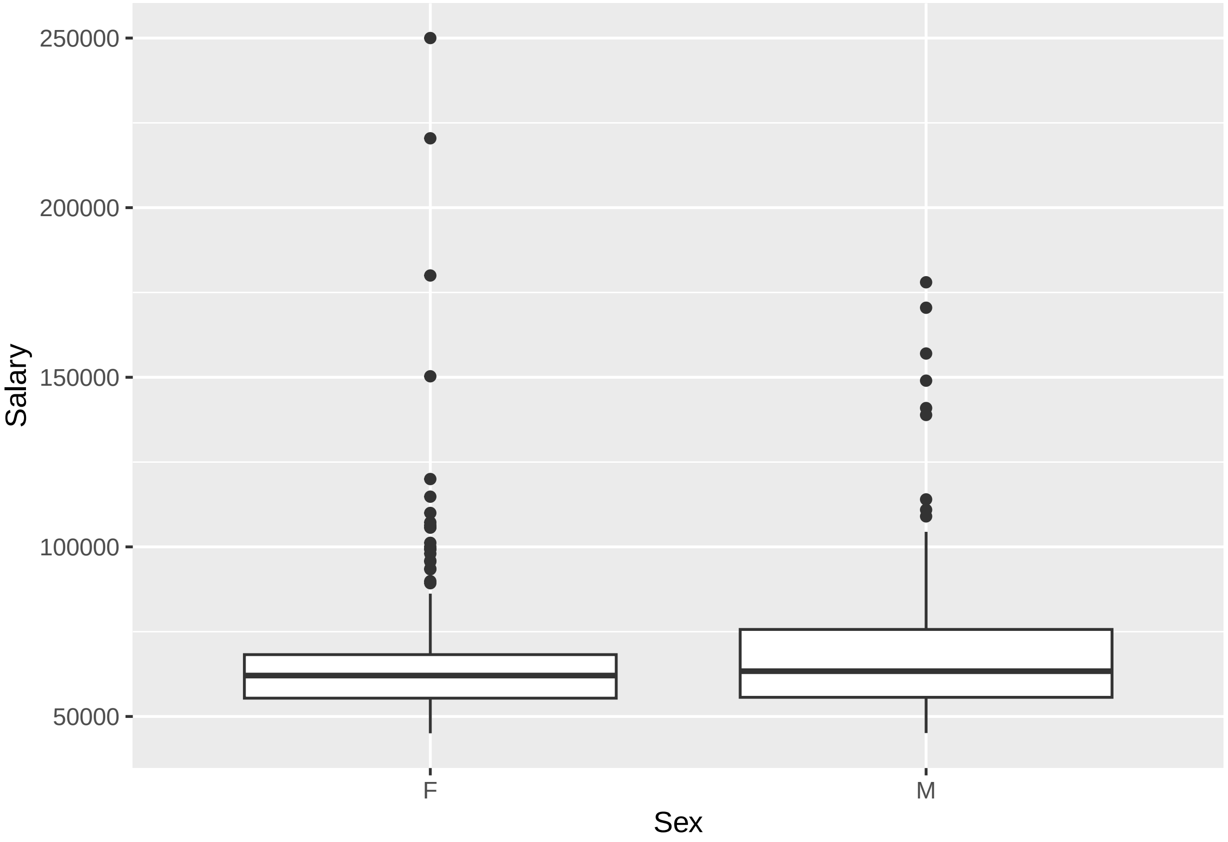
Salary Across Gender

Let's take a look at how salaries vary across gender.

```
# Summary statistics by gender
summary_by_gender <- hrdata %>%
  group_by(Sex) %>%
  summarise(mean_salary = mean(Salary),
            median_salary = median(Salary),
            min_salary = min(Salary),
            max_salary = max(Salary))

print(summary_by_gender)
```

```
## # A tibble: 2 x 5
##   Sex   mean_salary median_salary min_salary max_salary
##   <chr>       <dbl>         <dbl>      <dbl>      <dbl>
## 1 F          67787.        62066.      45046     250000
## 2 M          70629.        63353       45115     178000
```

```
# Create boxplot
ggplot(hrdata, aes(x = Sex, y = Salary)) +
  geom_boxplot() +
  labs(x = "Sex", y = "Salary")
```



From the visuals, it looks like Males may make more money overall as compared to females. Let's do a t test to find out for sure.

```r
# Perform t-test
t_test_result <- t.test(Salary ~ Sex, data = hrdata)

# Print t-test result
print(t_test_result)
```

```
##
##  Welch Two Sample t-test
##
## data:  Salary by Sex
## t = -0.9956, df = 296.39, p-value = 0.3203
## alternative hypothesis: true difference in means between group F and group M is not equal to 0
## 95 percent confidence interval:
##  -8461.792  2776.447
## sample estimates:
## mean in group F mean in group M
##        67786.73        70629.40
```

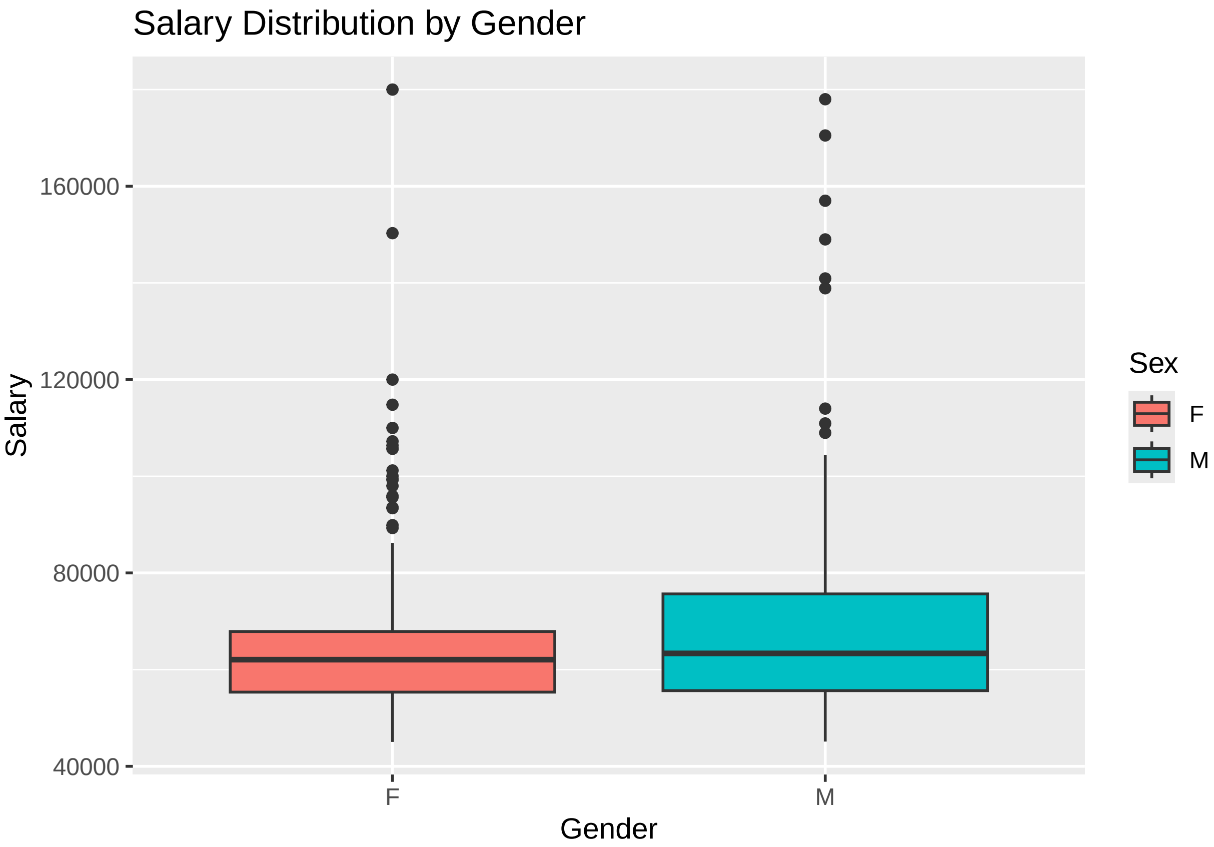
The p value does not indicate that there is a statistically significant difference in salary across sex. Something to keep in mind is that the President & CEO as well as the CIO make the highest salaries in the company and are both Female. We may want to eliminate them from the calculations, if we want to get a better idea of the rest of the company.

Let’s look at the same data, but without counting the President & CEO or CIO roles.

```r
# Filter out "President & CEO" and "CIO" roles
filtered_hrdata <- hrdata %>%
  filter(Position != "President & CEO" & Position != "CIO")

# Summary statistics by gender excluding "President & CEO" and "CIO" roles
summary_by_gender_filtered <- filtered_hrdata %>%
  group_by(Sex) %>%
  summarise(mean_salary = mean(Salary),
            median_salary = median(Salary),
            min_salary = min(Salary),
            max_salary = max(Salary))

print(summary_by_gender_filtered)
```

```
## # A tibble: 2 x 5
##   Sex   mean_salary median_salary min_salary max_salary
##   <chr>       <dbl>         <dbl>      <dbl>      <dbl>
## 1 F          65862.         62063      45046     180000
## 2 M          70629.         63353      45115     178000
```

```r
library(ggplot2)
library(dplyr)

# Box plot to visualize salary distribution by gender
ggplot(filtered_hrdata, aes(x = Sex, y = Salary, fill = Sex)) +
  geom_boxplot() +
  labs(x = "Gender", y = "Salary", title = "Salary Distribution by Gender")
```



```r
# Perform t-test
t_test_result <- t.test(Salary ~ Sex, data = filtered_hrdata)

# Print the result
print(t_test_result)
```

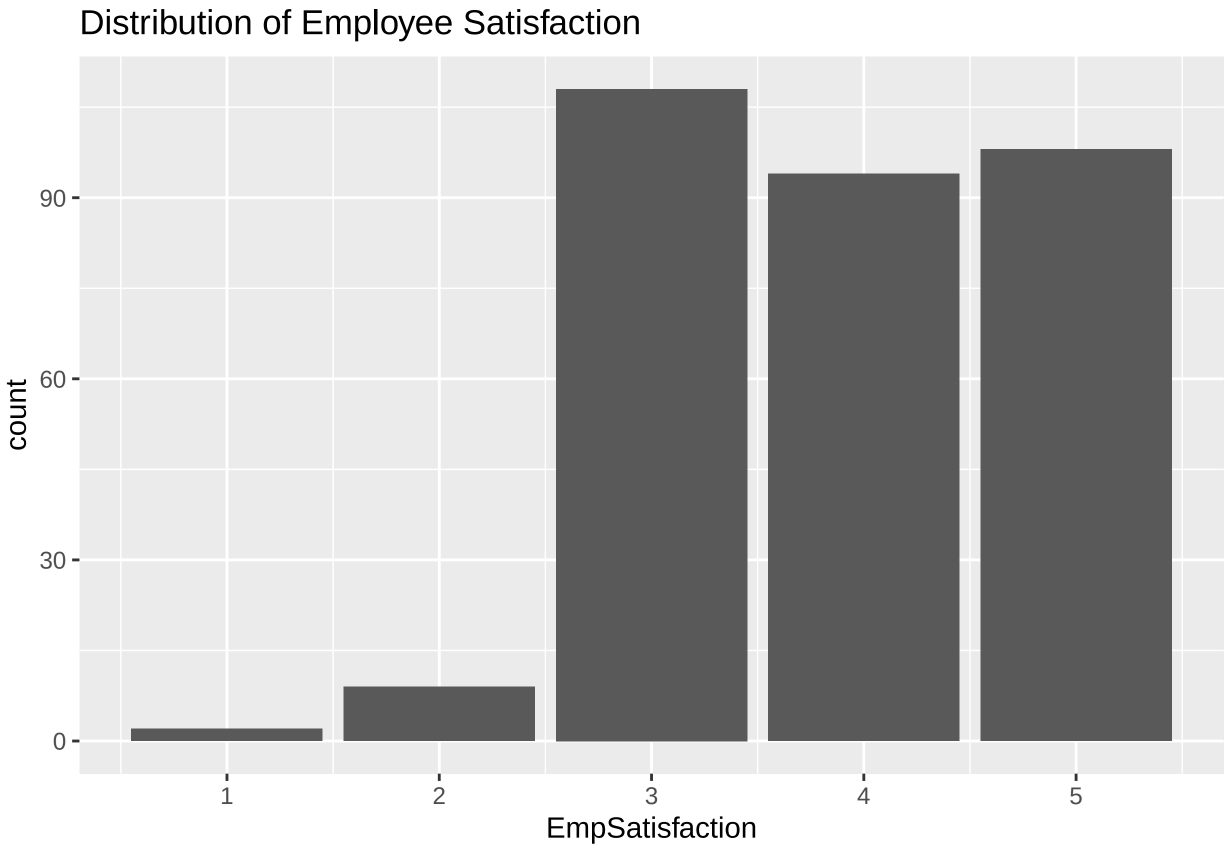
```
##
##  Welch Two Sample t-test
##
## data:  Salary by Sex
## t = -1.8931, df = 243.84, p-value = 0.05953
## alternative hypothesis: true difference in means between group F and group M is not equal to 0
## 95 percent confidence interval:
##  -9727.5297   193.0286
## sample estimates:
## mean in group F mean in group M
##        65862.15        70629.40
```

The results of this t-test still indicate that there is no statistically significant difference in the salaries between male and female employees. Great job to this company!

Employee Satisfaction

Let's take a look at overall employee satisfaction.

```r
library(dplyr)

ggplot(all_employees, aes(x = EmpSatisfaction)) +
  geom_bar() +
  labs(title = "Distribution of Employee Satisfaction")
```



```
mean(all_employees$EmpSatisfaction)
```

```
## [1] 3.890675
```

The mean employee satisfaction rate is 3.89 out of 5. This number can be compared to the organization's overall goal.

Employee Satisfaction and Performance Score

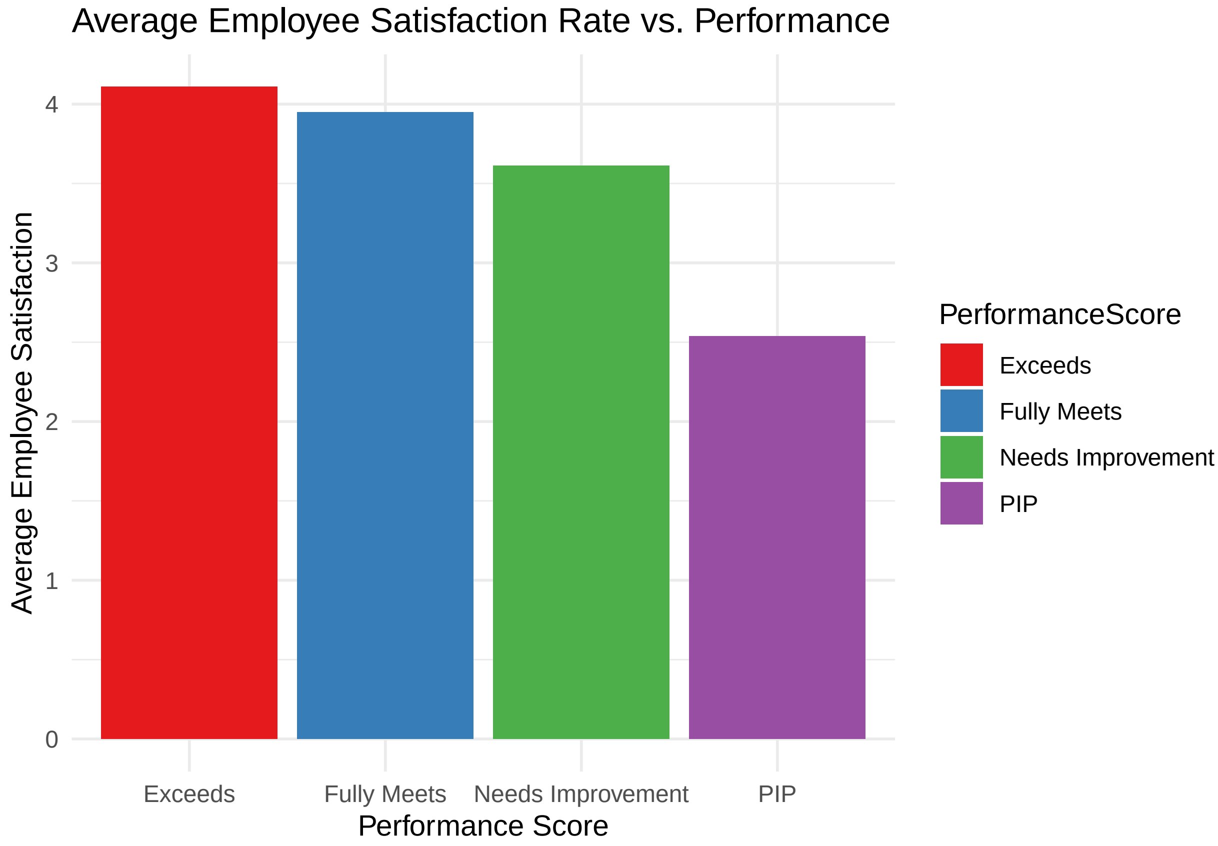
The average employee satisfaction score across the company is 3.89/5. Now let's take a look and see if there is any correlation between Performance Scores and Employee Satisfaction.

```
hrdata %>%
  group_by(PerformanceScore) %>%
  summarise(average_satisfaction=mean(EmpSatisfaction), min_satisfaction=min(EmpSatisfaction), max_sati
```

```
## # A tibble: 4 x 4
##   PerformanceScore  average_satisfaction min_satisfaction max_satisfaction
##   <chr>                            <dbl>            <dbl>            <dbl>
## 1 Exceeds                           4.11                3                5
## 2 Fully Meets                       3.95                2                5
## 3 Needs Improvement                 3.61                2                5
## 4 PIP                               2.54                1                5
```

```
avg_satisfaction <- hrdata %>%
  group_by(PerformanceScore) %>%
  summarise(avg_satisfaction = mean(EmpSatisfaction))
```

```r
ggplot(avg_satisfaction, aes(x = PerformanceScore, y = avg_satisfaction, fill = PerformanceScore)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(x = "Performance Score", y = "Average Employee Satisfaction") +
  scale_fill_brewer(palette = "Set1") + theme_minimal() +
  labs(title="Average Employee Satisfaction Rate vs. Performance")
```



```r
str(hrdata)
```

```
## spc_tbl_ [311 x 37] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ Employee_Name             : chr [1:311] "Adinolfi, Wilson  K" "Ait Sidi, Karthikeyan" "Akinkuolie
##  $ EmpID                     : num [1:311] 10026 10084 10196 10088 10069 ...
##  $ MarriedID                 : num [1:311] 0 1 1 1 0 0 0 0 0 0 ...
##  $ MaritalStatusID           : num [1:311] 0 1 1 1 2 0 0 4 0 2 ...
##  $ GenderID                  : num [1:311] 1 1 0 0 0 0 0 1 0 1 ...
##  $ EmpStatusID               : num [1:311] 1 5 5 1 5 1 1 1 3 1 ...
##  $ DeptID                    : num [1:311] 5 3 5 5 5 5 4 5 5 3 ...
##  $ PerfScoreID               : num [1:311] 4 3 3 3 3 4 3 3 3 3 ...
##  $ FromDiversityJobFairID    : num [1:311] 0 0 0 0 0 0 0 0 1 0 ...
##  $ Salary                    : num [1:311] 62506 104437 64955 64991 50825 ...
##  $ Termd                     : num [1:311] 0 1 1 0 1 0 0 0 0 0 ...
##  $ PositionID                : num [1:311] 19 27 20 19 19 19 24 19 19 14 ...
##  $ Position                  : chr [1:311] "Production Technician I" "Sr. DBA" "Production Technicia
##  $ State                     : chr [1:311] "MA" "MA" "MA" "MA" ...
##  $ Zip                       : num [1:311] 1960 2148 1810 1886 2169 ...
##  $ DOB                       : Date[1:311], format: "1983-07-10" "1975-05-05" ...
##  $ Sex                       : chr [1:311] "M" "M" "F" "F" ...
##  $ MaritalDesc               : chr [1:311] "Single" "Married" "Married" "Married" ...
##  $ CitizenDesc              : chr [1:311] "US Citizen" "US Citizen" "US Citizen" "US Citizen" ...
##  $ HispanicLatino           : chr [1:311] "No" "No" "No" "No" ...
##  $ RaceDesc                 : chr [1:311] "White" "White" "White" "White" ...
##  $ DateofHire               : chr [1:311] "7/5/2011" "3/30/2015" "7/5/2011" "1/7/2008" ...
##  $ DateofTermination        : chr [1:311] NA "6/16/2016" "9/24/2012" NA ...
##  $ TermReason               : chr [1:311] "N/A-StillEmployed" "career change" "hours" "N/A-StillEmpl
##  $ EmploymentStatus         : chr [1:311] "Active" "Voluntarily Terminated" "Voluntarily Terminated
##  $ Department               : chr [1:311] "Production" "IT/IS" "Production" "Production" ...
##  $ ManagerName              : chr [1:311] "Michael Albert" "Simon Roup" "Kissy Sullivan" "Elijiah G
##  $ ManagerID                : num [1:311] 22 4 20 16 39 11 10 19 12 7 ...
##  $ RecruitmentSource        : chr [1:311] "LinkedIn" "Indeed" "LinkedIn" "Indeed" ...
##  $ PerformanceScore         : chr [1:311] "Exceeds" "Fully Meets" "Fully Meets" "Fully Meets" ...
##  $ EngagementSurvey         : num [1:311] 4.6 4.96 3.02 4.84 5 5 3.04 5 4.46 5 ...
##  $ EmpSatisfaction          : num [1:311] 5 3 3 5 4 5 3 4 3 5 ...
##  $ SpecialProjectsCount     : num [1:311] 0 6 0 0 0 0 4 0 0 6 ...
##  $ LastPerformanceReview_Date: chr [1:311] "1/17/2019" "2/24/2016" "5/15/2012" "1/3/2019" ...
##  $ DaysLateLast30           : num [1:311] 0 0 0 0 0 0 0 0 0 0 ...
##  $ Absences                 : num [1:311] 1 17 3 15 2 15 19 19 4 16 ...
##  $ Age                      : num [1:311] 40 49 35 35 34 47 45 41 54 36 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   Employee_Name = col_character(),
##   ..   EmpID = col_double(),
##   ..   MarriedID = col_double(),
##   ..   MaritalStatusID = col_double(),
##   ..   GenderID = col_double(),
##   ..   EmpStatusID = col_double(),
##   ..   DeptID = col_double(),
##   ..   PerfScoreID = col_double(),
##   ..   FromDiversityJobFairID = col_double(),
##   ..   Salary = col_double(),
##   ..   Termd = col_double(),
##   ..   PositionID = col_double(),
##   ..   Position = col_character(),
##   ..   State = col_character(),
##   ..   Zip = col_double(),
##   ..   DOB = col_character(),
##   ..   Sex = col_character(),
##   ..   MaritalDesc = col_character(),
##   ..   CitizenDesc = col_character(),
##   ..   HispanicLatino = col_character(),
##   ..   RaceDesc = col_character(),
##   ..   DateofHire = col_character(),
##   ..   DateofTermination = col_character(),
##   ..   TermReason = col_character(),
##   ..   EmploymentStatus = col_character(),
##   ..   Department = col_character(),
##   ..   ManagerName = col_character(),
##   ..   ManagerID = col_double(),
##   ..   RecruitmentSource = col_character(),
##   ..   PerformanceScore = col_character(),
##   ..   EngagementSurvey = col_double(),
##   ..   EmpSatisfaction = col_double(),
##   ..   SpecialProjectsCount = col_double(),
##   ..   LastPerformanceReview_Date = col_character(),
##   ..   DaysLateLast30 = col_double(),
##   ..   Absences = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
anova_result <- aov(EmpSatisfaction ~ PerformanceScore, data = hrdata)
summary(anova_result)
```

```
##                   Df Sum Sq Mean Sq F value   Pr(>F)    
## PerformanceScore   3   27.8   9.266   12.45 1.05e-07 ***
## Residuals        307  228.5   0.744                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The ANOVA results indicate that there is a significant relationship between performance scores and employee satisfaction levels in the dataset because the p-value (1.05e-07) is less than 0.05.

Now let's dig deeper and see between which Performance Scores is there a statistically siginificant difference in Employee Satisfaction?

```
# Load the required library for post-hoc tests
library(stats)

# Perform Tukey's HSD test
tukey_result <- TukeyHSD(anova_result)

# Summary of Tukey's HSD test
summary(tukey_result)
```

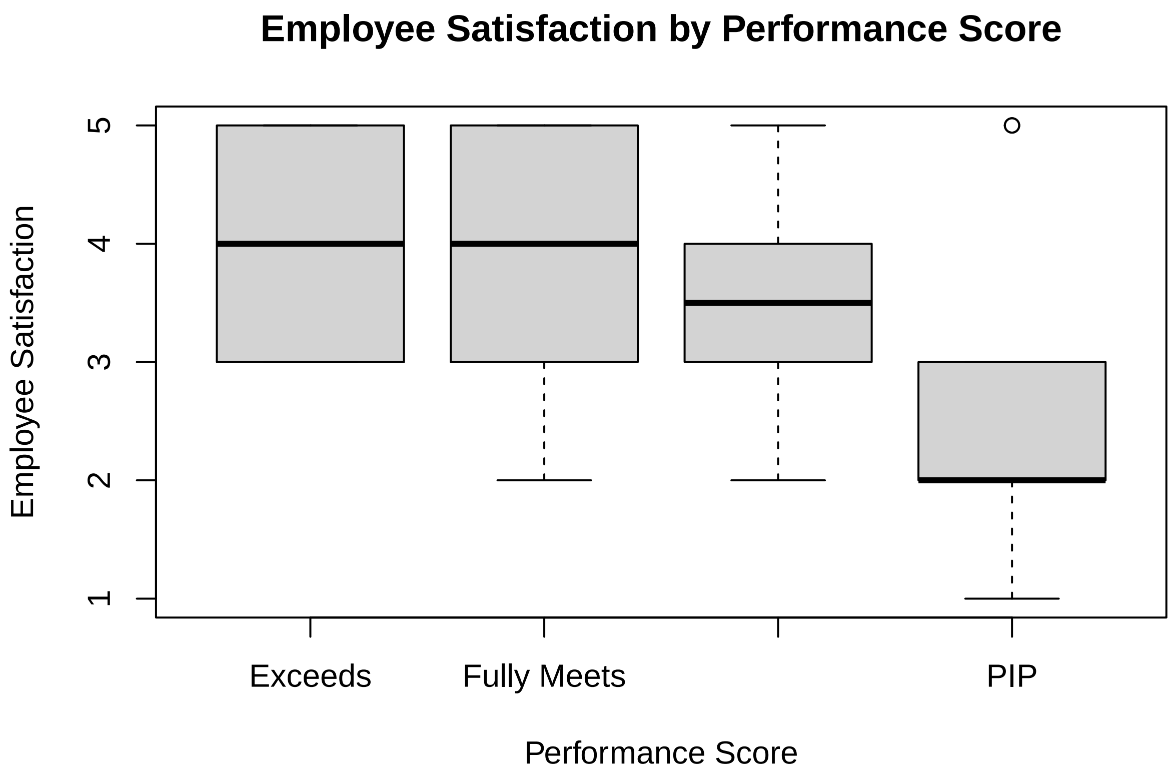
```
##                  Length Class  Mode   
## PerformanceScore 24     -none- numeric
```

The results suggest that:

- There is no statistically significant difference in employee satisfaction between the "Fully Meets" and "Exceeds" performance levels.
- There is no statistically significant difference in employee satisfaction between the "Needs Improvement" and "Exceeds" performance levels.
- There is no statistically significant difference in employee satisfaction between the "Needs Improvement" and "Fully Meets" performance levels.
- There is a statistically significant difference in employee satisfaction between the "PIP" (Performance Improvement Plan) and "Exceeds" performance levels.
- There is a statistically significant difference in employee satisfaction between the "PIP" and "Fully Meets" performance levels.
- There is a statistically significant difference in employee satisfaction between the "PIP" and "Needs Improvement" performance levels.

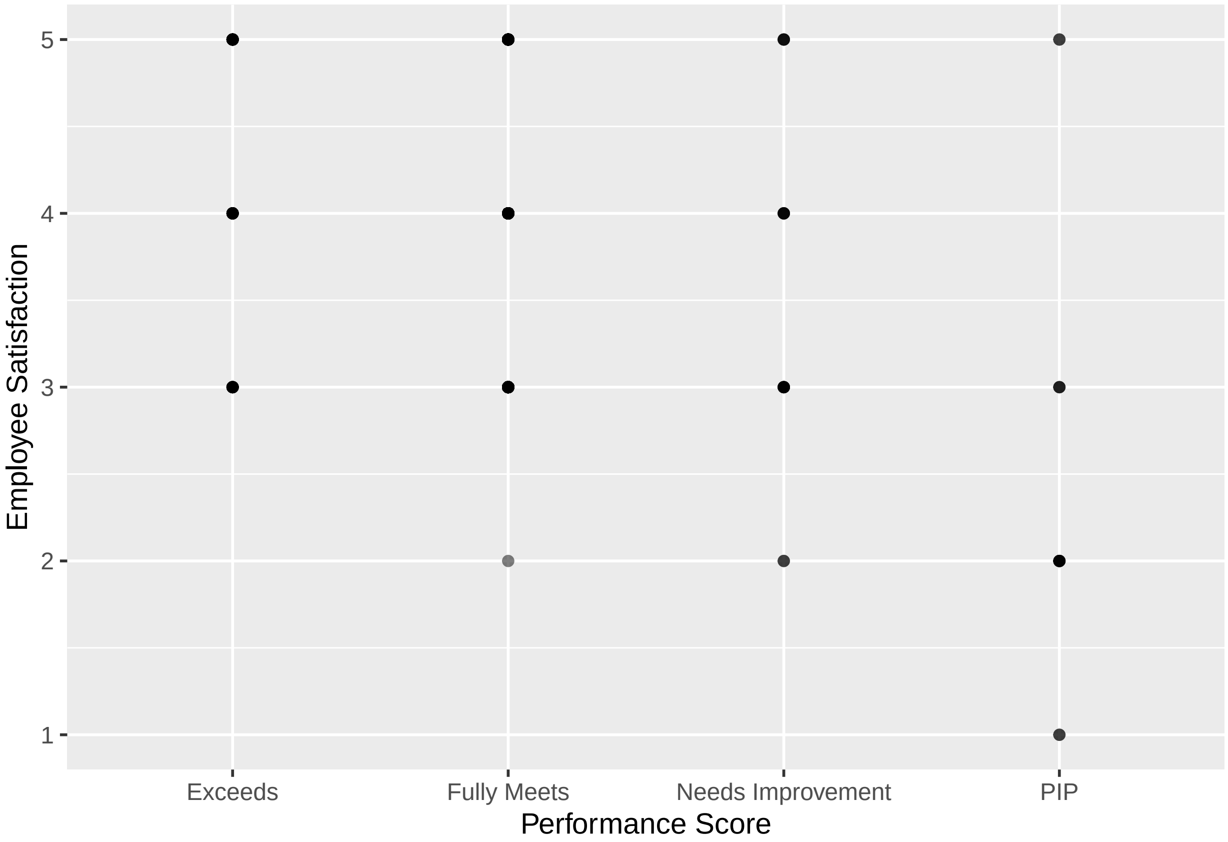
It looks like there are only statistically significant differences in satisfaction between employees in the PIP and the emplpyees in other performance score levels. The boxplot below indicates that those in PIP have lower overall Employee Satisfaction scores than those in other Performance Score Categories.

```
# Create a boxplot to visualize the difference in EmpSatisfaction between PIP and other performance sco
boxplot(EmpSatisfaction ~ PerformanceScore, data = hrdata,
        main = "Employee Satisfaction by Performance Score",
        xlab = "Performance Score", ylab = "Employee Satisfaction", cex.names = 0.5)
```



Let's perform a regression model to estimate how changes in performance score affect employee satisfaction while accounting for other factors.

```
# Create scatter plot
ggplot(hrdata, aes(x = PerformanceScore, y = EmpSatisfaction)) +
  geom_point(alpha = 0.5) +
  labs(x = "Performance Score", y = "Employee Satisfaction")
```



```r
# Perform linear regression
regression_model <- lm(EmpSatisfaction ~ PerformanceScore, data = hrdata)

# Summary of regression model
summary(regression_model)
```

```
##
## Call:
## lm(formula = EmpSatisfaction ~ PerformanceScore, data = hrdata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.95062 -0.95062  0.04938  1.04938  2.46154
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                        4.1081     0.1418  28.966  < 2e-16 ***
## PerformanceScoreFully Meets       -0.1575     0.1522  -1.034   0.3017
## PerformanceScoreNeeds Improvement  -0.4970     0.2479  -2.005   0.0459 *
## PerformanceScorePIP               -1.5696     0.2781  -5.643 3.79e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8627 on 307 degrees of freedom
## Multiple R-squared:  0.1085, Adjusted R-squared:  0.09976
## F-statistic: 12.45 on 3 and 307 DF,  p-value: 1.05e-07
```

The regression analysis suggests that Performance Score has a significant effect on EmpSatisfaction, with some performance categories having lower satisfaction scores compared to others. However, the model does not explain a large proportion of the variance in satisfaction, indicating that other factors may also be important.

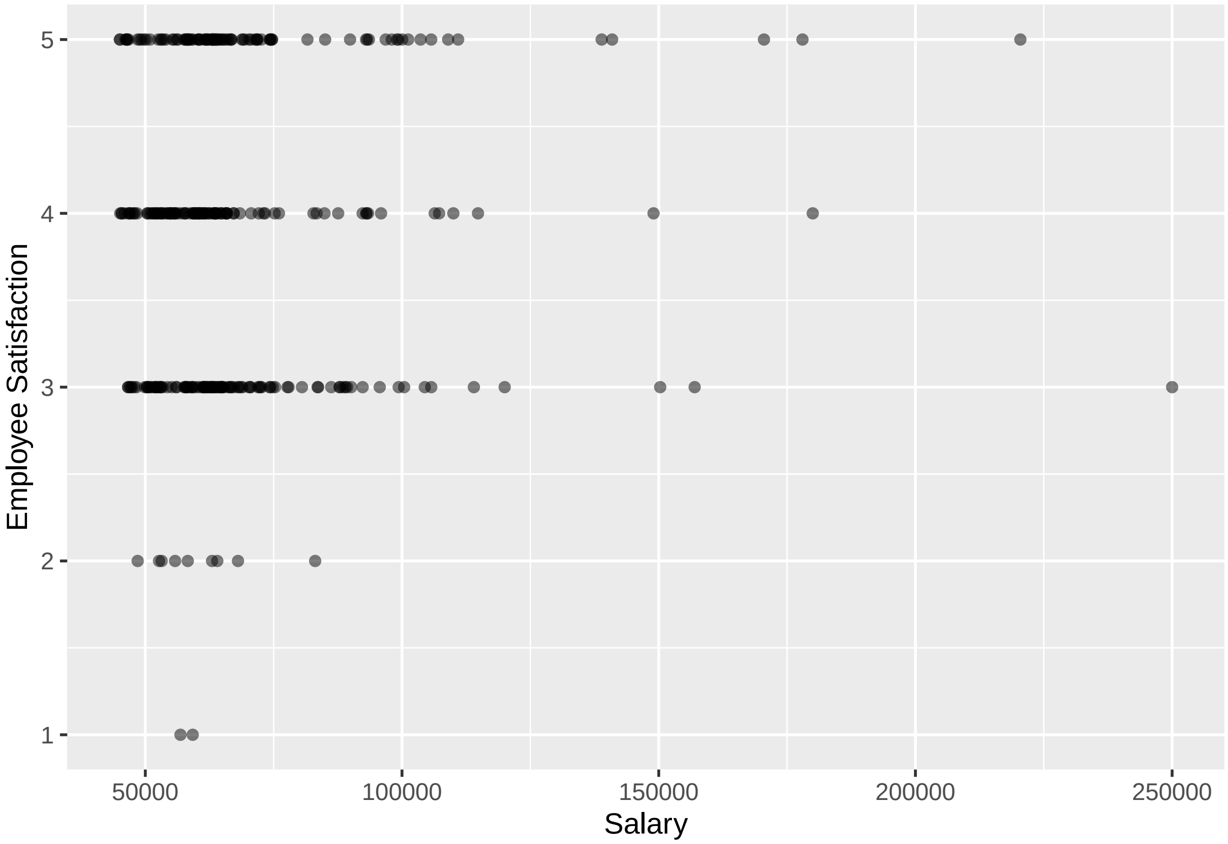
Employee Satisfaction and Salary

Since the Performance Score may not be the only factor with a significant impact on Employee Satisfaction, let's see if salary has any impact.

```r
# Perform linear regression
regression_model <- lm(EmpSatisfaction ~ Salary, data = hrdata)

# Summary of regression model
summary(regression_model)
```

```
##
## Call:
## lm(formula = EmpSatisfaction ~ Salary, data = hrdata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8685 -0.8733  0.1223  1.0769  1.1637
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.734e+00  1.507e-01  24.776   <2e-16 ***
## Salary      2.267e-06  2.052e-06   1.105     0.27
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9089 on 309 degrees of freedom
## Multiple R-squared:  0.003934,   Adjusted R-squared:  0.0007101
## F-statistic:  1.22 on 1 and 309 DF,  p-value: 0.2702
```

```r
# Create scatter plot
ggplot(hrdata, aes(x = Salary, y = EmpSatisfaction)) +
  geom_point(alpha = 0.5) + labs(x = "Salary", y = "Employee Satisfaction")
```



The regression analysis suggests that there is no statistically significant relationship between salary and employee satisfaction scores.

Employee Satisfaction and Managers

```
# Perform ANOVA
anova_result <- aov(EmpSatisfaction ~ ManagerName, data = hrdata)

# Summary of ANOVA
summary(anova_result)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## ManagerName  20  14.98  0.7488     0.9  0.588
## Residuals   290 241.31  0.8321
```

The regression analysis suggests that there is no statistically significant relationship between manager and employee satisfaction scores.

It is suggested that the company look into gathering more data on other factors that may have an affect on employee satisfaction to rule them out.

Terminations

It is important to look at employees who are no longer active and determine if there are any patterns in that data.

Termination Reasons

First, let's take a look at the reasons why employees were terminated.

```r
# Load the dplyr package
library(dplyr)

# Count the number of employees for each termination reason
termination_counts <- table(termed_employees$TermReason)

# Create a data frame with termination reasons and counts
termination_table <- data.frame(Termination_Reason = names(termination_counts),
                                Number_of_Employees = as.vector(termination_counts))

# Sort the table by the count column in descending order
termination_table <- termination_table %>%
  arrange(desc(Number_of_Employees))

# Print the sorted table
print(termination_table)
```

```
##                  Termination_Reason Number_of_Employees
## 1                  Another position                  20
## 2                           unhappy                  14
## 3                        more money                  11
## 4                     career change                   9
## 5                             hours                   8
## 6                        attendance                   7
## 7            relocation out of area                   5
## 8                  return to school                   5
## 9                          military                   4
## 10                 no-call, no-show                   4
## 11                      performance                   4
## 12                         retiring                   4
## 13 maternity leave - did not return                   3
## 14                   medical issues                   3
## 15                 Fatal attraction                   1
## 16                 gross misconduct                   1
## 17    Learned that he is a gangster                   1
```

It looks like the top three reasons why employees are leaving is for another position, they are unhappy, or more money. The company may want to drill down further into those who are leaving for another position to determine why the other role was better.

Let's take a look at those who left for another position.

```r
# Filter the dataset for employees who left for "Another position"
another_position <- termed_employees[termed_employees$TermReason == "Another position", ]

# Sort the filtered dataset by the "Salary" column in ascending order
another_position <- another_position[order(another_position$Salary), ]

# Display the salaries and departments for employees who left for "Another position"
print(another_position[, c("Salary", "Department")])
```

```
## # A tibble: 20 x 2
##    Salary Department
##     <dbl> <chr>
##  1  46799 Production
##  2  47434 Production
##  3  48513 Production
##  4  52505 Production
##  5  53180 Production
##  6  53492 Production
##  7  54670 Production
##  8  55578 Production
##  9  58062 Production
## 10  60754 Production
## 11  62659 Production
## 12  63515 Production
## 13  66074 Production
## 14  68407 Production
## 15  72202 Production
## 16  74326 Sales
## 17  74669 Production
## 18  80512 Production
## 19 108987 Software Engineering
## 20 120000 IT/IS
```

Next, let's look into those who left because they were unhappy.

```r
# Filter the dataset for employees who left because they were "unhappy"
unhappy_employees <- termed_employees[termed_employees$TermReason == "unhappy", ]

# Sort the filtered dataset by the "Salary" column in ascending order
unhappy_employees <- unhappy_employees[order(unhappy_employees$Salary), ]

# Display the salary, department, and EmpSatisfaction for employees who left because they were "unhappy
print(unhappy_employees[, c("Salary", "Department", "EmpSatisfaction")])
```

```
## # A tibble: 14 x 3
##    Salary Department EmpSatisfaction
##     <dbl> <chr>                <dbl>
##  1  46430 Production               5
##  2  47211 Production               3
##  3  52624 Production               4
##  4  55140 Production               3
##  5  55722 Production               4
##  6  55800 Production               2
##  7  60270 Production               5
##  8  61154 Production               4
##  9  62425 Production               4
## 10  64066 Production               5
## 11  68182 Production               3
## 12  71966 Production               3
## 13  74813 Production               3
## 14  83082 Production               2
```

Now let's look into the salary and department for those who left for more money.

```r
# Filter the dataset for employees who left for "more money"
left_for_more_money <- termed_employees[termed_employees$TermReason == "more money", ]

# Sort the filtered dataset by the "Salary" column in ascending order
left_for_more_money <- left_for_more_money[order(left_for_more_money$Salary), ]
```

```r
# Display the salaries and departments for employees who left for "more money"
print(left_for_more_money[, c("Salary", "Department")])
```

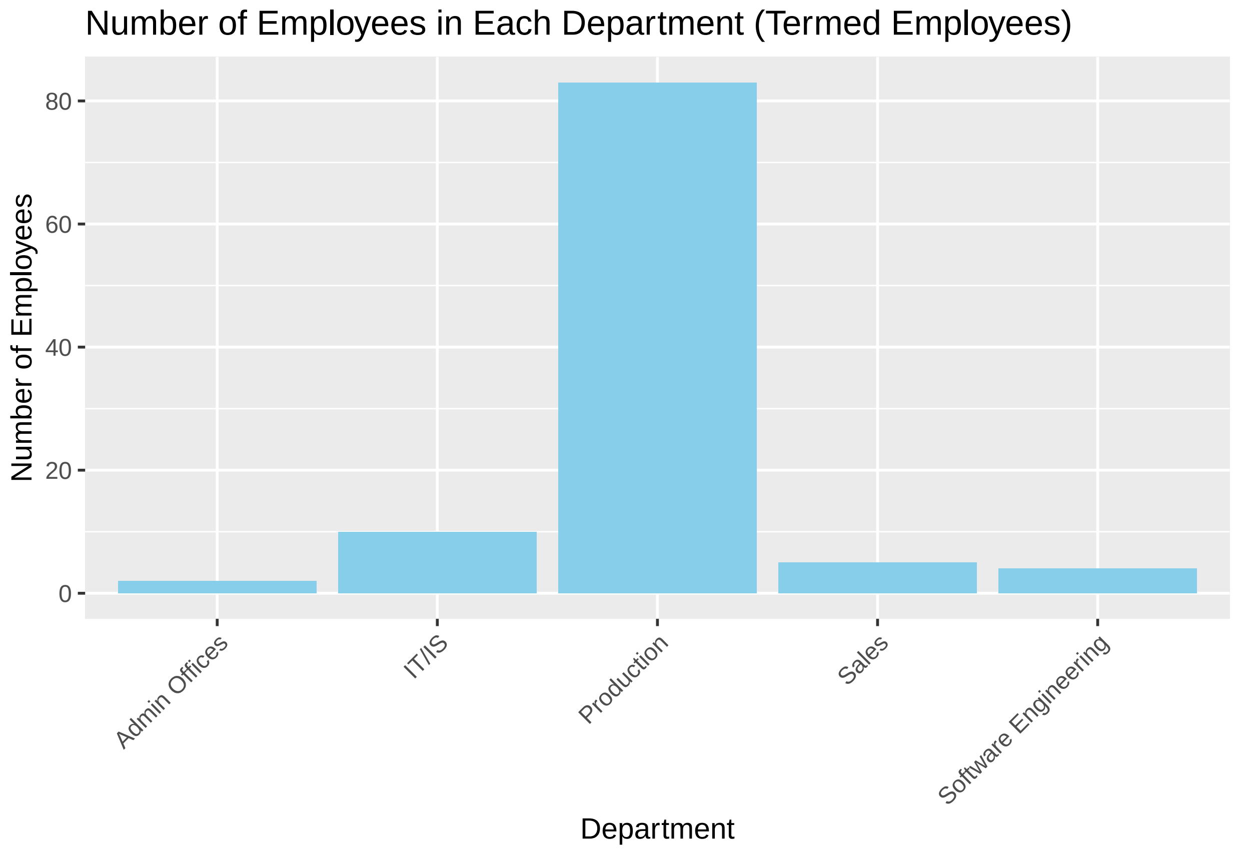
```
## # A tibble: 11 x 2
##    Salary Department
##     <dbl> <chr>
##  1  45433 Production
##  2  46664 Production
##  3  46837 Production
##  4  54005 Production
##  5  57954 Production
##  6  58275 Production
##  7  61729 Production
##  8  61962 Production
##  9  63813 Production
## 10  64724 Production
## 11  67237 Production
```

A majority of the emplpoyees who left for Another position were a part of the Production department. All of the employees who left because they were unhappy and for more money were a part of the Production department and made between 45,000 dollars to 84,000 dollars in salary. The company may want to consider increasing the salaries for lower-salary Production employees. The company may also want to look into the Production department more and determine what is making employees unhappy with their role.

Terminations by Department

Let's take a look at the distribution of terminations across departments.

```r
# Load necessary packages
library(ggplot2)

# Count the number of employees in each department from the termed_employees variable
department_counts <- table(termed_employees$Department)

# Convert department counts to data frame
department_data <- data.frame(Department = names(department_counts),
                              Number_of_Employees = as.vector(department_counts))

# Create a bar plot
ggplot(department_data, aes(x = Department, y = Number_of_Employees)) +
  geom_bar(stat = "identity", fill = "skyblue") +
  labs(title = "Number of Employees in Each Department (Termed Employees)",
       x = "Department", y = "Number of Employees") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



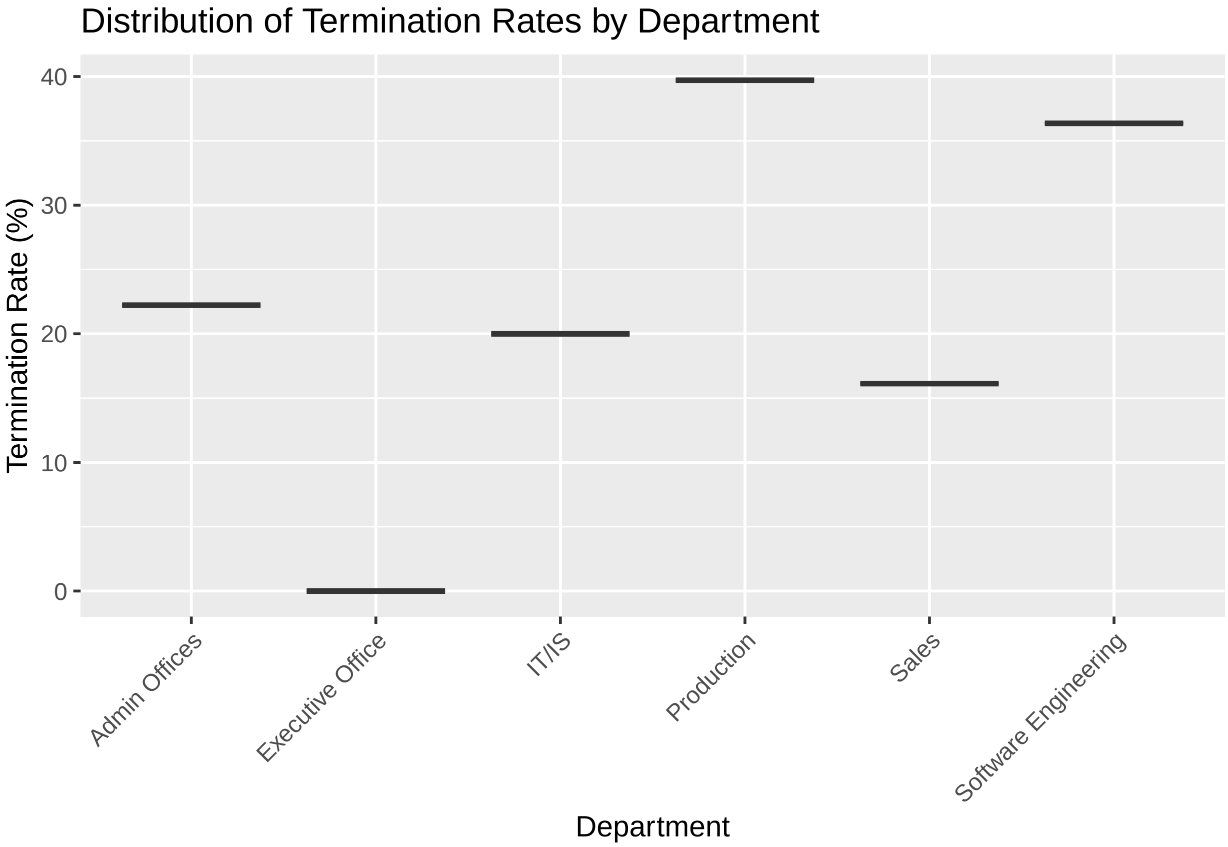
From the chart, it looks like the Production department has the highest number of terminations. Let's take a better look at the termination rates for each department and compare them with the overall termination rate.

```
# Calculate termination rates for each department in all_employees dataset
all_employees_termination_rates <- all_employees %>%
  group_by(Department) %>%
  summarise(Termination_Rate = sum(ifelse(TermReason != "N/A-StillEmployed", 1, 0)) / n() * 100)

# Display termination rates for all_employees dataset
print(all_employees_termination_rates)
```

```
## # A tibble: 6 x 2
##   Department           Termination_Rate
##   <chr>                           <dbl>
## 1 Admin Offices                    22.2
## 2 Executive Office                  0
## 3 IT/IS                            20
## 4 Production                       39.7
## 5 Sales                            16.1
## 6 Software Engineering             36.4
```

According to several online sources, the average annual turnover rate is 47 percent. It looks like rates are good as compared to the average but the Production and Software Engineering departments have significantly higher termination rates than the other departments within this company. Here is another visual to display the differences.

```
# Plot boxplot of Termination_Rate by Department
ggplot(all_employees_termination_rates, aes(x = Department, y = Termination_Rate)) +
  geom_boxplot() +
  labs(title = "Distribution of Termination Rates by Department",
       x = "Department", y = "Termination Rate (%)") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



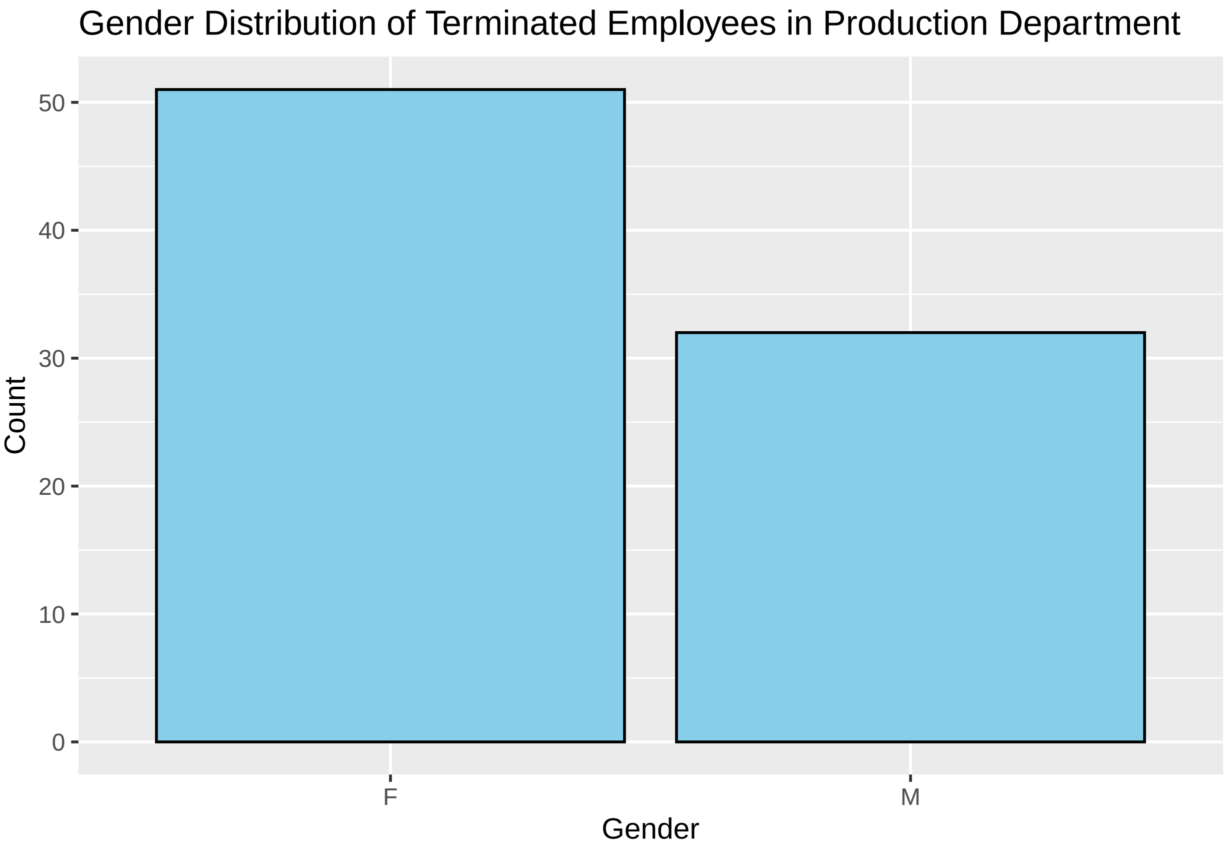
It may be worth it for the company to explore the termination rates in the Production and Software Engineering departments further, depending on the company's turnover rate goals for those departments.

Production Terminations

For now, let's dive deeper into the department with the highest turnover rate, Production.

Let's start by taking a look at the Gender of the terminated employees from the Production department.

```
library(ggplot2)

# Filter terminated employees in the Production department
terminated_production <- termed_employees %>%
  filter(Department == "Production")

# Plot gender distribution
ggplot(terminated_production, aes(x = Sex)) +
  geom_bar(fill = "skyblue", color = "black") +
  labs(title = "Gender Distribution of Terminated Employees in Production Department",
       x = "Gender", y = "Count")
```



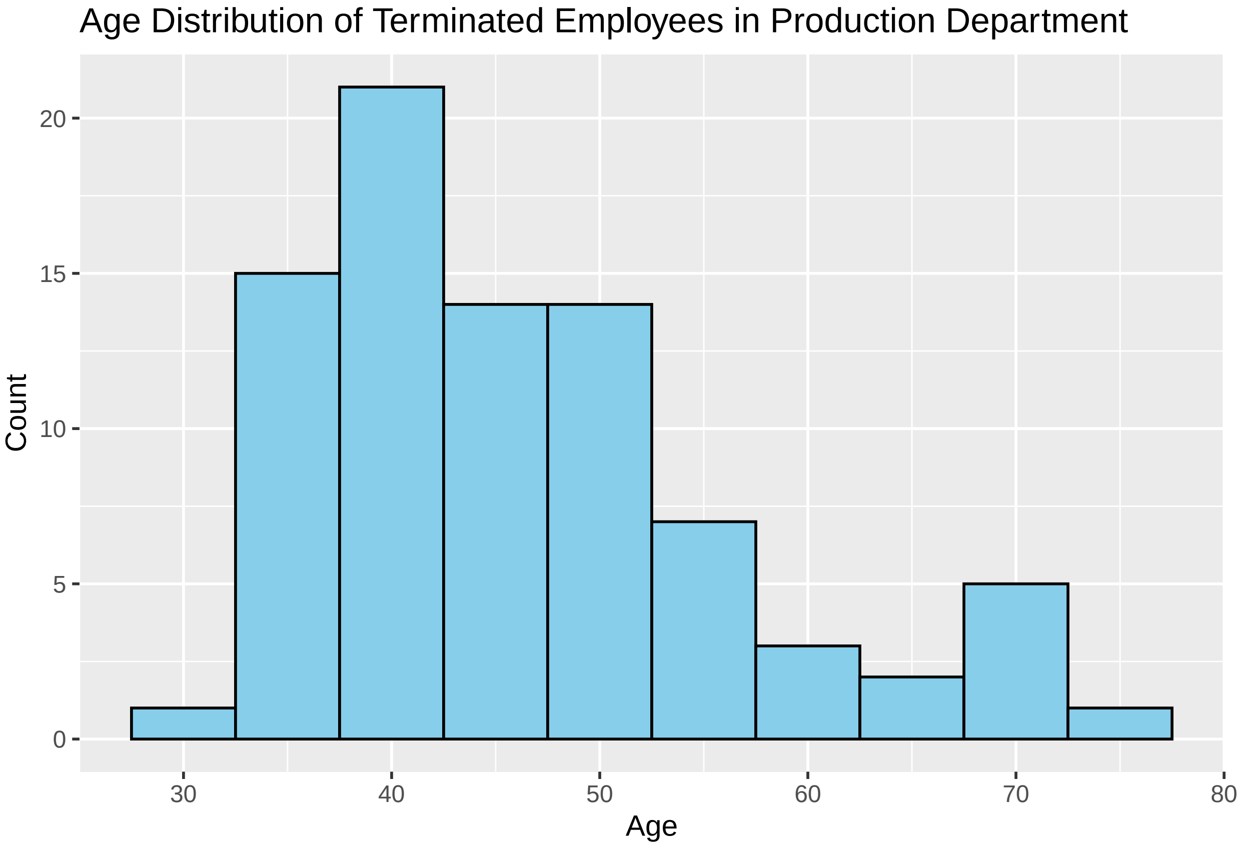
Let's use a chi-square test of independence to determine if there is a significant association between Sex (gender) and EmploymentStatus (terminated or not terminated).

```
library(dplyr)

# Create contingency table for terminated employees by gender in the Production department
terminated_production_table <- terminated_production %>%
  count(Sex)

# Create contingency table for all employees by gender in the Production department
all_production_table <- all_employees %>%
  filter(Department == "Production") %>%
  count(Sex)

# Perform chi-square test of independence
chi_squared_test <- chisq.test(rbind(terminated_production_table$n, all_production_table$n))

# Print the chi-square test result
print(chi_squared_test)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  rbind(terminated_production_table$n, all_production_table$n)
## X-squared = 0.0025016, df = 1, p-value = 0.9601
```

The results of the chi-square test indicate that there is no significant difference in the gender distribution of terminated employees compared to all employees in the Production department. Next, let's see if Age has any affect on terminations in this department. Maybe older employees leave Production more frequently if it is more hands-on work?

```r
library(ggplot2)

# Filter terminated employees in the Production department
terminated_production <- termed_employees %>%
  filter(Department == "Production")

# Plot age distribution
ggplot(terminated_production, aes(x = Age)) +
  geom_histogram(binwidth = 5, fill = "skyblue", color = "black") +
  labs(title = "Age Distribution of Terminated Employees in Production Department",
       x = "Age", y = "Count")
```



Now, let's compare this data to the ages across all employees in the Production department.

```r
# Perform Wilcoxon rank-sum test
wilcox.test(terminated_production$Age, all_employees$Age)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  terminated_production$Age and all_employees$Age
## W = 13972, p-value = 0.2473
## alternative hypothesis: true location shift is not equal to 0
```

Wilcoxon rank sum test indicate that there is no significant difference in the age distributions between terminated employees and all employees in the Production department. Which means, age is not a reason why employees are leaving Production.

We already know that the Production department makes statistically significantly less than the IT/IS and the Engineering departments, but we may want to compare this salary distribution to that of the market rate standards to see if we are on track.

Let's look to see if Employee Satisfaction is an indicator of terminations in the Production department.

```r
# Calculate the average EmpSatisfaction
production_avg_emp_satisfaction <- mean(terminated_production$EmpSatisfaction, na.rm = TRUE)

# Print the average EmpSatisfaction
print(production_avg_emp_satisfaction)
```

```
## [1] 3.855422
```

Let's see if this is statistically significantly different from the mean of the company's overall employee satisfaction score.

```r
# Calculate the mean EmpSatisfaction of all employees
mean_emp_satisfaction_all <- mean(all_employees$EmpSatisfaction, na.rm = TRUE)

# Perform a t-test
t_test <- t.test(terminated_production$EmpSatisfaction, all_employees$EmpSatisfaction)

# Print the t-test result
print(t_test)
```

```
##
##  Welch Two Sample t-test
##
## data:  terminated_production$EmpSatisfaction and all_employees$EmpSatisfaction
## t = -0.32444, df = 133.59, p-value = 0.7461
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.2501684  0.1796613
## sample estimates:
## mean of x mean of y
##  3.855422  3.890675
```

The p-value does not indicate a statistically significant difference in the employee satisfaction scores of terminated Production employees compared to the company overall.

It will be important for the company to research other factors that may be contributing to the high turnover rate in the Production department.

Thank you for reading!