IBM Attrition Dataset Analysis

Tony Kwang Hyun Kim

12/7/2020

If you came here to see something specific please go to the bookmark and . . .

1. For those who want to see how **I would set up and clean data (especially survey data)**, please go to my **Data Wrangling Section.**
2. For those who want to see how **I analyze and communicate data through data visualizations (or see data visualization codes)** please go to my **Exploratory Data Analysis Section.**
3. For those who are interested in the **final results of the analysis** please go to my **Conclusion Section.**

Thoughts on the Dataset & Project

The IBM HR Analytics Employee Attrition & Performance dataset has become one of the most well recognized datasets for those interested in people analytics. Although the dataset is a fictional, it includes various HR metrics commonly collected in various organizations today.

The data is perfect for those interested in practicing data analytics skills; however, it should not be used as a template in organizational settings. My reasons are as follows:

First, it is important to note that the dataset simplifies few metrics and does not provide additional information on how each construct was measured or what each construct means. For example, let's take **job satisfaction** which was measured on a scale of 1 to 4 (1 = low, 2 = medium, 3= high 4 = very high). In the organizational psychology literature, there various scales which measures and defines job satisfaction differently. Without knowing how an organization measures or defines these constructs, it may be difficult to understand why an effect is taking place.

Second, I would like to note how dangerous it can be to utilize some of the metrics included in the dataset for business decisions. Based on your region's labor regulations, you may be exposing you and your organization to potential discrimination charges. I advise you to consult with your legal team before making any decisions.

Overall, I loved playing with the dataset. I think it provides a glimpse of what people analytics can be like. Let me show you how I would've approached the dataset if I was in a real organizational setting.

Introduction

For organizations, turnover is often extremely costly. Resources must be allocated to finding a suitable replacement, and even after finding someone, the organization must invest in the replacement's learning and development. Because turnover comes at a premium, organizational psychologists often use turnover as a way to persuade company leaders to better care for their employees. This is one reason why I believe company culture, engagement, and well-being have recently become such hot topics.

Therefore with this dataset, I will strive to identify the reasons why individuals may be leaving the organizations. Therefore I will attempt the following:

1. **Data Wrangling**
2. **Exploratory Data Analysis**
3. **Provide Recommendations based on Organizational Psychology Literature**

Note that I will not be creating a prediction model. Rather than trying to predict, the goal is to analyze why employees have left.

Side Note Although the data is fictional, I have attempted to treat the data as a real-world dataset.

Loading Libraries & Dataset

Loading Libraries

```
library(readr)
library(ggplot2)
library(ggcorrplot)
library(dplyr)
library(ggthemes)
library(scales)
library(ggthemr)
library(fabricatr)
library(Hmisc)
library(forcats)
knitr::opts_chunk$set(message = FALSE, warning = FALSE, fig.width=8, fig.height =6)
knitr::opts_chunk$set(fig.path = "figures/", dev = c("pdf", "png"))
```

Adjusting Plot Theme

```
swatch_pal <- function() {
  function(n) {
    if(n > length(swatch()) - 1)
      warning(paste("Swatch only has",
        length(swatch())-1), " colours")
    color_list <- as.character(swatch()[2:length(swatch())])
    return(color_list[1:n])
  }
}

scale_colour_ggthemr_d <- function(...) {
  ggplot2::discrete_scale(
              "colour", "ggthemr_swatch_color",
              swatch_pal(),
```

```
            ...
          )
}

scale_color_ggthemr_d <- scale_colour_ggthemr_d
```

Loading Data

```
ibm_data <- read_csv("data/ibm_dataset/WA_Fn-UseC_-HR-Employee-Attrition.csv")
names(ibm_data)
```

```
##  [1] "Age"                      "Attrition"
##  [3] "BusinessTravel"           "DailyRate"
##  [5] "Department"               "DistanceFromHome"
##  [7] "Education"                "EducationField"
##  [9] "EmployeeCount"            "EmployeeNumber"
## [11] "EnvironmentSatisfaction"  "Gender"
## [13] "HourlyRate"               "JobInvolvement"
## [15] "JobLevel"                 "JobRole"
## [17] "JobSatisfaction"          "MaritalStatus"
## [19] "MonthlyIncome"            "MonthlyRate"
## [21] "NumCompaniesWorked"       "Over18"
## [23] "OverTime"                 "PercentSalaryHike"
## [25] "PerformanceRating"        "RelationshipSatisfaction"
## [27] "StandardHours"            "StockOptionLevel"
## [29] "TotalWorkingYears"        "TrainingTimesLastYear"
## [31] "WorkLifeBalance"          "YearsAtCompany"
## [33] "YearsInCurrentRole"       "YearsSinceLastPromotion"
## [35] "YearsWithCurrManager"
```

```
class(ibm_data)
```

```
## [1] "spec_tbl_df" "tbl_df"      "tbl"         "data.frame"
```

```
# setting up second dataset for wrangling and analysis
data1 <- ibm_data
```

Data Wrangling

Before any analysis, it is critical for the data to be cleaned. For anyone interested in what a clean data looks like, I highly recommend reading Hadley Wickham's paper on "Tidy Data" Link.

Overall, the data has already been cleaned in terms of format and follows the principle of "Tidy Data". However, there are additional checks that should be done prior to analysis. This is especially true if there are data collected from surveys. These are some of the key things I checked for:

1. **Missing Data:** are there missing data that we need to account for?
2. **Column Classes:** are all the column classes set to the appropriate class?
3. **Irrelevant Data:** are there any data that we don't need for the analysis?
4. **Data Error Checks:** are there any data errors (e.g. collection, recording, or entry errors) that should be accounted for?

Checking for Missing Data

Counting all NA values within each column.

```
colSums(is.na(data1))
```

```
##                      Age                Attrition           BusinessTravel
##                        0                        0                        0
##                DailyRate               Department         DistanceFromHome
##                        0                        0                        0
##                Education           EducationField            EmployeeCount
##                        0                        0                        0
##           EmployeeNumber  EnvironmentSatisfaction                   Gender
##                        0                        0                        0
##               HourlyRate           JobInvolvement                 JobLevel
##                        0                        0                        0
##                  JobRole          JobSatisfaction            MaritalStatus
##                        0                        0                        0
##            MonthlyIncome              MonthlyRate       NumCompaniesWorked
##                        0                        0                        0
##                   Over18                OverTime        PercentSalaryHike
##                        0                        0                        0
##        PerformanceRating RelationshipSatisfaction            StandardHours
##                        0                        0                        0
##         StockOptionLevel        TotalWorkingYears    TrainingTimesLastYear
##                        0                        0                        0
##          WorkLifeBalance           YearsAtCompany       YearsInCurrentRole
##                        0                        0                        0
##  YearsSinceLastPromotion     YearsWithCurrManager
##                        0                        0
```

Results: No missing data.

Checking and Changing Column Classes

Checking the classes of all the columns.

```
str(data1)
```

```
## tibble [1,470 x 35] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ Age                     : num [1:1470] 41 49 37 33 27 32 59 30 38 36 ...
##  $ Attrition               : chr [1:1470] "Yes" "No" "Yes" "No" ...
##  $ BusinessTravel          : chr [1:1470] "Travel_Rarely" "Travel_Frequently" "Travel_Rarely" "Travel
##  $ DailyRate               : num [1:1470] 1102 279 1373 1392 591 ...
##  $ Department              : chr [1:1470] "Sales" "Research & Development" "Research & Development"
##  $ DistanceFromHome        : num [1:1470] 1 8 2 3 2 2 3 24 23 27 ...
##  $ Education               : num [1:1470] 2 1 2 4 1 2 3 1 3 3 ...
##  $ EducationField          : chr [1:1470] "Life Sciences" "Life Sciences" "Other" "Life Sciences" ..
##  $ EmployeeCount           : num [1:1470] 1 1 1 1 1 1 1 1 1 1 ...
##  $ EmployeeNumber          : num [1:1470] 1 2 4 5 7 8 10 11 12 13 ...
##  $ EnvironmentSatisfaction : num [1:1470] 2 3 4 4 1 4 3 4 4 3 ...
##  $ Gender                  : chr [1:1470] "Female" "Male" "Male" "Female" ...
```

```
##  $ HourlyRate              : num [1:1470] 94 61 92 56 40 79 81 67 44 94 ...
##  $ JobInvolvement          : num [1:1470] 3 2 2 3 3 3 4 3 2 3 ...
##  $ JobLevel                : num [1:1470] 2 2 1 1 1 1 1 1 3 2 ...
##  $ JobRole                 : chr [1:1470] "Sales Executive" "Research Scientist" "Laboratory Technici
##  $ JobSatisfaction         : num [1:1470] 4 2 3 3 2 4 1 3 3 3 ...
##  $ MaritalStatus           : chr [1:1470] "Single" "Married" "Single" "Married" ...
##  $ MonthlyIncome           : num [1:1470] 5993 5130 2090 2909 3468 ...
##  $ MonthlyRate             : num [1:1470] 19479 24907 2396 23159 16632 ...
##  $ NumCompaniesWorked      : num [1:1470] 8 1 6 1 9 0 4 1 0 6 ...
##  $ Over18                  : chr [1:1470] "Y" "Y" "Y" "Y" ...
##  $ OverTime                : chr [1:1470] "Yes" "No" "Yes" "Yes" ...
##  $ PercentSalaryHike       : num [1:1470] 11 23 15 11 12 13 20 22 21 13 ...
##  $ PerformanceRating       : num [1:1470] 3 4 3 3 3 3 4 4 4 3 ...
##  $ RelationshipSatisfaction: num [1:1470] 1 4 2 3 4 3 1 2 2 2 ...
##  $ StandardHours           : num [1:1470] 80 80 80 80 80 80 80 80 80 80 ...
##  $ StockOptionLevel        : num [1:1470] 0 1 0 0 1 0 3 1 0 2 ...
##  $ TotalWorkingYears       : num [1:1470] 8 10 7 8 6 8 12 1 10 17 ...
##  $ TrainingTimesLastYear   : num [1:1470] 0 3 3 3 3 2 3 2 2 3 ...
##  $ WorkLifeBalance         : num [1:1470] 1 3 3 3 3 2 2 3 3 2 ...
##  $ YearsAtCompany          : num [1:1470] 6 10 0 8 2 7 1 1 9 7 ...
##  $ YearsInCurrentRole      : num [1:1470] 4 7 0 7 2 7 0 0 7 7 ...
##  $ YearsSinceLastPromotion : num [1:1470] 0 1 0 3 2 3 0 0 1 7 ...
##  $ YearsWithCurrManager    : num [1:1470] 5 7 0 0 2 6 0 0 8 7 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   Age = col_double(),
##   ..   Attrition = col_character(),
##   ..   BusinessTravel = col_character(),
##   ..   DailyRate = col_double(),
##   ..   Department = col_character(),
##   ..   DistanceFromHome = col_double(),
##   ..   Education = col_double(),
##   ..   EducationField = col_character(),
##   ..   EmployeeCount = col_double(),
##   ..   EmployeeNumber = col_double(),
##   ..   EnvironmentSatisfaction = col_double(),
##   ..   Gender = col_character(),
##   ..   HourlyRate = col_double(),
##   ..   JobInvolvement = col_double(),
##   ..   JobLevel = col_double(),
##   ..   JobRole = col_character(),
##   ..   JobSatisfaction = col_double(),
##   ..   MaritalStatus = col_character(),
##   ..   MonthlyIncome = col_double(),
##   ..   MonthlyRate = col_double(),
##   ..   NumCompaniesWorked = col_double(),
##   ..   Over18 = col_character(),
##   ..   OverTime = col_character(),
##   ..   PercentSalaryHike = col_double(),
##   ..   PerformanceRating = col_double(),
##   ..   RelationshipSatisfaction = col_double(),
##   ..   StandardHours = col_double(),
##   ..   StockOptionLevel = col_double(),
##   ..   TotalWorkingYears = col_double(),
```

```
##   ..   TrainingTimesLastYear = col_double(),
##   ..   WorkLifeBalance = col_double(),
##   ..   YearsAtCompany = col_double(),
##   ..   YearsInCurrentRole = col_double(),
##   ..   YearsSinceLastPromotion = col_double(),
##   ..   YearsWithCurrManager = col_double()
##   .. )
```

I found that all of factor variables are character variables. In addition, utilizing my background knowledge, I know that Education, JobLevel, and StockOptionLevel are also factor variables. However, before changing them, I will use the table function to make sure that they are factor variables

Checking if all of the character columns + Education, JobLevel, and StockOptionLevel are appropriate factor variables.

```
#won't evaluate. Run code if you would like
table(data1$Attrition)
table(data1$BusinessTravel)
table(data1$Department)
table(data1$EducationField)
table(data1$Gender)
table(data1$JobRole)
table(data1$MaritalStatus)
table(data1$Over18)
table(data1$OverTime)
table(data1$Education)
table(data1$JobLevel)
table(data1$StockOptionLevel)
```

I check the table for all of the character class columns because my hunch is that they are all factor (categorical) variables. Checking the table for the columns allows me to evaluate if they are indeed categorical variables (set number of levels)

Based on the table information of each of the character columns(), the character class columns are all factor variables and should be converted to such.

Changing character class columns into factor class.

```
for(i in 1:dim(data1)[2]){
    if(class(data1[[i]]) == "character"){
        data1[[i]] <- as.factor(data1[[i]])
    }
}
```

Changing Education, JobLevel, and StockOptionLevel into factor class

```
data1[["Education"]] <- as.factor(data1[["Education"]])
data1[["JobLevel"]] <- as.factor(data1[["JobLevel"]])
data1[["StockOptionLevel"]] <- as.factor(data1[["StockOptionLevel"]])
```

Checking for Irrelevant Data.

Often times, irrelevant data are peppered into the dataset. I like to remove irrelevant variables to keep my dataset neat and slim as much as possible (stylistic preference). However, if you or the organization are planning to add on to the dataset or foresee this being a long term project, I would recommend not taking any variables out. However, since this is a completed dataset with no plans to add more participants, I will be removing irrelevant variables Additionally, datasets often includes participant identity variables such as names, computer id, etc.; all of which should be removed to insure participant anonymity.

Irrelevancy should be determined with careful consideration and should be discussed with relevant stakeholders

Considering Irrelevancy

```
summary(data1)
```

```
##       Age        Attrition            BusinessTravel   DailyRate
##  Min.   :18.00   No :1233   Non-Travel       : 150   Min.   : 102.0
##  1st Qu.:30.00   Yes: 237   Travel_Frequently: 277   1st Qu.: 465.0
##  Median :36.00              Travel_Rarely    :1043   Median : 802.0
##  Mean   :36.92                                       Mean   : 802.5
##  3rd Qu.:43.00                                       3rd Qu.:1157.0
##  Max.   :60.00                                       Max.   :1499.0
##
##                   Department  DistanceFromHome Education
##  Human Resources       : 63   Min.   : 1.000   1:170
##  Research & Development:961   1st Qu.: 2.000   2:282
##  Sales                 :446   Median : 7.000   3:572
##                               Mean   : 9.193   4:398
##                               3rd Qu.:14.000   5: 48
##                               Max.   :29.000
##
##           EducationField EmployeeCount EmployeeNumber   EnvironmentSatisfaction
##  Human Resources : 27    Min.   :1     Min.   :   1.0   Min.   :1.000
##  Life Sciences   :606    1st Qu.:1     1st Qu.: 491.2   1st Qu.:2.000
##  Marketing       :159    Median :1     Median :1020.5   Median :3.000
##  Medical         :464    Mean   :1     Mean   :1024.9   Mean   :2.722
##  Other           : 82    3rd Qu.:1     3rd Qu.:1555.8   3rd Qu.:4.000
##  Technical Degree:132    Max.   :1     Max.   :2068.0   Max.   :4.000
##
##    Gender      HourlyRate     JobInvolvement JobLevel
##  Female:588   Min.   : 30.00   Min.   :1.00   1:543
##  Male  :882   1st Qu.: 48.00   1st Qu.:2.00   2:534
##               Median : 66.00   Median :3.00   3:218
##               Mean   : 65.89   Mean   :2.73   4:106
##               3rd Qu.: 83.75   3rd Qu.:3.00   5: 69
##               Max.   :100.00   Max.   :4.00
##
##                       JobRole    JobSatisfaction  MaritalStatus MonthlyIncome
##  Sales Executive          :326   Min.   :1.000   Divorced:327   Min.   : 1009
##  Research Scientist       :292   1st Qu.:2.000   Married :673   1st Qu.: 2911
##  Laboratory Technician    :259   Median :3.000   Single  :470   Median : 4919
##  Manufacturing Director   :145   Mean   :2.729                  Mean   : 6503
##  Healthcare Representative:131   3rd Qu.:4.000                  3rd Qu.: 8379
```

```
##  Manager                  :102   Max.   :4.000                  Max.   :19999
##  (Other)                  :215
##   MonthlyRate    NumCompaniesWorked Over18   OverTime   PercentSalaryHike
##  Min.   : 2094   Min.   :0.000      Y:1470   No :1054   Min.   :11.00
##  1st Qu.: 8047   1st Qu.:1.000               Yes: 416   1st Qu.:12.00
##  Median :14236   Median :2.000                          Median :14.00
##  Mean   :14313   Mean   :2.693                          Mean   :15.21
##  3rd Qu.:20462   3rd Qu.:4.000                          3rd Qu.:18.00
##  Max.   :26999   Max.   :9.000                          Max.   :25.00
##
##  PerformanceRating RelationshipSatisfaction StandardHours StockOptionLevel
##  Min.   :3.000     Min.   :1.000            Min.   :80    0:631
##  1st Qu.:3.000     1st Qu.:2.000            1st Qu.:80    1:596
##  Median :3.000     Median :3.000            Median :80    2:158
##  Mean   :3.154     Mean   :2.712            Mean   :80    3: 85
##  3rd Qu.:3.000     3rd Qu.:4.000            3rd Qu.:80
##  Max.   :4.000     Max.   :4.000            Max.   :80
##
##  TotalWorkingYears TrainingTimesLastYear WorkLifeBalance YearsAtCompany
##  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
##
##  YearsInCurrentRole YearsSinceLastPromotion YearsWithCurrManager
##  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
##
```

Take Out:

- EmployeeCount: all partipants are labeled as 1
- StandardHours: all partipants have 80
- Over 18: all participant are over 18

Taking out Irrelevant Data

```
data2 <- subset(data1, select = -c(EmployeeCount, StandardHours, Over18))
```

Checking for Data Error:

When dealing with people metrics some causes of errors are as follows:

1. **Data Collection, Entry, & Recording Errors:** When not using an online survey tool such as Qualtrics, SurveyMonkey, etc., an individual will need to collect, enter, and record the data. Errors

can occur in any of the three steps. Although the best way to inhibit error is to have a checking system before it reaches the data analyst, I can do the following to check:

- Check for extreme outliers. This will check if someone accidentally added or subtracted a number (e.g. putting 100 instead of 10)
- Check for numbers outside the expected range. (e.g. if a survey scale is from 1:4, checking if there are any numbers outside that range.)

2. **Laziness:** When dealing with survey data, it is critical to have a system in place to check for human laziness. For one, participants may "straightline" which is when participants select the same answer for every question. In addition, participants may hastily speed through the survey, providing inaccurate responses. I can do the following to check:

    - If using surveys, add multiple questions to check for attentiveness (e.g. have a question on the survey request the participant select choice #4)
    - Look at the average time it took for people to answer the survey. Remove participants who have finished at an "unreasonable" time span.
    - Check if people chose the same answer for each survey question. I will show a function below.

Checking for Data Collection, Entry & Recording Error

I use the summary function to see if the min and max values of each variable seem to make sense. Here I am checking for any unreasonable min and max values (extreme outliers + numbers outside expected range).

```
summary(data2)
```

Found all of the min and max values of factor variables and survey questions to be within their appropriate range. **However, without knowledge of the company, I could not properly evaluate the other variables.**

Checking for Laziness (Checking for straightlining)

```
#gathering survey data columns
survey_cols <- c(10, 13, 16, 24, 28)
survey_variables <- data2[,survey_cols]

which(apply(survey_variables, 1, function(x) length(unique(x))==1))
```

```
##  [1]  134  158  194  241  297  351  432  472  497  502  507  547  688  729  858
## [16]  932 1104 1108 1127 1142 1285
```

```
sum(apply(survey_variables, 1, function(x) length(unique(x))==1))
```

```
## [1] 21
```

Overall, I found 21 participants to have circled the same response for each. However, I will not remove these participants for the following reasons:

- All of the survey questions should theoretically be highly correlated with each other, meaning it is not unreasonable to see similar answers.
- I do not know the composition of the survey, meaning I do not know if all these questions were next to each other, making it difficult to identify "straightlining."
- There are too few questions for me to comfortably declare straightlining.

Setting up data to use for EDA

```
f_data <- data2
```

Therefore, I will be using f_data (final data) as my data set for EDA & inferential analysis

Exploratory Data Analysis

The purpose of the EDA is to help me understand the data and help me form my hypotheses. At an initial glance of the variables, employees are divided structurally by the following:

- Job Role
- Department
- Job Level.

I will initially take a look at the attrition distribution of those three divisions. Then I will investigate other potential variables affecting attrition. They are as follows:

- Demographic Variables
- Income Variables
- Satisfaction Variables
- Misc.

Attrition Distribution by Organization Structure

Summary: After looking at the attrition distribution as a whole and then by job role, department, and job level here are my following insights:

1. Overall, **16% of the company is leaving**. Please note we don't know how the organization has defined or calculated attrition. Also having comparative metrics such as competitor's attrition rates would help understand the value of this 16%.
2. **Sales reps had the highest within job role attrition rate at 40%.** Laboratory technicians and human resources had the next highest with 24% and 23%. Identifying why these job roles had such high turnover compared to other roles will be important.
3. In terms of departments, the attrition rate within departments did not vary as much seeing how the sales, hr, and r&d had an attrition rate of 21%, 19%, and 14% respectively.

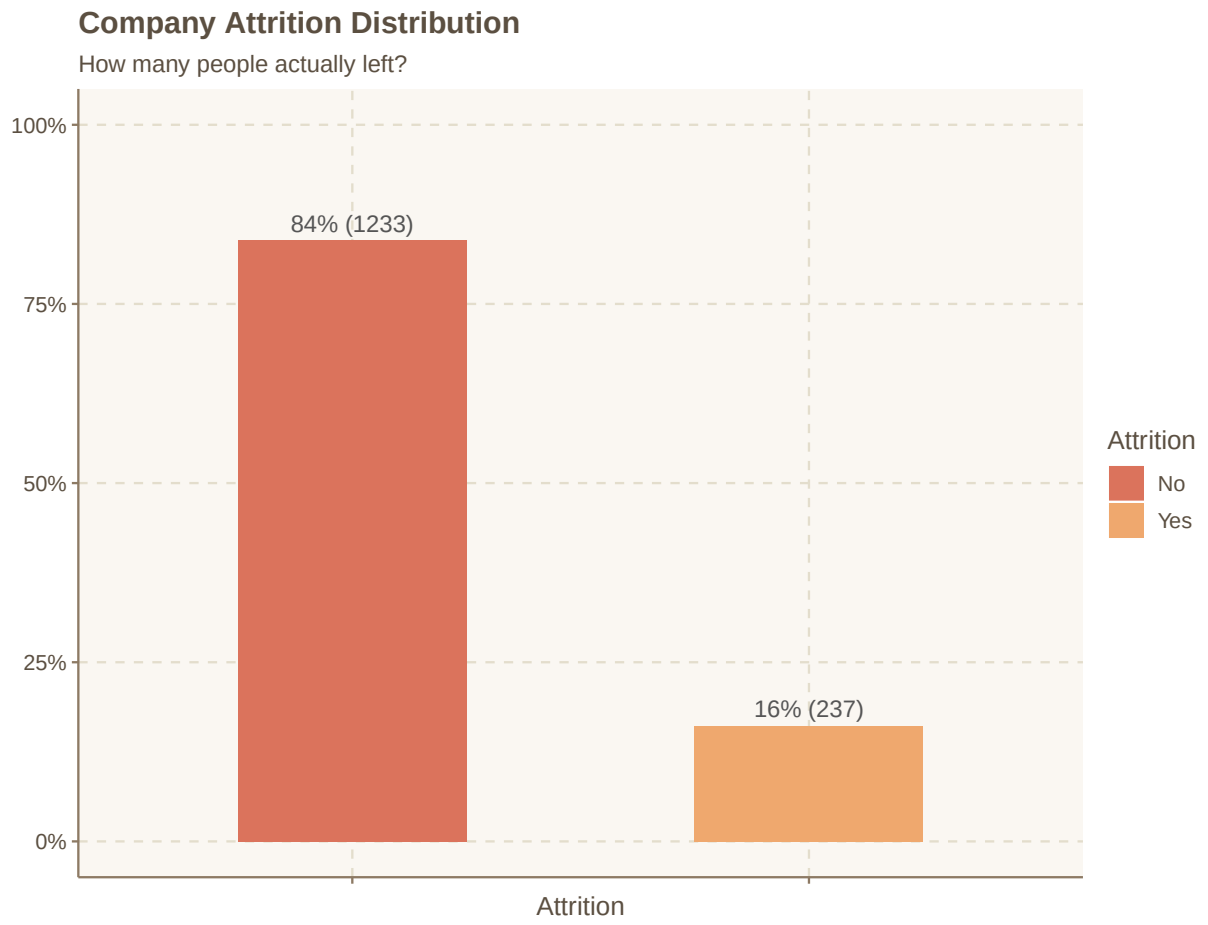
4.**Those in level 1 had the highest within attrition rate at 26%.** However, there was a slight increase in attrition rate from level 2 to level 3 with level 2 having a within attrition rate of 10% with level 3 at 15%.

Final thoughts: The most concerning irregularities comes when looking at individuals from different job roles. Sales reps leaving at 40% attrition is an alarming number. Although less shocking, individuals from level 1 are also leaving at a higher rate of 26%. Identifying the key drivers of attrition especially at the noted job role and levels will be critical. Therefore, once looking at the data, I decided to focus my analysis on job role and level.

Supporting Analysis for Overall Attrition Distribution

Overall Attrition Distribution

```
ggthemr('dust')
f_data %>%
  count(Attrition) %>%
  mutate(pct = prop.table(n)) %>%
  mutate(name = paste(round(pct,2)*100,"%", " (", n, ")", sep = "")) %>%
  ggplot(aes(x = Attrition, y = pct, fill = Attrition)) +
    geom_col(position = 'dodge', width = .5) +
    geom_text(aes(label = name), vjust = -.5) +
    scale_y_continuous(labels = scales::percent, limits = c(0,1)) +
    labs(x = "Attrition", y = "", title = "Company Attrition Distribution",
      subtitle = "How many people actually left?") +
    theme(axis.text.x = element_blank())
```

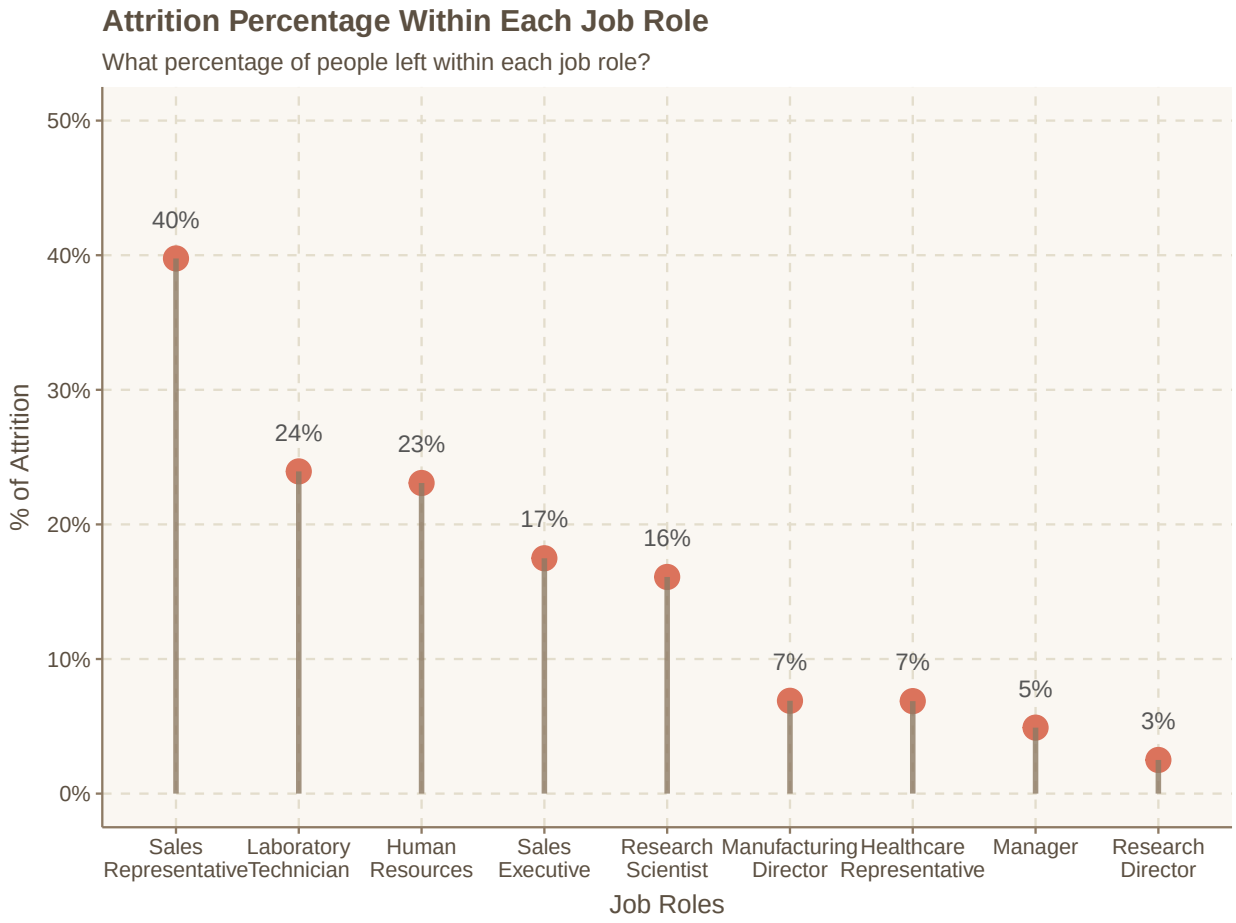


Quick Takeaways:

- Organization had a 16% attrition rate with 237 people having left.
- Imbalance between those who left and those who didn't.

Attrition Within Job Role

```
f_data %>%
  group_by(JobRole) %>%
  count(Attrition) %>%
  mutate(pct = prop.table(n)) %>%
  mutate(name = paste(round(pct,2)*100,"%", sep = "")) %>%
  mutate(JobRole = gsub(" ", "\n", JobRole)) %>%
  subset(Attrition == "Yes") %>%
  ggplot(aes(x = reorder(JobRole, -pct), y = pct)) +
  geom_point(size = 5, aes(y = pct)) +
  geom_segment(aes(x = JobRole, xend= JobRole, y = 0, yend = pct),
    size = 1.2, linetype = 1, alpha = .8, color = "#8d7a64") +
  labs(title = "Attrition Percentage Within Each Job Role",
    subtitle = "What percentage of people left within each job role?",
    x = "Job Roles",
    y = "% of Attrition") +
  geom_text(aes(label = name, x = JobRole, y= pct), vjust = -1.8) +
  scale_y_continuous(labels = scales::percent_format(accuracy = 1), limits = c(0, .5))
```

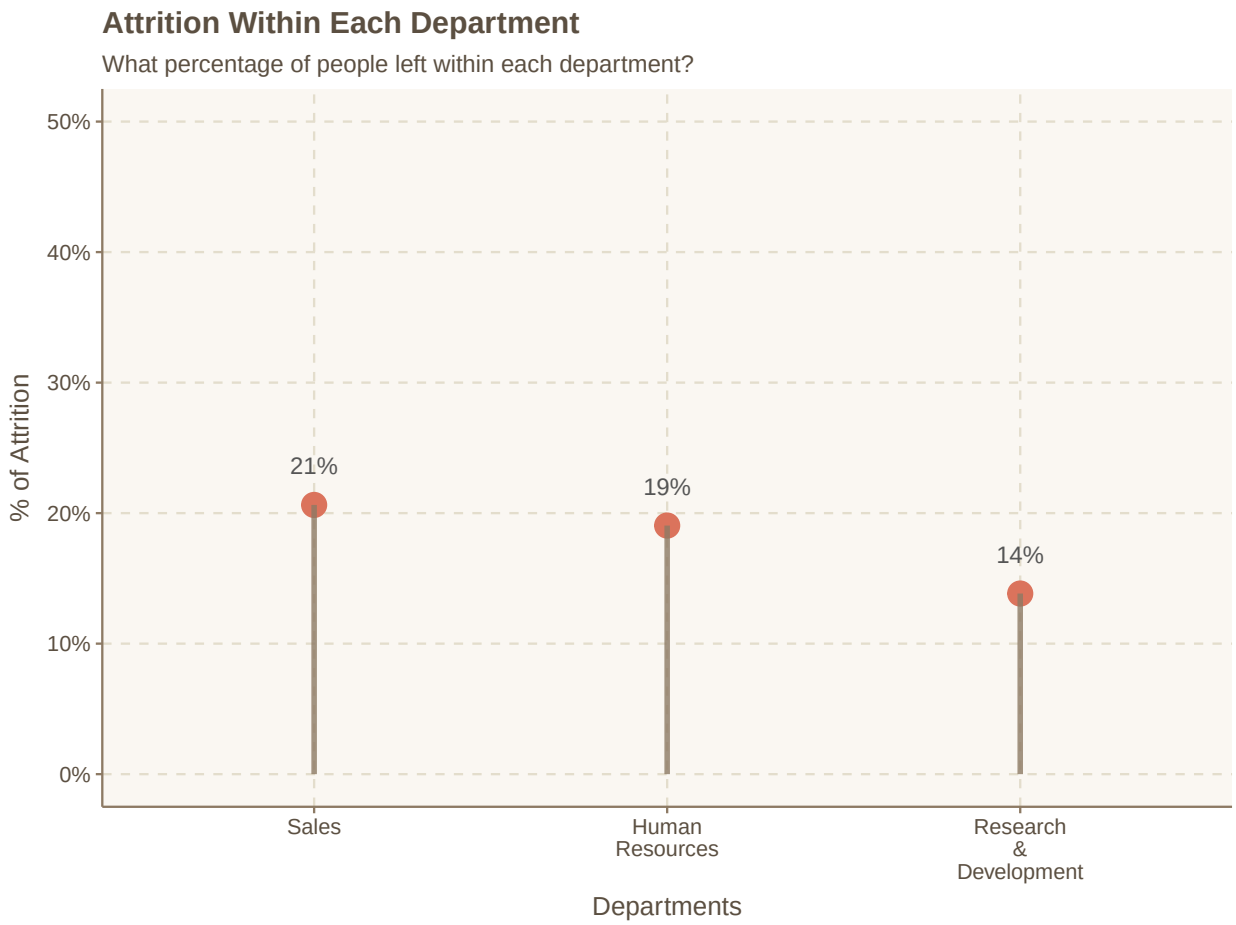


Quick Takeaways:

- Sales representatives had the highest percentage of people leaving while research director and managers had the lowest percentage.

Attrition Within Each Department

```
f_data %>%
  group_by(Department) %>%
  count(Attrition) %>%
  mutate(pct = prop.table(n)) %>%
  mutate(name = paste(round(pct,2)*100,"%", sep = "")) %>%
  mutate(Department = gsub(" ", "\n", Department)) %>%
  subset(Attrition == "Yes") %>%
  ggplot(aes(x = reorder(Department, -pct), y = pct)) +
  geom_point(size = 5, aes(y = pct)) +
  geom_segment(aes(x = Department, xend= Department, y = 0, yend = pct),
    size = 1.2, linetype = 1, alpha = .8, color = "#8d7a64") +
  labs(title = "Attrition Within Each Department",
    subtitle = "What percentage of people left within each department?",
    x = "Departments",
    y = "% of Attrition") +
  geom_text(aes(label = name, x = Department, y= pct), vjust = -1.8) +
  scale_y_continuous(labels = scales::percent_format(accuracy = 1), limits = c(0, .5))
```



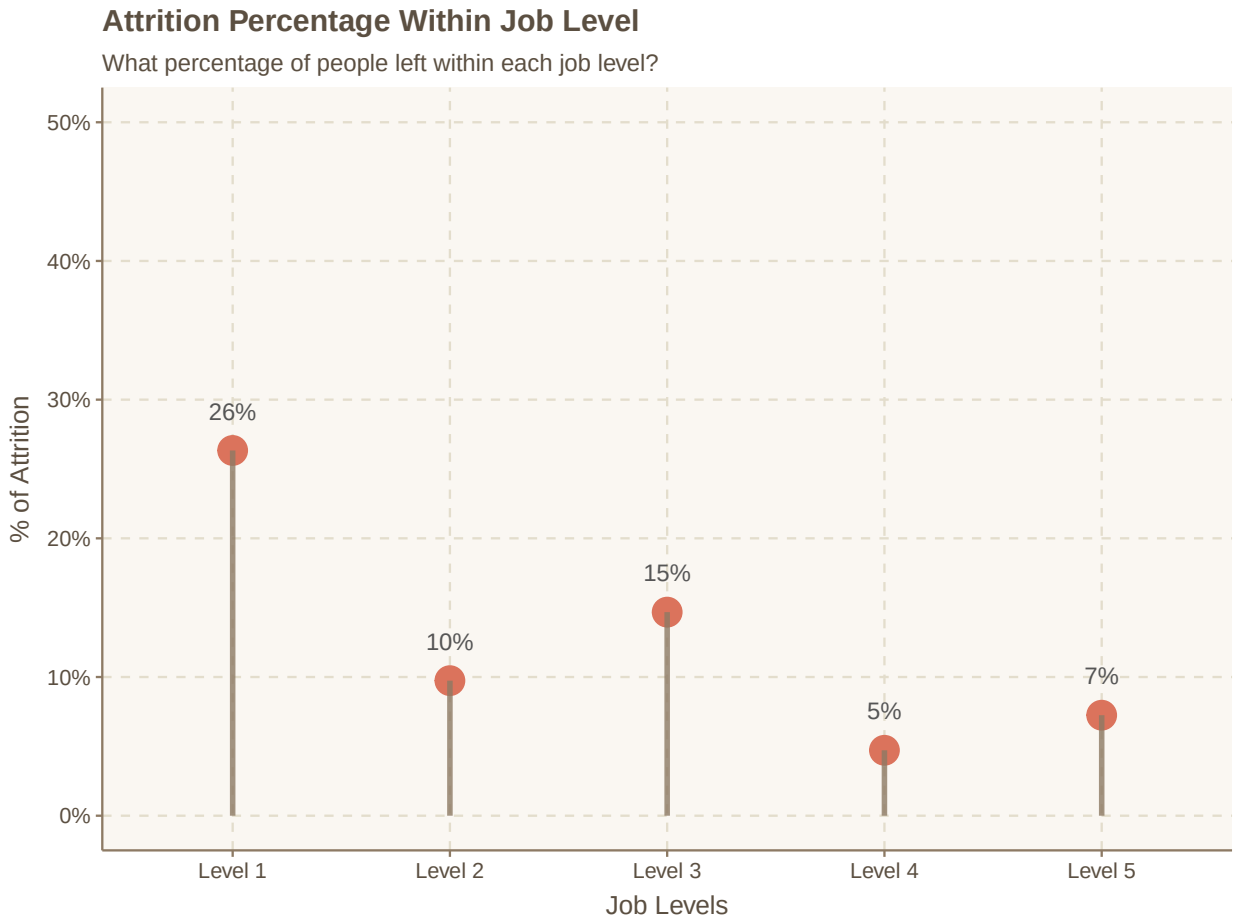
Quick Takeaways:

- Unlike job roles, each department didn't differ as much in terms of the percentage of people leaving. This leads me to think that other factors rather than departmental issues are more likely to affect attrition

Attrition Within Job Level

```
f_data %>%
  group_by(JobLevel) %>%
  count(Attrition) %>%
  mutate(pct = prop.table(n)) %>%
  mutate(name = paste(round(pct,2)*100,"%", sep = "")) %>%
  mutate(JobLevel = paste("Level", JobLevel)) %>%
  subset(Attrition == "Yes") %>%
  ggplot(aes(x = JobLevel, y = pct)) +
  geom_point(size = 6, aes(y = pct)) +
  geom_segment(aes(x = JobLevel, xend= JobLevel, y = 0, yend = pct),
    size = 1.2, linetype = 1, alpha = .8, color = "#8d7a64") +
  labs(title = "Attrition Percentage Within Job Level",
    subtitle = "What percentage of people left within each job level?",
    x = "Job Levels",
    y = "% of Attrition") +
```

```
geom_text(aes(label = name, x = JobLevel, y= pct), vjust = -1.8) +
scale_y_continuous(labels = scales::percent_format(accuracy = 1), limits = c(0, .5))
```



Quick Takeaways:

- As expected people from the lowest level are leaving than those in higher levels. However, it may be worth while to investigate why there was a slight jump in attrition from level 2 to level 3.

Demographics Analysis

Therefore, I will evaluate if people in different **age** and **gender** are leaving at different rates. I will ask the following:

- Are people in certain gender or age groups leaving at higher rates?
- Are people in certain gender or age groups in a particular job role or job level leaving at higher rates?

Demographics Analysis 1: Age Summary: After looking at the attrition distribution separated by age groups and then age groups within job roles, departments, and job levels here are my following insights:

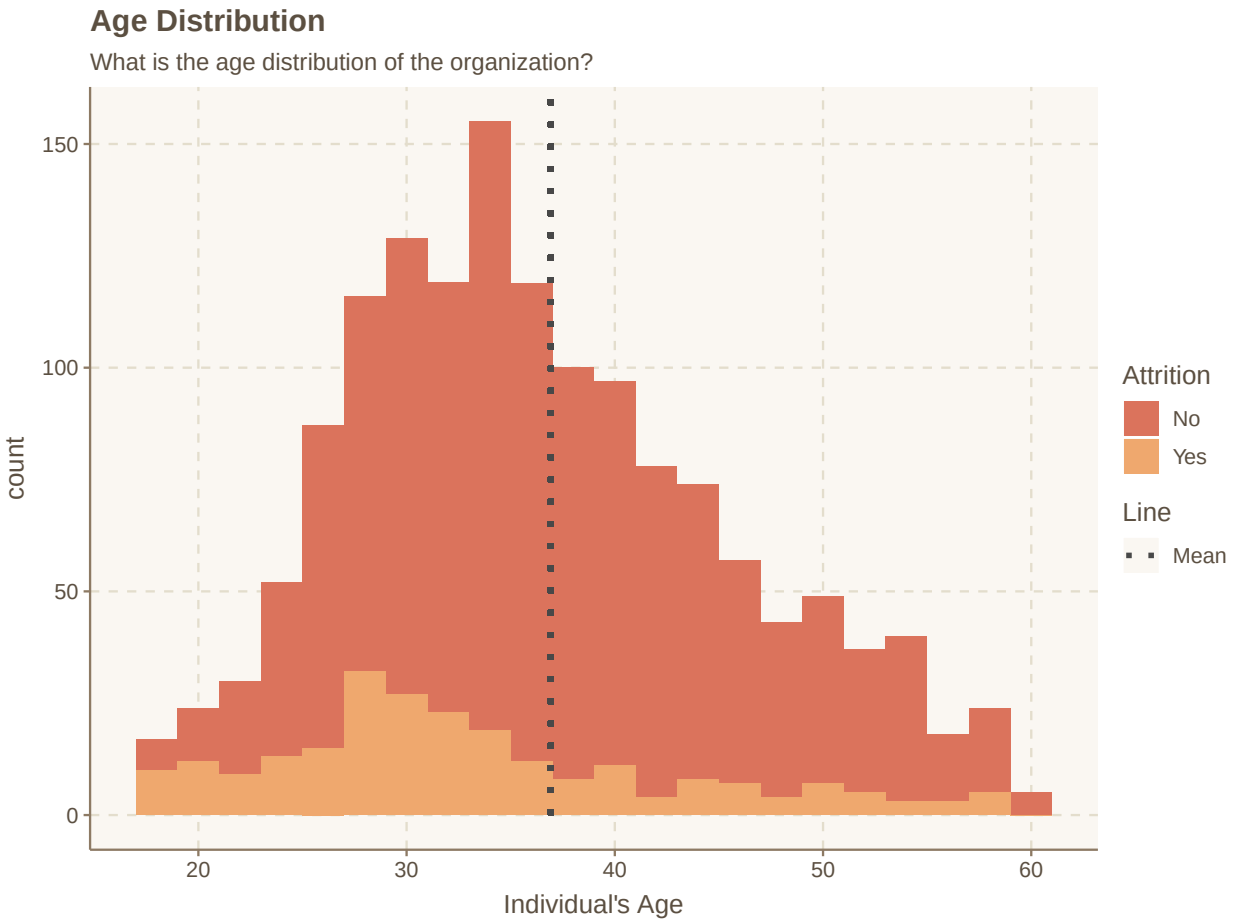
1. The average employee age is 39.
2. **Approximately 55% of all of the attrition occurred between individuals ages 18-32** This seems to indicate that the company is struggling to maintain younger talent than older talent.
3. The company lost **36% of employees within the 18-25 bracket** while losing **22% of all employees within the 26-32 bracket.**
4. The loss of young talent was **specially bad for those who are the sales representatives role and in for those in level 1.**

Final thoughts: The most concerning irregularities comes when looking at the **loss of young talent.** It may be possible that there is a culture or policy unfavorable to younger individuals. Further analysis on why younger talents are leaving will be evaluated when looking at income and satisfaction ratings. This high attrition rate may be fine if those who are leaving are lower performers. However, it is a serious issue if high performers are leaving as well. Unfortunately due to the way the performance ratings were done, it would be nearly impossible to tell (everyone scored either a 3 or 4 on performance rating which means not enough variability exists).

Supporting Analysis for Demgraphic Analysis 1: Age

Age Distribution

```
ggthemr("dust")
f_data %>%
  ggplot(aes(x = Age, fill = Attrition)) +
  geom_histogram(binwidth = 2) +
  geom_segment(aes(x = mean(Age), y = 0, xend = mean(Age), yend = Inf, linetype = "Mean"), col = "#4848
  labs(x = "Individual's Age", y = "count", title = "Age Distribution", subtitle = "What is the age dis
  scale_linetype_manual(name = "Line", values = c("Mean" = 3)) +
  guides(fill = guide_legend(order = 1), linetype = guide_legend(order = 2))
```



Quick Takeaways:

- The average age of the employees is approximately 39.
- Overall, the organization has a large number of employees that are in their late twenties to mid 40s.
- Although not definitive, there does seem to be a higher percentage of those in their 20s leaving the organization.

Creating Age Brackets

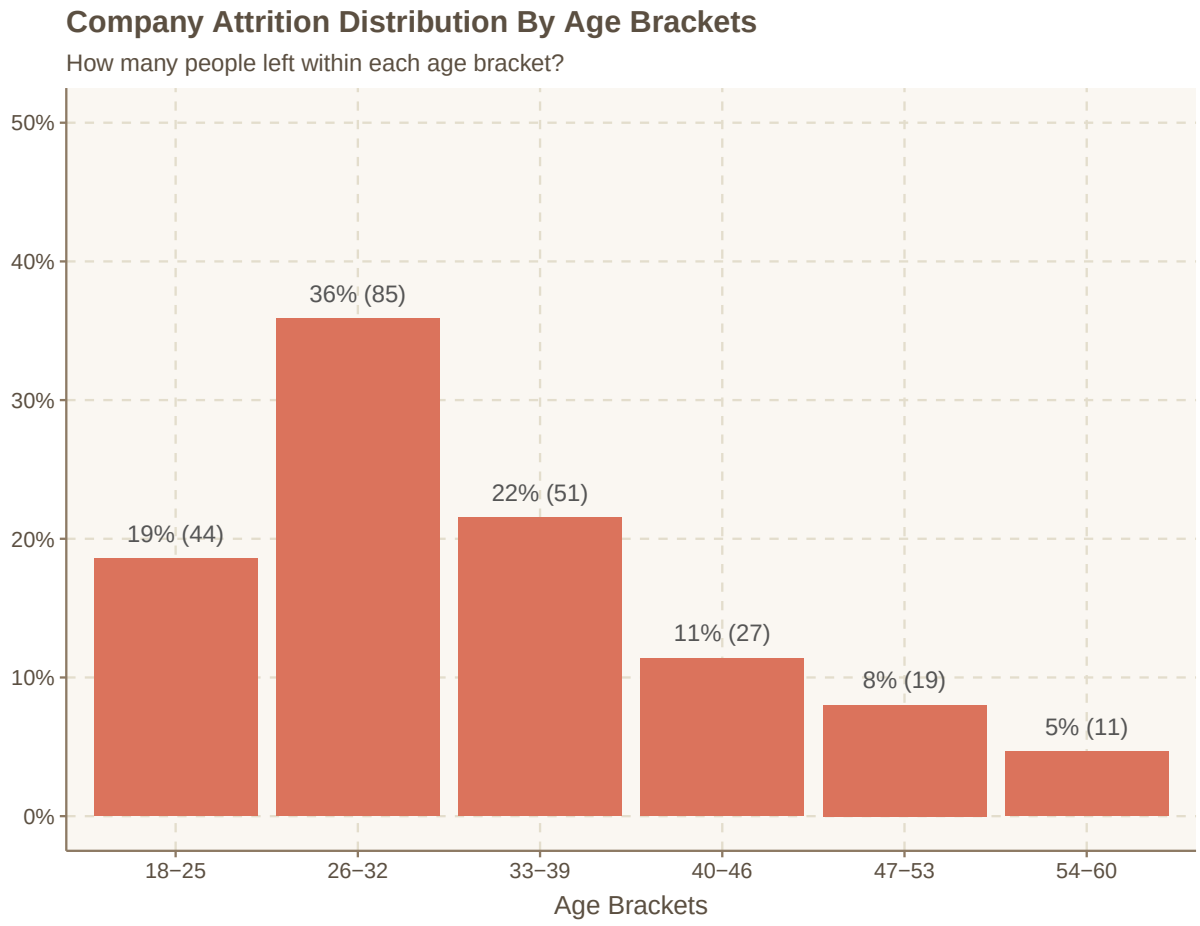
```
AgeQ <- split_quantile(f_data$Age, type = 4)
AgeQ <- as.factor(cut(f_data$Age, breaks = 6,
  labels = c("18-25", "26-32", "33-39", "40-46", "47-53", "54-60")))
table(AgeQ)
```

```
## AgeQ
## 18-25 26-32 33-39 40-46 47-53 54-60
##   123   393   432   282   153    87
```

Attrition by Age Brackets

```
f_data %>%
  select(Age, Attrition) %>%
```

```r
  mutate(AgeQ = AgeQ) %>%
  filter(Attrition == "Yes") %>%
  group_by(AgeQ) %>%
  summarise(count = n()) %>%
  mutate(pct = prop.table(count),
    label= paste0(round(pct*100,0), "%"," (", count, ")", sep = "")) %>%
  ggplot(aes(x = AgeQ, y = pct)) +
  geom_bar(stat = "identity") +
  geom_text(aes(x = AgeQ, y = pct, label = label),
    vjust = -1) +
  scale_y_continuous(labels = scales::percent_format(accuracy = 1), limits = c(0,.5)) +
  labs(title = "Company Attrition Distribution By Age Brackets",
    subtitle = "How many people left within each age bracket?",
    x = "Age Brackets",
    y = "")
```

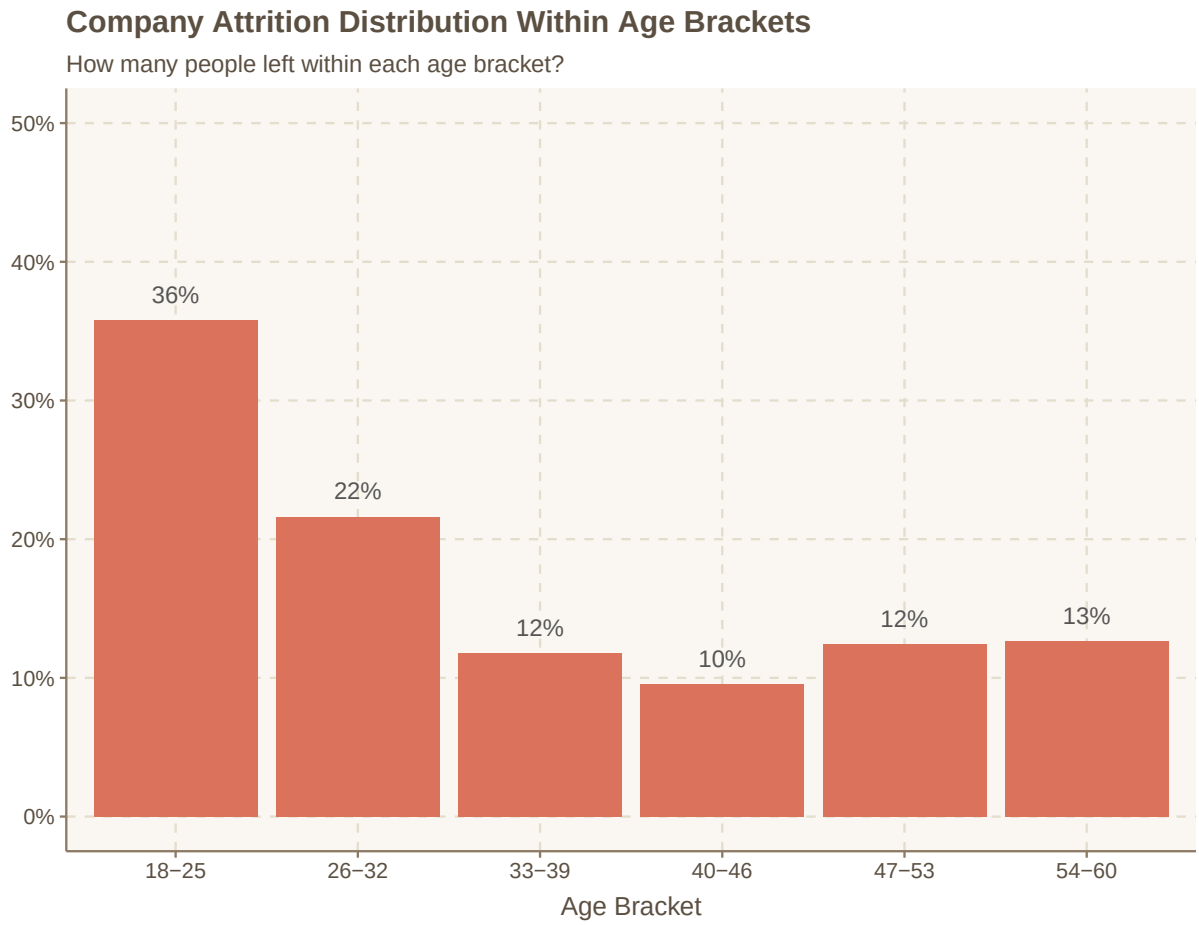


Quick Takeaways:

- Approximate 55% of all attrition comes from younger employees between 18-32. Overall, it seems as though the company is struggling to maintain younger talent verses older talent.
- The organization seems to be doing well in keeping employees between the age 33-39 considering only a 22% of attrition for the organization's largest group. The 33-39 age bracket contains their largest number of employees with 432 employees approximately 29.4% of the entire workforce.

Attrition Within Each Age Brackets

```
f_data %>%
  select(Age, Attrition) %>%
  mutate(AgeQ = AgeQ) %>%
  group_by(AgeQ, Attrition) %>%
  summarise(count = n()) %>%
  group_by(AgeQ) %>%
  mutate(pct = prop.table(count),
    label= paste0(round(pct*100,0), "%", sep = "")) %>%
  filter(Attrition == "Yes") %>%
  ggplot(aes(x = AgeQ, y = pct)) +
  geom_bar(stat = "identity") +
  geom_text(aes(x = AgeQ, y = pct, label = label),
    vjust = -1) +
  scale_y_continuous(labels = scales::percent_format(accuracy = 1),
    limits = c(0,.5)) +
  labs(title = "Company Attrition Distribution Within Age Brackets",
    subtitle = "How many people left within each age bracket?",
    x = "Age Bracket",
    y = "")
```

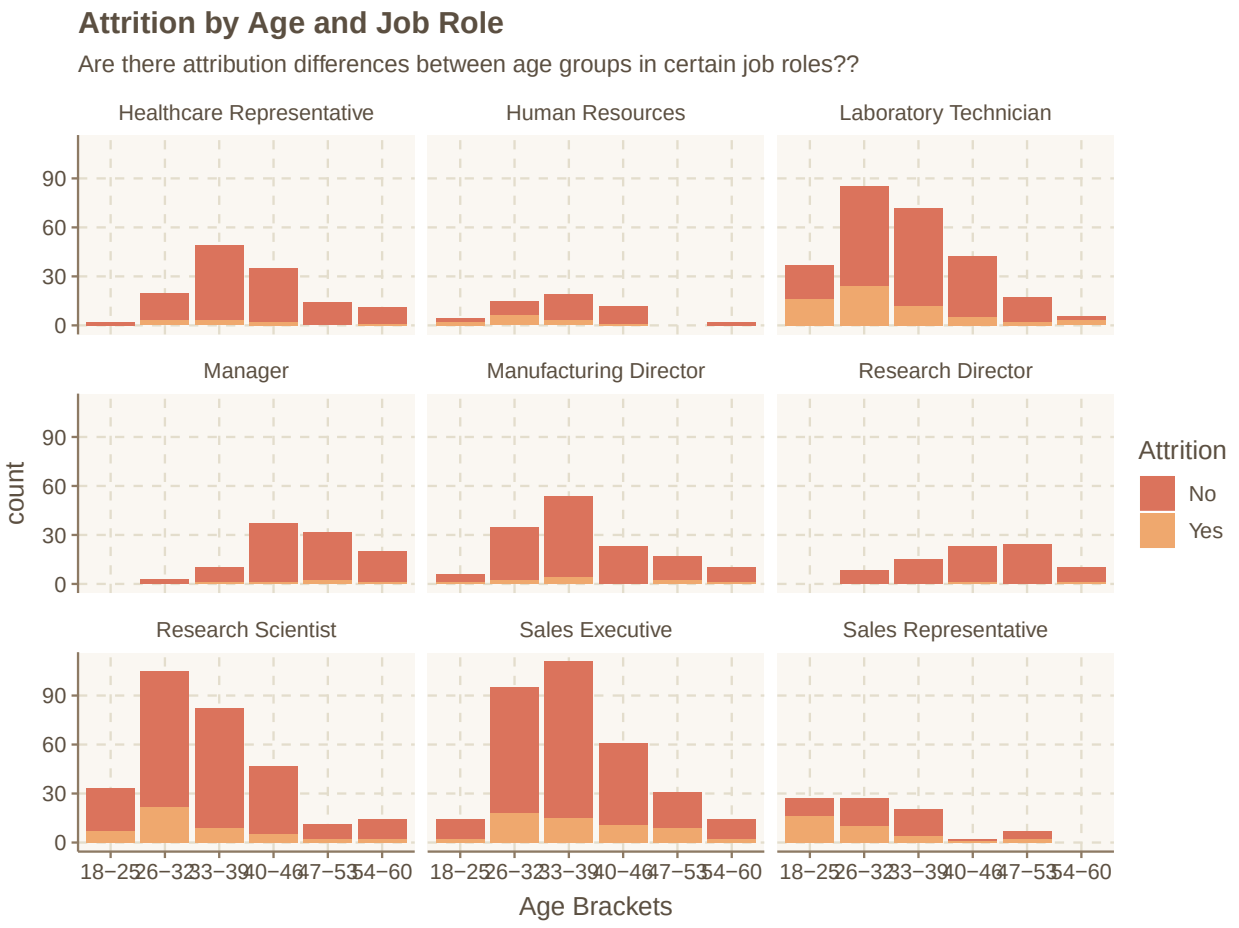


Quick Takeaways:

- Overall, 36% of all employees in the 18-25 age bracket left in the given year. That percentage drops to 22% when we go down to the next age bracket.
- Looking at the attrition rates by age bracket and within age bracket, the story continues to indicate that younger people are leaving at a higher rate.

Attrition Within Each Age Brackets

```
f_data %>%
  select(Age, JobRole, Department, JobLevel, Attrition) %>%
  mutate(AgeQ = AgeQ) %>%
  ggplot(aes(x = AgeQ, fill = Attrition)) +
  geom_bar() +
  facet_wrap(vars(JobRole)) +
  labs(title = "Attrition by Age and Job Role",
    subtitle = "Are there attribution differences between age groups in certain job roles??",
    x = "Age Brackets")
```

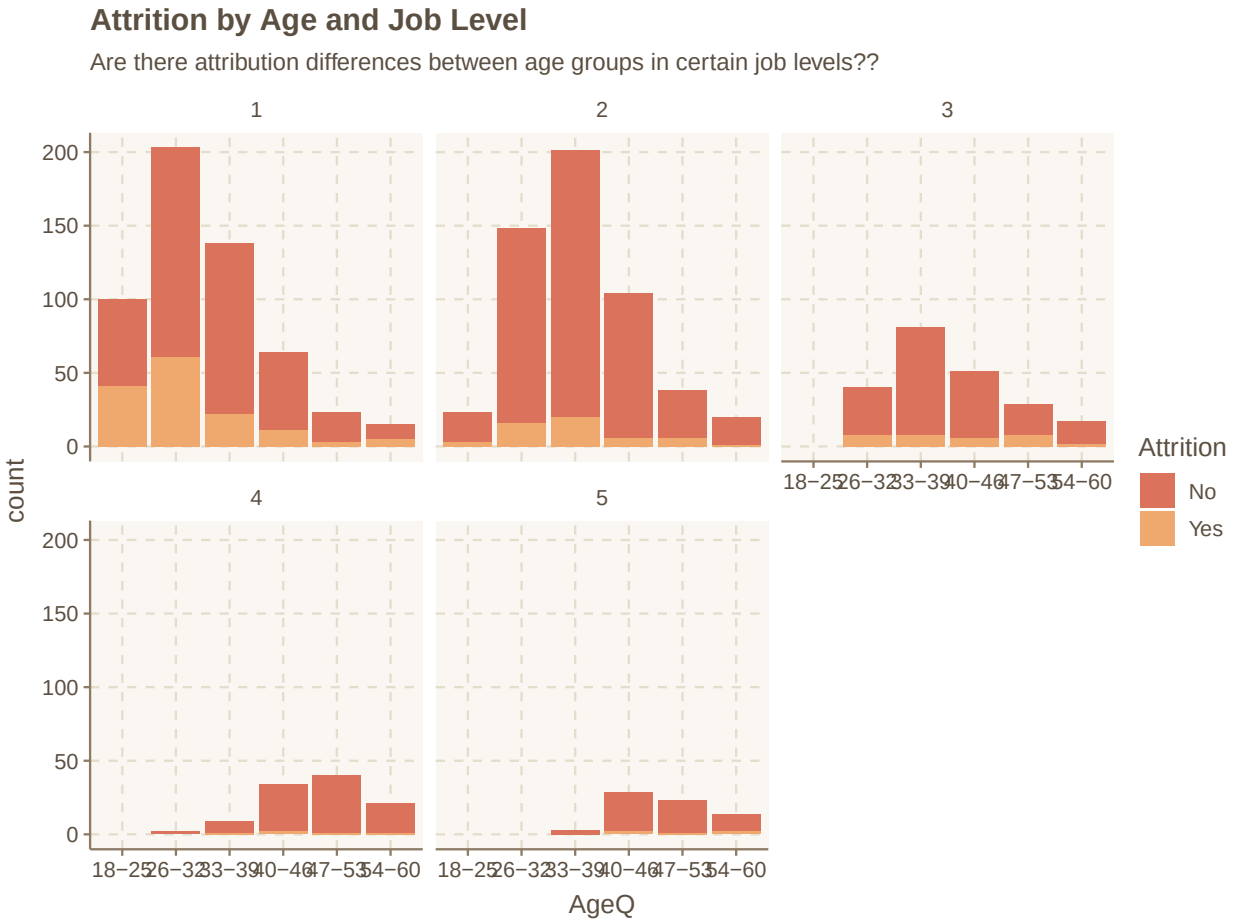


Quick Takeaways:

- When we look at attrition by age and job role, things seem relatively normal. People are leaving proportionally to the number of people within the age bracket and job role.
- More than 50% of sales rep between the ages of 18-25 are leaving.

Attrition Within Each Age Brackets and Job Level

```r
f_data %>%
  select(Age, JobLevel, Attrition) %>%
  mutate(AgeQ = AgeQ) %>%
  ggplot(aes(x = AgeQ, fill = Attrition)) +
  geom_bar() +
  facet_wrap(vars(JobLevel)) +
  labs(title = "Attrition by Age and Job Level",
    subtitle = "Are there attribution differences between age groups in certain job levels??")
```



Quick Takeaways:

- Individuals in the 18-25 age bracket + level 1 are leaving at a higher rate than those in the 26-32 age bracket + level 1.

Demographics Analysis #2: Gender

Summary: After looking at the attrition distribution separated by **gender** and then gender within job roles and job levels here are my following insights:

1. **Males within the organization seems to be leaving at a higher rate than female (17% to 14.8%)**

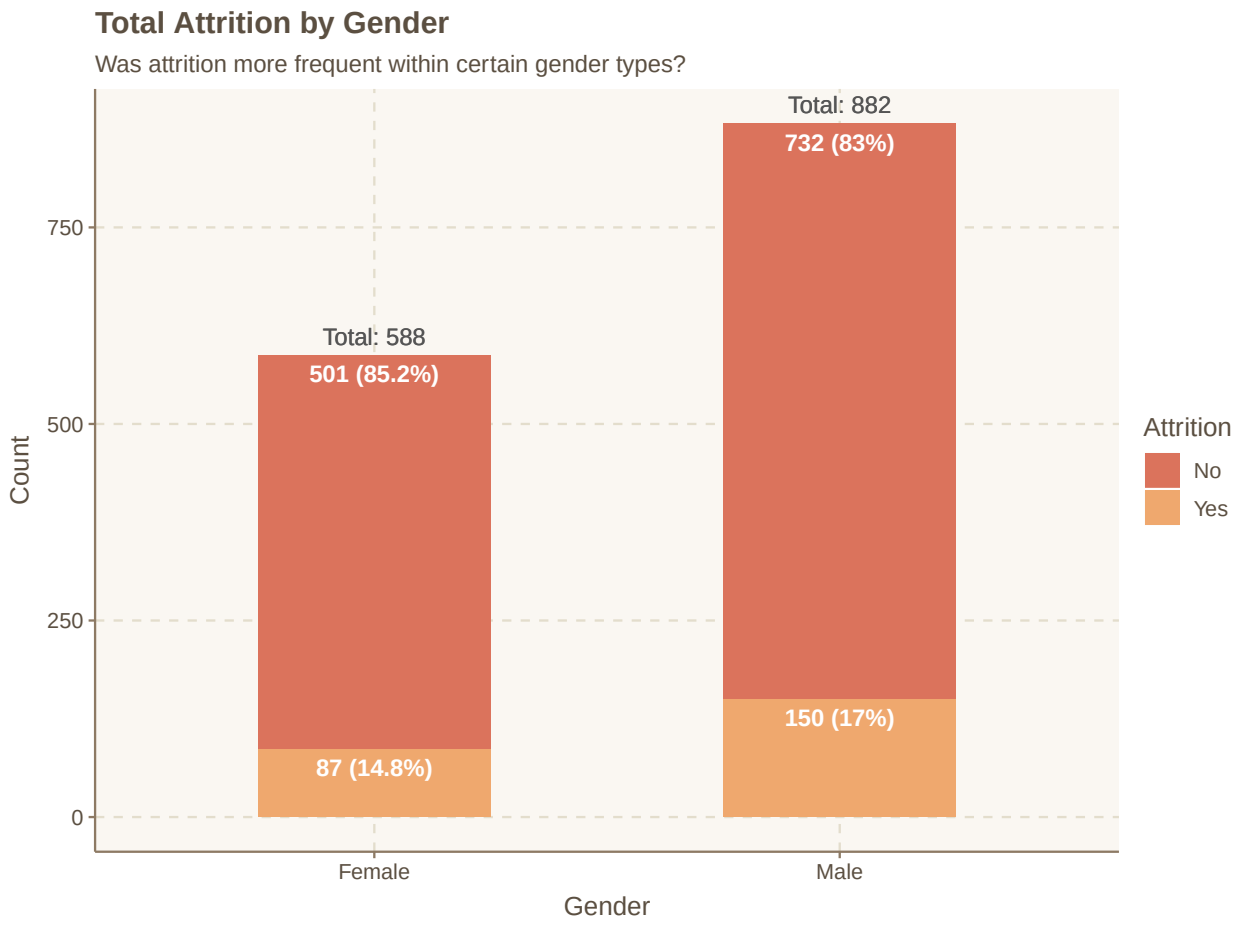
2. Overall, attrition rates between the two gender seem to be proportional within each job role and level.

Final Thoughts: Although not conclusive, the exploratory analysis does not seem to suggest that there were any gender differences in terms of attrition. However, further analysis may be required to ascertain the claim.

Supporting Analysis for Demographics Analysis 2: Gender

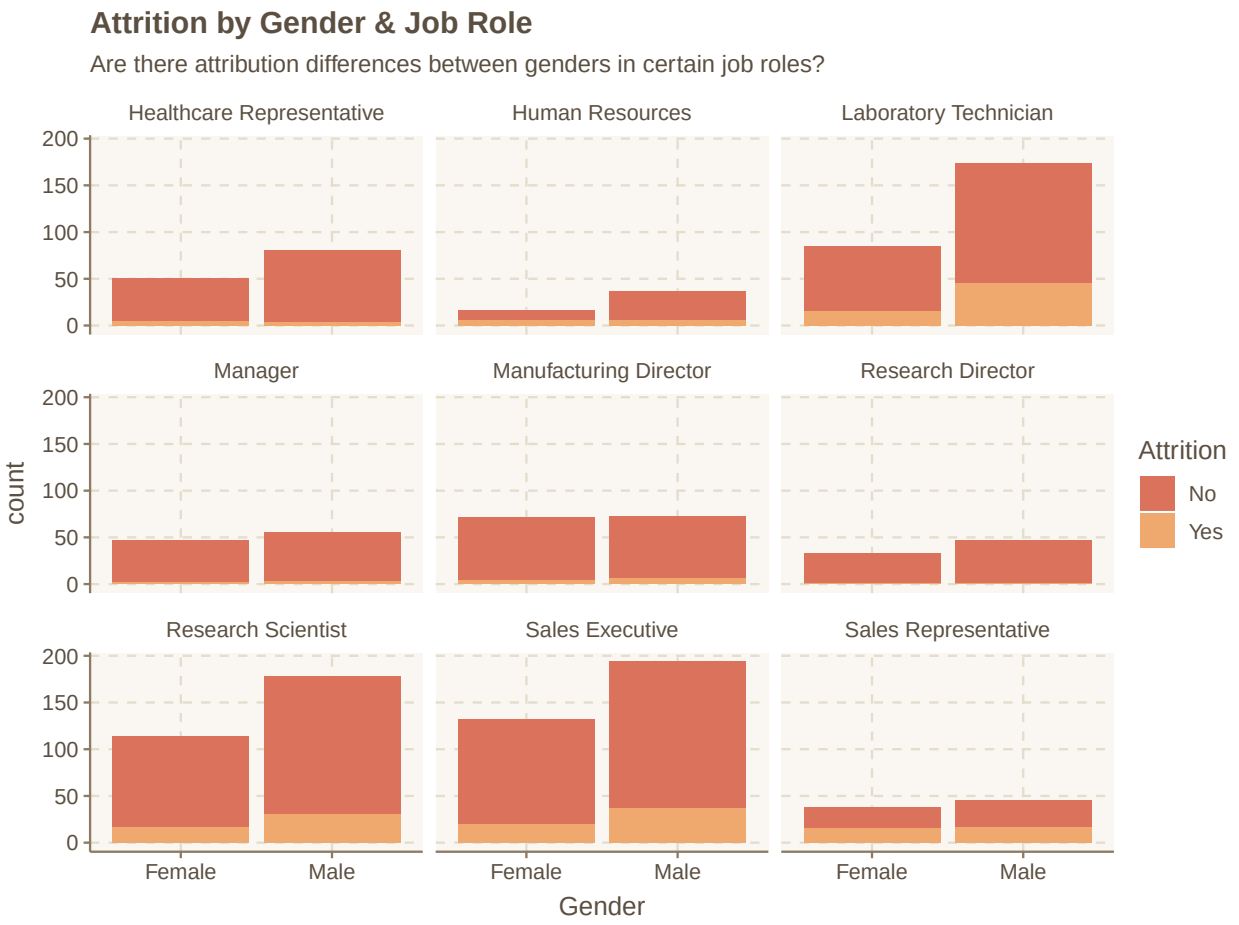
Attrition Distribution Between Gender

```
f_data %>%
  select(Attrition, Gender) %>%
  group_by(Attrition, Gender) %>%
  summarise(Count = n()) %>%
  group_by(Gender) %>%
  arrange(Count) %>%
  mutate(Count_total = cumsum(Count),
    CountPer = prop.table(Count),
    final = paste(Count, " (", round(CountPer*100,1), "%)", sep = ""),
    Totals = paste("Total:", sum(Count)),
    Top = sum(Count)) %>%
  ggplot(aes(x = Gender, y = Count, fill = Attrition)) +
  geom_bar(stat = "identity", width = .5) +
  geom_text(aes(label = final, x = Gender, y = Count_total),
    vjust = 1.6, color = "white", fontface = "bold") +
  geom_text(aes(label = Totals, x = Gender, y = Top), vjust = -.6) +
  labs(title = "Total Attrition by Gender",
    subtitle = "Was attrition more frequent within certain gender types?")
```



- **Quick Takeaways:**
  - There are 882 male employees to 588 female employees
  - There was a lower percentage of females leaving the organization than males.

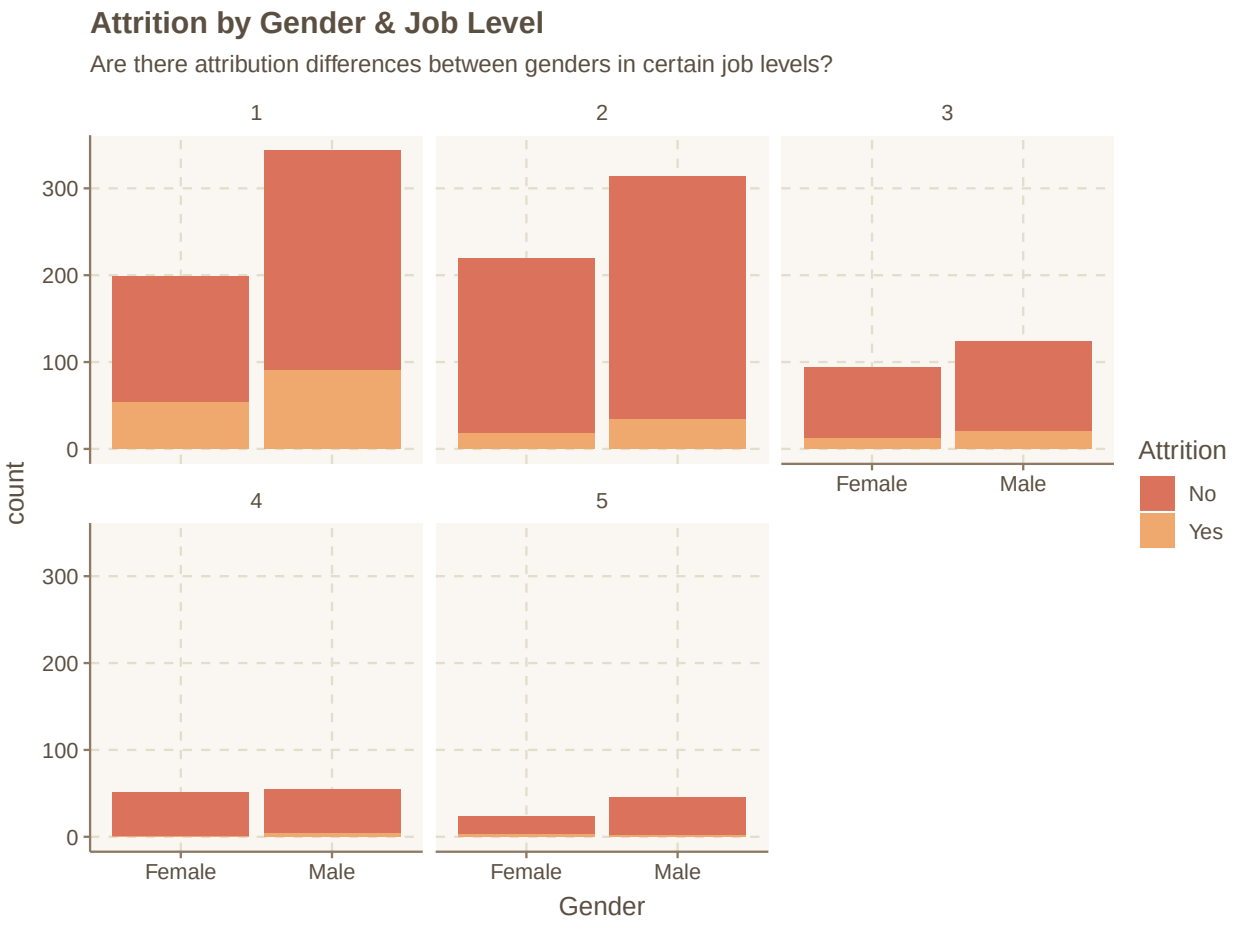
```
f_data %>%
  ggplot(aes(x = Gender,  fill = Attrition)) +
  geom_bar()  +
  facet_wrap(vars(JobRole)) +
  labs(title = "Attrition by Gender & Job Role",
    subtitle = "Are there attribution differences between genders in certain job roles?")
```



Quick Takeaways:

- Overall, relatively proportional attrition between genders with all the job role.

```
f_data %>%
  ggplot(aes(x = Gender,  fill = Attrition)) +
  geom_bar()  +
  facet_wrap(vars(JobLevel)) +
  labs(title = "Attrition by Gender & Job Level",
    subtitle = "Are there attribution differences between genders in certain job levels?")
```



Quick Takeaways:

- Overall, the attrition rate between gender seem relatively well distributed between job levels.

Potential Factor: Income Differences

Hypotheses to Test:

1. Individuals are leaving the organization due to low income.
2. Individuals are leaving the organization due to low income in comparison to their peers within their job role, department, or level. This may be caused by a low perception of organizational fairness.
3. Individuals who earn more are more satisfied with their jobs.

Summary:

1. After comparing the income of those who stayed and those who left, **there seems to be support to claim that individuals are leaving due to low income.** The median monthly income of all those who stayed was 5204 while the median monthly income of those who left was 3202.
2. **The first point is further supported when analyzing income between job roles.** Sales reps, research scientists, and laboratory technicians have the lowest income. These roles also have the highest attrition rates at 40%, 24%, and 16%. This shows that attrition may be correlated with low income.

3. **When looking within job roles and job levels, there was less conclusive support that individuals were leaving due to income disparities among their peers.** Although income disparities between ex and current employees within departments were high, the effect was most likely driven by income differences between job roles.
4. **Income does not seem to be a strong driver of job satisfaction.**

Final thoughts: The analysis has shown that **income seems to be a strong driver of attrition especially for low paying job roles.** Sales, which had the highest attrition rate was one of the least paid jobs. The analysis has also shown that **income disparities between peers is most likely not a primary reason for attrition.** It may be relevant to see if job roles such as sales reps, research scientists and laboratory technicians should have their income levels be reevaluated and readjusted based on industry and location standards.

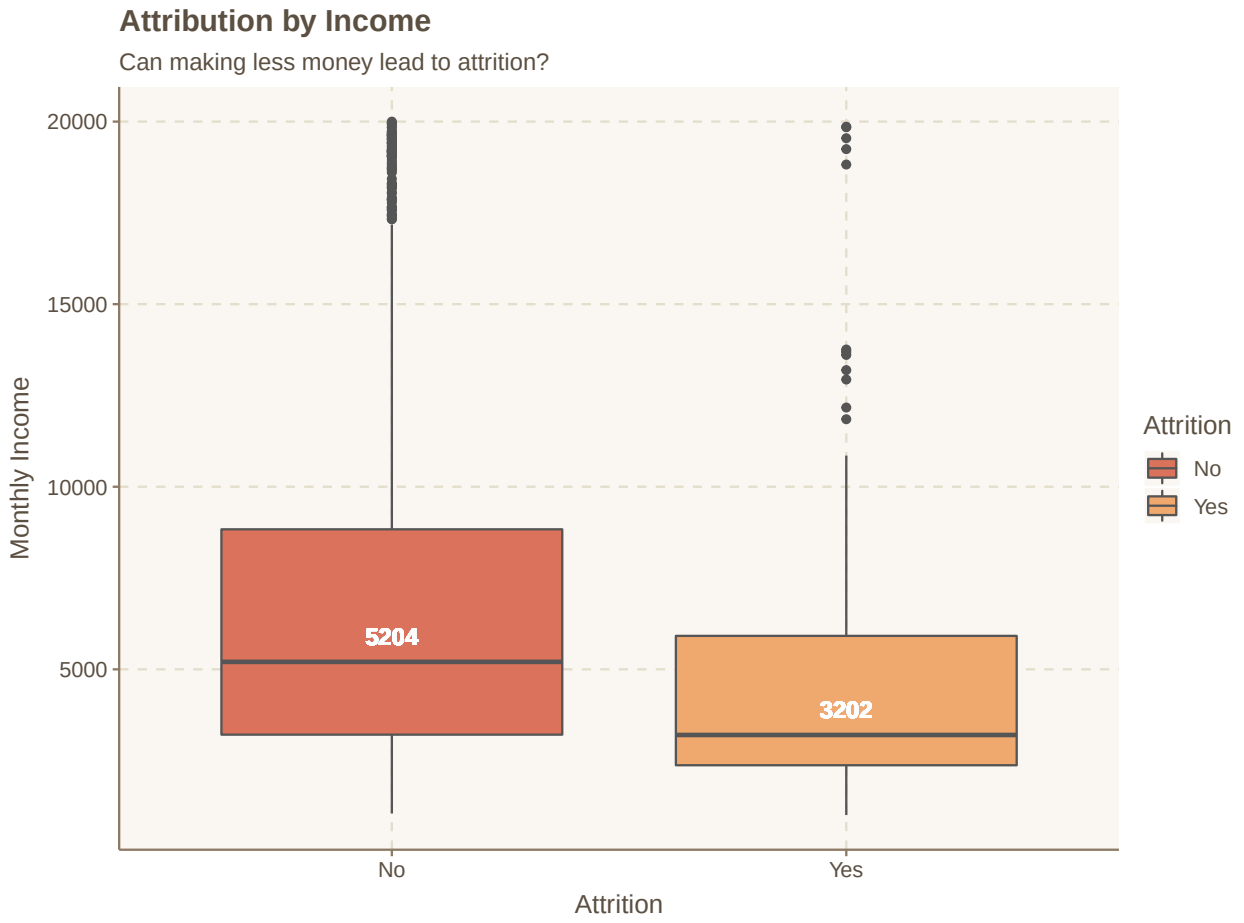
Supporting Analysis For Income Difference

Monthly Income and Attrition

```
ggthemr('dust')

medianvalues <- f_data %>%
  group_by(Attrition) %>%
  mutate(medians = median(MonthlyIncome))

f_data %>%
  ggplot(aes(x = Attrition, y = MonthlyIncome)) +
  geom_boxplot(aes(fill = Attrition)) +
  geom_text(data = medianvalues, aes(x = Attrition, y = medians, label = medians),
    color = "white", fontface = "bold",
    vjust = -1) +
  labs( x = "Attrition", y = "Monthly Income", title = "Attribution by Income", subtitle = "Can making
```

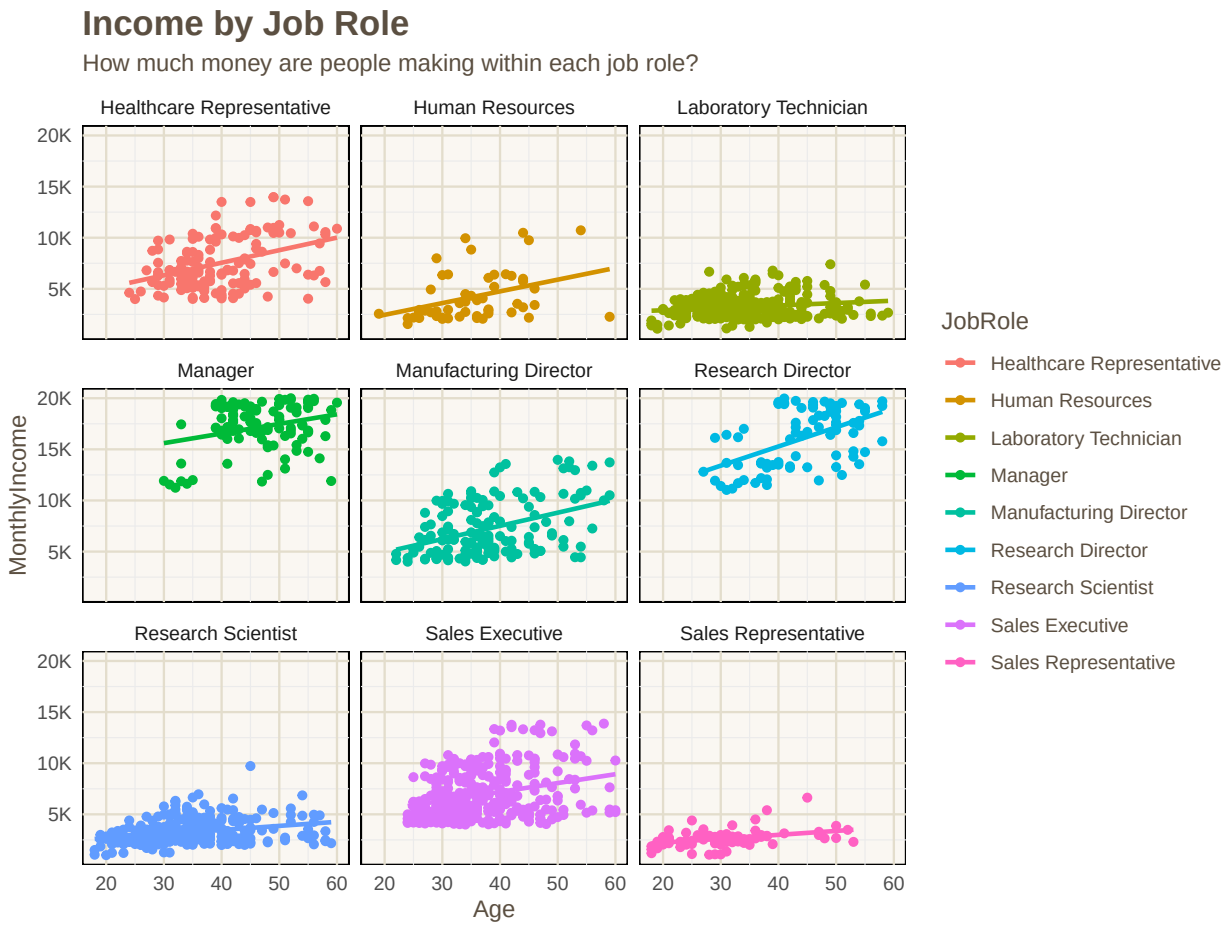


Quick Takeaways:

- There is a noticeable income difference between people who stayed (median = 5204) and left (median = 3202).

Income by Job Role

```
ggthemr_reset()
f_data %>%
  ggplot(aes(x = Age , y = MonthlyIncome, col = JobRole)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Income by Job Role", subtitle = "How much money are people making within each job role?
  scale_y_continuous(breaks = seq(0, 20000, 5000),
                     labels = paste0(as.character(seq(0,20,5)),"K")) +
  facet_wrap(vars(JobRole)) +
  theme_minimal() +
  theme(text = element_text( color = '#5b4f41'),
    plot.title = element_text(size = 16, face = "bold"),
    panel.background = element_rect(fill = "#FAF7F2"),
    panel.grid.major = element_line(colour = "#E3DDCC"))
```

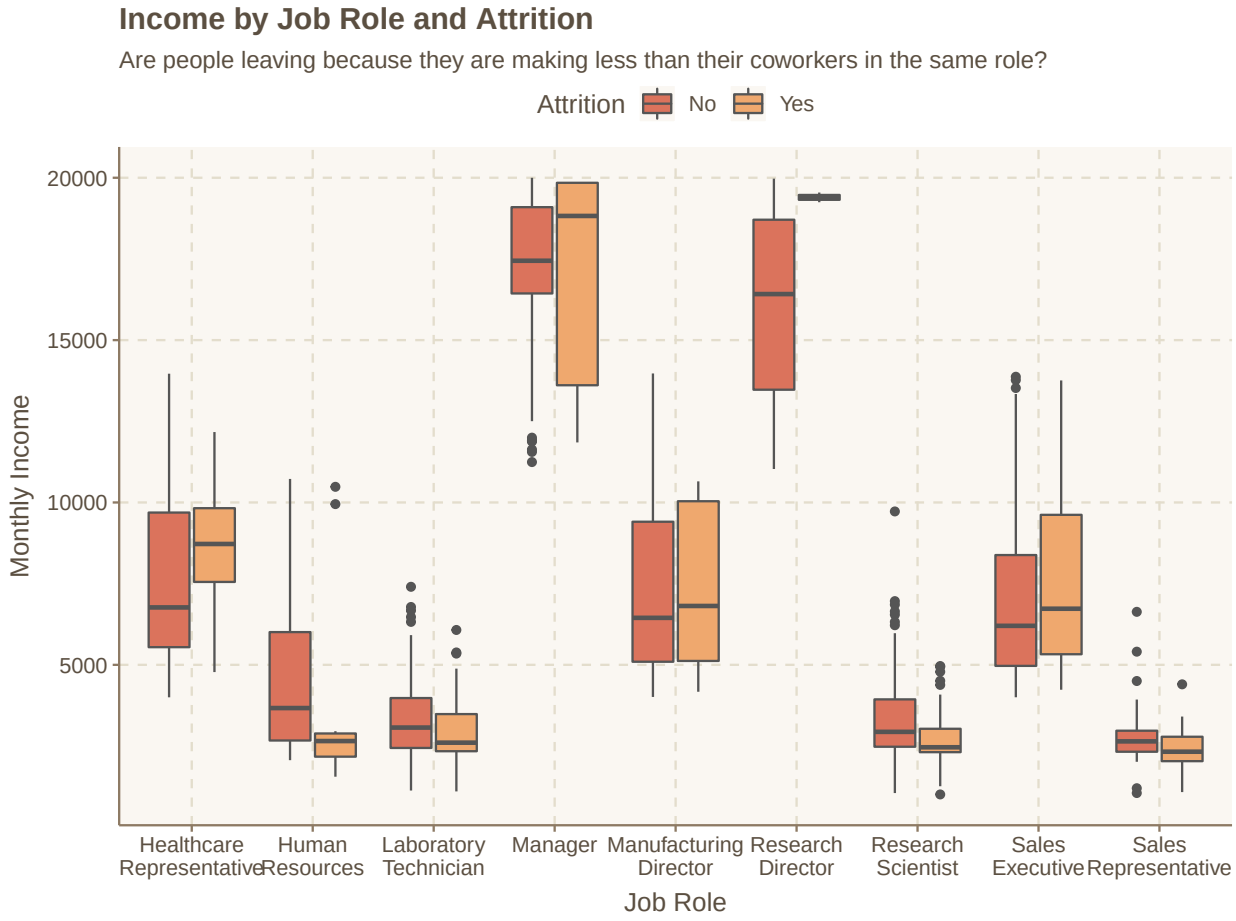


Quick Takeaways:

- Managers and Research Directors make the most income. The two positions also have the lowest attrition rate at 5% and 3% respectively.
- Sales reps, laboratory technicians, and research scientists have the lowest income. These three positions have high attrition rates at 40%, 24%, and 16% each.
- This seems to support my hypothesis that lower income leads to

Income Differences by Job Role and Attrition

```
ggthemr('dust')
f_data %>%
  mutate(JobRole = gsub(" ", "\n", JobRole)) %>%
  ggplot(aes(x = JobRole, y = MonthlyIncome, fill = Attrition)) +
  geom_boxplot() +
  theme(legend.position = "top") +
  labs( x = "Job Role", y = "Monthly Income", title = "Income by Job Role and Attrition", subtitle = "A
```

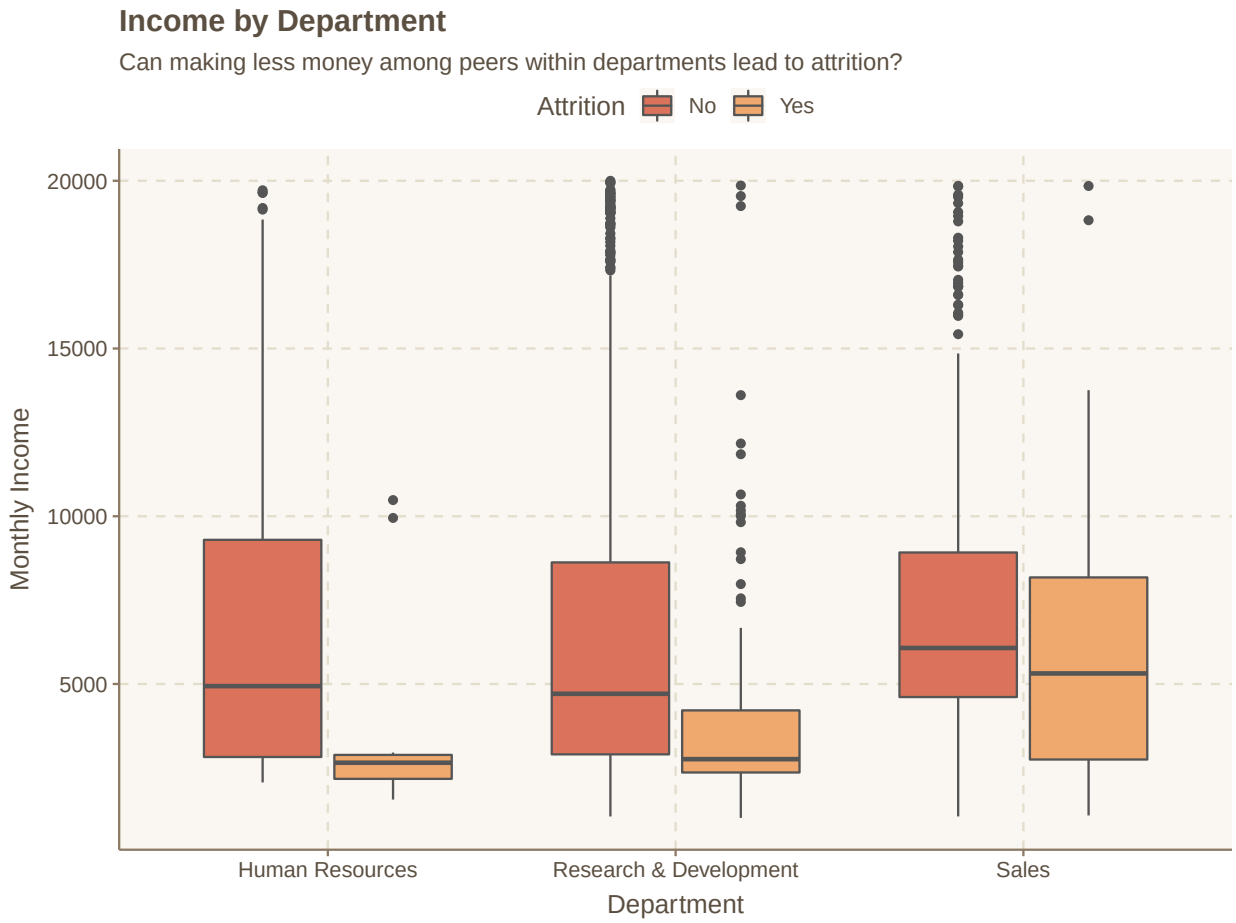


Quick Takeaways:

- Overall, there does not seem to be strong evidence supporting the claim that people are leaving due to income differences within roles.
- In fact, there are many instances (healthcare reps, managers, manufacturing director, research director, and sales executives) where the attrition group has a higher median income than the group that stayed.

Income Differences by Department and Attrition

```
ggthemr('dust')
f_data %>%
  ggplot(aes(x = Department, y = MonthlyIncome, fill = Attrition)) +
  geom_boxplot() +
  theme(legend.position = "top") +
  labs( x = "Department", y = "Monthly Income", title = "Income by Department", subtitle = "Can making
```

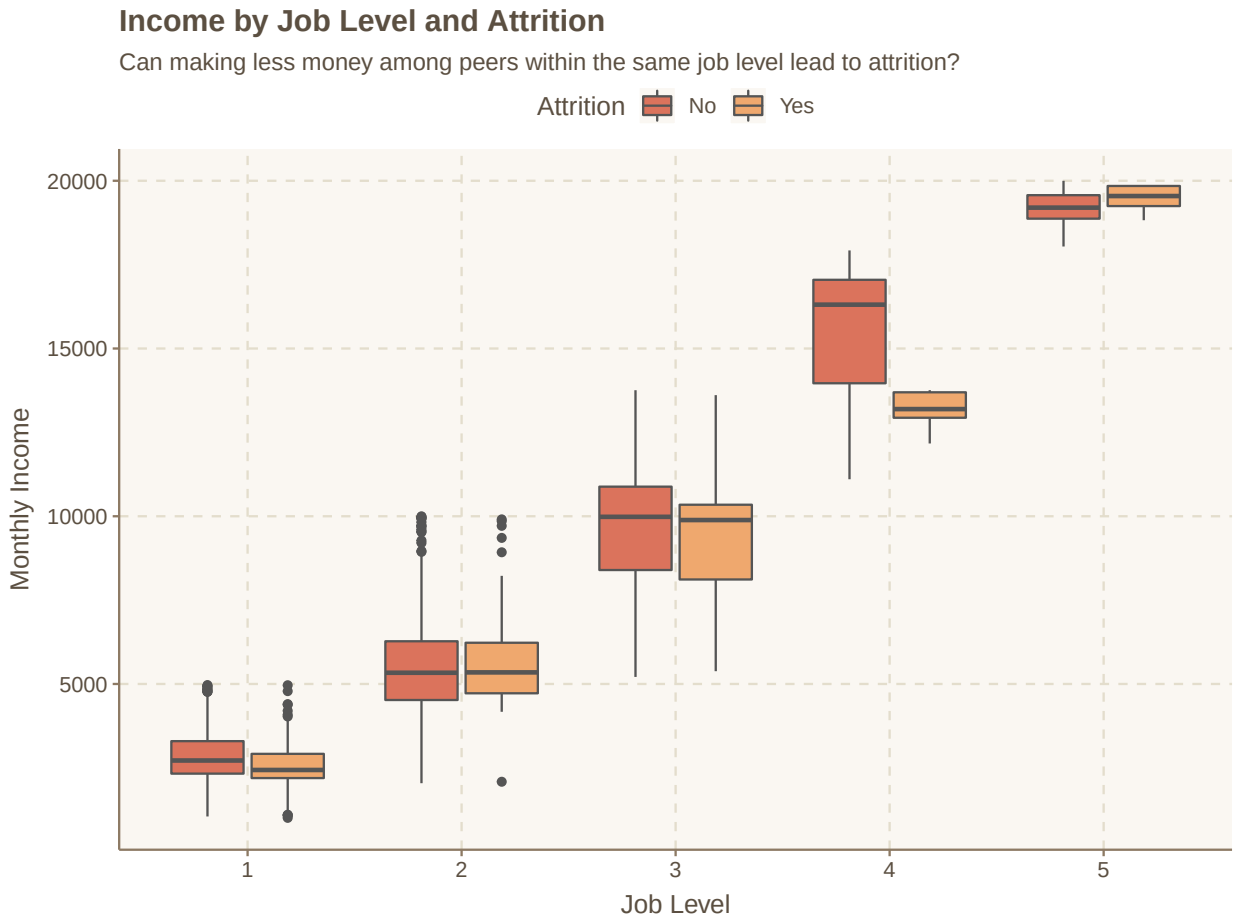


Quick Takeaways:

- Although there was no strong support for attrition due income differences within job roles, there does seem to be support when it comes to departments.
- This leads me to hypothesize that the problem exists within departments rather than job roles. For example, research scientists who make far less than research directors are more likely to leave due to the income disparity between the two roles.

Income Differences by Job Level and Attrition

```
ggthemr('dust')
f_data %>%
  ggplot(aes(x = as.factor(JobLevel), y = MonthlyIncome, fill = Attrition)) +
  geom_boxplot() +
  theme(legend.position = "top") +
  labs( x = "Job Level", y = "Monthly Income",
    title = "Income by Job Level and Attrition",
    subtitle = "Can making less money among peers within the same job level lead to attrition?" )
```



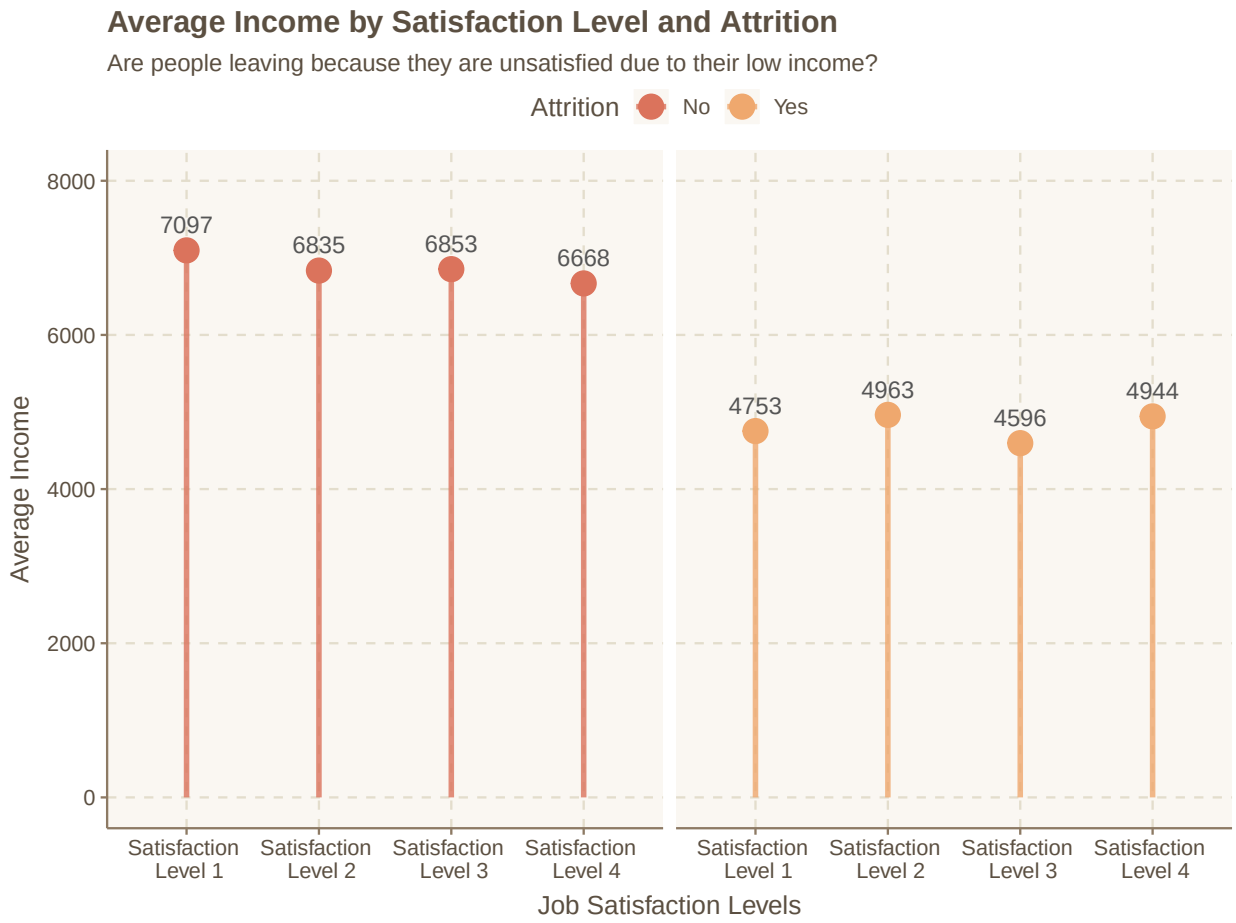
Quick Takeaways:

- Overall, there does not seem to be strong support for attrition due to income differences within job levels.
- The only level where huge income differences occurred was in level 4. Further analysis on level 4 may be required.

Income Differences by Job Satisfaction and Attrition

```
ggthemr('dust')
f_data %>%
  group_by(JobSatisfaction, Attrition) %>%
  summarise(AvgInc = mean(MonthlyIncome)) %>%
  mutate(JobSatisfaction = gsub("^", "Satisfaction \n Level ", JobSatisfaction)) %>%
  ggplot(aes(x = JobSatisfaction, y = AvgInc)) +
  geom_point(size = 5, aes(y = AvgInc, color = Attrition)) +
  geom_segment(aes(x = JobSatisfaction, xend = JobSatisfaction, y = 0,
    yend = AvgInc, color = Attrition), size = 1.2, linetype = 1, alpha = .8) +
  facet_wrap(vars(Attrition))+
  scale_y_continuous(limits = c(0,8000)) +
  labs(x = "Job Satisfaction Levels", y = "Average Income",
    title = "Average Income by Satisfaction Level and Attrition",
    subtitle = "Are people leaving because they are unsatisfied due to their low income?") +
```

```
geom_text(aes(x = JobSatisfaction, y = AvgInc,
  label = round(AvgInc,0)), vjust = -1) +
scale_color_ggthemr_d() +
theme(strip.text.x = element_blank(), legend.position = "top")
```



Quick Takeaways:

- Overall, there doesn't seem to be strong support showing that lower income is the reason for lower job satisfaction. For those who left and those who stayed, average income did not rise with rising job satisfaction. In fact, there is nearly 0 correlation between income and job satisfaction.

Potential Factor: Satisfaction Variables

Hypotheses to Test:

1. Individuals are leaving the organization due to low levels of job satisfaction
2. Individuals are leaving the organization due to levels of environmental satisfaction
3. Individuals are leaving the organization due to working overtime.

Summary:

1. After comparing the average job satisfaction by role and attrition, it is evident that for the jobs with the highest attrition rates (e.g. sales rep, lab tech, and human resources), there was a large gap in average job satisfaction for those who left and those who stayed.
2. This effect however is close to 0 for manufactoring director and healthcare representative.
3. After comparing the average environment satisfaction by role and attrition, healthcare representative and managers (two job roles that does not have high attrition rates) have the largest gaps between those who leave and those who stay. Understanding their environment and how their environment is affecting their work will be critical.
4. In addition, exploratory data analysis shows that that for sales reps and research scientists, the average environmental satisfaction gap between those who stayed and those who left were minimal. This indicates that for those roles, environment did not play a huge factor in determining leaving the organization.
5. 30.5% of individuals who worked overtime left the organization in comparison to 10.4% of individuals who did not.
6. However, overtime was relatively constant throughout levels and job role. It may be worth considering other factors that contribute to when and which employees work overtime.
7. Overtime did not seem to be a driver of job satisfaction and environment satisfaciton

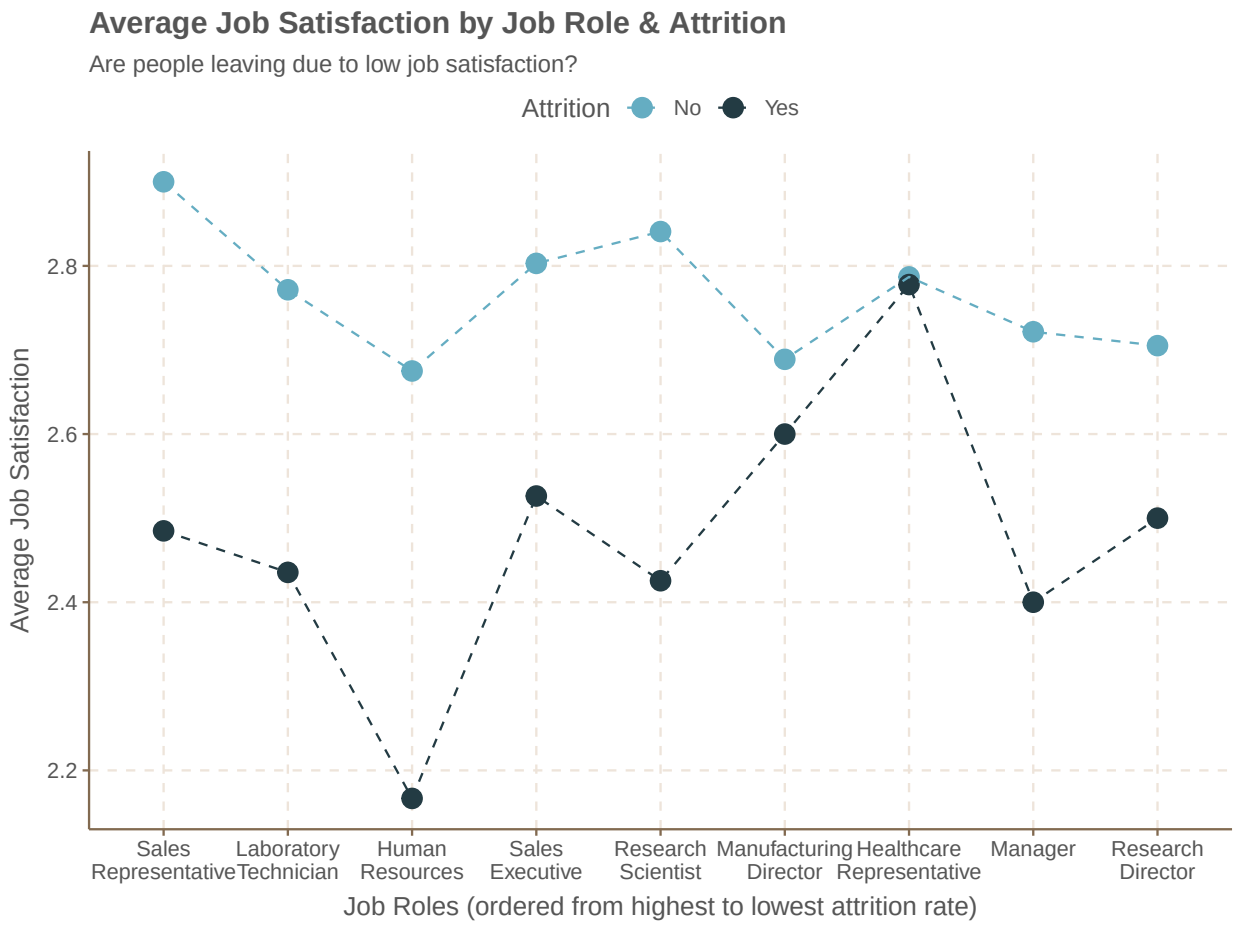
Final thoughts:

There is strong evidence that factors relating to satisfaction are key drivers to attrition. Job satisfaction seems to be a bigger factor for roles with high levels of attrition such as Sales Rep, Laboratory technicians, human resources, and sales executives. However environment satisfaction seems to be a bigger factor for roles with low attrition such as Healthcare reps, and managers. Managers seem to be the most impacted by ow Environment Satisfaction. Overtime also seems to play a factor in attrition. However, overtime rates seem relatively similar within roles and level.

Supporting Analysis for Satisfaction Differences

Job Satisfaction by Job Roles and Attrition (ordered by highest to lowest attrition rate)

```
ggthemr('fresh')
f_data %>%
  select(JobRole, JobSatisfaction, Attrition) %>%
  group_by(JobRole, Attrition) %>%
  summarise(avg = mean(JobSatisfaction)) %>%
  mutate(JobRole = gsub(" ", "\n", JobRole)) %>%
  mutate(JobRole = as.factor(JobRole)) %>%
  mutate(JobRole = fct_relevel(JobRole, "Sales\nRepresentative", "Laboratory\nTechnician",
    "Human\nResources", "Sales\nExecutive",
    "Research\nScientist", "Manufacturing\nDirector",
    "Healthcare\nRepresentative", "Manager",
    "Research\nDirector")) %>%
  ggplot(aes(x = JobRole, y = avg, color = Attrition)) +
  geom_point(size = 4) +
  geom_line(aes(group = Attrition), linetype = "dashed") +
  labs(title = "Average Job Satisfaction by Job Role & Attrition",
    subtitle = "Are people leaving due to low job satisfaction?",
    x = "Job Roles (ordered from highest to lowest attrition rate)",
    y = "Average Job Satisfaction") +
  scale_color_ggthemr_d() +
  theme(legend.position = 'top')
```



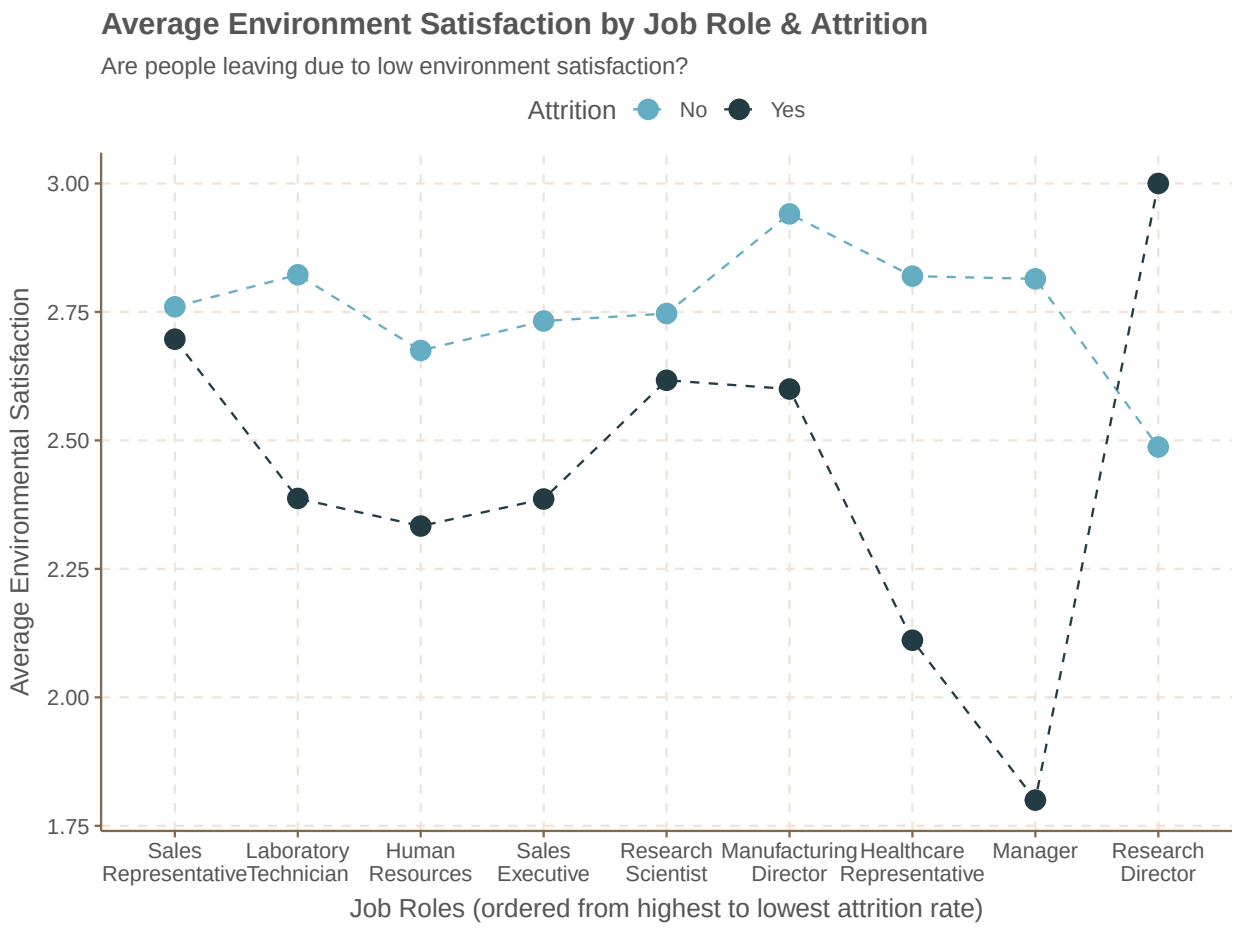
Quick Takeaways:

- Here we find something really interesting. Overall, those who leave the organization tend to have lower job satisfaction ratings.
- However, the effect of job satisfaction on attrition seems to be stronger for those who are in job roles with higher attrition rate.
- The difference in job satisfaction is smaller when you go to job roles with lower attrition rates.

Attrition by Environmental Satisfaction

```
f_data %>%
  select(JobRole, EnvironmentSatisfaction, Attrition) %>%
  group_by(JobRole, Attrition) %>%
  summarise(avg = mean(EnvironmentSatisfaction)) %>%
  mutate(JobRole = gsub(" ", "\n", JobRole)) %>%
  mutate(JobRole = as.factor(JobRole)) %>%
  mutate(JobRole = fct_relevel(JobRole, "Sales\nRepresentative", "Laboratory\nTechnician",
    "Human\nResources", "Sales\nExecutive",
    "Research\nScientist", "Manufacturing\nDirector",
    "Healthcare\nRepresentative", "Manager",
    "Research\nDirector")) %>%
  ggplot(aes(x = JobRole, y = avg, color = Attrition)) +
  geom_point(size = 4) +
```

```
  geom_line(aes(group = Attrition), linetype = "dashed") +
  labs(title = "Average Environment Satisfaction by Job Role & Attrition",
    subtitle = "Are people leaving due to low environment satisfaction?",
    x = "Job Roles (ordered from highest to lowest attrition rate)",
    y = "Average Environmental Satisfaction") +
  scale_color_ggthemr_d() +
  theme(legend.position = "top")
```

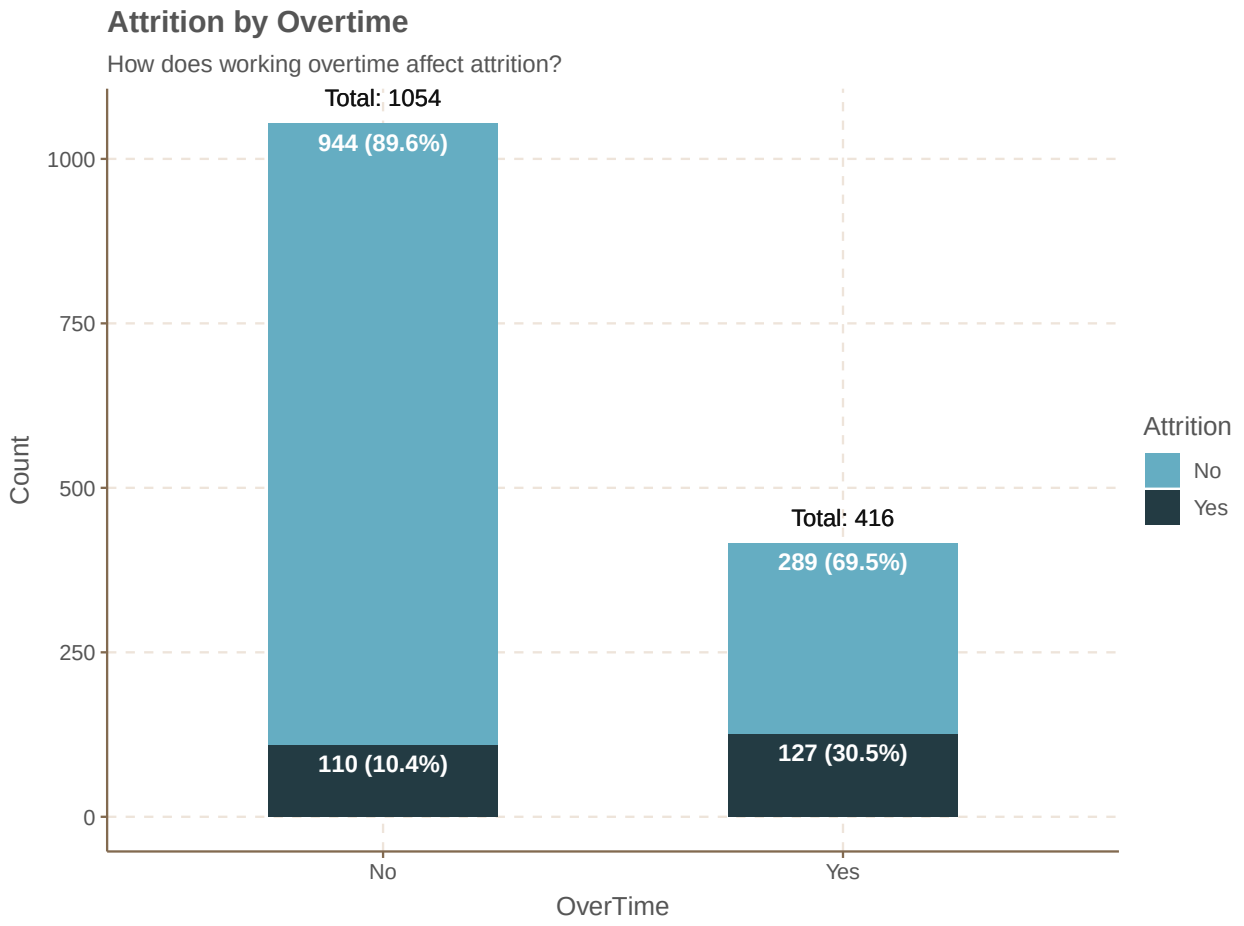


Quick Takeaways:

- Here we find something really interesting again. Overall, those who leave the organization tend to have lower environmental satisfaction ratings. This effect however is close to 0 for manufactoring director and healthcare representative.
- However, in an opposite trend to job satisfaction, the effect of environmental satisfaction ratings on attrition seems higher for job roles with higher attrition rates.
- This shows that while job satisfaction may have a stronger role in attrition for sales rep, laboratory technician, HR, environmental satisfaction may have a stronger role for positions like healthcare representative and managers.
- Strangely enough research director was the only position that had higher environmental satisfaction for those who left compared to those who stayed.
- The largest difference in environment satisfaction ratings were from Managers. There should be an investigation to see what is driving such low environmental satisfaction from Managers.

Attrition by Overtime

```r
ggthemr("fresh")
f_data %>%
  group_by(Attrition, OverTime) %>%
  summarise(Count = n()) %>%
  group_by(OverTime) %>%
  arrange(Count) %>%
  mutate(Count_total = cumsum(Count),
    CountPer = prop.table(Count),
    final = paste(Count, " (", round(CountPer*100,1), "%)", sep = ""),
    total = paste("Total:", sum(Count)),
    top = sum(Count)) %>%
  ggplot(aes(x = OverTime, y = Count, fill = Attrition)) +
  geom_bar(stat = "identity", width = .5) +
  geom_text(aes(x = OverTime, y = Count_total, label = final),
    vjust = 1.6, color = "white", fontface = "bold") +
  geom_text(aes(x = OverTime, y = top, label = total), vjust = -1) +
  labs(title = "Attrition by Overtime",
  subtitle = "How does working overtime affect attrition?")
```

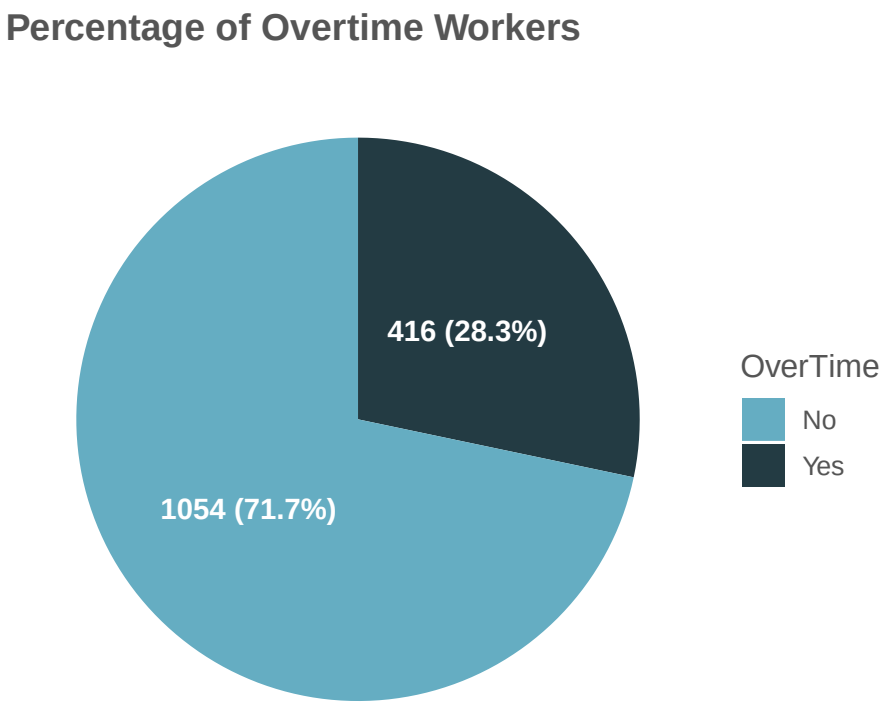


Quick Takeaways:

- Overall, 30.5% of workers who worked overtime left the company in comparison to the 10.4% of employees who never worked overtime.

Overtime Distribution Within Company

```
f_data %>%
  group_by(OverTime) %>%
  summarise(Count = n()) %>%
  arrange(desc(OverTime)) %>%
  mutate(ypos = cumsum(Count)-.5*(Count)) %>%
  mutate(pct = round(prop.table(Count)*100,1),
    label = paste(Count, " (",pct, "%", ")", sep = "")) %>%
  ggplot(aes(x = "", y = Count, fill = OverTime), color = "white") +
  geom_bar(stat = "identity", width = 1) +
  coord_polar("y", start =0) +
  geom_text(aes(y = ypos, label = label), color = "white",
    fontface = "bold", size = 4) +
  labs(title = "Percentage of Overtime Workers", x = "", y = "") +
  theme(axis.line = element_blank(),
    axis.text.x = element_blank(),
    axis.ticks = element_blank(),
    axis.text.y = element_blank(),
    panel.grid = element_blank())
```

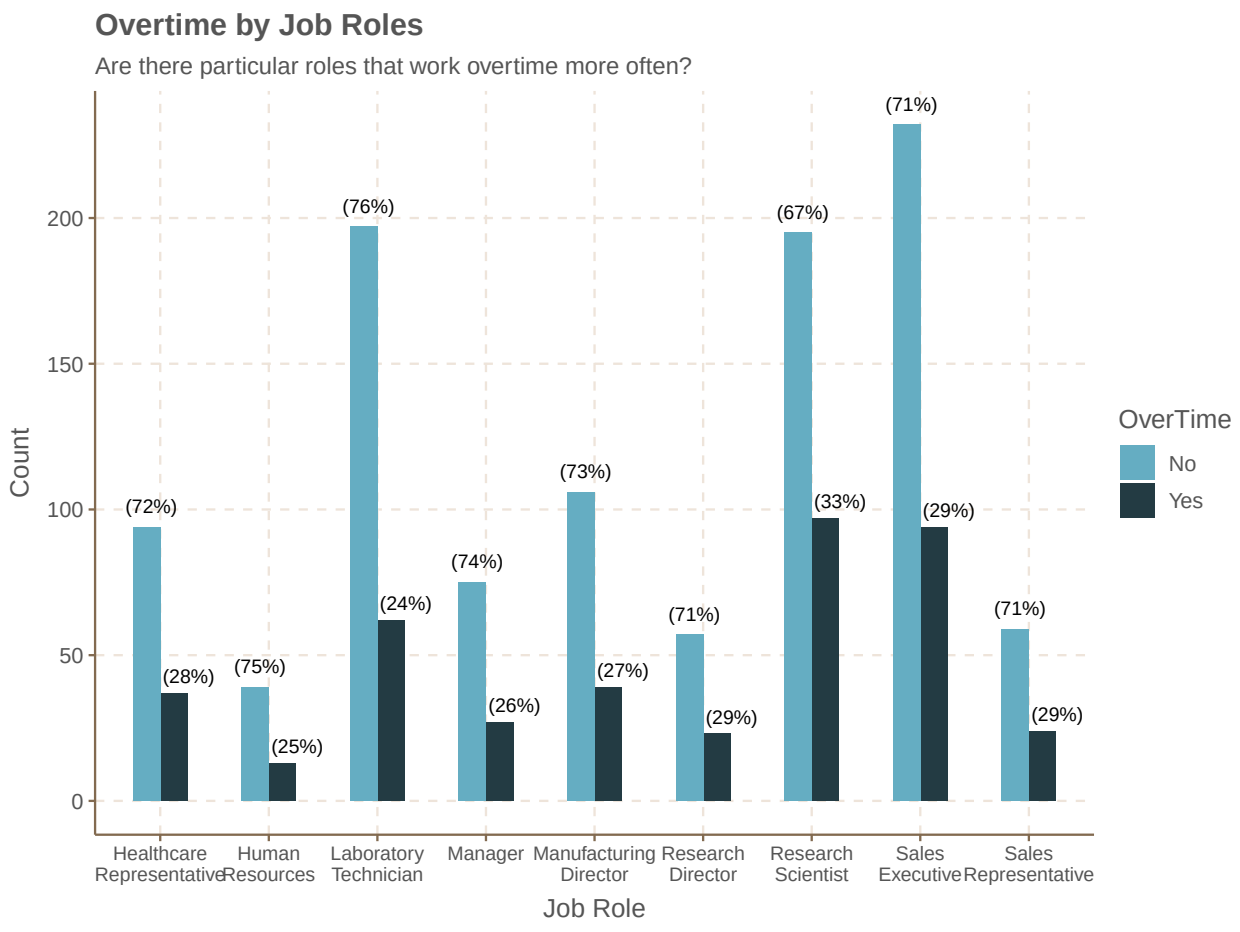


Quick Takeaways:

- 28.3% of the workforce worked overtime

Overtime by Job Roles

```
f_data %>%
  group_by(OverTime, JobRole) %>%
  summarise(Count = n()) %>%
  group_by(JobRole) %>%
  arrange(Count) %>%
  mutate(CountPer = prop.table(Count),
    label = paste(" (", round(CountPer*100,0), "%)", sep = ""),
    total = paste("Total:", sum(Count)),
    top = sum(Count),
    JobRole = gsub(" ", "\n", JobRole)) %>%
  ggplot(aes(x = as.factor(JobRole), y = Count, fill = as.factor(OverTime))) +
  geom_bar(stat = "identity", position = "dodge", width = .5) +
  geom_text(aes(x = JobRole, y = Count,
    label = ifelse(OverTime == "Yes", label, "")),
    vjust = -.7, hjust = .05, color = "black", size = 3.1) +
  geom_text(aes(x = JobRole, y = Count,
    label = ifelse(OverTime == "No", label, "")),
    vjust = -1, hjust=.70, color = "black", size = 3.1) +
  labs(title = "Overtime by Job Roles",
    subtitle = "Are there particular roles that work overtime more often?",
    x = "Job Role", fill = "OverTime") +
  theme(axis.text.x = element_text(size = 9))
```

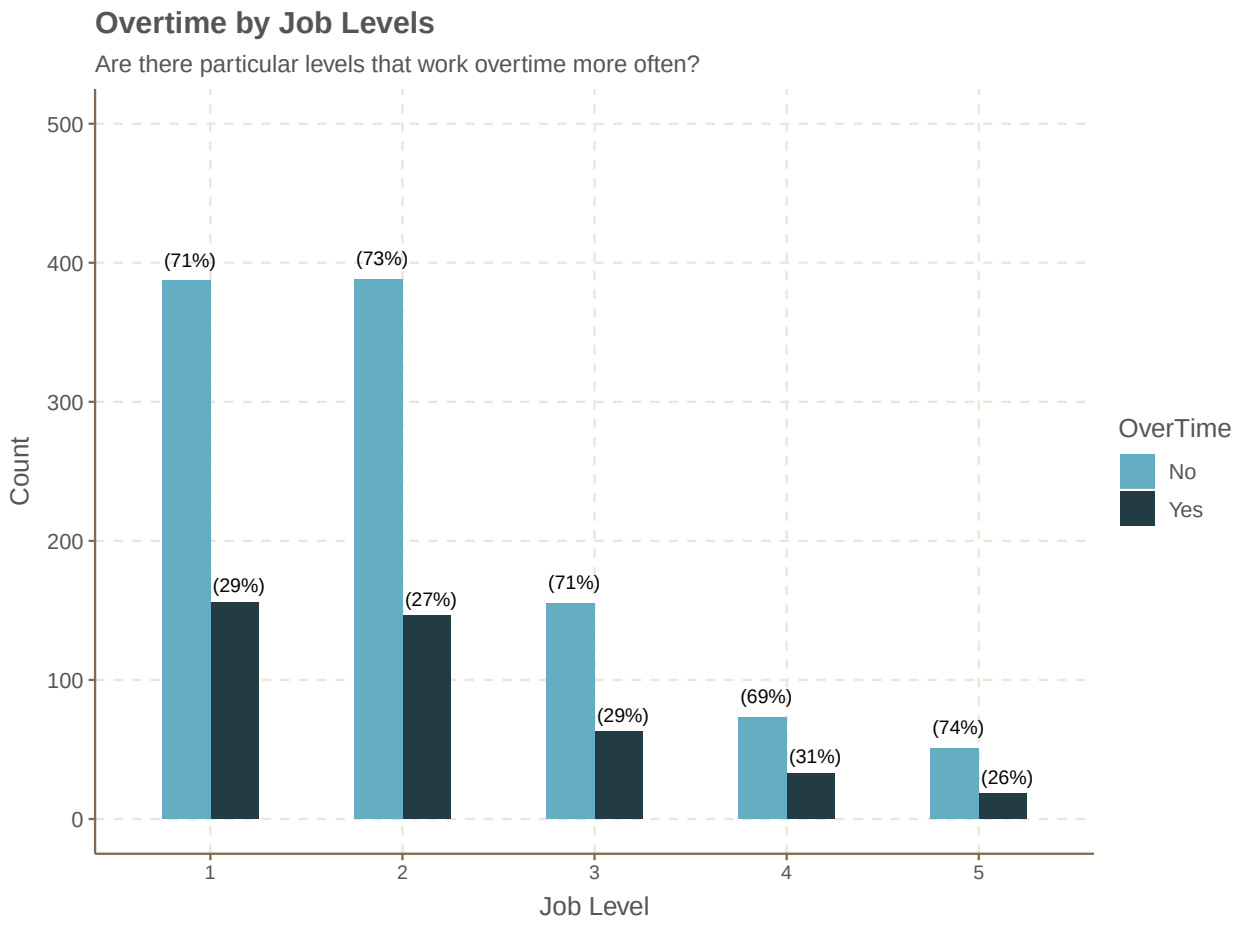


Quick Takeaways:

- Overall, all job roles had 25%-33% of their workforce work overtime with the highest being research scientists.

Overtime by Job Levels

```
f_data %>%
  group_by(OverTime, JobLevel) %>%
  summarise(Count = n()) %>%
  group_by(JobLevel) %>%
  arrange(Count) %>%
  mutate(CountPer = prop.table(Count),
    label = paste(" (", round(CountPer*100,0), "%)", sep = ""),
    total = paste("Total:", sum(Count)),
    top = sum(Count)) %>%
  ggplot(aes(x = as.factor(JobLevel), y = Count, fill = as.factor(OverTime))) +
  geom_bar(stat = "identity", position = "dodge", width = .5) +
  geom_text(aes(x = JobLevel, y = Count,
    label = ifelse(OverTime == "Yes", label, "")),
    vjust = -.7, hjust = .05, color = "black", size = 3.1) +
  geom_text(aes(x = JobLevel, y = Count,
    label = ifelse(OverTime == "No", label, "")),
    vjust = -1, hjust=.9, color = "black", size = 3.1) +
  labs(title = "Overtime by Job Levels",
    subtitle = "Are there particular levels that work overtime more often?",
    x = "Job Level", fill = "OverTime") +
  theme(axis.text.x = element_text(size = 9)) +
  scale_y_continuous(limits = c(0,500))
```

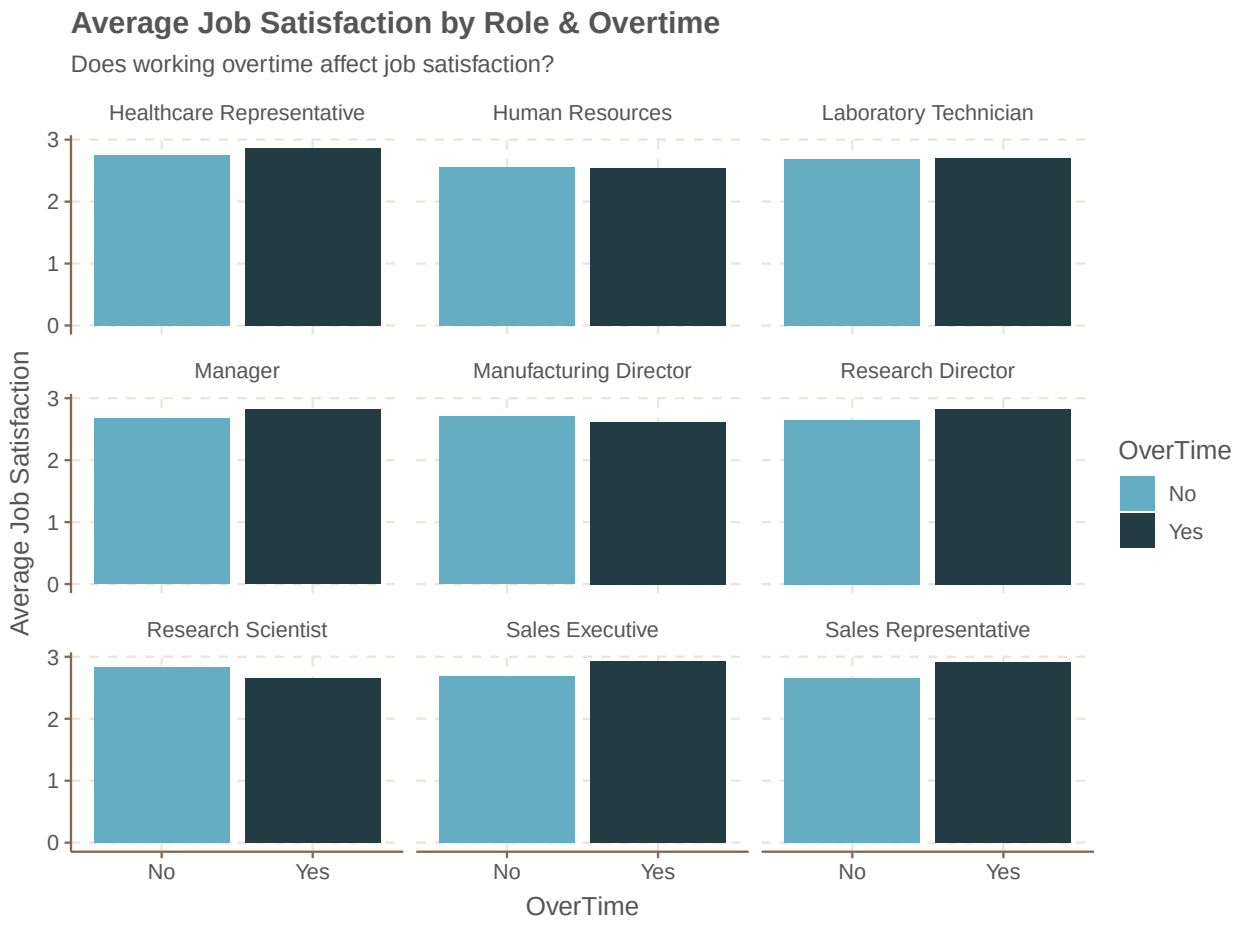


Quick Takeaways:

- Overall, all job levels had 25%-31% of their workforce work overtime with the highest being level 4.

Average Job Satisfaction by Role & Overtime

```
f_data %>%
  group_by(OverTime, JobRole) %>%
  summarise(avg = mean(JobSatisfaction))%>%
  ggplot(aes(x = OverTime, y = avg, fill = OverTime))+
  geom_bar(stat = "identity") +
  facet_wrap(vars(JobRole)) +
  labs(y = "Average Job Satisfaction",
    title = "Average Job Satisfaction by Role & Overtime",
    subtitle = "Does working overtime affect job satisfaction?")
```

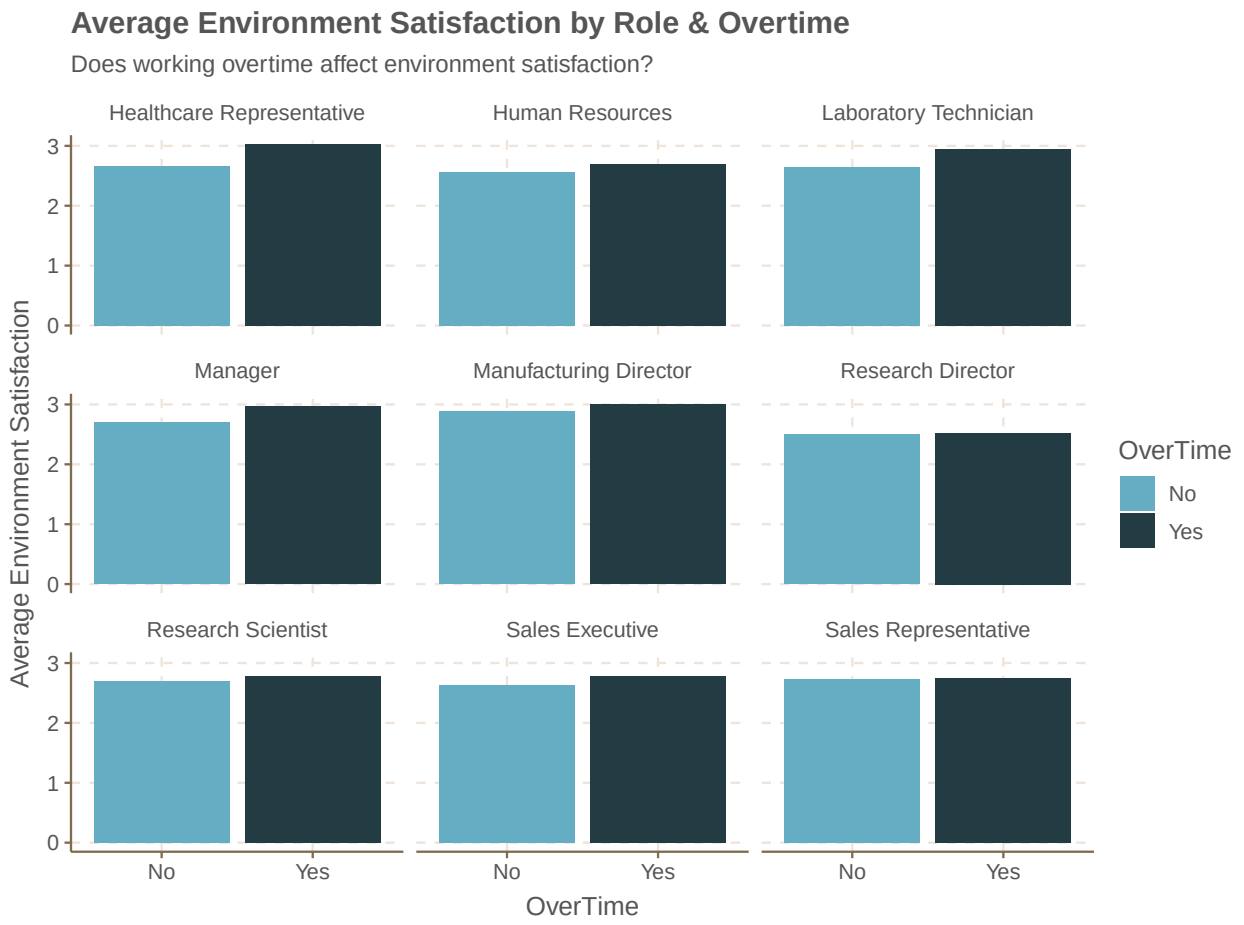


Quick Takeaways:

- Overall, job satisfaction does not seem to be impacted much by overtime.

Average Environment Satisfaction by Role & Overtime

```
f_data %>%
  group_by(OverTime, JobRole) %>%
  summarise(avg = mean(EnvironmentSatisfaction))%>%
  ggplot(aes(x = OverTime, y = avg, fill = OverTime))+
  geom_bar(stat = "identity") +
  facet_wrap(vars(JobRole)) +
  labs(y = "Average Environment Satisfaction",
    title = "Average Environment Satisfaction by Role & Overtime",
    subtitle = "Does working overtime affect environment satisfaction?")
```



Quick Takeaways:

- – Overall, job environment satisfaction does not seem to be impacted much by overtime.

Conlusions: & Next Steps

Conclusions

After a preliminary exploratory data analysis, the following conclusions seemed to be supported:

1. The company's is losing a **large amount of young talent** as 55% of all attrition came from employees between 18-32. Also, 36% of employees within the 18-25 age bracket as well as 22% of employees within the 26-32 bracket left indicating the company's struggles maintaining younger talent
2. Sales representatives had the highest within attrition at 40% with laboratory technician and human resources following at 24% and 23%. After investigation, **low income as well as low job satisfaction ratings may be strong contributors to the high attrition rate within sales representatives, laboratory technicians and human resources.**
3. **While environment satisfaction does seem to be a factor contributing to attrition across job roles, it was most evident for healthcare representatives and managers.** There was a large gap in environmental satisfaction between those healthcare representatives and managers who

stayed and those who left. Further investigation on what environmental issues are negatively affecting those roles will be critical.

4. **While job satisfaction does seem to be a factor contributing to attrition across job roles, it was most evident for sales reps, laboratory technicians, and human resources.** There was a large gap in job satisfaction between those sales reps, laboratory technicians, and human resources who stayed and those who left. Further investigation on what issues are negatively affecting those roles will be critical.

5. Analysis shows that for the employees at IBM, **income did not have a strong influence on job satisfaction.** This suggests that low job satisfaction and income are two different factors.

6. Although I hypothesized that **job and environment satisfaction would be lower for those who worked overtime, that did not seem to be the case.** Further analysis on why job and environment satisfaction are low for certain job roles should be evaluated.

7. Altough overtime did not seem to affect job and environment satisfaction, a larger percent of those who worked overtime left the organization compared to those who didn't.

Next Steps:

The purpose of the project was to conduct a preliminary analysis to understand the data to guide future investigations. In addition, I worked to create data visualizations to communicate the data to viewers. All of the conclusions should be validated through a series of point-biserial + pearson correlations, inferential statistics (t-tests), and eventually linear + logistic regression models.
In addition, gathering qualitative data will be important as well. Exit interview information as well as focus groups within job roles, levels, and departments will help investigate questions we still have. For example, qualitative data from exit interviews and focus groups may help in:

1. Understanding drivers of low job satisfaction especially within sales reps, laboratory technicians and human resources.
2. Understanding drivers of low environment satisfaction especially within healthcare representatives and managers.
3. Identifying potential culture issues or policies in place pushing younger talent away.
4. Confirming income as one of the most common drivers of attrition
5. Identifying reasons for why employees work overtime.