DIIG Data Challenge

Lily Li

Load packages & data

```
library(tidyverse)
library(knitr)
library(broom)
library(ggplot2)
library(dplyr)
library(tidyr)

library(ggridges)
library(forcats)

IBM <- read_csv("data.csv")
```

Guiding Questions

What factors contribute to employee satisfaction levels and what can IBM do to improve satisfaction? Do certain roles have greater employee churn? If so, what factors lead to this churn?

Aspects that we will be investigating:

- Basic Demographic Info
- Income Reward & Motivation
- Employee Churn
- Job Satisfaction
- Diversity & Inclusion

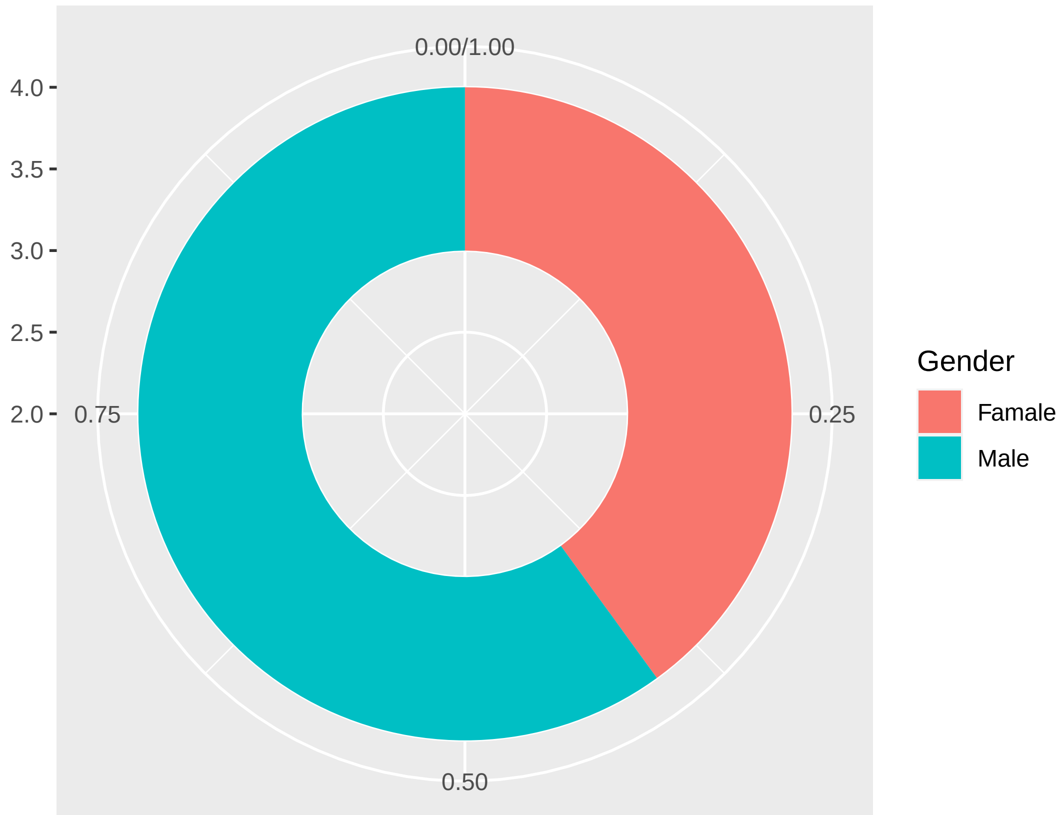
BASIC DEMOGRAPHIC INFO

```
# Gender Distribution
G_init <- IBM %>%
  count(Gender)
#G_init

G <- data.frame(
  Gender=c("Famale", "Male"),
  count=c(588, 882)
)

G$fraction = G$count / sum(G$count)
G$ymax = cumsum(G$fraction)
G$ymin = c(0, head(G$ymax, n=-1))
```

```r
ggplot(G, aes(ymax=ymax, ymin=ymin, xmax=4, xmin=3, fill=Gender)) +
     geom_rect() +
     coord_polar(theta="y") +
     xlim(c(2, 4))
```



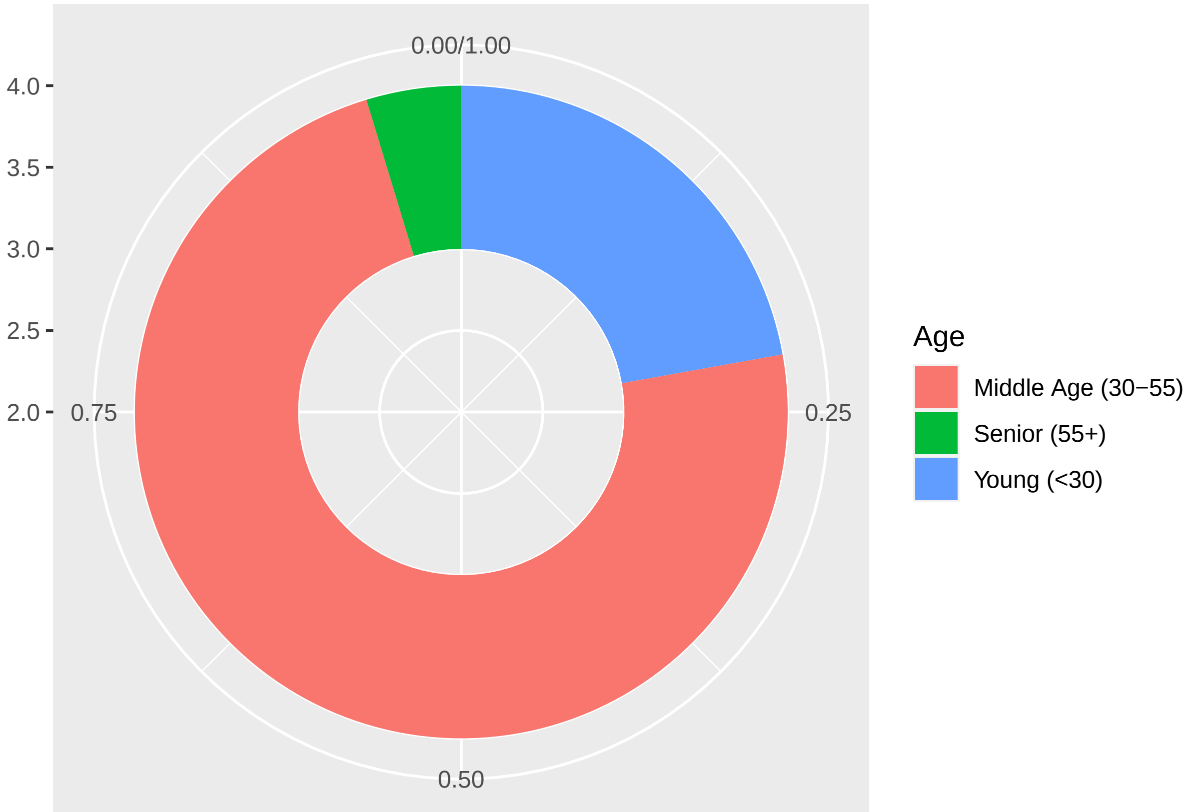

```r
# Age Distribution
A_init <- IBM %>%
  mutate(age = case_when(
    Age < 30 ~ "Young (<30)",
    Age >= 30 & Age < 55 ~ "Middle Age (30-55)",
    Age >= 55 & Age < 65 ~ "Senior (55+)",
  ))%>%
  count(age)
#A_init

A <- data.frame(
  Age=c("Young (<30)", "Middle Age (30-55)", "Senior (55+)"),
  count=c(326, 1075, 69)
)

A$fraction = A$count / sum(A$count)
A$ymax = cumsum(A$fraction)
A$ymin = c(0, head(A$ymax, n=-1))

ggplot(A, aes(ymax=ymax, ymin=ymin, xmax=4, xmin=3, fill=Age)) +
     geom_rect() +
     coord_polar(theta="y") +
```

```
    xlim(c(2, 4))
```

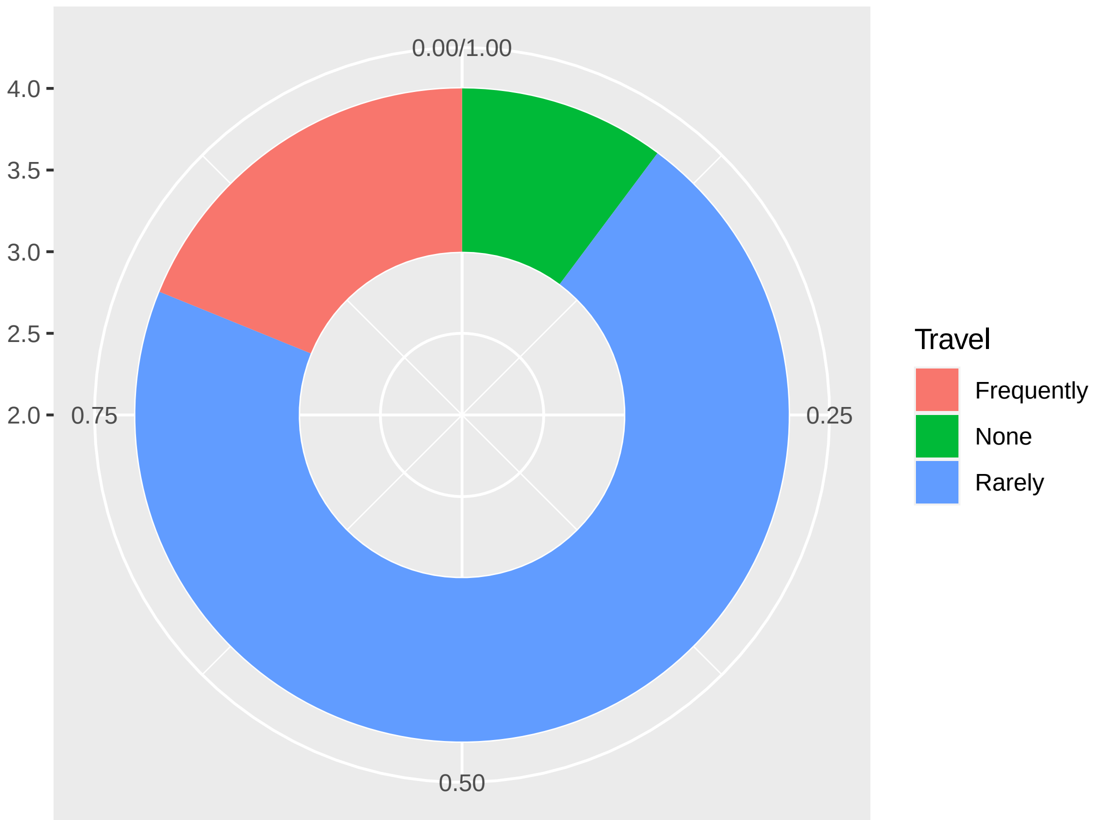



```
# Travel Distribution
T_init <- IBM %>%
  count(BusinessTravel)
#T_init

Tr <- data.frame(
  Travel=c("None", "Rarely", "Frequently"),
  count=c(150, 1043, 277)
)

Tr$fraction = Tr$count / sum(Tr$count)
Tr$ymax = cumsum(Tr$fraction)
Tr$ymin = c(0, head(Tr$ymax, n=-1))

ggplot(Tr, aes(ymax=ymax, ymin=ymin, xmax=4, xmin=3, fill=Travel)) +
     geom_rect() +
     coord_polar(theta="y") +
     xlim(c(2, 4))
```

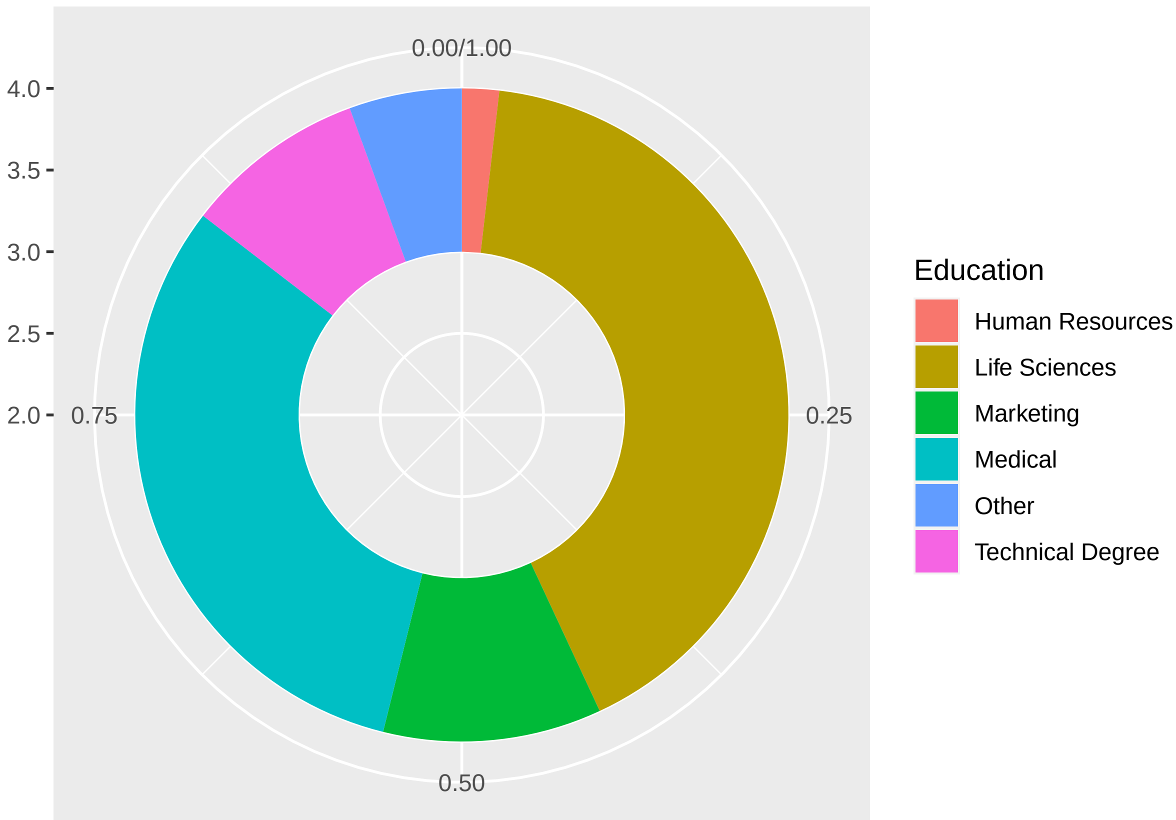



```
# Education Field
Ed_init <- IBM %>%
  count(EducationField)
#Ed_init

Ed <- data.frame(
  Education=c("Human Resources", "Life Sciences", "Marketing", "Medical", "Technical Degree", "Other"),
  count=c(27, 606, 159, 464, 132, 82)
)

Ed$fraction = Ed$count / sum(Ed$count)
Ed$ymax = cumsum(Ed$fraction)
Ed$ymin = c(0, head(Ed$ymax, n=-1))

ggplot(Ed, aes(ymax=ymax, ymin=ymin, xmax=4, xmin=3, fill=Education)) +
  geom_rect() +
  coord_polar(theta="y") +
  xlim(c(2, 4))
```

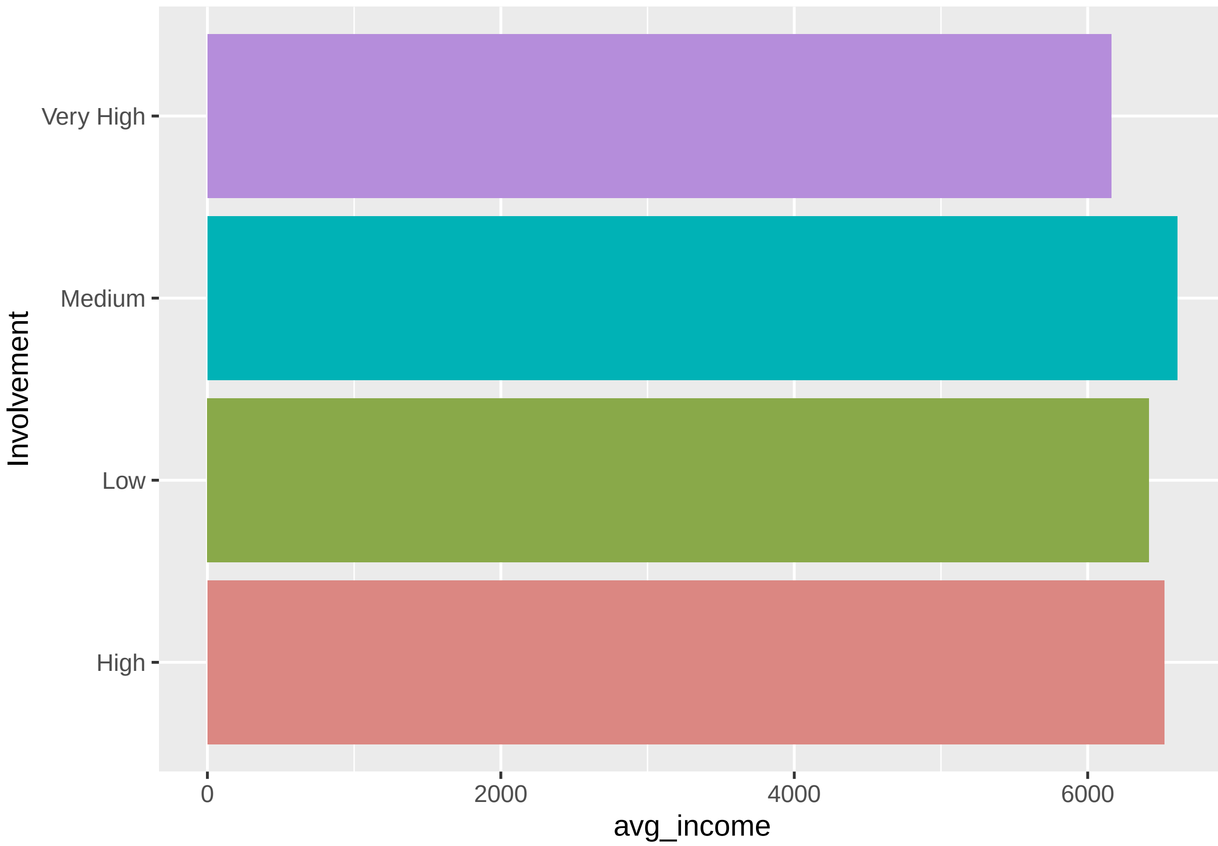


INCOME REWARD & MOTIVATION

Job Involvement vs. Monthly Income

```
involvement <- IBM %>%
  select(JobInvolvement, MonthlyIncome) %>%
  mutate(Involvement = case_when(
    JobInvolvement == 1 ~ "Low",
    JobInvolvement == 2 ~ "Medium",
    JobInvolvement == 3 ~ "High",
    JobInvolvement == 4 ~ "Very High",
  )) %>%
  group_by(Involvement) %>%
  summarize(avg_income = mean(MonthlyIncome))

ggplot(involvement, aes(x = Involvement, y = avg_income, fill = Involvement)) +
  geom_bar(stat = "identity") +
  scale_fill_hue(c = 60) +
  theme(legend.position="none") +
  coord_flip()
```



Job Involvement & Income: are people who are involved the most fairly rewarded? Not very differentiated – maybe can improve evaluation in terms of job involvement and reward those who have higher levels of involvement. Maybe can design monthly/seasonally evaluations&competitions to reward those with higher job involvement, so that employees can be more motivated.

Years At Company vs. Monthly Income

For those who have been working for a long time, if their income is not higher, then might need to think about what is causing this problem – is it that employees are having a hard time getting promotion (structure change)? Or is it that they don't have a lot motivation (innovation approaches)?

It would be better if employees know that if they work harder and stay at the firm longer, then they will be properly rewarded.

```
ggplot(IBM, aes(x = YearsAtCompany, y = MonthlyIncome, color = Age)) +
  geom_point() +
  geom_smooth(method=lm , color="red", fill="#69b3a2", se=FALSE) +
  scale_color_gradient(low="turquoise1", high="turquoise4")
```

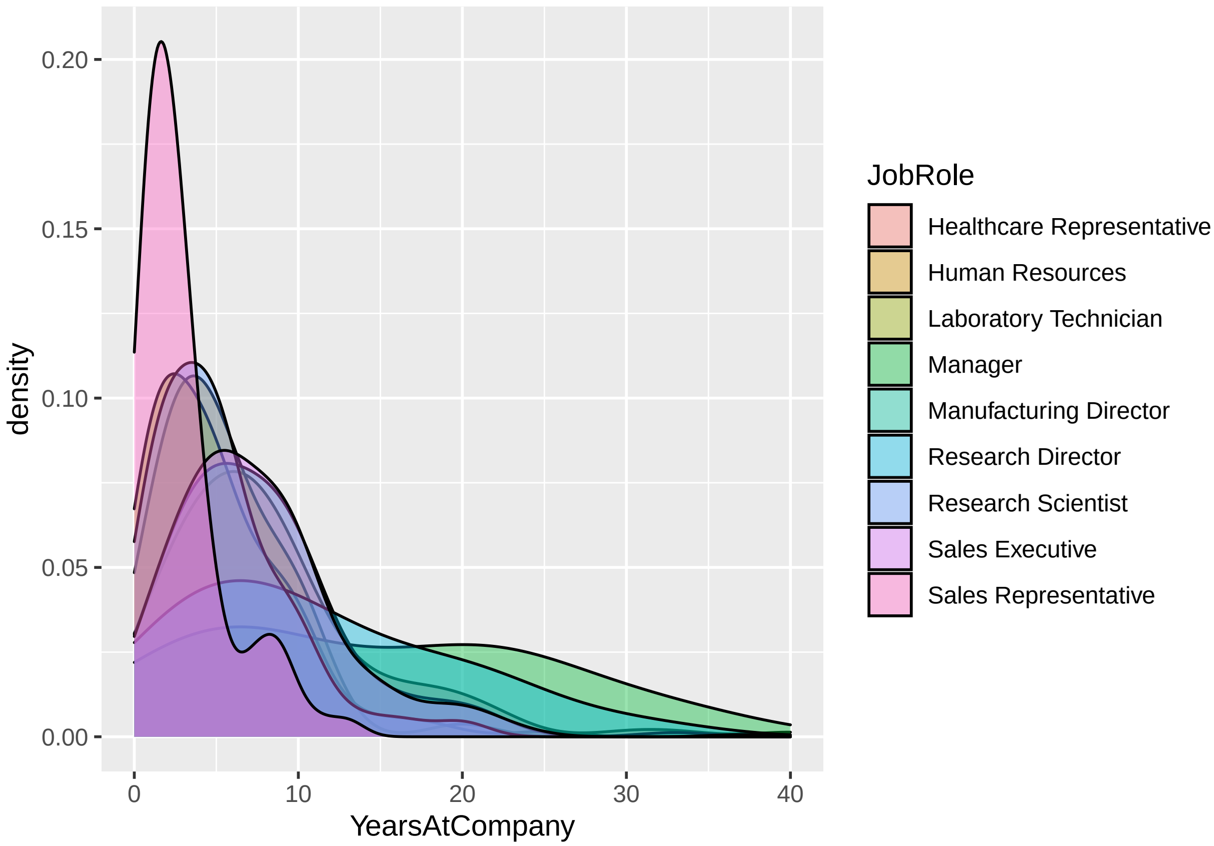

```r
ggplot(IBM, aes(x = TotalWorkingYears, y = MonthlyIncome, color = Age)) +
  geom_point() +
  geom_smooth(method=lm , color="red", fill="#69b3a2", se=FALSE) +
  scale_color_gradient(low="turquoise1", high="turquoise4")
```

EMPLOYEE CHURN

Education level vs. Years At Company

```
yearsVedu <- IBM %>%
  mutate(College = case_when(
    Education == 1 ~ "Below College",
    Education == 2 ~ "College",
    Education == 3 ~ "Bachelor",
    Education == 4 ~ "Master",
    Education == 5 ~ "Doctor"
  ), !is.na(Education)) %>%
  select(College, YearsAtCompany)

ggplot(data=IBM, aes(x=YearsAtCompany, group=JobRole, fill=JobRole)) +
  geom_density(adjust=1.5, alpha=.4)
```



Employee Churn

We can see that more senior roles have less employee churn.

Factors that might be leading to higher churn (both looking at p-value nd estimated coefficient): * Monthly Income * Years since last promotion

```
churn <- IBM %>%
  group_by(JobRole) %>%
  summarize(AverageWorkYear = mean(YearsAtCompany))

ggplot(churn, aes(x = JobRole, y = AverageWorkYear, fill = JobRole)) +
  geom_bar(stat = "identity") +
  scale_fill_hue(c = 40) +
  theme(legend.position="none") +
  coord_flip()
```


```r
Sales <- IBM %>%
  filter(JobRole == "Sales Representative")

churn_model <- lm(YearsAtCompany ~ as.factor(BusinessTravel) + DistanceFromHome + as.factor(EnvironmentS

model_sales <- lm(YearsAtCompany ~ as.factor(BusinessTravel) + DistanceFromHome + as.factor(EnvironmentS

tidy(churn_model)
```

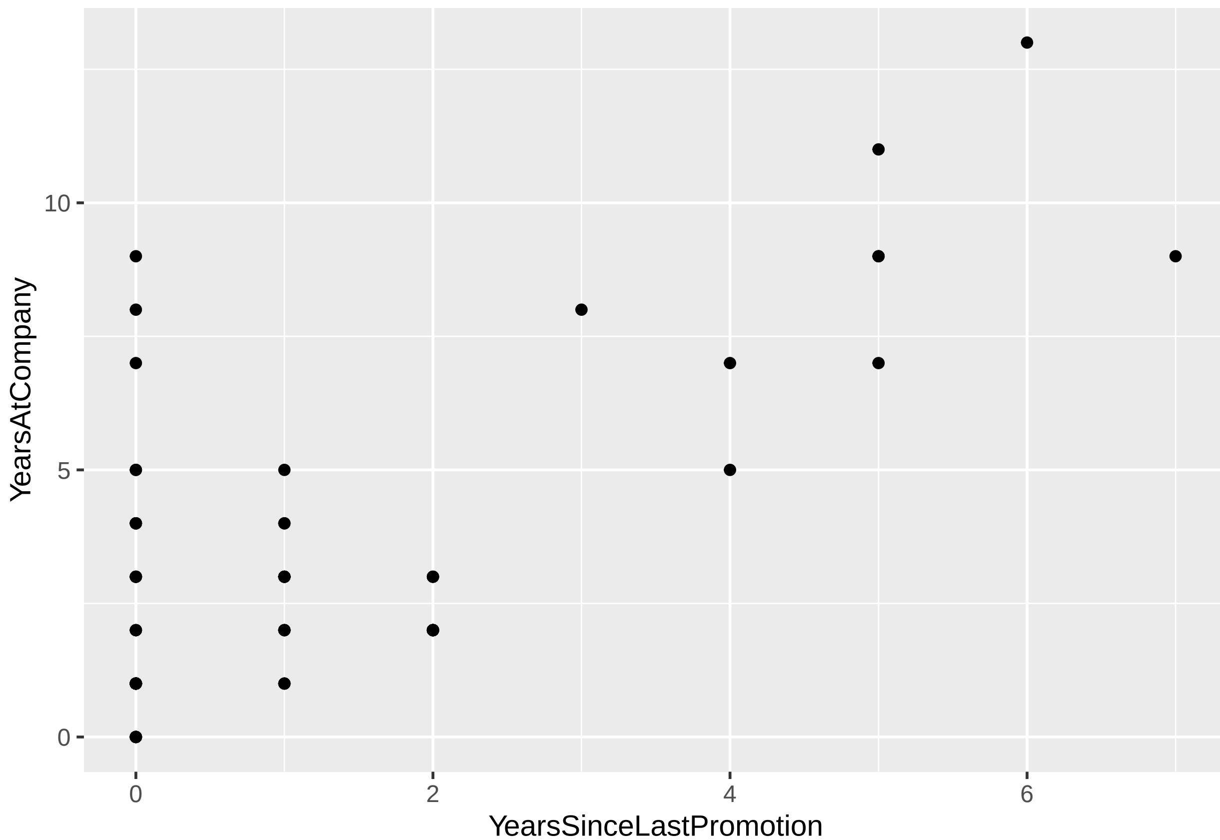
```
## # A tibble: 12 x 5
##    term                                  estimate std.error statistic   p.value
##    <chr>                                    <dbl>     <dbl>     <dbl>     <dbl>
##  1 (Intercept)                            1.81e+0 0.541        3.35   8.21e-  4
##  2 as.factor(BusinessTravel)Travel_Frequ~  1.24e-1 0.448        0.277  7.82e-  1
##  3 as.factor(BusinessTravel)Travel_Rarely -7.62e-2 0.386       -0.197  8.44e-  1
##  4 DistanceFromHome                        7.57e-3 0.0142       0.533  5.94e-  1
##  5 as.factor(EnvironmentSatisfaction)2    -2.11e-2 0.370       -0.0570 9.55e-  1
##  6 as.factor(EnvironmentSatisfaction)3     1.70e-1 0.335        0.507  6.12e-  1
##  7 as.factor(EnvironmentSatisfaction)4    -1.14e-1 0.335       -0.341  7.33e-  1
##  8 as.factor(JobSatisfaction)2             2.45e-1 0.370        0.663  5.08e-  1
##  9 as.factor(JobSatisfaction)3             2.17e-1 0.334        0.648  5.17e-  1
## 10 as.factor(JobSatisfaction)4             1.69e-1 0.331        0.510  6.10e-  1
## 11 YearsSinceLastPromotion                 9.50e-1 0.0381      24.9    1.74e-114
## 12 MonthlyIncome                           4.46e-4 0.0000261   17.1    9.08e- 60
```

```r
tidy(model_sales)
```

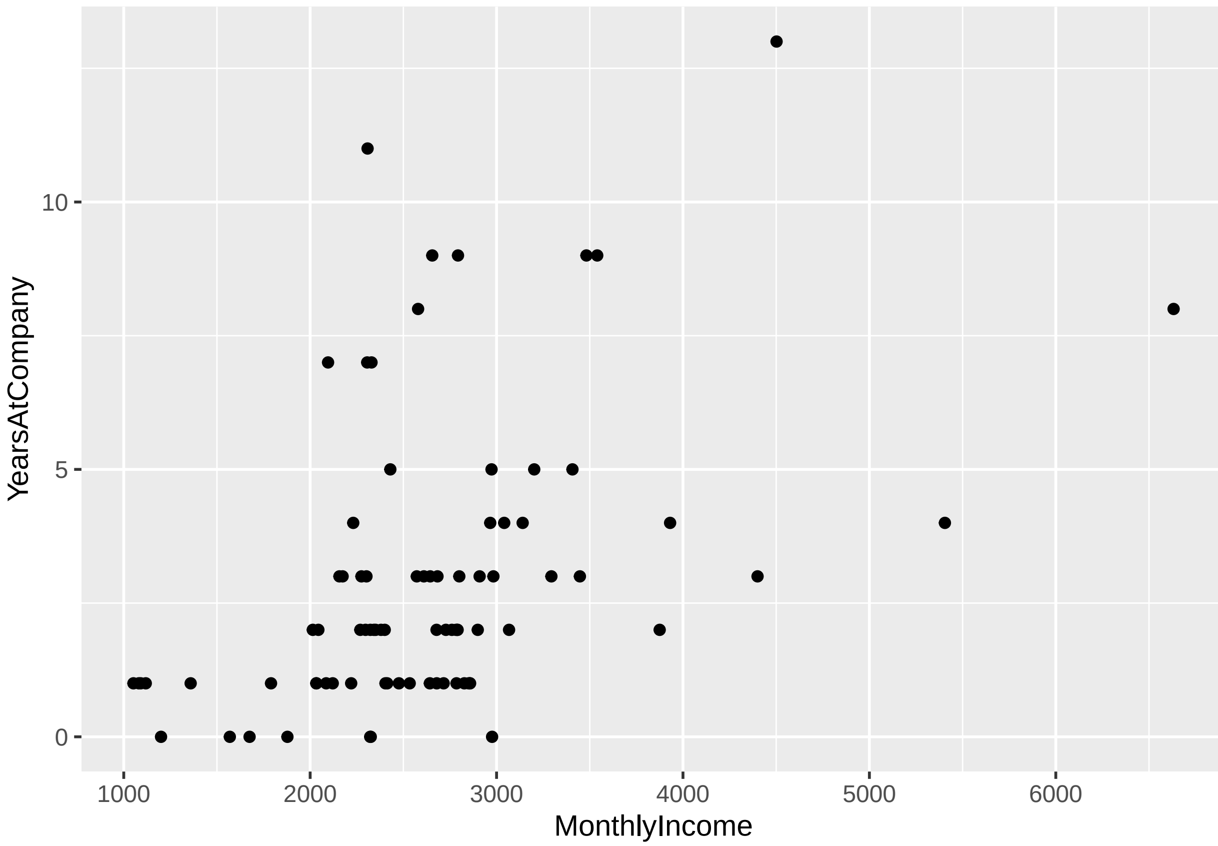
```
## # A tibble: 12 x 5
```

```
##    term                                       estimate std.error statistic  p.value
##    <chr>                                         <dbl>     <dbl>     <dbl>    <dbl>
##  1 (Intercept)                                -1.92e-1  1.27        -0.151 8.80e- 1
##  2 as.factor(BusinessTravel)Travel_Freque~ -5.59e-1  0.917       -0.610 5.44e- 1
##  3 as.factor(BusinessTravel)Travel_Rarely  -1.90e-1  0.888       -0.214 8.31e- 1
##  4 DistanceFromHome                          4.05e-2  0.0277       1.46  1.48e- 1
##  5 as.factor(EnvironmentSatisfaction)2     -7.15e-1  0.688       -1.04  3.02e- 1
##  6 as.factor(EnvironmentSatisfaction)3     -5.67e-1  0.667       -0.850 3.98e- 1
##  7 as.factor(EnvironmentSatisfaction)4     -1.31e+0  0.695       -1.88  6.42e- 2
##  8 as.factor(JobSatisfaction)2             -1.31e-1  0.683       -0.192 8.48e- 1
##  9 as.factor(JobSatisfaction)3              9.83e-1  0.664        1.48  1.43e- 1
## 10 as.factor(JobSatisfaction)4              6.06e-1  0.654        0.927 3.57e- 1
## 11 YearsSinceLastPromotion                  1.00e+0  0.137        7.33  2.89e-10
## 12 MonthlyIncome                            8.57e-4  0.000259     3.31  1.48e- 3
```

```r
ggplot(Sales, aes(x = YearsSinceLastPromotion, y = YearsAtCompany)) +
  geom_point()
```




```r
ggplot(Sales, aes(x = MonthlyIncome, y = YearsAtCompany)) +
  geom_point()
```



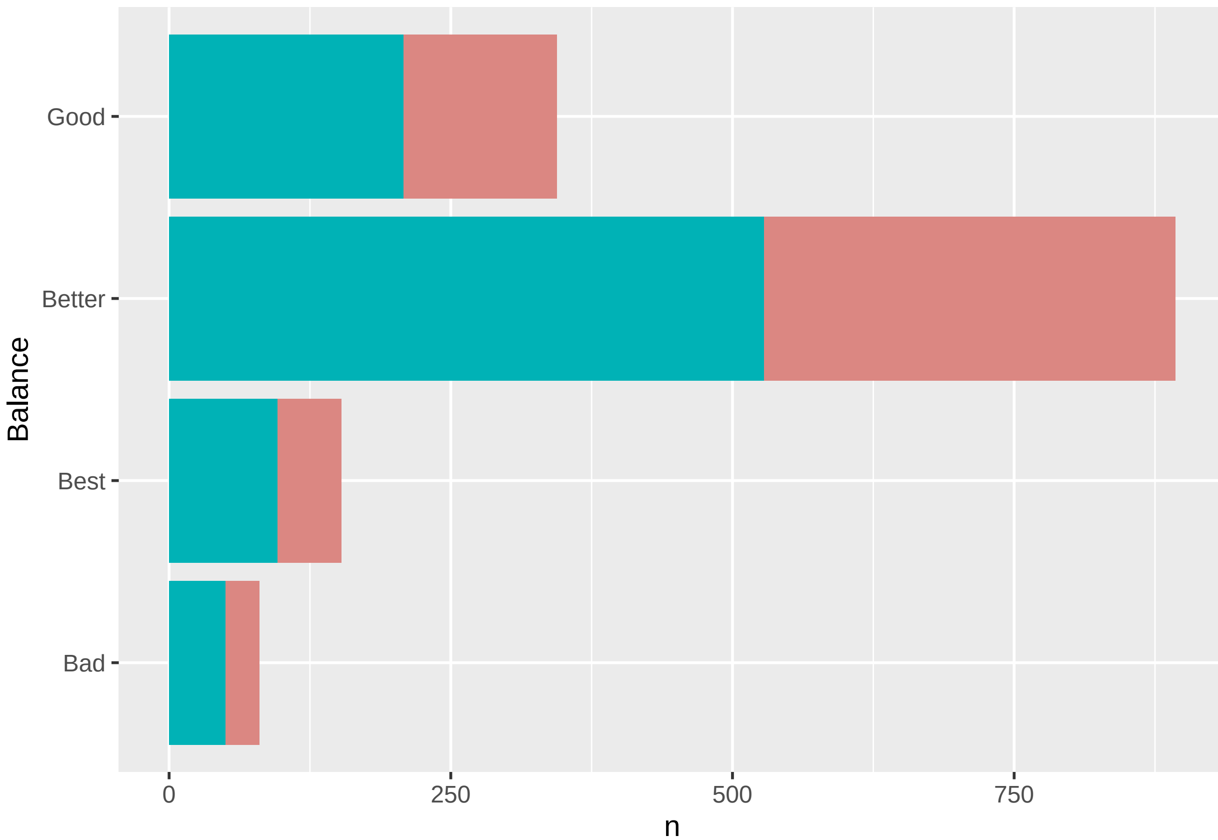
JOB SATISFACTION

Relationship Between Length at IBM & Work-life Balance

```
B <- IBM %>%
  select(WorkLifeBalance, YearsAtCompany, Gender, JobRole) %>%
  mutate(Balance = case_when(
    WorkLifeBalance == 1 ~ "Bad",
    WorkLifeBalance == 2 ~ "Good",
    WorkLifeBalance == 3 ~ "Better",
    WorkLifeBalance == 4 ~ "Best",
  )) %>%
  count(Balance, Gender)
B
```

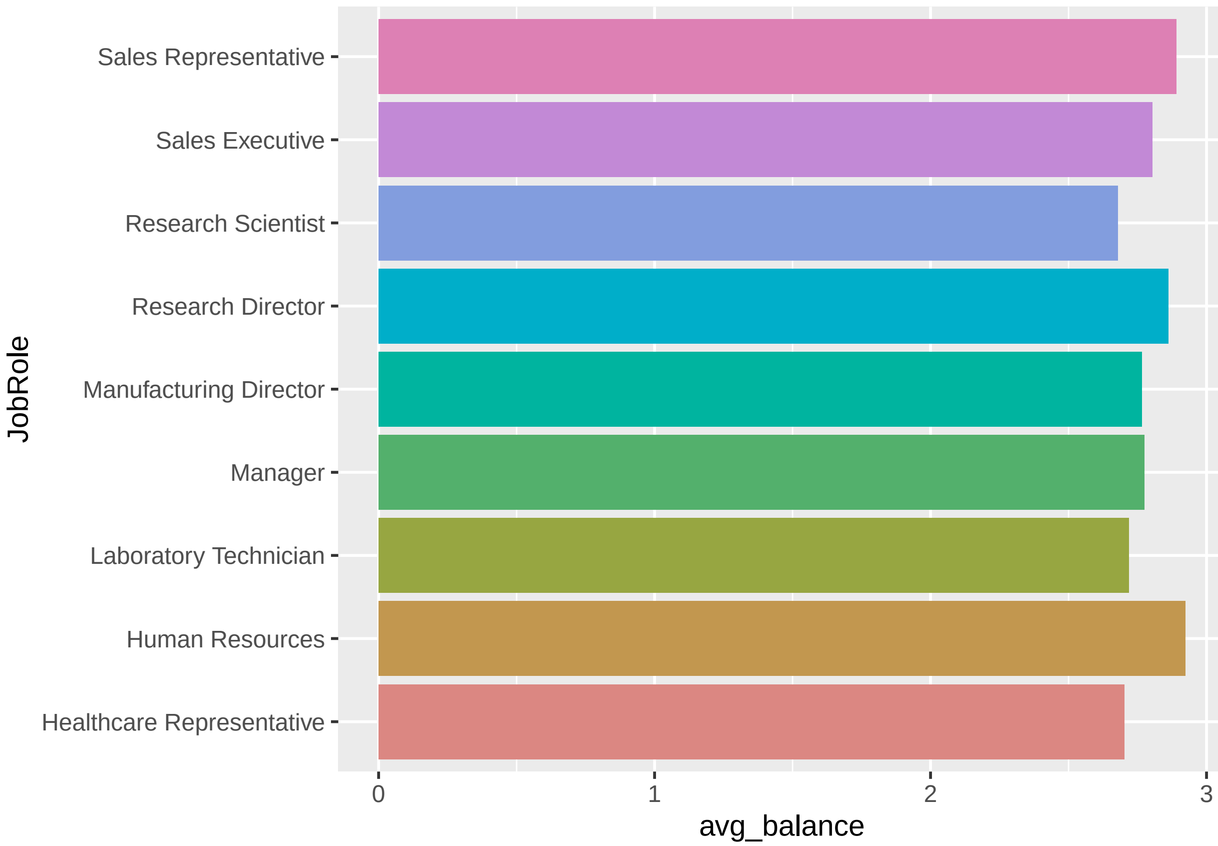
```
## # A tibble: 8 x 3
##   Balance Gender     n
##   <chr>   <chr>  <int>
## 1 Bad     Female    30
## 2 Bad     Male      50
## 3 Best    Female    57
## 4 Best    Male      96
## 5 Better  Female   365
## 6 Better  Male     528
## 7 Good    Female   136
## 8 Good    Male     208
```

```r
ggplot(B, aes(x = Balance, y = n, fill = Gender)) +
  geom_bar(stat = "identity") +
  scale_fill_hue(c = 60) +
  theme(legend.position="none") +
  coord_flip()
```




```r
B1 <- IBM %>%
  select(WorkLifeBalance, YearsAtCompany, Gender, JobRole) %>%
  group_by(JobRole) %>%
  summarize(avg_balance = mean(WorkLifeBalance))

ggplot(B1, aes(x = JobRole, y = avg_balance, fill = JobRole)) +
  geom_bar(stat = "identity") +
  scale_fill_hue(c = 60) +
  theme(legend.position="none") +
  coord_flip()
```



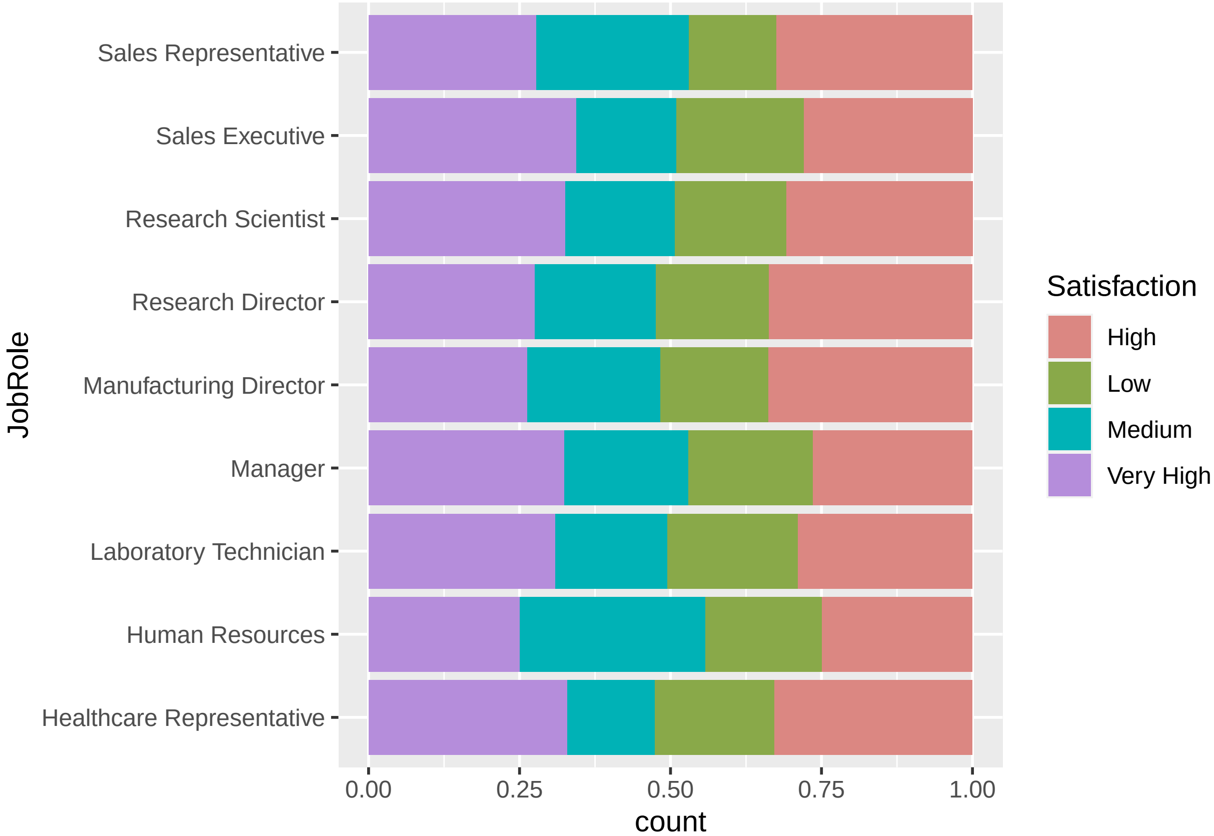

```
satisfaction <- IBM %>%
  select(JobRole, JobSatisfaction) %>%
  arrange(JobSatisfaction) %>%
  mutate(Satisfaction = case_when(
    JobSatisfaction == 1 ~ "Low",
    JobSatisfaction == 2 ~ "Medium",
    JobSatisfaction == 3 ~ "High",
    JobSatisfaction == 4 ~ "Very High"
  ))
satisfaction
```

Satisfaction

```
## # A tibble: 1,470 x 3
##    JobRole                JobSatisfaction Satisfaction
##    <chr>                            <dbl> <chr>
##  1 Laboratory Technician                1 Low
##  2 Manufacturing Director               1 Low
##  3 Sales Representative                 1 Low
##  4 Research Scientist                   1 Low
##  5 Research Scientist                   1 Low
##  6 Manager                              1 Low
##  7 Research Scientist                   1 Low
##  8 Sales Executive                      1 Low
##  9 Laboratory Technician                1 Low
## 10 Sales Executive                      1 Low
```

```
## # ... with 1,460 more rows
```

```
# satisfaction across roles
ggplot(satisfaction, aes(x = JobRole, fill = Satisfaction)) +
  geom_bar(position="fill", stat="count") +
  scale_fill_hue(c = 60) +
  coord_flip()
```




```
model_satisfaction <- lm(JobSatisfaction ~ as.factor(BusinessTravel) + DistanceFromHome  + YearsSinceLa
tidy(model_satisfaction)
```

```
## # A tibble: 6 x 5
##   term                                 estimate std.error statistic   p.value
##   <chr>                                   <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)                           2.81        1.04e-1   26.9    3.30e-130
## 2 as.factor(BusinessTravel)Travel_Fr~ -0.00328      1.12e-1   -0.0293 9.77e-  1
## 3 as.factor(BusinessTravel)Travel_Ra~ -0.0928       9.65e-2   -0.962  3.36e-  1
## 4 DistanceFromHome                    -0.000573     3.55e-3   -0.161  8.72e-  1
## 5 YearsSinceLastPromotion             -0.00678      9.53e-3   -0.711  4.77e-  1
## 6 MonthlyIncome                         0.000000246 6.53e-6    0.0376 9.70e-  1
```

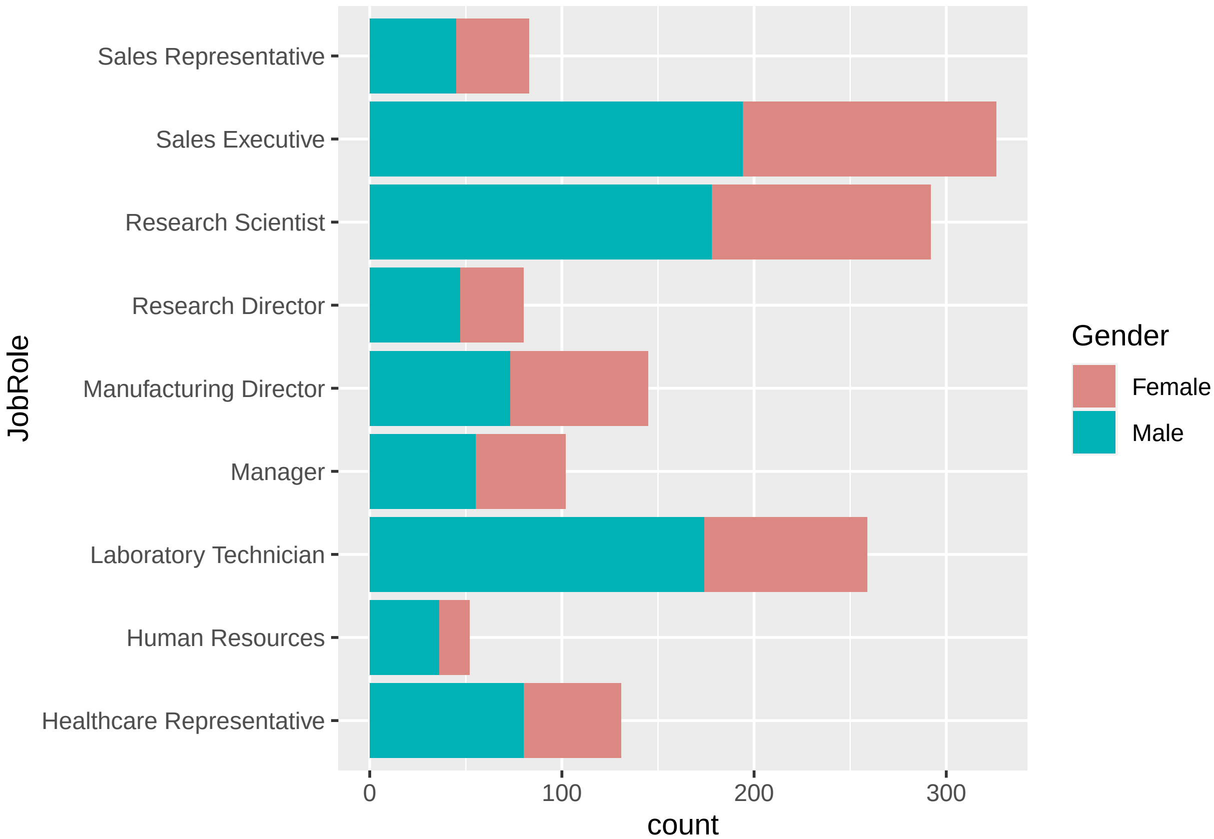
Suggestion: mostly high job satisfaction in all roles. None of the known factors is correlated.

DIVERSITY & INCLUSION

```
role <- IBM %>%
  select(JobRole, Gender)
```

```r
salary <- IBM %>%
  select(Gender, MonthlyIncome) %>%
  group_by(Gender) %>%
  summarize(avg_salary = mean(MonthlyIncome))

ggplot(role, aes(x = JobRole, fill = Gender)) +
  geom_bar(position="stack", stat="count") +
  scale_fill_hue(c = 60) +
  coord_flip()
```



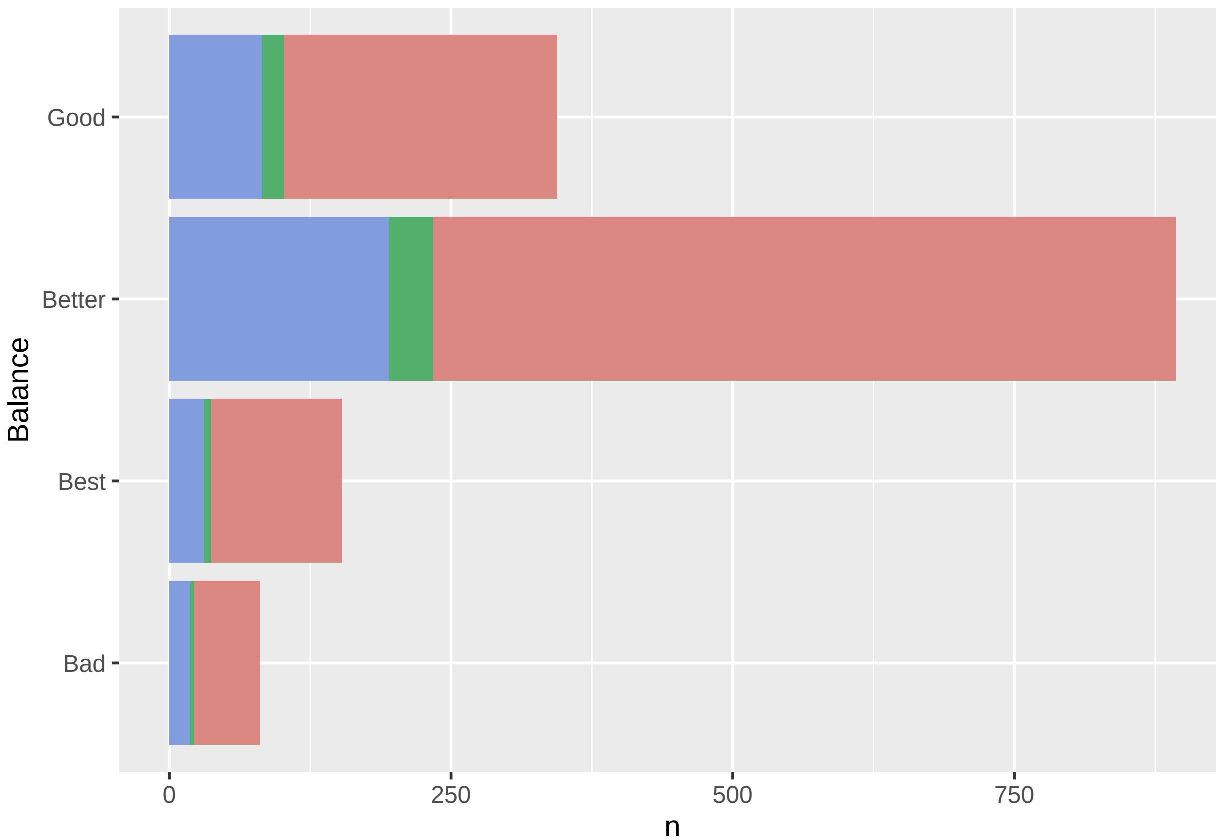

```r
ggplot(data=IBM, aes(x=YearsAtCompany, group=Gender, fill=Gender)) +
  geom_density(adjust=1.5, alpha=.4)
```


```
# Age
X <- IBM %>%
  select(WorkLifeBalance, Age) %>%
  mutate(Balance = case_when(
    WorkLifeBalance == 1 ~ "Bad",
    WorkLifeBalance == 2 ~ "Good",
    WorkLifeBalance == 3 ~ "Better",
    WorkLifeBalance == 4 ~ "Best",
  )) %>%
  mutate(AgeFactored = case_when(
    Age < 30 ~ "Young (<30)",
    Age >= 30 & Age < 55 ~ "Middle Age (30-55)",
    Age >= 55 & Age < 65 ~ "Senior (55+)",
  )) %>%
  count(Balance, AgeFactored)
X
```

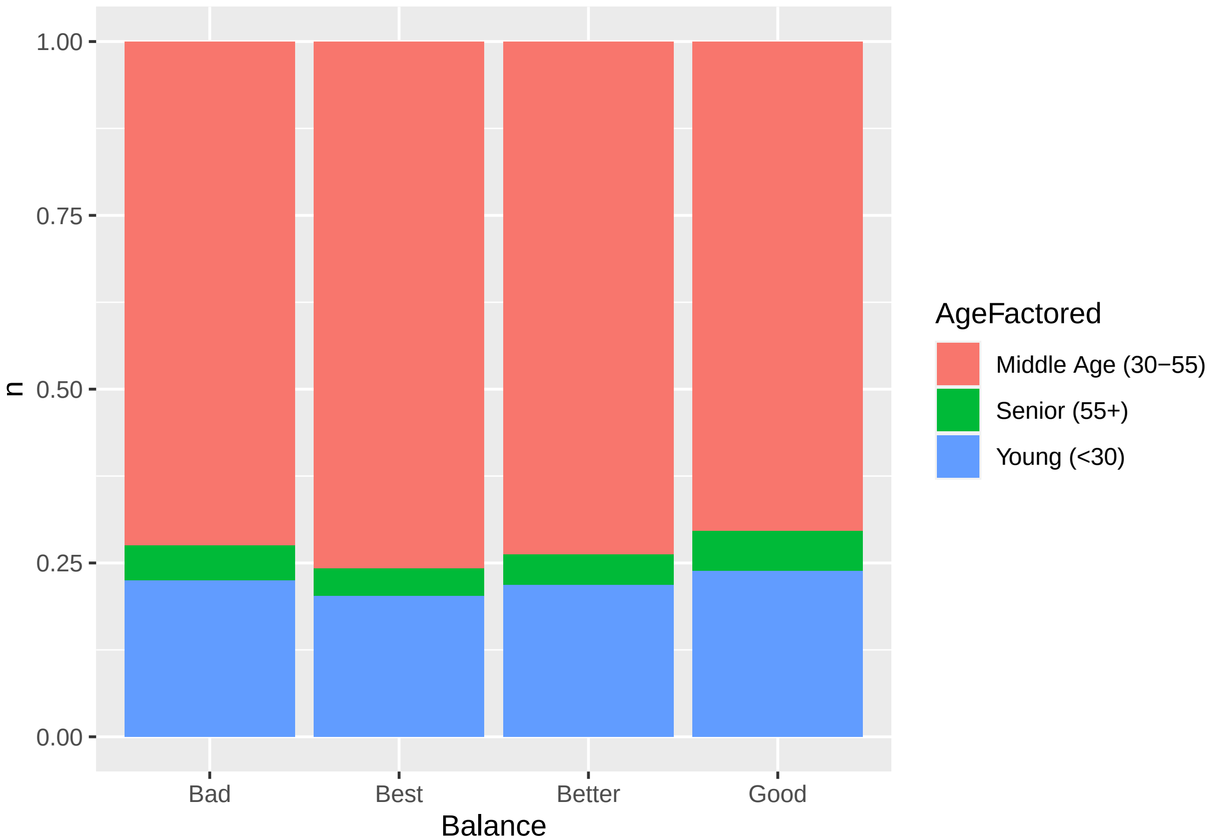
```
## # A tibble: 12 x 3
##    Balance AgeFactored            n
##    <chr>   <chr>              <int>
##  1 Bad     Middle Age (30-55)    58
##  2 Bad     Senior (55+)           4
##  3 Bad     Young (<30)           18
##  4 Best    Middle Age (30-55)   116
##  5 Best    Senior (55+)           6
##  6 Best    Young (<30)           31
##  7 Better  Middle Age (30-55)   659
```

```
##   8 Better  Senior (55+)         39
##   9 Better  Young (<30)         195
## 10 Good    Middle Age (30-55)   242
## 11 Good    Senior (55+)          20
## 12 Good    Young (<30)           82
```

```r
ggplot(X, aes(x = Balance, y = n, fill = AgeFactored)) +
  geom_bar(stat = "identity") +
  scale_fill_hue(c = 60) +
  theme(legend.position="none") +
  coord_flip()
```




```r
ggplot(X, aes(fill = AgeFactored, y = n, x = Balance)) +
    geom_bar(position="fill", stat="identity")
```



Comment: generally doing pretty well in terms of diversity & inclusion Suggestion: overall ok, but can higher more female in certain departments & roles. Noticed that there are only two categories. Maybe can expand the umbrella.