Mid-Term Data Analysis Project

Jiwon Mok

04/03/2022

## 1. Introduction.

The project aims to perform the initial data analysis for a U.S-based company to find the fundamental problem that employees are leaving the company and to come up with solutions to prevent people from leaving the company and further reduce the turnover rate in the end. In the recent past, during and after the pandemic, a lot of companies had their employees work remotely, and some expanded their business to hire more employees by increasing 1.5% or above salaries from their previous workplaces. In that regard, people come to think of the time they used to consume for commuting and start looking for companies where they can work remotely or located near their homes. Being a U.S-based company and a tech company, employee retention is challenging since there has been a flooding demand for tech positions in the job markets. Therefore, this report will analyze if there is any relationship between turnover rate and distance from the workplace to employees’ homes.

Appendix\_1

library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(scales)  
library(tidyverse)

## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.1 ──

## ✓ ggplot2 3.3.5 ✓ purrr 0.3.4  
## ✓ tibble 3.1.4 ✓ stringr 1.4.0  
## ✓ tidyr 1.1.3 ✓ forcats 0.5.1  
## ✓ readr 2.0.1

## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## x readr::col\_factor() masks scales::col\_factor()  
## x purrr::discard() masks scales::discard()  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

employee\_data <- read\_csv("~/Desktop/EmployeeData.csv")

## Rows: 1470 Columns: 13

## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (7): left\_company, department, job\_level, business\_travel, job\_satisfact...  
## dbl (6): salary, weekly\_hours, yrs\_at\_company, yrs\_since\_promotion, previous...

##   
## 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.

## 2. Data and Model

1. Is the distance from workplace to home the significant factor contributing to the increasing turnover rate to employees? (Divide the section into the following categories.)

No, proximity is not a significant factor contributing to the turnover rate of employees. Based on the numbers in the table, the average miles of employees who left the company are 10.6 while people who have not left live 8.9 on average from the workplace. Even though people who left the company lived 1.1 miles further away from the company, it is not a significant difference. However, it was true that there was more number of current employees proportioned to the points that they lived within 10 miles from the workplace at 71% while people who quit the job make up for 60%.

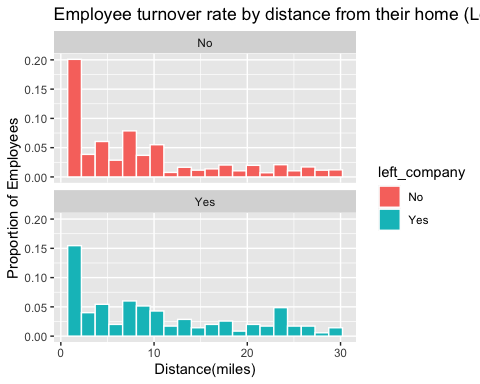
Appendix\_2(A) employee data

employee\_data %>% group\_by(left\_company) %>%  
 summarise(n\_employees = n(),  
 min\_miles = min(miles\_from\_home),  
 avg\_miles = mean(miles\_from\_home),  
 max\_miles = max(miles\_from\_home),  
 sd\_miles = sd(miles\_from\_home),  
 miles\_less\_10 = mean(miles\_from\_home <=10),  
 miles\_less\_20 = mean(miles\_from\_home <=20 & miles\_from\_home >10),  
 miles\_less\_30 = mean(miles\_from\_home <=30 & miles\_from\_home >20),  
 miles\_above\_30 = mean(miles\_from\_home >30))

## # A tibble: 2 × 10  
## left\_company n\_employees min\_miles avg\_miles max\_miles sd\_miles miles\_less\_10  
## <chr> <int> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 No 1233 1 8.92 29 8.01 0.715  
## 2 Yes 237 1 10.6 29 8.45 0.608  
## # … with 3 more variables: miles\_less\_20 <dbl>, miles\_less\_30 <dbl>,  
## # miles\_above\_30 <dbl>

Appendix\_2(B) visualization\_employee\_data

ggplot(data = employee\_data, aes(x = miles\_from\_home, fill = left\_company)) +  
 geom\_histogram(aes(y= ..density..), color = "white", bins = 20) +  
 facet\_wrap( ~ left\_company, nrow = 2) +  
 labs(title = "Employee turnover rate by distance from their home (Left the Company-Yes/No)",  
 x = "Distance(miles)", y= "Proportion of Employees")



Such a trend is also apparent even in the histogram graph above. If the distance is the most contributing factor, the numbers of longer miles should be proportioned more than the ones with shorter miles. However, the majority of employees living less than 10miles are ranged from 60 to 70 for both who have not left and left the company. Therefore, distance cannot be considered as the factor that people left the company.

1. If the distance is not a significant factor, how do employees in general value how far it is from home to company? Can you answer it without running any predictive models?

Employees in general live close to the company. The number of employees who have not left the company is 1233. The average number of miles from home to the company is 8.9, which is within 10miles. Moreover, the majority of them (70%) live less than 10 miles, and none of them live 30 miles further away from the workplace. It was clear to see in the table without running any predictive models. However, it is more obvious to see in the plot.

1. If distance is not a significant factor, can job-satisfaction be another contributing factor?

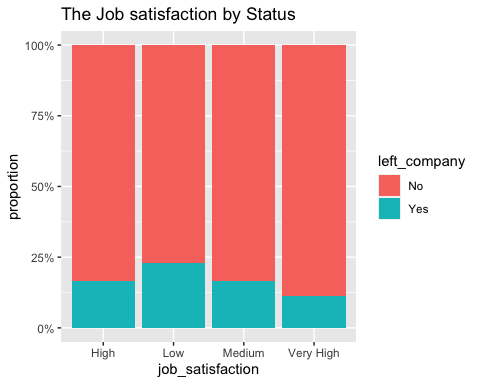
Appendix\_3(A) table\_relationship between job\_Satisfaction and left\_company

job\_satisfaction2 <- employee\_data %>% group\_by(left\_company) %>%  
 count(job\_satisfaction)  
job\_satisfaction2

## # A tibble: 8 × 3  
## # Groups: left\_company [2]  
## left\_company job\_satisfaction n  
## <chr> <chr> <int>  
## 1 No High 369  
## 2 No Low 223  
## 3 No Medium 234  
## 4 No Very High 407  
## 5 Yes High 73  
## 6 Yes Low 66  
## 7 Yes Medium 46  
## 8 Yes Very High 52

Appendix\_3(B) visualization ggplot\_job satisfaction

#Data Visualization   
ggplot(job\_satisfaction2, aes(job\_satisfaction, n, fill = left\_company)) +  
 geom\_bar(position = "fill", stat="identity") +  
 scale\_y\_continuous(labels=percent) +  
 labs( y="proportion",   
 title ="The Job satisfaction by Status")

 Yes. Job satisfaction can be a significant factor contributing to the turnover rate based on the numbers on the graph. Unlike the distance between home and workplace, job\_satisfaction shows the significant difference that people who left the company make up for more than 70% in high and very high parts. Even though there are about 75% of current employees who have not left the company have low job satisfaction while who left account for one-third of them.

1. If job satisfaction is one of the contributing factors, can you answer it without running any predictive models?

It is possible to answer that employees with a high satisfaction rate are more likely to stay at the company. However, it needs additional calculations for in-depth analysis. Since more employees left the company, their numbers are bigger. 407 employees answered “very high” in job satisfaction, and 369 said high. Considering the overall number of employees is 1470, they account for half, and job satisfaction is a huge contributing factor. However, it is more clear to see the proportion in the plot since no additional calculation is needed and different colors easily distinguish them.

1. If there a relationship between employees leaving the company and their current salary? (Scatter plot)

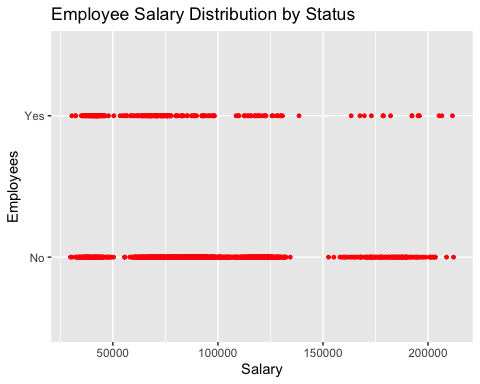
Appendix\_4(A)employee’s salary data

employee\_data %>% group\_by(left\_company) %>%  
 summarise(n\_employees = n(),  
 min\_salary = min(salary),  
 avg\_salary = mean(salary),  
 max\_salary = max(salary),  
 sd\_salary = sd(salary),  
 pct\_less\_60k = mean(salary <= 60000))

## # A tibble: 2 × 7  
## left\_company n\_employees min\_salary avg\_salary max\_salary sd\_salary  
## <chr> <int> <dbl> <dbl> <dbl> <dbl>  
## 1 No 1233 29849. 97431. 212135. 36470.  
## 2 Yes 237 30488. 76626. 211621. 38567.  
## # … with 1 more variable: pct\_less\_60k <dbl>

Appendix\_4(B) ggplot\_ employee’s salary data

ggplot(data = employee\_data, aes(x = salary, y= left\_company)) +  
 geom\_point(color = "red", size = 1)+  
 labs(title = "Employee Salary Distribution by Status", x = "Salary", y="Employees")

 Yes. It is quite clear to see in the plot that salary is the contributing factor leading to the high turnover rate of the company. In the scatter plot, in the salary range from 15k to 20k, employees who have not left the company are densely populated while only a few people who left are in the range. Moreover, there is a tendency that a lot of them who quit are mostly populated in the salary range from around 50k to 100k accounting for 34%. On the other hand, employees’ salary range seems to be constant from 50k to 130k making up for less than 10%(9.7%).

1. If the salary is one of the contributing factors, which range of salary is considered as low which likely leads to high turnover rate?

Appendix\_5 employee’s salary range

employee\_data %>% group\_by(left\_company) %>%  
 summarise(n\_employees = n(),  
 min\_salary = min(salary),  
 avg\_salary = mean(salary),  
 max\_salary = max(salary),  
 sd\_salary = sd(salary),  
 pct\_less\_60k = mean(salary <= 60000),  
 pct\_less\_80k = mean(salary > 60000 & salary <= 80000),  
 pct\_less\_100k = mean(salary > 80000 & salary <= 100000),  
 pct\_less\_120k = mean(salary > 100000 & salary <= 1200000),  
 pct\_above\_121k = mean(salary >1200000))

## # A tibble: 2 × 11  
## left\_company n\_employees min\_salary avg\_salary max\_salary sd\_salary  
## <chr> <int> <dbl> <dbl> <dbl> <dbl>  
## 1 No 1233 29849. 97431. 212135. 36470.  
## 2 Yes 237 30488. 76626. 211621. 38567.  
## # … with 5 more variables: pct\_less\_60k <dbl>, pct\_less\_80k <dbl>,  
## # pct\_less\_100k <dbl>, pct\_less\_120k <dbl>, pct\_above\_121k <dbl>

Since there is a significant difference in the number of employees who receive less than 60k, employees who have received more than 60k might be likely to go home. There was more proportion in salary range from 61k to 80k at 34% among people who left while 22% of employees who have not left receive 61k to 80k. Therefore, this salary cannot be the factor contributing to the turnover rate. There is a more significant difference from 81k to 120k as 67% of employees receive higher salaries while a third of them who left the company used to receive such salary.

## 3. Conclusion(s)/Discussion.

In analysis, people are more likely to stay at the company compared to when the job\_satisfaction is lower. Moreover, employees who quit their job tend to live 1.1 miles further away from the ones who currently work. Compared to distance, job\_satisfaction and salary are more likely to be the determining factor when it comes to the turnover rate. However, considering current employees’ low job satisfaction is higher than those who quit, the relationship between job\_satisfaction and turnover rate cannot be considered positive. When it comes to salary, a third of those who quit the job were proportioned in salary ranging less than 60k while a few current employees were receiving less than 60k. Most current employees at work receive 100k - 120k. Therefore, to prevent an increasing turnover rate, the company will have to figure out a way to improve employees’ welfare more than relocating their company to the center where a lot of employees will likely reside. Moreover, the company should higher the overall salary who currently receive less than 100k.

## 4. Appendix/Appendices.

This is the section, where you will provide the detailed information on the following.

Appendix\_1 - package Appendix\_2(A) employee data Appendix\_2(B) visualization\_employee\_data Appendix\_3(A) table\_relationship between job\_Satisfaction and left\_company Appendix\_3(B) visualization ggplot\_job satisfaction Appendix\_4(A)employee’s salary data Appendix\_4(B) ggplot\_ employee’s salary data Appendix\_5 table\_employee’s salary range