**Sarang Tirmanwar**

**2023-10-08**

**INTRODUCTION**

In this mid-semester project, we delve into the field of Human Resources Analytics, aiming to address the critical issue of employee attrition within a U.S.-based tech company. With a dataset containing 1,200 employee records, we embark on an initial data analysis to understand the factors contributing to employee departures. This is pivotal for the company, as high attrition rates can lead to significant costs and disruptions in operations.

**PROBLEM DESCRIPTION**

The company has experienced an uptick in employee departures, particularly in a competitive tech landscape where certain skill sets are in high demand. This has had a detrimental effect on the company’s ability to deliver quality products and maintain operational efficiency

**SUMMARY OF DATA**

The dataset comprises various attributes of employees, such as department, job level, salary, weekly hours, business travel, tenure at the company, years since last promotion, and more. These variables are crucial in understanding the dynamics surrounding employee attrition.

Employee attrition status (left\_company): Yes/No  
Department: Department within the company   
Job level: Associate - Vice President   
Salary: Employee yearly salary (US Dollars)   
Job satisfaction: Self-reported job satisfaction (company survey)  
Performance rating: Most recent annual performance rating   
Marital status: Single, Married, or Divorced   
Business travel: Level of required business travel   
Years at company: Tenure at the company (years)   
Years since promotion: Years since last promotion   
Previous companies: Number of previous companies the employee has worked for   
Miles from home: Distance from employee address to office location   
Weekly hours: Self-reported average weekly hours spent on the job (company survey)

**BIG QUESTIONS**

What factors contribute to employees leaving the company, based on summary statistics? Can we provide insights for top management, using plots alone and without running predictive models?

**EXPLORATORY DATA ANALYSIS**

Utilizing commands like dim(), str(), head(), glimpse(), skim(), and summary() for initial data exploration.

Formulating five questions exploring the relationship between employee attrition and other variables.

Investigating relationships between leaving the company and department, as well as department and job level combined.

Providing insights through summary tables and graphs, drawing conclusions on variables association with employee attrition.

**ANALYSIS AND RESULTS**

**Question 1: Is there a relationship between leaving the company and department?**

Answer: Yes, there does appear to be a relationship between leaving the company and the department. For instance, the “Research” department has a notably low attrition rate, with only 8 employees leaving. Conversely, the “Sales” department experiences a much higher departure rate, with 54 employees leaving

**Question 2: Is there a relationship between leaving the company and department and job level together?**

Answer: Yes, there is a clear relationship between leaving the company and the combination of department and job level. For example, among Associates, 5 stayed while 7 left. For Directors, 17 stayed and only 2 left. This pattern persists across various job levels. In short, the likelihood of an employee leaving is significantly influenced by their department and job level.

**Question 3: Are marital status and salary factors for leaving?**

Answer: The graph indicates that employees’ marital status doesn’t significantly influence their likelihood to resign. However, salary plays a crucial role. Both single and married employees appear to resign for similar salary ranges. On the other hand, even with higher salaries, divorced employees might still leave the company. This suggests that for them, salary isn’t the primary factor in their decision to leave. It implies that, at this stage of their

**Question 4: Is there a relationship between department and salary?**

Answer: It’s evident that marital status doesn’t show a significant correlation with employee resignations. Both single and married employees appear to leave their positions regardless of their marital status. However, it’s apparent that salary plays a pivotal role. Both single and married employees who resign seem to do so for similar salary considerations. Interestingly, even though divorced employees tend to have higher salaries, this doesn’t necessarily deter them from leaving the company. This suggests that salary might not be their primary focus at this stage in their post-divorce phase.

**Question 5: Is there any relationship between employees leaving the company and their weekly hours?**

Answer: Yes, there is a clear connection between employees departing from the company and their weekly working hours. The table demonstrates a noticeable disparity in average weekly hours between those who stayed and those who left. Employees who remained with the company worked less than those who chose to leave.

**Question 6: Is there a relationship between employees leaving the company and their job level?**

Answer: Yes, there is a correlation between employee attrition and their job position. The data depicted in the “Leaving-position” chart indicates that individuals in foundational roles such as associates and managers exhibit a greater likelihood of leaving the company. Conversely, employees in elevated positions are more inclined to remain, particularly evident in the notable disparity between managers and senior managers.

**Question 7: Is there a relationship between job\_satisfaction and salary?**

Answer: It is evident that contented employees are less inclined to seek opportunities elsewhere. Additionally, there’s a correlation between higher salaries, job satisfaction, and reduced likelihood of leaving. However, intriguingly, not all individuals in lower-paying and less satisfying roles express a desire to switch jobs. The majority of them are content to remain in their current positions. Identifying these employees and comprehending the reasons behind their reluctance to leave can furnish valuable insights for the HR department.

**Question 8: Question: Is there a relationship between yrs\_at\_company and salary? Does the relationship affect the employee’s willingness to leave their company?**

Answer: Looking at below chart, it’s clear that salary and weekly working hours don’t have a straightforward connection. This might be due to the diversity of job roles. However, in the bottom-left section, we can see that employees with lower salaries and less experience are more likely to leave. When salaries increase, even new employees show a stronger inclination to stay. So, it’s a good idea for the HR department to identify promising talents early on and offer competitive salaries to help retain them.

**Question 9: Is there a relationship between employees leaving the company and their current salary?**

Answer: The data shows a notable distinction between employees who stay with the company and those who leave. Those who left generally had lower salaries, while the remaining employees had higher average salaries. Specifically, a larger proportion of departing employees had salaries below a certain threshold compared to those who stayed.

**CONCLUSION**

The key findings from the Exploratory Data Analysis that stood out to me are:

I discovered that departmental affiliation strongly impacts attrition rates. “Sales” experienced higher departures, while “Research” had lower attrition. Job levels also play a crucial role; Associates and Managers were more likely to leave, emphasizing the need for clear career growth paths.

Marital status showed minimal correlation with resignations, but salary emerged as a key motivator. Both single and married employees who left cited similar salary concerns, highlighting the importance of competitive compensation packages.

Workload management and work-life balance initiatives may be needed, as employees who stayed tended to work fewer hours. Competitive entry-level salaries are crucial, as lower-paid employees were more likely to leave.

Additionally, job satisfaction is vital for retention. Contented employees are less likely to leave, and higher salaries are linked to greater job satisfaction. Addressing job satisfaction factors can be crucial in retaining talent.

These findings are important because they provide actionable insights for the company. By understanding the specific factors influencing attrition, executives can implement targeted strategies to improve retention rates, ultimately enhancing productivity and reducing operational costs associated with turnover.

Recommendations:

Develop department-specific retention strategies based on observed attrition rates. Provide targeted career development opportunities and recognize achievements.

Ensure competitive salaries, especially for entry-level positions. Conduct regular reviews and adjustments based on performance to retain valuable talent.

Implement policies promoting a healthy work-life balance, including flexible work arrangements, wellness programs, and workload management initiatives.

Offer clear paths for career advancement, mentorship programs, and training opportunities to empower employees and increase retention.

Foster a positive work environment through employee recognition programs. Address factors influencing job satisfaction to enhance employee engagement and retention.

Conduct thorough exit interviews to understand reasons for employee departures. Utilize feedback to improve HR processes and address underlying issues affecting retention.

#Loading the dataset  
employee <- read.csv("F:/GBUS 738/Emp.csv")  
# First five rows of the 'employee' dataset  
head(employee)

## left\_company department job\_level salary weekly\_hours  
## 1 Yes Sales Director 118680.74 56  
## 2 No Sales Senior Manager 85576.44 42  
## 3 Yes Product Development Associate 46235.79 56  
## 4 No IT and Analytics Director 117226.84 50  
## 5 No Sales Associate 36634.73 46  
## 6 No Marketing Senior Manager 83519.59 48  
## business\_travel yrs\_at\_company yrs\_since\_promotion previous\_companies  
## 1 Rarely 6 0 5  
## 2 Frequently 10 1 5  
## 3 Rarely 0 0 6  
## 4 Frequently 8 3 1  
## 5 Rarely 2 2 2  
## 6 Frequently 7 3 1  
## job\_satisfaction performance\_rating marital\_status miles\_from\_home  
## 1 Very High Meets Expectations Single 1  
## 2 Medium Exceeds Expectations Married 8  
## 3 High Minimally Effective Single 2  
## 4 High Exceptional Married 3  
## 5 Medium Exceeds Expectations Married 2  
## 6 Very High Meets Expectations Single 2

#importing ggplot library  
library(ggplot2)  
#importing dplyr library  
library(dplyr)

## Warning: package 'dplyr' was built under R version 4.2.2

##   
## 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

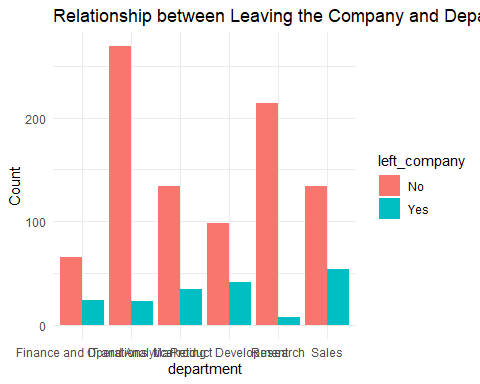
summary\_table\_dept <- employee %>%  
 group\_by(department, left\_company) %>%  
 summarise(count = n())

## `summarise()` has grouped output by 'department'. You can override using the  
## `.groups` argument.

# Print the summary table  
print(summary\_table\_dept)

## # A tibble: 12 × 3  
## # Groups: department [6]  
## department left\_company count  
## <chr> <chr> <int>  
## 1 Finance and Operations No 66  
## 2 Finance and Operations Yes 24  
## 3 IT and Analytics No 269  
## 4 IT and Analytics Yes 23  
## 5 Marketing No 134  
## 6 Marketing Yes 35  
## 7 Product Development No 98  
## 8 Product Development Yes 41  
## 9 Research No 214  
## 10 Research Yes 8  
## 11 Sales No 134  
## 12 Sales Yes 54

# Data Visualization  
ggplot(summary\_table\_dept, aes(x = department, y = count, fill = left\_company)) +  
 geom\_bar(stat = "identity", position = "dodge") +  
 labs(title = "Relationship between Leaving the Company and Department",  
 y = "Count") +  
 theme\_minimal()



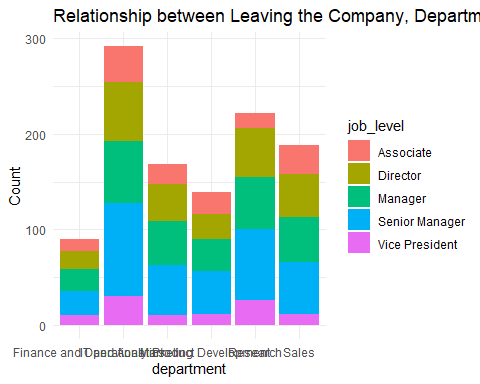
# Calculate the count of employees who left or stayed by department and job level  
summary\_table\_dept\_level <- employee %>%  
 group\_by(department, job\_level, left\_company) %>%  
 summarise(count = n())

## `summarise()` has grouped output by 'department', 'job\_level'. You can override  
## using the `.groups` argument.

# Print the summary table  
print(summary\_table\_dept\_level)

## # A tibble: 58 × 4  
## # Groups: department, job\_level [30]  
## department job\_level left\_company count  
## <chr> <chr> <chr> <int>  
## 1 Finance and Operations Associate No 5  
## 2 Finance and Operations Associate Yes 7  
## 3 Finance and Operations Director No 17  
## 4 Finance and Operations Director Yes 2  
## 5 Finance and Operations Manager No 16  
## 6 Finance and Operations Manager Yes 7  
## 7 Finance and Operations Senior Manager No 20  
## 8 Finance and Operations Senior Manager Yes 6  
## 9 Finance and Operations Vice President No 8  
## 10 Finance and Operations Vice President Yes 2  
## # … with 48 more rows

# Data Visualization  
# Create a stacked bar plot to visualize the relationship  
ggplot(summary\_table\_dept\_level, aes(x = department, y = count, fill = job\_level)) +  
 geom\_bar(stat = "identity", position = "stack") +  
 labs(title = "Relationship between Leaving the Company, Department, and Job Level",  
 y = "Count") +  
 theme\_minimal()



Q3 <- employee %>% group\_by(left\_company,marital\_status)%>% select(left\_company,marital\_status,salary) %>%   
 summarise( quantile1\_salarye = quantile(salary, 1 / 4),  
 mean\_salary = mean(salary),  
 quantile3\_salary = quantile(salary, 3 / 4)) %>%  
 arrange(marital\_status,desc(left\_company))

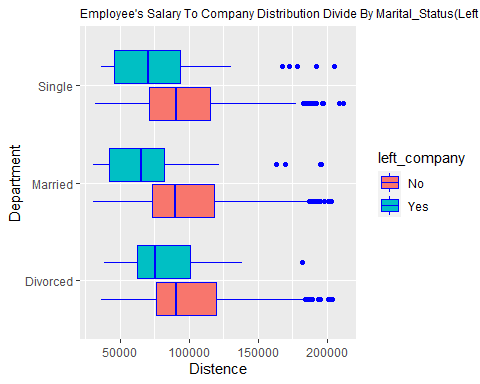
## `summarise()` has grouped output by 'left\_company'. You can override using the  
## `.groups` argument.

# Print the summary table  
print(Q3)

## # A tibble: 6 × 5  
## # Groups: left\_company [2]  
## left\_company marital\_status quantile1\_salarye mean\_salary quantile3\_salary  
## <chr> <chr> <dbl> <dbl> <dbl>  
## 1 Yes Divorced 62469. 82923. 100828.  
## 2 No Divorced 76750. 98984. 119584.  
## 3 Yes Married 42056. 72150. 81964.  
## 4 No Married 73372. 97506. 118522.  
## 5 Yes Single 45977. 78263. 94103.  
## 6 No Single 71197. 95048. 115487.

# Data Visualization  
ggplot(data = employee, aes(x=salary, fill = left\_company))+  
 geom\_boxplot(aes (y = marital\_status),color = "blue", bins = 30) +  
 labs(title = "Employee's Salary To Company Distribution Divide By Marital\_Status(Left the Comapny - Yes/No)",  
 x = "Distence", y = "Department")+  
 theme(plot.title = element\_text(size=9))

## Warning: Ignoring unknown parameters: bins



Q4 <- employee %>% group\_by(left\_company, department)%>% select(department,salary) %>%   
 summarise( quantile1\_salarye = quantile(salary, 1 / 4),  
 mean\_salary = mean(salary),  
 quantile3\_salary = quantile(salary, 3 / 4)) %>%  
 arrange(department,desc(left\_company))

## Adding missing grouping variables: `left\_company`

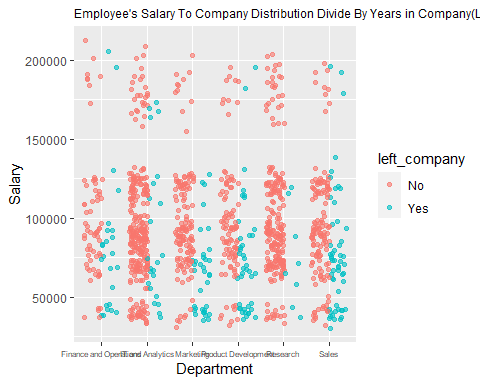
## `summarise()` has grouped output by 'left\_company'. You can override using the  
## `.groups` argument.

# Print the summary table  
print(Q4)

## # A tibble: 12 × 5  
## # Groups: left\_company [2]  
## left\_company department quantile1\_salarye mean\_salary quantile3…¹  
## <chr> <chr> <dbl> <dbl> <dbl>  
## 1 Yes Finance and Operations 45072. 82546. 92356.  
## 2 No Finance and Operations 76037. 101081. 117414.  
## 3 Yes IT and Analytics 57166. 89638. 117344.  
## 4 No IT and Analytics 70835. 94137. 115697.  
## 5 Yes Marketing 42151. 65817. 74551.  
## 6 No Marketing 78463. 98597. 119566.  
## 7 Yes Product Development 43108. 74717. 88983.  
## 8 No Product Development 77001. 98019. 116615.  
## 9 Yes Research 54563. 74904. 95352.  
## 10 No Research 73032. 99880. 119991.  
## 11 Yes Sales 42948. 77738. 91423.  
## 12 No Sales 72144. 94828. 117775.  
## # … with abbreviated variable name ¹​quantile3\_salary

# Data Visualization  
ggplot(data = employee, aes(x= department, color = left\_company))+  
 geom\_point(position=position\_jitterdodge(),alpha=.6,aes (y = salary), bins = 30) +  
 labs(title = "Employee's Salary To Company Distribution Divide By Years in Company(Left the Comapny - Yes/No)",  
 x = "Department", y = "Salary")+  
 theme(plot.title = element\_text(size=9),  
 axis.text.x = element\_text(size=6) )

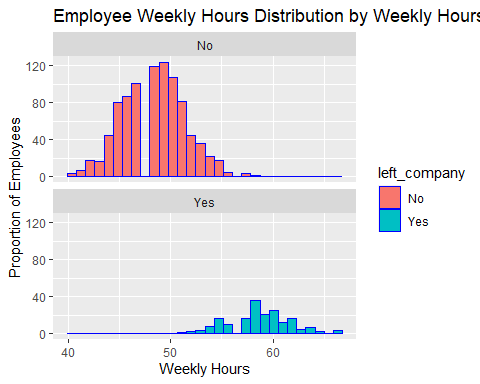
## Warning: Ignoring unknown parameters: bins



Q5 <- employee %>% group\_by(left\_company) %>%   
 summarise(n\_employees = n(),  
 min\_weekly\_hours = min(weekly\_hours),  
 avg\_weekly\_hours = mean(weekly\_hours),  
 max\_weekly\_hours = max(weekly\_hours),  
 sd\_weekly\_hours = sd(weekly\_hours))   
  
# Print the summary table  
print(Q5)

## # A tibble: 2 × 6  
## left\_company n\_employees min\_weekly\_hours avg\_weekly\_hours max\_weekl…¹ sd\_we…²  
## <chr> <int> <int> <dbl> <int> <dbl>  
## 1 No 915 40 48.4 58 3.05  
## 2 Yes 185 51 58.7 66 2.95  
## # … with abbreviated variable names ¹​max\_weekly\_hours, ²​sd\_weekly\_hours

# Data Visualization  
ggplot(data = employee, aes(x=weekly\_hours, fill = left\_company))+  
 geom\_histogram(color = "blue", bins = 30)+  
 facet\_wrap(~ left\_company, nrow = 2) +  
 labs(title = "Employee Weekly Hours Distribution by Weekly Hours (Left the Comapny - Yes/No)",  
 x = "Weekly Hours", y = "Proportion of Employees")



Q6 <- employee %>% group\_by(left\_company,job\_level)%>% select(left\_company,job\_level) %>% summarise(total = n())

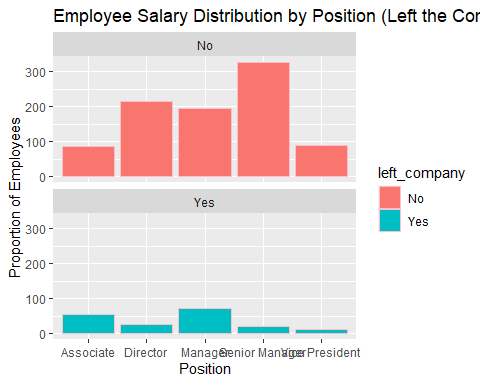
## `summarise()` has grouped output by 'left\_company'. You can override using the  
## `.groups` argument.

# Print the summary table  
print(Q6)

## # A tibble: 10 × 3  
## # Groups: left\_company [2]  
## left\_company job\_level total  
## <chr> <chr> <int>  
## 1 No Associate 85  
## 2 No Director 216  
## 3 No Manager 196  
## 4 No Senior Manager 328  
## 5 No Vice President 90  
## 6 Yes Associate 55  
## 7 Yes Director 25  
## 8 Yes Manager 72  
## 9 Yes Senior Manager 22  
## 10 Yes Vice President 11

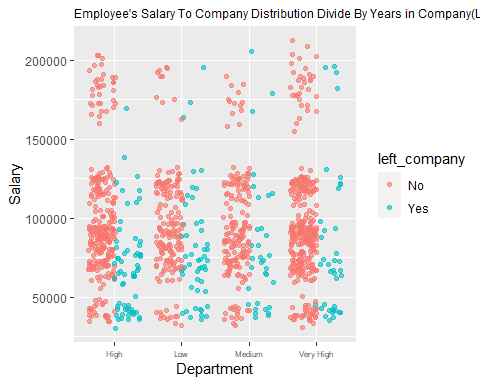
# Data Visualization  
ggplot(data = Q6, aes(x = job\_level, fill = left\_company)) +   
 geom\_histogram(stat = 'identity', aes(y = total), color = "pink") +  
 facet\_wrap(~ left\_company, nrow = 2) +  
 labs(title = "Employee Salary Distribution by Position (Left the Company - Yes/No)",  
 x = "Position", y = "Proportion of Employees")

## Warning: Ignoring unknown parameters: binwidth, bins, pad



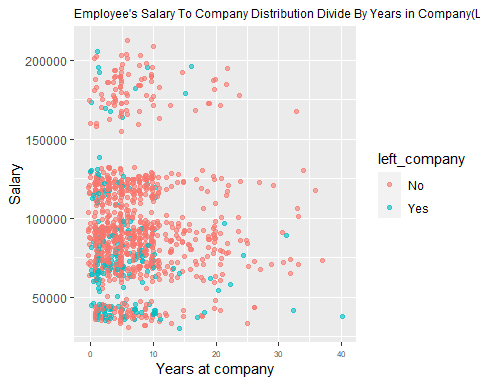
# 7 ## Data Visualization  
ggplot(data = employee, aes(x= job\_satisfaction, color = left\_company))+  
 geom\_point(position=position\_jitterdodge(),alpha=.6,aes (y = salary), bins = 30) +  
 labs(title = "Employee's Salary To Company Distribution Divide By Years in Company(Left the Comapny - Yes/No)",  
 x = "Department", y = "Salary")+  
 theme(plot.title = element\_text(size=9),  
 axis.text.x = element\_text(size=6) )

## Warning: Ignoring unknown parameters: bins



# 8 #Data Visualization  
ggplot(data = employee, aes(x= yrs\_at\_company, color = left\_company))+  
 geom\_point(position=position\_jitterdodge(),alpha=.6,aes (y = salary), bins = 30) +  
 labs(title = "Employee's Salary To Company Distribution Divide By Years in Company(Left the Comapny - Yes/No)",  
 x = "Years at company", y = "Salary")+  
 theme(plot.title = element\_text(size=9),  
 axis.text.x = element\_text(size=6) )

## Warning: Ignoring unknown parameters: bins



Q9 <- employee%>% 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))  
  
# Print Summary Table  
print (Q9)

## # A tibble: 2 × 7  
## left\_company n\_employees min\_salary avg\_salary max\_salary sd\_salary pct\_less…¹  
## <chr> <int> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 No 915 30559. 97151. 212135. 36875. 0.0995  
## 2 Yes 185 30488. 76794. 205267. 37626. 0.341   
## # … with abbreviated variable name ¹​pct\_less\_60k

#Data Visualization  
ggplot(data = employee, aes(x = salary, fill = left\_company)) +   
 geom\_histogram(aes(y = ..density..), color = "white", bins = 20) +  
 facet\_wrap(~ left\_company, nrow = 2) +  
 labs(title = "Employee Salary Distribution by Status (Left the Comapny - Yes/No)",  
 x = "Salary (US Dollars", y = "Proportion of Employees")

