Mid-Term Data Analysis Project Template

ZzYEEUN PARK

10/13/2022

# 1. Introduction. The introductions should include

Most companies want employees to work productively in the long run. As a company selects and hires employees, employees also often change jobs in consideration of the company’s welfare, salary, value, and so on. For businesses, employees are both human resources and valuable assets. If a company loses human resources, it suffers a lot of damage, such as the cost of replacing them, training new employees, and business interruption. However, employees are difficult to control and predict because they are people, not machines. Companies can conduct analysis through several assumptions to improve employee satisfaction and productivity and prevent employee decline. With data on employees, it is possible to find a correlation between some data through various assumptions and analyses. It will help identify employees’ points of view, psychology, and the cause of the turnover. In addition, companies can improve the reasons found through analysis and reduce and prevent employee turnover. Here are six possible assumptions the company can expect from employee data:

1. Is there a correlation between employee turnover and their working hours?
2. Is there a correlation between marital status and their work distance? Does satisfaction affect turnover?
3. Is there a correlation between employee turnover and job level?
4. Is there a correlation between employees’ years promotion and satisfaction? Does satisfaction affect turnover?
5. Is there a correlation between employees’ departments and salaries? Does the department affect the turnover?
6. Is there a correlation between employees’ marital status and salary? Did marital status affect the turnover?

In order to analyze these six assumptions, data on turnover status, working hours, marital status, workplace distance, employee satisfaction, position, department, and salary should be collected and compared first. Next, it is necessary to confirm their association by comparing and analyzing the data needed for each assumption. The degree of their association in the identification process allows us to determine whether the assumptions we anticipate are responsible for the consumption rate. For example, let’s say we analyze Assumption 1. We need data on employee turnover status and working hours, we can confirm the relationship between the two by comparing the two data. Suppose that during the confirmation process, employees with long working hours can confirm that the turnover rate is high. If so, it can be seen that the assumption is established and excessive working hours are the cause. Then, companies can adjust their employees’ working hours to avoid overwork, which can prevent staff cuts and improve employee well-being and productivity.

## 2. Data and Model

### 1. Is there a correlation between employee turnover and their working hours?

* Data:
  + left\_company shows whether employees have changed their job or not
  + weekly\_hours shows how many hours each employee worked per week
* Method: It shows their correlation by comparing employees’ weekly\_hours and left\_company. Within the employee data, employees’ working hours are separated by grouping them into turnover. Work hours are calculated by minimum, average, maximum, and standard deviation, and then the number of each employee is calculated. Finally, two histograms can be compared as a result through facet\_wrap.
* Analysis: The employees who have not changed jobs worked a minimum (40 hrs), average (48.37 hrs), maximum (58 hrs), and standard deviation (3.02). For employees who have changed jobs, minimum (51 hrs), average (58.61 hrs), maximum (66 hrs), and standard deviation (2.99). From these statistics, it can be seen that the higher the weekly working hours, the higher the probability of turnover. The number of people who have not changed jobs is 1,233, and the number of people who have changed jobs is 237. In addition, it shows that the number of people with longer working hours is significantly lower. In other words, it can be seen that working hours are not distributed equally.
* Results: Working hours are one of the big factors that directly affect employees. Long working hours make employees tired and decide to change jobs. And, as can be seen from the above analysis, it can be seen that the working hours of all employees are not the same. Only a relatively small number of employees have longer working hours. Companies should prevent the over-distribution of working hours to a small number of employees by redistributing them equally. Redistributing working hours so as not to be excessive will reduce the number of employees who were dissatisfied with working hours, reducing the number of employees who decided to change jobs due to excessive working hours.

### 2. Is there a correlation between marital Status and their work distance? Does satisfaction affect turnover?

* Data:
  + left\_company shows whether employees have changed their job or not
  + miles\_from\_home shows the distance each employee traveled from work to home
  + marital\_status shows whether employees are married, divorced or single
* Method: Dividing and grouping employees by marital status of employees show the correlation between the distance, marital status, and the turnover rate of employees. For example, grouping them shows the separation of those who moved their job and not depending on marital status. After that, various distances can be grouped into three forms by dividing the distance into quarters. Arranging into marital status makes it easier to see whether each group has changed jobs. Finally, the boxplot allows you to compare the turnover rate according to marital status at a glance.
* Analysis: Comparing the distance of divorced employees, the distance in the first quarter is the same between turnover status, the distance in the second quarter is three miles, and the distance in the third quarter is four miles different. Married employees and single employees similarly show only subtle differences in distance, ranging from two, four, and 5.25 miles.
* Results: The distance of employees who have changed jobs is farther than those who have not. However, the difference does not seem to have much impact because it seems so subtle. It can be seen that employees who were married or were single are more affected than divorced employees. In particular, single employees are most affected by the distance between work and home.

### 3. Is there a correlation between employee turnover and job level?

* Data:
  + left\_company shows whether employees have changed their job or not
  + job\_level shows the working position of each employee
* Method: The correlation between the two is revealed by comparing job\_level and left\_company data. Employees can be separated by job level(position) and turnover status as groups. Thereafter, they can be compared by counting the number of employees according to each group. Through the histogram table, the number of them can be compared at a glance. Finally, facet\_wrap will show the result of comparison separately depending on the turnover status.
* Analysis: There are five positions at the job level: Associate, Manager, Senior Manager, Director, and Vice President. For each position, 72 of the Associates have changed jobs and 113 have not. 93 of the Managers have changed jobs and 251 employees have not. 29 of the Senior Managers have changed jobs and 447 have not. 28 of the Directors have changed jobs and 303 employees have not. Finally, there are 15 Vice Presidents who have changed jobs and 119 who have not. It shows that the higher the rank, the greater the gap between those who have changed jobs and those who have not.
* Results: The attrition rate of low-ranking people is high, and that of high-ranking people is low. Companies should pay more attention to lower-level positions than to higher-level positions to reduce attrition rates.

### 4. Is there a correlation between employees’ years promotion and satisfaction? Does satisfaction affect turnover?

* Data:
  + left\_company shows whether employees are changed their job or not
  + yrs\_since\_promotion shows the total number of promotions for each employee
  + job\_satisfaction shows that each employee assessed their satisfaction with this company
* Method: Comparing the satisfaction of employees and each number of promotions can see whether the correlation between the two data affects the employee’s turnover status or not. First, grouping employees by satisfaction and separating them by turnover can categorize the data. After that, the number of various promotions of each employee can be divided into four quarters. Chaging the arrangement the listed data values from the order of turnover status to job satisfication makes the data easier to see. Finally, it is possible to compare the results according to job satisfaction at a glance through gom\_point.
* Analysis: At the geom\_point, it is looked like no significant difference between the four graphs because there is no difference at all in the first and second quarters of the number of promotions in data. However, there is a slight difference in the third quarter. In the third quarter, employees with high, low, and middle satisfaction differed from 1.0 to 2.0 between turnover status. In other words, it can be seen that although the promotion rate affects the satisfaction level, the degree is not large. Besides, in the very high satisfication, employees with lower promotion did not change their jobs whereas employees with higher promotion changed jobs. In other words, the promotion rate slightly affects satisfaction, but the satisfaction does not affect turnover.
* Results: Satisfaction and promotion affect turnover, but they do not appear to have a significant impact, so companies should care about them but not consider them to be the main causes

### 5. Is there a correlation between employees’ departments and salaries? Does the department affect the turnover?

* Data:
  + left\_company shows whether employees are changed their job or not
  + department shows the department to which each employee belongs
  + salary shows how much employees are paid
* Method: Grouping employees by department and turnover status can compare salaries by department. Dividing the salaries of various employees into quarters and then arranging them according to the department makes the data more simple and easier to see. Finally, using the geom\_boxplot can compare departmental data at a glance.
* Analysis: It can be seen that the fact that there is a clear difference in salary compared to turnover status has a significant impact on the turnover status. However, the turnover status according to the impact of salary varies from department to department. Comparing the salary difference between those who have changed jobs and those who have not changed jobs for each department, there is no difference in the salary of the IT and Analytics departments. In other words, IT and Analytics departments are the least affected by salaries. There will be other implications for determining their turnover. The Finance and Operations department has a large salary difference, but the turnover rate is almost the same as that of the IT and Analytics department. It can be seen that they are less sensitive to pay. On the other hand, the Marketing and Product Development departments have the highest turnover rates. They are very sensitive to salaries. Finally, the Research Department has only nine people who have left the company despite the biggest salary gap. It can be seen that they will not leave due to salary.
* Results: Turnover due to pay differences is obvious. However, since different departments have different influences, companies must compare and adjust salaries between departments. In addition, for departments that are not affected by salaries, it is necessary to analyze what other causes are.

### 6. Is there a correlation between employees’ marital status and salary? Did marital status affect the turnover?

* Data:
  + left\_company shows whether employees are changed their job or not
  + marital\_status shows whether employees are married, divorced or single
  + salary shows how much employees are paid
* Method: A relationship can be found by comparing the correlation between marital status and salary with the turnover rate. First of all, grouping employees by turnover status shows how many people have changed jobs depending on marital status. Then, various salaries of employees can be divided into quarters. Groups can compare each other by arranging based on marital status. Finally, using facet\_grid can check the results at a glance by comparing six histograms.
* Analysis: According to statistics, turnover has nothing to do with marriage. For divorced employees, there are 261 more employees who have not changed jobs, and there are also far more employees who have not changed jobs for married employees and single employees. In other words, marital status does not affect turnover. However, it can be seen that salaries are clearly causing. Regardless of marital status, they are paid a similar salary. However, the difference between those who changed jobs and those who didn’t is very large. In other words, marital status doesn’t matter, salaries greatly impact on turnover.
* Results: Marital status does not affect turnover, but salary has a strong impact, so companies should focus more on analyzing whether employees are changing jobs based on their salaries.

## 3. Conclusion(s)/Discussion.

There were more diverse factors than what we anticipated, affecting job turnover. Six factors were analyzed. Some factors were the causes of turnover. Whereas other factors were not causes of factors and need to be analyzed

1. Working hours had a lot of influence on employees’ turnover as expected. Employees’ working hours were not evenly distributed, and the difference in unequal working hours was large. All employees should be redistributed so that they have equal working hours and adjusted so that they do not work overtime.
2. The distance from employees’ homes to work did not affect us as much as we expected. However, this element should not be ruled out because the impact has existed at all. It may not be considered the main cause, but it can have a more positive effect by proceeding with welfare for employees far from home.
3. The turnover according to the position clearly existed. High-ranking people have more affection and lingering feelings for their positions, so they cannot easily decide to change jobs. However, low-ranking people can easily change jobs due to various factors because they have less affection and lingering feelings about their position. It is necessary to pay more attention to the welfare and satisfaction of employees with lower positions.
4. The promotion rate was not related to the satisfaction of employees more than what we expected. It cannot be said that there is no impact at all, but the difference was not significant. In addition, it could be seen that the promotion rate had little effect on the turnover. In fact, people with high promotion rates were more likely to change jobs. If so, it is necessary to further analyze what other factors are affected by employees.
5. Wage differences are the biggest factor in turnover. However, the impact of salary varies from department to department. Salary adjustments will be essential for payroll-sensitive departments. In addition, non-salary-sensitive departments should further analyze what factors are affected by by comparison with other data.
6. There was no correlation between marital status and salary. In addition, it was found that the marital status had no correlation with the turnover. However, it can still be seen that the impact of pay is significant. If so, it is necessary to analyze what other factors salary will be correlated with.

Through the six analyses, it was possible to know the parts that affect, the parts that do not affect, and the factors that require further analysis. The influencing parts were the distance to work, working hours, and salary. Employees were greatly affected by working hours and salaries, especially salaries. The impact on the distance to work was minimal. The areas that needed further analysis were position and departmental salaries. The difference in rank clearly affected the turnover, but it is impossible to give everyone a high position. If so, further analysis is needed to pay more attention to which sectors the lower positions are affected by. Each department was divided into a department that was greatly affected by salary and a department that was not very affected. If so, further analysis is needed on which sectors affected to departments that are not affected much by salaries. Finally, marriage status and promotion rates did not affect turnover. Companies should further refine the causes by organizing more diverse and specifically possible assumptions.

## 4. Appendix/Appendices.

library(tidyverse)

## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2 ──  
## ✔ ggplot2 3.3.6 ✔ purrr 0.3.5   
## ✔ tibble 3.1.8 ✔ dplyr 1.0.10  
## ✔ tidyr 1.2.1 ✔ stringr 1.4.1   
## ✔ readr 2.1.3 ✔ forcats 0.5.2   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()

employee\_data <- read\_csv("C:/Users/ssiba/Downloads/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.

employee\_data

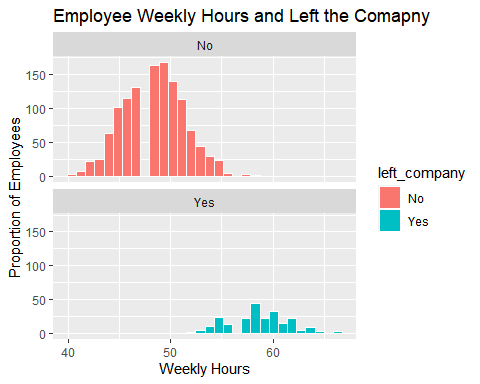
## # A tibble: 1,470 × 13  
## left\_company departm…¹ job\_l…² salary weekl…³ busin…⁴ yrs\_a…⁵ yrs\_s…⁶ previ…⁷  
## <chr> <chr> <chr> <dbl> <dbl> <chr> <dbl> <dbl> <dbl>  
## 1 Yes Sales Direct… 1.19e5 56 Rarely 6 0 5  
## 2 No Sales Senior… 8.56e4 42 Freque… 10 1 5  
## 3 Yes Product … Associ… 4.62e4 56 Rarely 0 0 6  
## 4 No IT and A… Direct… 1.17e5 50 Freque… 8 3 1  
## 5 No Sales Associ… 3.66e4 46 Rarely 2 2 2  
## 6 No Marketing Senior… 8.35e4 48 Freque… 7 3 1  
## 7 No Marketing Senior… 8.86e4 44 Rarely 1 0 3  
## 8 No Sales Direct… 1.22e5 47 Rarely 1 0 3  
## 9 No Finance … Senior… 9.46e4 50 Freque… 9 1 3  
## 10 No Product … Direct… 1.25e5 51 Rarely 7 7 3  
## # … with 1,460 more rows, 4 more variables: job\_satisfaction <chr>,  
## # performance\_rating <chr>, marital\_status <chr>, miles\_from\_home <dbl>, and  
## # abbreviated variable names ¹​department, ²​job\_level, ³​weekly\_hours,  
## # ⁴​business\_travel, ⁵​yrs\_at\_company, ⁶​yrs\_since\_promotion,  
## # ⁷​previous\_companies

**Question 1: Is there a correlation between employee turnover and their working hours?**

Q1 <- employee\_data %>%  
 group\_by(left\_company)%>%  
 summarise(n\_employees = n(),  
 min\_weekly\_hours = min(weekly\_hours),  
 max\_weekly\_hours = max(weekly\_hours),  
 avg\_weekly\_hours = mean(weekly\_hours),  
 sd\_weekkly\_hours = sd(weekly\_hours))  
Q1

## # A tibble: 2 × 6  
## left\_company n\_employees min\_weekly\_hours max\_weekly\_hours avg\_weekl…¹ sd\_we…²  
## <chr> <int> <dbl> <dbl> <dbl> <dbl>  
## 1 No 1233 40 58 48.4 3.02  
## 2 Yes 237 51 66 58.6 2.99  
## # … with abbreviated variable names ¹​avg\_weekly\_hours, ²​sd\_weekkly\_hours

ggplot(data = employee\_data, aes(x = weekly\_hours, fill = left\_company))+  
 geom\_histogram(color = "white", bins = 30)+  
 facet\_wrap(~ left\_company, nrow = 2) +  
 labs(title = "Employee Weekly Hours and Left the Comapny",  
 x = "Weekly Hours", y = "Proportion of Employees")



**Question 2: Is there a correlation between marital Status and their work distance? Does satisfaction affect turnover?**

Q2 <- employee\_data %>%  
 group\_by(left\_company, marital\_status)%>%  
 summarise(n\_employee = n(),  
 quantile1\_miles\_from\_home = quantile(miles\_from\_home, 1 / 4),  
 quantile2\_miles\_from\_home = quantile(miles\_from\_home, 2 / 4),  
 quantile3\_miles\_from\_home = quantile(miles\_from\_home, 3 / 4)) %>%  
 arrange(marital\_status, desc(left\_company))

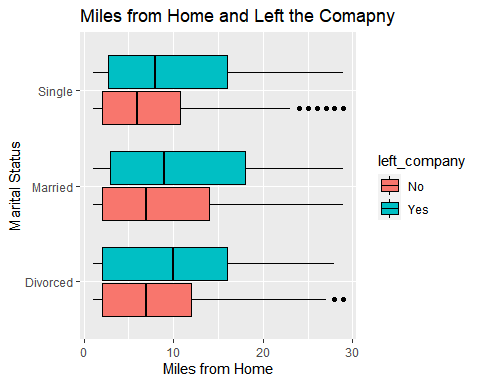
## `summarise()` has grouped output by 'left\_company'. You can override using the  
## `.groups` argument.

Q2

## # A tibble: 6 × 6  
## # Groups: left\_company [2]  
## left\_company marital\_status n\_employee quantile1\_miles\_from\_…¹ quant…² quant…³  
## <chr> <chr> <int> <dbl> <dbl> <dbl>  
## 1 Yes Divorced 33 2 10 16   
## 2 No Divorced 294 2 7 12   
## 3 Yes Married 84 3 9 18   
## 4 No Married 589 2 7 14   
## 5 Yes Single 120 2.75 8 16   
## 6 No Single 350 2 6 10.8  
## # … with abbreviated variable names ¹​quantile1\_miles\_from\_home,  
## # ²​quantile2\_miles\_from\_home, ³​quantile3\_miles\_from\_home

ggplot(data = employee\_data, aes(x = miles\_from\_home, fill = left\_company)) +  
 geom\_boxplot(aes(y = marital\_status), color = "black", bins = 30) +  
 labs(title = "Miles from Home and Left the Comapny",  
 x = "Miles from Home", y = "Marital Status")

## Warning: Ignoring unknown parameters: bins



**Question 3: Is there a correlation between employee turnover and job level?**

Q3 <- employee\_data %>%   
 group\_by(left\_company,job\_level)%>%   
 summarise(total = n())

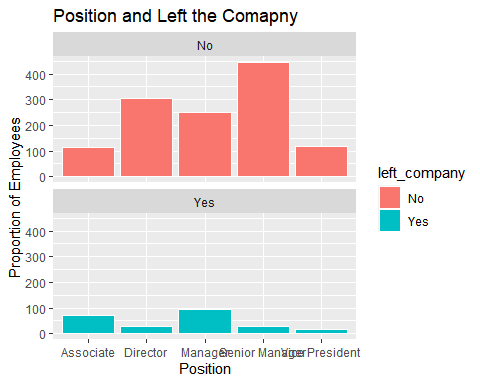
## `summarise()` has grouped output by 'left\_company'. You can override using the  
## `.groups` argument.

Q3

## # A tibble: 10 × 3  
## # Groups: left\_company [2]  
## left\_company job\_level total  
## <chr> <chr> <int>  
## 1 No Associate 113  
## 2 No Director 303  
## 3 No Manager 251  
## 4 No Senior Manager 447  
## 5 No Vice President 119  
## 6 Yes Associate 72  
## 7 Yes Director 28  
## 8 Yes Manager 93  
## 9 Yes Senior Manager 29  
## 10 Yes Vice President 15

ggplot(data = Q3, aes(x = job\_level, fill = left\_company)) +   
 geom\_histogram(stat = "identity", aes(y = total), color = "white") +  
 facet\_wrap(~ left\_company, nrow = 2) +  
 labs(title = "Position and Left the Comapny",  
 x = "Position",   
 y = "Proportion of Employees")

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



**Question 4: Is there a correlation between employees’ years promotion and satisfaction? Does satisfaction affect turnover?**

Q4 <- employee\_data %>%  
 group\_by(left\_company, job\_satisfaction)%>%  
 summarise(n\_employee = n(),  
 quantile1\_yrs\_since\_promotion = quantile(yrs\_since\_promotion, 1 / 4),  
 quantile2\_yrs\_since\_promotion = quantile(yrs\_since\_promotion, 2 / 4),  
 quantile3\_yrs\_since\_promotion = quantile(yrs\_since\_promotion, 3 / 4))%>%  
 arrange(job\_satisfaction, desc(left\_company))

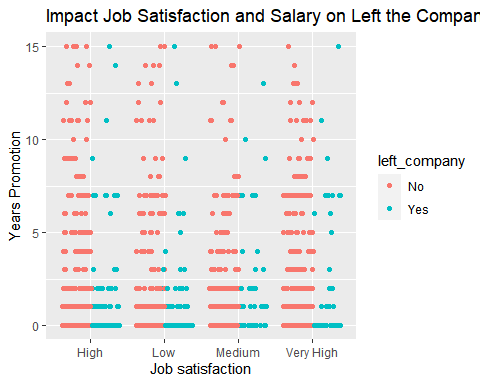
## `summarise()` has grouped output by 'left\_company'. You can override using the  
## `.groups` argument.

Q4

## # A tibble: 8 × 6  
## # Groups: left\_company [2]  
## left\_company job\_satisfaction n\_employee quantile1\_yrs\_since…¹ quant…² quant…³  
## <chr> <chr> <int> <dbl> <dbl> <dbl>  
## 1 Yes High 73 0 1 2   
## 2 No High 369 0 1 3   
## 3 Yes Low 66 0 0 2   
## 4 No Low 223 0 1 4   
## 5 Yes Medium 46 0 1 2   
## 6 No Medium 234 0 1 3   
## 7 Yes Very High 52 0 1 4   
## 8 No Very High 407 0 1 2.5  
## # … with abbreviated variable names ¹​quantile1\_yrs\_since\_promotion,  
## # ²​quantile2\_yrs\_since\_promotion, ³​quantile3\_yrs\_since\_promotion

ggplot(data = employee\_data, aes(x= job\_satisfaction, y = yrs\_since\_promotion, color = left\_company))+  
 geom\_point(position=position\_jitterdodge(), bins = 10) +  
 labs(title = "Impact Job Satisfaction and Salary on Left the Company)",  
 x = "Job satisfaction", y = "Years Promotion")

## Warning: Ignoring unknown parameters: bins



**Question 5: Is there a correlation between employees’ departments and salaries? Does the department affect the turnover?**

Q5 <- employee\_data %>%   
 group\_by(left\_company, department)%>%   
 summarise( n\_employees = n(),  
 quantile1\_salary = quantile(salary, 1 / 4),  
 quantile2\_salary = quantile(salary, 2 / 4),  
 quantile3\_salary = quantile(salary, 3 / 4)) %>%  
 arrange(department,desc(left\_company))

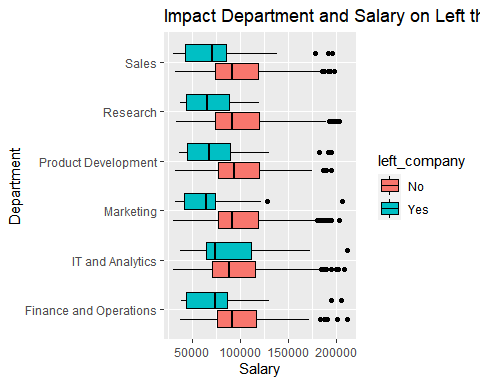
## `summarise()` has grouped output by 'left\_company'. You can override using the  
## `.groups` argument.

Q5

## # A tibble: 12 × 6  
## # Groups: left\_company [2]  
## left\_company department n\_employees quantile1\_s…¹ quant…² quant…³  
## <chr> <chr> <int> <dbl> <dbl> <dbl>  
## 1 Yes Finance and Operations 31 43974. 73595. 86624.  
## 2 No Finance and Operations 79 76067. 92074. 117184.  
## 3 Yes IT and Analytics 30 64350. 73836. 111705.  
## 4 No IT and Analytics 369 71212. 88930. 116099.  
## 5 Yes Marketing 40 41805. 64081. 74198.  
## 6 No Marketing 198 77468. 91317. 118799.  
## 7 Yes Product Development 50 44413. 67940. 89631.  
## 8 No Product Development 128 77108. 93703. 119504.  
## 9 Yes Research 9 43394. 65246. 88555.  
## 10 No Research 284 73670. 91535. 120201.  
## 11 Yes Sales 77 42927. 70501. 85275.  
## 12 No Sales 175 74150. 91795. 118849.  
## # … with abbreviated variable names ¹​quantile1\_salary, ²​quantile2\_salary,  
## # ³​quantile3\_salary

ggplot(data = employee\_data, aes(x= salary, y = department, fill = left\_company))+  
 geom\_boxplot(color = "black", bins = 30) +  
 labs(title = "Impact Department and Salary on Left the Company",  
 x = "Salary", y = "Department")

## Warning: Ignoring unknown parameters: bins



**Question 6: Is there a correlation between employees’ marital status and salary? Did marital status affect the turnover?**

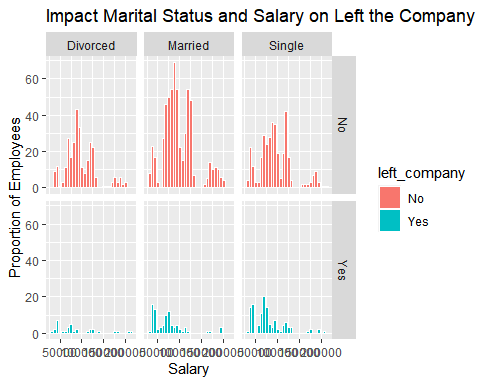
Q6 <- employee\_data %>%   
 group\_by(left\_company,marital\_status)%>%  
 summarise(n\_employees = n(),  
 quantile1\_salary = quantile(salary, 1 / 4),  
 quantile2\_salary = quantile(salary, 2 / 4),  
 quantile3\_salary = quantile(salary, 3 / 4)) %>%  
 arrange(marital\_status,desc(left\_company))

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

Q6

## # A tibble: 6 × 6  
## # Groups: left\_company [2]  
## left\_company marital\_status n\_employees quantile1\_salary quantile2\_s…¹ quant…²  
## <chr> <chr> <int> <dbl> <dbl> <dbl>  
## 1 Yes Divorced 33 43284. 75255. 117529.  
## 2 No Divorced 294 76083. 91076. 118256.  
## 3 Yes Married 84 42843. 66782. 80794.  
## 4 No Married 589 75146. 90858. 119617.  
## 5 Yes Single 120 45977. 70818. 92554.  
## 6 No Single 350 71438. 90651. 117280.  
## # … with abbreviated variable names ¹​quantile2\_salary, ²​quantile3\_salary

ggplot(data = employee\_data, aes(x = salary, fill = left\_company))+  
 geom\_histogram(color = "white", bins = 30) +  
 facet\_grid(left\_company ~ marital\_status) +  
 labs(title = "Impact Marital Status and Salary on Left the Company",  
 x = "Salary", y = "Proportion of Employees")



— End of Instructions —