Data Analysis on Employee Satisfaction

Payton Pruss

Minh Thu Pham

Final Paper

Math 268

Dr. Hamdan

Section 2

Overview

One of the most common and essential metrics for human resources is employee satisfaction, or commonly referred to as job satisfaction. Employee satisfaction is the level of contentment an employee feels with their job and often includes feelings towards the company. It has a large impact on their workplace experience as well as their attitude towards tasks and colleagues. In this paper, we will analyze the dataset regarding employee satisfaction to further understand the influences on workplace culture and to better understand the feelings and attitudes of employees.

Data

The table below describes the characteristics of each variable present in the dataset, there are different statistical measures of central tendency and variation. The attrition rate is equal to 24%, the satisfaction level is around 62% and the performance average is around 71%. We see that on average people work on 3 to 4 projects a year. Average monthly hours of work have a mean of 201. To put this into context, a 40-hour workweek, which is the common average of hours worked in a week, has a total monthly hour worked of 160. Therefore, for this company the employees work on average 50 hours a week.

```
##  satisfaction_level last_evaluation  number_project  average_montly_hours
##  Min.   :0.0900     Min.   :0.3600   Min.   :2.000   Min.   : 96.0
##  1st Qu.:0.4400     1st Qu.:0.5600   1st Qu.:3.000   1st Qu.:156.0
##  Median :0.6400     Median :0.7200   Median :4.000   Median :200.0
##  Mean   :0.6128     Mean   :0.7161   Mean   :3.803   Mean   :201.1
##  3rd Qu.:0.8200     3rd Qu.:0.8700   3rd Qu.:5.000   3rd Qu.:245.0
##  Max.   :1.0000     Max.   :1.0000   Max.   :7.000   Max.   :310.0
##  time_spend_company Work_accident         left
##  Min.   : 2.000     Min.   :0.0000   Min.   :0.0000
##  1st Qu.: 3.000     1st Qu.:0.0000   1st Qu.:0.0000
##  Median : 3.000     Median :0.0000   Median :0.0000
##  Mean   : 3.498     Mean   :0.1446   Mean   :0.2381
##  3rd Qu.: 4.000     3rd Qu.:0.0000   3rd Qu.:0.0000
##  Max.   :10.000     Max.   :1.0000   Max.   :1.0000
##  promotion_last_5years    sales              salary
##  Min.   :0.00000       Length:14999       high  :1237
##  1st Qu.:0.00000       Class :character   low   :7316
##  Median :0.00000       Mode  :character   medium:6446
##  Mean   :0.02127
##  3rd Qu.:0.00000
##  Max.   :1.00000
```

For our analysis, we used 8 of the 11 variables: how satisfied the employee is in their position (scale of 0 to 1) (satisfaction_level), how management rated employee performance during the last evaluation (scale of 0 to 1) (last_evaluation_score), the number of projects an employee is currently working on (number_of_projects), a binary variable that indicates whether the employee experienced an accident at work(yes= 1, no=0) (work_accident), the department that the employee works in with a total of 11 departments (department), the level of the employee' s salary(low, medium, high) (salary), and lastly the dollar range for the salary levels (less than $45,000, $45,000-$75,000, greater than $75,000) (salary_range). There are 15,000 rows of data in this dataset.

Methods

We use a chi-square test for independence when we want to formally test whether there is a statistically significant association between two categorical variables. The null hypothesis will state that there is no significant association between the two variables, with the alternative hypothesis stating that there *is* a significant association between the two variables. There are three assumptions that must be made about the data before being able to conduct a chi-square test of independence. The first being the data is random, a random sample or random experiment should be used to collect the data for both samples. The second being that the variables we are studying are categorical. Lastly, the expected number of observations at each level of the

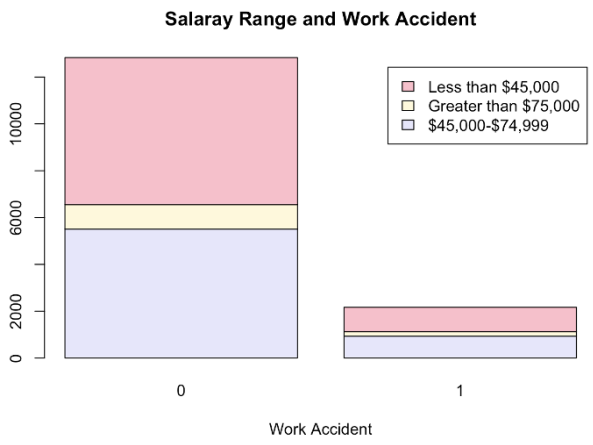
variable should be at least 5. If these assumptions are met, then we can then conduct the hypothesis test.

We use a t-test for a difference in means when we want to formally test whether there is a statistically significant difference between two population means. The null hypothesis will state that the two-population means are equal. The alternative will state that the two-population means are not equal. As will the chi-squared test, there are assumptions that must be made to ensure that test will be valid. Firstly, a random sample or random experiment should be used to collect data for both samples. Secondly, the sampling distribution is normal or approximately normal, Lastly, that the two samples are independent. If these assumptions are met, then we can then conduct the hypothesis test.

Analysis

For our analysis of employee satisfaction, three base questions were asked in order to establish associations between variables; followed by three questions that pertained to the bigger picture regarding satisfaction and employees. The base questions use the variables: salary, department, salary range, work accident, number of projects.

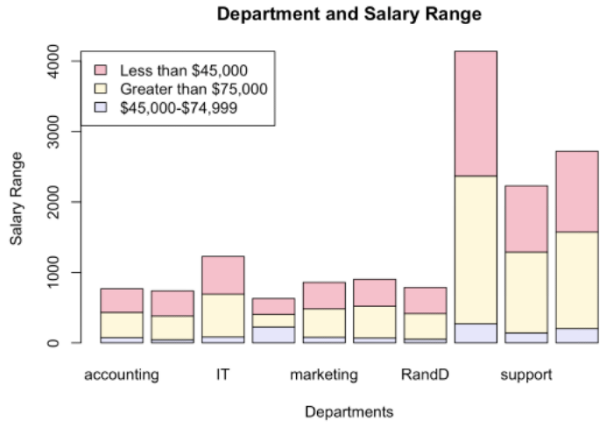
The second base question to review is asking if there is a statistically significant association between salary range and work accident occurrence? Salary range has three levels: less than $45,000, $45,000 to $75,000 and greater than $75,000. Work accident occurrence is a binary value of 0 and 1 representing an employee having had an accident (1) or not (0).



The "Salary Range and Word Accident" graph shows the distribution of work accidents separated by levels of salary. Those who make greater than $75,000 are a lower number of employees, however all levels are equally represented in both categories. In order to provide analysis on this question the Pearson's chi-squared test was performed. The p-value was 0.4685 which is larger than the alpha of 0.05. Therefore, we accept the null hypothesis that salary range and work accident occurrence are independent of each other. This is equally supported in the equal distribution of all salary ranges throughout the above graph. This is important to note that there is no association between how much an employee gets paid and whether they are in a work accident, so we can assume that the area in which a work accident can occur is not attributed to salary range.

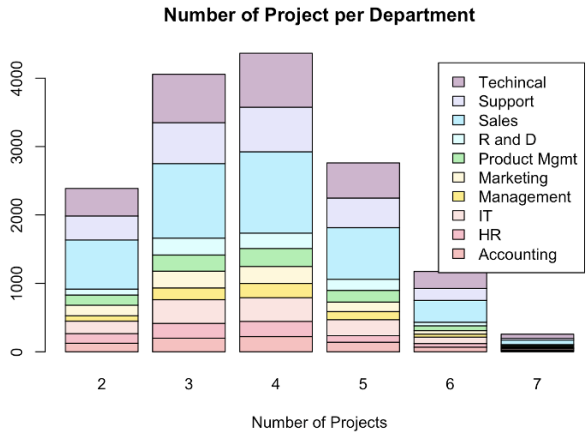
The third base question we want to understand is if there is a statistically significant association between salary and department? The salary range here is again separated into 3 levels: less than $45,000, $45,000 to $75,000 and greater than $75,000. For department, there are a total of 11 departments which include: technical, support, sales, research and development

(R&D), product management, marketing, management, IT, human resources (HR), and accounting.



The "Department and Salary Range" graphs' x-axis are the departments, which follow the pattern of listing previously stated in the listing of the variables. The bar plot tells us that Sales has the highest number of employees with varying representation of the various levels of salary range throughout the departments. Due to analyzing categorical variables, once again a Pearson's chi-squared test is used. The null hypothesis stated that salary range and department are independent. Moreover, the alternative hypothesis stated that salary range and department are dependent. The p-value here was less than 2.2e-16 which is less than the alpha of 0.05. Therefore, we reject the null hypothesis and have enough conclusive evidence to state that salary range and department are dependent. This is good to note for HR and recruitment purposes to recognize the departments in which salary ranges are higher and could potentially draw a conclusion of the affect responsibilities and job title have on salary range.

Lastly, we will consider the question of if there is a statistically significant association between department and number of projects. The mean number of projects is 3.8 with a range spanning from 2 projects to 7 projects. Department again includes technical, support, sales, research and development (R&D), product management, marketing, management, IT, human resources (HR), and accounting.

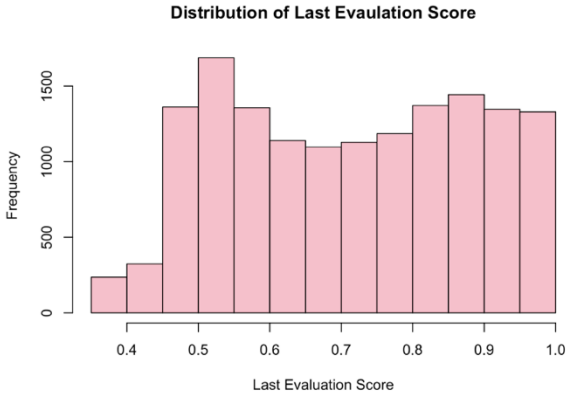


The "Number of Project per Department" bar chart shows the distribution of the departments and their representation within each number of projects. We see that sales, marketing, and product management are prevalent in 3 to 4 projects. For a total of 7 projects, we see only representation in technical, support, and sales. To analyze these categorical variables, a Pearson's chi-squared test was used with a null hypothesis of number of projects and department are independent, and the alternative hypothesis of number of projects and department are dependent. Our p-value for this analysis is 0.0002, which is less than 0.05, our alpha. We reject the null hypothesis and therefore, there is enough conclusive evidence to state that the number of projects and department are dependent. This tells us that depending on what department an employee is a part of has an

influence on how many projects they will work on. This can also tell us more about how different salary ranges are affected by both department and number of projects.

Now that we have outlined the basic understanding of associations between variables, we will analyze the areas related to satisfaction level, last evaluation score, and work accidents. These areas are more complex ideas and relate to the larger picture pertaining to employee satisfaction.

Firstly, analysis was conducted to determine whether there is there a statistically significant association between satisfaction level and last evaluation score? Satisfaction level is rated on a point scale from 0.09 to 1. The mean satisfaction level is 0.6128. To quantify this, we can analyze the decimals as percentages, so most employees are 61% satisfied with the company and their job. Evaluation scores are distributed on a point scale as well, with the mean evaluation score being 0.7161. To analyze that score in a percentage context, most employees receive an evaluation score of 70%. Due to the ordinal nature of a ranking point system these variables will be categorical in nature. We want to formally test whether there is a statistically significant association between these two categorical variables.



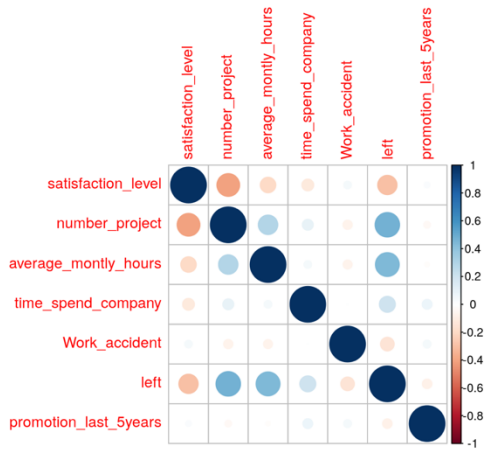
Above we can see further information regarding the distribution of these two variables. Satisfaction level is skewed slightly to the left, with evaluation score having a remarkably interesting distribution with a double curve being seen with one peak at 0.9 and one just above 0.5. The null hypothesis states that satisfaction level and evaluation score are independent, and the alternative hypothesis states that satisfaction level and evaluation score are dependent. The Pearson's chi-squared test resulted in a p-value of less than 2.2e-16 which is less than the alpha of 0.05. The chi-squared test resulted in enough conclusive evidence to state that satisfaction level and evaluation score and dependent. This concept seems altruistic in nature, as if an employee is told they are doing a good job, they would be more likely to feel satisfied in their company and job role, and the reverse can be said as well. This shows an important relationship between providing positive feedback and seeing positive feelings towards that entity.

Lastly, we will discuss if there is a statistically significant difference in mean evaluation score for those employees in an accident who left the company than those employees in an accident who did not leave the company? As stated previously, evaluation scores are distributed

on a point scale with the mean evaluation score being 0.7161 or 70%. The variable, left company, is a binary value of left the company (1) and did not leave the company (0). Work accident is a binary variable as well with being in a work accident (1) and not being in a work accident (0). To analyze the populations, the data was subset into those who were in an accident and those who were not. After this data was subset, the data regarding those who were in an accident was evaluated regarding if they had left the company and their last evaluation score. Below is a boxplot that shows the distribution of these the two groups of the variable "left_company" and shows the distribution of last evaluation scores.

Because we want to formally test whether there is a statistically significant difference between two population means, a Welch's Two-Sample T-Test was used. The null hypothesis stated that the true difference of the means in evaluation scores for employees who were in a work accident and left the company, and for employees who were in a work accident and did not leave the company is equal to 0. The alternative stated that the true difference of the means in evaluation scores for employees who were in a work accident and left the company, and for employees who were in a work accident and did not leave the company is not equal to 0. In summary, we are analyzing whether those in a work accident were influenced by their evaluation score in staying with the company. The p-value for the t-test performed was 0.8189 which is far larger than the alpha at 0.05. Therefore, we accept the null hypothesis that the true difference of the means in evaluation scores for employees who were in a work accident and left the company and those who were in a work accident and did not leave the company is equal to 0. In conclusion, the last evaluation score received by someone in a work accident does not affect whether an employee stays or leaves.

Due to the analysis performed above, more research of the correlation between the variables was conducted to further understand why employees leave the company. A correlation matrix was created and shows the correlations for each variable by each variable. The size of the bubbles in each box reveals the significance of the correlation, and the color represents the direction of the correlation: blue meaning positive and red meaning negative.



From this correlation matrix, we see under “left” there are three significant correlations. The first is a moderately negative correlation with satisfaction level, secondly there is a moderately positive correlation with the number of projects, and thirdly there is a moderately positive correlation with average monthly hours. This tells us that the main reasons an employee leaves the company are having to complete a high number of projects, working a high number of average monthly hours and rating their satisfaction at a lower level.