**Final Group Project**

A1

Ruyin Li, Fangrui Mei, Joel Esparza, Hyun Ji Lee

ACCY 512

December 10, 2023

1. Introduction
2. Research Validity Framework
3. Data Cleaning and Understanding
4. Statistical Analysis
5. Deeper Analysis
6. Conclusion
7. Limitations and Further Improvements

**1. Introduction**

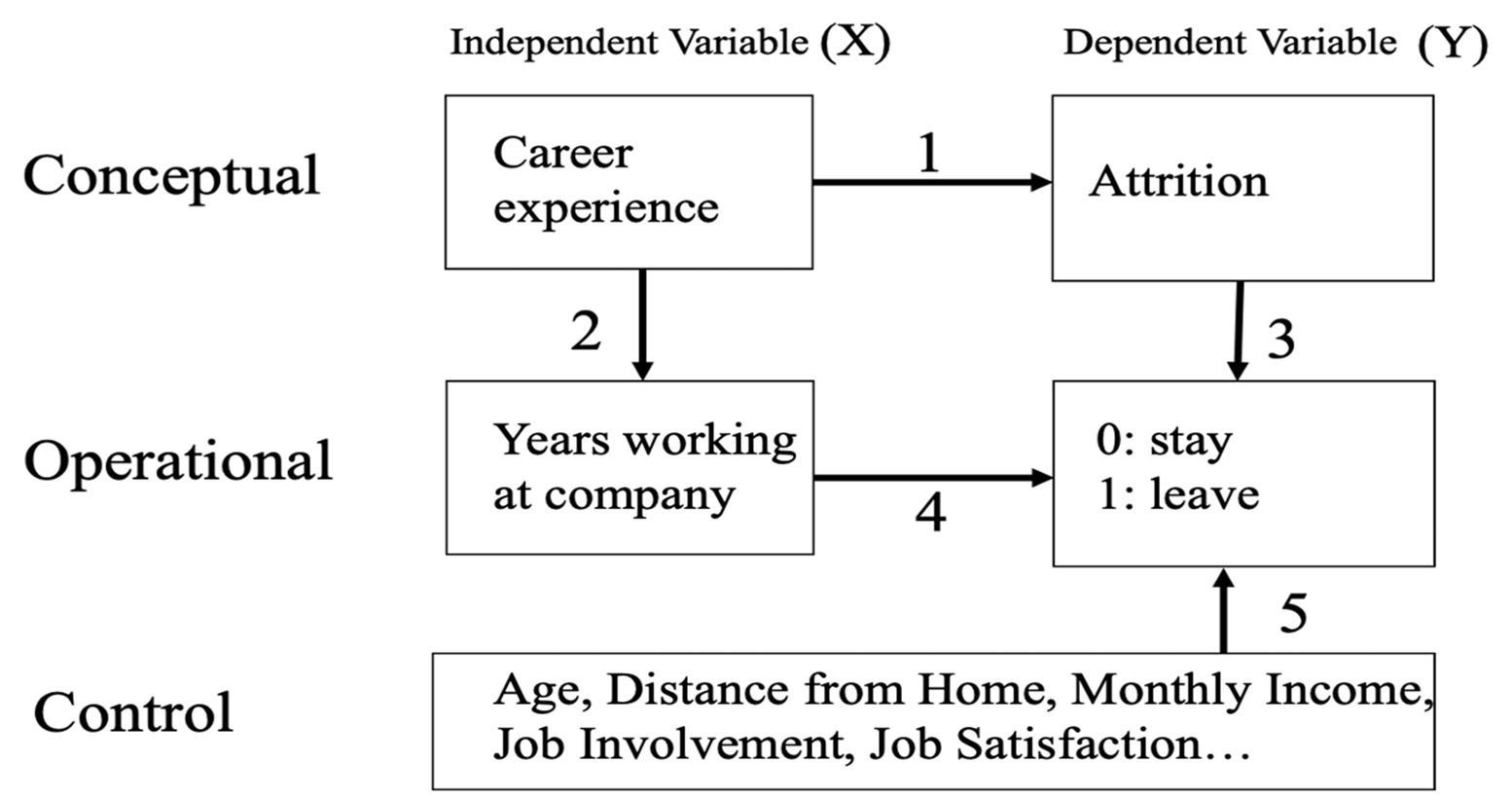
Employee attrition, measuring the departure of employees from a business, is a critical concern with far-reaching implications for productivity and organizational well-being. Our project addresses the underlying causes of employee attrition within the organization. High attrition rates not only affect overall productivity and employee morale but also lead to hidden costs, such as recruiting expenses, challenges associated with onboarding, and potential mis-hires. The reduction, even marginal, of attrition can result in substantial time and cost savings, making employee retention pivotal for organizational resilience.

Our objective is to provide actionable insights to the HR department, enabling them to implement strategies fostering employee retention and satisfaction. Our research centers on the question: How do career development and experience influence attrition rates? While intuition suggests these factors play a significant role, our approach moves beyond intuition to delve into the data. Leveraging HR analytics, specifically logistic regression, we aim to unveil concrete patterns and correlations that validate or challenge prevailing theories about employee attrition [1].

By pinpointing influential factors, we seek to develop a nuanced understanding of the dynamics at play. Our findings will offer the HR team data-driven recommendations for enhancing employee retention strategies and guiding decisions related to engagement, professional development, and retention policies.

**2. Research Validity Framework**

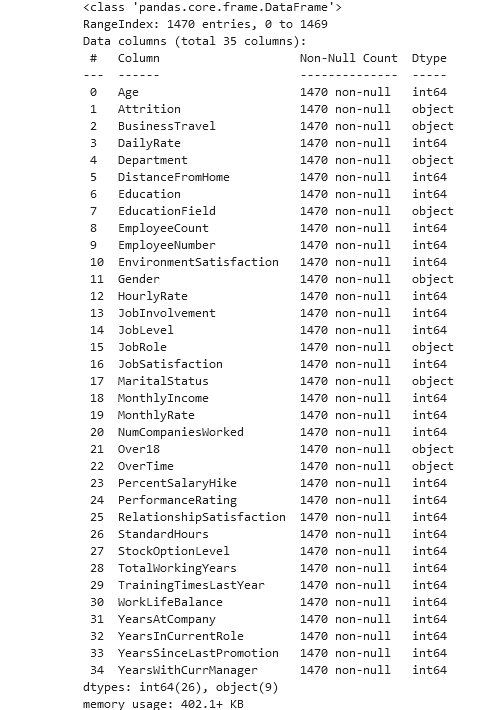
This project investigates the relationship between career and experience, measured by "years working at the company," and attrition. Career and experience are conceptualized as the broader professional journey and expertise gained in the workplace. Attrition is operationalized with a binary scale (0 for staying, 1 for leaving). Control variables include age, distance from home, monthly income, job involvement, and job satisfaction.



**3. Data Preparation and Understanding**

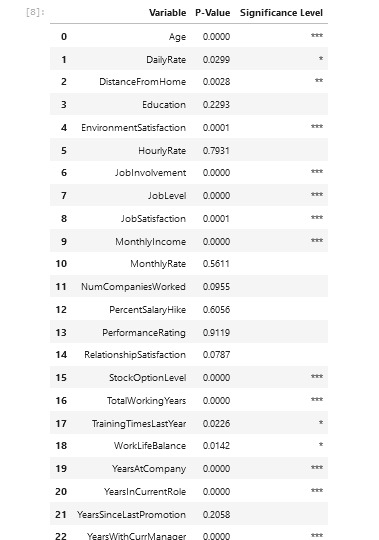
After exploring just a few of the thousands of datasets that Kaggle has to offer, we decided to settle with a data set containing information about employee attrition. This dataset provides a thorough analysis of an organization's workforce, delving into employee attrition, personal and job-related factors, and financial aspects. It encompasses a wide range of parameters, including demographic details such as Age, Gender, and Marital Status, as well as work-related factors like Business Travel Frequency, Daily Rate of Pay, and Departmental information. Job-related variables such as Job Involvement, Job Level, and specific Job Role are considered, along with total working hours, Percent Salary Hike, Performance Rating, and Relationship Satisfaction. The dataset also captures Monthly Income dynamics, accounting for overtime hours and previous work experiences. The Retirement Status, denoting whether employees intend to stay until retirement or experience earlier-than-expected attrition, is a key focus. Overall, this dataset offers valuable insights into contemporary workforce management philosophies influenced by technological advancements. Due to the wide range of predictors that this dataset offers, we decided to choose this one to help us answer our research question.

This dataset contains 1,470 unique observations with each observation representing an individual employee at a certain point in time. This dataset also contains 35 columns with each column containing different information regarding each employee. From these 35 columns, 34 will be used as predicting variables and 1 will be used as our target variable. Our target variable is a column called “Employee Attrition” that has a binary outcome: “Yes” if the employee left & “No” if the employee stayed with the company. For our other 34 variables, these were made of 25 numeric variables and 9 categorical. We can see in the image below that this dataset consisted of no null variables. This was great news for us as it allowed us to proceed to our variable selection.



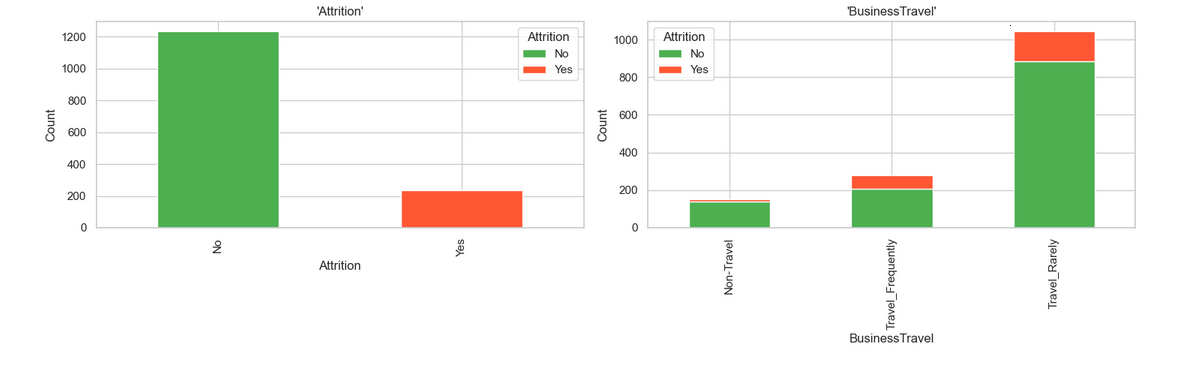
To start off our variable selection, we first began by viewing a table of the descriptive statistics for each numeric variable. By analyzing this table, we were able to identify the wide variety of employees that were present in this dataset. For example, employees were anywhere between 18 to 60 years old with their daily income ranging from $150 in a given day all the way to $1500 a day. In addition to giving us better insights on the distribution of our numeric variables, we were also able to identify several columns to drop from our logistic regression model. Employee count (indicating the number of employees working at the employee’s company) and working hours (indicating the amount of weekly hours the employee worked) both had a value of zero for its standard deviation. This means that there is absolutely no variance between observations when it comes to these columns. Due to each employee having the exact same value for each of these columns, we decided to drop them from our logistic regression. In addition to this, we dropped the column indicating the employee’s pin number. Since we are not joining any more data into this dataset, we do not need this column for our purposes. We dropped this column as well.

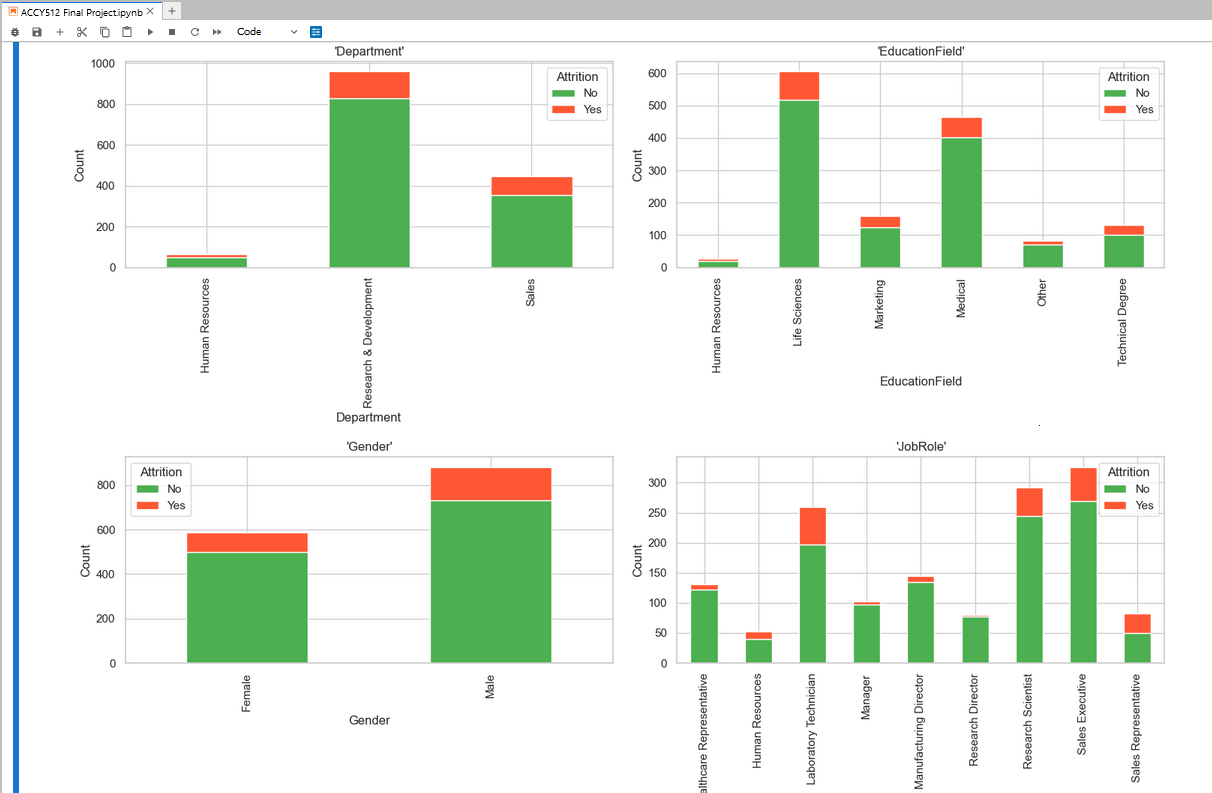
After dropping these three columns, we then wanted to identify the relevant variables that can accurately represent a difference between employees who left the company and those who stayed. For this reason, we decided to run a difference in means t-test to determine which variables would be worth exploring. After grouping by attrition and aggregating the mean for each numeric variable, we then calculated the difference between the two means. After doing so, we ran a t-test for each of these numeric variables to identify which mean differences were significant. These results are shown below. \* indicates a p-value less than .05, \*\* indicates a p-value less than .01, and \*\*\* indicates a p-value less than .001. We decided to only include variables that were significant at a .05 level for our model.

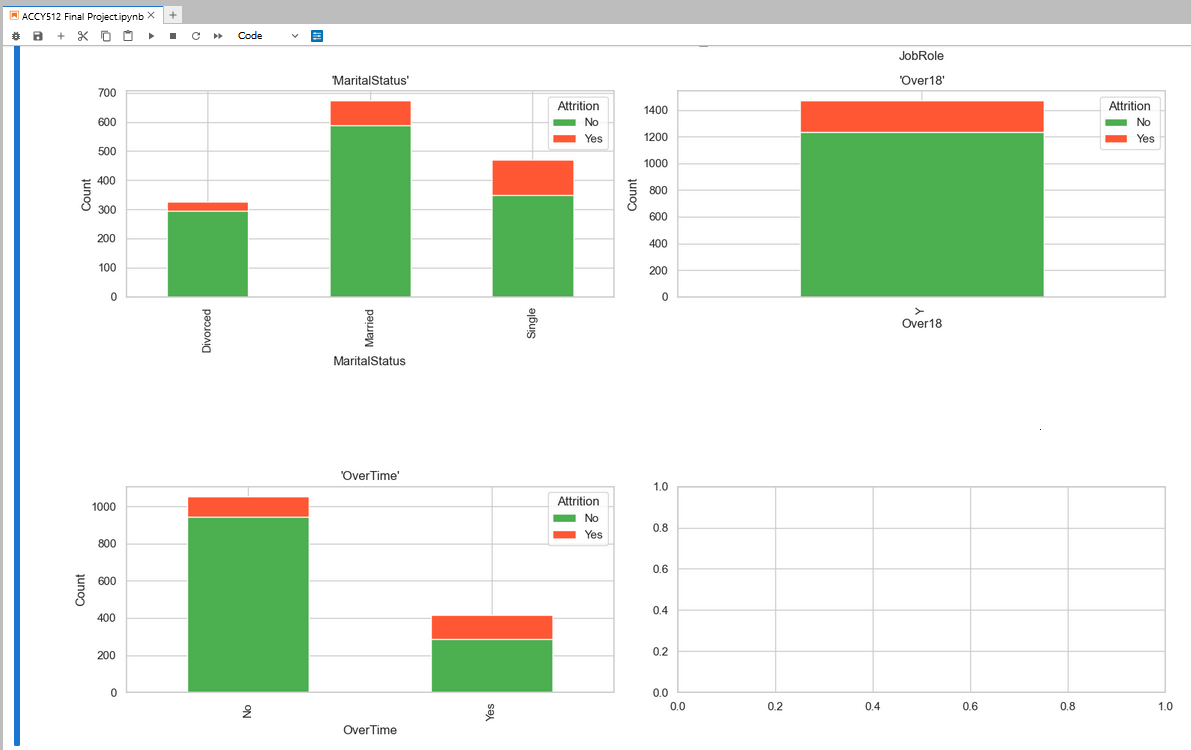


After analyzing the numerical variables, we then took a deeper dive to identify more information about the categorical variables in this dataset. We created a bar graph to represent the distribution of each unique variable within each column. From these visualizations, we are able to identify each unique value that is present for each categorical variable, the count of observations that falls within each categorical variable, and the proportion of employees who left the company for each unique value within each categorical variable. In these bar graphs, green represents those employees who stayed with the company while red represents those employees who left. Now that we fully understand what these graphs are showing us, we proceeded to analyze each of them.

The most insightful information that was displayed is the first bar graph that depicts the distribution of how many employees left and how many stayed. We can see that about 215 employees in this dataset ended up leaving the company. After discovering this, we wanted to see how these 215 employees were categorized for each of these variables. We can see that people with a sales job role tend to face higher proportions of attrition. In addition to this, it appears such factors such as overtime, traveling frequently, and being single results in higher employee attrition (based off these charts). This gave us some variables of interest when it comes to running our actual analysis. Lastly, we can notice that the Over18 variable has only one bar, indicating that each observation has the same value.Thus, we drop this variable from our model.



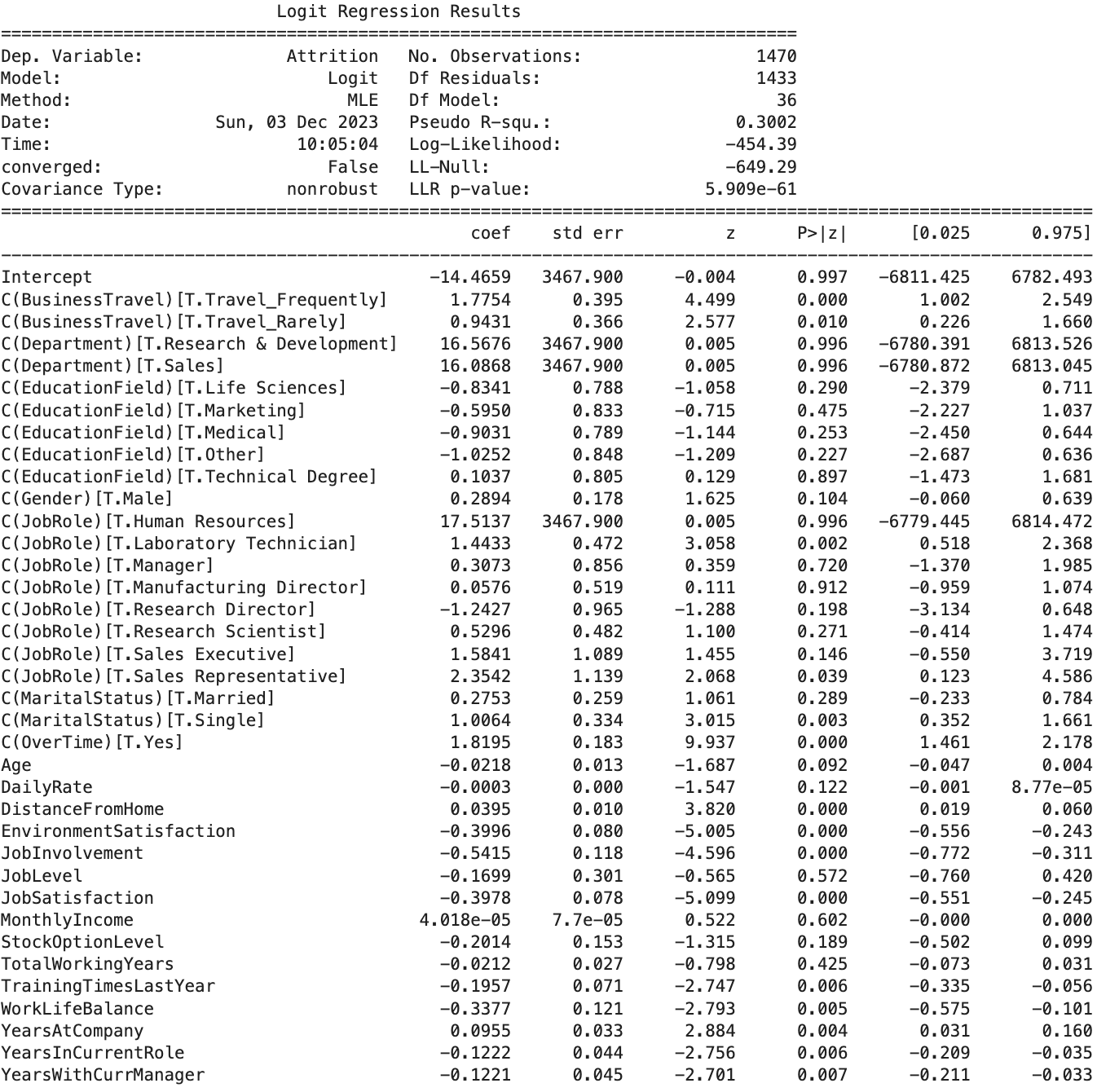




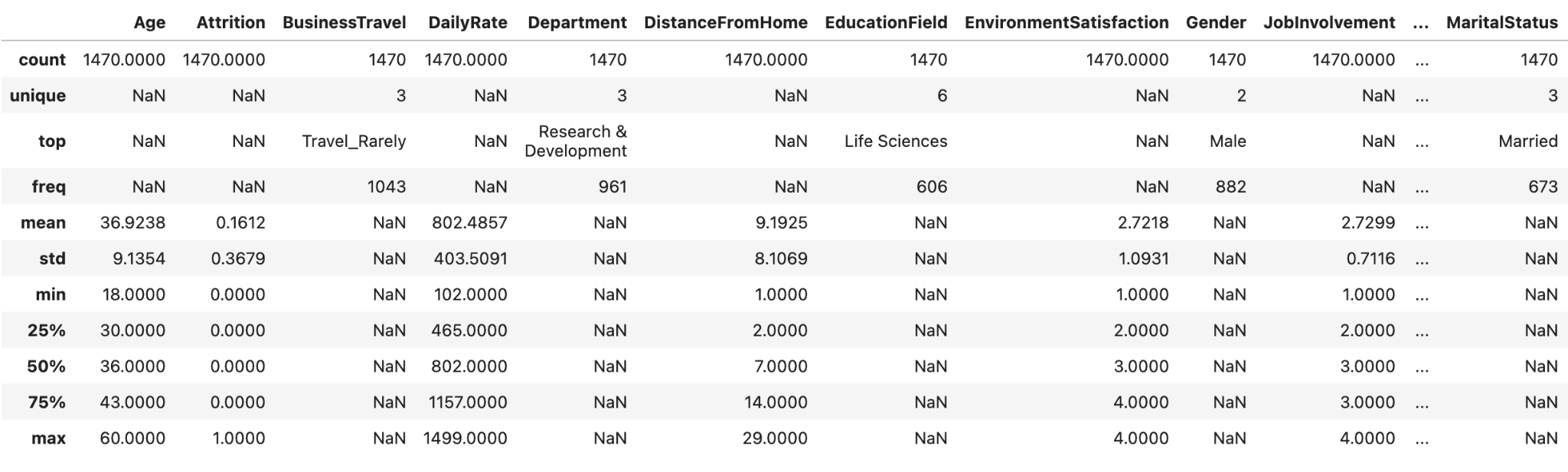
Now that we have our variables that we will include in our model, we must decide on a constant value for each variable so that we are able to isolate our variable of interest. For this reason, we calculated the mean for each of our numeric variables and the mode for each of our categorical variables. After calculating these values, we set these as the constant values for each variable. The only variation in the values of our logistic regression model will be that of our variable of interest. We will now take a look at the models we ran and their results.

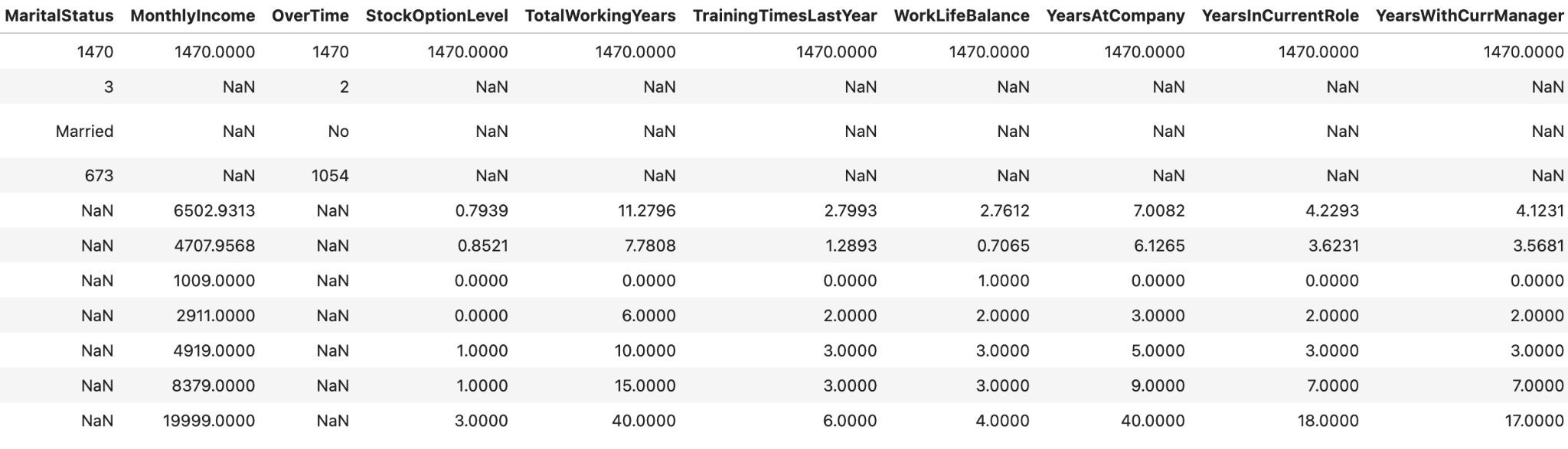
**4. Statistical Analysis**

Using the numeric and categorical variables mentioned previously, we basically ran the logistic regression. The value of y is shown as ln(p/(1-p)) and P is the probability that the employees would leave. According to the result (4-1), firstly, we could know the R-square ratio is 0.30 which means that about 30% of our 1470 observations could be explained by this model. Besides, the significance of independent and other control variables is also presented in the p-value columns. Then, when it comes to our research question, we must decide on a constant value for each variable so that we are able to isolate our independent variable, YearsatCompany. Therefore, we create a new dataframe (4-3), using the mean value of each variable in the mean value tables (4-2) and hold all the variables constant except for YearsAtCompany. Finally, we call the predict function and only keep YearsatCompany and Probability columns (4-4). From the result in the new dataframe, YearsatCompany and Probability have a positive relationship and the p-value in the logistic regression is also significant. Furthermore, we also use other years variables like YearInCurrentRole (4-5) and YearsWithCurrManager (4-6). However, it is interesting that they both have a negative relationship with Probability.

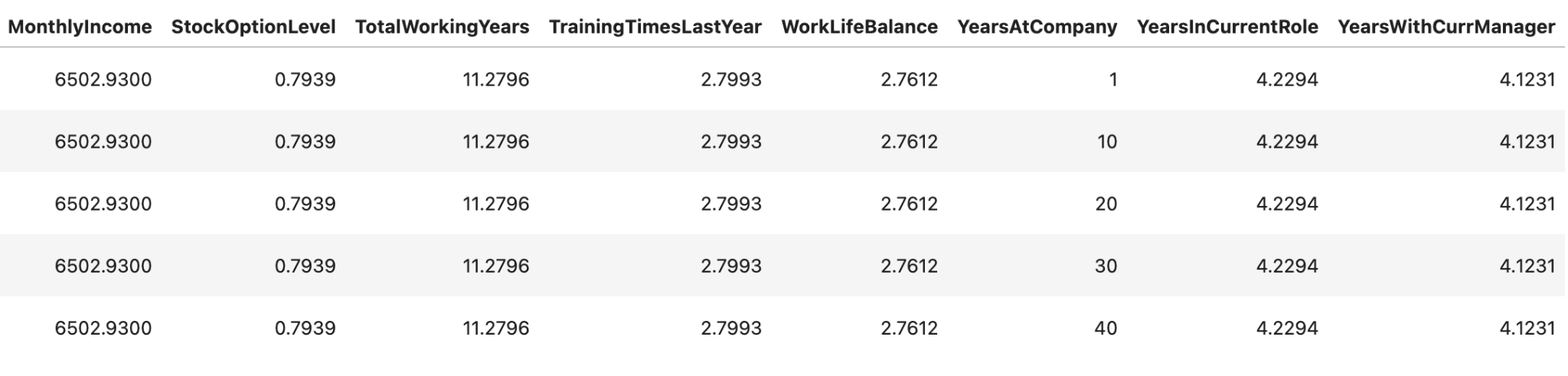
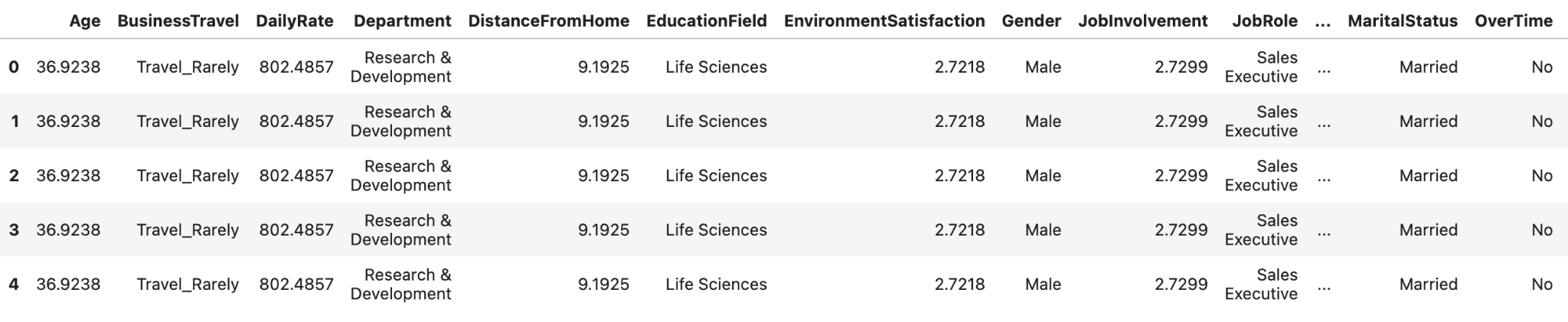


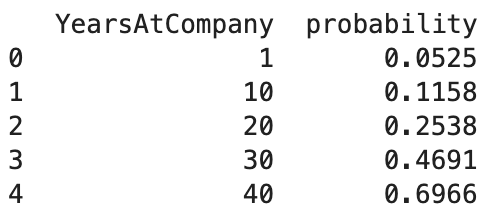
(4-1)



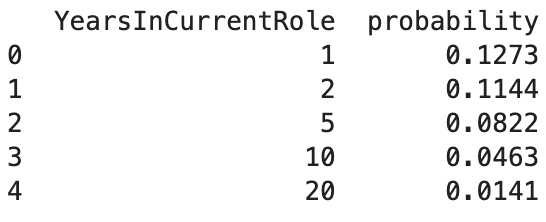


(4-2)

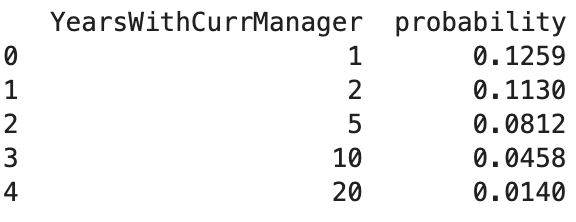
(4-3)



(4-4)

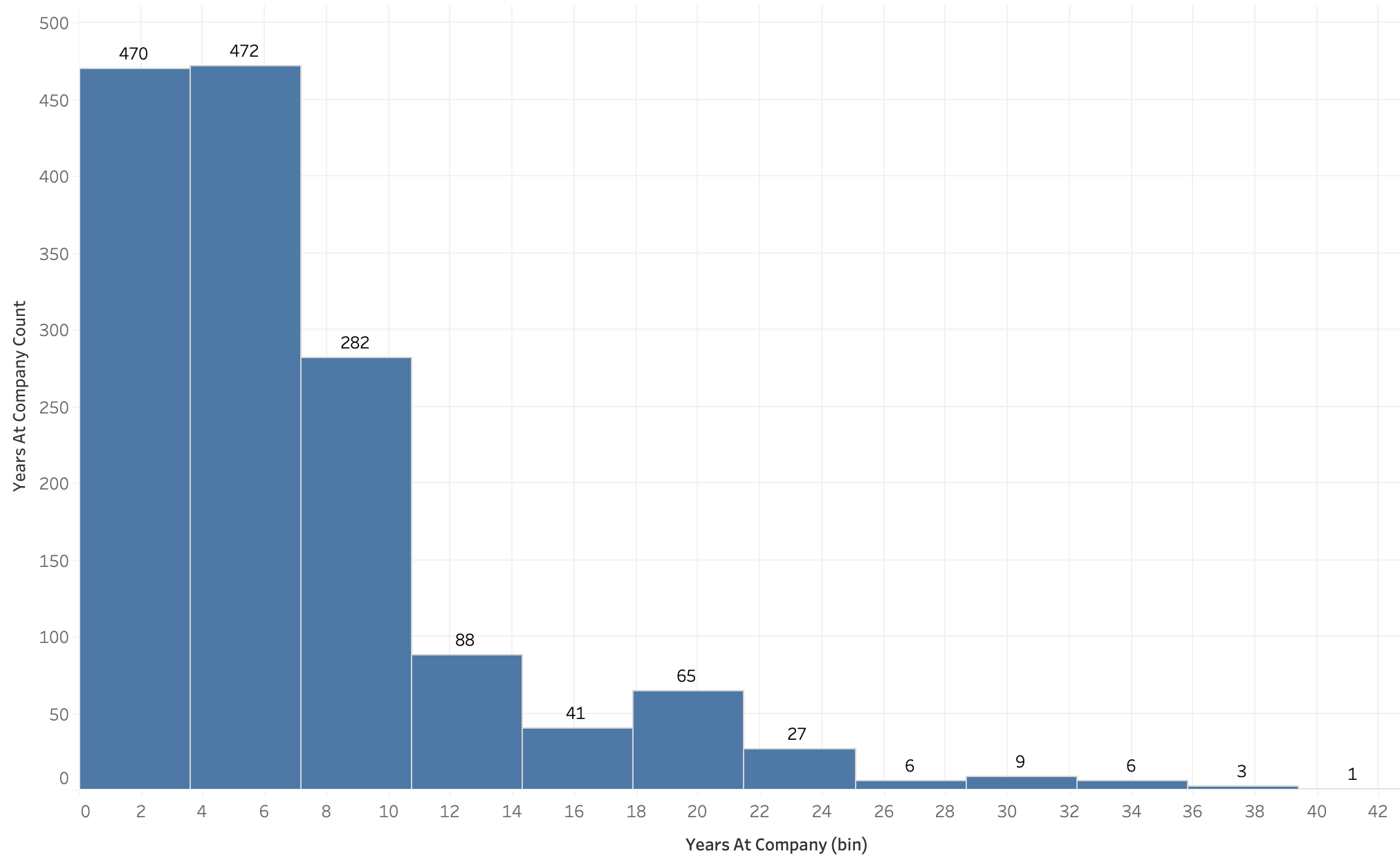


(4-5)

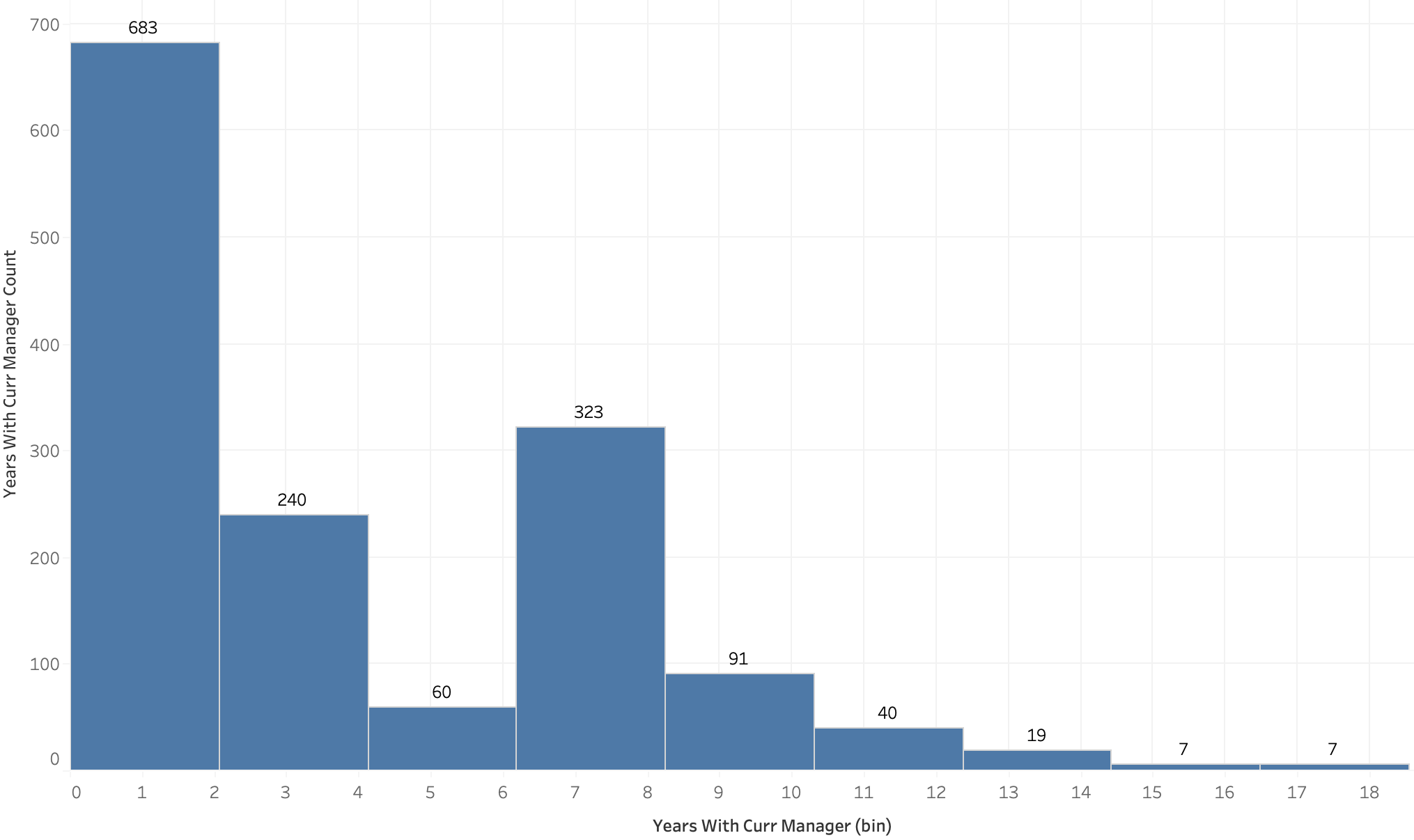


(4-6)

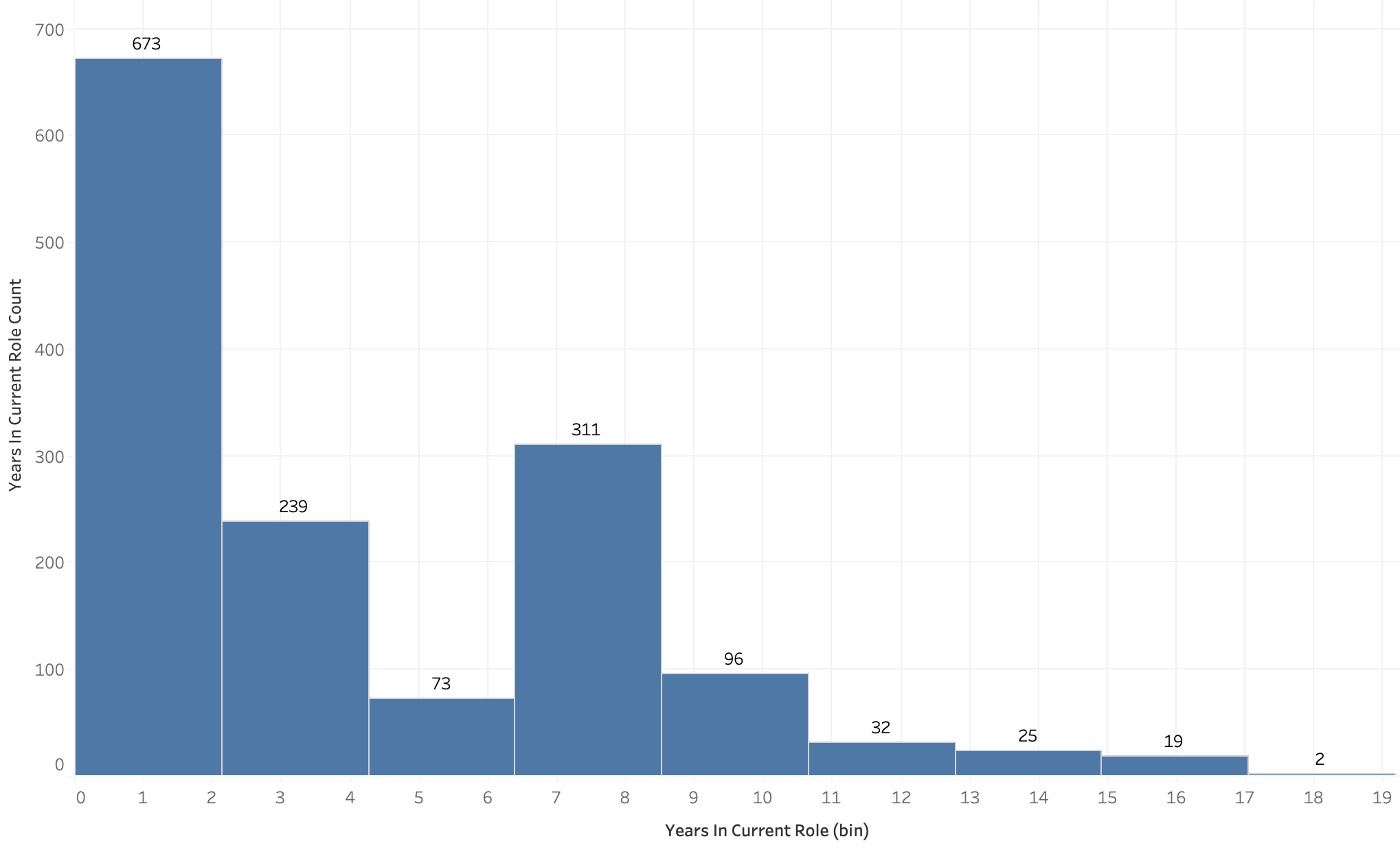
In our opinion, it is the retirement which affects this difference. To be specific, we write the distribution of these three Years variables. For YearsAtCompany (4-7), the max value of this variable is 40 and there are only a few employees staying in the company at this level. At the same time, these people are more likely to leave due to retirement, which results in a relatively high attrition rate. But for the other two variables (4-8) & (4-9), employees are comfortable with what they do and they could balance their life and work better when they stay at the company in the long term. So, they may be more likely to stay in the company and have a low attrition rate. Back to our research question, we conclude that the longer employees stay at the company, the lower the attrition rate is.



(4-7)



(4-8)

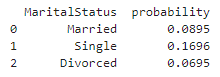


(4-9)

**5. Deeper Analysis**

Following our analysis, we identified a new business challenge: 'Junior-level employees exhibit higher attrition rates.' Our objective was to furnish the HR department with actionable solutions to address this issue. To delve deeper into the matter, we opted to conduct a more detailed analysis of the dataset. Intuitively, two strategic approaches were considered. Firstly, we aimed to identify key factors during the hiring process for junior-level employees that could contribute to lowering their attrition rates. Secondly, we sought to pinpoint variables influencing attrition and propose strategies to extend the tenure of junior employees within the company.

In exploring the first option, our analysis of the dataset revealed noteworthy distinctions in attrition rates based on marital status and gender attributes (5-1). Regression analysis demonstrated that, when holding all other variables constant, divorced employees displayed the lowest probability of leaving the company (0.0695). Married employees followed with a probability of 0.0895, while single employees exhibited the highest probability at 0.1696. Similarly, gender-based regression indicated that male employees had a lower attrition rate (0.007) compared to their female counterparts (0.0686). Although these results suggest a preference for hiring married or divorced males, adopting such discriminatory criteria in the hiring process is ethically untenable, and we strongly discourage its implementation.

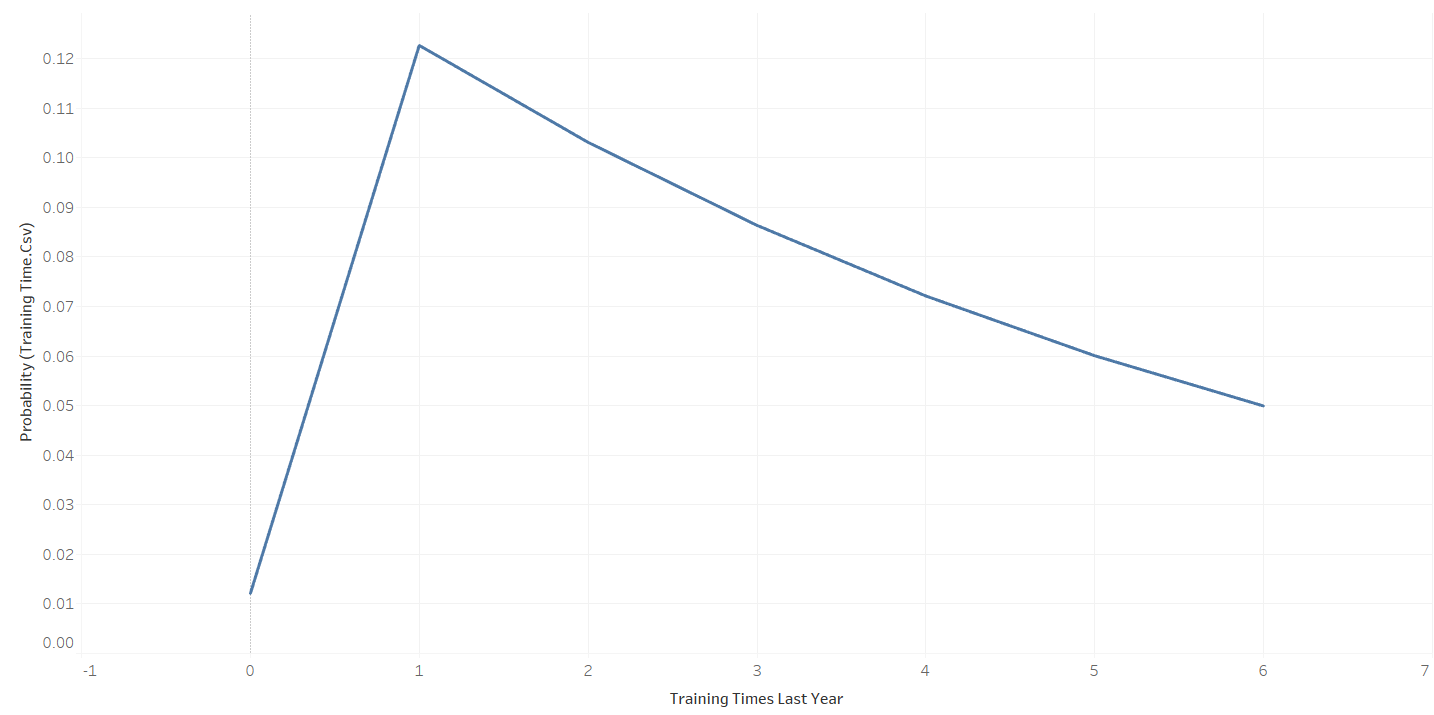
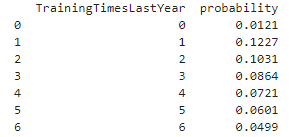




(5-1)

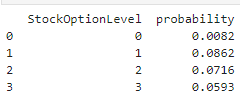
In our quest to understand the factors influencing employee retention within the company, our exploration into logistic regression produced some insights. Notably, attributes like maintaining a positive work-life balance and having higher job involvement were linked to lower attrition rates. However, the analysis also revealed some interesting findings for other variables. To dig deeper and gain more detailed insights, we've chosen to focus on three specific attributes from the dataset: training times last year, stock option level, and overtime hours.

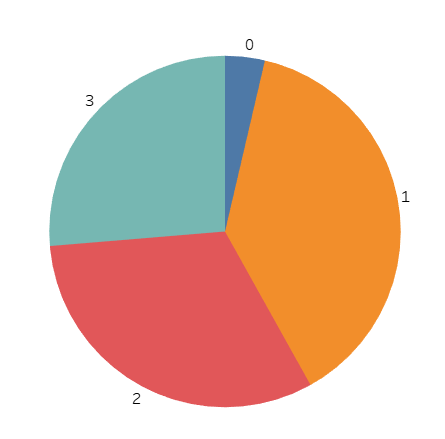
We found that employee training frequency is a key factor influencing attrition. Examining the table (5-2) below, it's evident that employees with no training exhibit minimal attrition, while those with training experience higher rates. Just one training session per year results in a substantial increase in the likelihood of leaving the company, with a probability of 0.1227. As the number of training sessions increases, the probability decreases. If your company plans to offer job training, it is essential to provide more than 3 hours of substantial training time, as this duration brings the attrition probability below 10%. If this isn't feasible, it may be better not to provide any training at all.



(5-2)

Another factor impacting attrition is the benefits provided to junior-level employees. Notably, for entry-level employees, stock options do not significantly contribute to retention. It is an ineffective benefit for retaining employees in the company. Analysis (5-3) indicates that retention is at its lowest when no stock options are offered. The attrition probability increases by 8% with the introduction of one level of stock option and does not decrease below the level observed with no stock option benefits as the level increases. We recommend that the HR department reserve stock options as incentives for promotions or higher-level roles.

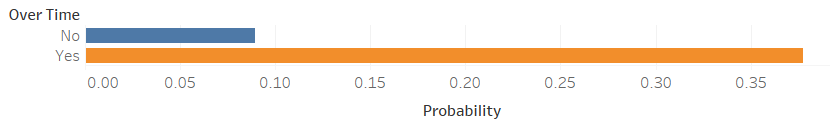


****

(5-3)

Lastly, overtime has emerged as a significant factor affecting attrition, displaying the most substantial difference in probabilities with a 30 percent point variance between employees working overtime and those who do not. The analysis (5-4) identified a 38% probability, the highest observed, associated with attrition when employees engage in overtime. Given the pronounced impact of overtime on employee attrition, we strongly recommend that companies exercise caution when imposing excessive overtime requirements on their employees.





(5-4)

**6. Conclusion**

In conclusion, our initial focus on the relationship between employee development and attrition led us to a nuanced understanding of various factors influencing employee retention within the organization. Utilizing logistic regression, we found that the longer employees stay at the company, the lower the attrition rate. This observation is particularly pronounced for YearsAtCompany, while YearsInCurrentRole and YearsWithCurrManager showed unexpected negative relationships with attrition probability. Further exploration revealed that retirement significantly impacts these differences, with the attrition rate increasing for employees with longer tenures, particularly at higher levels.

Upon deeper analysis, a new business challenge emerged—higher attrition rates among junior-level employees. Our investigation led to insights into the influence of marital status and gender on attrition rates, highlighting potential biases in hiring preferences. Ethical considerations aside, our exploration of logistic regression identified key attributes affecting attrition. Notably, maintaining a positive work-life balance and higher job involvement correlated with lower attrition rates. Additionally, employee training frequency, stock options, and overtime hours emerged as crucial factors. Recommendations include strategic planning for training sessions, reconsideration of stock options for entry-level employees, and caution in implementing excessive overtime requirements to mitigate attrition risks. Through these comprehensive analyses, we provide actionable insights to enhance employee retention strategies within the organization.

**7. Limitations and Future Improvements**

The research has some noteworthy limitations that deserve careful consideration. Firstly, the dataset's temporal scope may be restricted to a specific timeframe, potentially hindering a thorough examination of long-term trends or shifts in employee attrition patterns. Additionally, the dataset does not explicitly account for external factors such as economic conditions, industry trends, or global events, even though these variables could significantly impact attrition rates.

To enhance the research, a longitudinal analysis could be conducted by extending the temporal scope to incorporate data over multiple years. This approach would provide a more detailed understanding of trends and changes in attrition patterns over time. Additionally, to address the limitation related to external factors, it would be beneficial to integrate these influences into the analysis. By considering economic conditions, industry trends, and global events, the research can achieve a more holistic perspective, providing a comprehensive understanding of the dynamics influencing employee attrition.

**References**

[1] Assemble. (2023, March 21). Five Hidden Costs Of Employee Attrition. *Forbes*. https://www.forbes.com/sites/forbeseq/2023/03/21/five-hidden-costs-of-employee-attrition/?sh=2a9fce2062f4