Spring 2025 Writing Sample

Eli Vatsaas

2025-02-05

Contents

1 Introduction

I chose this data that looks at whether an employee stayed at or left a company based on a variety of factors. I chose this to broaden my portfolio while still doing work I find interesting, like looking at why people burnout. I hoped to learn what caused employees to leave companies. Although this dataset was synthetic, it was a worthwhile learning experience. This dataset was obtained from Kaggle, at www.kaggle.com/datasets/stealthtechnologies/employee-attrition-dataset/.

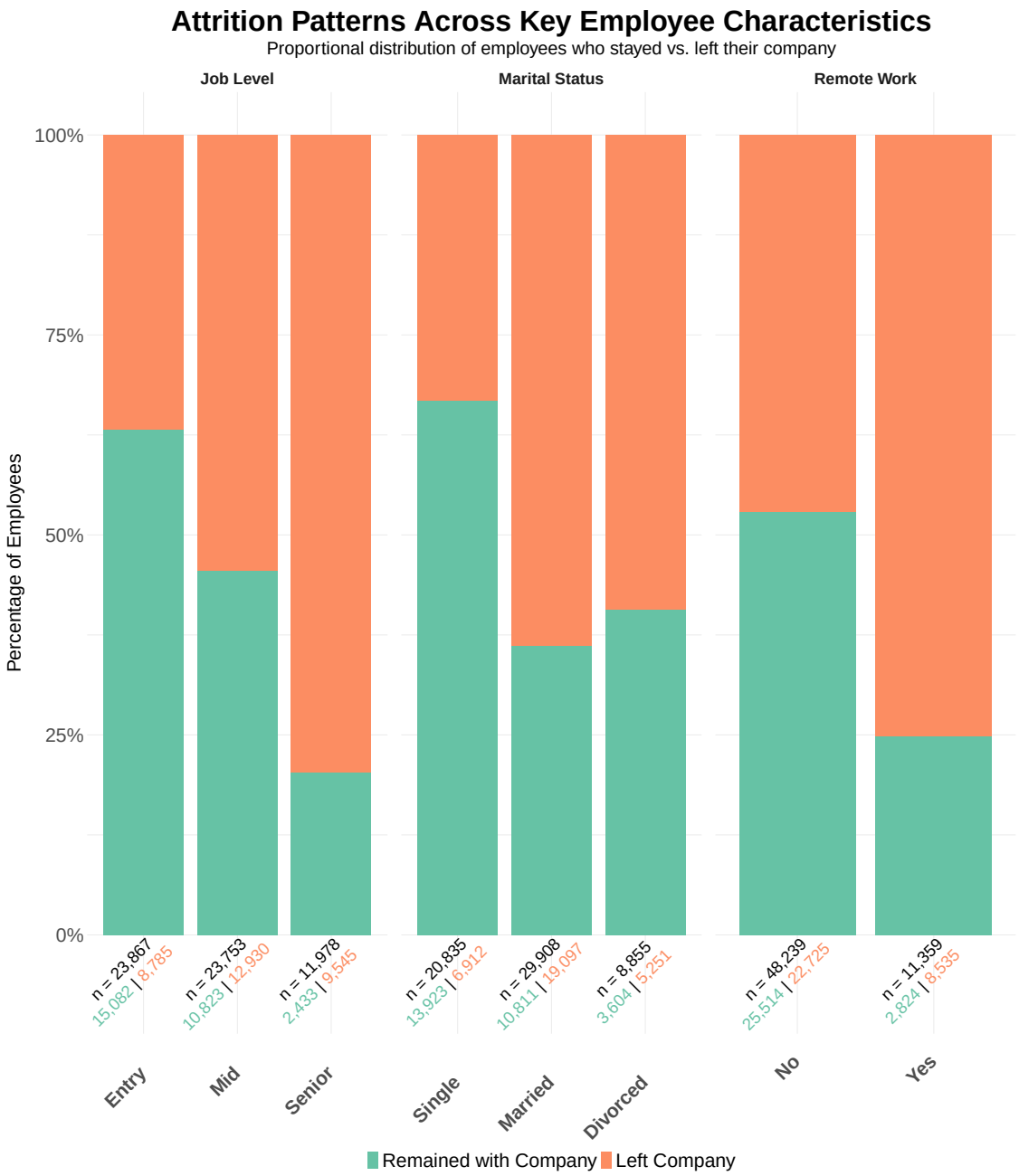
2 Data Cleaning

Since the data was synthetic, there was minimal cleaning to be done. I turned my character variables to factors, and ordered the ones where ordering mattered. I removed the employee id as it was not needed for this work. I finally removed all periods separating words in columns while adding spaces. This will allow for better visualization. For modeling one may want to turn these factors into dummy variables, and standardize the numerical data. Some of the key variables in the data set are **Attrition:** Whether an employee stayed at or left their company. *Factor (Stayed/Left).* **Job Level:** An employee's position importance in their company. *Factor (Entry/Mid/Senior).* **Age:** Age of employee in years. *Integer (Range:*18 *to* 59*).* As seen in the table below, there are no missing values, confirming tidy data.

Table 1: Summary of Variables in tidy.train Dataset

Variable	Type	Unique Values/Range	# Distinct	# Missing
Age	integer	18 to 59	42	0
Gender	factor	Female, Male	2	0
Years at Company	integer	1 to 51	51	0
Job Role	factor	Education, Finance, Healthcare, Media, Technology	5	0
Monthly Income	integer	1316 to 16149	9569	0
Work Life Balance	factor	Poor, Fair, Good, Excellent	4	0
Job Satisfaction	factor	Low, Medium, High, Very High	4	0
Performance Rating	factor	Average, Below Average, High, Low	4	0
Number of Promotions	factor	0, 1, 2, 3, 4	5	0
Overtime	factor	No, Yes	2	0
Distance from Home	integer	1 to 99	99	0
Education Level	factor	High School, Associate Degree, Bachelor's Degree, Master's Degree, PhD	5	0
Marital Status	factor	Divorced, Married, Single	3	0
Number of Dependents	factor	0, 1, 2, 3, 4, 5, 6	7	0
Job Level	factor	Entry, Mid, Senior	3	0
Company Size	factor	Large, Medium, Small	3	0
Company Tenure	integer	2 to 128	127	0
Remote Work	factor	No, Yes	2	0
Leadership Opportunities	factor	No, Yes	2	0
Innovation Opportunities	factor	No, Yes	2	0
Company Reputation	factor	Poor, Fair, Good, Excellent	4	0
Employee Recognition	factor	Low, Medium, High, Very High	4	0
Attrition	factor	Left, Stayed	2	0

This table shows us important information about the data. It shows the variable name followed by the data type that variable is. It then shows either the unique variables from that data (factor levels) or the range of the numerical variables. The second to last column shows the number of distinct values in the variable. The last column confirms the tidy data, showing no missing values. This table shows the data is ready for visualization without having to worry about handling NA's.



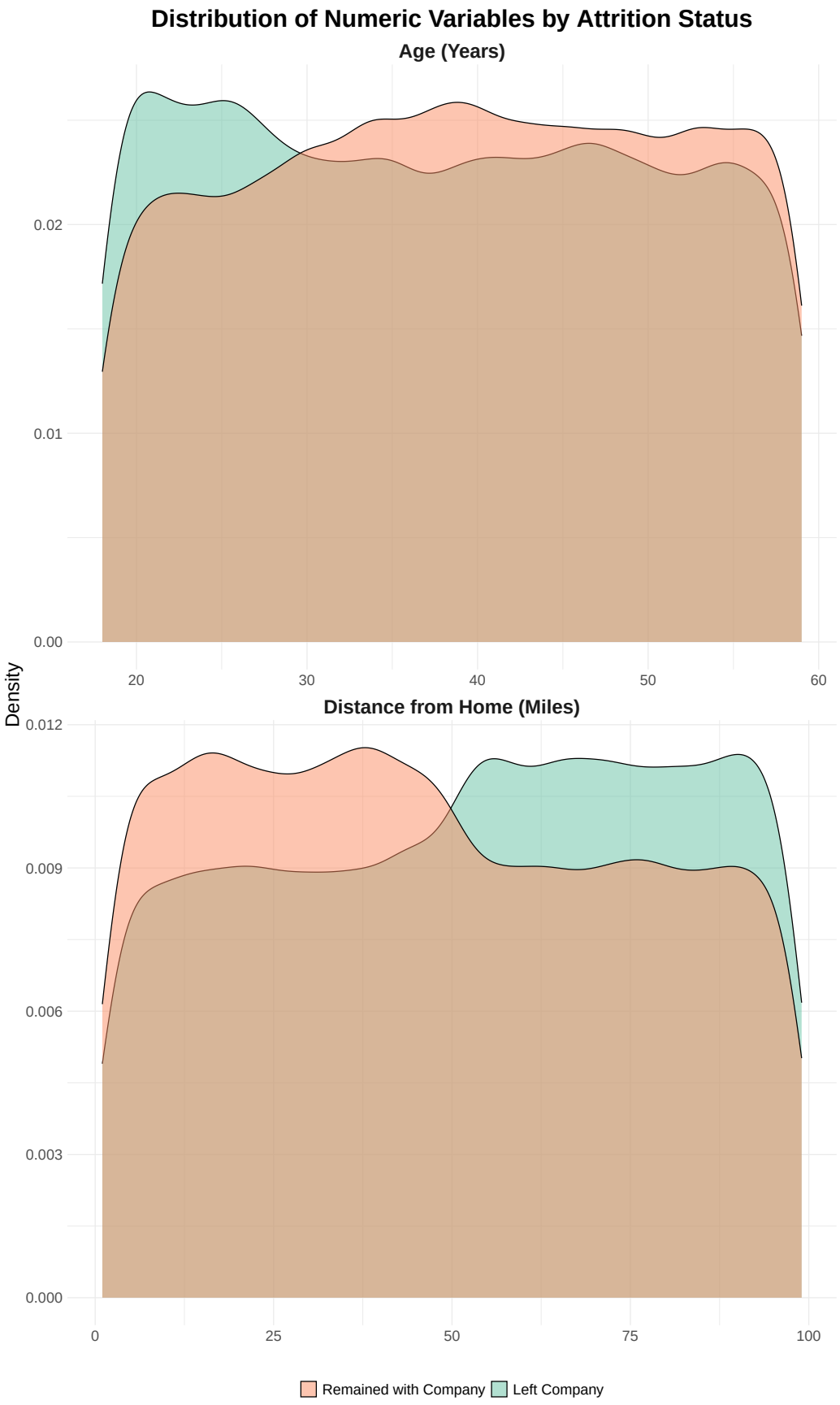
Statistical Significance:

Job Level: $X^2 = 5942.10$, $p < 2e{-}16$

Marital Status: $X^2 = 4829.50$, $p < 2e{-}16$

Remote Work: $X^2 = 2895.20$, $p < 2e{-}16$

Note: Chi−square tests indicate significant differences (critical value = 0.05)


Distribution of Numeric Variables by Attrition Status
Age (Years)
Distance from Home (Miles)
Density
0.02
0.01
0.00
0.012
0.009
0.006
0.003
0.000
20
30
40
50
60
0
25
50
75
100
Remained with Company
Left Company

3 Data Discussion

The first of the two graphs show the relationship between attrition and a few key categorical variables. Across the three variables, **Job Level, Marital Status, and Remote Work,** we see patterns that give us information about whether employees may stay. In **Job Role**, we see that *Entry* employees are likely to leave and *Senior* are likely to stay. In **Marital Status** we see that *Single* employees are likely to leave and *Married* employees may be likely to stay. In **Remote Work** we can infer remote employees are likely to stay and in office employees are more likely to leave. We can confirm these by locking at our Chi Square test standard residuals:

Table 2: Standardized Residuals from Chi-Square Tests

Group	Stayed	Left
Job Level		
Entry	-62.50	62.50
Mid	7.89	-7.89
Senior	66.78	-66.78
Marital Status		
Divorced	13.99	-13.99
Married	55.94	-55.94
Single	-69.09	69.09
Remote Work		
No	-53.82	53.82
Yes	53.82	-53.82

From these results, we confirm what was seen in the visualization. The weakeast assocation is *Mid* level employees at -7.8942759 and *Divorced* employees at -13.9850544. The rest of the grouping having strong associations with **Attrition**, the strongest being *Single* employees association with leaving. This is -69.0850705, confirming what was found in the visualizaiton.

The final two graphs show the pattern in attrition in two key numeric variables. The top graph shows the pattern in attrition in age, showing interesting patterns. This graph suggests that young employees are much more likely to leave their jobs, where around age 30 this switches to the opposite. When we run a simple ANOVA to test this, we seem to confirm this. After running an ANOVA on **Age** and **Attrition**, we see an F test statistic of 142.8468739 and p value of $6.9238073 \times 10^{-33}$. These show strong evidence for age as a factor of **Attrition.** With such a small p-value, we can believe age is a strong factor in attrition, although we can do some more checks. When we look at the mean and standard deviation of age across each attrition group, we see something else.

The mean age for those that left the company was 39.128279, while the mean age of those that stayed was 37.9454796. The corresponding standard deviations are 11.8886158, 12.2572973. This shows that the difference in age across these groups is small, with the means being just 1.1828 apart. The standard deviations are also both high around 12. This suggests that while age may be a factor in attrition, it is not the driving force, and there are other factors. We can do the same tests for the other visualized variable, distance from home, and fit an ANOVA.

With the ANOVA on **Distance From Home** and **Attrition**, we see an F test statistic of 549.3449837 and p value of $6.1788705 \times 10^{-121}$. Once again, we have a low p-value, suggesting that the distance an employee lives from the office is an important factor to whether they stay at their work. when looking at their mean and standard deviation, we see the mean and standard devaition for those who stayed at a company are mean = 52.8649869 and sd = 28.3069009. For those that left the corresponding stats are mean = 47.4174024 and sd = 28.3631104. The means show some distance, with the difference being 5.4475845, and the standard deviations hovering around 28.3350057 This suggests a lot of variablity in this data, but some evidence that distance to home is important to employees.

4 Conclusion

With some more time, I believe there would be several valuable analyses that could further enhance our understanding of employee attrition. A logistic regression model to predict **Attrition** could give more insight into these whether employees leave. Additionally, I think it would be interesting to explore what would motivate employees to stay, providing value for companies. Being particularly useful in retention strategies.

In previous projects, no single modeling method stood out as significantly superior, with accuracy rates ranging from 72% to 75%. However, key predictors such as senior job status, single marital status, and remote work status consistently emerged as strong indicators of attrition. These findings align with the patterns observed in the EDA above, reinforcing their importance in understanding employee turnover.

Another area providing promise is the interaction effects between variables.

Does the combination of remote work and distance from home have a stronger impact on attrition than either factor alone?

Does job satisfaction interact with leadership opportunities to influence employee retention?

Investigating these interactions could provide a stronger understanding of the data, though there could be some challenges. One of these challenges may be multicollinearity, which could be handled with VIF or regularization methods. Overall this dataset provides a lot of possibility for exploration and learning. Showing possible deep hidden relationships to be found.

5 Link to Github

elivatsaas/S25WriteUp