Ahti Holli 794222

Multivariate statistical analysis project

Introduction

For a company its employees are a vital part of their success. Companies often compete hard over the best talent, and they must keep the employees happy to retain them. Otherwise, they will switch to another company with a better salary and benefits. This is why it is important for HR to utilize their data to understand employee attrition better. If they can understand the features that lead to an employee switching companies, they can implement preventative actions.

This gets us to the following research question: **What are the common features of people who tend to switch jobs?**

By answering this question, we can hopefully understand the patterns that lead to employee attrition. We'll be utilizing data collected from 1470 employees, with demographic and job-related features. First, we'll do univariate analysis and try to identify some key numbers. Then we'll move to bivariate analysis, and lastly, we'll perform Multivariate Correspondence Analysis to understand the features that affect attrition.

Univariate Analysis

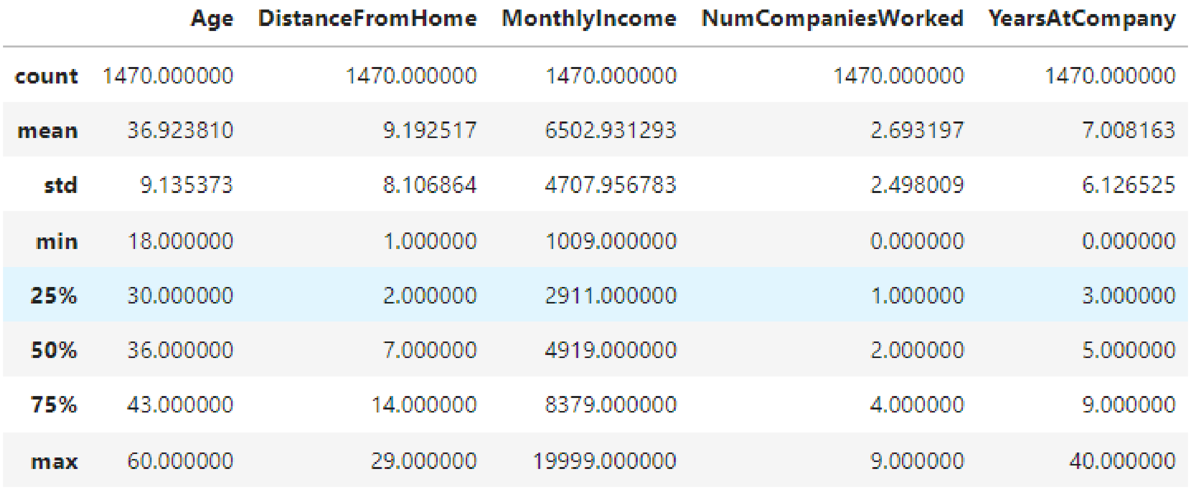
After narrowing down the dataset to 11 total variables, of which 6 are categorical and 5 are continuous.

Variable	*Modality*	*Frequency*	*Relative Frequency*
Attrition	Yes	237	83.9%
	No	1233	16.1%
Gender	Male	882	60%
	Female	588	40%
Work Life Balance	1	80	5.4%
	2	344	23.4%
	3	893	60.8%
	4	153	10.4%
Education	1	170	11.6%
	2	282	19.2%
	3	572	38.9%
	4	398	27%
	5	48	3.3%

Business Travel	Non-Travel	150	10.2%
	Travel Rarely	1043	70.9%
	Travel Frequently	277	18.9%
Marital Status	Single	470	31.9%
	Divorced	327	22.2%
	Married	673	45.9%

Table 1. Descriptive statistics of the categorical variables

From the table we can see that a very large proportion of the data hasn't actually left their respective companies. The gender column isn't heavily skewed to either side which means we'll get a good representation from both genders. All of the other variables seem to be quite evenly distributed, with a lot of the observations centered around the middle values like in the education variable. Next table will feature the continuous variables.



	Age	DistanceFromHome	MonthlyIncome	NumCompaniesWorked	YearsAtCompany
count	1470.000000	1470.000000	1470.000000	1470.000000	1470.000000
mean	36.923810	9.192517	6502.931293	2.693197	7.008163
std	9.135373	8.106864	4707.956783	2.498009	6.126525
min	18.000000	1.000000	1009.000000	0.000000	0.000000
25%	30.000000	2.000000	2911.000000	1.000000	3.000000
50%	36.000000	7.000000	4919.000000	2.000000	5.000000
75%	43.000000	14.000000	8379.000000	4.000000	9.000000
max	60.000000	29.000000	19999.000000	9.000000	40.000000

Table 2. Descriptive statistics of the continuous variables

Variable	*Skew*	*Kurtosis*
Age	0.413	-0.404
Distance from home	0.958	-0.225
NumCompaniesWorked	1.026	0.010
Years at Company	1.764	3.936
Monthly Income	1.369	1.005

Table 3. Skewness and kurtosis statistics of the continuous variables

From the table we can see that especially the columns MonthlyIncome and Years at Company have notably higher means than their median. This suggests that there are outliers that drive the mean significantly higher than the median. This can be later confirmed when looking at the kurtosis values that are clearly over 0 for these columns. To visualize this, boxplots were created to visualize these potential outliers.

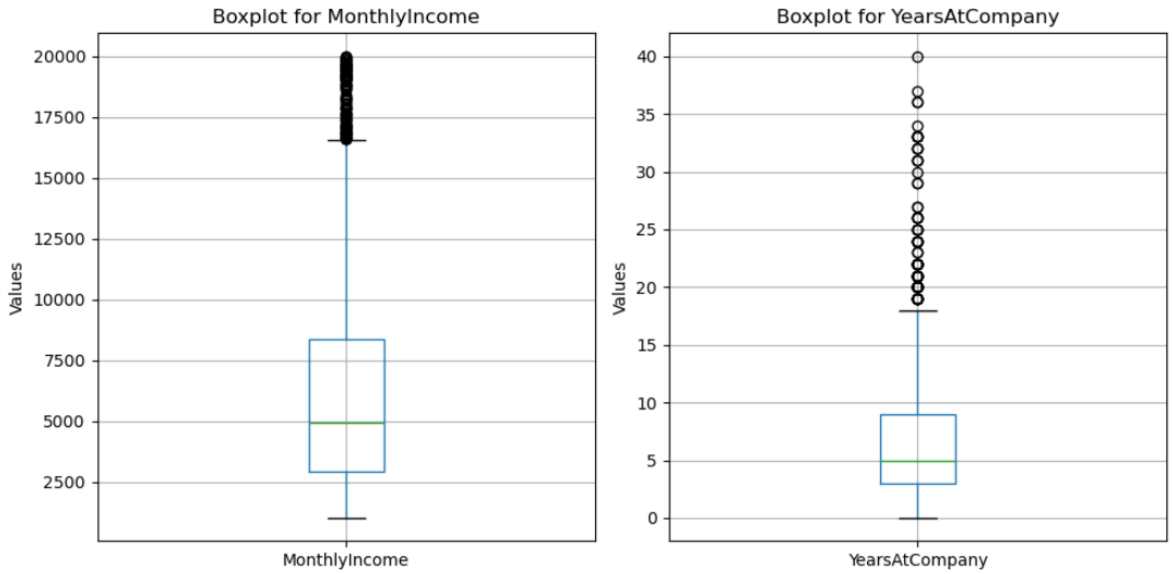


Figure 1. Boxplots of MonthlyIncome and YearsAtCompany

Lastly, I categorized the continuous variables so that each class has the equal number of observations, as Multivariate Correspondence Analysis is designed for categorical data, and with categorical data we'll also be able to calculate the attraction repulsion indices.

Bivariate analysis

Continuous variables

A correlation matrix was created to visualize the relationships between the continuous variables. The larger the absolute value number is, the larger the negative or positive correlation between the variables is. Monthly income seems to have the highest correlations with other variables. Age and years at the company seem to be moderately correlated with the persons income. Otherwise there aren't any notable correlations between the variables in the set.

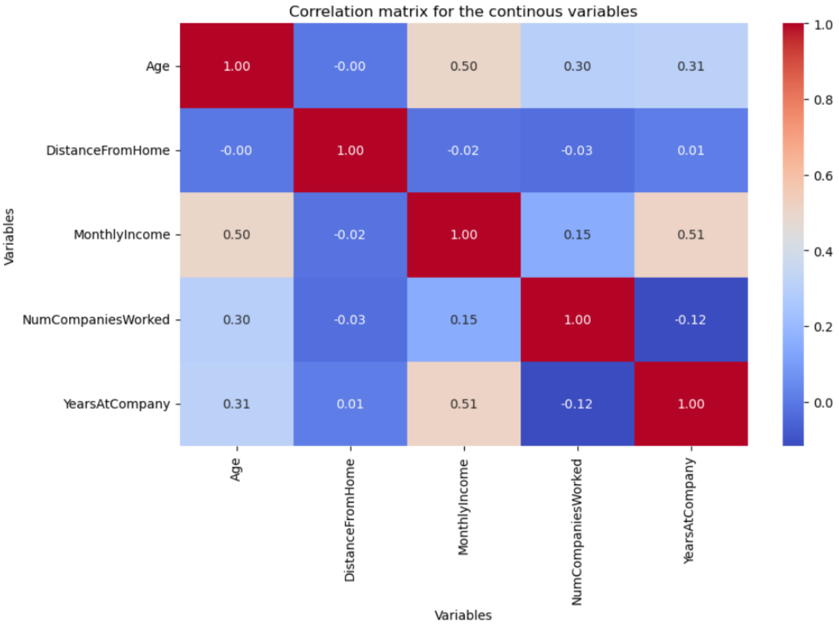


Figure 2. Correlation matrix of the continuous variables

Categorical variables

After transforming the continuous variables into categorical, attraction repulsion indices were created which imply how common features are in other features. Values below 1 imply that the variables repulse each other, while values over 1 signal that features attract each other.

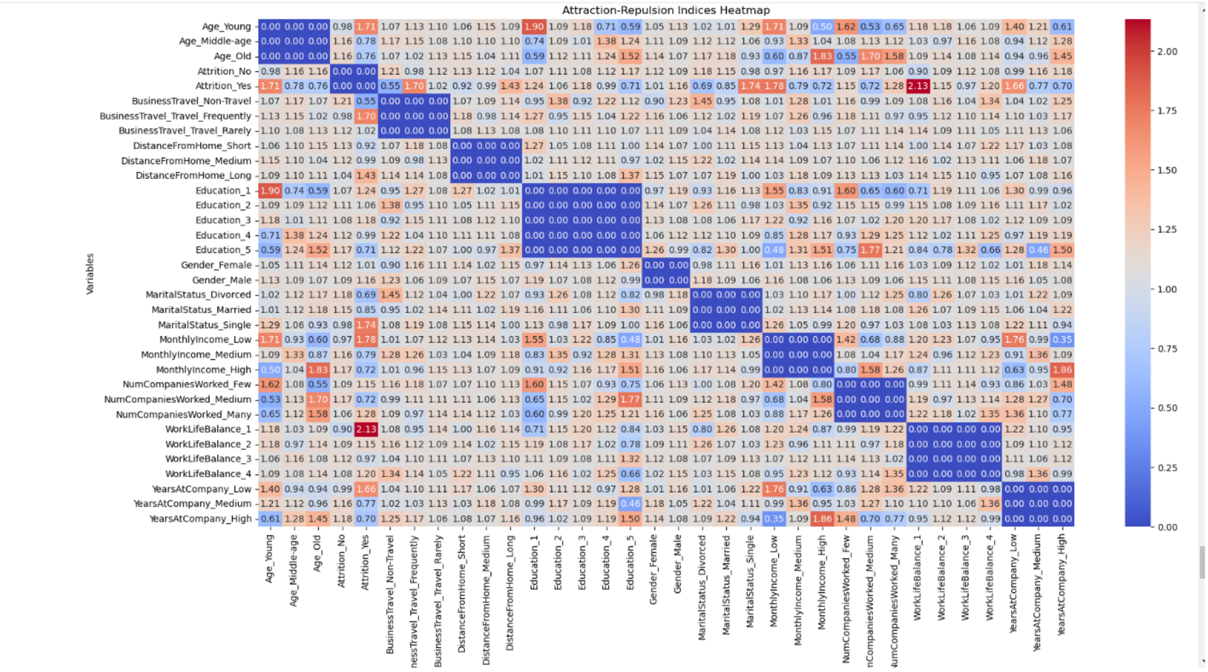


Figure 3. Attraction repulsion indices

From the matrix we can see that most notably, people with a bad perception of their work life balance tend to more often be the ones who leave the company. People who must travel a lot for work, are single, young and have a low income tend to be the ones who leave the company more often. Also, it can be noted that people who have spent a long time at the company are less likely to leave. So, a conclusion can be made that the people who haven't "settled down", are more often those who leave the company.

Multivariate Analysis

For the multivariate analysis part of the study, I performed a multiple correspondence analysis. Multiple correspondence analysis is a dimension reduction method, specifically designed for categorical variables. MCA allows us to visualize and interpret the relationships between categories and observations in high-dimensional categorical datasets. For this I used the categorized variables, since continuous variables aren't optimal for MCA.

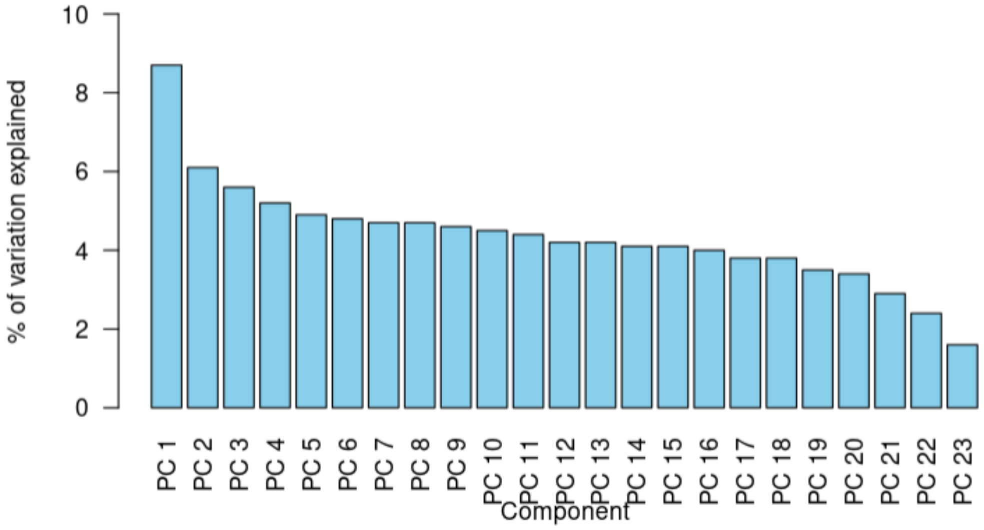


Figure 4. Scree plot of the MCA

The first two components explain 14.8% of the total variation in the dataset, which isn't a great amount, but still we can draw some sort of conclusions based on it. The drop-off after the first component is very gradual which means that each of the components explains almost as much of the variance as the previous ones.

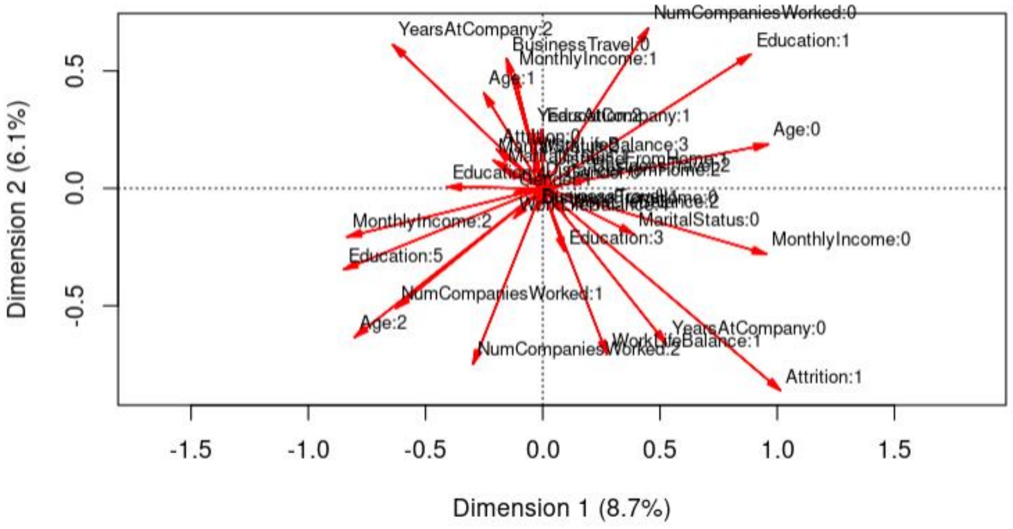


Figure 5. Biplot of the MCA

The MCA biplot is a visualization from which we can interpret categories and their correlation with each other. The proximity of points and direction of arrows suggest positive correlation, while the vice versa suggests negative correlation. More specifically, if the angle between arrows is less than 90, positive association can be assumed and if it’s more we can assume they repulse each other.

From the biplot we can see that the same results are present as in the bivariate analysis part. People who consider having bad work life balance, haven’t worked a long time at the company, and low income are the ones who tend to leave their job more often. It is to be noted that this biplot covers only roughly 14% of the total variation in the dataset, so it doesn’t tell the whole truth.

Conclusions

At the beginning of the project, a research question was stated: **What are the common features of people who tend to switch jobs?** The analysis provided some information on what things affect the event that an employee leaves their company. The most important factors for this were the number of years worked at the company, their perception of a bad work life balance and low monthly income. Furthermore, based on the analysis, the people that haven’t settled down in life tend to be those who leave their workplace.

Companies who would like to avoid their workers from leaving, they should invest more into their young and new employees. They should make sure their employees are happy and aren’t overloaded with work. This could be studied through various surveys done on the employees and offering support in various areas of life.

These results should still be considered with caution, as for example in the MCA part we saw that the two components with the most explanatory power on the variance covered only 14% of the total variance. This suggests that there is room for improvement in the analysis. For example, the choice of MCA might've not been optimal, the dataset was slightly skewed in terms of 1/6 of the observations covering those who've left a company and the variable selection process could've been more thoroughly researched. By addressing these things, we might be able to understand the reasoning behind employee attrition better.

The source of the dataset:

https://www.kaggle.com/datasets/thedevastator/employee-attrition-and-factors