Heart Disease

Ricardo Aguila
Joshua White

December 8, 2022

Executive Summary

For this project, we performed Hierarchical clustering on data from the Indian airline East-West Airlines. The data contains 12 variables and 3999 observations. We clustered these observations into 4 clusters that should be useful for establishing a more effective awards program and 1 cluster containing a single extreme outlier. The clusters are best summarized as:

1. Infrequent Flyers
2. Prime Incentive
3. Former Customers
4. Frequent Flyers
5. Extreme Outlier

Introduction

This data set came from the East-West Airlines passengers who belong to the airline's frequent flier program, with the goal of trying to cluster the passengers based off their similarities. This data set contains 11 variables that we analyzed, with the 12th variable being ID which was immediately removed from the data set. The variables in this data set are Balance (Number of miles eligible for award travel), Qual miles (Number of miles counted as qualifying for Topflight status), cc1 (Number of miles earned with freq. flyer credit card in the past 12 months miles), cc2 miles (Number of miles earned with Rewards credit card in the past 12 months), cc3 miles (Number of miles earned with Small Business credit card in the past 12 months), bonus trans (Number of miles earned from non-flight bonus transactions in the past 12 months), Flight miles 12 months (Number of non-flight bonus transactions in the past 12 months), flight trans (Number of flight miles in the past 12 months), days since enroll (the number of days since joining), and award. The first step was to analyze histograms of each variable to determine if the data needed to be scaled or transformed.

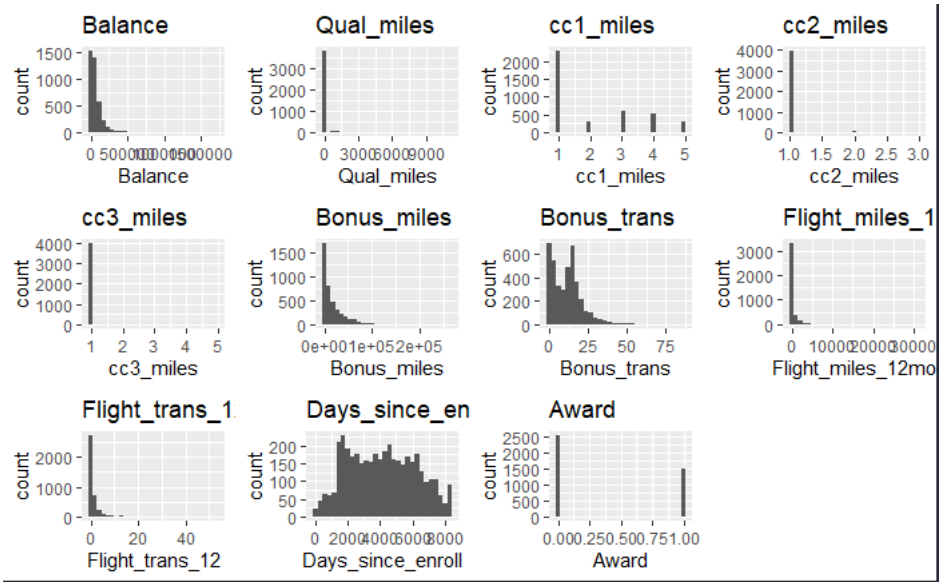

Figure 1: Histograms of each variable

Initially all of the variables were listed as numeric, but after looking at the variables a bit closer we found that several of them should be considered as categorical. The variables cc1 miles, cc2 miles, cc3 miles, and award were transformed into ordinal categorical variables. We will go into details on how we handled these variables later. As for the numeric variables we wanted to make sure that they were about normal. After displaying all of the variables in a histogram we found that the only variable that seemed normal was Days since enrolled. To prepare for cluster analysis, we needed to transform the other numeric variables to be approximately normal. We applied a log transformation to the numeric variables, and added a small constant value to each variable to avoid taking the logarithm of zero. In figure 2 is what the data looked after the transformation.

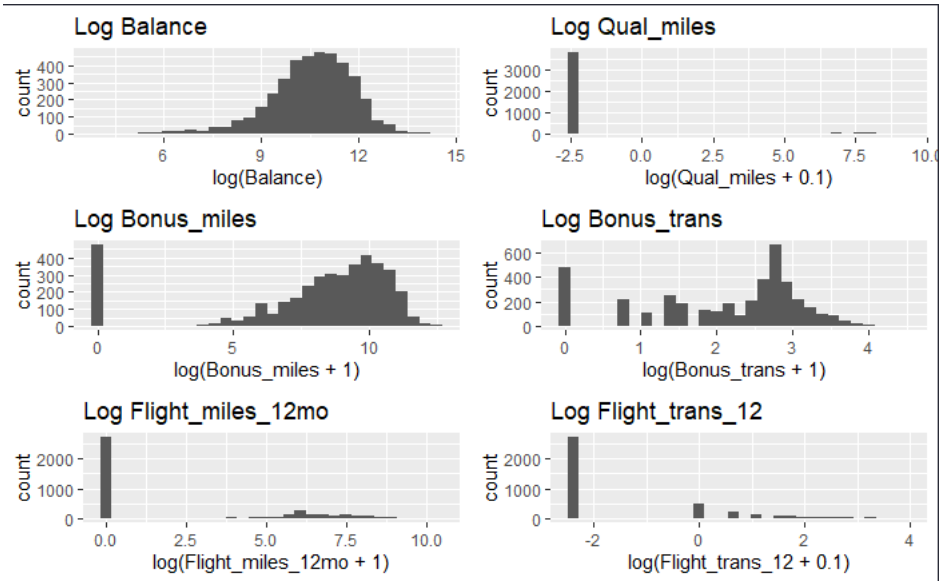

Figure 2: The numeric variables after the transformation

The log transformation improved the normality of variables like Balance and Bonus miles but did not work as well on the other numeric variables due to a high observation of zero values. The overwhelming number of zeros in each of these variables seems to make it quite difficult to scale the data to have a normal distribution. An idea that could be used to alleviate this issue is to find certain cutoff values that we deem important and change those numeric variables into categorical variables. Transforming numeric variables with a high number of zero values into categorical variables could be a useful alternative that preserves unique clusters while addressing non-normality. However, the approach that we decided to take instead was to see how the variables looked when all of those zero observations were removed.

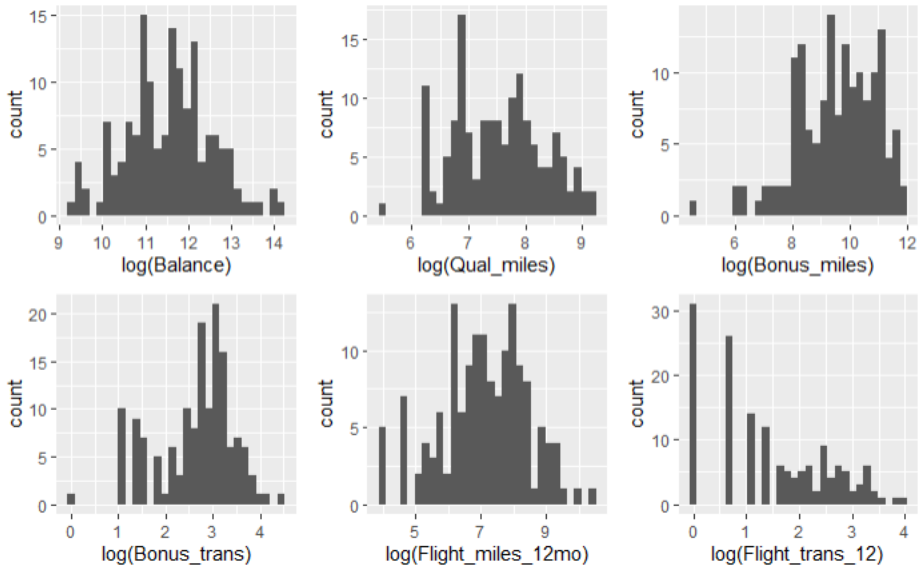

Figure 3: Histograms of the numeric variables after the zeros were removed and transformed

After all the zeros were removed from each numeric variable it made a large improvement in the normality of the data, with the exception of the Flight trans variable which looks better now but is still not normal. This decision had the greatest impact on the variables Qual miles and Flight miles 12 months, which before

were so overwhelmingly filled with zeros that the rest of the data seemed irrelevant. Now we can see that these variables are approximately normal after this transformation and we will now proceed to the clustering aspect of the paper.

Clustering Method

Although this data was presented as a hierarchical clustering problem, we wanted to make sure there was a clear reason to use hierarchical clustering over something more familiar like K-means clustering. To do this we compared scree plots of K-means vs Hierarchical clustering. It should be noted that these two plots are not comparable quantitatively, but they are still visually useful in determining an optimal number of clusters.

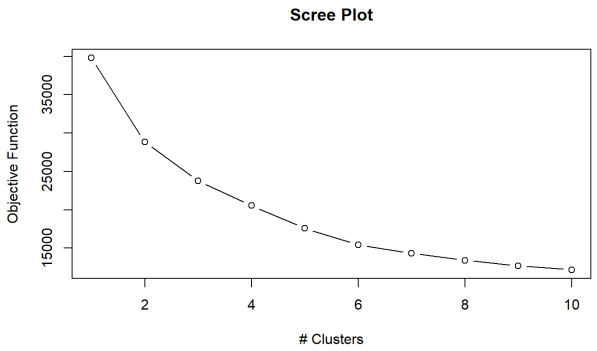

Figure 4: K-Means Scree Plot

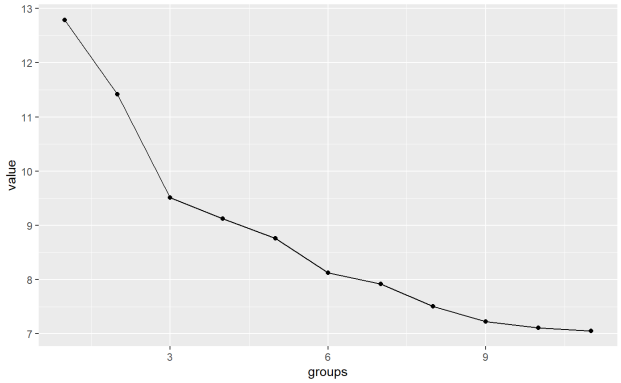

Figure 5: Hierarchical Scree Plot

Immediately, it was apparent that the elbow method would not be a useful way to decide the number of clusters for K-means. Alternatively, using a plateau approach on the Hierarchical data yields distinct numbers of clusters to consider, particularly 5, which is what we decided to go with.

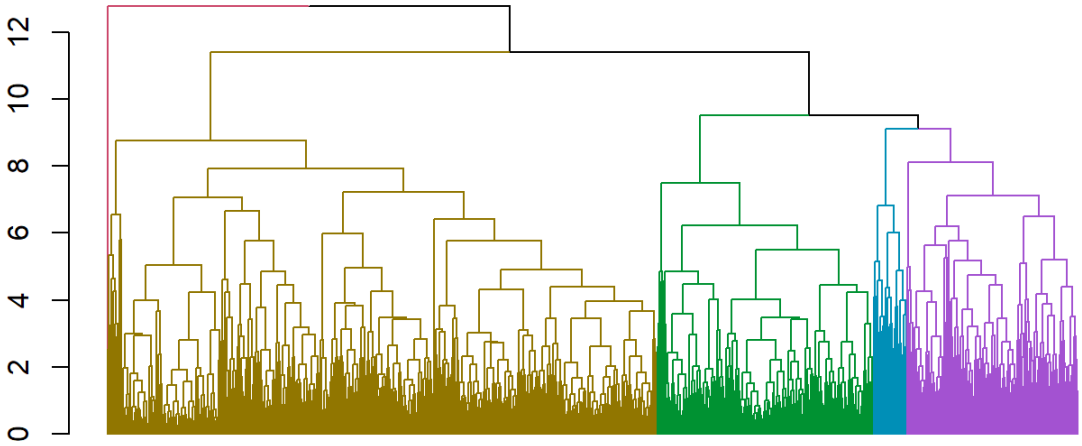

Figure 6: A dendrogram visualizing the 5 clusters.

We can observe that the clusters are not of similar size, but the size of the clusters does not determine their usefulness for designing a new rewards program. After creating these 5 clusters, we visualized them using principal component analysis:

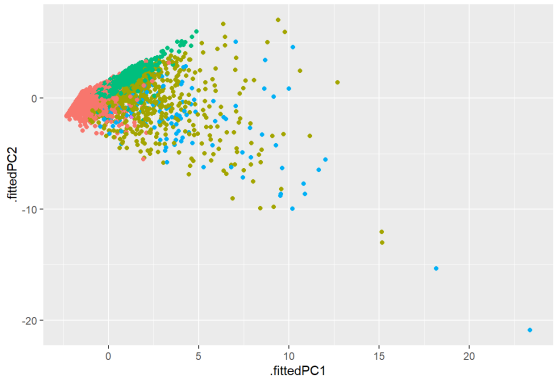

Figure 7: Plot of PC1 and PC2

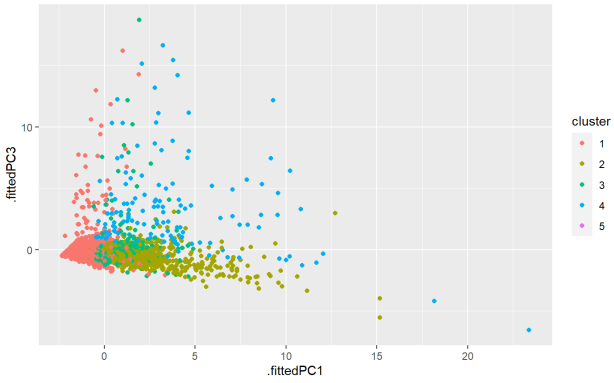

Figure 8: Plot of PC1 and PC3

Between the two plots, we can observe that the data is very easily separable with only three principal components. For detailed cluster analysis and to define our groups, we utilized the boxplots shown in the next section.

Detailed Comparison of Groups

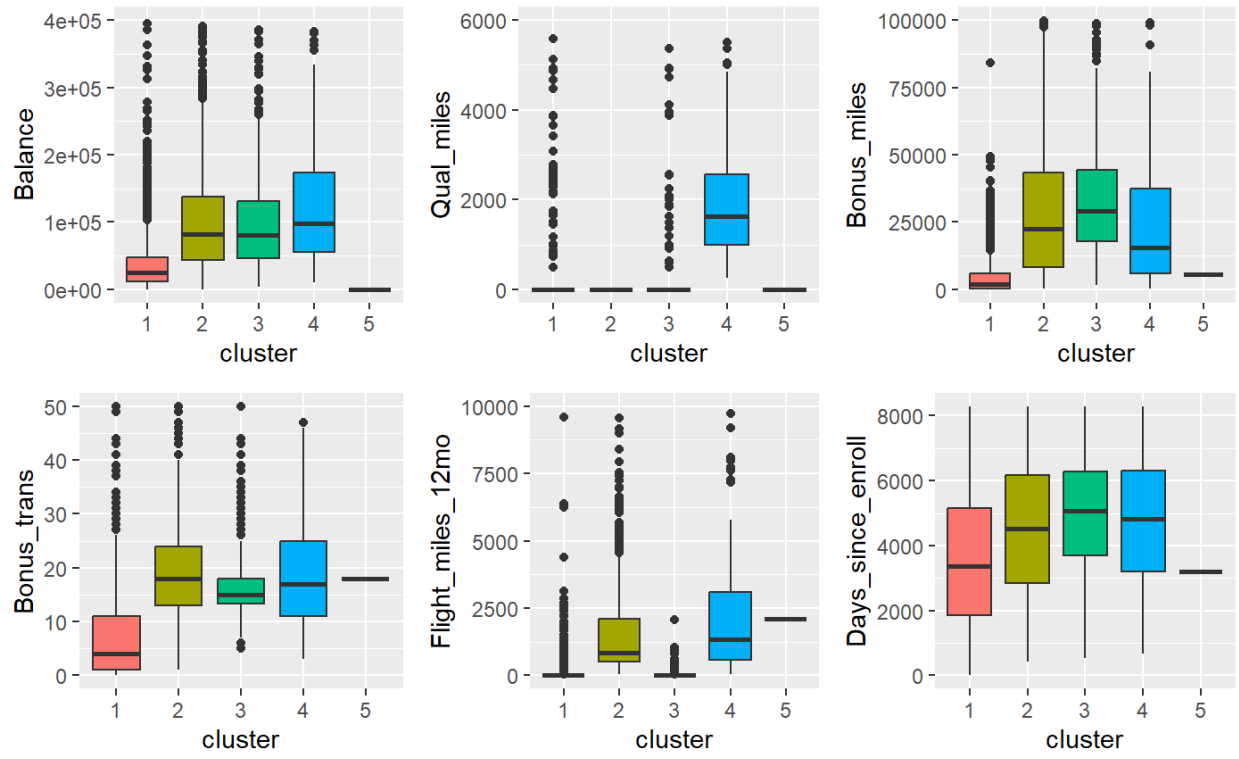

Figure 9: Box Plots showing the meaningful differences between the clusters.

1. In many ways, this cluster is the "Infrequent Flyers Group". This group has low median values for all of the award accumulation methods. Other than a few outliers, the individuals in this cluster did not receive any award flights. We believe that this category is not a prime target for award flights.

2. This cluster is the "Prime Incentive Group." The median individual in this cluster received an award flight. This cluster is extremely similar to cluster 3, but the main difference between the two clusters is this cluster has a much higher average flight miles within the past 12 months. This cluster is a prime target for awards, as they travel frequently, but many of them are still not taking advantage of award flights. We recommend offering further incentives to this group, so that they might see the benefits of loyalty towards East-West Airlines.

3. This is the "Former Customers Group." The median individual in this group received an award flight. This group is very similar to cluster 2, but the individuals in this group have on average traveled almost no East-West Airline miles in the past 12 months. These individuals may be incentivized to return, but we advise caution since most of them have received rewards already. These individuals may have changed situations resulting in them no longer needing to fly frequently, or they may have been enticed by another airline.

4. This is the "Frequent Flyers Group."Almost everyone in this group received an award flight. We believe that this group is likely already at peak saturation. Further flight awards for this group will likely not result in meaningful benefit to East-West.

5. This cluster is a single extreme outlier. Somehow, this individual received an award flight despite not having any miles eligible for award travel. They may have received their flight through credit card incentives alone. The airline may wish to examine the economic viability of the method this individual used to receive and award flight, because they are not contributing economically to the airline.

Conclusion

After our analysis we came to the conclusion that the data could be grouped into 5 different clusters, with one of these clusters only containing a single observation and representing an extreme outlier from the rest of the data set. As for the other four clusters we labeled them as follows: Infrequent Flyers Group, Prime Incentive Group, Former Customers Group, and Frequent Flyers Group. These groups should hopefully enable the airline to understand how their members choose to use their memberships and update their rewards program accordingly.