Matija Susic and Millie Chen
CSCI-UA 475 Predictive Analytics
Due October 23, 2024

Short Term Paper 2

The first clustering approach that we took was K-Means. We determined the number of clusters to use by comparing silhouette scores across different models (using values from 2 to 20 for the n_clusters hyperparameter). Next, we plotted the number of clusters and their respective silhouette scores and found that having 2 clusters results in the highest silhouette score of 0.1133. We then trained a KMeans model with 2 clusters and assigned the predicted cluster number to each observation. These clusters will be referred to as cluster 0 and cluster 1. Note: in this clustering approach, variables like revisit intent and fairness have 3 features associated with it (they have a larger “weight” compared to others) and that may lead to a large difference in the values for these features between the 2 clusters.

One thing that stood out was that for cluster 0, there was a higher proportion of frequent customers compared to first time customers whereas for cluster 1, there was a slightly higher proportion of first time customers. From the “Nightclub_Frequency” column, it appears that overall, customers in cluster 0 tend to go to nightclubs less often compared to customers in cluster 1 (Figure 1). For cluster 0, around 33% of customers go 2-3 times a month and around 26% go less than once a month. For cluster 1, around 33% of customers go once a week, 24% go 2-3 times a month, and 20% go 2-3 times a week. In terms of the employment status, there are a lot of customers that are employed (43%), retired(19%) or students(33%) for cluster 0 compared to the majority of customers being students(49%) or full time employees(33%) for cluster 1 (Figure 2). In terms of income, cluster 0 has customers that are roughly evenly distributed across income groups whereas cluster 1 has more customers in the lower ranges (under $49.999) or in the higher ranges($100.000 or over) and less in between.

There were also different trends for the pricing strategy between the 2 clusters: for cluster 0, about 29% of customers made “Reservations in advance” while for cluster 1, a large proportion of customers used “Day of the Week” or “Time of a Day” pricing strategy (about 23% and 21% respectively). The proportion of customers using pricing strategy “VIP entrance” and “Flat pricing” were around 20% for each category for both clusters (Figure 3). From this, we would suggest the nightclub to tailor “Reservations in advance” for customers in cluster 0 and “Day of the Week” or “Time of a Day” pricing strategies for customers in cluster 1. For the fairness, word of mouth, and revisit intent ratings, customers in cluster 0 tend to give lower ratings compared to customers in cluster 1 (see figures 4-6). From this, the nightclub can improve customer satisfaction by specifically targeting customers in cluster 0 and asking them more questions to learn more about what specific areas they were dissatisfied with. Customers in cluster 0 (who gave lower pricing fairness ratings) are also less familiar with the nightclub’s pricing strategy compared to cluster 1: the average rating for “FAM1” and “FAM2” were 3.420290 and 3.028986 for cluster 0 and 4.607735 and 4.165746 for cluster 1 (the ratings are on a scale from 1 to 7). The nightclub can address this issue (and potentially increase their fairness ratings) by making their pricing strategies more clear, especially to customers in cluster 0.

Figure 1

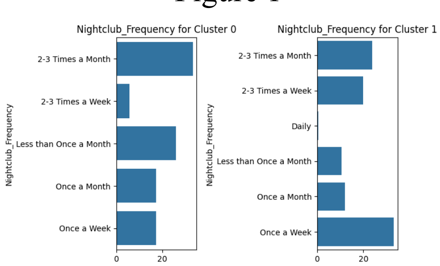


Figure 2

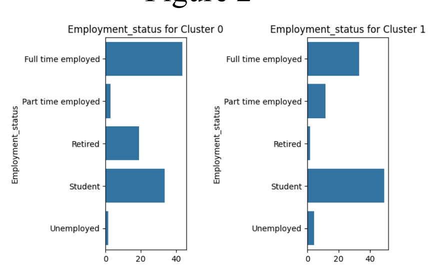


Figure 3

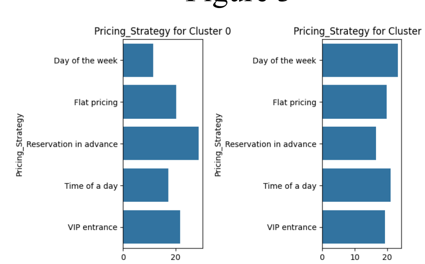


Figure 4

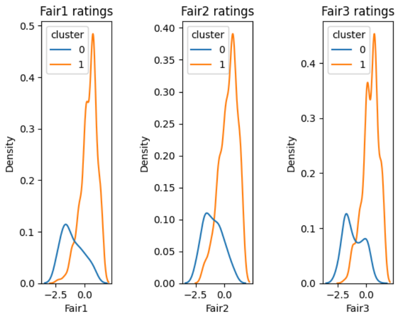


Figure 5

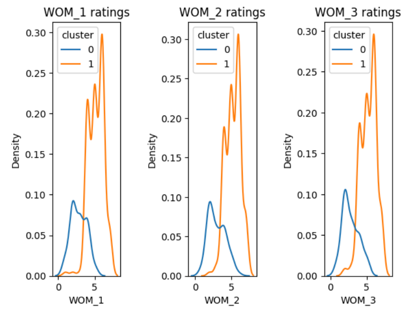
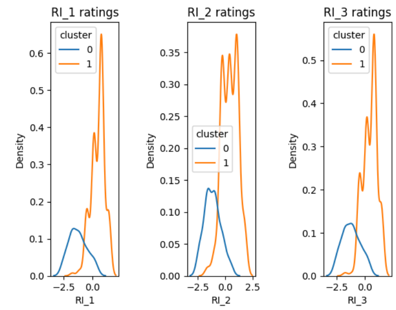
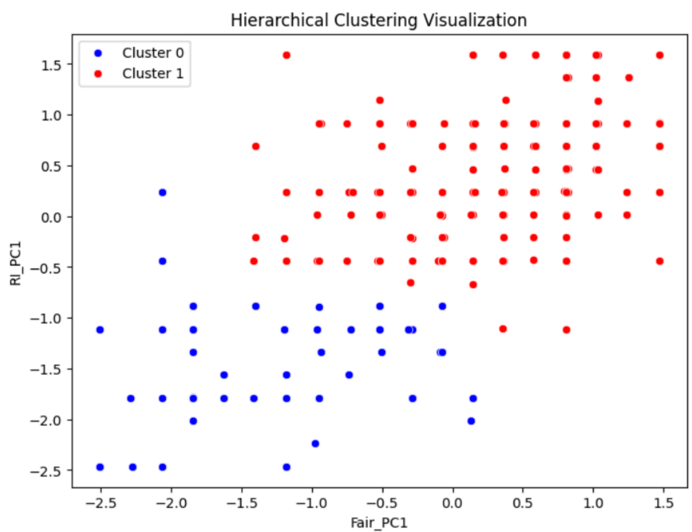


Figure 6



The second clustering approach that we took was Hierarchical Clustering. We determined the number of clusters to be 2 by looking at the dendrogram. We opted for the average linkage as the distance metric between clusters as that gave the best silhouette score (0.5279). We then trained a Agglomerative model with 2 clusters and assigned the predicted cluster number to each observation. These clusters will also be referred to as cluster 0 and cluster 1 as in the previous model. Note: in this clustering approach, variables like revisit intent and fairness which have 3 features associated with them, were reduced to a single variable each by doing PCA to help deal with the curse of dimensionality. Here's how the clusters ended up looking like:



Interestingly, hierarchical clustering results closely mirror the findings from the K-Means clustering approach. Specifically, the same patterns emerge in both analyses: Cluster 0 shows a higher proportion of frequent customers, with a tendency to visit nightclubs less frequently, and a wider distribution across employment and income groups. Cluster 1, on the other hand, has a slightly higher proportion of first-time customers, with more frequent nightclub visits and a higher concentration of students and full-time employees. In terms of pricing strategy, both methods found similar usage patterns for 'Reservations in advance' in Cluster 0 and 'Day of the Week' or 'Time of a Day' strategies in Cluster 1, along with comparable proportions for 'VIP entrance' and 'Flat pricing'.

A final idea that we'd propose to the nightclub manager is lowering the price/giving discounts for regular customers (as they seemed moderately less satisfied with the fairness of prices).

Millie - KMeans code and corresponding portion in report, Data Preprocessing
Matija - PCA, Hierarchical Clustering code and corresponding portion in the report, Data Preprocessing