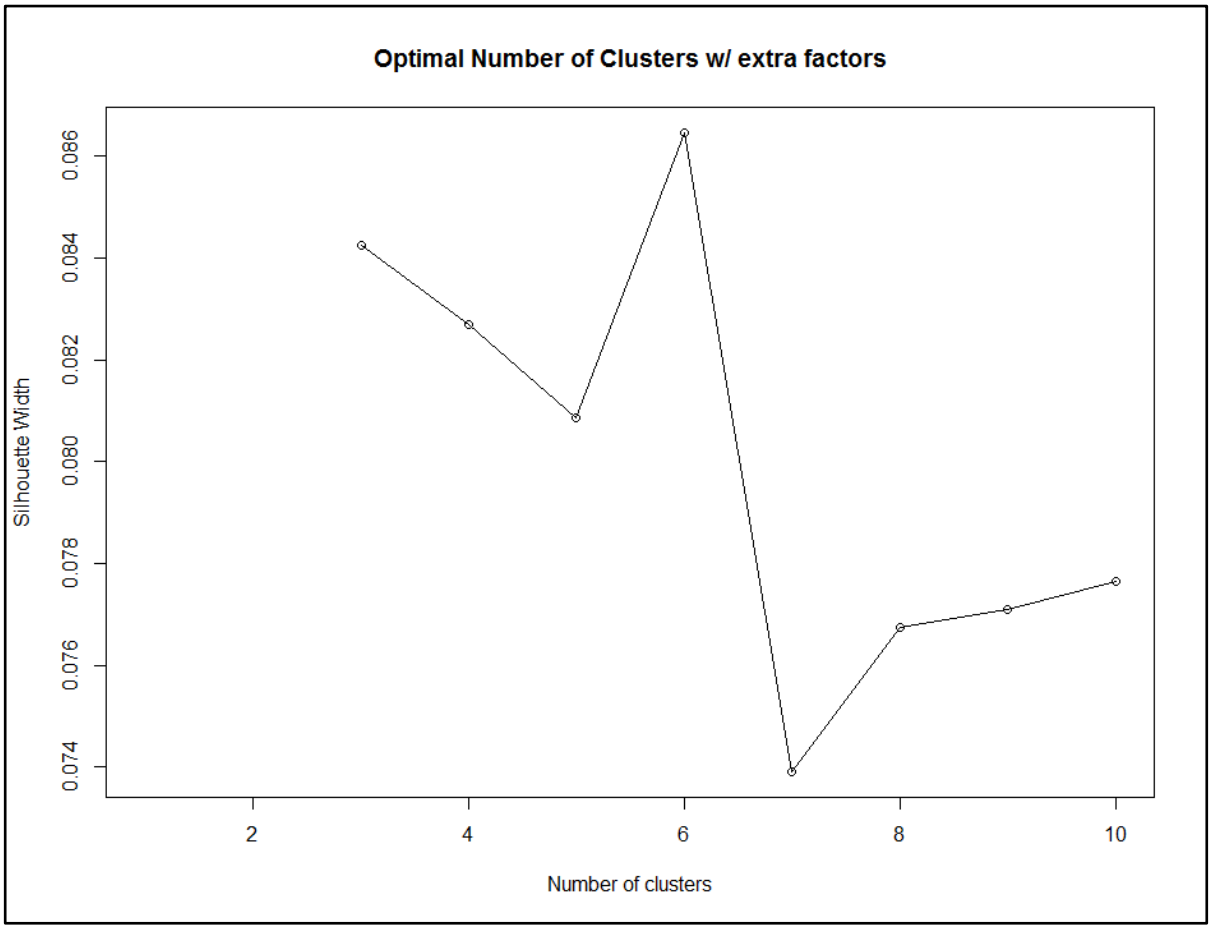
Informatics 172: Mining ER Data
Analysis and Results
Mar 24, 2017
Group LAX

I. Results of Data Analysis

Our initial dataset consisted of observations from nearly 11,000 frequent flyer visits to UCI ED. Our aim was to classify the patients who made these visits into distinct groups, which could then be the beneficiaries of targeted interventions based on their characteristics. To this end, we condensed the observations from the 11,000 visits into a dataset that profiled the 1,783 unique patients who made them.

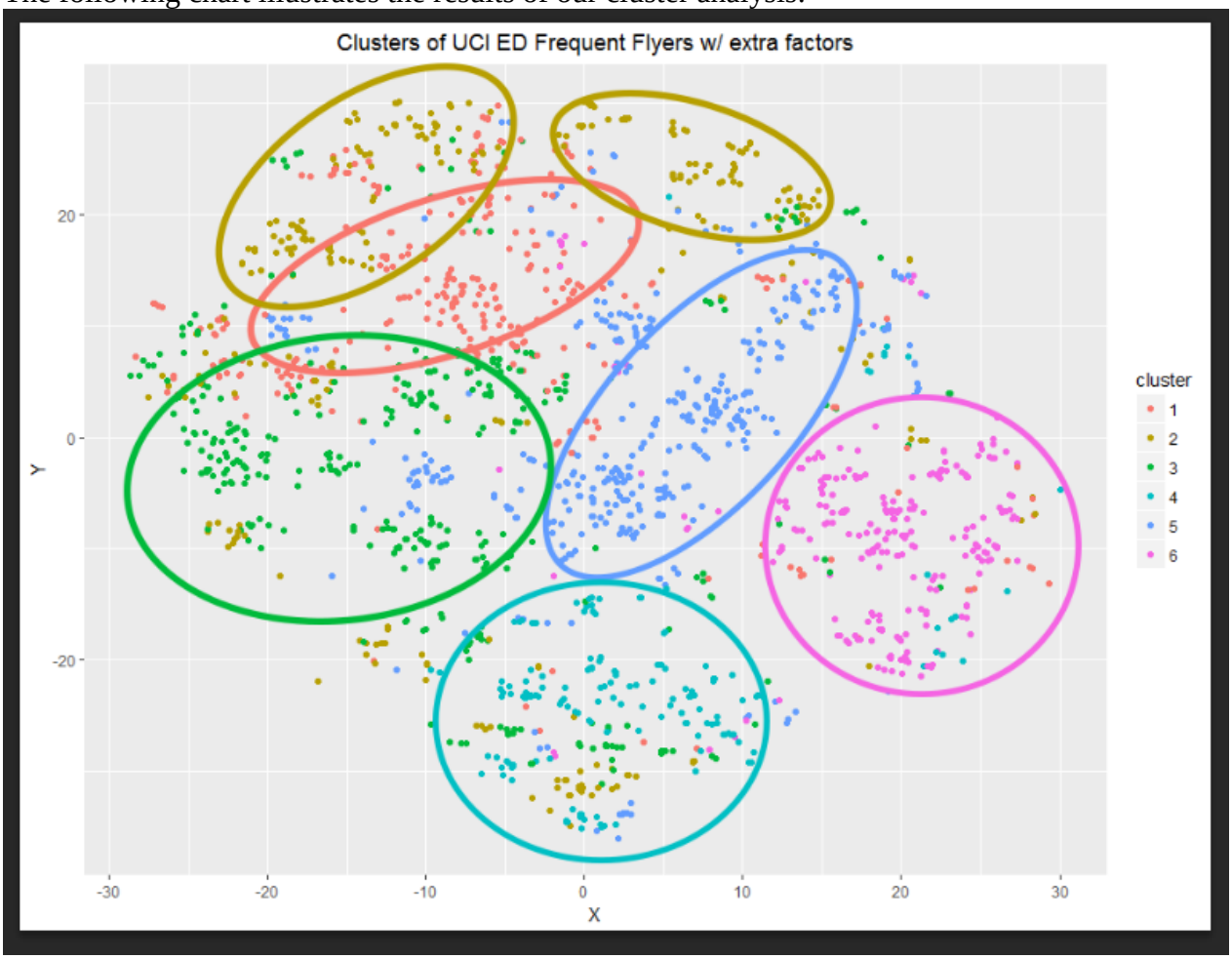
Once we produced the new dataset of the 1783 unique visitors to UCI ED, we generated descriptive statistics for it in order to assure ourselves that the population of UCI ED frequent flyers generally resembled the frequent flyer populations we observed in our literature review. We found that the population of UCI ED frequent flyers is largely unremarkable compared to the other frequent flyer populations. For example, similarly to the frequent flyer population described by SCPHI, over a third of our frequent flyers had mental disorders noted in their past medical history.

After assuring ourselves that the UCI ED frequent flyer population was not unusual compared to the populations we observed in our literature review, we conducted a cluster analysis to determine the major groups of frequent flyers. Our cluster analysis was accomplished by implementing t-distributed stochastic neighborhood embedding (t-SNE) in R. We used a custom Gower distance function to determine the optimal number of clusters in which to partition our data.



Silhouette value is a measure of how similar any given data point in a cluster is to the data points around it. Higher values mean that the data in the cluster is more similar; lower values mean it is more different. Our distance function indicates that the highest silhouette width is at six clusters. Therefore, our cluster analysis divided the UCI ED frequent flyers into six groups.

The following chart illustrates the results of our cluster analysis:



Cluster 1: Hispanic Smokers and Drinkers

Cluster 1 was 70% Hispanic, 80% male, and 90% English-speaking. The average age of this cluster was 40 years old. 80% were unemployed. 60% were binge drinkers and 75% were regular smokers. Members of this cluster used the ED an average of 6.5 times.

Cluster 2: White Smokers

Cluster 2 was 75% white, 62% male, and 94% English-speaking. The average age of this cluster was 50 years old. 78% were unemployed. Only 37% were binge drinkers, but 63% were regular smokers. Members of this cluster used the ED an average of 7 times.

Cluster 3: Substance-free Whites

Cluster 3 was 61% white, 77% male, and 94% English-speaking. The average age of this cluster was 45 years old. 77% were unemployed. However, 66% never binge drink and 93% never use tobacco. Members of this cluster use the ED an average of 5.8 times.

Cluster 4: Retirees with Cancer

Cluster 4 was 55% white, 74% female, 80% English-speaking. The average age of this cluster was 73 years old. 76% were retired. Despite their age, this cluster was fairly

healthy. 82% did not have diabetes, only 40% were prehypertensive, 80% never binge drink, and 67% never use Tobacco, Uniquely, however, 76% of this cluster has cancer. Members of this cluster visited the ED an average of 5.3 times.

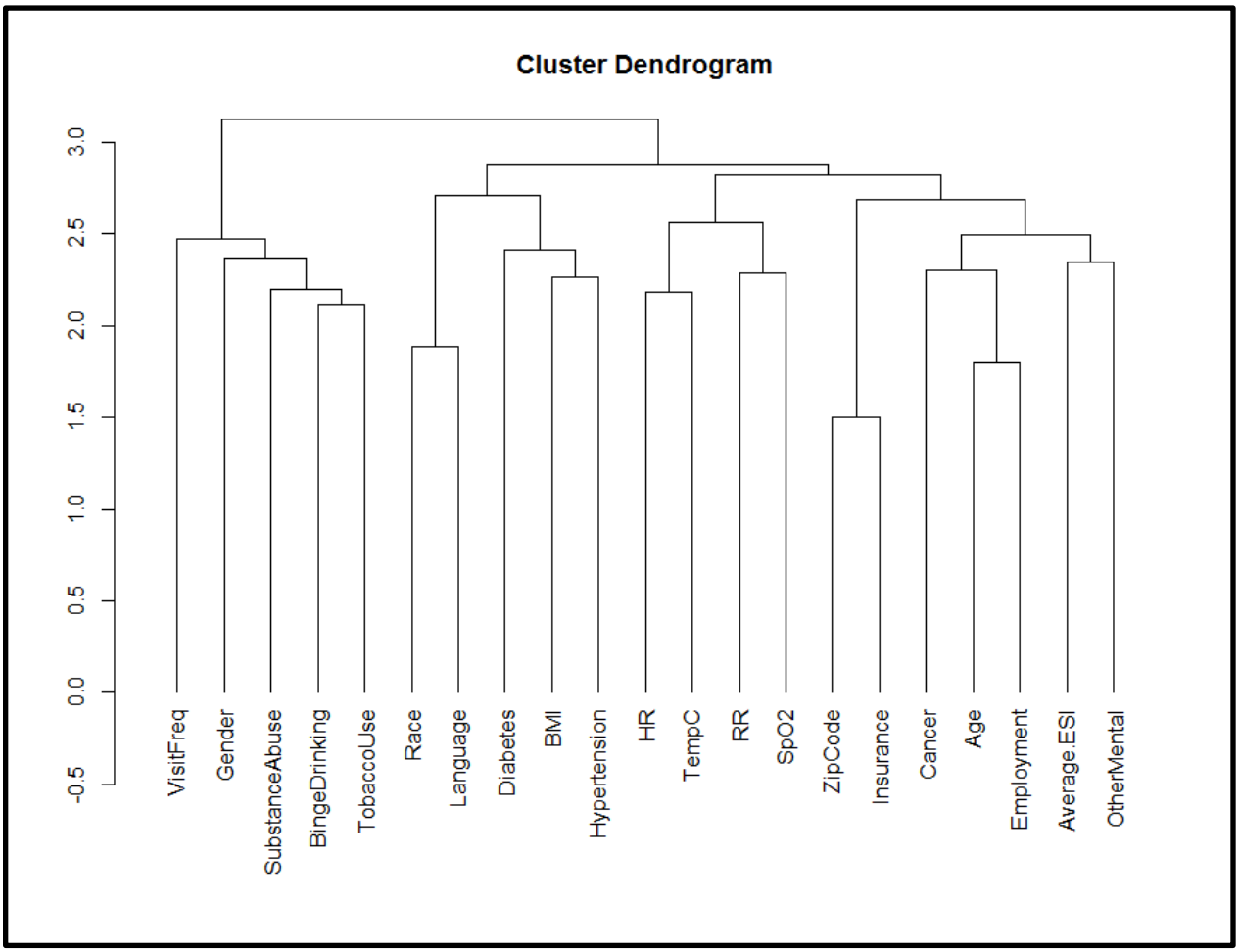
Cluster 5: Young Hispanic Women

Cluster 5 was 67% Hispanic, 86% female, and 98% English-speaking. The average age of this cluster was 37 years old. 62% never binge drink, and 81% didn't use tobacco. 43% were unemployed. Members of this cluster used the ED an average of 5.8 times.

Cluster 6: Older Hispanic Women

Cluster 6 was 95% Hispanic, 73% female, and 90% Spanish-speaking. The average age of this cluster was 54 years old. 80% never binge drink, 86% never used tobacco, and 97% no substance abuse. 60% were unemployed. Members of this cluster used the ED an average of 5.4 times.

We also produced a cluster dendrogram to illustrate which variables are most correlated:



The most dispositive factors in our frequent flyer analysis (as was prominently evident in our clusters) were binge drinking, tobacco use, and substance abuse. These factors were strongly correlated with gender, and together they were strongly correlated with how often frequent flyers visited the ED. Most of the frequent flyer population did not have cancer, but elderly frequent flyers were much more likely to have cancer.

II. Lessons Learned

One of the main lessons we learned was that ER doctors have a very demanding schedule and as a result, they do not always respond immediately or at all. If another team were to do the same project again, we would advise them to get in contact with the doctors as soon as possible and ask for the specific data they want as soon as they can because of how long it takes for the hopsital IT to de-identify and gather all the data. Looking back, it would have been beneficial to find a way to stay in contact with the doctors and hospital IT staff more often. Not all of our cells were straightforward and easy to analyze. Some, like the past medical history cells, contained a

large amount of information that was difficult to subdivide and analyze. Had we found a way to remain in contact with the doctors and the hospital IT staff, we might have been able to ask them the best way to break down this data.

III. Limitations and Future Plans

Given the time frame of this class, our project faced some limitations. For example, we were not able to meet with our clients until week 5 and did not receive our data to work with until week 7. If the data had been obtained earlier, we might have been able to conduct a more thorough analysis. Moreover, we might have been able to outline or begin implementing a targeted intervention for UCI ED. Our future plans for this project are to continue to refine our analysis and to develop and test targeted group-by-group solutions based on the clusters of frequent flyers we find. These solutions include. but are not limited to, implementing educational or outreach programs to help frequent flyers assess the ED alternatives they have available, or partnering with county health officials to reduce the amount of substance abuse in the neighborhoods around UCI ED.