COGS 118B FINAL PROJECT

TEJ QU NAIR, PRAMIT MAZUMDER, ISHAAN H. KAVOORI, CONNIE CHANG, AND ALAN CAO

1. Introduction and Motivation

Money is one of the center aspects of consumerism and is heavily valued by buyers and sellers. Product owners want to sell their product at a maximum profit, whereas consumers would want to buy products at the minimum cost amount. There is a usually a compromising line between how much a customer is willing to spend and how much a product will sell for in order to make profit. Based on consumer spending habits, we are able to predict how much a customer is willing to spend and price products accordingly. We are motivated to do this because in the worst case scenario, a transaction might not be optimal due to it being too profitable or not profitable at all. Products should be priced optimally and it is better to have a customer buy a product and make very little profit than to never buy it at all. Given the gender, age, annual income, and individual spending scores, we are motivated to predict what are the best kinds of products to sell customers. This would be helpful in removing certain items that are not profitable and are undesirable. These predictions can help evaluate what to replace in order to maximize money spending and profit.

2. Related Work

2.1. **Variety Shopping Centers.** Shopping centers like malls are not all built with the same consumer base in mind. Some malls are built mainly to focus on luxury goods while other malls focus on discount sales and cheaper material. Based on the variety of different types of shopping, consumers change what they value in products and may have different lifestyles and spending habits around what is available to them (Jeffrey J. Stoltman, James W. Gentry, and Kenneth A. Anglin, 1991)

2.2. **Consumer Confidence.** A person's annual income dictates how much money they can spend and how much they are willing to spend. This income can fluctuate due to a variety of factors such as inflation, raises, or unemployment. For example, during times of economic hardships such as unemployment cause consumers to have less confidence in spending which can be used to predict transactions in the near future (Ludvigson, Sydney C., 2004). Raw annual income can be used to predict generally how much a consumer is willing to spend, but there are external factors that can cause the confidence of a consumer to fluctuate.

3. Methods

3.1. **Data Preparation.** We read our data from the csv file given by this Kaggle dataset. We noticed that some of the data columns were categorical data rather than numerical data. Determining what we could do with the categorical data, we thought the two possibilities were to arbitrarily assign the categorical data with numerical values or to just drop the categorical data from the dataset entirely. We decided to go with the former assigned values to the categorical data.

3.2. **PCA.** PCA was utilized in order to reduce the dimensionality of the dataset, and to increase interpretability. We did a standard Principal Component Analysis of the data, finding the sorted eigenvalues and eigenvectors of the data and graphing the scree plot. The scree plot then informed us on the significant principal components with which we used to graph the data in the proper principal component coordinates. This dimensionally reduced data is what we used for our following clustering algorithms.

3.3. **K-Means.** K-means clustering was performed on the dimensionally reduced dataset utilizing values of $K \in \{1, \ldots, 10\}$ to test a range of cluster numbers. Initial μ's were randomized and K-means was run until cluster centers converged.

Date: March 2022.

3.4. **Gaussian Mixture Model.** We utilized a Gaussian Mixture model, trained using Expectation Maximization (EM). Similar to K-Means, we ran clustering, and re-projected the clustering using the first two principal components for ease of interpretability. The model was also run for values of $K \in \{1, \ldots, 10\}$, and the initial cluster centers were obtained from the centroids in the earlier K-Means runs for the same K. The algorithm was run for 100 iterations.

3.5. **Error Analysis.** Optimal values of K for both methods were chosen using the "elbow method" with the Sum of Squared Errors (SSE) for each clustering. This was calculated by taking the squared distance between every point in a cluster and the centroid that it was assigned to.

4. Results

4.1. **PCA.** After running PCA on our dataset, we generated a scree plot for the explained variance of the principal components. For the purposes of graphing, we only utilized the first two components. However,

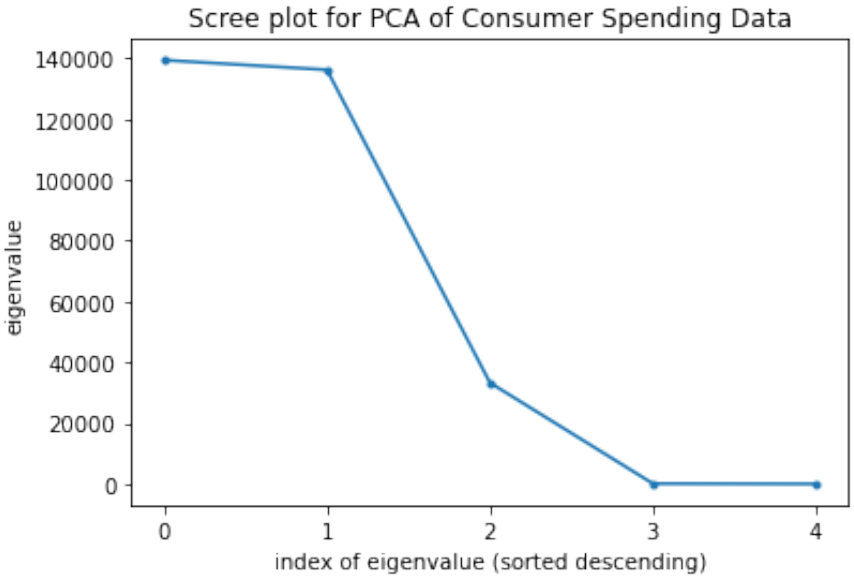

Figure 1. PCA Scree Plot

looking at Figure 1 and using the traditional "elbow" method, we can see that the first four components explain most of the variance in the dataset. Figure 2 below shows our the first two principal components of our transformed data.

4.2. **K-Means.** K-Means formed reasonable looking clusters. Figure 3 below shows that error strictly decreases as cluster count increases. This is as expected. The lack of smoothness in the decrease may be attributed to the fact that K-means is initialized with random cluster centers every run, so the error from one value of K to another is not directly comparable. One fix for this would be setting the initial centroids constant as K increases to have a consistent baseline.

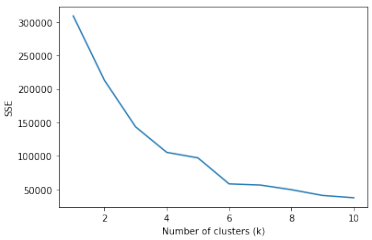

Figure 3. K-Means SSE vs Cluster Count

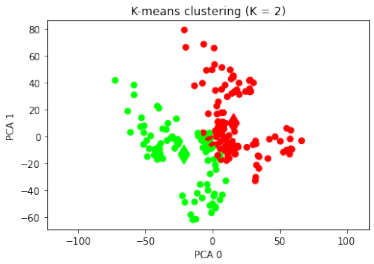

Figure 4. GMM SSE vs Cluster Count

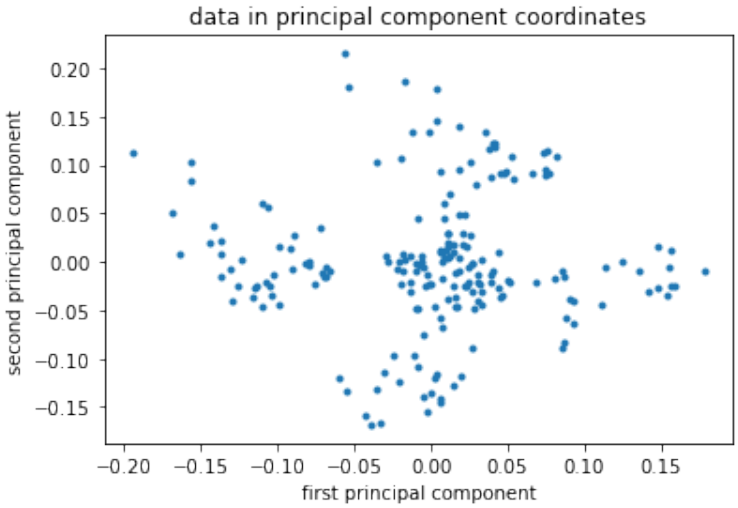

FIGURE 2. PCA Scatter Plot

4.3. **EDA.** One thing we attempted but did not end up performing was Exploratory Data Analysis (EDA). We wanted to see if we could display our data in various ways and notice patterns that we hadn't seen before. However, we had trouble getting our code for it to work and given the remaining time to complete the project and the fact that we didn't need to do it, we decided to abandon this analysis.

4.4. **GMM.** We visually find the optimal value of $K = 6$ through the SSE graph. Figure 6 shows the final clustering obtained for this value, and it looks very reasonable. This indicates that there are optimally 6 different demographics of consumers in our spending database, as transformed by the PCA axes.

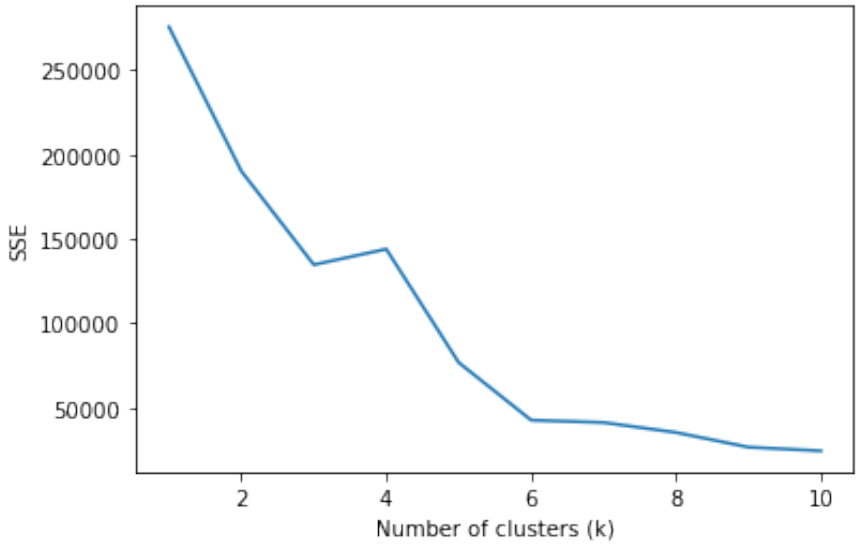

FIGURE 5. GMM Clustering for K=6

One concern is that our amount of error doesn't strictly decrease as our number of clusters increase. This may be due to an error in the way that we calculate error. Since the model performs soft clustering, calculating error solely based off the most responsible cluster for a point may not be the most accurate.

5. DISCUSSION

We learned how to use PCA to reduce the dimensionality of the data set and more easily visualize high dimensional data. We were able to cluster the data into up to 10 clusters to find common behavioral patterns in consumers

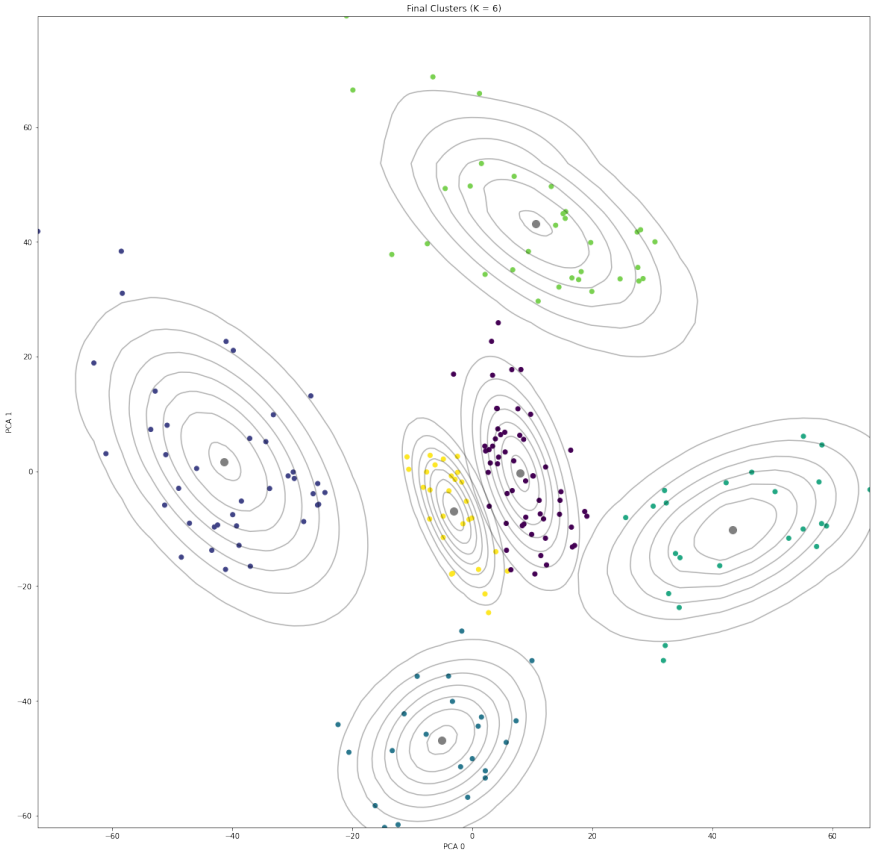

FIGURE 6. GMM Clustering for K=6

One thing we could have done better for this experiment is gathering more info about other possible factors that can contribute to consumer spending. Our current data set is limited to four columns, but increasing the amount of factors could form better predictions. For example, adding the percent inflation that year or another individual variable like consumer confidence is one way to figure out whether someone will buy or sell a product. Another thing we could have improved on is using a data set that had more data points. This would allow us to create less of a generalized prediction. Also, by utilizing PCA for clustering, we reduced the interpretability of our data. Since our clusters are in terms of principle coordinate axes instead of the original variables, it is more difficult to differentiate the factors that make the clusters themselves. Some possible next steps would be to see how the different variables contribute to each cluster by doing pairwise plots or other analysis. For purposes of data visualization, we only utilized the first two PCA components, however as seen in Figure 1, the third component explains a significant amount of variance. A possible improvement we could have made would be to keep the third component, and visualized it as either another dimension, or as color. Another thing worth mentioning is that initially, our project was going to use a dataset about heart failure prediction and we were going to do clustering based on medical attributes. However, the data wasn't the best for unsupervised clustering as it became obvious that the heart failure dataset would be better applied to supervised learning algorithms. So we went in a different direction for the project.

6. CONTRIBUTIONS

Tej - Data preprocessing, PCA code
Ishaan - GMM and K-means code, Slides
Alan - Introduction, Related Work Research, Discussion, Slides
Connie - Methods, Results, Slides
Pramit - GMM code, Methods, Error calculation and charts
everyone - help with debugging, polishing paper and slides

7. Code

The GitHub repo used to obtain the results in this paper can be found here

References

[1] Jeffrey J. Stoltman, James W. Gentry, and Kenneth A. Anglin (1991) ,"Shopping Choices: the Case of Mall Choice", in NA - Advances in Consumer Research Volume 18, eds. Rebecca H. Holman and Michael R. Solomon, Provo, UT : Association for Consumer Research, Pages: 434-440.

[2] Ludvigson, Sydney C. (2004), "Consumer Confidence and Consumer Spending", Journal of Economic Perspectives Volume 18 (2), Pages: 29-50, 10.1257/0895330041371222.