DNSC 6279
Spring 2019
Assignment 4

In this assignment we will use the training data from assignment 3 to build a reasonable number of customer clusters. We will use principal components analysis (PCA) to explore the data and confirm our results. We will use k-means clustering to make our clusters.

You may find the following notebook helpful in addition to resources we discussed in class: https://github.com/h2oai/h2o-tutorials/blob/master/training/h2o_algos/src/py/clustering_and_pca.ipynb

Removing outliers:

Since we have been working on our data for a few months now, it is reasonably clean. Also, we will standardize the data internally for each PCA and k-means run. So, the main data quality issue we are concerned about is outliers, as outliers dramatically affect the squared error functions used in both k-means and PCA. To identify outliers, project all the numeric inputs onto the first 3 principal components. Be sure to use a seed of 12345 and to standardize the training data. You should be able to create a plot similar to Figure 1 below, where there are 2 obvious outliers on the PC1 axis. The input variables to use for all calculations are: offervalue, avg_category_quantity, avg_category_amount, exact_item_bought, month, week, and day.

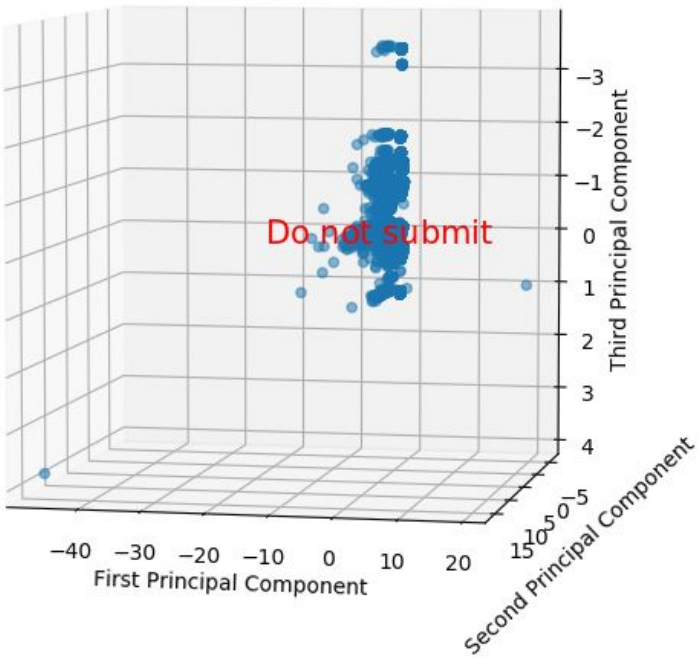

Figure 1: The first three principal components of the training data *with* outliers.

DNSC 6279
Spring 2019
Assignment 4

Remove any large outliers. For me this was two observations where:

- PC1 > 10
- PC1 < -40

Confirm that the outliers have been removed by retraining and re-plotting the first 3 principal components of the filtered training data. You should be able to create a plot like figure 2 below.

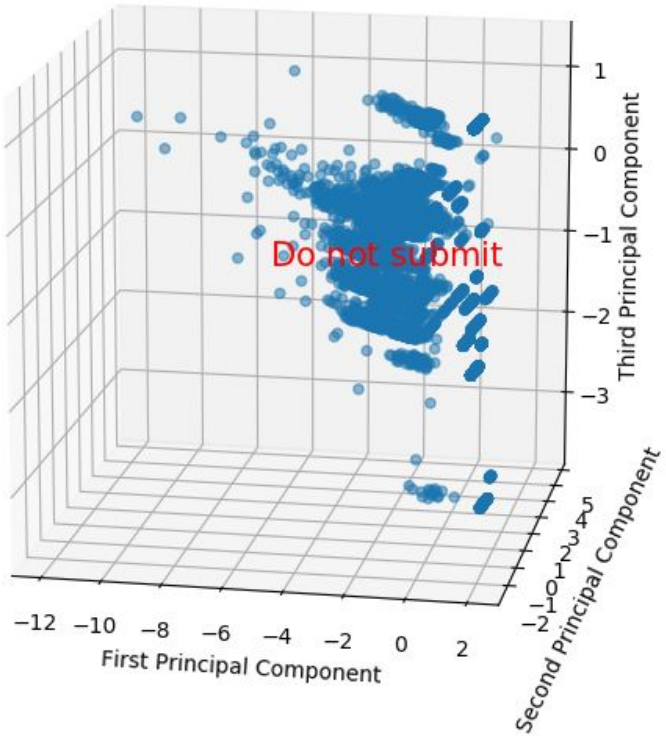

Figure 2: The first three principal components of the training data *without* outliers.

Determine the number of clusters:

Now that the outliers have been removed, run k-means clusters for k=1 to k=10. Plot the total within-cluster error vs. the number of clusters. Look for a good “elbow” to determine the number of clusters. You should be able to generate a plot like Figure 3. Pick a number of clusters where the curve flattens out, or if you get really lucky, where it dips. For me this is 5 clusters.

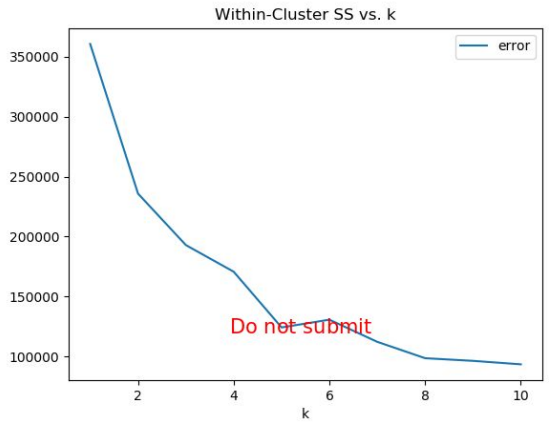

Figure 3: The number of clusters, k, vs. the total within-cluster error.

We will confirm this number of clusters by plotting them in the reduced dimensional space in the next step of the assignment.

Confirm the number of clusters:

Merge your cluster labels onto the 3 PCA projection *without* outliers and plot the projection colored by cluster labels. The clusters should appear reasonably neat and separate in the projection as in Figure 4.

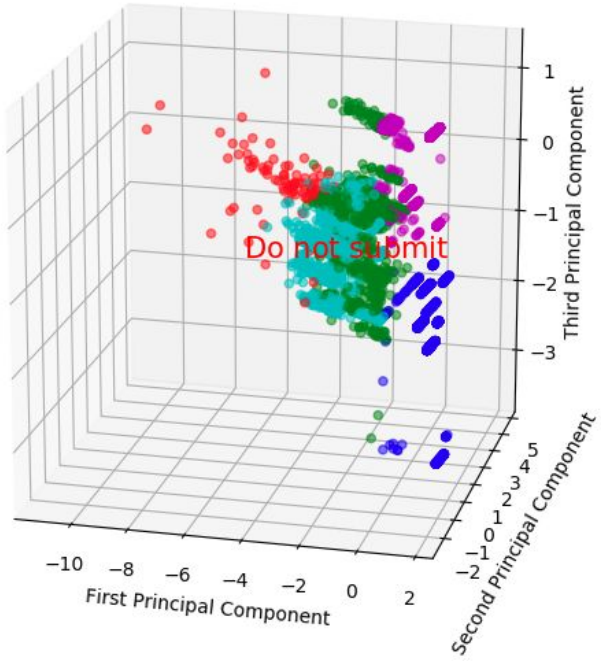

Figure 4: The first three principal components of the training data *without* outliers colored by cluster label.

DNSC 6279
Spring 2019
Assignment 4

Profile clusters:

Profile your clusters by examining the mean of each original input variable in each cluster. The combination of means should be somewhat different for each cluster, as in Figure 5. Come up with a brief, one sentence description for each of your clusters.

predict	mean_offervalue	mean_avg_category_quantity	mean_avg_category_amount	mean_exact_item_bought	mean_month	mean_week	mean_day
0	1.98868	0.0121355	0.0299874	0.0124084	0.722154	6.40432	22.5429
1	1.27502	1.31836	4.62313	0.996452	0.646387	6.39645	24.2474
2	1.12364	3.33443	10.3454	0.991833	0.853902	7.19964	23.1733
3	1.09815	0.939881	2.91457	0.719917	0.847232	4.5667	5.36753
4	0.787661	0.21914	0.543999	0.184691	0.10527	4.25917	25.8372

Figure 5: Centroids for k-means clusters in the training data.

Submission:

In a document, not a notebook, please turn in the following:

- Your plot of the first 3 principal components with outliers (1 pt.)
- Your plot of the first 3 principal components without outliers (1 pt.)
- Your plot of k vs. total in-cluster error (2 pts.)
- Your plot of the first 3 principal components without outliers colored by cluster label (2 pts.)
- Your cluster centroids and with brief profiles (2 pts.)

Bonus: Describe the expected effect of forming 1 to 10 clusters in a reference distribution and subtracting the within-cluster error in the reference distribution from the actual within-cluster error of the corresponding number of clusters in the training data. (1 pt.)

To receive full credit for this assignment, turn in your group's results and your R or Python script/notebook (2 pts.) in one zipped file to Blackboard.