Community Segmentation using K-Means

David N Campbell
The University of Texas at
Austin
Austin, TX
campbeda@utexas.edu

Jeffrey D Nelson
The University of Texas at
Austin
Austin, TX
jdn2283@utexas.edu

***Abstract*—Deriving groups from data and exploring feature relationships in an unsupervised learning environment can be accomplished using the K-Means algorithm. This paper explains the K-Means algorithm and explores its use on three datasets. The datasets are explained through a hands-on implementation that focuses on the process and the lessons learned.**

***Keywords*—k-means, clustering, machine learning, segmentation**

I. Introduction

The K-Means algorithm is an unsupervised learning algorithm used to partition a dataset into *k* clusters. The partitioning is done by calculating *k* cluster centers that minimize the average distance between a data point and its nearest cluster center. Distance is defined as the Euclidean distance between two points, therefore K-Means can only operate on numeric features. Since K-Means derives groups from unlabeled data, it is often used to define market segments, identify anomalies, or extract relationships from complex datasets. In this paper, three datasets are explored. A high-level description of the K-Means algorithm and lessons learned from its use are also presented.

II. Background

At a high-level, the K-Means algorithm can be divided into three steps:

1. Initializing cluster centers
2. Determining cluster membership
3. Updating cluster centers

Initialization of cluster centers can be done at random or by using a heuristic-based initializer such as *k-means++*. The implementation of the *k-means++* initializer is outside the scope of this paper, so a reference is included below [1]. Note that determining the ideal clustering is an NP-hard problem. The optimality of the derived clusters, as well as the convergence time, is sensitive to cluster initialization.

Cluster membership is calculated for each point by finding the cluster with the minimum Euclidean distance. Each cluster center is then moved to the geometric mean of its membership set. The algorithm then returns to step #2 and repeats until all cluster centers do not move or no cluster center moves more than a user-defined threshold.

The above algorithm runs for a fixed number of clusters, *k*. Determining the ideal value for *k* requires running with multiple values and determining which has the lowest silhouette score. The silhouette score measures the average distance every point is from its cluster center and from nearby clusters. The value ranges from 0 to 1, where a higher value means a better clustering. When a dataset has a large number of features, K-Means can become computationally infeasible. To speed up convergence, one can derive the principal features from the dataset using Principal Component Analysis (PCA) and run K-Means on those principal features. A general rule of thumb is that the principal components should capture at least 80% of the variance in the data set.

III. DataSets

A. Mall Customer Segmentation

The goal for the first dataset was to identify different groups of Mall customers in order to develop a marketing strategy geared towards each specific group. There were only four features in this data set, so PCA was not necessary. The reference Jupyter Notebook[2] visualizes these feature relationships and shows that the strongest relationship exists between "Annual Income" and "Spending Score".

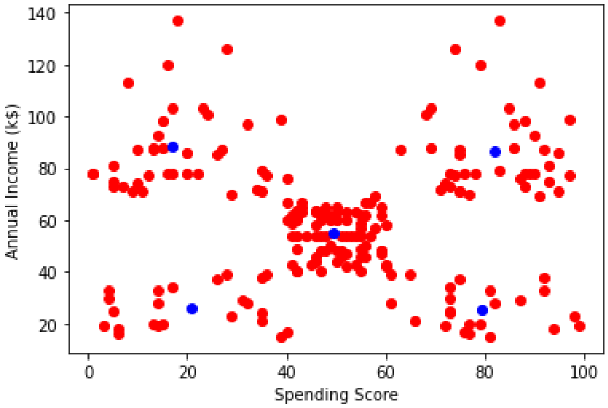


Figure 1 - Clusters for Spending Score and Annual Income

From the figure above, five clusters are visible, and evaluating silhouette scores confirmed that five is the ideal value for *k*. Including additional features in the model, such as "Gender" and/or “Age”, did not yield a better or more interpretable clustering, so we decided to keep the model based only on “Annual Income” and “Spending Score”. Our recommendation to the marketing department is as follows:

- To those with high annual income and high spending score, market aggressively, as they have the money to spend and the willingness to spend it.
- To those with low annual income and low spending score, marketing is likely not worth the effort.
- To those with median annual income and median spending score, we need a more targeted marketing strategy. This cluster represents the largest number of people, so the reward is high if one determines which products will result in this group spending more.

B. Country Profiles

The country data set includes features related to the health and wealth of nations around the world. The goal for this data set is to determine which nations are most in need of health and financial aid.

1) Data Preprocessing

Before training any K-Means model, this data set required preprocessing. Cluster centers are calculated based on mean Euclidean distance, so outliers and large numeric values can have an outsized impact. We decided to normalize all features using *StandardScaler* and then construct a heat map to visualize feature correlations. Most relationships are intuitive, such as income and GDP being positively correlated with life expectancy and negatively correlated with child mortality rate.

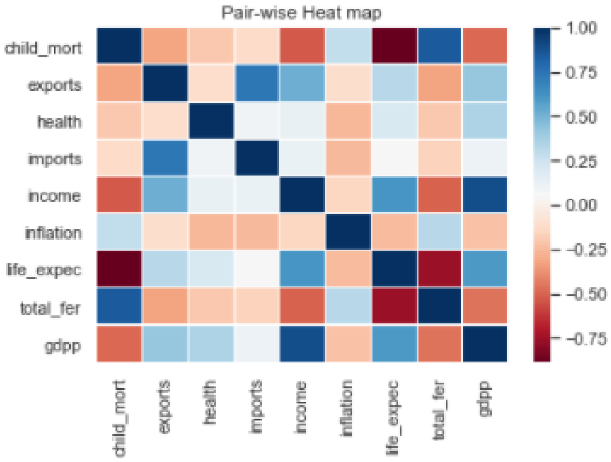


Figure 2 - Pairwise Heat Map of Country Profiles Dataset

2) PCA

To decrease the complexity of K-Means processing, we decided to reduce the dimensionality of the data with PCA. From the figure below, one can see that greater than 80% of the variance can be captured by the first three principal components. The table to the right shows the contribution of each original feature to the principal components.

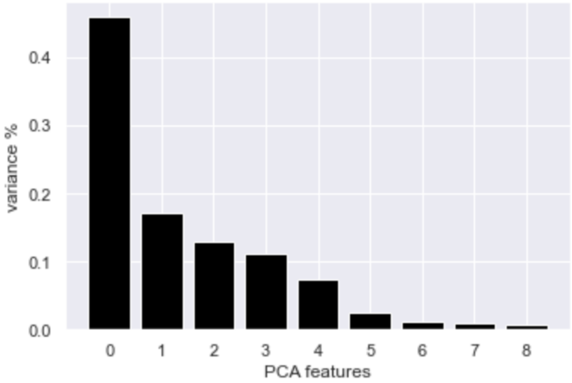
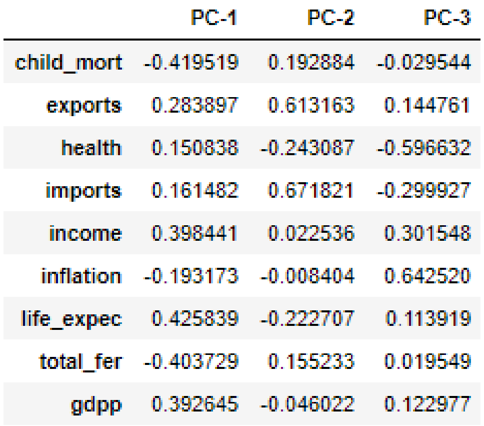


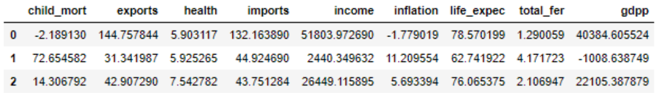
Figure 3 - Principle Component Variance



	PC-1	PC-2	PC-3
child_mort	-0.419519	0.192884	-0.029544
exports	0.283897	0.613163	0.144761
health	0.150838	-0.243087	-0.596632
imports	0.161482	0.671821	-0.299927
income	0.398441	0.022536	0.301548
inflation	-0.193173	-0.008404	0.642520
life_expec	0.425839	-0.222707	0.113919
total_fer	-0.403729	0.155233	0.019549
gdpp	0.392645	-0.046022	0.122977

Figure 4 - Largest 3 Principal Components

Using silhouette scores to determine the optimal *k* value yielded three, so we trained a K-Means model using three clusters and identified the following cluster centers:



	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
0	-2.189130	144.757844	5.903117	132.163890	51803.972690	-1.779019	78.570199	1.290059	40384.605524
1	72.654582	31.341987	5.925265	44.924690	2440.349632	11.209554	62.741922	4.171723	-1008.638749
2	14.306792	42.907290	7.542782	43.751284	26449.115895	5.693394	76.065375	2.106947	22105.387879

Figure 5 - Cluster Centers of Country Profile Dataset

The groups can be interpreted as follows:

1. *The ultra-wealthy countries*: the first cluster contained a small number of countries (about 5%) that have a very high income, GDP and life expectancy.
2. *The generally poor countries*: the second cluster contained about 40% of the countries, which are characterized by high child mortality rate, low income, and low GDP.
3. *The generally wealthy countries*: the third cluster contained about 55% of the countries, which are characterized by low child mortality rate, generally high income and high life expectancy.

While *k*=3 yielded the highest silhouette score, a charity operating on this data likely needs to work with a smaller number of countries than the second cluster provides. Our options included increasing the value of *k* or training a model with only the countries in the second cluster. We decided to increase the value of *k* to 10, which yielded two clusters that better identified the poorest of these countries. The countries present in those two clusters are visible in the reference Jupyter notebook [2].

C. Credit Card Clustering

The goal for the final data set was to segment credit card customers into different groups and develop marketing strategies for each of those groups.

1) Data Pre-Processing

Before training any models, this data set also required preprocessing. First, missing feature values were replaced with the feature median. Next, the data needed to be appropriately scaled. This dataset contained a large number of features with high skew, which the figure below shows for the *CASH_ADVANCE* feature:

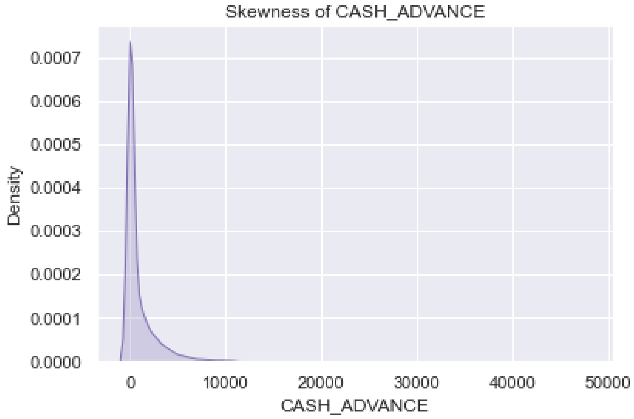


Figure 6 - Skewness of the Cache Advance Feature

We eventually decided to log transform features with high skew and apply *StandardScaler* to the rest. This led to improved convergence times and higher average silhouette scores for K-Means. We also visualized the feature correlations using a heat map and derived the principal components using PCA. The heat map can be found in the reference Jupyter notebook [2], while the PCA results are captured below:

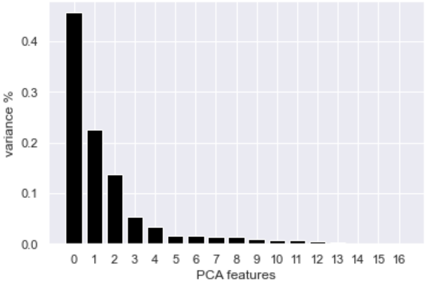
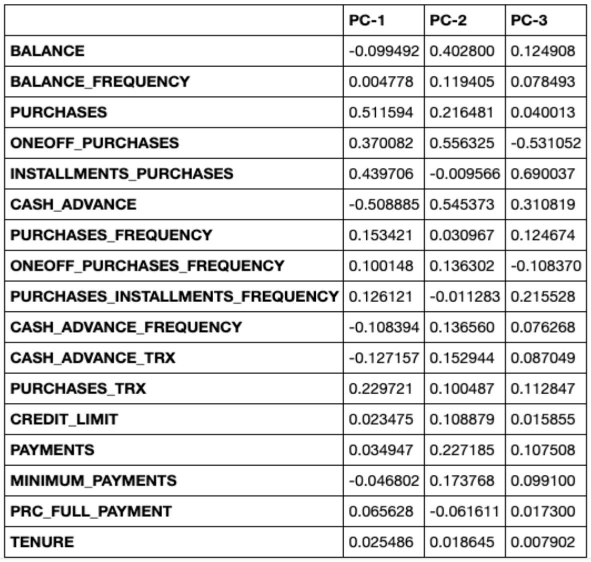


Figure 7 - Principle Component Variance



	PC-1	PC-2	PC-3
BALANCE	-0.099492	0.402800	0.124908
BALANCE_FREQUENCY	0.004778	0.119405	0.078493
PURCHASES	0.511594	0.216481	0.040013
ONEOFF_PURCHASES	0.370082	0.556325	-0.531052
INSTALLMENTS_PURCHASES	0.439706	-0.009566	0.690037
CASH_ADVANCE	-0.508885	0.545373	0.310819
PURCHASES_FREQUENCY	0.153421	0.030967	0.124674
ONEOFF_PURCHASES_FREQUENCY	0.100148	0.136302	-0.108370
PURCHASES_INSTALLMENTS_FREQUENCY	0.126121	-0.011283	0.215528
CASH_ADVANCE_FREQUENCY	-0.108394	0.136560	0.076268
CASH_ADVANCE_TRX	-0.127157	0.152944	0.087049
PURCHASES_TRX	0.229721	0.100487	0.112847
CREDIT_LIMIT	0.023475	0.108879	0.015855
PAYMENTS	0.034947	0.227185	0.107508
MINIMUM_PAYMENTS	-0.046802	0.173768	0.099100
PRC_FULL_PAYMENT	0.065628	-0.061611	0.017300
TENURE	0.025486	0.018645	0.007902

Figure 8 - Largest 3 Principle Components

2) K-Means

The optimal *k* value determined using silhouette scores was 7, however this is a large number of groups for a marketing department to manage, so we decided to use *k*=5, which still had a silhouette score greater than 0.5. Training a K-Means model on the principal components with *k*=5 yielded the cluster centers present in the reference Jupyter notebook[4] and allowed identified the following groups:

- those with high balance that make the median level of purchases
- those with median balance that make next to zero purchases
- those with very low balance that make few purchases and have a low credit limit
- those with below average balance that make a large number of purchases
- those with below average balance that make less than average purchases

Our recommendation to a marketing department is to focus on advertising to the first and fourth clusters, as they purchase the most frequently. Those in the first cluster represent the largest group and have the highest balance, so they deserve attention as well.

IV. Related Work

The three datasets listed above were all part of Kaggle competitions[5]. The approaches that we took for each data set were similar to other Kaggle notebooks at a high-level. In general, employing K-Means leads to a pretty standardized set of steps. Since this is an unsupervised learning task, the differentiation comes from the analysis, which is why we focused on visualization, interpretability, and explaining the cluster representations in real-world terms.

V. Conclusion

The three datasets that we explored show how K-Means can be utilized to solve real-world problems. Overall, we identified a few key takeaways from our experiments:

- **Data preprocessing is critical**: K-Means convergence and cluster location is highly influenced by the scalar value of a feature, so it is important to scale features appropriately.
- **Visualizing helps debuggability**: tools like heat maps and graphs help in understanding the data set, interpreting the results, and debugging the outputs.
- **Use PCA on high-dimensional data**: transforming the data set into a lower dimensional space improves K-Means convergence time and visualization. One must remember to transform results back to the original space before interpreting, though.
- **Use judgment in addition to silhouette scores**: the optimal number of clusters based on silhouette scores may not be optimal for the real-world problem one is trying to solve. The scores should be used as a guide to be used in conjunction with human intuition.

The datasets used in this research can be found on GitHub along with the Jupyter notebooks at https://github.com/jdn5126/CommunitySegmentation. In the future, we would like to explore how different cluster initialization algorithms impact our K-Means results, both in convergence time and silhouette scores.

References

[1] David Arthur and Sergei Vassilvitskii. k-means++: The Advantages of Careful Seeding. In 2007: *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 1027-1035.

[2] Jupyter Notebook for Mall Customer Segmentation: https://github.com/jdn5126/CommunitySegmentation/blob/main/CustomerSegmentation.ipynb

[3] Jupyter Notebook for Country Profiles: https://github.com/jdn5126/CommunitySegmentation/blob/main/CountryProfiles.ipynb

[4] Jupyter Notebook for Credit Card Clustering: https://github.com/jdn5126/CommunitySegmentation/blob/main/CreditCardClustering.ipynb

[5] Mall Customer Segmentation on Kaggle:

[6] https://www.kaggle.com/vjchoudhary7/customer-segmentation-tutorial-in-python

[7] Country Profiles on Kaggle: https://www.kaggle.com/rohan0301/unsupervised-learning-on-country-dat

[8] Credit Card Clustering on Kaggle: https://www.kaggle.com/arjunbhasin2013/ccdata