Analysis and Launch Plan for Retail Startup

Sanjeev Narasimhan

Introduction

This report aims to provide an understanding of the different kinds of markets we may be able to launch a retail location, as well as identify 10 choice markets to move forward with.

Data Files Provided (.xlsx & .csv):

- DMA Data set (PI-18803 DMA.xlsx): demographic dataset at the DMA (Designated Market Area aka major city or metropolitan area)
- MSA Excel files (south, northeast, midwest, west.xlsx): Metropolitan Statistical Areas provide input into consumer expenditure in major categories.
- Sales Data Set (sales data-set.xlsx): historical training data including store numbers, weekly sales, etc.
- Features Data Set (features.xlsx): additional data related to the store, department, and regional activity for specific dates including markdowns, CPI, unemployment, store number, etc.
- Stores Data Set (stores dataset with DMA.xlsx): information about the 45 stores, indicating the type and size of store and location.

Analysis

Data Preprocessing & Setup

1. The first steps taken were to go through the different sources of information (excel sheets, kaggle data card, outside google searches to understand variables)
2. The DMA Dataset was kept as is with the only change being the creation of a variable called HHldsWithVehicles, which was the sum of the households with 1 and 2+ vehicles.
3. A new dataset, Store_Data was created that linked the store data with some of their variables in the DMA dataset
4. Some of the other datasets (features, store sales, and MSA files) also had elements appended to the Store_Data and other elements were used for a "scorecard-based approach".
5. Subsets of the data were made as needed to append to master data and compare trends within stores.

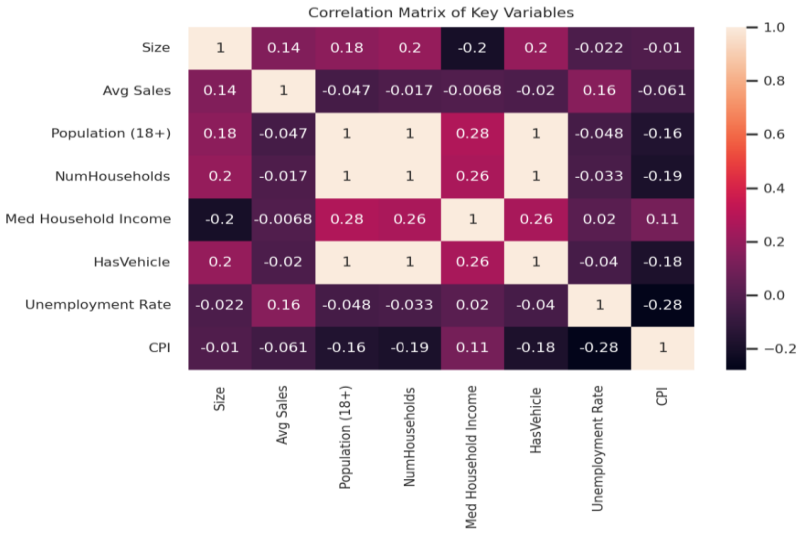
Analytical Plan

- Store_Data
  - Once all the data preprocessing and setup was completed, a correlation matrix was constructed to look at the relationships between some of the key variables within the Stores_Data dataset (Size, Avg Sales, Population (18+), NumHouseholds, Med Household Income, HasVehicle, Unemployment Rate, and CPI).

  - From there a Gaussian Mixture Model was to be used to look for stores that had similar characteristics. This would allow us to gauge the current performance and characteristics of established stores to not only provide a baseline to compare DMAs to, but also see if a new store could capitalize on DMAs. After finding the optimal number of clusters to work with based on the silhouette scores, we would then fit the model and check the metrics.
  - Once this was done the cluster assignments would be appended to the original stores and conduct a within cluster analysis to return summary statistics and categorize the kinds of stores that existed within the markets.
- DMA_Data (Part 1)
  - Once the analysis is done for the store data, the same will be done for the DMA data, determine optimal number of clusters, clustering using a Gaussian Mixture Model, and returning the characteristics of cluster assignments. Since there isn't any store data to compare to, and we don't want to extrapolate the findings of specific stores to explain the trends of stores in each DMA, we could compare the characteristics of those DMAs to those that have existing stores.
- DMA_Data (Part 2)
  - Another consideration for the data is that we wanted the retail store to "compete with Wal-Mart & Target for middle income to high income households", the quartile distribution for median household income was checked and all DMA observations that have a median household income greater than or equal to the median of the data column would be considered.
  - From here, the same process as above would be used and then again return the characteristics of each cluster. With this round of analysis, we are looking to see which markets are to be considered based on the condition that we wanted to compete primarily for middle to high income households.
- Comparisons
  - After the above rounds are completed, the results can then be compared to see which markets are built up of the more desirable characteristics. The current store DMAs can be used as a baseline and see which trends in household income, ownership of vehicles, unemployment rates, CPI, number of households, and demographic attributes are most similar to the current high performing DMAs. The markdowns can also be used to understand if the trends in store performance have to do with time of year, or perhaps because of underperforming/ low volume movement of inventory. The MSAs can also be used again to compare DMAs based on attributes of individuals and their spending/ expense habits.
- Choosing 10 DMAs
  - Finally, after filtering down to the better potential DMAs a scorecard-based approach can be used to make the final determinations of the best 10 DMAs.

Introductory Analysis and Dataset Breakdowns

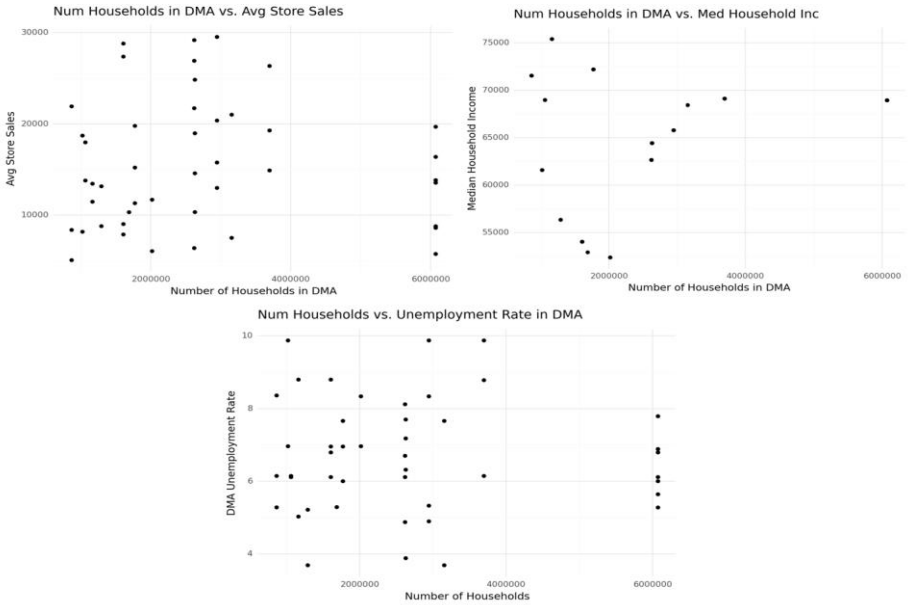
- 211 Potential DMAs
- 45 Existing Stores



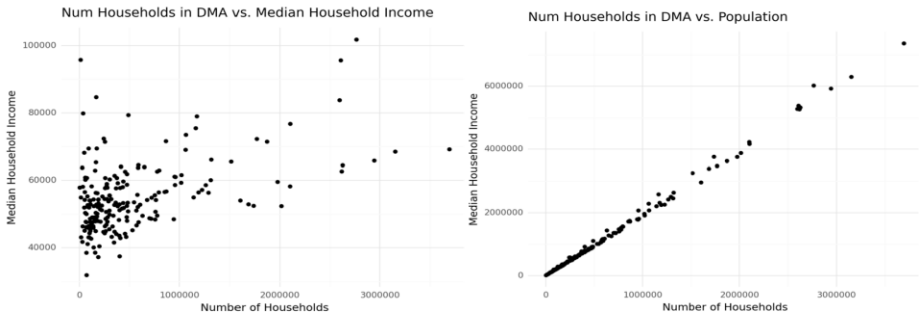
Correlation Matrix of Key Variables in Existing Store Data

The correlation matrix above helps visualize some of the key relationships within the stores data. As expected, variables like Population and NumHouseholds are perfectly correlated, as well as HasVehicle. The matrix also illustrates that most variables that describe the individuals/ households in the market are negatively correlated with CPI, aside from the median household income, which is positively correlated with CPI. Another notable relationship is that the average sales of stores is also positively correlated with the unemployment rate.

This correlation matrix allows for an understanding of how stores tend to perform based on the characteristics of the DMAs in which they exist. This will also aid in determining how the attributes of DMAs may translate to store performance for those in which there are no currently existing stores.



Scatterplots of Store Data Variables



Scatterplots of DMA Data Variables

The scatterplots above will be used to visualize the cluster assignments from the models used below.

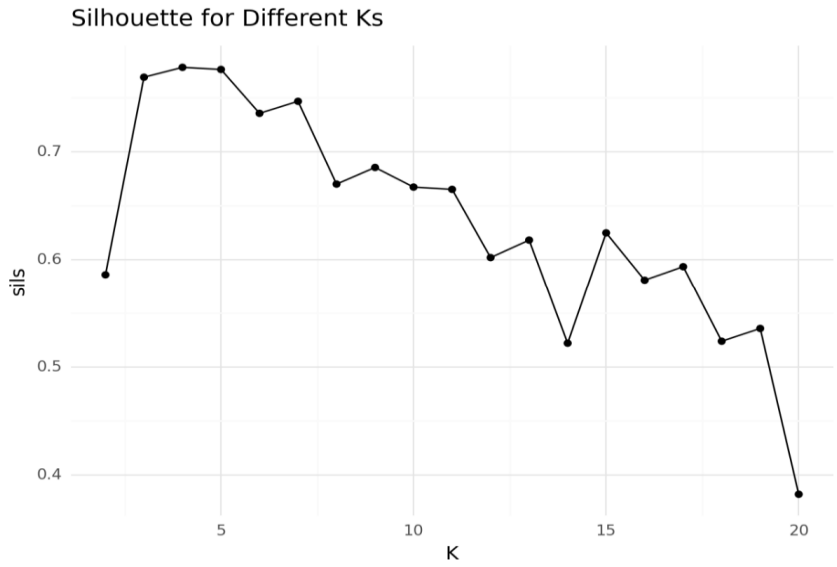
The Store Data plots illustrate the relationship between the number of households with median income, average store sales, and the unemployment rate.

The DMA Data plots illustrate the relationship between the number of households with median income and population.

Methods

Model 1 – Gaussian Mixture Model for Store Data

This model was made to cluster datapoints based on the following variables: Size, Avg Sales, Population (18+), NumHouseholds, Med Household Income, HasVehicle, Unemployment Rate, CPI

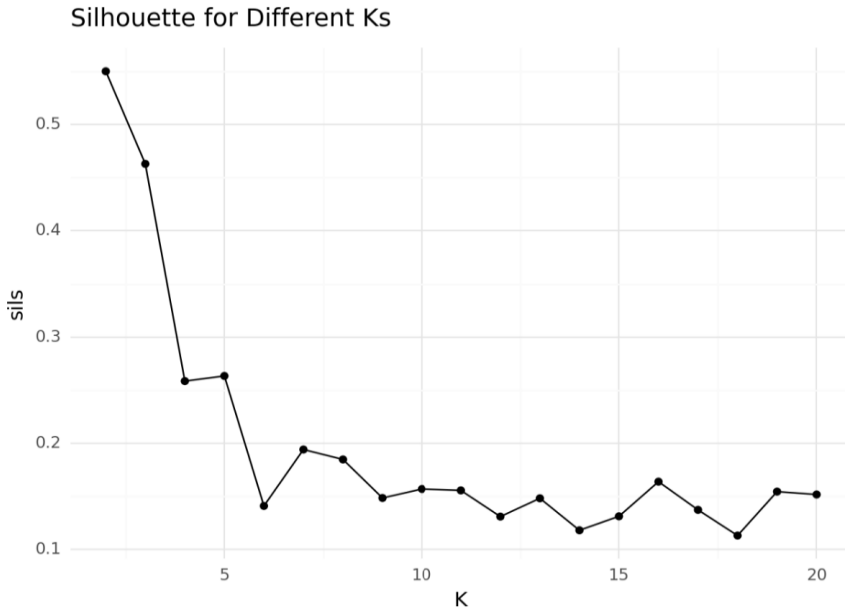


Plotting Silhouette Scores for Different Number of Clusters

After determining the optimal number of clusters to be 4, the model was fit and the final silhouette score of ~0.77813 was returned. While this silhouette score is not exceptional, it shows that the clusters have proper separation and cohesion and ensures that the within cluster analysis will have relatively distinct values.

Model 2 – Gaussian Mixture Model for DMA Data (No Filtering)

This model was made to cluster datapoints based on the following variables: Population 18+, Household Count, Med HHld Income, HHlds 1-2 Vehicles, HHlds 2+ Vehicles, White, African American, American Indian, Asian, Hawaiian/Pacific Islander, Hispanic

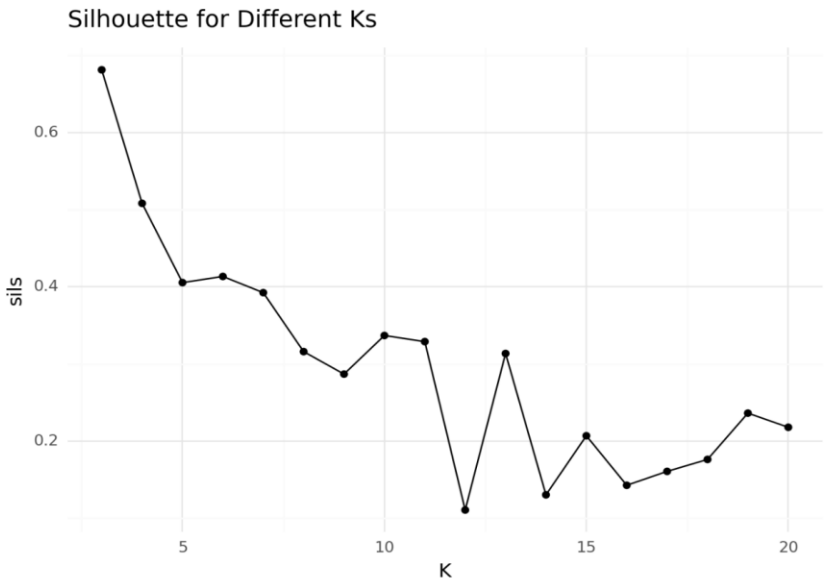


Plotting Silhouette Scores for Different Number of Clusters

After determining the optimal number of clusters to be 2, the model was fit and the final silhouette score of ~0.55022 was returned. While this is a rather poor silhouette score, it tells us that there isn't significant enough variation between the different DMA to be assigned a different cluster. Regardless, a within cluster analysis will need to be used to properly assess the characteristics of DMAs for each cluster.

Model 3 – Gaussian Mixture Model for DMA Data (Filtered Data to have Median Household Income >= Median Value)

This model was made to cluster datapoints based on the following variables: Population 18+, Household Count, Med HHld Income, HHlds 1-2 Vehicles, HHlds 2+ Vehicles, White, African American, American Indian, Asian, Hawaiian/Pacific Islander, Hispanic

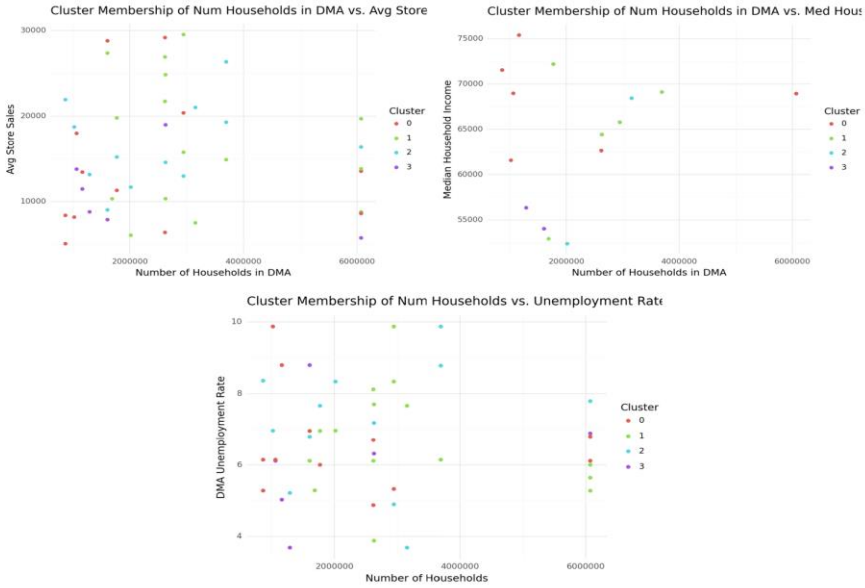


Plotting Silhouette Scores for Different Number of Clusters

After determining the optimal number of clusters to be 3, the model was fit and the final silhouette score of ~0.51415 was returned. While this is also a poor silhouette score, like the previous model a within cluster analysis will need to be used to properly assess the characteristics of DMAs for each cluster.

Results

Model 1 Results



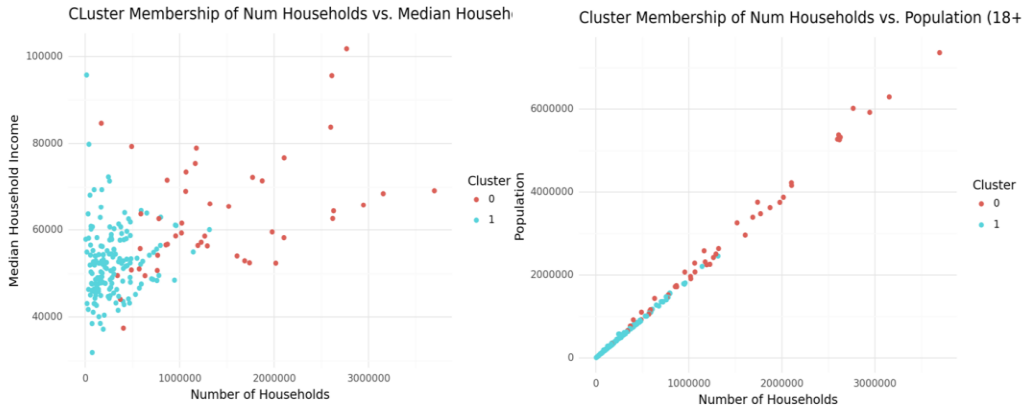
Scatterplots of Store Data Variables with Cluster Assignments

Cluster	Size	Avg Sales	Population (18+)	NumHouseholds	Med Household Income	HasVehicle	Unemployment Rate	CPI
0	43585.583333	14252.771995	5.096856e+06	2.390213e+06	67025.083333	2.248261e+06	6.581917	189.730097
1	203183.133333	17137.500173	6.826325e+06	3.235461e+06	64114.666667	3.017328e+06	6.669200	184.160047
2	114991.500000	16673.144398	5.228751e+06	2.563546e+06	64497.750000	2.361784e+06	7.126083	156.782246
3	152045.000000	11090.106329	4.944090e+06	2.303961e+06	64691.500000	2.162362e+06	6.135667	181.618947

Average Characteristics of Clusters

By comparing the scatterplots above to the original scatterplots from the store data, the cluster assignments can be visualized to the different stores. Since these DMAs have established stores, the within cluster data will be used to compare DMAs, based demographic/ household characteristics, rather than directly predict store performance.

Model 2 Results



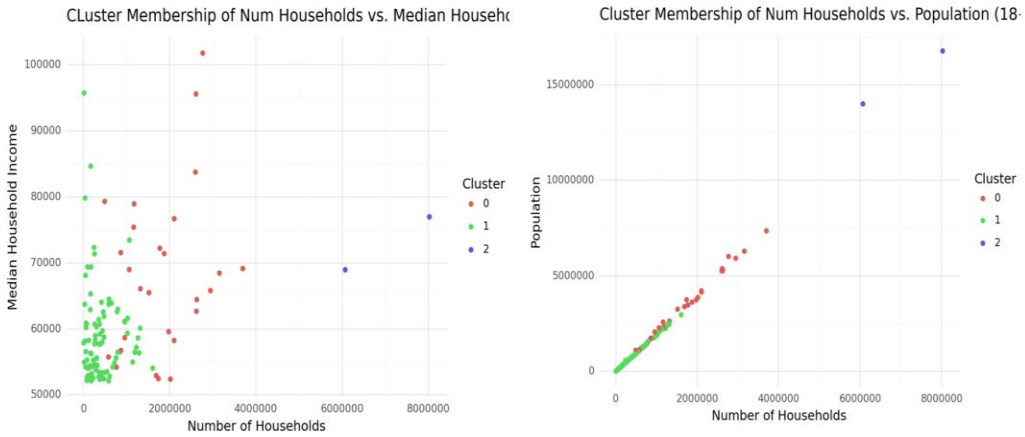
Scatterplots of Store Data Variables with Cluster Assignments

Cluster	Population 18+	Household Count	Med HHld Income	HHlds 1-2 Vehicles	HHlds 2+ Vehicles	White	African American	American Indian	Asian	Hawaiian/Pacific Islander	Other Race	Multi-Race	Hispanic
0	2.819142e+06	1.406243e+06	63395.00000	469888.913043	832699.652174	2.114293e+06	482085.26087	22172.413043	243299.630435	7364.586957	8356.195652	92290.695652	735862.173913
1	5.462504e+05	2.868225e+05	52430.90184	93745.705521	173491.785276	5.241771e+05	89932.95092	4230.625767	15225.067485	388.030675	852.417178	14381.036810	64112.411043

Average Characteristics of Clusters

By comparing the scatterplots above to the original scatterplots from the store data, the cluster assignments can be visualized to the different stores. Looking at the scatterplots, Cluster 0 (red datapoints) should be one of our focuses for deciding optimal DMAs, as the number of households is higher, and on average also have the higher median household income.

Model 3 Results



Scatterplots of DMA Data Variables with Cluster Assignments

Cluster	Population 18+	Household Count	Med HHld Income	HHlds 1-2 Vehicles	HHlds 2+ Vehicles	White	African American	American Indian	Asian	Hawaiian/Pacific Islander	Other Race	Multi-Race	Hispanic
0	3.611422e+06	1.782422e+06	67677.428571	5.980468e+05	1.050705e+06	2.636377e+06	6.143913e+05	22173.035714	3.563259e+05	11073.500000	10828.500000	117897.214286	9.710979e+05
1	7.748131e+05	4.059919e+05	58827.172043	1.313890e+05	2.463797e+05	7.514250e+05	1.125564e+05	5343.451613	2.805153e+04	598.236559	1293.591398	21632.924731	9.127108e+04
2	1.539158e+07	7.048974e+06	72950.000000	2.212000e+06	3.482883e+06	8.038506e+06	2.464570e+06	39467.500000	2.460030e+06	23955.500000	73847.500000	371763.500000	6.632986e+06

Average Characteristics of Clusters

By comparing the scatterplots above to the original scatterplots from the store data, the cluster assignments can be visualized to the different DMAs. Similar to the takeaways from the above model, clusters 0 (red datapoints) and 2 (blue datapoints) should be our focus for deciding optimal DMAs. These DMAs have many households, and even among our upper 50% quartile filter, are among the higher median household incomes.

Choosing Markets

How Markets Were Determined

As written earlier, the DMAs to consider are those in which stores already exist, and those that exist with clusters 0 and 2 in the third model. Of those DMAs that are still up for consideration, some key aspects were considered.

One of the most important factors is the number of households, as we want to ensure that opening a new location within the DMA will generate large amounts of traffic.

The next factor to consider is the median household income, and while all current DMAs fit the preferred requirements, taking another look at the leading DMAs within the high number of households can show which locations will deal with larger transaction volumes/ frequent customers.

The next thing to consider is the number of households with vehicles which can show consumers that have the means to make the trip to the new location as long as it is within a reasonable distance.

If applicable, the DMAs that had also had more specific MSA data were looked into to consider consumer expenditures and demographics at a closer level.

Lastly, for DMAs where stores were already present, we must consider the store's weekly sales, unemployment rates, and CPI.

Final Selection

Based on the criteria above, the 10 DMAs to choose had the following characteristics: comparatively higher number of households, comparatively higher median household income, and comparatively higher number of vehicles per household.

For DMAs that already had existing stores, the weekly sales, unemployment rates, and CPI were also compared to decide if they were meaningful locations.

By starting with a strong set of DMAs from modeling and returning the average characteristics of DMAs within clusters, and then filtering by preference/ fine tuning conditionals, and comparing MSA and feature data, 10 DMAs were found where opening a store may perform well.

The following 10 DMAs are locations that the data suggests would be best to open a store:

- **New York:** 8,025,869 households; 76,955 med income; 5,725,931 households with vehicles; 67,782 avg annual expenditures
- **Los Angeles:** 6,072,078 households; 68,945 med income; 5,663,836 households with vehicles; 66,971 avg annual expenditures; ~5.2 – ~7.8 unemployment rate; ~122 - ~140 CPI
- **Chicago:** 3,695,954 households; 69,122 med income; 3,261,278 households with vehicles; 60,582 avg annual expenditures; ~6.1 - ~9.8 unemployment rate; ~132 - ~223 CPI
- **Philadelphia:** 3,154,381 households; 68,438 med income; 2,787,085 households with vehicles; 65,436 avg annual expenditures; ~3.6- ~7.6 unemployment rate; ~139 CPI
- **Dallas-Fort Worth:** 2,944,691 households; 65,788 med income; 2,818,246 households with vehicles; 63,207 avg annual expenditures; ~4.8 - ~9.8 unemployment rate; ~132 - ~228 CPI
- **San Fracisco-Oakland-San Jose:** 2,766,578 households; 101,748 med income, 2,502,281 households with vehicles; 79,291 avg annual expenditures
- **Houston:** 2,628,496 households; 64,432 med income; 2,499,789 households with vehicles; 67,304 avg annual expenditures; ~3.8 - ~7.6 unemployment rate; ~132 - ~225 CPI
- **Washington D.C.:** 2,609,994 households; 95,570 med income; 2,365,982 households with vehicles; 79,921 avg annual expenditures
- **Boston-Manchester:** 2,598,094 households; 83,728 med income; 2,299,859 households with vehicles; 74,316 avg annual expenditures
- **San Diego:** 1,163,765 households; 75,397 med income; 1,104,003 households with vehicles; 79,585 avg annual expenditures; ~5.0 - ~8.7 unemployment rate; ~216 - ~228 CPI

Through the metrics listed before, all 10 of the DMAs were decided. Each location has strong prospects for growth by taking advantage of the characteristics and demographic variety within them, face minimal competition, as well as have many households that are doing well for themselves financially. Based on MSA data these locations also had relatively high expenditures,

which points to willingness to spend, and preferable consumer characteristics (earners, homeowners, spending habits).

The only outlier in the consideration is Los Angeles. While Los Angeles already houses established stores, the weekly sales of the stores, even after taking markdowns into account, perform particularly well. Furthermore, the number of households/ median household income is still among the top and the other demographics are also strongly distributed. As such, even with all the competition it is still worth considering it to be a potential location.

Additional Considerations

Unfortunately, the DMA data even after including the MSAs, DMA store trends, etc. doesn't tell as clear of a story about the characteristics of the respective DMAs and how a store in their location may perform. Having more complete DMA data with unemployment and CPI would have been very useful in determining the characteristics of the locations. Furthermore, when clustering with the DMA information, the silhouette scores were quite low, which is likely because there weren't many significant stand-out differences between some of the DMAs. This analysis provides a good foundation for the characteristics that make a particular market more appealing to open a location, but before a definite choice is made, there must be more considerations about the markets and competitor presence.