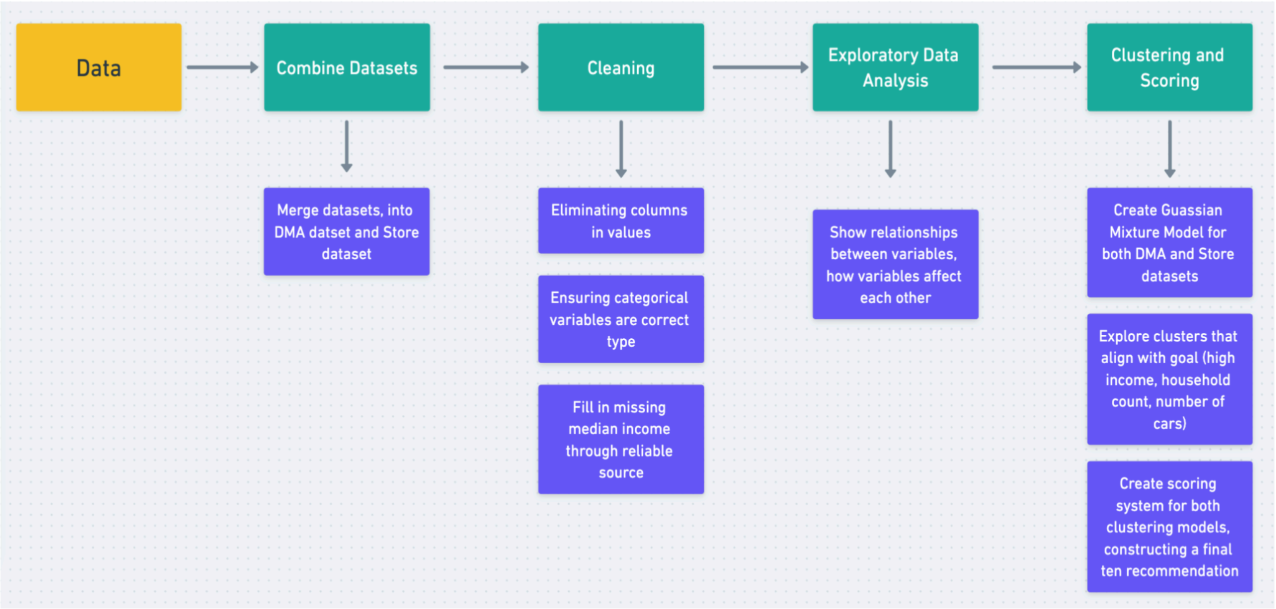
Using DS to Find Areas of Growth for Store Locations

Omair Memon

Analysis Plan

Presented with eight datasets, it is the following project's goal to effectively utilize the given data to quantitatively find the top ten places to locate a Target-'light' store. It is important to note that the competitors of this store are Walmart and Target for middle and high-income households, which are large retail chains with large and loyal customer bases. An analysis plan / data workflow has been included below, to give a better understanding of my analysis process.


Data
Combine Datasets
Cleaning
Exploratory Data Analysis
Clustering and Scoring
Merge datasets, into DMA datset and Store dataset
Eliminating columns in values
Ensuring categorical variables are correct type
Fill in missing median income through reliable source
Show relationships between variables, how variables affect each other
Create Guassian Mixture Model for both DMA and Store datasets
Explore clusters that align with goal (high income, household count, number of cars)
Create scoring system for both clustering models, constructing a final ten recommendation

To start it was imperative that I created consolidated datasets that integrated sales, store characteristics, and demographics for strategic analysis, focusing on average values and disregarding date specific information. This lengthy process was done all through excel, where I mapped data onto the DMA sheet to include location, population data, household data, vehicle data, transportation information, and demographic data. In addition, I created another dataset with store data and size and location, and mapping household vehicle data, population, number of houses, and median household income. Also on this dataset, I took the average sales, CPI, and unemployment rate from the date-time dataset so I could place it on the stores data set. This left me with two datasets with all the necessary data I planned on using for this analysis.

To be more precise, the final **DMA Dataset** included the following variables: DMA, Population 18+, Household Count, Med HHld Income, HHlds No Vehicles, HHlds 1-2 Vehicles, HHlds 2+ Vehicles, HHld Exp - Public Transport, HHld Exp - Intercity Bus Fare, HHld Exp - Mass Transit, HHld Exp - Taxi, HHld Exp - Other Public Transportation, White, African American, American

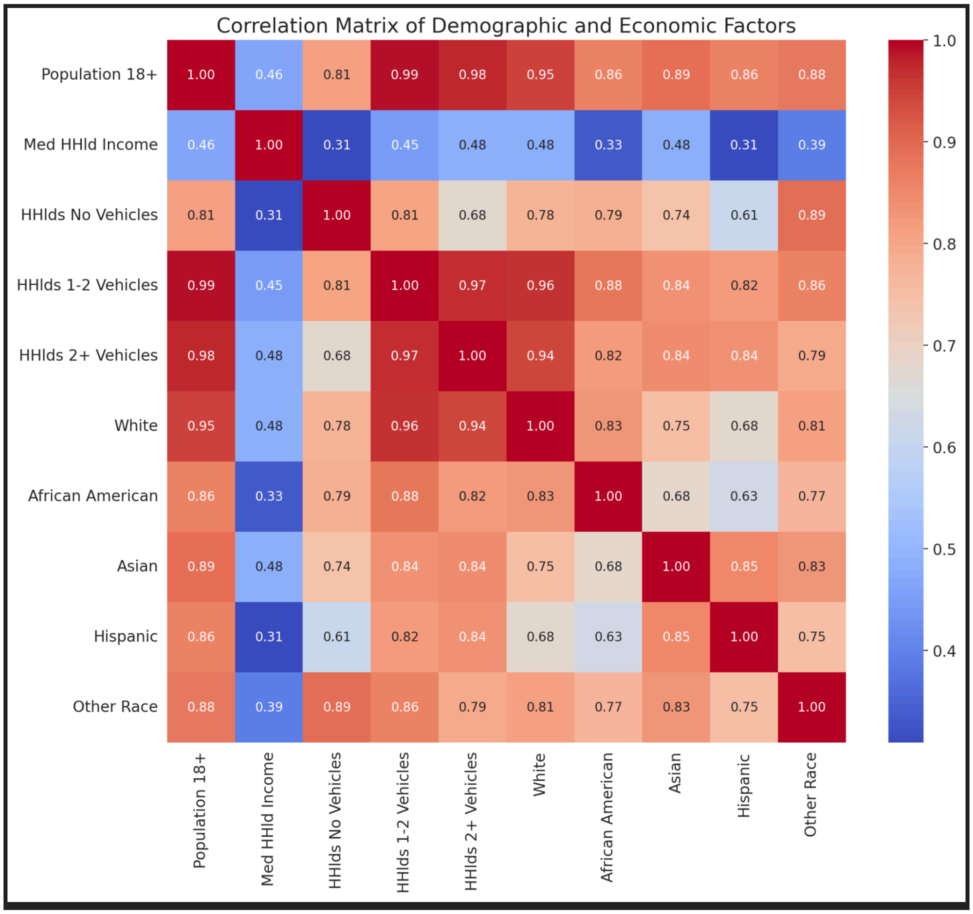
Indian, Asian, Hawaiian/Pacific Islander, Other Race, Multi-Race, Hispanic, HHLDS 1-2 Vehicles2.

The final **Store Dataset** included the following variables: Store, Type, Size, DMA, Households with Vehicles, Average Sales, Population (18+), NumberHouseHolds, Median Household Income, and CPI.

For the data cleaning process, I standardized column names across the datasets to ensure consistency during the merge, I used pivot tables to take the averages of the data stated above for a smooth integration into the consolidated store dataset, and converted data types where necessary, eliminating commas in certain values and ensuring categorical types are correct. In addition, I filled in missing income data with information from online (data.census.gov). After the datasets had been cleaned, I proceeded with an exploratory data analysis to have a clear understanding of the variables relationship to one another. Afterwards, I decided that it would be best to use two clustering techniques to group markets together, and create a ranking score for the stores based on the information that the clusters give. Using the clustering outputs for both datasets, I combined their scores to give a final top ten market recommendation. For the location of our stores, the main variables we want to prioritize are high income, high population (18+), and areas where people have at least 1-2 vehicles. This is because we know that high income enables people to spend more, high population allows for more customers, and vehicles allow for better accessibility, all leading to higher average sales.

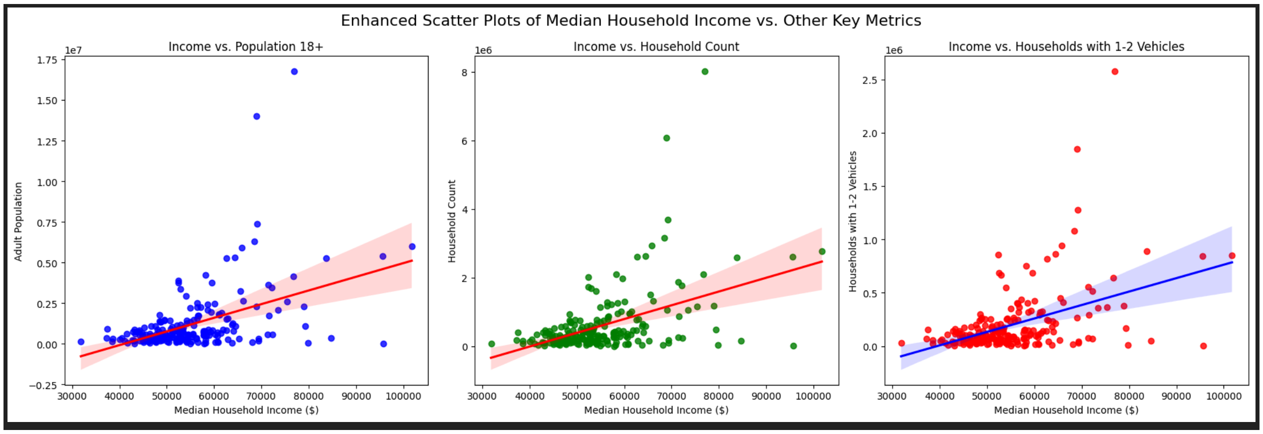
Exploratory Data Analysis

The EDA was a crucial part of this analysis, as it allowed for a comprehensive understanding of the variables presented. Mean adult population across all DMAs is about 1.18 million, with a standard deviation of 1.91 million, which indicated significant variability between DMAs. The average income across DMAs is approximately $55,016, with incomes ranging from $31,859 to $101,748, again showing a ton of variability. Transportation cost data was analyzed, showing that public transport is the largest expenditure out of the other costs, illustrating that areas with this type of transportation can make for a lack of household vehicles. A correlation matrix was created to portray that there is a high correlation between population and vehicle ownership, and income and vehicle ownership. In addition there is a clear correlation between demographics and income, with ‘Asian’ showing a high correlation and other demographics such as ‘Hispanic’ showing a lower correlation but are higher in number. This is imperative because as stated in the analysis plan, the goal is to find markets with the highest income and population, and the relationships shown below give us key insights into what to look for.



Correlation Matrix of Demographic and Economic Factors

	Population 18+	Med HHld Income	HHlds No Vehicles	HHlds 1-2 Vehicles	HHlds 2+ Vehicles	White	African American	Asian	Hispanic	Other Race
Population 18+	1.00	0.46	0.81	0.99	0.98	0.95	0.86	0.89	0.86	0.88
Med HHld Income	0.46	1.00	0.31	0.45	0.48	0.48	0.33	0.48	0.31	0.39
HHlds No Vehicles	0.81	0.31	1.00	0.81	0.68	0.78	0.79	0.74	0.61	0.89
HHlds 1-2 Vehicles	0.99	0.45	0.81	1.00	0.97	0.96	0.88	0.84	0.82	0.86
HHlds 2+ Vehicles	0.98	0.48	0.68	0.97	1.00	0.94	0.82	0.84	0.84	0.79
White	0.95	0.48	0.78	0.96	0.94	1.00	0.83	0.75	0.68	0.81
African American	0.86	0.33	0.79	0.88	0.82	0.83	1.00	0.68	0.63	0.77
Asian	0.89	0.48	0.74	0.84	0.84	0.75	0.68	1.00	0.85	0.83
Hispanic	0.86	0.31	0.61	0.82	0.84	0.68	0.63	0.85	1.00	0.75
Other Race	0.88	0.39	0.89	0.86	0.79	0.81	0.77	0.83	0.75	1.00

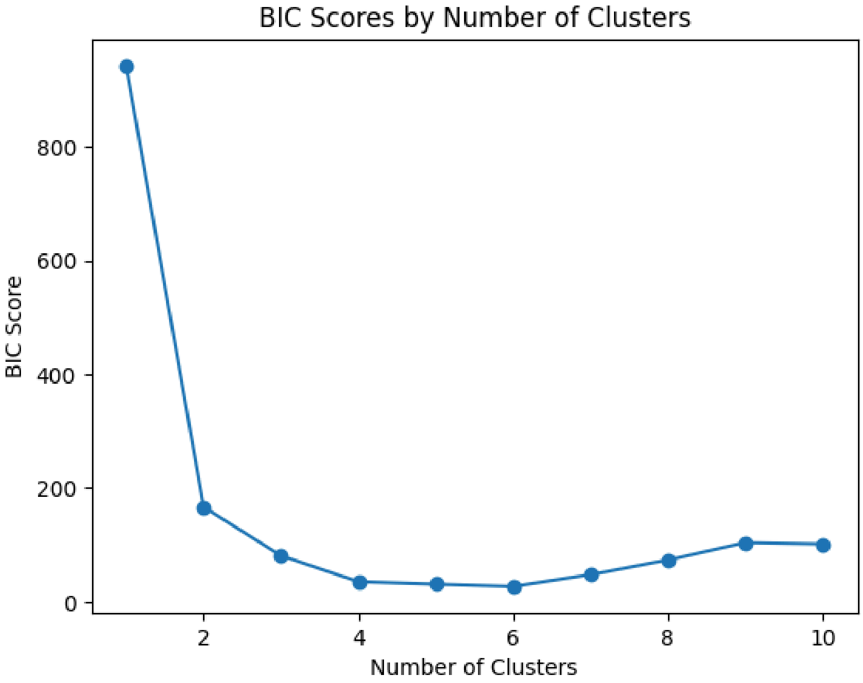


Enhanced Scatter Plots of Median Household Income vs. Other Key Metrics

The scatter plots above represent median household income and the adult population across different DMAs, how median household income correlates with the number of households in each DMA, and the relationship between median household income and households with 1-2 vehicles. As stated earlier these relationships are the most important representations for the ranking scorecard that will be created. As illustrated, there is a positive correlation for all three scatter plots, and they all look extremely similar. This shows that DMAs that combine high income, a substantial number of households, and good vehicle ownership are ideal candidates for new store locations. These factors typically contribute to higher consumer spending and greater accessibility to the stores, both critical components for the success of new retail locations targeting middle to high-income households.

Clustering

The bulk of the analysis was spent constructing multiple Gaussian Mixture Models to help get an idea of not only what crucial variables play into store success, but also where those numbers are being seen. To start, the first Gaussian Mixture Model was built using the DMA dataset described above, specifically focusing on the following variables: population data, median household income, and household 1-2 vehicles. A BIC score graph was used in concatenation with the elbow method to determine the optimal number of clusters. The optimal number of clusters turned out to be two and the graph is shown below.



Cluster Characteristics:

Cluster 0:

- Population 18+: Averages around 2.96 million, indicating these are larger metropolitan areas.
- Median Household Income: Averages about $65,162, which is higher compared to Cluster 1.
- Households with 1-2 Vehicles: Averages around 480,870, suggesting good transportation or higher dependency on vehicles, which might correlate with suburban areas.

Cluster 1:

- Population 18+: Averages around 492,463, indicating these are smaller cities or less densely populated areas.
- Median Household Income: Averages about $51,077, which is lower than Cluster 0, possibly pointing to less affluent areas.
- Households with 1-2 Vehicles: Averages around 85,184, which is significantly lower than Cluster 0, potentially indicating either urban areas with less need for multiple vehicles or rural areas with lower household counts.

These clusters present us with interesting and strategic insights that can be used to help determine where store locations will be the most competitive. Cluster 0 seems more suited for stores that aim to attract a wealthier customer base with larger family units or a need for vehicles. Cluster 1 could be targeted for expansion in areas with a smaller population and potentially lower competition, focusing on markets where store presence might be limited. For this specific problem, it is clear that Cluster 0 makes for a better deciding factor for store locations. This is because median household income, population, and households with 1-2 vehicles are the main drivers of average sales, and in order to compete with the likes of Walmart and Target for mid to high income markets, these variables need to be emphasized.

This cluster included the following locations as promising store targets based on the DMA dataset clustering:

City	Population	Median Household Income
New York	16.78 million	$76,955
Los Angeles	14.00 million	$68,945
Chicago	7.36 million	$69,122
Philadelphia	6.30 million	$68,438

San Francisco-Oakland-San Jose	6.02 million	$101,748
Dallas-Fort Worth	5.92 million	$65,788
Washington DC	5.38 million	$95,570
Houston	5.32 million	$64,432
Boston-Manchester	5.28 million	$83,728
Atlanta	5.26 million	$62,663

In addition, it is clear that the stores should be located in areas where most people have cars. This is because to be able to compete with stores with size such as Walmart, it is imperative that people are able to drive to Target-'light'. With a considerable product selection, a loyal customer base can be created that may be willing to drive an extra mile, passing Walmart and Target and coming to our store. However, if they do not have a car, then instead of going the extra distance they will settle on walking to one of the bigger retail stores as they are most likely closer to them. Moreover, it is imperative that the store locations are in areas with 1-2 vehicles prioritized. Therefore, a deeper analysis of Cluster 0 was employed to find the locations with most optimal number of vehicles, these locations were outputted:

Los Angeles, CA: 1,845,727 households with 1-2 vehicles, 3,818,109 with 2+ vehicles
New York, NY: 2,578,274 households with 1-2 vehicles, 3,147,657 with 2+ vehicles
Chicago, IL: 1,277,861 households with 1-2 vehicles, 1,983,417 with 2+ vehicles
Dallas-Fort Worth, TX: 942,743 households with 1-2 vehicles, 1,875,503 with 2+ vehicles
Philadelphia, PA: 1,078,603 households with 1-2 vehicles, 1,708,482 with 2+ vehicles

Markets like Los Angeles and New York, despite their higher competition, offer substantial customer bases with significant vehicle ownership, making them promising locations for new stores. These areas are extremely beneficial locations as they offer better accessibility for customers.

In addition, a Gaussian Mixture Model was created for the store dataset as well to group markets based on similarities in demographic and competitive characteristics. As done earlier, a BIC score was calculated to determine four as the optimal number of clusters.

Cluster Characteristics:

Cluster	Average Store Size (sq ft)	Median Household Income	Average CPI	Unemployment Rate
0	160,981	$60,340	158.25	7.56%
1	119,613	$85,217	211.28	7.34%
2	119,743	$80,695	130.49	11.47%
3	144,423	$85,298	134.97	7.39%

Clusters 1 and 3 appear to have higher median household incomes, which aligns with our goal of targeting middle to high income households. These clusters offer attractive markets due to their economic characteristics. Cluster 2 has a higher unemployment rate which might indicate economic challenges, illustrating a market that we would want to avoid. Cluster 0 has the lowest median household income, which does not align as well with our target demographic. Therefore, finding the specific DMA's within clusters 1 and 3 would indicate strong market potential.

These clusters included the following locations as promising store targets based on the store dataset clustering:

City	Median Household Income	Average Sales	Unemployment Rate	Cluster
Denver	$105,790	$10,575	6.68%	1
San Diego	$99,931	$13,274	8.95%	3
Atlanta	$90,502	$12,604	6.87%	1
Philadelphia	$88,974	$19,776	7.62%	1
Houston	$80,250	$22,720	7.90%	3
Dallas-Ft.Worth	$78,842	$16,034	7.61%	3
Chicago	$76,639	$11,513	7.57%	3
Los Angeles	$76,471	$17,723	7.97%	1

Picking the Ten Most Promising Markets

To determine the top ten locations for the launch of new retail locations, a comprehensive scorecard approach was employed, integrating data from the two distinct cluster analyses described above. The scoring approach was designed to combine insights from both the DMA dataset and the Store dataset to create a unified ranking of potential locations. For the DMA clustering, a score was given based on their demographics and economic indicators. Metrics such as population, median household income, and households with 1-2 vehicles were scored on a scale from 1 to 10, with higher scores indicating a more favorable environment for a retail store. For the clustering algorithm using the store dataset, markets were evaluated based on data reflecting store performance including average sales and economic health indicated by the unemployment rate. These metrics were also scored on a scale from 1 to 10. The top ten locations as well as their scoring can be seen below.

Rank	City	DMA Score	Store Score	Total Score
1	Los Angeles	22.28	21.0	43.28
2	Houston	13.38	25.0	38.38
3	Philadelphia	14.66	22.0	36.66
4	Chicago	16.14	19.0	35.14
5	Atlanta	13.95	20.0	33.95
6	Dallas-Fort Worth	13.65	17.0	30.65
7	San Diego	12.3	18.0	30.3
8	New York	27.56	-	27.56
9	San Francisco-Oakland -San Jose	23.28	-	23.28
10	Washington DC	20.36	-	20.36

A table of the key metrics for each location can be viewed below.

Market	Population	Median Household Income	Average Sales	Unemployment Rate	Households with 1-2 Vehicles
Los Angeles	14,000,000	$68,945	$17,723	7.97%	1,845,727

Houston	5,320,000	$64,432	$22,720	7.90%	1,000,000
Philadelphia	6,300,000	$68,438	$19,776	7.62%	1,078,603
Chicago	7,360,000	$69,122	$11,513	7.57%	1,277,861
Atlanta	5,260,000	$62,663	$12,604	6.87%	1,200,000
Dallas-Fort Worth	5,920,000	$65,788	$16,034	7.61%	942,743
San Diego	2,583,107	$99,931	$13,274	8.95%	1,290,536
New York	16,780,000	$76,955	N/A	N/A	2,578,274
San Francisco-Oakland-San Jose	6,020,000	$101,748	N/A	N/A	2,500,000
Washington DC	5,380,000	$95,570	N/A	N/A	2,000,000

This list provides a strategic blend of demographic strengths and strong market potential. The key metrics that have been emphasized so heavily in this report are the main drivers for this selection. High population centers offer a larger base of potential customers and higher median household income provides insight into the purchasing power prevalent in each location, which in turn leads to higher sales. Lower unemployment rates lead to more income which can influence consumer spending patterns, specifically promoting spending. Households with 1-2 vehicles show the accessibility and the likelihood that residents have easy access to Target-'light'. Markets such as Los Angeles, Houston, Philadelphia, Chicago, Atlanta, and Dallas all have above average median household incomes compared to the full DMA list, and due to their exposure to the other key drivers, they are great markets for store locations. New York, San Francisco, and Washington DC all offer high populations, median income, and households with 1-2 vehicles, making them perfect locations for Target 'light'. However, due to their lack of store data, they are missing scores for that cluster, and would most definitely be ranked higher if that data was available. Furthermore, it is clear that the above locations have the best market potential to infiltrate and compete in based on the reasons emphasized above.