Retail Analytics Report for Track 2: The Analysis of the Relationship Between Store Location and Revenue Based on Dominick's Finer Foods Category Specific File Data

By Cassie Liu, Xuanyi Zhu, Lehan Dong, Yunhong Wang
Professor: Prof. Mitsukuni Nishida

1. Executive Summary

This study aims to analyze the relationship between store location and revenue using Dominick's Finer Foods category-specific data. By examining key variables such as store density per capita ('density'), shopping ability index ('shopindx'), and demographic factors like the percentage of detached houses ('sinhouse'), we seek to identify retail market gaps and inform new store location decisions. Additionally, we evaluate the competitive environment by assessing the suitability of areas with highly active shoppers ('shopbird') or constrained shoppers ('shopcons') for high-end or discount retail stores. Our findings will provide actionable recommendations for optimal store location selection, balancing market potential and competitive dynamics to maximize revenue.

2. Research Business Problems

2.1. New Store Location Decision

- Combining 'density' (store density per capita) and 'shopindx' (shopping ability index) to identify retail market gaps.
- Is a high 'sinhouse' (percentage of detached houses) area suitable for opening high-end supermarkets?
- Assuming the Dominick has not closed, we will provide our suggestions for the location based on the analysis of hypotheses.

2.2. Competitive Environment Assessment

- Are areas with a high 'shopbird' (highly active shoppers) suitable for opening more competitive retail stores?
- Are areas with a high 'shopcons' (constrained shoppers) more suitable for discount stores rather than high-end retail?

3. Dataset Introduction and Preprocessing

3.1. Dataset Overview

The dataset used in this analysis consists of store-specific demographic data, originally sourced from the 1990 U.S. Census for the Chicago metropolitan area. The data was further processed by Market Metrics to generate detailed demographic profiles for each of the Dominick's Finer Foods stores.

This dataset initially contained 108 observations and 510 variables, covering a wide range of demographic and economic characteristics relevant to retail decision-making.

However, after removing irrelevant and entirely empty columns, we narrowed the focus to 54 key variables and addressed missing values to ensure data quality.

3.2. Data Cleaning and Imputation

To address missing values in key variables, we employed two primary imputation strategies: mean-based imputation, mode-based imputation, and K-Nearest Neighbors (KNN) imputation.

3.2.1. City-Wise Mean Imputation

Missing values for zip code, latitude (lat), longitude (long), and weekly sales volume (weekvol) were imputed using the mean values within each city, ensuring consistency across stores in the same region.

3.2.2. K-Nearest Neighbors (KNN) Imputation

The KNN imputation method (k=3) was applied to estimate missing values based on geographical proximity, using latitude (lat) and longitude (long) as reference points. After KNN imputation, temporary _imp variables were removed to maintain dataset cleanliness.

3.2.3. Mode-Based Imputation for city Using ZIP Code

Since **ZIP codes** are strongly associated with specific cities, missing city values were filled using the most frequently occurring city within each ZIP code group.

Through these steps, we ensured data completeness and accuracy, significantly improving the reliability of subsequent analyses. Finally, we got our clean dataset with 97 observations and 54 variables a comprehensive. Summary of all variables is provided in Appendix 1.

4. Cluster and Correlation Analysis

4.1. Cluster Analysis of Consumer Shopping Behavior

In this study, K-means clustering is employed to classify different shopping areas, aiming to identify distinct patterns of shopping behavior and retail strategy formulation.

To segment consumer shopping areas, we utilize four key indicators: Shopbird (shopping activity), Shopavid (targeted shopping index), Shophurr (impulse shopping index), and Shopindx (shopping ability index). Prior to analysis, missing values are removed, and all variables are standardized to ensure comparability. Given the complexity of consumer behavior, we set the number of clusters at K=4, obtaining four distinct types of shopping areas. However, due to the challenges posed by high-dimensional data in K-means clustering results, t-distributed stochastic neighbor embedding (t-SNE) is applied for dimensionality reduction, enabling a more interpretable visualization of clustering patterns. A convex hull is constructed to delineate cluster boundaries, enhancing the clarity and interpretability of the clustering outcomes (Appendix 2).

To better understand the characteristics of various shopping areas, we restore the dimensions of the four categories. After calculating the max-min range of each cluster, shopping areas can be divided into four categories (Details in Appendix 3).

Category	Shopping Bird Index	Shopping Avid Index	Shopping Hurr Index	Shopping Index	Consumer Characteristics

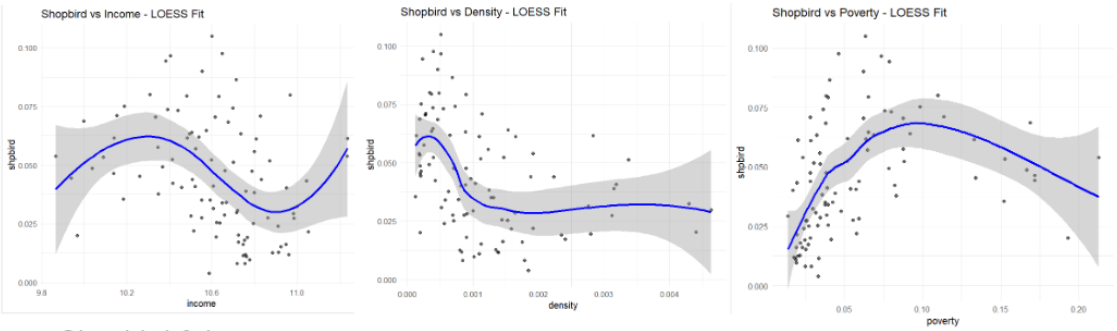
High-Activity Commercial Hubs	Low	High	Medium-High	Medium-High	Strong shopping ability Enthusiasm for shopping Suitable for brand loyalty marketing
Low-Engagement Retail Zones	Medium	Medium-High	Low	Low	Low shopping ability Conservative consumption Suitable for discount marketing
Impulse-Driven Shopping Districts	Medium	Low	Medium	Medium-High	Strong economic ability Easy to be affected by promotions Suitable for limited-time discount strategy
Balanced Retail Areas	High	Medium	Low-Medium	Medium	Balanced shopping habits Like shopping Suitable for experiential shopping

4.2. Correlation Analysis

To investigate the relationship between 'shopbird' and key socioeconomic factors—'income' (income), 'density' (population density), and 'poverty' (poverty rate)—we first conducted Pearson correlation tests. The results indicated that 'shopbird' exhibited significant correlations with all three variables (p-value < 0.01).

Factor	Correlation Coefficient	P-value	Interpretation
Income	-0.31	0.002***	Negative
Density	-0.29	0.004***	Negative
Poverty	0.32	0.001***	Positive

To visualize these relationships more intuitively, we plotted both linear regression and LOESS models, fitting two different trend lines. Given the potential presence of outliers in 'density,' we removed its maximum value to minimize its impact on the analysis. The LOESS model revealed that the relationships between 'shopbird ' and three variables were nonlinear, suggesting more complex interactions.



Shopbird & Income Shopbird & Density Shopbird & Poverty

Furthermore, to gain deeper insights into these relationships, we conducted a grouping analysis for 'income' and 'density.' 'Income' was divided into three categories (low, medium, and high) based on tertiles, and a boxplot was created to illustrate variations in 'shopbird' across different income groups. We then performed a one-way analysis of variance (ANOVA) to determine whether significant differences existed in the mean values of 'shopbird' across these groups. Similarly, 'density' was categorized into three groups— low, medium, and high—based on its range. A boxplot was generated to visualize changes in 'shopbird' across different density levels, and ANOVA was applied to assess statistical differences between these groups (Appendix 4).

The results revealed a significant effect of income group on shopping behavior, $F(2, 93) = 10.98$, $p < 0.001$, indicating substantial differences among the groups. Similarly, the density group also showed a significant effect, $F(2, 94) = 14.65$, $p < 0.001$. These findings suggest that both income level and population density have a meaningful impact on shopping behavior patterns.

The results of ANOVA and visual analysis show that shopping activity (Shopbird) varies significantly among groups with different incomes and population densities. Specifically, people whose income is lower than the middle would like shopping more, and areas with dense stores would like shopping more, but there is little difference between medium and sparse, and poor people may like shopping more.

Based on the findings, retail site selection strategies should consider the significant impact of income levels and population density on shopping behavior. The company can target low-income, low-density areas with frequent, accessible retail options and develop experience-oriented shopping centers in high-income, high-density areas to improve their strategies.

5. Location Selection Potential Analysis: Retail Suitability Index

5.1. Purpose of Retail Score

The main purpose of calculating the retail score is to evaluate the retail suitability of each store by quantifying the impact of factors such as business district activity, consumer income, population density, and transportation convenience. By calculating the retail score, we can evaluate the potential of each store, find the most suitable location for retail business, and improve the accuracy of decision-making.

5.2. Calculation and weight selection

The calculation formula for Retail Score is as follows:

*Retail Score = 0.2537 * (0.5 * shpbird + 0.5 * shopindx) + 0.2366 * (0.5 * income + 0.5 * hsizeavg) + 0.1944 * (0.5 * density + 0.5 * shpcons) + 0.1813 * (1 - nocar) + 0.1803 * (0.5 * sinhouse + 0.5 * hvalmean)*

The weights used in this calculation are based on studies by Kuo et al. (2002), Huizhong Ye et al. (2009), and Cagri Tolga, A. et al. (2013). These studies utilized the Fuzzy AHP method combined with expert questionnaires and focus group data to quantify the factors affecting location selections and calculate the weight. Specific coefficients are as follows:

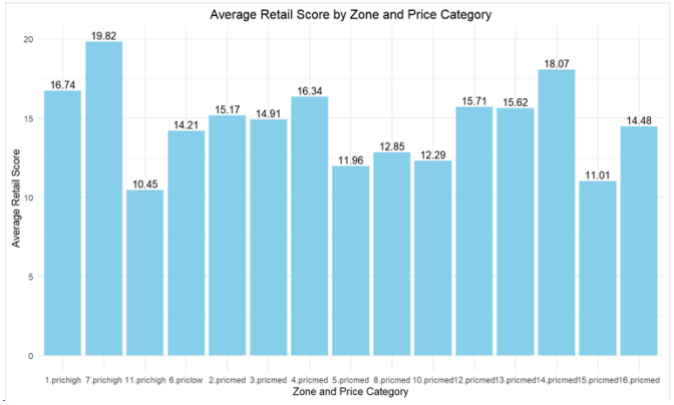
Factor	Variables	Description	Weight
Business Attractivenes	shpbird shopindx	Reflects the activity and attractiveness of the business district. Higher activity typically leads to more customer traffic.	0.2537
Consumer Income	income, hsizeavg	Income level affects purchasing power. Higher areas tend to have higher sales potential.	0.2366
Market Competition	density shpcons	The level of competition impacts profitability. Less competition means a larger market share.	0.1944
Traffic Conditions	1-nocar	Traffic convenience is key to customer access to the store. Better traffic conditions are ideal.	0.1813
District Conditions	sinhouse hvalmean	Reflects family income and property value, impacting purchasing power and profitability.	0.1803

Through the weighted calculation of these coefficients, we can obtain the Retail Score of each store and evaluate the retail suitability of the store accordingly.

5.3. Retail Score analysis results

Average Retail Score by Zone and Price Category: We grouped the data by zone and price category and calculated the average Retail Score. Zone 7 had the highest average Retail Score, and the prichigh category stores had significantly higher scores compared to other categories, indicating that high-price areas are likely located in more favorable commercial areas, attracting more high-purchasing-power customers.

	zone <dbl>	pricerange <chr>	avg_retail_score <dbl>
1	7	prichigh	19.8
2	14	pricmed	18.1
3	1	prichigh	16.7
4	4	pricmed	16.3
5	12	pricmed	15.7
6	13	pricmed	15.6
7	2	pricmed	15.2
8	3	pricmed	14.9
9	16	pricmed	14.5
10	6	priclow	14.2
11	8	pricmed	12.8
12	10	pricmed	12.3
13	5	pricmed	12.0
14	15	pricmed	11.0
15	11	prichigh	10.5



Average Retail Score by Price Category: Based on the previous analysis, we further analyzed the average Retail Score of stores in different price categories. Stores in the prichigh category showed higher Retail Score values, aligning with expectations—high-price stores are typically in high-income, competitive areas and have higher appeal.

pricerange <chr>	avg_retail_score <dbl>
prichigh	16.9
pricmed	14.6
priclow	14.2

Top 10 Stores Analysis: By sorting stores by Retail Score, we identified the top 10 stores. Most of these stores were in the prichigh category, reflecting the best geographic locations and market conditions.

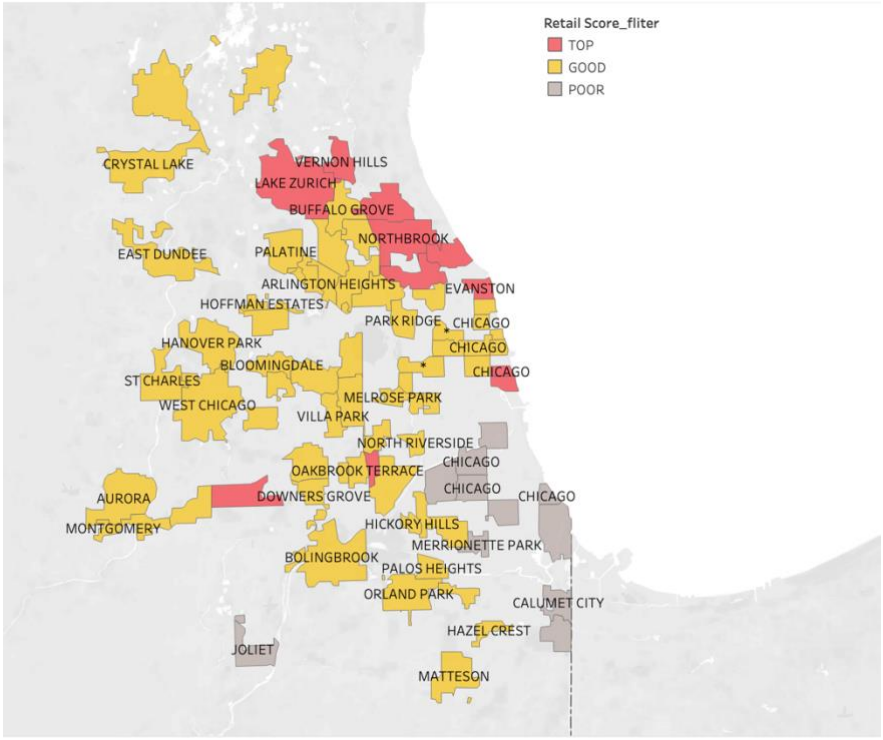
	store	Retail_Score	zone	city	pricerange
	<dbl>	<dbl>	<dbl>	<chr>	<chr>
1	62	26.1	1	NORTHFIELD	prichigh
2	109	25.4	7	BANNOCKBURN	prichigh
3	137	24.9	1	EVANSTON	prichigh
4	33	24.0	7	CHICAGO	prichigh
5	52	22.6	1	NORTHBROOK	prichigh
6	14	21.7	1	GLENVIEW	prichigh
7	129	21.2	12	LAKE ZURICH	pricmed
8	44	20.1	2	WESTERN SPRINGS	pricmed
9	77	20.1	6	VERNON HILLS	priclow
10	54	19.8	2	NAPERVILLE	pricmed

5.4. Conclusion and Recommendations

From the analysis of the Retail Score, it is evident that stores in high-price areas tend to have higher suitability scores, driven by the presence of high-income customers, competitive business districts, and convenient transportation. Retailers should prioritize these factors when selecting new locations, especially in prichigh areas, to ensure they can attract enough high-end customers.

6. Geographical analysis and Recommendation for location selection

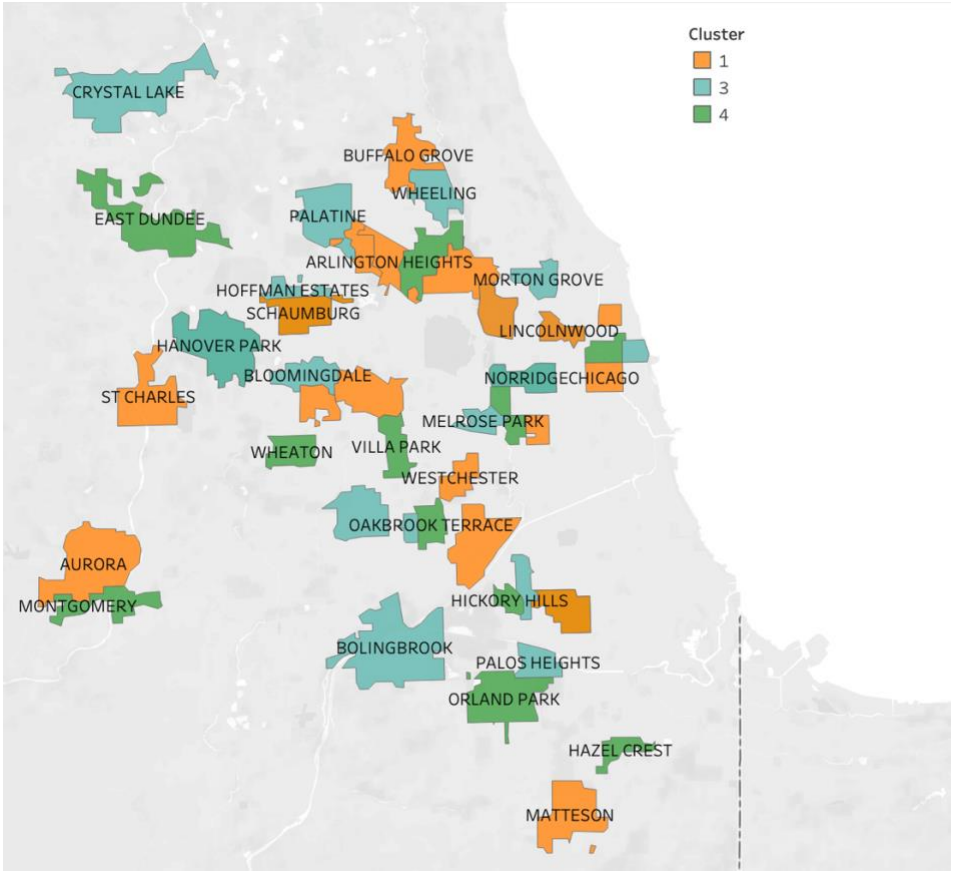
(For viewing our Interactive Tableau Dashboard, click here.)



Based on the above analysis of the retail scores, we can see that the retail scores for top 10 regions range from 19.7 to 26. Based on these scores, we categorized the regions into three distinct groups: Top (Score ≥ 19.7), Good (11 ≤ Score < 19.7), and Poor (Score ≤ 10).

For regions in the “Top” group, we recommend opening new stores, as these areas demonstrate strong retail potential and favourable market conditions. These areas also belong to the clusters with strong shopping or economic ability. Regions in the “Good” group may be suitable for specific store formats, such as discount or outlet stores, which align with the

moderate retail activity and consumer behavior observed in these areas. Finally, for regions in the "Poor" group, we advise against opening new stores due to limited shopping activity, lower consumer capability, and insufficient consumption potential.



When delving into the "Good" group, we interpret the result from clustering and try to find their intersection. As what we have suggested, the regions in Cluster 2 have the lowest shopping index and overall weak shopping behaviour; we excluded Cluster 2, and the above picture illustrates the specific area of regions that we suggest opening discount or outlet stores. It is highly recommended to viewing our Interactive Tableau Dashboard here to get more detailed information. The region names, their corresponding cluster, and retail score can be found in the Appendix 5.

7. Conclusion

Based on the analysis, our report highlights the significant influence of store location on revenue, with key factors such as shopping activity, income levels, population density, and competition shaping retail performance. By employing clustering techniques, correlation analysis, and a Retail Suitability Index, we identified optimal locations for new store openings and tailored retail strategies for different market segments. The findings emphasize the importance of aligning store types with consumer behaviour and economic conditions to maximize profitability. Retailers should leverage these insights to make data-driven decisions, ensuring strategic expansion and enhanced market positioning.

References

Cagri Tolga, A., Tuysuz, F., & Kahraman, C. (2013). A fuzzy multi-criteria decision analysis approach for retail location selection. *International Journal of Information Technology & Decision Making*, 12(04), 729-755.

Kuo, R. J., Chi, S. C., & Kao, S. S. (2002). A decision support system for selecting convenience store location through integration of fuzzy AHP and artificial neural network. *Computers in industry*, 47(2), 199-214.

Ye, H., Cai, C., & Zhang, H. (2009). Applying Fuzzy AHP in Selecting Supermarket Chain store Location's important factor. *Department of Industrial Engineering*, University of Kaohsiung, Kaohsiung, Taiwan.

Appendices

Appendix 1: Dominick's Dataset Summary

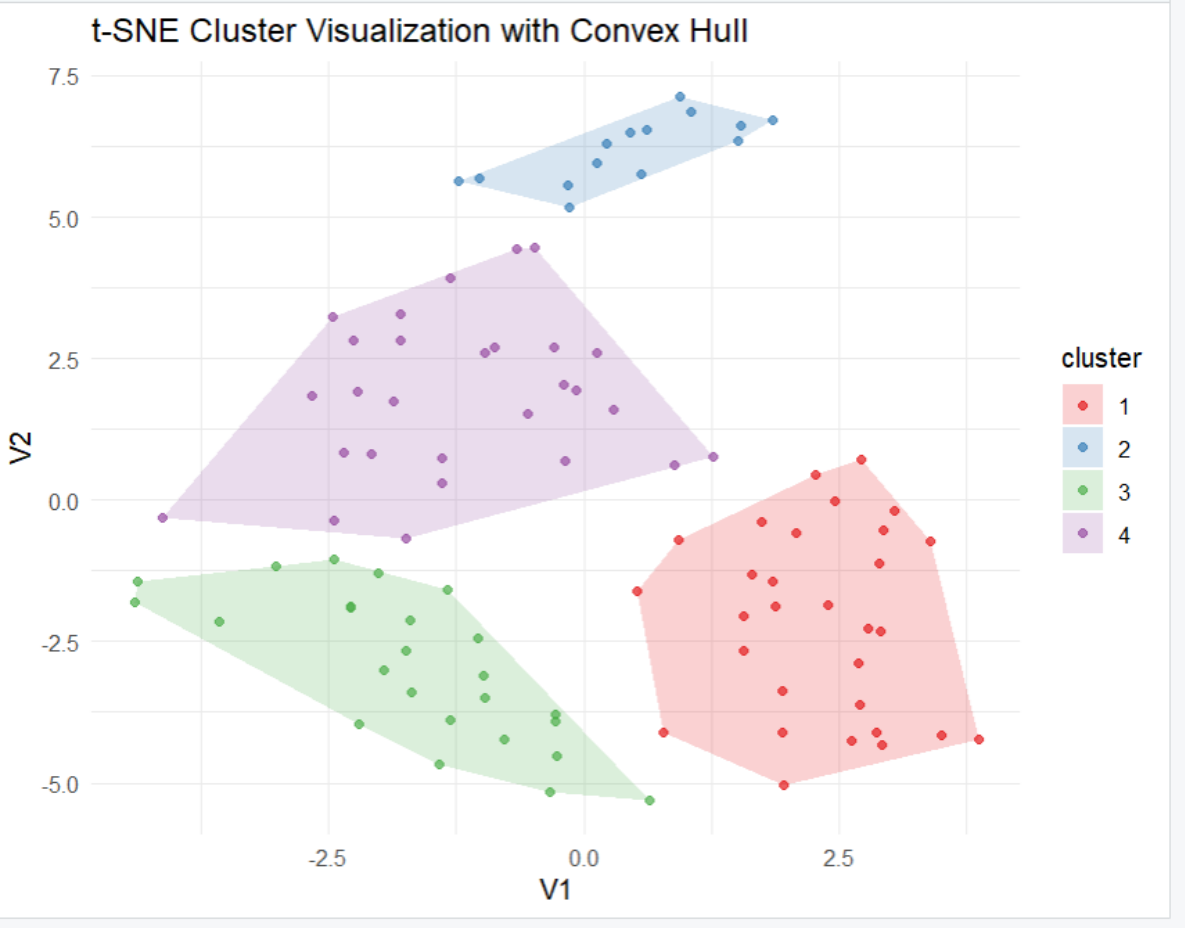
Category	**Variable Name**	**Description**	**Min**	**Max**	**Mean**	**Median**
Geographical Information	city	City name where the store is located	/	/	/	/
	zone	Trading area classification	/	/	/	/
	store	Store number of Dominicks	2	315	107.9	92
	zip	ZIP code of the store	60004	60662	60344	60435
	lat	Latitude of the store	415052	422413	419081	419364
	long	Longitude of the store	875394	883625	878907	878511
Retail Sales Volume	weekvol	Weekly sales volume for each store	250	875	464.8	450
Demographics	income	Log of Median Income	9.867	11.236	10.609	10.635
	incsigma	Std dev of Income Distribution (Approximated)	18857	30277	24671	24346
	density	Trading Area in Sq Miles per Capita	0.000122	0.007727	0.001213	0.000885
	hvalmean	Mean Household Value	64.35	267.39	146.15	148.95

		(Approximated)				
	age9	% Population under age 9	0.04607	0.20016	0.14024	0.13988
	age60	% Population over age 60	0.05805	0.31293	0.17005	0.16041
	ethnic	% Blacks & Hispanics	0.02283	0.99569	0.15362	0.07466
	educ	% College Graduates	0.04955	0.52836	0.21881	0.22079
	nocar	% With No Vehicles	0.01237	0.55065	0.1063	0.04995
	single	% of Singles	0.2031	0.5935	0.2774	0.259
	retired	% of Retired	0.05597	0.23611	0.14722	0.15377
	unemp	% of Unemployed	0.1403	0.2446	0.1827	0.1805
	nwhite	% of population that is non-white	0.02855	0.99497	0.19697	0.12998
	poverty	% of population with income under $15,000	0.01373	0.21296	0.05718	0.03928
	sinhouse	% Detached Houses	0.01671	0.83061	0.55319	0.59279
Household Characteristics	hsizeavg	Average Household Size	1.554	3.309	2.68	2.676
	hsize1	% of households with 1 person	0.1145	0.614	0.2411	0.2294
	hsize2	% of households with 2 persons	0.2188	0.3686	0.3087	0.3065
	hsize34	% of households with 3 or 4 persons	0.09203	0.4455	0.33213	0.33295
	hsize567	% of households with 5 or more persons	0.01351	0.21635	0.1181	0.11327
	hh3plus	% of households with 3 or more persons	0.1055	0.6504	0.4502	0.4517
	hh4plus	% of households	0.04052	0.4428	0.27713	0.27443

		with 4 or more persons				
	hhsingle	% of households with 1 person	0.1145	0.614	0.2411	0.2294
	hhlarge	% of households with 5 or more persons	0.01351	0.21635	0.1181	0.11327
	hval150	% of Households with Value over $150	0.002509	0.9167	0.338482	0.342017
	hval200	% of Households with Value over $200	0.000646	0.780819	0.176016	0.133652
	telephn	% of households with telephones	0.8387	0.9976	0.9766	0.9879
	mortgage	% of households with mortgages	0.4431	0.9605	0.713	0.7179
Employment & Socioeconomic Status	workwom	% Working Women with full-time jobs	0.2445	0.4723	0.3585	0.355
	wrkch5	% of working women with children under 5	0.02422	0.11755	0.05739	0.05272
	wrkch17	% of working women with children 6 - 17	0.04123	0.19809	0.12527	0.12502
	nwrkch5	% of non-working women with children under 5	0.03011	0.16852	0.08659	0.08418
	nwrkch17	% of non-working women with children 6 - 17	0.01762	0.12167	0.06967	0.07041
	wrkch	% of working women with children	0.07134	0.29348	0.18265	0.17509
	nwrkch	% of non-working women with children	0.04773	0.25009	0.15626	0.15935

	wrkwch	% of working women with children under 5	0.02399	0.11467	0.0565	0.05217
	wrkwnch	% of working women with no children	0.1566	0.4595	0.2553	0.2561
Shopping Behavior Segments	shopcons	% of Constrained Shoppers	0.01882	0.27926	0.08188	0.05702
	shphurr	% of Hurried Shoppers	0.02635	0.28596	0.15325	0.15548
	shpavid	% of Avid Shoppers	0.06131	0.30965	0.19396	0.19444
	shopstr	% of Shopping Stranges	0.16301	0.5577	0.2846	0.2769
	shopunft	% of Unfettered Shoppers	0.1451	0.3914	0.2414	0.2348
	shopbird	% of Shopper Birds	0.003979	0.104839	0.044968	0.043251
	shopindx	Ability to Shop (Car and Single Family House)	0.000001	0.986333	0.74593	0.817941
	shpindx	Ability to Shop (Car and Single Family House)	0.000001	0.986333	0.74593	0.817941

Appendix 2: t-SNE Cluster Visualization with Convex Hull



Appendix 3: Details of four clusters

Cluster	Shopping Bird Index (min, max)	Shopping Video Index (min, max)	Shopping Hurr Index (min, max)	Shopping Index (min, max)	Consumer Characteristics
1	0.00398 - 0.0570	0.211 - 0.310	0.0928 - 0.286	0.759 - 0.970	High shopping index, moderate video engagement, low bird index
2	0.0201 - 0.0752	0.0613 - 0.258	0.0264 - 0.0937	0.00000010 8 - 0.397	Lowest shopping index, overall weak shopping behavior
3	0.0110 - 0.0613	0.104 - 0.197	0.118 - 0.280	0.670 - 0.986	High shopping index, relatively high hurr index, low bird index
4	0.0476 - 0.105	0.126 - 0.229	0.0978 - 0.215	0.443 - 0.955	High shopping index, balanced engagement across all categories

Appendix 4: ANOVA Results

Factor	Sum of Squares	Mean Square	F-value	p-value	Significanc

Income Group	0.01055	0.005276	10.98	5.24e-05 ***	Significant
Density Group	0.01319	0.006595	14.65	2.9e-06 ***	Significant

Appendix 5: Cities Suggestions for discount or outlet stores, with their corresponding Cluster and Retail Score

Cluster	City Name	Retail Score
1	COUNTRYSIDE	19.446
	WESTCHESTER	18.677
	ARLINGTON HEIGHTS	18.035
	PARK RIDGE	17.606
	ROLLING MEADOWS	17.473
	BUFFALO GROVE	16.581
	CHICAGO	16.141
	OAK PARK	15.848
	ST CHARLES	15.827
	CHICAGO (Partial)	15.670
	SCHAUMBURG	15.461
	DES PLAINES	15.320
	ADDISON	15.123
	CHICAGO (Partial)	13.359
	CHICAGO (Partial)	12.314
	MATTESON	11.776
	AURORA	11.563

	OAK LAWN	10.735
	CHICAGO (Partial)	10.708
3	NAPERVILLE (Partial)	19.813
	NAPERVILLE (Partial)	19.616
	WILLOWBROOK	19.424
	PALATINE	18.586
	LINCOLNWOOD	18.368
	CHICAGO	18.142
	PARK RIDGE	17.441
	MORTON GROVE (Partial)	17.122
	DOWNERS GROVE	17.117
	MORTON GROVE (Partial)	16.562
	BLOOMINGDALE	15.832
	WHEELING	15.415
	PALOS HEIGHTS	14.910
	CRYSTAL LAKE	14.835
	HOFFMAN ESTATES	14.480
	HANOVER PARK	12.810
	NORRIDGE	12.629
	BOLINGBROOK	11.374
	BRIDGEVIEW	11.238
	MELROSE PARK	10.979
4	WHEATON	19.138
	RIVER FOREST	17.755
	ORLAND PARK	16.573
	OAKBROOK TERRACE	16.341
	MOUNT PROSPECT	16.333
	SCHAUMBURG	15.327
	HICKORY HILLS	13.941
	EAST DUNDEE	13.621
	HANOVER PARK	13.051
	CHICAGO (Partial)	12.985
	VILLA PARK	12.436
	OAK LAWN	12.330
	RIVER GROVE	11.959
	HAZEL CREST	11.488
	CHICAGO (Partial)	11.374
	MONTGOMERY	11.160