Where Is the Next Whole Foods in Los Angeles County?

Study on factors affecting the local presence of Whole Foods' Market using the logistic regression

Econ 584 Economic Consulting and Applied Econometrics

Group 9: Amanda Rago, Mohammed Alshlash, Micky Sun,

Wanlin Chen, Jianian Hua, Peiyao Li

Final Project

Apr 26, 2023

Executive Summary

Whole Foods is a multinational supermarket chain in the United States that focuses on selling grocery products that are mostly organic, free of preservatives, artificial colors, and flavors. In the Los Angeles county area, there are only 29 stores. With this study, we set out to determine what factors are seemingly significant to Whole Foods when they decide to open a new store, and which locations would be a great option for a new Whole Foods. The question we aim to answer is: where should Whole Foods open next, and why?

To answer this question, we performed a logistic regression analysis. A dummy variable for whether a neighborhood has a Whole Foods was the dependent variable. Initially we included many independent variables including demographic information, number of competitors, crime, and apartment rent prices. We aggregated our data at the zip code level, but model performance improved when we pivoted to a city level aggregation instead. Thus, our model consisted of 116 observations. The significant variables from this model were used to create the model that would forecast new locations for Whole Foods.

In the model for predictions, the variables included are the logarithm of the population density, the number of Ralph's in a neighborhood, the number of Trader Joe's in a neighborhood, the logarithm of the number of violent crimes in a neighborhood, and the percent of residents in a neighborhood that had a bachelor's degree or higher. Our results from this regression showed that the population density, number of Ralphs, number of violent crimes, and bachelor's degree education were positively correlated with having a Whole Foods, and only Trader Joe's was negatively correlated with the existence of Whole Foods. The results from our forecast indicate that a Whole Foods should open in the Pacific Palisades and Culver City.

Background

In the Los Angeles county area, there are 29 Whole Foods. It's safe to say that much of the population in our study area focuses on things like "clean eating", "artificial free", and "organic", among other diets. Whole Foods similarly places a lot of emphasis on maintaining a large selection of these products that meet those dietary needs. Thus, we concluded that there could be more Whole Foods in the area, but we were left with the question of determining which location. The answer to this is inherently reliant on a multitude of factors, which we set out to determine to therefore make the most informed prediction of where the next Whole Foods should open in the Los Angeles area.

This study looks into the presence of Whole Foods Market in Los Angeles County as displayed in Figure 1 in the Appendix. We utilized the data on demographics of the varying neighborhoods including the population density, education level (the proportion of population having bachelor's degree or higher), and the number of violent crimes. Also, we looked at the number of competing grocery stores like Trader Joe's and Ralph's, and included other variables on the various socioeconomic conditions of a city. With logistic regression models, the study is able to understand important factors on the existence of a Whole Foods Market in the neighborhood, and use the results to build a prediction model to suggest where future Whole Foods should open.

Literature Review

Rincón et al. (2020) explored the use of demand metrics in selecting suitable locations for a supermarket. They included demographic, economic, and behavioral data to develop the demand model. Rincón et al. (2020) identified the possible locations for the supermarket, and

determined the optimal number of stores within a specific market area in his research. Both studies shared the importance of incorporating local market knowledge and stakeholder input in the site selection process. In conclusion, the study highlighted the importance of utilizing data-driven insights to inform business decisions in launching supermarkets under a competitive environment.

Smith and Sanchez (2003) examined the methods used by a supermarket chain to evaluate the possible sites for retailing. They analyzed data from over 1,000 possible sites for the retail stores to identify factors that impact store performance: demographics, competition, and characteristics of the retail stores. They found sales per unit area, market penetration, population size, household income, and ethnicity are important factors. The study reinforces the importance of using data analytics to provide insights on business decisions as well as the need for continuing evaluation of store performance to optimize operations. Generally speaking, the article provided valuable insights into the detailed process of assessing potential sites and reinforced the importance of using data-driven analysis in decision-making.

Baviera et al. (2016) explored the use of geomarketing models in determining the most suitable locations for supermarkets to earn profits. They discussed how these models incorporate data from various resources:demographic and socioeconomic data to analyze the potential of the market, identifying the best locations for stores. They also discussed the importance of competition, accessibility, and governmental regulations in the process of selecting retailing sites.

Amparo et al. (2016) shared opinions upon the economic factors which could impact the selecting location of supermarkets and grocery stores. To examine several factors of how grocery stores succeed, the authors explored the details of store location, consumer behavior, and market

competition. They also utilized statistical factors such as population density, income levels, and proximity to competitors which can possibly impact the success of a store in a particular location. Urbanization and the change in demographics are also very important factors for the success of grocery stores in choosing location. The authors argued that understanding these economic factors is essential for retailers to make informed decisions about store location and to remain competitive in a rapidly evolving market.

This study would test the factors in the literature in the Los Angeles County area and complement the previous study with more affecting factors and the location strategy.

Data Description

1. Zip code level

This project utilized cross-sectional data collected at the zip code level in Los Angeles County. The county comprises 528 zip codes, numbered from 90001 to 93599. However, some of these zip codes correspond to unpopulated land or areas with limited social development. Following data cleaning, we were left with 264 zip codes that contained valid data for our analysis. To improve our analysis, we further aggregated the zip codes to the city/neighborhood level for our model.

2. City/Neighborhood level

Table 1 is the data description of all variables used in this project. We aggregated zip code data based on the supervisorial district of Los Angeles County (County of Los Angeles, 2023) to identify major cities and neighborhoods in the area. Eventually, 116 observations were obtained. To be specific, the dependent variable is the Whole Foods dummy variable, which indicates whether a city or neighborhood has a Whole Foods store or not. We also collected 12

independent variables, which we grouped into four categories: competitors level, demographic information, socioeconomic condition, and community safety and accessibility.

In terms of competitors, Whole Foods and Trader Joe's both offer a unique shopping experience that caters to health-conscious consumers. Ralphs and Trader Joe's both offer lower prices than Whole Foods, which appeal to budget-conscious consumers. We considered Trader Joe's and Ralphs as major competitors of Whole Foods, based on their pricing and target consumers. There are 48 Trader Joe's and 71 Ralphs in Los Angeles County, compared to 29 Whole Foods stores. Trader Joe's has more stores in coastal cities such as Santa Monica and Long Beach, while Ralphs tends to open stores in the northern part of the county, such as in Burbank, Glendale, and Pasadena.

Regarding demographic information, we examined population size, median age, and education level in each city/neighborhood. Population is a significant factor to consider when selecting a site for a supermarket, as the size and composition of the local population can greatly impact the demand for Whole Foods Market. A larger population in a specific area typically indicates a higher demand for supermarkets, which makes population an important variable to take into account when analyzing site selection for Whole Foods Market. Age is also an important indicator to consider when analyzing the demand for Whole Foods Market. We believe younger generations tend to prioritize health-focused and organic food options, while older generations may prefer more traditional products. According to Kashino (2015), Whole Foods primed to attract well-educated people, so we focused on the percentage of the population with a bachelor's degree or higher. Through spatial analysis, the City of Los Angeles and Long Beach have the largest population among all cities in Los Angeles County. In addition, coastal cities

such as Pacific Palisades, Manhattan Beach, and Hermosa Beach have a higher percentage of residents with bachelor's degrees or higher.

Within the socioeconomic condition category, we collected and analyzed variables including unemployment rate, household median income, and average housing prices and average apartment rents.Unemployment is an important indicator of the local economy, as it can affect the potential customer base for the supermarket. Our hypothesis was that housing prices reflect the purchasing power of individual households and that higher average housing prices would indicate a greater willingness among homeowners to shop at Whole Foods Market. Similarly, household income is often a measure of neighborhood affluence, and we expected that higher-income households would be more likely to shop at Whole Foods Market on a regular basis. The apartment rent variable, like the housing price variable, can also indicate the spending power of the residents in the area. Spatial analysis showed that coastal cities, such as Pacific Palisades, Manhattan Beach, and Malibu, generally have higher household median income compared to inland cities. Housing prices in Pacific Palisades, Manhattan Beach, and Malibu are also more expensive than other cities/neighborhoods. Beverly Hills and West Hollywood have particularly high housing prices.

Finally, under the community safety and accessibility group, we analyzed violent crimes, traffic volumes, and public transit ridership. Crime rate is an important factor to consider as it can impact the safety and security of both customers and employees at the supermarket. More importantly, Whole Foods prefers to open stores in areas with easy access to major roads (Kashino, 2015). Violent crimes are clustered in the northwestern part of Los Angeles County, such as in Van Nuys, North Hollywood, and Canoga Park. Public transit ridership shows

clustering patterns in the northwestern and southern part of Los Angeles County, including Van Nuys, North Hollywood, Huntington Park, and the City of Los Angeles.

Table 1. Data description of all variables.

Variable	Description (Source)	Min	Mean	StDev	Max
Dependent Variable					
Whole Foods	Whether a city/neighborhood has a Whole Foods or not. (Google Map, 2023)	0	0.25	0.98	9
Competitors Level					
Ralphs	Number of Ralphs in a city/neighborhood. (Google Map, 2023)	0	0.61	2.15	22
Trader Joes	Number of Trader Joes in a city/neighborhood. (Google Map, 2023)	0	0.41	1.29	12
Demographic Information					
Population	Total population in a city/neighborhood. (US Census Bureau, 2021)	5,114	85,579	226,351	2,420,705
Age	Median age in a city/neighborhood. (US Census Bureau, 2021)	31	40	4	51
Education	Percentage of population with bachelor's degrees or higher. (US Census Bureau, 2020)	4.20%	35.09%	19.71%	78.54%
Socioeconomic Condition					
Households Income	Households median income in a city/neighborhood. (US Census Bureau, 2021)	$49,554	$93,458	$31,766	$212,115
Unemployment Rate	Unemployment rate in a city/neighborhood.	2.85	6.48	1.47	9.92

	(US Census Bureau, 2021)				
Housing Price	Average housing price in a city/neighborhood. (Los Angeles Almanac, 2023)	$89,705	$1,175,492	$769,734	$4,623,281
Apartment Rent	Average apartment rents in a city/neighborhood. (Los Angeles Almanac, 2023)	$1,942	$2,439	$439	$3,244
Community Safety and Accessibility					
Violent Crimes	Number of violent crimes in a city/neighborhood. (Los Angeles Almanac, 2021)	14	28,838	151,968	1,632,191
Traffic Volume	Average annual daily traffic of freeway/highway segments in a city/neighborhood. (Los Angeles Almanac, 2016)	28,000	2,157,306	8,037,348	86,666,500
Public Transit Ridership	Percentage of population using public transit service in a city/neighborhood. (Neighborhood Data for Social Change, 2019)	0.40%	3.71%	2.68%	13%

Method

The method employed to answer our research questions in this paper relies mainly on the logistic regression model (logit model) which estimates the probability and odds of occurring event or taking a Yes/No decision based on the effect of the set of independent variables included in the model as in the following equation:

$$g(E[Y]) = \alpha + \beta X \Rightarrow ln(\frac{p}{1-p}) = \alpha + \beta X \quad (eq.1)$$

Where g(E[Y]) represents the associated odds ratio of the dependent variable and is the estimated coefficients of the included set of independent variables X.

To answer the first research question, what influences the decision to open a new Whole Foods store in a new location, the study implemented the logit model where the dependent variable is an indicator of Whole Foods' presence in that specific area. For the independent variables, most of the associated variables that were expected to affect that decision or a log transformation of the variable were included in the model. Table 1 provides a description of the independent variable used in the logit model.

After answering the first question, we used the obtained results and built another logit model where only the significant factors are included to answer the second part of our research question: Where should Whole Foods open a new store or close an existing one?

$$Whole\ Foods = + \beta_1\ log(Total\ Population) + \beta_2\ Ralphs + \beta_3\ Trader\ Joes + \beta_4\ log(violent\ crime) + \beta_5\ Education \quad (eq.2)$$

The study implemented a performance check before concluding the results to ensure the accuracy of the model with out-of-sample data where the total observation has been randomly splitted into a training data set and a testing data set. Training data comprised 80% of observations (92 observations) while testing data comprised 20% of observations (24 observations) as shown in figure 2. The performance of the model will be discussed in the results section.

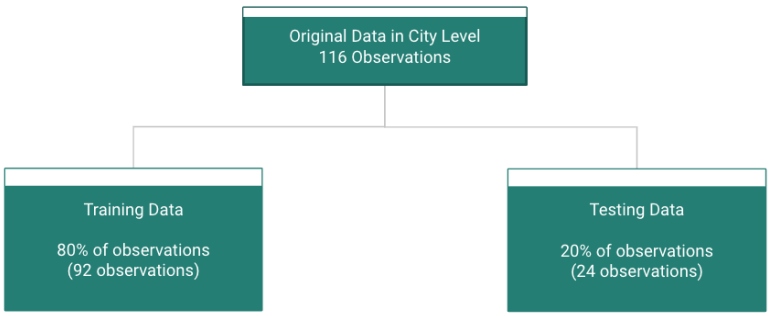


Figure 2. Method workflow

Finally, the logit model in equation 2 has been implemented and we concluded the research by providing insightful recommendations on the potential locations for a new Whole Foods store and the locations that might be underperforming locations and should be closed depending on the sales data.

Results

Table 2 is the results of our initial model on zip code level. The root mean square error (RMSE) is 0.26. And R square is low with only 0.345. There were only four significant variables which are log of number of households, four years population growth, number of ralphs and log of average housing price. The non-ideal performance is probably because of the non symmetrical distribution of some of the variables in the collected data So we aggregated the collected data into a neighborhood/city level.

Table 2. Regression results on zip code level

Logit Regression Results

	Dependent variable:
	Whole Foods presence
const	-61.7317
	(-24.23)
Ralphs	0.6993*
	(0.421)
Trader Joes	-0.2578
	(0.524)
Bristol Farms	0.4608
	(0.879)
Log(Total population)	-2.4703
	(1.812)
Log(Median Age)	0.6537
	(2.978)
Log(Number of Households)	4.0652**
	(1.947)
Log(Households Median income)	0.5616
	(1.58)
Unemployment Rate	-0.1689
	(0.178)
Log(Violent Crime)	0.1192
	(0.14)
Log(Average Apartment Rent)	2.7122
	(3.144)
Log(Average Housing Price)	1.1887*
	(0.707)
Four Years Population Growth	0.0939**
	(0.046)
Four Years Income Growth	-0.0046
	(0.027)
Pseudo R-squared	0.3454
Observations	263
Log Likelihood	-56.969
Note	*P<0.1; **P<0.05; ***P<0.01

Table 3 shows our model results on the neighborhood/city level. The RMSE is 0.19 which is lower than the model on zip code level, indicating higher accuracy level and the obtained R square is 0.697 which is almost twice as much as the previous model. This model can explain 69.7% of the variance in the log odds of Whole Foods. There are 5 significant variables in this model which will be used in a new model to answer the second question of our research as explained in the method section above.

Table 3. Regression results on neighborhood level

Logit Regression Results

	Dependent variable:
	Whole Foods presence
const	-22.917
	(74.811)
ralphs	2.296*
	(1.21)
Trader Joes	-2.5036*
	(1.318)
Bristol Farms	0.9176
	(2.195)
Log(Total population)	3.0199**
	(1.543)
Log(Median age)	-12.3676
	(10.841)
Log(Households Median income)	-4.3858
	(5.352)
Unemployment Rate	-0.0772
	(0.604)
Log(Violent Crime)	0.5955*
	(0.321)
Log(Average apt price)	6.7631
	(7.654)
Log(Average housing price)	1.0547
	(2.544)
Education	0.2536*
	(0.13)
Public transportation traffic	-0.5517
	(0.511)
Total traffic volume	5.93E-07
	(4.48E-07)
Pseudo R-squared	0.6971
Observations	116
Log Likelihood	-13.53
Note	*P<0.1; **P<0.05; ***P<0.01

Table 4 is the regression results of the model that only includes the significant variables and splitting the data into training data and testing data. The obtained R square is 0.65 which is slightly lower than the previous model due to lower number of observations compared to the original model. This model has excellent accuracy as the RMSE is 0.28. Trader Joe's is the only variable that is negatively associated with the odds ratio of Whole Foods presence. On average, one more Trader Joes' store is associated with an 88% decrease in the odds of whole foods presence, which indicates that Trader joes is a competitor of whole foods. However, Ralphs has a positive effect on Whole Foods which might be because they have a similar site selection strategy and Trader Joes may not be a strong competitor of Whole Foods due to the demographic

differences in targeted customers between the two. The violent crime has a slightly positive impact, which is surprising but it might be because a neighborhood with a higher total population tends to have more violent crimes and that might be fixed by using crime rates instead of the total number of crimes.

The effect of total population and education rate on the probability of Whole Foods Market's presence is aligned with our expectation. 1% increase of population is associated with a 3.92% increase of the odds ratio. Finally, a one unit increase in the education rates will multiply the odds ratio of Whole Foods presence by 1.26.

Table 4. Regression results on training data

Logit Regression Results	
	Dependent variable:
	Whole Foods presence
const	-59.1578
	(22.048)
Log(Total population)	3.9247**
	(1.683)
Ralphs	1.7822*
	(1.00)
Trader Joes	-2.0878*
	(1.133)
Log(Violent Crime)	0.4249*
	(0.219)
Education	0.2327***
	(0.083)
R-squared	0.6566
Observations	116
Log Likelihood	-13.53
Note	*P<0.1; **P<0.05; ***P<0.01

Figure 3 shows the evaluation on the model performance with out of sample data (testing data). The true positive rate (TPR) is 0.67. The true negative rate (TNR) is 0.90. The TPR is not very high, possibly because our testing data is small. And we can see the overall accuracy level of this model is also high with 0.88.

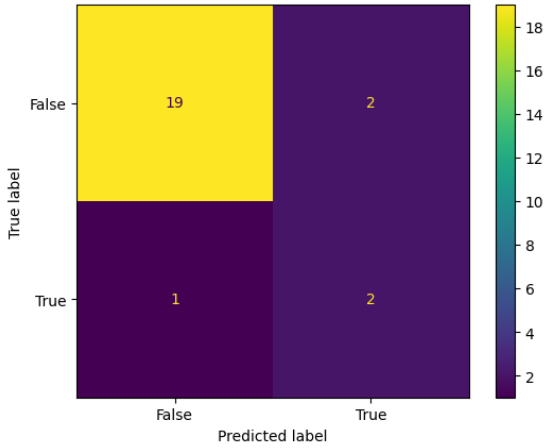


Figure 3. Prediction evaluation on testing data

Implication

This study also uses the best prediction model to compare location strategies of different grocery stores. We regressed a dummy dependent variable of whether there are stores in the city/neighborhood area on variables in the best prediction model and got the results of Whole Foods, Trader Joe's and Ralphs (Table 5).

Table 5. Strategy Comparison

	Whole Foods	Trader Joe's	Ralphs
Constant	-45.4359 (.002)	-36.7858 (.001)	-11.9961 (.052)
Log of Total Population	2.6472 (.013)	2.6540 (.003)	0.8609 (.096)
Log of Violent Crime Number	0.5823 (.003)	0.0012 (.993)	-0.0013 (.990)
Percentage of Bachelor's Degree or Higher	0.2028 (.001)	0.1492 (.000)	0.0314 (.151)

Number of Whole Foods		-.1.2180 (.128)	1.5884 (.073)
Number of Trader Joe's	-1.4932 (.058)		1.3735 (.024)
Number of Ralphs	1.5521 (.051)	0.9473 (.035)	

The total population has been an important factor for Whole Foods and Trader Joe's location strategy. It is not as important for Ralphs. The number of violent crimes has a positive and very significant relationship with Whole Foods market allocation while it has not much relationship with Trader Joe's and Ralphs. The positive relationship is counterintuitive and is probably because Whole Foods likes to locate in busy downtowns which have more violent crimes. The education level of the population is a significant factor for both WholeFoods and Trader Joe's while Whole Foods market values it more. One percentage higher in the proportion of the population having bachelor degrees is associated with 20% higher probability of WholeFoods' presence. This number is 15% for Trader Joe's. As for the effect of competitors, trader joes does not mind that much as Whole Foods when choosing locations. The coefficient of the number of Whole Foods is negative but not significant. Ralphs does not take the number of Trader Joe's or Whole Foods as a negative impact at all.

To provide some insights on which city/neighborhood area should Whole Foods Market expand, the study fits the prediction model with all the data in Los Angeles county. The study constructs an underserve level variable by subtracting the real Whole Foods dummy by the predicted probability. The lower the value, the more underserved this area is. In the visualization map, Pacific Palisades and Culver City have the lowest value saying that Whole Foods market should consider opening chains there (Figure 4). El Segundo and Venice are of the deepest blue, meaning Whole Foods should consider quitting the market there.

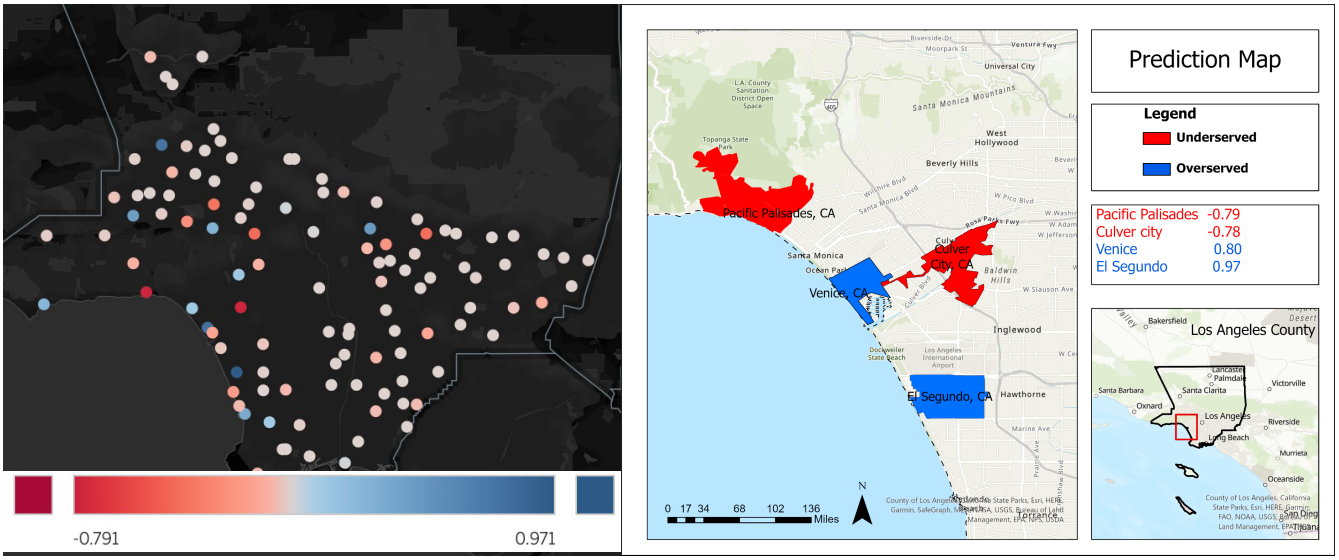


Figure 4. Prediction map

Conclusion

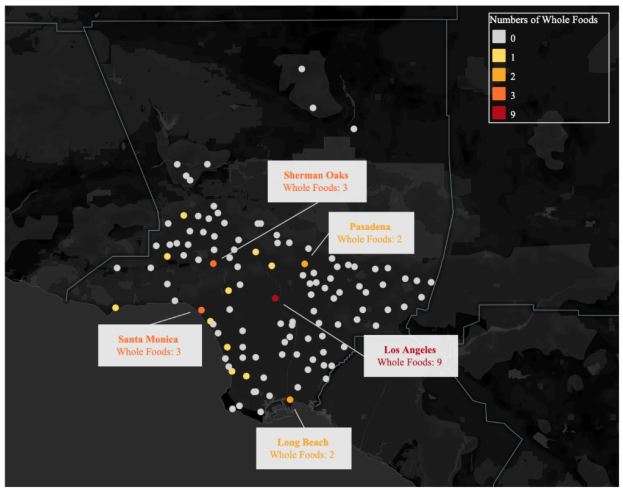
This study investigates the location strategy of the Whole Foods Market, a retail grocery store, in Los Angeles County. It finds that the total population of the city/neighborhood area, violent crime number, residents' education level, and the number of Trader Joe's are important factors to Whole Foods Markets' location. Other demographic factors including the median age, household median income, unemployment rate as well as the housing price are not as significant as hypothesized.

Compared to other grocery stores, Whole Foods Market's location strategy is unique. Trader Joes only take the population and the education level into consideration while none of the factors involved have significant impacts on Ralphs' location. The prediction model using only significant variables shows 88% prediction accuracy rate. The model suggests that Whole Foods Market should consider entering Pacific Palisades and Culver City while Venice and El Segundo are the most overserved.

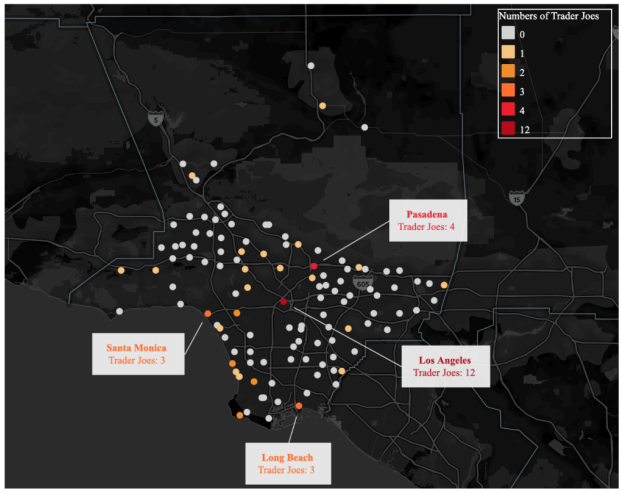
The research is limited by the accessibility of the data as it includes only 116 major neighborhoods and cities in Los Angeles County. Further studies could increase the robustness of the analysis by integrating more variables like more grocery brands and commercial rent data or by expanding the study area to the West Coast.

Appendix

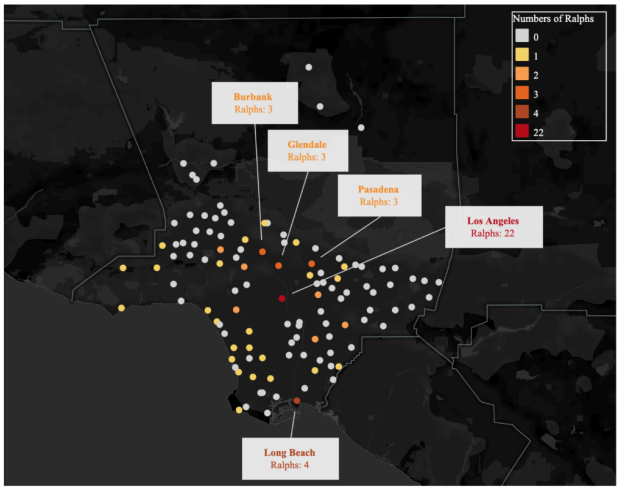
Distribution of Whole Foods



Distribution of Trader Joes

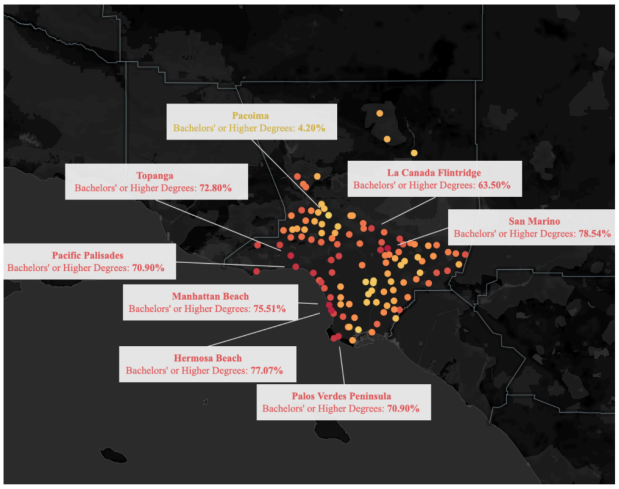


Distribution of Ralphs

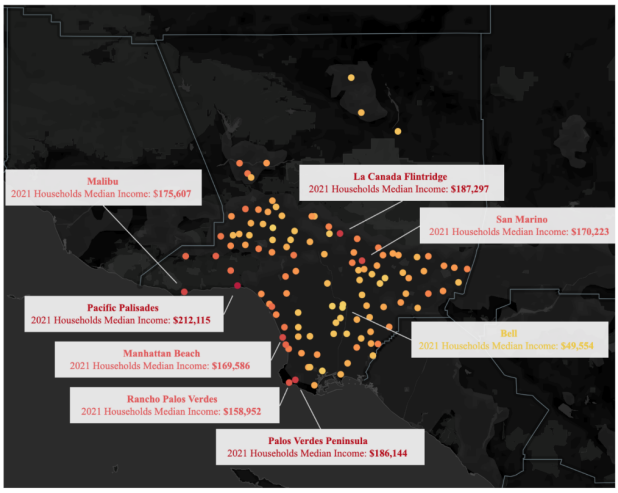


Distribution of Population

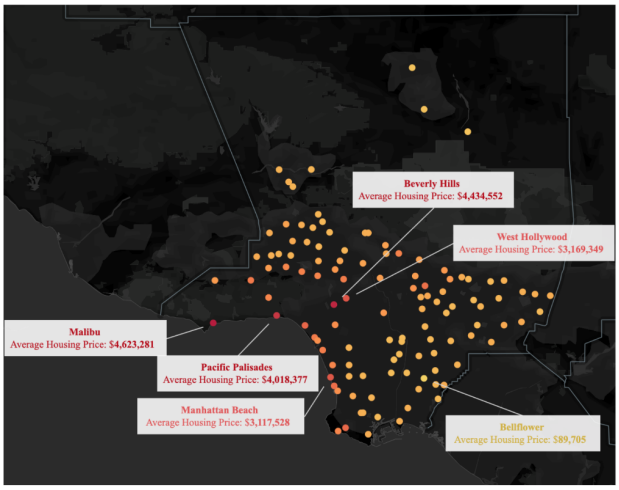
Distribution of Bachelor's Degrees or Higher



Distribution of Household Median Income



Distribution of Average Housing Price



Distribution of Violent Crimes

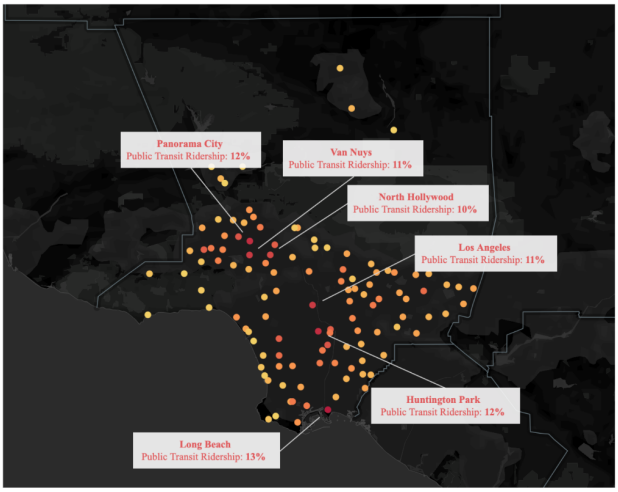


Figure 1. Maps of variables

References

County of Los Angeles. (2023). *Zip Codes by Supervisorial District*. Lacounty.gov. https://file.lacounty.gov/SDSInter/lac/1031552_MasterZipCodes.pdf

Kashino, M. (2015, July 14). *How Whole Foods Decides If Your Neighborhood Is Worthy | Washingtonian (DC)*. Washingtonian. https://www.washingtonian.com/2015/07/14/how-whole-foods-decides-if-your-neighborhood-is-worthy/

Rincón, & Tiwari, C. (2020). *Demand Metric for Supermarket Site Selection: A Case Study*. Papers in Applied Geography, 6(1), 19–34. https://doi.org/10.1080/23754931.2020.1712555

Assessment of business potential at retail sites: empirical findings from a US supermarket chain. Retrieved April 19, 2023, from https://www-tandfonline-com.libproxy2.usc.edu/doi/abs/10.1080/0959396032000051684

Smith, & Sanchez, S. (2003). Assessment of business potential at retail sites: empirical findings from a US supermarket chain. The International Review of Retail, Distribution and Consumer Research, 13(1), 37–58. https://doi.org/10.1080/0959396032000051684

The economics of supermarket and grocery store location - USDA. (n.d.). Retrieved April 19, 2023, from https://www.ers.usda.gov/webdocs/publications/42711/12705_ap036f_1_.pdf?v=0

Geomarketing models in supermarket location strategies - researchgate. (n.d.). Retrieved April 19, 2023, from

https://www.researchgate.net/publication/311809703_Geomarketing_models_in_supermarket_location_strategies