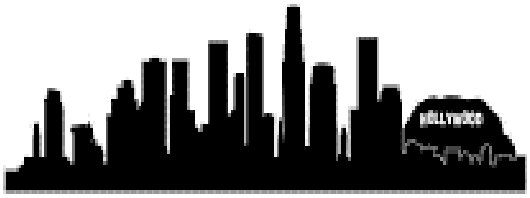
FOOD DESERTS BY ZIPCODE

How Big is the Problem in Three Large U.S. Cities?



Los Angeles

Chicago

New York

FOOD DESERT ANALYSIS BY ZIP CODE FOR 3 MAJOR U.S. CITIES

Michael V. Nugent

September 25, 2020

1. Introduction

1.1. Background

Healthful diets have a significant positive effect on reducing the occurrence of heart disease, diabetes, and even cancer. Yet, healthful diets are dependent on access to healthy food, which is generally available in supermarkets and generally not available in alternative stores like convenient stores. If the grocery stores are too few and far between, an area called a food desert, the purchase of healthy food is confounded.

The food desert concept was first discussed in the 1990s. The metaphor is built on a desert paradigm. The wandering person is in the desert and looking for water at an oasis. If the person cannot find the desert, the person's chance of survival are limited.

In the food desert, the wandering person is in the neighborhood (the desert) looking for healthy food at a supermarket (the oasis). If the person cannot find the supermarket, his alternatives for healthy food purchases and consumption are limited.

Research has shown that when food deserts are found, either convenient stores or fast food chains will provide nutritional support, both of which marginalize nutrient consumption.

1.2. Problem

The purpose of this research is to determine if a simple zip code analysis using Foursquare information can describe this situation in three large cities in the United States – Los Angeles, Chicago, and New York. Unlike other research, this is a comparative study looking to compare high-income areas' store counts with low-income areas' store counts to gauge the disparity rather than determining an arbitrary value for the food desert definition.

1.3 Interest

If the analysis shows food deserts still exist in some zip codes, then city, municipal, county, state, and federal agencies can use this information to set policy. These policies could be transportation based - to improve access to distant stores, incentive based – incenting new store construction, or even price based – promoting better for you foods when additional stores are not possible.

2. Data Section

2.1. Data Sources

There were 4 major sources of data used for this analysis. They are described below.

2.1.1. Income by Zip Code

This analysis will be using the 2017 income tax file available at "*www.irs.gov/statistics/soi-tax-stats-individual-income-tax-statistics-2017-zip-code-data-soi*". The IRS data contains information by 6 major income groups for each zip code and indicates the individual line item information on U.S. Income Tax forms 1040, 1040A, 1040EZ. The information will be

used to create an average income by zip code for correlation and comparative purposes. The Data Dictionary for this file can be downloaded "*www.irs.gov/pub/irs-soi/17zpdoc.docx*". Some rows from the Data Dictionary are below:

VARIABLE NAME	DESCRIPTION	VALUE/LINE REFERENCE	Type
STATEFIPS	The State Federal Information Processing System (FIPS) code	01-56	Char
STATE	The State associated with the ZIP code	Two-digit State abbreviation code	Char
ZIPCODE	5-digit Zip code		Char
AGI_STUB	Size of adjusted gross income	1 = $1 under $25,000 2 = $25,000 under $50,000 3 = $50,000 under $75,000 4 = $75,000 under $100,000 5 = $100,000 under $200,000 6 = $200,000 or more	Num
N1	Number of returns		Num
MARS1	Number of single returns	Filing status is single	Num
MARS2	Number of joint returns	Filing status is married filing jointly	Num

2.1.2. Zip Codes in target cities

To retrieve zip codes for each city, this analysis will be using information from for "*www.zip-codes.com*" using the city names as keys for the data. The data will be scraped from the web site and refined as needed. A sample from the web site appears below:

2.1.3. Grocery Stores near a Zip Code

To identify grocery stores by Zip code, FourSquare® data will be used utilizing the following *categoryId* keys:

grocery_store = '4bf58dd8d48988d118951735'
organic_grocery = '52f2ab2ebcbc57f1066b8b45'
supermarket = '52f2ab2ebcbc57f1066b8b46'

2.1.4. 2010 Census by Zip Codes

To identify population used in calculating per capita income, a file developed by Splitwise will be used. It can be reviewed at "*https://blog.splitwise.com/2013/09/18/the-2010-us-census-population-by-zip-code-totally-free/".*

A sample of the data rows is below:

Zip Code ZCTA	2010 Census Population
01001	16769
01002	29049
01003	10372
01005	5079
01007	14649
01008	1263
01009	741
01010	3609

2.2. Data Cleaning

2.2.1. FourSquare Venue Clean Up

The use of the category_ID field in the FourSquare data retrieval created extra rows that were not grocery stores. These included food service arenas like coffee shops and restaurants. The data were further parsed looking for *Grocery Store, Supermarket,* and *Organic Grocery* in the "Venue Category" field. In essence, the category_ID field and the Venue Category were not synchronized.

2.2.2. Zip Code Data

Zip Codes retrieved for LA, NY, and CHG did not always return correct results. Zip codes were selected when city name matched the selected city.

In LA, the UCLA campus is considered Los Angeles, but is well away from the rest of the city's zip codes and was removed.

In New York, the bureau names themselves were used to retrieve the zip codes for New York rather than "New York".

2.3. Data Limitations

Census data is from 2010, income tax data is from 2017, and grocery store data is from 2020.

2.3.1. 2010 Census Data v. 2017 Tax Information

The information for populations comes from the 2010 census. The Census Bureau has estimated population changes from 2010 to 2019 for New York, Los Angeles, and Chicago as 1.98%, 4.93%, and -0.06%. (See data tables found at "*https://www.census.gov/data/tables/time-series/demo/popest/2010s-total-cities-and-towns.html#tables".)*

The analysis for Los Angeles could be impacted by this increase. There may be growth in population in a zip code causing that area's average income to be misstated. This would cause an area to appear to be more or less affluent than it is actually is.

2.3.2. Census Bureau ZCTA v. USPS Zip Code Data

The Census Bureau uses Zip Code Tabulation Areas (ZCTA) that are slightly different in size from the USPS Zip Codes that are used with the other data files. Because this analysis uses the center of a Zip Code area to determine distance, it is not believed the differences in boundaries will vary the information significantly.

(A general discussion about these differences can be found at "*http://gis.washington.edu/phurvitz/zip_or_zcta/index.html*".)

3. Methodology

3.1. Exploratory Data Analysis

Determining income level breaks was conducted at 3 different levels using the qcut feature of pandas library. Using smaller quantiles accumulated the data in too large of baskets yielding cloudy results. As the quantiles were increased, the same issues occurred until the data were divided into 20% segments or 5 quantiles. The distribution of income is unevenly distributed and therefore, the quantiles used were more about separating the lower income values than accommodating the higher income amounts.

3.2. Programming Approach

Due to retrieval restrictions for screen scraping of Zipcode.com and FourSquare, files were created once and stored as a local csv file. The data were then loaded for subsequent runs. This is noted in the code. These data were significantly static during the testing and greatly decreased program development time.

Income levels by zip code were determined by using the Average Gross Income line from the IRS tax data for each zip code. The census population was used to determine income by individuals in the area. This allowed family size to influence the calculation of the gross income as it accounts for additional expense associated with the support of dependents. In essence, a family of 2 making $35,000 is better off than a family of 5 making the same amount.

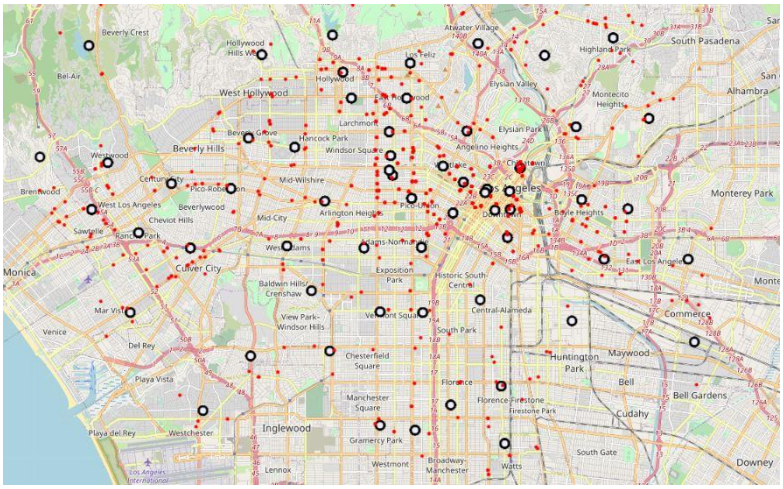
Distance to grocery stores was considered as 1 mile. The distance measurements were from the longitude and latitude of the zip code center to the longitude and latitude of the grocery store.

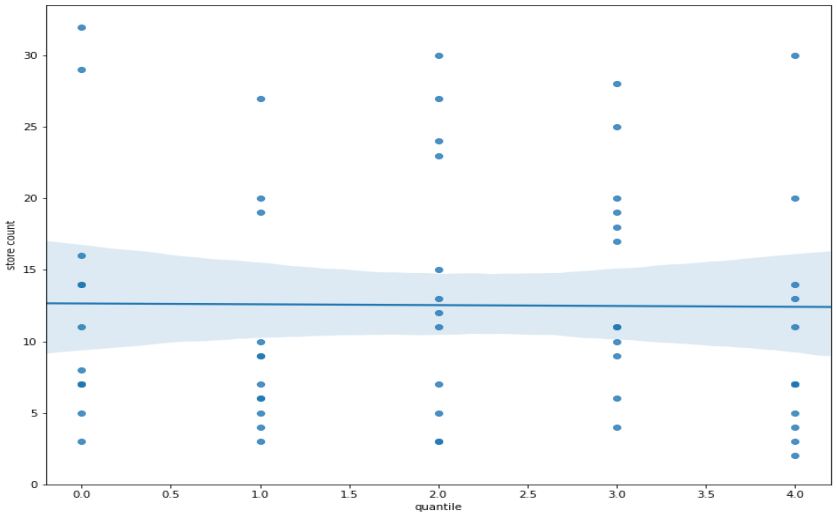
4. Results

4.1. Los Angeles results

Population Density is fairly low in LA at 6,900 per mile^2. This is revealed in the mapping of the grocery stores for each zip code area. Without significant concentration of population, a concentration of stores would be unsound economically.

NNot surprisingly then, and relative to the other cities, the concentration of grocery stores is low. Even in the highest income level, access to stores is not particularly easy. Given the distribution across all income quantiles, the data struggle to characterize the lowest income quantiles as food deserts.

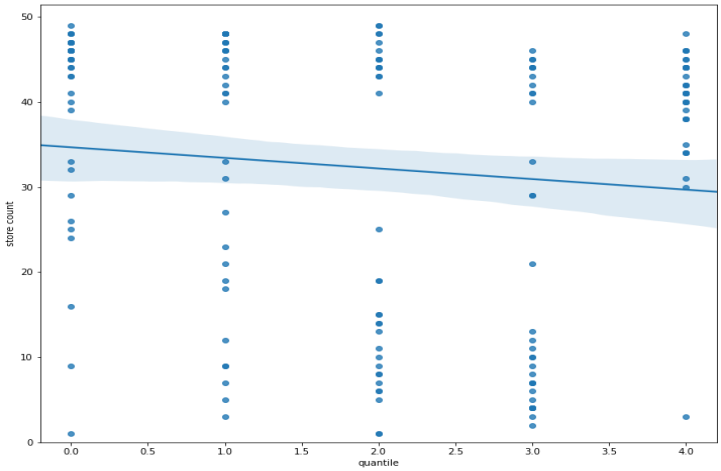




4.2. New York results

Population Density is highest in NY at 27,000 per mile^2. The number of grocery stores shown in the map is plentiful in certain parts of the city such as Manhattan that is both significantly populated and affluent. There are zip code areas with limited grocery stores such as on the peninsulas. With this level of density, the income quantiles become more muddled as both affluent and poor intersect side by side at times.

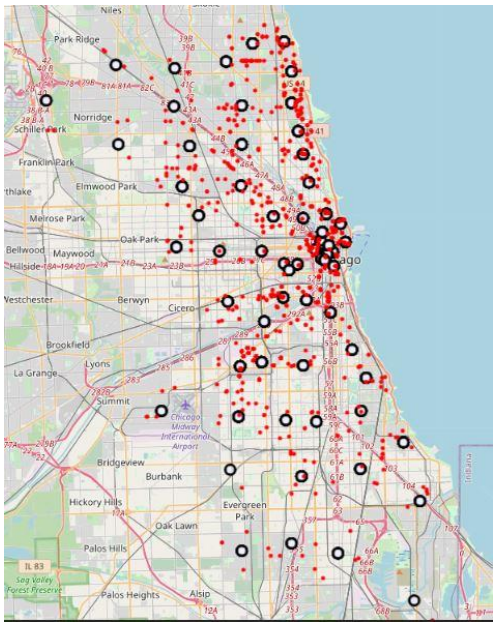
Grocery store concentration shows only a few low-income zip codes with accessibility problems. While some upper income quantiles also show low cross over, these people are likely living in high rent and therefore the densest areas. Only a few grocery stores would be needed to support them.

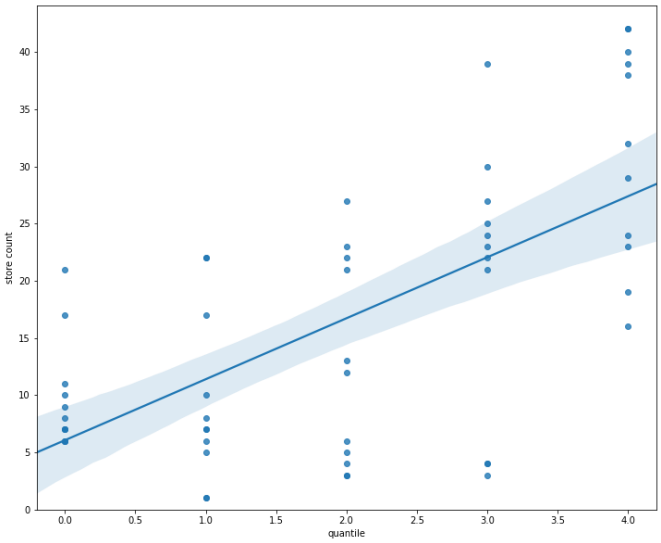


4.3. Chicago results

Chicago's population density falls between LA and NY at 11,900 per mile^2. The sparsity of some of the grocer stores becomes apparent even in this depiction with the south and west sides of Chicago, noted for their lower income, showing much fewer grocery stores.

Chicago shows the largest trend with food deserts appearing in the lowest income quantile and with significant access for all zip codes of the highest quantile. Chicago then would appear to have the most problems with food deserts relative to the low-income groups.





5. Discussion

Food Desert research has self-constrained to only look at income levels and grocery store availability. This research has clearly demonstrated that overall density may be an independent variable. Results from at least two cities demonstrated lack of access equally for both high income and low income quantiles. Additional research should be conducted to understand the general population density on the economics of grocery store construction which may operate independently of income levels.

6. Conclusions

The largest cities in the U.S. vary in their food-*desertness*.

New York has an inherent advantage to provide food access because of its significant population density, roughly twice that of Chicago and 4 times that of Los Angeles. This density likely provides economic surety for the construction of new grocery stores.

A city like Los Angeles is so spread out as shown with the lowest population density shows all income quantiles having equal lack of access to food grocery stores. However, keep in mind more affluent areas can more easily support grocery stores, so this equivalency in access is unusual. Further research of the data however showed several grocery chains that were specific to heritage such as Vietnamese and Hispanic chains. Further research should determine if ethnic concentrations foster ethnic grocery store construction displacing what would be food deserts.

Chicago has the highest correlation of grocery stores to income levels. There are clearly some food deserts and less affluent areas. Additionally, Chicago does not have the influx of ethnic groups driving a different economic. Further study should be conducted to determine what the impact is of ethnic change on the economics that support grocery store construction.

7. Python Program Web Link

https://dataplatform.cloud.ibm.com/analytics/notebooks/v2/c51db696-98c3-4274-ba45-9656d8d42938/view?access_token=adb0a2d797fad52f603a39c8738b32ed3775b24e1d0025034824bb4f13c9ba45