Analyzing Quality Grocery Options in Wake County, NC

By Jon Puryear

Author Note

This has been prepared as a part of the Coursera Applied Data Science Capstone

Business Problem

Wake County, NC is one of the fastest growing areas in the United States. However, there has been concern about access to grocery options in some parts of the county. Some parts of the county have been called a "Food Desert". a Food Desert is an urban area where it is difficult to buy affordable food or good-quality fresh food.

This analysis will look at the Grocery Options in Wake County, the Quality of the Grocery Options, and the Mean Incomes by Zip code to determine the following:

1) a Food Desert does exist

2) if said desert is tied to average income

3) Where a good place would be to open a grocery store that had the potential to both make money and provide food services to an under-served community

Interested Parties: This Analysis is targeted at investors looking to both start a profitable business in a location where there is demand for services and close a vital socio-economic gap.

Description of Data Used in this Analysis

The Data used in this analysis has been sourced directly from various gold data sources. Venue information from Foursquare has been used to determine the Quantity and Quality of Grocery Options for a given Zip Code. US IRS Tax Return Data has been used to determine mean income and number of households by Zip Code. Wake County Zip Codes has been used to determine the scope of Zip Codes to include in this analysis. Geospatial Data for Wake County Zip Codes has been sourced from the US Census Database.

Table 1 - Descriptions and Sources of Data Used

Data Source	Location	Key Data Points
Foursquare Places API	https://developer.foursquare.com/docs/places-api/	Venue Name, Location, Category
US Internal Revenue Service	https://www.irs.gov/statistics/soi-tax-stats-individual-income-tax-statistics-2017-zip-code-data-soi	Households and Income by Zip Code
Wake County Open Data	http://data-wake.opendata.arcgis.com/datasets/zip-codes/data	Zip Code Data in Wake County
US Census	https://www.census.gov/geographies/reference-files/time-series/geo/gazetteer-files.html	Geospatial Data for Zip Codes

Each of the data sources and datasets are described in detail below.

Foursquare Venue Information

Location and User Recommended Venue Data collected by Foursquare has been used to determine the Quantity and Quality of Grocery Options for a given Zip Code. This information was acquired by accessing the Foursquare places API. This API offers real-time access to Foursquare's global database of rich venue data and user content. Instructions for accessing the API and documentation of the results are available directly from the Foursquare developer portal: https://developer.foursquare.com/docs/places-api/. Relevant data points used in this analysis includes User Recommended Venues, Venue Names, Locations, and Type of Venue.

IRS Tax Return Data

Individual Income Tax Statistics - 2017 ZIP Code Data (SOI) has been used to identify the number of households present per Zip Code in scope as well as the Income of said households. This data is available from https://www.irs.gov/statistics/soi-tax-stats-individual-income-tax-statistics-2017-zip-code-data-soi in XLS format. Significant processing was necessary to extract the information needed for this analysis.

Wake County Zip Codes

Zip Codes that are wholly or partially in Wake County will be considered for this analysis. The listing of Zip Codes has been collected from the Official Wake County government data website located at http://data-wake.opendata.arcgis.com/datasets/zip-codes/data . This was provided as a CSV file and is being used as the core dataset used in this analysis.

Geographical Data for Zip Codes

The centroids of a given Zip Code has been sourced from the official website of the US Census. This is information is available to the public here: https://www.census.gov/geographies/reference-files/time-series/geo/gazetteer-files.html . The data points used are the latitude and longitude of the center point, or centroid of a given Zip Code.

Description of Methodology Used

The overall analysis combines the key data points from the data sources and combines them into a single dataset that is used to measure and observe the information. This model looks at the Quantity of Food Stores, recommended by foursquare users, in a given Zip Code and identifies if there is any correlation between population, income or the number of quality food stores. The basic analysis that was done using python and Jupyter Lab and consisted of two major steps:

1. Collect, Clean and Combine the base data
2. Calculate, Explore and Visualize Key Indicators

Collect, Clean, and Combine the base data

Data collection began with downloading the source datasets from Wake County Open Data, IRS, and the US Census. Extraneous data was removed and the IRS and US Census information was filtered and joined to the Wake County Dataset.

Each record in this combined dataset was passed through a custom python function which collected information from the foursquare API and appended that information into the combined dataset. The dataset was then filtered again to only include food stores such as Grocery Stores and Supermarkets. OneHot encoding was then used to convert the categorical Venue data into numeric data for ease of measurement.

This final dataset was used to calculate the Key Indicators used in the Analysis. The Data Points are:

- ZIPNAME: Name of the City that the Zip Code is in
- ZIPCODE: US Postal Zip Code
- LAT: Latitude of the ZIPCODE central point

- LON: Longitude of the ZIPCODE central point
- Neighborhood: ZIPNAME and ZIPCODE concatenated together for Display Purposes
- Grocery Store: Count of Grocery Stores recommended by Foursquare users within a 5 mile radius of a given ZIPCODE's central point
- Supermarket: Count of Supermarkets recommended by Foursquare users within a 5 mile radius of a given ZIPCODE's central point
- returns_count: Count of the number of households filing returns in a given zip code
- total_income: Sum of the Total Income reported to the IRS in dollars

Table 2: Top 5 Rows of the Cleaned and Combined Dataset

	ZIPNAME	ZIPCODE	LAT	LON	Neighborhood	Grocery Store	Supermarket	returns_count	total_income
0	RALEIGH	27601	35.773633	-78.634458	RALEIGH - 27601	1	0	4340	289082000
1	DURHAM	27713	35.895047	-78.923747	DURHAM - 27713	0	2	26260	198918000
2	RALEIGH	27604	35.822921	-78.562045	RALEIGH - 27604	4	0	21720	113071900
3	WAKE FOREST	27587	35.981382	-78.557584	WAKE FOREST - 27587	4	3	30920	299355600
4	DURHAM	27703	35.959272	-78.806808	DURHAM - 27703	0	2	27620	157682000

Calculate, Explore and Visualize Key Indicators

Calculate

Several Key Indicators were calculated based on the final dataset.

- food_store_count = sum of Grocery Store and Supermarket. This shows the total number of recommended Food Store options for a given Zip Code
- income_per_household = total_income divided by returns_count. This shows the average income per household for a given Zip Code
- households_served = returns_count divided by food_store_count. This shows the average number of households that each recommended Food Store services in each Zip Code

The intent of these measures is that we can determine if there is any correlation between income, population, and quality food stores, as well as identifying areas that might be underserved by quality food options.

Explore Datasets

Based on a description of the dataset, a given Zip Code has 3.87 Food Store Options, and Each store services ~5,200 Households on average.

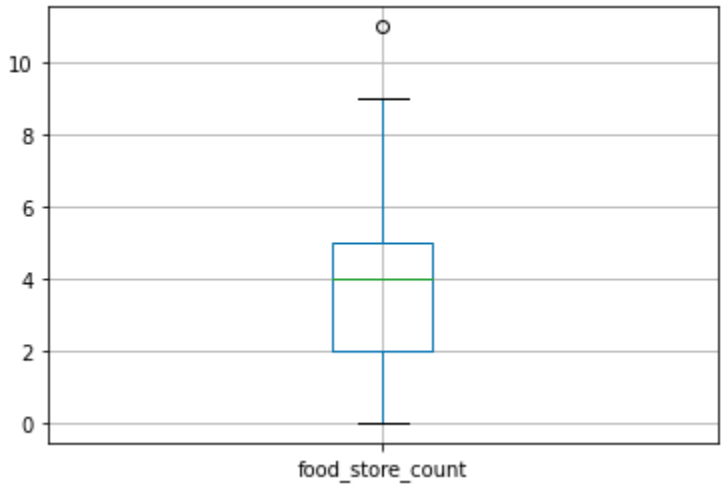
Table 3: Distribution of Key Indicators

	food_store_count	households_served
count	39.00	39.00
mean	3.87	5210.57
std	2.34	4375.82
min	0.00	0.00
25%	2.00	2859.17
50%	4.00	4005.00
75%	5.00	5754.17
max	11.00	22920.00

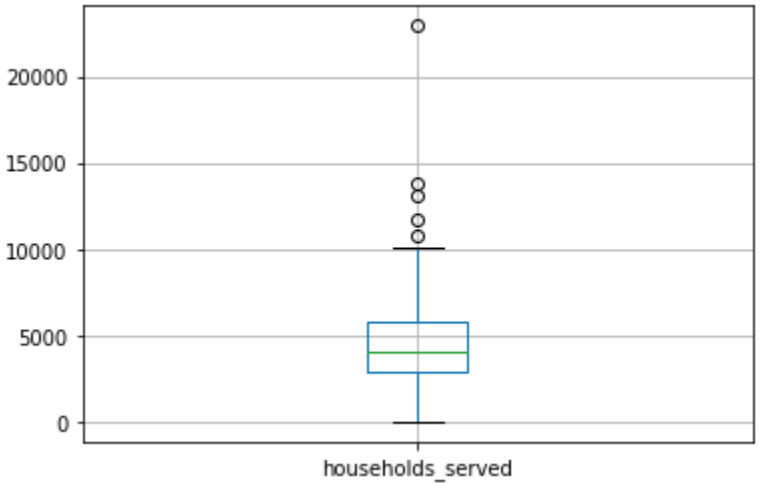
Visualize Key Indicators

Here we look at the information presented by the Key Indicators

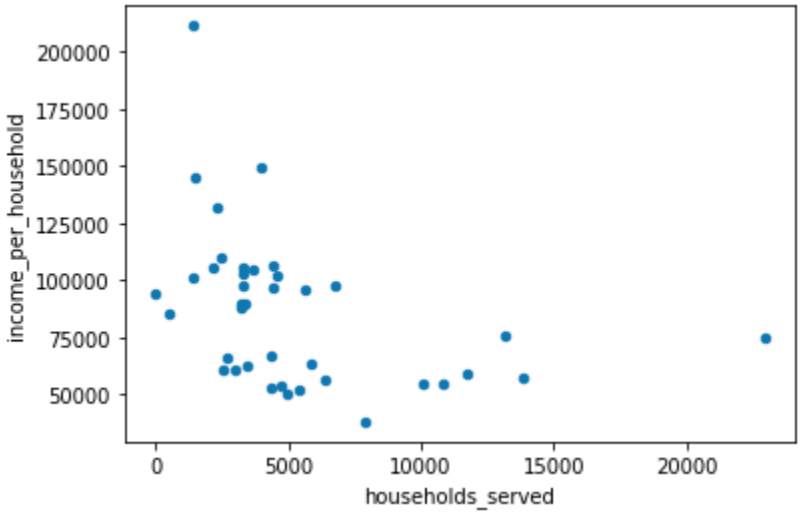
Distribution of Food Store Counts



Distribution of Households Served



Plot of Households Served and Income per Household



Graph of Number of Households Served per Food Store by Neighborhood

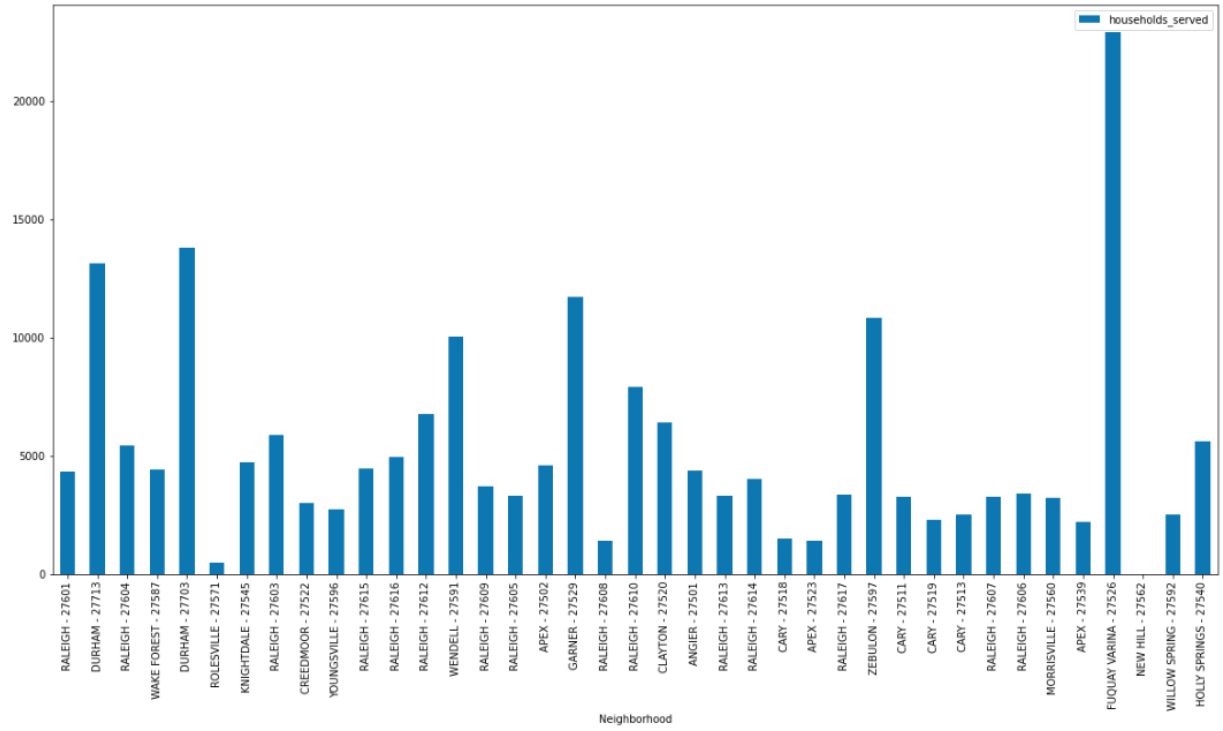
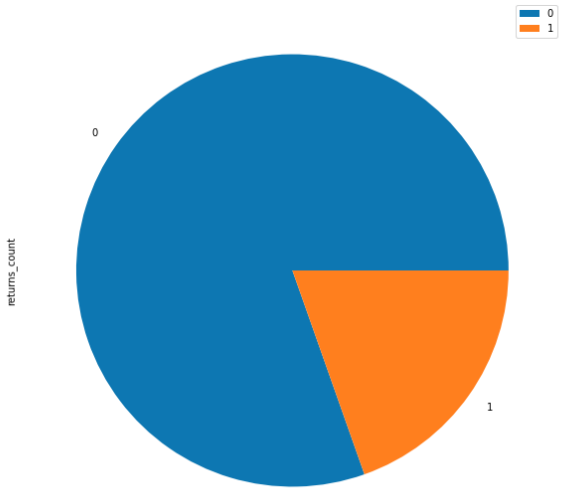


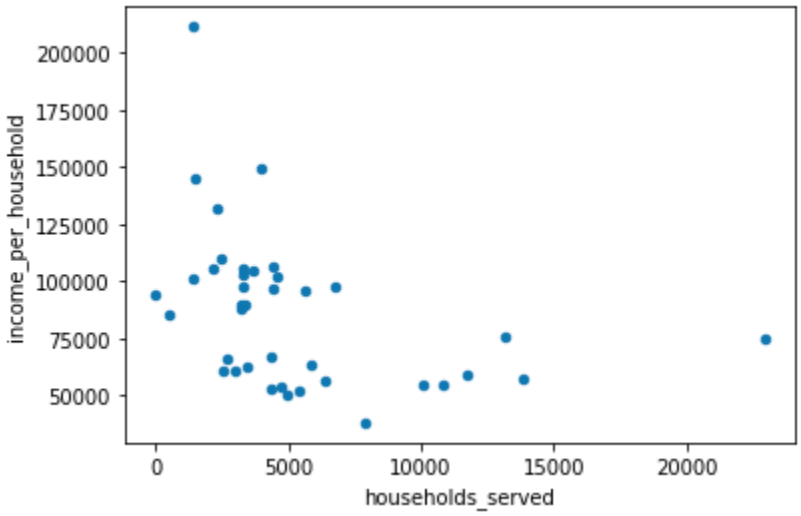
Chart of Served vs Under Served Neighborhoods by Population

0 = Served, 1 = Under Served

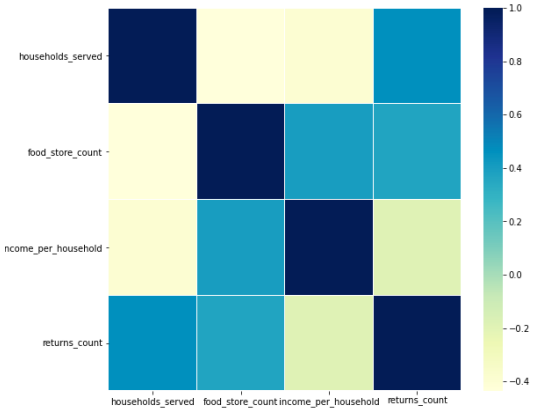


Discussion and Observations

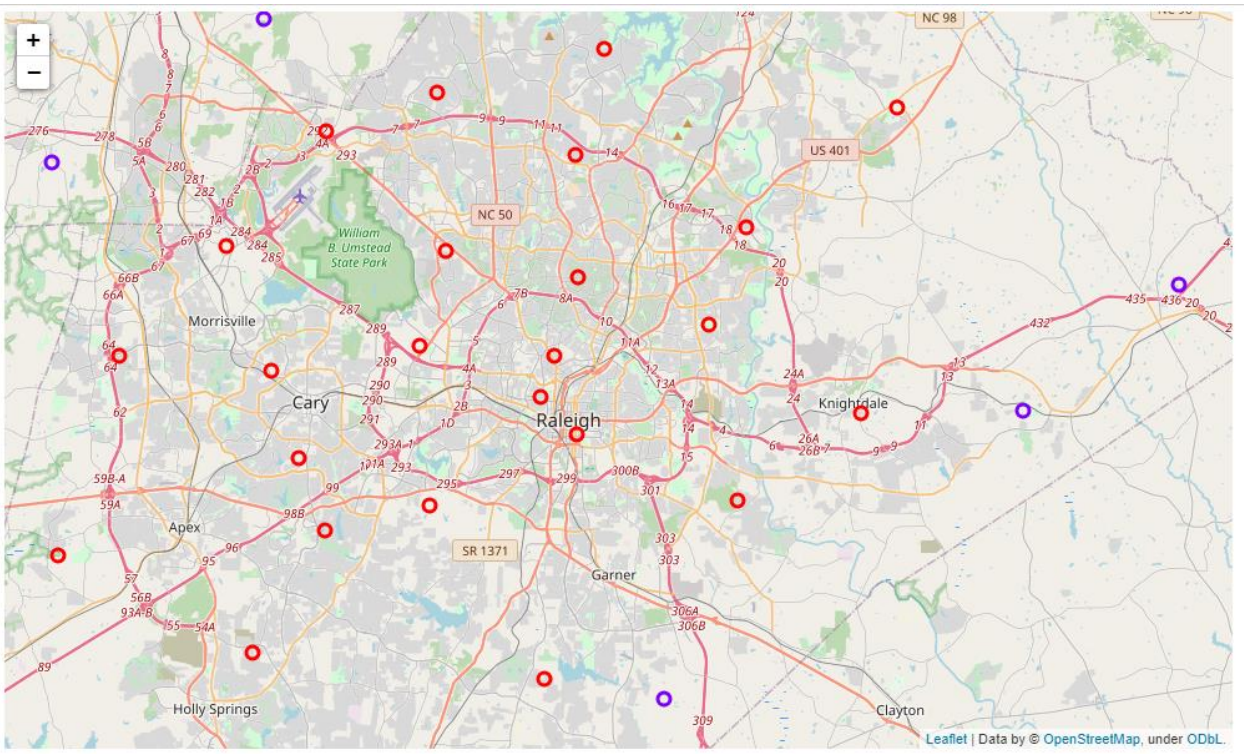
According to the Plot of Households Served and Income per Household, there does not appear to be a relationship between Quality Food stores and Income as all income ranges are clusters right around the 5,000 households per food store range.



A Correlation Analysis Evaluation shows this lack of correlation as well



There does seem to be a few areas of opportunity for new food stores to come in and help the market however on the outer edges of the County. The Red Circles indicate where Zip Codes are Well Served and the Purple where they are underserved.



Conclusion

Based on our analysis of the data, the questions in our initial business problem can now be answered.

- A food desert does appear to exist in certain zip codes where the population is "underserved" by quality food store choices.
- This food Desert does not appear to be tied to average income, but to geography. The underserved zip codes appear to be in the Rural areas of the county
- Any of the Zip codes that are under-served would warrant further investigation into provisioning additional food stores by an enterprising company willing to invest in providing high-quality food supplies to a rural environment