Food Accessibility in Erie County, NY

Erin Gregoire
Fall 2024

INTRODUCTION

As an Erie County native, I have heard anecdotal relics of food deserts within the county, especially in the City of Buffalo. In this project, I have performed analysis on the available food stores, income, and population within the municipalities located in Erie County to answer some key questions. Are there any areas of Erie County that experience shortages of quality grocery stores? Is there a correlation between the income in a zipcode and their access to these quality grocery stores? Does living in the City of Buffalo differ from the Suburbs in terms of food store quality and accessibility? Can we predict whether a person is a resident of the City of Buffalo or a Suburb based on their income and local food stores?

THE DATA

The two datasets used in this analysis are “Retail Food Stores” from the New York State Department of Agriculture and “Individuals, ZIP Code Data” from the US Department of the Treasury. In preparation for analysis, I cleaned the data by removing missing values, outliers, and unnecessary data including extraneous columns and zipcodes other than the ones in Erie County. Within the Food Store data, some highlights include the store’s name, zipcode, city name, establishment type, and store focus. Establishment type denotes the type of business, such as store, bakery, wholesale, etc. The store focus column provides the store’s quality indicator. A grocery store is classified by focusing primarily on food products, including food from all food groups (Produce, Dairy, Meat, and Dry/Canned Goods). A convenience store ranges from prepared food to snacks. They may contain some groceries but do not have the full scope of products. Multi-purpose stores are typically large name stores where groceries are only a small percent of products sold. Specialty stores focus on only one food group or product, such as candy shops, meat/fish stores, or other artisan goods. Lastly, a pharmacy may carry a small selection of grocery products but it is not the focus and is unreliable. A store may be categorized as “Other” if they do not easily fall into one of these aforementioned categories. The income

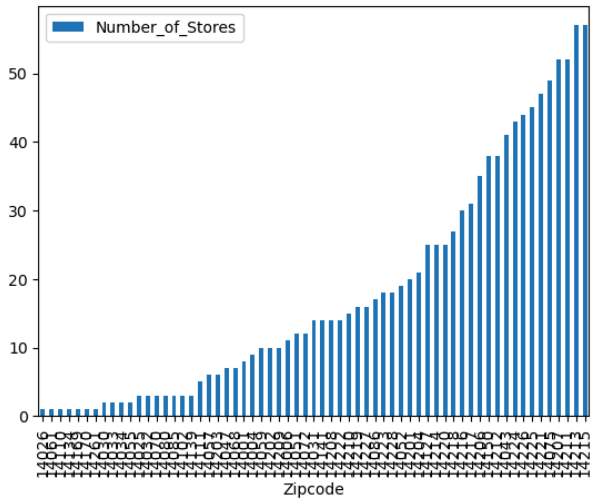
dataset provides standard income brackets and the number of returns conducted within each bracket for each zipcode. This dataset also includes a column for a total number of returns which was used as an approximation for the population in that area. Overall, the biggest impact of cleaning the data is that the size of the working files was reduced significantly, removing excess data that is irrelevant to my analysis. Also, after the data cleaning process, all of the data is very accessible and in easy-to-understand formats.

SQL DATABASE DESIGN

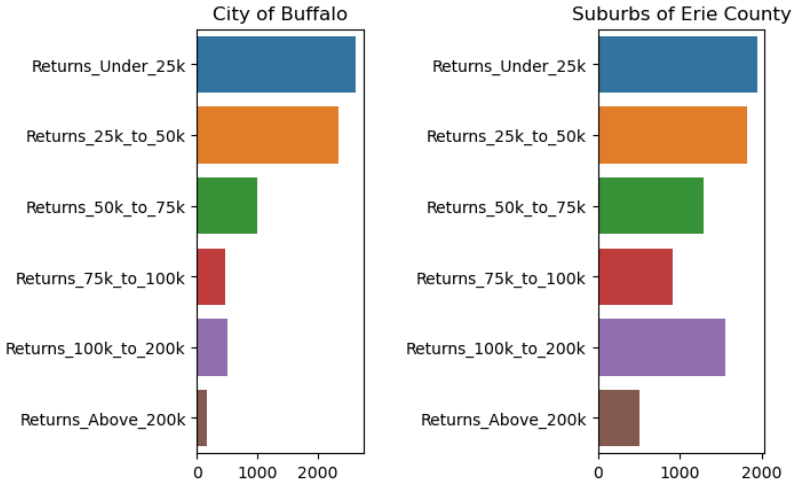
At this step in the project, I created a database called Erie County DB to hold all of the cleaned data in organized tables. I also performed normalization, which is the process of ensuring that data is well-structured and prevents anomalies. To perform 1NF, I created foreign keys for each Establishment Code that a food store has. For 2NF, each of the income brackets received their own table to prevent partial dependencies. Lastly, 3NF was achieved by creating a new table called Location due to City having a transitive dependency on Zipcode. Also, a new table was created for Establishment Type due to a transitive dependency of the establishment type on the establishment code. In the end, the following tables make up the Erie County database: Location, Food Store, Establishment Type, Income Total, Income for Under 25k, Income from 25k to 50k, Income from 50k to 75k, Income from 75k to 100k, Income from 100k to 200k, and Income Above 200k.

EXPLORATORY DATA ANALYSIS

During this step, I highlighted key sections that the data falls into, performing both visual and aggregate analysis, before combining the results into an interactive dashboard. The first section is looking at the number of food stores across Erie County.



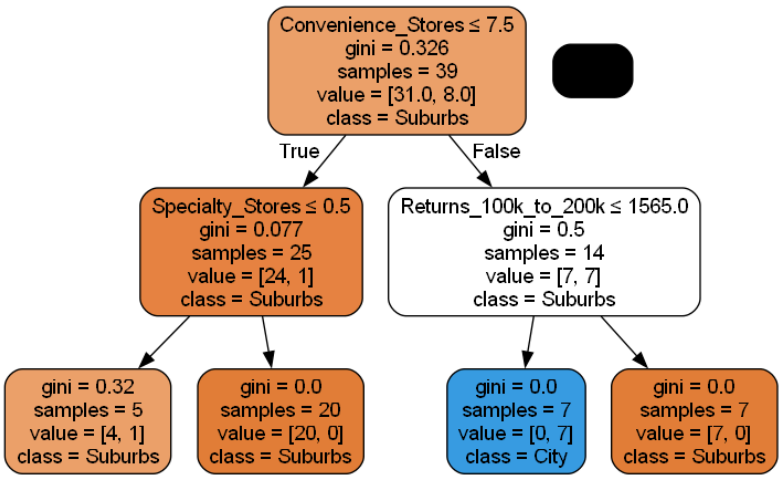
This graph shows how many food stores are present in each of the zipcodes of Erie County. Although this graph is a bit challenging to read, it is a good starting place with the data. We can see that the two zipcodes with the most food stores are 14213 and 14215. The food stores with the least stores are 14026, 14061, and 14110. It is important to keep our eye on these zipcodes especially moving forward. The next section of the analysis is about the quality of the food stores. To measure this, I looked at how many grocery stores of each focus are available in each zipcode. To view these graphs in depth, please visit the dashboard section of the project. My findings here were that convenience stores in all zipcodes were significantly higher than expected. No matter whether the zipcode is in the City of Buffalo or the Suburbs, one can expect to see approximately thirty percent or more of the food stores made up of convenience stores. While this is true, the City of Buffalo definitely experiences fewer instances of stores categorized as grocery, specialty, and multi-purpose, denoting that they have lower-quality food stores available. In the next section of the analysis, I took a deeper look into income trends.



Taking a look at the breakdown of income in the City of Buffalo versus in the Suburbs, we can see that although both have a similar number of returns in the lowest income brackets, the higher income brackets take a very high leap when moving to the Suburbs. When combining all of these findings, we can see that in areas where there is lower income, such as the City of Buffalo, the quality of food stores significantly decreases.

MACHINE LEARNING

Now it's time to determine if the region of Erie County can be predicted based on the income and food store variables. To do so, I build a decision tree model for classification. First, I separated the data into City of Buffalo (1) or Suburbs (0). Then I split the data into test and training so that I can validate my model. Next, I trained the model and then tuned the original model to its optimal version by enhancing its parameters with the gini index and a maximum depth. In the end, the decision tree model was able to accurately depict the City of Buffalo zipcodes from the Suburb zipcodes approximately 78% of the time. The graph below shows the final model. According to this model, the most important variables for predicting whether a record is in the City of Buffalo or the Suburbs include the number of Convenience Stores, the number of Specialty stores, and the number of returns within the 100 to 200k income bracket.



DASHBOARD

The final step of the project is to pull all of the data together to showcase the most important findings. This was done in a dashboard. Its first feature includes a dropdown menu of all zipcodes. Shown are statistics such as the populations, and visualizations of the income brackets and quality of grocery stores. I believe that this is the most effective aspect due to being able to directly see what incomes have access to what quality food stores. Also provided are the income bracket statistics for the regions of Erie County, and the quality of food stores by region.

CHALLENGES

While brainstorming during step one, I believed that my biggest obstacle would be combining the data from different files. This proved to be manageable due to the zipcode field on both of the datasets that acted as primary keys and allowed for a cohesive merge. However, another challenge arose organically. There were some zipcodes that were duplicated due to improper spelling of city names or were listed under “Buffalo” although their true name is something else, such as Lackawanna. Fortunately, I was able to sort this out by verifying duplicates with the correct spelling and town name. One last challenge that was not fully tackled in this analysis is that the zipcodes in Erie County were only categorized into two groups. Further analysis could be completed by adding a “rural” or “outskirts” category.

CONCLUSION

In conclusion, this analysis did find that there is a correlation between the quality of grocery stores, the regions of Erie County, and the average income in those areas. The City of Buffalo has lower income levels than the Suburbs of Erie County and also fewer high-quality food stores. Convenience stores make up significantly more of the food store industry than expected, including in both the City and the Suburbs.

REFERENCES

Department of the Treasury. “Individuals, ZIP Code Data.” *Statistics of Income (SOI),* 22 Aug. 2024. https://catalog.data.gov/dataset/zip-code-data

New York State Department of Agriculture and Markets. “Retail Food Stores.” *NY Open Data,* 9 Sept. 2024. https://data.ny.gov/Economic-Development/Retail-Food-Stores/9a8c-vfzj/about_data