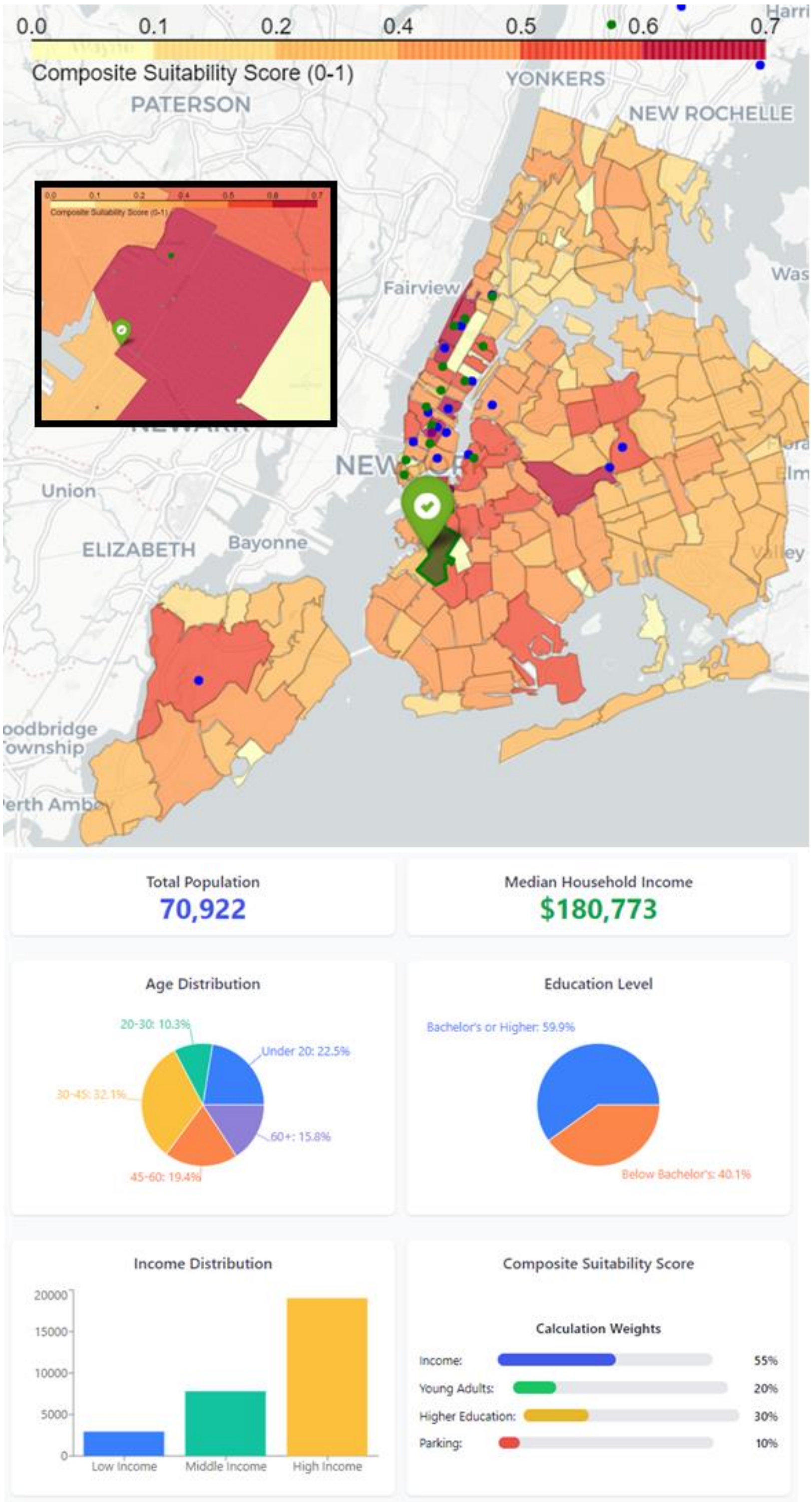
Trader Joe’s Point72 Case Study



Total Population
70,922

Median Household Income
$180,773

Calculation Weights

Income: 55%
Young Adults: 20%
Higher Education: 30%
Parking: 10%

This figure shows the recommended new location for Trader Joe’s, as it meets the requirement of a high-income, well-educated neighbourhood that performs well on a combined metric of income, age, education and parking (a key metric for a Trader Joe’s location).

It is located near 5 stations and will be competing with a Whole Foods in the area, which is something Trader Joe’s regularly does and acts as an indicator of demand.

Initial thoughts

First, the problem scope and definition of the problem should be formalised. What criteria should be used when advising where the Trader Joe's (TJ) store should be?

As a business, Trader Joe's goal when opening their new store will be to maximise profits. Note that **this is an assumption,** as they could be maximizing revenue, sales, or market share instead. When conferring with Trader Joe's, I would clarify this in a real use case.

As profit maximisation is the goal, I need to find methods to approximate revenue and costs so that profitable locations are identifiable. Additionally, understanding the demographics of TJ is key to finding successful locations.

Understanding TJ's business model is also key, so I began researching the company, learning about its unique approach that led to its "cult-like" following. I found that its streamlined approach allows it to provide unique, higher-end products at a more affordable price point compared to other upscale grocery chains, such as Whole Foods, which has previously stated that TJ is its main competition in the market.

The TJ selection process is data-centric, slow, and deliberate. It focuses on population density accessibility and focuses on business metrics over expressed customer demand (such as petitions).

Their general target demographic is reported to be in the middle-upper income bracket, educated, and younger, which closely matches the Whole Foods demographic.

Below is a list of some of the links I visited initially:

https://retailwire.com/discussion/trader-joes-discusses-painstaking-site-selection-process

https://www.grocerydive.com/news/trader-joes-store-growth-2024-mapping-expansion/

https://blog.osum.com/trader-joes-expansion-plans

https://www.foodnavigator-usa.com/Article/2014/04/15/Quirky-cult-like-aspirational-affordable-The-rise-of-Trader-Joe-s/

https://azira.com/blogs/trader-joes-recently-expanded-into-a-new-market-what-data-shows-about-their-site-selection

Possible approaches

After establishing a fundamental understanding of the problem, the solution's objectives, and an overview of the existing selection process, I began formulating different approaches.

The following were the approaches I considered:

1. One approach would be to visualise and develop an understanding of the features, working toward formulating a method of quantifying optimal locations using various collated data and some human interpretation.
2. A machine learning approach could more algorithmically attempt to model successful locations. However, to model successful locations there were two approaches I considered.
    a. Linear regression (this could also fall under the data science approach). The features could be in a linear regression (or even Deep Neural Net) model to approximate whether the following location (e.g., ZIP code) would be a good location for a TJ
    b. Clustering. This unsupervised approach would cluster the locations. This method could be formulated such that each ZIP code (a row in a table) has feature vectors from census data and other sources (see datasets mentioned below) and attempts to cluster zip codes using these features as similarity indexes; this could be achieved with a variety of clustering approaches such as KNN or DBSCAN. Once the clusters are acquired, the clusters with the most TJs in them (or near them) could serve as indicators of good locations for a new TJ. For example, if one cluster held 90% of the TJs then other nodes (correlating to ZIP codes) could be considered more likely to be a good location

However, considering the larger picture, I concluded that an ML approach isn't suitable for the problem. We are unsure of what we are trying to predict for a regressive predictor, with the simplest being linear regression. There is no clear ground truth against which to evaluate. We could attempt to use the existing TJ locations as labels to form a classifier instead or measure the distance to the closest TJ for a regressive model; however, this does not solve the question. At best, such models would learn to approximate an imperfect version of the existing TJ strategy for picking locations, and at worst, they could overfit and learn the current locations and develop a bias toward the existing locations.

Finally, taking one step further towards the bigger picture, if I were to advise Trader Joe's on their new location, it would run counter to the point to predict how they already predict locations to form my advice. Additionally, any proposals should be clearly evidenced by transparent and explainable proof, such that it can give both us and the client confidence in our proposals.

Thus, I approached this from a data science perspective, focusing on human interpretable insights and visualisations.

Data processing and visualisation

I collated a variety of data to approximate combinations of features that could indicate a successful location:

Census data tables:
B03002,BO8301,B11001,B15003,B17001,B19001,B19013,S0101

These census tables cover income, household information (and income), income brackets, age, gender distribution, education, poverty, commute data, population counts.

Other datasets:

Automated traffic volume counts, Bi-annual pedestrian counts, Licensed parking lots, MTA Subway stations, Retail food stores, Modified zip code tabulation areas (MZCTA), (Non-building) Zip codes

Why zip codes? I chose to use zip codes to segment the data not just because it is a **familiar** and **practical** divider for NYC such that the visualisations and results would be **easily interpretable** for the **client**.

Note: While working with the datasets, I discovered that the data for Bi-annual pedestrian counts and Automated traffic volume counts were geolocated in a few sparse locations. These locations were too few to approximate surrounding counts, so I didn't use these datasets.

I first explored the census datasets and identified key columns to extract and use, such as median income per household, specific ethnicity data, and academic achievement (e.g., highest qualification—BSc). I merged these datasets by zip code (referenced by geoid at points in the code).

The other geometry (longitude latitude or geometry-based) datasets were also processed to aggregate data (e.g., number of stations in each zip code, summing age groups to get "under 20" from other columns, getting aggregate counts of each grocery store by filtering and grouping retail store data). These datasets were then converted into geopandas datasets either using their co-ordinate data or geometric data, then combined into one geopandas dataset and finally combined with the census dataset using the MZCTA dataset to convert geographical data into a zip code tabulated dataset of all features combined. After some final processing (adjusting income columns to include low, middle, and high-income brackets, age groups, etc), the data was ready for visualisation.

Note: During exploration, I found that there were often entire rows of ZIP codes with missing or very small counts of data. Upon investigation (not only into the dataset but also contextualising with ZIP code lookups on specialised map sites) I discovered ZIP codes were assigned to buildings as well, this fed into my decision to use MZCTA as it mirrors the ZIP code information without these unnecessary rows. Aside from this, since most of my data was count-based, null, or nan values referred to a zero count and were usually set to 0, this was dealt with using .fillna(0) where necessary.

Before visualisation, I also **normalised** each selected factor (this was possible as they were numeric features), such as total population, median household, etc., so that the importance of each feature was scaled appropriately when weighing it.

Finally, I **visualised** the complete dataset with **interactive** widgets using Streamlit to create a Streamlit app, **generating sliders for each feature** to **dynamically** set a value between 0 and 1 for each parameter correlating to its importance. This allowed for a **composite score calculation** used as the heat value for a zip-code segmented heatmap using Folium visualisation of the dataset. I then added an overlay on the heatmap with the store locations for TJs, Whole Foods and Wegmans.

The entire visualisation is built as a Streamlit app that is fully interactable. It is attached to my submission, and **you should be able to run it** using this script:

```
streamlit run app.py
```

If it is not working please ensure the correct geojson is in the same directory as the app and that your environment has the packages imported in the app.

From then on, I had a fully interactable tool that produced an easily interpretable heatmap output using transparent weights that I could freely adjust.

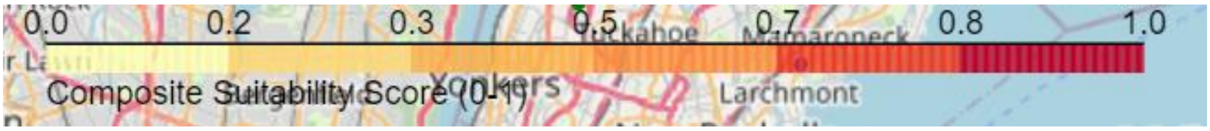
I created an interactive app instead of a series of quicker-to-code manually set parameters in Folium for a static heat map because, in a real case study, this would be powerful for our use and the client. It would be **generalisable to other use cases and augmentable with third-party data**.

Furthermore, I added code that calculates the “score” for the zip code relating to each TJ location so I could see the score existing TJ locations get for each combination of feature weightings I chose. However, I did not use this as a success metric but rather as a sanity-check measure.

Visualised data exploration

With my visualisation set up, I began systematically creating visualisations for features and combinations of features, first checking the heatmap for individual features (this could be done by weighting all other features at zero and the chosen feature at any positive value).

Key:



Stores: **Trader Joe's**, **Whole Foods**, **Wegmans**

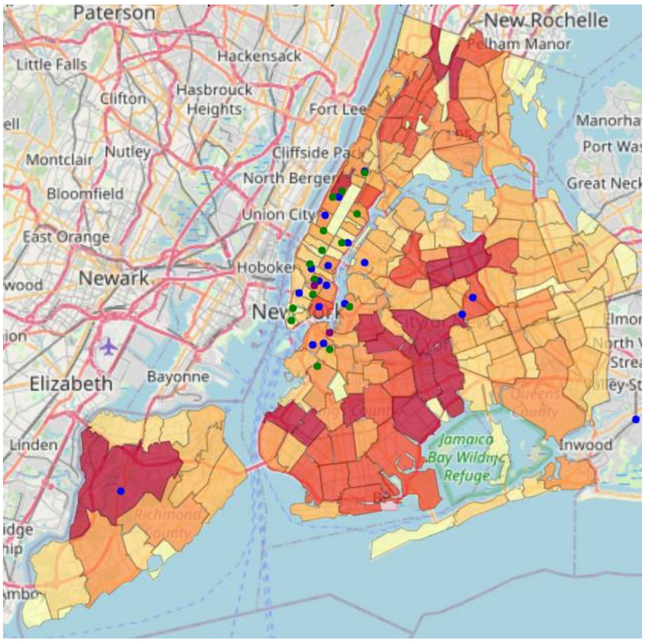


Figure 1- total population

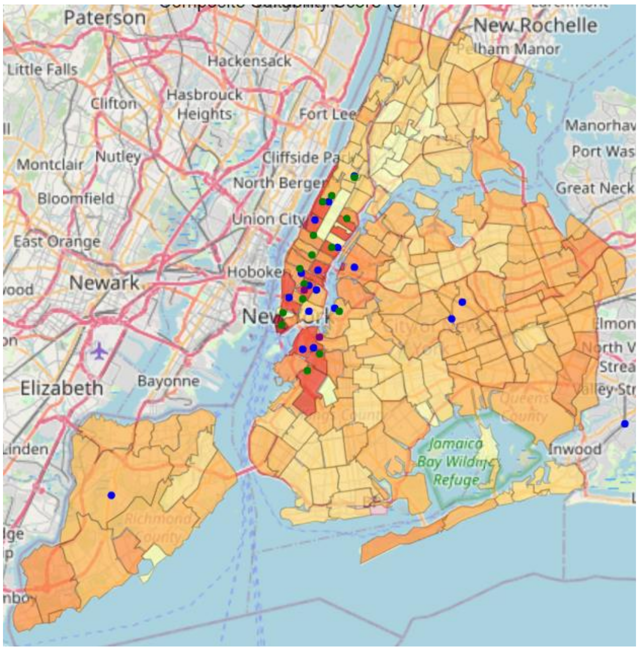


Figure 2 - median household income

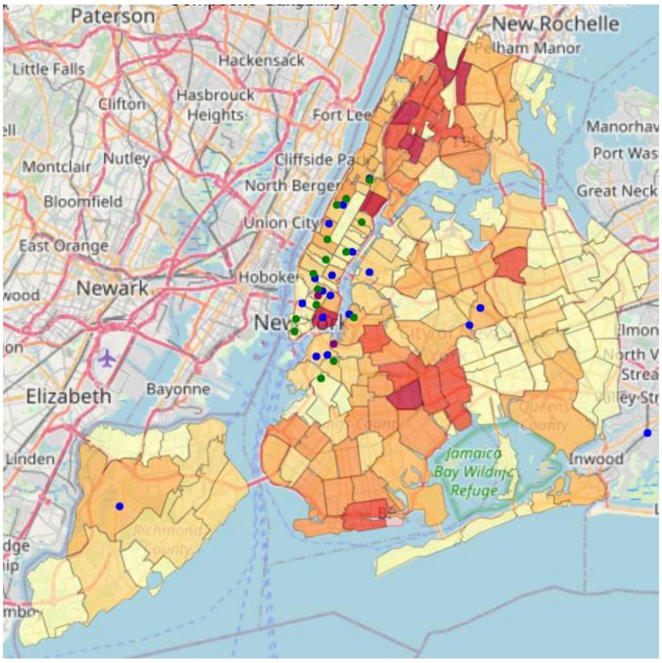


Figure 3 - low income

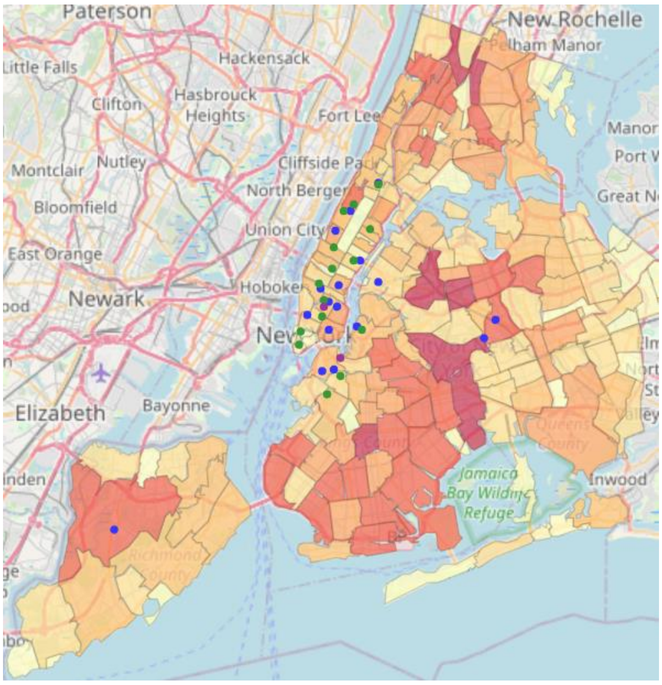


Figure 4 - medium income

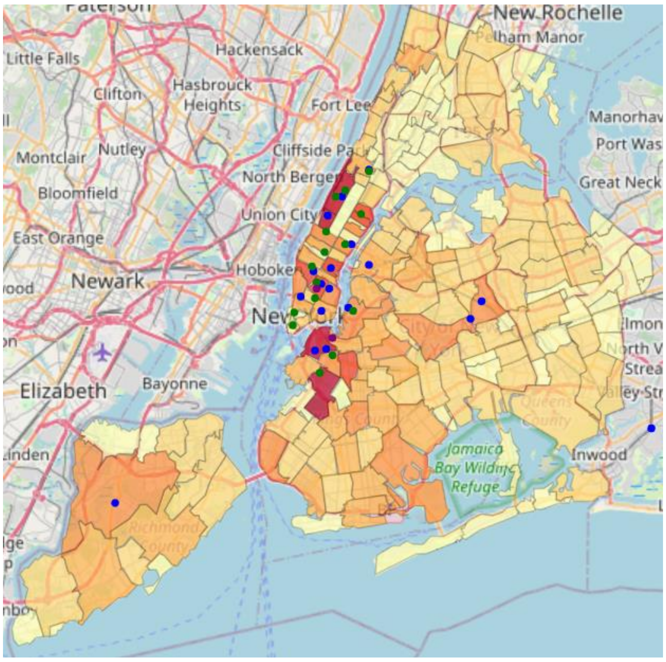


Figure 5 - high income

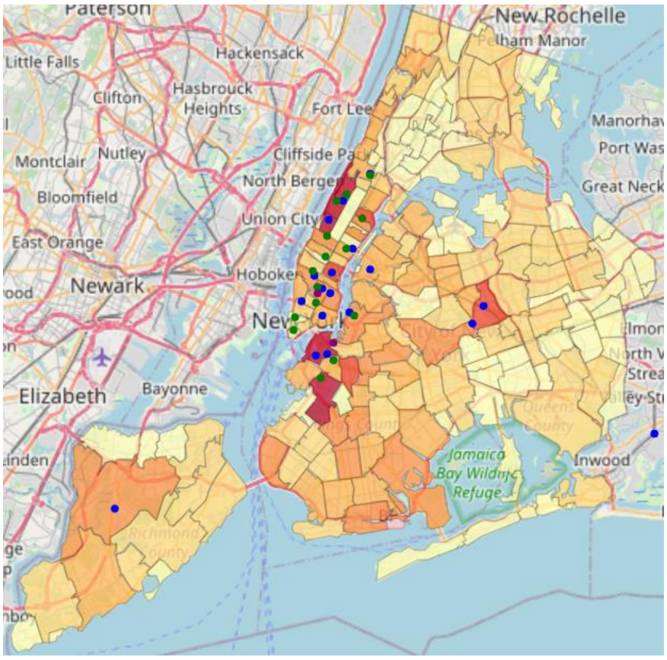


Figure 6 - BSc or above

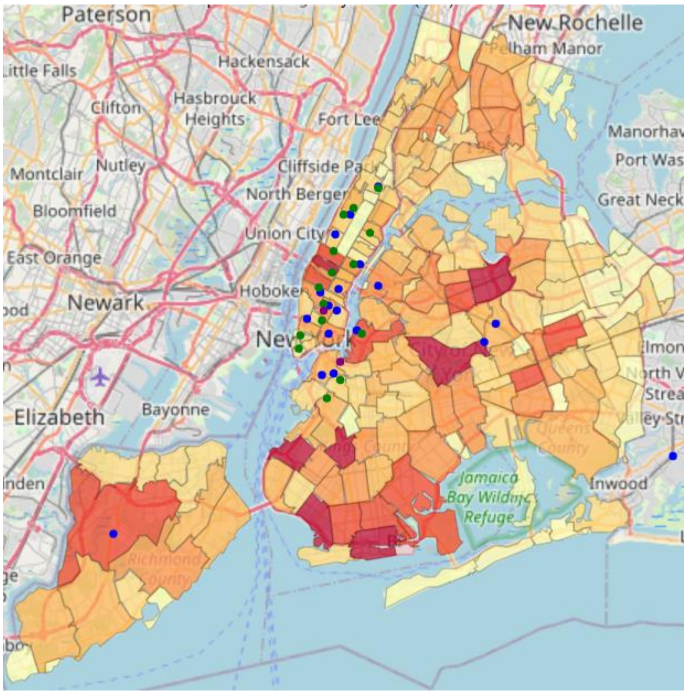


Figure 7 – parking space

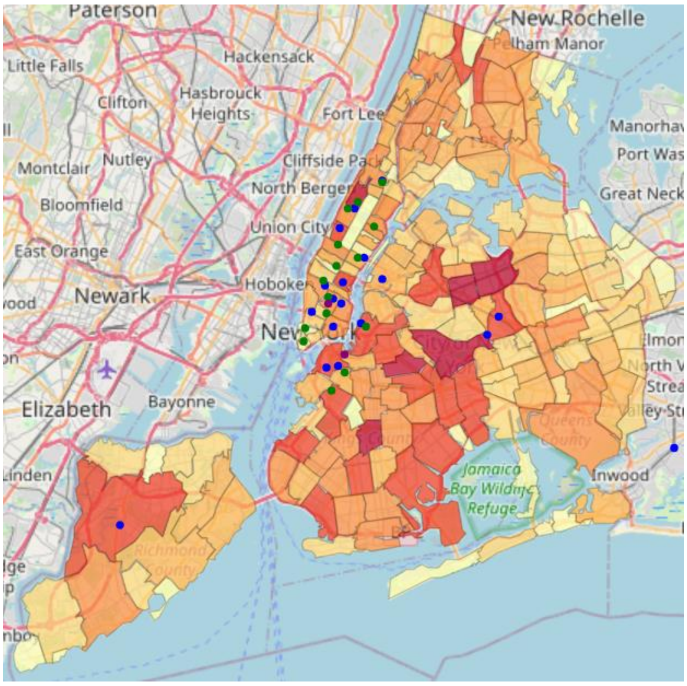


Figure 8 - commuters

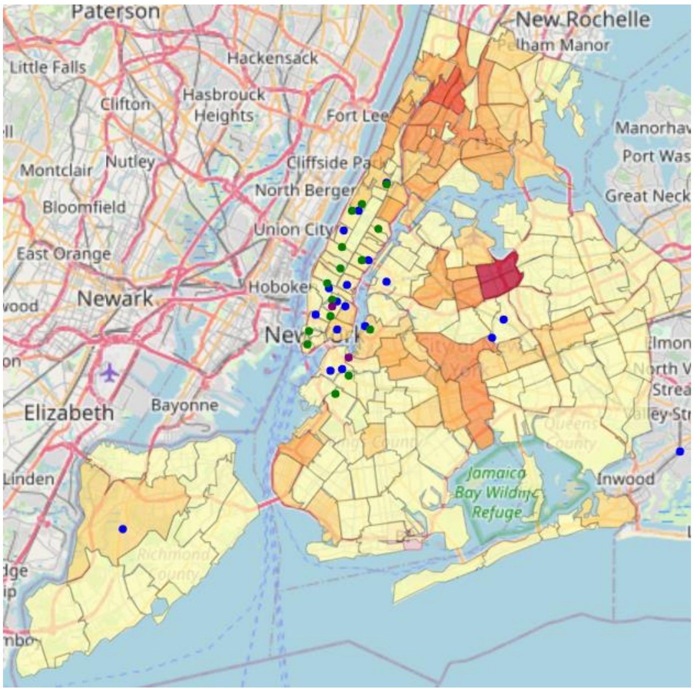


Figure 9 – Hispanic or Latino population

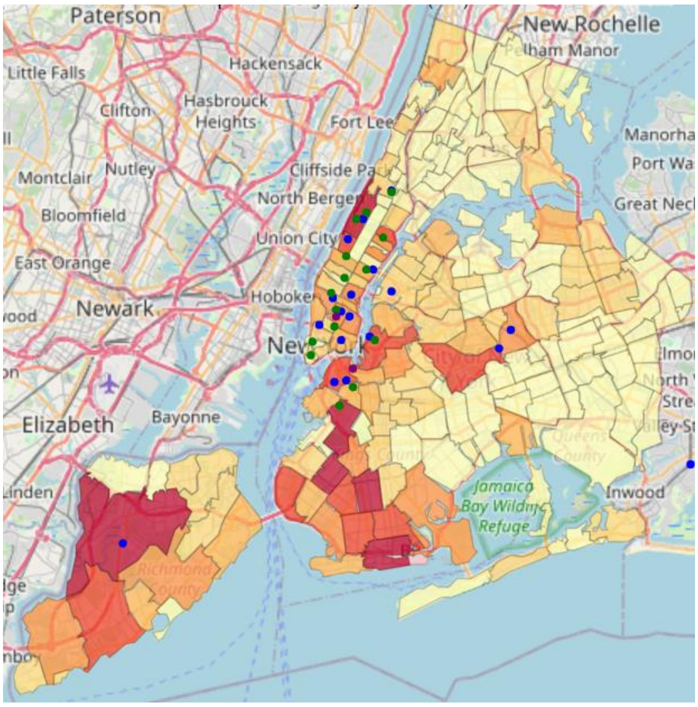


Figure 10 – White population

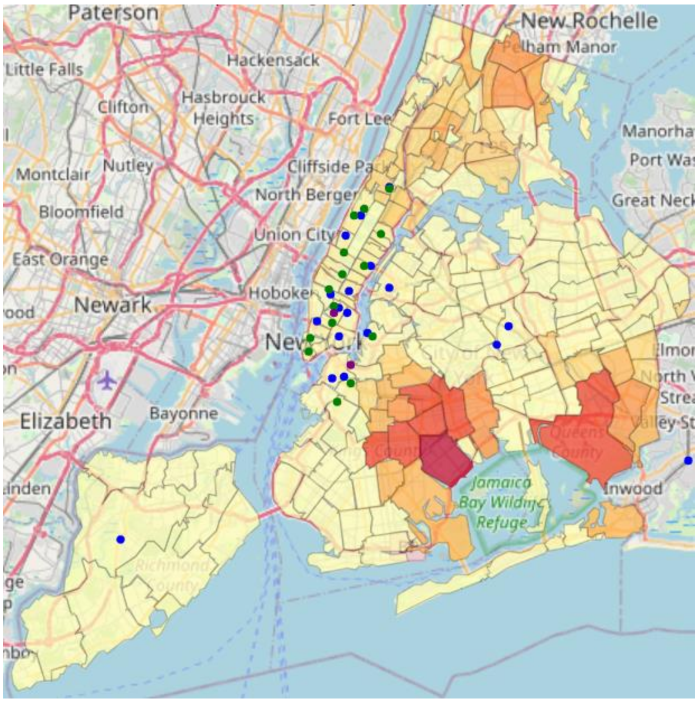


Figure 11 - Black or African American population

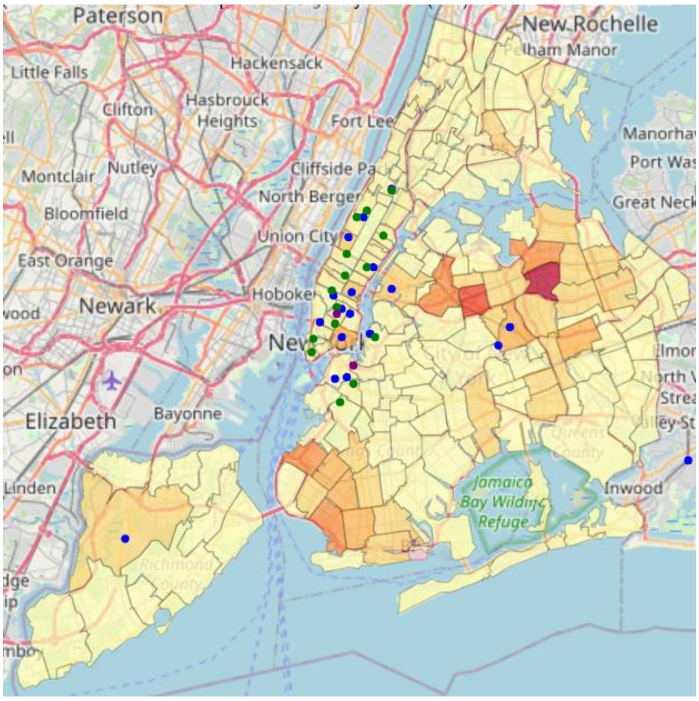


Figure 12 - Asian population

From the single-variable heatmaps, we can observe that the wealth distribution across the city, median income, and high income closely correlate with the existing stores. This is expected as these stores are generally considered upmarket grocery stores with reportedly middle-upper-income shoppers.

The minimum-BSc variable also matches our expectations of closely matching the stores since the average shopper in TJ and WF (Whole Foods) is more educated than the base population.

Having validated our assumptions from our initial research (**which is still based on the assumption that these locations are profitable and optimal**), I continued researching to find indicators for TJ shoppers.

Given more time and access to buy third-party data from sources such as:

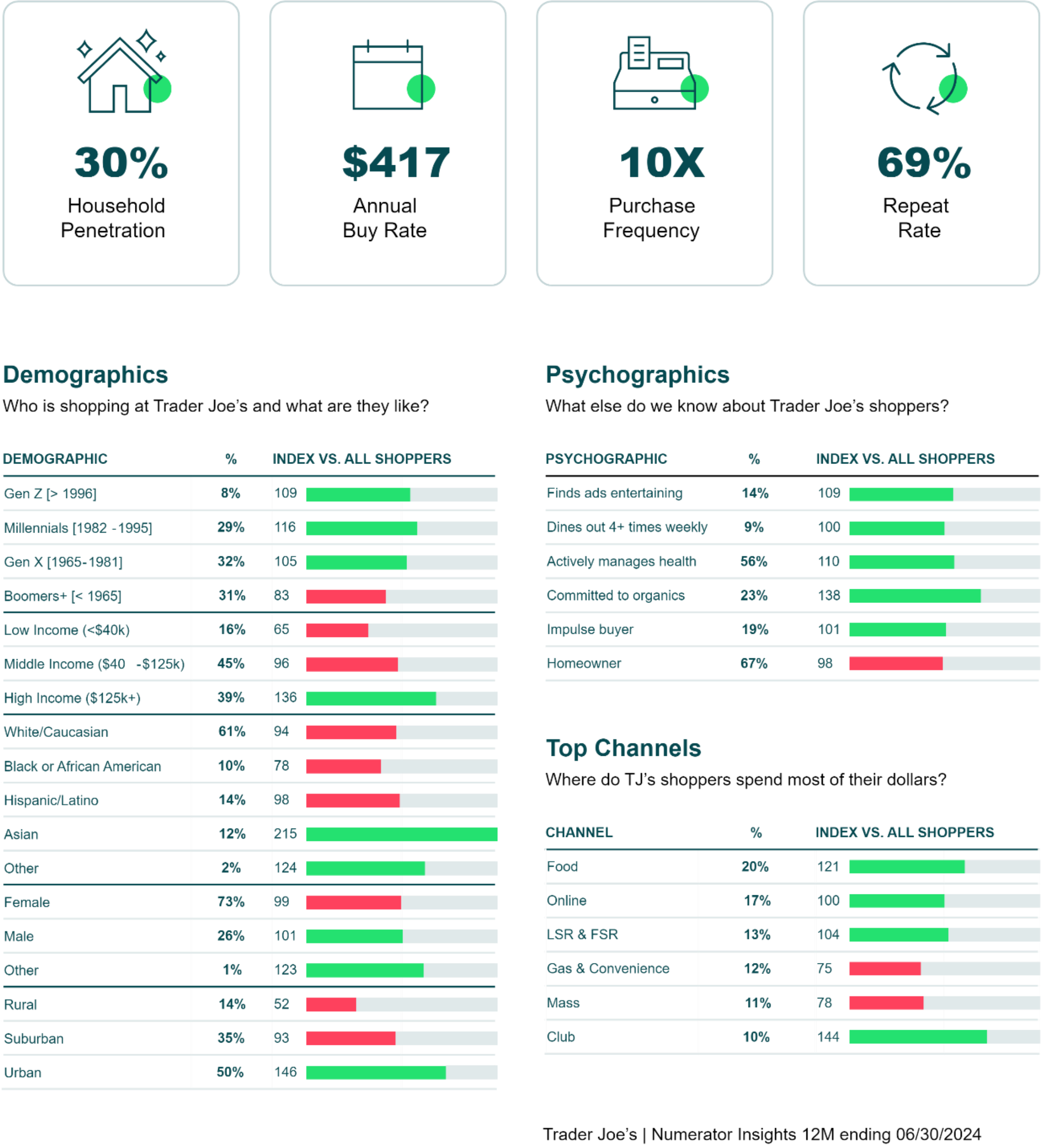
- Nielsen Solutions (customer purchasing habits, media consumption, lifestyle)
- SafeGraph (granular foot traffic data and visitation metrics)
- Brandwatch (Social media interests)
- Epsilon data (Demographic data, purchase behaviour, psychographic data, lifestyle insights)
- CoStar Group (Industry standard for commercial real estate data, which is necessary for modeling costs more effectively)
- Simmons consumer study (comprehensive lifestyle/brand preference study)

Adding these datasets to find correlations over a more extended period – especially social media usage and time-based foot traffic, commuting times, and turnstile information combined with Trader Joe's data on credit card transactions and profits by half-hourly data (granular) could lead to calculating how important commuting-based locations are, when shoppers shop, what kind of shops occur at what times (we could find where people regularly shop on the way home from work and thus find key commuting locations which are most important). With time to create and analyse a sufficiently large and varied dataset, the possibility of finding optimal locations should increase drastically.

However, since I didn't have access to such data and could only approximate an ideal location using second-hand information online and vaguely observing existing locations (to avoid re-implementing a distilled version of their strategy), the closest approximation of the TJ demographic I found was an infographic sample published by Numerator.

Using a demographic TJ model for visualisation

This source allowed me to have an approximation of the distribution by the statistics in the figure (sourced from their website).



30% Household Penetration

$417 Annual Buy Rate

10X Purchase Frequency

69% Repeat Rate

Demographics

Who is shopping at Trader Joe's and what are they like?

DEMOGRAPHIC	%	INDEX VS. ALL SHOPPERS
Gen Z [> 1996]	8%	109 🟩
Millennials [1982 -1995]	29%	116 🟩
Gen X [1965-1981]	32%	105 🟩
Boomers+ [< 1965]	31%	83 🟥
Low Income (<$40k)	16%	65 🟥
Middle Income ($40 -$125k)	45%	96 🟥
High Income ($125k+)	39%	136 🟩
White/Caucasian	61%	94 🟥
Black or African American	10%	78 🟥
Hispanic/Latino	14%	98 🟥
Asian	12%	215 🟩
Other	2%	124 🟩
Female	73%	99 🟥
Male	26%	101 🟩
Other	1%	123 🟩
Rural	14%	52 🟥
Suburban	35%	93 🟥
Urban	50%	146 🟩

Psychographics

What else do we know about Trader Joe's shoppers?

PSYCHOGRAPHIC	%	INDEX VS. ALL SHOPPERS
Finds ads entertaining	14%	109 🟩
Dines out 4+ times weekly	9%	100 🟩
Actively manages health	56%	110 🟩
Committed to organics	23%	138 🟩
Impulse buyer	19%	101 🟩
Homeowner	67%	98 🟥

Top Channels

Where do TJ's shoppers spend most of their dollars?

CHANNEL	%	INDEX VS. ALL SHOPPERS
Food	20%	121 🟩
Online	17%	100 🟩
LSR & FSR	13%	104 🟩
Gas & Convenience	12%	75 🟥
Mass	11%	78 🟥
Club	10%	144 🟩

Trader Joe's | Numerator Insights 12M ending 06/30/2024

The following on their site explains the index metric:

"**Demographics & Psychographics**: Percent of a retailer's shoppers who fall into each demographic/psychographic breakout, indexed against the percent of the general population in that breakout. An index of 120 would indicate the retailer's shoppers are 20% more likely to be part of a given breakout than the average shopper (i.e., 20% more likely to be Gen Z)."

Using this, we can now estimate the number of shoppers that would visit a Trader Joe's, adjusted with the index vs. all shoppers' data from the source above.

Note: This **assumes that** this data is reliable and has been taken from a sample size that is large enough. This also **assumes** that these demographic statistics are applicable specifically to NYC. Ideally, I would acquire a third-party dataset with detailed consumer spending habits ranging from grocery shopping to renting likelihood to identify strong correlations to extract useful potential indicators for a successful Trader Joe's. Such a dataset would be especially useful if it were very varied (e.g., the Simmons national consumer study) so unexpected identifiers (e.g., social media usage) could potentially show correlations, allowing predictions based on indicators Trader Joe's (or any other client) may not have considered themselves.

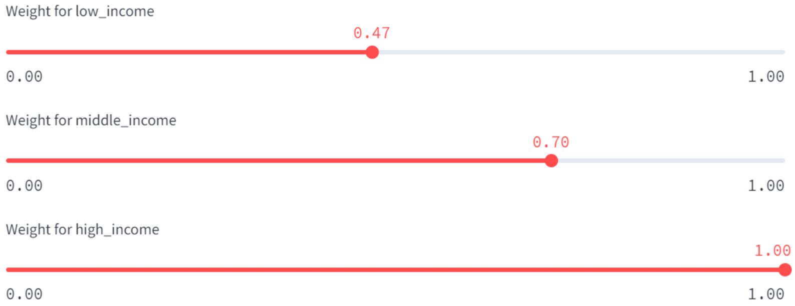
Income

Combining the census data collected for the number of residents in each income tier (low, middle, high) with the propensity of each income bracket to shop at Trader Joe's allows us to estimate the number of potential Trader Joe shoppers in each zip code contains.

Heatmap distribution of the estimated number of shoppers adjusted using the likelihood of shoppers to visit based on their income:

Note: Income weights were determined by scaling the index values to a maximum of 1 (therefore, each income index was divided by the highest value in the high-income index).

Looking at the generated heatmap, we already observe many estimated customers in the Brooklyn/Queen boundary when solely observing an income-adjusted customer estimate score.



Weight for low_income

0.47

0.00 1.00

Weight for middle_income

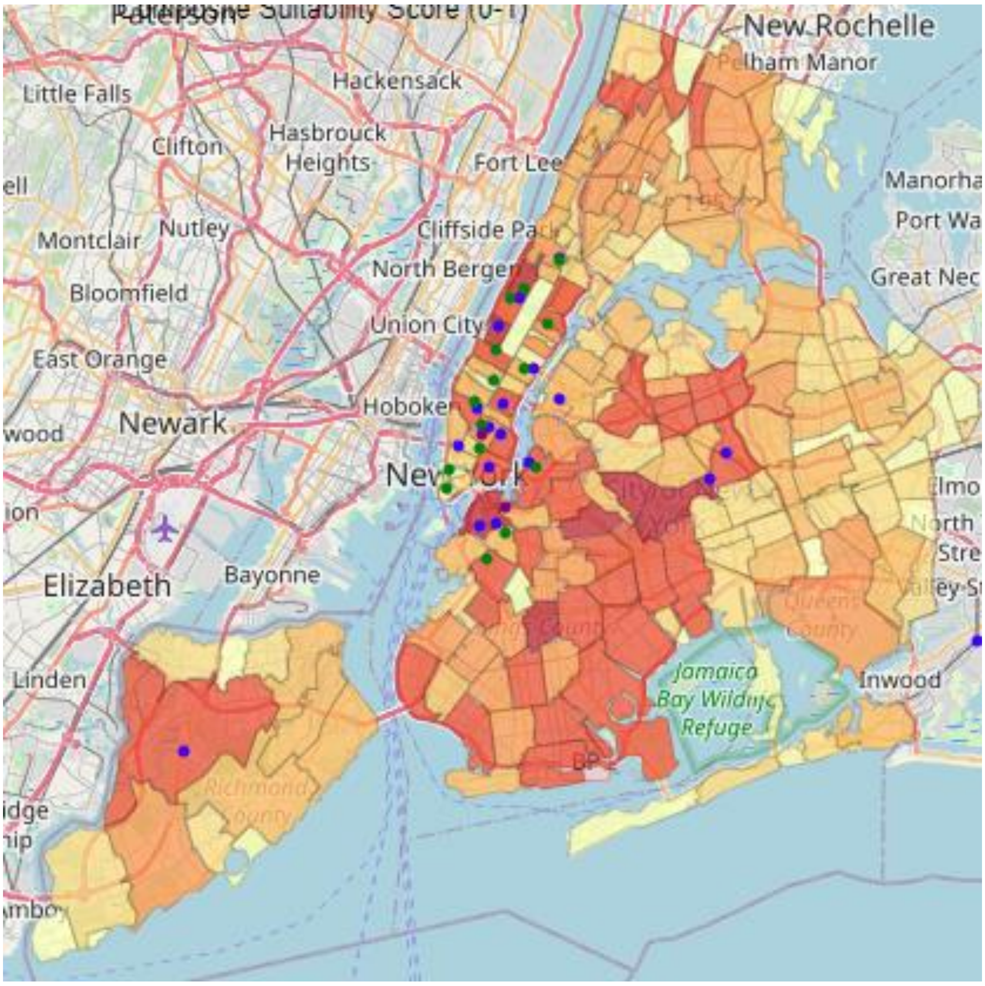
0.70

0.00 1.00

Weight for high_income

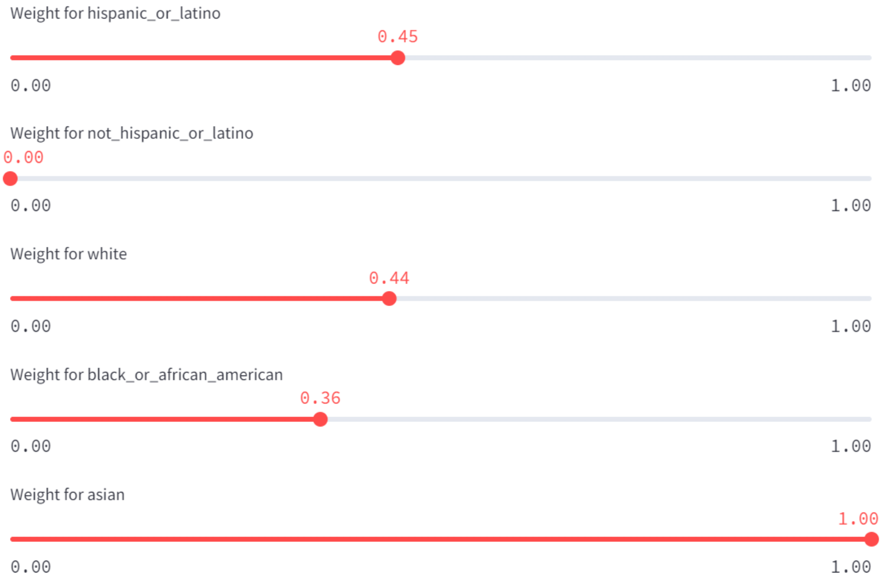
1.00

0.00 1.00



Ethnicity

Using the same approach, we scale the weights according to ethnicity in this case to observe our ethnicity-based customer estimation. This time, there is a relatively equitable score across the heatmap, with central Queens and west Brooklyn showing some pockets with high scores due to a denser Asian population (reaching over 50% in some cases).



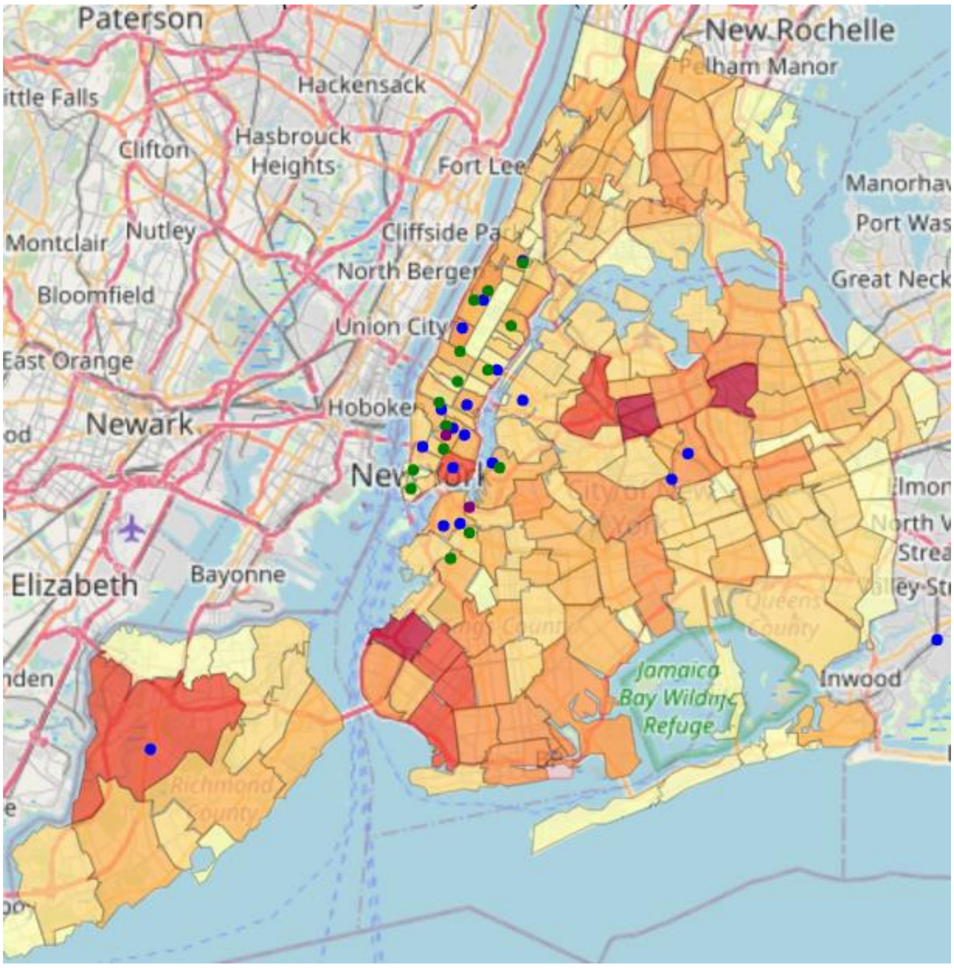
Weight for hispanic_or_latino
0.45
0.00 1.00

Weight for not_hispanic_or_latino
0.00
0.00 1.00

Weight for white
0.44
0.00 1.00

Weight for black_or_african_american
0.36
0.00 1.00

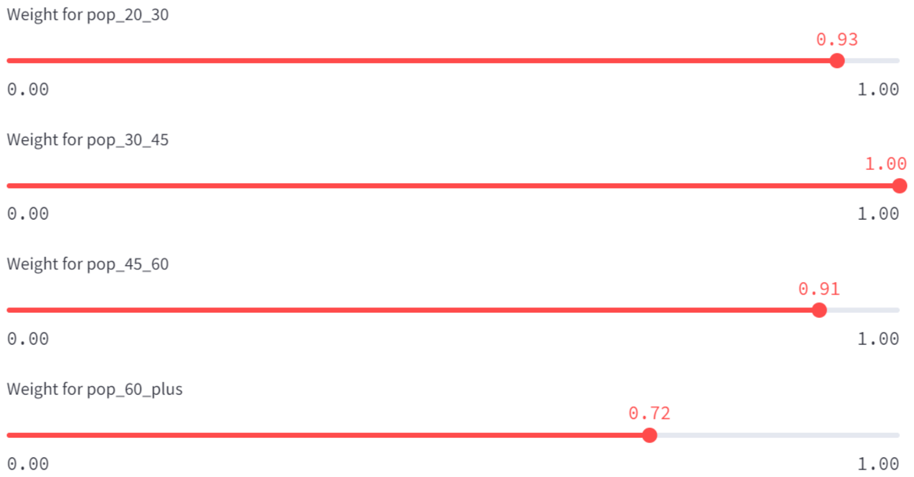
Weight for asian
1.00
0.00 1.00



Age

Using the same approach, the scaled age values produce the heatmap below. In this case, the heatmap has less concentrated high-scoring areas, although we observe that the Brooklyn/Queen boundary is once again scoring highly as it did with the income-adjust heatmap.

Note: The ages for this calculation don't exactly match the demographic categories from the numerator; however, they have been aligned within 3 years at most on either side of the range.

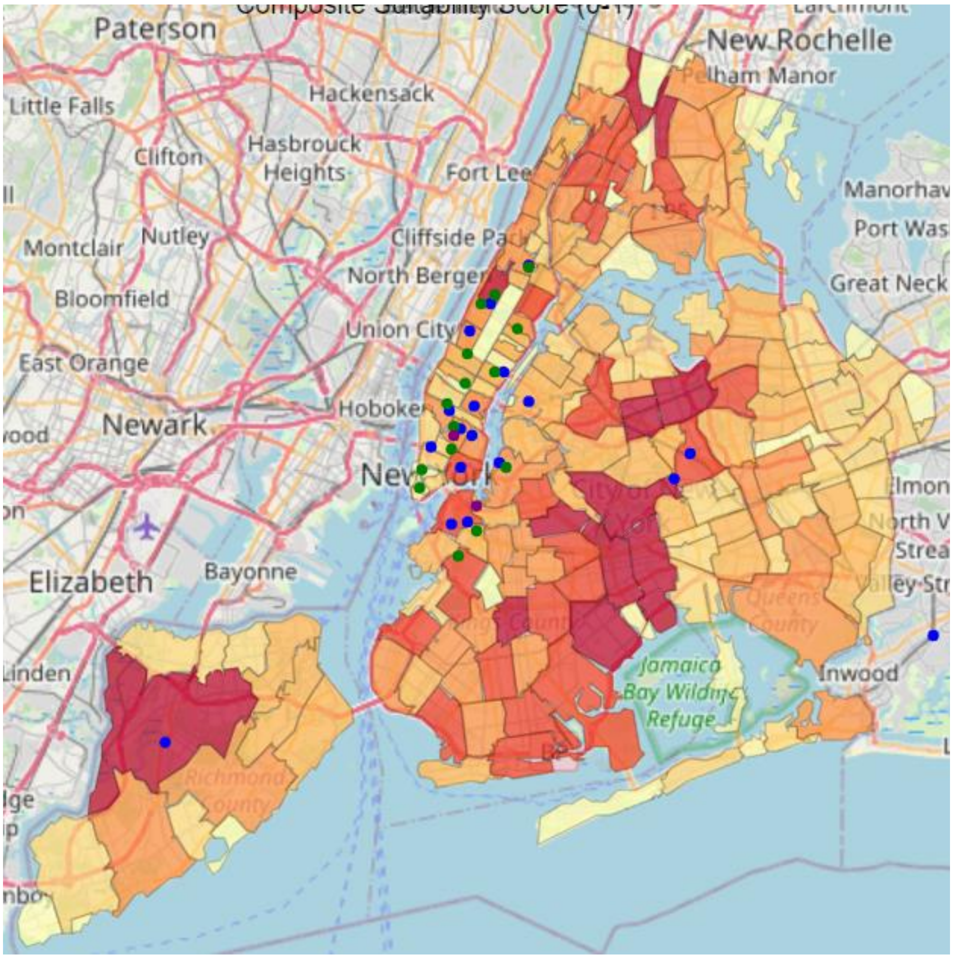


Weight for pop_20_30
0.93
0.00 1.00

Weight for pop_30_45
1.00
0.00 1.00

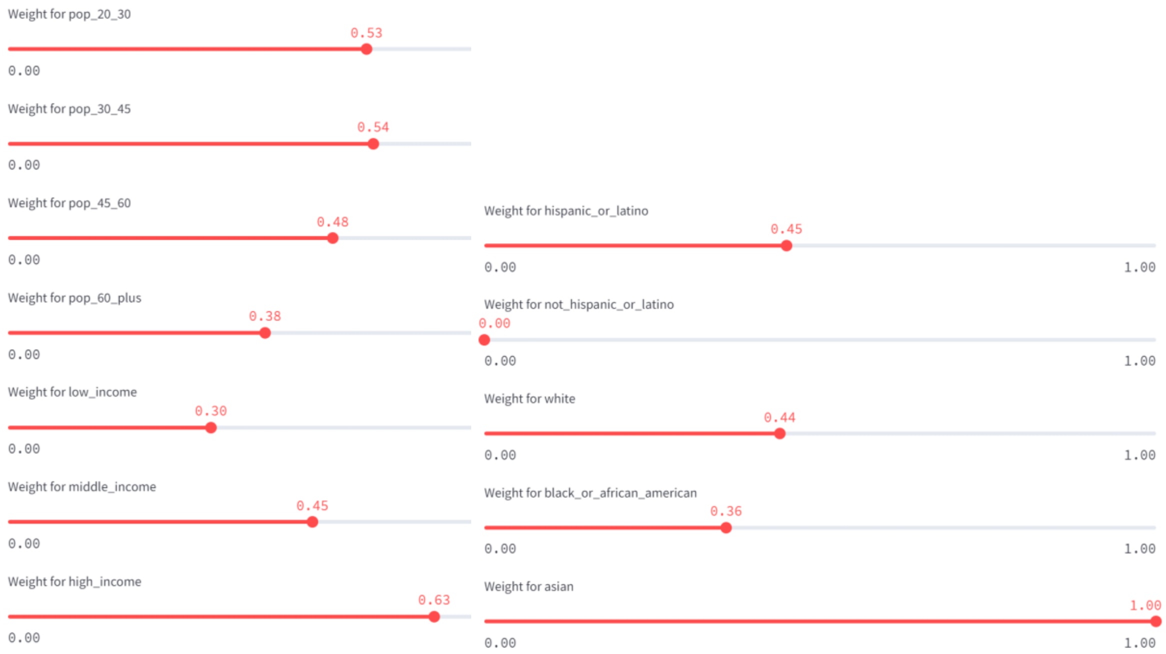
Weight for pop_45_60
0.91
0.00 1.00

Weight for pop_60_plus
0.72
0.00 1.00



Combined (age, income, ethnicity)

Once I had a base understanding of the index-based predictions based on age, income, and ethnicity individually, I created a combined heatmap



Weight for pop_20_30
0.53
0.00

Weight for pop_30_45
0.54
0.00

Weight for pop_45_60
0.48
0.00

Weight for pop_60_plus
0.38
0.00

Weight for low_income
0.30
0.00

Weight for middle_income
0.45
0.00

Weight for high_income
0.63
0.00

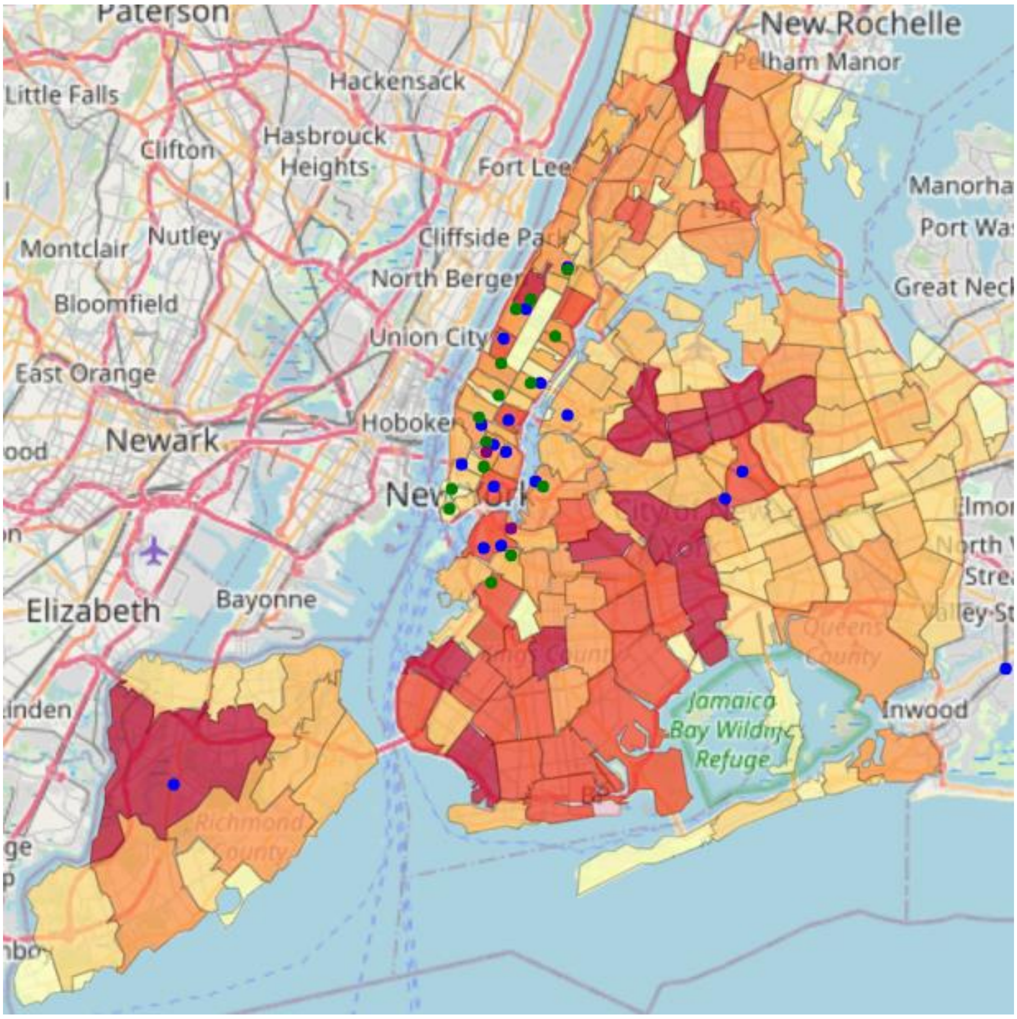
Weight for hispanic_or_latino
0.45
0.00 1.00

Weight for not_hispanic_or_latino
0.00
0.00 1.00

Weight for white
0.44
0.00 1.00

Weight for black_or_african_american
0.36
0.00 1.00

Weight for asian
1.00
0.00 1.00

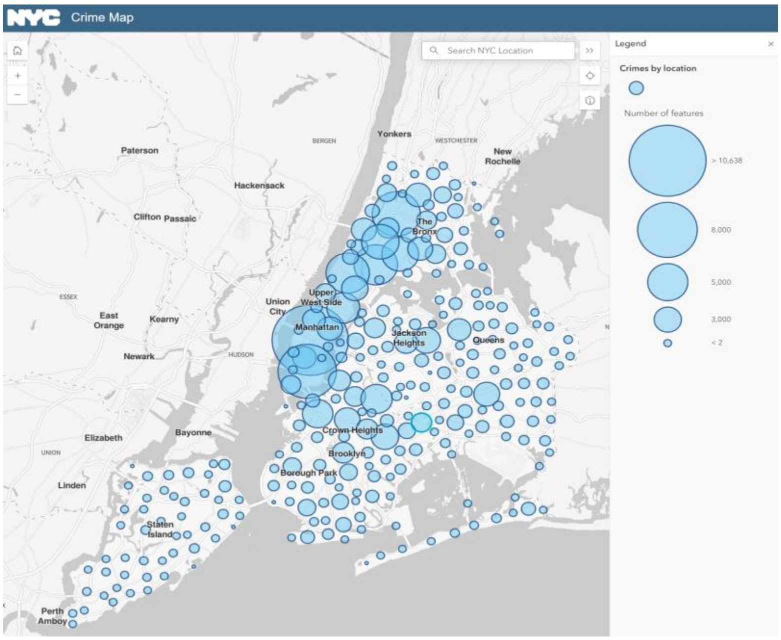


Given the initial assumptions (that the index data was applicable and accurate), I expected to see the heatmap focus on the existing TJ stores, ideally with some additional hotspots indicating they could be successful locations. Of course, that would also assume that the existing TJ stores are located in places where they maximise their success, which I believe is reasonable considering their incredibly profitable reports (ranking highest for profits per square foot).

In conclusion, this approach seems to be a dead end, as it is relatively unlikely that the many “hot” locations are “hidden gems.”

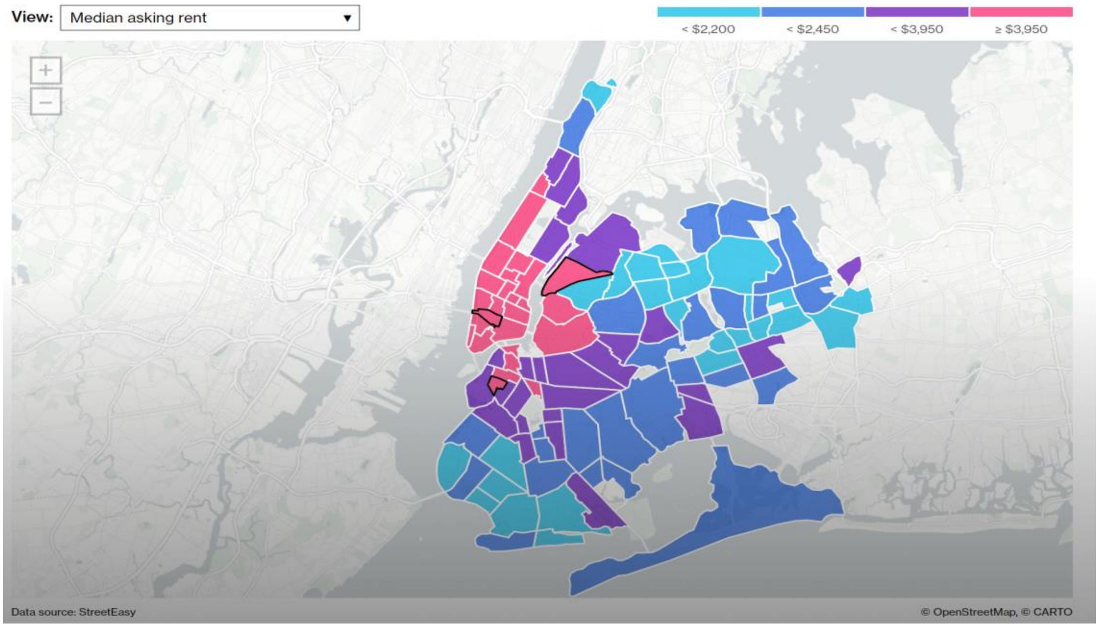
Returning to the original approach

With the demographic modeling seemingly failing to match our expectations from the heatmap, I returned to iterating through the sliders, combining TJ's regular approach of population density, higher income areas, and younger populations while cross referencing with crime rates and approximate rental prices:



Asking rent (residential) – 2022

While this metric of residential properties is two years old, it can still serve as a vague indicator of rental prices across sections of New York. This is based on the assumption that residential and business rent are related and that this two-year-old data is still representative of current rent, hence why I only claim it as a vague indicator.

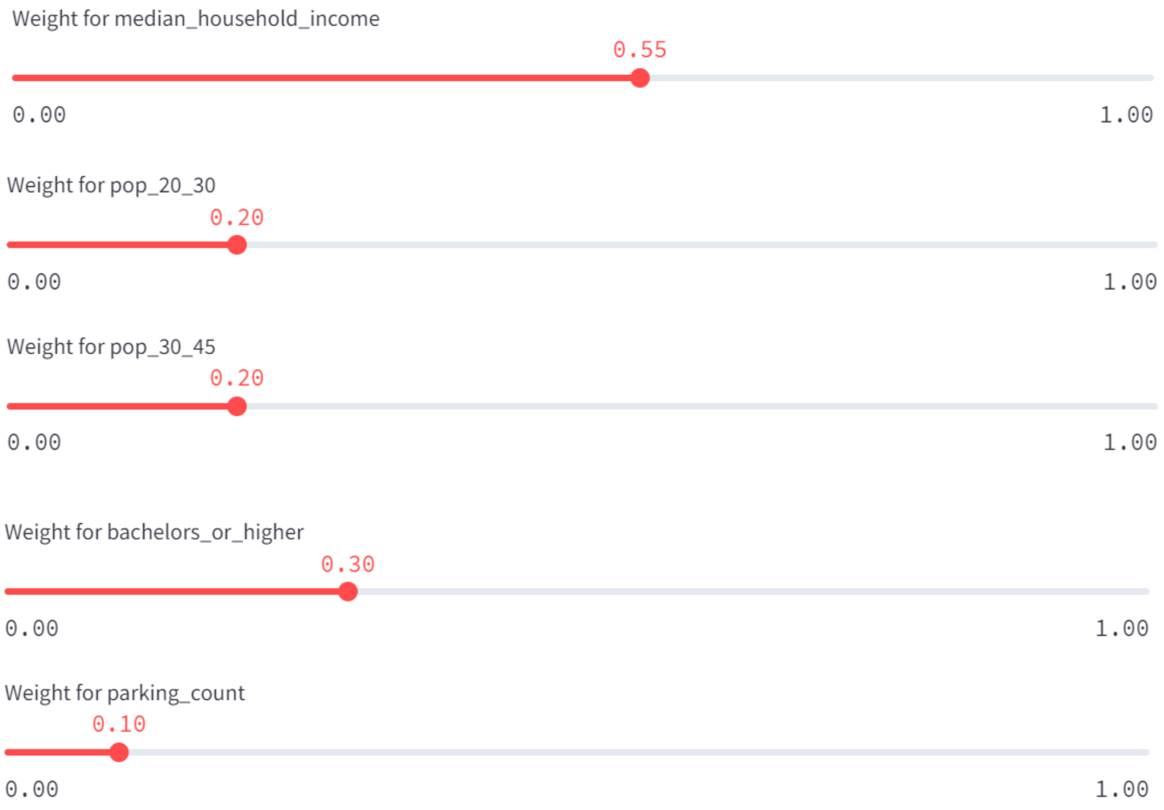


Data source: StreetEasy

© OpenStreetMap, © CARTO

Solution

After experimenting with key features including income (and income variants), age, ethnicities, population and education I found the following slider settings ideal:



Weight for median_household_income

0.55

0.00 1.00

Weight for pop_20_30

0.20

0.00 1.00

Weight for pop_30_45

0.20

0.00 1.00

Weight for bachelors_or_higher

0.30

0.00 1.00

Weight for parking_count

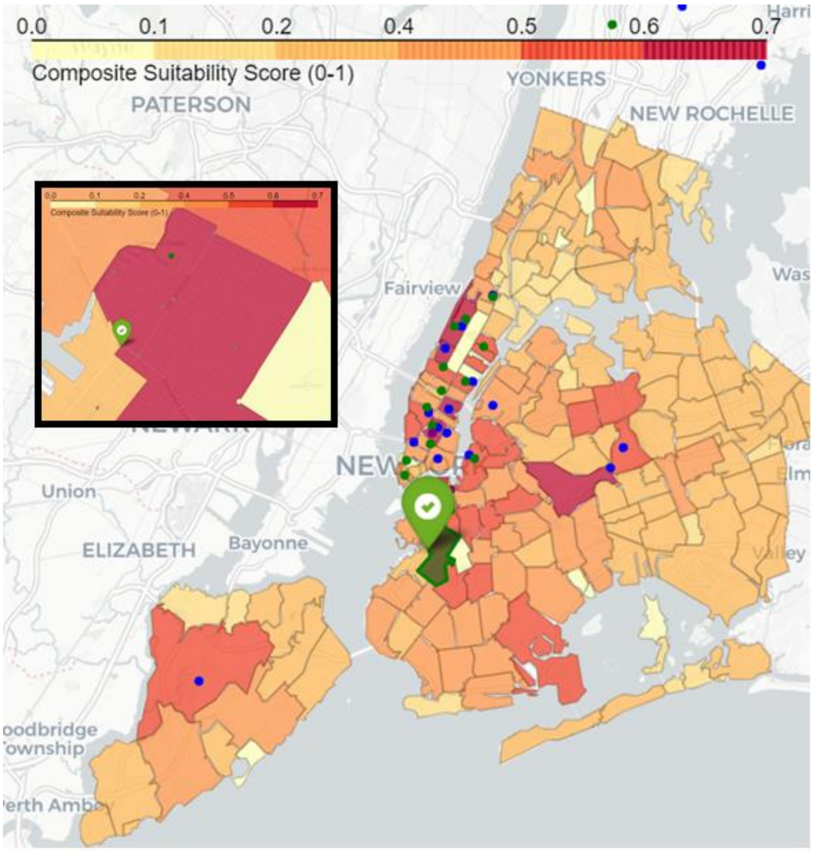
0.10

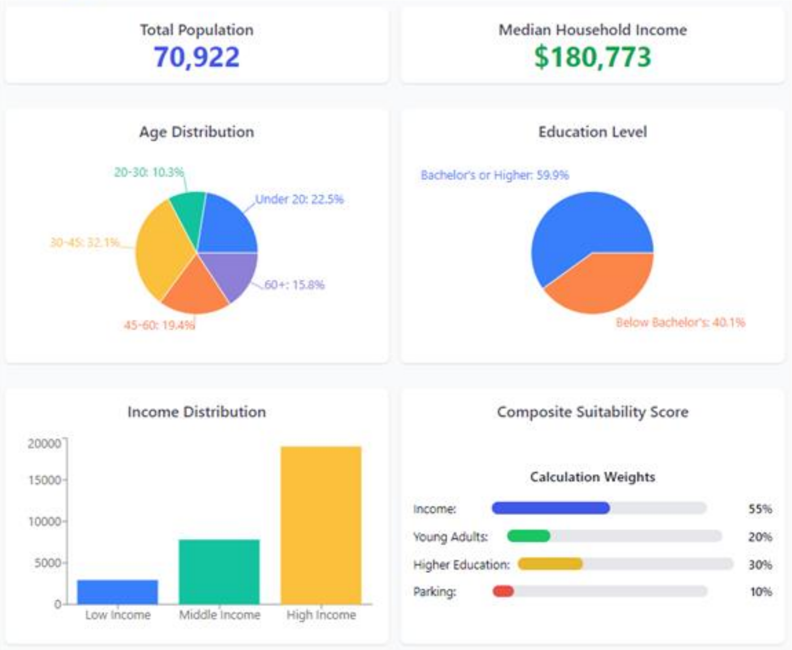
0.00 1.00

The existing Trader Joe's scored 0.56 (out of 0.74) with a standard deviation of 0.11, placing the majority of existing TJs in the red/dark red areas of the heatmap for this as well.

ZIP code 11215 showed great results in this model and matched the desirable criteria for a TJ in terms of income, age distribution, education, low crime rates (cross referenced) and with (approximate) rent below the Manhattan prices and is hence my recommendation given the resources and time at my disposal. The location highlighted on the map is close to 5 stations as well as next to an expressway so accessibility (reportedly another key criteria for TJ) is met.

Solution heatmap and figure (same as page 1)





Total Population
70,922

Median Household Income
$180,773

Age Distribution

Education Level

Income Distribution

Composite Suitability Score

Calculation Weights

Income: 55%
Young Adults: 20%
Higher Education: 30%
Parking: 10%

Question 3 has been answered throughout this document, as I intersparsed answering questions 2 and 3 after page 1 since they link together. However, for formality, here is a snippet from above referring to one answer to q3:

Given more time and access to buy third-party data from sources such as:

- Nielsen Solutions (customer purchasing habits, media consumption, lifestyle)
- SafeGraph (granular foot traffic data and visitation metrics)
- Brandwatch (Social media interests)
- Epsilon data (Demographic data, purchase behaviour, psychographic data, lifestyle insights)
- CoStar Group (Industry standard for commercial real estate data, which is necessary for modeling costs more effectively)
- Simmons consumer study (comprehensive lifestyle/brand preference study)

Adding these datasets to find correlations over a more extended period – especially social media usage and time-based foot traffic, commuting times, and turnstile information combined with Trader Joe's data on credit card transactions and profits by half-hourly data (granular) could lead to calculating how important commuting-based locations are, when shoppers shop, what kind of shops occur at what times (we could find where people regularly shop on the way home from work and thus find key commuting locations which are most important). With time to create and analyse a sufficiently large and varied dataset, the possibility of finding optimal locations should increase drastically.

However, since I didn't have access to such data and could only approximate an ideal location using second-hand information online and vaguely observing existing locations (to avoid re-implementing a distilled version of their strategy