Livability Assessment Using Big Data Analytics

Animesh Ramesh(ar8006) Efe Kalyoncu(ek2608)
Suryakiran Sureshkumar(ss16030) Prasanna A P(pa2490)

Abstract

As one of the most populous and diverse cities in the world, New York City presents a wide array of neighborhoods for residents to call home. However, quantifying the livability of different areas can be challenging due to the many factors that contribute to quality of life. This project aims to leverage open data sources to generate a data-driven livability score for neighborhoods across New York City's five boroughs. Datasets spanning transportation access, food availability, safety, public health, and local amenities will be integrated and analyzed. Key data sources include NYC open data on subway stations, bus stops, Citi Bike, restaurant inspections, recognized shop locations, 311 complaints, and health facility information. Geospatial analysis will be conducted to calculate metrics such as distance to transit, food access, green space, health resources, and more for each neighborhood. A weighted algorithm will be developed to generate an overall livability score on a scale from 1-100 for each neighborhood based on these underlying metrics. This work will demonstrate the power of unifying disparate open datasets to provide novel and actionable insights for public benefit. The methodology could be extended to other cities to generate similar livability indices.

1 Introduction

New York City is a massive metropolitan area, with varying neighborhoods of different advantages. However for an incoming person, the diversity of the city can feel overwhelming. This makes it very hard for newcomers to choose where to live, and combined with regulations regarding renting, it becomes very hard to choose where to live. To make the process of finding out where to live easier, this paper will attempt to generate a score that represents the quality of a neighborhood based on publicly available statistics about New York.

2 Data Sources

Data was sourced from:

- NYC Open Data: Subway stations, bus stops, and restaurant inspection results.
- Lyft City Bike: Trip data and station locations.
- NYPD: Complaint and shooting incident data.

3 Data

3.1 Bus Stop Shelters data

This dataset contains the location of Bus Stop Shelters in New York City provided by NYC Department of Transportation.

3.2 MTA Subway Stations data

A dataset listing all subway and Staten Island Railway stations, with information on their locations, Station Master Reference Number (MRN), Complex MRN, GTFS Stop ID, the services that stop there, the type of structure the station is on or in, and their ADA-accessibility status.

3.3 Citi Bike Trip data

This dataset contains the Citi bike trip data and station location which is publicly available on Lyft City bike site.

3.3.1 New York City Bus Data

This dataset originates from the New York City Metropolitan Transportation Authority (NYC MTA) bus data streaming service. It updates approximately every 10 minutes and includes information on bus locations, routes, stops, and more. Additionally, it features the scheduled arrival times from the bus timetable to indicate whether buses are running late, on schedule, or ahead of time.

3.4 New York City Restaurant Inspection Results data

This dataset contains food safety inspections of restaurants in NYC provided by the state from its Open Data NYC program.

3.5 Recognized Shop Healthy Stores data

This dataset contains information on bodegas and grocery stores recognized by the Shop Healthy NYC program for promoting healthier food options, including details on store names, addresses, zip codes, recognition years, and program waves. Managed by the NYC Department of Health, it highlights efforts to make healthy choices more accessible in specific neighborhoods, with annual updates on stores maintaining program standards.

3.6 NYPD Complaint Data Historic

This dataset contains NYPD Complaint Incident Level Data, providing an up-to-date snapshot of criminal complaints with details on offenses, locations, times, and participant demographics. It employs specific geocoding for accuracy and excludes unfounded or voided complaints, ensuring a focus on valid criminal incidents reported in New York City.

3.7 NYPD Shooting Incident data

This dataset contains NYPD Shooting Incident Level Data, capturing real-time approximations and updates of shooting incidents. It details handling multiple victims per incident, geocoding practices, and inclusion criteria focused on injury-resulting incidents. It also details the shooting's borough, precinct, jurisdiction, and includes demographic information about perpetrators and victims, alongside geographic coordinates in specific coordinate systems.

3.8 Health Facility Map

This dataset from NYC Open Data features up-to-date inspection results for active NYC restaurants and college cafeterias, covering violations, grades, and types of inspections over the past three years. It reflects the adjudication outcomes where scores may be revised and notes a temporary pause in letter grading due to COVID-19 from March 2020 to July 2021.

3.9 Air Quality data

This dataset, managed by the NYC Department of Health and Mental Hygiene, provides comprehensive information on air quality indicators in New York City, including pollutants like PM2.5, NO2, SO2, and O3. It offers insights into neighborhood-level air quality variations and the associated health burdens, including estimated deaths, hospitalizations, and emergency visits due to air pollution.

3.10 Facilities Database

This dataset contains information on over 30,000 facilities and program sites related to City, State, or Federal services in New York City, compiled by the Department of City Planning into the City Planning Facilities Database (FacDB). It supports planning activities and allows New Yorkers to explore government resources in their neighborhoods.

3.11 US Zip Codes from 2013 Government Data

This dataset contains the epicenters of ZIP codes across the United states. It uses 2013 government data.

4 Data Processing and Cleaning

4.1 Indexing the Data

As there were many data sources we had to use a common index to combine the data together, that would also be expressive enough to choose a neighborhood off of. ZIP codes were a natural choice here as some apartment information alredy contained ZIP codes, the area that a singular zip code encapsulates is small enough that we can expect addresses that share the same ZIP code to show similar characteristics, and it is generally useful for searching as all third party applications have support for ZIP codes.

4.2 ZIP codes

Some datasets included ZIP codes as a column. However ZIP codes were not naturally applicable for all databases so we had to match location information with a zip code. For the databases that did not include ZIP codes, a common location information provided was latitude and longitude coordinates. We combined the given coordinates with a database that contained centers of areas included by a zip code, to translate the coordinates to which ZIP code they corresponded to.

For databases that were too large to join with a ZIP code database in a naive fashion, locations were normalized to 3 mile grid, and ZIP codes were broadcast to every grid point within 15 miles. Then points at the grid points would select the ZIP code whose original (non grid) coordinates were closest to itself.

4.3 Boroughs

In creating the BoroughZipMapping, zip codes from diverse city datasets are aggregated and standardized to ensure they are in a uniform format. Each zip code is then assigned to one of New York City's five boroughs—Manhattan, Bronx, Brooklyn, Queens, or Staten Island—based on predefined numerical ranges. Zip codes that do not align with these ranges are categorized as 'Unknown'. This mapping enables effective data analysis and decision-making by allowing data with just boroughs to be mapped with zipcode.

Figure 1: Pipeline

5 Methodology

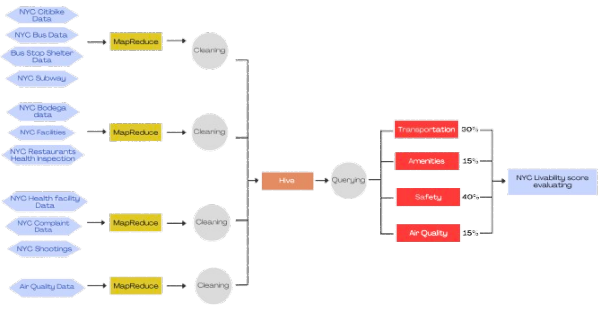

Figure 2: Scoring metric

We initially took the raw data, and cleaned it and did processing through map reduce, so that the data contained fields that would give us valuable information about how good a particular neighborhood is.

This gave us one cleaned and processed smaller dataset, per initial dataset. Afterwards, we readjusted the output of the mapreduce programs to fit the hive tabular form. We calculated the scores for each of the tables

We then divided data we found to four categories. These four categories were transportation, amenities, safety and air quality.

The individual scores are then combined to create a final score for each area, prioritizing transportation

and safety due to their significant impact on daily life and well-being, providing a comprehensive measure of the overall quality and livability of each neighborhood.

5.1 Transportation

This category aggregated data from five data sources to assess how available transportation options were for a given zip code. For this category we have aggregated total number of stop for both busses and subway stations, to see how easy it would be for an average resident to get to a shelter. We then also aggregated the data of busses, subways and citi bikes available in the area, and calculated how often transportation tools were in the area of the zip code. The point of this aggregation was that assuming vehicle location recording timings did not have an inherent bias, it would give us a good Monte Carlo estimator of how often public transportation tools pass through the neighborhood.

We considered subways to be the most convenient means of transportation, hence gave the availability of subways, and how often they pass through the neighborhood to be of highest importance for this subscore. The busses were given the second highest priority with availability of citi bikes having the least importance as the other two methods would be more useful for traveling long distances.

subway.zipcode	subway.stopname	subway.burough	subway.latitude	subway.longitude
11220	53 St	brooklyn	40.645069	-74.014034
11236	Canarsie-Rockaway Pkwy	brooklyn	40.646654	-73.90185
11220	8 Av	brooklyn	40.635064	-74.011719
11209	Bay Ridge-95 St	brooklyn	40.616622	-74.030876
11215	4 Av-9 St	brooklyn	40.670847	-73.988302
11212	Junius St	brooklyn	40.663515	-73.902447
11201	Clark St	brooklyn	40.697466	-73.993086
11421	Cypress Hills	brooklyn	40.689941	-73.87255
11219	50 St	brooklyn	40.63626	-73.994791
11204	Avenue N	brooklyn	40.61514	-73.974197

Figure 3: Subway Data in Hive

5.2 Amenities

For amenities, we have aggregated data about the facilities that are available in a zip code as well as data about restaurant inspections, and shops in the area. For this score, we gave the utmost importance to facilities, as we believed that important facilities like generic grocery stores, laundromats and such would mean that most needs of individuals could be met within the neighborhood. We put the second most importance to the restaurants and their health inspection scores. We believed that existence if restaurants with high scores would mean that there are places for people to have fun in the area. Also ensuring that the restaurants had high health score ratings made us more confident that the restaurants had to put out high quality food.

facilities.zipcode	facilities.burough	facilities.facilityname	facilities.facilitytype
10310.0	staten island	BBM - SI STREET MAINTENANCE	MAINTENANCE, MANAGEMENT, AND OPERATIONS
11235.0	brooklyn	NW MALCOLM X BOULEVARD & RAINBRIDGE STREET FOOD DROP-OFF SITE	COMPOST
11234.0	brooklyn	SUNKEN PLAYGROUND	PLAYGROUND
11231.0	brooklyn	DESTINATION AUTOMOTIVE & RECOVERY INC.	TOW TRUCK COMPANY
10460.0	bronx	CROTONA PARKWAY MALLS	MALL
11101.0	queens	MURRAY PLAYGROUND	NEIGHBORHOOD PARK
11216.0	brooklyn	BEST BUY ELECTRONICS DROP-OFF SITE	ELECTRONICS
11203.0	brooklyn	BROOKLYN EXPRESS TOWING INC.	TOW TRUCK COMPANY
10471.0	bronx	WOODLY AVENUE PRIMARY CARE CENTER	HOSPITAL EXTENSION CLINIC
11354.0	queens	LAZ PARKING NEW YORK/NEW JERSEY, LLC	COMMERCIAL GARAGE

Figure 4: Facilities in Hive

5.3 Safety

We used three metrics to assess the safety of a neighborhood. These were violent crimes, NYPD complaints and availability of hospitals. We have further classified the complaint report based on the nature of the complaint, where complaints that were classified as more severe, were weighted more. We have prioritized the weight of the shooting data, because harsher crimes are a bigger detriment to living in a neighborhood.

complaint.burough	complaint.lawcatcd	complaint.reportdate	complaint.latitude	complaint.longitude	complaint.zipcode
QUEENS	MISDEMEANOR	07/01/2022	40.650928	-73.736827	11422
QUEENS	VIOLATION	08/02/2022	40.77933	-73.750192	11363
QUEENS	MISDEMEANOR	10/15/2022	40.76564	-73.730088	11429
QUEENS	VIOLATION	05/25/2022	40.654003	-73.744662	11422
QUEENS	FELONY	03/07/2022	40.702502	-73.747998	11412
QUEENS	MISDEMEANOR	11/30/2022	40.60277623007 81	-73.75011736483 31	11691
QUEENS	VIOLATION	05/05/2022	40.730525	-73.724721	11426
QUEENS	FELONY	06/23/2022	40.718883	-73.737606	11428
QUEENS	VIOLATION	04/04/2022	40.749444	-73.709759	11426
QUEENS	VIOLATION	03/05/2022	40.675542	-73.754422	11413

Figure 5: Complaint Data in Hive

5.4 Air Quality

Air quality scores are determined by normalizing and inverting the pollutant levels of PM 2.5, NO2, and O3 against their respective health-based threshold values to create a scale where a higher score indicates better air quality. These scores are then weighted (50% for PM 2.5, 30% for NO2, and 20% for O3) to reflect their relative importance to health outcomes.

The overall score for each borough is computed as a weighted average of these scores. The decision to use only the latest year available, 2022, ensures the data's relevance and accuracy, reflecting the most current environmental conditions and regulatory standards, which is crucial for effective public health analysis and policy making.

5.5 Final Score

To calculate the final score we combined all the previous scores as well as air quality data. For the final score, majority of the score was allotted to the safety and transportation score. Both lack of transportation as well as lack of safety, are vital for day to day lives of most people. Also other amenities or restaurants being good would be inconsequential if the streets of a neighborhood did not feel safe.

6 Result and Analysis

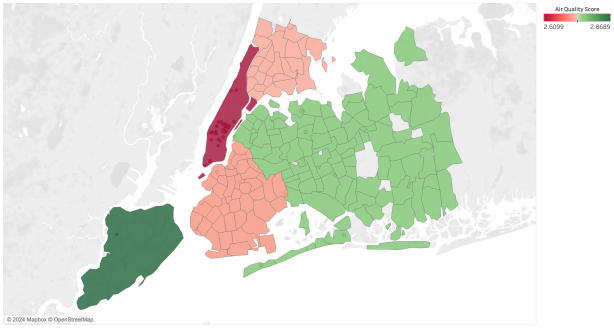

Figure 6: Air Quality Heat Map

The air quality map showed that areas with low population density has much better air quality. Manhattan ZIP codes with high population density tend to have worse scores in air quality whereas more open areas around Queens and to a lesser extent Brooklyn have better air quality.

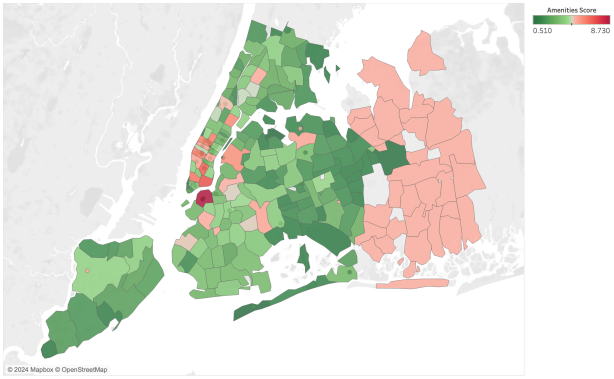

Figure 7: Amenities Heat Map

We see that there are relatively more amenities outside of city center, with Manhattan having the least amount of amenities, whereas Queens and Staten island having the most. However this can be attributed to the fact that the zip codes in these areas cover a larger area, meaning that for the two ZIP codes in Manhattan and Queens to have the same convenience, the one in the Queens may have needed more amenities. This might show that additional datasets that represent the area, as well as population density of a given ZIP code could be used to further normalize the data against these variables to create a database with more informative queries.

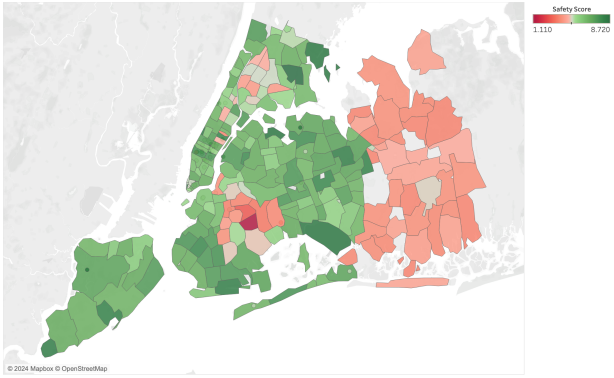

Figure 8: Safety Heat Map

Looking at the safety map, we see that majority of Brooklyn and Manhattan as well as Staten Island being safe. Certain areas of Bronx and Queens also appear to perform well in this metric, however places that are too far off the center start to suffer in safety score.

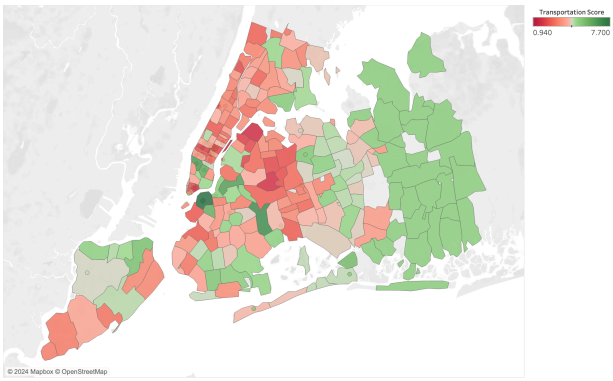

Figure 9: Transportation Heat Map

Transportation map is one where trends are a lot harder to observe. We see certain central transportation hubs having high scores, such as Herald Square area and Downtown Brooklyn. Furthermore neighborhoods that connect different areas such as Staten Island seaside toward New Jersey also doing well. Lastly, eastern Queens also performs really well in this metric. However his is another map where size of the ZIP code matters. If the neighborhood is just bigger, a transportation vehicle is more likely to exist there for longer.

Furthermore, subways are inherently likely to spend disproportionate times at their stops compared when they are moving. This means that if there are a lot of subways that pass through an area, which would typically indicate subways stops nearby the neighborhood, it is very hard to infer this information from the data that we have. As a result Manhattan which has been divided finely by neighborhoods have areas that suffer in this metric, despite there being available transportation options at neighboring neighborhoods for each ZIP code.

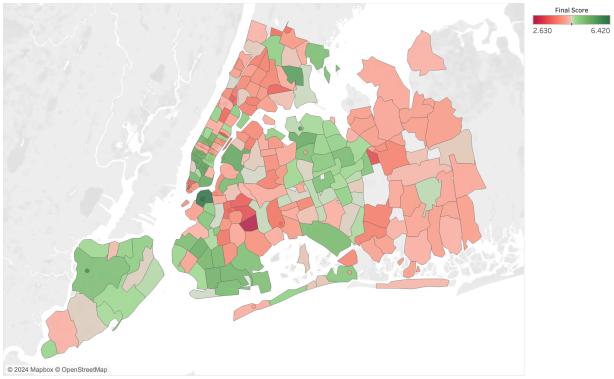

Figure 10: Livability Heat Map

Looking at our weighted cumulative scores, the clear winner is Downtown Brooklyn/Williamsburg areas. Lower Manhattan Staten island, and certain neighborhoods of Queens also score fairly well. In particular the top 5 neighborhoods based on our metric are:

1. Downtown Brooklyn
2. Staten Island
3. Flushing
4. Parkchester
5. Greenwich

Comparing these finding to the AreaVibes rankings we see that 3/5 of the ones we selected were also in top 10 at AreaVibes. However Flushing and Staten Island have over performed when using our weighted score. This can mostly be attributed to the fact that these areas had high transportation scores due to availability of public transportation, however the high commute times for people living in these areas were not factored in when processing the data.

7 Future Work

While our study made good use of publicly available data to create robust metrics to judge neighborhoods in New York, we could further improve these metrics by factoring in geological data. In particular proximity to other ZIP codes typically deflate transportation and amenity scores because neighborhoods that are more isolated need to be more sufficient in these areas

for them to function, however in general isolation of a neighborhood would be considered a negative.

A similar problem also comes up with transportation scores where some areas have high transportation scores, because there needs to be a lot of public transportation in the area for its populace to be able to get to the city center, however the fact that these neighborhoods have high commute time is not represented in the database.

Also we considered complaint and shooting data to create a safety score, however this metric is also inversely corralated with population density, because if two areas are similarly safe, but one of them is twice as dense, there will be more crime in the area, which would reduce the safety score.

References

https://opendata.cityofnewyork.us/ - Data Source

https://areavibes.com - Used to compare our results.