Table of Contents

I. Summary

The Metropolitan Transportation Authority New York City allows New York to move 24 hours a day and seven days a week across the boroughs of Manhattan, Brooklyn, Queens, and the Bronx, with an additional MTA Staten Island Railway that links 22 communities. As one of the largest public transportation agencies in the world, the average ridership is 2.4 million in a day in the five boroughs. In lieu of being an epicenter of a pandemic, questions arise regarding access to health facilities. Therefore, an analysis has been conducted to explore and answer these questions. How accessible are health facilities to every New Yorker who relies on public transportation (subway stations and bus stop shelters)? Do the five boroughs have a sufficient number of bus stop shelters compared to health facilities? Is there a relationship between the population size of a zip code and the number of health facilities? How is the number of health facilities in Midtown Manhattan, the center of attraction, different from the median number of health facilities?

II. Data Description

- **Bus Stop Shelters** (NYC Open Data) contains the longitude, latitude, brough name, and street name for every bus stop shelter in New York City.
- **Health Facility** (New York State Department of Health) contains the facility name, facility longitude, facility latitude, and facility zip code for every health facility in New York State.
- **Subway Stations** (NYC Open Data) contains the station name and geometry points for every subway station in New York City. *The limitations of this dataset are the lack of zip code or borough information and combining of a longitude and latitude points into one column.*
- **Zip Codes** (Open Data Soft) contains the zip code and assigns it to a borough. *The limitations of this dataset is the lack of longitude and latitude information.*
- **NY Population** (New York Demographics) contains the zip code and the population in each zip code for New York State.

III. Data Manipulation

- **Step 1: Data Cleaning**

(a) Health Facility and (b) NY Population contain New York State information, which requires anything outside of the five boroughs to be excluded. This was performed by matching every zip code to the Zip Code dataset and assigning the borough. Entries without an identified borough counts as outside the boundaries. (c) Subway Stations do not contain columns for longitude and latitude, requiring these two fields to be extracted from the column of coordinate points.

- **Step 2: Matching**

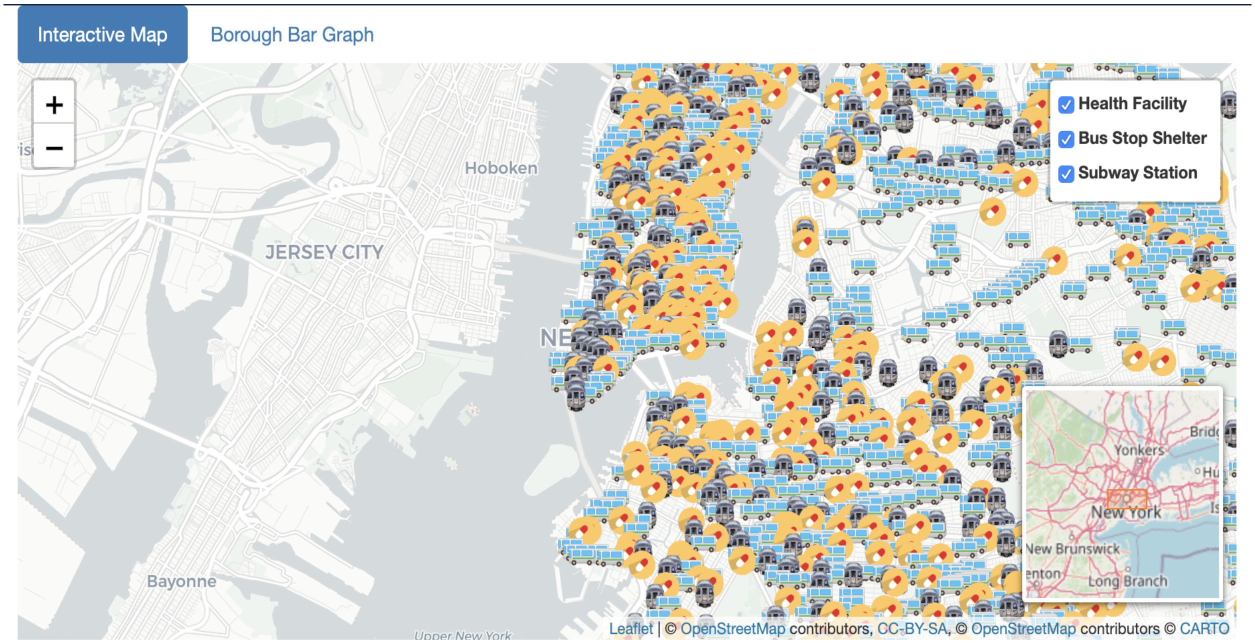
(a) Bus Stop Shelters and Health Facility contain zip code information, which are matched with Zip Codes to assign a borough. These were combined in a barplot to visualize the ratio between the number of health facilities and number of bus stop shelters in the five boroughs.

- **Step 3: Combining**

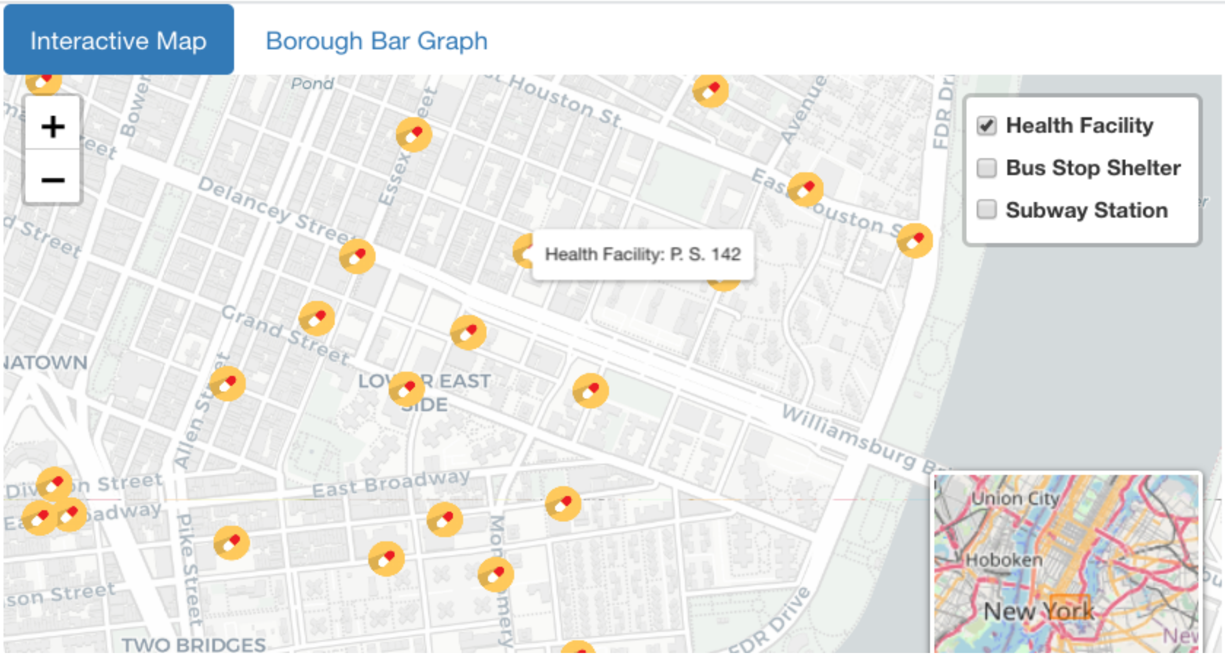
(a) Bus Stop Shelters, Health Facility, and Subway Stations are combined to create an interactive map to allow a user to identify a facility and its nearest shelters and stations when hovered.

(b) Health facility zip codes are grouped and tallied to determine the number of health facilities in each zip code. A left join was performed to merge zip codes with their respective population counts. There are zip codes that do not have population information, which result in NA. These are removed through na.omit(dataset_name) before creating the visualization.

Data Analysis: Data Visualization
(a) R Shiny: Interactive Map



☑ Health Facility
☑ Bus Stop Shelter
☑ Subway Station



☑ Health Facility
☐ Bus Stop Shelter
☐ Subway Station

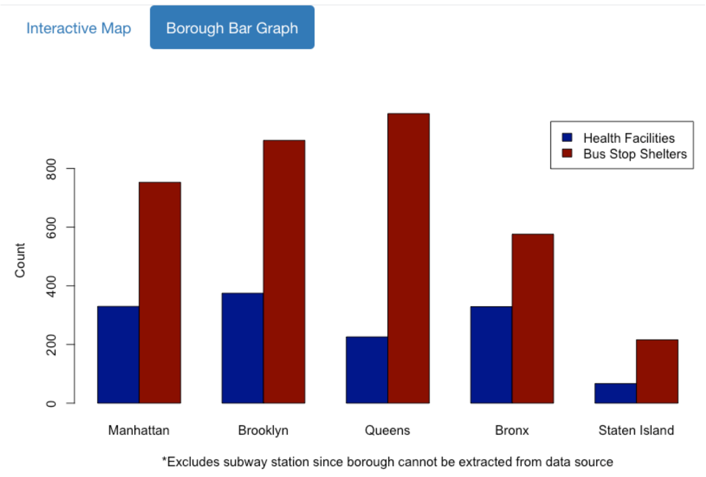
Health Facility: P. S. 142

This is an interactive map created mainly with leaflet and R Shiny. It shows the location of health facilities, bus stop shelters, and subway stations, one at a time or everything at once. There is also a mini map to allow the user to have a glimpse of a bigger map if they are too close to a certain location on the map.

The purpose of this map is to understand the distance between health facilities and public transportation and to answer the question: *How accessible are health facilities to every New Yorker who relies on public transportation (subway stations and bus stop shelters)?*

(b) R Shiny: Borough Bar Graph

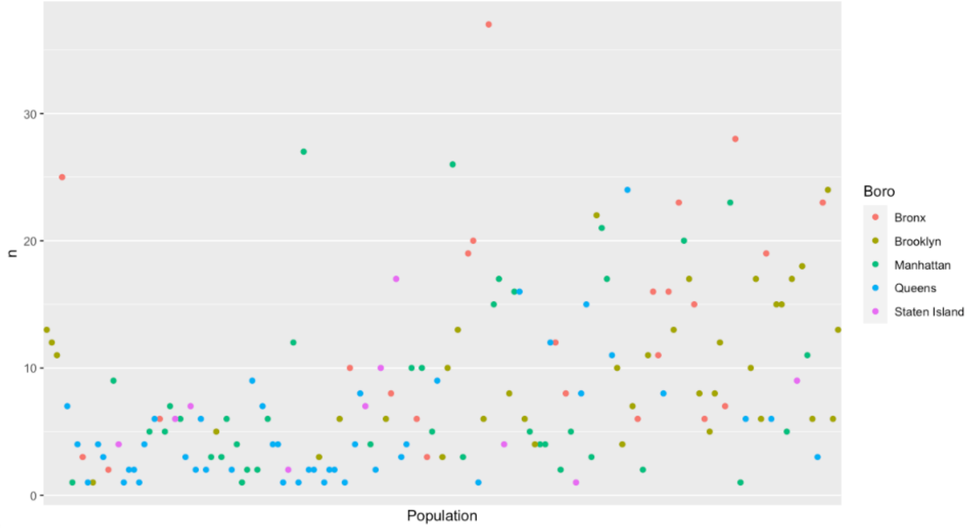
This is a static bar graph merging information from Health Facility and Bus Stop Shelters into a side-by-side comparison of how many health facilities there are compared to the number of health facilities in each borough. Evidently, there are more bus stop shelters in Queens yet the highest number of health facilities is Brooklyn. However, it can be observed that there are more buses than health facilities. When accompanied by the interactive map, there are busses near every health facility in New York City.



*Excludes subway station since borough cannot be extracted from data source

The figure answers the question: *Do the five boroughs have a sufficient number of bus stop shelters compared to health facilities?*

(c) Scatter Plot



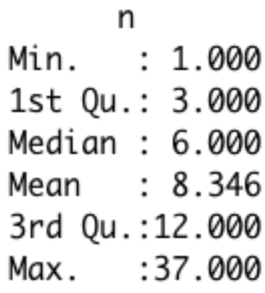
This figure shows the relationship between the population size of each zip code in the five boroughs and the number of health facilities in each zip code. The colors identify which borough each point is from with the hopes of inferring a relationship. Evidently, population size does not determine the number of health facilities. For example, zip code 10461 has 37 health facilities in the Bronx. However, 11368 in Queens has the highest population size.

This figure answers the question: *Is there a relationship between the population size of a zip code and the number of health facilities?*

IV. Data Analysis: Data Summary

The figure displays the summary statistics for the number of health facilities per zip code. The minimum number of health facilities is 1, an example is 10004 in Bowling Green.

A function to identify the mode was created, which equates to 6. There are 20 zip codes with 6 health facilities.

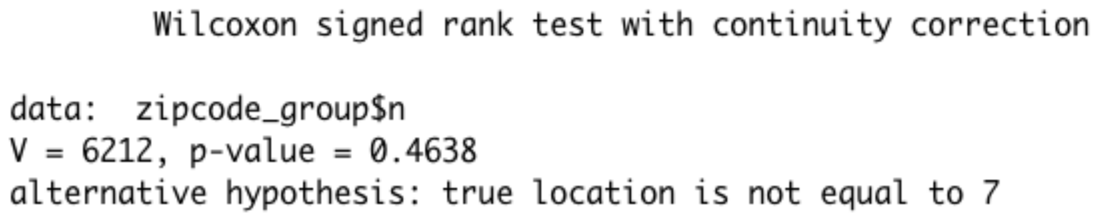


```
       n
 Min.   : 1.000
 1st Qu.: 3.000
 Median : 6.000
 Mean   : 8.346
 3rd Qu.:12.000
 Max.   :37.000
```

V. Data Analysis: Wilcoxon Signed-Rank Test

A Wilcoxon Signed-Rank Test was performed to avoid the assumption of the data on zip codes and health facility count are normally distributed.

Question: Since Midtown Manhattan (10001) has the most inflow and outflow of tourists and commutes from different states on a daily basis, a researcher believes that it reflects the median number of health facilities in the five boroughs of New York. The number of health facilities is 7. Is this significantly different from the true location?



```
	Wilcoxon signed rank test with continuity correction

data:  zipcode_group$n
V = 6212, p-value = 0.4638
alternative hypothesis: true location is not equal to 7
```

The p-value is 0.4638, which is greater than 0.05. This means that it is not statistically different/significant and indicates a strong evidence for the null hypothesis. Therefore, Midtown Manhattan (10001) is not statistically different from the true location.

VI. Data Analysis: Model

(a) R Shiny: Interactive Map

How accessible are health facilities to every New Yorker who relies on public transportation (subway stations and bus stop shelters)?

Model Building and Final Model: When creating an interactive map mainly incorporating three datasets and matching each dataset into a fourth dataset, the main limitation is with the subway station data since it does not contain a column for longitude and latitude. A solution was to use regular expressions to extract the longitude and latitude from one column. This successfully allowed subway station data to be plotted in the interactive map.

Model Fit and Pitfalls: It accurately depicts the location of each entity as well as providing the user the ability to select which aspects of the map should be displayed. The data is also cleaned and there are no points that exceed the boundaries of the five boroughs. It accurately paints the picture for each geographic location. A potential pitfall is that some health facilities did not have coordinate points that had to be excluded.

Model Interpretation: Every point on the map can be hovered to display the name of the bus stop shelter, subway station, or health facility. Depending on the health facility of choice or an individual's closest subway station or bus stop shelter, they can easily locate the nearest health facility. The model answers the question since most of the health facilities are covered by subway stations and bus stop shelters. An attempt to identify the exact distance, finding the nearest entities was done yet did not work since a way to exclude using bodies of water and illegal pathways could not be determined.

(b) R Shiny: Borough Bar Graph

Do the five boroughs have a sufficient number of bus stop shelters compared to health facilities?

Model Building and Final Model: However, subway station data do not contain zip codes and boroughs to incorporate it in a bar graph. A solution was to identify the longitude and latitude boundaries for each borough and aggregate the subway station data accordingly. However, the attempt failed and it could not be incorporated with the given data. Other datasets were incorporated yet provided the same outcome. Therefore, the bar graph is able to produce a side-by-side comparison of bus stop shelter data and health facility data by borough.

Model Fit and Pitfalls: This is nearly perfect, other than the pitfall of not having data on subway stations. Other than that, this accurately depicts the ratio between boroughs and between health facilities and bus shelter stations. A potential pitfall is that some entries did not have zip codes that had to be excluded.

Model Interpretation: In an aggregated perspective, the five boroughs have a sufficient number of bus stop shelters to provide access to health facilities.

(c) Scatter Plot

Is there a relationship between the population size of a zip code and the number of health facilities?

Model Building and Final Model: The overall goal of the project is to have a glimpse of access to health facilities in the New York City area, which means that it is necessary to incorporate the population. A scatter plot serves such a purpose since it can provide the relationship between the number of health facilities in a zip code and the population size. In addition, a scatter plot allows grouping based on boroughs to show possible hidden relationships.

Model Fit and Pitfalls: A pitfall of the scatterplot is its inability to display every point on the graph since there are outliers in the data. A potential pitfall is that the data for the population size comes from a source outside of the New York State and New York City data resources that it could not contain some data and vice versa especially when the left join was conducted.

Model Interpretation: There is a weak relationship between the population size and the number of health facilities in a zip code.

(d) Wilcoxon Signed-Rank Test

Model Building and Final Model: A one-sample t-test was the first attempt in hypothesis testing. However, it was giving incorrect information and an unreliable p-value when compared to the summary statistics. A solution to see the problem was creating a histogram, which showed that the data is not skewed. This is not displayed in the report since it was a rough glimpse of the data without data cleaning. Since it is important to not assume the normalcy of the data, a Wilcoxon Signed-Rank Test is best to use.

Refer to **Data Analysis: Wilcoxon Signed-Rank Test** for the test results.

VII. Overall Limitations

- There is an attempt to identify the nearest bus stop shelter from every health facility. However, the resulting data did not reflect the actual bus stop shelter closest to the health facility since it is not based on the streets or lines that are actually able to be used. It included the bodies of water and the parks that may or may not be used to walk.
- There is no data on every bus stop in New York City that could be implemented with the other datasets available. Therefore, this provides a smaller number of bus access since it only includes bus stop shelters.
- The subway station data does not contain zip code or borough information, which limits its ability to be matched and incorporated with other datasets and figures.
- The zip code data and the population data do not contain the same number of zip codes when merged due to NA.

VIII. Potential Improvements

- Deploy the R Shiny web application
- Incorporate subway lines and bus lines to determine the distance

IX. Conclusion

In conclusion, the number of health facilities in a zip code does not rely on the population size, according to the scatter plot. However, with New York City's vast public transportation system, health facilities are accessible to New Yorkers who rely on public transportation, specifically subway stations and bus stop shelters, according to the interactive map. In addition, the five boroughs have a sufficient number of bus stop shelters compared to health facilities based on the ratio provided by the bar graph. However, a potential improvement is to identify the distance between a facility and its nearest bus stop to accurately determine if each facility is indeed accessible via public transportation. An attempt to do so was conducted yet unable to fully capture the intended purpose since it did not account for the bodies of water and areas where it is not suitable to walk, drive, and etc.

It can also be inferred from the hypothesis test that Midtown Manhattan, which is the center of attraction in New York City, contains a number of health facilities not significantly different from the median number of health facilities in New York City. This is beneficial to know since New York is a popular tourist destination and the number of people who truly occupy Midtown Manhattan constantly changes, which means that the median number does reflect.

Further research can be conducted based on a number of inferences taken from the models created. For example, since there is a weak relationship between the number of health facilities and the population size for each borough, what could be the reasoning behind how health facilities are established? In addition, numerous bus stop shelters evidently pass a health facility. However, how accessible is a health facility and what is the average number of transfers one must take to get to a health facility in a different borough? Lastly, how do you determine the distance between a health facility and a bus stop shelter excluding the areas that are not suitable for transportation? These are questions that arise from the data analyses conducted and could lead to more analysis on this topic.

X. Appendix

Data Sources:
Bus Stop Shelters in New York City
https://data.cityofnewyork.us/Transportation/Bus-Stop-Shelters/qafz-7myz

Health Facilities in New York State
https://health.data.ny.gov/Health/Health-Facility-Map/875v-tpc8

Subway Stations in New York City
https://data.cityofnewyork.us/Transportation/Subway-Stations/arq3-7z49

Zip Codes in New York State
https://public.opendatasoft.com/explore/dataset/us-zip-code-latitude-and-longitude/export/?q=&refine.state=NY

New York State Population
https://www.newyork-demographics.com/zip_codes_by_population

R Sources:
R Shiny
https://spatialanalysis.github.io/workshop-notes/interactive-maps-with-shiny.html

Leaflet
https://www.rdocumentation.org/packages/leaflet/versions/2.0.3/topics/leafletOutput