Developing a Social Determinants of Health Model to Unravel Asthma Drivers in New York City

A Multidisciplinary Approach to Investigate the Effects of Urban Green Spaces on Asthma, and its Association with Economic and Temperature metrics

Final Project Report for p8105

Team Members:

Meritxell Mallafré Larrosa (mm5951)

Tim Hauser (th2899)

Emma Warshaw (ew2718)

Niklas Hess (nh2706)

Esther Kim (ei2262)

Table of contents

Introduction

Motivation & Project Goals

Roofs planted with vegetation — known as green roofs — can help cities adapt to a changing climate by absorbing storm water, lowering local temperatures, and providing insulation that cuts indoor heating and cooling costs. Yet green roofs cover less than 0.1% of New York City's 1 million buildings, according to a new analysis performed by colleagues at Columbia University.

Our motivation for this project revolves around understanding how rooftop greenspace availability relates to asthma rates, economic realities, and temperature metrics in each borough. This project allows us to use a social determinants of health model, utilizing the following key themes that we've been learning as public health students:

1. Location Matters: Anyone active in Public Health will know that location is one of the most important determinants of health. One of the key examples of this is redlining - referring to a discriminatory US practice that saw predominantly black neighborhoods being disadvantaged by withholding loans, public spending, and more. While redlining as a "known" practice might be an atrocity of the past, the underlying importance of a person's living location continues to matter. Just think about implicit biases, access to education, environmental factors, and density of healthcare providers. As such, it was very clear to this group that our project needed to incorporate some type of location factor.

2. Public Health is EVERYTHING: One of the most influential lectures for this group at Mailman was with Professor Merlin. He started by describing what Public Health really meant. Just healthcare? Gun violence? Income distribution? Access to work? Clean water? The correct answer: All of the above.

   Public Health is a unique field, as it is much broader than most people imagine - everything a person sees in daily life is connected to Public Health in one way or another. As such, it was important to this group to touch upon a number of different factors while working on this project. To us, only by considering a number of factors, could we do justice to Professor Merlin and the education we have received at Mailman.

3. Systems Thinking: Of course, when considering a number of different Public Health factors, some key challenges arise. Most importantly, "how do those factors interact with one another?"

   Within Public Health, Systems Thinking models tackle this very nuanced question. Simply put, your access to education influences your access to work, which in turn influences your income and investment within your community, which ultimately influences your neighborhood's access to education.

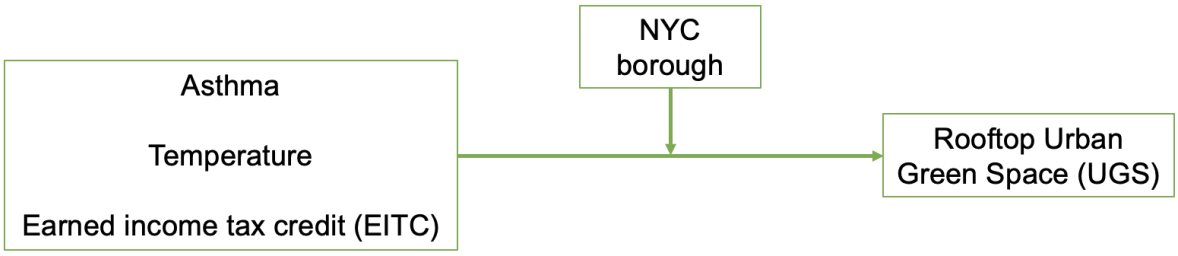
For this group, applying systems thinking meant looking beyond individual analysis, but combining all datasets to understand how green spaces, income, health, and temperature could all influence one another.

Related Work

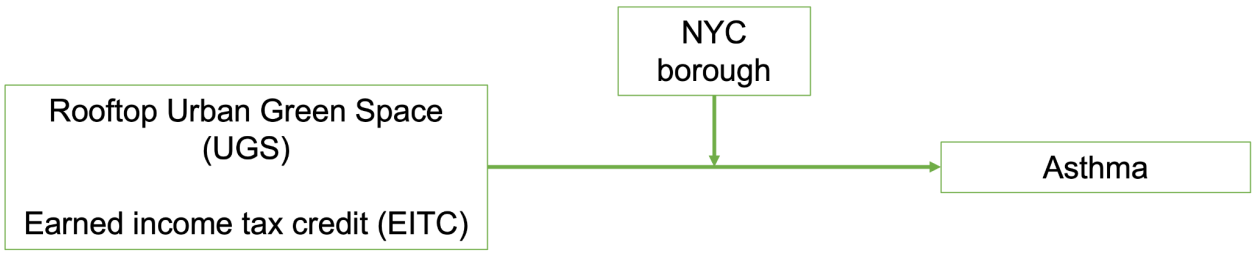
This project was inspired by the data set from Examining the distribution of green roofs in New York City through a lens of social, ecological, and technological filters. This study examined the presence of rooftop greenspaces in New York City, noting that although about 20% of NYC landscape is covered in buildings, rooftop greenspace is only detected in a small fraction of those areas. We were interested in comparing rooftop greenspace to additional variables that can give a better overall understanding of the social determinants of health in the five boroughs.

Research Question(s)

Our initial goal was to understand how rooftop urban green space correlated with 3 other predictors individually: asthma rates, temperature data, and earned income tax credit (EITC) amounts in each borough. We originally hoped to use temperature, asthma rates, and EITC as predictors of cumulative green space area, stratified by borough.



As we progressed through the project, we began to rethink our outcome variable for the overall model. Since we're interested in public health outcomes, we decided to use asthma as the outcome variable in a full regression model. Because of data limitations and a univariate comparison between our temperature data and greenspace proved to not be significant, we excluded temperature from the overall model.



Manipulating these data sets taught some valuable lessons, namely how much effort goes into data collection and wrangling. It also showed us that studying the social determinants of health

is extremely challenging. We were able to piece together data sets that showed us the individual elements we wanted to model, but the reality is that it would be better if all of these factors were coming from a single unified, health-systems informed study. That being said, we were still able to find meaningful insight with what we had to work with.

Methods

The methods section is structured into 4 subsections, entailing the data source, goals of the analysis, wrangling, and codebook, on : (1) Urban Green Space, (2) Asthma, (3) Temperature, and (4) Tax. The results section includes a latter (5) Joint Analysis.

1. Urban Green Space

Data Source

We are presenting a **secondary analysis of the publicly available dataset**, which includes data from n=736 UGS as of 2016-2018. In complementary analyses found within the project website, we **draw links to economic, clinical and environmental factors** which could potentially be associated with the distribution of UGS within New York City (see subsections 2-4).

The primary dataset was developed by The Nature Conservancy's New York City Program (Mike Treglia and Emily Maxwell) with contributions (data aggregation and additional support) from Timon McPhearson of The Urban Systems Lab at The New School, Eric Sanderson of The Wildlife Conservation Society, and Greg Yetman of CIESIN at Columbia University.

Treglia, Michael L., McPhearson, Timon, Sanderson, Eric W., Yetman, Greg, & Maxwell, Emily Nobel. (2018). Green Roofs Footprints for New York City, Assembled from Available Data and Remote Sensing (Version 1.0.0) Available here. Zenodo. http://doi.org/10.5281/zenodo.1469674

Goals of the Analysis

The **primary goal** of this analysis is to provide a primary dataset of NYC Urban Green Spaces upon which regression analysis can be conducted with potentially associated dimensions (temperature, econometrics, and health outcomes).

Secondly, this analysis aims to provide:

- a **descriptive analysis** of the NYC Urban Green Spaces (UGS) primary dataset, as well as **stratified analysis** at the City Council District and Borough levels;
- relevant **data visualizations** of the above (notably including a map using the leaflet package);
- an attempt to fit the dataset into a **linear regression model** to elucidate differences in the distribution of UGS surfaces at the City Council District and Borough levels.

Data Load & Wrangling

The **primary Urban Green Spaces (UGS) dataset** is loaded and wrangled to generate a **UG_df dataframe**. The dataset is available in the "data" folder within the project's repository. The following procedure is then applied:

- select relevant variables using select()
- rename certain variables for comprehension using rename()
- reorder the variables using arrange()
- drop entries with now total green area available using drop_na()
- calculate percentage of green coverage of each building, as well as recode() borough names and ownership status, and convert sqft into sqm, using mutate()
- note owner types are aggregated into "Public", "Private" and "Other" as per mapPLUTO definitions (refer to the codebook for more details). Given the presence of "na", the replace_na() function is employed.
- coerce "borough" into a factor variable and apply fct_relevel() to set Manhattan first (then becoming the reference category for regression analyses below)

Notably, the same authors provide an aggregated dataset at the City Council District which is loaded as a **secondary UGS dataset** under the **UG_agg_df** name. A similar procedure as described above is used.

Data Codebook

Primary Dataset: Variables codebook

ID and location data

- fid - Unique identifier
- address - Address based on MapPLUTO, joined to the dataset based on bbl.
- borough - Borough abbreviation pulled from MapPLUTO.
- ownertype - Owner type field pulled from MapPLUTO (C: City ownership, M: Mixed city & private ownership, O: Other – owned by either a public authority or the state or federal government, P: Private ownership, X: Fully tax-exempt property that may be owned by the city, state, or federal government; a public authority; or a private institution; blank: Unknown (usually private ownership))
- zonedist1 - Zoning District 1 type pulled from MapPLUTO.
- longitude - Longitude in decimal degrees.
- latitude - Latitude in decimal degrees.

Urban Green Space data

- green_area - Total area of the footprint of the green roof as per this data layer, in square meter, calculated using the projected coordinate system (EPSG 2263).
- building_area - Total area of the footprint of the associated building, in square meter, calculated using the projected coordinate system (EPSG 2263).

- prop_gr - Proportion (%) of the building covered by a green roof according to this layer (gr_area/bldg_area).
- cons_year - Year the building was constructed, pulled from the Building Footprint data.
- heightroof - Height of the roof of the associated building, in meter, pulled from the Building Footprint Data.
- groundelev - Lowest elevation at the building level, in meter, pulled from the Building Footprint Data.

NYC building identifier

- bin - NYC Building ID Number based on overlap between green roof areas and a building footprint dataset for NYC from August, 2017. (Newer building footprint datasets do not have linkages to the tax lot identifier (bbl), thus this older dataset was used). The most current building footprint dataset should be available at: https://data.cityofnewyork.us/Housing-Development/Building-Footprints/nqwf-w8eh. Associated metadata for fields from that dataset are available at https://github.com/CityOfNewYork/nyc-geo-metadata/blob/master/Metadata/Metadata_BuildingFootprints.md.
- bbl - Boro Block and Lot number as a single string. This field is a tax lot identifier for NYC, which can be tied to the Digital Tax Map and PLUTO/MapPLUTO. Metadata for fields pulled from PLUTO/MapPLUTO can be found in the PLUTO Data Dictionary. All joins to this bbl were based on MapPLUTO version 18v1.

Aggregated Dataset : Variables codebook

The same authors provide an additional dataset aggregated at the City Council District, which is used for regression purposes.

- coun_dist - Unique Identifier for each City Council District
- borough - Name of the borough containing the respective City Council District
- num_greenroofs - Number of Green Roofs Contained in the respective City Council District
- green_area - total area (square meters) of green roofs in each City Council District
- num_bldgs_total - Number of total buildings contained within the respective City Council District
- building_area - Total area (square meters) of buildings in each City Council District
- prop_gbuild - Proportion (%) of the buildings in each City Council District with a green roof
- prop_gr - Proportion (%) of the rooftop area (building footprint area) covered by green roof within each City Council District

2. Asthma

Data source

The asthma data set was sourced from the New York State Department of Health website. We downloaded the publicly-available CSV file then read it into R.

Goals of the analysis

The primary goal of this analysis is to provide an overview of asthma rates across the different boroughs of NYC. We would like to understand how the asthma burden is distributed across the different boroughs, as well as analyze whether there have been any changes over time.

Data Load & Wrangling

In the following analysis, we will be drawing on a dataset of asthma hospitalizations and ED visits between 2000 and 2019 in New York State, broken down by regions. Asthma ED visits and hospitalization rates tell us about the effect of asthma in a community due to environmental and household triggers, access to medical care, and the quality of disease management for asthma. ED visits and hospitalization shows us only the most severe cases of asthma; most people with asthma suffer the health effects without needing to go to hospital.

We load and wrangle the ‘secondary NY Asthma dataset’ to generate a ‘asthma_df dataframe’. The original .csv dataset is available in the "data" folder within the project's repository. The following procedure is applied to clean the data:

- filter for relevant observations (aka the boroughs of New York City from years 2006-2018) using `filter()`
- select relevant variables using `select()`
- rename certain variables for comprehension using `remame()`, as well as certain borough names and years using `mutate()`and `recode`()`
- re-order the variables using `arrange()`

Lastly, we created a second asthma dataframe with the goal of having a severity measure, which we defined as the number of ED visits divided by the number of Hospitalizations (both per 10k inhabitants), for each borough and year separately. We achieved this by leveraging the use of `pivot_wider()`, `mutate()` (i.e., calculating the new measure) and `pivot_longer()`.

Data Codebook

Asthma in NYC

The dataset provides the number and rate of asthma hospitalizations and emergency department (ED) visits in New York State. These data are presented by age groups, sex, month

of year, and total population for 3 selected year intervals (ex. 2000-2002, 2001-2003, etc...). These data show only a small proportion of people who have asthma, those whose condition is serious enough to require an emergency department (ED) visit or hospital stay. Data show the number of asthma ED visits and hospitalizations rather than the number of patients who were admitted or seen for asthma related illness. For example, if a person was hospitalized twice they would be counted twice.

Asthma Age-Adjusted Hospitalization Rates

Rates are calculated as the number of asthma hospitalizations per 10,000 residents in 5-year age groups using age-specific weights based on the estimated 2008-2010 population. The maps show average asthma age-adjusted hospitalization rates for the selected three year period. Users can compare asthma hospitalizations across counties in the following regions: New York State (whole state), New York City, New York State Excluding New York City. Users can see more clearly the differences in asthma hospitalization rates between the remaining New York State counties when New York City counties are removed from the display.

Asthma Age-Adjusted Emergency Department Visit Rates

Rates are calculated as the number of asthma ED visits per 10,000 residents in 5-year age groups using age-specific weights based on the estimated 2008-2010 population. The maps show average asthma age-adjusted ED visit rates for the selected three year period. Users can compare asthma ED visits across counties in the following regions: New York State (whole state), New York City, New York State Excluding New York City. Users can see more clearly the differences in asthma ED visits rates between the remaining New York State counties when New York City counties are removed from the display.

The following variables are present in the wrangled and tidied `asthma_df` (and `asthma2_df`) dataframes:

- `borough` = New York City boroughs
- `indicator` = shows what metric `aa_rate10ky` stands for: either age adjusted ED Visits (see explanation above), age adjusted Hospitalizations (see explanation above), or a self-created Severity Metric (ED visits divded by Hospitalizations, only present in `asthma2_df`)
- `year` = year the measure was recorded
- `aa_rate10kpy` = value of indicator per 10k inhabitants per year within a given borough
- `count` = absolute value of indicator per year within a given borough
- `daily_average` = absolute value of indicator per day within a given borough and year

3. Temperature

Data Source

To complete the exploratory analysis for temperature, datasets were retrieved from NYC.gov Environmental and Health Data Portal. The datasets used were "Daytime Summer Surface Temperatures (°F)" and "Heat Vulnerability Index (NTA)". Neighborhood tabulation areas (NTA) codes were used to merge the two datasets for 188 neighborhoods in New York City. NTAs are neighborhoods that were created by the NYC Department of Planning by aggregating census tracts into 195 neighborhood-like areas. Some neighborhoods were combined to reach the minimum population of 15,000 per NTA.

The heat vulnerability index data were collected from 188 neighborhoods across the 5 boroughs of NYC and represent the community-level heat impacts due to extreme heat events. A statistical model using social and environmental factors, such as surface temperature, green space, air conditioning access, poverty, and Black population, was used to estimate the risk of heat-related death and develop the 5-point index score scale, with 1 being the lowest risk and 5 being the highest risk. The Black population was specifically included in the model because it identified as the most excluded from heat resources. Residents who are at highest heat risk also did not have or use air conditioning, were older or had chronic health conditions or serious mental illnesses.

The daytime surface temperature data were collected in Fahrenheit and varied based on vegetative covering and materials that retain heat, such as paved roads, sidewalks, and buildings. In this analysis, the daytime surface temperature was re-coded to represent temperature in Celsius. The neighborhoods were identified by NTA codes to match the neighborhoods captured in the heat vulnerability index dataset.

The heat vulnerability index is a collection of data sources: American Community Survey (2013 - 2017, 5-year estimates), New York City Department of Parks and Recreation (2017), U.S. Geological Survey LandSat (2018), and United States Census Housing and Vacancy Survey (2017). The daytime surface temperature data source is the U.S. Geological Survey (2018).

New York City Department of Health, Environment & Health Data Portal. "Climate" data. Daytime summer surface temperature. Available here.

New York City Department of Health, Environment & Health Data Portal. "Climate" data. Heat vulnerability index (NTA). Available here.

Goals of the Analysis

The **primary goal** of this analysis is to explore the daytime summer surface temperature and heat vulnerability index scores of neighborhoods across New York City. A dataset that contains both daytime summer surface temperature and heat vulnerability index scores (“envo_health”) was used.

Secondly, this analysis aims to provide:

- a **descriptive analysis** of the daytime summer surface temperature and heat vulnerability index scores in New York, as well as **stratified analysis** at the Borough levels;
- relevant **data visualization** of the above (notably a map using the leaflet package);
- an attempt to fit the dataset into a **linear regression model** to elucidate differences in the daytime summer surface temperatures and heat vulnerability index scores at the Borough levels.

Data Load & Wrangling

To create the dataset for the temperature analyses, we combined the “Heat Vulnerability Index” dataset and the “Daytime Surface Temperature” dataset. Neighborhood tabulation areas (NTA) codes were used to merge the two datasets for 188 neighborhoods in New York City. NTAs are neighborhoods that were created by the NYC Department of Planning by aggregating census tracts into 195 neighborhood-like areas. Some neighborhoods were combined to reach the minimum population of 15,000 per NTA.

- Relevant variables were selected and rearranged using select()
- Temperatures were converted from Fahrenheit to Celsius using mutate()
- For the univariate models, mutate() and fct_relevel() was used to relevel the categorical variable “Borough” and place Manhattan as the reference group

To create the dataset for the joint analysis investigating the influence of total urban green spaces on daytime summer surface temperature and heat vulnerability index scores, we created the “envo_greenroof” dataset. The following procedure was then applied:

- Relevant variables were selected and rearranged using select()
- The variable “total_green_area” was created by using mutate() and using cum()
- The variables “avg_daytime_temp” and “avg_heat_index” was created by using mutate() and using mean()
- The merged and tidied dataset was exported as a “csv” file and saved as “envo_greenroof” in the Group Project data folder using write_csv()
- Temperatures were converted from Fahrenheit to Celsius using mutate()
- For the univariate models, mutate() and fct_relevel() was used to relevel the categorical variable “Borough” and place Manhattan as the reference group

Data Codebook

Envo_health Dataset: Variables codebook

ID and location data

- borough = New York City boroughs
- neighborhood = 188 neighborhood tabulation areas identified by neighborhood names
- longitude - Longitude in decimal degrees.
- latitude - Latitude in decimal degrees.

- geo_id = geography ID for this entry (coded in the original dataset)
- year = year of data entry

Note: Some coordinates may not accurately represent the combined NTAs due to multiple neighborhoods being grouped together

Daytime Surface Temperature and Heat Vulnerability Index Score data

- daytime_surface_temp = daytime landsat thermal data on surface collected on July 17, 2018 (averaged spatially), Farenheit
- heat_vulnerability_index = score ranging from 1 (lowest risk) to 5 (highest risk)

Envo_greenroof Dataset: Variables codebook

- borough = New York City boroughs
- total_green_area = total area of the footprint of the green roof per Borough, in square meter
- avg_daytime_temp = average daytime surface temperature, Fahrenheits
- avg_heat_index = average heat vulnerability index score

4. EITC Tax

Data Source

The data used for this element of the project comes from two different data sources:

1. The EITC Data provided by New York State Open Data, pulled in using an API.
2. The Population Estimates, provided by the U.S. Census Bureau in a csv file.

Goals of the Analysis

The primary goal of this analysis is to understand the financial status of residents in each borough. To that end, we will first create a data-set upon which regression analysis can be conducted with potentially associated dimensions (urban greenspaces, temperature, and health outcomes).

Secondly, this analysis aims to provide:

- a descriptive analysis of the NYC EITC primary dataset, as well as stratified analysis at the Borough levels;
- relevant data visualizations of the above;
- an attempt to fit the dataset into a linear regression model to elucidate differences in the distribution of EITC data at the Borough levels.

Data Load & Wrangling

The EITC Data set required moderate cleaning for our purposes. The original data set includes information for all counties in New York State. For purposes of this project, we're only interested in New York City, and the five counties that correspond to each borough.

Once the boroughs were isolated, there were still several columns that weren't needed for this analysis. The 'notes', 'place_of_residence', and 'place_of_residence_sort_order' were dropped because they either didn't contain any data (i.e. notes) or they were redundant of other columns.

Next, we filtered this data to only show the counties within New York City: Bronx, Kings, Manhattan, Queens and Richmond. For consistency across the project, we then converted the county names to their corresponding borough names. There are two that need to change: Kings County corresponds to Brooklyn and Richmond County corresponds to Staten Island.

Next, we cleaned up the 'credit_type' variable names for ease of use, then coerced character columns to numeric as needed.

Finally, the 'credit_amount_claimed_in_thousands' needed to be multiplied by 1000 to get the actual dollar amount claimed in each borough. We also renamed 'county' to 'borough' for consistency across the project.

We then wanted to understand what percentage of each borough's populace filed a claim each year. The US Census Bureau extrapolates estimated county populace each year, basing their projections on the most recent Census data and vital statistics. We pulled these estimates into a dataframe and merged it with the tax data, limiting the time frame from 2016-2018, the same years that the `Greenspace` data were collected.

Data Codebook

Once cleaned and optimized for use in this project, the variables of interest are as follows:

- tax_year (numeric): The tax year, from 1994 - 2020
- credit_type (character): The type of credit received, either `City EITC`, `State EITC` or `Noncust. EITC`
- borough (character): Borough of New York City, either Bronx, Brooklyn, Manhattan, Queens or Staten Island
- number_of_claims (numeric): The number of claims filed
- boro_credit_total (numeric): The total amount, in dollars, received in EITC credit
- average_credit (numeric): The average dollar amount of the EITC credit
- popestimate (numeric): The population estimate for each borough, from 2016-2018

Results

The results section is structured into 4 subsections, entailing the descriptive analysis, data visualization, univariate regression analyses, and summary of results, on: (1) Urban Green Space, (2) Asthma, (3) Temperature, and (4) Tax. The results section includes a latter (5) Joint Analysis.

1. Urban Green Space

First, we perform a descriptive analysis of the NYC Urban Green Spaces (UGS) primary dataset, as well as stratified analysis at the City Council District and Borough levels.

Descriptive Analysis

Overall UG_df Dataset

Overall, the “UG_df” contains 16 variables related to 736 urban green spaces in NYC. The total UGS area in NYC sums to 2.4620772^{5} square meters. The following table summarizes the average UGS in NYC, including its size (in square meters), height (in meters) and percentage of green coverage (that is, the proportion of green space within the total building area). Key variables are summarized in the table below using knitr:kable().

Mean surface	Median surf.	Minimum surf.	Maximum surf.	Mean building height	Percent of UGS
334.521	109.06	0.939	25763.66	43.474	19.635

Overall, within New York City, UGS have a **mean size of 334.52 sqm (range 0.94-25,763.66 sqm)**. The broad range of values, as well as the median situated at 109.06 sqm (way below the mean), suggests this variable’s **right-skewed distribution**. We will explore this more in the “data visualization” section below. On average, UGS in NYC are situated at **43.47 m of height**, and **cover 19.63% of the building’s total surface**.

Stratified Analysis by Borough

When stratified by borough, differences in the number and dimensions of UGS become apparent. Following a similar procedure as above, we generate a summary table below, arranged according to the amount of UGS present in each borough.

borough	n	perc	tot_green	mean_green	median_green	min_green	max_green	mean_height	perc_green
Manhattan	465	63.179	145265.382	312.399	90.500	0.939	25763.656	54.242	15.236

Brooklyn	135	18.342	50211.591	371.938	86.516	3.168	11906.341	21.947	21.534
Bronx	82	11.141	25476.308	310.687	174.847	21.429	2053.387	20.839	29.356
Queens	50	6.793	23427.543	468.551	214.848	7.359	2531.361	39.633	39.635
Staten Island	4	0.543	1826.896	456.724	449.312	94.747	833.525	19.450	17.983

Manhattan contains over 63% of all NYC's UGS (n=465), situated at a considerably higher altitude (mean height 54.24 m). Nevertheless, its UGS are also the smallest in percentage terms when assessing the proportion of building coverage (15.24%). Conversely, **Queens** and **Staten Island** have the lowest amount of UGS (n=50 and n=4, respectively), but their UGS area is the largest on average (mean sizes of 468.55 and 456.72 sqm, respectively). These results could be linked to the property prices in each area (further analysis would be needed to assess this).

Stratified Analysis by City Council District

Although NYC city council districts typically correspond to borough sub-divisions, the equivalence is not perfect (note council district #8 belongs to both Manhattan and Bronx, more info here). In that sense, the same authors provide a similar dataset aggregated at the City Council District level, which we have loaded with the UG_agg_df name. A similar descriptive analysis as above is performed. Notably, in this dataset the total number of buildings per district is available, and therefore we compute the percentage of buildings with an UGS.

coun_dist	borough	n	perc_UGS	perc_building	mean_green
3	Manhattan	131	17.799	1.645	42.923
1	Manhattan	107	14.538	1.912	19.200
4	Manhattan	80	10.870	1.364	18.859
33	Brooklyn	53	7.201	0.356	29.525
2	Manhattan	44	5.978	0.893	23.836
6	Manhattan	28	3.804	0.686	28.200
8	Manhattan/ Bronx	26	3.533	0.472	33.407
5	Manhattan	25	3.397	0.747	13.309
26	Queens	24	3.261	0.158	41.056
39	Brooklyn	24	3.261	0.112	15.793

7	Manhattan	20	2.717	0.617	39.401
16	Bronx	19	2.582	0.496	16.535
17	Bronx	18	2.446	0.198	30.240
9	Manhattan	15	2.038	0.258	55.090
35	Brooklyn	15	2.038	0.115	94.903
15	Bronx	11	1.495	0.140	19.469
34	Brooklyn	10	1.359	0.064	10.449
14	Bronx	9	1.223	0.201	10.827
11	Bronx	8	1.087	0.077	65.061
29	Queens	7	0.951	0.034	42.294
38	Brooklyn	7	0.951	0.039	111.036
31	Queens	5	0.679	0.014	18.058
36	Brooklyn	5	0.679	0.032	8.180
42	Brooklyn	5	0.679	0.030	7.462
47	Brooklyn	4	0.543	0.017	13.764
49	Staten Island	4	0.543	0.010	42.431
13	Bronx	3	0.408	0.010	31.165
18	Bronx	3	0.408	0.022	74.748
20	Queens	3	0.408	0.014	128.468
27	Queens	3	0.408	0.006	12.653
40	Brooklyn	3	0.408	0.028	12.985
43	Brooklyn	3	0.408	0.009	34.771
21	Queens	2	0.272	0.012	67.457
46	Brooklyn	2	0.272	0.005	27.857
48	Brooklyn	2	0.272	0.008	38.630

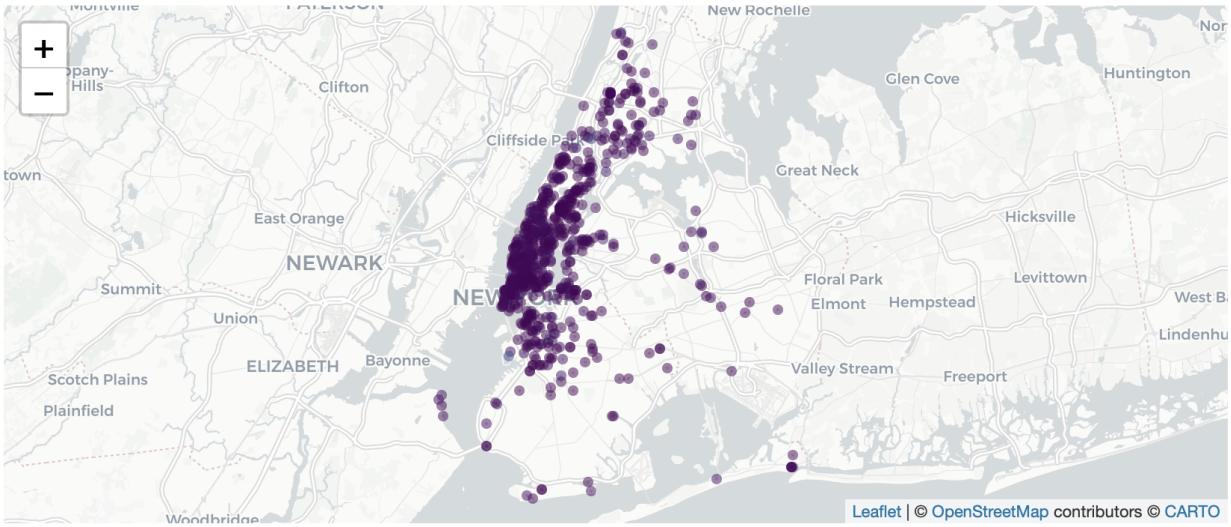
10	Manhattan	1	0.136	0.038	18.895
19	Queens	1	0.136	0.002	76.946
22	Queens	1	0.136	0.004	11.561
24	Queens	1	0.136	0.004	18.606
25	Queens	1	0.136	0.007	2.549
28	Queens	1	0.136	0.002	22.227
32	Brooklyn	1	0.136	0.002	114.641
37	Brooklyn	1	0.136	0.005	1.324

We observe that **3 City Council Districts in Manhattan concentrate over half of the borough's UGS (n=318)**, with over 1% of its buildings covered by green spaces. Conversely, **8 districts within Bronx, Queens, Brooklyn and Staten Island have no UGS at all**. These results suggest the existence of distribution differences, which will be addressed in the last section of this analysis.

Data visualization

Mapping UGS in NYC

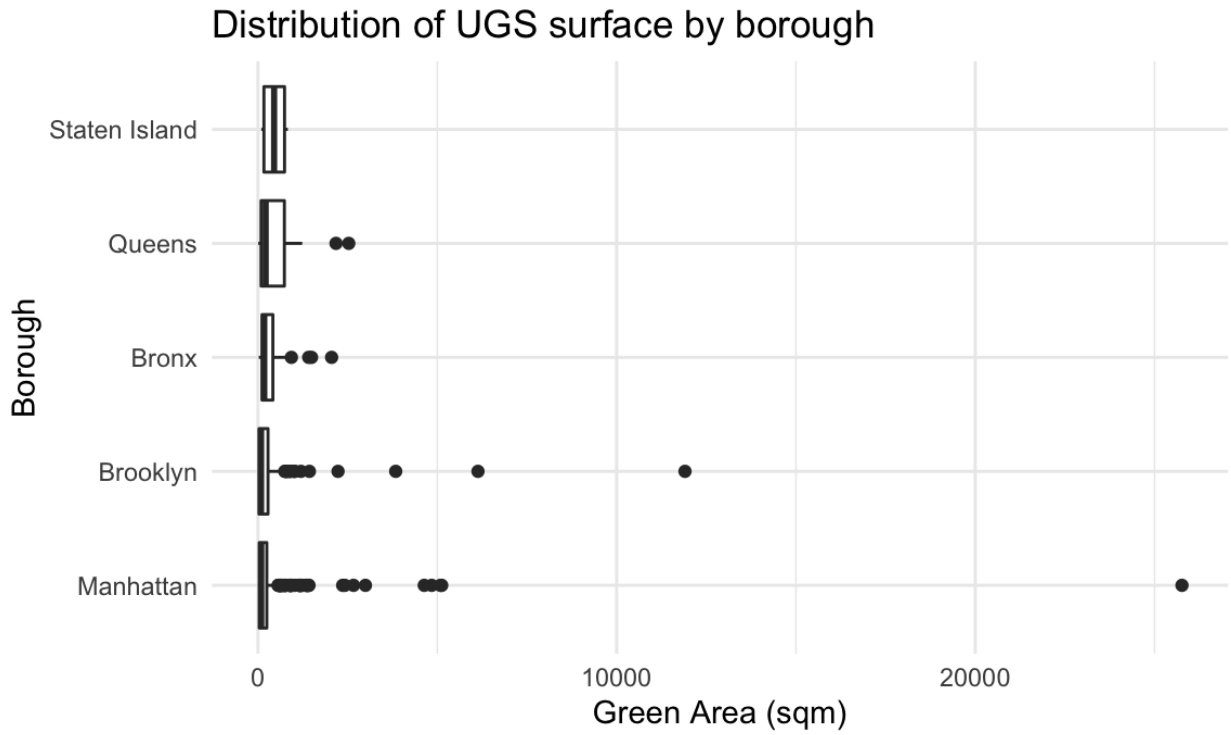
Using the leaflet package, we plot each UGS on a NYC map. A pal() function is generated to establish a color according to the size of the UGS. Please note an interactive version of the map is available on the project's website.



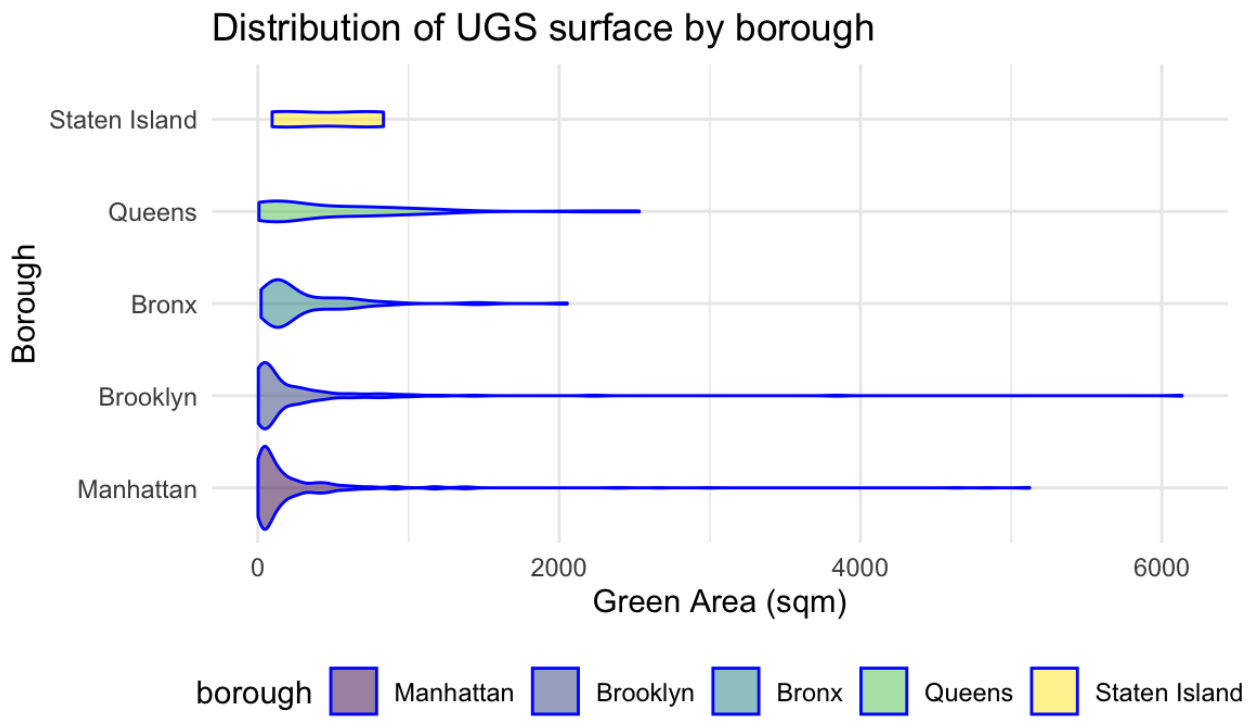
As described above, we observe a concentration of UGS in southern Manhattan. Hardly any green spaces are detected in southern Brooklyn, Staten Island, Queens or the Bronx.

Distribution of UGS surface by borough

Next, the distribution of UGS surface by borough is investigated. The UG_df is first plotted on a boxplot to have a better understanding of outliers (as we have indication of a right-skewed distribution according to the descriptive analysis above).



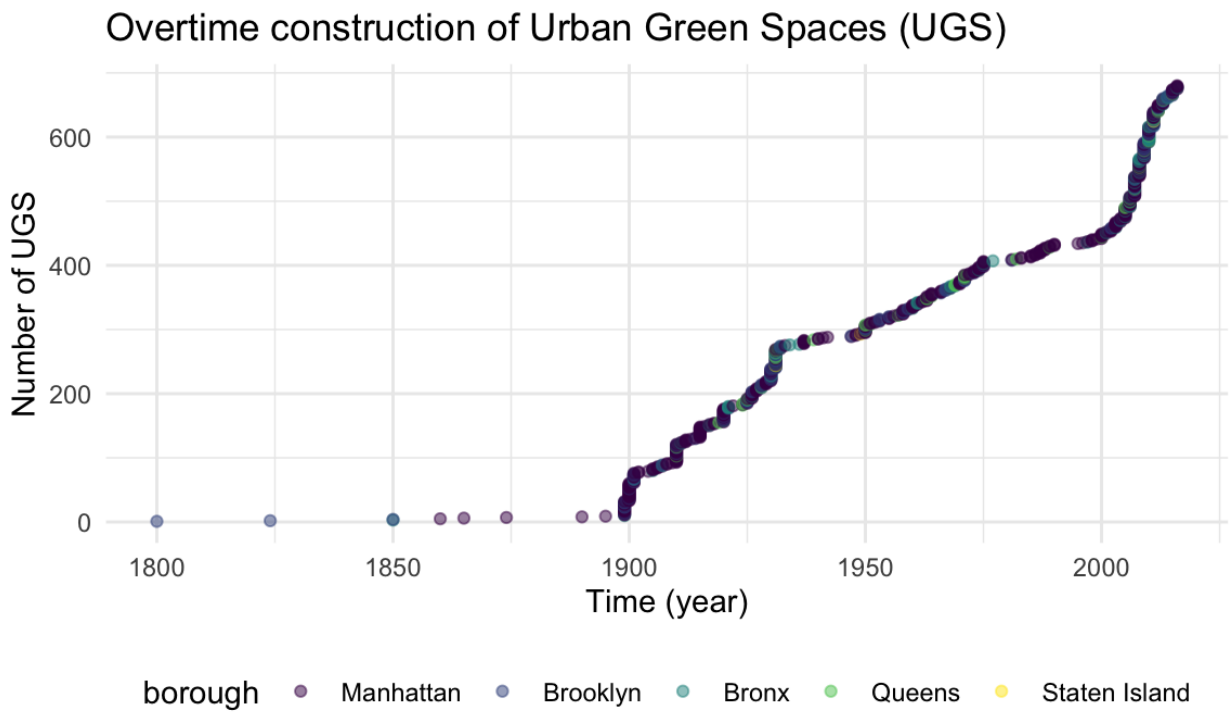
Given the presence of considerable outliers, the plotted green area surface is then limited to 10,000 sqm and visualized through a violin plot.



Overall, these visualizations confirm the sparse distribution of UGS surface area (sqm) across boroughs, as well as suggest differences according to its location (see map above). We will investigate this further through regression in the upcoming section.

UGS development overtime in NYC

Finally, the construction of UGS trends is investigated overtime. To do so, the cumsum() function is used over a generated dummy variable that allocates the value “1” to all UGS. Note that the observations with an invalid or missing construction year are omitted.



We observe a progressive construction of UGS overtime, with a rapid upward trend as of early 2000s (potentially linked to urban planning regulations and greater interest in climate change mitigation strategies).

Regression Analyses: Univariate Models

In this section, we investigate the geographical differences we have noted above graphically. We perform statistical tests and univariate regression analyses to understand whether the location within NYC (x: predictor variables, both “borough” and “city council district”) influences either the size or number of UGS (y: outcome variable, both “green_area” and “num_greenroofs”).

UGS number by borough and city council district

We first investigate whether there are differences in the amount of UGS by borough and city council district (using the UG_agg_df). Given that we are testing significance of differences for multiple groups and independent samples, the appropriate test is ANOVA (Analysis of Variance).

DF	Statistic	p-value
5	6.54487	0.00012
45	NA	NA

Next, we run a similar ANOVA analysis at the city council district level.

DF	Statistic	p-value
1	17.29994	0.00013
49	NA	NA

Both at the borough and city council district levels, resulting p-values from the ANOVA tests are far below our accepted significance level of 0.05, so we know there is at least one proportion that is different among the 5 boroughs and 51 city council districts of NYC.

To better understand the distribution of UGS at the borough level, we run a linear regression lm() model. For the purposes of this project, we are using Manhattan as our universal reference group, so we will then see which proportions differ from Manhattan.

Term	Estimate	p-value	Lower Confidence Interval	Upper Confidence Interval
(Intercept)	50.11111	0.00000	36.33637	63.88585
boroughBronx	-41.23611	0.00022	-61.31608	-21.15615
boroughQueens	-46.34188	0.00001	-64.26127	-28.42249
boroughBrooklyn	-42.16993	0.00002	-59.20507	-25.13479
boroughStaten Island	-48.77778	0.00116	-76.32726	-21.22829
boroughManhattan/ Bronx	-24.11111	0.28374	-67.67067	19.44845

Again, the model yields a p-value which is far under our accepted significance level of 0.05, indicating that when comparing NYC boroughs to Manhattan, the number of UGS are significantly different. To be noted is the entry marked as “Manhattan/Bronx” borough, which is not significant. That corresponds to district 8 in NYC, which includes southern Bronx and northern Manhattan.

Note that we did not run a similar linear model at the city council district level given the presence of n=51 groups which would yield results of difficult and questionable interpretation. From now onwards, our team will focus on assessing differences at the NYC borough level.

UGS surface by borough

Next, we investigate differences on the total UGS surface area by borough using a similar approach as above. We first run a ANOVA test to understand whether there are differences in the mean UGS surface. To do so, we generate a UG_sum_df that includes an “agg_surface” variable at the borough level. Next, we run the ANOVA test.

DF	Statistic	p-value
4	2.65557e+29	0

731	NA	NA

The results are far above our accepted significance level of 0.05. In fact, the p-value is rounded to zero despite indicating to show up to 70 digits, indicating a very significant finding (that is, differences in the total surface of UGS by borough). The same results are depicted in the model below investigating boroughs as a predictor of the aggregated borough surface.

Term	Estimate	p-value	Lower Confidence Interval	Upper Confidence Interval
(Intercept)	145265.38	0	145265.38	145265.38
boroughBrooklyn	-95053.79	0	-95053.79	-95053.79
boroughBronx	-119789.07	0	-119789.07	-119789.07
boroughQueens	-121837.84	0	-121837.84	-121837.84
boroughStaten Island	-143438.49	0	-143438.49	-143438.49

Summary of Results

In conclusion, throughout this exploratory analysis, plotting and regression modeling attempts of n=736 Urban Green Spaces (UGS) present in NYC as of 2018, we conclude:

- Overall, within New York City, UGS have a mean size of 334.52 sqm (range 0.94-25,763.66 sqm). On average, UGS in NYC are situated at 43.47 m of height, and cover 19.63% of the total building's surface.
- Manhattan contains over 63% of all NYC's UGS (n=465), situated at a considerably higher altitude (mean height 54.24 m). Nevertheless, its UGS are also the smallest in percentage terms when assessing the proportion of building coverage (15.24%). Conversely, Queens and Staten Island have the lowest amount of UGS (n=50 and n=4, respectively), but their UGS area the biggest on average (mean sizes of 468.55 and 456.72 sqm, respectively).
- Such uneven distribution of UGS becomes apparent when visualizing it, as well as running statistical tests. We observe significant differences in terms of the number of green spaces as well as its surface at the borough level.

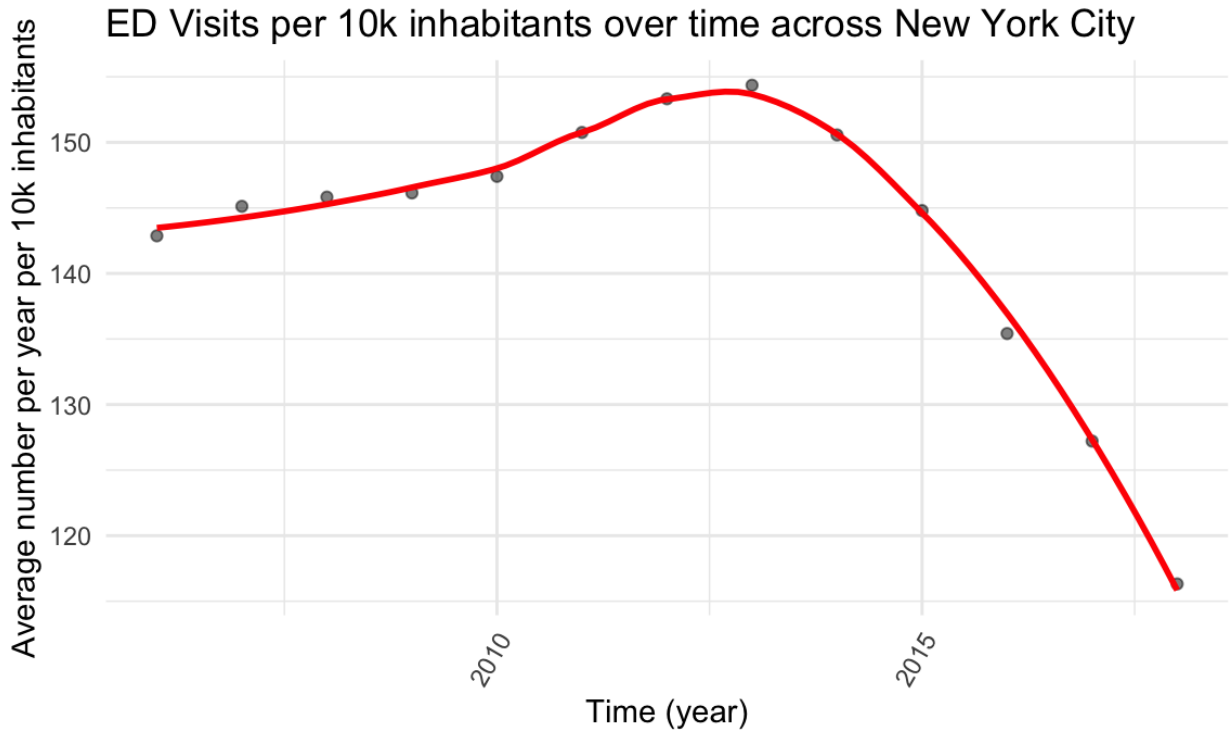
2. Asthma

Descriptive Analysis and Data Visualizations

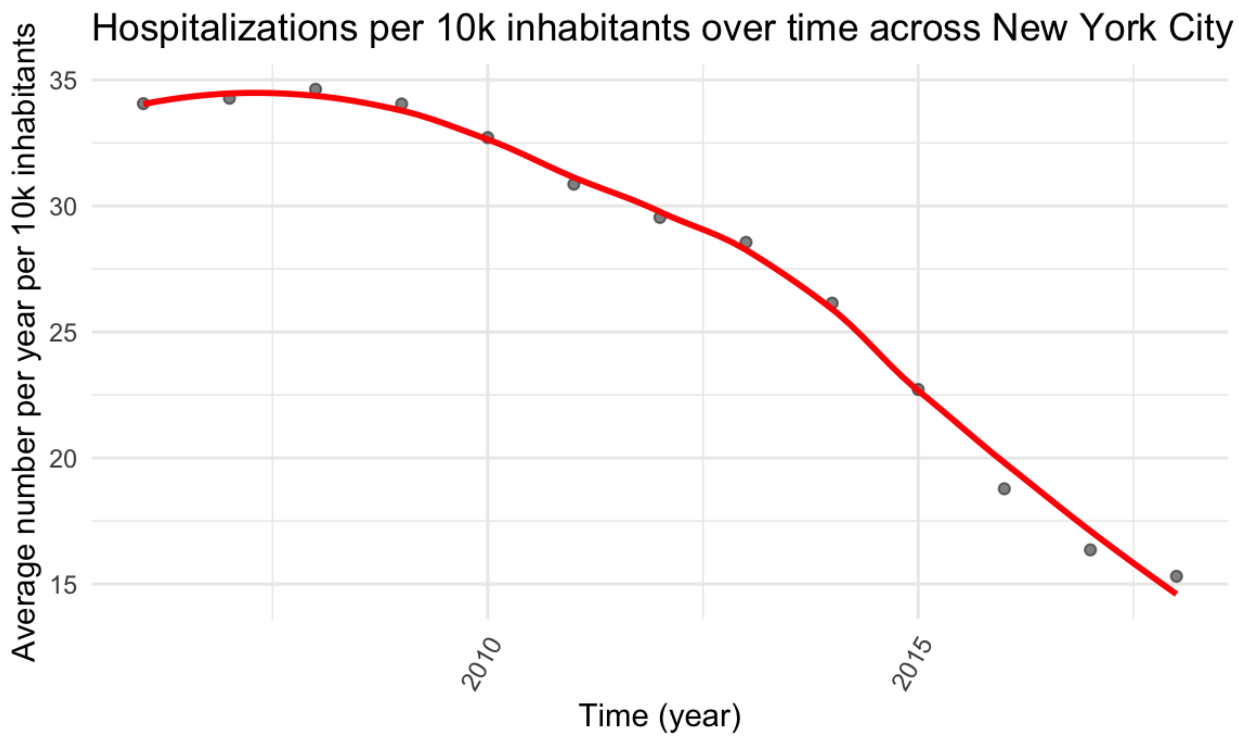
Next, we perform a descriptive analysis of the NY Asthma secondary dataset, as well as stratified analysis at the borough level. Overall, the ***asthma_df** contains 6 variables related to 156 measures of asthma in NYC.

Developments on New York City Level

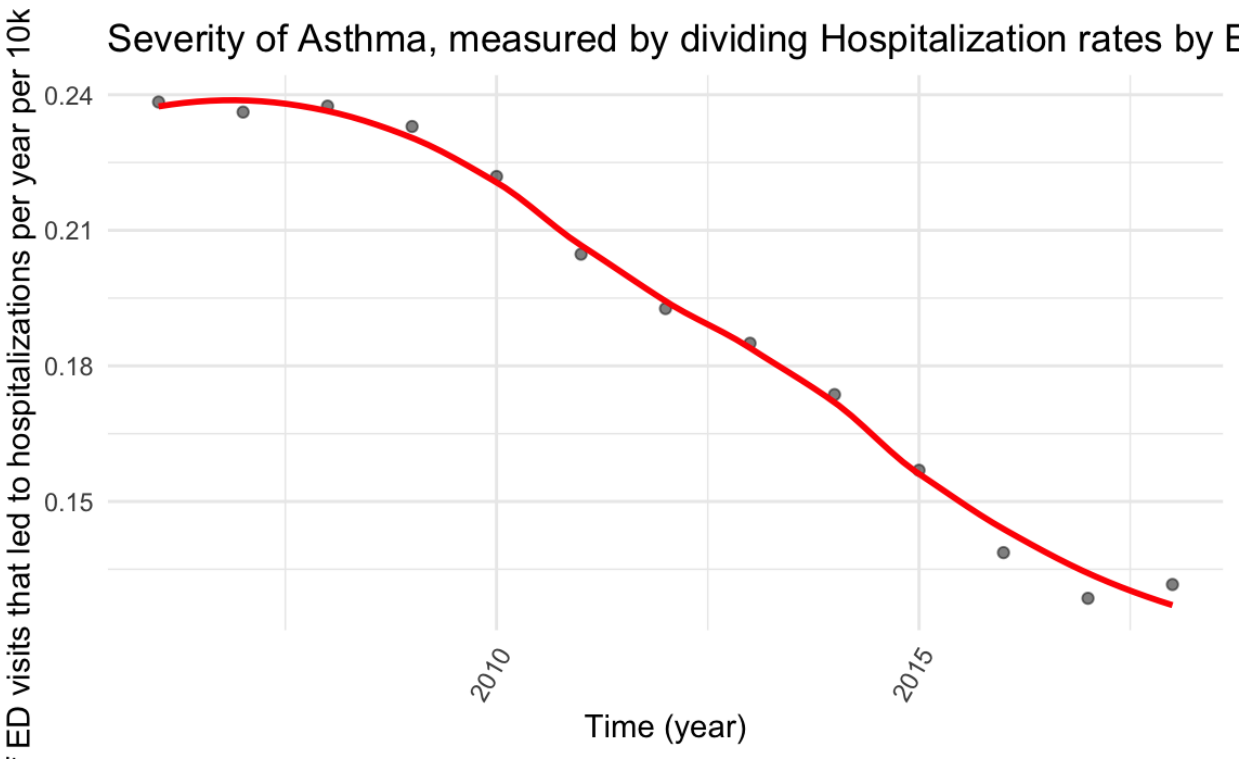
The graph below shows the development of ED visits per 10k inhabitants between 2006-2018 across all of New York City. Rates increased until 2013 to a peak of close to 160 recorded ED visits per 10k inhabitants, after which a stark drop was recorded, a trend that seems to continue to this day. In 2018 recorded ED visits per 10k inhabitants were down to close to 110 per 10k inhabitants.



Next, we look at a graph that shows the development of Hospitalizations per 10k inhabitants between 2006 and 2018 across all of New York City. After remaining stable at approximately 35 hospitalizations per 10k inhabitants, the rate started to drop continuously and has reached less than half of that value, at approximately 15 hospitalizations per 10k inhabitants. At first glance, it seems as if hospitalizations have decreased more than ED visits, indicating that while both rates and severity of asthma have dropped, severity has declined more.



To test this we next plotted the severity measure we calculated manually in the previous section: the rate of Hospitalizations per 10k inhabitants divided by the rate of ED Visits per 10k inhabitants. We conclude that our above hypothesis seems to have been correct: in 2006 roughly every fourth (or approximately 24%) of ED visits resulted in hospitalizations; that number has gradually decreased to or approximately 13%. For this calculation, we assumed that every Hospitalizations had a prior ED visit.



Developments on Borough Level - Static

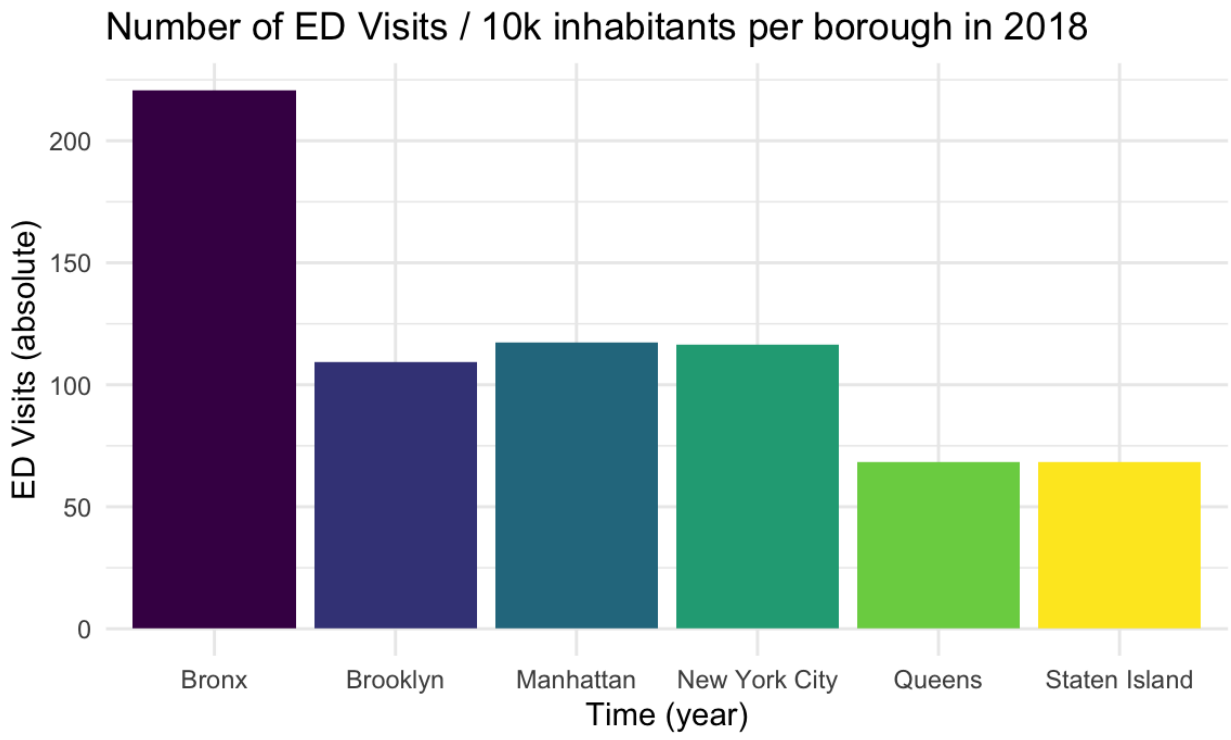
The table below shows ED visits and Hospitalizations per 10k inhabitants for each borough, and for New York City overall as averages across 2006-2018. The Bronx is a clear outlier, with 264 / 10k ED visits and 56 /10k Hospitalizations, compared to the city average of 143 / 10k ED visits and 28 / 10k ED visits. Manhattan is slightly above the average, while Brooklyn is slightly below. Queens and Staten Island are below the average. In summary, the asthma burden does not look to be distributed evenly across the different boroughs of New York City.

borough	ED Visits	Hospitalizations
Bronx	263.830	55.731
Brooklyn	135.996	26.165
Manhattan	151.509	22.836
New York City	143.085	27.538
Queens	84.315	17.105
Staten Island	82.445	16.911

Let's look at whether these inequalities persisted in the most recent measurements. The table below shows ED visits and Hospitalizations per 10k inhabitants for each borough, and for New York City overall in 2018. Asthma rates across all boroughs have decreased. While Bronx rates went down to 221 / 10k ED visits and 34 / 10k Hospitalizations, the New York City average lowered to 116 / 10k and 15 / 10k respectively. In conclusion, the inequities observed across the different boroughs seem to have persisted.

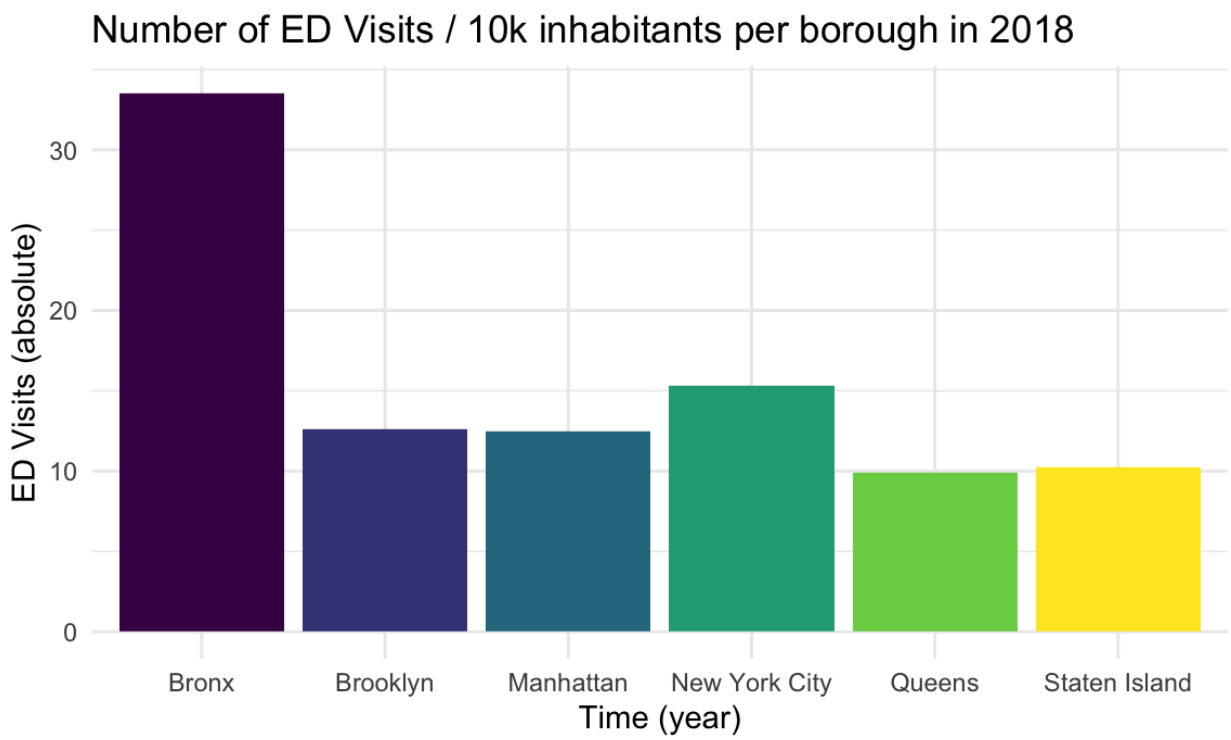
borough	ED Visits	Hospitalizations
Bronx	220.84	33.54
Brooklyn	109.20	12.64
Manhattan	117.33	12.45
New York City	116.32	15.31
Queens	68.12	9.90
Staten Island	68.30	10.24

The following graph visualizes the above table: ED visits / 10k inhabitants by borough in 2018. The Bronx asthma rates are particularly high.



A very similar pattern is observed for hospitalizations rates / 10k inhabitants by borough in 2018. The unequal burden is even more accentuated, as the Bronx tops the list by a wide margin. The Bronx

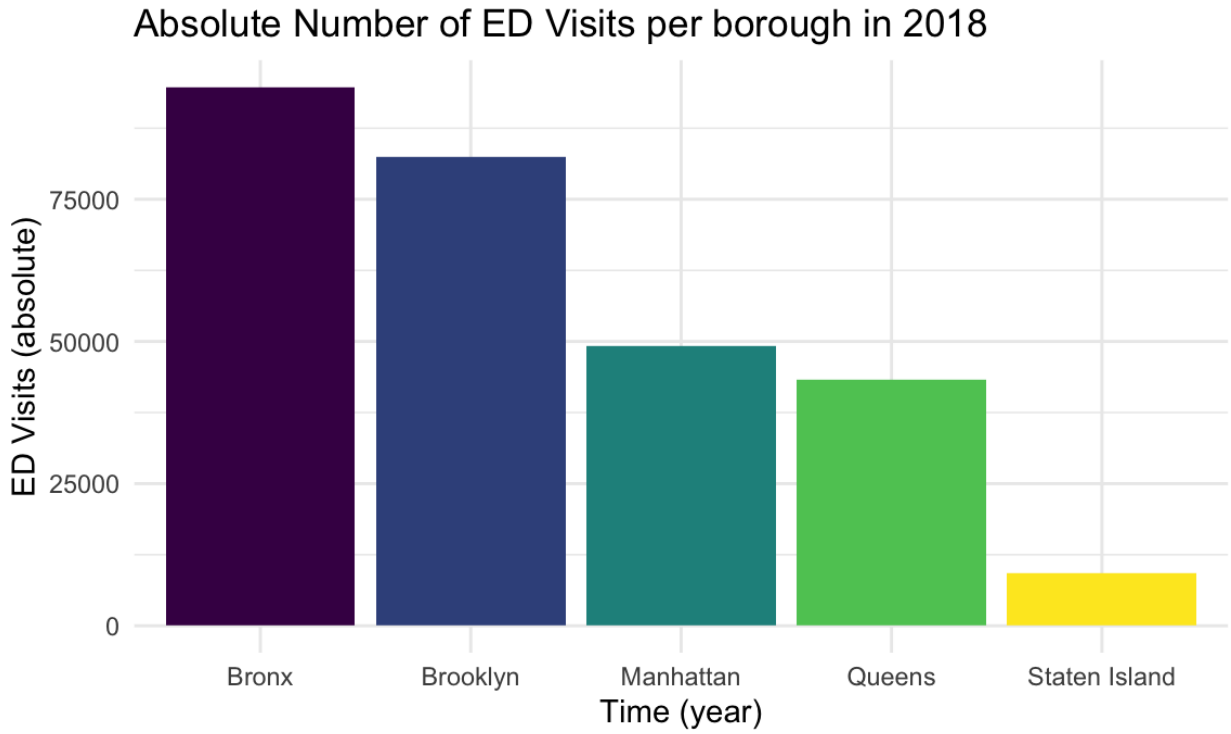
not only seems to be affected more by higher asthma rates, the asthma cases also seem to be more severe compared to the other boroughs.



Before we move to visualizations over time, we wanted to take a look at absolute burden across the different boroughs. The following table shows the total number of ED Visits and Hospitalizations per borough (i.e., not adjusted per 10k inhabitants) in 2018. The Bronx tops the list with 95k ED visits, followed by Brooklyn with 82k - compared to the total number in NYC of 279k this means that those two boroughs make up roughly two thirds of all asthma related ED visits. The same holds true for hospitalizations, where Bronx recorded 14k and Brooklyn 10k in 2018 (compared to 37k across NYC).

borough	ED Visits	Hospitalizations
Bronx	94758	14441
Brooklyn	82412	9547
Manhattan	49141	4980
New York City	278838	36918
Queens	43355	6546
Staten Island	9172	1404

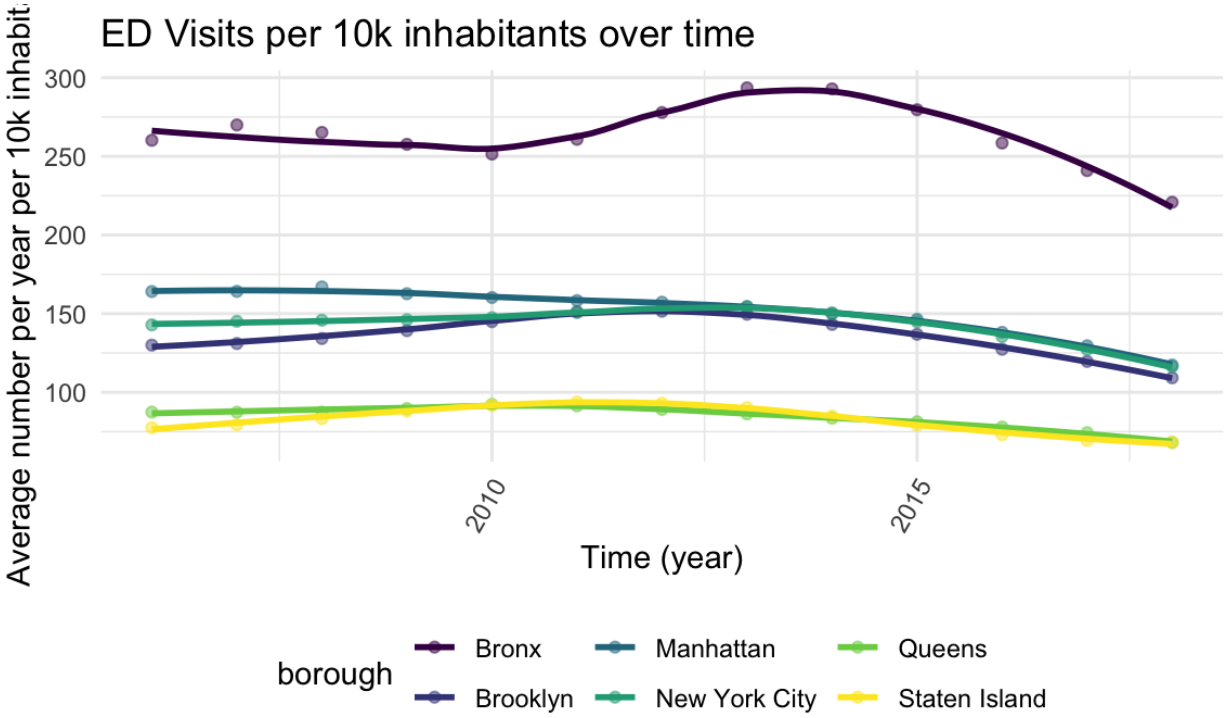
The following graphic depicts what was discussed above, showing the absolute number of ED Visits in 2018. The Bronx and Brooklyn are clearly carrying the largest burden of asthma ED visits in New York City.



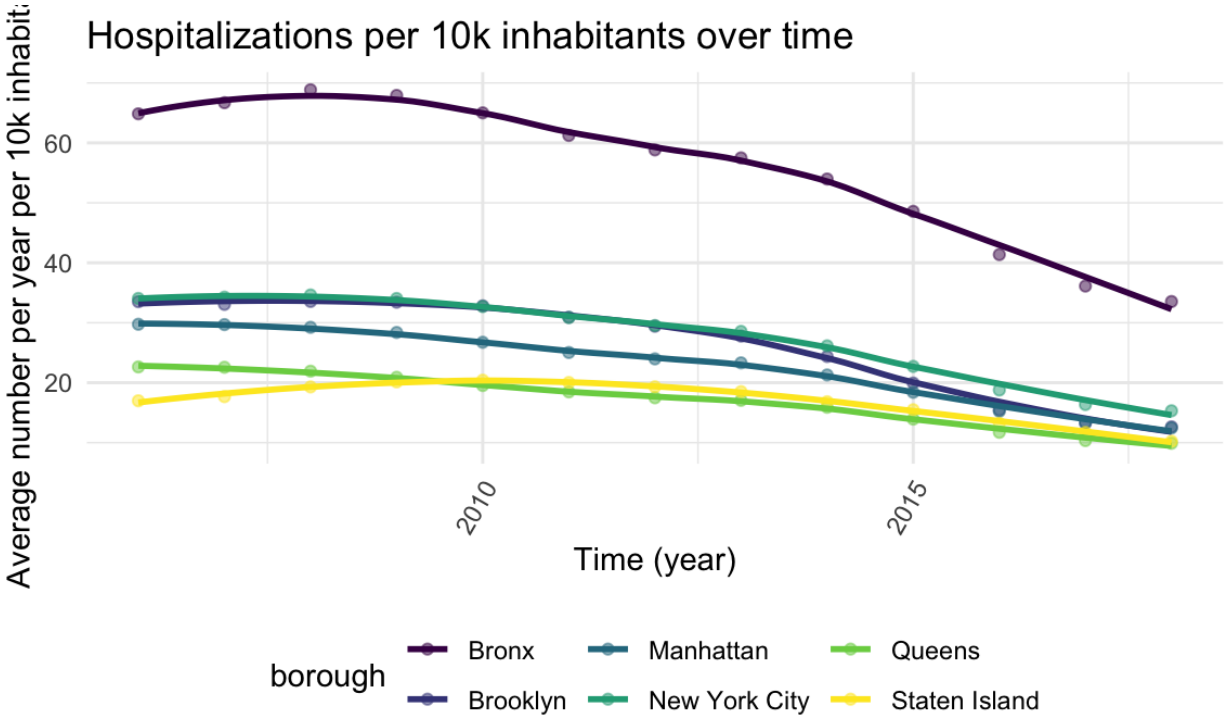
Developments on Borough Level - Over Time

This section is the most critical one, as we analyze how rates have changed over time.

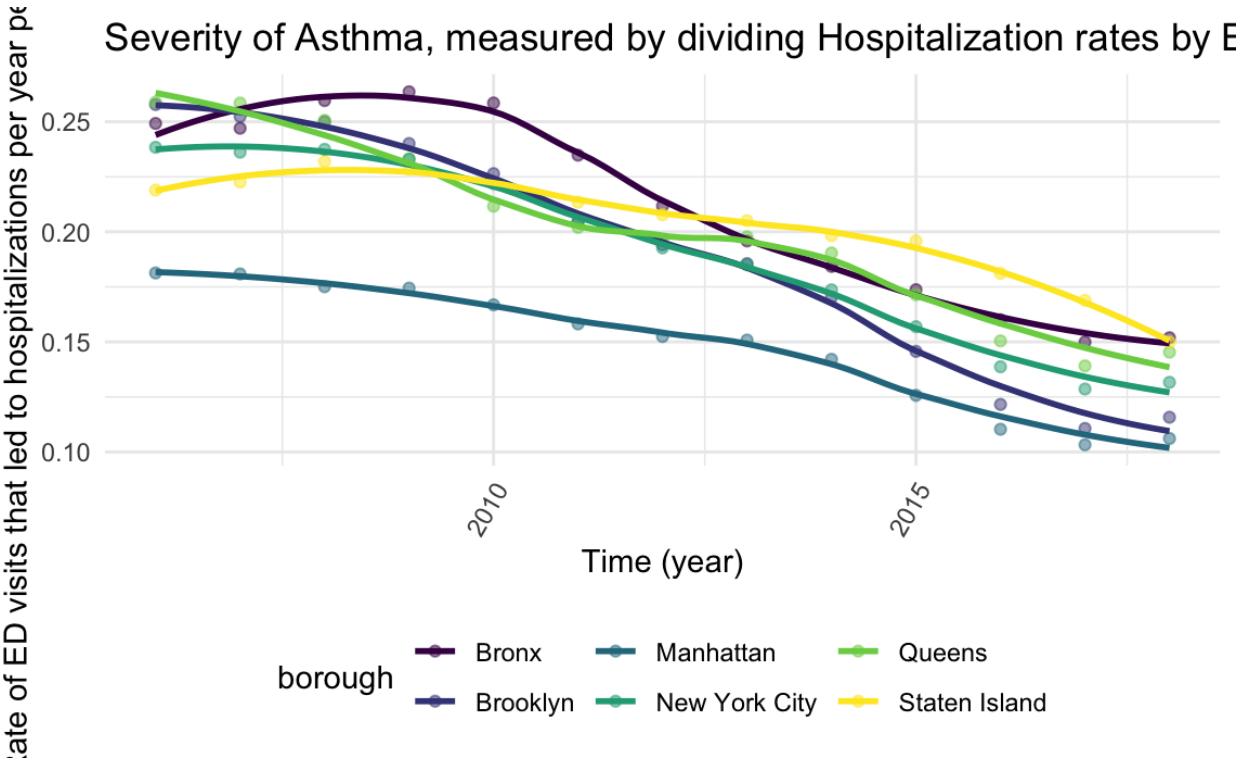
The graph below visualizes ED visits / 10k inhabitants per borough from 2006 to 2018. The patterns over time are relatively consistent: the Bronx has always been the most affected borough, and Queens / Staten Island always the least, while Manhattan and Brooklyn were always close to the city average. The Bronx experienced more fluctuation in their rates: a decrease from 2006-2010 was followed by an increase until 2014 and a subsequent stark decline until 2018. All boroughs seem to experience sinking ED visit rates over between 2014-2018.



The graph below visualizes Hospitalizations / 10k inhabitants per borough from 2006 to 2018. The general pattern is different to ED visits, as nearly all boroughs experienced a consistent decline in Hospitalization rates since 2006. The Bronx tops the list across all years, while all other boroughs have gotten pretty close to the city average.



Lastly, we look at severity of asthma in the boroughs across time by using our measure of hospitalizations divided by ED visits. Interestingly, this is the one measure where the Bronx (while still remaining among the highest on the list) has managed to close the gap to the other boroughs. Still, in the Bronx about 15% of ED visits lead to hospitalizations, while that number is only 10% in Manhattan.



Regression Analyses

Next, we will be analyzing whether linear models can provide further insights into asthma rates across NYC's boroughs.

The first step is to create three dataframes (one for each indicator) that do not include New York City overall and set Manhattan as the comparator borough.

Running a linear regression for ED visits, all values are significant: the Bronx is predicted to have 112 / 10k ED more visits compared to Manhattan, while Brooklyn, Queens, and Staten island are predicted to have 16 / 10k, 67 / 10k, 69 / 10k less than Manhattan respectively.

term	estimate	std.error	statistic	p.value
(Intercept)	151.509	3.757	40.327	0.000
boroughBronx	112.321	5.313	21.140	0.000
boroughBrooklyn	-15.514	5.313	-2.920	0.005

boroughQueens	-67.194	5.313	-12.647	0.000
boroughStaten Island	-69.064	5.313	-12.999	0.000

Running a linear regression for Hospitalizations, all values except for Brooklyn and Queens are significant: the Bronx and Brooklyn are predicted to have 33 / 10k and 3 / 10k hospitalizations more respectively when compared to Manhattan, while Queens and Staten Island are predicted to have approx 6 / 10k less each than Manhattan.

term	estimate	std.error	statistic	p.value
(Intercept)	22.836	2.097	10.890	0.000
boroughBronx	32.895	2.966	11.093	0.000
boroughBrooklyn	3.328	2.966	1.122	0.266
boroughQueens	-5.731	2.966	-1.933	0.058
boroughStaten Island	-5.925	2.966	-1.998	0.050

Also the severity measures are all significant. The actual values are a bit difficult to interpret, but all boroughs seem to experience higher severity compared to Manhattan.

term	estimate	std.error	statistic	p.value
(Intercept)	0.148	0.011	13.388	0.000
boroughBronx	0.063	0.016	3.992	0.000
boroughBrooklyn	0.042	0.016	2.689	0.009
boroughQueens	0.052	0.016	3.331	0.001
boroughStaten Island	0.055	0.016	3.523	0.001

Summary of Results

This exploratory analysis has made evident the inequalities that continue to persist across NYC's boroughs in terms of asthma rates. It has shown that asthma rates are generally improving, but that a lot of work still needs to be done, especially in the Bronx, which continues to carry the largest asthma burden. These results were confirmed to be statistically significant, with regression models as well as manually calculated severity measures.

Secondly, we would like to understand whether there might be any links between the UGS dataset and the asthma dataset. For example, did increases in the primary dataset on NYC Urban Green Spaces also correlate with a simultaneous decrease in asthma rates?

3. Temperature

First, we performed a descriptive analysis of the daytime summer surface temperature and heat vulnerability index (“envo_health”) dataset.

Descriptive Analysis

Overall envo_health Dataset

Overall, the “envo_health” dataset contains 8 variables related to 188 neighborhoods in NYC, which are defined by NTA codes. The mean daytime summer surface temperature in NYC is 36.92°C. The mean heat vulnerability index score in NYC is 3.

Mean Temperature (°C)	Mean Heat Vulnerability Index Score
36.924	3

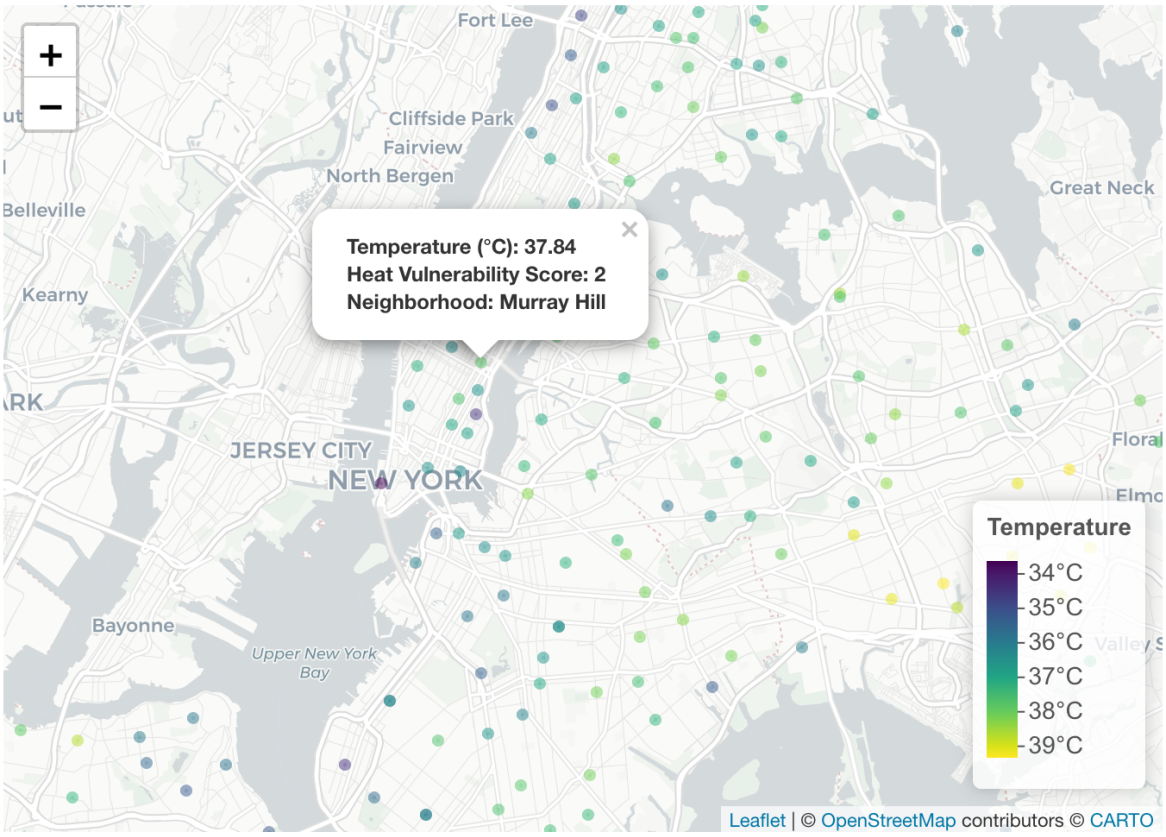
When stratified by borough, there are no differences in the mean daytime surface temperature but there are appreciably different mean heat vulnerability index scores.

Borough	Mean Temperature (°C)	Mean Heat Vulnerability Score
Bronx	37.027	4.086
Brooklyn	36.760	3.408
Manhattan	36.132	2.429
Queens	37.622	2.754
Staten Island	36.233	1.526

Bronx has the highest mean heat vulnerability score, which makes it the borough with the highest at-risk score of heat-related illness and heat-exacerbated deaths. Staten Island has the lowest mean heat vulnerability score, which makes it the borough with the lowest at-risk score of heat-related illness and heat-exacerbated deaths.

Data Visualizations

Using the leaflet package, we mapped the daytime summer surface temperature/heat vulnerability index on a NYC map. A pal() function is generated to establish a color according to the daytime surface temperature of each neighborhood. Please note an interactive version of the map is available on the project’s website.



Based on the map, Manhattan appears to have lower daytime summer surface temperatures and lower heat vulnerability index scores compared to Bronx and Queens.

Regression Analyses

In this section, we investigated if the differences in daytime summer surface temperature and heat vulnerability index were significant across boroughs. We performed univariate regression analyses to understand whether the borough (x: predictor variable, “borough”) influences either the daytime summer surface temperature or the heat vulnerability index score (y: outcome variables, “daytime_surf_temp” or “heat_vulnerability_index).

Univariate Model 1: Daytime Surface Temperature and Borough

In the first univariate model, we investigated the influence of boroughs on daytime summer surface temperature in NYC. The descriptive analyses showed that the mean daytime summer surface temperature was not appreciably different across boroughs. A linear model was created to further explore if there was an association between borough and daytime summer surface temperature.

Term	Estimate	p-value	Lower Confidence Interval	Upper Confidence Interval
(Intercept)	36.1322	0.0000	35.7479	36.5166
boroughBronx	0.8952	0.0008	0.3796	1.4108

boroughBrooklyn	0.6278	0.0115	0.1460	1.1096
boroughQueens	1.4900	0.0000	1.0207	1.9593
boroughStaten Island	0.1006	0.7446	-0.5038	0.7051

The linear model shows that there is a significant difference (p-value≤0.05) in daytime summer surface temperature comparing Queens to Manhattan (reference group). This may explain why the daytime summer surface temperatures in Manhattan are lower than those in Queens. There may be multiple factors that contribute to this, such as the amount of urban green space.

Univariate Model 2: Heat Vulnerability Index and Borough

In the second univariate model, we investigated the influence of boroughs on the heat vulnerability index in NYC. The descriptive analyses showed that the heat vulnerability index score was appreciably different across boroughs, with Bronx having the highest at-risk score. A linear model was created to further explore if there was an association between borough and heat vulnerability index score.

Term	Estimate	p-value	Lower Confidence Interval	Upper Confidence Interval
(Intercept)	2.4286	0.0000	1.9797	2.8774
boroughBronx	1.6571	0.0000	1.0549	2.2594
boroughBrooklyn	0.9796	0.0008	0.4169	1.5423
boroughQueens	0.3258	0.2455	-0.2223	0.8739
boroughStaten Island	-0.9023	0.0131	-1.6082	-0.1963

The linear model shows that there is a significant difference (p-value≤0.05) in heat vulnerability index score compared to Manhattan (reference group). This may explain why the heat vulnerability index scores in Manhattan are lower than those in Bronx and Brooklyn, but higher than Staten Island. There may be multiple factors that contribute to this, such as the amount of urban green space.

Summary of Results

In conclusion, throughout this exploratory analysis, plotting, and regression modeling attempts of n=188 Daytime Summer Surface Temperature and Heat Vulnerability Index Dataset (“envo_health) as of 2018, we conclude:

- Overall, within New York City, the average daytime surface temperature is 36.92°C and the average heat vulnerability index score in NYC is 3.
- The average daytime summer surface temperature is not appreciably different across the 5 boroughs, but the heat vulnerability index scores vary. The Bronx has the highest at-risk score of heat-related illness and heat-exacerbated deaths. Staten Island has the lowest at-risk score of heat-related illness and heat-exacerbated deaths.

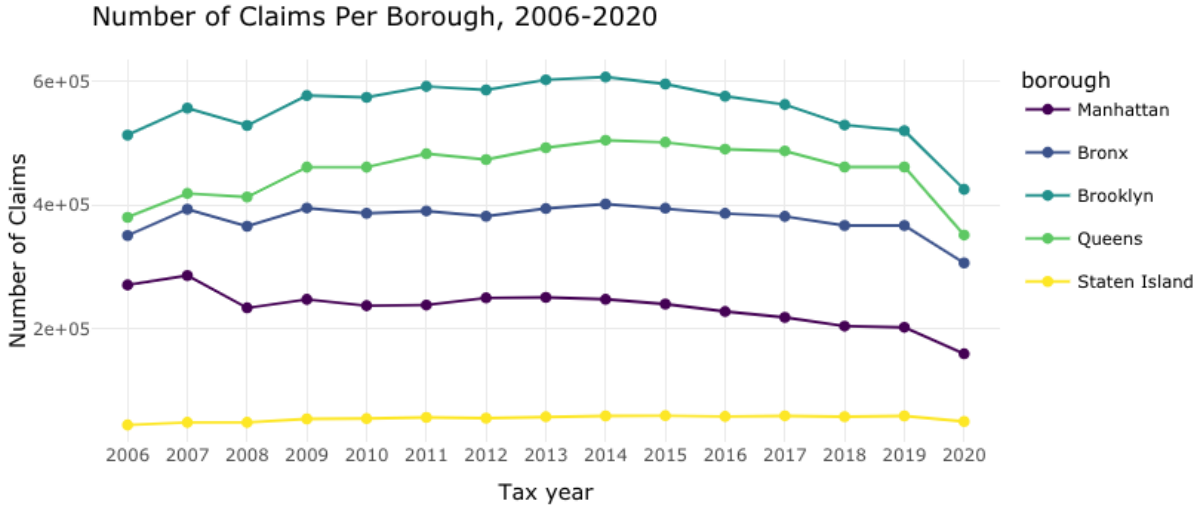
- The differences in daytime summer surface temperatures are significant between Queens and Manhattan (reference group). The differences in heat vulnerability index scores are significant between Bronx and Manhattan, Brooklyn and Manhattan, and Staten Island and Manhattan. These significant differences may be influenced by multiple factors, including the amount of urban green space available to bring down the daytime surface temperature and lower the heat vulnerability index score.
- There are limitations to the temperature analyses performed. The daytime surface temperature dataset was collected on a single-day in July (July 17, 2018), which may not be an accurate representation of the daytime summer surface temperature. This may have created a false association seen in the univariate linear regression model investigating the influence of boroughs on daytime summer surface temperature. A dataset that has collected daytime summer surface temperature each day of summer in 2018 may have been a more accurate representation. The heat vulnerability index dataset was collected from multiple data sources. It is not clear how each social and economic predictor was collected and coded, which may affect the true association of borough and heat vulnerability index. Moving forward, we recommend that daytime summer surface temperature and heat vulnerability index scores are collected similarly, with data on comparable social and economic predictors. This would allow a more accurate representation of the differences and influences of each borough in New York City and further assist the New York City Department of Health in creating policies that benefit the most vulnerable populations.

4. Tax

Descriptive Analysis & Visualisations

Overall EITC Dataset

First, we'll create a graph to show the number of claims in each borough from 2006 onward. Although the data goes back as far as 1994, the City EITC wasn't introduced until 2004 and the Noncustodial Parent EITC wasn't introduced until 2006. For consistency, we'll look at 2006 onward for an idea of the overall number of claims made each year in each borough.

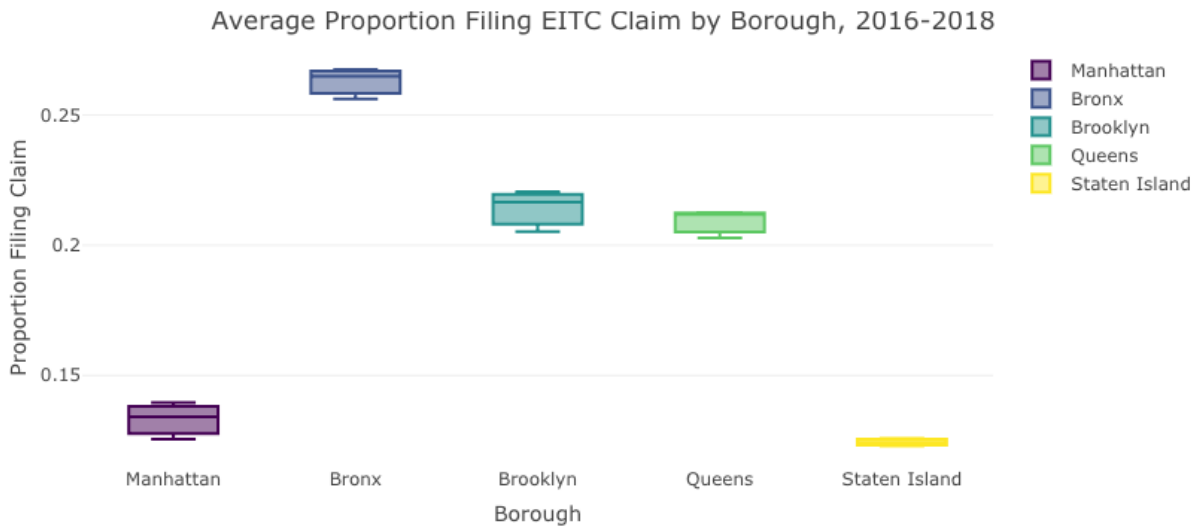


Stratified Analysis by Borough

The graph above shows the overall trend in the number of claims filed per borough, but this relationship scales with the population of each borough. For instance, Brooklyn is the most populous borough, so it logically follows that the most claims were filed in that borough.

Instead of looking at the trend, it would be helpful to understand what percentage of each borough's population files a claim. Each year, The US Census Bureau extrapolates estimated county populace, basing their projections on the most recent Census data and vital statistics. We're going to pull these estimates into a dataframe, merge it with our tax data, then find the proportion of claimants in each borough.

We'll limit the time frame from 2016-2018, the same years that the Greenspace data were collected. The code book for this data can be accessed here.



Regression Analyses

Proportion of Claimants

Graphically, we can see that there is a difference in the proportion of claimants in each borough. Statistically, we have to establish the significance of this finding.

First we'll run an ANOVA F-test to determine if the group proportions are different between boroughs. For purposes of this project, we're using Manhattan as our universal reference group, so we'll then see which proportions differ from Manhattan.

term	estimate	std.error	statistic	p.value
(Intercept)	0.1331000	0.0034701	38.35666	0.0000000

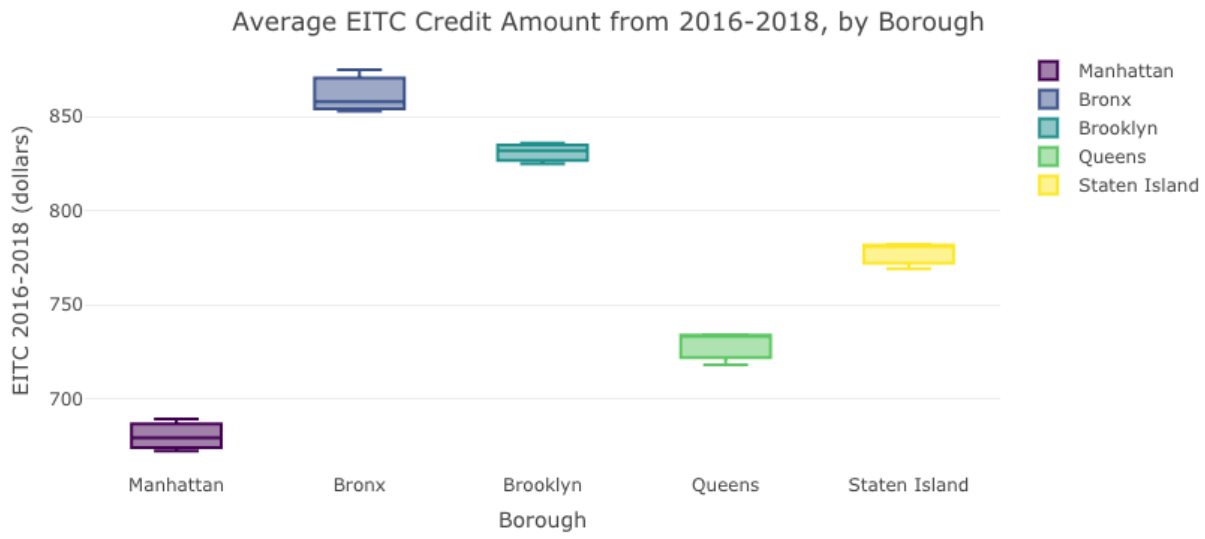
boroughBronx	0.1297667	0.0049074	26.44301	0.0000000
boroughBrooklyn	0.0810333	0.0049074	16.51245	0.0000000
boroughQueens	0.0760667	0.0049074	15.50037	0.0000000
boroughStaten Island	-0.0087000	0.0049074	-1.77283	0.1066616

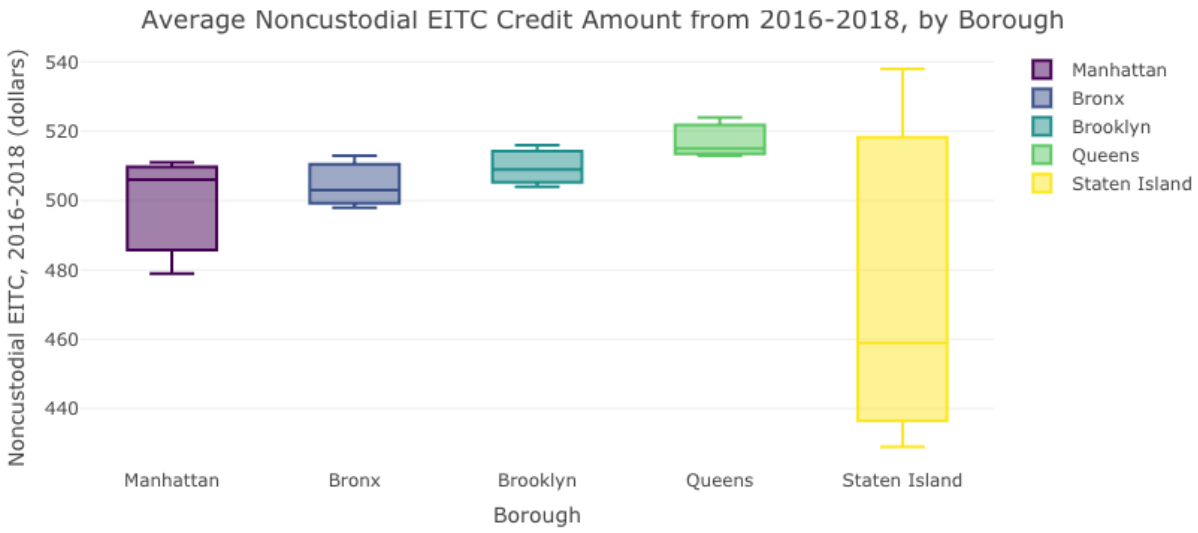
The F-test gives a p-value far less than our accepted significance level of 0.05, so we know there is at least one proportion that is different among the 5 boroughs.

When we analyze against a reference group, we see that Bronx, Brooklyn and Queens are all statistically different from Manhattan. Using our visualizations, we see that the proportion of claimants is higher in these boroughs.

Average Claim Amount

Next, we'll look at the average claim amount 2016-2018 for each borough, stratifying by credit type. Since our primary Greenspace data set only includes data from 2016-2018, we're going to limit our analysis to these years. Since the qualification criteria is the same for City and State EITC, we're going to add the amounts to create a single category called EITC.





Average Credit Amount

Similar to the population proportions, we need to test if these differences are significant.

We'll start with the EITC data and run an ANOVA F-test to determine if the amount claimed is different in between boroughs. We'll then use Manhattan as our reference group to see which boroughs are paying a different amount.

term	estimate	std.error	statistic	p.value
(Intercept)	680.00000	4.964317	136.977551	0.00e+00
boroughBronx	182.00000	7.020605	25.923693	0.00e+00
boroughBrooklyn	151.00000	7.020605	21.508119	0.00e+00
boroughQueens	48.33333	7.020605	6.884497	4.27e-05
boroughStaten Island	97.33333	7.020605	13.863953	1.00e-07

The F-test gives a p-value far less than our accepted significance level of 0.05, so we know there is at least one average amount that is different among the 5 boroughs.

When we analyze against a reference group, we see that Bronx, Brooklyn, Queens and Staten Island are all statistically different from Manhattan. Using our visualizations, we see that the average amount of EITC credit is higher in these boroughs

Repeating the same process, we'll now run the same model for Noncustodial EITC:

term	estimate	std.error	statistic	p.value
(Intercept)	498.66667	15.48261	32.2081853	0.0000000

boroughBronx	6.00000	21.89571	0.2740263	0.7896415
boroughBrooklyn	11.00000	21.89571	0.5023815	0.6262767
boroughQueens	18.66667	21.89571	0.8525261	0.4138742
boroughStaten Island	-23.33333	21.89571	-1.0656576	0.3116315

The F-test gives a p-value of 0.4237. For Noncustodial EITC, at the 5% level of significance, there is no statistical difference in the average amount of credit received among the 5 boroughs.

Summary of Results

Based on this tax data, we can see that the overall trend in the number of claims filed between 2006-2020 scales directly with the population of each borough. When finding the proportion of claimants in each borough, however, we see that there is a difference. When compared to Manhattan, the Bronx, Brooklyn, and Queens all file a higher proportion of claims.

Moving on to the average dollar amount of the credit in each borough, there's no statistical difference in Noncustodial EITC among the boroughs.

For EITC between 2016-2018, however, the Bronx, Brooklyn, Queens and Staten Island all claimed a higher dollar amount in credit, on average.

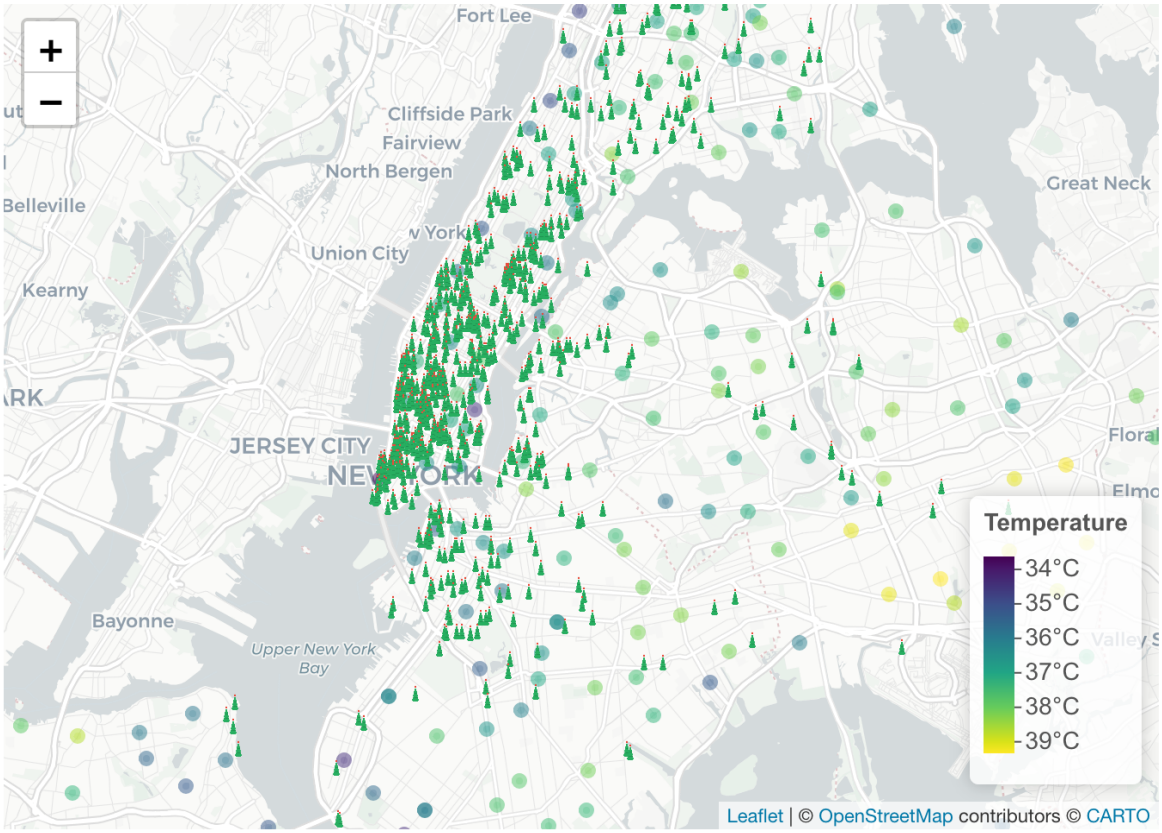
From our visualizations, we can see that Manhattan has the lowest proportion of residents filing an EITC claim, and that the credit amount is generally lower compared to other boroughs.

5. Joint Multivariate Analysis

Mapping UGS and Daytime Surface Temperature/Heat Vulnerability Index in NYC

Using the leaflet package, we combined the UGS and daytime surface temperature/heat vulnerability index on a NYC map. A pal() function is generated to establish a color according to the daytime surface temperature of each neighborhood. A addMarkers() function was used to generate *Christmas trees*, which represent the urban green space. **Please note an interactive**

version of the map is available on the project's website.



Regression Analyses: Univariate Models

In this section, we investigated if there is an association with urban green space and daytime surface temperature/heat vulnerability index. We performed univariate regression analyses to understand whether the total number of urban green space (x: predictor variable, "total_green_area") influences either the average daytime surface temperature or the average heat vulnerability index score (y: outcome variables, "avg_daytime_temp" or "avg_heat_index).

Univariate Model 1: Total Urban Green Space and Average Daytime Surface Temperature

In the first univariate model, we investigated the influence total urban green space has on the average daytime surface temperature in NYC.

Term	Estimate	p-value	Lower Confidence Interval	Upper Confidence Interval
(Intercept)	37.18428514	0.00000312	36.3662725	38.00229782
total_green_area	-0.00000246	0.28700715	-0.0000062	0.00000127

We found that there is no significant association (p-value≤0.05) between average daytime surface temperature and total urban green space.

Univariate Model 2: Total Urban Green Space and Average Heat Vulnerability Index

In the second univariate model, we investigated the influence total urban green space has on the average heat vulnerability index in NYC.

Term	Estimate	p-value	Lower Confidence Interval	Upper Confidence Interval
(Intercept)	2.91245543	0.03900296	1.28933447	4.535576
otal_green_area	-0.00000041	0.92015460	-0.00000782	0.000007

We found that there is no significant association (p-value≤0.05) between average heat vulnerability index and total urban green space.

We have decided to move forward without temperature and heat vulnerability index in the joint analyses investigating the influence of urban green spaces due to the insignificant findings in univariate models 1 and 2.

Comparison of Asthma Rates with UGS

Next we will compare the UGS dataset with the asthma dataset, first visually and then also based on linear models.

We load and wrangle the **two primary Urban Green Spaces (UGS) dataset** to generate a **UG_df dataframe**, as well as the **asthma dataset** to generate a **asthma_df** dataframe, as described in the respective pages dedicated to the exploratory analysis of the datasets.

Next, we create a value of green spaces that we can compare to asthma rates, namely the cumulative surface area of vertical greenspace constructed between 2006-2018 in each borough.

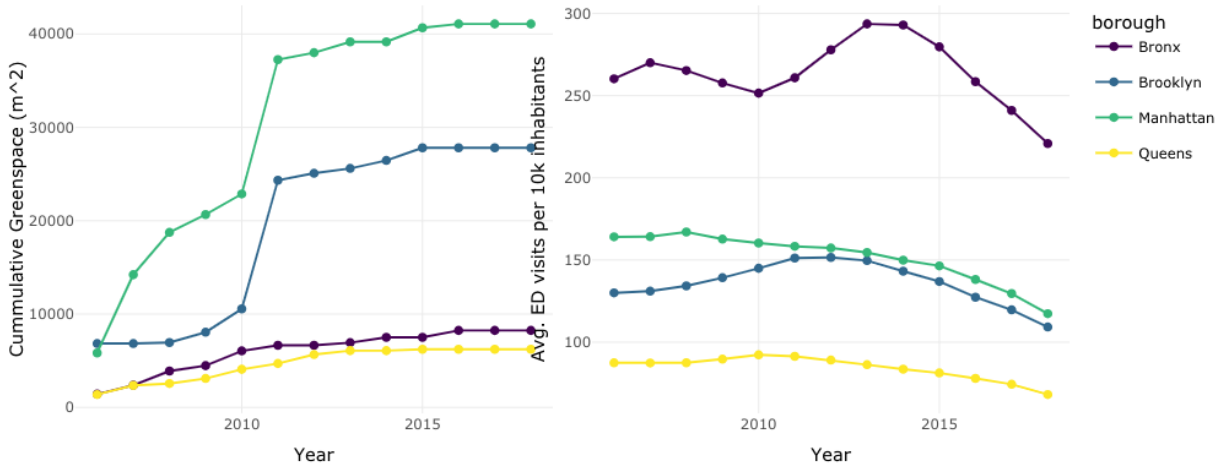
borough	cons_year	green_area_sum	green_area_cum
Bronx	2006	1430.45	1430.45
Bronx	2007	942.75	2373.20
Bronx	2008	1525.18	3898.38
Bronx	2009	574.71	4473.08
Bronx	2010	1579.57	6052.65
Bronx	2011	606.97	6659.62
Bronx	2013	276.50	6936.12

Bronx	2014	566.23	7502.36
Bronx	2016	741.25	8243.60
Brooklyn	2007	6843.79	6843.79
Brooklyn	2008	103.05	6946.84
Brooklyn	2009	1111.39	8058.23
Brooklyn	2010	2496.56	10554.80
Brooklyn	2011	13776.28	24331.08
Brooklyn	2012	752.78	25083.86
Brooklyn	2013	524.56	25608.41
Brooklyn	2014	854.42	26462.83
Brooklyn	2015	1344.79	27807.62
Manhattan	2006	5838.59	5838.59
Manhattan	2007	8377.50	14216.09
Manhattan	2008	4532.25	18748.34
Manhattan	2009	1912.66	20661.00
Manhattan	2010	2205.54	22866.53
Manhattan	2011	14391.25	37257.78
Manhattan	2012	742.04	37999.82
Manhattan	2013	1165.55	39165.37
Manhattan	2015	1507.17	40672.54
Manhattan	2016	411.37	41083.91
Queens	2006	1379.60	1379.60
Queens	2007	980.11	2359.72
Queens	2008	202.25	2561.97
Queens	2009	541.88	3103.84
Queens	2010	978.16	4082.00
Queens	2011	619.90	4701.90

Queens	2012	961.81	5663.71
Queens	2013	416.90	6080.61
Queens	2015	151.63	6232.24
Staten Island	2011	709.47	709.47

Visualizations on UGS & Asthma Rates

We first visualize the result and compare it to the graph created on ED visits (taking that as a proxy for asthma rates). One trend is clear: while asthma rates seem to have declined, there seems to have been an increase in UGS. The patterns are not consistent across borough, e.g., there has been only gradual improvement in UGS in the Bronx, while the decline in asthma was more drastic. In short, there must surely be many other factors at play that have caused asthma rates to decline, but potentially UGS did play an effect in reducing the numbers.



Regression Analysis of UGS and Asthma Rates

Let's run a couple of linear regressions to see whether they produce any statistically significant results.

First, we are creating a combined df consisting of both UGS and asthma datasets by using a (double) leftjoin based on borough and year. We are using na.locf() to replace missing UGS values with the value from the last year (i.e, if there was a year with no greenspaces added, then that year

must have had the same total number of greenspaces as the last year). Due to the way the dataframe is structured, this requires a manual insertion of a value at two points of the dataset.

First we are creating separate dataframes for each indicator and set Manhattan as the comparator.

The table below show the results of linear regression of the indicator (either ED visit rate, Hospitalization rate or severity measure) on a combination of borough and total UGS surface area.

Table 1: ED visit rate vs. borough and cumulative UGS surface area

term	estimate	std.error	statistic	p.value
(Intercept)	169.474	8.610	19.684	0.000
green_area_cum	-0.001	0.000	-2.332	0.024
boroughBronx	97.871	8.244	11.872	0.000
boroughBrooklyn	-22.158	6.138	-3.610	0.001
boroughQueens	-82.420	8.497	-9.700	0.000

Table 2: Hospitalization rate vs. borough and cumulative UGS surface area

term	estimate	std.error	statistic	p.value
(Intercept)	41.361	4.235	9.767	0.000
green_area_cum	-0.001	0.000	-4.889	0.000
boroughBronx	17.994	4.055	4.438	0.000
boroughBrooklyn	-3.523	3.019	-1.167	0.249
boroughQueens	-21.433	4.179	-5.129	0.000

Table 3: Severity vs. borough and cumulative UGS surface area

term	estimate	std.error	statistic	p.value
(Intercept)	0.258	0.020	12.779	0.000
green_area_cum	0.000	0.000	-6.070	0.000
boroughBronx	-0.026	0.019	-1.327	0.191
boroughBrooklyn	0.002	0.014	0.110	0.913
boroughQueens	-0.041	0.020	-2.045	0.046

To our surprise, all three models return statistically significant values for the effect of cumulative UGS surface area on the studied indicators, while accounting for borough differences. The direction of the effect is what we would expect: increase in cumulative UGS surface area leads to a reduction in ED visit rate, hospitalization rate, and our severity measure. We conclude that urban green spaces might indeed have played a role in reducing asthma rates across New York City.

Comparison of Earned Income Tax Credit (EITC) with UGS

When trying to find some association between greenspace data and average EITC tax credit amount, we will first look at the outcomes over time visually to get an understanding of the overall trend.

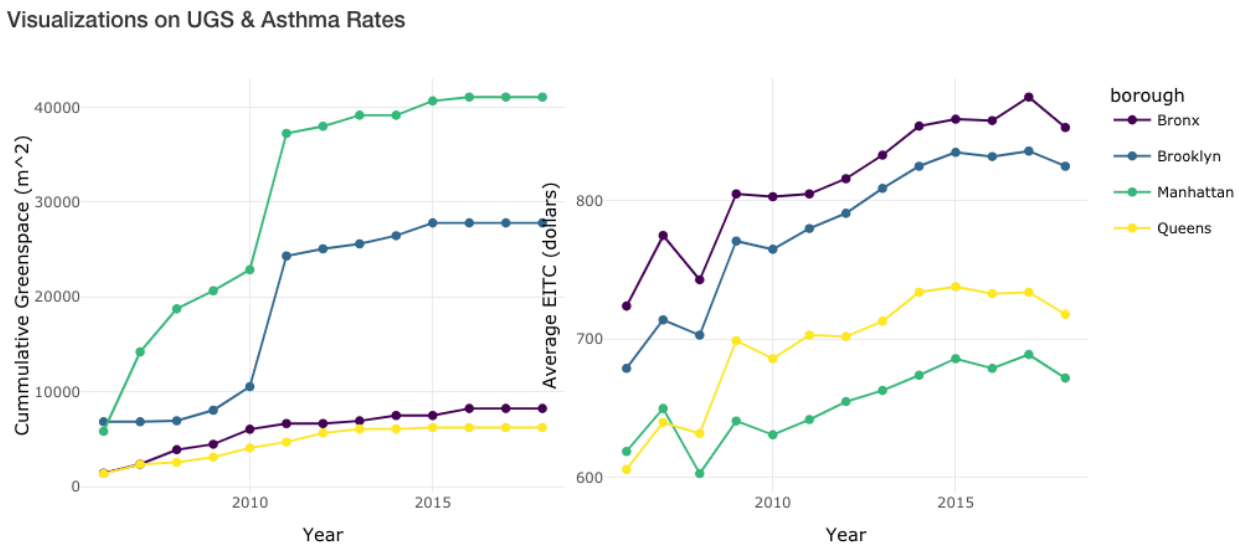
Our hypothesis is that there will be more rooftop greenspaces present in boroughs where the average EITC credit amount is lower.

To create this visualization, we have to make a few assumptions.

The Greenspace data does not specify when rooftop greenspaces were constructed. We know that the data was collected between 2016-2018 and have information on when the building was constructed but that's all the detail we're given. Meanwhile, EITC tax data for all EITC categories is accounted for from 2006 onward.

We are going to assume that buildings constructed from 2006 onward were constructed initially with their rooftop greenspace. Given the cost of construction and the logistics involved with residential building in New York City, we're assuming that developers did not retroactively add greenspace after initial construction of these modern buildings.

We're also going to eliminate Staten Island from this visualization. Staten Island doesn't have enough data on Greenspace to make it reasonable to include in a visualization.



The visualization shows what we initially thought: Manhattan rooftop greenspace cumulative area has rapidly increased in the time frame while their average EITC credit amount has stayed the lowest of all the boroughs. Meanwhile, the Bronx has consistently had the highest EITC credit amount while having a very low rate of increase in rooftop greenspace. **Please note that an interactive version of this visualization is available on the project website.**

Again, this is correlation, not causation but it does speak to the idea that rooftop greenspaces tend to appear in wealthier areas.

Regression Analysis of UGS and EITC

term	estimate	std.error	statistic	p.value
(Intercept)	-46565.675	13448.911	-3.462	0.001
eitc	118.203	20.392	5.796	0.000
boroughBronx	-43825.559	4084.993	-10.728	0.000
boroughBrooklyn	-26477.998	3554.540	-7.449	0.000
boroughQueens	-30924.661	2558.839	-12.085	0.000

Although making a regression model with UGS as the outcome and EITC as a predictor requires several assumptions, we see generally that our hypothesis holds true: as EITC increases, the amount of UGS present in a borough trends downward.

Full Model: Regression

term	estimate	std.error	statistic	p.value
(Intercept)	11165.658	1913.304	5.836	0.000
eitc	0.336	0.076	4.437	0.000
boroughBronx	64.616	14.571	4.435	0.000
boroughBrooklyn	-55.426	10.945	-5.064	0.000
boroughQueens	-74.117	9.141	-8.108	0.000
green_area_cum	0.000	0.000	0.903	0.372
tax_year	-5.587	0.972	-5.746	0.000

Discussion

The full regression model uses ETIC, borough and cumulative green area as predictors of ED visits per 10,000 inhabitants. The model shows statistically significant results for two boroughs, using Manhattan as a reference group: the Bronx and Queens. This is an effect we have already observed in univariate analysis of the asthma data set. Unsurprisingly, the model did not yield statistically significant results for the effects of cumulative green space and tax on asthma rates. However, the direction of the effect is as expected: the model does show lower asthma rates for higher cumulative greenspace. This effect was not statistically significant, however, the p value of 0.098 is moderately strong. What is unexpected is the effect of tax on asthma rates: asthma rates decrease as EITC increases in this model. We would have expected the opposite relationship, however the effect in this model was not statistically significant, with a p-value of 0.91.

Limitations

As already discussed, this was a hypothetical public health model that could have been applied in the real world. Because of the limitations of our data (we had to make do with what we could source online) as well as the fact that asthma is a multifactorial disease (meaning many factors contribute to its development), this isn't a perfect representation of the real world.

We have to also acknowledge that these data sets were collected from different studies using different methods. Ideally, for a social determinants of health model like we were trying to build, the data would have been collected in one unified study. We didn’t have that luxury, so we did the best we could with what was available.

The timeline of the various data sets also varied. EITC Tax data was incredibly consistent, as one would expect for tax information. It spanned from 1994-2020, with each element of the EITC system included as they were introduced (i.e. City EITC’s introduction in 2004). Asthma data was similarly consistent. It measured ED visits and Hospitalizations between 2006-2018.

The temperature data was relatively consistent and thorough, but we were only able to access records from 2018. This is what ultimately led us to not include temperature as a predictor in our model.

The greenspace data timeline was also tricky to work with. As noted in the study, the greenspace area was calculated via satellite imaging taken between 2016-2018. The researchers were able to match the greenspace to coordinates and to building addresses, but they were not able to provide information on when the greenspace was constructed. For instance, there were buildings in lower Manhattan that were constructed in 1893 and had greenspace, but we were making an educated guess that the greenspace was not constructed at the time of the building. With the data provided, we had no way of knowing this for sure. For our analysis, and based on our group member’s expertise in real estate development, we

decided to assume that buildings constructed from 2006 onward were likely built with greenspace in their original plan, which allowed us to calculate cumulative space in each borough from that point onward and compare it with our other predictors.

Ideally, we would have also liked more granularity in the Greenspace data. The study was very Manhattan-centric, which is why we used Manhattan as our universal reference group throughout the project. As students in Manhattan, we know that there are deep disparities from neighborhood to neighborhood within the borough. When looking at the leaflet of greenspace, it's evident that most are concentrated below 14th street which also corresponds to some of the wealthiest neighborhoods in the city. It would have been interesting to break Manhattan down further into zip code across all predictors to be able to make deeper comparisons.

Conclusions

This project has given us the opportunity to apply the tools and theories that we've learned this semester and that are very relevant to cutting-edge research within the public health world. It also showed us that good, consistent data is difficult to collect and even more difficult to manage.

Ultimately, our project showed the need for more studies that take a health systems-related approach to their data collection and methodology. While we were able to establish some significance surrounding greenspace and EITC credits effects on asthma rates, we had to make several assumptions with this data.

As we referenced in our introduction, this group was fascinated by Dr. Merlin's lecture about how public health is truly everything. We know that decisions like redlining created the city that we're familiar with today, but modeling the lingering effects of these technically defunct policies is difficult when individual studies examine one predictor at a time. Health, and the maintenance of it, is a complex outcome with numerous predictors. We tried to show that train of thought in this project, and although not perfect, we did begin to tell that story.