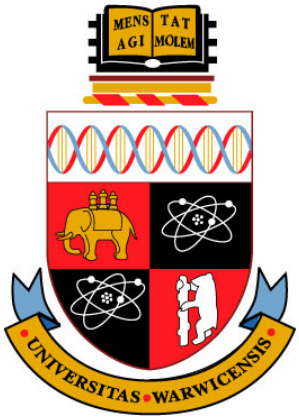


Understanding Public Resource Equity in New York City Using Open Data

CS310 Computer Science Project

Final Report

Emma Dutton

Supervisor: Dr. Matthew Leeke

Department of Computer Science
University of Warwick

2016-17

Abstract

This project documented in this report aimed to understand the equity of public resources in the context of New York City. A measure for this equity was derived by quantifying the accessibility that citizens had to a variety of state provided facilities across all five boroughs. Key themes were identified to guide analysis, namely educational standard, environmental impacts, healthcare provisions, transport quality and accessibility to public services. Data representing these key themes was sourced to provide scope of the project, and in turn served as input to the system developed. This system contained a number of analytical scripts that utilised current practices in the data science discipline. In particular, a variety of visualisations were created using the Plotly graphing library. These visualisations were interesting and dynamic, and told a story about the lives of New York residents through emerging patterns and clusters of behaviour. The results gained from this highlighted that there was areas of inequity in parts of the city. For example, Manhattan had a disproportionally low number of fire fighting provisions in comparison to its population, especially downtown. Additionally, it was observed that Brooklyn residents had the lowest access to mental health facilities compared to any other borough. These results inspired a set of recommendations to be made, that indicated possible changes to public policy which could realign the balance of equity in favour of more people across the city of New York.

Keywords: *Science of Cities, Data Analytics, Data Visualisation, Public Policy*

Acknowledgements

The project could not have been completed without acknowledging certain individuals that were key to its success. Firstly, the largest amount of appreciation is extended to the project supervisor, Dr. Matthew Leeke. The reasoning for this regard is two-fold; primarily, for initially sparking the ideas that led to the project's inception, which, over the course of the last nine months has reaffirmed my enthusiasm for using technology for social good. Utilising technology for the betterment of people around the world has provided me with great enjoyment throughout the project and something I aspire to continue into my career. Secondly, his constant support cannot go unnoticed. For the duration of the year he has sacrificed a large amount of his time by providing advice, direction and encouragement both in the project setting and other academic endeavours. Without this advocacy, my experience throughout this year and the time of my degree would not have been as as exceptional as it has been. In addition to this, I would also like to thank Dr. Alexander Tiskin, who was the second assessor in this project. His comments in the presentation were insightful and considerate, allowing a different perspective of the work to be gained. The observations he contributed were extremely beneficial to the project's direction, and resulted in the problem definition being refined to provide additional clarity around key themes. Finally, I also acknowledge the camaraderie of my peers during this time, who provided support and perspective throughout the duration of the project.

Abbreviations

DOB	Department of Buildings
DOE	Department of Education
EDC	Economic Development Corporation
FDNY	Fire Department of New York
HPD	Housing Preservation and Development
NYC	New York City
NYPD	New York Police Department
UN	United Nations
CSV	Comma-separated Values
JSON	JavaScript Object Notation

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

The earliest cities grew out of a common need to provide physical protection, import resources and export production [10]. As cities have developed, so have the needs of those who reside within them. With the global population reaching 7.4 billion, more than half of whom living in urban areas, questions relating to the sustainability of cities have never been more urgent [65]. Chief among issues in urban sustainability is the question of resource equity. Simply put, does being born in a certain region of a city make it more likely that you will receive a better education, a higher wage, or better healthcare? Many argue that, by combining the power of modern data science techniques with newly published civic data, municipal policies can be recommended that will benefit the lives of thousands of people [16]. Substantiating this, recent studies have demonstrated that data-centric approaches can be used in identifying areas of social deprivation and propose practicable solutions in addressing the associated challenges [81].

1.1 Motivation

Data science is an emerging new area of science that promotes the interdisciplinary study of many subjects, such as Computer Science, Mathematics, Economics, and Sociology. Due to the ever increasing amount of data produced by systems and devices, it is a discipline that in the UK alone, has grown more than ten-fold in the past five years [73]. The aim of data science is to solve complex problems by exploring a variety of data using statistical measures to identify trends and patterns. The rising urbanisation prospects shown in Figure 1.1 highlights the immense growth of urban areas, and shows how attention is required to utilise statistical measures to improve the lives of people residing in populous areas. Data science has been successfully applied to the study of cities to *'enable future cities to deliver services effectively, efficiently and sustainably, whilst keeping their citizens safe, healthy and prosperous'* [74]. There has been many successes in using data science techniques to solve civic issues, such as reducing crime and influencing municipal budgeting [14].

The largest metropolitan areas in the world differ vastly in wealth, culture and politics. However a similar set of challenges can be identified in different cases, indicating that the issues

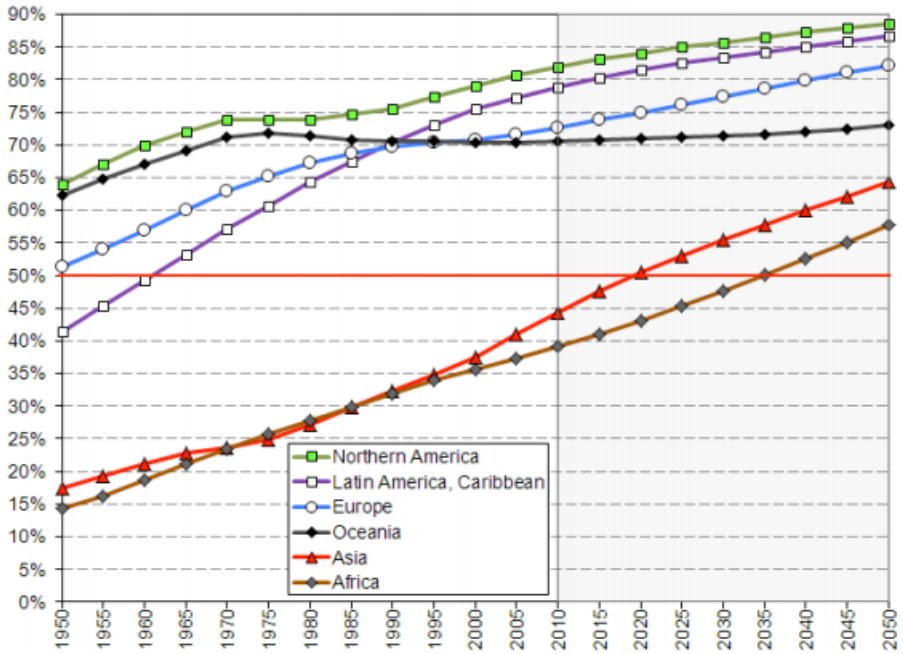


Figure 1.1: The World Urbanisation Prospects [69]

surrounding mass population are a homogeneous on a global scale [1]. Preliminary research identified that New York City was a unique subject due to its diversity across a multitude of metrics such as ethnicity, linguistics and age [66, 55]. To elucidate this claim, Figure 1.2 highlights the geographical layout of the city, where wealthy boroughs such as Manhattan are situated next to poorer boroughs such as the Bronx. This geographical wealth distribution allows comparisons to be made on the basis of locality. New York City strives to contend with many issues synonymous to those of other large global cities. Challenges such as crime, pollution and disease are at the forefront of governmental discussion and continue to consume public sector resources dedicated to identifying solutions for the current urban landscape. The project was proposed to contribute to the integration between data analytics and civic problem solving by investigating whether public sector services are equally accessible to residents residing in all areas on the city. The project utilised the wealth of preceding research to derive statistic methods that evaluated the state of fairness in resource allocation amongst citizens of different sociological backgrounds.

1.2 Project Aims

The overarching aim of the project was to assess resource allocation of public services in New York City. Resources in this context were defined as state funded facilities, such as schools, hospitals or transport systems. Research has shown that improving accessibility to these amenities creates happier and more successful communities [70]. By evaluating each resource area, a measure can be define describing how equitable the service is to the citizens of New York City, regardless of location or social standing. This measure may indicate unintentional partisanship that has lead to an unfair distribution of public services favouring certain individuals.

Figure 1.2: Map of New York Boroughs [59]

The project aimed to identify if this hypothesis was correct, and if so, recommend policies to rebalance the impartially of these provisions.

The project aimed to utilise the results of analysis by providing evidence-based recommendations that would improve the distribution of resources. Previous research on equity provision in New York City has focused on specific subject areas such as education or healthcare [79, 15]. Whilst it was important to examine these factors, the project differed in approach by adopting a holistic mechanism to explore a broad range of possible contributing factors. It was assumed that by evaluating a range of issues that affect the majority of citizens on a daily basis would facilitate a greater understanding of the resource allocation across the city. Due to the political nature of the discussion, the project did not aim to argue a 'right way' to allocate resources, but instead provide insight into areas of possible disadvantage amongst current policies.

1.3 Stakeholders

The project had two main stakeholders. The first was the project supervisor, Dr. Matthew Leeke, as he actively contributed to the direction of the research throughout the duration of the project and was invested in the success of the results. Secondly, the project leader, Emma Dutton, was also a stakeholder in the process as developed the analytical system in hope that the findings were rewarding and beneficial to the wider community.

1.4 Report Structure

The following report will document the inception, development and results of the project. This document has been written to provide the audience with a thorough understanding of how the project was undertaken, and highlights the successes that the project achieved. The report is structured in the following way.

Research and Design
Chapter 1 of the report has introduced the problem in hand; are resources across New York City distributed equitably amongst its citizens? Here, background material has been discussed to highlight the relevance of the question to the real world. Chapter 2 extends the discussion of background material and provides extensive contextual research on the project domain. Initially, the research focuses on providing a breadth of information about the geography, demographics and government of the city. This knowledge is utilised throughout the project and provides a set of prior beliefs about the current state of the city. Additionally, this research allows the identification of important factors that affect the majority of the citizens who reside in the city. Finally, the research section is concluded with a discussion of the most appropriate technical tools to utilise in the development of the system. Chapter 3 considers the legal, ethical, social and professional issues surround the topic. After the initial research was conducted on both context and professional issues, the design of the system was conceived. Chapter 4 discusses the functional and non-functional requirements of the proposed system, providing goals that allowed the progression to be monitored throughout the development phases. These requirements also served as measures of success in the evaluation phase of the project. Chapter 5 discusses the methodology of the system. The methodology builds on the initial requirements set out in the previous chapter, by proving development phases to meet all required functionality of the system. The methodology further elucidates how the development process will be orchestrated, allowing time to be managed as efficiently as possible. This structure is then mirrored throughout the rest of the report, with the following chapters exploring each phase of the methodology in turn.

Analysis
Chapter 6 highlights the first phase of the methodology which is Data Ingest. In this chapter, the processes for discovering relevant data are explored, with consideration given to standardisation techniques. Chapter 7 reviews the analytically phase of the methodology through single factor analysis. Here, each of the factors are analysed from a wealth of datasets that explore different civic data. In each case, a data visualisation is provided and a discussion is made into the impact of these results. This iterative inspection allowed results from each run of the analysis to be explored, with the most interesting and fruitful findings summarised at the end of the chapter. Subsequently, Chapter 8 extends the discussion of the analytical component of the system by further exploring key questions raised in the previous chapter. The analysis presented in Chapter 8 expands on the analytical techniques used previously, by highlighting the importance of fusing datasets to gleam further insight. Chapter 9 utilises the results found in previous chapters by discussing their context and limitations in the society of New York City. This discussion will evaluate the outcome of the results and provide recommendations to improve the equity of public resource provisions.

Evaluation
Chapter 10 identifies the validity of the system through rigorous testing strategies. The different types of testing are highlighted in this chapter and results are discussed. Chapter 11 expresses the project management procedures that were put in place to manage the quantity of work, whist ensuring quality was maintained. This chapter discusses the project timeline, management tools and methodologies used to keep the project on track. Additionally, the risks of the project are noted in this chapter, with mitigations given that reduced the chance of failure.The penultimate Chapter 12 evaluates the project by discussing the successfulness of the results. Evaluation is also given to the project management of the system, to ensure that work was made efficiently. The section summaries the challenges encountered and provides and oversight of the authors assessment of the project. Finally, Chapter 13 concludes the project by reassessing the aims of the project, whilst considering the successes explored in the previous chapter. This chapter continues to explore further work that could be undertaken to improve the results of the system, and ideas that could expand the future use of the developed system.

CHAPTER 2

Research

Due to the multidisciplinary breadth of the project, it was necessary to conduct comprehensive and thorough research of a variety of areas. To enable this, research was broken down into topics to gain a broad understanding. These topics were the contextual study of New York City, reviewing literature of research conducted utilising data science techniques to produce city analytics and technical platform evaluation. A diagram depicting how research was structured is shown in Figure 2.1. Only after these facets were explored thoroughly could the system be designed and results understood. The following section will describe the understanding gained from the research carried out, with recognition of the literature reviewed.

2.1 Geography and Demographics

New York City is a single city that is a part of the wider New York State, situated on the east coast of America. Its location between the Hudson River and the Atlantic Ocean provides a small land mass positioned in a naturally sheltered harbour. Throughout history New York City has been home to a diverse immigrant community, which is documented as far back as 1609 when Dutch settlers took residence on the land. The cities strategic location on the seafront between Britain and America made it advantageous for the import and export trade of the 1700's. This prolific trade industry coupled with the industrial boom following the Great Depression allowed New York City to become one of the most important urban areas in the world, drawing in millions of tourists and migrants each year [60, 85, 21].

The scarcity of land in New York City has lead to it being one of the most densely packed cities in the United States, giving rise to density management being an important environmental issue. However in attempts to counteract this difficulty, the continuous efforts of the government have enabled it to become one of the most efficient cities in America thorough proficient waste management systems, affordable housing and accessible transportation. An article written by David Owen explains how *'the key to New York's relative environmental benignity is the very thing that makes it appear to be an ecological nightmare: its extreme compactness'*. Evidence supports this by highlighting how the close proximity of destinations in the city has reduced the

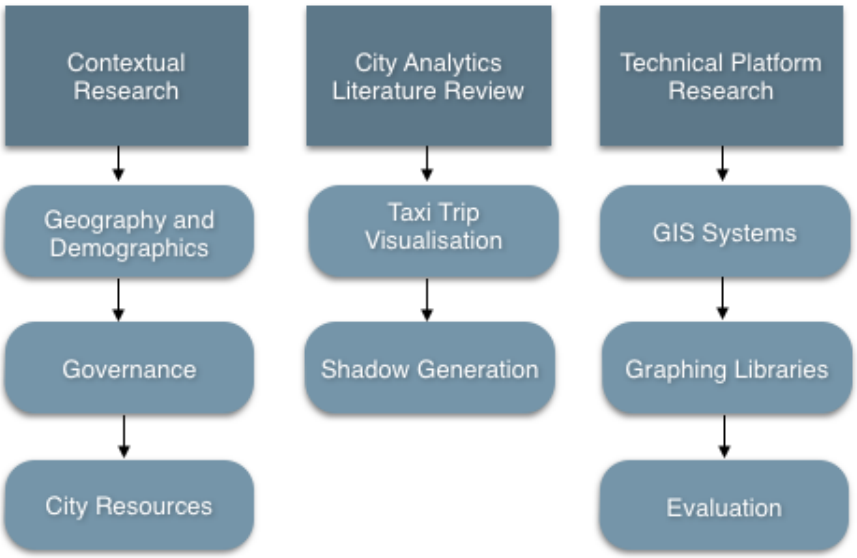


Figure 2.1: The Research Structure Followed to Gain Understanding of the Problem Scope

need for automobiles and encouraged citizens to walk or utilise public transport. These factors have lead to New Yorkers having the smallest carbon footprint in the United States [24]. Noting these surprising correlations between the geographic environment of New York City and energy consumption was important, as common factors such as land based proximity may play a role in the projects analysis and discussion of resource allocation.

New York City houses over 16 million citizens across 5 boroughs; namely the Bronx, Brooklyn, Manhattan, Queens and Staten Island. These boroughs are extremely diverse in nature, combining different cultures to produce the cosmopolitan atmosphere that New York is famous for. This chapter will further explore details of each borough, which provided the contextual understanding for the project [17].

2.1.1 The Bronx

The Bronx is situated in the north of New York City. It is historically know as the birth place of rap and hip hop culture due to its high populous of African American residents during the 1990's [76]. The Bronx has been home to many nationalities, with the borough going through rapid growth after World War One during which many Irish, Italian and Jewish people began to settle in the area [76]. A decade later, in the prohibition days surrounding 1926, it was known for having one of the highest crime rates, with many speakeasies selling illicit alcohol.

Currently, the Bronx is the site of the Yankee Stadium; a 50,000 seated stadium home to the New York Yankee's baseball club [75]. Additionally, the Bronx holds the United State's largest cooperatively owned housing development, aptly named 'Co-op City' [5]. Co-op City provides affordable housing to residents, as well as offering three shopping centres, six schools, a weather station and a planetarium [5, 86]. Co-op City is so vast that if the area was classed as a city, it

would be the 10th largest city in the United States [86].

2.1.2 Brooklyn

Brooklyn is the located in the south west part of the city, sharing a border with Queens. It has the largest number of citizens compared to the other boroughs, with the population being marked at 2.592 million as of 2013 [72, 28]. Brooklyn is know for its cultural diversity, that has given rise to a unique architectural heritage and independent art scene [19]. As of 2007, the top 5 ethnicities in Brooklyn were African American (15.2%), Religious Responses (7.4%), Puerto Rican (6.0%) Italian (5.8%) and Chinese (4.7%) [72]. It is also notable that the top 5 languages spoken in Brooklyn were English, Spanish, Chinese, Russian and Yiddish, contributing to a diverse linguistic environment [27, 72].

2.1.3 Manhattan

The most populous borough in New York City is Manhattan. It is located centrally to the city and houses the business and financial districts. This has lead to the borough being lined with city scrapers, the largest being the One World Trade Center [63]. Additionally, Manhattan was the location of the Twin Towers, which were disastrously bombed in 2001. This has lead the area to have increased security and police presence [4].

Manhattan also hosts many of the tourist attractions that are iconic to New York City, such as the Empire State Building, Rockefeller building and Central Park. These high levels of tourism contribute to the economy, and worldwide presence of New York City. Due to the large footfall provided by tourism and industry, Manhattan contributes to 90% of the taxi demands of the whole city [3].

2.1.4 Queens

Queens is north of Brooklyn, with the largest land mass in comparison with the other boroughs. Queens is mainly a residential area for middle class citizens but also the most ethically diverse county in the United States. Its population consists of a diverse range of people, with 50% being white, 28% Hispanic, 24% Asian, 21% black, and 3% mixed race [72]. Additionally, Queens also has a relatively low education rate compared to other boroughs, with only 27% having a Bachelors Degree level education [17].

2.1.5 Staten Island

Staten Island is a detached from the main body of New York City and separated by water, situated in the south west of the region. Due to its distance from the main centre of Manhattan, it has grown more of a residential suburban character. It is connected to Brooklyn by the Verrazano-Narrow bridge, and to Manhattan by the free Staten Island ferry, proving it to be popular for commuters to the centre of the city. Staten Island was home to one of the largest landfill sites in the world, named Freshkills Landfill, which is currently undergoing transformation into one of the largest urban parks in America, almost three times the size of Central Park. This park is due for opening in 2036 [39, 18].

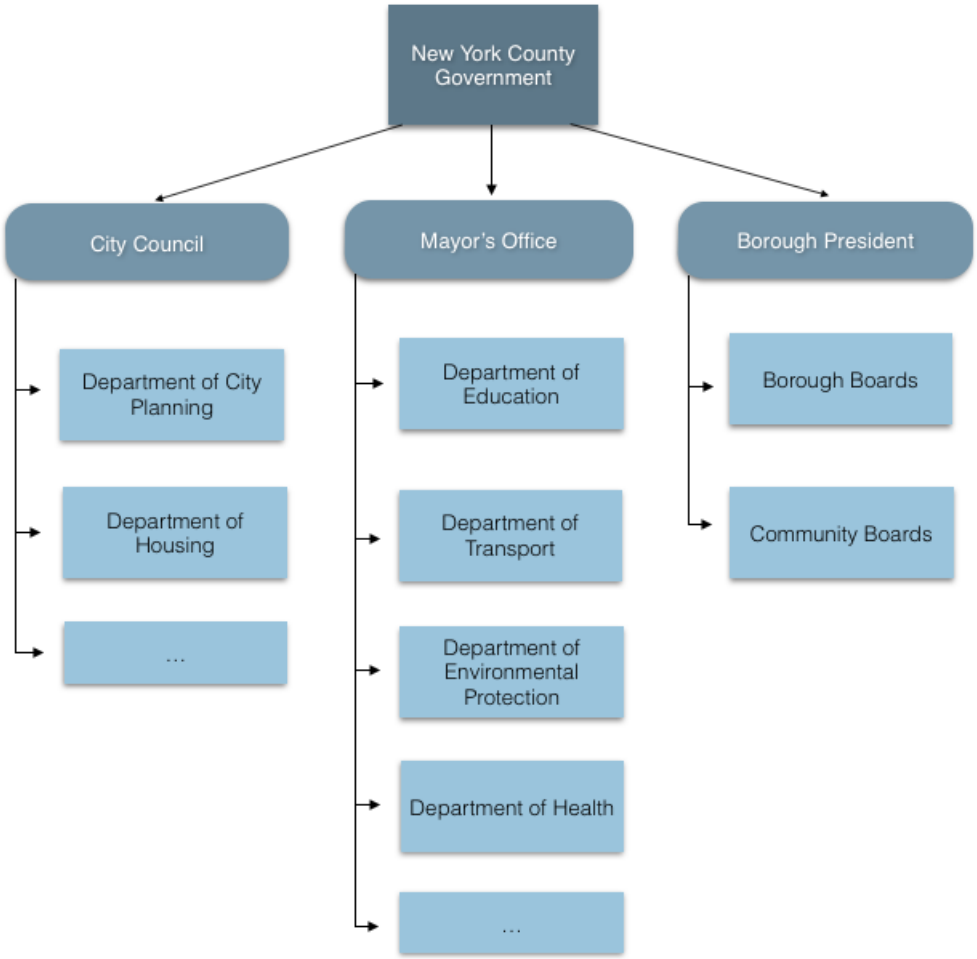


Figure 2.2: A Subsection of the Organisational Structure in the Governance of New York [61]

2.2 Governance

To understand how the city of New York is managed it was important to be aware of the governmental structure that serves the city. The governance of New York is split across three entities that have specific legislative powers, headed by the county government. These entities the City Council, Mayorship and Borough Presidents, as shown in Figure 2.2.

2.2.1 New York County Government

In New York State every county has a governing power. As New York City is part of New York State, it is presided by the New York County Government. The role of the county government is to provide state mandated services, such as management of the police force, maintaining roads and transport and providing economic development assistance [45].

2.2.2 Legislative Branch - New York City Council

New York City has one council that is responsible for the legislative branch of the government structure. City charters are passed to enable an organisational structure of the city, for example, the name, boundaries and administrative processes. The city council is ran by elected council members which represent their own borough. The council have responsibility for departments that are tasked with the management of the city, such as the Department of City Planning, City Housing Authority and Department of Parks and Recreation [49].

2.2.3 Executive Branch - Mayorship

The elected Mayor heads the city's executive branch from city hall, and has jurisdiction over the five boroughs of New York City. There are around 50 departments that are appointed by the Mayor, for example, the Department of Education who managed the city's schools [50]. The Mayor of New York City is reelected every four years, with the position currently held by Bill de Blasio.

2.2.4 Borough President

There is one borough president for each of the five boroughs in New York City; the Bronx, Brooklyn, Manhattan, Staten Island and Queens. The borough presidents are voted in by the residents of the respective borough. Their role is to advise the Mayor on executive issues, and analyse borough needs through the annual budget process. Additionally, the Borough President will appoint the community board who serve as advocates of the citizens of the borough. This board will evaluate the needs of the local communities, and make recommendations to the President from their findings [47].

The understanding gained from researching the governmental structure of New York City identified the agencies that are responsible for current policies that manage resource equity. From this review, it was decided that the project would focus on issues that were under the management of the presiding Mayor of the city. This was chosen as the position of Mayor was elected based on proposed policies, and analysing the results of these policies would provide an indication of how successful they had been implemented.

2.3 City Resources

As explained in Chapter 1, the project aimed to analyse whether there is inequity in the resource allocation of New York City. To achieve this understanding, a number of factors needed to be chosen to base later analysis on. These factors were chosen by evaluating the departments managed by the Office of the Mayor as discussed above, and referring to literature that highlighted factors that improved the contentment of societies, such as the United Nations' 'Handbook on Social Indicators' [38]. From this material, it was decided that the project would focus on the following issues; educational standard, environmental influences, healthcare provisions, transportation quality and accessibility to public services. These factors would be used in the measure of equity to ensure that residents from each borough a fair access to resources. The following section will explore each one in turn.

2.3.1 Healthcare

There is no public healthcare system in the United States, requiring individuals to pay insurance companies to ensure they can pay for medical costs. However, there are government schemes in place to make certain that underprivileged groups can still access health insurance so they can be treated by a medical professional if necessary. Medicaid [29] is a social healthcare programme aimed at low income individuals and families, providing free healthcare to those whose income is below a certain threshold. Medicare [30] is an alternative government backed social insurance programme, that covers citizens over the age of 65 that have paid into the scheme through payroll whilst they were in employment. These schemes aim make healthcare accessible for all citizens, regardless of their social situation. However, even though these schemes are in place, there is evidence to show that *'lack of timely and effective ambulatory care may have a significant impact on hospitalisation rates in the low income areas'* [2]. This suggests that lower income neighbourhoods may be disadvantaged when in need of hospital care.

2.3.2 Transport

New York City has a broad transport network. Its vastly connected subway system is one of the largest subway systems in the world, built between 1913 and 1931 [68]. The subway connects every borough apart from Staten Island, and runs 24 hours a day, seven days a week [68]. The subway therefore runs overnight, however the arrival times of trains are less frequent. The convenience of the subway system enables residents to commute into central Manhattan for work, or tourists to explore different areas of the city.

An iconic image of New York City is that of the yellow medallion taxis. Currently, there are over 13,000 licensed taxicabs that transport citizens and tourists around the city. The demographics of taxi driver ethnicities mirror the diversity of the city, with 82 percent of drivers being foreign born. These taxis are managed by the Taxicab and Limousine Cooperation, a public agency of the New York City government. After being unrivalled for many years, these drivers caused unrest when they were rivalled by the private sector transportation company Uber [82]. Uber undercut the demand for yellow taxis by offering cheaper fares to customers. Yellow taxi drivers submitted official complaints stating that Uber drivers faced fewer regulatory burdens in comparison to them, resulting in medallion cab pickups falling by 3.83 million between April to June 2015 [33, 67, 3].

2.3.3 Education

The Department of Education is responsible for running all of the public schools across New York City. Their reach spans over 1.1 million students in over 1,800 schools, making the Department of Education one of the largest school districts in the United States [12]. It has been argued that education is a fundamental factor in the socioeconomic development of a population in a paper by Burchi [7]. The paper draws the conclusion that when schools in third world countries are better attended, the quantity of people affected by food insecurity decreases, which can be used as an indication of poverty levels. Further analysis can be undertaken to identify whether this theory is applicable in such a developed city as New York, and to what extent the age someone leaves education could be a factor of their food insecurity and therefore household income for the future.

Table 2.1: A Comparison of Average Household Income Across the Boroughs of New York City

Borough	**Population [17]**	**Average Household Income (Per Annum) [17]**
Bronx	1,446,350	$50,306
Brooklyn	2,528,061	$64,217
Manhattan	1,563,897	$132,754
Queens	2,282,534	$78,438
Staten Island	478,652	$88,637

2.3.4 Environment

New York City has a diverse environment. From corporate skyscrapers that house global conglomerates to the vast array of parks [39], there are many challenges in sustaining the environmental equilibrium in the city. To ensure this stability, the Department of Environmental Protection aims to *'protect public heath and the environment by supplying clean drinking water, collecting and treating wastewater, and reducing air, noise, and hazardous material solution'* [52]. Other agencies contribute to the environmental management go the city too. The 311 agency was set up to provide residents a facility to make non-emergency complaints. This is primary managed through a call system, however there is now a mobile and online version too. 311 calls can cover an array of complaints from broken street lights to noise complaints. The 311 agency are able to use these calls to assign the correct department to make resolutions where possible. This system aims to improve the environmental impact of citizens by managing any issues quickly and effectively [46].

It has been previously discussed how the economic environment of New York City differs vastly across boroughs. The data show in Table 2.1 draws some interesting comparisons. The average yearly household income in Manhattan is more than any other borough with significant difference between the difference between that and the Bronx, which has the lowest. Interestingly, both boroughs have a similar number of citizens. Differences like this could provide curtail insight into possible disparity of public services across the city, and can be analysed further in the analytical phase.

2.3.5 Public Services

New York City offers many public services to citizens, in order to assist and improve their lives. These services range from fire departments to libraries to social housing [53, 54, 51]. One such service is the Fire Department of New York City, which is split into divisions that serve each of the five boroughs. As the work of the fire department is so integral the the emergency response system of the city, their responsibilities are divided into several 'core competencies'. Some of these competencies include fire suppression, pre-hospital emergency care and fire prevention inspections.

Due to the terror attacks of September 2001, the Citywide Incident Management System [48] was established to provide better care for citizens in emergency situations, such as natural disasters or the threat of terror groups. This organisation aimed to combine the efforts of the New

York Fire Department and the New York Police Department, allowing a more effective response to emergency situations. The structure of the inter-agency operations has been criticised [11], claiming that coordinated response does not occur as often as such be expected, with protocols often not adhered to. This research illuminates where further investigation can take place in order to identify what steps could be made to better improve the cross collaboration between the public services in New York City.

Exploring the main factors that were to be analysed by the system provided greater contextual understanding that would benefit the understanding of the results. Once these factored had been studied in depth, research was then undertaken previous academic work in the science of cities.

2.4 City Analytics

A number of institutions have been appointed to develop the understanding of science in cities. These institutions are collectives of interdisciplinary work that span a range of analytical problems faced in civic planning and sustainability. The following section will review key papers that utilised city analytics to provide meaningful results.

2.4.1 Visual Exploration of New York City Taxi Trips

The current trend towards big data analytics has resulted in interesting results. The paper by Ferreira et al created a model to allow the exploration of publicly available taxi trip data. This propriatorey software was utilised to investigate taxi demand patters and identify city dynamics. For example, the paper discusses the impact of taxi activity in the week of Hurricane Sandy. Analysis from heat maps show a disproportionate reduction in demand for taxis especially around downtown Manhattan that only reemerged days after the hurricane had passed [13]. This exposure to citizens decisions to stay indoors on a comprehensive scale provides a pattern of behaviour in reaction to an emergency situation. Insight like this can be utilised to improve the preparedness for future disaster situations.

2.4.2 Shadow Generation In Manhattan

An interactive article published in the New York Times maps the shadows of Manhattan buildings. These generated visualisations show the dappled light effect that lines the streets of the city, due the frequency of high rise buildings. An example of this is shown in Figure 2.3, which shows the shadow cast by the One World Trade Center. Although seemingly insignificant at first, this insight into the light effect of New York City's public spaces can be used to infer knowledge about where people will congregate to enjoy the sunshine. *'In most parts of America, sunlight is not debated the way it is in New York, where the city's thirst for living space, working space and economic growth has turned the sun into a virtual commodity'.* This analysis shows that identifying patterns in light can predict movement patterns, as well as the desirability of an area which in turn can increase property prices. This environmental effect of light has been observed by the environmental control agencies, and have in turn placed development regulations to ensure that light is preserved for the enjoyment of the citizens. This distinction

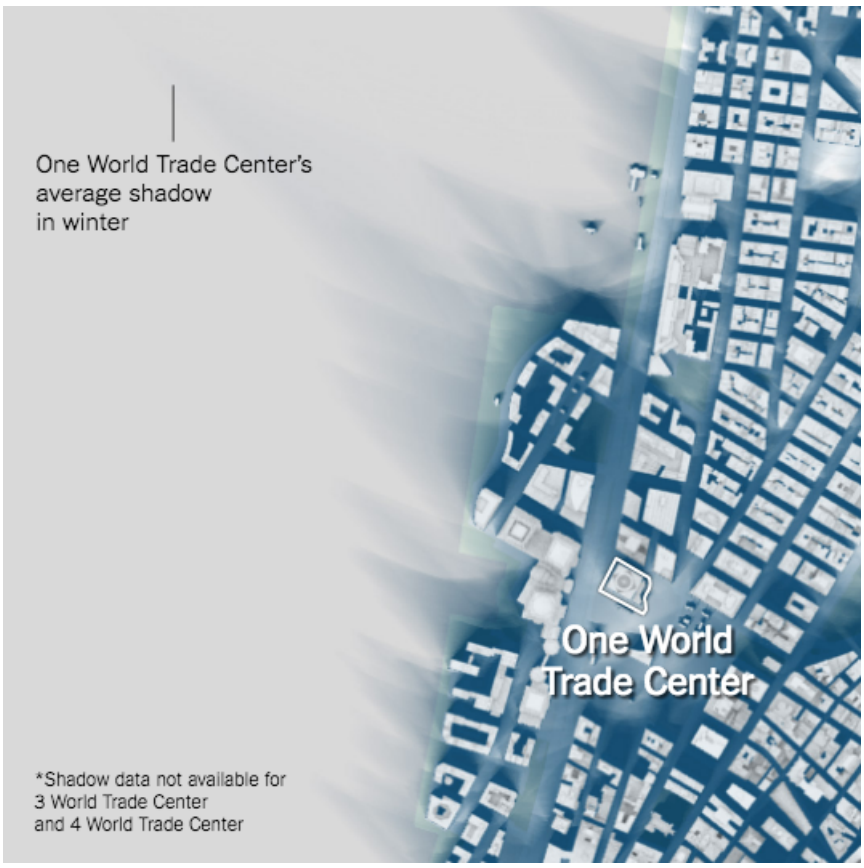


Figure 2.3: The Shadow Generated by the One World Trade Center [6]

illustrates how light and air are valued by the people and law makers of the city [6].

Reviewing related research that utilised city analytics showed the importance of identifying emerging patterns in big data. The results from the papers discussed showed how evidence from data could be applied to current civic issues to gleam novel and compelling insights. As the research component had now reviewed a variety of thoughtful evidence and literature, progress was moved on to evaluate possible technical platforms that could be used to build the analytical system.

2.5 Platform Research

A key component of the project was the use of modern data science technologies. At the time of the project, there were multiple technical platforms that could have been utilised to develop analytics and visualisations. To decide which platform to apply, time was spent investigating the usefulness of each one. The ideal system would be expected to fulfil the following requirements:

1. **Low-cost or Free:** Due to the nature of the project there was not a large budget to spend on commercial technology.

2. **Easy to Use:** The platform should be intuitive for a user with a background in Computer Science. It should provide out of the box solutions for simple data plots such as bar charts, scatter plots, etc.

3. **High Quality Graphics:** The platform should produce high quality professional standard graphics.

4. **Ability to Produce Geographical Plots:** The platform should be able to plot data onto a map using location data.

5. **Useful Documentation:** The platform should be supported by extensive documentation to allow the developer to learn how to use the system.

The platforms that were explored will be discussed below, with considerations to how they could be utilised in the project. The platforms assessed were divided into two categories; online Geographic Information Systems and technical graphing libraries. All the data utilised in the project was freely available and provided by the New York City Government. The online web portal NYC Open Data provided agency specific datasets on a variety of civic topics [57].

2.5.1 Online Geographical Information Systems

There was a variety of companies offering online Geographical Information Systems (GIS) to plot and visualise data. These products offered simplicity and required low technical ability to produce high quality data visualisations. To evaluate each GIS platform, a dataset containing the locations of subway stops was sourced from NYC Open Data and used as input to each system, allowing a direct comparison to be made [58].

ArcGIS
The ArcGIS platform offered an online tool that allowed users to upload data in CSV format and create visualisations on preloaded maps. The software provided an extensive amount of analytical mathematical functionalities, however computing this analysis took a long time to load. The results of this are shown in Figure 2.4, where the first image shows the subway stations and the second shows the density of these points. This exploration took the online system a few minutes to run and the result was not intuitive. When loading data into ArcGIS, settings had to be configured to enable it to plot the information on a map. This required filling out a large form which threw errors several times. Overall the visualisations produced were to a good standard, however the usability of the system was lacking [25].

Carto
Carto was similar to ArcGIS as it offered an online platform for users to explore geographic data. The user interface of the platform made using the system very simple. It was straightforward to upload the subway entrances dataset, and instantly the system identified which attribute contained the latitude and longitude coordinates. Carto was able to plot the data automatically unlike ArcGIS, and gave the resulting visualisation shown in Image (a) in Figure 2.5. Configuration in Carto was easy to achieve and resulted in visually attractive maps. Image (b) shows the map after configuring it to show categories, colouring each station based the line it serves. Additionally, Image (c) shows a heat map of the dataset, which allows the user to

interpret that there is a higher density of subway stations in Manhattan compared to Brooklyn or Staten Island [8].

2.5.2 Graphing Libraries

It was necessary to consider the use of graphic libraries in the project, as they offered more technical control over the resulting data visualisations. Research into possible libraries produced many options, so a variety of the most popular were analysed.

D3
D3 is a free Javascript library which creates interactive data visualisations for web applications. This technology is utilised in many industries where displaying data to users is of chief importance. An example of this is in journalism, where the Huffington Post created an interactive voting map that displayed results of a senate election in Massachusetts [41]. A still image of this visualisation is shown in Figure 2.6. D3 creates professional looking visualisations, however it comes with certain drawbacks. If used in the project, graphics would have to be rendered on client side. With datasets that contain over 1 million points this would cause the computation to run extremely slowly, if at all. For this reason, it seemed that D3 would not be appropriate for the needs of the project [37, 9].

Plotly
Plotly is a plotting library that can be used in a variety of languages, including Javascript and Python. It contains simple out of the box charts such as pie charts, line charts and chloropeth maps. It can be used in conjunction with other technologies such as MapBox to create geographical plots. Plotly's plots are interactive and can be embedded easily into web applications. Additionally, Plotly has detailed documentation and an a strong level of technical support online. Once a visualisation is created in Plotly, it is stored in the cloud which can be managed through a user account. This was advantageous in this project, as many iterations of data exploration would take place. Plotly has a free version with a limit to the number of calls per day, but provide a low-cost unlimited version. An example of geographical plots using Plolty is shown in Figure 2.7 [35, 36, 64].

Matplotlib
Matplotlib is a Python plotting library that provides the functionality to plot a variety of graphs. It is a powerful graphing language that allows the user to configure a multitude of attributes, however the high level of control adds complexity when trying to create simple and effective plots. As Matplotlib is an older graphing language compared to the others reviewed, it has a more traditional design that is less visually appealing. Within the library, geographic plotting functionality utilises the Basemap package [43], which is not as initiative to use as the other technologies evaluated. However, some examples were created in Matplotlib to assess its suitability in the project. The graph in Figure 2.8 shows a scatter graph marking the latitude and longitude locations of taxi pickup points across New York City. Due to the density of the data, the resulting graph is easily identified as Manhattan without the use of a base map. By configuring the background and point colour it was displayed in a contrasting black and white pattern, which improved the visual appeal from the standard Matplotlib appearance. It was decided that Matplotlib would be suitable for dense datasets, however due to the difficulty of

Platform Scoring System: 1 = Not Suitable, 2 = Unlikely to be Suitable, 3 = Neutral, 4 = Somewhat Suitable, 5 = Highly Suitable

Table 2.2: A Table Showing the Comparison of Different Platforms

Platform	Cost	Ease of Use	Quality of Graphics	Documentation	Geographic Plots	Total
ArcGIS	3	2	2	3	4	**14**
Carto	3	5	5	4	5	**22**
D3	5	2	5	4	5	**21**
Plotly	4	5	5	5	5	**24**
Matplotlib	5	2	2	3	2	**14**

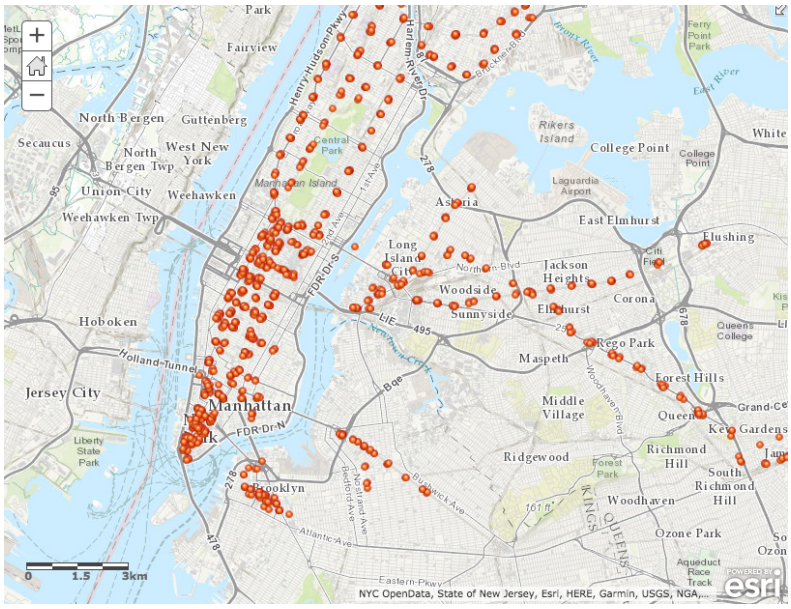
configuration a different technology would be more effective [42].

2.5.3 Evaluation of Platforms

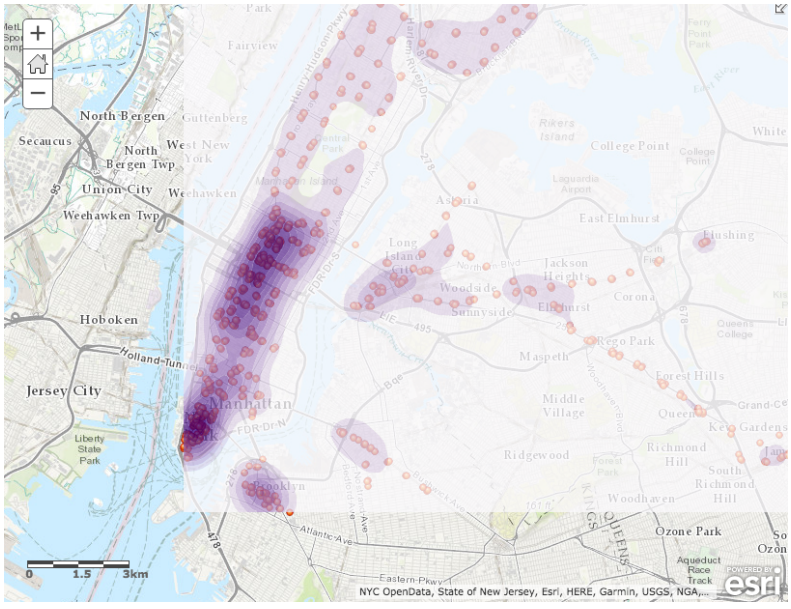
After an assortment of platforms were trialled, comparisons were drawn to decide which technology would be utilised throughout the project. As documented previously, the ideal platform would offer simple out of the box graphs, yet provide the ability to refine and configure specific attributes. Most importantly, the solution should be able to support data with a high volume of attributes. To assess the suitability of each platform, a numeric scoring system between 1 and 5 was devised, where 1 indicated **Not Suitable** and 5 indicated **Highly Suitable**. This allowed a direct comparison to be made between each platform across the range of requirements specified at the beginning of the section.

Table 2.2 shows the evaluated scores of each platform. Ploty achieved the highest rating of 22, due to its ease of use and high standard of visualisations. As Plotly could be accessed in a range of languages, the decision was made to use it in collaboration with the Python programming language. Python was chosen to be the main language utilised in the project due to its mathematical libraries and simple syntax. It was decided Anaconda would be utilised as the package manager for the project, as it supplied additionally functionality in providing Jupyter Notebook to write and run scripts. The use of these technical components constructed the backbone of all data analytics and visualisations created during the project.

The thorough research conducted in this chapter allowed vital understanding to be gained on the problem scope before a system was designed to assess the accessibility of resources in New York City. From this process, a contextual understanding of the city was gained by studying the geography, demographics and government of the city. Additionally, a number of important factors that affected the lives of the majority of citizens were identified; education quality, environmental factors, healthcare provision, transport and quality of public services. Evidence showed that these factors were aspects that touched the every day lives of residents across all five boroughs. It was decided that the proposed system would aim to analyse each of these factors, resulting in a broad insight to resource allocation. To ensure the system was built to a professional standard, it was necessary to also consider the legal, social, ethical and professional ramifications of the project.

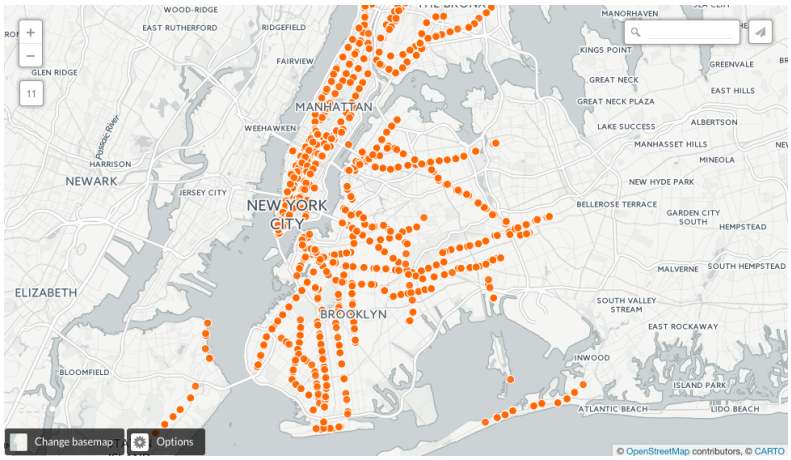


(a) Visualisation Created Using the ArcGIS Data Library



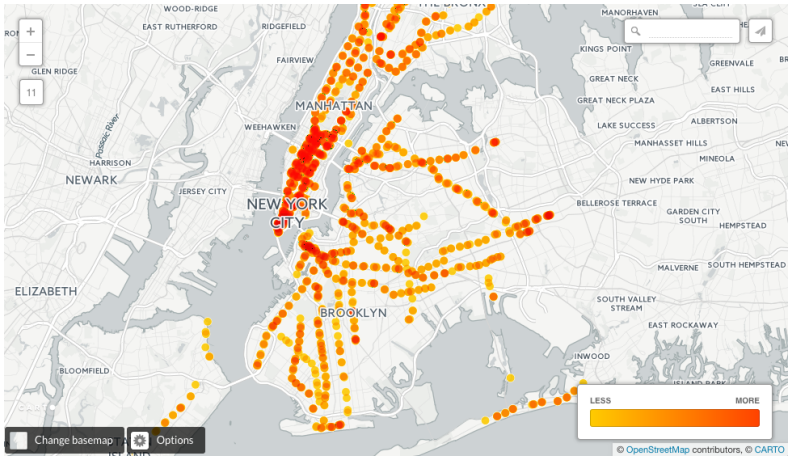
(b) Configuration to Show Density

Figure 2.4: Analysing ArcGIS by Exploring the Subway Entrances Dataset



(a) Automatic Visualisation Created by Carto

(b) Configuration to Colour Nodes By Subway Line



(c) Configuration to Display Heatmap Intensities

Figure 2.5: Analysing Carto by Exploring the Subway Entrances Dataset

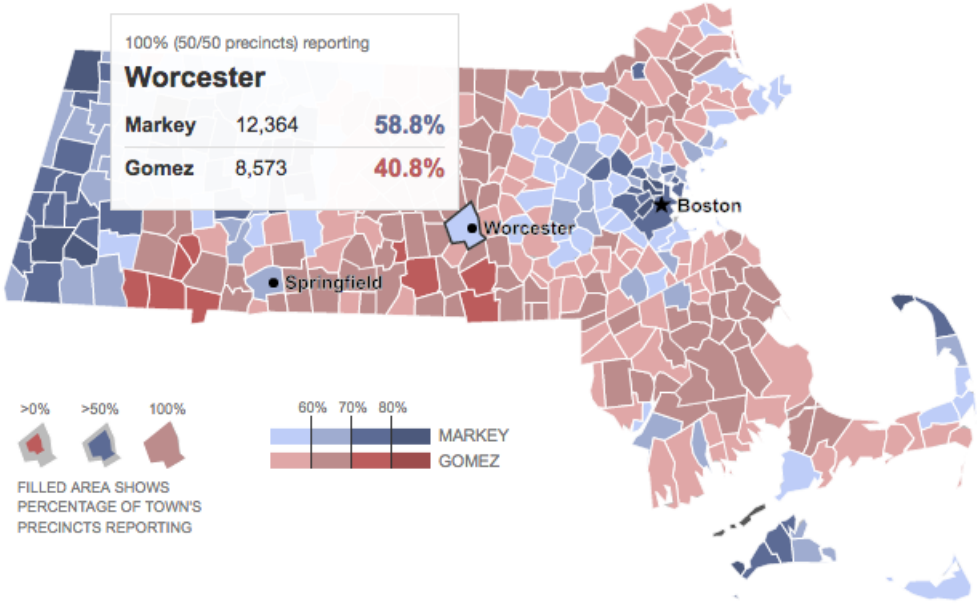


Figure 2.6: D3 Technologies Utilised by Huffington Post to Show Voting Trends

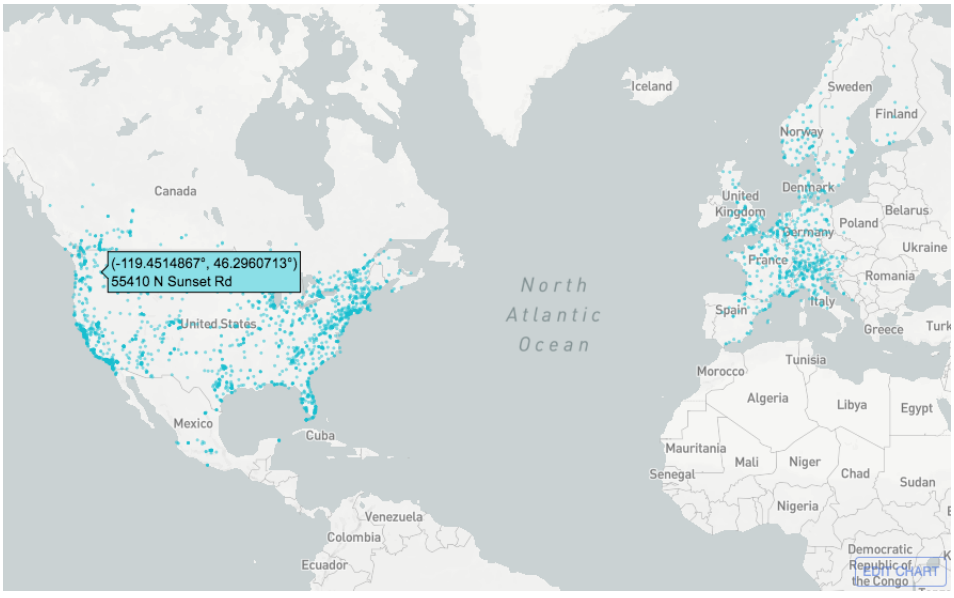


Figure 2.7: A Plotly Map Showing Tesla Gas Stations Around the World

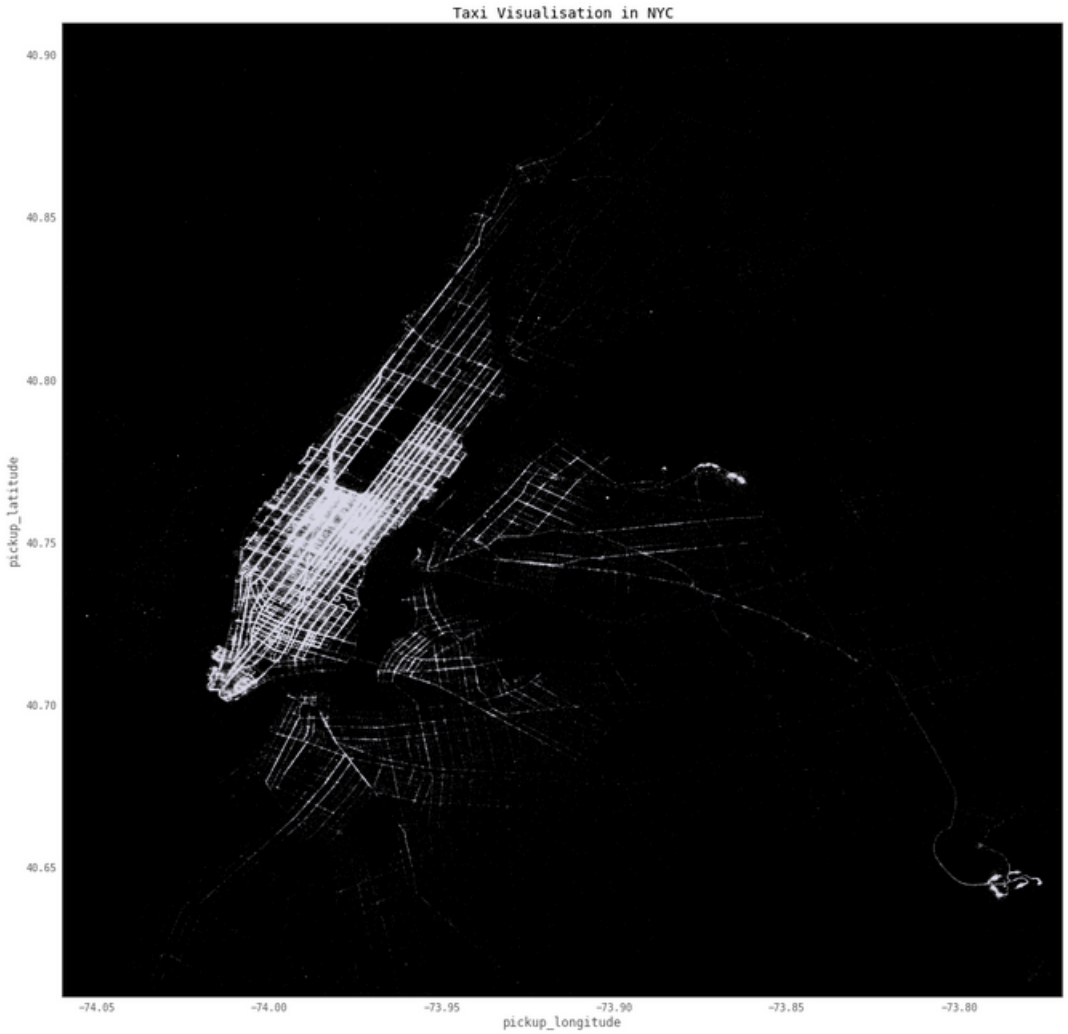


Figure 2.8: A Matplotlib Scatter Chart of Taxi Pickup Locations Across Manhattan

CHAPTER 3

Legal, Ethical, Social and Professional Issues

To ensure the project was carried out to the highest standard of professionalism, consideration was taken place to address the legal, ethical, social and professional issues surrounding the subject. These issues were important to discuss as the analysis of socioeconomic data had the potential to explore sensitive issues. To assist in understanding good practise guidelines, the British Computing Society's Code of Practise was referenced and discussed to provide a guideline of professional practice [26]. The following chapter will discuss each issue in turn.

3.1 Legal Issues

As previously discussed, all data used in this project was obtained freely from the NYC Open Data portal. This ensured that the data had already been pre-processed to make certain that all information was of a public nature. By using this data, the project abided by the Terms of Use set out by the New York City government [56]. The data would only be used to facilitate analysis into a variety of factors that pose problems to the citizens of New York City. In agreeing to these terms of use, the project ensured that the data was not used for any illegal purpose, for example to commit a crime or engage in any conduct that would result in civil liability. The project also serviced the use of a number of management and development tools. In keeping with professional standards, these tools were utilised within the bounds of the Terms of Service.

3.2 Ethical Issues

The project touched on many social and ethical issues, however they were be treated with upmost professionalism and transparency. As the data had already been published, it was be expected that due diligence had taken place and therefore results were anonymous. No attempt was made to identify individuals or groups of people. The exploration into resource allocation focused on generalised issues that effected the majority of citizens, in order to provide results that would benefit the wider community. It was important that the project resulted information

that was not politically biased, and worked on the assumption that all people across the five boroughs should have equal and fair access to services provided by the government.

3.3 Social Issues

To ensure the project represented a broad view of the social environment of New York City, a set of factors were decided on to further analysis. These factors included education quality, environmental standards, healthcare, transport accessibility and the quality of public services. Research into each of these areas indicated that they were issues that affected citizens regardless of age, race or gender. This allowed the project to focusing on identifying areas that would benefit the largest number of people, and would not discriminate against any groups. The datasets chosen for analysis were fully representative of these social issues, as categorised by the United Nation's publication of global issues [70]. The project aimed to explore these issues in a scientific and unbiased way to produce results that were representative in the context of New York City.

3.4 Professional Issues

Research was a crucial part of the project as literature was evaluated by a range of sources, with all work referenced to allow further reading. During the analytic development process, the details of all data sources were made available and findings were reported in the scope of the project to avoid data bias. As data was collected from the NYC Open Data portal, due diligence was taken to ensure that datasets were not used unless they contained reliable, transparent and useful data. Analytical software produced throughout the project maintained professional standards by being well structured, modular and test driven, to ensure that it maintained the standard set out by the BSC guidelines. The project was only feasible due to the wealth of online data provided by the New York City government, and in keeping the open source philosophy all visualisations were made publicly available online via the Plotly cloud platform.

Thorough discussion of the legal, ethical, social and professional issues highlighted areas that needed consideration in the conception of the system. Specially, ensuring that the use of public data was used appropriately and without bias was imperative in a project that assesses equity. This awareness allowed a comprehensive system to be designed to derive fair and meaningful results.

CHAPTER 4

Project Requirements

To achieve the overarching aims of the project, a set of requirements were outlined which described the functionality of the proposed system. Due to hybridity of the project between research and development, the term system was adopted to describe the process of utilising datasets as an input to series of analytical scripts that identified patterns and correlations, resulting in insightful recommendations. This holistic nature of the proposed system necessitated a clear scope of expected functionality to be defined. This chapter will provide that scope, by highlighting the requirements for the aforementioned functionality. Additional optional functional requirements were noted in the case that work was completed earlier than expected. Additionally, non-functional requirements of the system were also defined to ensure the system was developed to a high standard.

4.1 Functional Requirements

The functional requirements documented what the capabilities of the system should be, describing its behaviour in different environments. In this project, the system was the analytical component that generated data-driven results by identifying patterns in large datasets. The system was designed to take data as input and compute analysis as an output. The resulting analysis would provide evidence to evaluate whether resources are equitably distributed in New York City. Due to this interdisciplinary approach between research and development, it was important highlight clear requirements at the start of the project.

F1: *The system should accept CSV datasets as input.*

F2: *The system should create a model based on the inputted datasets.*

F3: *The system should produce graphical visualisations of data using mapping software.*

F4: *The system should produce results that provide recommendations.*

F5: *The system should be comprehensible to a data scientist.*

These functional requirements outline the basic functionality that was required of the system. If the developed system met these prerequisites, it was expected that it would be able to derive meaningful insight about resource equity based on locality when fed a dataset of multiple attributes.

4.1.1 Optional Functional Requirements

If the case arose where all functional requirements were met before the project deadline, it was necessary to construct additional requirements that would serve as stretch tasks. These requirements would build on the existing system and aim to improve the analysis of data.

OF1: *The system should combine data from a range of agencies.*

OF2: *The system should utilise social media data.*

OF3: *The system should make predictions of how policy change would effect conditions.*

4.2 Non-functional Requirements

The system also had a set of non-functional requirements which described other integral needs that should be met that were not directly related to the functionality. These non-functional requirements were still equally as important and expected to be achieved in addition to the functional requirements.

NF1: *The system should follow the licensing agreements of open source data.*

NF2: *The system should be maintainable.*

NF3: *The system should be testable.*

NF4: *The system should be extendable to alternative data sources.*

Outlining these functional and non-functional requirements at the start of the project was an important part of the system design. Creating these small subgoals allowed development to be manageable and attainable. The evaluation phase of the project utilised these requirements to assess whether the system had met these goals to a sufficient standard, discussed in more detail in Chapter 12. To ensure the proposed system came to fruition, it was important to define a structured methodology that guided the workflow of the project.

CHAPTER 5

Methodology

Having a clear framework was important in this project, due to the breath of factors considered for analysis. To ensure this structure was maintained, the following methodology provided phases that outlined the workflow, as illustrated in Figure 5.1. A benefit of defining this structure was that it allowed time to be managed effectively and ensured that all project aims described in Chapter 1 were completed to a high standard. The following section will describe each phase of the methodology and highlight the expected outcome. Referring to the methodology throughout the project ensured accountability and provided focus to obtain results.

5.1 Research

The project was focused on how data science can be utilised to improve the everyday life of citizens in New York City. The research was broken down into sections to allow a broad coverage of interdisciplinary topics. The broadest research area was regarding the behaviour of people, which was undertaken by researching into demographics and government. Analysing the people of New York City on a wider scale gave light to important factors such as educational standards, environmental influences, healthcare accessibility, transport quality and access to public services. It was decided that these factors would be used by the proposed system to assess the accessibility of public services.

5.2 Data Ingest

The research phase identified the single factors that would become the focus of analysis. The aim of data collection was to identify data that provided a diverse representation of each factor, in the hope that this would lead to a rich analysis of resource equity. All data was sourced from NYC Open Data, spanning different timeframes and locations across the city.

Once data had been gathered, it was necessary to clean and standardise it. As the importance of location specific data was previously identified, it was essential that all data collected had

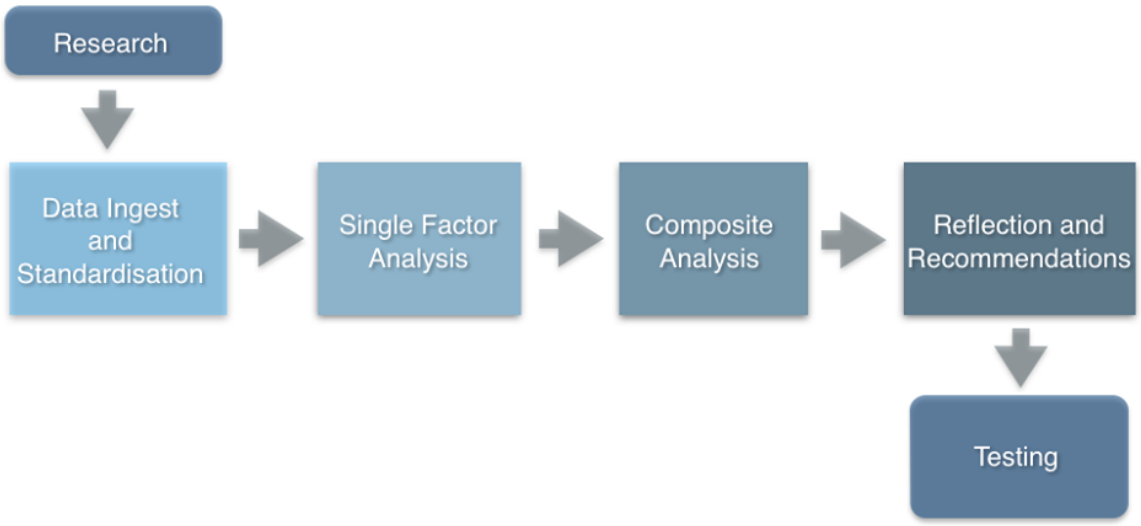


Figure 5.1: A Diagram of the Project Methodology

a geographical reference of latitude and longitude coordinates. Other data cleaning techniques regarding missing values or anomalies were implemented and are discussed in more detail in Chapter 6. The Data Ingest phase resulted in a range of standardised datasets representing each factor, that were utilised as input to the proposed system.

5.3 Single Factor Analysis

With standardised datasets that represented the factors previously identified, analysis could begin on each issue in turn. During this phase each individual factor was analysed independently using data visualisation techniques. Utilising locality attributes allowed data to be plotted on a map, identifying patterns or clusters. Once each factor was fully explored, questions were raised regarding the equity of resources in the given factor. The factors that produced the most interesting questions were chosen to continue to the Composite Analysis phase, where statistical techiques aimed to provide answers these queries.

5.4 Composite Analysis

The predecessing Single Factor Analysis phase shed light on possible areas of inequity, which required further focus during the Composite Analysis phase. Composite analysis was the process of fusing datasets from different factors to analyse whether more insight could be gained. In doing so, possible correlations were considered to understand the dynamics of resource provision in the city.

During this phase, data was standardised by population density to account for areas that had a greater number of residents. Normalising the data was important to show whether trends identified in the Single Factor Analysis phase still persisted once equalised across boroughs. If

the results of this analysis still indicated inequity in resources, it was clear that recommendations could be made to improve the balance for all citizens.

5.5 Reflection and Recommendations

Once the recommendations had been generated, a discussion was undertaken to measure the practicality of the results. This further study allowed a conclusion to be drawn on the viability of each result produced by the system. The efficacy of the results was measured by identifying literature to support or contradict the outcomes reached by the system. Limitations of the results were also discussed, to ensure a fair and unbiased appraisal. The results that were deemed appropriate in the discussed context of the city were then provided as recommendations. The recommendations implicitly gave an answer to the question set out in the initial project aims by providing a suggestion of how better to structure public resources allocation to improve the accessibility for all citizens. Answering this question contributed to a wider discussion of equity across the city.

5.6 Testing

It was important to exhaustively test the system to ensure that results collected were accurate and representative. The strategy was broken down into unit, integration and system testing that evaluated the system to ensure proficient software development practice were employed. This ensured that the results and therefore recommendations produced could be trusted and were substantiated in evidence.

The methodology outlined in this chapter highlighted the highly streamlined workflow of the project. Subsequent chapters will follow this structure, describing the implementation and results of each phase.

CHAPTER 6

Data Ingest

The data that was used as an input to the proposed system could be sourced once research had identified what factors would be used in analysing resource allocation in New York City. As discussed in Chapter 2, these factors were Education, Environment, Healthcare, Transportation and Public Services. To explore these factors, a multitude of datasets were collected to represent their impact on citizens. These datasets spanned different timeframes and were derived from different agencies to provide a rich and holistic understanding of the socioeconomic landscape. Time was devoted to ensure that all data was standardised by creating methods that dealt with inconsistencies. The following chapter describes these procedures in more detail.

6.1 Data Sourcing

As aforementioned, all the data used in throughout the project came from the data repository, NYC Open Data. NYC Open Data is an online web portal that was created to provide the people of New York City more accessibility to their civic public data collected by their government. It was instantiated as part of the 'Open Data Law' passed in 2012, which aimed to improve transparency and engagement between officials and citizens of the city. A statement from the Commissioner of NYC Information Technology and Communications highlighted the importance of this law, by stating that *'true engagement, equitable engagement, isn't only about releasing data, it's about expanding opportunities for collaboration to all corners of the five boroughs'* [78].

At the time of the project, NYC Open Data hosted a range of datasets that could be searched via category or agency. This functionality was useful to identify data that represented the factors required. Once a dataset was chosen, it generally came in a variety of formats, with the most common being a CSV file. It was decided that all datasets would be downloaded as CSV files as the format was easily read into Python using the Pandas library. A table depicting the data collected for each factor is displayed in Table 6.1.

Although all the files were in the same format, there was still a large range of disparities in data attributes collected for each factor. Some factors such as transport provided a wealth of

Table 6.1: A Table Showing the Datasets Collected For Each Factor

Factor	Dataset	Agency	Summary
Education	Graduation Outcomes	Department of Education	Graduation outcomes of cohorts from 2001 to 2006
Environment	311 Service Requests from 2010 to Present	311	All 311 calls from 2010
Healthcare	Health and Hospitals Corporation Facilities	Health and Hospitals Corporation	List of all HCC facilities in NYC
Healthcare	Mental Health Facilities Finder	Health and Hospitals Corporation	List of all mental health facilities in NYC
Transport	2015 Yellow Taxi Trip Data	Taxi & Limousine Commission	All trip records from January to June 2015 in Yellow Cabs
Transport	2015 Green Taxi Trip Data	Taxi & Limousine Commission	All trip records from January to June 2015 in Green Cabs
Public Services	FDNY Firehouse Listings	Fire Department of New York	Listing of all NYC fire houses with addresses
Public Services	Emergency Response Incidents	Office of Emergency Management	Type and address of emergency incident

datasets with a fine granularity of geographical attributes, such as precise latitude and longitude coordinates. Others, such as education generally just offered address or borough. Due to these differences in format, a method for standardising the data was required to ensure that all data collected could be plotted and analysed.

6.2 Standardisation

A number of decisions needed to be made on how the collected data was going to be standardised. There were a number of possible inaccuracies in the data, for example, missing values or outliers. Consideration needed to be given to how they would effect the analysis procedure, and what could be done to ensure they were providing a correct representation of the landscape of New York City. These considerations will be discussed individually.

Geographical Reference
The most important standardisation technique was converting all geographical references into the same type. It was decided that latitude and longitude coordinates served the purpose for this. Fortunately, the majority of datasets came with latitude and longitude attributes, however there were others that primarily used street address or zip code. To convert these attributes to latitude and longitude coordinates, the Google Geocoder API was utilised, highlighted in Listing 1. For every row in the dataset, a query containing the address was sent to the Geocoder server and a JSON file was returned. This file was then parsed to identify the latitude and lon-

gitude coordinates which could then be stored alongside the original data. This was a lengthly process as the API could only be queried once every two seconds, so a cache was employed to collect results whilst script was left to run in the background. This step resulted in all datasets containing latitude and longitude coordinates.

Listing 1 Utilising the Geocoder API to Return Latitude and Longitude Coordinates

```
# method to return a JSON file containing lat and long coordinates
def nycGeo(street_name, neighborhood, borough, state):

    #base provided by API
    base = 'https://maps.googleapis.com/maps/api/geocode/json?address='
    mid = street_name + ',+' + neighborhood + ',+' + borough + ',+'
            + state + '&key=' + api_key
    end = mid.replace(' ','+')

    # query to pass to API
    url = base + end
    r = requests.get(url, verify=False)

    # JSON file contains lat and long coordinates
    r_json = r.json()

    #r_json is later parsed to give lat and long variables used in model
```

Missing Values

As discussed, the most important attributes in the datasets were the geographical coordinates. If these were missing, the data was unusable as it could not be plotted on a map. It was decided that if there were missing values in other attributes, the sentiment of the data could still be understood, however not if it was in regards to geographical location. This required a simple script to determine whether the latitude or longitude attributes contained null or missing values. If they did, the row would be removed from the table. A sample of this script is shown in Listing 2.

Listing 2 Dropping Missing Values from Latitude or Longitude Attributes

```
# dropping any row that has a nan value for borough, lat or long attributes
df = df.dropna(subset = ['Borough', 'Lat', 'Long'])
```

Outliers

In this project, outliers were considered to be instances that were not located in New York City. When analysing the raw data, it was difficult to assess whether the latitude and longitude coordinates were within the bounds of the city. However, when data was plotted on a map it became apparent if there were any outliers as the plots would stretch to accommodated these unexpected results. Using this technique, this outliers were identified and removed from the datasets.

6.3 Workflow

The process of standardising each dataset individually lent itself to an agile workflow to ensured that data could be sourced, standardised and stored concurrently. After each dataset had been standardised, it was necessary to store it in the cloud due to the large size of files that couldn't be held locally. Maintaining an organised workflow for data collection was important due to the high volume of datasets needed for the project. Figure 6.1 shows how this workflow was structured. With datasets collected and standardised, it was possible to begin developing the system.

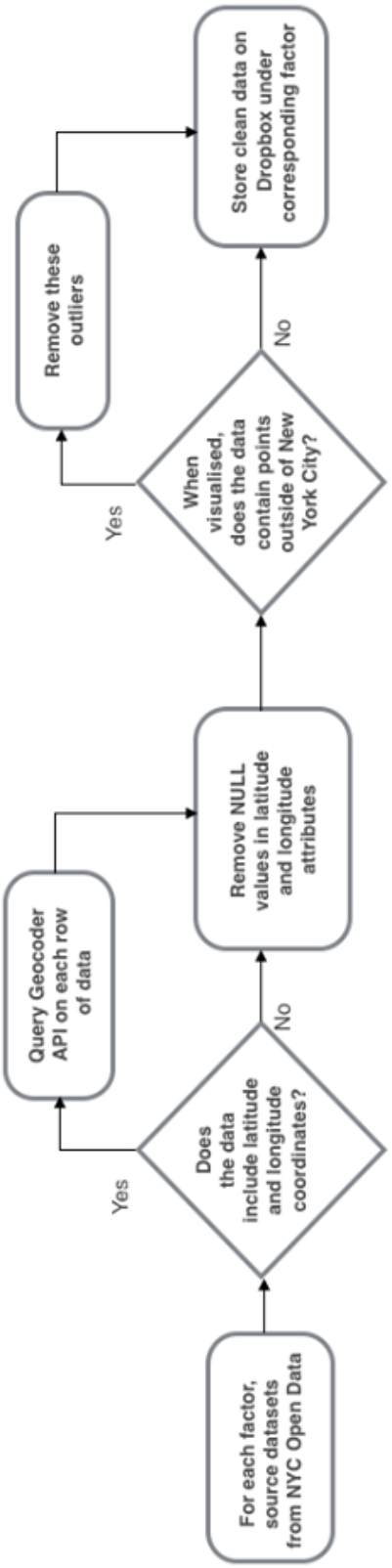


Figure 6.1: The Process Followed When Collecting Data

CHAPTER 7

Single Factor Analysis

Once the data had been gathered and standardised, the development of the proposed system could begin. As discussed the in Project Requirements chapter, the word *system* was used holistically to describe the process that derived results from an inputted dataset. Following the structure of the methodology, each factor was individually analysed by the system in turn. This ensured that each factor was analysed only using the data specifically collected to represent it. This segmentation of factors allowed the results of each factor to be independently observed, in an attempt to mitigate incorrect assumptions of correlations between factors.

The development of this part of the system required the first technical component to be produced. This was implemented by writing a script that initially read in a CSV dataset and stored it as a Pandas data frame. The data frame could then be manipulated to produce the resulting visualisations displayed in this chapter. This was achieved by utilising the Plotly graphing library to plot the data in a variety of graphs, such as scatter plots, bar charts and histograms. By evaluating the visual output in the form of graphs, questions were raised that would later be analysed in the following Composite Analysis chapter, by fused together datasets from a variety of factors to gain further insight into resource provision.

7.1 Education

As documented in Chapter 2, the Department of Education operates the largest school system in the United States [12]. The data collected to represent this factor was the Graduate Outcomes dataset for the graduating population between 2005 and 2010, provided by the Department of Education. This data was a rich source of information, and contained a variety of insight into the educational environment of New York City. For each borough, the dataset contained values for the total number of graduating students, in addition to the total number of students to drop out and not complete their course. This data was utilised by the system to calculate if students in a specific borough were more likely to drop out of education compared to students who lived in other boroughs.

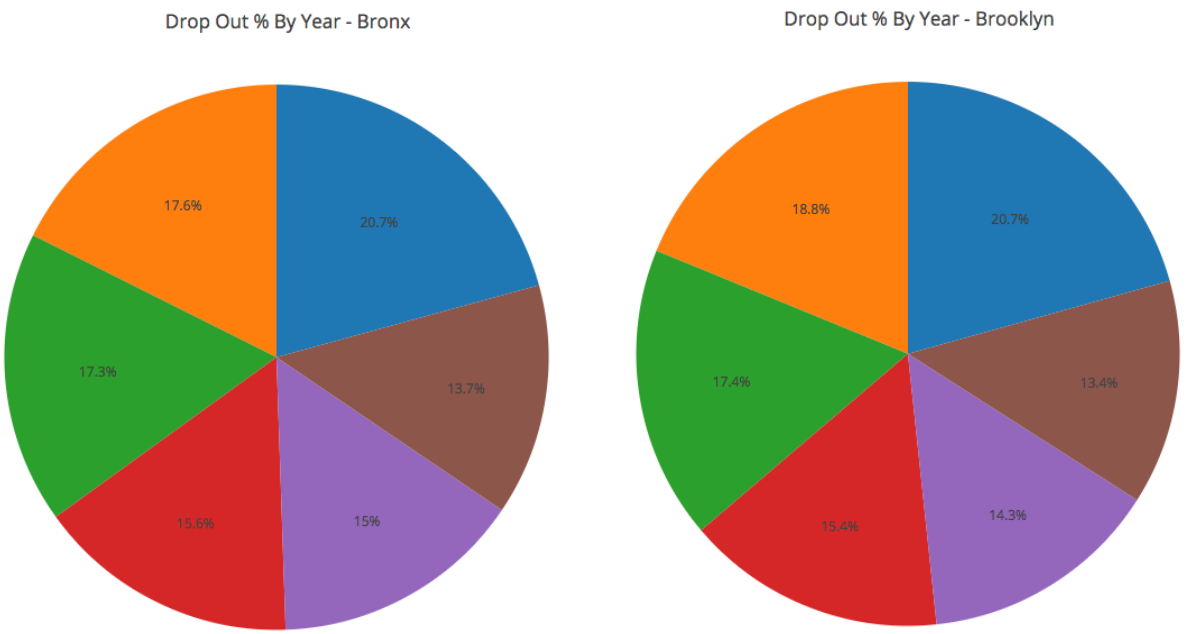


Figure 7.1: Drop Out Percentage by Year for the Bronx and Brooklyn

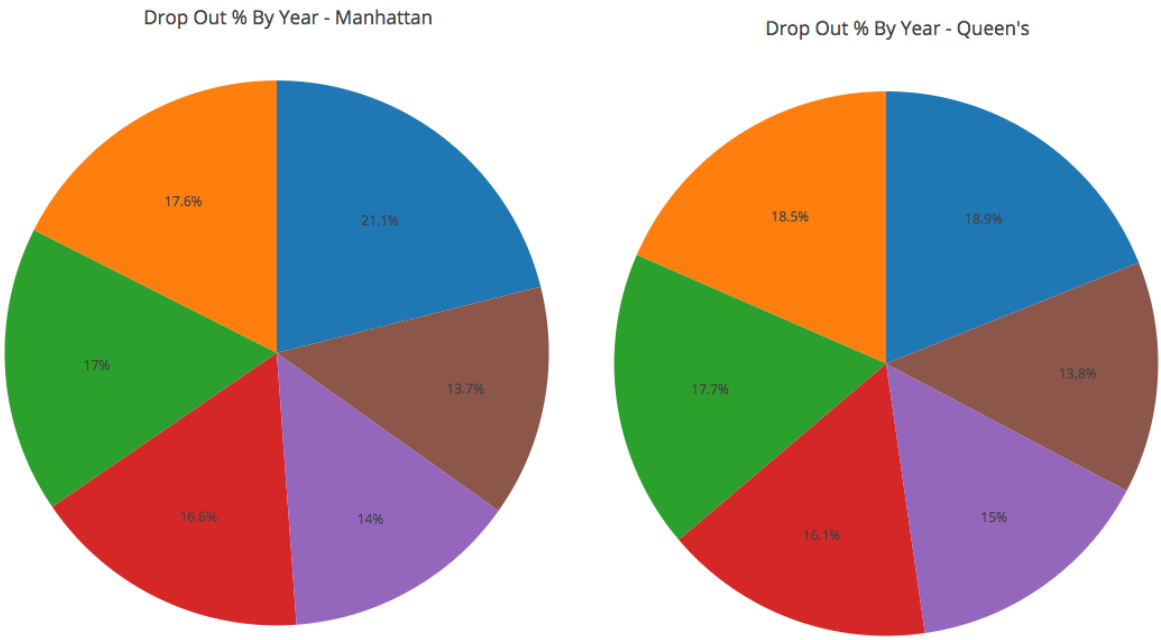


Figure 7.2: Drop Out Percentage by Year for Manhattan and Queens

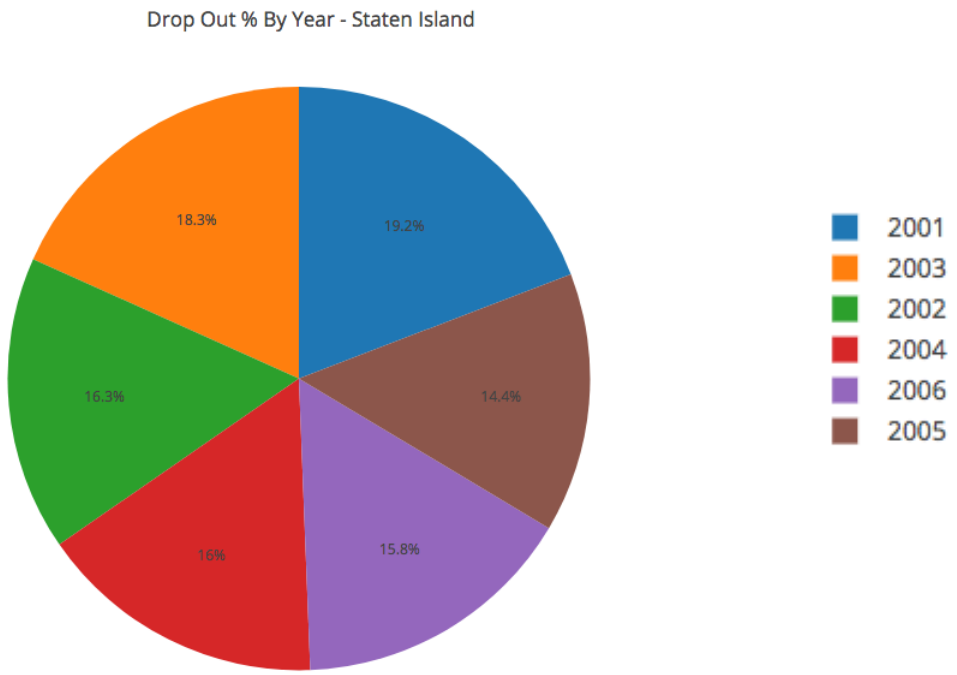


Figure 7.3: Drop Out Percentage by Year for Staten Island

The pie charts shown above represent the number of students who decided to terminate their education before the end of their course. This was known as the drop out rate, and given as a percentage of the number of students to drop out compared to the total student population in each borough for every cohort between 2005 and 2010. The key shown in Figure 7.3 corresponds to the year that the cohort enrolled on their course. The data across all five boroughs followed a similar trend of decreasing numbers in drop out rates between the first analysed cohort who graduated in 2005 and the last in 2010. The similarities in numbers across the charts suggested that the probability of a student dropping out of education reduced as time went on. In addition, this data suggested that successful strategies were employed by the Department of Education to retain students, between the five years of these observations.

When the same dataset was analysed by year instead of borough, a different picture emerged. The pie charts in Figures 7.4 and 7.5 displayed the results when the drop out rate was calculated as a percentage of students who failed to complete their education in the entirety of New York City for the cohorts who begun studying in 2003 and 2005. These two plots represented a trend that was sustained across the other cohorts analysed. It displayed that the Bronx consistently had the largest number of students who dropped out before completing their education. This conclusion suggested that a student who was educated in the Bronx was more likely to drop out compared to their peers who were educated in other boroughs of the city. To understand why this pattern emerged, further analysis was needed to combine a multitude of factors to gleam further understanding on this trend.

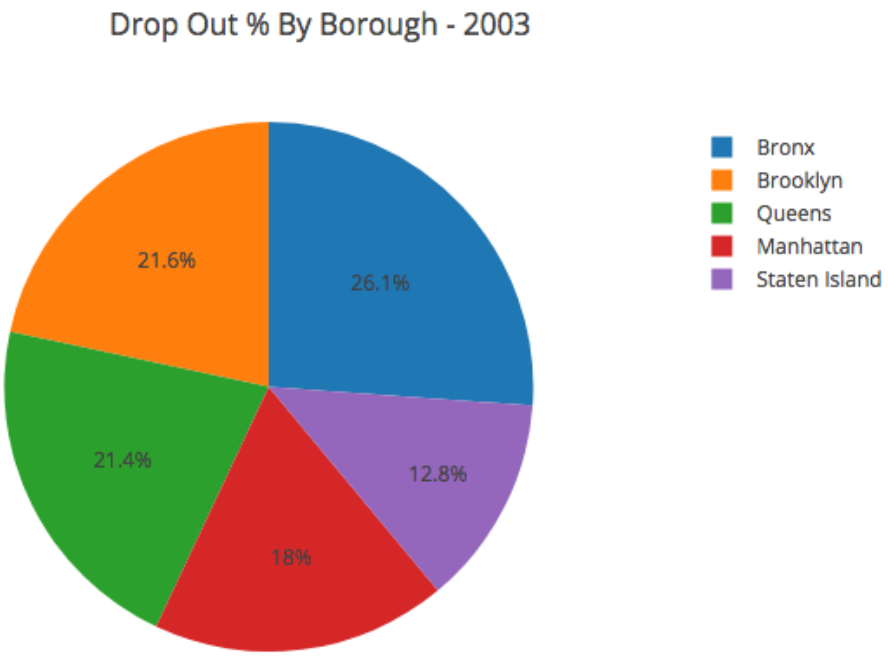


Figure 7.4: A Pie Chart Showing the Drop Out Percentage by Borough in 2003

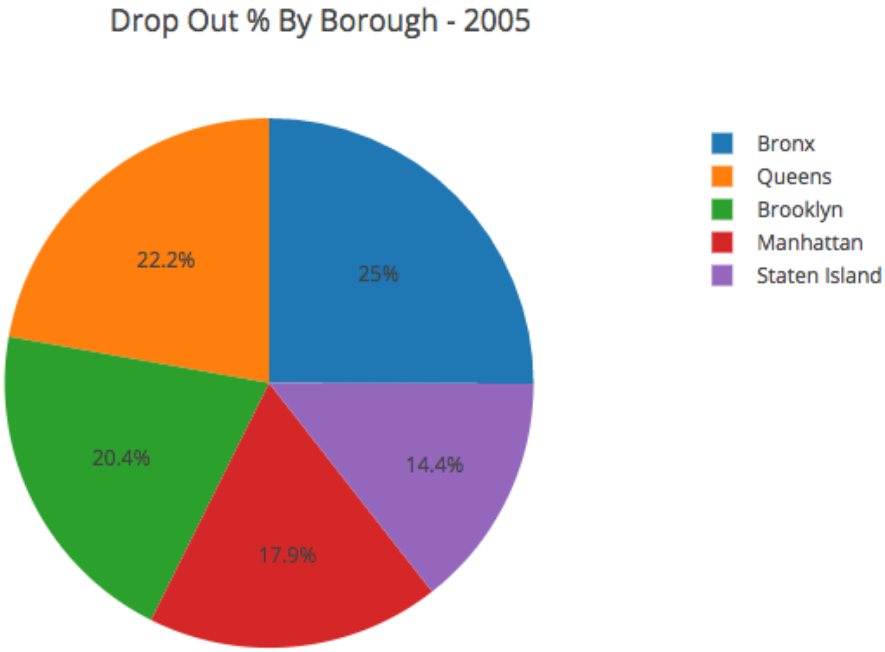


Figure 7.5: A Pie Chart Showing the Drop Out Percentage by Borough in 2005

Summary of Education
The results identified by analysing the Education factor showed that the Bronx had the highest number of student drop out rates. It was shown in Table 2.1 that the Bronx had the lowest average income compared the other boroughs. By combining this data with the statistics identified from researching into the demographics of the area, further insight could be derived as to the cause of this disparity. This analysis would be furthered in the Composite Analysis chapter.

7.2 Environment

To analyse the environment of New York City, the 311 dataset was employed as it described a variety of issues that citizens were discontent about. The dataset contained calls that spanned from 2010 to 2017, making it one of the larger datasets processed in the project. Due to its size, it was not possible to read the data straight into Jupyter Notebook as it exceeded the memory allocations of the program, so Sublime Text was used instead. To mitigate this problem, it was first split into separate files based on the 'Borough' attribute, which represented where the call was taken from. This was achieved by writing a Python script that iterated through each row in the dataset, and copied the row to a new file depending on which borough it belonged to.

The 311 dataset was a comprehensive record of all the calls that were taken by the 311 agency. The attributes contained a variety of information such as date and time of call, a description of the issue raised, the agency that was responsible for following up the the issue, and the outcome. This depth of information allowed rich analysis and manipulation of a range of attributes. The following section will explore the 311 data in a variety of ways, such as scatter plots of geographical location, histograms of complain frequency and pie charts of complaint proportions.

7.2.1 Investigating Locality Using Scatter Plots

To create a representation of what the call data displayed, scatter plots were utilised to show the geographic locations of calls. The latitude and longitude positions of each call were used as the x and y axis respectfully. The colour of the point was dictated by which agency was employed to resolved the complaint call. This gave an indication of what type of complaint it was. Utilising a scatter plot enabled patterns to emerge that would highlight whether some areas were experiencing more discontent about a particular issue than others. Plotly had the functionality to plot onto a base map, although due to the size of the 311 data it was unusable in this case. Instead the visualisations were created with the background set to being transparent. However, as the data was so dense, the boundaries of each borough were easily identified.

The Bronx
The plots in Figure 7.6 show how the densely packed 311 data created a visual representation of the streets in the Bronx. The scatter plot was made up of a variety of different coloured points, with the most populous being orange. As shown in Figure 7.6 and 7.7, these points correlate to calls whose complaint was handled by the New York Police Department (NYPD). Additionally, Figure 7.7 shows the interactive element of Plotly, which was engaged by rolling the cursor over the points. In doing so, the agency associated with the point highlighted was displayed.

From these graphs it is clear that the highest frequency of calls were handled by the NYPD. By further analysing the data, more information could be gained as to what issues those calls were centred around. This information was later identified using histograms, and will be explored more fully in the next section.

Figure 7.6: Scatter Graph of The Bronx

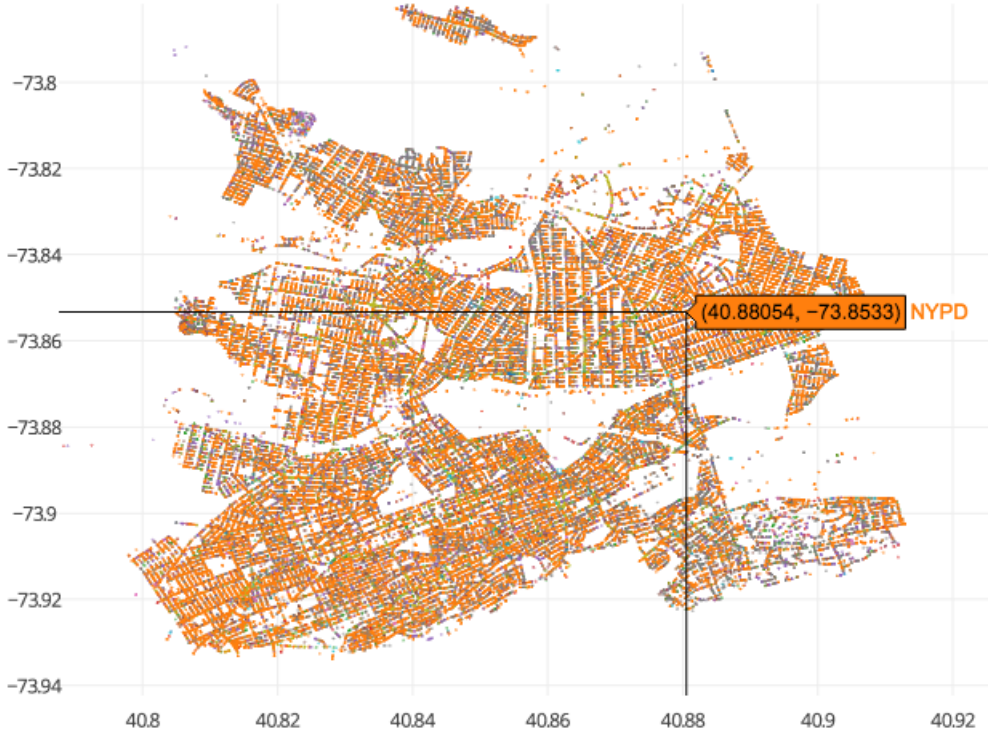


Figure 7.7: Scatter Graph of The Bronx With an Interactive Element

Brooklyn

In the 311 dataset, the data collected from Brooklyn was sufficiently large that it could not be visualised using Plotly . This was principally due to the large number of residents residing in borough, which caused a higher frequency of calls to the 311 service. However, insight on the types complaints the citizens had could still be gained by analysing the contents of this data is other visual ways. For example, the pie chart in Figure 7.16 was created to understand what proportion of complaints were most frequent amongst the citizens of Brooklyn. This analysis provided the same insight into discontent in the borough without the use of a scatter map. The disadvantage of this limitation was that understanding could not be gained on whether particular areas had complained more about certain things than others, however later work in the Composite Analysis phase remedied this by using smaller and more specific datasets to measure resource equity in Brooklyn.

Manhattan

With Manhattan being at the centre of New York City, it drew some interesting observations. The graph in Figure 7.8 shows the result when the 311 data was visualised using a scatter plot. Similarly to the Bronx, it contained a majority of orange points corresponding to NYPD managed calls, however there was a larger number of other colours that came through in addition.

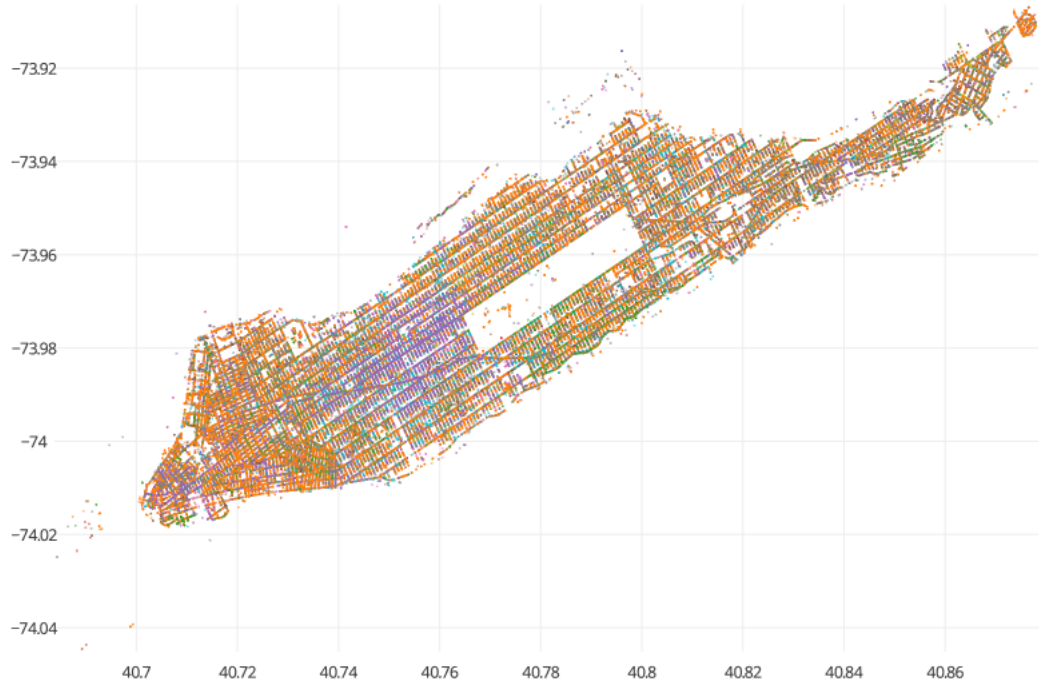


Figure 7.8: A Scatter Graph of Manhattan

In the east of the borough, there were a number of green points that correspond to calls managed by the Economic Development Corporation (EDC). This can be seen in Figure 7.10. Additionally, the south west of the borough showed a large cluster of purple points. These were calls that were handled by the Fire Department of New York. It seemed unusual that there would be such a large number of points concerning the Fire Department all highly centralised to one area of downtown Manhattan. This was an interesting conclusion drawn from this visualisation, which could be further investigated in later phases of analysis.

Queens

As shown in Figure 7.11, the borough of Queens was heavily covered in calls that were managed by NYPD. The shape of the area is very clearly seen, showing that there are a high number of calls from all parts of the borough. To the east of the plot, there is a small patch of red points, corresponding to calls directed to the Department of Buildings (DOB). It was identified in the Research chapter that Queens is a primarily residential area. However, the calls complaining about buildings are specifically localised to a small area of the borough. This question could be answered through further analysis of the 311 dataset.

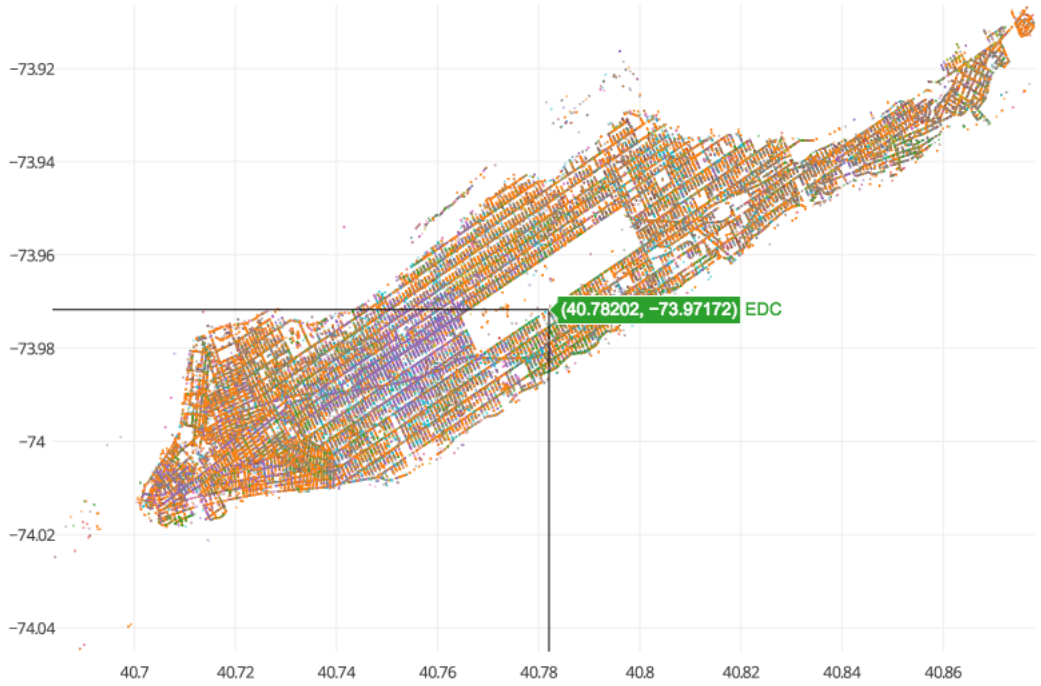


Figure 7.9: Patterns Identified in the Manhattan Scatter Plot Showing EDC

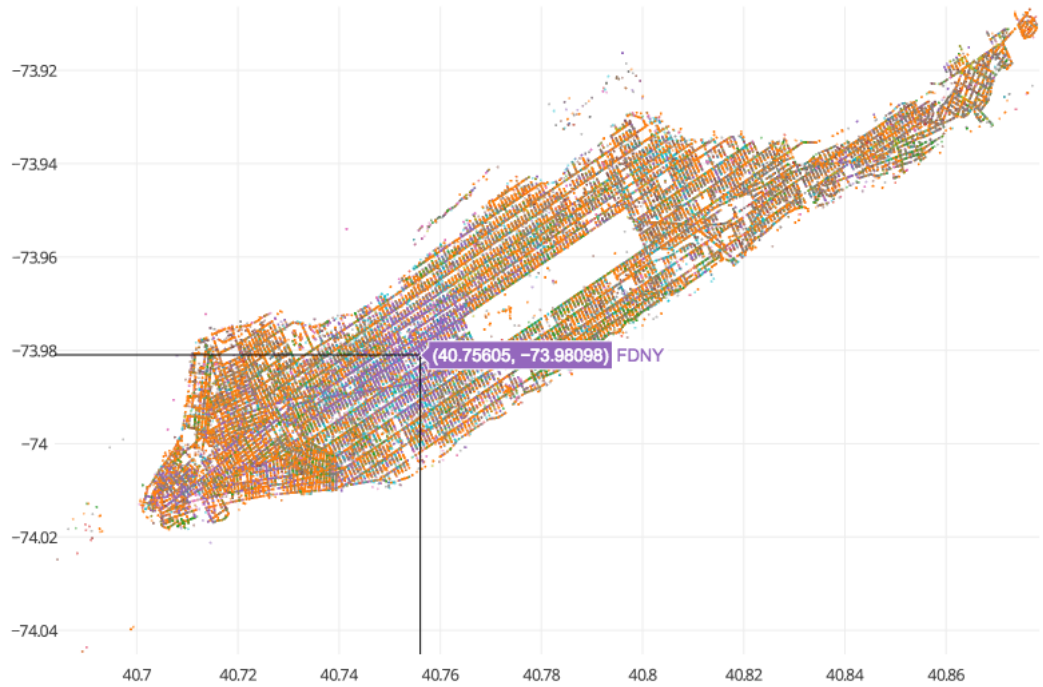


Figure 7.10: Patterns Identified in the Manhattan Scatter Plot Showing FDNY

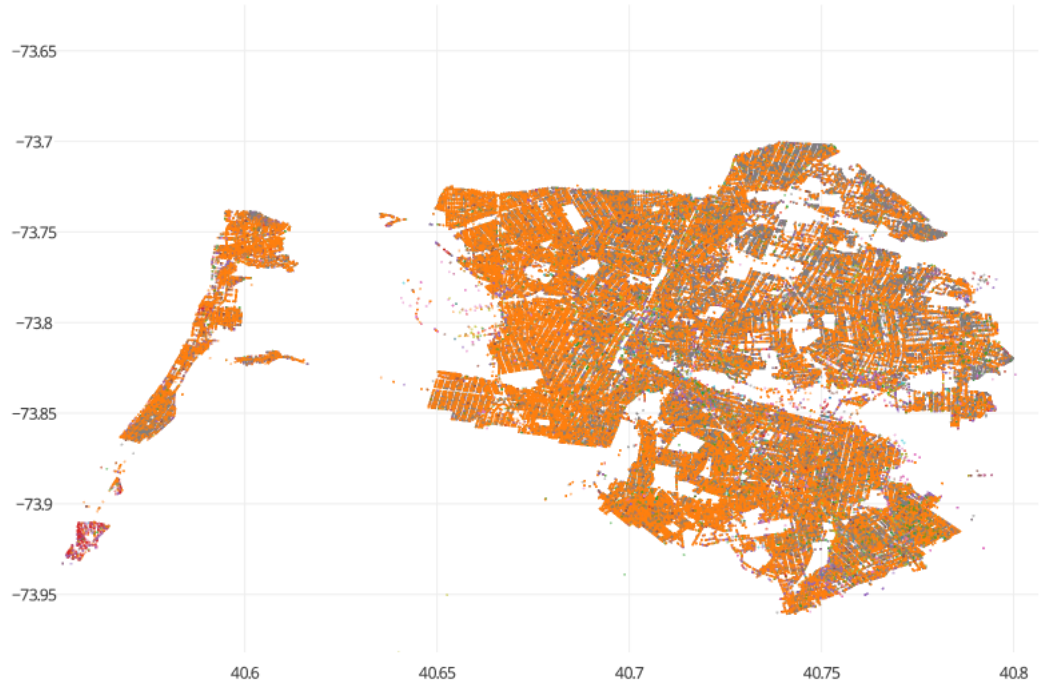


Figure 7.11: Scatter Plots of Queens

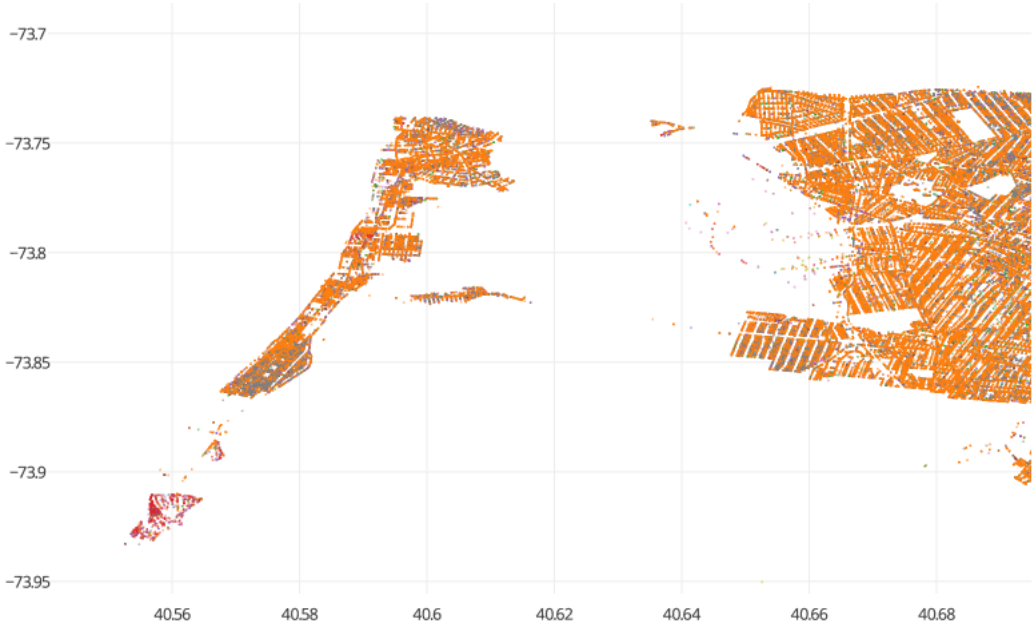


Figure 7.12: Scatter Plots of Queens Highlighting a Cluster of DOB Points

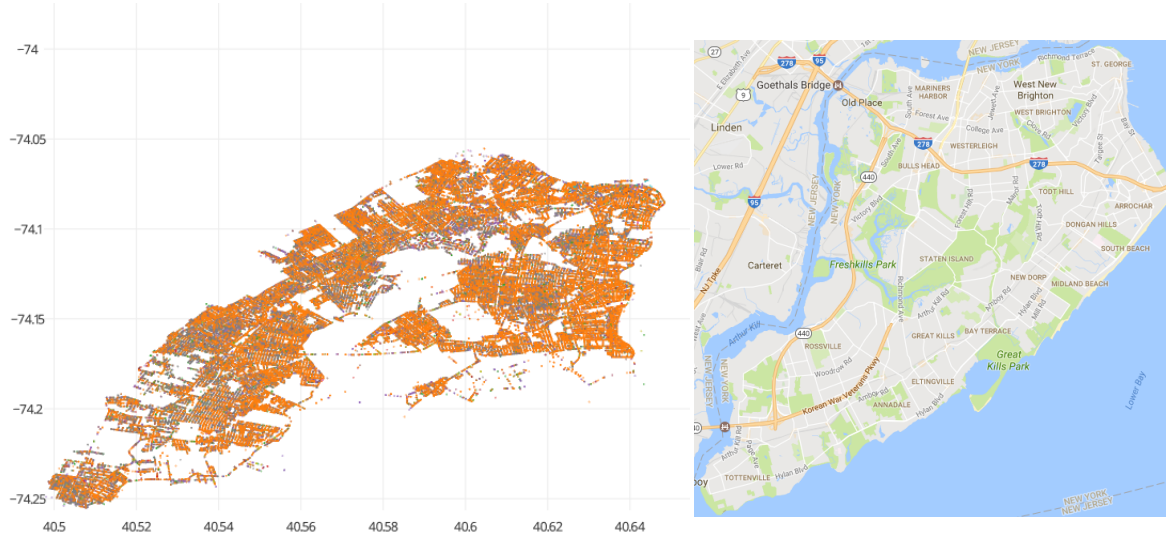


Figure 7.13: Scatter Plots of Staten Island Compared to a Map of the Borough [40]

Staten Island
The visualisation of the Staten Island dataset is much sparser in comparison to the other boroughs analysed. The left image in Figure 7.13 shows the result of the data visualisation compared to a map of the borough [40]. This sparsity was assumed to be due to the large parkland areas in the borough, where less calls are taken from. In a similar pattern to the other boroughs, the majority of calls in Staten Island were in reference to issues investigated by NYPD. However, there were no specific questions raised by analysing the Staten Island visualisation.

Overview of Scatter Plots
By visualising the 311 data in scatter plots, a number of questions were raised that would otherwise not have become apparently by just inspecting the raw data. A summary of these questions are described below.

- Why were there such a large number of calls managed by the NYPD? What were these calls complaining about?
- Why does Manhattan have a number of densely located points corresponding to calls managed by the FDNY?
- What are the high frequency of complaints to the DOB in Queens in regards to?

After the completion of the Single Factor Analysis phase of the system, these questions were revisited with the most interesting taken into the Composite Analysis phase for further investigation.

7.2.2 Investigating Complaint Frequency Using Histograms

From the scatter plot analysis, it was clear to see that some issues were more common than others. For example, the majority of calls were about issues that were handled by the NYPD.

This allowed some insight, however there was still limited understanding as to what these issues these calls were in regards to. By constructing histograms to measure the frequency of complaint types, further information was gained from the 311 dataset.

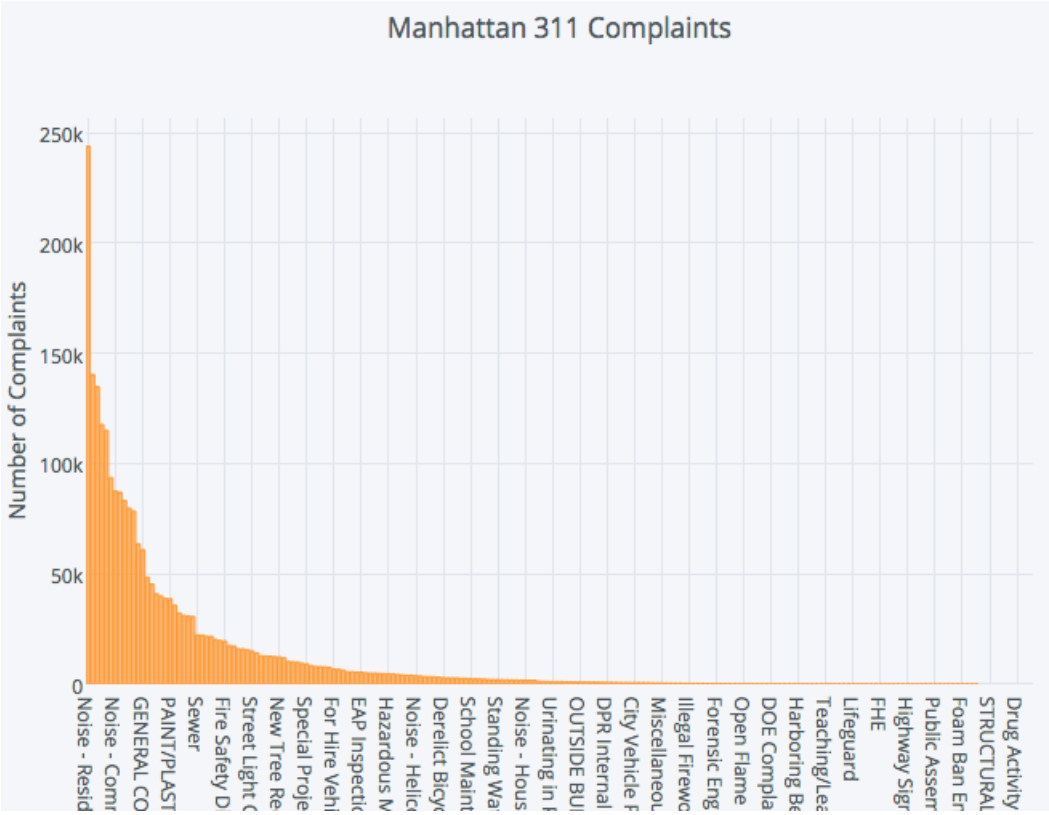


Figure 7.14: A Histogram of Complaint Types from 311 Data for Manhattan

The histogram shown in Figure 7.14 shows how the most recurrent complaint type was that of noise, both residential and commercial. This was followed by complaints about heating, hot water and street condition. By inspecting the results from other boroughs like those shown in Figure 7.15, it could be identified whether this trend was unique to Manhattan or mirrored in other locations.

Table 7.1: A Table Highlighting the Top 5 Complaints in Different Boroughs

Borough	**1**	**2**	**3**	**4**	**5**
The Bronx	Noise Residential	Heating	Hot Water	Plumbing	General Construction
Manhattan	Noise Residential	Noise Commercial	Heating	Hot Water	Street Condition
Queens	Noise Residential	Blocked Driveway	Street Condition	Street Light Condition	Water System
Staten Island	Street Condition	Street Light Condition	Water System	Noise Residential	Sewer

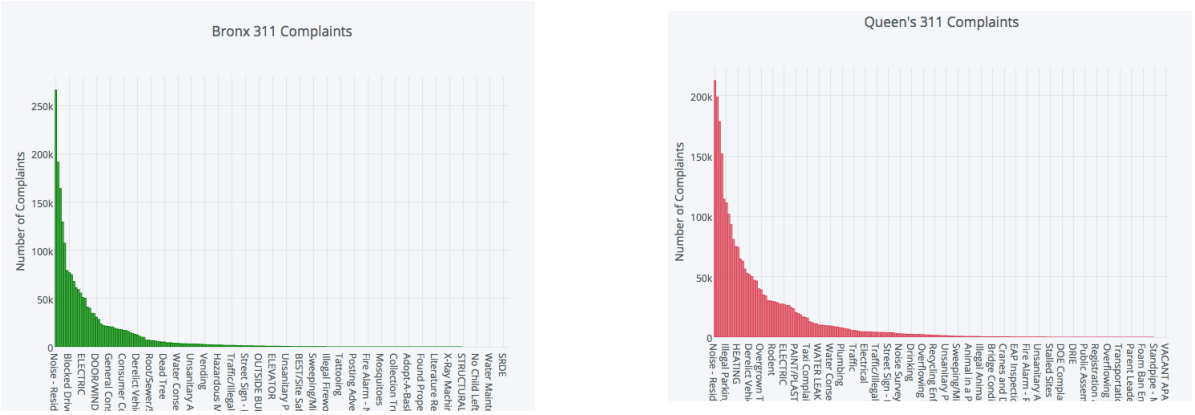


Figure 7.15: Using Histograms to Count Frequency of Complaint Types

Table 7.1 highlights how many of the issues are common place in all five boroughs. Issues such as residential noise consistently have high numbers of calls, along with complaints surrounding heating and hot water problems. This result seemed to indicated that the quality of buildings of New York City are a lower standard across all boroughs.

7.2.3 Investigating Proportionality of Complaints Using Pie Charts

To further gain understanding of what type of issues were present across the boroughs, pie charts were utilised to measure what proportion of calls were managed by each agency. This validated the speculated results from the scatter plots and gave a quantified measurement to base conclusions on.

From the pie charts shown in Figure 7.16, 7.17 and 7.18 it was clear to see that all the boroughs followed a similar trend, with the Housing Preservation and Development (HPD) department and NYPD as the two largest agencies that managed complaints. The only case where this differed was in the Bronx, which had almost twice as many complaints aimed at HPD than any of the other boroughs. This information was interesting, as the scatter graph indicated most calls were targeted to NYPD. Although the pie chart visualisations provided validation of previous conclusions, they didn't provide insight into what was causing these issues in the first

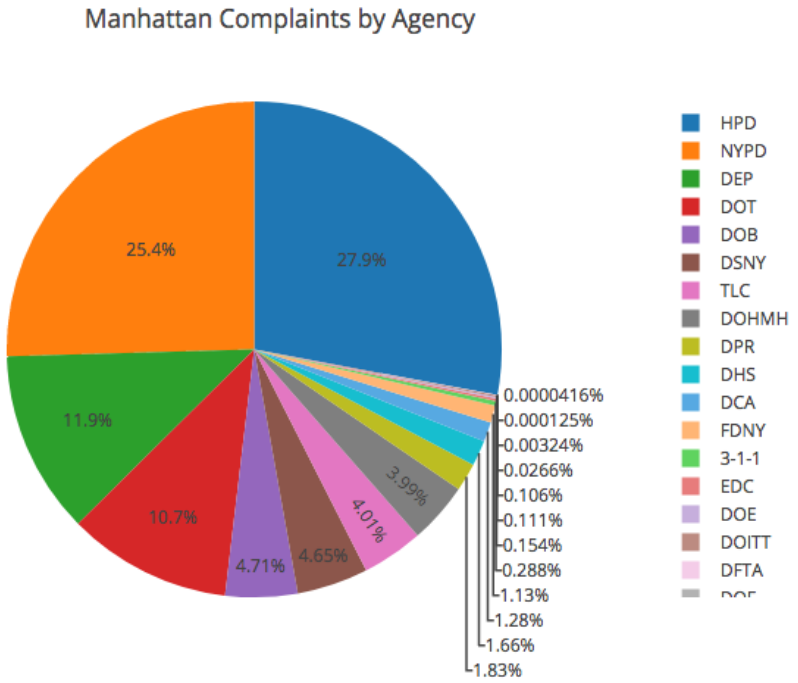


Figure 7.16: Pie Charts Showing Complaints by Agency for Manhattan

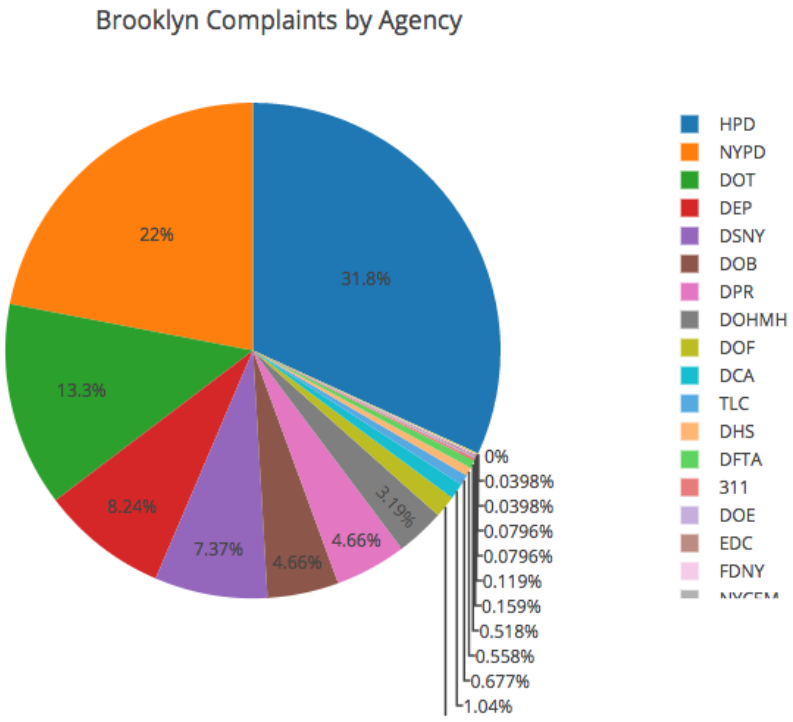


Figure 7.17: Pie Charts Showing Complaints by Agency for Brooklyn

instance. This was a limit of the Single Factor Analysis phase, however following the methodology described in Chapter 5, these results could be further explored in the following Composite Analysis phase of the project.

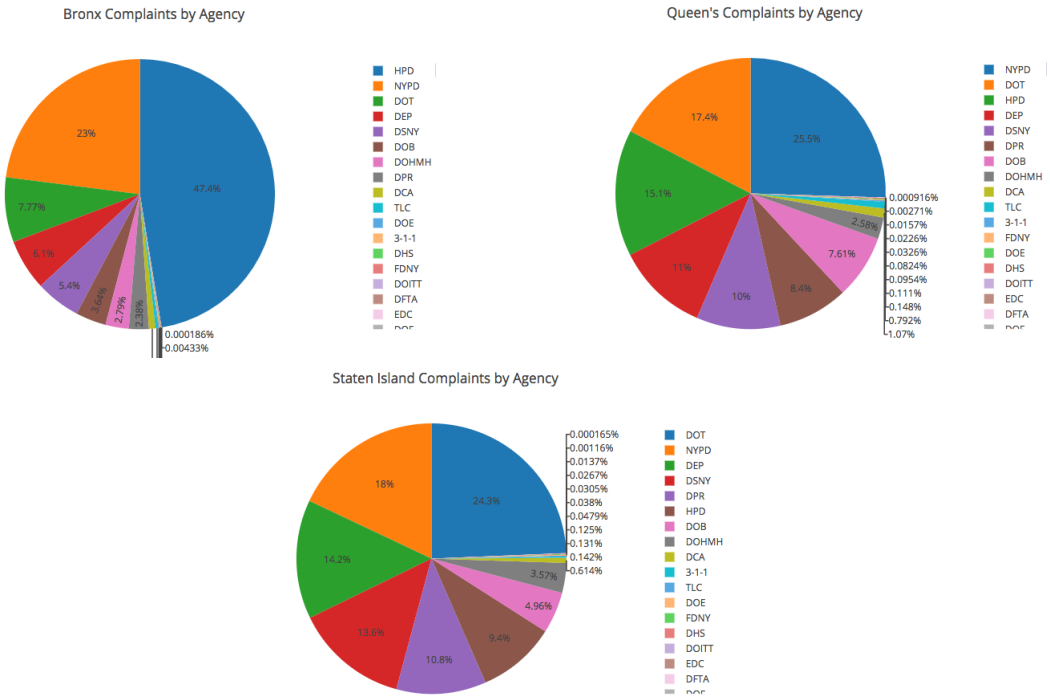


Figure 7.18: Pie Charts Showing Complaints by Agency for the Bronx, Queens and Staten Island

Summary of Environment

By analysing the Environment factor through the use of the 311 dataset, a variety of questions were raised. Inconsistencies were identified from examining scatter plots, such as the high frequency of calls in Manhattan to the FDNY, along with other clusters of DOB complaints in the Bronx. The complains were then visualised using histograms to measure the frequencies of complaint types, with the most common being complaints about noise and building quality. Finally, the proportion of complaint types per borough was analysed by using pie charts. These graphs suggested that most boroughs followed a samiliar trend of having a high proportion of calls regarding building quality, resulting in the majority of complaints managed by HPD.

Throughout the analysis of this factor, key themes were repeatedly raised. The unsatisfactory quality of buildings was been identified a through a multitude of analysis, indicating housing quality was a city wide problem. Additionally, as inconsistency was observed in FDNY calls in Manhattan, suggesting further analysis could be made into the availability of fire previsions in the area. The results identified in this section could be further explored in the Composite Analysis phase, by fusing together data to provide a richer picture of the issues explained above.

7.3 Healthcare

As healthcare is a prerogative of most New York City citizens, it is an area where multidisciplinary research has been trialled to produce more effective solutions to hospitals and emergency departments [15]. The analysis for this factor will take a similar approach by inspecting a range of healthcare information. The Health and Hospitals Corporation (HCC) was the primary agency that provided health related data on NYC Open Data, and therefore was used to investigate the healthcare previsions of New York City.

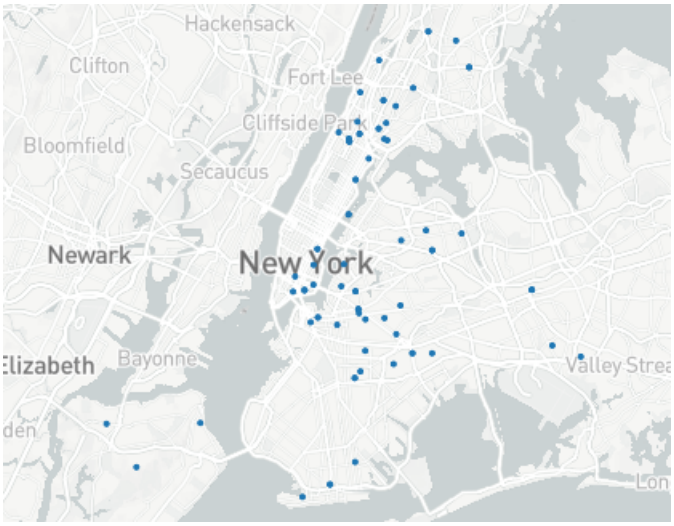


Figure 7.19: A Scatter Plot Showing the HCC Facilities in New York City

The scatter plot in Figure 7.19 showed that the majority of health facilities were situated in the Bronx or north-east Brooklyn. Areas such as Staten Island seemed to have less access to hospital facilities, with only three plotted in the north of the borough. This analysis suggested that some citizens would have to travel further to seek medical help, which could highlight possible unfairness for citizens like those in Staten Island.

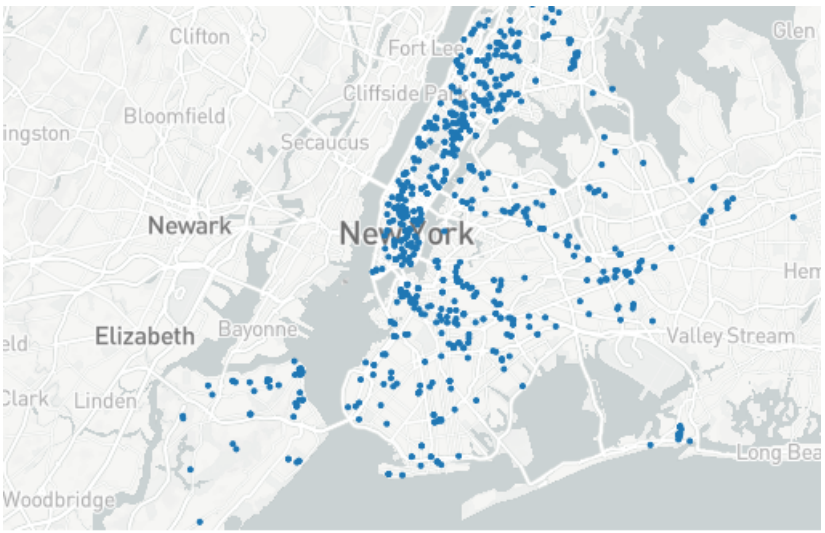


Figure 7.20: A Scatter Plot Showing Mental Health Services Across New York City

To gain a different view of the healthcare factor, the specific area of mental health was investigated. The scatter plot in Figure 7.20 shows the resulting plot using the Mental Health Service Finder data from NYC Open Data. This plot displays a similar trend to that shown in Figure 7.19 with sparsity in the suburban boroughs of Staten Island and Queens. However, this data showed a larger network of mental health facilities in Manhattan.

Summary of Healthcare
It seemed that there were strong provisions for hospital access in the centre of the city, however this access appeared to dissipate in the outlying suburban areas. Similarly, the accessibility to mental health facilities seemed strong around Manhattan and the Bronx. Without further analysing the request for these services in other boroughs, it was difficult to tell if a proportion citizens located in these further a field locations were being disadvantaged by the central placement of facilities. These findings raised suggestions that would be further studied in the Composite Analysis phase of the project.

7.4 Transport

When initial research was undertaken in Chapter 2, it was discovered that the Taxi and Limousine Commission (TLC) was responsible for all public transport in New York City. By searching NYC Open Data for datasets published by TLC, the 2015 Yellow Taxi Trip data was identified. This dataset was a comprehensive list of all the taxi trips taken between January to June 2015, which contained a variety of attributes such as pick-up location, pick-up time, drop-off location and drop-off time. Using these values, a comprehensive picture of the taxi travel in New York City could be created by visualising the data in a scatter plot.

The scatter plots in Figure 7.21 and 7.22 utilised the latitude and longitude coordinates to plot a point where the taxi either picked up a passenger or dropped them off, respectfully. Figure 7.21 showed how the pick-up points were highly localised to the centre of Manhattan with a

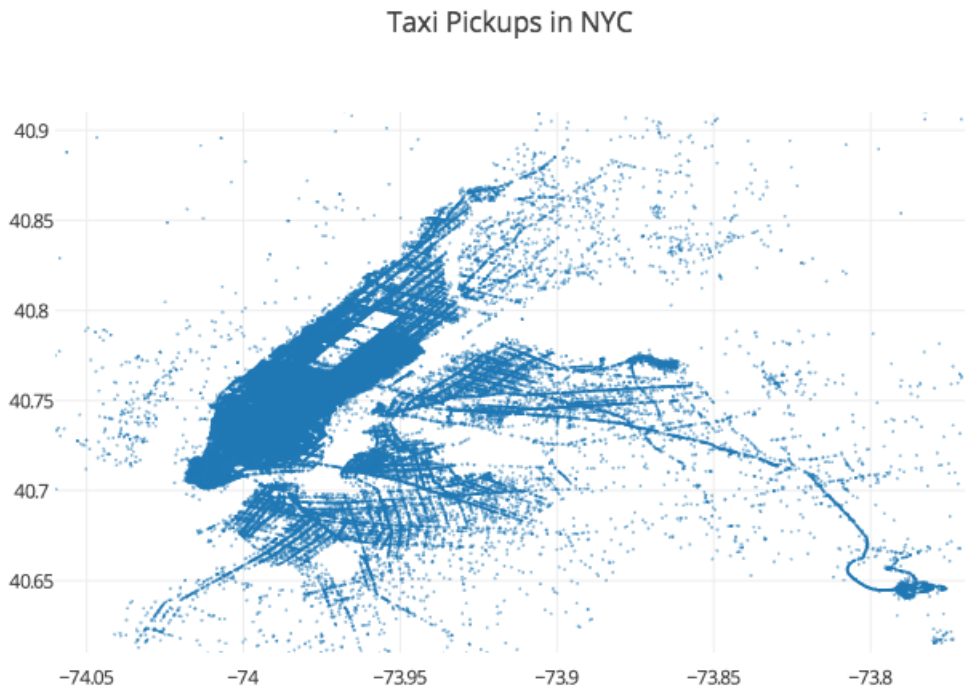


Figure 7.21: Scatter Plots Showing the Taxi Pick-up Points

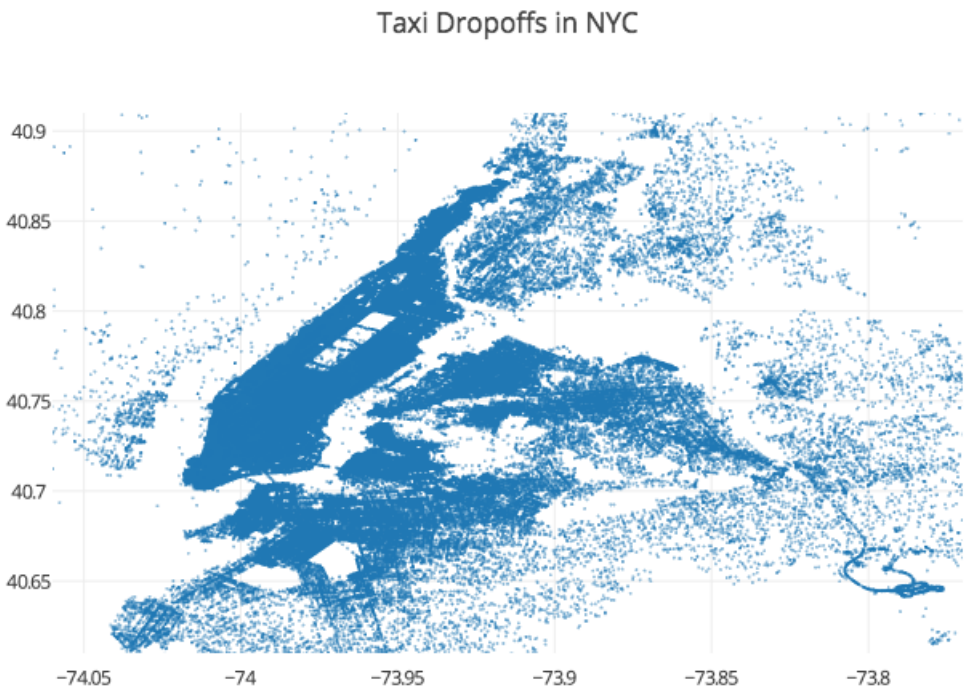


Figure 7.22: Scatter Plots Showing the Taxi Drop-off Points

small number of specific areas in west Brooklyn highlighted. However, Figure 7.22 showed that the drop-off points were much more sparsely spread across the whole region. If the number of passengers taking trips into the city was the same as the number leaving the city, these representations would be identical. However, this difference indicates that more citizens were taking taxis out of Manhattan and into the residential suburbs. The visualisation of the taxi data in this way allowed this interesting insight to be observed. It is unclear whether the is evidence to substantiate this claim, but it is a question that could be further analysed with composite analysis.

Summary of Transport
As discussed, there was an unbalance in the number of passengers that were being picked up in Manhattan and traveling out of the city, compared to those being picked up in the outer boroughs and travelling into the centre. It can be speculated that this was due to a lack of other public transport options available. For example, trips could have been taken late at night when the subway services were running at less frequent intervals, as identified in the Research chapter. In order to develop this analysis, more information was needed to conceptualise what other factors had been an influence to result in this pattern of data.

7.5 Public Services

Research into the City Council highlighted how there were many public services available to the people of New York City, managed primarily through the Mayor's Office. A large proportion of these services included facilities covered by other factors, such as schools, hospitals and public transportation systems which were discussed in the Education, Healthcare and Transportation sections respectively. To narrow the scope of the analysis within Public Services, it was decided that an specific area of these facilities would be chosen to represent the factor and create data visualisations. This area was chosen to be emergency services, derived from the research carried out in Chapter 2, which explained the importance of the Fire Department of New York. The following section describes the results found when evaluating the emergency services datasets, and draws conclusions regarding possible inequity in fire previsions in some boroughs.

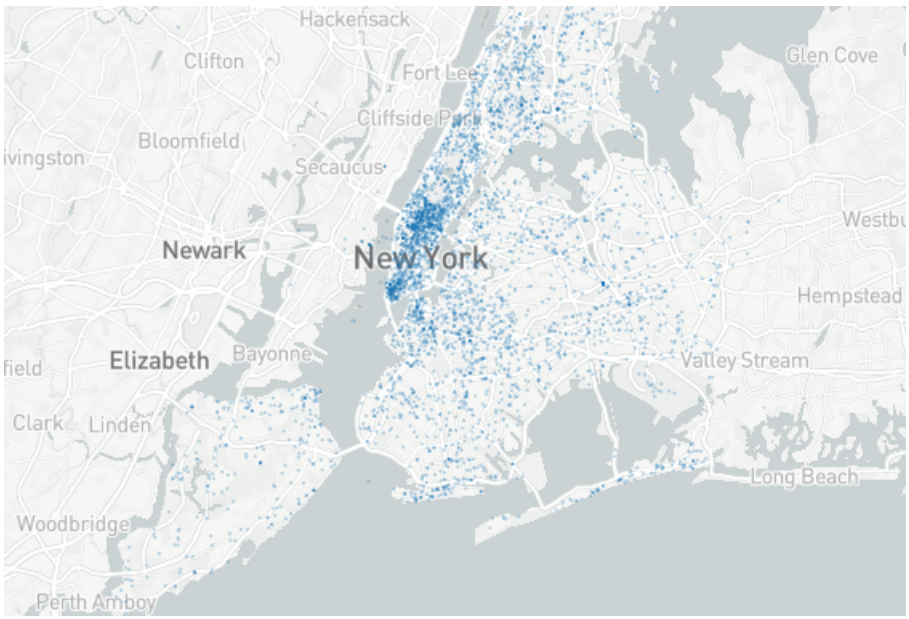


Figure 7.23: A Scatter Plot of Emergency Response Incidents in New York City

The scatter plot in Figure 7.23 showed large number of points localised in the centre of the city. These points represented the latitude and longitude coordinates of where emergency response incidents were reported from. The visualisation illustrated that more incidents occurred around the downtown Manhattan displayed by densely clustered points. From the research collated about this location, it was inferred that there was a higher level of activity in this area due to the location of offices, bars and restaurants, resulting in a higher frequency of footfall. This increased number of people could have been the cause of more emergency response incidents in the area. To validify this assumption, further data was needed. For example, population density would determine whether downtown Manhattan was an area of more congestion. By collecting footfall and population statistics, analysis could further extrapolate information as to why there was a greater number of incident calls from Manhattan than in Brooklyn or Queens.

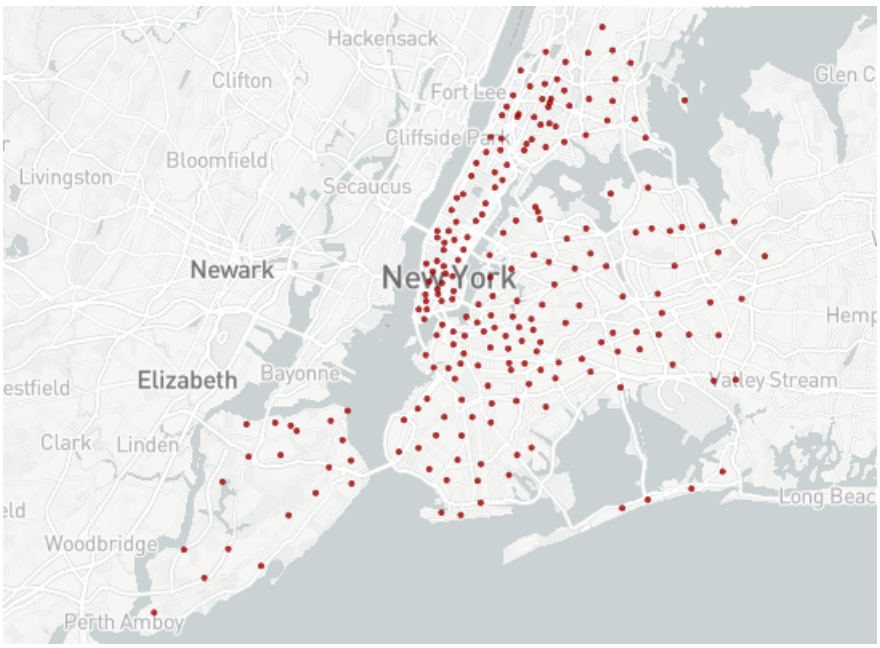


Figure 7.24: A Scatter Plot of FDNY Facilities Across New York City

Further study was undertaken to explore where fire fighting provisions were positioned. Operational research conducted by the New York City-Rand Institute described investigations made to identify the most effective locations for fire fighting facilities by developing algorithms to assign fire companies to fire houses. In doing this, they generated mathematical formulas that identified the optimum locations of tower ladders to assist in emergency situations [83]. These studies indicated that the council and Fire Department had provided a lot of thought about the placement of these services. To explore locations of current facilities, data containing the locations of FDNY facilities was collected from NYC Open Data and plotted on a base map. This resulted in the visualisation shown in Figure 7.24.

The FDNY facilities plot in Figure 7.24 showed a large number of fire fighting provisions in a structured order across Manhattan and the Bronx. This pattern was extended to the more suburban areas in the east of the city around Brooklyn and Queens. This dataset alone could not provide analysis as to whether these provisions were adequate for the defence against fire in these regions. However, combining this data with information sourced from the 311 data discussed in the Environment factor could allow conclusions to be drawn about the success of FDNY resources.

Summary of Public Services

Analysis into the provisions of public services highlighted that there were a large number of emergency response incidents located in the centre of the city. It was discussed that more data would be needed to understand whether this cluster was proportionate to the number citizens in this area. The FNDY facilities replicated a similar view to the emergency incidents plot, which indicated that suburban boroughs may have longer wait times for fire fighting services. If wait time data could be sourced, this claim could be further analysed.

7.6 Overview of Findings

The first part of the system provided a large number of results from analysing each single factor independently. A variety of visualisations were produced to identify patterns in the datasets for each factor. From the range of visualisations that were created, the most telling graphs were the scatter plots. They utilised the latitude and longitude coordinates to provide a geographical representation of the factors distribution around the city, which highlighted many nuances in the data. By combining these plots with quantitate analysis from the pie charts and histograms, a variety of questions were raised surrounding the equity of resource distribution around the city. A summary of the questions raised from this analysis is shown below. These results are labeled for future use in testing and evaluating the system.

Education

A1: *What caused the reduction in drop out rates between the cohort of 2001 to 2005?*

A2: *Why did the Bronx have the highest rates of students dropping out of education compared to other boroughs?*

Environment

A3: *Why is there a higher number of complaints to the Housing Preservation and Development agency in the Bronx?*

A4: *Why are there a cluster of calls to the Fire Department of New York in downtown Manhattan?*

Healthcare

A5: *Are mental health provisions less accessible for citizens in suburban boroughs?*

Transport

A6: *Why are there more taxi trips taken out of the city compared to those taken into the city?*

Public Services

A7: *Are Manhattan citizens more at risk of fire due to a higher level of footfall?*

A8: *Does each borough have equal access to fire prevention provisions?*

The analysis undertaken in this section raised some interesting discussions, however answers to all questions could not be found due to the approach of keeping the factors separate. Following the methodology devised in Chapter 5, a number of factors were chosen for further investigation. These factors were selected by evaluating how promising the results from the Single Factor Analysis phase had been. It was decided that the factors to continue to the next phase of the project would be Education, Healthcare and Public Services. By combining the original datasets with other newly sourced additional data, further insight was gained to substantiate some of the questions raised in this chapter.

CHAPTER 8

Composite Analysis

The next component of the system that aimed to analyse resource equity as developed in by making use of composite analysis. The objective of the Composite Analysis phase was to further the investigation made using the analysis from the single factors by fusing together additional datasets. The outcome of the Single Factor Analysis phase provided three factors that were chosen due to an expectation of promising results. These factors were Education, Healthcare and Public Services. The pipeline in Figure 8.1 shows an overview of the system and highlights the transition between the Single Factor Analysis and Composite Analysis phase. For each of these factors, the previous phase identified questions which focused the direction of composite analysis and gave guidance when sourcing additional datasets. The need for this additional data was highlighted in the Single Factor Analysis, as researching single factors alone limited the amount of understanding that could be achieved. The following chapter will discuss the supplementary data that was sourced for this phase of development, and continue to describe the results identified for each area of investigation.

8.1 Supplementary Data

In this phase of the methodology, supplementary data needed to be sourced to facilitate composite analysis. The supplementary data needed to be related to the three factors identified for further analysis, and so it was decided that it would also be sourced from the NYC Open Data portal to maintain consistency. The information shown in Table 8.1 describes what data was sourced to gain further insight into the issues raised in Education, Healthcare and Public Services.

8.2 Composite Analysis

Once the additional supplementary data was sourced, it could be utilised using composite analysis in conjunction with the original datasets. The following section will reiterate the questions raised from the Single Factor Analysis phase to review what areas needed further

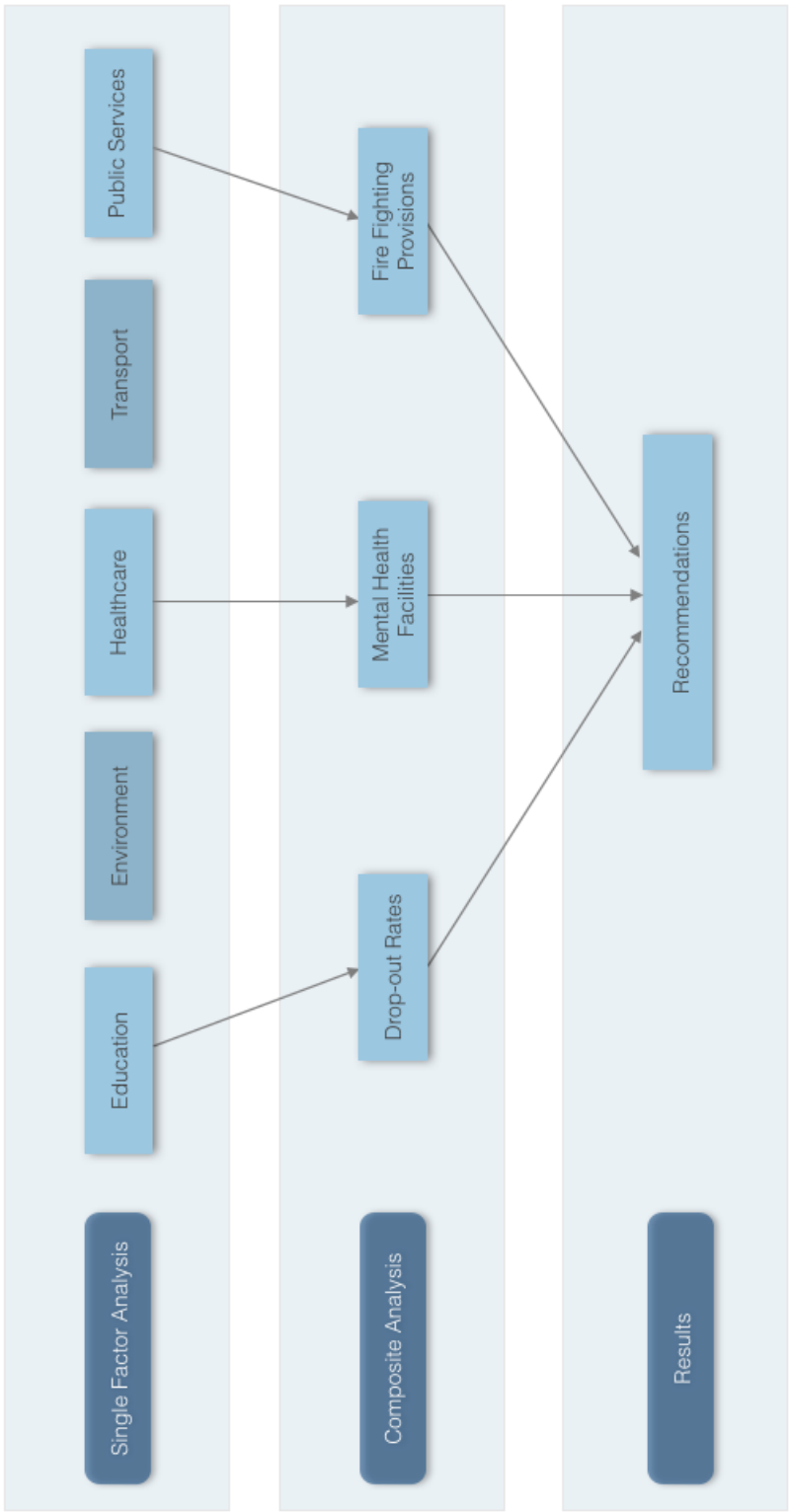


Figure 8.1: The Pipeline of the System

Table 8.1: A Table Showing a Sample of Additional Datasets Sourced for Composite Analysis

Dataset	Agency	Summary
New York City Population By Census Tracts	Department of City Planning	Population figures from the 2010 census by borough
Fire Companies	Department of City Planning	The boundaries of fire companies across the city
Fire Incident Dispatch Data	Fire Department of New York	Occurrences of fire incidents where resources were dispatched

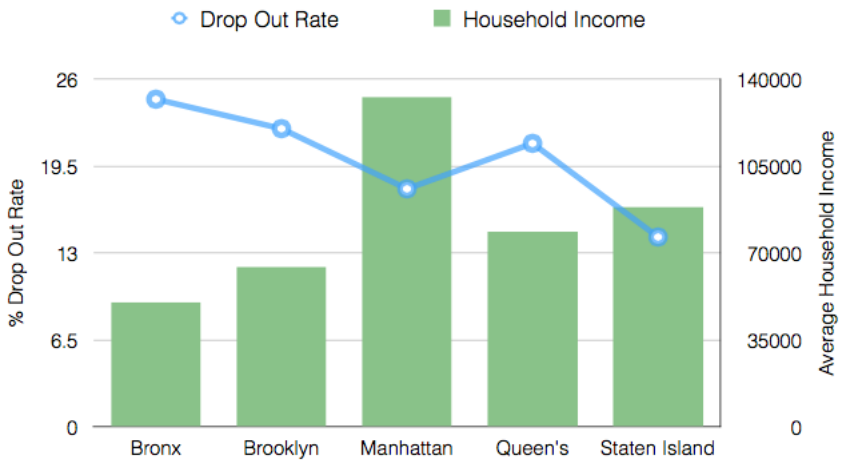


Figure 8.2: A Plot Showing Average Household Income Compared to Dropout Rates

consideration . These questions formed the direction of each investigation and ascribed the high level aims that this phase intended to answer.

8.2.1 Education

The findings from the Single Factor Analysis phase indicated that there were disadvantages for students educated in specific boroughs. In particular, students from the Bronx were found to be more likely to drop out of school before graduation compared to those studying in other boroughs. By combining this information with preliminary research conducted in Chapter 2 allowed potential contributing factors of this issue to be identified.

In the Research chapter, Table 2.1 showed the average household income for each borough. The plot in Figure 8.2 showed the results of combining this data with the original Graduate Outcomes dataset. The pattern that emerged from this visualisation showed that the Bronx had the highest drop out rate and also the lowest average household income. In addition, boroughs such as Brooklyn and Queens followed a similar pattern. This indicated that a low income was an aspect that could have caused students to be more likely to terminate their education.

8.2.2 Healthcare

The analysis into the Healthcare factor resulted in the premise that medical facilities were more densely populated around Manhattan. This raised the question surrounding accessibility to healthcare for citizens in suburban boroughs such as Staten Island. To test this hypothesis, it was necessary to gain a measurement of how many citizens resided in borough. For this purpose, the 2010 Census data was sourced to provide a population figure for each of the boroughs. This allowed the number of facilities to be normalised by the number of citizens living in a given area, which quantified how much access these citizens had to healthcare facilities. It was decided that the access to mental health facilities would be measured by utilising the Mental Health Service Finder dataset explored in Chapter 7. The data in Table 8.2 shows these results.

Table 8.2: A Table Showing the Normalised Number of Mental Health Facilities Per Borough

Borough	**Normalised Number of Mental Health Facilities**
Bronx	0.000105
Brooklyn	1.11789×10^{-5}
Manhattan	0.000182
Queens	0.00147
Staten Island	0.000123

These results show that citizens in Brooklyn did not have equitable access to mental health facilities. When the number of resources in each borough were divided by the population value from the 2010 Census data, the data showed that Brooklyn had a much smaller in comparison. This result could indicate that it is more difficult for a resident of Brooklyn to see a medical professional about their mental health concerns which may discourage them from seeking advice. Alternatively, patients may have to travel to other boroughs for medical appointments, due to a lack of provisions around their local neighbourhood.

8.2.3 Public Services

An interesting observation that was drawn from the Single Factor Analysis phase identified a cluster of 311 calls in the south of Manhattan which directed to the FDNY department. This prompted enquiry as to whether Manhattan was at more risk of fire outbreaks compared to any other borough. The analysis of the Public Service factor in Chapter 7 utilised the Emergency Response Incidents dataset by plotting a scatter map to identify areas that had high levels of incident frequencies. As the Composite Analysis phase aimed to narrow down the scope of the investigation into Public Services, the Fire Incident Dispatch dataset was sourced to provide FDNY specific data. This was used in collaboration with the Fire Companies dataset, which gave guidance on the areas each fire company protected. These new datasets were utilised by creating composite visualisations shown below.

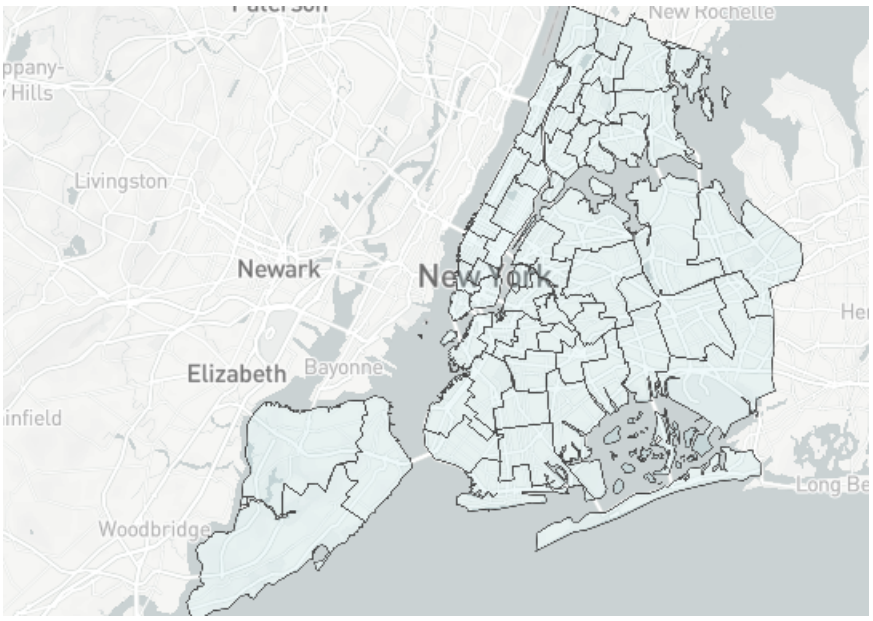


Figure 8.3: A Plot Showing Fire Company Boundaries Across New York City

The initial plot created in Figure 8.3 utilised only the new Fire Companies dataset. It clearly mapped the areas which were managed by subsections of FDNY across New York City. By combining this with the Fire Incident Dispatch data, an additional visualisation was generated shown in Figure 8.4. In this plot, the points was positioned at each zip code around the city. The size of the points was dictated by the frequency of incidents that occurred in that area, meaning the larger the point, the more FDNY resources were called out to tackle incidents that happened in that location. This analysis showed larger points around downtown Manhattan, the Bronx and Brooklyn. However, from this data it was unclear to know whether the boundaries of the fire companies ensured these high frequency incident areas still had the same accessibility to firefighting provisions than other areas.

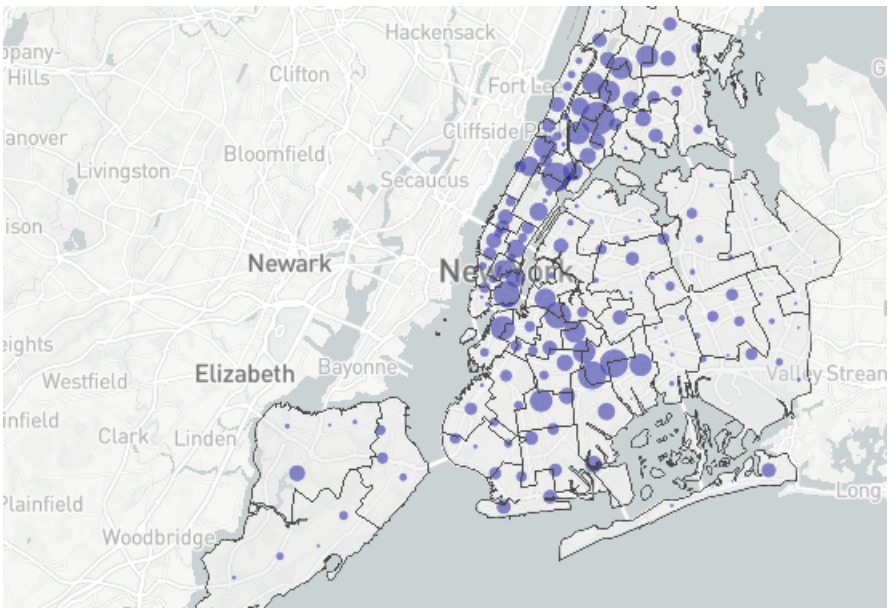


Figure 8.4: A Plot Showing Fire Company Boundaries and FNDY Request Density

Table 8.3: A Table Showing the Normalised Number of Incidents Compared to the FNDY Facilities per Borough

Borough	**Normalised Number of Incidents**	**FDNY Facilities per Borough**
Bronx	0.000374	34
Brooklyn	0.0000406	66
Manhattan	0.000942	48
Queens	0.000299	50
Staten Island	0.000456	20

To further this investigation, the frequency of occurrences in the Fire Incident Dispatch dataset was normalised by the population density given in the 2010 Census data. This produced the results shown in Table 8.3. The normalised values showed that Manhattan had the highest number of incidents out of all five boroughs. This was contrasted with the number of FDNY facilities in each borough. Interestingly, Manhattan did not have the largest number of fire fighting provisions. These results indicated that citizens living in Manhattan were more at risk of fire than those living in other boroughs, will suggests there is some inequity in public service resources for those residents.

8.3 Results

The Composite Analysis phase of the project aimed to deepen the understanding that was gained during the Single Factor Analysis phase. This was delivered by identifying the factors with that showed the most promising results. These factors, Education, Healthcare and Public Services were analysed in turn with the use of new supplementary data. The findings of this phase

Table 8.4: A Table Summarising the Results Identified Through Composite Analysis

Factor	**Result**
Education	The Bronx had the highest student drop out rate, and additionally the lowest average household income.
Healthcare	Brooklyn had the smallest number of mental health facilities per person, indicating that residents had less accessibility to seeing a mental health professional in their neighbourhood.
Public Services	Manhattan had the largest number of fire incidents per resident that the other boroughs, however it did not have the largest number of FNDY previsions. This suggests that citizens living in Manhattan were more at risk of fire and related emergencies.

described areas were some citizens had lower accessibility to services than their counterparts residing in different boroughs, which suggests that there is some inequity in the allocation of resources around the city. A summary of these results produced by the system are described in Table 8.4. In order to utilise these results, it was necessary to evaluate their standing against further contextual research and possible limitations. This assessment would highlight whether specific changes to governing policies could be recommended to improve the equity of public resource provisions across the city.

CHAPTER 9

Reflection and Recommendations

The aim of the analysis undertaken in this project was to gain insight into the equity of resource allocation across New York City. This equity allowed an assessment to be made in determining whether all citizens of the city had the same accessibility to a variety of state funded factors, namely Education, Environment, Healthcare, Transport and Public Services. From the initial analysis in Chapter 7, visualisations were created that raised questions from nuances displayed in the data. These questions were then further explored through fusing multiple datasets in the Composite Analysis phase. This system provided a set of results which allowed recommendations to be made based on creating equal access to these factors for all citizens across the five boroughs of the city. The following section will discuss the findings of the project which led to the formation of a set of recommendations. Finally, due consideration will be given to the restrictions that would limit these suggestions given the context of the problem in question.

9.1 Discussion

The system developed throughout the project provided a variety of results that measured how equitable resources were allocated across New York City. The questions raised during the Single Factor Analysis phase were furthered explored by combining similar supplementary datasets to gain a deeper understanding of issues surrounding Education, Healthcare and Public Services. This holistic approach identified results that supported a variety of hypotheses. For example, it was found that a student in the Bronx was more likely to drop out of their education before reaching graduation, based on historical data for cohorts of students between 2001 and 2005. By combining this data with the average household income, a pattern emerged that showed not only did the Bronx have the highest student drop out rates, but also the lowest household income. This indicates that there may be some correlation between poverty and retainment in schools.

The result produced by the system discovered that areas of poverty caused an increase in the drop out rate amongst students. Research confirmed this hypothesis, further identifying that one in ten students living in the south of the Bronx were homeless [80]. The impact of poverty

on education levels was quantified by studies showing that students facing homelessness are four times more likely to drop out of school. In addition, low income has been correlated to high levels of absenteeism, meaning that children who are economically disadvantaged are more likely to miss school due to poor attendance. However, the figures of high poverty levels in the Bronx are shown in other areas of the city, with an Independent Budget Office report showing a 63 percent rise in student homelessness in 2013. Efforts have been made to curb this disadvantage, with Mayor Bill de Blasio announcing an educational reform as one of his major policies during his election campaign in 2014. Recent data suggests these new strategies are improving the educational outlook for students across the city, with the drop out rate falling to lower than ever before, 8.5%, with the highest increase in graduation rate in the Bronx [22, 44, 23, 62]. This is supported by the results identified from the system.

Public opinion highlighted that citizens believed another area in need of reform, in addition to education was was mental health. This was an area of inequity that was successfully identified by the system. The results highlighted that citizens in Brooklyn had dramatically less accessibility to mental health facilities in their own borough. A current initiative, Thrive NYC, issued by the Department of Health and Mental Hygiene described that '*at least one in five adult New Yorkers suffer from depression, substance abuse, suicidal thoughts or other psychological disorders every day*'. The accompanying report outlined a road map to improve mental health provisions and support the wellbeing of residents through six guiding principles; changing culture surrounding mental health discussions, acting early to offer initiatives in schools, closing treatment gaps, partnering with communities, using better data to identify issues and strengthening the governments responsibility to coordinate and support the mental health of all citizens. This road map was part of a mental health initiative to improve the current issues faced by citizens. The 150-day update summarised the achievements Thrive NYC has made since it came into fruition in 2015, which highlighted positive outcomes such as an additional 2,300 citizens trained to identify signs and symptoms of mental illness. This initiative contributed to a step forward in advocating better mental health practises for all citizens of New York City [20, 31, 71, 32].

Results into the fire provisions across the city identified that citizens in Manhattan had considerably more fire incidents than any other borough. This was correlated to a small number of fire stations, suggesting that Manhattan residents were more at risk than citizens residing in other boroughs. Research into the locality of these fire stations has shown that *'the present number and arrangement of fire companies in most cities are based more on historical facts, such as where volunteer companies were first organised, than on a careful analysis of actual needs'* [84]. This indicates that due to the raising population of inner city residential properties coupled with slow degradation of building quality has led downtown Manhattan to show a disproportional high number of fire incidents, compared to other locations around the city. A number of parametric models have been proposed to suggest better allocations of fire companies to neighbours to improve the equity for citizens in this factor [77, 83].

The previous discussion allowed reflection on the results produced by the system. This required research into additional contextual material, to highlight whether there were current strategies in place that had recognised this disadvantage, and what was currently being done to improve it.

9.2 Recommendations

The discussion affirmed the project's findings with contextual information about current government policies. These recommendations were based on the analysis produced by the system thorough the Single Factor Analysis and Composite Analysis phase. Under the circumstances discussed throughout the project, the following recommendations were made.

- Current efforts in educational reform should be encouraged, with a view to identify and prioritise children from a low income background.
- In keeping with the ethos of the Thrive NYC programme, there should be more mental health facilities for citizens in suburban areas, particularly Brooklyn which has substantially less access per person.
- There should be more fire stations situated in Manhattan, as analysis has show residents are more at risk of fire from historical data.

9.3 Limitations

As with any research project, there were certain limitations to the results provided by the system. It was responsible practise to acknowledge these limitations to understand how much they impacted on the insight that was gained. The rest of this chapter will discuss each of these limitations in turn, highlighting possible mitigations that could be undertaken with future work.

Data Integrity
The first limitation was a systematic problem of the project, which questioned the genuineness of the data. This was the most important limitation, as the data was used as input to the system to provide results. If the data was not genuine and reflective of current factors in the city, results would be screwed to falsely show incorrect patterns that didn't exist in real life. This limitation was acknowledged in the inception of the project and efforts were made to only collect data from official sources. The government managed data repository NYC Open Data offered protection against possible data genuity issues as it had previously been pre-processed by agent officials.

Historic Data
The project aimed to build up a picture of the socioeconomic landscape of New York City using a variety of data sources over different timeframes. This lead to results that identified areas that were particularly more disadvantaged than others. A restraint from using this method of working is that as datasets are ranging over a ten year period, issues that are identified could have already been recognised by local law makers and new policies could have been introduced to improve them. This would suggest that re-identifying these areas again would't be necessary, however this is based on an assumption that all problems of inequity are recognised by the government in the first instance. Additionally, adding to the discussion was an important part of the project meaning that recommendations were not only limited to finding novel results, but also contributing to existing known issues such as mental health, as highlighted in the Discussion section.

Software Limitations

The project utilised a graphing library, Plotly, to create dynamic visualisations. These visualisations were inferred to identify clusters of abnormal behaviour, for example a group of calls complaining about building quality in the east of the the Bronx. This process of human rational was facilitated by the graphing library, assuming that the visualisations it presented were an accurate representation of the underlying data. However, consideration must be made to the possibility that Plotly graphs had some limitations. For example in the 311 data analysis, the colours on the scatter plot point depicted what agency the complaint was in reference to. It could have been possible that these colours overlapped each other and could be hiding additional results underneath. An opacity value was set to 0.5 to attempt to mitigate this possibility, however due to the large size the possibility of overlapping was still present. It was decided that this particular limitation wouldn't have a large affect on the results, as the scatter plots were used to gain a broad understanding of each factor with later further analysis looking at particularities.

The current chapter has discussed the results of the project, which have led to a set of recommendations being defined. These recommendations aimed to redistribute the allocation of public resources around New York City to improve accessibility for all citizens. These results were discussed in light of their context in the city, and possible limitations. The following document will analyse the results of testing the system, and evaluate the success of the project.

CHAPTER 10

Testing

Proper insight into the datasets could only gained if all components of system worked as expected. The hybrid nature of the project necessitated assurance from the testing phase, as the results produced by the system were used to make civic recommendations and needed to be accurate. To ensure this, it was necessary to test the code strategically and robustly. The following chapter will discuss various testing methods applied to the system, such as unit, integration and system testing.

10.1 Unit Testing

The nature of the system was appropriate for thorough unit testing due to its compartmentalised structure. Focusing the unit tests on small isolated pieces of code ensured that the system was working as intended and returning the required outputs. Ensuring that these outputs were as expected allowed the functionality of the system to be observed. The requirements of this functionality were described in Chapter 3, which structured the systematic unit tests performed on each script in the system. An example such as empty variables caused by issues reading in the data could be identified using unit tests that asserted the status of each variable. These problems may have not caused explicit errors during development but would have hindered the collection of results. The three most important areas of the system were focused on during unit testing, as highlighted by the predefined project requirements. These areas were data importing, manipulation and visualisation. Tables 10.1, 10.2 and 10.3 describe a summary of the unit tests carried out, highlighting the input, output and expected output. Each unit test in the table is referenced to one of the original system requirements and indicates whether the system passed or failed the test. This process ensured high quality and sound code.

10.2 Integration Testing

The successful completion of unit tests ensured that the individual components of the system were performing as expected. To maintain confidence that these components functioned prop-

Figure 10.1: A Sample of Data Import Unit Tests

F#	**Description**	**Input**	**Expected Output**	**Actual Output**	**Result**
F1	The system should read in the data from CSV format	Dataset in CSV format	A variable containing the dataset	A pandas data frame object	Pass
F1	The dataset should contain values	A variable containing the dataset	The variable to not be equal to null	The variable was not equal to null	Pass
F1	The shape of the data frame should be the same as the original data	Pandas data frame object and the original dataset	Both datasets to be equivalent in shape	Both datasets were equivalent in shape	Pass
F1	Latitude and longitude attributes should be of type long	A true assertion that the type of latitude and longitude variables are of type long	The truth assertion to return true	A true assertion was returned	Pass
F1	All datasets representing a factor should have latitude and longitude attributes	Truth assertion that a variety of datasets have latitude and longitude variables	The truth assertion to return true	A true assertion was returned	Pass
F1	The latitude and longitude attributes to not be null	Datasets containing latitude and longitude coordinates	No null values to be found	No null values were found	Pass

Figure 10.2: A Sample of Data Manipulation Unit Tests

F#	**Description**	**Input**	**Expected Output**	**Actual Output**	**Result**
F2	The model should contain latitude and longitude coordinates	A sample model for the healthcare factor	The model to contain latitude and longitude coordinates	The model did contain latitude and longitude coordinates	Pass
F2	The model should provide an attribute to be plotted	A sample model for the healthcare factor	The model to contain an attribute that could be plotted	The model contained the complaint type which allowed plotting	Pass
F4	The model should produce statistical analysis	A sample model for the healthcare factor	The model to produce an output	The model correctly outputted statistical result	Pass
F2	The model should combine multiple datasets to produce a result	Two sample datasets	The model to not throw errors	The model successfully utilised the two datasets	Pass
F4	The model should provide accurate results	The results by the system that normalised datasets by population density	The correct values for normalising by population density	The two values were equal	Pass

Figure 10.3: A Sample of Data Visualisation Unit Tests

F#	**Description**	**Input**	**Expected Output**	**Actual Output**	**Result**
F3	The system outputted a data visualisation	A sample dataset	A data visualisation	A data visualisation was produced	Pass
F3	The data visualisation was dynamic	A sample dataset	A dynamic data visualisation that could be manipulated by a user	A dynamic data visualisation was observed that allowed scaling and hover labels	Pass
F3	The data visualisation should be plotted on a map	A sample dataset	A data visualisation plotted on a map	The resulting visualisation was plotted with a base map	Pass
F3	The data visualisations should render	A sample dataset that included 10,000 points	A data visualisation of these points	A visualisation was produced, even with the large number of points	Pass
F3	The data visualisation to identify clusters of points	A dataset with similar values	A visualisation that illustrated how these points were clustered together	A visualisation that displayed the points on a scatter chart, highlighting the cluster of points	Pass
F3	The data visualisation to identify anomalies in the data set	A dataset with points outside of New York City	A visualisation that highlighted points that were outside of the city bounds	The resulting visualisation illustrated points that were outside of New York City	Pass

Table 10.1: A Sample of Integration Tests

Factor	**Investigation**	**Errors Thrown?**	**Result Provided?**	**Evaluation**
Education	Drop out rate per borough	None	Yes	Pass
Education	Drop out rate per year	None	Yes	Pass
Environment	Locality of 311 calls	None	Yes	Pass
Environment	311 complaint frequency	None	Yes	Pass
Environment	Proportionality of 311 complaints	None	Yes	Pass
Healthcare	Locality of HCC facilities	None	Yes	Pass
Healthcare	Locality of mental health provisions	None	Yes	Pass
Transport	Patterns in pick up location	None	Yes	Pass
Transport	Patterns in drop off location	None	Yes	Pass
Public Services	Locality in emergency response incidents	None	Yes	Pass
Public Service	Locality in fire emergencies	None	Yes	Pass
Public Service	Locality in fire company boundaries	None	Yes	Pass

erly in the whole system, it was necessary to perform integration tests. The aim of these tests were to ensure the pipeline of functions that started from the raw data to the visualisation and analytical output was working properly to derive meaningful and accurate results. These integration tests were performed on each analytical script that provided a result of the system. The results in Table 10.1 shows an overview of the tests performed to ensure scripts ran properly and produced results.

10.3 System Testing

The holistic nature of the system resulted in analytical measures that provided insight into the state of resource prevision in New York City. Therefore, the requirements of the system differed substantially from a typical software development project. To ensure the system was functioning, the outputted results were assessed against further research to validate the accuracy of findings. The discussion of these results is provided in more detail in Chapter 9, however to formalise this process for testing it was necessary to create a systematic check to certify results were reliable. The results in Tables 10.4 and 10.5 describes these tests by referencing analysis that was identified at the end of the Single Factor Analysis chapter. For each of these analyses,

the result from the system is described and in addition to evidence using real world context. This evidence will either supports or contends it the result from the system. This allowed an evaluation to be made that assessed whether the system passed the test.

As the developed system was not common of a typical development project, a different approach to testing was taken as discussed in this chapter. The code portions were easily unit and integration tested due to their modular structure, however the results of the system could only be tested using wider knowledge of the project context. This universalistic approach was novel, as it different from the typical concrete approach to system testing. However, without identifying evidence to back up the claims of the results, it was difficult to know if the recommendations were accurate and relevant to the current sociological landscape of New York City. Concluding the testing phase determined that the results were relevant, giving weight to the recommendations made based on them. As the project had now evaluated the developed system, it was necessary to critique the project management and methodological decisions that managed the work flow.

Figure 10.4: System Testing I

Analysis	**Description of Analysis Provided by System**	**Real World Context and Evidence**	**Evaluation**
A1	The number of students dropping out of school between 2001 and 2005 decreased	Educational reform has been a chief concern in the Mayor's office for the last decade. Policy changes to help under performing schools have been implemented that has seen a rise in education attainment.	The system correctly observed the city wide reduction in drop out rates. However, as this information was not utilised in the Composite Analysis phase, no conclusion was made as to what specific factors had influenced this reduction.
A2	The Bronx drop out rates are higher than any other borough and also the lowest average household income. This indicates that there is a relationship between poverty and education standard.	The Bronx has the highest level of child poverty resulting in the lowest number of students reaching graduation across all five boroughs.	The contextual information provides evidence that this was a successful observation from system.
A3	There was a higher number of complaints to the HPD in the Bronx	Researched showed that the Bronx was one of the largest residential areas in the city with the lowest average household income.	The system did not identify what the cause of these complaints was, as it was deemed a low priority result and no further analysis was undertaken.
A4	There were an abnormally large cluster of calls to the FDNY in downtown Manhattan	Research showed that there were fire service cuts in 192 to 1978 which moved many fire fighting provisions out of the the centre of the city. Since these cuts, matters of city ecology has changed with the increase of housing overcrowding and deterioration of quality. It has been suggested that the current boundaries of fire companies do not serve the city to their maximum potential.	The Composite Analysis phase showed that there were disproportionately fewer fire fighting resources in Manhattan than there was in the other four boroughs, leading to inequity in fire control provisions. This observation was based in evidence and therefore successfully identified by the system.

Figure 10.5: SystemTesting II

Analysis	**Description of Analysis Provided by System**	**Real World Context and Evidence**	**Evaluation**
A5	Mental health provisions were less accessible for citizens who lived in more suburban areas of the city	Investing in mental health provision is a major policy of the current government of the city. It has been recognised as a growing problem to citizens of the city due to lack of services, which a new initiative aims to fix.	The need to create a new initiative describes how lacking mental health facilities was a problem. This meant that the system had correctly identified it as an area where resources in the city were not fairly accessible to all citizens.
A6	There were more taxi trips taken out of the city compared to those taken into the city	It was noted that subway lines run 24 hours a day, possibly indicating a discontentment with the service late at night. Additionally, preliminary research identified that the rise in private taxi companies such as Uber had effected the number of people taking yellow taxi cabs.	As this topic was not furthered after the Single Factor analysis phase, it did not result in conducive evidence that the original observation was correct.
A7	Manhattan citizens had a higher number of emergency service calls than any other borough	As discussed in A4 and A5, it was discovered that there was evidence to support lacking provisions for firefighting provisions and hospital access.	This evidence confirms the results of the system, as it correctly identified that residents in Manhattan had a higher rate of emergency situations with a lower accessibility to fire fighting provisions and hospitals.
A8	There was not equitable access to firefighting provisions.	Again, this result was highlighted by the system many times and the previous research highlighted in A4 and A7 provided research that suggested fire stations were not located in the centre of the city.	The system correctly identified the locality of fire controls was not equitable based on population statistics across the city.

CHAPTER 11

Project Management

The project was undertaken over a period of nine months, using a data-driven approach to achieve results that contributed to the discussion of resource equity within New York City. A complexity of this project resulted in the management of a multitude of different phases highlighted in the Methodology chapter, ranging from the breath of research to the analysis of a variety of factors. The process of managing this work was complex, and required stringent project management practices to ensure time was being utilised effectively. Once a methodology was defined, management tools were identified to ensure the projects progress would not be hindered by lack of organisation. By utilising these tools, both the project leader and the supervisor were able to monitor the pace of work, and adjust the direction of focus to match the results identified in each phase of the project.

The following section will discuss the design approaches that were taken and explore how the methodology was devised. The software methodology that was utilised during the development of system will be described, with practical examples of how it was undertaken. The chapter will also cover how the project timeline has evolved through the course of the project, and explain the management tools that were necessary in producing high quality work. Finally, the risks of the project are discussed with comments on how they were mitigated to ensure best practise was adhered to.

11.1 Design Approach

Due to the breath of investigation in the project, a structured well-organised methodology was necessary to ensure progress was on track for completion. As explained in the methodology section, five phases of the project were defined. These phases were Research, Data Ingest, Single Factor Analysis, Composite Analysis and Recommendations. Breaking the work down into small measurable goals allowed tasks completed and reviewed with the project supervisor. This kept up a timely pace and ensured all deadlines were met.

The analytical element of this project could not have been understood without the extensive research component during the projects inception. Because of this, it was integral to structure the work with consideration to exploring the data throughout the Research, Data Ingest and Single Factor Analysis phases. In this way, it was decided that an agile methodology would be utilised to update goals as the data was investigated. This allowed the methodology to be flexible in focusing on a broad range of factors, such as those identified in the research section. It was unknown when the project was underway as to which factor would yield successful results, so this flexibility allowed the direction of the project to focus on where the most interesting discoveries were found.

The success of the project hinged on the analytical component. Because of this, a proportion of the research time was allocated to investigating the numerous packages and scripts available for data analytics. It was necessary that the chosen software was low-cost or free, could manage large data sets, and produce high quality visualisations. Utilising geographical visualisations allowed the data to be assessed by borough which indicated areas that needed further research.

11.2 Software Development Methodology

As previously discussed, it was important that the software development methodology was agile to support the changing demands of the project. This was due to the decision to analyse a broad range of factors, knowing that not all of them would be fruitful in providing results. The flexibility that an agile methodology offered reduced the risk of this situation becoming problematic, as whatever factor was offering more promising results could become to focus of further development.

In carrying out the software development, an agile methodology was followed by attending regular weekly meeting between the project supervisor and the developer. This encouraged discussion of the results so far, and gave indication of what direction should be followed in further work. This continual feedback ensured the project met its initial aims by creating smaller measurable subgoals that were easier to maintain with a single developer.

11.3 Project Timeline

The original project timeline shown in Figure 11.1 was devised in the Project Specification, before a concrete methodology was decided on. Because of this, it displays different phases compared to the ones explained in Chapter 5. When identifying the breadth of research and analysis that was to be undertaken, the methodology was adapted to the one explained in this document. The revised project timeline is shown in Figure 11.2. Here, the new phases of work are highlighted, with the expected duration in blue. The red cells highlight where the work was extended as work had not been fully completed. The areas that needed more time were mostly during the research and analytics phases, due to sizeable amount of data that was necessary to maintain the projects broad scope.

A task thats time requirement was underestimated was data-preprocessing, during the Data Ingest phase. As the data came from many different agencies within the Mayor's office, it often

had multiple types of geographic references, such as address, Block-Borough-Lot code or latitude and longitude coordinates. To ensure that all data was following the same standard, a lengthly amount of time was taken to convert datasets attributes to the right type. However due to the agile nature of the project, work was maintained by focusing on different factors in parallel, so as one dataset was undergoing pre-processing another could be explored and visualised. This workflow ensured that the project was still completed on time.

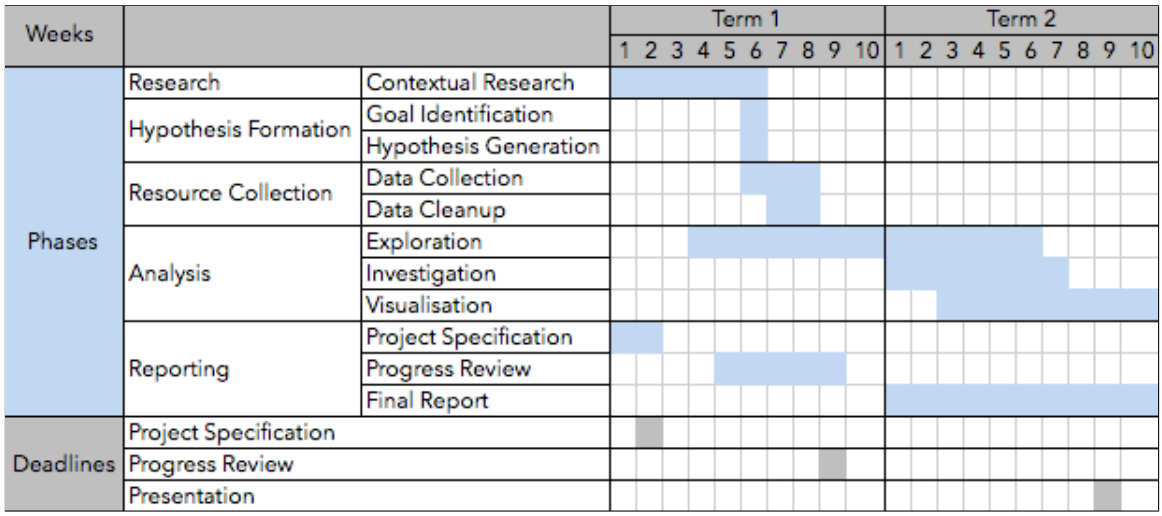


Figure 11.1: Original Timeline of Workload

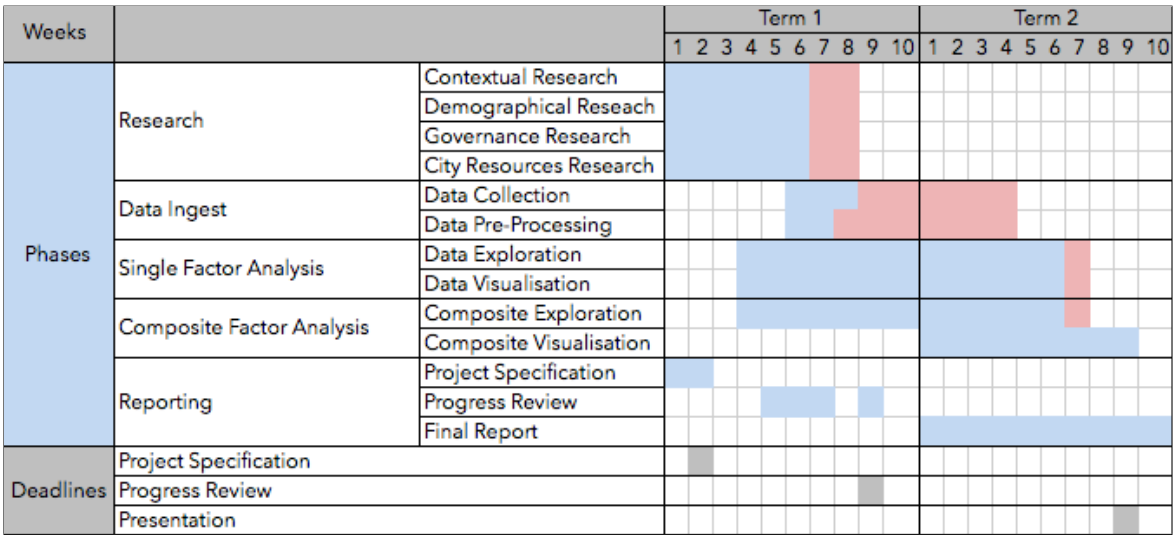


Figure 11.2: Revised Timeline of Workload

11.4 Development and Management Tools

As project was undertaken over a period of nine months, there was great importance in managing both time and resources effectively. There were large organisational demands when managing

tens of datasets, meaning it was imperative that a system was created and adhered to. The following tools were a huge benefit to the projects progression, allowing time to be spend on improving the algorithms for analysis rather than managing data.

11.4.1 Development Tools

The following development tools were utilised during the project to analyse datasets. These tools were imperative for gaining insight into how equitably resources were allocated in New York City.

Python
The only programming language used in the project was Python. This was due to its simple grammar coupled with its power in data analytics. The simplicity of writing Python scripts was necessary due to the breath of datasets gathered, all of which needed individual pre-processing and analysis. Another benefit of Python is its strong mathematical packages, such as NumPy and SciPy, which were used during the data analytics phase of the project. Python version 2.7 was used in all development.

Anaconda
Anaconda, a Python package manager was heavily utilised during the project. By downloading Anaconda many popular packages such as the aforementioned NumPy and SciPy came preinstalled and dependencies were managed. Anaconda removed the need to set up an a specific environment for the project, which reduced the set up time needed before development.

Jupyter Notebook
Another benefit of the Anaconda package manager was the preinstalled Juypter Notebook. This was chosen as the main editor because it offered the ability to run scripts in short segments, holding variables and data in memory. This functionality was extremely useful in trying different visualisation techniques to understand what provided the best information about the data. However, in some cases the memory of Jupyter Notebook was exceeded by the size of the datasets. In these instances Sublime Text was used to write scripts that split the data into sections that were small enough to then read into Jupyter Notebook.

Plotly
After the analysis discussed in the research section, Plotly was utilised as the main API to produce data visualisations. With any new API, there was a steep learning curve to understand how to build useful analytics. To aid this process, the Plotly tutorials were referenced extensively to gain the knowledge that allowed to project to continue.

Google Geocoder
Another API that was important to the progression of the project was Google Geocoder. This allowed the Google Maps server to be queried with an address and would return a latitude and longitude coordinate pair. Due to the dependency on geographical data, it was important in the standardisation process to ensure that all data had latitude and longitude reference points.

Figure 11.3: Development Tools used Throughout the Project

Figure 11.4: Management Tools used Throughout the Project

11.4.2 Management Tools

To ensure the project was managed effectively and efficiently, the use of the following tools was employed. The following section will describe the functionality of each tool and how this was utilised in the project.

Dropbox
To maintain the breadth of analysis following the identification of the single factors, a number of datasets were collected. These datasets ranged in size from 50MB to 15GB. Therefore, it was imperative to effectively manage the large quantity of data as it could not be held locally at one time. Dropbox allowed the data to be stored on the cloud which meant the computers local memory was freed up. A system was devised where a dataset would be uploaded to Dropbox after it was pre-processed, and subsequently downloaded when it was required for analysis. This system also required an organised file system to be implemented to maintained to ensure time was not lost looking for files.

Apple Notes
During the lifetime of the project, there were many directions discussed between the project supervisor and the developer. To ensure these discussions were documented, extensive notes were taken in each meeting using the Apple Notes application. These notes were crucial due to the agile nature of the project, and allowed milestones to be set and reviewed. The meeting notes also served as a to-do list for the developer, outlining the outstanding tasks to be accomplished before the next progress review with the supervisor.

Email
As the importance of communication has been highlighted, it was necessary to stay in touch with the project supervisor even when a meeting in person was possible. In these cases, email was utilised to maintain the weekly meetings to ensure the project progressed at an appropriate pace. This mitigated the risk of the project becoming stagnant over vacation periods such as Christmas and Easter, and ensured the project timeline was adhered to.

11.5 Risk Management

The success of the project hinged on work being completed to a high standard in a timely manner. Due to the length and breadth of the project, it was therefore necessary to assess any risks which could lead to potential problems throughout. Each risk that was identified was given a severity rating; high, medium or low. This rating would indicate the potential impact the scenario would have to the projects timeline and therefore completion. To understand how likely each risk was, a probability was also given. This information was utilised by making decisions on how to mitigate potential risks. Information about the risk management process are displayed in Table 11.5.

The management of the project was crucial in defining its success. Without intelligent and structured organisational practises, results would not have been yield. This was primarily due to the number of moving parts within the system, ranging from the analysis of five factors that each possessed multiple datasets and two phases of development. This complexity required thorough organisational skill to undertake research, development, analysis and testing in each iterative cycle. To assess the success of the project management, the evaluation phase analysed areas that worked well whilst also indicating additional methods that could have improved current processes.

Figure 11.5: Risk Identification and Proposed Mitigations

Risk	**Description**	**Severity**	**Probability**	**Mitigation**
Hardware failure	During the development phase, a single machine was used to create all the analytical scripts required. Because of this, there was a risk that work could be lost due to computational failure.	High	Low	All code was backed up utilising cloud storage. As the scripts were short and usually only used once, it was not necessary to use git repositories. Instead, the scrips were saved to Dropbox to ensure work was always accessible.
Single developer	The project was being led and developed by an individual, which posed potential risks if illness occurred and work could not be completed.	High	Medium	The agile approach chosen offered the flexibility that was necessary with regards to a single developer. Through regular meetings, the progress could be tracked and subgoals could be defined. In the instance of illness, the subgoals could be reduced or paused until a time when the project leader had recovered.
Legal, social, ethical and professional issues	Due to the nature of the project, the data contained many sensitive topics such as gender, ethnicity and race. These topics could cause offence or discrimination if used inappropriately.	Medium	Low	By referencing the British Computing Societies Code of Conduction, best practise guidelines were adhered to in the collection and processing of data. The data collected was anonymised, and no attempted to exploit individuals or groups of people was made.
Inadequate data collected	The project hinged on the successful collection of data. Without this, analysis could have not taken place and the insight into the equity of resource allocation in New York City would not have been identified.	High	Low	When researching the project it was identified that NYC Open Data was the official governmental data repository. Most datasets in this repository used location as an identifying factor, which was used heavily during the project.

CHAPTER 12

Evaluation

It was possible to evaluate the project once the findings had been reported and recommendations given. In this chapter the evaluation will be discussed, allowed the strengths of the project to be noted whist identifying areas that could be improved if further work was undertaken. This evaluation was formed by primarily assessing whether the functional and non-functional requirements had been met, providing a qualitative measure to define the success of the system. Subsequently, the methodology of the project was also evaluated, by critiquing management techniques and approaches. This comprehensive discussion will be summarised by author's assessment of the project, answering a variety of questions that highlight both achievements and limitations encountered.

12.1 Functional Evaluation

The system was built to satisfy the functional requirements defined in Chapter 4. These requirements highlighted the important features of the system, such as the inputs and expected outputs. In this project, it was imperative that the system took a variety of datasets as an input, and outputted data visualisations and analytical results. By ensuring the behaviour of the system following these functional requirements allowed the project achieved the aims highlighted in Chapter 1. The results in Table 12.1 reiterated the original functional requirements described, and discusses whether the system met those demands. In the case of functional requirements, all the needs were met within the project timeframe. Additionally, optional requirements were defined that would stretch the capabilities of the system if the functional requirements were fulfilled before the end of the development lifecycle. These optional requirements are shown in Table 12.2. Fortunately, one of the three optional requirements was completed, which added improvement to the overall system.

Table 12.1: An Evaluation of the Function Requirements of the System

	Functional Requirement	**Evaluation**	**Outcome**
F1	The system should accept csv datasets as an input.	Using the Pandas framework, the csv files were read into the system using the `read_csv()` method. There were then stored as Pandas data frames which could be sorted, searched and sliced.	Pass
F2	The system should create a model based on the inputted datasets.	The data frame that contained the attributes for analysis was passed into a mathematical model that normalised values and plotted a visualisation. This model was then reused for each individual dataset to ensure a consistent standard across results.	Pass
F3	The system should produce graphical visualisations of data using mapping software.	The mapping software was chosen to be the graphing library Plotly, as described in the Platform Research section. The computational model created a visualisation for each unique dataset that was read into the system.	Pass
F4	The system should produce results that provide recommendations	The system used visualisations and analysis to output potential problem areas where equity of public resources did not seem fair. These areas were then thoroughly researched to provide recommendations that could improve currently policies.	Pass
F5	The system should be comprehensible to a data scientist	The system utilised a variety of technical components, chiefly Python for writing scrips, Pandas for storing and analysing data, and Plotly for creating dynamic visualisations. The code was written in a manner that followed software engineering guidelines to ensure it could be easily understood by a technical audience.	Pass

Table 12.2: An Evaluation of the Optional Functional Requirements of the System

	Optional Requirement	**Evaluation**	**Outcome**
OF1	The system should combine data from a range of agencies	The system utilised data from NYC Open Data, where data from a variety of agencies were stored. Methods were written to normalise agency specific attributes, such as location. This ensured that a variety of data could be used in analysis.	Pass
OF2	The system should utilise social media data	The integration of social media sentiment analysis was trialled by mining data from Twitter, however there wasn't enough time to integrate it with the current model. Future improvement of the work would allow this to enrich the breadth of data investigated.	Fail
OF3	The system should make predictions of how policy change would effect conditions	Again, as time was limited in the project the goal of combining recommendations with a prediction model was not accomplished due to the amount of time it would have required. Instead the recommendations were analysed using a variety of contextual research. If the timeframe of the work was extended, a predictive system could allow accuracy of recommendations to be measured.	Fail

12.2 Non-Functional Evaluation

In addition to the functional requirements of the system, the initial specification of the project defined non-functional requirements to judge the operation of the system. This allowed the appropriateness of particular attributes of the system to be measured, such as testing and extensibility. Table 12.3 documents the original non-functional requirements proposed and an describes the evaluation of each one. The project met all of the non-functional requirements specified.

Table 12.3: An Evaluation of the Non-Function Requirements of the System

	Requirement	**Evaluation**	**Outcome**
NF1	The system should follow the licensing agreements of open source data	The research ensured that open source data laws were acknowledged in Chapter 3, and that these rules were not broken by any parts of the system. This included not utilising the developed software for criminal purposes, to commit a crime, or identify any individuals.	Pass
NF2	The system should be maintainable	The system was developed in a modular way to ensure that work could be maintained or extended if necessary. This meant following current software development principles were adhered to, such as using clear method names, writing comments and following exhaustive testing practices. This would allow another developer to utilise the system in the future.	Pass
NF3	The system should be testable	As described in Chapter 11, extensive testing of the system was undertaken through unit testing, integration testing, system testing and user acceptance testing. The system was built following test driven development practices to ensure code was of high quality throughout.	Pass
NF4	The system should be extendable to alternative data sources	Due to the flexibility of the Pandas data frame work, the system can take any csv file as input. This allows extendibility to alternative data sources outside of the NYC Open Data portal.	Pass

12.3 Legal, Social, Ethical and Professional Evaluation

In addition to the requirements discussed above, a number of legal, social, ethical and professional issues were highlighted in Chapter 3. It was important that the developed system honoured these issues to respect the integrity of the data used throughout the project. When evaluating the success of the project, it was also necessary to evaluate how work was carried out in regards to each issue.

12.3.1 Legal Issues

All the data sourced in this project originated from the online NYC Open Data portal. It was important that the use of this data abided by the terms of use set out by the New York City Government. These terms specified that the data could not be used to commit a criminal offence or encourage others to do so. This was ensured by detailing the aims of the project before data collection, and only utilising data that was relevant to the factors being observed. It was important that the use of the collected data was transparent, and could not be utilised

to impersonate any agency parties or utilise municipal information in a deceitful way.

The technical component also adhered to the legal requirements of API's used throughout the project. The graphing library Plolty was used under the the student license, which provided the full service at a low cost. It was important to note that any visualisations created using the Plotly software gave the company full rights to use and reproduce created material. As the analysis undertaken was created to provide insight and was not generating revenue, these terms of use were deemed acceptable.

12.3.2 Social Issues

As the project was centred around social issues faced by citizens in New York City, it was important to ensure all findings were representative of all people in the city. This was of particular importance when discussing the recommendations found in Chapter 8. The measure of resource equity in the project was defined as accessibility to state provided facilities for all citizens across each of the five boroughs. In doing so, the assumption was made that citizens in suburban areas should have the same access to resources as those who live in the centre of the city, deeming that all citizens had equal rights. This assumption ensured that the project did not have a political agenda, and that any findings were not biased to particular government parties. To ensure there was a balance of opinions, a range of factual sources were used during the discussion of results to maintain the use of sound facts rather than opinions. In all cases, references were given to the original material.

12.3.3 Ethical Issues

The project had the potential to touch on ethical issues throughout the data collection and analysis. It was important to ensure that the single factors for analysis chosen were representative of issues that effected a diverse group of people. With this in mind, the chosen factors were areas that affected all citizens of New York City, for example healthcare and education. By ensuring that the project maintained transparency throughout data collection, it avoided possible conflict of ethical issues. The use of open source data also protected the project against issues surrounding anonymity of data, as all due diligence was taken before the datasets were uploaded to the internet. This made it difficult for the system to identify individuals or groups of people, and no attempt was made by the software to extrapolate this information.

12.3.4 Professional Issues

To ensure a high standard of professionalism thorough the project, institutes such as the BCS were researched for guidelines on software development practises. This guidelines were abided throughout development by to ensure the code was created to a high standard. It was also necessary to respect the third-party services utilised throughout the project, such as the Google Geocoder API. This product allowed the translation between addresses and latitude-longitude coordinate pairs, however had strict limits on usage policies. For example, it was stated that a call could only be made to the server once every two minutes, to stop potential malicious attacks. To respect these boundaries, code was created that employed a wait clause in order to maintain the limits set by the Geocoder terms of service. However, this wait clause lead the programs run time to be over 72 hours, so a cache was utilised incase of failure. This ensured

that collected results could still be accessed even if the script terminated, and ensured there was no need to send multiple requests of the same data to the API.

12.4 Project Management Evaluation

The project undertaken required efficient management techniques to be employed to ensure that deadlines were met and work was produced to a high quality. Although a full discussion of the project management approach is given in Chapter 11, it was deemed necessary to evaluate these standards to identify areas that could be improved with future work. The rest of this section will give an appraisal of the organisation of the project.

12.4.1 Approach to Design and Development

The methodology was structured in an organised way, highlighting the six phases of the work. These phases were Research, Data Ingest, Single Factor Analysis, Composite Analysis and Recommendations and Reflection. This formation followed a similar process as a waterfall model, with the intention of each phase beginning only once the previous phase had been completed. However, it was discovered early in the project that this style of working was not efficient. The chief reason for this was because within each phase there were a multitude of factors being investigated, which required different workloads in terms of standardisation and analysis. This made it difficult to see continuity within a factor when work had to be delayed until all factors had completely a certain phase. After this realisation, it was decided that the developmental phases of the project would be iterative, following an agile way of working. This suited the data exploration much better, as it allowed factors to be visualised and analysed in a continuous way, which provided the most insight into the results. It was therefore agreed that an agile approach would be used for the project, which allowed the project leader and supervisor to meet regularly to discuss each iteration of analysis. This style of working ensured that communication of the project direction was strong, and that results could be created in a quick turn around time.

12.4.2 Project Timeline

The original project timeline shown in Figure 11.1 underestimated the time it took to collect and standardise relevant datasets from the NYC Open Data portal. This was due to the differences in data publishing standards between each agency, and therefore a lot of time was taken to ensure that all data contained had the same format for important attributes such as location. This delay caused a knock on effect to the rest of the work as the analysis of a factor could not begin until data had been sourced and standardised. However, the decision follow an agile methodology ensured that this would not hinder the project, and that work could be made up in the timeframe original allocated for analysis. Rather than working on a specific phase in one sitting, the approach was adapted to focus on a particular factor and work on it until completion. This style of working allowed to project to still meet the aims set out in Chapter 1 before the deadline. The revised timetable of work is shown in Figure 11.2.

12.4.3 Tools and Techniques

The project could not have been undertaken without the use of specific management and development tools. Some of these tools were used to a greater extent that others, enabling success in meeting the initial project aims. Each of those tools are evaluated in Table 12.4, with an explanation of how their functionality was utilised.

12.5 Technical Evaluation

The project was centred around creating an system that utilised current data science approaches to analyse a variety of datasets. With the technical component at the heart of the project, it was necessary to evaluate the outcome of the system. This section will discuss the technical challenges that were encountered and how they were overcome.

12.5.1 Technical Challenges

A variety of technical challenges were met throughout the duration of the project which were not anticipated when the system was designed. By maintaining an agile and flexible approach, it was possible to mitigate these issues without them impinging on the progress of the project. The following section explores each one of these challenges and describes how it was overcame.

Data Without Location Reference
There was a risk that the data sourced would be suitable for analytics. This would occur if the data didn't contain any geographical reference, such as latitude and longitude coordinates or addresses. By choosing to use utilise the NYC Open Data repository, access to specific geolocation datasets were provided.

Datasets Too Large to Analyse
Some datasets, such as 311 calls, contained millions of rows due to the high frequency at which calls were taken. This caused problems as visualisation was not possible in such a large dimension. A script was written in Python to sort through large files and break them down into smaller files that were more manageable to load into memory and analyse.

Storage of Large Quantities of Data
Many datasets were needed for analysis and therefore were required to be stored. The quantity of this data exceeded the storage on the local machine used throughout the duration of the project. Utilising cloud storage through applications such as Dropbox ensured that memory space was not a concern. By maintaining consistency in naming conventions and organisation it made it possible to decipher between raw and processed data.

Limitations on API Usage
Research identified that most popular API's have a usage limit. In this project, the Google Geocoder only allowed making one request every two seconds. Any more requests would return null values that would have terminated the script resulting in invalid data. This was mitigated by writing scripts that contained a delay sequence of two seconds between each call. These scripts also cached the results so they could be retrieved if the script crashed.

Table 12.4: Evaluation of the Tools used Throughout the Project

Tool	**Evaluation of Use**
Python	Python was utilised exclusively to write the analytical scripts and therefore used extensively throughout the project. Its ease of use allowed code to be written quickly without needing to create instances of objects. Additionally, it supported the functionality to import a range of libraries that were of use throughout the project, such as Pandas for data storage and NumPy for mathematical analytics.
Anaconda	The Anaconda package was used to ensure that all package dependencies were managed externally, which freed up time in the initial set up of the project. This was a useful tool although not imperative to the projects success.
Jupiter Notebook	Jupiter Notebook was installed alongside Anaconda. It was the main editor for creating analytical scripts as it allowed small chunks of code to be tested whilst maintaining results in memory for later use. Although it was used extensively throughout the project, it was sometimes problematic when datasets were large, causing scripts to run slower than if the they were ran in the terminal. If work was continued, it would only be used for small scale exploration and testing, with IDE such as Sublime being favoured to write scripts.
Plotly	The Plotly graphing library was imperative to the project as it allowed large datasets to be translated into succinct data visualisations. This tool allowed a lot of insight to be gained about the collected data, such as outliers and clusters of points. However, using Plotly did have some constraints due to the size of the data, and would only be recommended for datasets with less that 40,000 points.
Google Geocoder	The Geocoder API was necessary when standardising the location attribute in the sourced datasets. Although it did not have a large part in the project, it was necessary to allow further analysis to continue.
Dropbox	Dropbox was extremely important for data storage due to the large spacial requirements of the project. The necessity of a cloud storage platform was not fully appreciated until the Data Ingest phase of the project, where local storage was not sufficient for carrying out data pre-processing tasks.
Apple Notes	The Notes app provided on Apple computers was utilised to take notes in meetings between the supervisor and the project leader. It allowed the work required in each development iteration to be documented, and served as a to-do list for what tasks needed to be complete. The use of this particular note taking application was not necessary as it only offered basic functionality, however it was preinstalled and easily used.
Email	The use of email was important in vacation times to ensure the progression of the project, however it wasn't a tool that was used heavily as the supervisor and project leader were co-located for the majority of the project.

12.6 Author's Assessment of the Project

To conclude the evaluation of the project, a brief summary of the authors assessment of the project was undertaken. The following questions explore the technical contribution and limitations of the project.

What is the technical contribution of this project?
This project aimed to understand the accessibility of public resource equity across New York City, and therefore measure how equitable these services are to citizens residing in different boroughs. To achieve this, the project required that a system was developed to analyse a broad range of data representing a number of key factors. This system was created by writing a number of analytical scripts that utilised current practices in the data science discipline. In particular, a variety of visualisations were created using the Plotly graphing library. These visualisations were interesting and dynamic, and told a story about the lives of New York residents through emerging patterns and clusters of behaviour. This technical contribution of visualisations and statistics allowed recommendations to be made to rebalance the fairness in current public policy. To gain this understanding, a lengthy process of data collection, standardisation and analysis was necessary, utilising a variety of technical skill and problem solving aptitude.

Why should this contribution be considered relevant and important for the subject of your degree?
My degree in Computer Science has provided me with the skill and understanding to utilise technology for social good. Without the thorough use of data analytics and data management, this project would not have been able to produce results that contribute to a wider discussion of civic equity. The importance of these results are necessary to the wider context of city science, by impacting on policy making, civic planning and resource management. On a high level, contributing to the preexisting research on the science of cities is necessary of the sustainability of cities for future generations, however the results in this context have a deeper immediate effect on the every day lives of thousands of people.

How can others make use of the work in this project?
The work conducted in this project lends itself to a number of applications. In the first instance, the recommendations made can be used to analyse specific areas of public resource equity, such as fire fighting provisions. These recommendations have been made to benefit majority of citizens across New York City. In addition, a number of future work options exist. The broad range of analysis in this project highlights areas that could be further researched in more depth, such as fire fighting previsions or accessibility to mental health facilities. By providing a broad overview that contributes to the wider conversation of public services, the results of the project have many real-world applications.

Why should this project be considered an achievement?
The achievement of the project is shown in acquiring results that indicated there is inequity in resource allocation across New York City, and therefore answering the question initially set out in the project aims. By providing evidence that there are instances where location disadvantages a group of citizens, the project has highlighted possible changes that could be made to mediate this. This is beneficial in contributing to wider research in resource allocation.

What are the limitations of this project?
Due to the nature of the system, analysis cannot be performed without the necessary datasets. Therefore, this could be a limitation of the project if the datasets provided did not contain geographical references. Additionally, the integrity of this data has to be assumed, that it is accurate and reliable in the first instance. As this project only utilised data from a government managed portal, it was assumed that all refections found were genuine issues.

CHAPTER 13

Conclusion

To conclude the work of this project, a summary will provide an overview of the work completed. This chapter will summarise the aims of the project set out in Chapter 1, and describe the actions taken to fulfil them. Additionally, this chapter will provide an outline of how the work could be continued further, with suggestions of additions to improve the analytical system and utilise the recommendations achieved.

13.1 Project Summary

As set out in the project aims, work was carried out to assess the equity of resource allocation in New York City. In this case, equity was seen as all citizens having equal and fair access to public provisions supplied by the government. The project aimed to test whether locality was an influencing factor in this access, by testing whether living in a particular borough lowered the accessibility to resources. If this was the case, it would show that resources were not equitably distributed. To undertake this analysis, the use of modern data science approaches were employed, most notably visualisations. The data for these visualisations was collected over a range of key factors that touched the lives of many people in the city such as education, healthcare and transport. The graphing library Plotly was extensively utilised to create dynamic and visualisations based on this data, to identify emerging trends and patterns. From this analysis, results were gathered. The results showed that there were areas of inequity across the city, highlighting a lack of mental health facilities in Brooklyn and a disproportionate low number of fire fighting provision in Manhattan. These results were discussed in the wider context of the city, allowing a concrete set of recommendations which indicated possible changes in future policy to facilitate a more equal distribution of public resources for all citizens, stemming from results based in evidence.

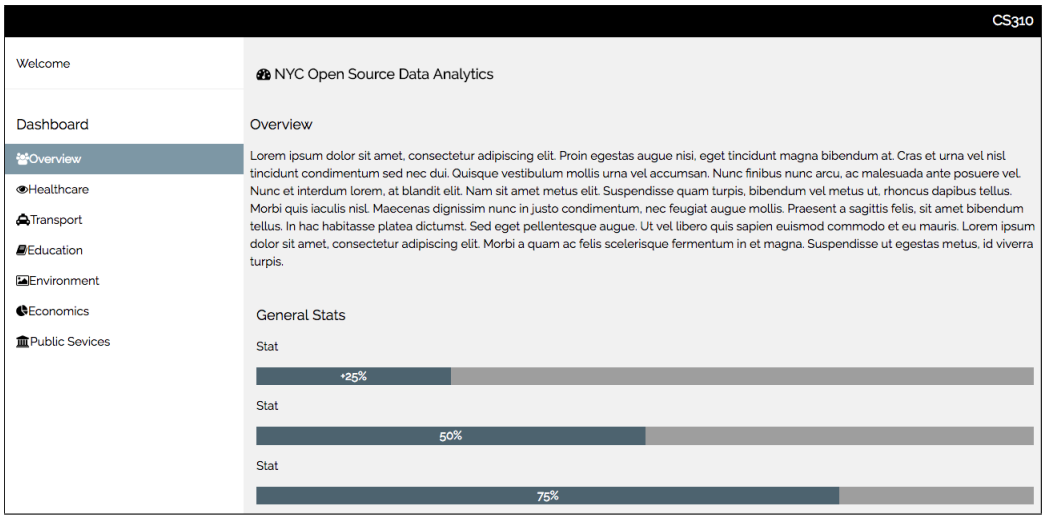


Figure 13.1: A Mock Up of a Proposed Web Application

13.2 Future Work

The work that was undertaken in this project differed from current literature available due to the breadth of factors analysed. This breadth provided a rich picture of the sociological landscape of New York City to be observed, which led interesting results. This work could be used as a substantial base to a variety of projects. The rest of this chapter will provide discussion and ideas for potential future work.

13.2.1 Web Application

A possible future extension of this project would be to present the data visualisations in a web application. Plotly provided the functionality to embed graphs directly into HTML, so producing a simple website that displayed these results would not take much more further work. A mock up of a possible design was created, as shown in Figure 13.1. To extend the complexity and usefulness of this website, visualisations could be responsive to queries. For example, a user may live in New York City and want to search for their zip code to analyse what public resources are available locally. This feature would require the use of Ploty's streaming service, to create new visualisations on the client side. This proposed web application could produce further functionality by giving a score for the equity measure in different locations, as inspired by the Robin Hood Poverty Tracker that monitors disadvantage in New York [34] illustrated in Figures 13.2 and 13.3. Creating this scoring system would entail the need to derivate an expression to quantify equity, that could therefore be applied to different locations over different attributes. Providing a score would further encourage residents to engage with their local government to improve their equity score resulting in a more cohesive community.

13.2.2 Predictive Modeling

A machine learning model could be built to provide more accurate and detailed recommendations into areas that were more disadvantaged than others. This would entail finding a large number of additional datasets with fitting attributes to train a model on the correlation between

Figure 13.2: An Example of a Current Web Application Surrounding Poverty Statistics I

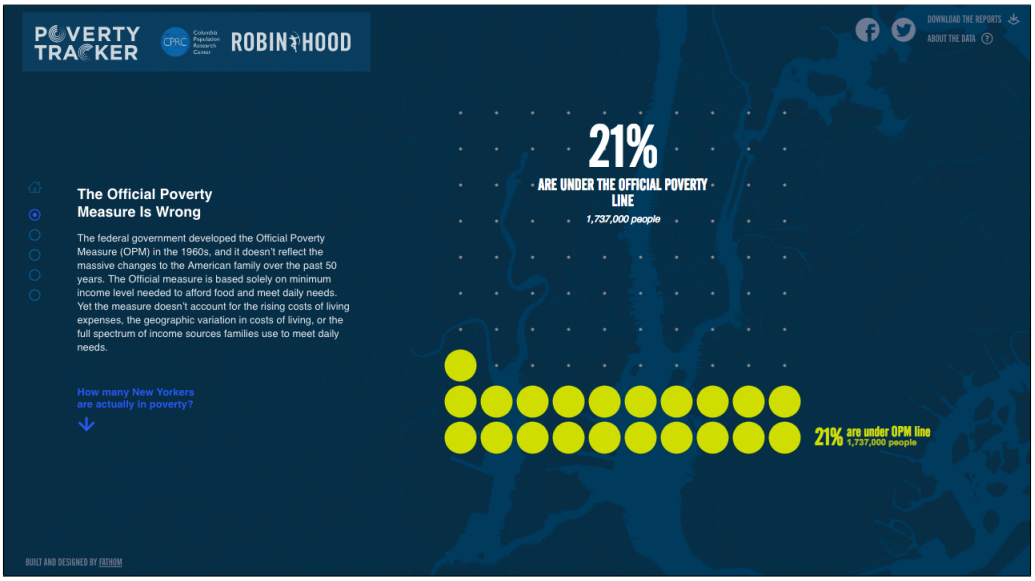


Figure 13.3: An Example of a Current Web Application Surrounding Poverty Statistics II

certain factors and previous historic inequity. A difficult would emerge due to the holistic nature of the problem, as equity can be regarded as semantic in nature, in turn limiting the ease for data to capture this. By following a similar pattern to this project, a model could be trained on a number of factors and draw correlations between particular influences. Researching into this work highlighted that a good starting point could be to create a linear regression model, as it would lead to adequately accurate results and could utilised the preexisting data sourced for this project.

13.2.3 Richer Data

The analysis of the system would improve in accuracy with additional datasets. In the scope of the current project, this additional data would have caused storage and capacity problems, however ideally a much larger quantity of data would have been used. Subsequently, a more people-centric picture could be provided by using data mining techniques and sentiment of popular social media websites such as Twitter. Using this mechanism, tweets originating from or mentioning New York City could be mined and analysed. Leveraging this data was a possibility in this project but was rejected due to the time constraints, however future work could utilise this extra datasource to provide a more in-depth analysis.

Bibliography

[1] L. M. A. Bettencourt, J. Lobo, D. Helbing, C. Kuhnert, and G. B. West. Growth, innovation, scaling, and the pace of life in cities. *National Academy of Sciences of the United States of America*, 2007.

[2] J. Billings, L. Zeitel, J. Lukomnik, T. S. Carey, A. E. Blank, and L. Newman. Impact of socioeconomic status on hospital use in new york city. *Health Affairs 12*, 1993.

[3] M. R. Bloomberg and D. Yassky. 2014 taxicab fact book. `http://www.nyc.gov/html/tlc/downloads/pdf/2014_taxicab_fact_book.pdf`, 2014.

[4] W. Bloss. Escalating u.s. police surveillance after 9/11: an examination of cause and effects. *Surveillance and Society*, 2007.

[5] E. Brenner. Everything you need, in one giant package. `http://www.nytimes.com/2008/04/06/realestate/06live.html`, April 2008.

[6] Q. Bui and J. White. Mapping the shadows of new york city: Every building, every block. `https://www.nytimes.com/interactive/2016/12/21/upshot/Mapping-the-Shadows-of-New-York-City.html?_r=0`, 2016.

[7] F. Burchi. Indentifying the role of education in socio-economic development. *International Conference on Human and Economic Resources*, 2006.

[8] Carto. Carto. `https://carto.com/`, 2016.

[9] D3. Data-driven documents. `https://d3js.org/`, 2015.

[10] L. Duhl and A.K.Sanchez. Healthy cities and the planning process. *World Health Organisation*, 1999.

[11] J. M. Esposito. *New York City Fire Department Chief Officer's Evaluation of The Citywide Incident Management System As It Pertains To Interagency Emergency Responce.* PhD thesis, Naval Prostgraduate School, 2011.

[12] C. Farina. Department of education. `http://schools.nyc.gov/AboutUs/default.htm`, 2016.

[13] N. Ferreira, J. Poco, H. T. Vo, J. Freire, and C. T. Silva. Visual exploration of big spatio-temporal urban data: A study of new york city taxi trips. *IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS*, 2013.

[14] J. Freire. Data science for cities. `http://cosmo.nyu.edu/hogg/research/2014/04/28/freire.pdf`, 2014.

[15] R. Heffernan, F. Mostashari, D. Das, A. Karpati, M. Kulldorff, and D. Weiss. Sydromic surveillance in public health practice, new york city. *Emerging Infectious Diseases*, 10, November 2004.

[16] J. Hochtl, P. Parycek, and R. Schollhammer. Big data in the policy cycle: Policy decision making in the digital era. *Journal of Organizational Computing and Electronic Commerce*, 26:147–169, December 2015.

[17] P. . Homes. New york city demographics. `http://www.point2homes.com/US/Neighborhood/NY/New-York-City.html`, 2016.

[18] K. Jacobs. Nyc curbed. `http://ny.curbed.com/2016/9/13/12891320/freshkills-park-nyc-staten-island-engineering-design`, September 2016.

[19] M. Kahn. An art scene grows in brooklyn. `http://www.cntraveler.com/stories/2016-02-12/an-art-scene-grows-in-brooklyn`, February 2016.

[20] L. Keaney. New york city finds one in five adults has mental health problems. `http://www.reuters.com/article/us-new-york-mentalhealth-idUSKCN0T12OO20151112`, 2015.

[21] D. P. M. Nathan Glazer. *Beyond the Melting Pot: The Negroes, Pureto Ricans, Jews, Italians and Irish of New York City.* Massachusetts Institute of Technolody Pres, 1963.

[22] K. Nauer, N. Mader, G. Robinson, and T. Jacobs. A better picture of poverty: What chronic absenteeism and risk load reveals about nyc's lowest income elementary schools. Technical report, Center for New York City Affairs, 2014.

[23] C. O'Connor. In effort to ensure sound education, city battles student homlessness. `http://www.gothamgazette.com/government/5937-in-effort-to-ensure-sound-education-city-battles-student-homelessness`, 2015.

[24] D. Owen. The greenest place in the u.s. may not be where you think. `http://e360.yale.edu/features/greenest_place_in_the_us_its_not_where_you_think`, 2009.

[25] ArcGIS. Arcgis features. `https://www.arcgis.com/features/index.html`, 2016.

[26] British Computing Society. British computing society - code of practise. `http://www.bcs.org/upload/pdf/cop.pdf`.

[27] Brooklyn Community Foundation. Brooklyn insights. `https://issuu.com/studybrooklyn/docs/all-brooklyn-brooklyn-neighborhood-report`, 2012.

[28] Brooklyn Online. Brooklyn ny history. `http://www.brooklynonline.com/history/`, 2010.

[29] Centers for Medicare and Medicaid Services. Medicaid. https://www.medicaid.gov/, 2016.

[30] Centers for Medicare and Medicaid Services. Medicare. https://www.medicare.gov/, 2016.

[31] City Of New York. Nyc thrive. https://thrivenyc.cityofnewyork.us/wp-content/uploads/2016/03/ThriveNYC.pdf, 2015.

[32] City Of New York. Thrivenyc: 150-day update. https://thrivenyc.cityofnewyork.us/wp-content/uploads/2016/06/Thrive150_report_fnl_singlepages.pdf, 2016.

[33] CNBC. Taxi owners, lenders sue new york city over uber. http://www.cnbc.com/2015/11/17/new-york-city-sued-over-uber-by-taxi-owners-say-livelihood-under-threat.html, November 2015.

[34] Columbia Population Research Center. Monitoring disadvantage in new york city. http://povertytracker.robinhood.org/, 2017.

[35] Constructive. 6 great interactive data visualization tools in 2016. https://constructive.co/insights/6-best-data-visualization-tools-2016-pt-1/, 2016.

[36] Creative Bloq. The 38 best tools for data visualization. http://www.creativebloq.com/design-tools/data-visualization-712402, February 2017.

[37] Dashing D3.js. Why build data visualisations with d3.js. https://www.dashingd3js.com/why-build-with-d3js, 2016.

[38] Department of Internation Economic and Social Affairs - Statistical Office. *Handbook On Social Indicators*. United Nations, 1989.

[39] Department of Parks and Recreation - New York Government. New york city department of parks and recreation. https://www.nycgovparks.org/, 2016.

[40] Google Maps. Staten island map. https://www.google.co.uk/maps/place/Staten+Island,+NY,+USA/@40.564521,-74.2869025,11z/data=!3m1!4b1!4m5!3m4!1s0x89c245ef79f4d4e7:0x50271f8534babc78!8m2!3d40.5795317!4d-74.1502007, 2017.

[41] Huffington Post. Live results: Massachusetts senate special election. http://elections.huffingtonpost.com/2013/massachusetts-senate-results, June 2013.

[42] Matplotlib. Matplotlib. http://matplotlib.org/, 2017.

[43] Matplotlib. Plotting data on a map. http://matplotlib.org/basemap/users/examples.html, 2017.

[44] New York City Independent Budget Office. *New York City Public School Indicators: Demographics, Resources, Outcomes*, 2015.

[45] New York Goverment. New york city counties. http://www1.nyc.gov/nyc-resources/service/2123/new-york-city-counties, 2016.

[46] New York Goverment - 311 Agency. 311 agency. http://www1.nyc.gov/311/index.page, 2017.

[47] New York Goverment - Borough Presidents. New york borough presidents. http://www1.nyc.gov/nyc-resources/service/3083/contact-a-borough-president, 2016.

[48] New York Goverment - Emergency Management. Citywide incident management system. https://www1.nyc.gov/site/em/about/citywide-incident-management-system.page, 2016.

[49] New York Goverment - New York City Council. The new york city council. http://council.nyc.gov/html/home/home.shtml, 2016.

[50] New York Goverment - Office of the Mayor. The office of the mayor of new york. http://www1.nyc.gov/office-of-the-mayor/index.page, 2016.

[51] New York Government - Department of Affordable Housing. Department of housing preservation. http://www1.nyc.gov/nyc-resources/service/1021/affordable-housing, 2016.

[52] New York Government - Department of Environemental Protection. Department of environmental protection. http://www.nyc.gov/html/dep/html/about_dep/mission_statement.shtml, 2016.

[53] New York Government - Fire Department of New York City. Fire department city of new york. http://www1.nyc.gov/site/fdny/index.page, 2016.

[54] New York Public Library. New york public library. https://www.nypl.org/, 2016.

[55] NYC Gov. New york census. http://www.census.gov/quickfacts/table/PST045215/36, 2016.

[56] NYC Gov. Nyc terms of use. http://www1.nyc.gov/home/terms-of-use.page, 2016.

[57] NYC Open Data. Nyc open data. https://nycopendata.socrata.com/, 2016.

[58] NYC Open Data. Subway entrances dataset. https://data.cityofnewyork.us/Transportation/Subway-Entrances/drex-xx56, March 2017.

[59] NYCGo. Boroughs of new york. http://www.nycgo.com/boroughs-neighborhoods, 2017.

[60] NY.com. Geography and origins of new york city. https://www.ny.com/histfacts/geography.html, 2017.

[61] Office of the Mayor. Nyc organisational chart. http://www1.nyc.gov/office-of-the-mayor/org-chart.page, 2016.

[62] Office of the Mayor. Mayor de blasio, chancellor farina annouce highest-ever graduation rate. http://www1.nyc.gov/office-of-the-mayor/news/076-17/mayor-de-blasio-chancellor-fari-a-highest-ever-graduation-rate#/0, 2017.

[63] One World Trade Center. One world trade center. https://oneworldobservatory.com/, 2016.

[64] Plotly. Plotly. https://plot.ly/, 2017.

[65] Population Reference Bureau. 2016 world population data sheet. http://www.prb.org/Publications/Datasheets/2016/2016-world-population-data-sheet.aspx, 2016.

[66] The Economist. Linguistics - say what? http://www.economist.com/node/21528592, 2016.

[67] The Telegraph. A history of the new york cab. http://www.telegraph.co.uk/news/worldnews/northamerica/usa/8491507/A-history-of-the-New-York-cab.html, 2011.

[68] Transit Museum Education. History of public transportation in new york city. http://www.transitmuseumeducation.org/trc/background, 2016.

[69] United Nantions Department of Economic and Social Affairs. World urbanisation prospects, 2011.

[70] United Nations. Un global issues. http://www.un.org/en/globalissues/, 2016.

[71] United Nations. New york city thrive mental health initative presented at the united nations. http://iaapsy.org/united-nations/current-reports/articleid/42/report-on-new-york-city-thrive-mental-health-initiative-presented-at-the-united-nations 2017.

[72] United States Government. United states cenus. http://www.census.gov/popest/about/terms.html, 2016.

[73] University of London. University of london data science course. http://www.city.ac.uk/courses/postgraduate/data-science-msc, 2017.

[74] University of Warwick. Warwick institude for the science of cities. http://www.wisc.warwick.ac.uk/, 2017.

[75] Yankee Stadium. Yankee stadium reference guide. http://newyork.yankees.mlb.com/nyy/ballpark/information/index.jsp, 2016.

[76] Yes The Bronx. History of the bronx. http://www.yesthebronx.org/about/history-of-the-bronx/, 2015.

[77] K. L. Rider. A parametric model for the allocation of fire companies in new york city. *Management Science*, 1976.

[78] A. Roest. Nyc open data overview. https://opendata.cityofnewyork.us/overview/, 2017.

[79] D. Rogers. *110 Livingston Street: Politics and Bureaucracy in the New York City Schools*. Random House, 1968.

[80] S. Ryley. Parents in south bronx school district, nyc's worst, struggle to find promising options. http://www.nydailynews.com/new-york/education/failing-south-bronx-schools-affected-student-life-home-article-1.2150189, 2015.

[81] J. Stark, K. Neckerman, G. S. Lovasi, J. Quinn, C. Weiss, M. D. M. Bader, K. Konty, T. G. Harris, and A. Rundle. The impact of neighbourhood park access and quality on body mass index among adults in new york city. *Elsevier*, 2014.

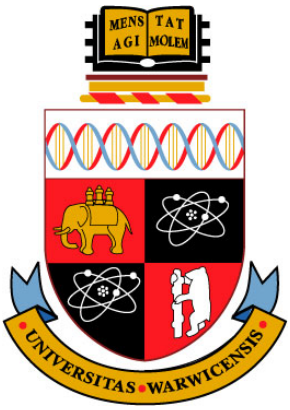
[82] Uber. Uber. `https://www.uber.com/en-GB/`, 2017.

[83] W. Walker. Using the set-covering problem to assign fire companies to fire houses. *Operations Research*, 22(2):275–277, 1974.

[84] W. Walker. Fire department deployment analysis. *The Rand Corporation*, 1979.

[85] M. Wallace. *Gotham: A History of New York City to 1898*. Oxford University Press, 1999.

[86] R. Whitsett. Urban mass: A look at co-op city. `http://cooperator.com/article/urban-mass/`, December 2006.



Understanding the Socioeconomics of New York City Using Open Data

CS310 Computer Science Project

Progress Review

Emma Dutton

Supervisor: Dr. Matthew Leeke

Department of Computer Science
University of Warwick

2016-17

Contents

List of Figures

1 Introduction

The earliest cities grew out of a common need to provide physical protection, import resources and export waste [11]. As cities have developed, so the needs of those who reside within them have evolved. With the global population reaching 7.4 billion [46], more than half of whom live in urban areas, questions relating the sustainability of cities and the models we have for supporting their continued development have never been more urgent.

Chief among issues in urban sustainability is the question of resource equity. Simply put, does being born in a certain region of a city make it more likely that you will receive a better education, higher wage, or better healthcare? Many argue that, by combining the power of modern data analytics with newly published city data, municipal policies can be developed to benefit the lives of thousands of people [16]. Substantiating this, recent studies have demonstrated that data-centric approaches can be used in identifying areas of social deprivation and proposing practicable policies in addressing the associated challenges [55].

1.1 Project Aim

The overarching aim of this project is to explore resource equity in New York City, one of the most developed and densely populated cities on Earth. This will result in the identification of areas for improvement, with a practicable municipal policy recommendations being made in these areas. The exploration will be conducted with an emphasis on open data, not least because of the tremendous growth in the volume and availability of this data over the past decade [10].

New York City was selected for analysis due to its diverse and distinctive demographic profile across a multitude of categories, including those relating to ethnicity, linguistics and age [41, 48, 51]. New York City also has an abundance of publicly available data sources online, such as NYC Open Data [42], which means analysis can be performed without the need for a potentially costly and time consumption period of data collection.

1.2 Project Progress

The Project Specification document broke down the problem statement into five project phases; research, hypothesis formation, data collection, analysis and reporting. This document describes the areas where significant progress has been made, most notably in contextual research, data set identification and analytic approach exploration.

1.3 Report Structure

Section 2 will identify key findings in the research that has been undertaken, such as geography, governance, and social equity. Section 3 will introduce the datasets that have been identified, and explain how they will be utilised with regard to the aims of the project. Section 4 will elaborate on the progress in identifying the key data analytics techniques that will be used to manipulate the data. The report will conclude with discussions on project management and future work in Section 5, and finally a conclusive summary in Section 6.

Figure 1: Map of New York Boroughs [43]

2 Research

Since the Project Specification was written, further investigation of the domain, i.e., New York City, of the project has taken place. This section will highlight the key research areas that provide a firm basis for the project. This research will also be used to quantify the success of any project findings, allowing comparisons to be drawn between current policies and those proposed through the results of this project.

2.1 Geography

New York City is a single city that is a part of the wider New York State, situated on the east coast of America. Its position between the Hudson River and the Atlantic Ocean provides a small land mass positioned in a naturally sheltered harbour. The scarcity of land means that it is one of the most densely packed cities in the United States. This has lead to density management being an important environmental issue in New York City, though the tinuing efforts of the government have enabled it to become one of the most efficient cities in America thorough proficient waste management systems, affordable housing and accessible transportation.

New York City houses over 16 million [17] citizens over 5 boroughs; namely the Bronx, Brooklyn, Manhattan, Queens and Staten Island. These boroughs are extremely diverse in nature, combining different cultures to produce the cosmopolitan atmosphere that New York is famous for. The rest of this section will discuss each borough in order to provide a high-level overview of the context of this project.

2.1.1 The Bronx

The Bronx is situated in the north of New York City. It is historically know as the birth place of rap and hip hop culture due to its high populous of African American residents during the 1990's [53]. The Bronx has been home to many nationalities, with the borough going through rapid growth after World War One during which many Irish, Italian and Jewish people began to settle in the area [53]. A decade later, in the prohibition days surrounding 1926, it was known for having one of the highest crime rates, with many speakeasies selling illicit alcohol.

In current days, the Bronx is the site of the Yankee Stadium; a 50,000 seated stadium home to the New York Yankee's baseball club [52]. Additionally, the Bronx holds the United State's largest cooperatively owned housing development, aptly named 'Co-op City' [4]. Co-op City provides affordable housing to residents, as well as offering three shopping centres, six schools, a weather station and a planetarium [4, 56]. Co-op City is so vast that if the area was classed as a city, it would be the 10th largest city in the United States [56].

2.1.2 Brooklyn

Brooklyn is the located in the south west part of the city, sharing a border with Queens. It has the largest number of citizens compared to the other boroughs, with the population being marked at 2.592 million as of 2013 [50, 24]. Brooklyn is know for its cultural diversity, that has given rise to a unique architectural heritage and independent art scene [20]. As of 2007, the top 5 ethnicities in Brooklyn were African American (15.2%), Religious Responses (7.4%), Puerto Rican (6.0%) Italian (5.8%) and Chinese (4.7%) [50]. It is also notable that the top 5 languages spoken in Brooklyn at the same time were English, Spanish, Chinese, Russian and Yiddish [23, 50].

2.1.3 Manhattan

The most populous borough in New York is Manhattan. It is located centrally and houses the business and financial districts. This has lead to the borough being lined with city scrapers, the largest being the One World Trade Center [45]. Additionally, Manhattan was the location of the twin towers, which were disastrously bombed in 2001. This has lead the area to have increased security and police presence [3].

Manhattan also hosts many of the tourist attractions that are iconic to New York, such as the Empire State Building, Rockefeller building and Central Park. These high levels of tourism contribute to the economy, and worldwide presence of New York. Additionally, due to the large footfall provided by tourism and industry, Manhattan contributes to 90% of the taxi demands of the whole city [2].

2.1.4 Queens

Queens is north of Brooklyn, with the largest land mass in comparison to the other four boroughs. Queens is mainly a residential area for middle class citizens but also the most ethically diverse

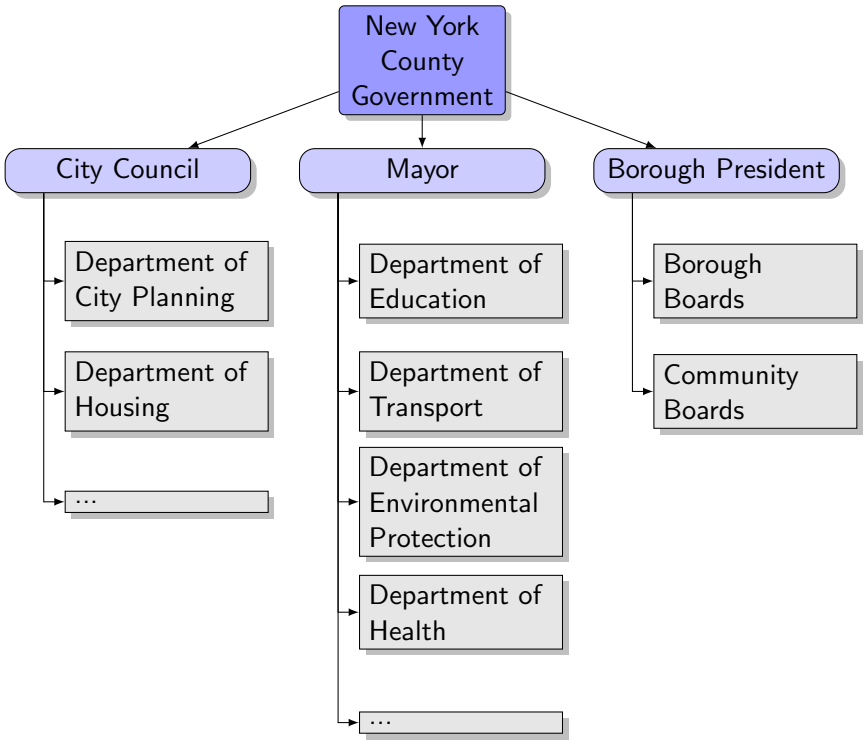


Figure 2: A subsection of the organisational structure in the governance of New York [44]

county in the United States. Its population consists of a diverse range of people, with 50% being white, 28% Hispanic, 24% Asian, 21% black, and 3% mixed race [50]. Additionally, Queens also has a relatively low education rate compared to other boroughs, with only 27% having a Bachelors Degree level education [17].

2.1.5 Staten Island

Staten Island is a detached from the main body of New York city, in the south west of the region. Due to its distance from the main centre of Manhattan, it has grown more of a residential suburban character. It is connected to Brooklyn by the Verrazano-Narrow Bridge, and to Manhattan by the free Staten Island ferry, proving it to be popular for commuters to the city. Staten Island was home to one of the largest landfill sites in the world, named Freshkills Landfill, which is undergoing transformation into one of the largest urban parks in America, almost three times the size of Central Park. This park is due for opening in 2036 [29, 18].

2.2 Governance

The governance of New York is split across four entities that have specific legislative powers. These entities are the county government which presides over the city council, mayorship and borough

presidents, as shown in Figure 2.

2.2.1 New York County Government

In New York State every county has a governing power. As New York City is part of New York State, it has its own New York County Government. The role of the county governments is to provide state mandated services, such as management of the police force, maintaining roads and transport and providing economic development assistance [30].

2.2.2 Legislative Branch - New York City Council

New York City has one council that is responsible for the legislative branch of the government structure. City charters are passed to enable an organisational structure of the city, for example, the name, boundaries and administrative processes. The city council is ran by elected council members which represent their own borough. The council have responsibility for departments that are tasked with the management of the city, for example the Department of City Planning, City Housing Authority and Department of Parks and Recreation [33].

2.2.3 Executive Branch - Mayorship

The elected Mayor heads the cities executive branch from city hall and has jurisdiction over the five boroughs of New York. There are around 50 departments that are appointed by the Mayor, for example, the Department of Education who managed the city's schools, or the Department of Transport who maintain and upgrade New York's extensive transport system [34].

2.2.4 Borough President

There is one borough president for each of the five boroughs in New York; The Bronx, Brooklyn, Manhattan, Staten Island and Queens. The borough presidents are voted in by the people. Their role is to advise the Mayor on executive issues, and analyse borough needs through the annual budget process. Additionally, the borough president will appoint the community board who serve as advocates of the citizens of the borough. This board will evaluate the needs of the local communities, and make recommendations to the president from their findings [31].

2.3 City Resource Categories

As explained in Section 1, the project aims to improve current policies surrounding day-to-day issues facing the citizens of New York. In order to understand the environment in which they are currently living, key areas such as health, transportation, education, environmental protection and economic stability must be understood. These areas were considered following research from United Nations' "Handbook on Social Indicators" [28], and categorised using departments that span the mayors office.

2.3.1 Healthcare

There is no public healthcare system in the United States, requiring individuals to pay insurance companies to ensure they can pay for medical costs. However, there are government schemes in place to ensure that underprivileged groups can still access health insurance so they can be treated if they need to be. Medicaid [25] is a social healthcare programme aimed at low income individuals and families, providing free healthcare to those whose income is below a certain threshold. Medicare [26] is an alternative government backed social insurance programme, that covers citizens over the age of 65 that have paid into the scheme through payroll whilst they were in employment. These schemes aim to ensure that all citizens have access to health care, regardless of their social situation. However, even though these schemes are in place, there is evidence to show that "lack of timely and effective ambulatory care may have a significant impact on hospitalisation rates in the low income areas" [1].

2.3.2 Transport

New York City has a broad transport network. Its vastly connected subway system is one of the largest subway systems in the world, built between 1913 and 1931 [49]. The subway connects every borough apart from Staten Island, and runs 24 hours a day, seven days a week [49]. This subway system enables residents to commute into central Manhattan for work, or tourists to explore different areas of the city.

2.3.3 Education

The Department of Education is responsible for running all of the public schools across New York City. Their reach spans over 1.1 million students in over 1,800 schools, making the NYC Department of Education one of the largest school districts in the United States [14].

It has been argued that education is a fundamental factor in the socioeconomic development of a population in a paper by Burchi [5]. The paper draws the conclusion that when schools in third world countries are better attended, the quantity of people affected by food insecurity decreases, which can be used as an indication of poverty levels. Further analysis can be done to identify whether this theory is applicable in such a developed city as New York, and whether the age someone leaves education could be a factor of their food insecurity for the future.

2.3.4 Environment

New York City has a diverse environment. From corporate skyscrapers that house global conglomerates to the vast array of parks around the area [29], there are many challenges in sustaining the environmental equilibrium in the city. To ensure this stability, the Department of Environmental Protection aims to"protect public heath and the environment by supplying clean drinking water, collecting and treating wastewater, and reducing air, noise, and hazardous material solution." [36]

Borough	Population [17]	Average Household Income (Per Annum) [17]
The Bronx	1,446,350	$50,306
Brooklyn	2,528,061	$64,217
Manhattan	1,563,897	$132,754
Queens	2,282,534	$78,438
Staten Island	478,652	$88,637

Figure 3: A comparison of average household income across the five boroughs of New York City

2.3.5 Economics

To analyse an aspect of the economics in New York City, the employment statistics of citizens in each borough can draw potential insights, however it is difficult to find specific employment data. Instead, average household income statistics can be utilised. Due to the vast variety between each borough, it can be useful to make some direct comparisons to analyse where there are noticeable differences between areas.

The data show in Figure 3 draws some interesting comparisons. The average yearly household income in Manhattan is more than any other borough with significant difference between the difference between that and the Bronx, which has the lowest. Interestingly, both boroughs have a similar about of citizens. Differences like this can be analysed further in the analytical phase.

2.3.6 Public Services

New York City offers many public services citizens, in order to assist and improve their lives. These services range from fire departments [37] to libraries [38] to social housing [35].

The Fire Department of New York City is split into divisions that serve each of the five boroughs. As the work of the fire department is so integral the the emergency response system of the city, their responsibilities are divided into several 'core competencies'. Some of these competencies include fire suppression, pre-hospital emergency care and fire prevention inspections.

Due to the terror attacks of September 2001, the Citywide Incident Management System [32] was established to provide better care for citizens in emergency situations, such as natural disasters or the threat of terror groups. This organisation aimed to combine the efforts of the New York Fire Department and the New York Police Department, allowing a more effective response to emergency situations. The structure of the inter-agency operations has been criticised [13], claiming that coordinated response does not occur as often as such be expected, with protocols often not adhered to. This research illuminates where further investigation can take place in order to identify what steps could be made to better improve the cross collaboration between the public services in New York City.

3 Data Collection

Many factors of New York City's policies were identified in the research section. Through the identification of public resources, it is clear that there are a lot of elements that can be considered when looking for beneficial policy reforms. In particular, the research grouped areas of policy into healthcare, transport, education, environment, ecconomics and public services, following from departmental structure under the Mayor.

In order to allow analysis of these policies, representative datasets must be identified to quantify historic trends in data. It is important to remember that the datasets cannot be assumed to be representative of an entire issue, such as healthcare, but more of a snapshot representing part of the discussion. Each of the research topics is represented by at least one dataset, with some having more depending on the breadth of the topic. This approach allows the breadth of the project to be wide, with a general focus on many contributing factors to New York City policies in order to suggest where improvements can be made. These suggestions can then be the focus of further study.

3.1 Data Sets

The following datasets were chosen to represent the main topics identified through research. Data of this nature is extremely accessibly from the government managed website NYC Open Data [42]. This website provides accurate data on a range of topics, providing the diversity that the project requires.

3.1.1 311 Data

311 is a telephone number used in the United States for non-emergency situations. Common 311 calls are concerning noise complaints, non-working state provided facilities (such as street lamps, parking meters, traffic lights), or potholes in roads. The 311 data has "proved useful not just at detecting reliable patterns but also at providing insight when the normal patterns are disrupted" [9]. This will allow for interesting analysis later on in the project.

3.1.2 School Survey

Ran annually by the Department of Education, the school survey is one of the largest surveys conducted nationally [27]. The results of this survey help leaders in the education sector understand how to develop the schooling system and "supports a dialogue amongst members of the school community about how to make the school a better place to learn" [27].

3.1.3 Directory of Parks

One of the most popular public resources in New York is the abundance of parks. The use of the directory of parks data set will allow analysis to evaluate whether each borough has equal access to outdoor space. This could be an interesting metric when combined with other statistics

Data Set	Research Topic	Year
311 Data	Environment	2010
School Survey	Education	2008
Yellow Taxi Data	Transport	2015
Emergency Responses	Public Services	2014
Community Health Survey	Healthcare	2010 - 2014
Cooperative and Condominium Comparables	Economics	2008 - 2012

Figure 4: A table showing the data collected

such as health. Similar research has already found a correlation between the access to parks and BMI, finding that the more access to outdoor space in a neighbourhood lowers the average BMI of citizens [55].

3.1.4 Yellow Taxi Data

As previously mentioned, the yellow taxi cabs are an icon of New York City with over 12,000 on the roads as of 2006 [54]. In order to analyse the trips of these yellow taxis, a data set containing pick-up and drop of points, times and locations will be utilised. This data is provided by the NYC Taxi and Limousine Commission.

3.1.5 Emergency Responses

The emergency response incident data set provides incident type and locations of incidents reported organised by date. This data was chosen as it has geo-locational coordinates so it can be plotted onto a map, and is also already separated by borough. The locational information allows the incidences to be categorised by borough, to see whether there are more emergency situations in a particular location, or if a specific borough has a higher rate of a particular incident, for example a bomb threat. This data can then cross referenced with the location of emergency service stations, to see whether there are more optimal locations to ensure a quicker response time.

3.2 Community Health Survey

The Community Health Survey data is released by the Department of Health and Mental Hygiene. This survey is conducted annually over the phone and captures the health traits and borough of the responders. This data can be used to predict healthcare trends and indicate possible neighbourhood-wide issues that could be addressed by policy changes.

3.3 Cooperative and Condominium Comparables

A New York State law requires cooperative and condominium properties to be valued as if they were rental apartment buildings. In doing so, the size, age, and location of the property is evaluated.

This data is especially relevant as it is extensive in date range and also categorised by borough. These estimated property values are released by the Department of Finance, and can be used as an indication of wealth in a given area. This data set can be used in comparison to the income data shown in Table 3 to build up a picture of the economic value in different boroughs.

4 Data Analytics

Following the original plan laid out in the Project Specification, time has been devoted to identifying different data analysis techniques. This section will highlight different software that could assist the analytical developments of the project.

4.1 Analytics

To allow the identification of patterns in the data sets identified, various analytic approaches will be used, and their results analysed with respect to possible policy changes. Before work on this analysis can commence, it is necessary to research the techniques and languages that could need to be utilised, a selection of which are shown below.

4.1.1 Python

Python [15] will be used as the primary coding language in the project, used to develop the analytic algorithms that will identify patterns in the data sets. Python has been chosen due to its flexibility and ease of use, but also its extensive and widely supported packages for data analytics, namely pandas [39] for storage and NumPy [40] for mathematical analysis.

4.1.2 Anaconda Distribution

The Anaconda distribution [7] is a package manager for Python which provides an environment specifically for predictive analytics. Anaconda is attractive as a prospective software platform as the time it would take to install all the necessary python packages such as pandas [39], NumPy [40], SciPy [47] and many others, would take much longer than simply installing Anaconda. The Anaconda distribution also provides a testing environment which allows for more efficient debugging, as well as Juypter notebook for data visualisations.

4.2 Visualisation

Once the data has been explored algorithmically, the results of the data analysis can be visualised to allow patterns to be displayed interactively. This allows the results to be displayed in different ways, such as Carto [6] and ArcGIS [22] which will be discussed below. Alternatively, using web development packages may be a more interactive way to explore the data. An example of this is the Guardian Interactive Team's data visualisation that tracked the spread of rumours posted on Twitter throughout the London Riots, such as a tiger escaping from London Zoo. The rumours were categorised into those that were supporting the claim, opposing the claim, querying or commenting. A timeline demonstrates how these rumours spread with the majority of initial tweets supporting, and then overtime as the rumour becomes dispelled the majority becomes opposing [21]. The visual output displayed can be seen in Figure 6.

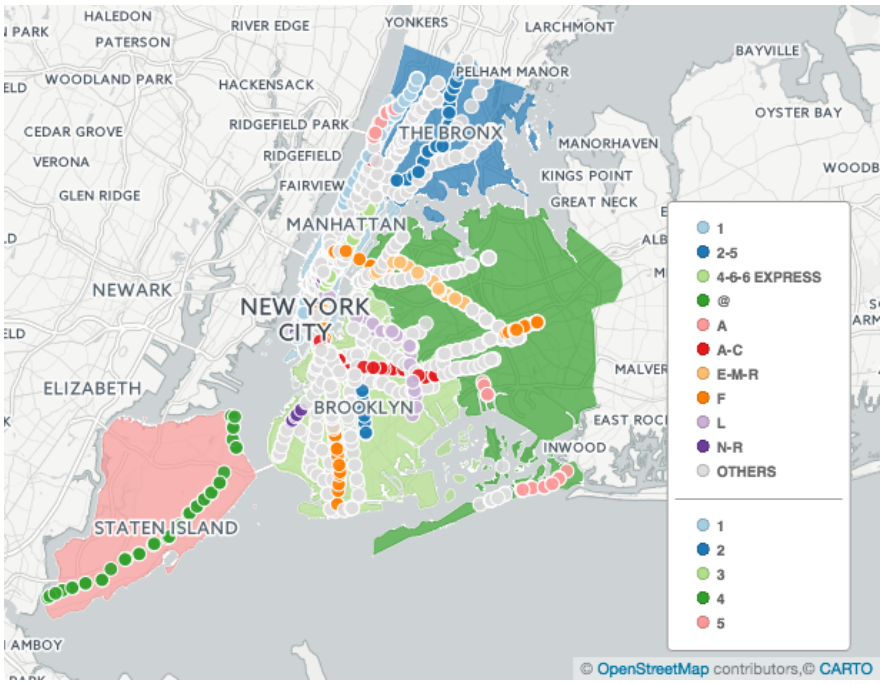


Figure 5: Using Carto [6] to produce a map of New York City, showing boroughs and train routes

4.2.1 Geographic Information System

Many ways of displaying geographical information have been explored, such as Carto [6] and ArcGIS [22]. Through experimenting with Carto, a map of New York City was created show in Figure 6. The map is made up of two layers; the lower layer mapping the different boroughs using geolocation data and shading them in different colours, and the higher layer showing the train routes across these boroughs. As a proof of concept, this figure was a simple test to analyse whether some boroughs had more access to train lines than others, and shows that Staten Island only has one train route compared to multiple through Manhattan.

4.2.2 Programming Language Support

In order to produce dynamic and visually appealing data visualisations hosted on a website, the appropriate languages will be utalised. A combination of HTML5, CSS, and JavaScript [19] will provide a clean interface between the user and the data. Additionally, specific JavaScript packages such as D3 [8] have been identified to enable responses to user input, similarly to the timeline in the Guardian example.

As the visualisation phase requires the preceding algorithmic phase to have been completed successfully before it can begin, there may not be enough time to develop a sufficient visualisation solution from scratch. If this is the case, a simpler solution will be implemented to allow the over-

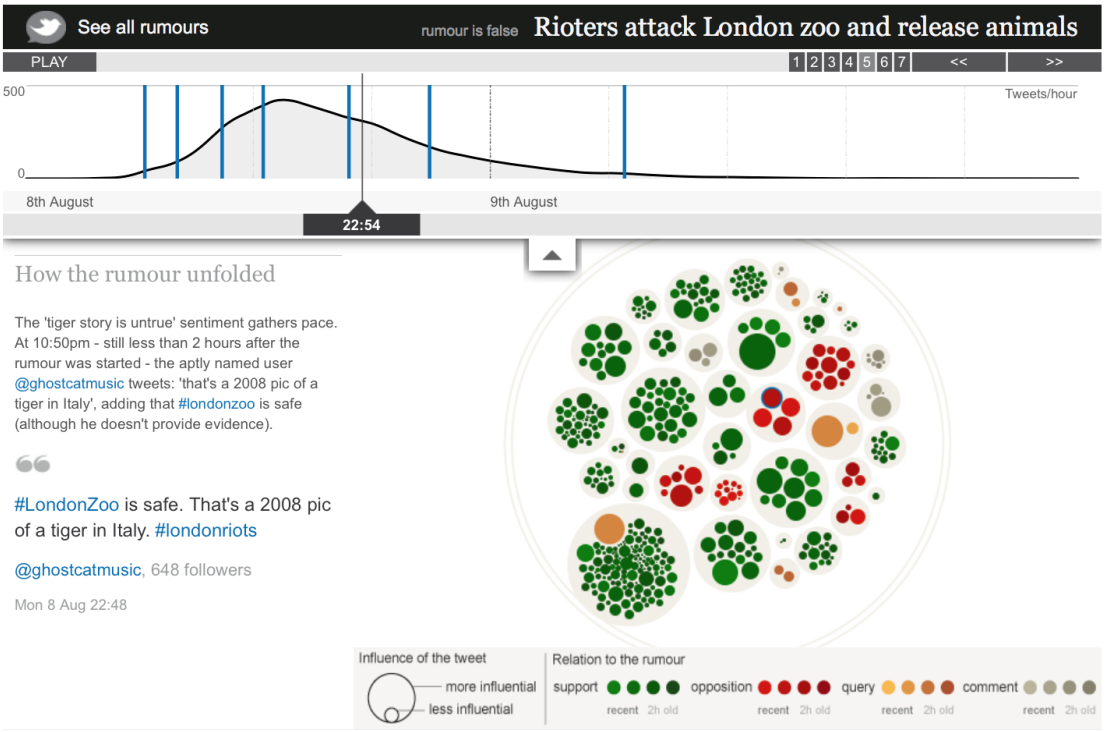


Figure 6: A visualisation of how the rumour about a tiger escaping from London Zoo unfolded on twitter during the London Riots in 2011 [21]

arching goal of the project to be met within an acceptable timeframe. This simpler solution will use geographical information systems, such as Carto and ArcGIS, for analysis and visualisation.

5 Project Management

The Project Specification highlighted different phases of the project, allowing the workload to be broken down into smaller quantities. This allowed preliminary planning to be made to ensure that all deadlines were met. As the project is now underway, some of these plans have been altered now further insight has been gained.

5.1 Project Phases

The project was broken down into phases to simplify the methodology and allow for a streamlined workflow. Each phase has specific tasks and provides an output that the next phase requires to start. This design is based loosely around the waterfall method and lends itself to the project due to the large research content. The extracts show in italics below were taken from the Project Specification [12], and commentary has been written to explain how work has progressed.

5.1.1 Research

"The project will begin by looking at current work surrounding a large scope of areas. Firstly, research into the city of New York will be beneficial in order to learn about the day-to-day issues, such as transport, government and municipal laws. This will provide a strong base to identify areas of inequality in state facilities such as parks, schools, or links to transport systems. In parallel, research will be conducted into socioeconomics topics such as state welfare, health and poverty. These findings will be useful in later stages to identify a suitable hypothesis to evaluate. Finally, research of socioeconomic analysis will provide exposure to relevant work using analytics to identify and improve social and economic issues."

This work has been completed but took longer than anticipated. In order to understand the breadth of services available in New York City, such as the departments under the Mayorship, a significant amount of time was needed to investigate academic resources in sufficient depth to make a judgement on areas for analytic exploration . Once the knowledge of these are was gained, more time was taken to look for academic articles that discussed the socioeconomic and equity issues surrounding these topics, and was therefore a lengthly task. Rather than being done in parallel, most of the topics discussed in the Research Section were explored individually as there was a necessary fundamental level of understanding needed before continuing to the next topic, for example it was necessary to learn about the geography before issues of governance could be appreciated.

5.1.2 Research Question Formulation

"After collecting valuable resources through the research phase, it will then be possible to identify a specific claim to evaluate. This will require a firm understanding of current government policies in New York City and therefore time will be devoted to exploring this. Once the background has been investigated, a hypothesis can be generated with the goal of finding new ways to improve the current environment to the benefit of citizens. A test case will be developed to test the solution to

the proposed hypothesis following the principles of test driven development."

The direction of this aspect of the project has evolved since the Project Specification. Rather than a specific hypothesis being analysed with regard to resource equity, it was decided that identifying research questions relating to a board range of topics in resource would allow a greater spectrum of potential issues to be targeted. This need for this change of direction was identified early during the research phases of the project. The change is motivated by the fact that there is no single measure of resource equity, which motivated the development of a more holistic methodology to support the proposed aims of the project.

5.1.3 Resource Collection

"At this point, the project will be focused on a set of metrics identified in the hypothesis formation section that can be explored using online datasets. Time will be taken to identify a selection of related data to be used in later analysis. Additionally, the data will be refactored into a common type."

Due to the breadth of the project and the change is scope, the resource collection was actually completed before the formulation of research questions. It was necessary to see what reliable and accurate data was available which would represent aspects of the research topics identified. Not much time was needed to clean the data as it was all downloaded from NYC Open Data in CSV format.

5.1.4 Analysis

"The majority of set up will now have been completed for the project, and analysis can begin using the clean data from the previous phase. Data exploration will look at correlations between metrics and try possible analytical algorithms and machine learning techniques. From this proof of concept, further analysis will take place to produce concrete findings. These findings will be tested thoroughly using the testing strategy decided in the hypothesis formation phase."

This phase has not begun and is still on track to start in January following the original project timeline as shown in Figure 7.

5.1.5 Reporting

"Finally, all the results will be collated and recommendations will be drawn. The report will evaluate the impact these recommendations would have on New York and discuss how successful the project has been."

This phase has also not begun but will commence once analysis is finished as documented in the Project Specification.

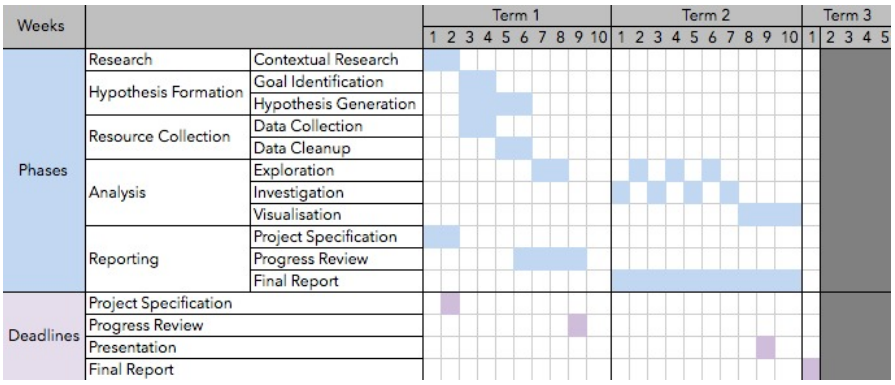


Figure 7: The original project timeline

5.2 Project Timeline

The Project Specification highlighted the need for an organised and structured plan to enable timely completion of the project. The original timeline produced can be seen in Figure 7.

"The project will be split into phases highlighted in the methodology section and run between October and April. The first two weeks will be spend on researching related work on New York, socioeconomics and civil analytics. Following this, time will be taken to find the correct data sets for analysis, and then refactoring them to the right format. The results of the research can be evaluated to identify the scope of the project. This will allow data exploration and analysis to begin. Additionally, the timeline also marks the deadlines for the following deliverables; project specification, progress report, presentation and the final report."

As progress has now been made on the project, it is clear that some of these assumptions were underestimated. Most notably, the research section took almost three weeks longer than anticipated due to the necessary breadth that was required. This meant that none of the other tasks could begin until the research was properly gathered. A revised version of the project timeline can be seen in Figure 8.

With the perspective gained from the first ten weeks of the project, it was decided that the data exploration, investigation and visualisation would be undertaken in parallell, rather that serially as the Project Specification described. This allows quicker iterations between development and feedback times to produce a better solution.

5.3 Management Tools

As the project lends itself to discussion of direction from the supervisor, weekly meetings have been imperative to allow both parties to discuss ideas and how to move forward. The minutes from these meetings have been uploaded to a blog for easy retrieval when the final report is being written.

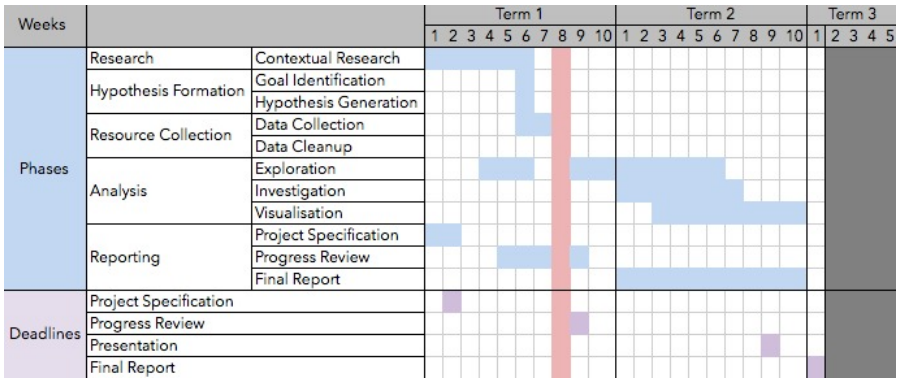


Figure 8: A revised timeline of work to ensure the projects completion

Additionally, due to the large quantity of reading that was required in the Research Section, BibTex was utilised to save all the references that have been cited in documentation.

6 Conclusion

To conclude the Progress Review, a lot of improvement has been made on the project since the Project Specification document was written. The continuation of work will now involve sophisticated data analytics techniques to allow trends and insights in the data sets to be identified. These trends will allow potential policy reforms to be acknowledged and tested accordingly. The deadline of the project will be met, as the timeline has been adjusted to reflect the change in scope, allowing time to represent the findings in interesting and novel web data visualisations.

References

[1] J. Billings, L. Zeitel, J. Lukomnik, T. S. Carey, A. E. Blank, and L. Newman. Impact of socioeconomic status on hospital use in new york city. *Health Affairs 12*, 1993.

[2] M. R. Bloomberg and D. Yassky. 2014 taxicab fact book. `http://www.nyc.gov/html/tlc/downloads/pdf/2014_taxicab_fact_book.pdf`, 2014.

[3] W. Bloss. Escalating u.s. police surveillance after 9/11: an examination of cause and effects. *Surveillance and Society*, 2007.

[4] E. Brenner. Everything you need, in one giant package. `http://www.nytimes.com/2008/04/06/realestate/06live.html`, April 2008.

[5] F. Burchi. Indentifying the role of education in socio-economic development. *International Conference on Human and Economic Resources*, 2006.

[6] Carto. Carto. `https://carto.com/`, 2016.

[7] Continuum. Anaconda distribution. `https://docs.continuum.io/anaconda/`, 2016.

[8] D3. Data-driven documents. `https://d3js.org/`, 2015.

[9] J. Daly. Nyc's 311 call center is a big data gold mine. `http://www.statetechmagazine.com/article/2014/03/nycs-311-call-center-big-data-gold-mine`, March 2004.

[10] A. Deshpande and D. Riehle. The total growth of open source. *The International Federation for Information Processing*, 2008.

[11] L. Duhl and A.K.Sanchez. Healthy cities and the planning process. *World Health Organisation*, 1999.

[12] E. Dutton. Understanding the socioeconomics of new york city using open data. Project Specification, October 2016.

[13] J. M. Esposito. *New York City Fire Department Chief Officer's Evaluation of The Citywide Incident Management System As It Pertains To Interagency Emergency Responce*. PhD thesis, Naval Prostgraduate School, 2011.

[14] C. Farina. Department of education. `http://schools.nyc.gov/AboutUs/default.htm`, 2016.

[15] P. S. Foundation. Python. `https://www.python.org/`, 2016.

[16] J. Hochtl, P. Parycek, and R. Schollhammer. Big data in the policy cycle: Policy decision making in the digital era. *Journal of Organizational Computing and Electronic Commerce*, 26:147–169, December 2015.

[17] P. . Homes. New york city demographics. `http://www.point2homes.com/US/Neighborhood/NY/New-York-City.html`, 2016.

[18] K. Jacobs. Nyc curbed. http://ny.curbed.com/2016/9/13/12891320/freshkills-park-nyc-staten-island-engineering-design, September 2016.

[19] JavaScript. Javascript. https://www.javascript.com/, 2016.

[20] M. Kahn. An art scene grows in brooklyn. http://www.cntraveler.com/stories/2016-02-12/an-art-scene-grows-in-brooklyn, February 2016.

[21] R. Procter, F. Vis, and A. Vos. Reading the riots. https://www.theguardian.com/uk/interactive/2011/dec/07/london-riots-twitter, 2011.

[22] ArcGIS. Arcgis features. https://www.arcgis.com/features/index.html, 2016.

[23] Brooklyn Community Foundation. Brooklyn insights. https://issuu.com/studybrooklyn/docs/all-brooklyn-brooklyn-neighborhood-report, 2012.

[24] Brooklyn Online. Brooklyn ny history. http://www.brooklynonline.com/history/, 2010.

[25] Centers for Medicare and Medicaid Services. Medicaid. https://www.medicaid.gov/, 2016.

[26] Centers for Medicare and Medicaid Services. Medicare. https://www.medicare.gov/, 2016.

[27] Department of Education - New York Government. Nyc school survey. http://schools.nyc.gov/Accountability/tools/survey/default.htm, 2016.

[28] Department of Internation Economic and Social Affairs - Statistical Office. *Handbook On Social Indicators*. United Nations, 1989.

[29] Department of Parks and Recreation - New York Government. New york city department of parks and recreation. https://www.nycgovparks.org/, 2016.

[30] New York Goverment. New york city counties. http://www1.nyc.gov/nyc-resources/service/2123/new-york-city-counties, 2016.

[31] New York Goverment - Borough Presidents. New york borough presidents. http://www1.nyc.gov/nyc-resources/service/3083/contact-a-borough-president, 2016.

[32] New York Goverment - Emergency Management. Citywide incident management system. https://www1.nyc.gov/site/em/about/citywide-incident-management-system.page, 2016.

[33] New York Goverment - New York City Council. The new york city council. http://council.nyc.gov/html/home/home.shtml, 2016.

[34] New York Goverment - Office of the Mayor. The office of the mayor of new york. http://www1.nyc.gov/office-of-the-mayor/index.page, 2016.

[35] New York Government - Department of Affordable Housing. Department of housing preservation. http://www1.nyc.gov/nyc-resources/service/1021/affordable-housing, 2016.

[36] New York Government - Department of Environemental Protection. Department of environmental protection. http://www.nyc.gov/html/dep/html/about_dep/mission_statement.shtml, 2016.

[37] New York Government - Fire Department of New York City. Fire department city of new york. http://www1.nyc.gov/site/fdny/index.page, 2016.

[38] New York Public Library. New york public library. https://www.nypl.org/, 2016.

[39] NUMFocus. Pandas. http://pandas.pydata.org/, 2016.

[40] NumPy. Numpy. http://www.numpy.org/, 2016.

[41] NYC Gov. New york census. http://www.census.gov/quickfacts/table/PST045215/36, 2016.

[42] NYC Open Data. Nyc open data. https://nycopendata.socrata.com/, 2016.

[43] NYC Tourist. Map of nyc. http://www.nyctourist.com/map1.htm, 2016.

[44] Office of the Mayor. Nyc organisational chart. http://www1.nyc.gov/office-of-the-mayor/org-chart.page, 2016.

[45] One World Trade Center. One world trade center. https://oneworldobservatory.com/, 2016.

[46] Population Reference Bureau. 2016 world population data sheet. http://www.prb.org/Publications/Datasheets/2016/2016-world-population-data-sheet.aspx, 2016.

[47] SciPy. Scipy. http://scipy.org/, 2016.

[48] The Economist. Linguistics - say what? http://www.economist.com/node/21528592, 2016.

[49] Transit Museum Education. History of public transportation in new york city. http://www.transitmuseumeducation.org/trc/background, 2016.

[50] United States Government. United states cenus. http://www.census.gov/popest/about/terms.html, 2016.

[51] Wikipedia. Demographics of new york city. https://en.wikipedia.org/wiki/Demographics_of_New_York_City#cite_note-3, 2016.

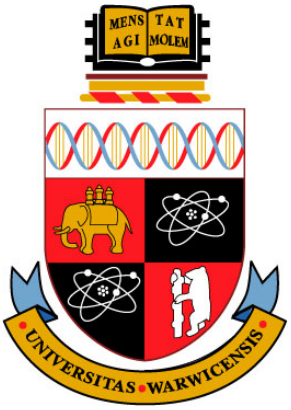
[52] Yankee Stadium. Yankee stadium reference guide. http://newyork.yankees.mlb.com/nyy/ballpark/information/index.jsp, 2016.

[53] Yes The Bronx. History of the bronx. http://www.yesthebronx.org/about/history-of-the-bronx/, 2015.

[54] P. Schenkman. The state of the nyc taxi. http://www.nyc.gov/html/tlc/downloads/pdf/state_of_taxi.pdf, 2006.

[55] J. Stark, K. Neckerman, G. S. Lovasi, J. Quinn, C. Weiss, M. D. M. Bader, K. Konty, T. G. Harris, and A. Rundle. The impact of neighbourhood park access and quality on body mass index among adults in new york city. *Elsevier*, 2014.

[56] R. Whitsett. Urban mass: A look at co-op city. http://cooperator.com/article/urban-mass/, December 2006.



Understanding the Socioeconomics of New York City Using Open Data

CS310 Computer Science Project

Project Specification

Emma Dutton

Supervisor: Dr. Matthew Leeke

Department of Computer Science
University of Warwick

2016-17

Contents

1 Introduction

It has been shown that cities were historically built to import food and water and export waste [5], but with the population now at 7.4 billion in 2016 [18], do these traditional city designs benefit the mass of modern living? Does being born in a certain borough mean you are more likely to receive a good education, higher wage, or better health? If we combine the power of cutting edge data science techniques with new open datasets, could municipal policies be developed to benefit the lives of thousands of people? With the expanse of the population ever growing, it is becoming crucially important to use computational analysis to identify areas of social deprivation and provide solutions.

Over the last decade there has been tremendous growth in the amount of open source data available [3]. This project will use such data to identify if there are correlations between the equity of state facilities and deprivation in citizens. A lack of equity could be the absence of a local park, for example, which may lead to a neighbourhood having higher BMI measurements and therefore a lower quality of health [25].

1.1 Problem Evolution

To answer these high level questions, it is logical to look at one particular sample. Preliminary research identified New York City an interesting case due to its diversity in a multitude of metrics such as ethnicity, linguistics and age [24] [22] [13]. An example that highlights this is the geographical layout, where wealthy boroughs such as Manhattan are situated next to poorer boroughs such as the Bronx, as shown in Figure 1. This will allow comparisons to be made on the basis of locality, enabling the use of geolocation datasets. New York also has an abundance of publicly available data sources online, such as NYC Open Data [15], meaning analysis can occur without the need of data collection, and therefore allowing the project timeline to focus on discovery rather than collection.

The proposed project will evaluate current research into socioeconomic issues in New York, and make use of public data to analyse a specific hypothesis with the aim to improve current policies. The project will look at a range of socioeconomic factors, such as wealth, transportation and health.

1.2 Aims

The overarching aim of the project is to identify areas of improvement in New York City's governing policies that will benefit the majority of citizens. The research will focus on equity and deprivation; identifying the level of accessibility to public resources and whether there exists a relationship between equitableness and socioeconomics.

This research will utilise the abundance of open data sets available online such as NYC Open Data [15]. These datasets will provide the foundations to derive meaningful insights and develop recommendations to improve the socioeconomic environment of New York.

Figure 1: Map of New York Boroughs [16]

1.3 Stakeholders

The project will have two main stakeholders. The first will be the project supervisor, Dr. Matthew Leeke, as he will be invested in the success of the results, and will be actively contributing to the direction of the research throughout the duration of the project. Secondly, the researcher, Emma Dutton, will also be a stakeholder as she hopes the findings of the research will be rewarding and also beneficial to the wider community.

1.4 Document Roadmap

This document will expand on the project evolution and forecasted work. The preliminary research and project scope will be discussed in section two, with considerations given to related work on analysis of civic data. In section three, the research methodology will be outlined, explaining how the project will be broken down into smaller aims. Section four will illustrate the requirements of the project, both functional and non-functional, which will be referenced again in the evaluation phase. The legal, ethical, social and professional issues will be discussed in section five, which will be of upmost importance due to the sensitive nature of the research. The management of the project will be explained in section six, touching on software development methodology and project timeline. Section seven is an particularly important, where the success measurement and testing strategies of the project will be discussed with note being taken of the holistic approach to the topic. Finally, section eight will conclude the document and round up the many topics considered.

2 Research

To understand the many complexities involved in the socioeconomics of a city, a wide range of research has been conducted since the problem conception. In order to structure the preliminary research, it has been broken down into categories of contextual research, socioeconomic research and related work.

2.1 City Context

New York is split into five boroughs: Manhattan, the Bronx, Queens, Brooklyn and Staten Island. New York residents constitute a range of ethnicities, with 33% being white, 26% Hispanic, 26% black and 13% Asian, according to the 2010 decennial census [12]. When understanding socioeconomic factors affecting the lives of New York citizens, it is useful to group people into communities where there are similarities. Research shows that "majority white neighbourhoods have, on average, the highest average income, share of college educated residents, and home ownership rates" [10].

2.1.1 Governance of New York City

New York City is split into the five boroughs of the Bronx, Manhattan, Queens, Brooklyn and Staten Island. These boroughs are governed by the New York City Mayor's Office [11]. Within each borough there is also a borough president who will work with the mayor to define policies for their specific borough.

2.2 Socioeconomic Research

The aim of this project is to produce statistical analysis that produces better results than policies that already exists, however an understanding of socioeconomics is needed to appreciate the need for such results. 23% of people living in New York have suffered severe health problems [21] and when coupled with the issue of no public health care system, this can be a debilitating issue for many Americans. Poverty In New York is also on the rise, with the poorest areas having 44% of people living in below the federal poverty line [17]. These metrics will prove useful in the analytics stage of the project, and provide a metric to test possible solutions against.

2.3 Related Work

There is a lot of related work on the analysis of civic data. One such article is "Visual Exploration of Big Spatio-Temporal Urban Data: A Study of New York City Taxi Trips" [7] explains how the use of publicly attained data about taxi trips can be used to create data-driven analytics that "improve the lives of citizens through evidenced-based decision making". This paper will be used as a foundation to this project, as it has a lot of similarities which lend themselves to the NYC Open Data sources. Additionally the paper "Identifying urban crowds using geo-located Social media data: a Twitter experiment in New York City" [2] will be utilised to inform the choice of

algorithms when analysing the open source data, as it illustrates interesting density-based clustering techniques. Finally, the book "Visual Analytics of Movement" [1] will prove useful in the data visualisation section of work, and therefore this piece of work will prove beneficial.

The above sources will aid the direction of the project, however, not a lot has been researched in using open source data to inform specific government policies. This project will look at the holistic socioeconomic impact of current policies, and use the open source data in conjunction with the techniques gained from the related work above. Hopefully, this will provide insights into how policies can be changed to benefit the citizens of New York.

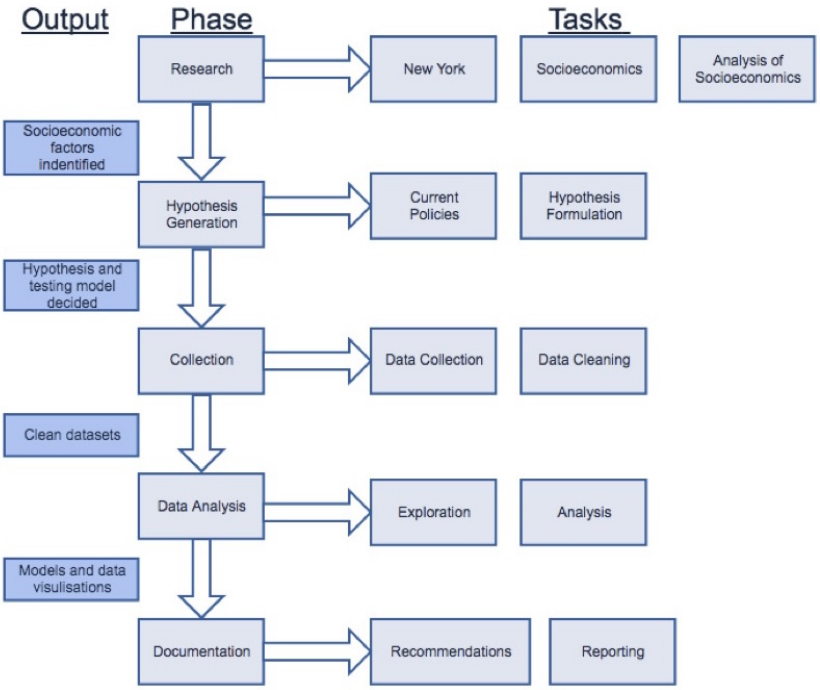


Figure 2: Methodology Phases

3 Methodology

As this topic lends itself strongly to a research project, a well-devised research methodology will be crucial to produce useful findings.

3.1 High Level Overview

The problem will structured using a top down approach, by analysing a set of specific social issues. After some preliminary research, a group of socioeconomic issues will be selected and subsequent relevant datasets will be identified. This data will then allow models to be built to explore a variety of possible hypotheses. The results of such findings will aid the direction of the project and lead to further computational analysis. Finally, the analytics will be visualised and recommendations will be drawn to improve current policies governing New York City, aiming to improve social mobility and equity for citizens.

3.2 Project Phases

The project was broken down into phases to simplify the methodology and allow for a streamlined workflow. Each phase has specific tasks and provides an output that the next phase requires to start. This design is based loosely around the waterfall method and lends itself to the project due to the large research content.

3.2.1 Research

The project will begin by looking at current work surrounding a large scope of areas. Firstly, research into the city of New York will be beneficial in order to learn about the day-to-day issues, such as transport, government and municipal laws. This will provide a strong base to identify areas of inequality in state facilities such as parks, schools, or links to transport systems. In parallel, research will be conducted into socioeconomics topics such as state welfare, health and poverty. These findings will be useful in later stages to identify a suitable hypothesis to evaluate. Finally, research of socioeconomic analysis will provide exposure to relevant work using analytics to identify and improve social and economic issues.

3.2.2 Hypothesis Formation

After collecting valuable resources through the research phase, it will then be possible to identify a specific claim to evaluate. This will require a firm understanding of current government policies in New York City and therefore time will be devoted to exploring this. Once the background has been investigated, a hypothesis can be generated with the goal of finding new ways to improve the current environment to the benefit of citizens. A test case will be developed to test the solution to the proposed hypothesis following the principles of test driven development.

3.2.3 Resource Collection

At this point, the project will be focused on a set of metrics identified in the hypothesis formation section that can be explored using online datasets. Time will be taken to identify a selection of related data to be used in later analysis. Additionally, the data will be refactored into a common type.

3.2.4 Analysis

The majority of set up will now have been completed for the project, and analysis can begin using the clean data from the previous phase. Data exploration will look at correlations between metrics and try possible analytical algorithms and machine learning techniques. From this proof of concept, further analysis will take place to produce concrete findings. These findings will be tested thoroughly using the testing strategy decided in the hypothesis formation phase.

3.2.5 Reporting

Finally, all the results will be collated and recommendations will be drawn. The report will evaluate the impact these recommendations would have on New York and discuss how successful the project has been.

4 Project Requirements

The project will aim to fulfil the following requirements. These requirements will be used to track progress and measure the success of the project in its evaluation stage.

4.1 Functional Requirements

F1: *The system should accept open datasets as input.*

F2: *The system should create a model based on the inputted datasets.*

F3: *The system should produce graphical visualisations of data using mapping software.*

F4: *The system should produce results that will inform policy.*

F4: *The system should be comprehensible to a data scientist.*

4.2 Optional Functional Requirements

OF1: *The system should combine data from a range of sources.*

OF2: *The system should utilise social media data.*

OF3: *The system should make predictions of how policy change would effect conditions.*

4.3 Non-functional Requirements

NF1: *The system should follow the licensing agreements of open source data.*

NF2: *The system should be maintainable.*

NF3: *The system should be testable.*

NF4: *The system should be extendable to alternative data sources.*

4.4 Hardware and Software

The projects main stages will be data discovery, analysis, and reporting. These tasks will be implemented on a Mac OSX, as the processing power will be adequate for the size of the datasets used.

The visualisations will utilise current graphical representation and business intelligence technologies, such as Carto [9] and Qlikview [20]. These industry standard programs allow large datasets to be correlated and displayed professionally, to allow insights to be identified.

Python [19] will be used to analyse the datasets and identify trends, as it has access to mathematical libraries and data structures such as pandas. This functionality will be beneficial in the analytics phase of the project.

4.5 Expected Challenges

As previously noted, the data analysis will be computed using OSX. Through preliminary research, it was identified that the visualisation software Qlikview is only available to use in Windows. This will require additional set up time to configure a virtual machine to a Windows platform.

Another challenge will be the learning curve of understanding the appropriate data analytics techniques. The specifics of the data analytics cannot be forecasted yet as data collection has not begun; so extra time has been added to the Gantt chart to accommodate investigating data analytics more thoroughly.

5 Legal, Ethical, Social and Professional Issues

To ensure the project is carried out to the highest standard of professionalism, consideration has taken place to address the legal, ethical, social and professional issues surrounding the subject. These issues are important to address as the analysis of socioeconomic data may cause sensitive issues to arise. To assist in understanding good practise, the British Computing Society's Code of Practise was acknowledged and discussed [8].

5.1 Legal Issues

All data used in this project will be obtained freely from the Internet and abide by the Terms of Use set out by the New York government [14]. This means that the data will not be used for any illegal purpose, for example to commit a crime, or engage in any conduct that would result in civil liability.

5.2 Ethical Issues

The project will be touching on many social and ethical issues, however they will be treated with upmost professionalism and transparency. As the data has been already been published, it will be expected that due diligence has taken place, and results are anonymous. No attempt will be made to identify individuals. The analytics software created will not be used to spread virus' or malware and will not be loaded onto any other machines other than for testing purposes with the consent of the machine owner.

5.3 Social Issues

The datasets chosen for analysis will be fully representative of all social issues, as categorised by the United Nation's publication of global issues [23]. These issues include but are not exhaustive to, children, environment, food, governance and health. The project will aim to explore these issues in a scientific and unbiased way to produce results that are representative in the context of New York.

5.4 Professional Issues

As this work will evaluate a range of existing work, all relevant literature will be referenced for further reading. During the software development process, the details of all data sources will be available and findings will be reported in the scope of the project to avoid data bias. As data is collected through the NYC Open Data [15] portal, all due diligence will be taken to ensure that datasets will not be used unless they contain reliable, transparent and useful data.

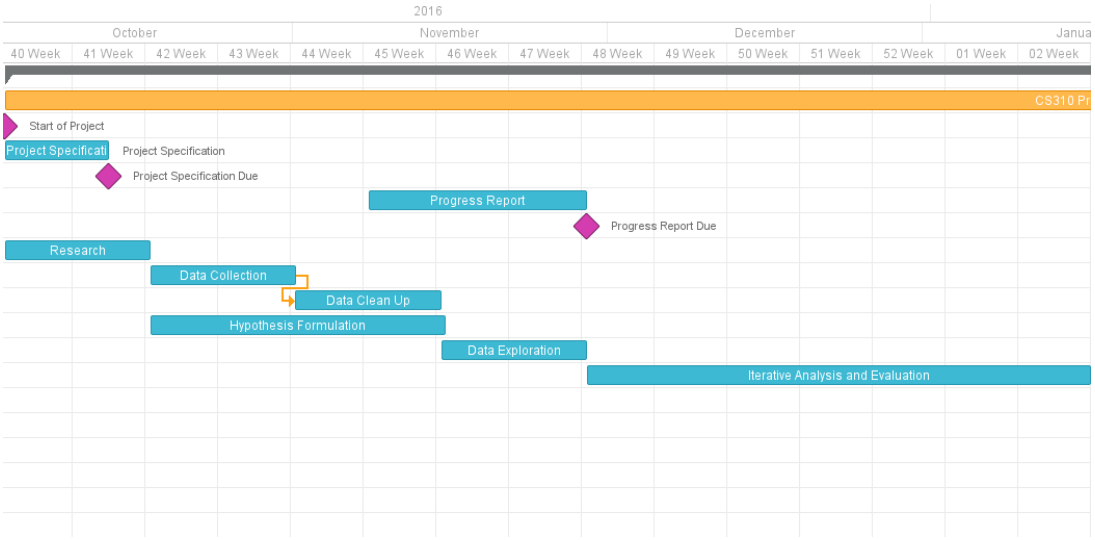


Figure 3: Project Gantt Chart Oct - Jan

6 Project Management

Due to the size of the project, it will been useful to break down the overarching aims into smaller goals, as illustrated below. This will allow the researcher to focus on specific goals at a given point in the project, ensuring time can be taken for reflection at each stage for discussions about the projects direction.

6.1 Software Development Methodology

As the evolution of the project depends on data exploration and analysis, an agile approach will be taken to the software design phase. This is contrasting to the research methodology, which will be more structured and follow a waterfall approach. The advantage of being agile in writing the analysis software will be the ability to improve after iterations and the rapid pace of development. To abide by good software design principles, thorough documentation will be maintained to improve the code standard (ref).

6.2 Project Timeline

To allow the project to run successfully, the following timeline is proposed.

The project will be split into phases highlighted in the methodology section and run between October and April. The first two weeks will be spend on researching related work on New York, socioeconomics and civil analytics. Following this, time will be taken to find the correct datasets for analysis, and then refactoring them to the right format. The results of the research can be evaluated to identify the scope of the project. This will allow data exploration and analysis to

begin. Additionally, the timeline also marks the deadlines for the following deliverables; project specification, progress report, presentation and the final report.

6.3 Project Organisation

In addition to the phases highlighted above, a weekly meeting with the project supervisor will be arranged to discuss current developments and future direction. A Wordpress blog has been created [6] to document these meetings, and log the work of the researcher. This blog will be beneficial when writing the final report, and serve as a reminder of how the project unfolded.

The project will utilise a variety of productivity aiding software. Dropbox [4] will be used to organise the files shared between the researcher and supervisor, and Github will be used to store all of the software components. The final report will be written in LaTeX as it is an industry standard text editor, combined with BibTex to organise references.

7 Testing

During the analytics phase, software will be created with a variety of functions to analyse the datasets and testing these functions will be an integral part of the development lifecycle. In this section, some of the methods of testing will be explored. The analytics software created will be referred to as the system, and will be tested in the following ways.

7.0.1 Unit

Each function written will be unit tested. This will be managed by creating a set of sample data. This data will be passed into the function, allowing an assertion to be made if the outcome agrees with what was expected. This will allow small bugs to be identified before new functionality is integrated with the working solution.

7.0.2 Integration

Integration testing will take place to ensure that different parts of the system will work together without throwing errors. A top-down testing approach will be taken, started at the top of the system and traversing down the program hierarchy towards its branches.

7.0.3 System

System testing will be undertaken to confirm that the program fulfils the specified requirements documented in the Project Requirements section. It will assess the scalability of the system, using datasets of different sizes to see whether any errors are thrown.

7.1 Success Measurement

The testing method outlined above will allow the functionality of the software to be evaluated. However, there are more considerations that will define the success of the results due to the holistic nature of the project, therefore quantifying it's success requires additional steps. A successful solution will be one that informs current policy and benefits a group of people, which is a harder metric to asses. To address this, predictive software will be used to forecast the success of the new proposed policies and evaluate how much benefit citizens of New York gain from the changes proposed by this project. Additionally, reflection will be used to explore the resilience of the results, through seeking advice from domain experts to provide their perspective on the outcome of the project.

8 Conclusion

This document has looked at many aspects of the project life cycle, such as methodology, project requirement and testing. These stages will all be important to fulfil the aim of deriving valuable insight for policy change from open data. This document emphases the large research component, as the ability to create good analytical software will depend on a firm understanding of current governing policies and local municipal laws. Once this understanding in theory has taken place, then the software can be developed with the aim to produce results that will benefit the lives of New York residents, with the intentions to improve the socioeconomic environment of New York.

References

[1] Gennady Andrienko, Natalia Andrienko, Peter Bak, Daniel Keim, and Stefan Wrobel. *Visual Analysis of Movement*. Springer, 2013.

[2] Mohamed ben Khalifa, Rebeca P. Diax Redondo, Ana Fernandez Vilas, and Sandra Servia Rodriguez. Identifying urban crowds using geo-located social media data: a twitter experiment in new york city. *Springer*, 2015.

[3] A Deshpande and D Riehle. The total growth of open source. *The International Federation for Information Processing*, 2008.

[4] Dropbox. `http://www.dropbox.com`, 2016.

[5] L.J. Duhl and A.K.Sanchez. Healthy cities and the planning process. *World Health Organisation*, 1999.

[6] E. Dutton. Cs310 wordpress blog. `www.cs310ed.wordpress.com`.

[7] Nivan Ferreira, Jorge Poco, Huy T. Vo, Juliana Freire, and Claudio T. Silva. Visual exploration of big spatio-temporal urban data: A study of new york city taxi trips. *IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS*, 2013.

[8] British Computing Society. British computing society - code of practise. `http://www.bcs.org/upload/pdf/cop.pdf`.

[9] Carto. `https://carto.com/`.

[10] Furman Center. `http://furmancenter.org/files/sotc/The_Changing_Racial_and_Ethnic_Makeup_of_New_York_City_Neighborhoods_11.pdf`, 2016.

[11] NYC Gov. `http://www1.nyc.gov/office-of-the-mayor/`.

[12] NYC Gov. `https://www1.nyc.gov/site/planning/data-maps/nyc-population/census-2010.page`, 2010.

[13] NYC Gov. New york census. `http://www.census.gov/quickfacts/table/PST045215/36`, 2016.

[14] NYC Gov. Nyc terms of use. `http://www1.nyc.gov/home/terms-of-use.page`, 2016.

[15] NYC Open Data. `https://nycopendata.socrata.com`, 2016.

[16] NYC Tourist. `http://www.nyctourist.com/map1.htm`.

[17] Phys Org. `http://phys.org/news/2015-11-york-poorest-area-poverty.html`, 2016.

[18] Population Reference Bureau. `http://www.prb.org/Publications/Datasheets/2016/2016-world-population-data-sheet.aspx`.

[19] Python. http://www.python.org.

[20] Qlikview. http://www.qlik.com/en-gb/.

[21] Robin Hood Poverty Tracker. http://povertytracker.robinhood.org/, 2016.

[22] The Economist. Linguistics - say what? http://www.economist.com/node/21528592, 2016.

[23] United Nations. Un global issues. http://www.un.org/en/globalissues/, 2016.

[24] Wikipedia. Demographics of new york city. https://en.wikipedia.org/wiki/Demographics_of_New_York_City#cite_note-3.

[25] J.H. Stark, K. Neckerman, G. S. Lovasi, J. Quinn, C.C. Weiss, M. D. M. Bader, K. Konty, T. G. Harris, and A. Rundle. The impact of neighbourhood park access and quality on body mass index among adults in new york city. *Elsevier*, 2014.