A4: Final Project Preliminary Proposal

Title: Analyzing Service Accessibility and Response Patterns in NYC 311
Author: Sagorika Ghosh
Course: DATA 512: Human-Centered Data Science
Submission Date: November 2, 2025

Motivation and Problem Statement

I spent this summer interning in NYC and during those 2.5 months, I noticed how often people mentioned calling or filing complaints through **311**, whether it was for noise, trash pickup, or heat issues. I used it a few times and was struck by how different the experiences felt depending on where from and how you reported it.

This project builds on that experience. I want to explore **who reports issues, how they report them (via phone, app or web), and how quickly those issues are resolved**. From a human-centered perspective, this is about everyday access to public services since 311 is like the invisible interface between the residents and their city. From a scientific standpoint, it offers a structured way to study how digital reporting channels and geography intersect with urban service response.

Understanding patterns in NYC 311 data could help reveal operational efficiency and equity in visibility and responsiveness. If certain neighbourhoods rely on channels that lead to slower closures or lower follow-up, I feel it's important to surface that issue responsibly. My goal is not to assign blame but to highlight data-driven insights that can inform fairer and more transparent city services.

Ultimately, I hope to learn how complaint patterns and response times vary across boroughs and submission channels, and how to present findings in ways that respect privacy and avoid stigmatizing communities.

Data Selected for Analysis

Source: NYC Open Data Portal
Dataset: 311 Service Requests from 2010 to Present
Dataset Link: NYC 311 Service Requests
Dataset ID: erm2-nwe9

Summary:
Each row is a 311 service request. Key fields include created_date and closed_date, complaint_type and descriptor, agency, borough, latitude and longitude, open_data_channel_type, and a unique_key. These fields will let me study timing, place, and channel in a direct way.

License and terms:
The dataset is available under the NYC Open Data Terms of Use. I will cite the source, link to the dataset page, and record the date and query used for my snapshot.
Terms:https://www.nyc.gov/main/terms-of-use, https://opendata.cityofnewyork.us/open-data-law/

Why does this dataset fit my goal?
It is public, current and very large, so there is enough signal for stable summaries. My quick probes also showed strong coverage in 2024. About 98% of records have closed_date and proper coordinates. This would support resolution time analysis and safe spatial aggregation at the borough or community district level. The topic also connects to my own summer experience in the city, which would help me keep the work grounded and real.

Ethical considerations:
There are several ethical considerations in using the NYC 311 dataset. First, the data reflects who reports an issue, not who experiences it, which introduces ***reporting bias*** across neighbourhoods and demographics. The dataset contains location and text fields that could raise privacy concerns, so I will exclude free-text data and only display results in aggregated form, such as by borough or community district. Maps or language that compare neighbourhoods can unintentionally create stigma, so I will describe patterns carefully and avoid labeling any area as “problematic.”

Unknowns and Dependencies

There are a few factors outside my control that could affect the pace or scope of this project. The NYC 311 dataset is extremely large, with over 40 million records, so downloading or exporting it directly from the portal can fail. To manage this, I plan to use the Socrata API and work with 2024 year data in smaller batches, while also keeping a sample in my repository for a smooth notebook execution. Another potential issue is schema drift. Since field names or values might change over time, I’ll handle this by checking the schema early and documenting the exact snapshot date in my README file. The API also has rate limits that could slow down data collection, so I’ll try to use an app token and pagination to stay within those limits. Finally, because the 2025 data is still being updated, I’ll focus my analysis on the complete 2024 year to ensure consistent and reliable results.

A5: Final Project Plan

Submission Date: November 10, 2025

Research Questions

1. How do average 311 response times differ across boroughs in 2024?
    a. How do median and mean response times compare across boroughs, and what patterns emerge in their distributions?
    b. How do these response time patterns relate to complaint volume in each borough?

2. Which ZIP codes have statistically significant response time problems, and do certain complaint types cluster in slow-responding neighborhoods?
    a. Do response times vary significantly across ZIP codes within the same borough, based on statistical testing?
    b. Which ZIP codes consistently show the slowest responses, and do they form geographic clusters?
    c. Do certain complaint types cluster within slow-responding ZIP codes, and are some areas slow across multiple complaint types?

3. Do digital channels lead to faster responses than phone reports, and which complaint types rely most on phone reporting?
    a. For the same complaint type within the same borough, do digital channels (App/Web) receive faster responses than phone submissions?
    b. Which complaint types rely most heavily on phone reporting versus digital channels?
    c. Do phone-dominant complaint types experience systematically slower closure times?

4. Does complaint volume or seasonal workload amplify delays, or do some boroughs and complaint types remain slow even during low-demand periods?

5. Can we predict whether a complaint will be slower than the norm for its issue, and which structural factors drive that risk?

These questions are focused for a short research window but I believe it still connects to a bigger theme of access and fairness in city services. I want to understand whether using a digital channel actually helps people get quicker responses, and if those benefits reach all the neighborhoods equally.

Background and Related Work

Below are few of the articles/papers that have explored various aspects using the 311 NYC Service Request data:

1. **Equity in 311 Reporting: Understanding Socio-Spatial Differentials in the Propensity to Complain (Kontokosta et al., 2017)**
   Shows that complaint data reflect who chooses or is able to report, not just where problems occur. The authors estimate "propensity to complain" for heat and hot water by predicting likely violations and comparing to observed 311 reports. This warns me not to treat raw counts as true incidence and to analyze channels and geography with care.

2. **Structure of 311 service requests as a signature of urban ecology (Wang et al., 2017)**
   Compares NYC, Boston, and Chicago and finds that mixes of complaint types track neighborhood context and socioeconomic patterns. This supports my choice to compare like with like within complaint categories, rather than mixing categories across places or channels.

3. **How socio-spatial disparities in 311 complaint behavior bias smart city governance (Kontokosta & Hong, 2021)**
   Documents that complaint behavior varies with demographics and income, which can bias any downstream "smart city" use of 311 data. This backs my plan to present borough-level aggregates, to run tests within complaint types, and to avoid claims about true problem incidence.

4. **Estimating reporting bias in 311 complaint data (Annals of Applied Statistics, 2025)**
   Provides a formal approach to under-reporting for NYC heat and hot water issues. The takeaway is that "no complaint" often means "not reported," not "no problem." I will be explicit that my study is about reporting and response patterns, not the full incidence of issues.

5. **NYU press release: Tool to estimate 311 under-reporting (2025)**
   Summarizes an automated modeling tool the City can use to assess under-reporting of heat and hot water complaints. I am not using the tool, but it reinforces the caution above and motivates careful language in my findings.

6. **NYC 311 Language Access Implementation Plan (2024, PDF)**
   Describes interpretation and translation support for 311, including phone interpretation. It shows that channel usability and language access can differ across phone, web, and app. That is relevant for who uses which channel and how that might relate to response times.

7. **NYC 311 testimony to City Council on technology and access (2024)**
   Notes how the mobile app connects to live agents with Language Line and outlines ongoing improvements to digital access. I treat these as context for why channel choice could vary across communities.

These findings help me shape my plan and inform my study design and questions in the following ways:

- I will analyse 2024 data only and aggregate at the borough level to reduce the risk of neighbourhood-level stigma and to reflect the known socio-spatial reporting differences. This also aligns with prior warnings about bias and context as the study would be on an aggregated borough level. (Refer to 3)
- I will compare response times across channels within the same complaint categories so that category mix does not get masked as a channel effect. (Refer to 2)
- I won't interpret counts as the true incidence of problems. Under-reporting and channel accessibility mean I'm just studying patterns in reporting and response, not the full universe of issues. (Refer to 4)

Methodology

Data Scope & Plan at a glance:

Scope the data, pull a clean 2024 subset, build derived fields, do comparisons within complaint types and boroughs, test differences by channel, fit a small model, and report results with clear limits.

Data Cleaning:

The raw 2024 data has several cleaning steps:

1. Date Conversion: Convert created_date and closed_date from strings to datetime format
2. Response Time Calculation: Calculate response_time_days = closed_date minus created_date
3. Invalid Record Removal: Remove records with missing created_date, closed_date, or borough, and records with negative or zero response times (data entry errors)
4. Channel Standardization: Standardize open_data_channel_type to three values:
   • PHONE to Phone
   • ONLINE/UNKNOWN/OTHER to Web
   • MOBILE to App
5. Duplicate Removal: Remove duplicate records using unique_key
6. Outlier Treatment: Winsorize response times at the 99th percentile (273.21 days) to limit outlier influence
7. Complaint Type Selection: Select top 10 complaint types by volume for reliable within-category comparisons

Analysis Methods

Descriptive Statistics: I will calculate median, mean, standard deviation, minimum, and maximum response times by borough, channel, and complaint type. I will focus on medians because response times are heavily right-skewed.

Statistical Testing (RQ2): To statistically validate the observed disparities in response times across boroughs, I will employ the Kruskal-Wallis H-test, a non-parametric alternative to ANOVA suitable for non-normally distributed data. Significant findings might indicate that at least one borough's response time distribution differs from the others.

Predictive Modeling (RQ5): I will build a Random Forest classifier to understand what factors are associated with slow service. I will define the target as a binary label based on whether response time exceeded the global median, and include features such as complaint type, borough, channel, and time of submission. I will evaluate the model using the ROC-AUC score and use feature importance to identify which predictors most strongly influence delays. This will help distinguish between procedural bottlenecks and potential structural patterns.

Visualizations: To explore patterns visually, I will create bar charts showing borough-level averages and medians, box plots illustrating distribution and spread, geographic maps built with Plotly to show spatial patterns of response times, and interactive maps that group complaints into small grid cells for clearer visualization of dense regions.

Why Are These Methods Appropriate?

Median-based analysis provides a more stable representation of response patterns when outliers are present. Non-parametric tests like Kruskal-Wallis are well suited for non-normal service data, and the model choice allows handling interactions between multiple features without heavy assumptions. Comparing channels within the same complaint type avoids misleading cross-type comparisons since different complaints have inherently different resolution times. Borough-level analysis maintains privacy while still revealing useful insights, and ZIP-level results are used cautiously. Winsorizing the top 1 percent of values limits extreme outlier influence while preserving distribution shape, and focusing on top complaint types ensures that comparisons are based on reliable sample sizes.

Expected Findings

I expect I will see small raw gaps in response times across channels and across boroughs. I will treat those as a starting point only. After comparing channels within the same complaint types and adding a simple model that controls for borough, I expect most gaps to shrink. If any adjusted differences remain, I will report their size with confidence intervals and describe them as associations, not causes. Because time to close often has long right tails, I will summarize with medians and show sensitivity to outliers. If the distribution looks roughly symmetric, I will switch to means and standard parametric tests. Overall, I will describe patterns in reporting and response for 2024 rather than claim the true incidence of problems.

Why does this study matter?

311 is the front door to city help, so who uses which channel and how quickly cases close is a real access issue, not just a data question. If web and app reports move faster, and those channels are used unevenly across boroughs or languages, residents can experience different speeds for the same kinds of problems. By comparing channels within the same complaint types and controlling for place, this study will separate signal from mix effects and point to practical fixes. It could give clearer guidance on which channel to use for specific complaints, stronger language support in digital tools, smarter phone triage, and targeted outreach where phone is still the main entry point. The focus still stays on fairness and usability rather than judging neighborhoods, and the workflow will be simple to repeat for later years or other cities.