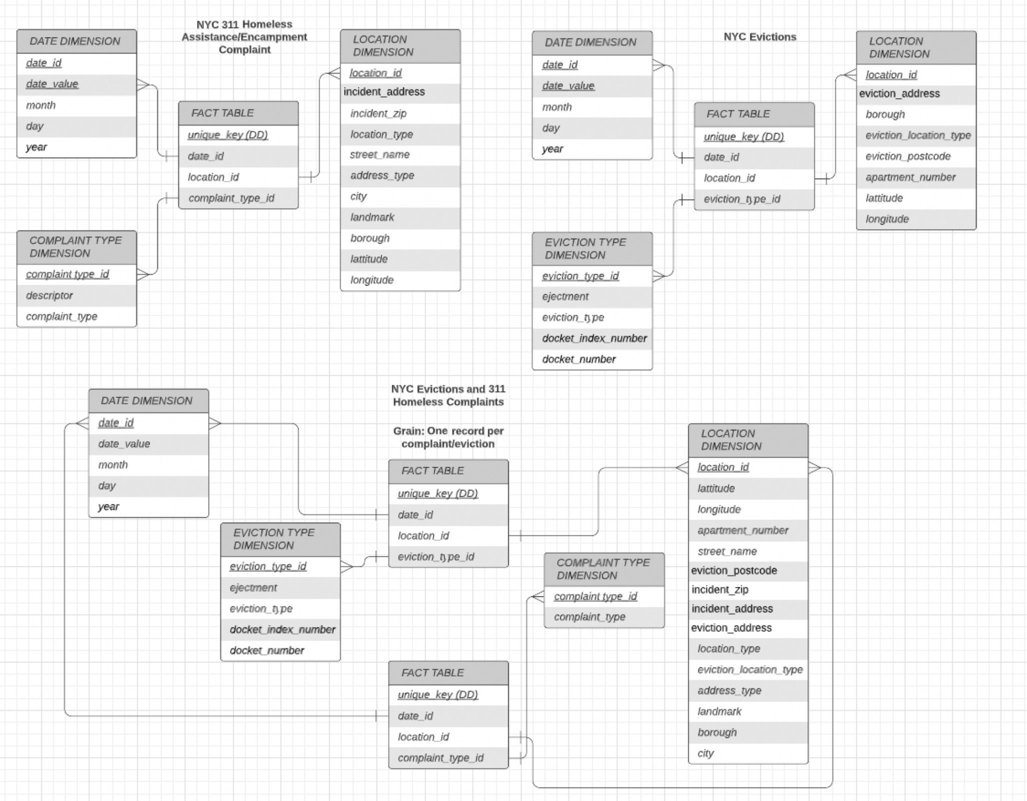
INTRODUCTION

The organization that our data warehouse would be created for would be Coalition for the Homeless, the nation's oldest advocacy and direct service organization helping homeless individuals and families in New York City. The issue we hope to address is the increase in homelessness in NYC. In recent years, homelessness in New York City has reached the highest levels since the Great Depression of the 1930**s,** according to an article published by Coalition for Homelessness in 2022. Our source data consists of NYC Open Data: 311 Service Requests from 2010 to Present, filtered for "Homeless Encampment" and "Homeless Assistance". This dataset contains 311 service request data from 2010 to present, and can be found here. Our second dataset used is also from NYC Open Data, which contains pending, scheduled and executed evictions. In the data warehouse, we will use the 311 data regarding "Homeless Encampment" and "Homeless Assistance" to understand the patterns of 311 complaints. We then hope to compare this data to eviction rates in NYC. By understanding the factors that may be contributing to this increase in homelessness, we may understand how to better help the homeless population, with the goal of slowing down the rate. Our key performance indicators that we will use to assess the data include: the number of complaints per year, the number of evictions per year, number of complaints per location (zipcode), evictions per location (zip code), complaints by location type (Subway, street, etc), eviction type (residential or commerical), and number of complaints and evictions by location (borough).

I. Dimensional Model Diagram



II. ETL Processes

To complete the ETL we first downloaded the data from NYC open data using API Socrta. Then, we completed Data Profiling where we noticed some missing data. We initially wanted to include “descriptor” from the 311 complaint data, but contained 94.2% missing values, so we decided to exclude it. After that, we uploaded our data to Google BigQuery. Finally, we sourced that data from BigQuery to dbt, cleaned the data with staging process, and created the dimension and fact tables.

The code for the project is hosted on GitHub https://github.com/Woys/homelessness_in_nyc_dbt

1. Data Profiling.

Homeless Encampment Data.

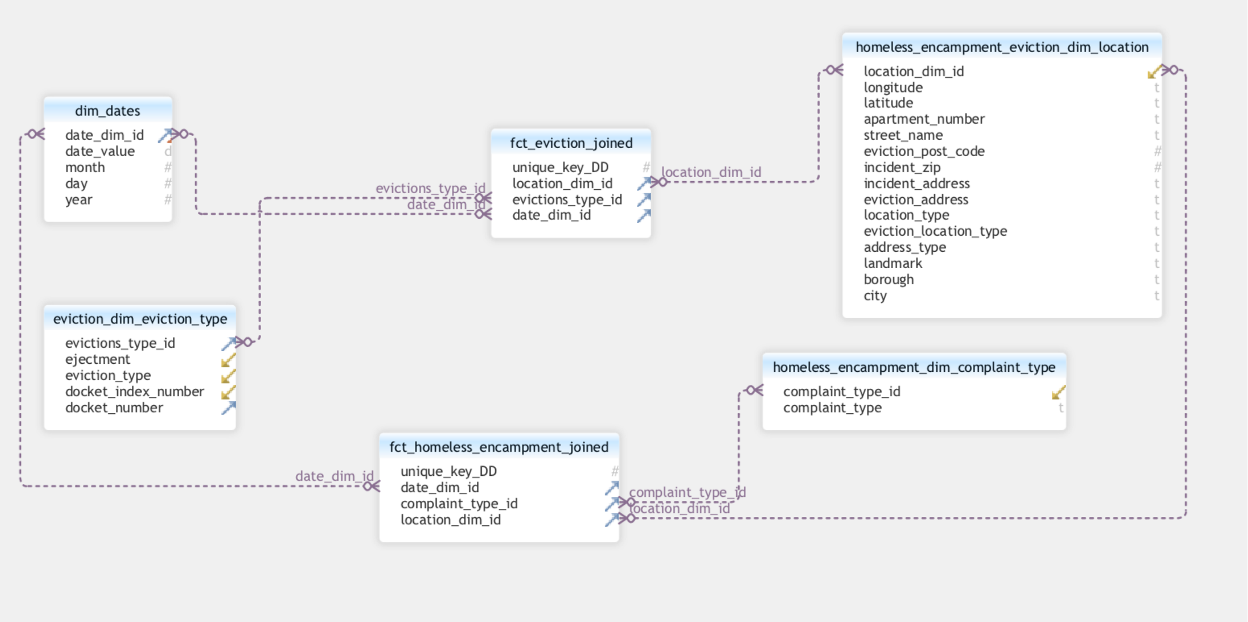
descriptor has 237473 (94.2%) missing values.
location_type has 60838 (24.1%) missing values.
incident_zip has 24893 (9.9%) missing values.
incident_address has 47121 (18.7%) missing values.
street_name has 47135 (18.7%) missing values.
address_type has 8133 (3.2%) missing values.
city has 26362 (10.5%) missing values.
landmark has 175884 (69.8%) missing values.
latitude has 3096 (1.2%) missing values.
longitude has 3096 (1.2%) missing values.

Eviction Data

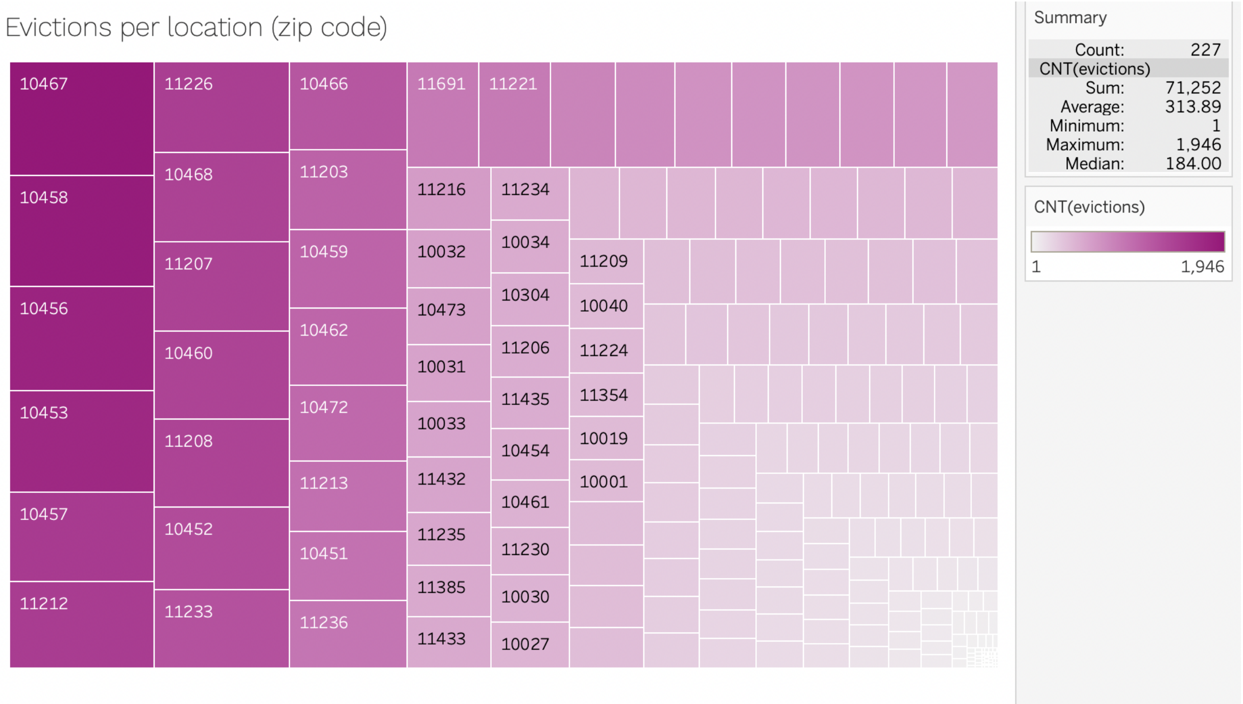
eviction_apt_num has 11913 (16.7%) missing values.
latitude has 6845 (9.6%) missing values.
longitude has 6845 (9.6%) missing values.

Later, this infurmatin will be used for cleaning step.
In this step, we also chose dates based on our data, '2017-01-03' AND '2021-07-10'.
This time period contains in both datasets.

III. Final Dimensional Schema



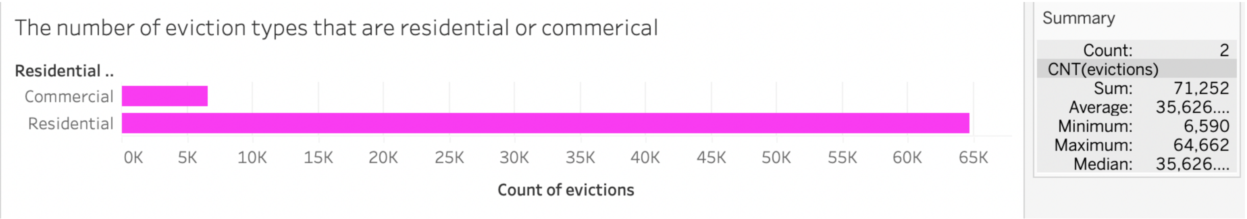
IV. KPI Visualizations



Summary

Count:	227
CNT(evictions)	
Sum:	71,252
Average:	313.89
Minimum:	1
Maximum:	1,946
Median:	184.00

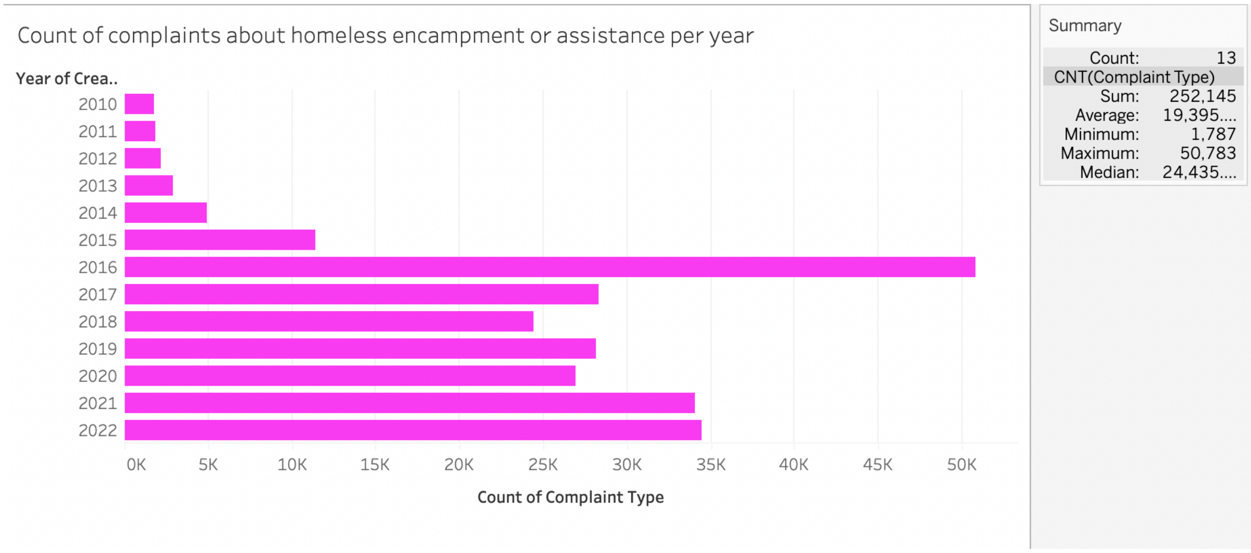
- The top 5 zip codes that have the most evictions are all located in the Bronx, NY.



Summary

Count:	2
CNT(evictions)	
Sum:	71,252
Average:	35,626....
Minimum:	6,590
Maximum:	64,662
Median:	35,626....

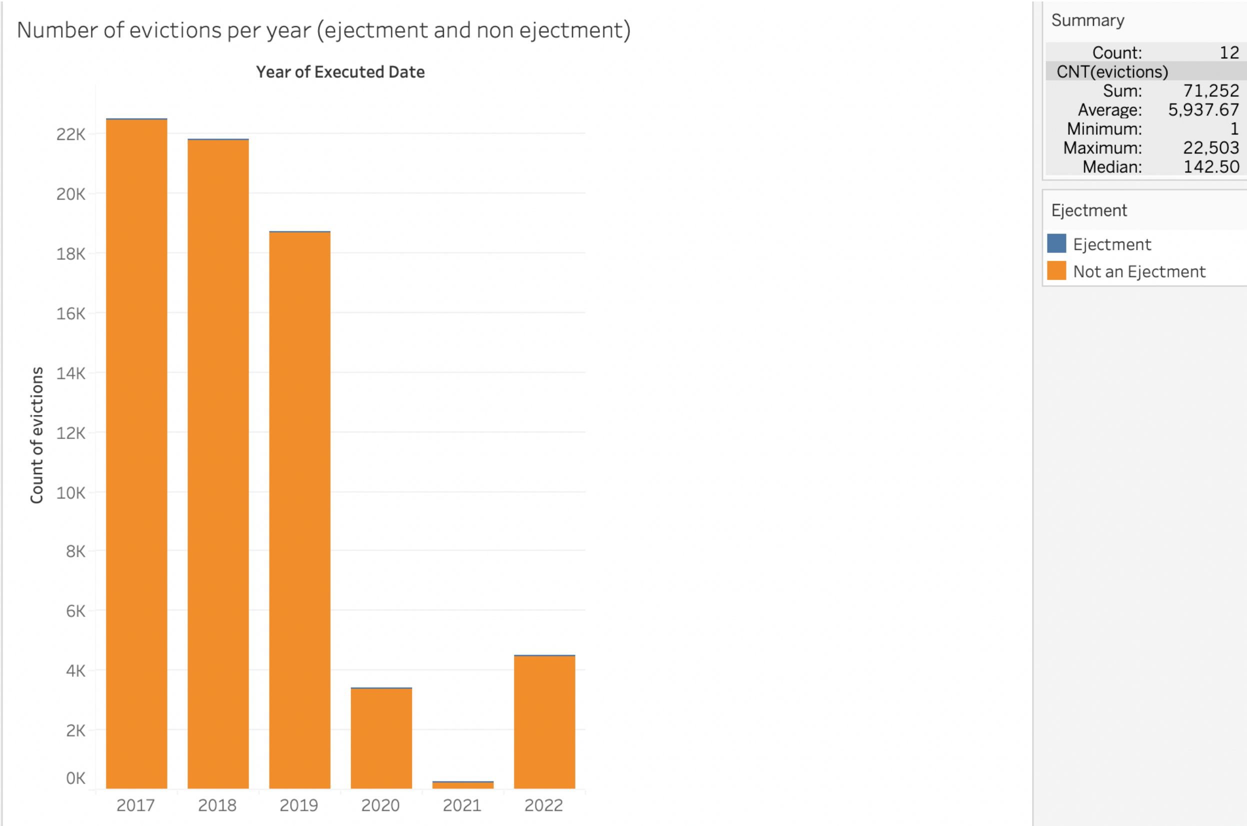
- There are almost 13 times more residential evictions than commercial in NYC.



Summary

Count:	13
CNT(Complaint Type)	
Sum:	252,145
Average:	19,395....
Minimum:	1,787
Maximum:	50,783
Median:	24,435....

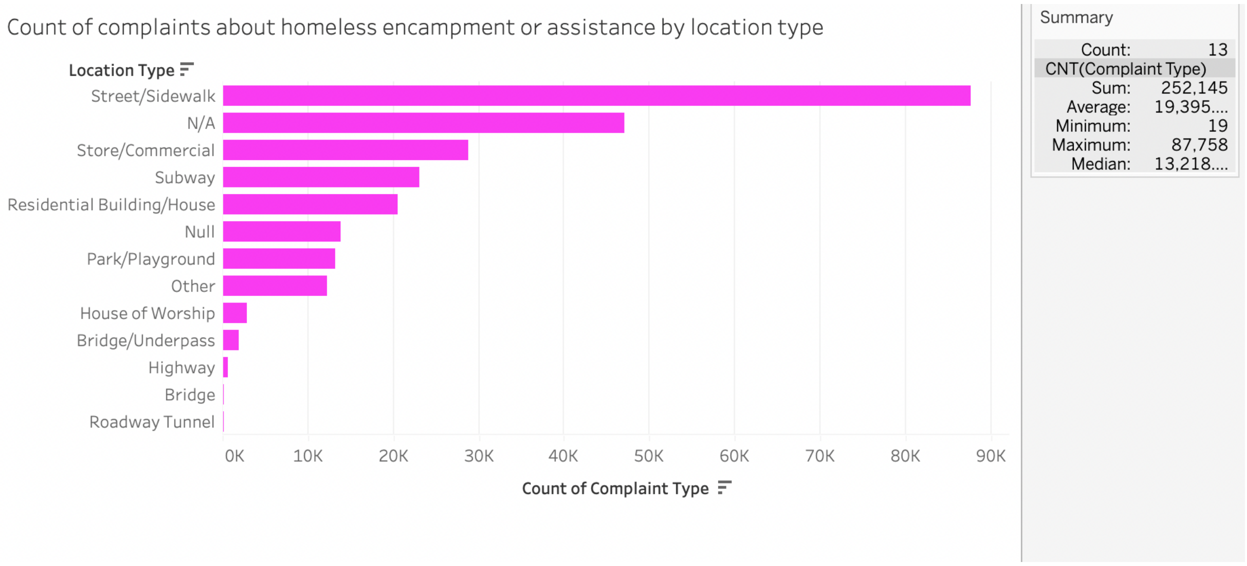
- Almost 35,000 calls to 311 concerning homeless assistance/encampment in 2022 alone.



Summary

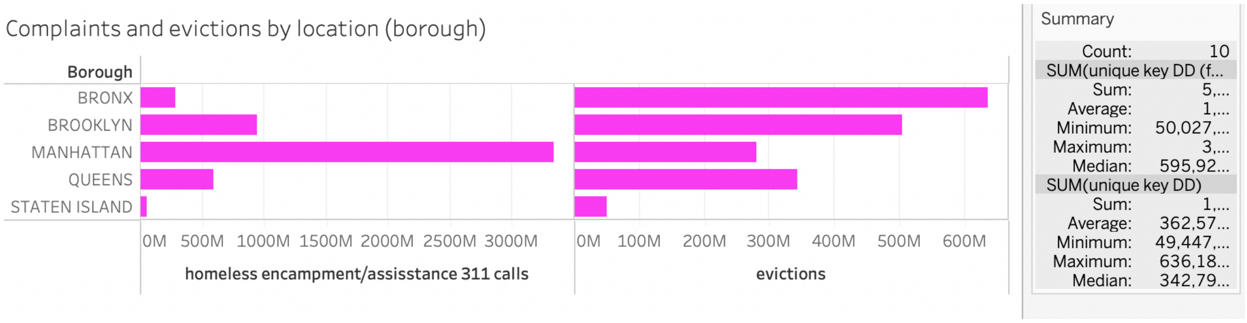
Count:	12
CNT(evictions)	
Sum:	71,252
Average:	5,937.67
Minimum:	1
Maximum:	22,503
Median:	142.50

- The number of evictions had been decreasing since 2017, but are starting to rise again. Most evictions are non-ejecment evictions.



Summary

Count:	13
CNT(Complaint Type)	
Sum:	252,145
Average:	19,395....
Minimum:	19
Maximum:	87,758
Median:	13,218....



Summary

Count:	10
SUM(unique key DD (f...	
Sum:	5,...
Average:	1,...
Minimum:	50,027,...
Maximum:	3,...
Median:	595,92...
SUM(unique key DD)	
Sum:	1,...
Average:	362,57...
Minimum:	49,447,...
Maximum:	636,18...
Median:	342,79...

- The Bronx has the 2nd lowest amount of complaints made to 311 about homeless assisstance/encampent, yet the highest eviction rates. Manhattan has the highest homeless encampment/assisstance 311 complaints, yet the 2nd lowest eviction rate.

V. Tools Used

- LucidChart (dimensional model)
- Google BigQuery (target dbms)
- Github (hosts the ETL code)
- Dbt (ETL tool)
- DbSchema (dimensional schema)
- MySQL (ETL programming language)

VI. Conclusion

a)

- LucidChart (dimensional model)
  - We used LucidChart to create the dimensional model to plan the structure of our data warehouse.
- Google BigQuery (database)
  - We used Google BigQuery as our target DBMS.
- Github (ETL coding)

        - We used github to write SQL queries and share between the team.
    - DbSchema (dimensional schema)
        - We used DbSchema to create the final dimensional model for our data warehouse.
    - DBT
        - We used DBT as our ETL tool.
    - Tableau
        - We used Tableau for our dashboard application and made visualizations.
    - MySQL (ETL programming language)
        - Used MySQL to carry out the ETL programming.
    - Google Docs
        - We used Google docs to to document our process and progress.

b) The easiest part of the assignment was figuring out which 311 complaint was most interesting, and what we could potentially do with that data. It was fun and a chance to be creative. The most difficult part of the assignment was turning the dimensional model into a data warehouse using SQL. Aside from the general hardships that generally come with programming, it was a challenge to decide what to do with nulls and missing data. It was also challenging creating the visualizations; learning how to use Tableau to get generate knowledge through our dashboard application and visualizations that addressed our KPIs was challenging but successful.

c) In summary, we extracted insightful knowledge through our KPIs. Concerning the evictions data, we were able to see which zipcode has the most evictions, which location types are more commonly evicted, and the yearly trend in number of evictions per year. The Coalition for the Homeless, our intended organization to use our data warehouse, would be able to utilize this data to understand the demand for help with eviction rights, demand for shelter, food, and more necessities that may be challenging for a person facing eviction from their residence to obtain. As for the 311 homeless encampment/assistance data, we were able to see the trend in 311 calls concerning homeless encampment/assistance over several years, as well as the total yearly amounts. We were also able to easily understand which locations are most common for complaints to be made. The Coalition for the Homeless can potentially use this data to understand how many hot meals need to be provided on a daily basis, see which zipcodes need more assistance with providing food/shelter to those struggling, and so much more. Finally, we saw the comparison between evictions and homelessness per location. Although there does not seem to be much of a meaningful correlation as we anticipated between 311 complaints regarding the homeless and eviction rates, the proposed benefits of the data warehouse listed prior can be realized by the new system we developed.

d) In terms of final comments, it can be said that this project was very daunting at first; it seemed impossible that we would be able to create a data warehouse. Yet, week by week we came closer. The most satisfying moment of this project was actually creating the visualizations that addressed the KPIs because the data finally felt digestible. It was real data before our eyes that told a real story about our city. It was quite shocking to see the numbers; we were so shocked that we did research to see if our figures match up with those of the NYC Gov. data-- and they did. It was definitely mind blowing and eye opening. This project has been a positive experience overall.

VII. References

- https://dol.ny.gov/statistics-new-york-city-labor-force-data
    - Labor force data for 2022 in NYC
- https://comptroller.nyc.gov/newsroom/new-york-by-the-numbers-monthly-economic-and-fiscal-outlook-no-65-may-2nd-2022/
    - NYC Labor Markets, NYC Real Estate Markets, NYC Return to Office, NYC Business and Tourism, City Finances
- https://www.bls.gov/regions/new-york-new-jersey/news-release/consumerpriceindex_newyorkarea.htm
    - NY CPI
- https://dol.ny.gov/labor-statistics-new-york-city-region
- https://www.bls.gov/regions/new-york-new-jersey/news-release/consumerpriceindex_newyorkarea.htm
- https://www.data2go.nyc/map/?id=401*36081006300*unemployment_tract!undefined!ns*!other_pop_cd_506~ahdi_puma_1~sch_enrol_cd_112~age_pyramid_male_85_plus_cd_20~median_household_income_puma_397~median_personal_earnings_puma_400~dis_y_perc_puma_102~poverty_ceo_cd_417~unemployment_cd_408~pre_k_cd_107!*air_qual_cd~ahdi_puma*family_homeless_cd_245#17/40.76487/-73.92249