# Urban Traffic

Group Members: Jin Chen, Bosen Li, Juliana Ma, Zhiying Zhu

# Past Two Week Summaries

- Data Preprocessing
  - Clean and format the NYC yellow taxi data
    - Time Period:  March to May for both 2019 and 2020
  - Region matching
    - Taxi zone with zip code and neighborhood
  - Clean the NYC Covid-19 cases by zip code
    - Time Period: 3/31/2020 to 5/17/2020
- Data Visualization
  - Scatter plot for analyzing correlations between different attributes
  - Bar chart for analyzing traffic through different time periods
  - Heatmap for see the traffic in the NYC during selected time periods
- Report
  - Introduction and Related Work

# Data Preprocessing

# Yellow Taxi Trip Data Cleaning

Original Yellow Taxi Trip Data (Jan 2009 – Jun 2020)

- Included Attributes:
  - Pickup and drop-off: Time and taxi zone location id
  - Payments: fare-amount, tip, toll, improvement surcharge, total  amount, payment type
  - Others: Trip distance, Passenger count, Vendor ID

## Data Cleaning

- Find and remove the trip with wrong data
  - Filter Requirements
    - Trip distance must be greater than 0
    - Total cost of the trip also need to be greater than 0
    - Trip drop-off time must be greater than pick-up time
    - Trip duration must be greater than 5 min
    - Only consider payment using cash or credit card
    - All required field should not be missing
    - Trip must be within the required time periods
    - The price per each mile travel must be less than 500

# Yellow Taxi Trip Data Aggregation

- Data Aggregation
  - New Fields for grouping trip by hour and taxi zone
    - **Number of Pickup**
    - **Number of Dropoff**
    - **Number of cash payments**
    - **Number of credit card payments**
    - **Average number of passengers per trip**
    - **Average travel speed(mph) per trip**
    - **Average distance(mile) travelled per trip**
    - **Average total price per trip**
    - **Average price cost per each mile travelled**
- About **51 millions** trips in total

| time | zone | num_pickup | ave_trip_passenger | avg_trip_speed_mph | avg_trip_distance | avg_total_price | Cash | Card | num_dropoff | avg_price_per_mile |
|---|---|---|---|---|---|---|---|---|---|---|
| 2019-03-01 00:00:00 | 4 | 56 | 3.2857142857142856 | 29.776383978256 | 7.1728571428571435 | 38.29285714285714 | 16 | 40 | 106 | 5.338577972515434 |
| 2019-03-01 00:00:00 | 7 | 10 | 3.2 | 26.50234258319495 | 3.4360000000000004 | 18.82 | 6 | 4 | 156 | 5.477299185098952 |
| 2019-03-01 00:00:00 | 10 | 4 | 5 | 55.5201948708521 | 30.52 | 164.33999999999997 | 0 | 4 | 12 | 5.384665792922673 |
| 2019-03-01 00:00:00 | 13 | 70 | 3.4285714285714284 | 41.07962704668685 | 9.050285714285717 | 44.896 | 6 | 64 | 90 | 4.960727364566232 |
| 2019-03-01 00:00:00 | 16 | 2 | 2 | 48.07169529499626 | 17.88 | 72.6 | 2 | 0 | 10 | 4.060402684563758 |
| 2019-03-01 00:00:00 | 17 | 10 | 3.6 | 20.76574219752783 | 3.7560000000000002 | 22.42 | 6 | 4 | 38 | 5.969116080937168 |
| 2019-03-01 00:00:00 | 18 | 2 | 2 | 34.94462540716613 | 5.96 | 29.52 | 0 | 2 | 8 | 4.953020134228188 |
| 2019-03-01 00:00:00 | 24 | 30 | 4.133333333333334 | 24.731196394866856 | 5.050666666666667 | 29.29600000000001 | 14 | 16 | 50 | 5.800422386483634 |

# Covid-19 Data

- Original Dataset
- Clean Data
  - Remove missing value or have negative value for number of cases
- Final Data
  - Contain time period from 2020/3/31 – 2020/5/17
  - Not all NYC zip code has data for each day in above time periods

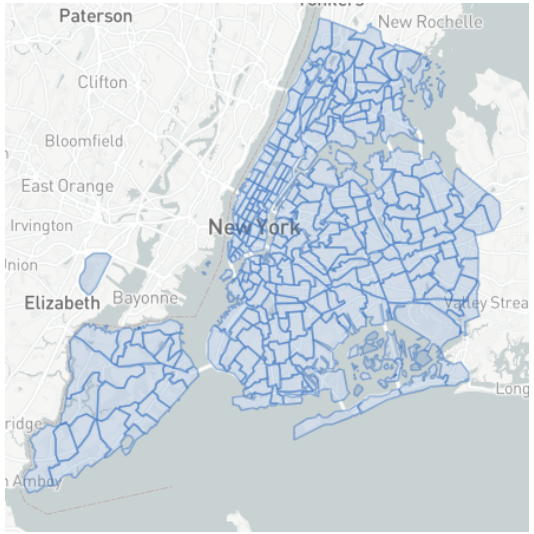
| month | day | zipcode | daily_case | total_case |
|---|---|---|---|---|
| 3 | 31 | 10001 | 113 | 265 |
| 3 | 31 | 10002 | 250 | 542 |
| 3 | 31 | 10003 | 161 | 379 |
| 3 | 31 | 10004 | 16 | 38 |
| 3 | 31 | 10005 | 25 | 81 |
| 3 | 31 | 10006 | 6 | 24 |
| 3 | 31 | 10007 | 26 | 67 |
| 3 | 31 | 10009 | 181 | 450 |
| 3 | 31 | 10010 | 101 | 282 |
| 3 | 31 | 10011 | 222 | 487 |
| 3 | 31 | 10012 | 68 | 183 |
| 3 | 31 | 10013 | 122 | 255 |
| 3 | 31 | 10014 | 140 | 305 |
| 3 | 31 | 10016 | 288 | 581 |
| 3 | 31 | 10017 | 45 | 138 |
| 3 | 31 | 10018 | 66 | 151 |
| 3 | 31 | 10019 | 187 | 451 |

# Region Matching

- Problem
  - Taxi zone and zip code cannot be directly match, as some taxi zone regions are across multiple zip code regions.

- Solution
  - Calculate the centroid location of the taxi zone and zip code region, then match by find the nearest point.

Taxi zone region




Zip code region

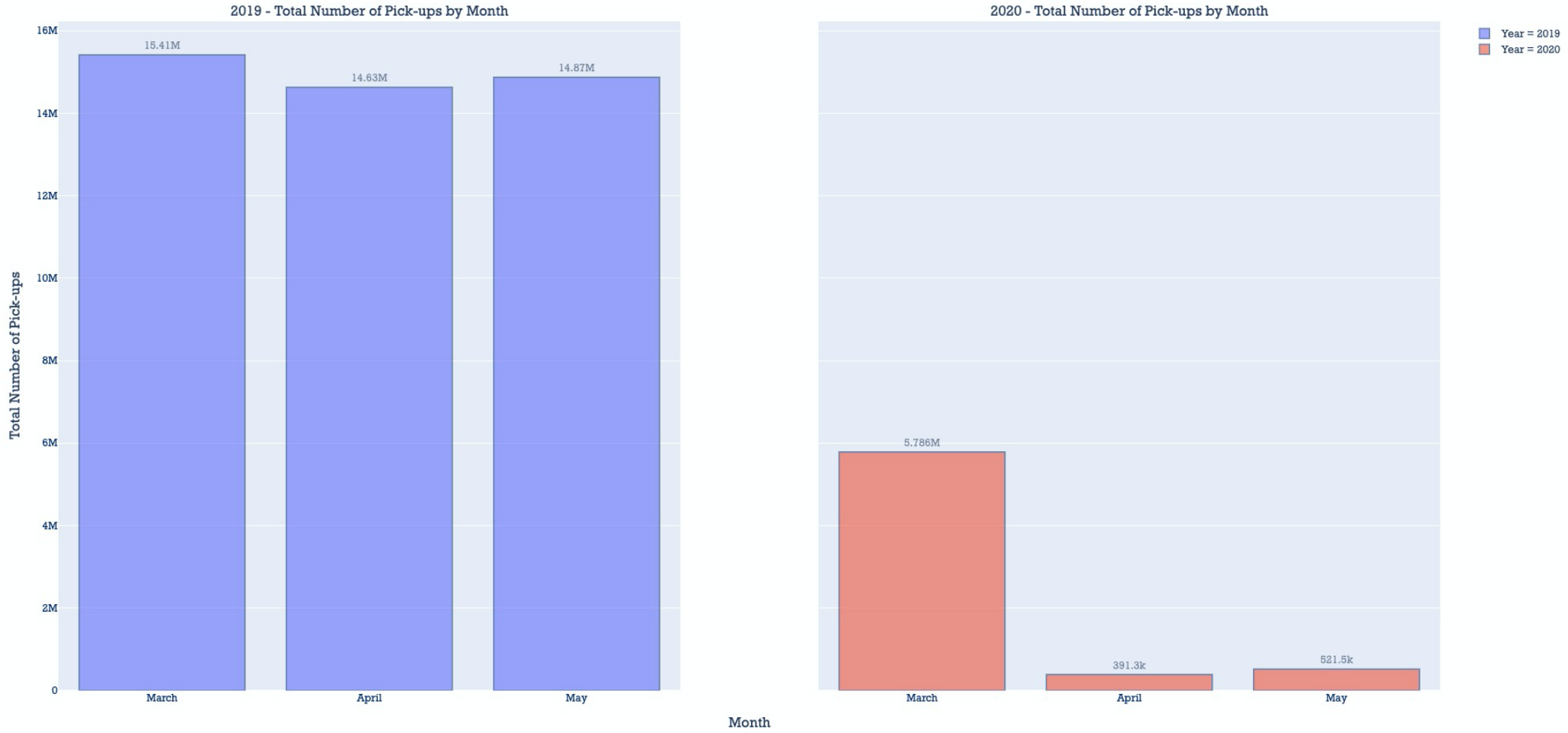
# Zip code region Information

- Find the additional information about the zip code region
  - Neighborhood name
    - From health.ny.gov
  - Population and income information
    - From uszipcode package in python
      - up-to-date database

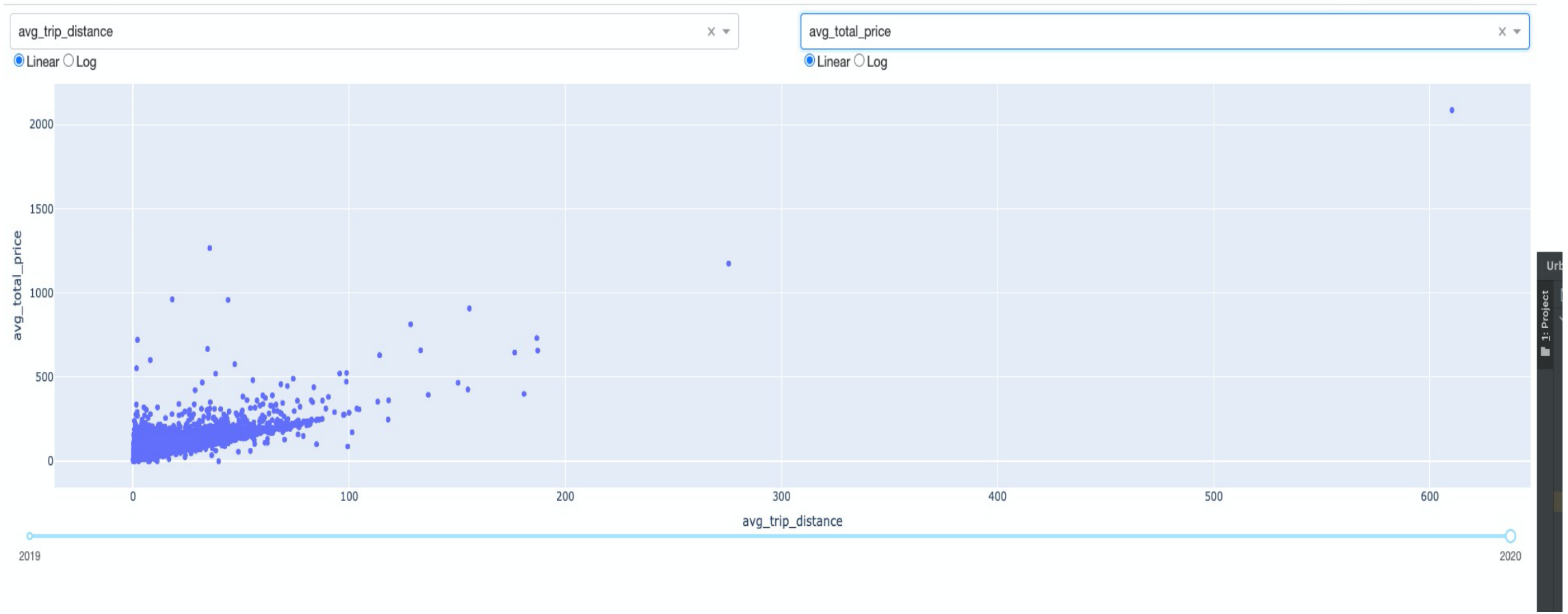
| location_id | zone_name | borough | centroid_x | centroid_y | zipcode | county | neighborhood | median_household_income | median_home_value | population | population_density |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Newark Airport | EWR | -74.174270275862 | 40.69024330172404 | 7114 | Essex County | unknow | 19639 | 253200 | 14748 | 1980 |
| 2 | Jamaica Bay | Queens | -73.8298447614632 | 40.61112041531353 | 11693 | Queens County | Rockaways | 50570 | 323100 | 11916 | 11950 |
| 3 | Allerton/Pelham Gardens | Bronx | -73.8465098633635 | 40.864294038404864 | 10469 | Bronx County | Northeast Bronx | 57776 | 427100 | 66631 | 26903 |
| 4 | Alphabet City | Manhattan | -73.97520908904802 | 40.723853149707395 | 10009 | New York County | Lower East Side | 59929 | 672800 | 61347 | 99492 |
| 5 | Arden Heights | Staten Island | -74.18980260697641 | 40.55667805421754 | 10312 | Richmond County | South Shore | 85324 | 467300 | 59304 | 7723 |
| 6 | Arrochar/Fort Wadsworth | Staten Island | -74.07174675600425 | 40.60111748548471 | 10305 | Richmond County | Stapleton and St. George | 70758 | 415300 | 41749 | 9819 |
| 7 | Astoria | Queens | -73.9193650454889 | 40.7614329365109 | 11103 | Queens County | Northwest Queens | 55129 | 648900 | 38780 | 54537 |
| 8 | Astoria Park | Queens | -73.9233958241691 | 40.77807743974382 | 11102 | Queens County | Northwest Queens | 49924 | 597700 | 34133 | 42255 |
| 9 | Auburndale | Queens | -73.79018835662427 | 40.750302438936394 | 11365 | Queens County | Central Queens | 55492 | 580300 | 42252 | 16923 |

# Data Visualization

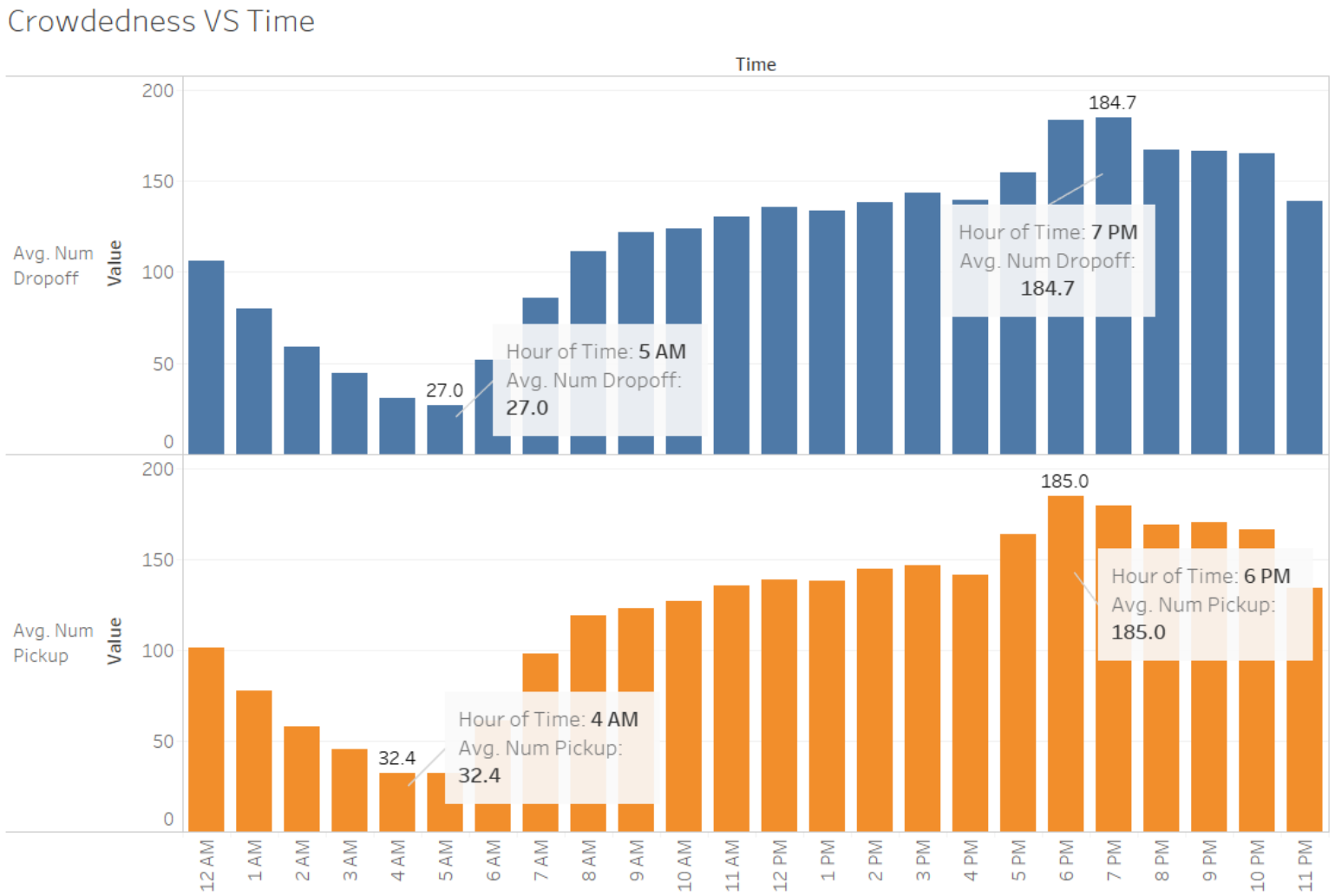
# Quick Comparison of Traffic – 2019 vs 2020



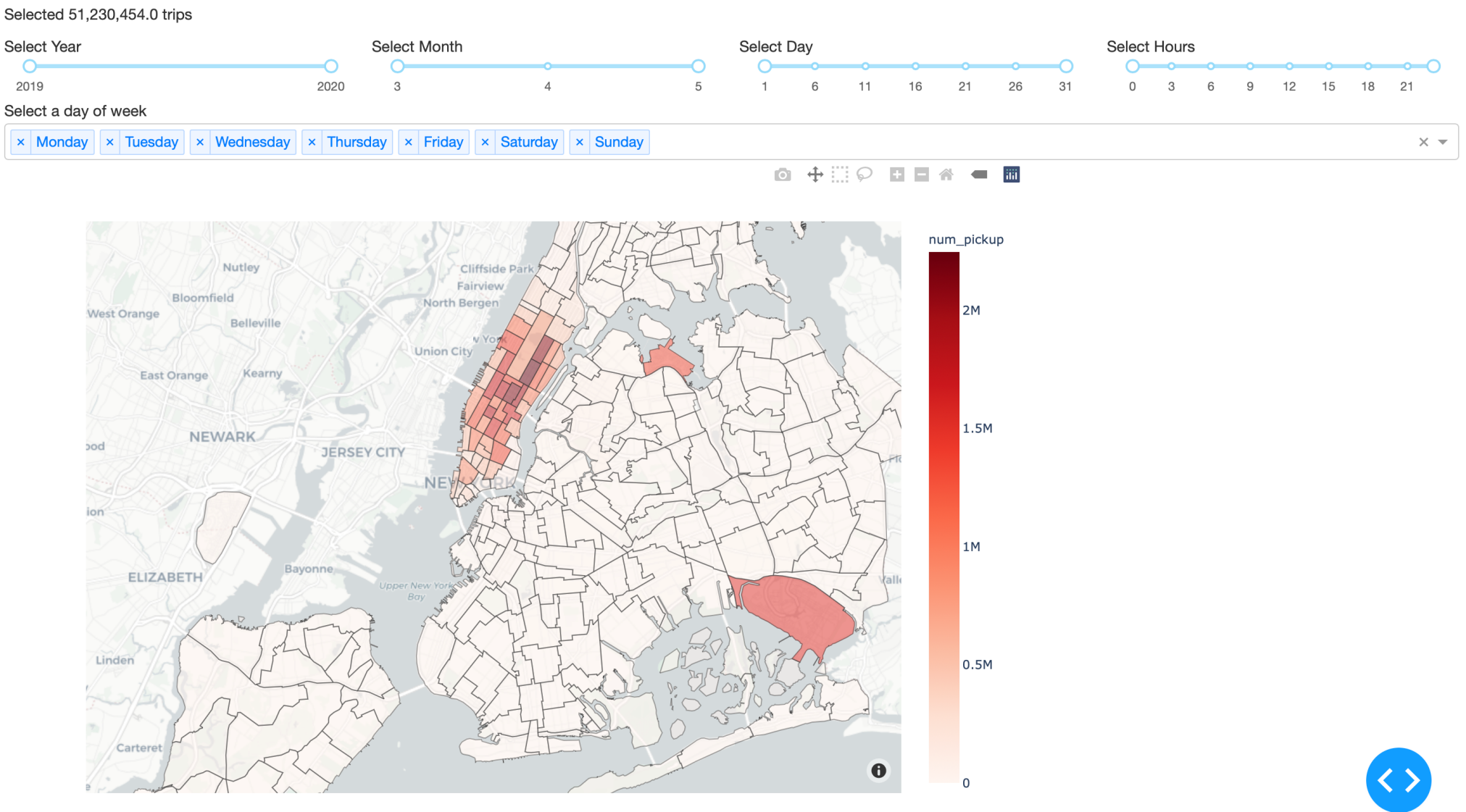
# Attributes Comparison – Trip Distance vs Price



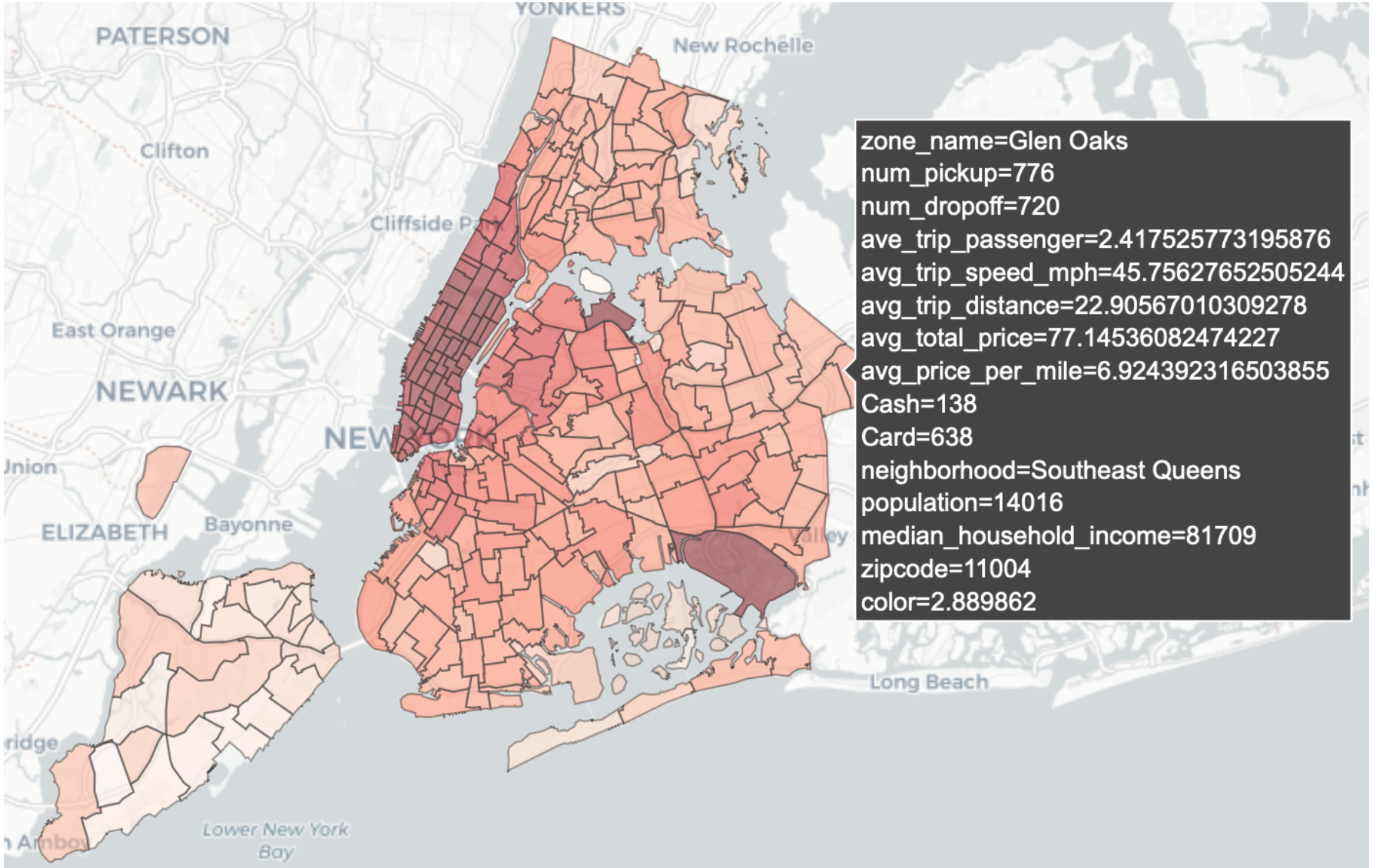
# Traffic through the day



# Heatmap for NYC traffic



# Heatmap for NYC traffic – Log Scale



# Report writing

**Analysis of Covid impacts on New York City urban traffic and taxi routes management based on spatiotemporal and time big data**

Jin Chen, Zhiying , Juliana Ma, Bosen Li

*Abstraction* — The recent COVID-19 pandemic has forced humanity to experience an unprecedentedly expansive lockdown around the world, which resulted in the urban transport systems under a near standstill. In this study, we took five urban areas in the center of New York City as an example, used Python to process, analyze, and used tableau, Python Plotly to visualized the data from 112 millions of NYC taxi data records that can help to study the pre-covid-19 and during covid-19 urban traffic congestion/situation, and then establish [what, such as principle component analysis] model to find out the impact of that shows how the covid-19 has impacted urban traffic flow. Finally, we found that [how many factors], such as [detail factors], have significant impact on road urban traffic congestion, and predicted what traffic

# Future Task Timeline

# This Week Task

- Data Process
  - Inquiry covid-19  data and zip code info based on pass-in requirement -> Jin
- Visualization (Plotly)
  - Compare heatmaps (traffic vs covid-19) -> Jin
    - Adjust map by zipcode
    - Allow selection for different attributes
  - Find and analyze correlation between traffic and region info -> Zhiying
    - Need summaries the findings
  - Find and analyze relationship between traffic and covid-19 -> Juliana
    - Need summaries the findings
  - Exploring current figures to find any relationships -> Bosen & Zhiying & Juliana
- Write Report -> Bosen
  - Write Method Section
    - Include Data Preprocess (e.g., data clean and aggregation)
    - Graphs
      - Purpose and reason of selection figure type

# Next Week Task

- Visualization (Plotly) -> All Team Members
  - Finalized the visualization
    - Add any additional elements if needed
    - Check for any bugs
  - Explore the visualization result

- Write Report -> All Team Members
  - Write result discussion and conclusion, abstract
  - Finalized the report