01 DS-GA 1007 | FALL 2022

NYC Traffic Danger

PLUCKY PYTHONS

02

Questions

Question 1: What were the most dangerous times of year/week/day?

Question 2: What were the most dangerous areas in NYC in 2021?

Question 3: Which areas had the most bike accidents in NYC in 2021?

Question 4: Are e-bikes more dangerous than regular bikes?

03

Data

NYC OpenData Motor Vehicle Collisions

Variables of Interest: Crash Date, Crash Time, Zip Code, Latitude, Longitude, Vehicle Type(5), Contributing Factor(5) Persons Injured, Persons Killed, Cyclists Injured, Cyclists Killed

Cleaning:
Initial Data: 1,946,561 rows
Convert date to pandas datetime object
Filtered dataframe to 2021
Dropped rows with missing location values
Dropped unused columns
All Data: 101,757 rows
Bike Data: 7,331 rows
Filled in all missing zip code and boroughs using geopy API with longitude/latitude data

04

Methodology

Workflow

Question 1: Converted date and time columns into Pandas Datetime index to group number of traffic accidents in specified periods of time and generated line plots for comparison. Compared traffic accidents for various holidays using bar charts.

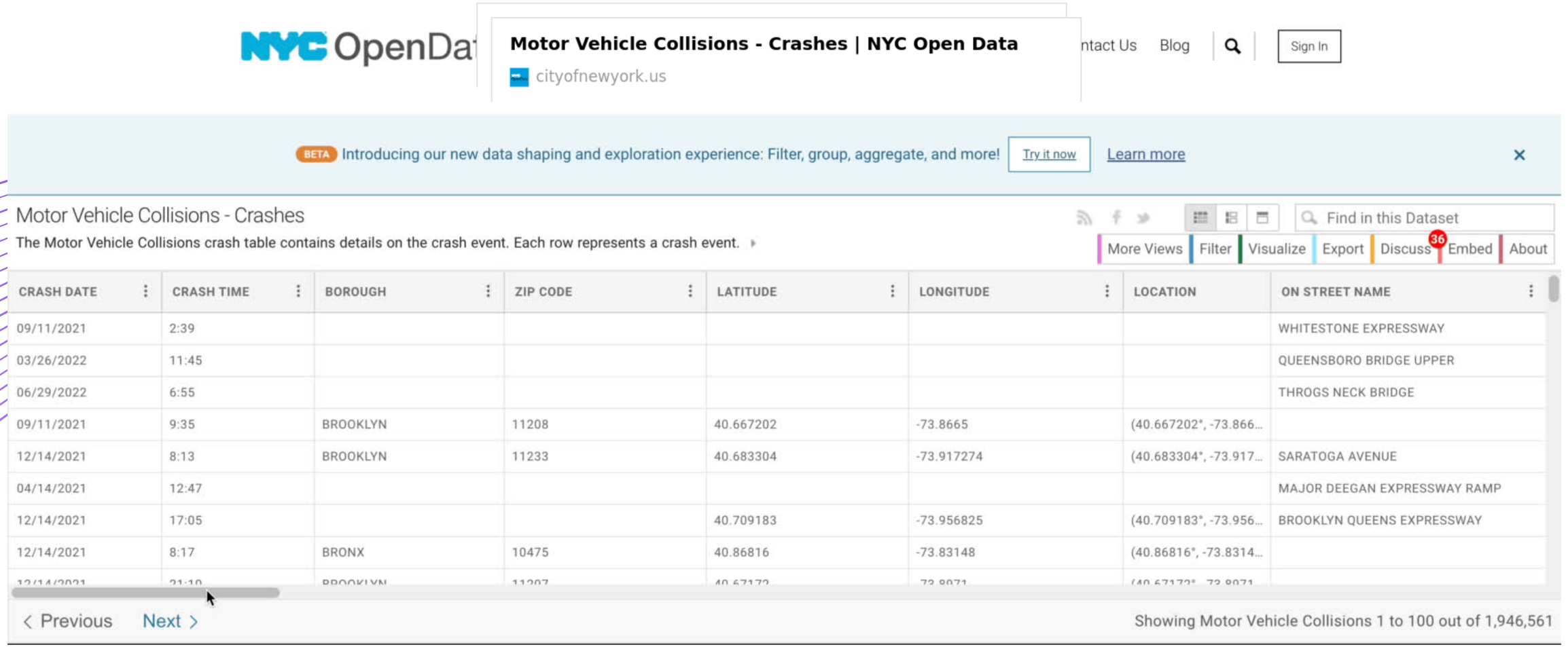
Question 2: Mapped latitude and longitude on folium maps, grouped by zipcode and graphed with geojson, and clustered with DBSCAN to find most dangerous intersections.

Question 3: Divided data by borough and zip code. Included data on bike commuters in each borough to estimate percentages of injuries and fatalities. Used folium to visualize hot spots of collisions with heat map.

Question 4: Extracted data on injuries and deaths for bikes and e-bikes only. Separated the groups and compared them using bar charts and resampling methods.

05

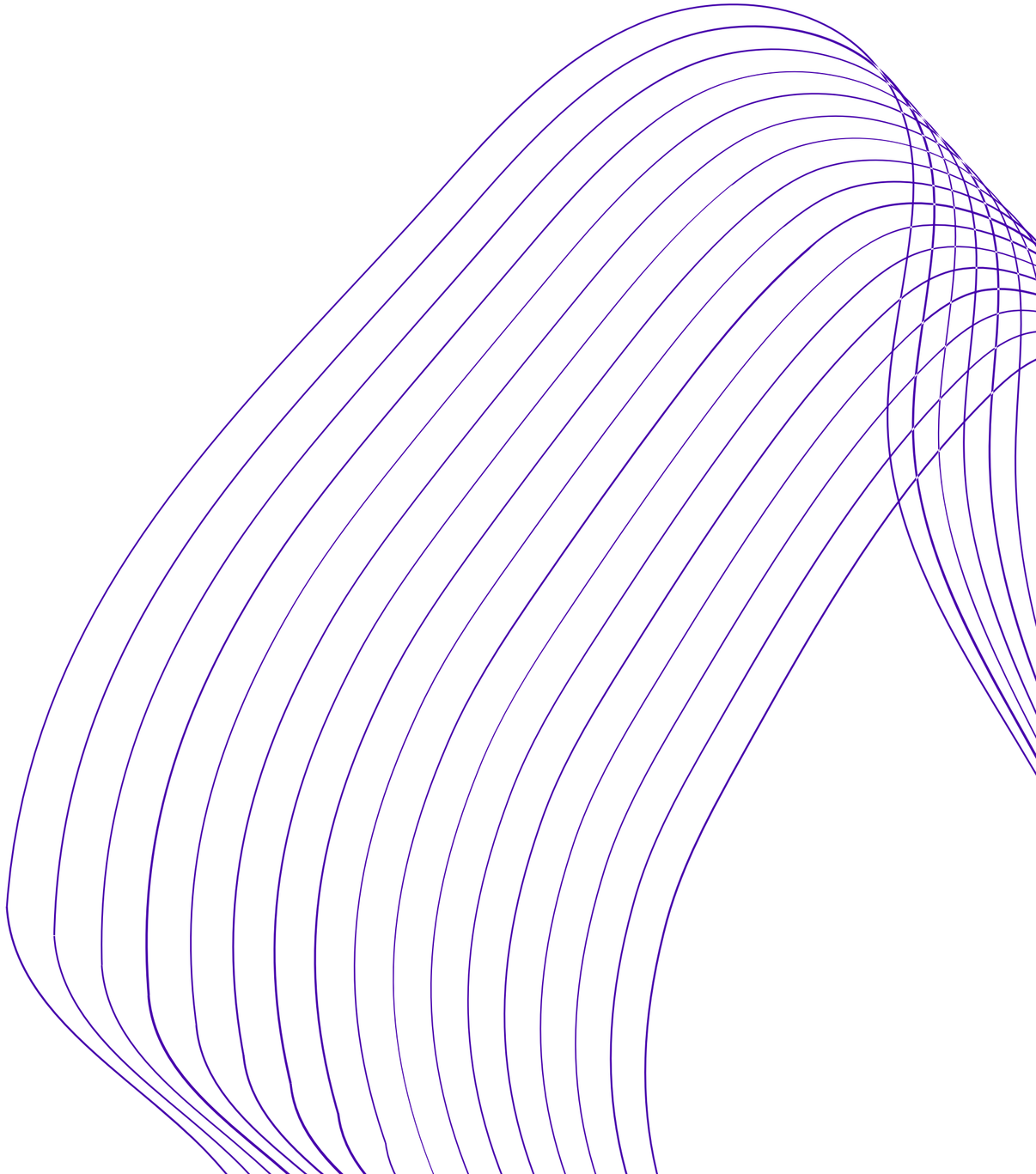
Original Dataset



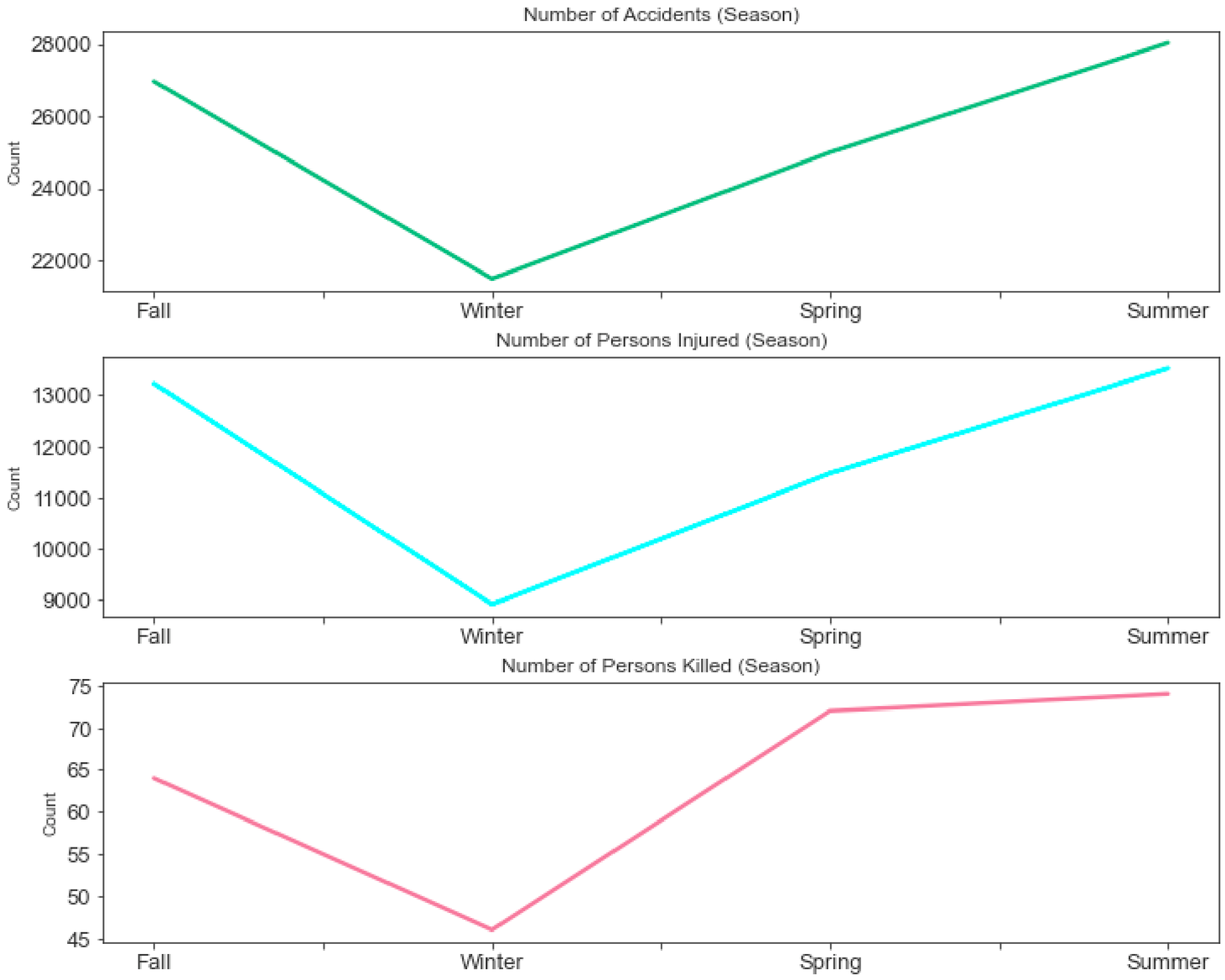
LINK: HTTPS://DATA.CITYOFNEWYORK.US/PUBLIC-SAFETY/MOTOR-VEHICLE-COLLISIONS-CRASHES/H9GI-NX95

What were the most dangerous times of year/ week/day?

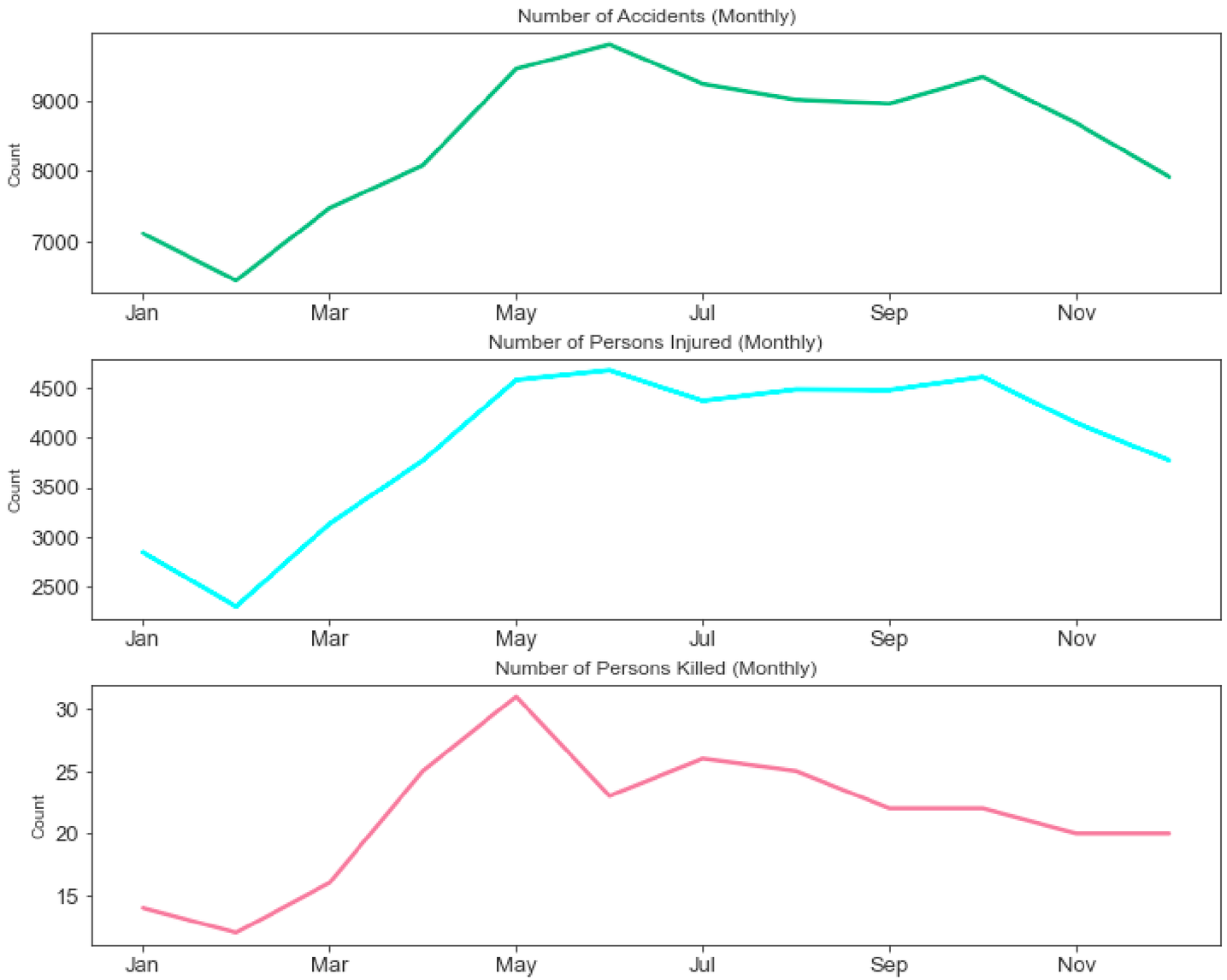
Method: Converted date and time columns into Pandas Datetime index to group number of traffic accidents in specified periods of time and generated line plots for comparison. Compared traffic accidents for various holidays using bar charts.



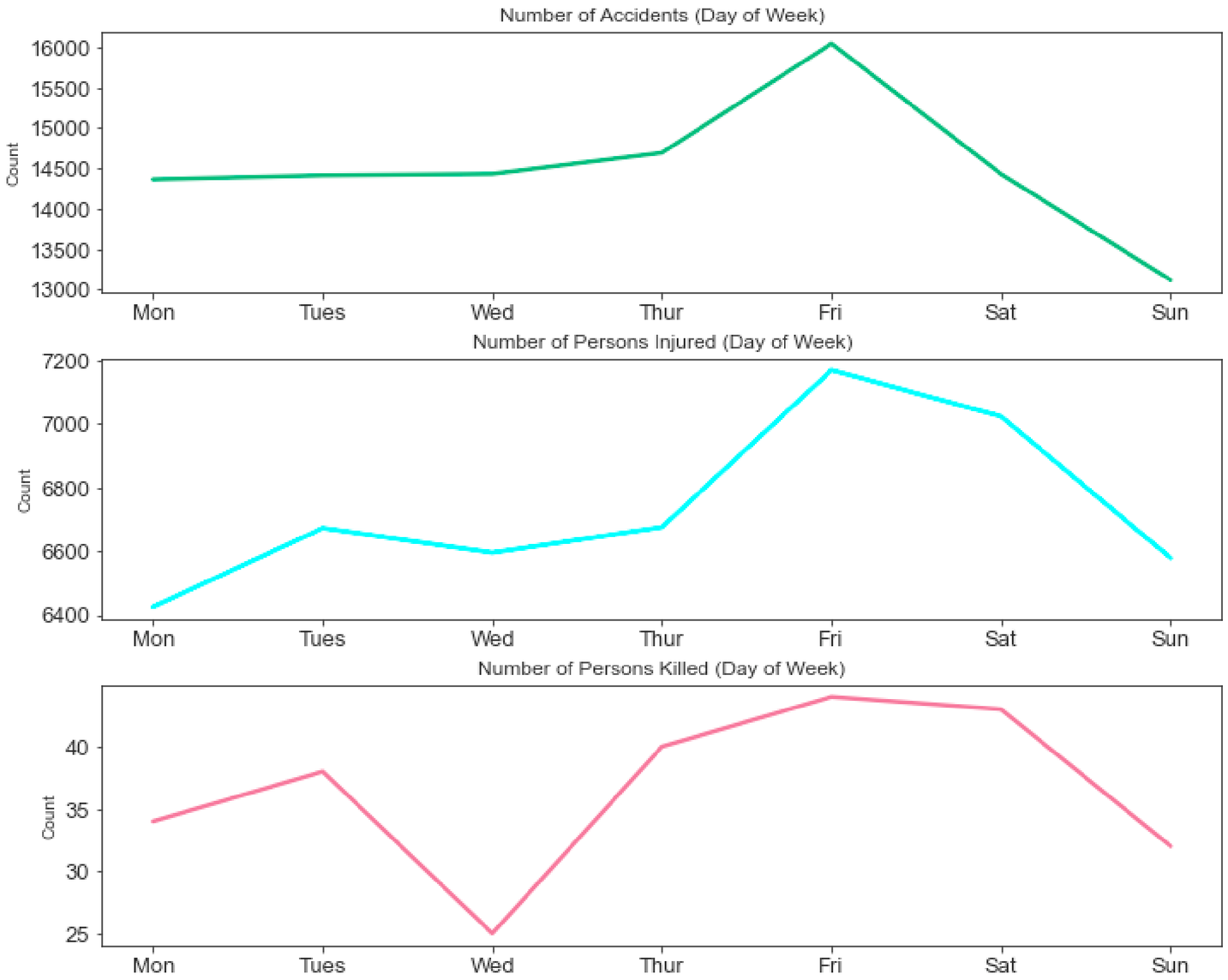
Seasons


Number of Accidents (Season)
Count
28000
26000
24000
22000
Fall
Winter
Spring
Summer
Number of Persons Injured (Season)
Count
13000
12000
11000
10000
9000
Fall
Winter
Spring
Summer
Number of Persons Killed (Season)
Count
75
70
65
60
55
50
45
Fall
Winter
Spring
Summer

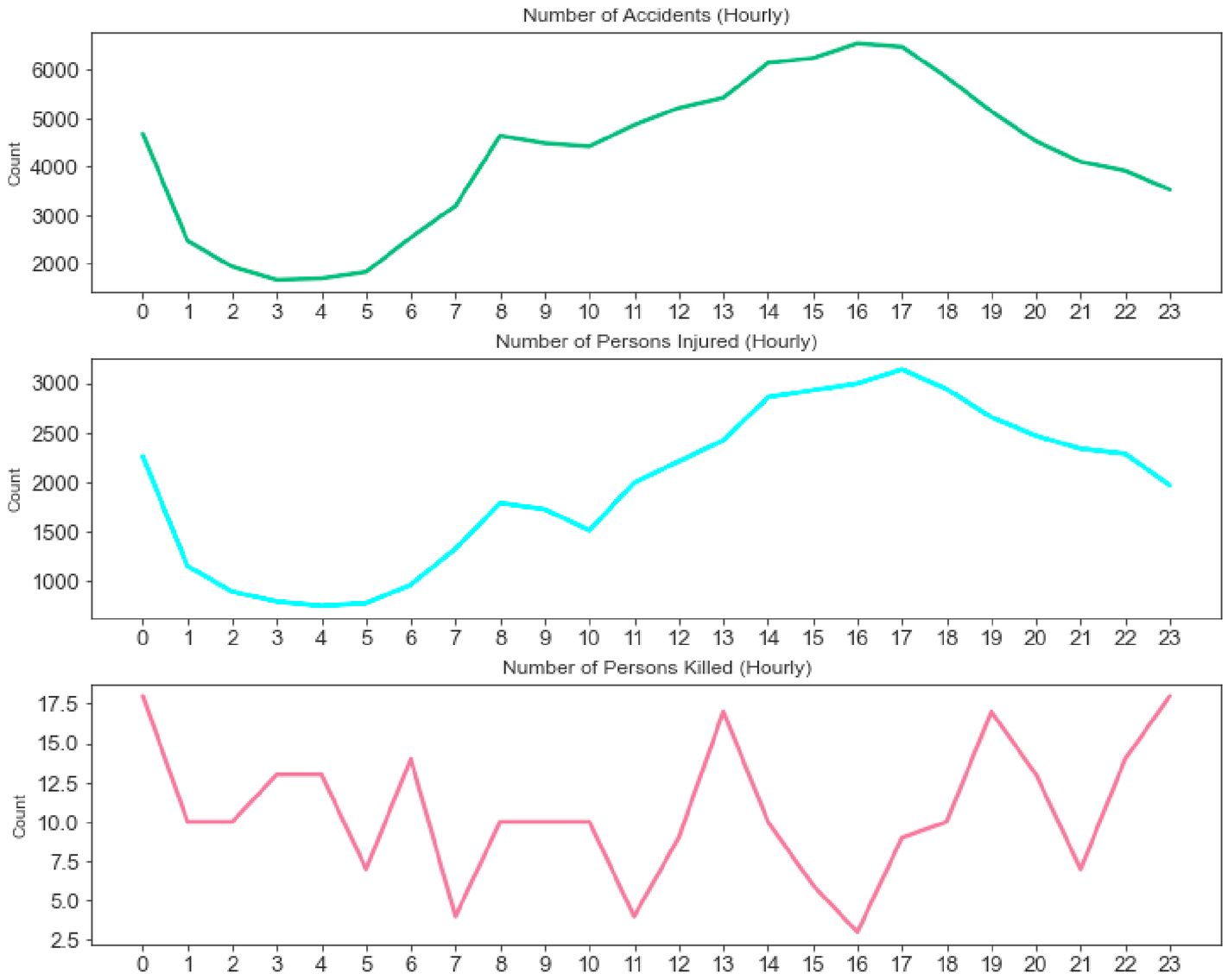
Months


Number of Accidents (Monthly)
Count
9000
8000
7000
Jan
Mar
May
Jul
Sep
Nov
Number of Persons Injured (Monthly)
Count
4500
4000
3500
3000
2500
Jan
Mar
May
Jul
Sep
Nov
Number of Persons Killed (Monthly)
Count
30
25
20
15
Jan
Mar
May
Jul
Sep
Nov

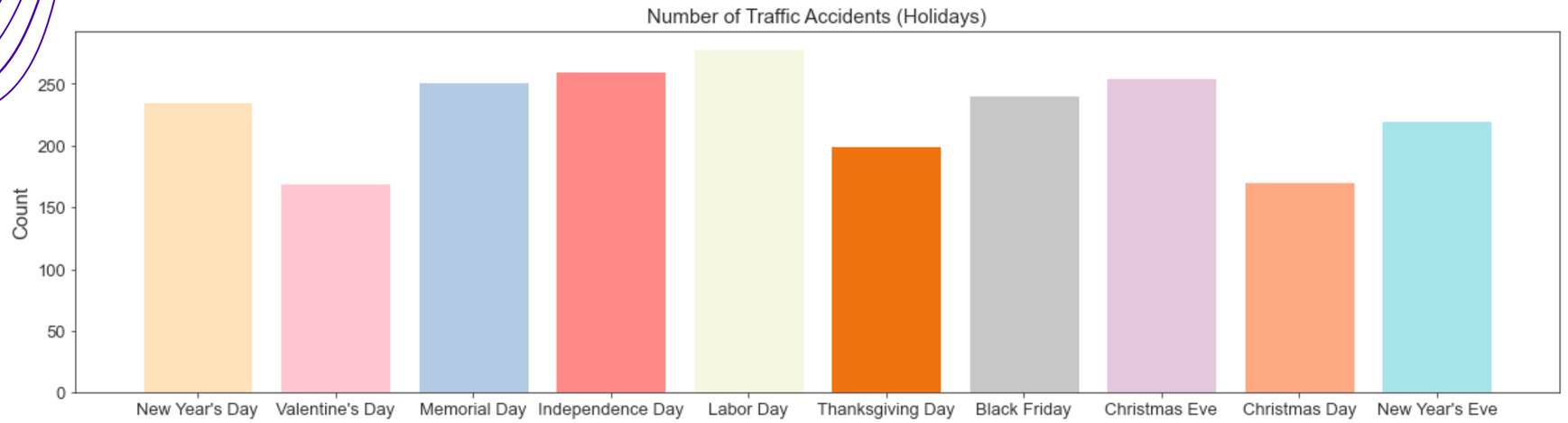
Weekdays

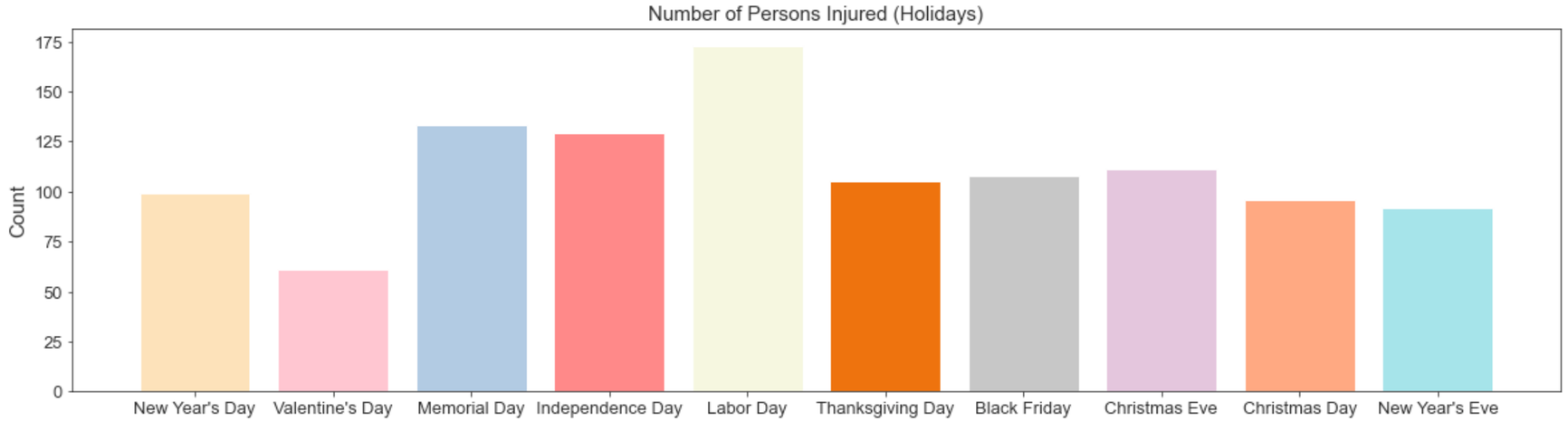

Number of Accidents (Day of Week)
Count
16000
15500
15000
14500
14000
13500
13000
Mon
Tues
Wed
Thur
Fri
Sat
Sun
Number of Persons Injured (Day of Week)
Count
7200
7000
6800
6600
6400
Mon
Tues
Wed
Thur
Fri
Sat
Sun
Number of Persons Killed (Day of Week)
Count
40
35
30
25
Mon
Tues
Wed
Thur
Fri
Sat
Sun

Time of Day


Number of Accidents (Hourly)
Count
6000
5000
4000
3000
2000
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
Number of Persons Injured (Hourly)
Count
3000
2500
2000
1500
1000
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
Number of Persons Killed (Hourly)
Count
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23

Holidays

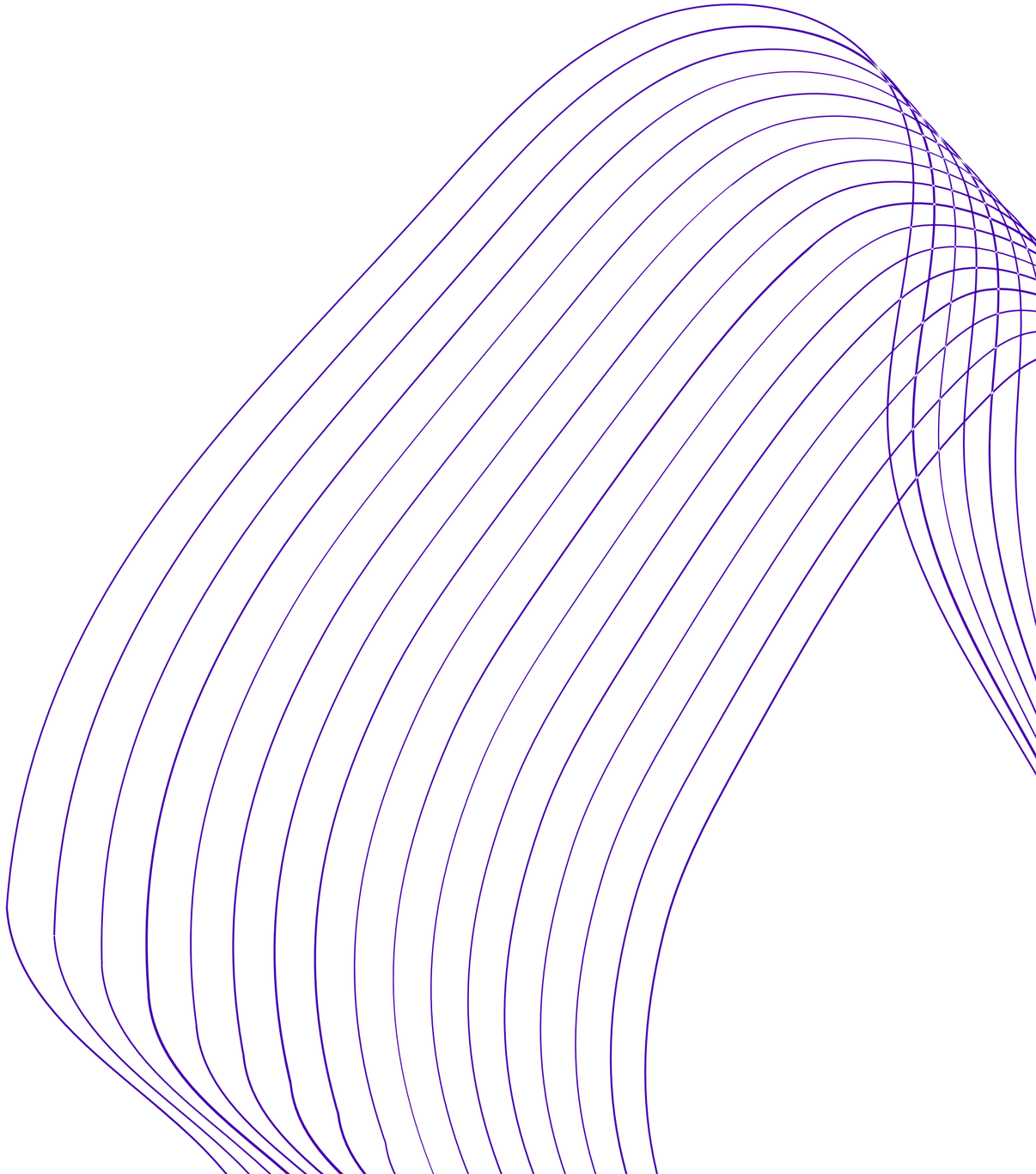

Number of Traffic Accidents (Holidays)
Count
0
50
100
150
200
250
New Year's Day
Valentine's Day
Memorial Day
Independence Day
Labor Day
Thanksgiving Day
Black Friday
Christmas Eve
Christmas Day
New Year's Eve


Number of Persons Injured (Holidays)
Count
0
25
50
75
100
125
150
175
New Year's Day
Valentine's Day
Memorial Day
Independence Day
Labor Day
Thanksgiving Day
Black Friday
Christmas Eve
Christmas Day
New Year's Eve


Number of Persons Killed (Holidays)
Count
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
New Year's Day
Valentine's Day
Memorial Day
Independence Day
Labor Day
Thanksgiving Day
Black Friday
Christmas Eve
Christmas Day
New Year's Eve

What were the most dangerous areas in NYC in 2021?

Methods: Mapped latitude and longitude on folium maps, grouped by zipcode and graphed with geojson, and clustered with DBSCAN to find most dangerous intersections.



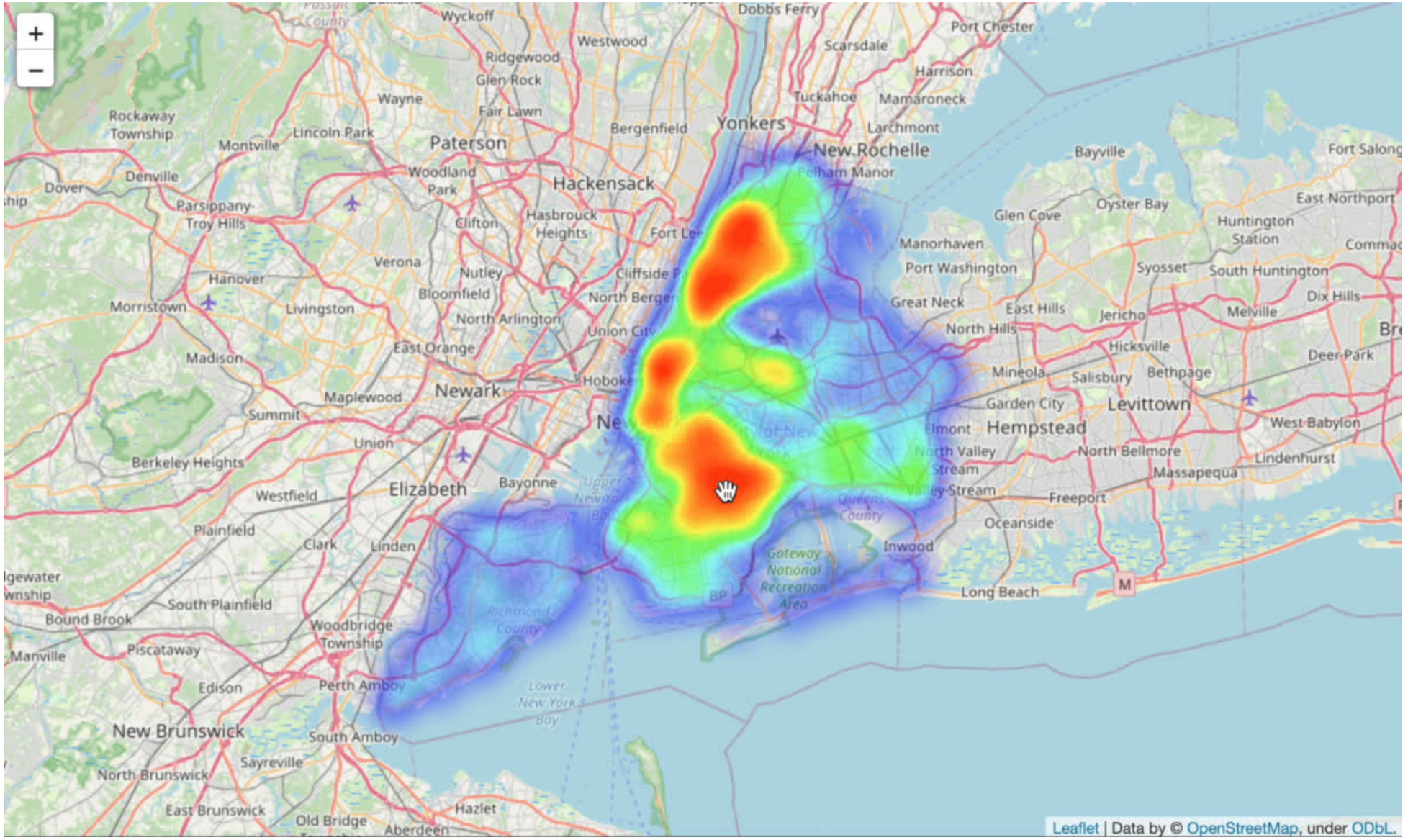
All Deaths Folium Map

```
if val == 'all persons':
    val = 9
elif val == 'pedestrians':
    val = 11
elif val == 'cyclists':
    val = 13
elif val == 'motorists':
    val = 15
else:
    print("unacceptable entry, please try again")

#create death database based on the column number above
death = all_data[all_data.iloc[:,val] > 0]
death

#iteratively create markers in folium map, with lat/long and contributing factor vehicle as popup
for i in range(len(death['ZIP CODE'])):
    folium.Marker(
        location=[death['LATITUDE'].iloc[i],death['LONGITUDE'].iloc[i]],
        popup=str(death['CONTRIBUTING FACTOR VEHICLE 1'].iloc[i]),
        icon=folium.Icon(color="red", icon="info-sign"),
    ).add_to(m)

m
```

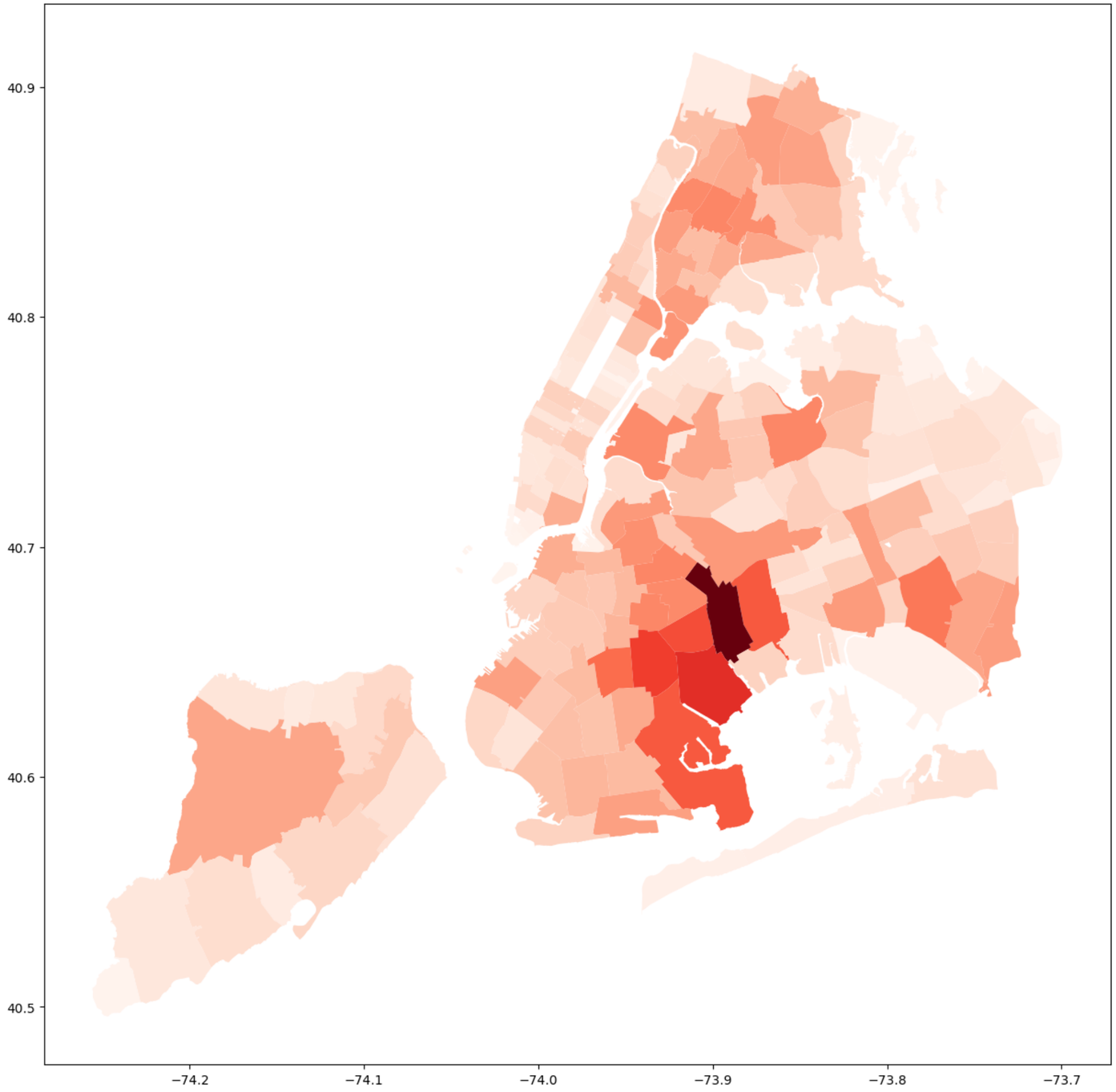
```
Enter one of the following to see a map of the death:
 all persons
 pedestrians
 cyclists
 motorists
motorists
```

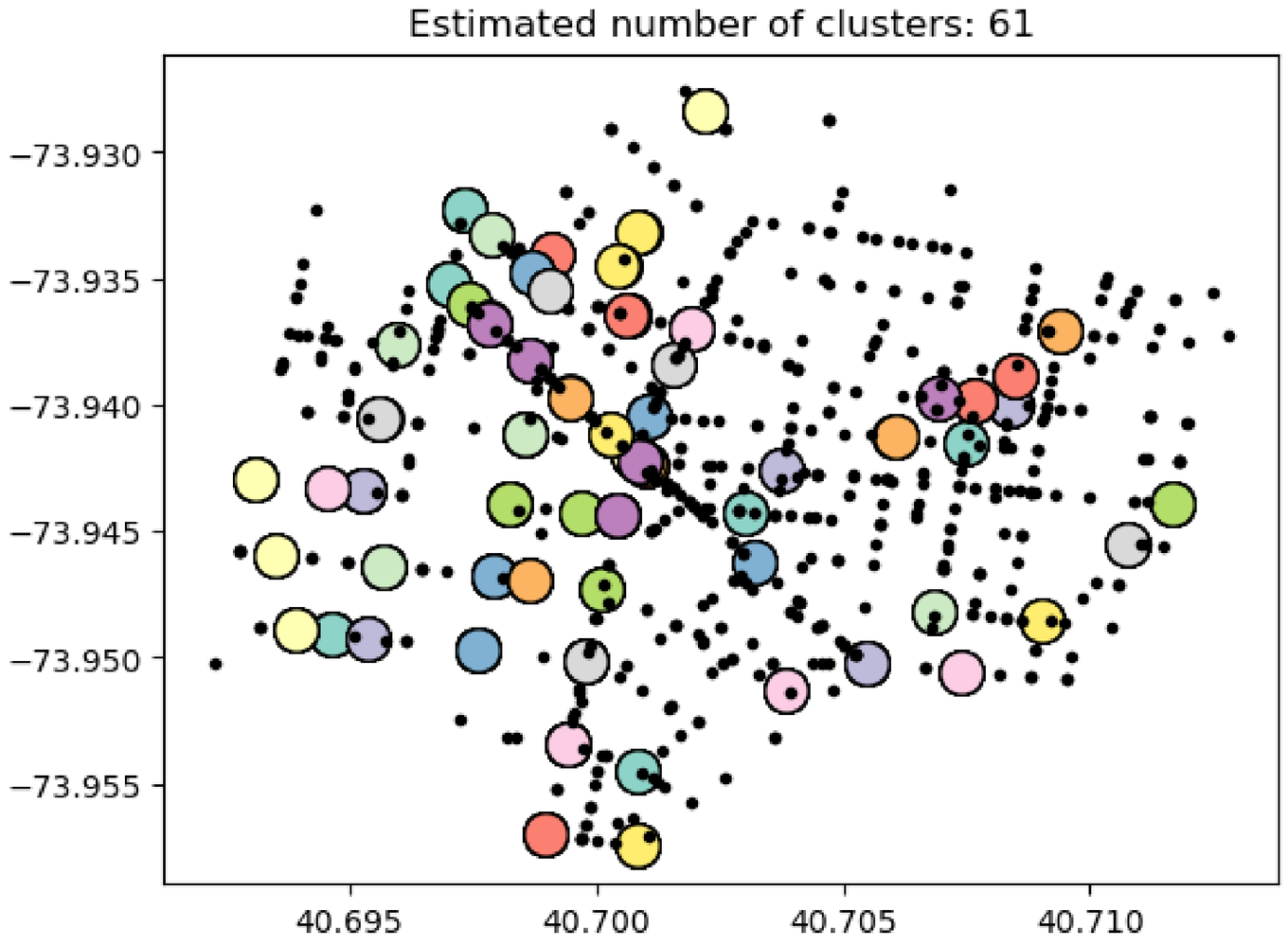
Out[6]:

All Injuries Heat Map



Map of Zipcodes Colorized by Injuries





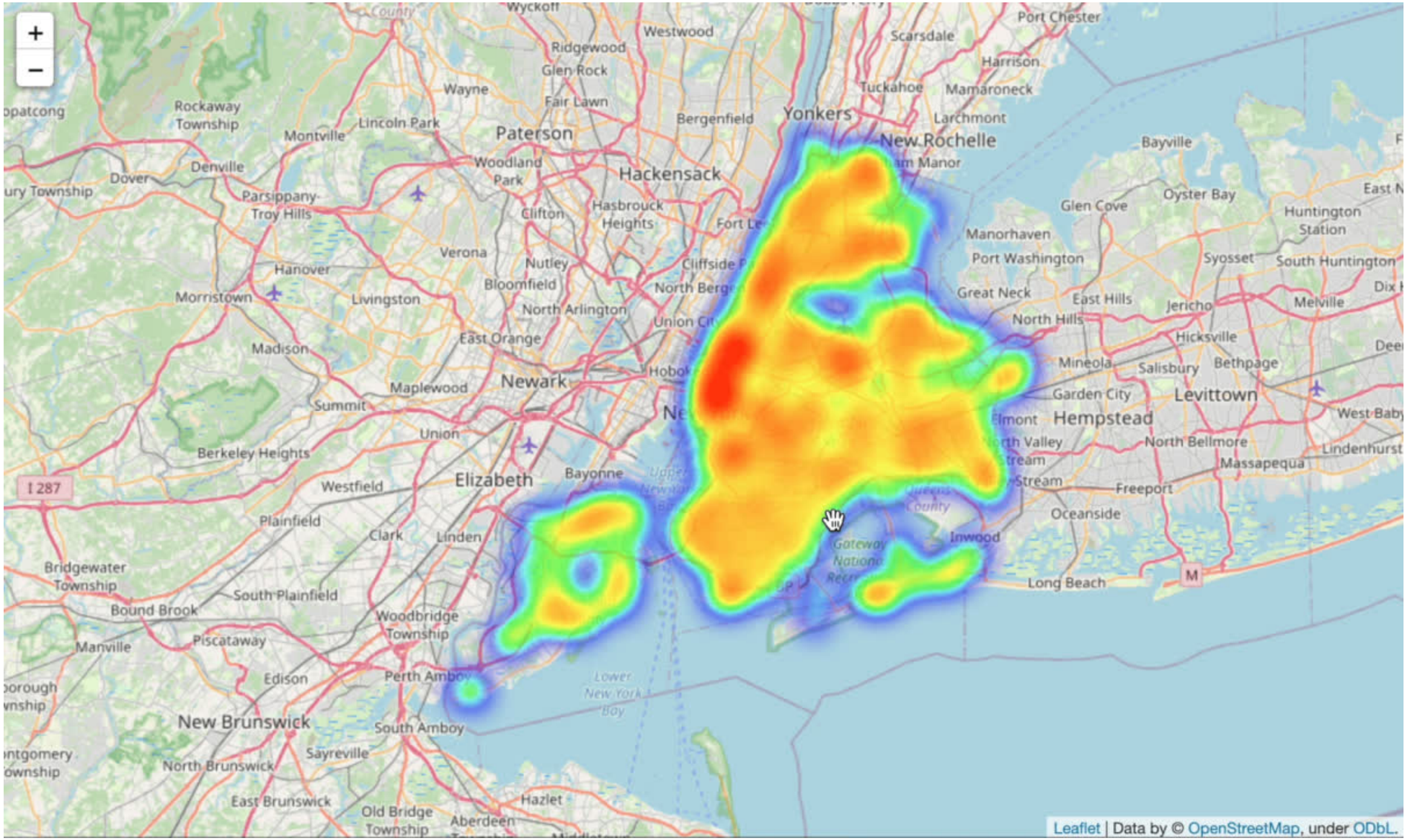
Plot of Clusters Found Using DBSCAN Zipcode:11206

Cluster Markers Map for 11206

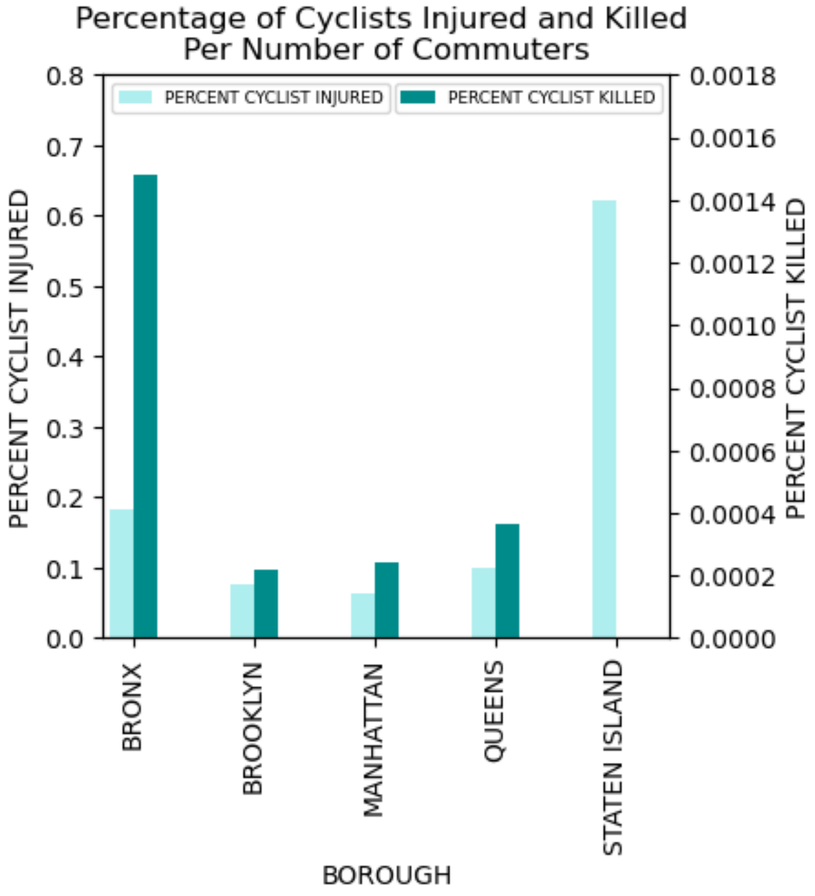
Which areas had the most bike accidents in NYC in 2021?

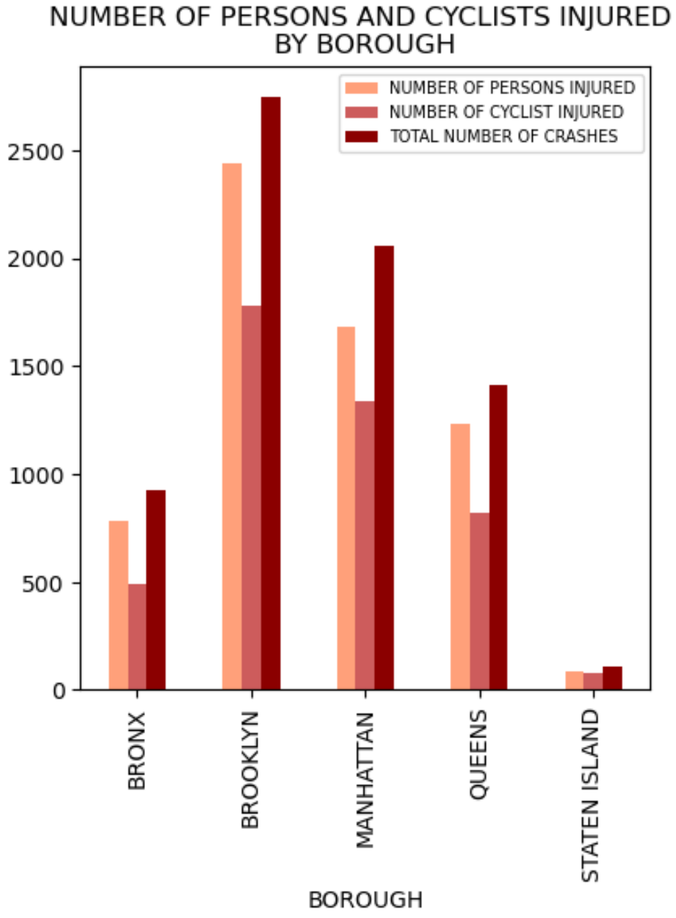
Methods: Divided data by borough and zip code. Included data on bike commuters in each borough to estimate percentages of injuries and fatalities. Used folium to visualize hot spots of collisions with heat map.

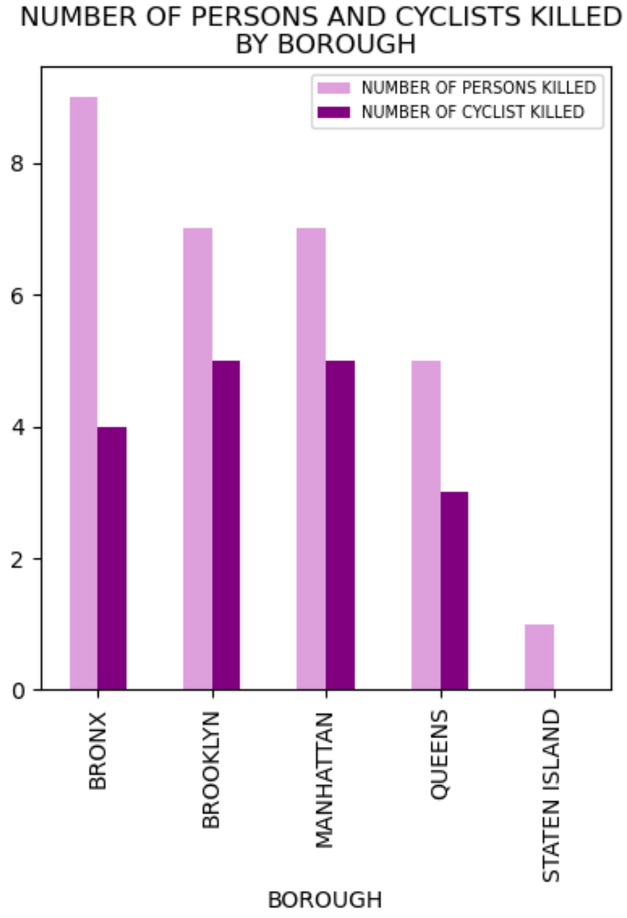
All Crashes Involving Bikes Heat Map



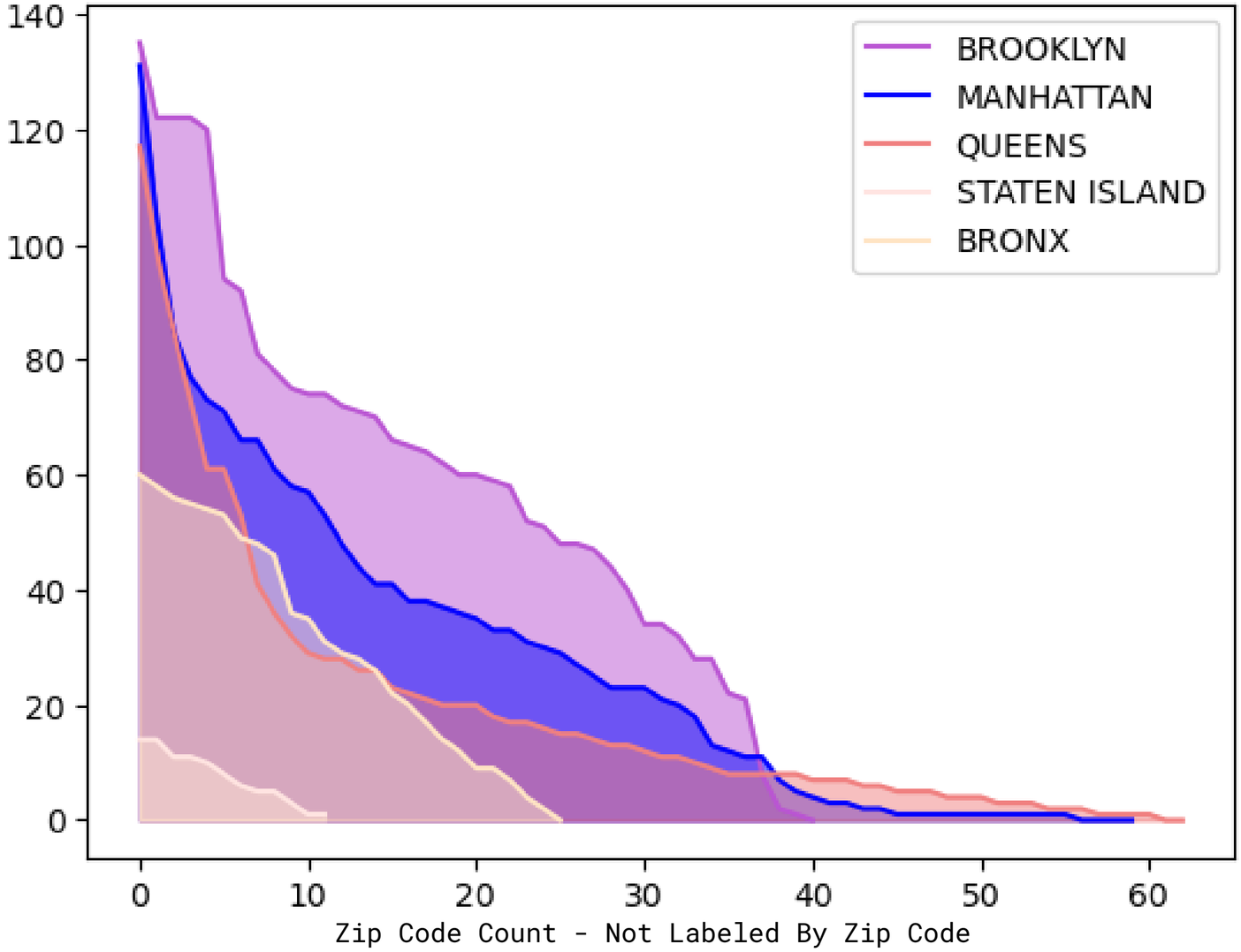
Injuries and Deaths by Borough for Bike Collisions





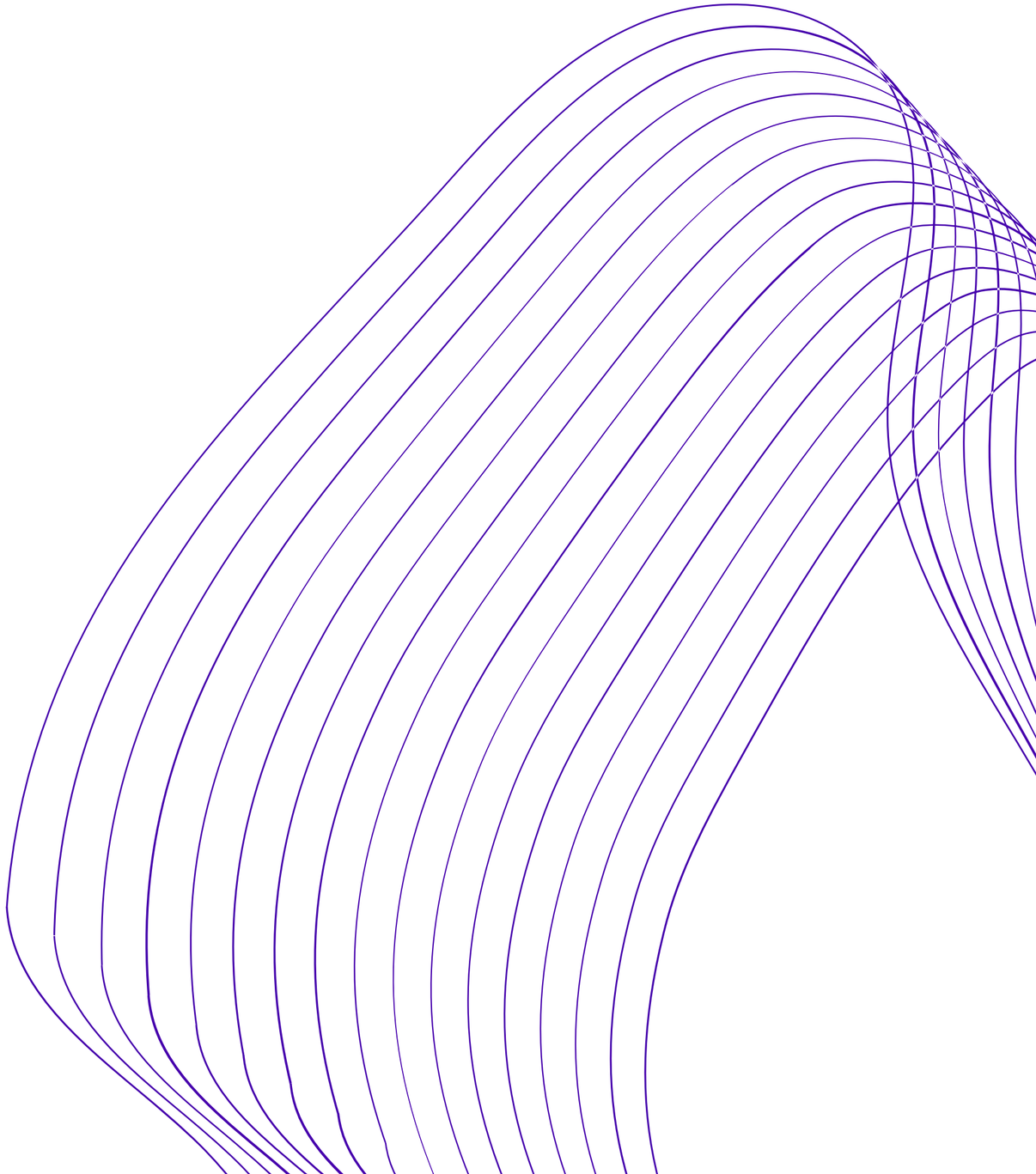


Persons Injured by Zipcode for Bike Collisions

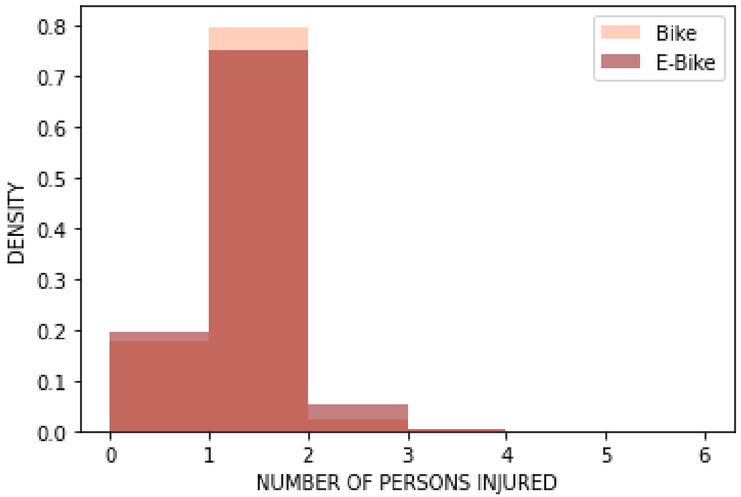


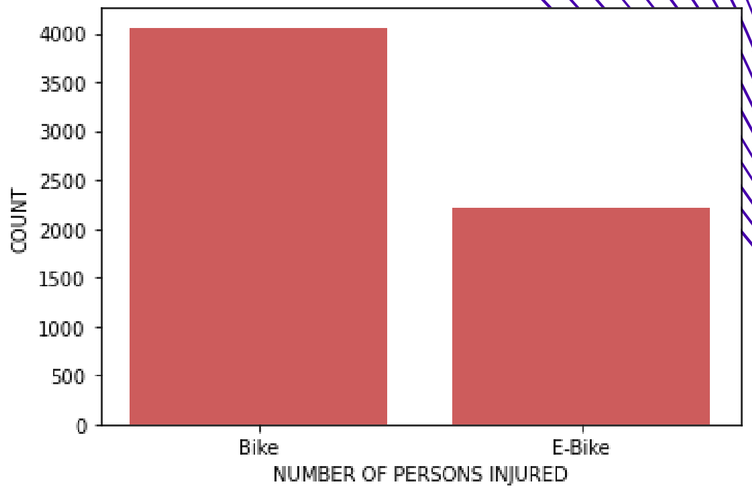
Are e-bikes more dangerous than regular bikes?

Method: Extracted data on injuries and deaths for bikes and e-bikes only. Separated the groups and compared them using bar charts and resampling methods.

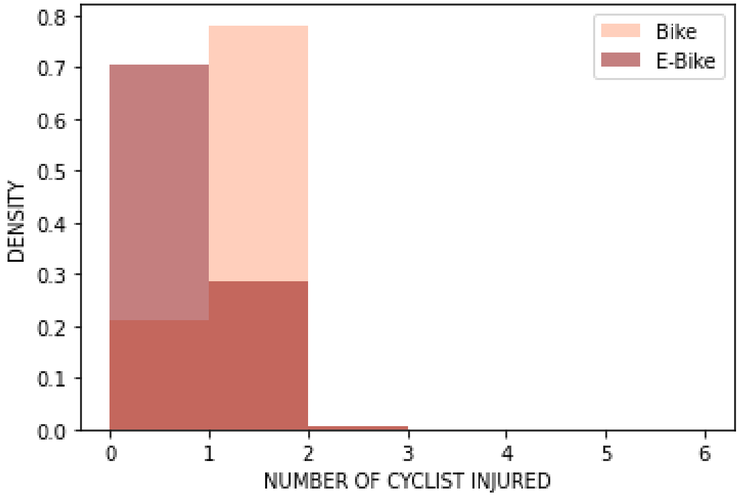


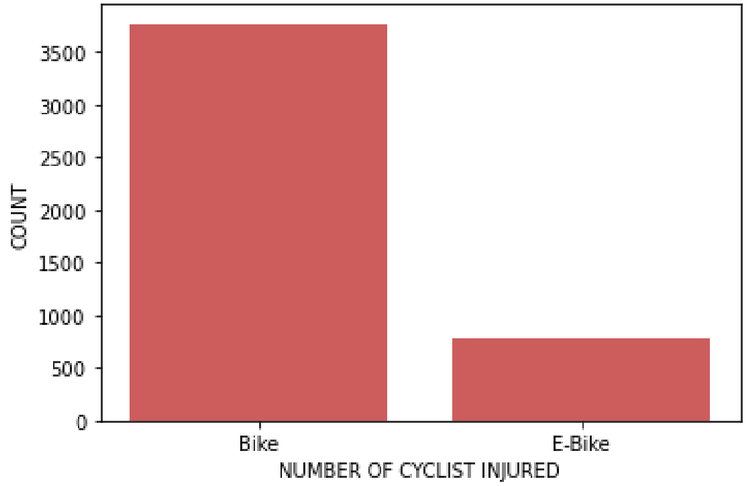
Number Injured





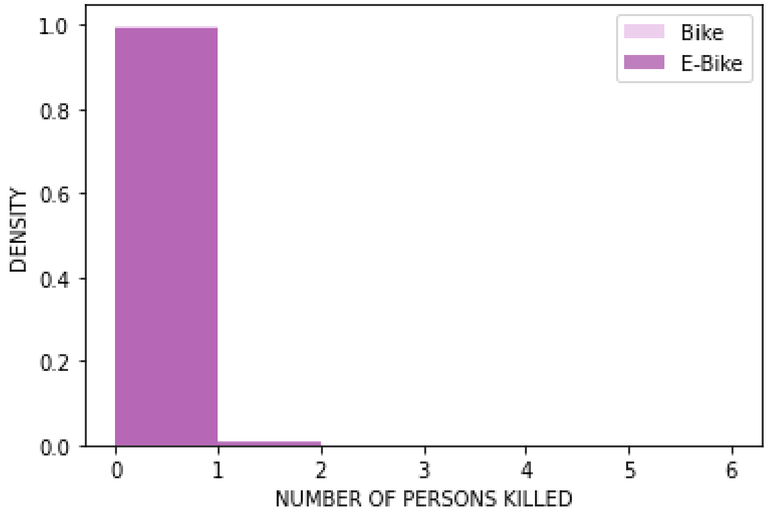
NUMBER OF PERSONS INJURED Welch's t-test p-value: 0.48282999332773846

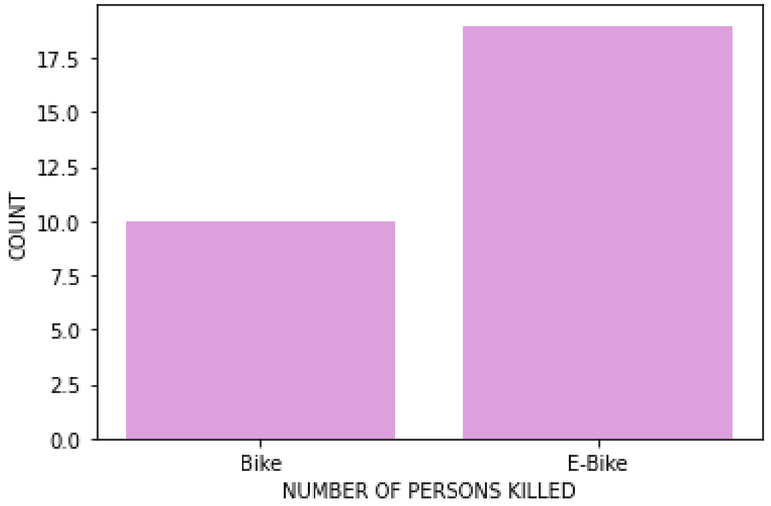




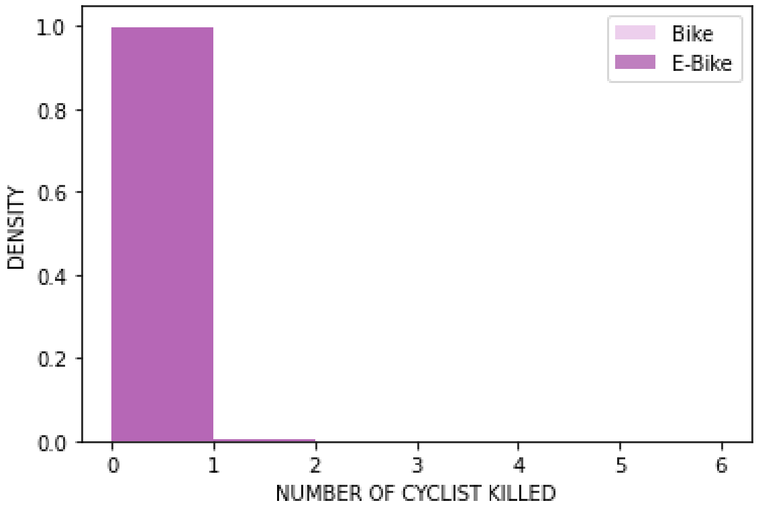
NUMBER OF CYCLIST INJURED Welch's t-test p-value: 0.0

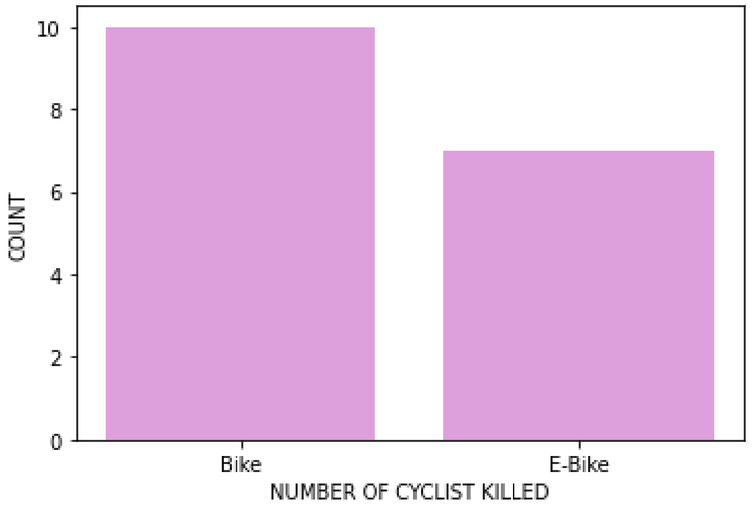
Number Killed





NUMBER OF PERSONS KILLED Welch's t-test p-value: 0.005328408716739479





NUMBER OF CYCLIST KILLED Welch's t-test p-value: 0.6116547798822987

Under and Over Sampling Analysis

```
NUMBER OF PERSONS INJURED:
Undersample count: 3
Undersample significance ratio: 0.003
Oversample count: 16
Oversample significance ratio: 0.016
NUMBER OF PERSONS KILLED:
Undersample count: 981
Undersample significance ratio: 0.981
Oversample count: 1000
Oversample significance ratio: 1.0
NUMBER OF CYCLIST INJURED:
Undersample count: 1000
Undersample significance ratio: 1.0
Oversample count: 1000
Oversample significance ratio: 1.0
NUMBER OF CYCLIST KILLED:
Undersample count: 7
Undersample significance ratio: 0.007
Oversample count: 2
Oversample significance ratio: 0.002
```

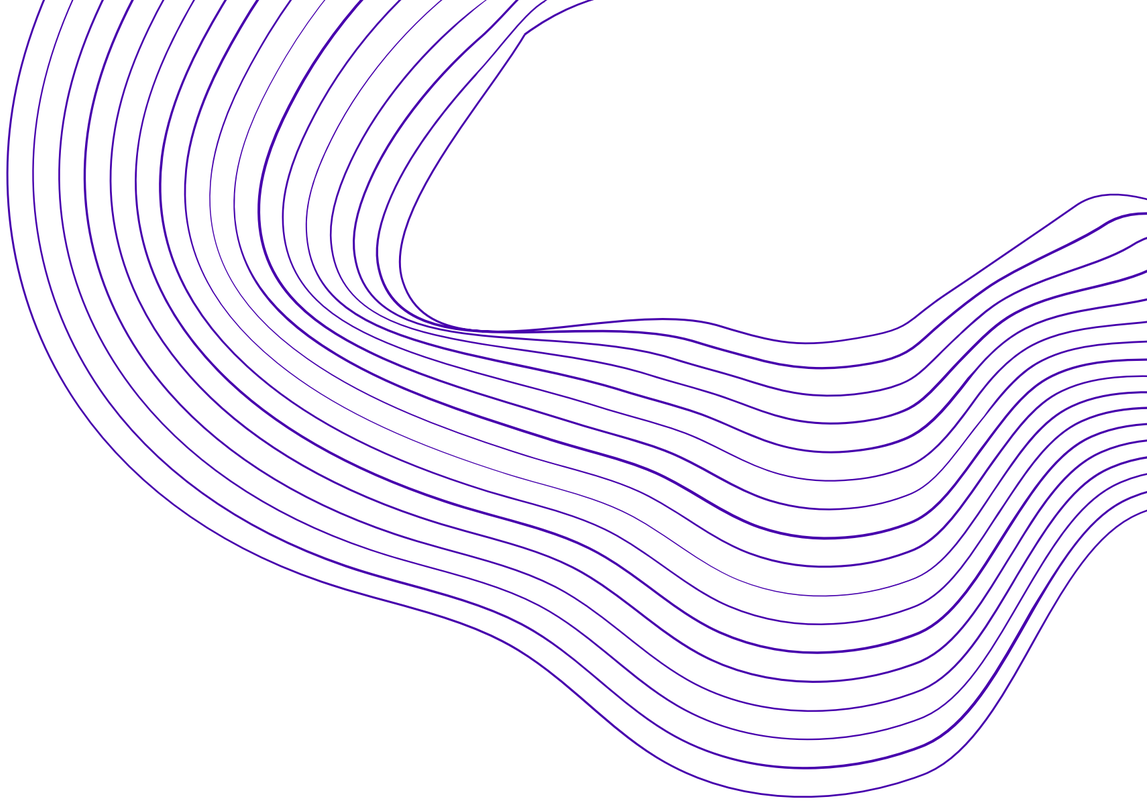
100% of the time there is a significant difference between cyclist injuries for bikes and e-bikes involved in motor vehicle collisions.

Bikes tend to be significantly more 'dangerous' - characterized as having more cyclist injuries per crash - than e-bikes.

Most of the time there is a significant difference between total people killed for bikes versus e-bikes.

E-Bikes tend to be more 'dangerous' - characterized as more total deaths occurring when one is involved in a crash - than bikes.

Thank you!



Questions?