NYC Motor Vehicle Crashes

Syracuse University IST 687 Fall 2021

By Patrick Furlong, Nora Lin,
Michael Morrey & Ashley Silverstien

02

About the Data

The original dataset contains about 1.8M observations. Our team has subset this data into a two-year period, 2018-2019. The data is available through the following link: https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Crashes/h9gi-nx95

Our dataset contains 443,045 observations with 29 variables.

The "Motor Vehicle Collisions – Crashes" dataset contains details about crash events in New York City. Each row represents a crash. A police report (MV104-AN) is required when the crash involves a person(ie pedestrian, cyclists, or motorist) is injured or killed or $1,000 worth in damages.

Table Preview

View Data | Create Visualization

CRAS...	CRAS...	BOR...	ZIP C...	LATI...	LON...	LOCA...	ON S...	CROS...	OFF S...	NUM...	NUM...	NUI
04/13/2021	0:00	BROOKLYN	11222	40.726444	-73.95233	(40.72644...			745 MAN...	0	0	0
04/13/2021	0:00	BROOKLYN	11231	40.678524	-74.0021	(40.67852...			667 HEN...	0	0	0
04/13/2021	0:00	QUEENS	11420	40.677498	-73.8276	(40.67749...	111 STREET	ROCKAW...		0	0	0
04/13/2021	0:00						ALEXAND...			0	0	0
04/13/2021	0:00			40.877235	-73.91781	(40.87723...	JOHNSON...			0	0	0
04/13/2021	0:00			40.90021	-73.8865	(40.90021...	MOSHOL...			1	0	0
04/13/2021	0:07	BRONX	10456	40.82942	-73.91226	(40.82942...	EAST 166 ...	CLAY AVE...		1	0	0
04/13/2021	0:15	QUEENS	11378	40.727432	-73.90713	(40.72743...	LONG ISL...	MAURICE ...		0	0	0
04/13/2021	0:45			40.843956	-73.8978	(40.84395...	CROSS B...			1	0	0
04/13/2021	10:00	BRONX	10463				Ft indepe...	Bailey pla...		0	0	0
04/13/2021	10:00	BROOKLYN	11207	40.686226	-73.9124	(40.68622...	BUSHWIC...	COVERT S...		0	0	0

03 Variable

Below are a list of all the variables within the dataset. We have a good variety of variables to explore here which include geographic information, date and time information as well as more detailed crash information.

Time and Location Information:

- Crash Time
- Crash Date
- Borough
- Zip Code
- Longitude
- Latitude
- Location
- Off street name
- On street name

Crash Details:

- Collision ID
- Number of persons injured
- Number of persons killed
- Number of pedestrians injured
- Number of cyclists injured
- Number of motorist injured
- Number of pedestrians killed
- Number of cyclists killed
- Number of motorists killed

Cash Factors:

- Vehicle Code Type 1-5
- Contributing Factor 1-5

04

Our Process

Picking our data

- Thought about questions we wanted to ask
- The original data set was clean

Changed our data set

- Changed our dataset because it the data was too clean
- Revised our questions

Cleaning pt. 1

- Changed white spaces to underscores
- Separate the hours, minute and second columns

Cleaning pt. 2

- We found a lot of missing data (N/A)
- Clean up the borough, zip code, latitude and longitude data.
- Checked for outliers

Data exploration/ Mapping

- Looked into what data we wanted to explore
- Figure out what type of maps and graphs we wanted to display

Type of information

- Chose to focus on boroughs, time of day, year to year data, type of incident, total incidents, incidents by boroughs

Regression

- Tried 3 regression models and came out with similar R-values

Predictions

- How did the pandemic affect our trends?

05

Final Cleaned Dataset

- Out of our 443,045 data point entries, we started off with **156,791 NA values** for Borough.
- With our code, we were able to use the min max of longitude and latitude to fill in these NA values.
- Ultimately, we filled in **132,516 NA values** for the Borough Column.

	YEAR	MONTH	DAY	HOUR	MINUTE	SECOND	BOROUGH	ZIP_CODE	LATITUDE	LONGITUDE	LOCATION
1	2018	1	1	15	0	0	STATEN_ISLAND	10307	40.65906	-73.88459	(40.65906, -73.88459)
2	2018	1	1	4	16	0	MANHATTAN	10025	40.80180	-73.96108	(40.8018, -73.96108)
3	2018	1	1	20	30	0	QUEENS	11374	40.74397	-73.88510	(40.743973, -73.8851)
4	2018	1	1	23	41	0	STATEN_ISLAND	10307	40.72143	-73.89275	(40.72143, -73.892746)
5	2018	1	1	15	30	0	MANHATTAN	10025	40.80174	-73.96477	(40.80174, -73.96477)
6	2018	1	1	15	0	0	MANHATTAN	10025	40.66622	-73.80086	(40.666225, -73.80086)
7	2018	1	1	12	10	0	QUEENS	11374	40.76307	-73.81634	(40.763073, -73.816345)
8	2018	1	1	18	35	0	BRONX	10462	40.82030	-73.89083	(40.820305, -73.89083)
9	2018	1	1	13	50	0	BROOKLYN	11220	40.65892	-73.88982	(40.65892, -73.889824)
10	2018	1	1	1	37	0	BROOKLYN	11220	40.66228	-73.91078	(40.662277, -73.91078)
11	2018	1	1	11	50	0	STATEN_ISLAND	10307	40.57440	-74.09930	(40.5744, -74.099304)
12	2018	1	1	10	15	0	STATEN_ISLAND	10307	40.66528	-73.82955	(40.665276, -73.82955)
13	2018	1	1	12	50	0	QUEENS	11374	40.72236	-73.73689	(40.722363, -73.736885)
14	2018	1	1	3	0	0	MANHATTAN	10025	40.73416	-73.98926	(40.734158, -73.98926)
15	2018	1	1	2	46	0	QUEENS	11374	40.74757	-73.88462	(40.747566, -73.88462)

06

Data Exploration: Total Incidents

We decided to start by looking at Total Incidents, a broad examination of incidents that include Motorists, Cyclists and Pedestrians.

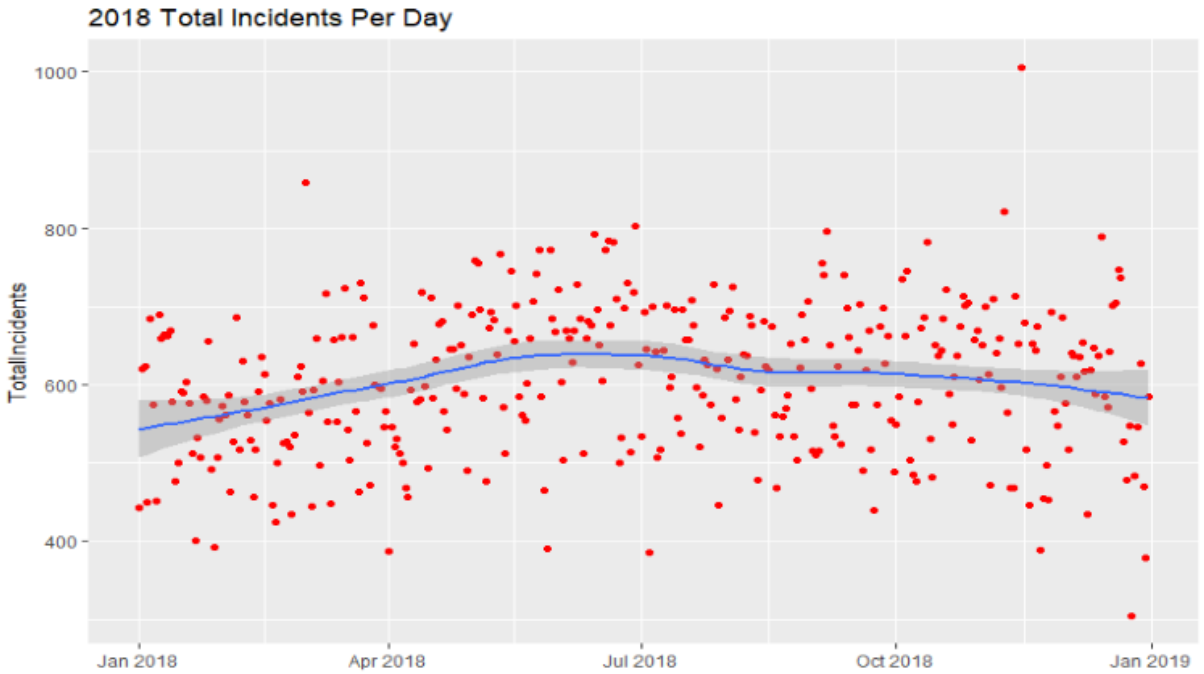
What we are looking for:

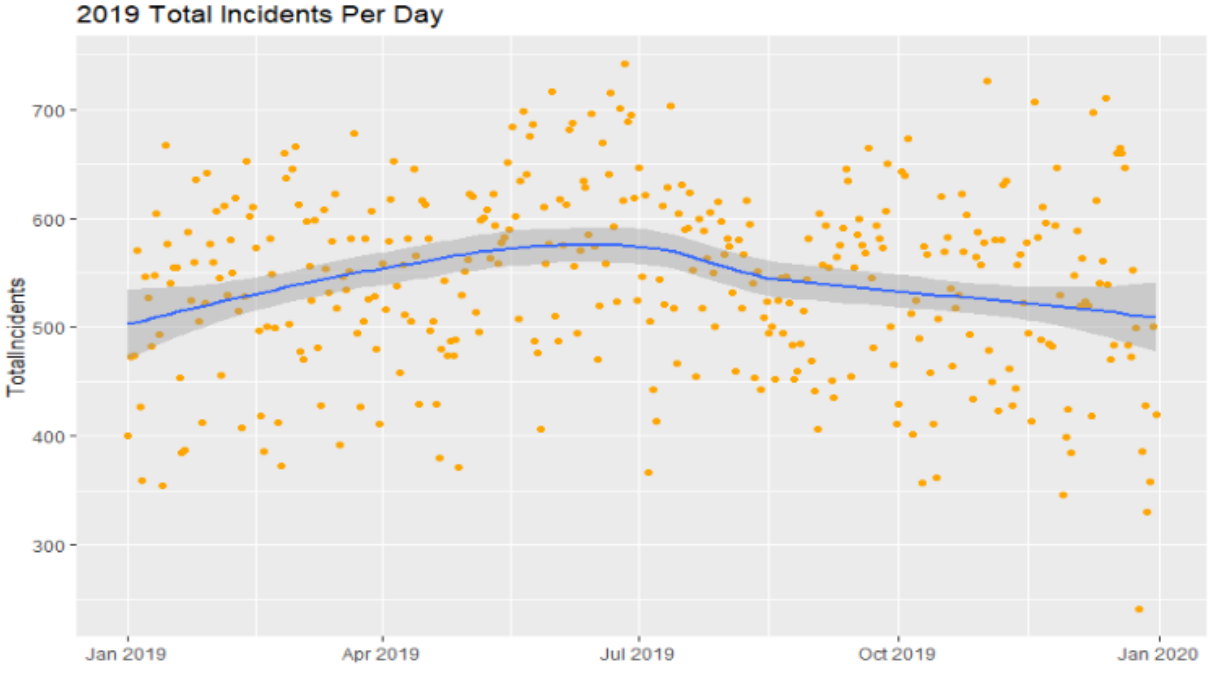
- Trend of total incidents from 2018 to 2019
- How Total incidents compare each Month?
- How Total incidents compare each Day of the Week?
- Which hours of the day have the most incidents?

07

Trend of Total Incidents per Year

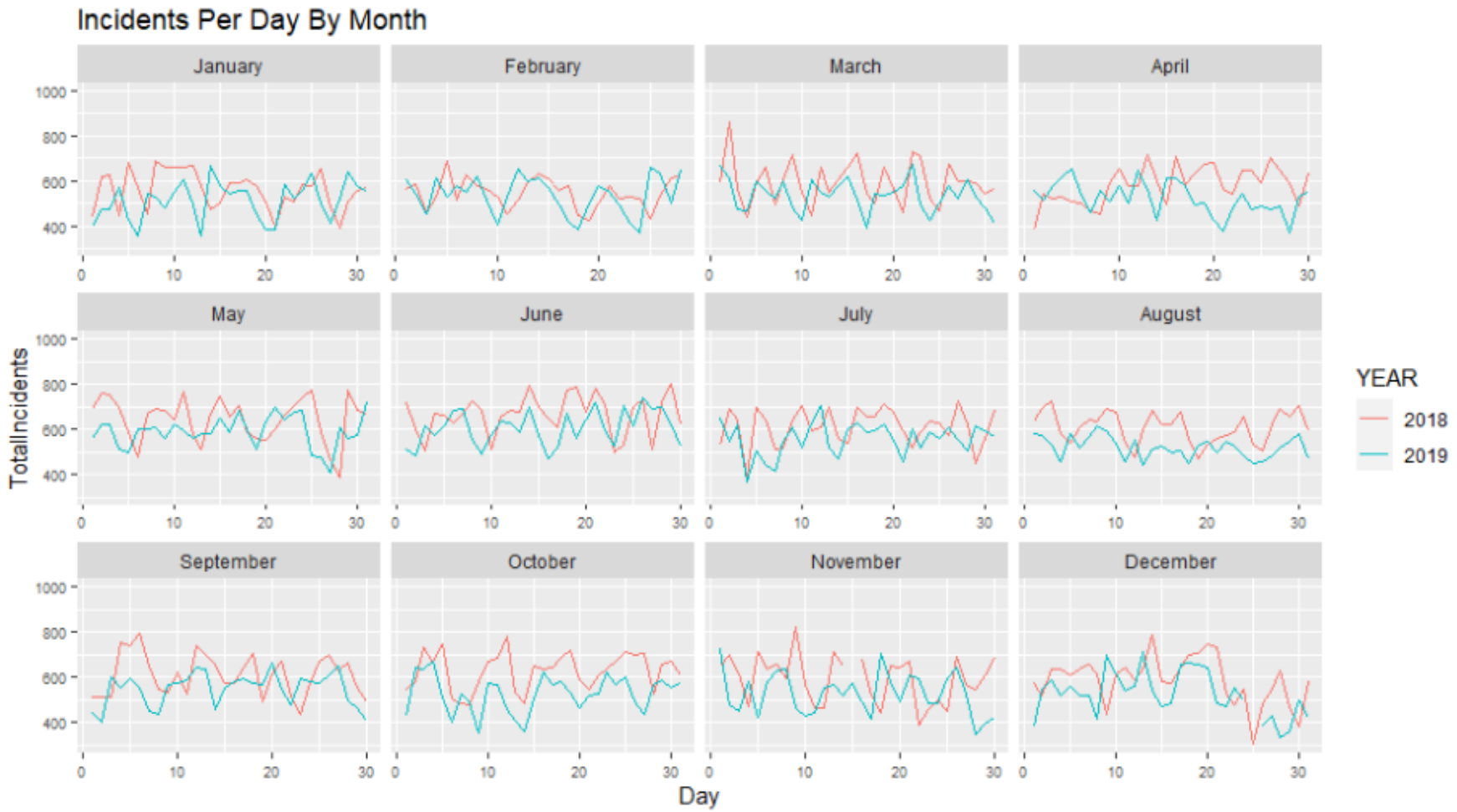
- The dataset presents a relative trend of incidents by year. We used lm_smooth() to show this trend.
- Each dot represents a day and the total number of incidents that occurred.
- The number of total incidents peaks between April and July in both 2018 and 2019.
- The total number of incidents through the year remains consistent between 400 and 800 number of incidents.



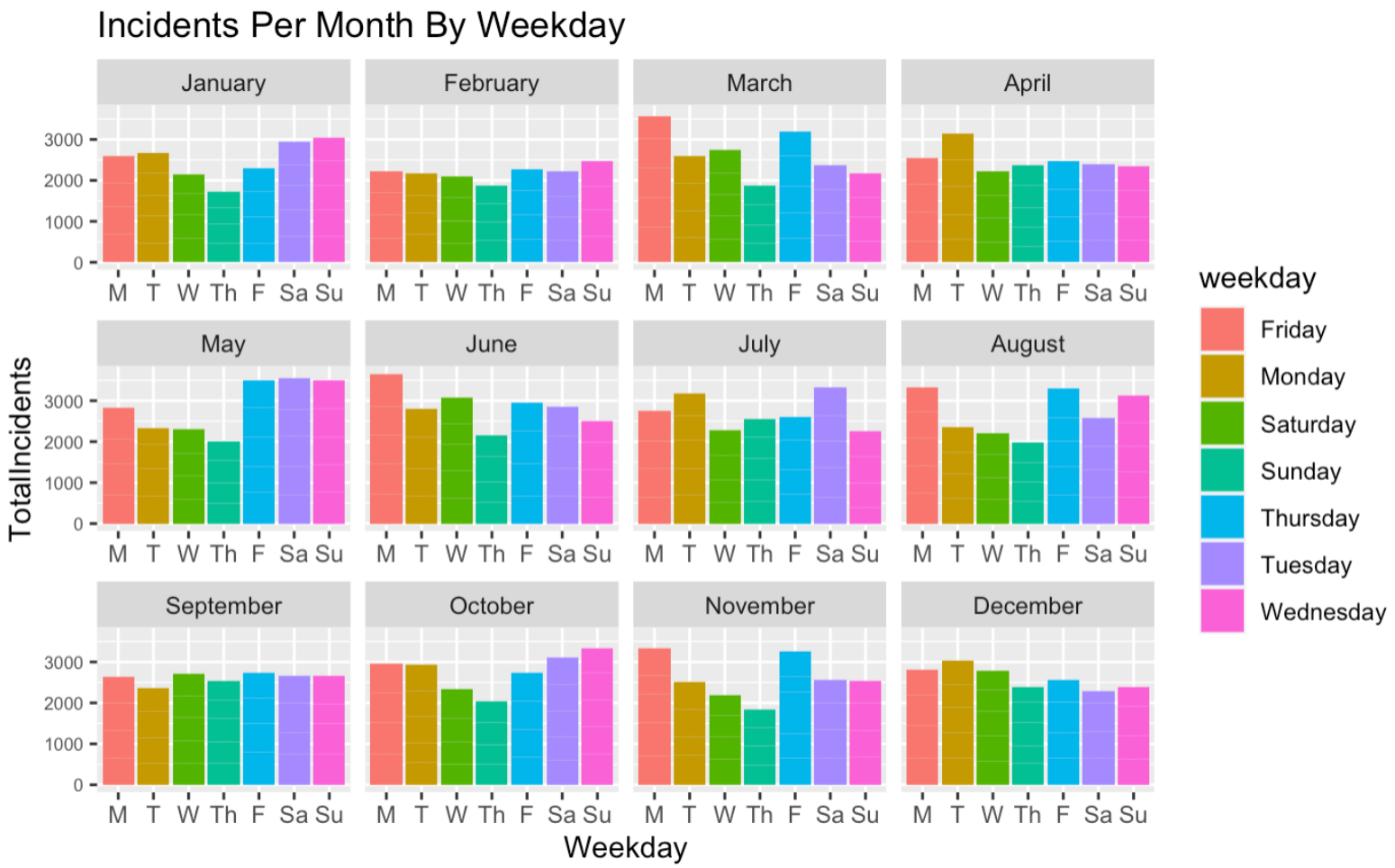


Total Incidents: Monthly Comparison

- Plotted days on the x-axis and number of incidents on the y-axis for each month.
- Across all 12 months, the trend in 2019 follows the trend in 2018.
- Many months follow almost the exact same pattern.



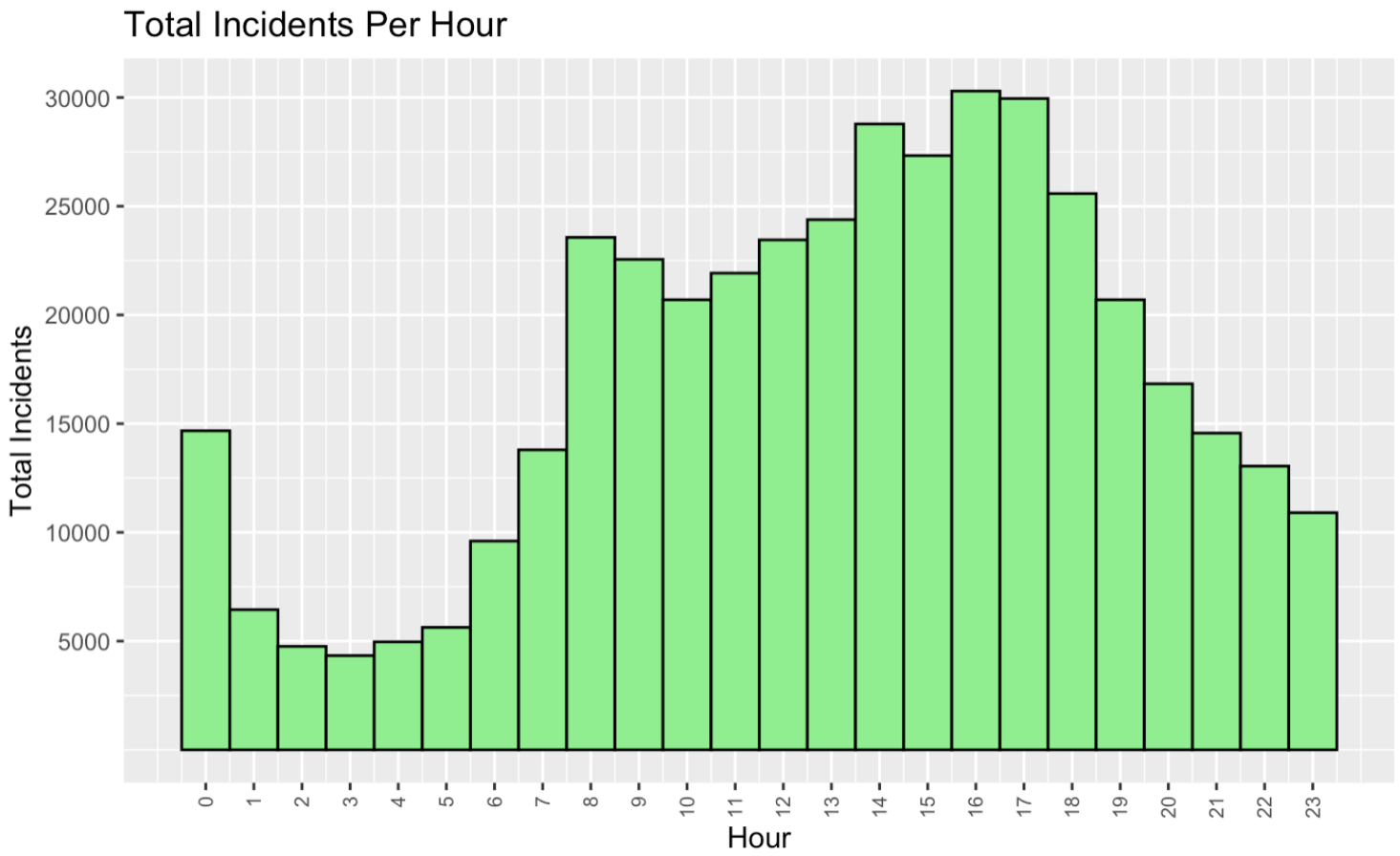
Total Incidents: Daily Comparison



- This chart shows an aggregate of all incidents for each day of the week (2018 and 2019 combined).
- You could look at this chart and see which days incidents are most likely to occur.
- To point out a couple of days, Monday in March and weekends in May have over 3,000 recorded incidents in 2018 and 2019.
- Most days of the week do not change much from month to month and remain close to the average of about 2,500.

Total Incidents by the Hour

- Since we cleaned our data from a single variable that contained the date and time into separate variables for Month, Day, Year, Hour, Minute, and Second, we were able to explore which hours the incidents occurred.
- The graph to the left shows a bimodal distribution.
- The number of total incidents between 7am and 11pm presents closely to a normal distribution.
- As expected, we see incidents increase and remain higher during hours during the day since more people would be awake.
- We can see a sharp spike in the number of incidents between 7am and 8am which could be the result of the morning commute.



Data Exploration: Total Incidents by Borough

After looking at the numbers for total incidents, we decided to look more closely at the incidents for each Borough of the City.

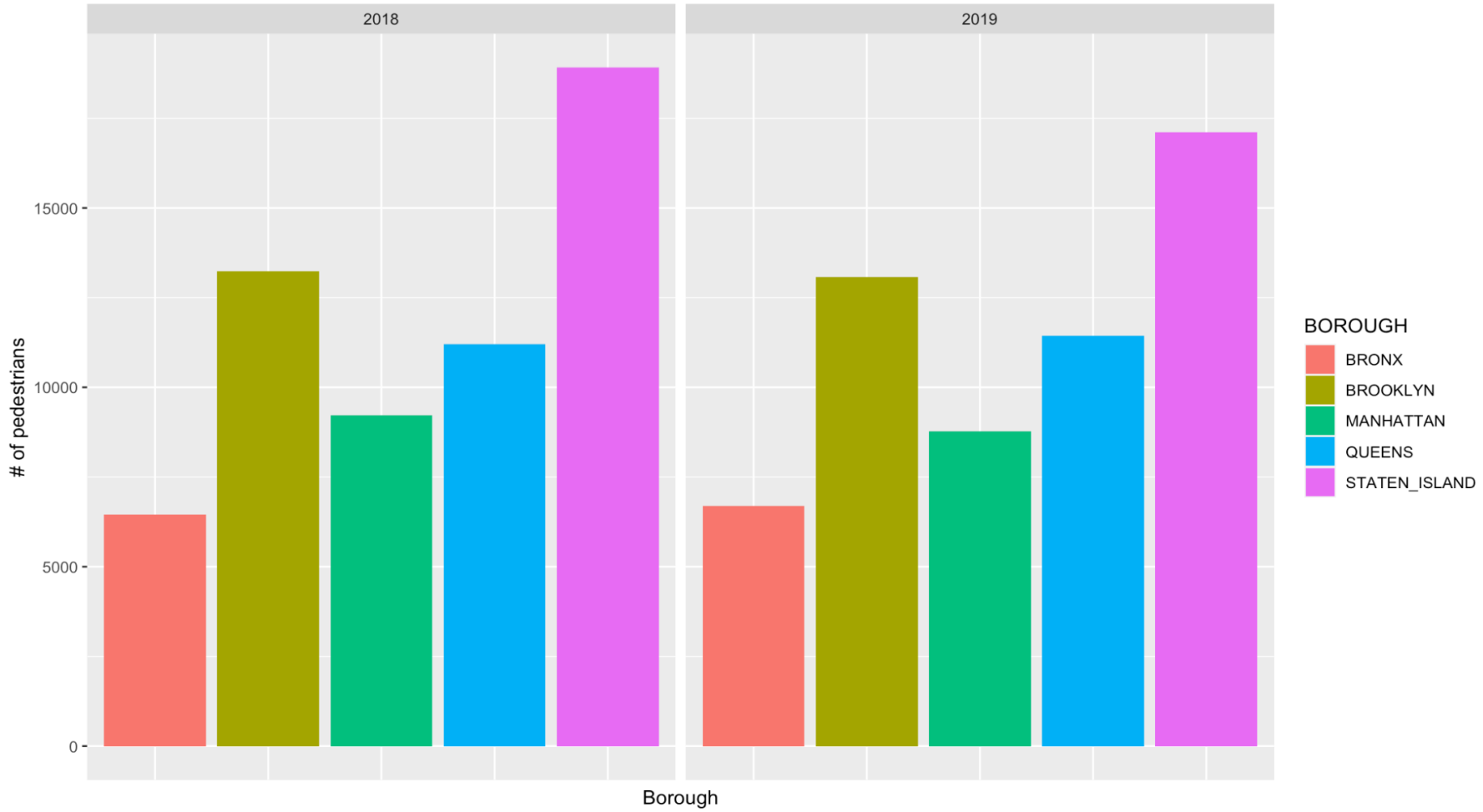
What we are looking for:

- How do the incidents vary per Borough?
- Do different Boroughs have different hourly highs?

Total Incidents by Borough

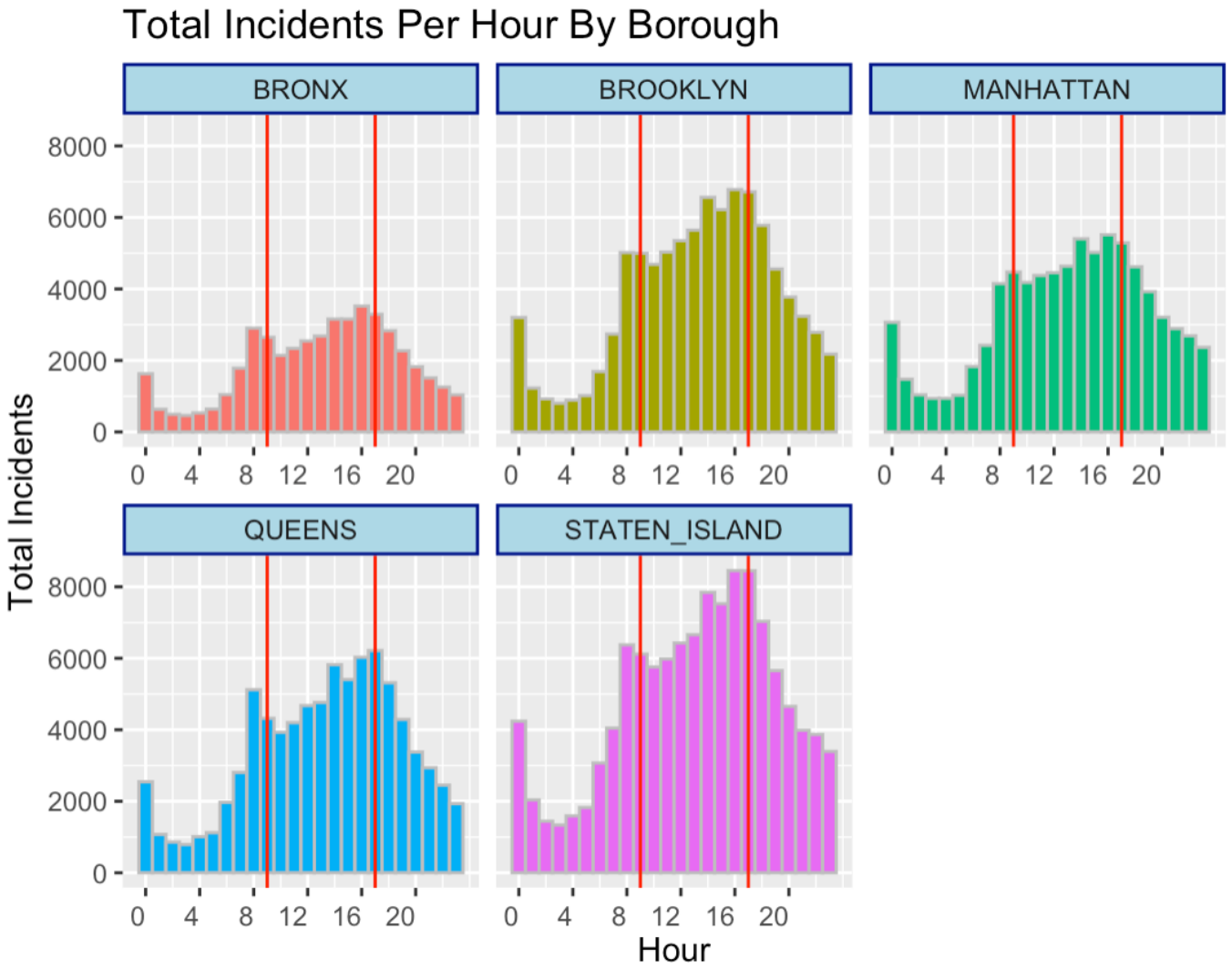
- We can see clear ranking of the different borough with regards to the total number of incidents in 2018 and 2019.
- Staten Island has the most number of total incidents and the Bronx has the least number of total incidents.
- Trend is almost identical from 2018 to 2019.

Insight: This trend is not surprising. Staten Island, queens and Brooklyn are larger than Manhattan and the Bronx



Total Incidents by the Hour

- Breakdown of incidents per hour for each Borough.
- The vertical red lines represent 8am to 5pm, a typical workday.
- The majority of total incidents tend to occur during the middle of the day/after-mid rather than the early and later hours.
- This is expected because most of the population is active during these hours.



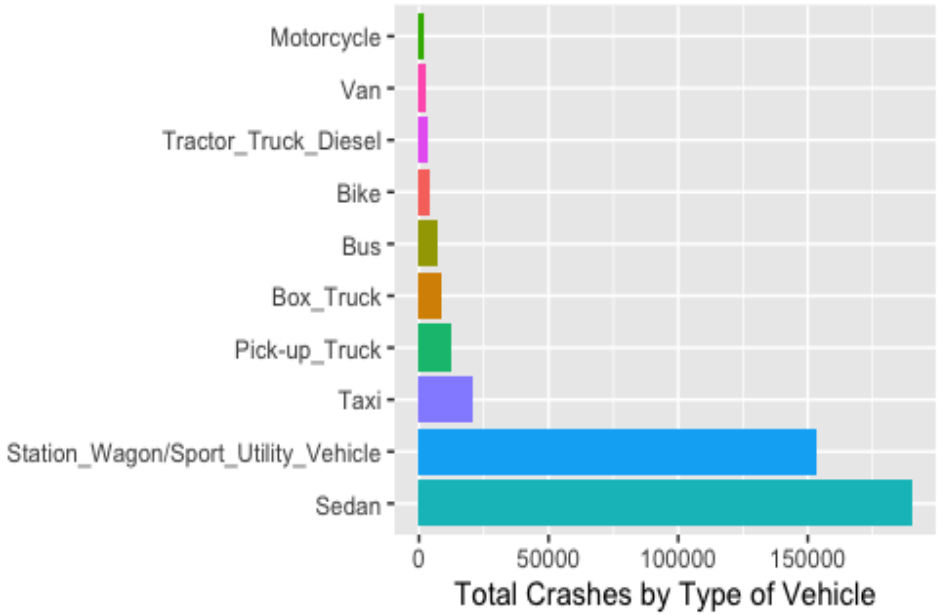
Data Exploration: Pedestrians vs Cyclist & Motorist

We wanted to narrow down our analysis to once type of incident category and decided to focus on Pedestrians. This is because:

- A majority of NYC residents don't own a car in the city.
- Motorist category had a lot of clutter and potentially bad data due to manual recording of vehicle codes.

Motorist Vehicle Codes

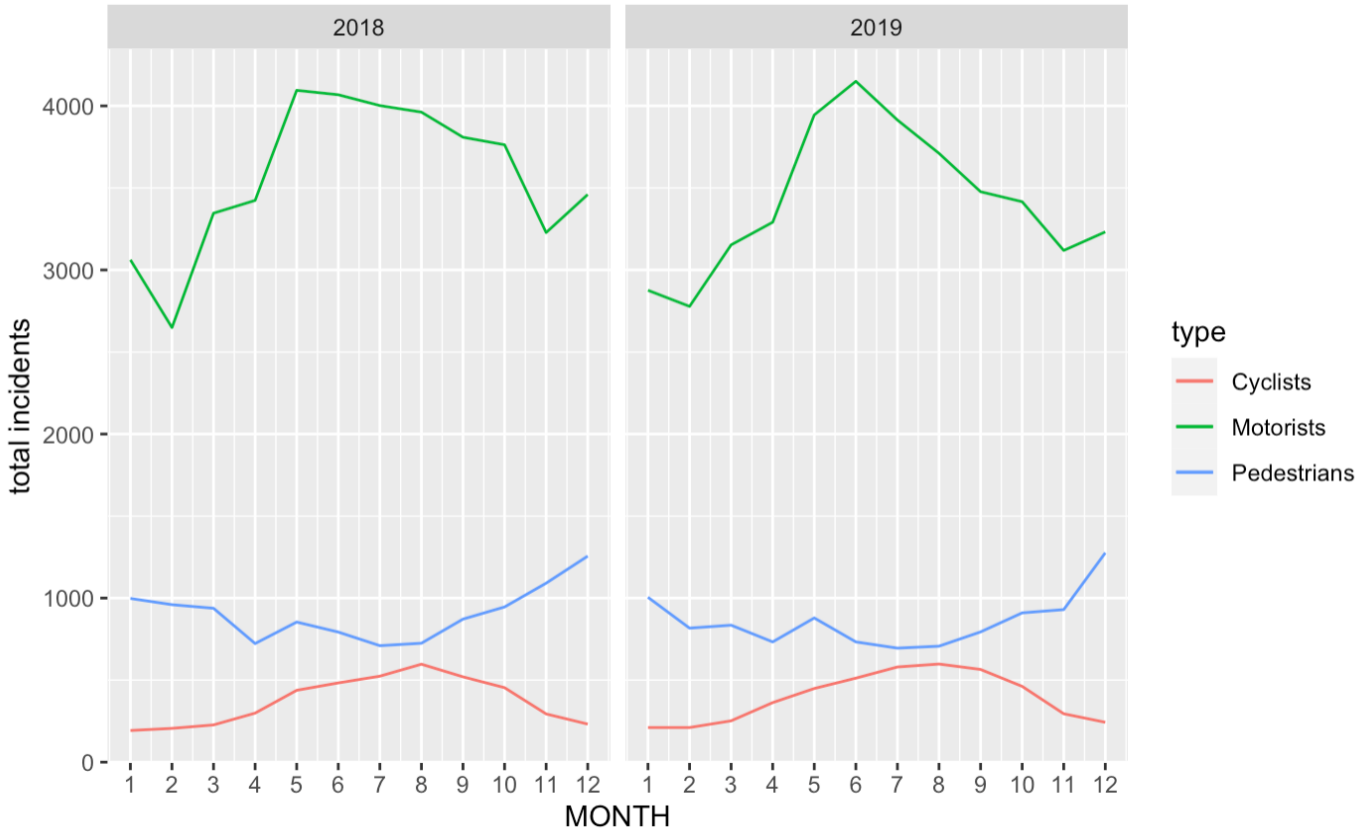
- There are 562 unique values within the Vehicle Code Type column (we have highlighted some of our favorites below).
- This messiness in data is potentially due to the human error when filling out the incident report form.
- Most incidents are caused by Sedans and SUVs however there are hundreds of other vehicle types listed.
- Decided to stay away from vehicle type and therefore also Motorist incidents for now due to messiness.



[151]	"HORSE_CARR"	"Vanette"
[185]	"scoot"	"NYPD"
[235]	"VAN_T"	"gator"
[387]	"Fork"	"Grain"
[489]	"Golf"	"back"
[519]	"DOT_T"	"BOX"

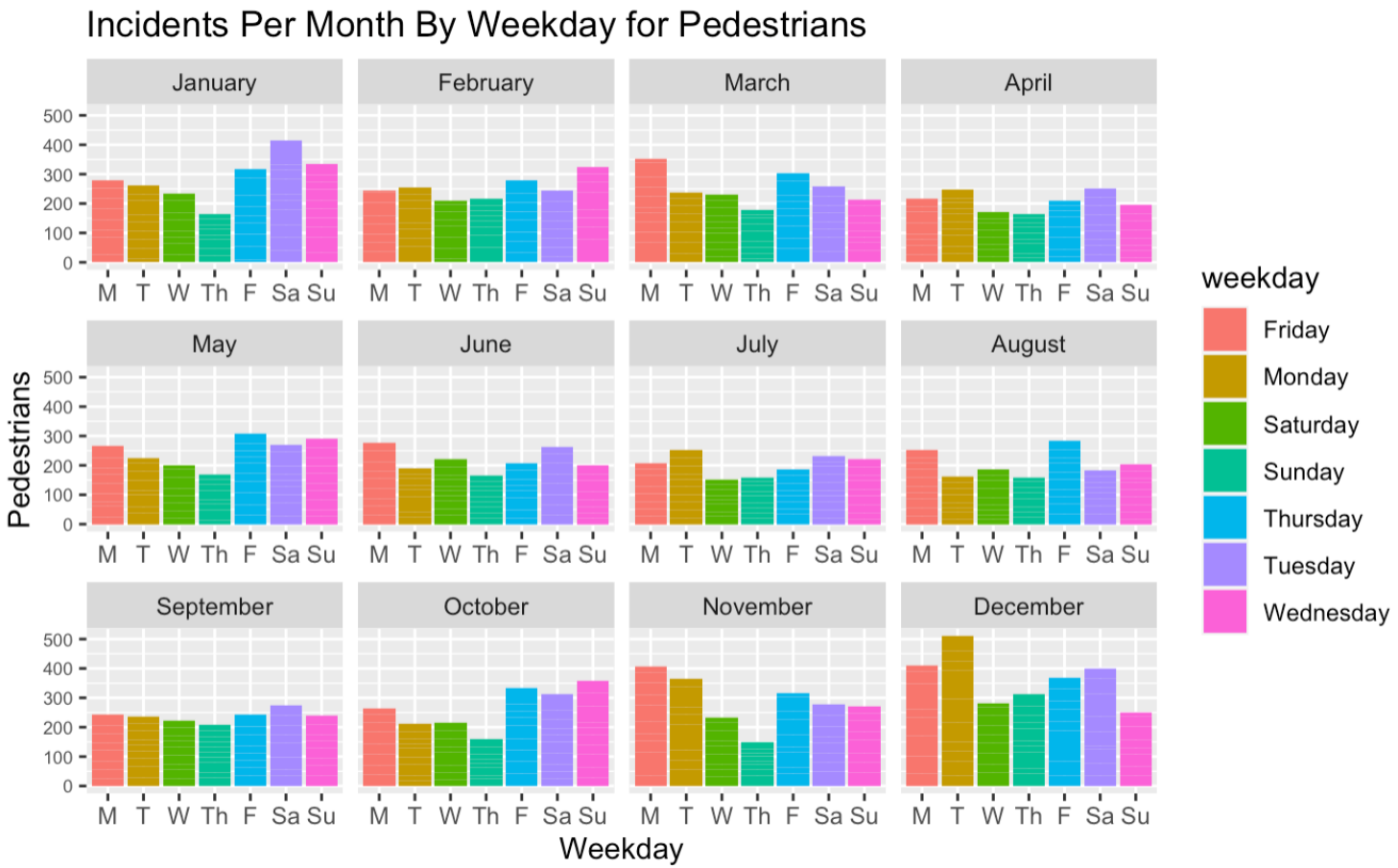
Insight: One would expect Taxis to be the greatest frequency in the Motor Vehicle Types. However, in NYC, taxis primarily operate in Manhattan. Since our data includes otther boroughs, the high frequency of Sedans and Station Wagons is not surprising.

Pedestrians, Cyclists and Motorists



- The dataset shows relation of total number of incidents by type: Motorists, Cyclists and Pedestrians.
- For the purpose of our data investigation, we want to focus in on pedestrians.
- However, we would like to note that our dataset has the potential to provide insights for cyclists and motorists.

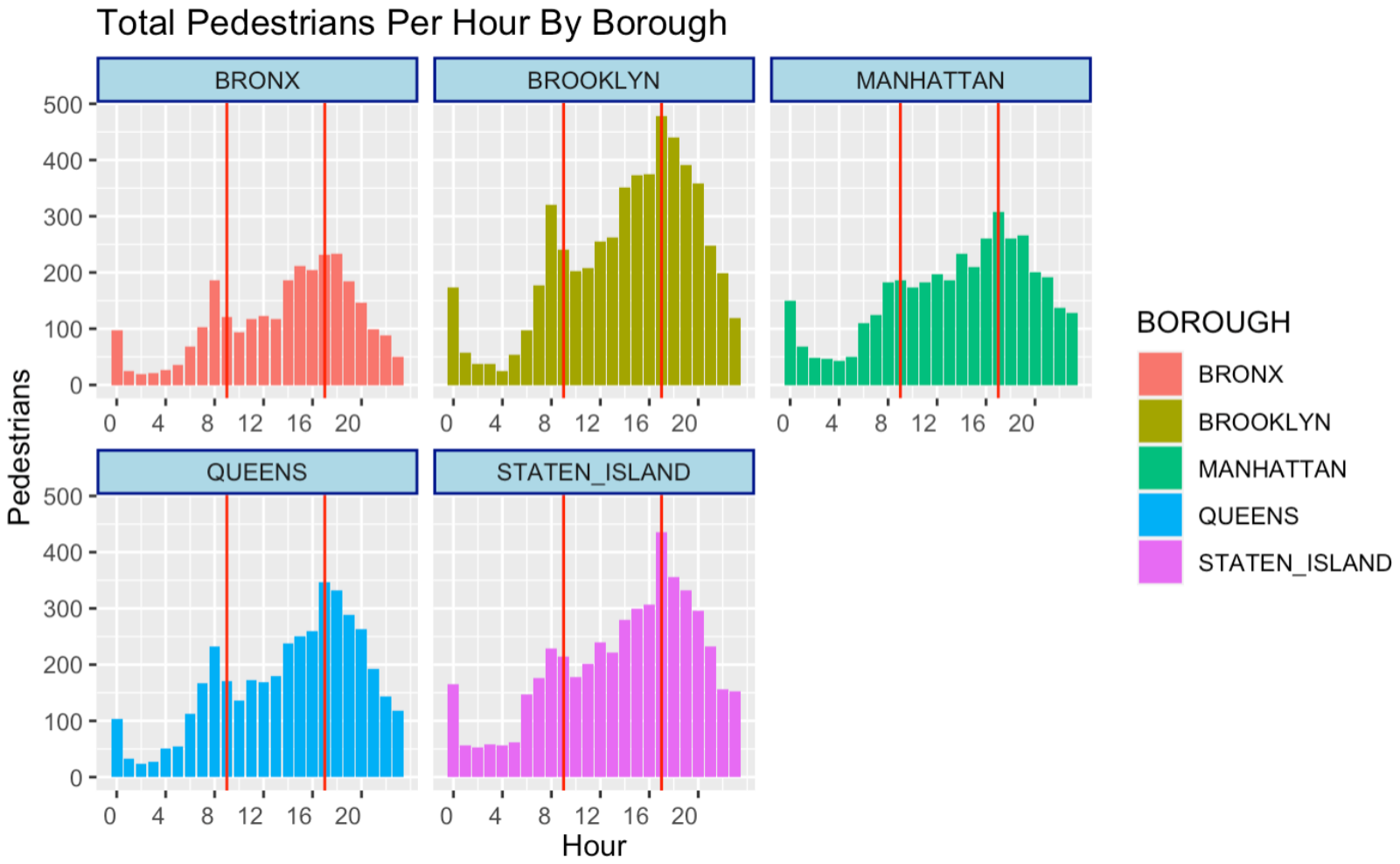
Pedestrian Incidents: Daily Comparison



- This chart shows an aggregate of all Pedestrian incidents for each day of the week (2018 and 2019 combined).

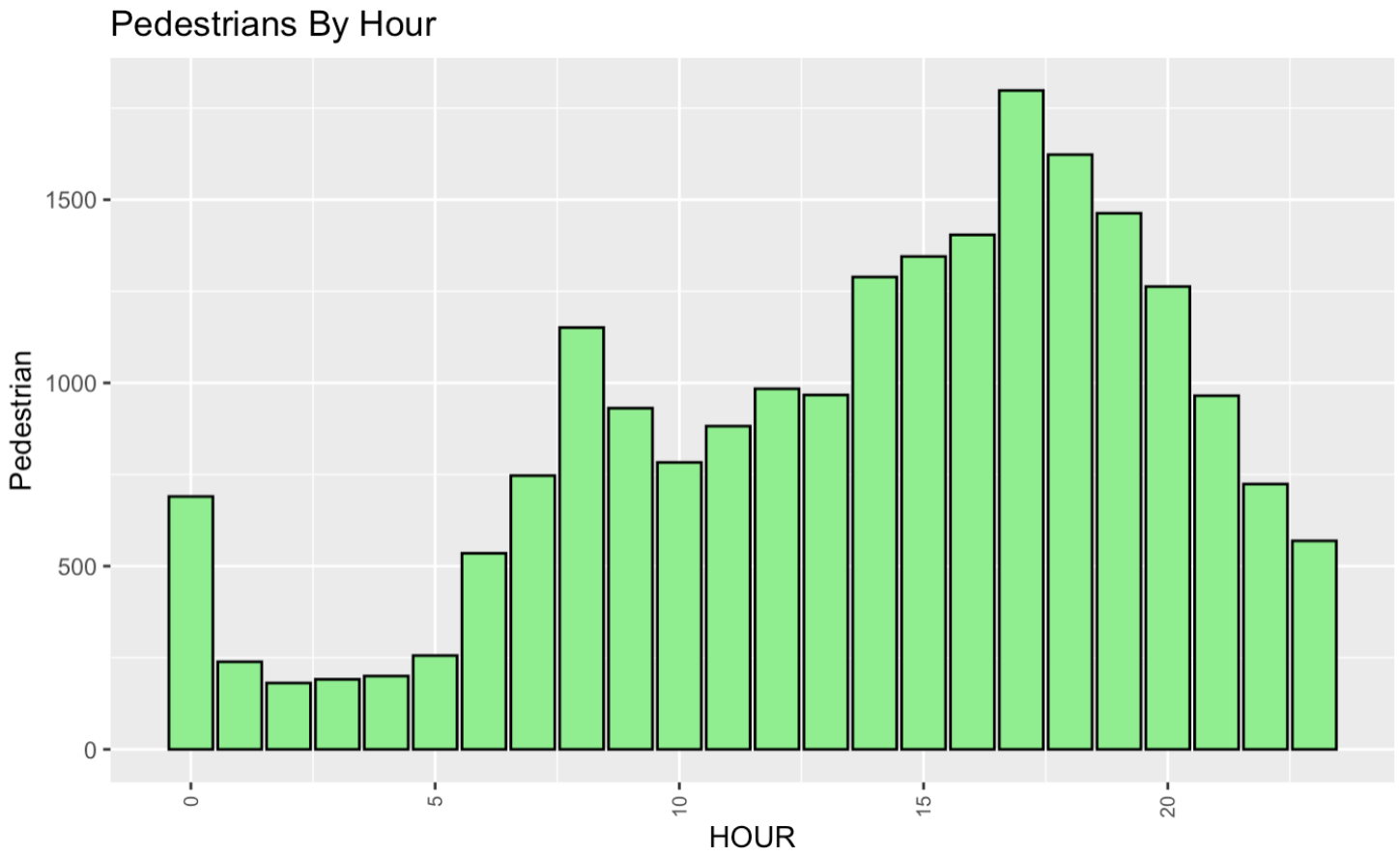
- Like we saw in the daily comparison of total incidents, there isn't much variation between days of the week per month.

Pedestrians Incidents per hour by Borough



- Breakdown of Pedestrian incidents per hour for each Borough.
- Insights remain the same: The majority of total incident tend to occur during the middle of the day/after-mid rather than the early and later hours.
- This is expected because most of the population is active during these hours.

Pedestrians per hour

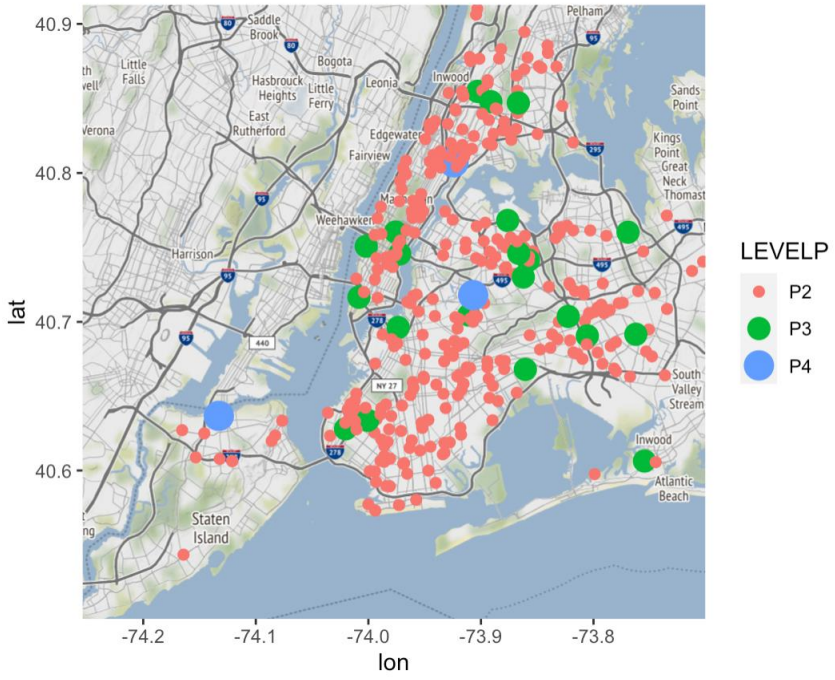


- As expected, we see incidents increase and remain higher during hours during the day since more people would be awake.
- We can see a sharp spike in the number of incidents between 7am and 8am which could be the result of the morning commute.
- Pedestrian incidents per hour follow the total incidents trend

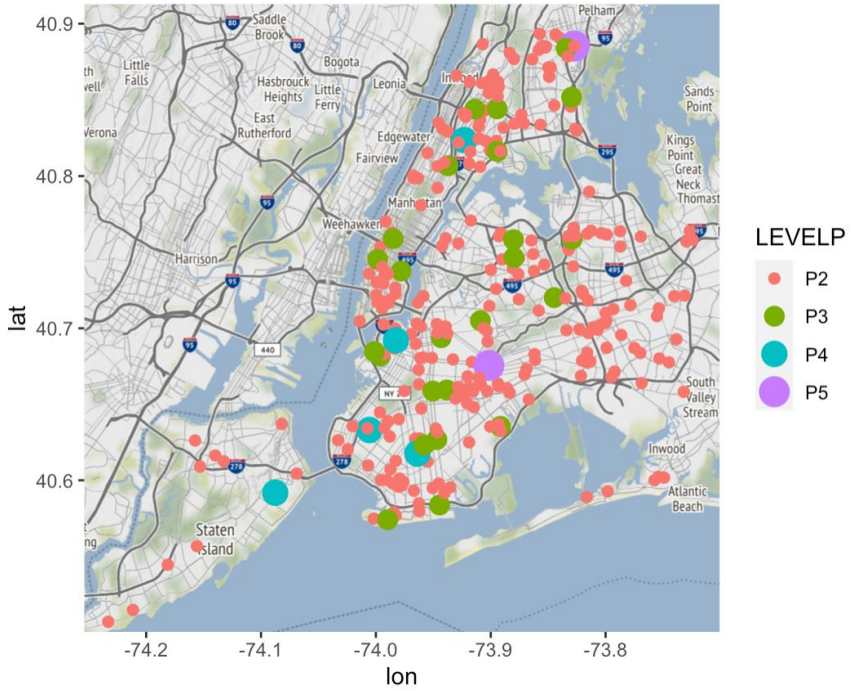
Geographic Pattern: Pedestrians Injured Levels

- The maps below show the number of pedestrians injured. Levels: 1, 2, 3, 4, and 5.
- Levels 1-4 represent the number of pedestrians injured.
- Level 5 represents 5 or more pedestrians injured.
- We can see that in 2019, there were a couple of incidents with 5 or more pedestrians injured while there were no such incidents in 2018

2018



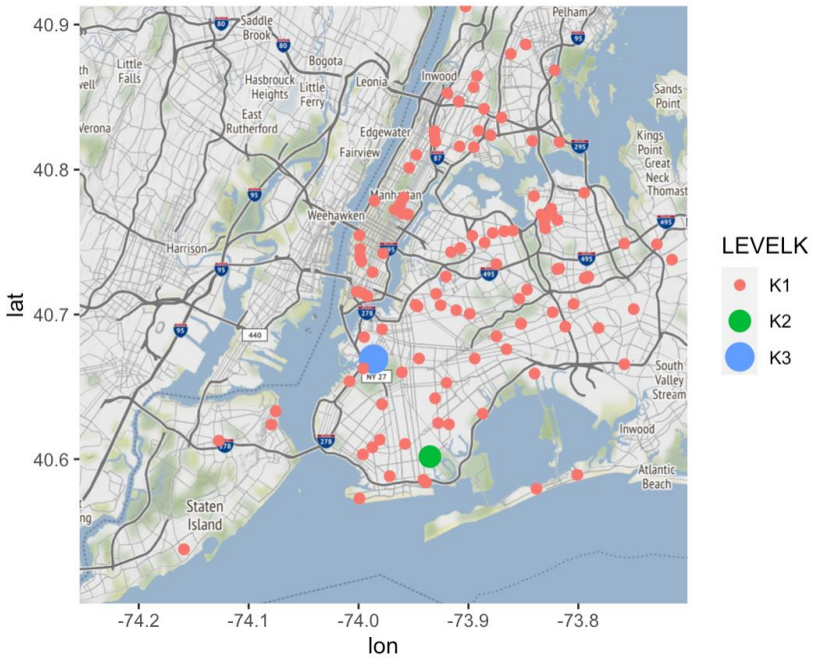
2019



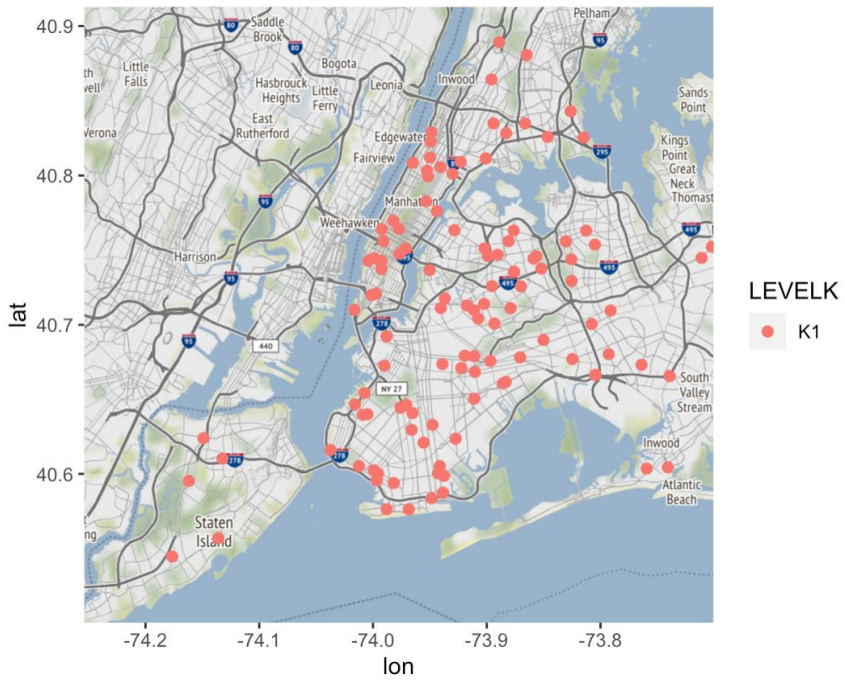
Geographic Pattern: Pedestrians Killed Levels

- The maps below show the number of pedestrians killed.
- Levels: 1, 2 and 3. Levels 1 and 2 represent the number of pedestrians killed, respectively.
- Level 3 represents an incidents when a pedestrian is injured, and another is killed.
- We can see that in 2019, there were a couple of incidents with 5 or more pedestrians injured while there were no such incidents in 2018

2018



2019



Models – Linear Regression and SVM

For our models, we wanted to:

- Explore linear relationship between pedestrian injuries/deaths and our other variables using regression models
- Predict whether a pedestrian injury/death would occur using SVM

Regression

```
Call:
lm(formula = Pedestrian ~ con + bor + YEAR + MONTH + DAY, data = r2)

                                                        Estimate Std. Error t value Pr(>|t|)
conAggressive_Driving/Road_Rage                        8.802e-02  1.606e-02   5.480 4.26e-08 ***
conFailure_to_Yield_Right-of-Way                       1.307e-01  1.525e-02   8.573  < 2e-16 ***
conFatigued/Drowsy                                    -3.708e-02  1.873e-02  -1.979  0.04778 *
conFell_Asleep                                        -3.551e-02  1.647e-02  -2.156  0.03110 *
conFollowing_Too_Closely                              -4.006e-02  1.524e-02  -2.629  0.00856 **
conGlare                                               1.719e-01  1.719e-02  10.001  < 2e-16 ***
conPedestrian/Bicyclist/Other_Pedestrian_Error/Confusion  2.760e-01  1.585e-02  17.417  < 2e-16 ***
conVehicle_Vandalism                                  -4.414e-02  3.820e-02  -1.155  0.24790
conView_Obstructed/Limited                             6.301e-02  1.564e-02   4.028 5.63e-05 ***
conWindshield_Inadequate                               2.768e-02  6.117e-02   0.453  0.65089
bor2                                                  -1.110e-02  1.122e-03  -9.887  < 2e-16 ***
bor3                                                   8.932e-04  1.096e-03   0.815  0.41527
bor4                                                   5.869e-03  1.332e-03   4.406 1.06e-05 ***
bor5                                                  -8.577e-03  1.031e-03  -8.320  < 2e-16 ***
YEAR                                                   9.843e-04  6.954e-04   1.415  0.15693
MONTH                                                  5.342e-04  1.022e-04   5.227 1.72e-07 ***
DAY                                                   -2.530e-05  3.962e-05  -0.638  0.52317
---

Residual standard error: 0.2217 on 408974 degrees of freedom
  (1375 observations deleted due to missingness)
Multiple R-squared:  0.04436,   Adjusted R-squared:  0.04422
F-statistic: 311.3 on 61 and 408974 DF,  p-value: < 2.2e-16
```

- Trying to fit chaos into a linear model
- R-squared is very low
- Con --> contributing factor
- Incidents and x variables aren't linear. When running linear regression, we don't expect adjusted r-squared would be high or significant.
- Started with just Pedestrian and contributing factor and got adjusted-r squared of 4.4

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 4529   16
         1    0  246

               Accuracy : 0.9967
                 95% CI : (0.9946, 0.9981)
    No Information Rate : 0.9453
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.9667

 Mcnemar's Test P-Value : 0.0001768

            Sensitivity : 1.0000
            Specificity : 0.9389
         Pos Pred Value : 0.9965
         Neg Pred Value : 1.0000
             Prevalence : 0.9453
         Detection Rate : 0.9453
   Detection Prevalence : 0.9487
      Balanced Accuracy : 0.9695

       'Positive' Class : 0
```

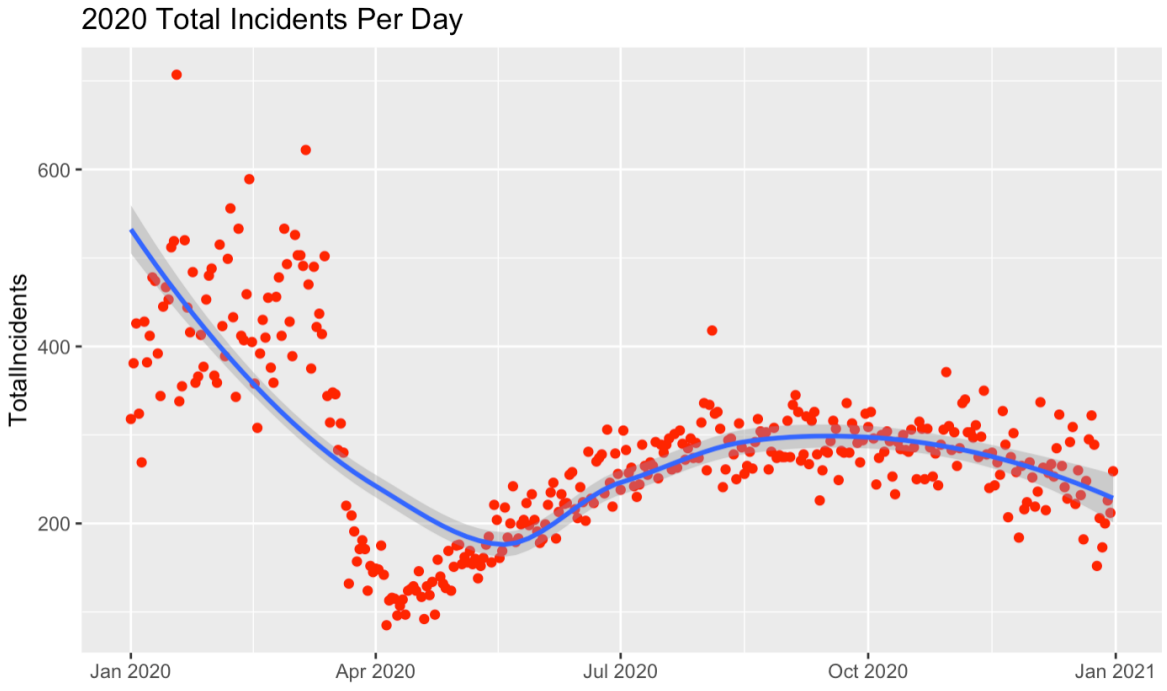
Predict Pedestrian Injury or Death

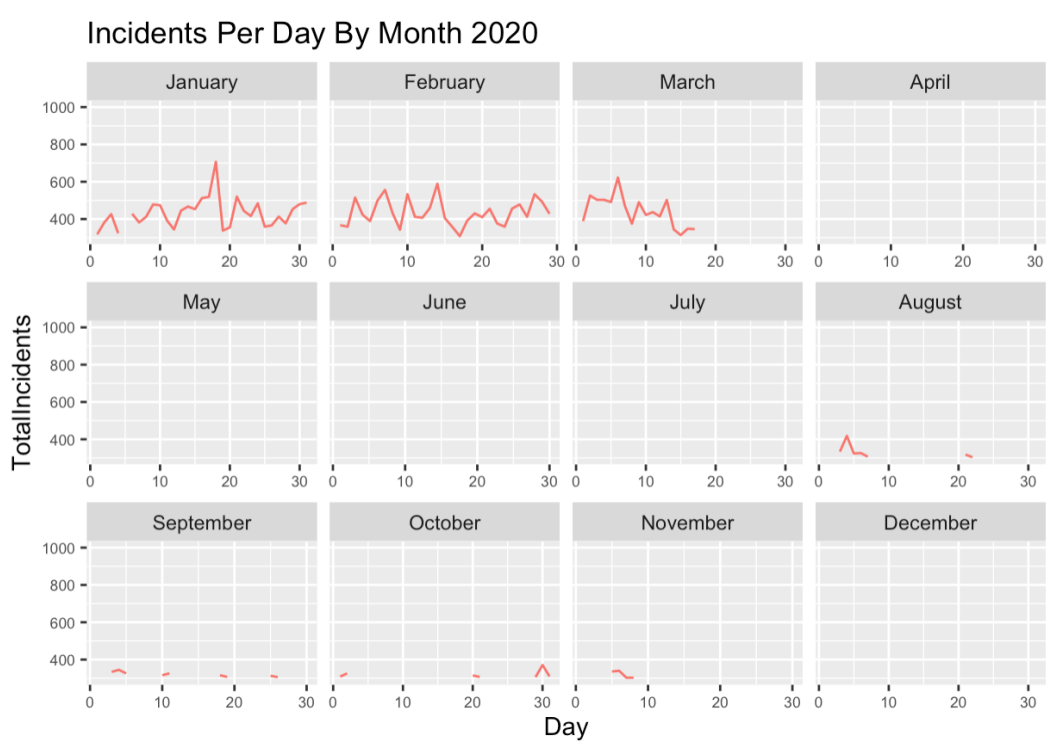
- SVM model
- Y variable: Pedestrian injury/death
  - 0 = no injuries/death
  - 1 = 1 or more injuries/death
- X variables: Motorist deaths, motorist injuries, cyclist deaths, cyclist injuries
- P-value: Statistically significant
- No Information Rate: 94.5%
- Accuracy: 99.6%
  - Model raises accuracy by 5%

Going beyond our scope: What about 2020?

What about 2020?

- Our data cleaning code works for any time section taken from the NYC Motor Vehicle Crashes. Using the same code, we imported in crash data for 2020. The data contains 112,893 incidents. We can see a huge drop in the total incidents between January 2020 and April 2020, then the total incidents follow an inverse parabola from July to Dec 2020. Looking closer, we can see that in the middle of March, around March 18th is when the total number of incidents drops off.

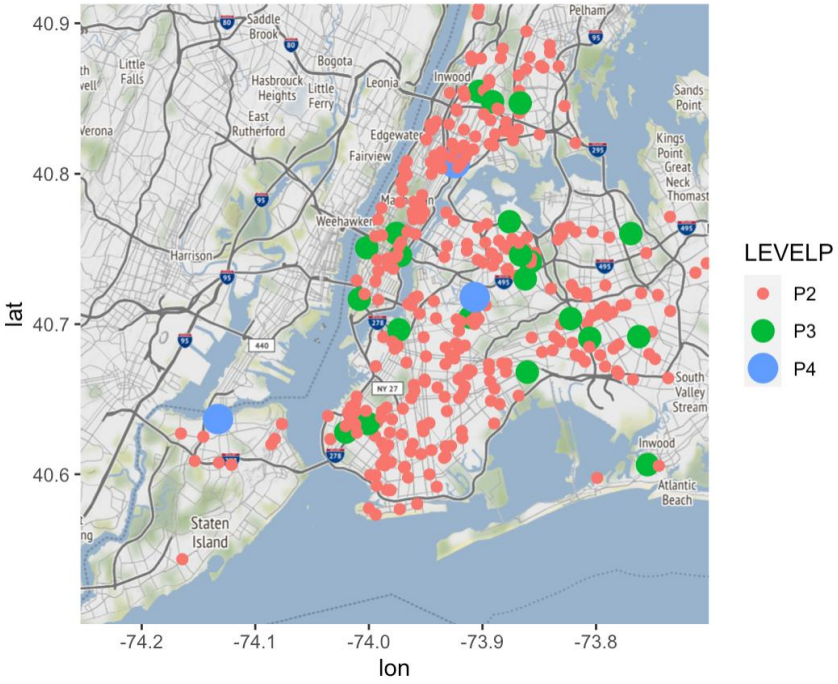




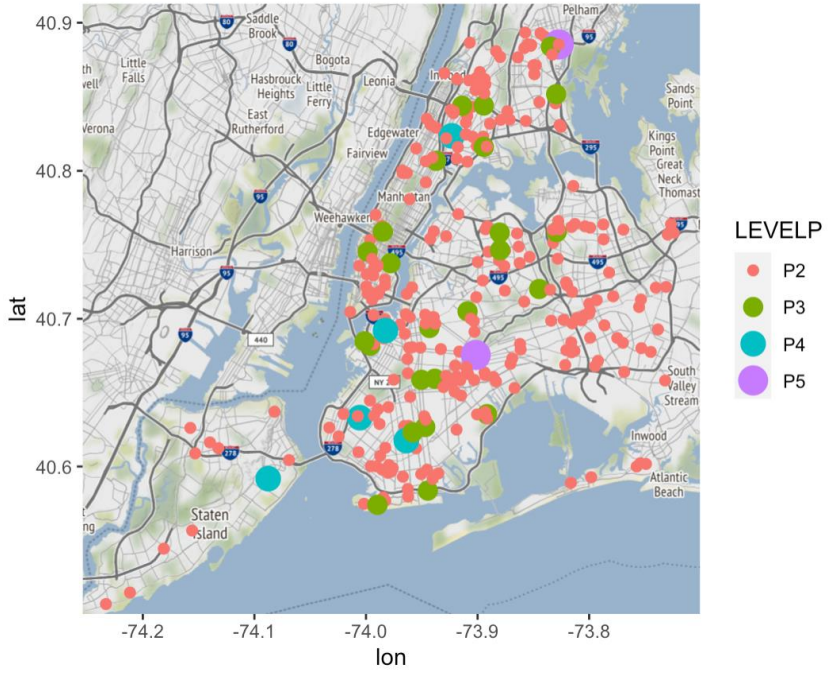
2018, 2019, and 2020

- The maps below show the number of pedestrians injured. Levels: 1, 2, 3, 4, and 5. Levels 1-4 represent the number of pedestrians injured. Level 5 represents 5 or more pedestrians injured. Even though there were fewer incidents in 2020 from April to July, where the crashes occur still mirror that of 2018 and 2019

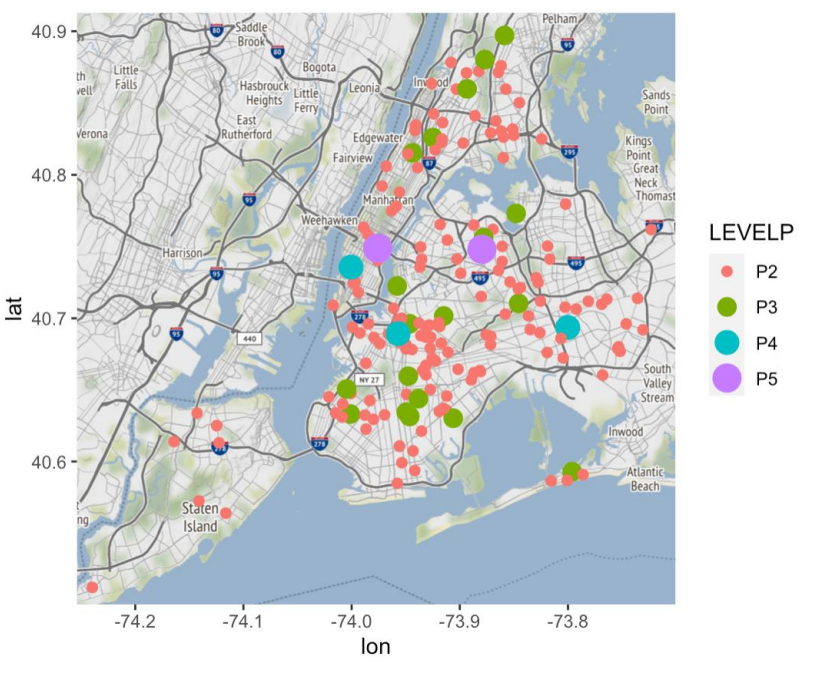
2018



2019



2020



Final Insights

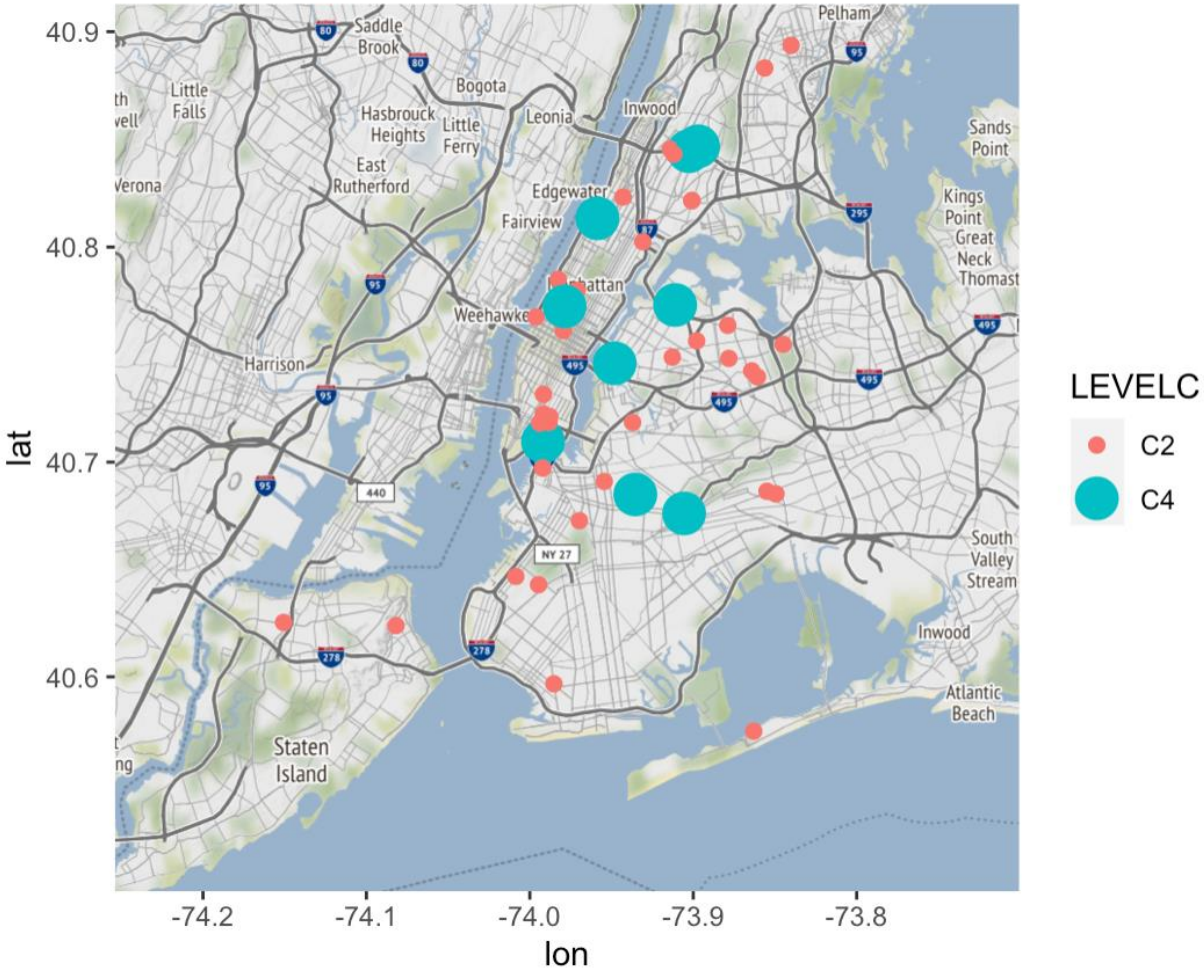
- 2018 and 2019 trends for month, day, year, borough is all follow the same trends without many outliers.
- Data is chaotic so linear predictions are weak – should try to merge with other traffic datasets to potentially strengthen predictors.
- Our analysis provides a good foundation for extended analysis into 2020 to view changes caused by COVID-19.

Appendix

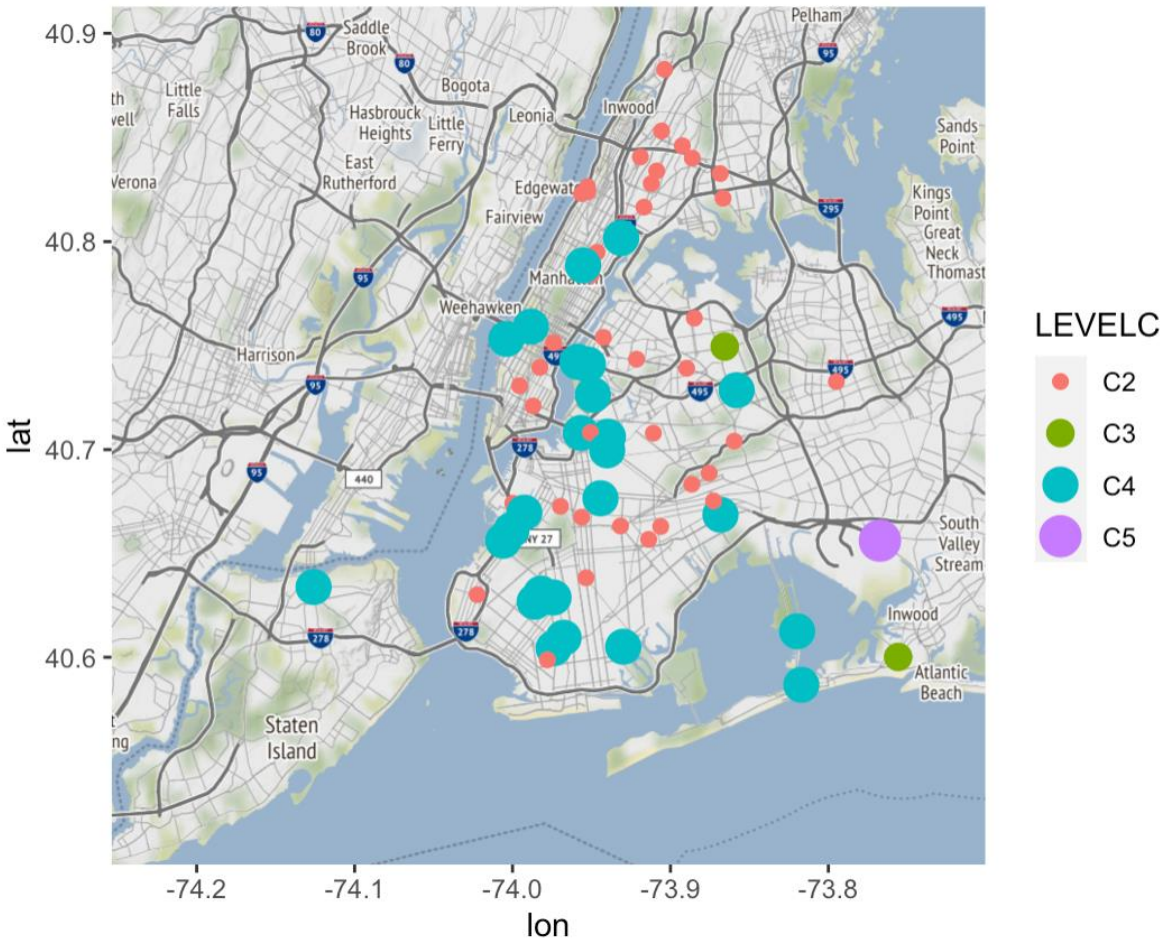
AP 01

Cyclists Danger Levels

2018



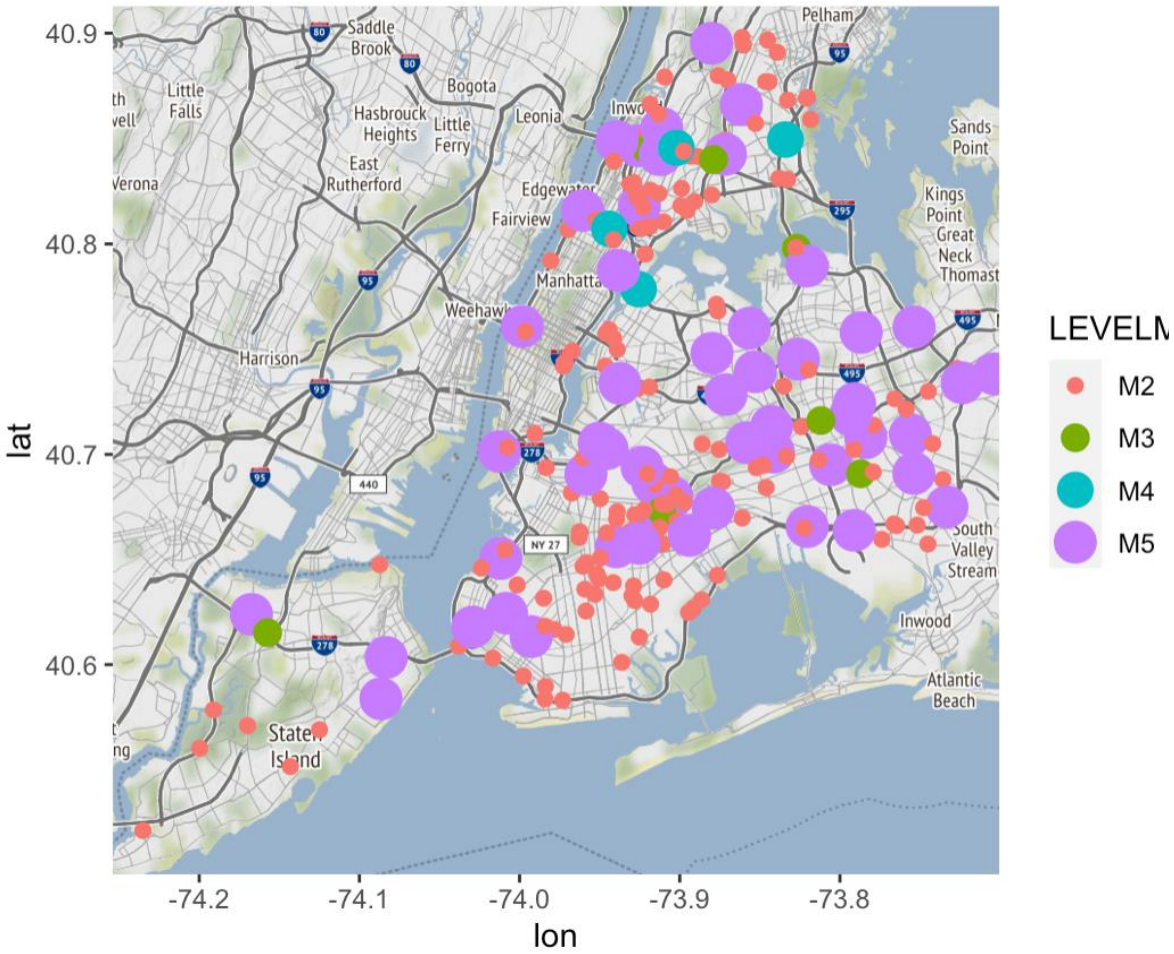
2019



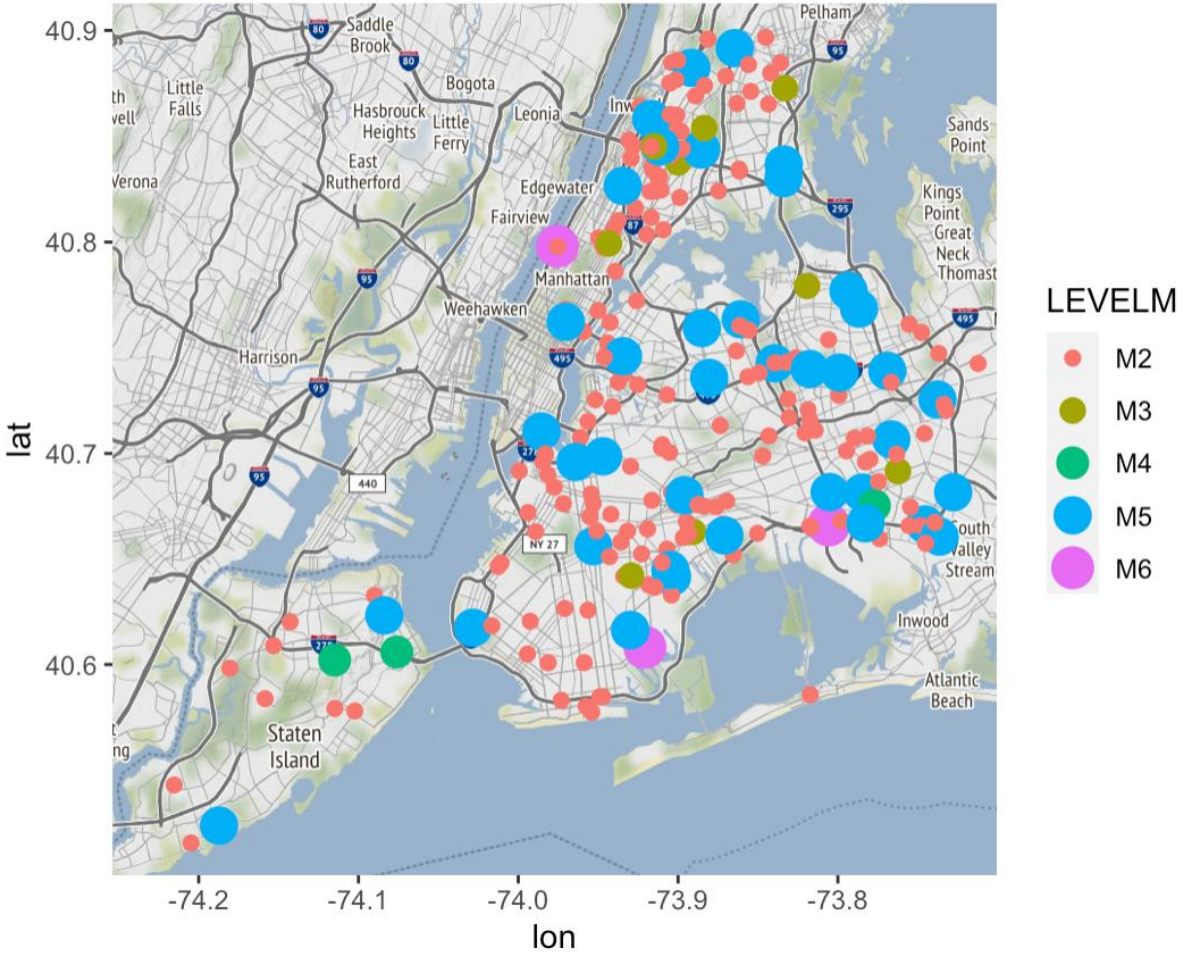
AP 02

Motorists Danger Levels

2018



2019



AP 03

Regression 1

```
Call:
lm(formula = Pedestrian ~ con, data = r2)

                                                          Estimate Std. Error t value Pr(>|t|)
conPedestrian/Bicyclist/Other_Pedestrian_Error/Confusion  0.277929   0.015854  17.531  < 2e-16 ***
conTinted_Windows                                         0.130160   0.043897   2.965  0.00303 **
conView_Obstructed/Limited                                0.063695   0.015648   4.071 4.69e-05 ***
conUnsafe_Lane_Changing                                  -0.039599   0.015318  -2.585  0.00973 **

Residual standard error: 0.2218 on 408981 degrees of freedom
  (1375 observations deleted due to missingness)
Multiple R-squared:  0.04363,   Adjusted R-squared:  0.0435
F-statistic: 345.5 on 54 and 408981 DF,  p-value: < 2.2e-16
```

AP 04

Regression 2

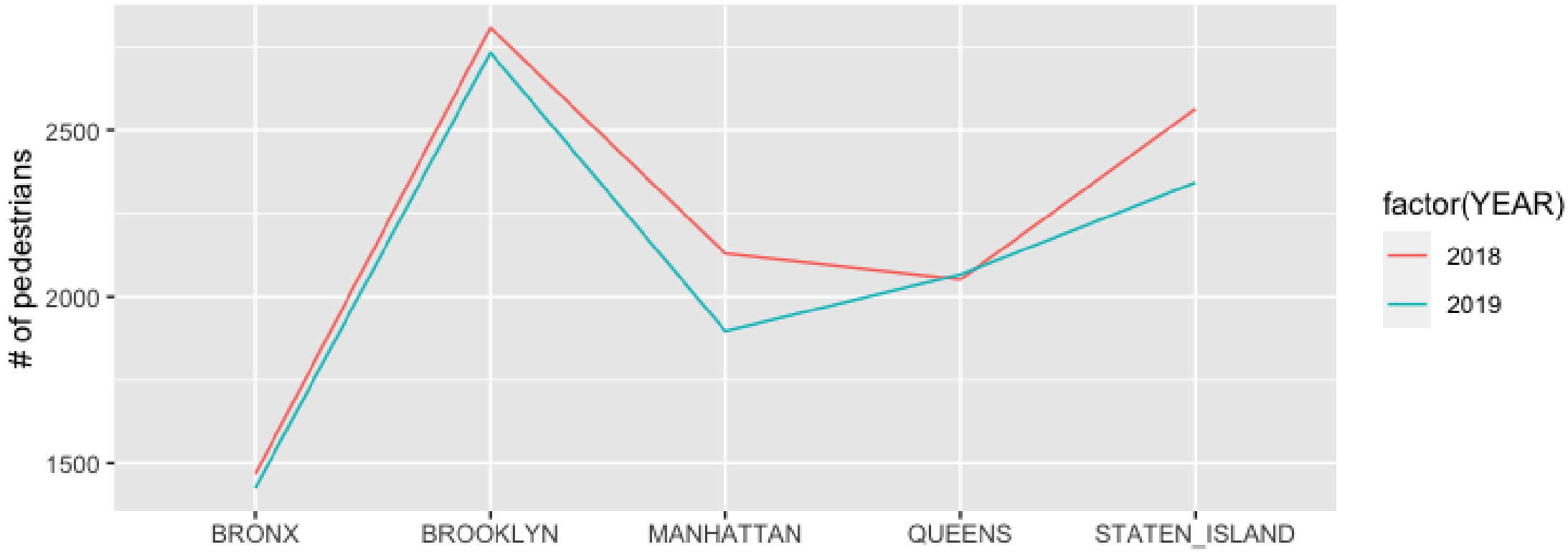
```
Call:
lm(formula = Pedestrian ~ con + bor, data = r2)

Coefficients:
                                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                                      0.0465888  0.0152125   3.063  0.00219 **
conAggressive_Driving/Road_Rage                  0.0879519  0.0160627   5.476 4.36e-08 ***
conFollowing_Too_Closely                        -0.0401760  0.0152368  -2.637  0.00837 **
conGlare                                         0.1715630  0.0171861   9.983  < 2e-16 ***
conPedestrian/Bicyclist/Other_Pedestrian_Error/Confusion  0.2760402  0.0158493  17.417  < 2e-16 ***
conVehicle_Vandalism                            -0.0440579  0.0382042  -1.153  0.24882
conView_Obstructed/Limited                       0.0629457  0.0156425   4.024 5.72e-05 ***
conWindshield_Inadequate                         0.0264339  0.0611689   0.432  0.66564
bor2                                            -0.0110520  0.0011222  -9.849  < 2e-16 ***
bor3                                             0.0009162  0.0010964   0.836  0.40335
bor4                                             0.0059055  0.0013321   4.433 9.28e-06 ***
bor5                                            -0.0086082  0.0010308  -8.351  < 2e-16 ***

Residual standard error: 0.2217 on 408977 degrees of freedom
  (1375 observations deleted due to missingness)
Multiple R-squared:  0.0443,    Adjusted R-squared:  0.04416
F-statistic: 326.8 on 58 and 408977 DF,  p-value: < 2.2e-16
```

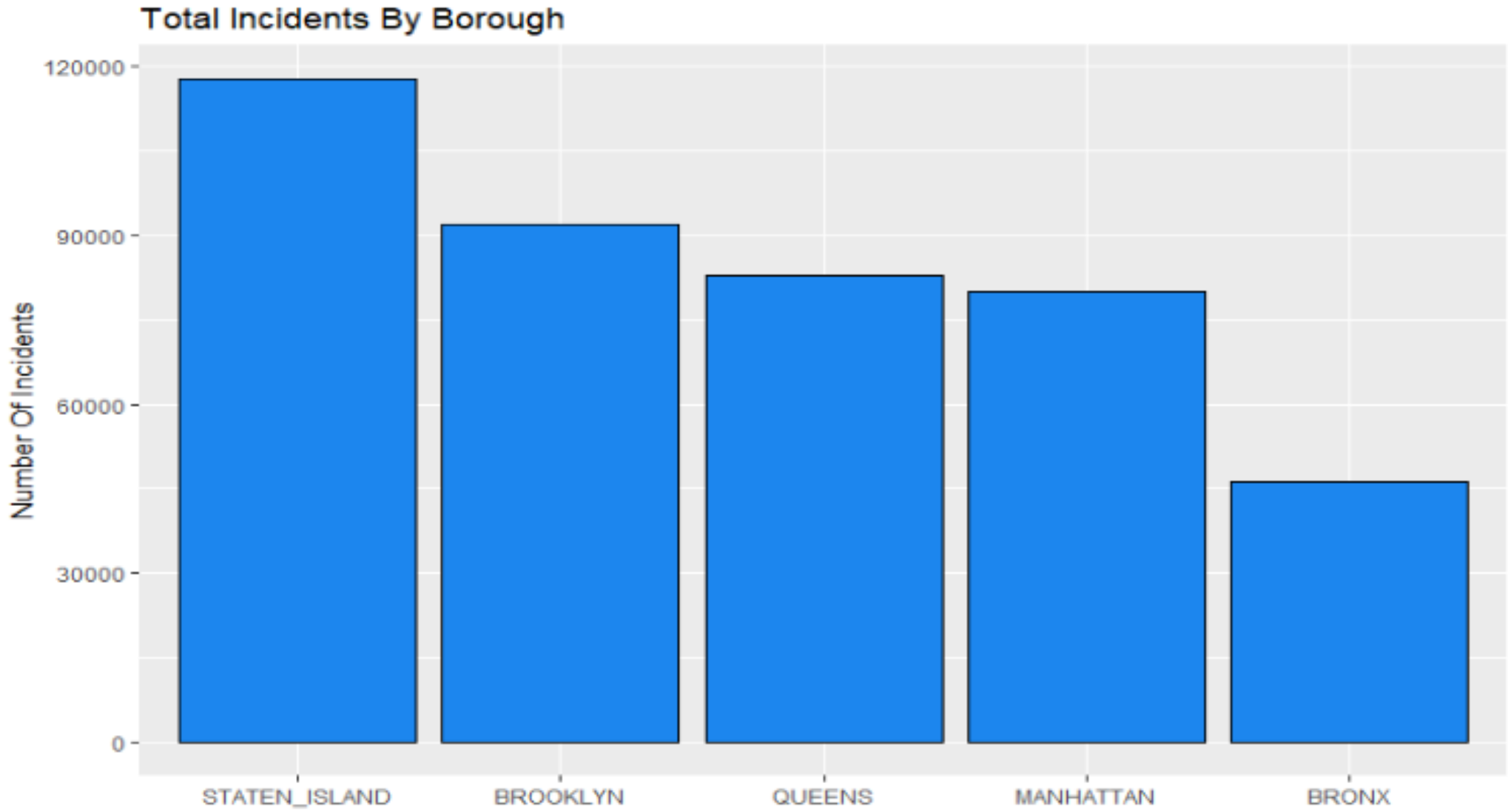
- to do

AP 06 Trend of pedestrians injured by borough across 2018-2019



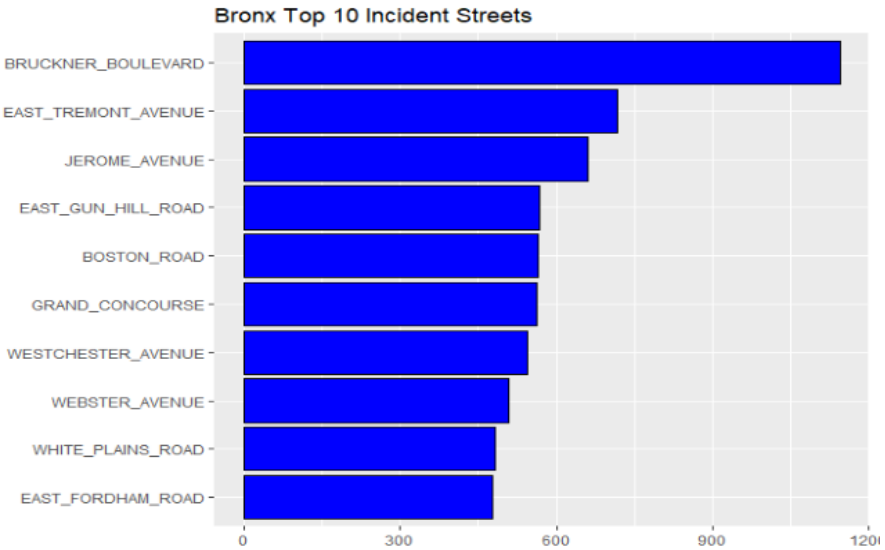
AP 07

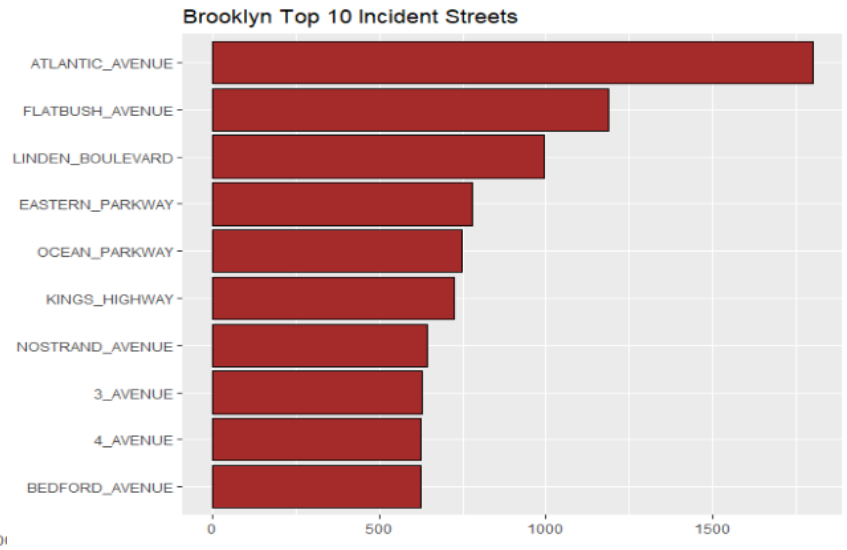
Incidents By Borough

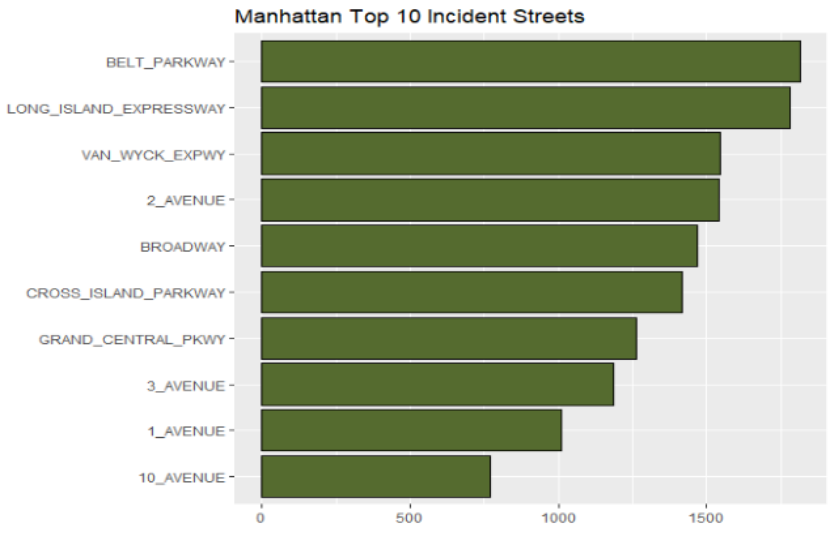


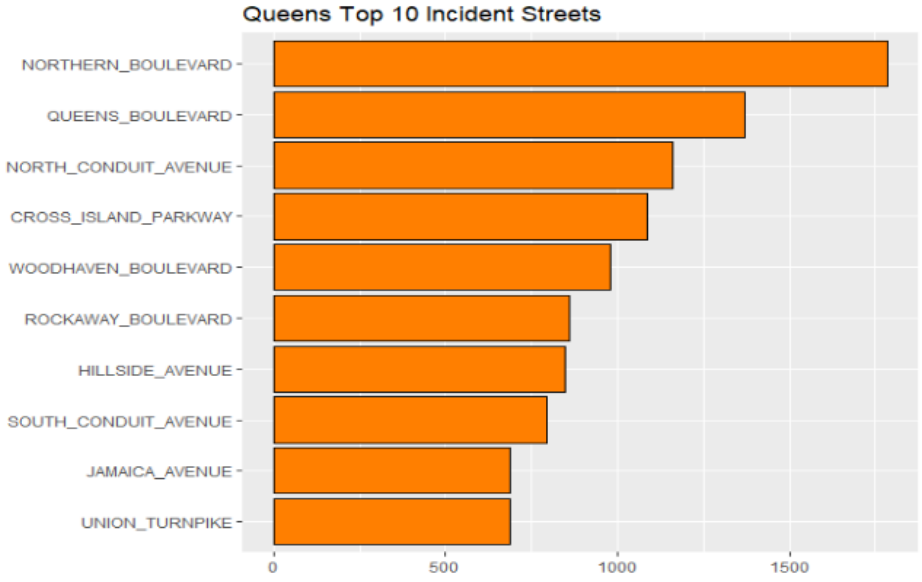
AP 09

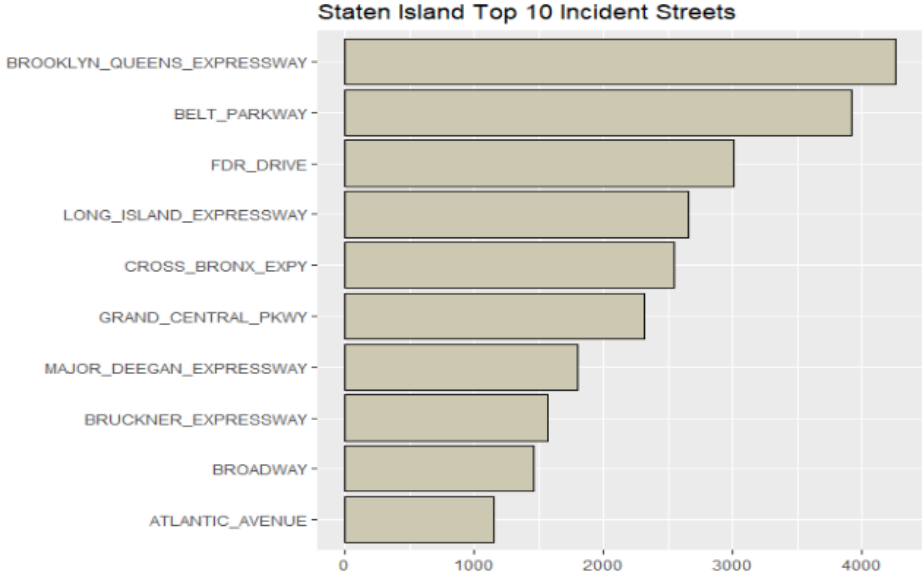
Top 20 Incident Streets Per Borough



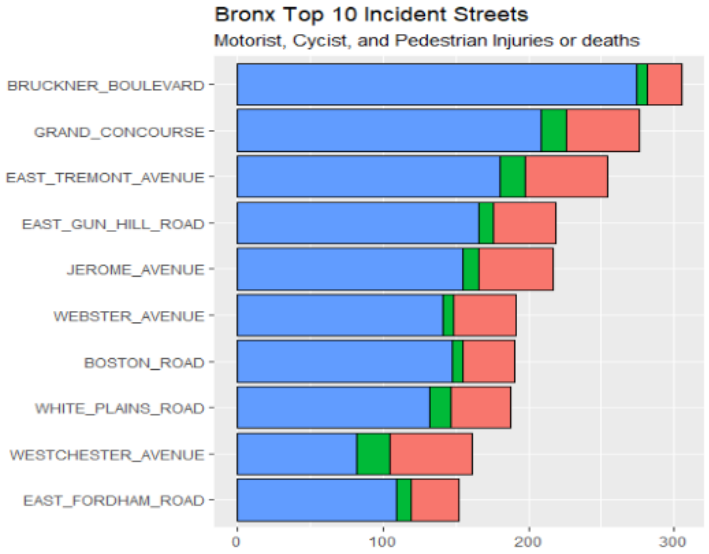


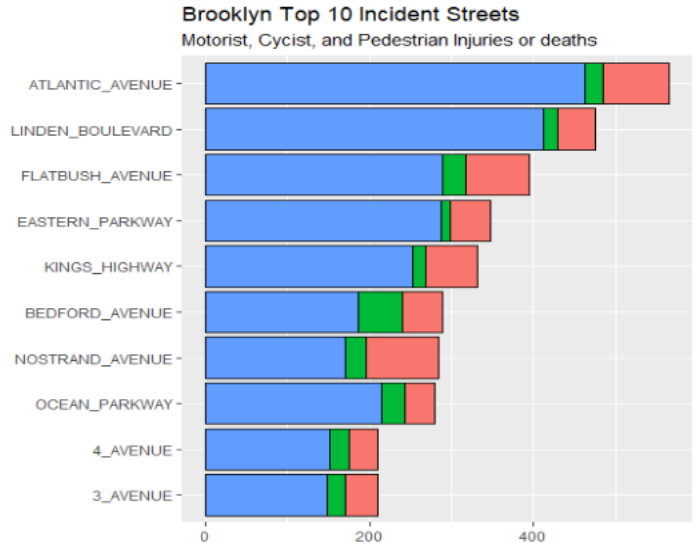


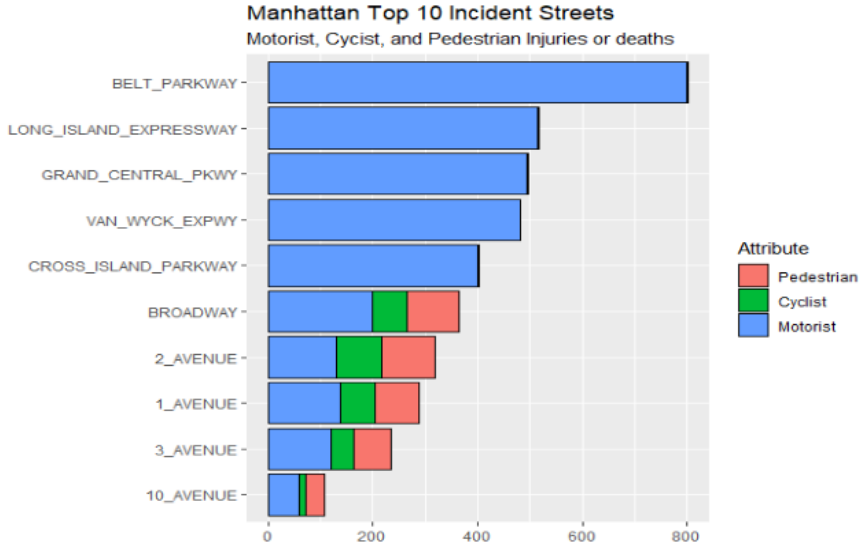


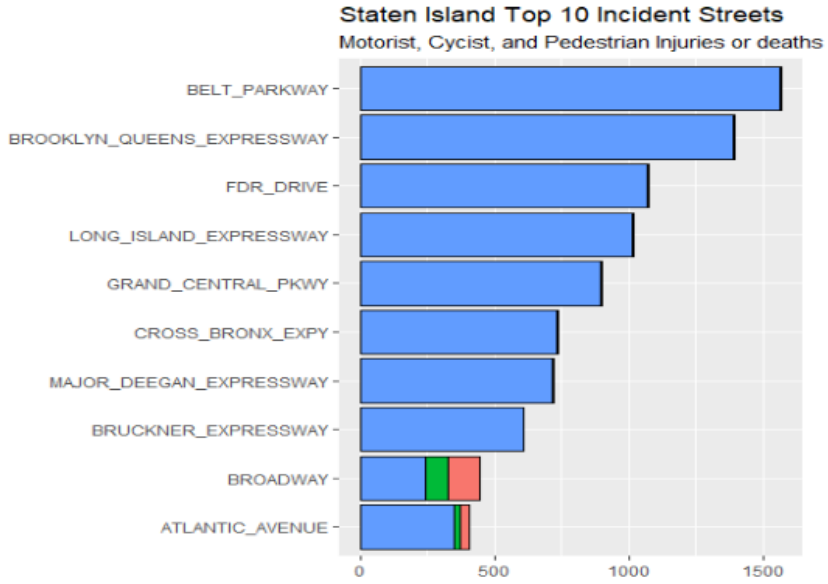


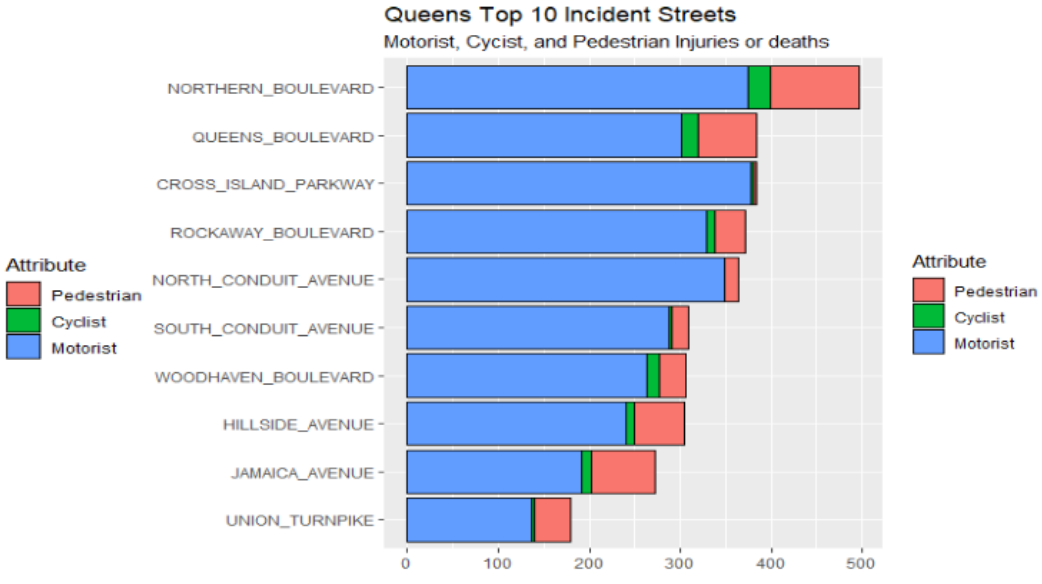
AP 10 Top 20 Incident Streets – What type of deaths/injuries?





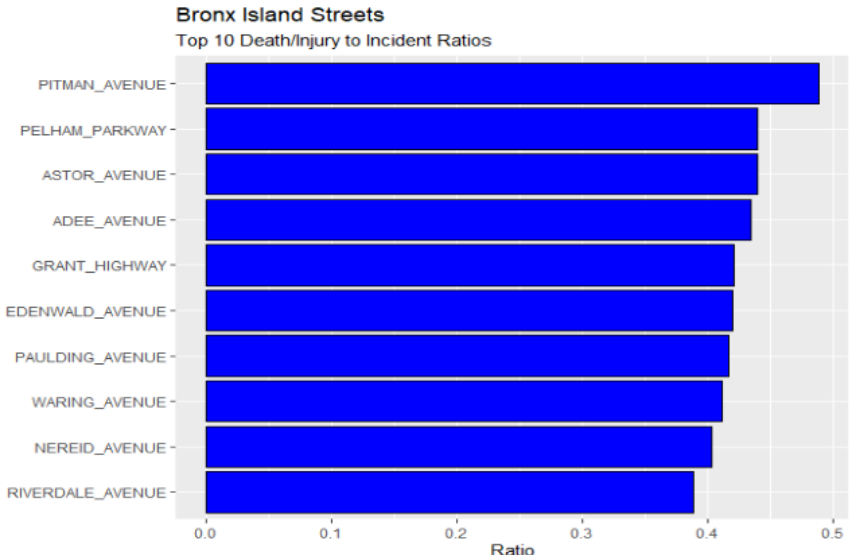


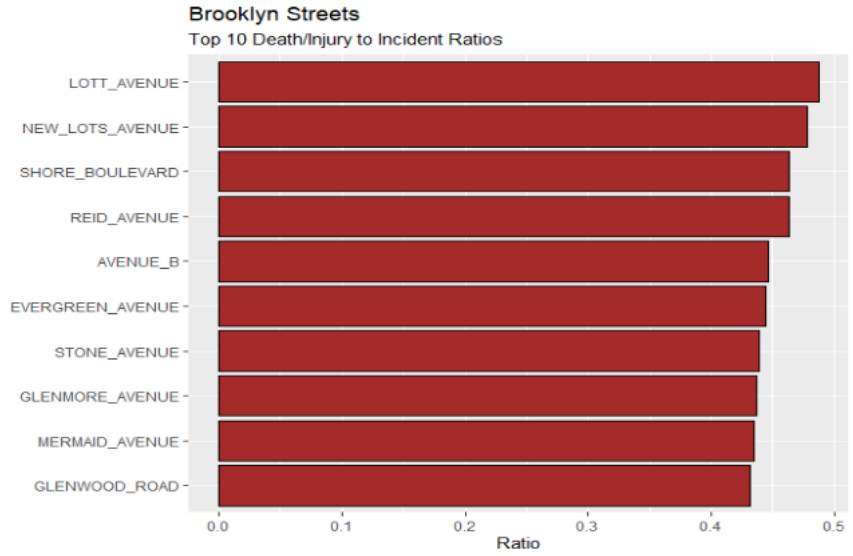


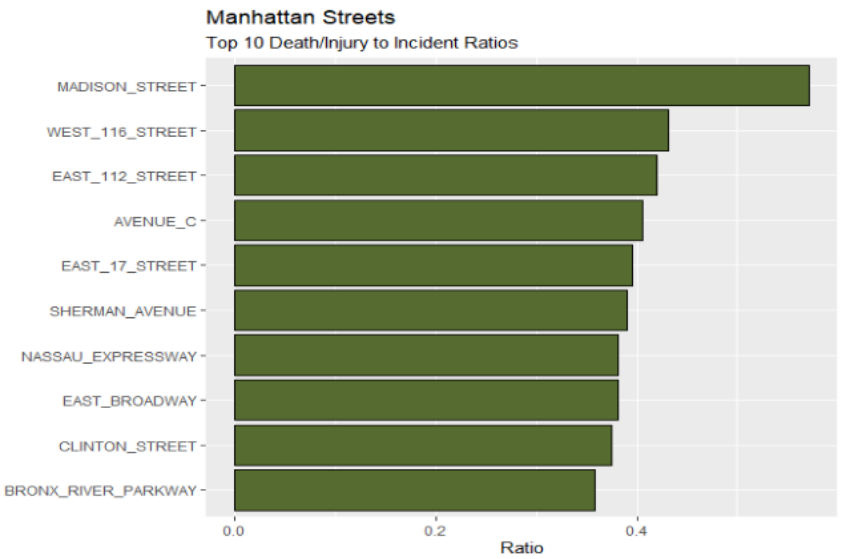


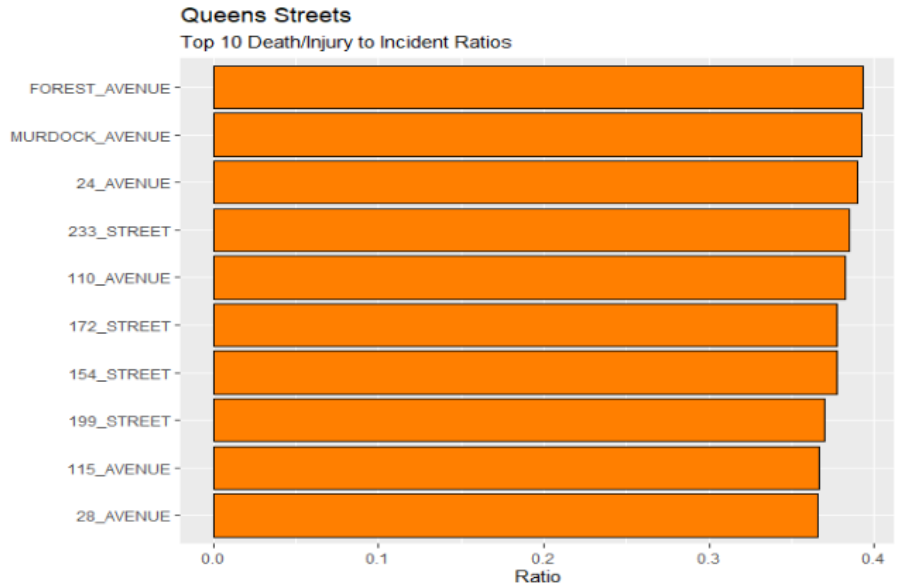
AP 11

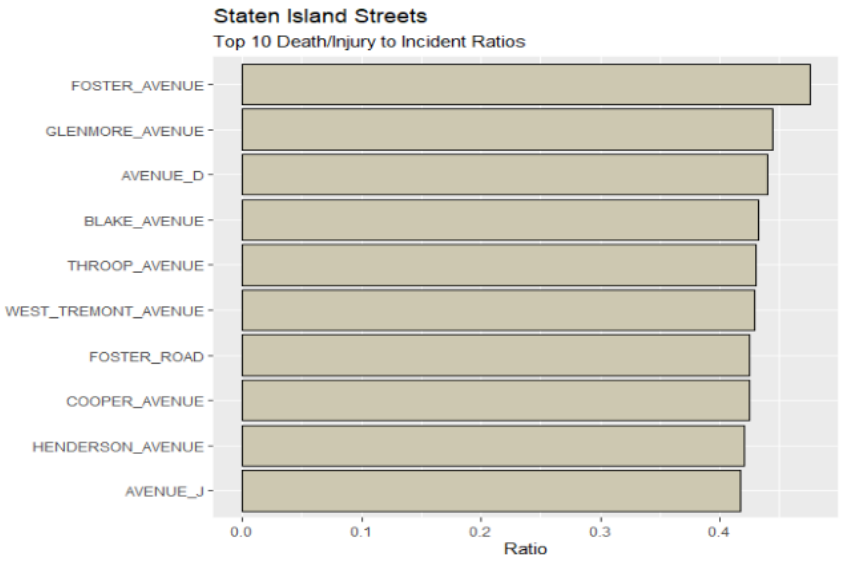
Top 20 Streets: Death/Injury to Incident Ratios











AP 12

Top 20 Death/Injury to Incident Ratios