Team : Chaitanya Patel and Herat Patel

# *Introduction:* We have performed investigations on the New York City car crash data. The data has records until early months of 2021 hence we can use the it to analyze the accidents and also observe how the trends in the data change with the COVID-19 pandemic and the lockdowns applied to to it. The goal is to analyze the the crash data and provide some actionable information to the city which will be used to avoid these accidents and also to observe any changes the city needs to do in the existing measures in place in view of the pandemic.

# *Data cleaning:*

After reviewing the investigations we have to carry out on the data, we decided to filter out some attributes as we were not going to be using them and only keep certain relevant fields like boroughs, month, year, type of accident, latitude, longitude, etc.

Since a lot of our evaluations are related to the date, we have created 3 new attributes, ‘CRASH YEAR’, 'CRASH MONTH’ and 'CRASH DAY OF WEEK’.

We decided to quantize the ‘CRASH TIME’ attribute to the hour value in the 24 hour format so that we can analyze the crash times easier by the hour.

For all the attributes about the number of people, we decided to group them each injured and killed numbers together because for our analysis we just need the number of people involved in each category.

* ‘NUMBER OF PERSONS INJURED’ and ‘NUMBER OF PERSONS KILLED’ is summed for each data record to ‘PERSONS INVOLVED’
* ‘NUMBER OF PEDESTRIANS INJURED’ and ‘NUMBER OF PEDESTRIANS KILLED’ is summed for each data record to ‘PEDESTRIANS INVOLVED’
* ‘NUMBER OF CYCLIST INJURED’ and ‘NUMBER OF CYCLIST KILLED’ is summed for each data record to ‘CYCLIST INVOLVED’

- ‘NUMBER OF MOTORIST INJURED’ and ‘NUMBER OF MOTORIST KILLED’ is summed for each data record to ‘MOTORIST INVOLVED'

In order to explore the accidents by the intersections they took place near, we created an attribute ‘INTERSECTION’ by appending the ‘ON STREET NAME’ and the ‘CROSS STREET NAME’ with a ‘X’ in the middle.

Many attributes are unused for our investigations and we decided to drop them,

* COLLISION\_ID
* 'ZIP CODE', 'OFF STREET NAME', ‘LOCATION’
* 'CONTRIBUTING FACTOR VEHICLE 2', 'CONTRIBUTING FACTOR VEHICLE 3', 'CONTRIBUTING FACTOR VEHICLE 4', 'CONTRIBUTING FACTOR VEHICLE 5’
* 'VEHICLE TYPE CODE 2', 'VEHICLE TYPE CODE 3', 'VEHICLE TYPE CODE 4', 'VEHICLE TYPE CODE 5’

And there are the attributes we have dropped since they have been transformed to other meaningful attributes,

* ‘CRASH DATE'
* 'NUMBER OF PERSONS INJURED', 'NUMBER OF PERSONS KILLED', 'NUMBER OF PEDESTRIANS INJURED', 'NUMBER OF PEDESTRIANS KILLED', 'NUMBER OF CYCLIST INJURED', 'NUMBER OF CYCLIST KILLED', 'NUMBER OF MOTORIST INJURED', 'NUMBER OF MOTORIST KILLED’
* 'ON STREET NAME', 'CROSS STREET NAME'

# *Data processing platform:*

We have loaded the data from the CSV file into a Pandas dataframe. We have modified the dataframe using various Pandas and other standard python libraries like Calender and Datetime. The code for the data processing and cleaning operations is in `crash\_investigation.py` in the read\_data.py function.

# *Tools:*

We have used Pandas library for handling and manipulating the data and matplotlib library to plot graphs.

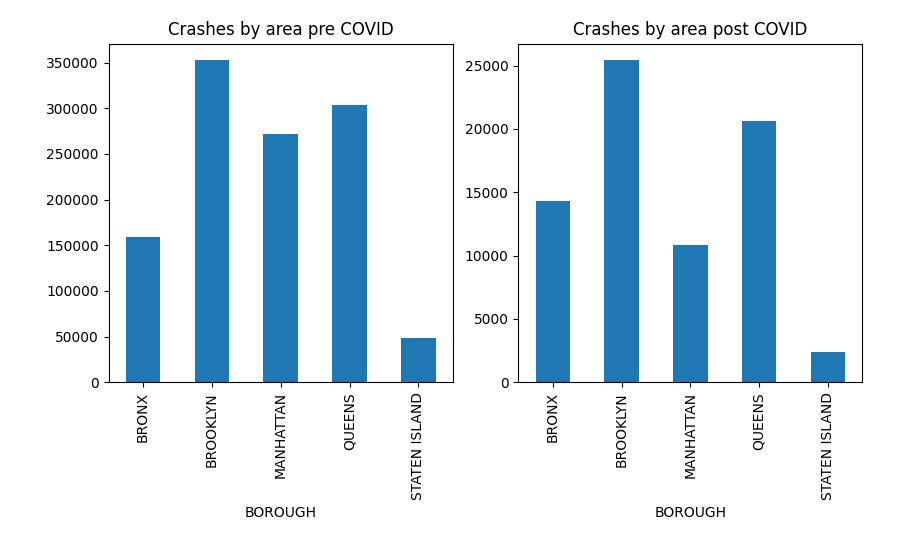
# *Investigations:*

After the preprocessing and cleaning, as our project focuses on investigating the impact of the pandemic on the accidents in New York City we have split the data into 2 dataframes based on the new attribute ‘CRASH YEAR’. Although the pandemic had started before 2020 we have assumed the start of year 2020 as the time the pandemic started impacting USA and the lockdowns were placed at the start of 2020.

* Pre-Covid data which includes years before 2020
* Post-Covid data which includes the year of 2020   
  (Since the dataset is being regularly updated, for the year 2021, the data records of the early months of January could add a bias for the early months as the records don’t exist for the rest of the months of 2021. Hence we are not including any records from the year 2021)

### *For the years before 2020, which boroughs had the most accidents? Did this distribution change during 2020?*

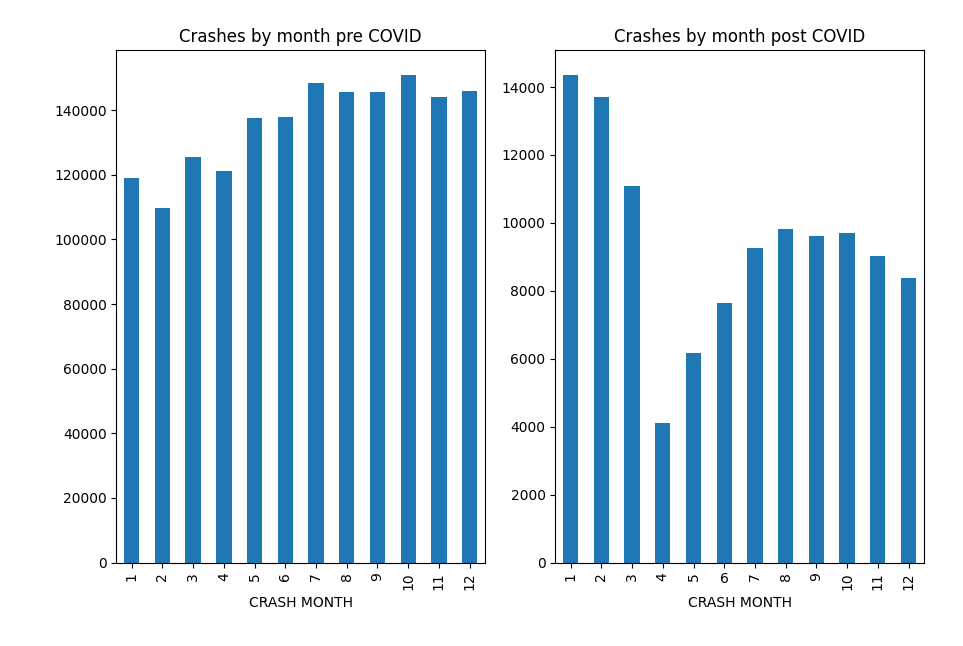
We create these graphs by grouping pre and post Covid data on the BOROUGH attribute.



We observe from the two graphs that there is a significant dip in the cases in Manhattan for the post Covid data. A reason for this could be the lockdown imposed in 2020 due to which people aren’t traveling to their offices and as Manhattan is primarily an office district of NYC there is huge dip in the traffic there which results in less crashes.

The borough of Brooklyn and Queens also see drop in accidents but the accidents in Bronx stay the same.

### *2. For the years before 2020, which months had the most accidents? Students in the past have said they thought there were 10% fewer accidents in February then in January. Is this true or is this bogus? Did this distribution change during 2020?*

We create these graphs by grouping pre and post Covid data on the new CRASH MONTH attribute that we created.

For pre Covid data, the months of July and October have the most number of crashes.

The trend of a February dip is indeed observed here as for the pre Covid data we see January has more cases that February. In fact February is the month with the least number of crashes. But this result could be because of a bias with February as there are fewer days in it than any other month. So to make a statement that February has a dip is mostly meaningless.

We observe from the two graphs that there is a huge dip in April for the post Covid data and the numbers slightly increase since April but stay low compared to the pre Covid data.

Observing this trend further supports the analysis that the dip in car crashes in the whole NYC region has occurred due to the Covid lockdowns which were put in place starting April 2020.

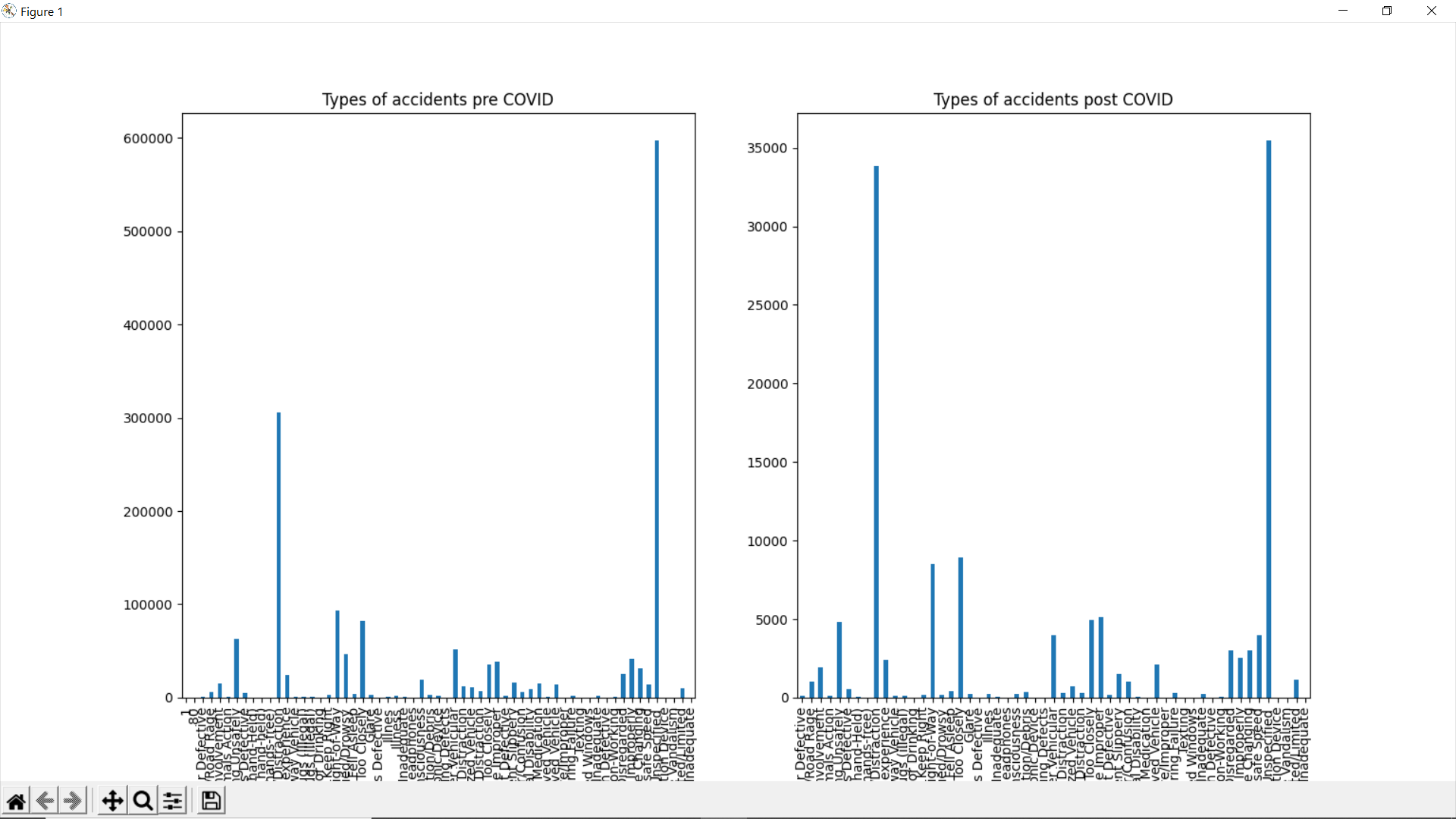
### *3. For the years before 2020, which types of accidents were most prevalent? Did this distribution change during 2020?*

Accidents can be categorized on many different parameters, we have chosen two ways,

* By the cause of the accident
* By the vehicle type involved in the accident

1. *By the cause of the accident*

We create these graphs by grouping pre and post Covid data on the CONTRIBUTING FACTOR VEHICLE 1 attribute.



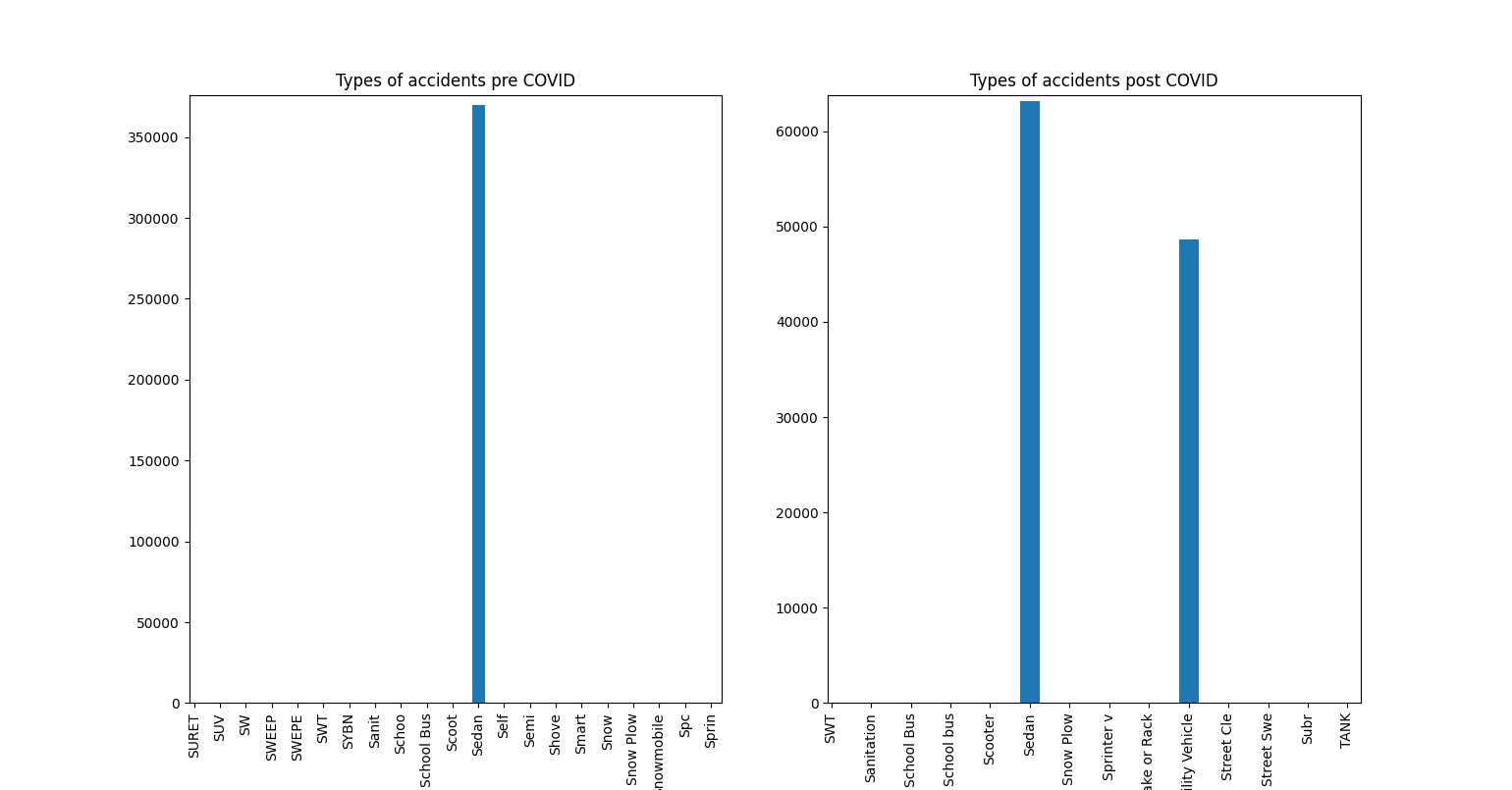
Driver In-attention and Distraction is the main cause of accidents before covid and after covid the trend remains the same. But after 2020, its proportion is more than any other types of accidents.

The reasons for this could be varied from drivers assuming there is less traffic on the road which gives them the liberty to text while driving or that people are distracted and overthinking due to the pandemic. They are worried because their friends, families have covid which makes them less attentive during driving.

Another explanation for this could just be that all the other causes of crashes have decreased in the pandemic due to less traffic on the roads and hence we see the proportion of Distraction causes have increased.

*B. By the vehicle type involved in the accident*

We create these graphs by grouping pre and post Covid data on the VEHICLE TYPE CODE 1 attribute.

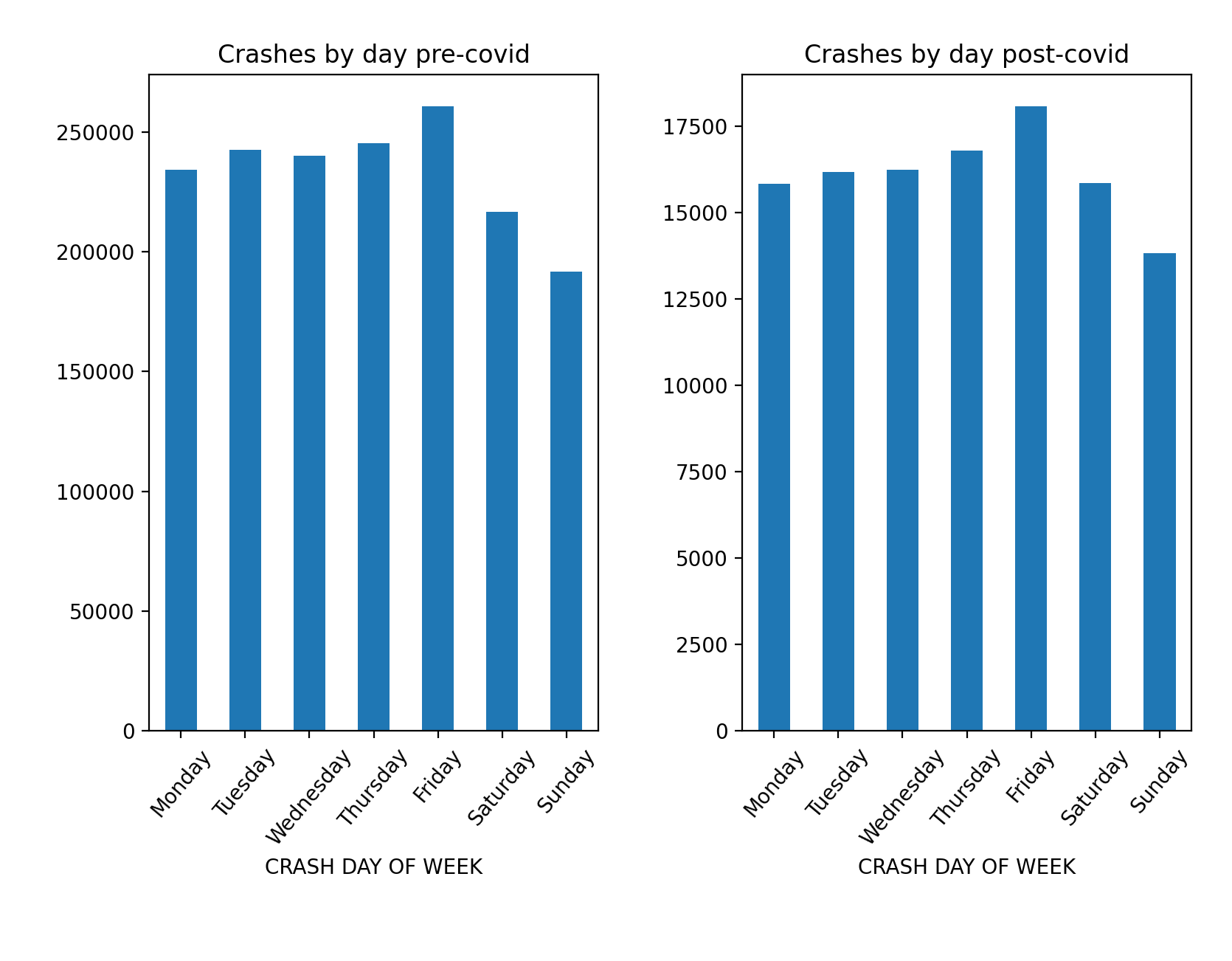


Sedan cars had more accidents pre Covid. It still has the highest number of accidents post Covid. However, the number of crashes in utility vehicles have increased very much. They are more than 4000 accidents in utility vehicle compared to 6000 in Sedan cars.

This could be simply because there are a lot of Sedan cars on the roads compared to any other cars. But we don’t have the data of the types of cars being driven and hence we can only guess the reason.

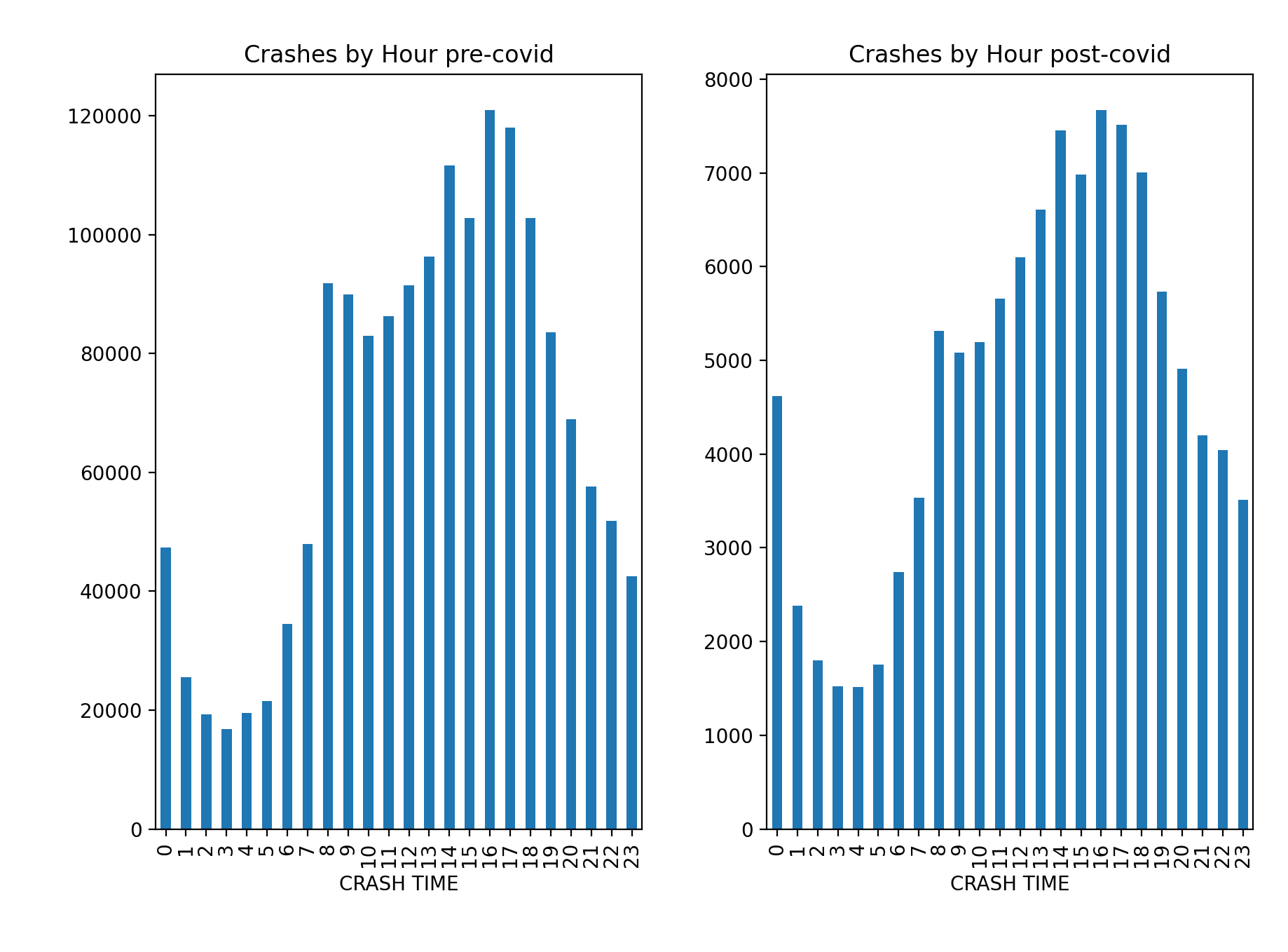
### *4. For the years before 2020, which days of the week had the most accidents? Did this distribution change during 2020?*

We create these graphs by grouping pre and post Covid data on the CRASH DAY OF WEEK

attribute that we created.

Friday has the most number of accidents before covid and after covid also. The relative change is constant overall for all days in 2020. That means that the accidents have reduced for all weekdays but the proportion is same. Sunday before also had the lowest number of accidents and it still has lowest number of accidents. On Fridays, there were around 2,50,000 crashes and after covid there are slightly more than 20,000 crashes.

### *5.For a typical year before 2020, how given a seven day calendar that starts at 12:01 AM on Saturday and goes until 11:59 PM midnight on Sunday, when are accidents most likely to happen? Which day of the week is most likely to have an accident? What time of the day is most likely to have an accident. Does this change in 2020?*

We create these graphs by grouping pre and post Covid data on the CRASH TIME attribute that we created.

Accidents are most likely to happen at 4:00 p.m before 2020. This does not change

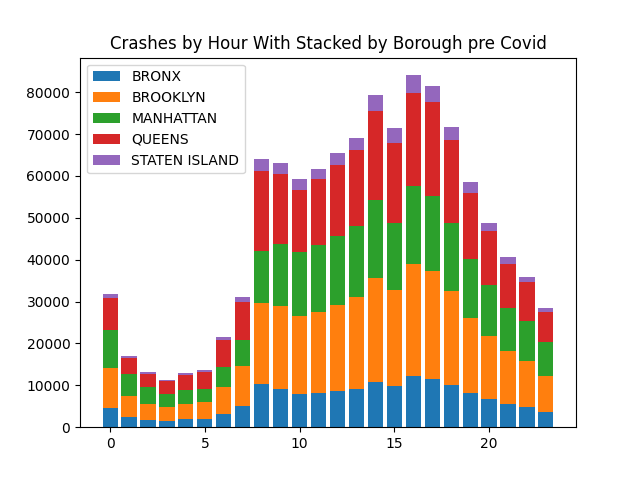
in 2020. In general most of the crashes are around 2:00 to 6:00 p.m. One of the reasons for this peak could be this could be people are feeling a slump in energies during the afternoon and are feeling sleepy after lunch. It could also be that’s the peak time and there are more number of vehicles on the road during this time period.

According to the National Highway Traffic Safety Administration (NHTSA) also, most

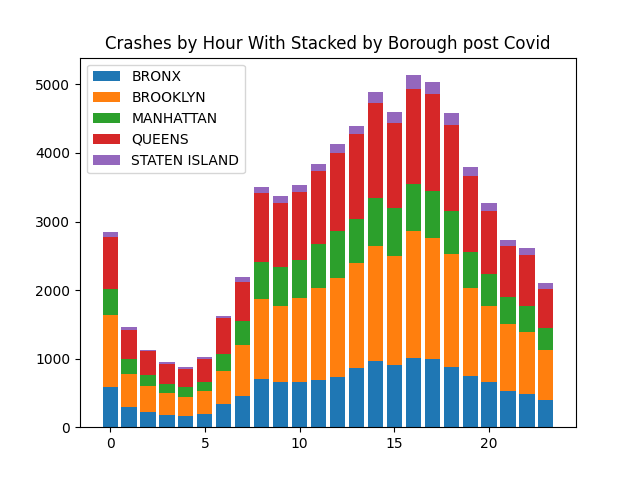
accidents occur between 3 pm and 6 pm. Fridays have a bit more accidents than other days. Then comes Thursday. Lowest number of accidents are seen on Sundays. The statistics does not change in 2020.

### *6. Does the timing of when accidents happen depend on the borough of NY City? Does the amount of change vary from year to year?*

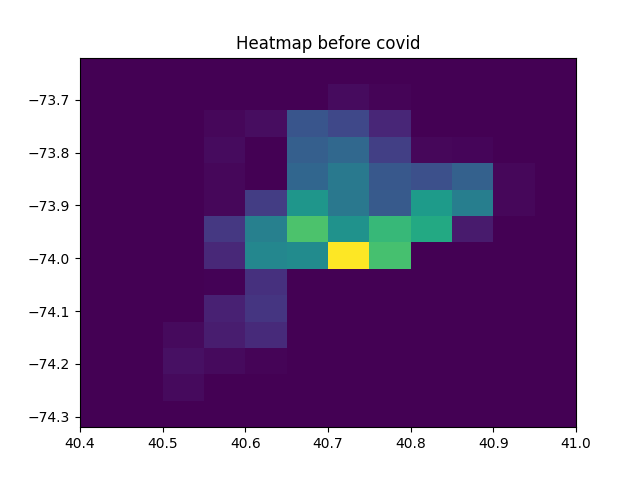
We create these graphs by filtering data by each of the 6 boroughs. We group each of these filtered data by the CRASH TIME attribute which is the hour of the day.

We have created a stacked bar graph which shows the amount of crashes by each hour in each borough. We have created 2 of these graphs for pre and post 2020 data. 

For all the boroughs the time of 4pm and 5pm has the highest number of crashes. We see Brooklyn has the most number of crashes followed by Queens. This could be due to the number of highways in these areas and different trucks and cars coming in and out of NYC from here at high speeds with sleepy drivers could be a factor for this.

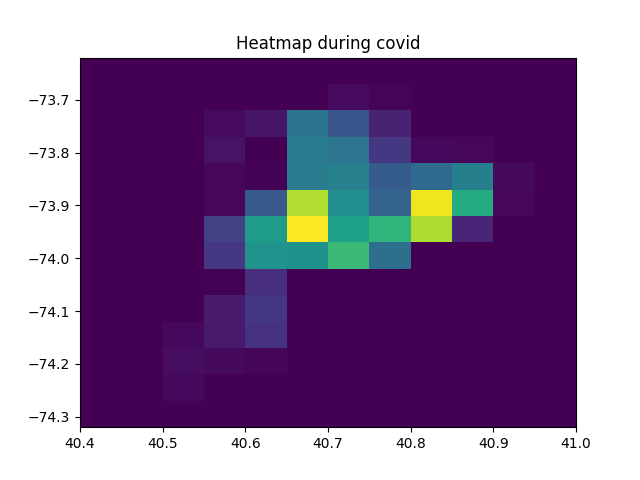
The total number of crashes has decreased overall. Although the decrease in crashes have been mostly in the frame of 8 to 20 hours. But the general trend remains the same for post Covid data. The time of 4 and 5 pm remains the peak across all boroughs.

### *7. For the years before 2020, given the entire region of NY City, which regions are “hot spots,” or places where accidents are most likely to occur? You will need to do some Parzen density estimation on this. You will need to use the GPS coordinates. You may need to convert the GPS coordinates from degrees:minutes:seconds, to degrees.fractions of degrees (fractional degrees). You do not need to use the Haversine distance. Assume that longitude and latitude is Euclidean for NY City. Create a “heat map” of where accidents occur and overlay it on a map of the city. Do the hot spot locations change in 2020?*

  
  
  
  
  
Before the pandemic, most accidents happened around the Manhattan and empire state building(i.e approximately latitude(40.71 – 40.75) and longitude(-73.96 - -73.41)). The second highest accidents are also around this region as you can see in the map(diagonal up and adjacent right side).



These shows the area of maximum accidents from the google map. They are being found by right clicking on the map manually, so there may be minor errors.



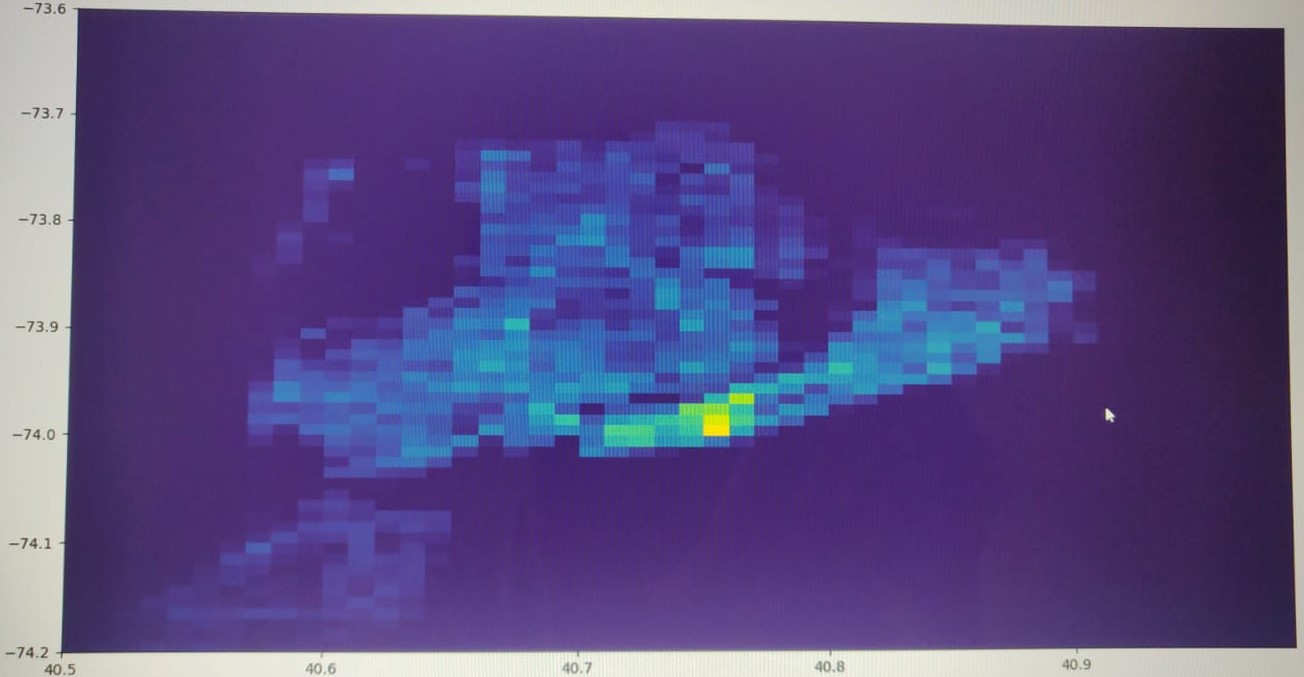
Similarly during 2020, the maximum number of accidents are at 2 places. One is around latitude(40.675) and longitude(-73.95) and other is latitude(40.83) and longitude(-73.9). These regions represent area in Brooklyn. These plots are created using 2d histogram(hist2D) in matplotlib. Such type of graphs can also be created using hexbin().



Areas like Bushwick, East New York on google maps is shown in the above figure:-



During the pandemic, there are more accidents in Crown Heights, Flatbush and nearby regions.



The above diagram is plotted with a different configuration. And it is the heatmap for maximum accidents in New York city with all the data combined(before 2020 and after 2020). It can be zoomed by matplotlib functionality to see for the exact hot spots in the map.

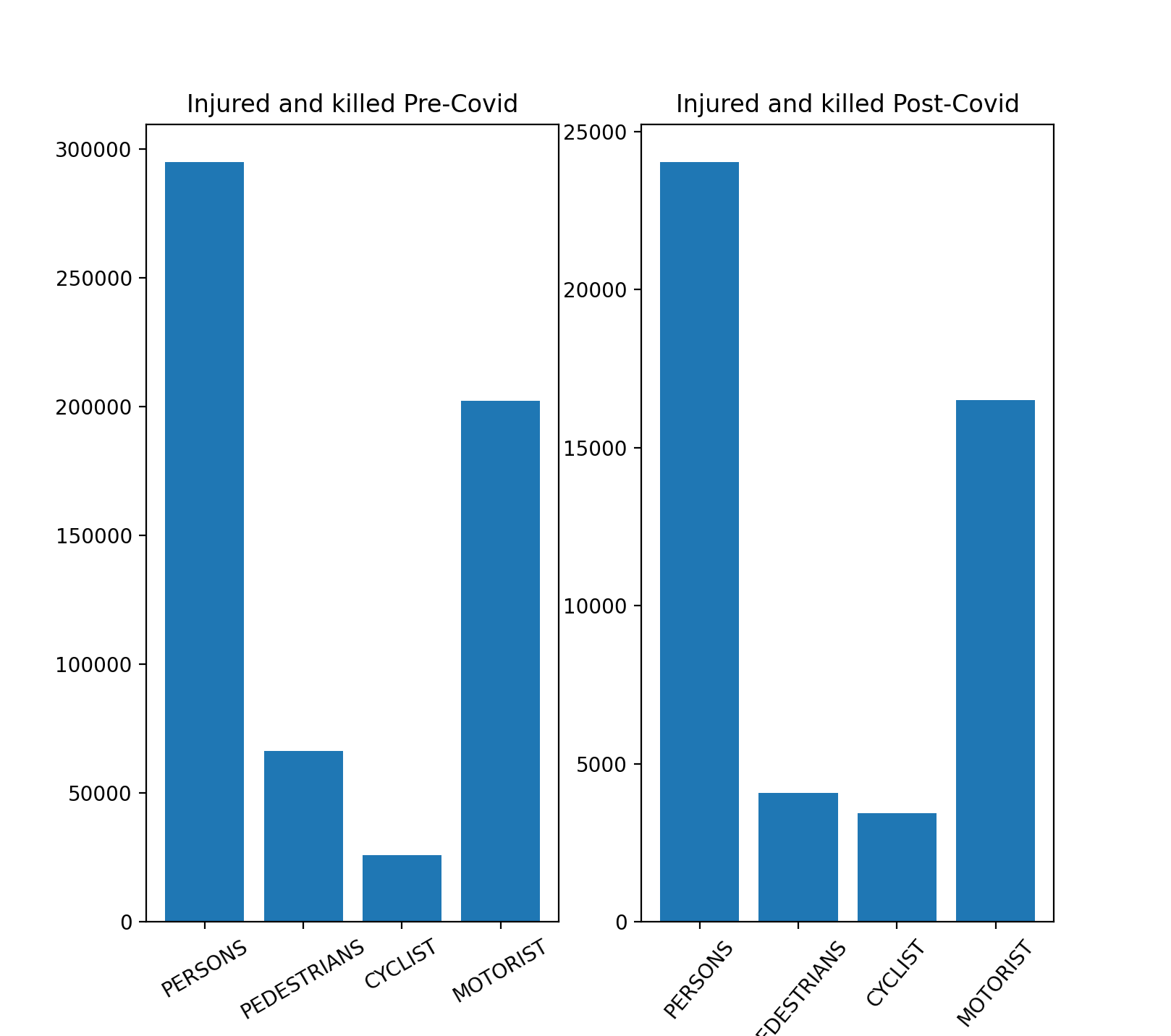
### *8. Compare the number of car-only accidents (car and car or car and obstacle) with car-pedestrian accidents (car and person or car and bicycle). Do these proportions change in 2020? Do they cha**nge in any particular location?*

We have created these graphs by plotting the counts of accidents which had any persons involved in the accidents or not. If there weren’t any persons involved we have assumed it was only a car to car crash.

We observe that the Car only crash counts leads the way. Although post Covid the number of Car-Pedestrian accidents have appeared to remain the same as pre 2020 but the Car only accidents have decreased post Covid.

### *9. Did you discover anything else you wanted to explore?*

One aspect of this data we wanted to explore was the type of people injured or killed by traffic accidents overall.

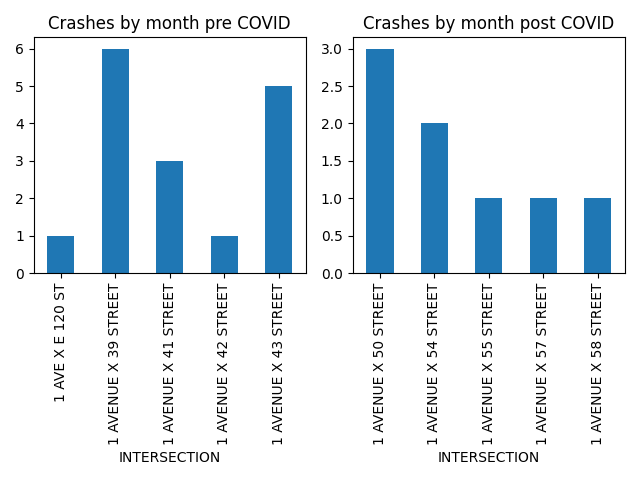
The number of cyclists injured and killed have relatively increased during the 2020. The pandemic saw a steep rise in the number of cyclists as many people took up to cycling while other activities were shut down. This could be a reason for the increase in the number of cyclist related accidents post 2020.

The number of pedestrians injured and killed have decreased. This could be because people are not going out of the house much. The number of persons and motorist injured and killed got reduced by same percentage (i.e. 88%)

Another aspect we wanted to explore in the data was the intersection which are most dangerous. As by intuition we thought most accidents would be occurring at intersections.

We created a new attribute ‘Intersection’ by appending the values of ON STREET NAME and CROSS STREET NAME. These values had to be cleaned thoroughly as many values were empty.

What we observed was very peculiar that the top 5 dangerous intersections were all on the 1st Avenue. The traffic authorities could change some rules like speed limits, etc to reduce the traffic on the 1st Avenue.



The Street got moved up higher away from the busy office areas in the post 2020 data.

### 10.

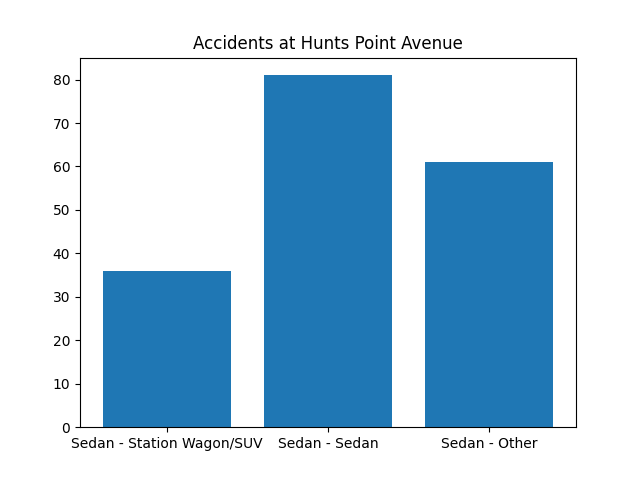
If there is an accident at Hope Street, even before emergency vehicles are dispatched we

can say that it would be a car-station wagon/sport utility vehicle or a car – other accident.

The number of car-car accidents are quite low. It may be the case that there are more

number of COVID cases in Hope Street as there are more number of accidents of vehicles

carrying patients.

In Hunts Point Avenue, the number of car-car accidents are highest than any other type of

accident. That may mean that this region has less number of covid cases, as there are less

number of big vehicles(which carry dead bodies) accidents and so there are less big vehicles

on road.

Filtering the records as per the term of the area name was a challenge as we don’t have exact attribute to map towards.   
  
A classifier could be built using the street names and finding the query word in them.

If we are given a particular location, then we can gain some useful insights from the data like how many people are killed or injured at that place, what are the contributing factors for accidents at the location, what type of vehicles have more accidents, whether there is an intersection near that place which is prone to accidents and so on.

We can build a classifier for this by using sklearn. We’ll first preprocess the data, build and evaluate the model(like Random Forest, Support Vector Machine etc..) using the library. We’ll calculate accuracy by splitting into training and test data set. We can do n fold-cross validation. We’ll take as input the locations(i.e latitude and longitude) and use different features from the data like vehicle type, people killed/injured, contributing factor vehcicle, time and so on. We can create output labels. Thereafter we can do classification and check the results.

## *Discussion:*

The scale of the data was challenging to deal with considering the amount of missing and unclean data amongst these records. It causes a major hinderance in arriving at accurate conclusions. This is a lesson we learnt about real world data that the major task is about cleaning the data and a lot of preliminary issues stem from issues with data not pre processed correctly.

We were facing issues about creating the heat map of the crashes across the city and mapping the results on the city map.

The part about making a classifier to predict the type of accident based on the area is challenging as the area name may not match with the records that have street names. So it was difficult to first find records that ares supposed to match with the query we enter. Also, choosing the appropriate classifier model and modeling the data is difficult. Testing this classifier is also a challenge.

## *Future Work:* We will be exporting the dataframe to a MongoDB database to avoid loading times and possibly have advanced filtering on the data. We plan to merge the database with the weather data of the same time as the accident to find the relation between the accidents and the weather. The city authorities’s main focus would be to decrease casualties from accidents so we would further like to explore the areas where the persons are getting killed or injured frequently could also be identified and this information could be used to warn the people of it being an accident prone zone. For a more thorough analysis we could explore the NYPD Motor Vehicle Collisions page which has more detailed statistics of Motor Vehicle Collisions. This data has the number of collisions, people involved, the contributing factor and vehicle types involved divided by precinct. Also there exists borough wise data about each intersection and the type of collision and people involved are. This highly refined and detailed reports could help in making accurate investigations and draw deeper insights in the data.

## *Conclusion:*

We performed data visualization for NY City car crash data. We analyzed which boroughs had more accidents, which month and time had more accidents, compared the accidents of car with car and other vehicles, found relationship between different causes of casualties and were injured in accidents, what type of accidents usually occur at a particular location, compared crashes by hour of the accident and day of the week.

An important lesson we learnt about dealing with real world data is that the major task is about cleaning the data and a lot of preliminary issues stem from issues with data not pre processed correctly.

We learned how to pre-process the data. In that we converted the date and time to more usable format using split() and weekday().  Then we combined some columns, deleted useless columns. We divided the whole data in 2 dataframes(preCovid and postCovid). We learned how to use groupby and plot bar graph using subplot. We learned how to visualize data, plot graph and gain useful insights from the data. We learned a lot about matplotlib library also while implementing the project.

We learnt that 2 results can’t simply be compared, each value and the context have to be taken into consideration along with the trend the visualization presents.

We did all these investigations separately for data before covid and after covid. The comparison of pre and post Covid data for each question provides greater insights into the data and also the wide encompassing impact the pandemic has had on all our lives.