Chinmay Bhate (cb4490)
Maitreya Kocharekar (mk1651)

Data Mining Project

Team Members:

1. Chinmay Bhate - cb4490
2. Maitreya Kocharekar - mk1651

Data Preparation:

a. **How clean is the data?**

The dataset appears to be relatively clean overall. There are no duplicate entries, which suggests that each row represents a unique record. Additionally, the absence of null values in columns related to crash details, such as the time of the crash and the number of people injured, indicates that these critical pieces of information are consistently recorded.

However, there are null values in columns related to the location of the crash and the vehicles involved. So for performing analysis on locations of crashes, we will need to drop rows having null columns.

Following screenshot showcases the count of duplicates in dataset

```
Number of Duplicates:  0
```

Following screenshot displays the number of null values in each column of the dataset

Chinmay Bhate (cb4490)
Maitreya Kocharekar (mk1651)

```
Null values in dataset by each column:
CRASH DATE                            0
CRASH TIME                            0
BOROUGH                          635526
ZIP CODE                         635770
LATITUDE                         231535
LONGITUDE                        231535
LOCATION                         231535
ON STREET NAME                   431621
CROSS STREET NAME                767979
OFF STREET NAME                 1703656
NUMBER OF PERSONS INJURED            18
NUMBER OF PERSONS KILLED             31
NUMBER OF PEDESTRIANS INJURED         0
NUMBER OF PEDESTRIANS KILLED          0
NUMBER OF CYCLIST INJURED             0
NUMBER OF CYCLIST KILLED              0
NUMBER OF MOTORIST INJURED            0
NUMBER OF MOTORIST KILLED             0
CONTRIBUTING FACTOR VEHICLE 1      6526
CONTRIBUTING FACTOR VEHICLE 2    313561
CONTRIBUTING FACTOR VEHICLE 3   1897568
CONTRIBUTING FACTOR VEHICLE 4   2010232
CONTRIBUTING FACTOR VEHICLE 5   2034057
COLLISION_ID                          0
VEHICLE TYPE CODE 1               13087
VEHICLE TYPE CODE 2              385052
VEHICLE TYPE CODE 3             1902725
VEHICLE TYPE CODE 4             2011339
VEHICLE TYPE CODE 5             2034325
dtype: int64
```

b. Which data did you ignore?

The analysis was specifically targeted at the data related to Queens Borough for the years 2019 and 2020. We ignored all the records that fall outside this scope to maintain the focus on selected subset of data

There were certain columns that were excluded from consideration. For instance, the "Location" column, which contains a tuple of latitude and longitude, was overlooked since these geographic coordinates were already available as distinct columns in the dataset. Additionally, the "Collision ID" column was disregarded as it is not a measured attribute.

c. What data did you focus on?

The analysis was focused on the data for Queens Borough for years 2019 and 2020. Moreover, we focused on the data related to the time of crash to

understand when were the crashes occurring more oftenly. And the geolocation data to understand the patterns in which areas were more prone to crashes.

d. Did you quantize the data into regions?

Since we were working on data specific for Queens Borough. We filtered the data for the specified borough and we did not quantize the data further into specific regions.

e. Are there any issues with the data?

There are some null entries in the dataset in the location details. Therefore, for performing geolocation analysis required cleaning of the dataset for removing null entries.

There is noise in the data. While working on generating heat maps it was discovered that there were some records with noisy latitude and longitude. For example, there are records in Queens borough with latitude and longitude as 0, 0.

f. Is the data from the two years comparable?

Yes, the data from 2 years namely 2019 and 2020 is comparable. Data analysis for the given 2 years indicates that the data for both years share similar attributes and features which can be used to find trends, patterns and gain insights from it.

Chinmay Bhate (cb4490)
Maitreya Kocharekar (mk1651)

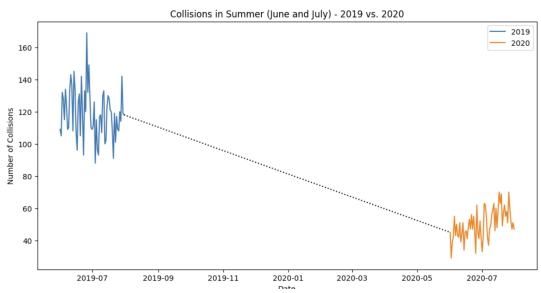
g. Are there any other issues you found?

Apart from the issues reported above, there were no issues identified during the analysis.

Answers to the asked questions:

1. **For the two years given, figure out what has changed in the summer from one year to the next. Figure out how to visualize the difference, in some way.**

   For the given two years 2019 and 2020, there is a significant drop in the number of crashes was observed. For addressing this query, dataset was filtered to get data for a given time period for Queens borough. Furthermore, the crash records were grouped on a daily basis to observe the pattern of crash occurrence throughout the summer periods of both years. Following is the line graph visualizing the number of crashes that occur in summer 2019 and 2020. And scatter plot shows accidents in 2019 and 2020. Blue point signifies accident in year 2020 and red point signifies accident in 2019



   A plausible inference from this observation is that the impact of the COVID-19 pandemic in 2020 led to a reduction in the overall vehicular traffic on the roads.

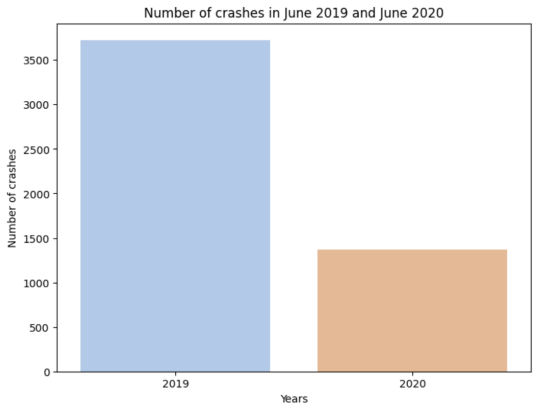
This reduction in traffic volume is likely associated with a corresponding decrease in the number of reported crashes during the specified period. We have also attached a map based representation of the accidents data and applied DB Scan on it to cluster data in each summer based on the location or rather the latitude and longitude of the recorded accident. The two markers, i.e the red and blue represent the where the mean or where do accidents happen on average, so we can infer here that they are predominant in the central part of the borough and later in the 2020 they shifted to the southern part of the borough. This could be caused because in 2020, 73rd street to 80th street in Jackson Heights area in Queens were closed because of the outbreak of the Coronavirus (as per our research), hence the epicenter shifted further down near the airport area, where the Queens Hospital Centre lies.

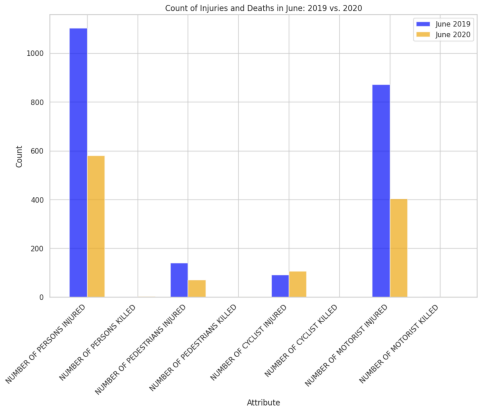
2. **How was June of 2019 different then June of 2020? Figure out how to show or demonstrate the difference.**

   There was a noticeable decrease in the number of crashes from June 2019 to June 2020. Following is the bar plot showcasing the number of crashes in June 2019 and June 2020. Looking at the visualizations we can clearly see that the number of crashes in 2020 have reduced more than 50%.

CSCI 720 - Big Data Analysis

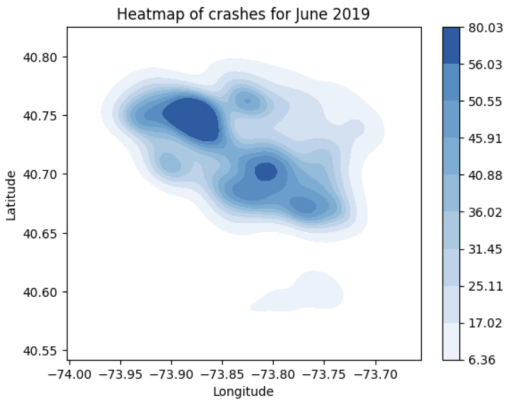
Chinmay Bhate (cb4490)
Maitreya Kocharekar (mk1651)



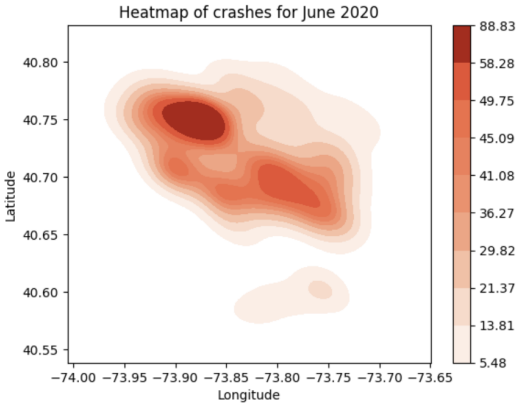


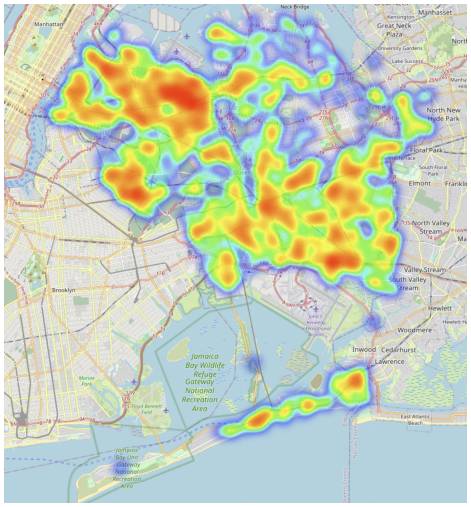
The graph on the left compares the total number of crashes in June 2019 and June 2020. The graph on the right compares the number of people injured and killed in June 2019 and June 2020.

June 2019:



June 2020:





Furthermore, we applied kernel density estimation plots from the seaborn library to generate the heat maps to perform geolocation analysis. The KDE plots utilizes a convolution algorithm with a kernel function to analyze the density of crash occurrences across the area. Additional processing required for this plot was to clean the geolocation data. Hence we ignored the records having null entries in the latitude and longitude columns.

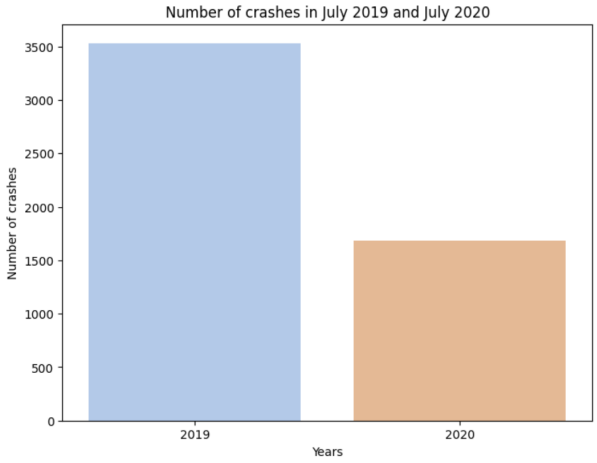
We can observe that while the accident prone area which is the central region of Queens borough remained relatively constant, areas surrounding Liberty Avenue, home to York College, experienced a notable reduction. This can be linked to the COVID-19 pandemic, which led to the closure of colleges, resulting in a decreased presence of drivers, particularly young adults, and consequently contributing to the observed decline in crash incidents.

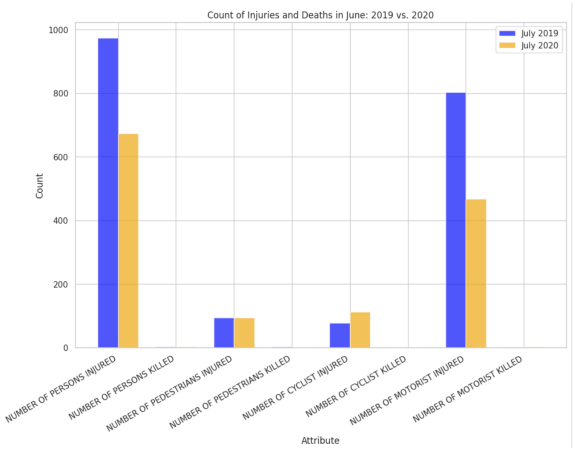
3. **How was July of 2019 different then July of 2020? Figure out how to show or demonstrate the difference.**

Building on the findings from the previous comparison between June 2019 and 2020, a consistent decline in the number of accidents during 2020 was observed. The primary focus of accidents remains concentrated in the accident-prone central area of Queens. However, notable changes are apparent across other regions, marked by an overall reduction in accidents in 2020. Particularly intriguing is the reduced occurrence of accidents in an area around latitude 40.68 and longitude -73.77 in 2020, which corresponds to locations near supermarkets, home depot stores, and other commercial areas. Additionally, zones adjacent to

college areas exhibit fewer accidents in 2020 compared to the corresponding period in 2019. These shifts in accident patterns suggest an impact, potentially influenced by changes in traffic flow and human activity, such as those brought about by external factors like the COVID-19 pandemic

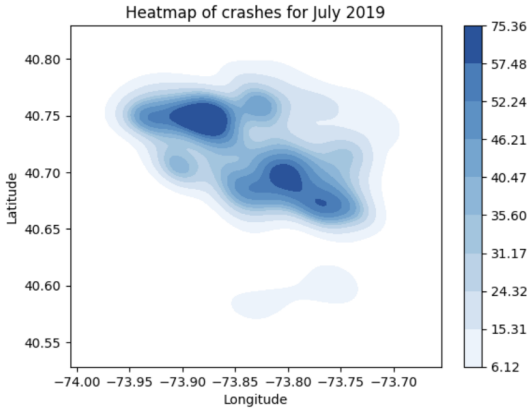
.

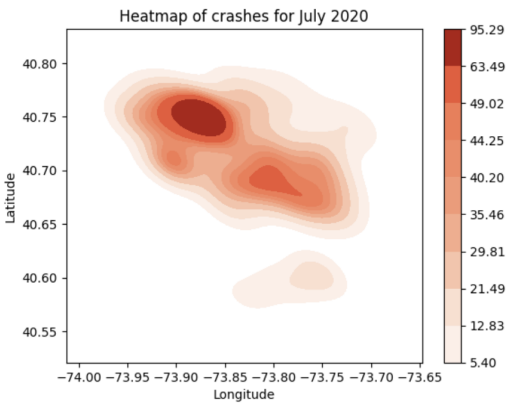




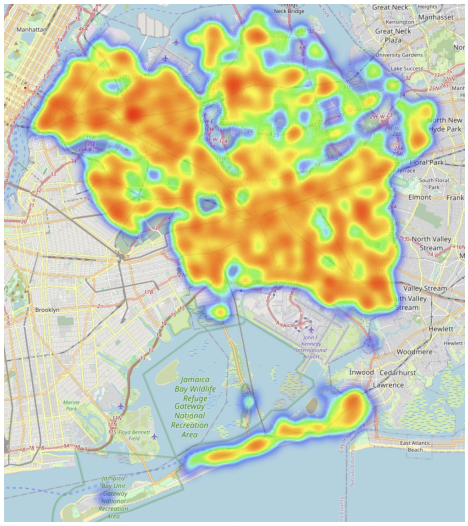
The graph on the left compares the total number of crashes in July 2019 and July 2020. The graph on the right compares the number of people injured and killed in July 2019 and July 2020.

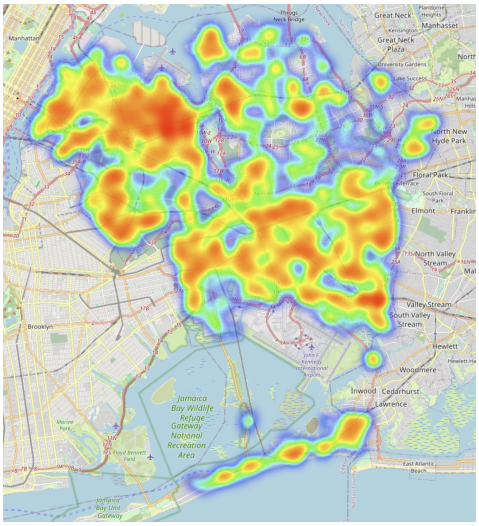
Following graph compares the density map for crash regions in given time frame:





Chinmay Bhate (cb4490)
Maitreya Kocharekar (mk1651)





4. **For the year of January 2019 to October of 2020, which 100 consecutive days had the most accidents?**

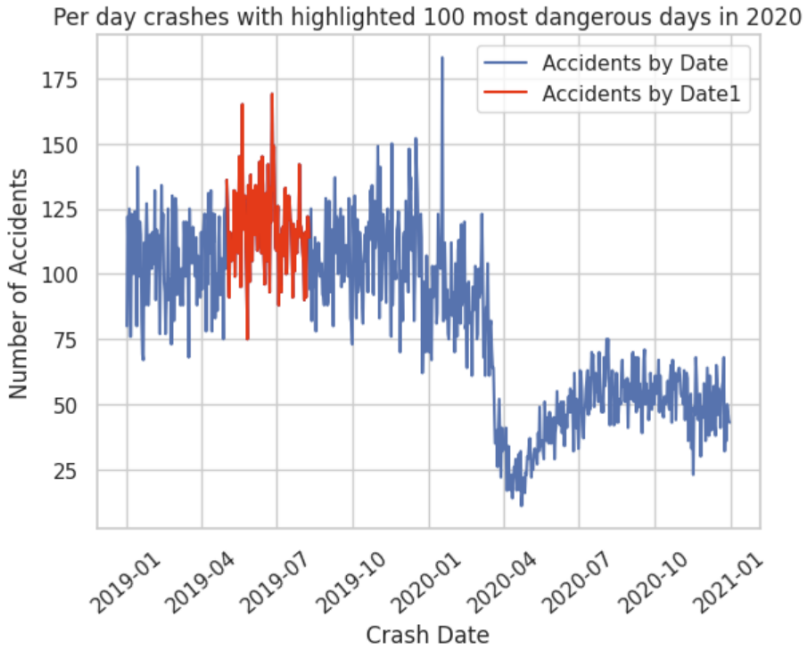
   Upon investigating the data for the 100 consecutive most dangerous days, a significant surge in the number of crashes was identified during the summer of 2019, spanning from May to August, totaling 11,739 incidents. Following screenshot displays the exact dates of the 100 most dangerous days along with the count of accidents recorded during that period.

```
Start Date :  2019-05-02 00:00:00
End Date:  2019-08-10 00:00:00
Max Crashes:  11739
```

   One reasonable explanation for this observed increase is attributed to the summer vacation period. As individuals visit and explore New York City during this time, increased number of vehicles on the road, combined with factors like increased tourism and recreational travel, may contribute to a greater likelihood

of accidents. Another reason could be as the students are on summer break, there might be more students engaging in the recreational activities.

Following is the line graph highlighting 100 most dangerous days in year 2020.



5. **Which day of the week has the most accidents?**

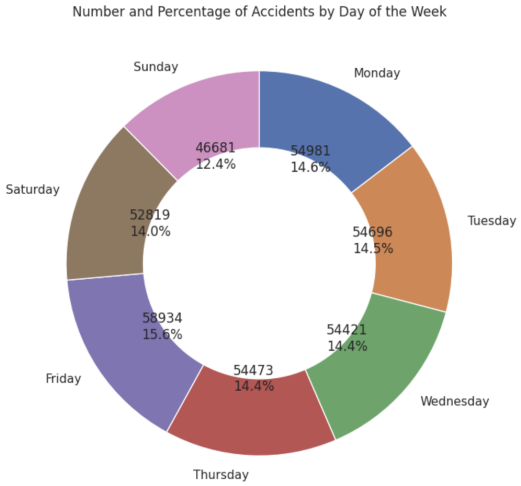
Following is the screenshot of the output of the program highlighting the day of the week with the most accidents. Maximum accidents were recorded on **Friday**. With the record of 58935 accidents.

```
The day of the week with the most accidents is: Friday
Number of accidents on Friday: 58935
```

We can infer that since Friday is the start of the weekend, people might go for

parties or for picnics which may result in more vehicles on the road and eventually more accidents.

To get more proper results we have performed analysis on the complete data available for Queens borough for this question. Following is the pie chart for the number of accidents that occur on each day of the week.



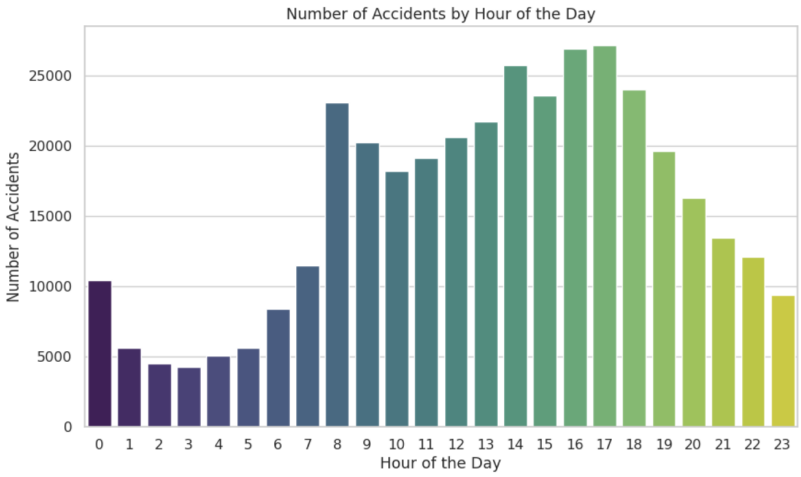
6. **Which hour of the day has the most accidents?**

The hour of the day with the most number of accidents is 5pm-6pm with 27166 crashes recorded in the time frame.

To get better results, we have executed queries on the complete data available for Queens borough to understand the peak hours having most accidents throughout. Following is the screenshot of the output of the query for finding the hour having the most accidents.

```
The hour of the day with the most accidents is: 17 o'clock
Number of accidents at 17 o'clock: 27166
```

Chinmay Bhate (cb4490)
Maitreya Kocharekar (mk1651)



The above plot showcases the hourly accident rates. Even if the most number of accidents occur between 5pm - 6pm, the number of crashes that occur between 4pm - 5pm is also high. This might be the reason that most of the employees return home during this time frame.

7. **In the year 2020, which 12 days had the most accidents? Can you speculate about why this is?**

   Following is a screenshot showcasing the top 12 days of the year 2020 having most accidents recorded along with the number of crashes recorded on that particular day.

```
Top 12 days with the most accidents in 2020:
CRASH DATE
2020-01-18    183
2020-01-13    123
2020-03-06    123
2020-02-14    120
2020-02-10    119
2020-02-07    113
2020-01-21    112
2020-01-29    112
2020-01-14    110
2020-02-06    109
2020-01-15    105
2020-02-03    104
dtype: int64
```

Chinmay Bhate (cb4490)
Maitreya Kocharekar (mk1651)

The analysis of the 12 days in 2020 with the highest accident counts reveals interesting patterns related to holidays, special occasions, and day-of-the-week dynamics. Notably, long weekends and holidays, such as Martin Luther King Day on January 18th and Valentine's Day on February 14th, correspond with elevated accident rates, likely due to increased travel and leisure activities. Mondays consistently emerge as days with higher accident counts, emphasizing the need for increased safety measures during the beginning of the workweek.

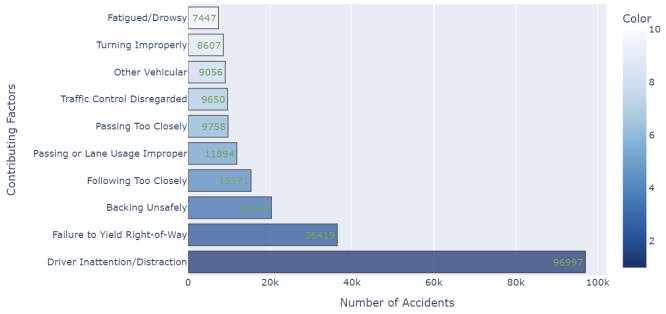
Additionally, Fridays before significant events, like the start of Valentine's week on February 7th and the Friday before daylight savings on March 6th, exhibit increased accident frequencies, suggesting potential anticipatory behaviors or altered traffic patterns.

Furthermore, a notable observation is the concentration of the top 12 days with the most accidents in the first two months of 2020. This pattern coincides with the emergence of the COVID-19 pandemic. The confirmation of the first case on March 1st, announced on March 3rd, marked the beginning of a series of containment measures initiated on March 10th. These measures, including the creation of containment areas, contributed to a decrease in vehicles on the road, resulting in a simultaneous reduction in the number of crashes

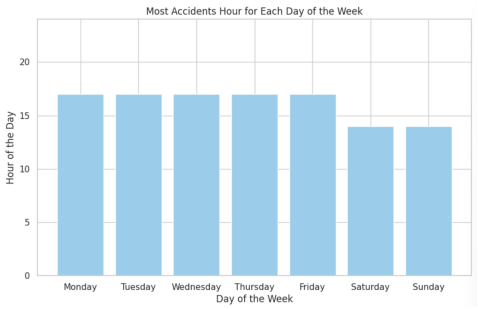
Following are some of the interesting facts about the data:

Chinmay Bhate (cb4490)
Maitreya Kocharekar (mk1651)

1. Driver Inattention was the main factor for most of the crashes. Following are the top 10 main reasons contributing to crashes.



2. Mostly the accidents happen between 5pm - 6pm for Weekdays and around 2pm - 3pm for Weekends



Conclusions:

In concluding this data analysis project focused on the examination of NY crash data. The analysis was further narrowed down to Queens Borough for the years 2019 and 2020, several valuable insights have emerged. The exploration of temporal patterns, geographical hotspots, and external influences such as holidays and the COVID-19 pandemic has provided a comprehensive understanding of factors influencing traffic accidents. Notably, the visualizations revealed a decline in accidents during 2020, especially in areas associated with containment measures and changes in daily routines due to the pandemic.

Chinmay Bhate (cb4490)
Maitreya Kocharekar (mk1651)

Challenges in the project included data preprocessing complexities, particularly in handling missing values and noisy data. Initially for geolocation analysis, due to noisy data the algorithms for generating heat maps were failing but once the data was further cleaned, geolocation analysis of data to gain insights on the accident prone regions was possible

What made this project particularly interesting was the relationships between temporal, geographical, and societal factors and their impact on accident occurrences was vividly visible in the graphs. The ability to discern patterns related to holidays, long weekends, and the unprecedented effects of the pandemic were discovered during the analysis. We were able to learn about geolocation visualizations which made this project interesting.

Various algorithms were employed throughout the project, kernel density estimation for visualizing hotspots. Implementing this algorithm on geolocation data such as latitude and longitude helped to reveal the crash prone regions during the years 2019 and 2020. In terms of data mining, this project reinforced the significance of context in interpreting results. Understanding the nuances of the data, considering external influences were key takeaways. Overall, this project not only improvised our technical skills but also deepened our appreciation for the nature of data exploration and its real-world implications.