CSCI 720. Big Data Analytics Project

Aakash Anil Khatu

Rutvi Jiten Bheda

# Analyzing Patterns for Brooklyn Traffic Accidents

## 3. Data Preparation

a. The data was not completely clean. There were many fields that had unspecified values so we replaced them with nan for better data analysis. Functions in python libraries like numpy and pandas can automatically handle nana by ignoring them from the calculations. Hence they would not include those fields and no manual filters will be needed in that scenario. Also, using nan would maintain consistency throughout the data. We also converted the date format to day week and year for better analysis.

b. Since we were supposed to take only one borough for our mining and analysis, hence we selected Brooklyn and ignored data of all the other boroughs.

c. The main focus was on the Brooklyn borough and also between the years 2019 to 2022. The main columns that needed focus were crash date, time, contributing factors, and the vehicle types.

d. Since, We are only using the data from one borough, we chose not to split the data further and quantize it into regions. Instead, we chose to use k means to create clusters from the data to find out patterns.

e. As such there are not many issues. One of the issues we encountered was the unspecified missing data making it difficult to navigate. Many of the boroughs too had missing values. However, overall the features of the data give us all necessary information needed.

f. Yes the data from two years is comparable and it can be seen from our findings and visualization graphs below.

g. Except for the ones mentioned above, there were no more issues found with the data. All the necessary columns were provided that aided us in answering the questions below.

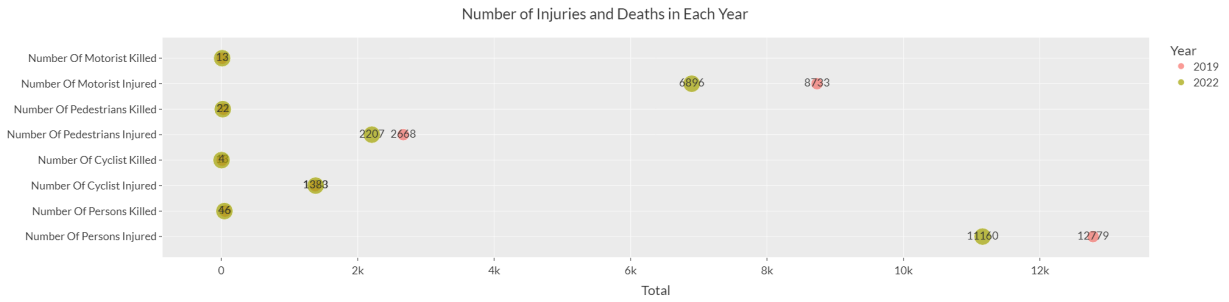
h. The main features important to use after selecting the borough was the crash date, crash time, the location, the five contributing factor vehicles and the vehicle type codes that tell us the type of vehicle it was.

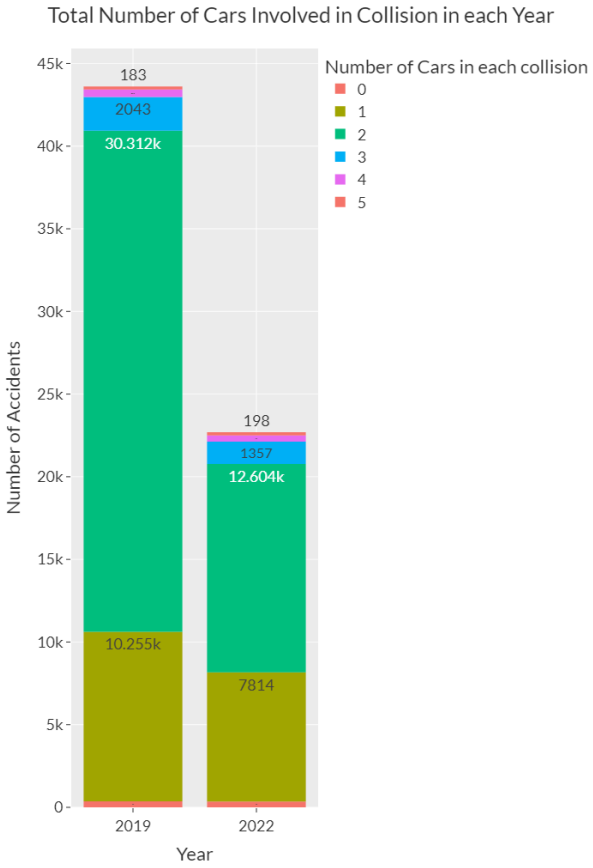
## 4. Answers to the mentioned questions

**1.What ethical considerations are there? Suppose you find a neighborhood that has many accidents, and you publish this. Could you be sued? Is it just data?**

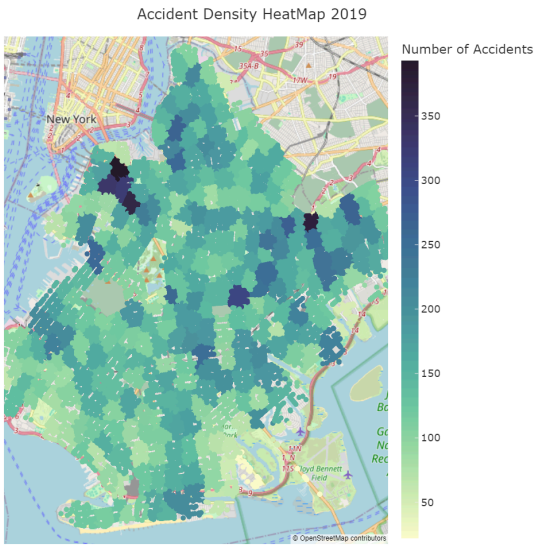
Some of the ethical considerations to keep in mind would be firstly privacy. It is important to ensure that no individuals are identified from the data that is published. Another consideration would be the impact on the communities. If we highlight a particular neighborhood by saying and giving the information that it has a high rate of accidents then it could have negative consequences on the community. It might not only affect the residents staying there but also the property values and that neighborhood overall in different aspects. Also it is important to make sure that the data is accurate and gives the true information without misleading in any way. To conclude, it is important to anonymize the data by removing any personal identifiers if any. A context and background should also be given to ensure accuracy of the data.

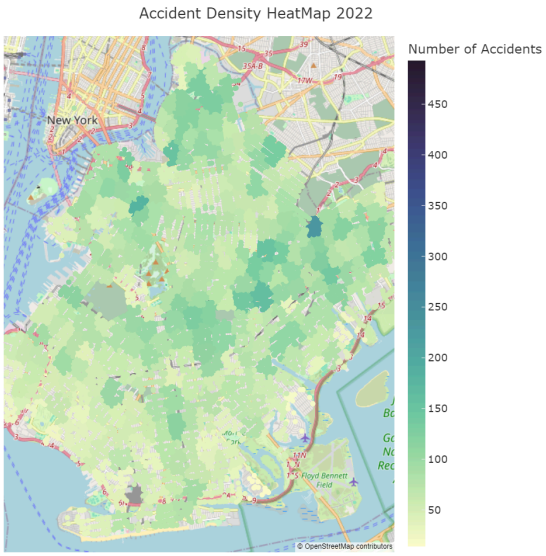
**2. Pick two regions of time, say two years. Figure out what has changed from one year to the next. Figure out how to visualize the difference, in some way.**



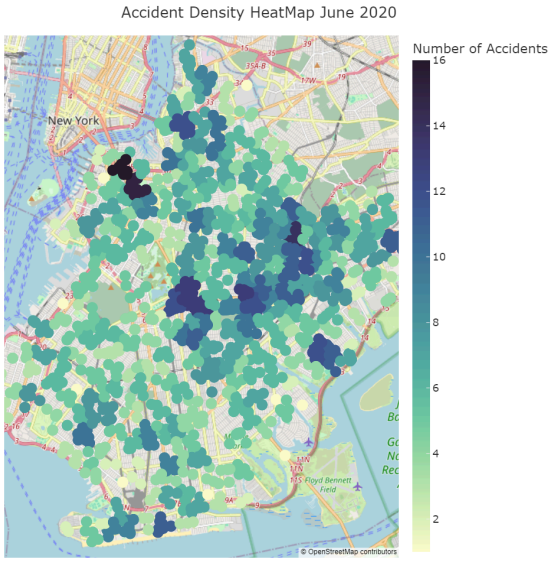



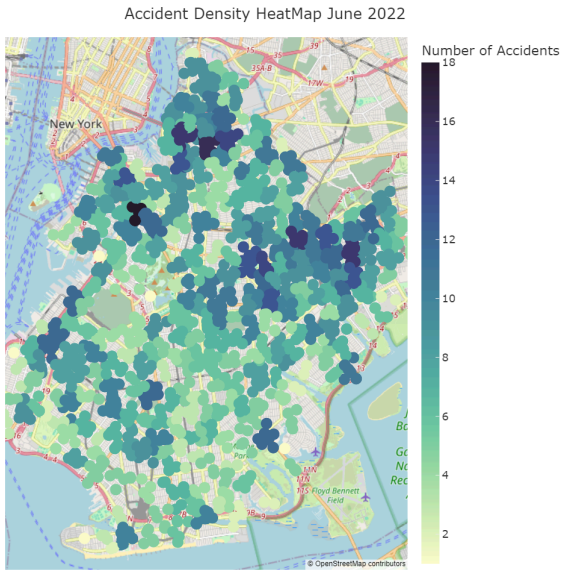

The two years picked up are 2019 and 2022. As seen from the graphs above, the number of accidents in 2019 were more as compared to that of in 2022. The colors represent the number of cars that were involved in the collision. The majority of the accidents had taken place in both the years by two cars. Below given are the heat maps showing the accidents. The darkest color shows the maximum number of accidents in that area.

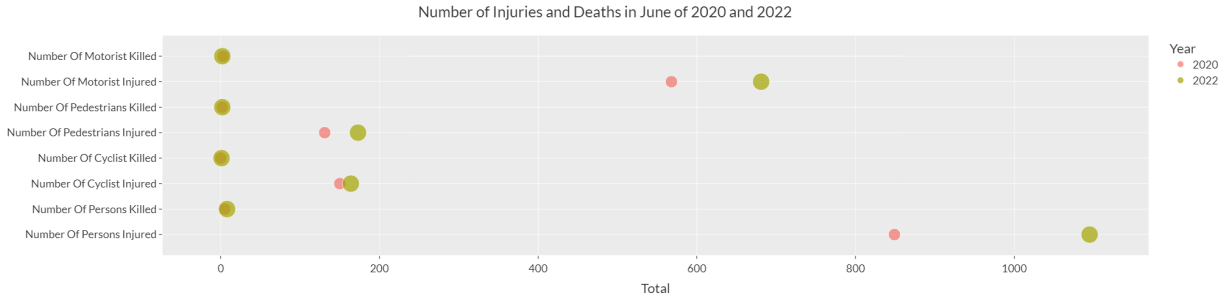


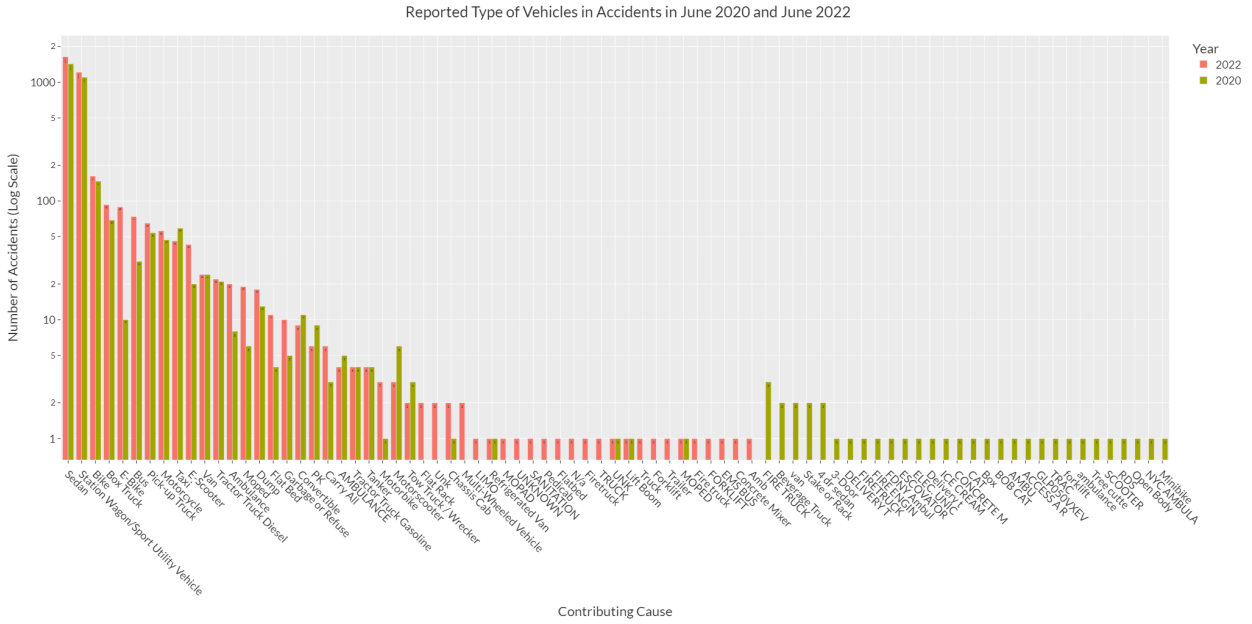




**3. How was June of 2020 different then June of 2022? Figure out how to show or demonstrate the difference. Were there more pedestrian accidents? Were there more accidents involving delivery vehicles?**



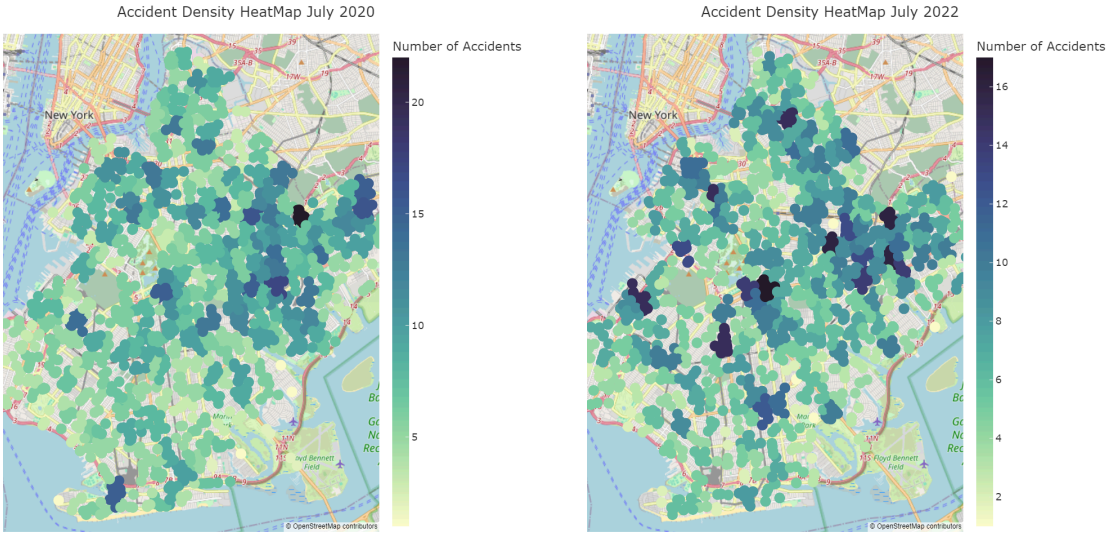



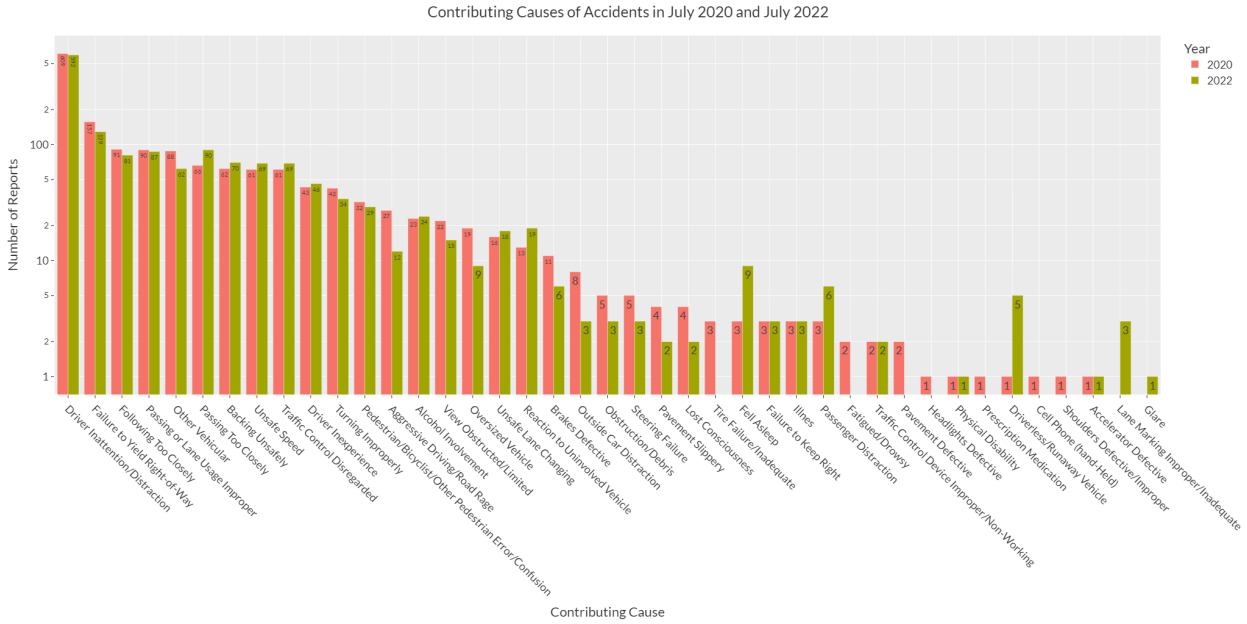






As seen from the above graphs, there were more accidents in June 2020 as compared to June 2022. In 2022, there were more people who were injured as compared to that in 2020. The maximum contributing cause of the accidents was Sedan followed by Sport Utility Vehicles overall. There were accidents caused by delivery trucks in 2020 and not in 2022.

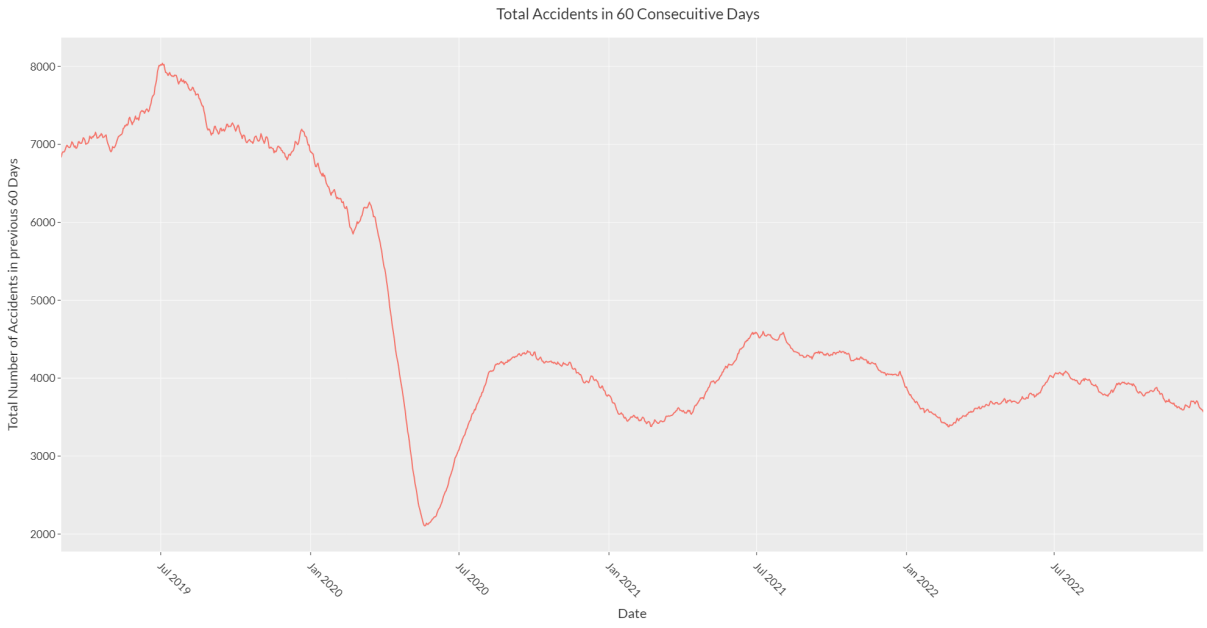
**4. How was July of 2020 different then July of 2022? Figure out how to show or demonstrate the difference. What was the reported cause of the accidents?**






July 2020 had more accidents compared to July 2022. The major cause of accidents for July 2020 and July 2022 was due to driver inattention/ distraction as seen in the graphs above, followed by the failure to yield right of way.

**5. For the year of January 2020 to October of 2022, which 60 consecutive days had the most accidents? The Automobile Association of America (AAA) says they are in the summer. Can you verify this?**



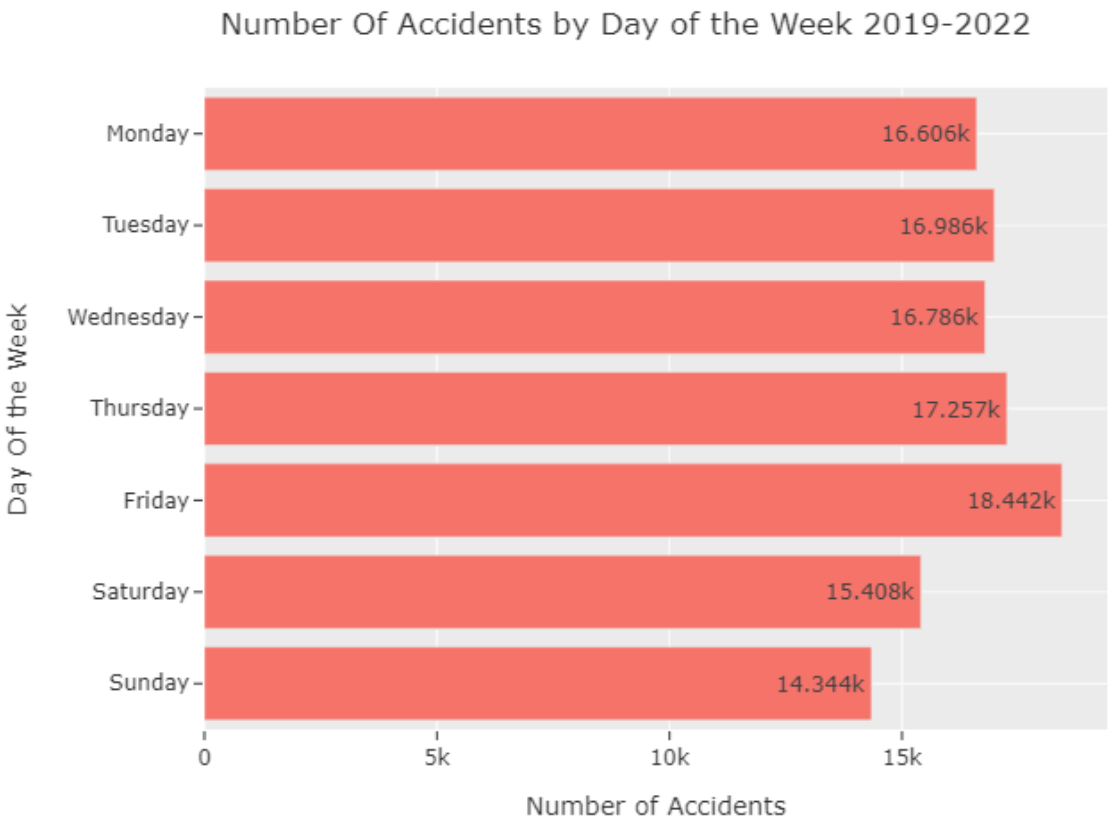
```
consecuitive_accidents_df[(consecuitive_accidents_df.index ≥ '2020-03-01')&(consecuitive_accidents_df.index<'2022-10-01')].sort_values(ascending=False)
```

0.0s

```
CRASH DATE
2020-03-13    6258.0
2020-03-14    6230.0
2020-03-12    6229.0
2020-03-15    6216.0
2020-03-11    6199.0
                ...
2020-05-23    2127.0
2020-05-24    2123.0
2020-05-19    2112.0
2020-05-21    2110.0
2020-05-20    2105.0
Name: CRASH DATE, Length: 944, dtype: float64
```

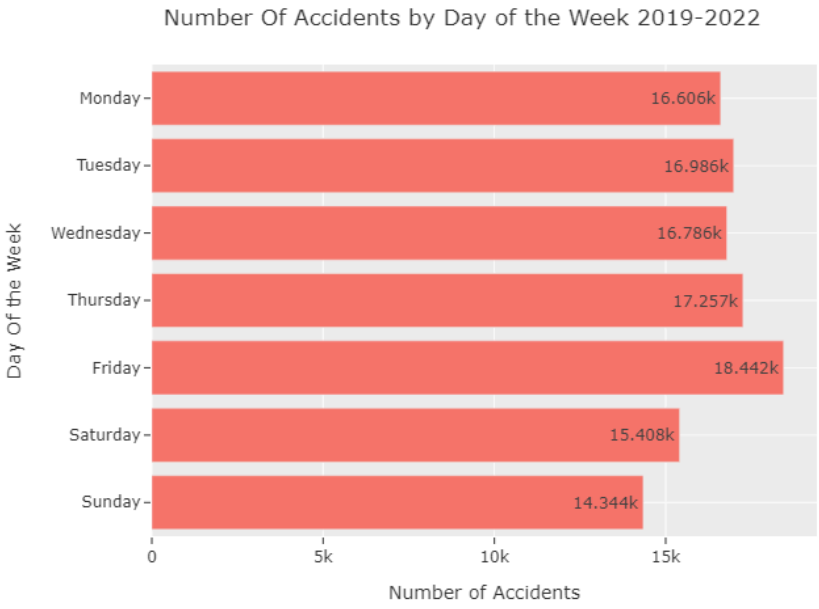
From January 2020 to October 2022, the 60 consecutive days that had the most number of accidents were from mid March (13th March) till 20th May 2020. The graph shows the maximum accident for that timeline to be starting from mid March as well. This continues till May 20, hence it can be verified that the Automobile Association of America were correct that most accidents took place in summer. In this case, it was early summer as we went in almost May.

**6. Which day of the week has the fewest accidents?**



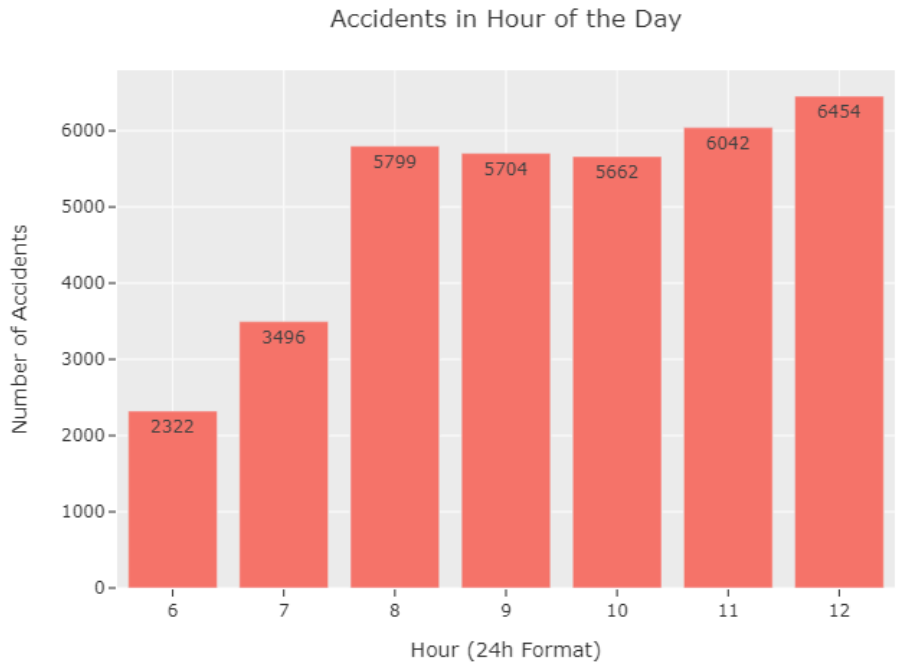

As seen from the graph, the day that had the least number of accidents was Sunday.

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



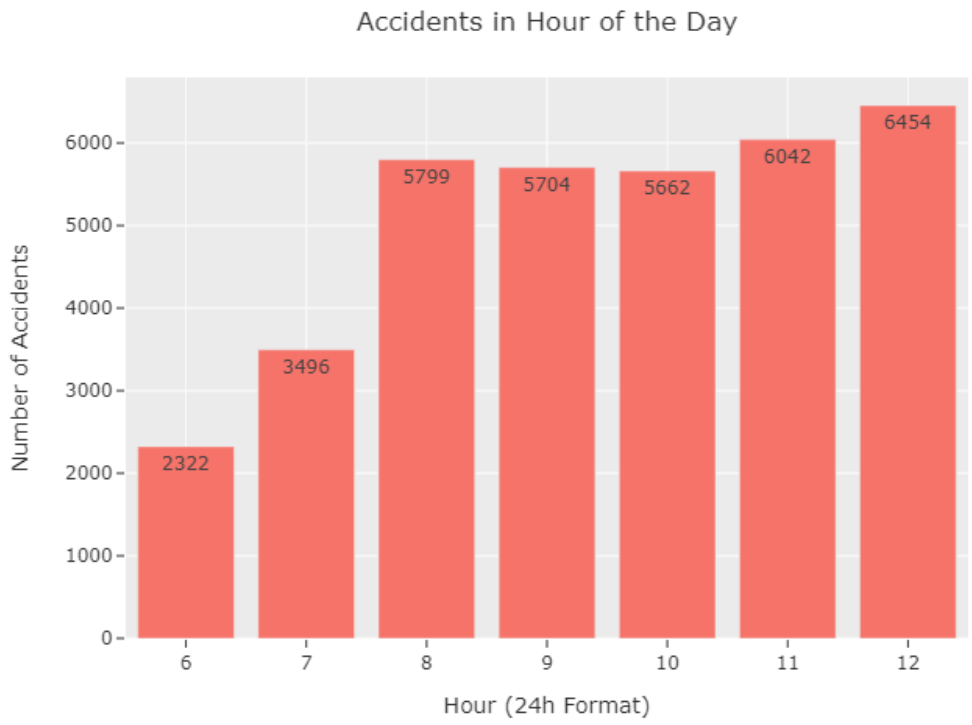

As seen from the graph, the day that had the most number of accidents was Friday, probably due to more vehicles on the road following the weekend.

**8. From 6 AM to 12PM, which hour of the day has the fewest accidents?**



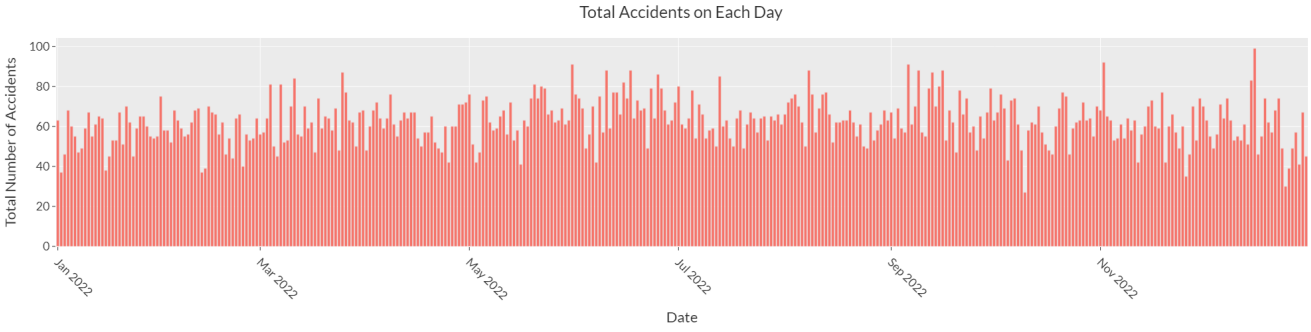

From 6AM to 12PM, the hour of the day with the fewest number of accidents was at 6AM.

**9. From 6 AM to 12PM, which hour of the day has the most accidents?**




From 6AM to 12PM, the hour of the day with the most number of accidents was at 12PM.

**10. In the year 2022, which 10 days had the most accidents? These are not consecutive days. For example, are there more accidents before the December holidays? Is this true? Can you speculate about why the worst days are the worst days?**



```
brooklyn_accidents.loc[
    brooklyn_accidents["YEAR"] == "2022", "CRASH DATE"
].value_counts().sort_values(ascending=False)[:10]
```

✓ 0.0s

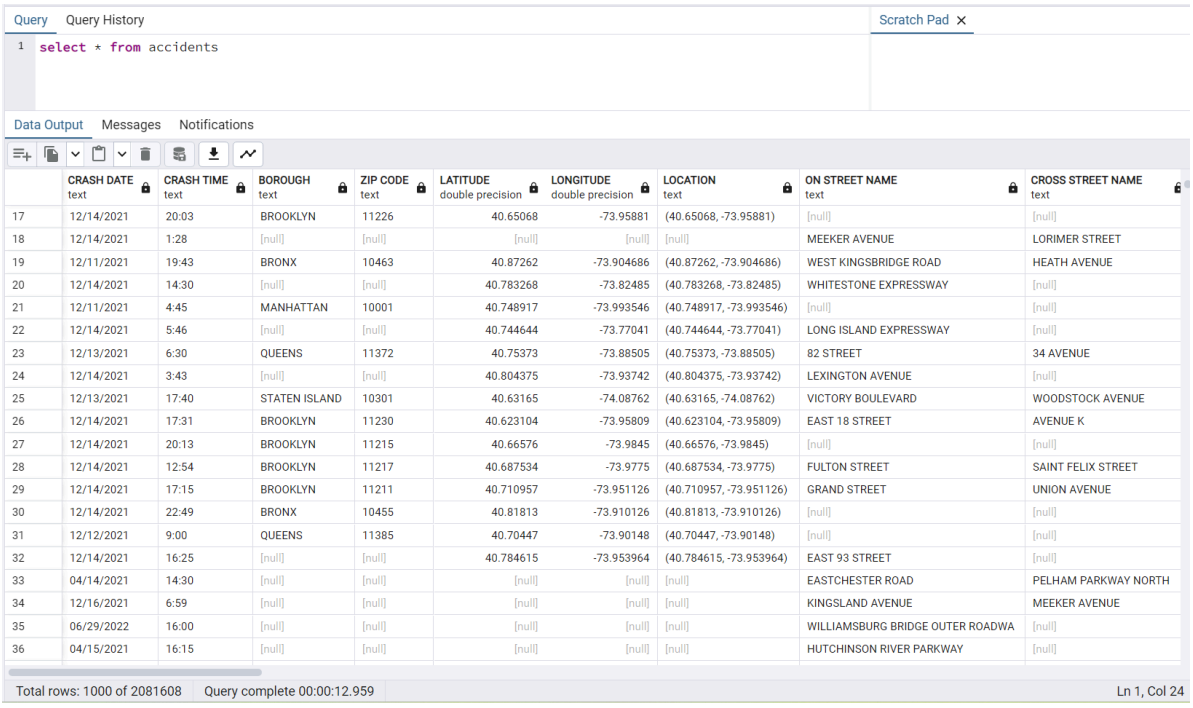
```
CRASH DATE
2022-12-16    99
2022-11-02    92
2022-05-31    91
2022-09-06    91
2022-09-16    88
2022-08-08    88
2022-06-17    88
2022-06-10    88
2022-09-09    88
2022-03-25    87
```

The above data shows the 10 days that had the most accidents. Starting with 16th December, the possible reason for the most accidents could be the pre Christmas time, where families would probably travel for the upcoming Christmas holidays. The next date fell on 2nd November just two days after Halloween which fell on October 30th in 2020. Next maximum accidents come on 31st May just after Memorial Day(30th May) followed by 6th September which comes just after Labor Day(on the 5th September). The later accidents cover most of the days in September which indicate that there might be some events that took place in that area around that time.

## 5. Conclusion and Discussion

Overall, we learnt the importance of data mining. We were given a huge dataset with the accidents that had taken place in different boroughs, after selecting one borough we found out the details of the accidents as to how they took place, at what time and such other important details. Some of the challenges we faced was to first ignore the missing values, since many places were unspecified and that could not be handled well hence we converted them to nan for simplicity. Next was loading the data into the database. Since the dataset was too huge, creating a python dataframe would take a lot of time hence we loaded the complete dataset from python to postgres database where we selected the Brooklyn borough with the years starting from 2019 to 2022. After that, we loaded the final Brooklyn accidents data back to python where we further

continued with our data mining. Below attached is the complete dataset queried and then the brooklyn accidents from 2019 to 2022 queried.



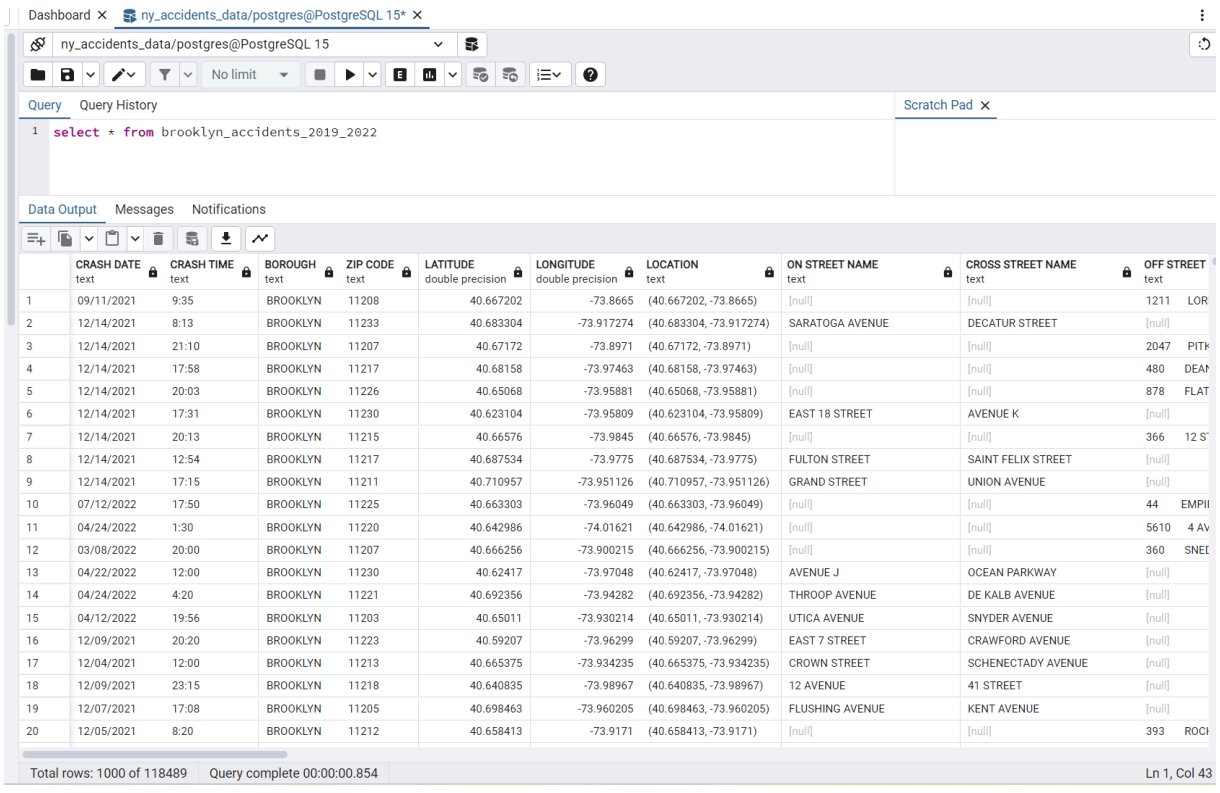
Query | Query History | Scratch Pad

```
1 select * from accidents
```

Data Output | Messages | Notifications

| | CRASH DATE text | CRASH TIME text | BOROUGH text | ZIP CODE text | LATITUDE double precision | LONGITUDE double precision | LOCATION text | ON STREET NAME text | CROSS STREET NAME text |
|---|---|---|---|---|---|---|---|---|---|
| 17 | 12/14/2021 | 20:03 | BROOKLYN | 11226 | 40.65068 | -73.95881 | (40.65068, -73.95881) | [null] | [null] |
| 18 | 12/14/2021 | 1:28 | [null] | [null] | [null] | [null] | [null] | MEEKER AVENUE | LORIMER STREET |
| 19 | 12/11/2021 | 19:43 | BRONX | 10463 | 40.87262 | -73.904686 | (40.87262, -73.904686) | WEST KINGSBRIDGE ROAD | HEATH AVENUE |
| 20 | 12/14/2021 | 14:30 | [null] | [null] | 40.783268 | -73.82485 | (40.783268, -73.82485) | WHITESTONE EXPRESSWAY | [null] |
| 21 | 12/11/2021 | 4:45 | MANHATTAN | 10001 | 40.748917 | -73.993546 | (40.748917, -73.993546) | [null] | [null] |
| 22 | 12/14/2021 | 5:46 | [null] | [null] | 40.744644 | -73.77041 | (40.744644, -73.77041) | LONG ISLAND EXPRESSWAY | [null] |
| 23 | 12/13/2021 | 6:30 | QUEENS | 11372 | 40.75373 | -73.88505 | (40.75373, -73.88505) | 82 STREET | 34 AVENUE |
| 24 | 12/14/2021 | 3:43 | [null] | [null] | 40.804375 | -73.93742 | (40.804375, -73.93742) | LEXINGTON AVENUE | [null] |
| 25 | 12/13/2021 | 17:40 | STATEN ISLAND | 10301 | 40.63165 | -74.08762 | (40.63165, -74.08762) | VICTORY BOULEVARD | WOODSTOCK AVENUE |
| 26 | 12/14/2021 | 17:31 | BROOKLYN | 11230 | 40.623104 | -73.95809 | (40.623104, -73.95809) | EAST 18 STREET | AVENUE K |
| 27 | 12/14/2021 | 20:13 | BROOKLYN | 11215 | 40.66576 | -73.9845 | (40.66576, -73.9845) | [null] | [null] |
| 28 | 12/14/2021 | 12:54 | BROOKLYN | 11217 | 40.687534 | -73.9775 | (40.687534, -73.9775) | FULTON STREET | SAINT FELIX STREET |
| 29 | 12/14/2021 | 17:15 | BROOKLYN | 11211 | 40.710957 | -73.951126 | (40.710957, -73.951126) | GRAND STREET | UNION AVENUE |
| 30 | 12/14/2021 | 22:49 | BRONX | 10455 | 40.81813 | -73.910126 | (40.81813, -73.910126) | [null] | [null] |
| 31 | 12/12/2021 | 9:00 | QUEENS | 11385 | 40.70447 | -73.90148 | (40.70447, -73.90148) | [null] | [null] |
| 32 | 12/14/2021 | 16:25 | [null] | [null] | 40.784615 | -73.953964 | (40.784615, -73.953964) | EAST 93 STREET | [null] |
| 33 | 04/14/2021 | 14:30 | [null] | [null] | [null] | [null] | [null] | EASTCHESTER ROAD | PELHAM PARKWAY NORTH |
| 34 | 12/16/2021 | 6:59 | [null] | [null] | [null] | [null] | [null] | KINGSLAND AVENUE | MEEKER AVENUE |
| 35 | 06/29/2022 | 16:00 | [null] | [null] | [null] | [null] | [null] | WILLIAMSBURG BRIDGE OUTER ROADWA | [null] |
| 36 | 04/15/2021 | 16:15 | [null] | [null] | [null] | [null] | [null] | HUTCHINSON RIVER PARKWAY | [null] |

Total rows: 1000 of 2081608 | Query complete 00:00:12.959 | Ln 1, Col 24



Dashboard | ny_accidents_data/postgres@PostgreSQL 15*

ny_accidents_data/postgres@PostgreSQL 15 | No limit

Query | Query History | Scratch Pad

```
1 select * from brooklyn_accidents_2019_2022
```

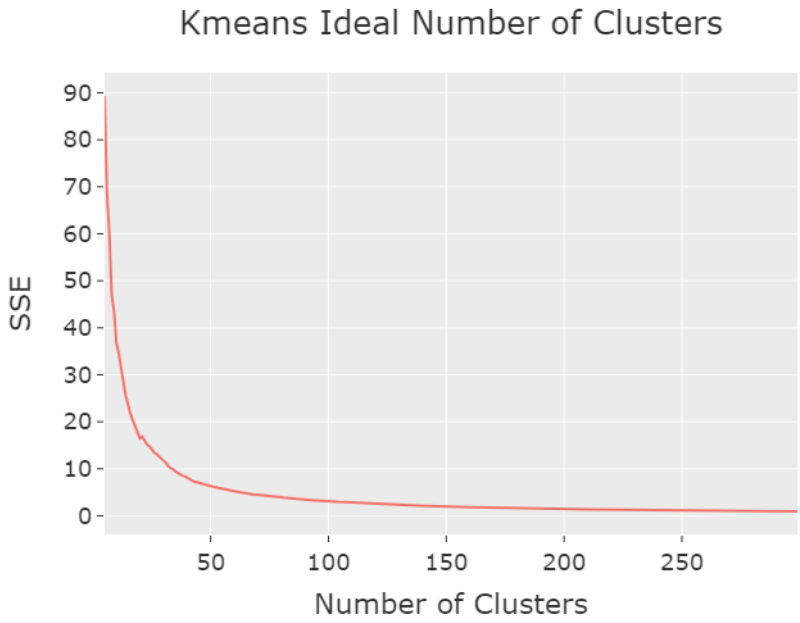
Data Output | Messages | Notifications

| | CRASH DATE text | CRASH TIME text | BOROUGH text | ZIP CODE text | LATITUDE double precision | LONGITUDE double precision | LOCATION text | ON STREET NAME text | CROSS STREET NAME text | OFF STREET text | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 09/11/2021 | 9:35 | BROOKLYN | 11208 | 40.667202 | -73.8665 | (40.667202, -73.8665) | [null] | [null] | 1211 | LOR |
| 2 | 12/14/2021 | 8:13 | BROOKLYN | 11233 | 40.683304 | -73.917274 | (40.683304, -73.917274) | SARATOGA AVENUE | DECATUR STREET | [null] | |
| 3 | 12/14/2021 | 21:10 | BROOKLYN | 11207 | 40.67172 | -73.8971 | (40.67172, -73.8971) | [null] | [null] | 2047 | PITK |
| 4 | 12/14/2021 | 17:58 | BROOKLYN | 11217 | 40.68158 | -73.97463 | (40.68158, -73.97463) | [null] | [null] | 480 | DEAI |
| 5 | 12/14/2021 | 20:03 | BROOKLYN | 11226 | 40.65068 | -73.95881 | (40.65068, -73.95881) | [null] | [null] | 878 | FLAT |
| 6 | 12/14/2021 | 17:31 | BROOKLYN | 11230 | 40.623104 | -73.95809 | (40.623104, -73.95809) | EAST 18 STREET | AVENUE K | [null] | |
| 7 | 12/14/2021 | 20:13 | BROOKLYN | 11215 | 40.66576 | -73.9845 | (40.66576, -73.9845) | [null] | [null] | 366 | 12 S |
| 8 | 12/14/2021 | 12:54 | BROOKLYN | 11217 | 40.687534 | -73.9775 | (40.687534, -73.9775) | FULTON STREET | SAINT FELIX STREET | [null] | |
| 9 | 12/14/2021 | 17:15 | BROOKLYN | 11211 | 40.710957 | -73.951126 | (40.710957, -73.951126) | GRAND STREET | UNION AVENUE | [null] | |
| 10 | 07/12/2022 | 17:50 | BROOKLYN | 11225 | 40.663303 | -73.96049 | (40.663303, -73.96049) | [null] | [null] | 44 | EMPII |
| 11 | 04/24/2022 | 1:30 | BROOKLYN | 11220 | 40.642986 | -74.01621 | (40.642986, -74.01621) | [null] | [null] | 5610 | 4 AV |
| 12 | 03/08/2022 | 20:00 | BROOKLYN | 11207 | 40.666256 | -73.900215 | (40.666256, -73.900215) | [null] | [null] | 360 | SNEI |
| 13 | 04/22/2022 | 12:00 | BROOKLYN | 11230 | 40.62417 | -73.97048 | (40.62417, -73.97048) | AVENUE J | OCEAN PARKWAY | [null] | |
| 14 | 04/24/2022 | 4:20 | BROOKLYN | 11221 | 40.692356 | -73.94282 | (40.692356, -73.94282) | THROOP AVENUE | DE KALB AVENUE | [null] | |
| 15 | 04/12/2022 | 19:56 | BROOKLYN | 11203 | 40.65011 | -73.930214 | (40.65011, -73.930214) | UTICA AVENUE | SNYDER AVENUE | [null] | |
| 16 | 12/09/2021 | 20:20 | BROOKLYN | 11223 | 40.59207 | -73.96299 | (40.59207, -73.96299) | EAST 7 STREET | CRAWFORD AVENUE | [null] | |
| 17 | 12/04/2021 | 12:00 | BROOKLYN | 11213 | 40.665375 | -73.934235 | (40.665375, -73.934235) | CROWN STREET | SCHENECTADY AVENUE | [null] | |
| 18 | 12/09/2021 | 23:15 | BROOKLYN | 11218 | 40.640835 | -73.98967 | (40.640835, -73.98967) | 12 AVENUE | 41 STREET | [null] | |
| 19 | 12/07/2021 | 17:08 | BROOKLYN | 11205 | 40.698463 | -73.960205 | (40.698463, -73.960205) | FLUSHING AVENUE | KENT AVENUE | [null] | |
| 20 | 12/05/2021 | 8:20 | BROOKLYN | 11212 | 40.658413 | -73.9171 | (40.658413, -73.9171) | [null] | [null] | 393 | ROCI |

Total rows: 1000 of 118489 | Query complete 00:00:00.854 | Ln 1, Col 43

Looking at the features available in the data we decided to primarily use Latitude and Longitude to perform clustering on the data, there were two reasons for this choice. Firstly, the alternative to using latitude and longitude is using street names but we found out that the street name data could be unreliable and it was missing in a lot more reports than the latitude and longitude data. Secondly, using latitude and longitude with a clustering Algorithm would allow us to find out the relation between two intersections which are close to each other but which would get missed out had we used the other approach.

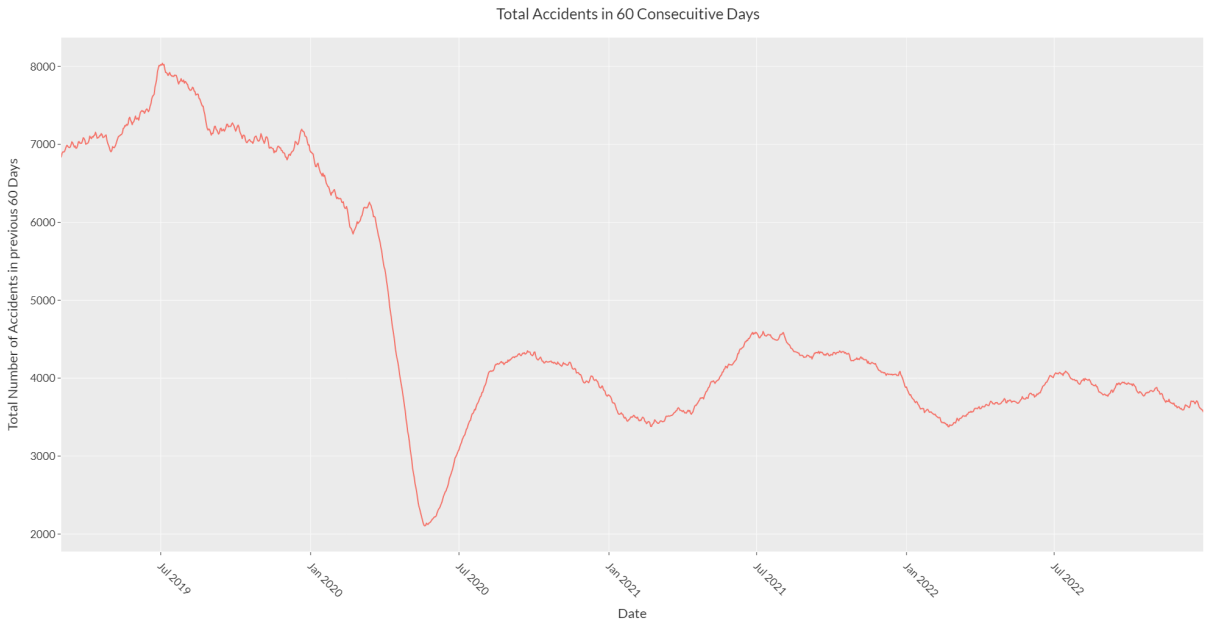
The Algorithm we finalized on was K-Means. We chose K Means entirely based on its efficiency. For finding out useful insights from the given data we had to iteratively create a large number of clusters testing out a number of features seeing if there was a correlation between the features. The best features we figured out were the location of the reported accident itself. According to our testing for finding out the number of clusters we should use, we chose to use a K value of 300.




We decided to use a k value greater than the elbow point because, plotting it on a graph allowed us to more accurately pinpoint the location of intersections which are accident prone. Also, using a lower K value meant that the clustering algorithm would club together internal intersections which are not accident prone with problem intersections just because they are closer in distance to larger highways.

Apart from the clustering Algorithm, in this assignment we wanted to compare the older maps from 2019 with the latest maps to see how the city has changed its infrastructure to increase safety. This proved to be a lot more difficult than we thought it would be, as it is really difficult to get a hold of older maps of the city with the resolution that we needed to notice small changes in the intersections and roads.

Another interesting point we noticed was that during the covid 19 pandemic, there was a sharp decrease in the accidents reported and even after the pandemic, the daily number of accidents sharply decreased.




And even after the pandemic, the total number of accidents over 60 consecutive days never went back up to the pre pandemic levels. We speculate that the main cause of this could be that there are a lot less number of cars on the roads and people traveling daily than before the pandemic, possibly due work from home jobs being common nowadays.

Overall, this project taught us the importance of mining a huge dataset to get meaningful insights. Data Mining helps us in extracting the useful information from huge amounts of data, and the mined information helps us in making informed decisions and identifying patterns. We saw how data mining helped us in making decisions by revealing the hidden patterns. Data mining essentially turns raw data into actionable insights The other thing we learnt was the importance of visualization. We have generated a number of visualizations in the form of heat maps, barplots to show the comparison between different years. This gives an immediate visual along with information as to how the data is. Visualization is hence extremely important when it comes to dealing with huge datasets.