CSCI-720 (Big Data Analytics)

Project

Aditi Karad (ak2298)

Pratyush Jena (pj1384)

Data Preparation

Q1.How clean is the data?

The data cleanliness presented several challenges:

- Missing Values: Significant missing or incomplete information in crucial fields like street names and accident details affected the analysis.
- Inconsistent Entries: There were inconsistencies in data entries, especially in textual fields like contributing factors or vehicle types, which required careful handling to standardize for analysis. For eg the first empty column was given as Unspecified and from there onwards the columns had empty values rather than having Unspecified in all the columns related to contributing factors

Overall, while the dataset provided valuable information, its cleanliness required considerable attention to ensure that the analysis was based on accurate and reliable data.

Q2 Which data did you ignore?

In the analysis of the NYPD Motor Vehicle Collisions dataset, the following outlines the data that we chose not to include in the analysis:

- Collisions Outside of Brooklyn: We exclusively analyzed incidents occurring in Brooklyn. Therefore, collisions in other boroughs, such as Queens, Manhattan, the Bronx, and Staten Island, were not considered.
- Non-2019 and 2020 Data: Our analysis was restricted to crash data from the years 2019 and 2020. As a result, collision records from other years were excluded.
- Months Other Than June and July: For certain questions we specifically targeted data from June and July of 2019 and 2020. Incidents from other months were not part of the analysis.
- Zip Code Information: The 'ZIP CODE' field in the dataset was not utilized in our analysis, as we are focusing on only one borough
- Collision ID: We dropped the collision id table as it serves its purpose as primary key for storing the tables in SQL or NoSQL databases. Since we are not using any database here we can drop the column
- Unspecified Data Fields: In instances where data fields were marked as 'Unspecified', such as some contributing factors or vehicle types the data is not added to the aggregated array.
- Location: We dropped the column named LOCATION due to redundancy as it was the combination of the column LATITUDE and LONGITUDE.
- Street Names: We didn’t focus much attention on the various types of street names that were given in our data because for the majority of data Street Names were an empty field. Nonetheless we have still kept the columns for Street Names in the data as it can be

populated using data for some 3rd party source by specifying the LATITUDE and LONGITUDE.

Q3 What data did you focus on?

In our analysis of the NYPD Motor Vehicle Collisions dataset, we concentrated on specific data elements that were most relevant to understanding traffic collision patterns and trends in Brooklyn for the years 2019 and 2020, particularly in the months of June and July. The data we focused on included:

- Temporal Data: We included crash data specifically from the years 2019 and 2020, with a further focus on the months of June and July for certain questions. This allowed for a targeted examination of collision trends and patterns during these specific timeframes.
- Geographical Focus on Brooklyn: The analysis was centered on collisions that occurred in the borough of Brooklyn. This geographical focus was essential for understanding local patterns and identifying potential hotspots within Brooklyn.
- Crash Date and Time: The 'CRASH DATE' and 'CRASH TIME' data were crucial for analyzing when collisions most frequently occurred, which helped in identifying potential temporal patterns in crash occurrences.
- Injury and Fatality Counts: We paid particular attention to data related to injuries and fatalities, where we aggregated all the column related to fatalities in accident with the name 'NUMBER OF PERSONS KILLED', and all the columns related to injured in accidents with the name 'NUMBER OF PERSONS INJURED'. This data provided insights into the severity and impact of the collisions.

- Location Details: We utilized the columns ‘LATITUDE’,’LONGITUDE’ to understand the specific locations of crashes within Brooklyn. This geographic data was instrumental in mapping collision hotspots.
- Contributing Factors: The 'CONTRIBUTING FACTOR VEHICLE' fields were analyzed to identify common causes of collisions. Understanding these factors is key to recognizing preventable causes and proposing potential safety improvements.
- Vehicle Types Involved: Information on the types of vehicles involved in collisions, captured in 'VEHICLE TYPE' fields, was included in the analysis. This helped in understanding if certain types of vehicles were more frequently involved in incidents.

By focusing on these specific elements of the dataset, we aimed to gain a comprehensive understanding of traffic collision trends in Brooklyn for the designated period, identifying key patterns, contributing factors, and potential areas for intervention to enhance road safety.

Q4 Did you quantize the data into regions?

For this project, the data was not further quantized into smaller regions within Brooklyn. The entire analysis was conducted at the borough level, focusing exclusively on Brooklyn as a whole, without subdividing it into neighborhoods, ZIP codes, or other smaller geographical units. This approach allowed for a broad overview of traffic patterns and accident trends across Brooklyn, but it did not delve into more localized or granular spatial analysis within the borough.

Q5 Are there any issues with the data?

Yes we encountered the following issues with the data

- Missing Street Names: The absence of street names for many entries, despite having latitude and longitude coordinates, can complicate the process of location-based analysis. Without specific street names, it becomes challenging to accurately determine the exact location of accidents, which is crucial for identifying high-risk areas and implementing targeted safety measures.
- Incomplete Latitude and Longitude Data: Some rows did not contain latitude and longitude data which can hinder spatial analysis. Since we are finding the hotspots of accident based on Latitude and Longitude this can create inaccuracies in identifying the hotspots as we might miss some places..
- Unpopulated Contributing Factors and Vehicle Types: Incompletedata in 'CONTRIBUTING FACTOR VEHICLE' and 'VEHICLE TYPE CODE' fields can significantly impact the analysis of crash causes and the types of vehicles involved.

Q6 Is the data from the two years comparable?

Yes, the data from the two years is comparable, but with considerations. The consistent format and collection methods across both years provided a solid basis for comparison. However, it was crucial to account for unique factors affecting each year, especially the impact of the COVID-19 pandemic in 2020. This required careful interpretation of trends and anomalies, ensuring that any observed changes in traffic patterns and accident frequencies were contextualized within the broader societal and environmental changes unique to each year.

Q7Are there any other issues you found?

In the data preparation phase, several issues were identified:

- Missing or Incomplete Data: Key fields like street names and contributing factors had missing or incomplete entries, complicating the analysis of specific accident locations and causes.
- Inconsistencies in Data Recording: Variations in how data was recorded, such as differing formats or terminology used in the contributing factors, posed challenges for consistent analysis.

Assignment Answers

(For Below questions GREEN color is associated with 2019 and RED color is associated with 2020)

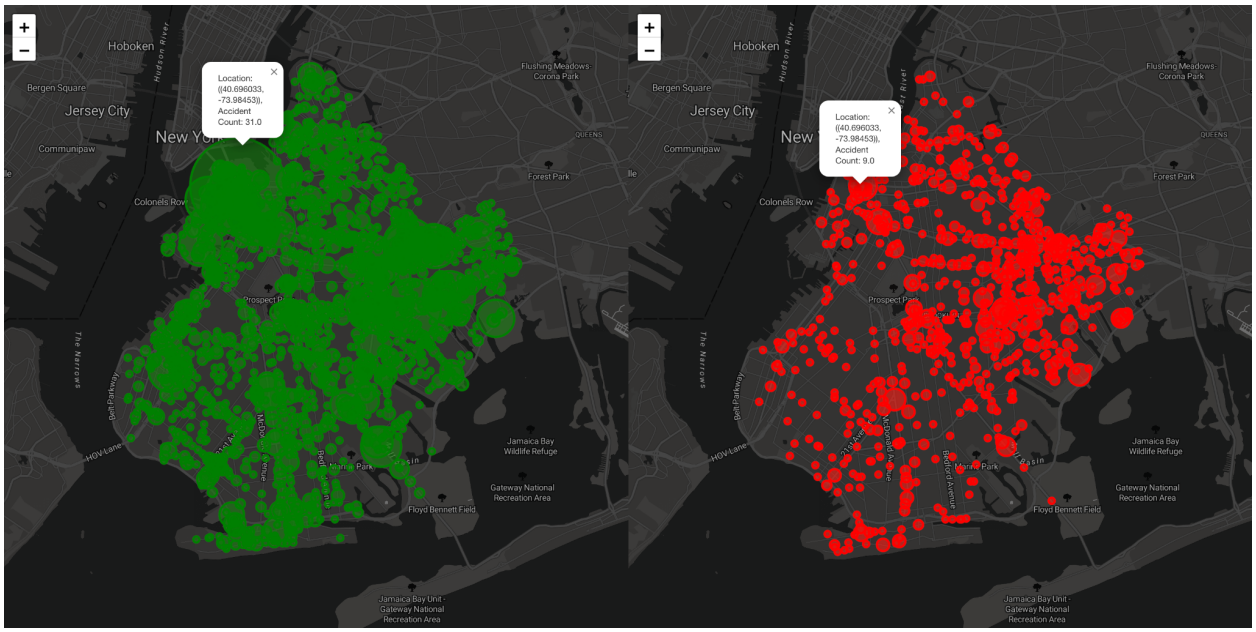
Q1 For the two years given, figure out what has changed in the summer from one year to the next. Figure out how to visualize the difference, in some way

Below are the differences we observed between the Summer of 2019 and Summer of 2020.

```
Question 1
-----------------------------------------------------------
Summer 2019
---------
Total number of accidents:11515
Max number of accidents: 31
Location with max accidents: (40.696033, -73.98453)
Number of persons killed: 36
Number of persons injured: 7238
Most frequent types of accident:-
    Vehicle Types: ('Sedan', 'Station Wagon/Sport Utility Vehicle'), Count: 3500
    Vehicle Types: ('Sedan', 'Sedan'), Count: 2421
    Vehicle Types: ('Station Wagon/Sport Utility Vehicle', 'Station Wagon/Sport Utility Vehicle'), Count: 1858
    Vehicle Types: ('Sedan', 'Sedan', 'Station Wagon/Sport Utility Vehicle'), Count: 407
    Vehicle Types: ('Sedan', 'Station Wagon/Sport Utility Vehicle', 'Station Wagon/Sport Utility Vehicle'), Count: 393
Most probable types of accidents:-
    ('Pick-up Truck',) → ('Sedan',), accident probability (confidence): 0.5213270142180095
    ('Tractor Truck Diesel',) → ('Sedan',), accident probability (confidence): 0.5220588235294118
Most common contributing factor: ('Driver Inattention/Distraction', 'Driver Inattention/Distraction')
```

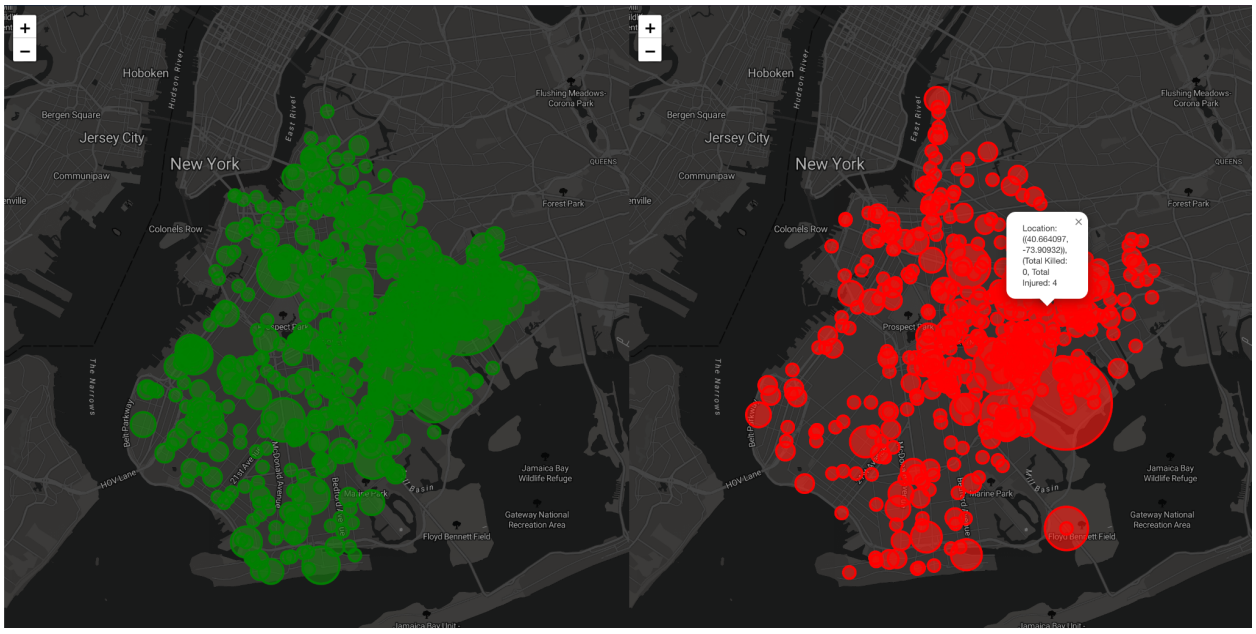
```
---------
Summer 2020
---------
Total number of accidents:6098
Max number of accidents: 9
Location with max accidents: (40.696033, -73.98453)
Number of persons killed: 32
Number of persons injured: 5706
Most frequent types of accident:-
    Vehicle Types: ('Sedan', 'Station Wagon/Sport Utility Vehicle'), Count: 1633
    Vehicle Types: ('Sedan', 'Sedan'), Count: 1580
    Vehicle Types: ('Station Wagon/Sport Utility Vehicle', 'Station Wagon/Sport Utility Vehicle'), Count: 851
    Vehicle Types: ('Sedan', 'Sedan', 'Sedan'), Count: 377
    Vehicle Types: ('Sedan', 'Sedan', 'Station Wagon/Sport Utility Vehicle'), Count: 360
Most probable types of accidents:-
    ('Pick-up Truck',) → ('Sedan',), accident probability (confidence): 0.5658536585365853
    ('Box Truck',) → ('Sedan',), accident probability (confidence): 0.522633744855967
Most common contributing factor: ('Driver Inattention/Distraction',)
```

We see that there are significantly more accidents during Summer 2019 than Summer 2020. We can also see the maximum number of accidents on a single day in 2019 was 31 while the maximum number of accidents in 2020 was 9. All of this can be associated with the reason that in 2020 we had COVID-19 and summer 2020 was the period of lockdown. Thus we see such few accidents. But we also see that the hotspot of accidents for 2019 and 2020 has not changed in between years. This indicates that this street is a very dangerous region for drivers as well as pedestrians and thus we might need stricter traffic rules for the given street. Below is the visualization of the number of accidents that happened in each location. I am only visualizing those locations where 2 or more than 2 accidents have happened.



PLOT ON THE BASIS OF ACCIDENT COUNT (2019 vs 2020)

PLOT ON THE BASIS OF SEVERITY (2019 vs 2020)



We have also visualized the severity of accidents below. As seen below the hotspot with most accidents is different than where most people have been injured or killed in the accidents. We also observe that the most common type of accident involves a Sedan and a Station Wagon which are very common in USA as these cars are usually owned by families in USA. We also observe that most common contributing factor for accidents was Driver Inattention/Distraction.

Q2 How was June of 2019 different then June of 2020? Figure out how to show or demonstrate the difference.

Below are the differences we observed between the Summer of 2019 and Summer of 2020.

```
June 2019
---------
Total number of accidents:4067
Max number of accidents: 13
Location with max accidents: (40.696033, -73.98453)
Number of persons killed: 16
Number of persons injured: 2372
Most frequent types of accident:-
    Vehicle Types: ('Sedan', 'Station Wagon/Sport Utility Vehicle'), Count: 1266
    Vehicle Types: ('Sedan', 'Sedan'), Count: 867
    Vehicle Types: ('Station Wagon/Sport Utility Vehicle', 'Station Wagon/Sport Utility Vehicle'), Count: 662
    Vehicle Types: ('Sedan', 'Station Wagon/Sport Utility Vehicle', 'Station Wagon/Sport Utility Vehicle'), Count: 139
    Vehicle Types: ('Sedan', 'Sedan', 'Station Wagon/Sport Utility Vehicle'), Count: 132
Most probable types of accidents:-
    ('Station Wagon/Sport Utility Vehicle',) → ('Sedan',), accident probability (confidence): 0.4425026214610276
    ('Taxi',) → ('Sedan',), accident probability (confidence): 0.4407894736842105
    ('Pick-up Truck',) → ('Sedan',), accident probability (confidence): 0.47085201793721976
    ('Motorcycle',) → ('Sedan',), accident probability (confidence): 0.46774193548387094
    ('Bus',) → ('Sedan',), accident probability (confidence): 0.5267857142857143
    ('Box Truck',) → ('Sedan',), accident probability (confidence): 0.4563758389261745
    ('Bike',) → ('Sedan',), accident probability (confidence): 0.4117647058823529
    ('Bus',) → ('Station Wagon/Sport Utility Vehicle',), accident probability (confidence): 0.4107142857142857
    ('Tractor Truck Diesel',) → ('Sedan',), accident probability (confidence): 0.5238095238095238
```

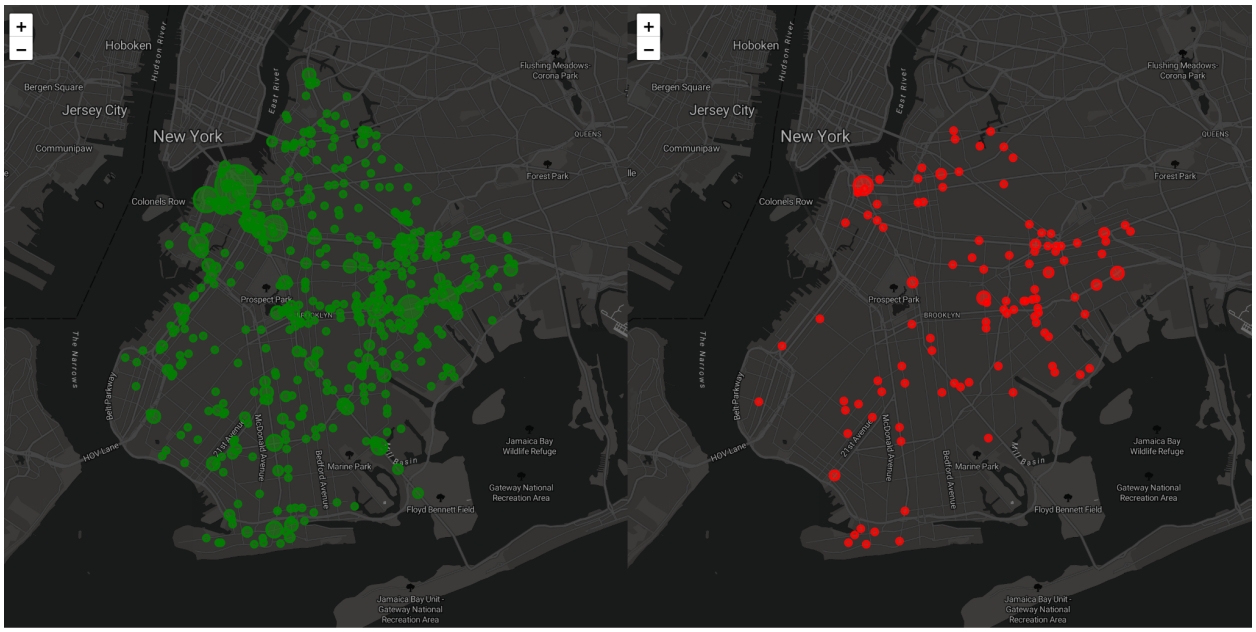
```
---------
June 2020
---------
Total number of accidents:1750
Max number of accidents: 6
Location with max accidents: (40.696033, -73.98453)
Number of persons killed: 10
Number of persons injured: 1698
Most frequent types of accident:-
    Vehicle Types: ('Sedan', 'Station Wagon/Sport Utility Vehicle'), Count: 486
    Vehicle Types: ('Sedan', 'Sedan'), Count: 434
    Vehicle Types: ('Station Wagon/Sport Utility Vehicle', 'Station Wagon/Sport Utility Vehicle'), Count: 270
    Vehicle Types: ('Sedan', 'Station Wagon/Sport Utility Vehicle', 'Station Wagon/Sport Utility Vehicle'), Count: 104
    Vehicle Types: ('Sedan', 'Sedan', 'Station Wagon/Sport Utility Vehicle'), Count: 99
Most probable types of accidents:-
    ('Station Wagon/Sport Utility Vehicle',) → ('Sedan',), accident probability (confidence): 0.44181818181818183
    ('Bike',) → ('Station Wagon/Sport Utility Vehicle',), accident probability (confidence): 0.40816326530612246
    ('Taxi',) → ('Sedan',), accident probability (confidence): 0.4915254237288136
    ('Pick-up Truck',) → ('Sedan',), accident probability (confidence): 0.6666666666666666
    ('Box Truck',) → ('Sedan',), accident probability (confidence): 0.42028985507246375
    ('Motorcycle',) → ('Station Wagon/Sport Utility Vehicle',), accident probability (confidence): 0.48936170212765956
```

Comparing June 2019 to June 2020, there was a significant reduction in the total number of accidents, from 4,067 to 1,750. This decrease is likely a direct effect of the COVID-19 pandemic and the associated restrictions on movement. The location with the maximum accidents remained the same, indicating a persisting hotspot that may require targeted safety measures regardless of overall trends.

The number of persons injured remained relatively high despite the fewer accidents, suggesting that while there were fewer incidents, the severity of those that did occur may not have diminished proportionally. This could be due to a variety of factors, including changes in traffic patterns or the nature of the accidents during the pandemic.

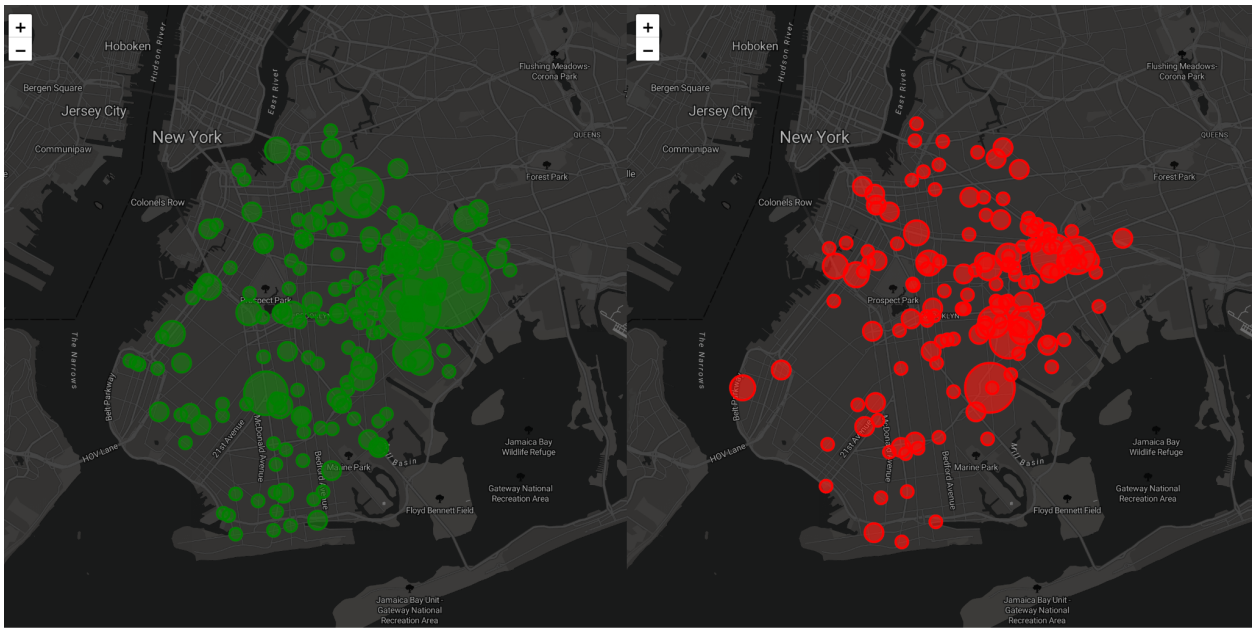
The types of vehicles involved in the most frequent accidents remained consistent, with sedans and station wagons/sport utility vehicles being the most common. However, the counts for these combinations dropped by more than half, in line with the overall reduction in accidents**.**

The probabilities of certain types of accidents occurring given the involvement of specific vehicle types—what the data refers to as "most probable types of accidents"—showed some variations. Notably, the confidence for accidents involving a 'Pick-up Truck' leading to a 'Sedan' increased significantly in 2020.



PLOT ON THE BASIS OF ACCIDENT COUNT (2019 vs 2020)

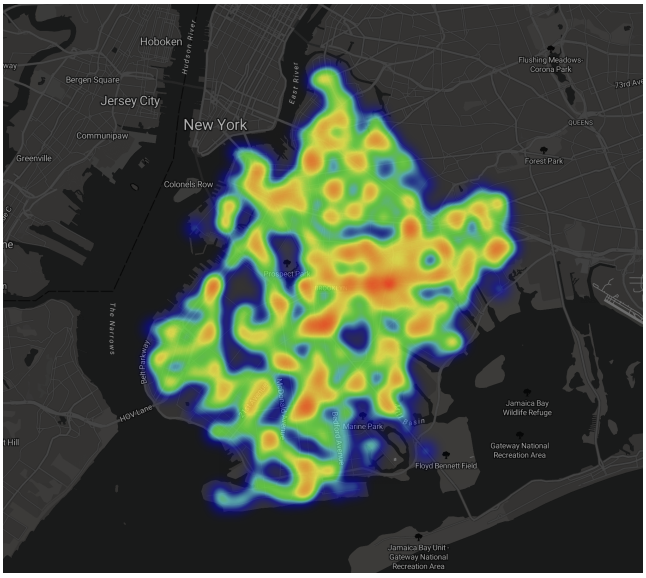
PLOT ON THE BASIS OF SEVERITY (2019 vs 2020)



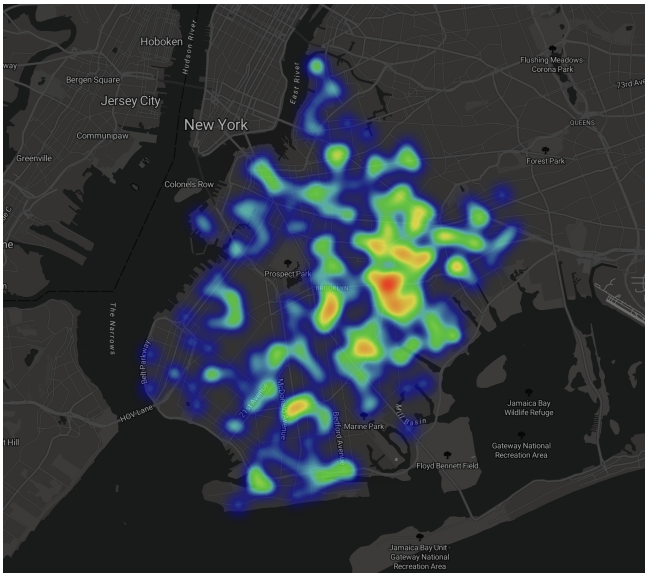
As seen in Question 1 we observe a similar trend between the hotspot for frequency of accidents and hotspots for severity of accidents.

We have also plotted the heatmap of where the most common type of accident happens (Sedan , Station Wagon) for both years 2019 and 2020

JUNE 2019 (Sedan , Station Wagon)



JUNE 2020 (Sedan , Station Wagon)



Q3 How was July of 2019 different then July of 2020? Figure out how to show or demonstrate the difference?

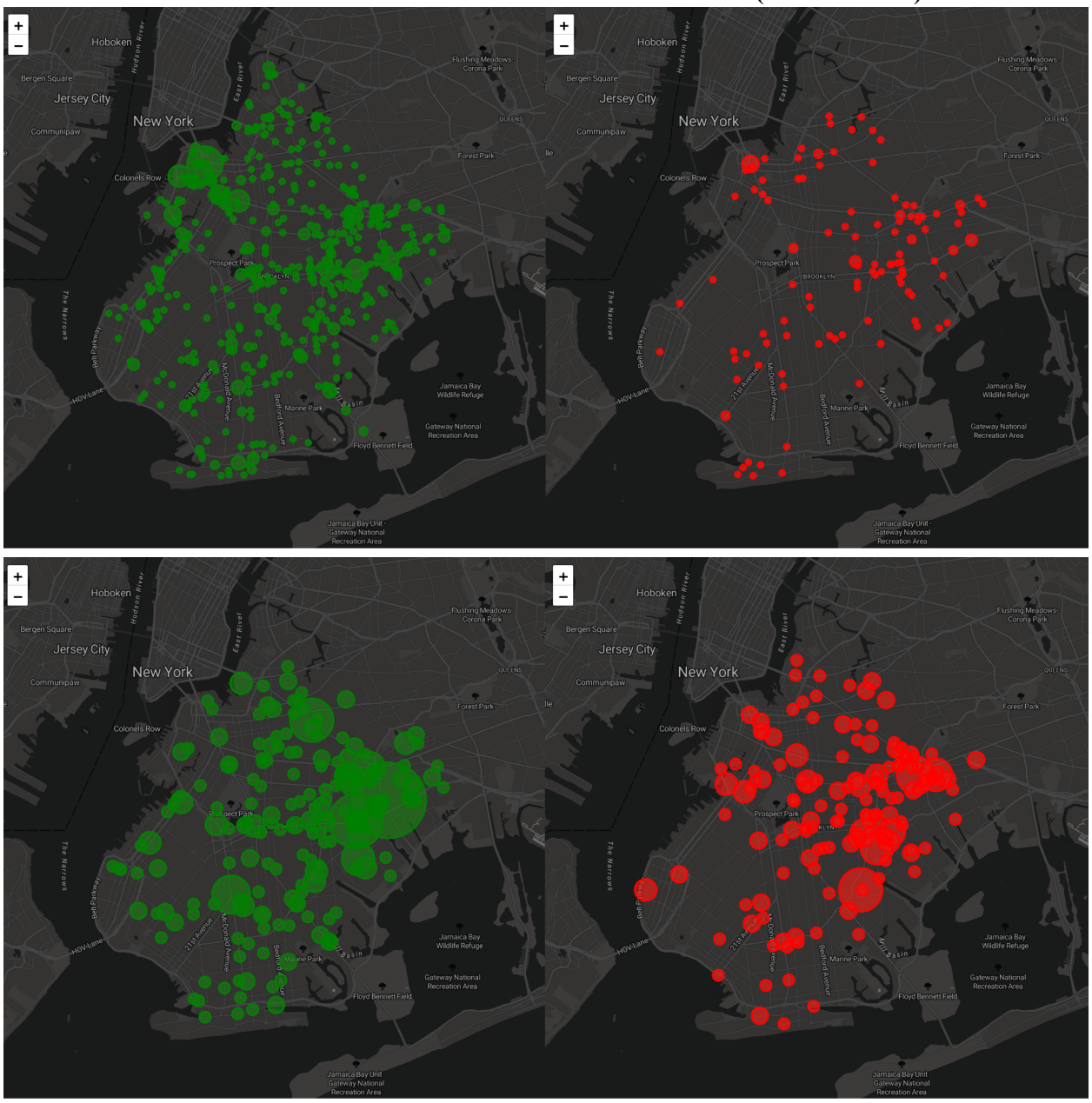
Comparing July 2019 and July 2020, the total number of accidents decreased from 3,837 to 2,145. The location with the maximum number of accidents also shifted, which may indicate changes in traffic flow or possibly construction and road closures that altered driving patterns.

Despite a lower overall accident count in July 2020, there was an increase in fatalities from 12 to 16. This could suggest that accidents were more severe, which might be due to various factors, including changes in driving behavior during the pandemic.

The most frequent types of accidents involved sedans and station wagons/sport utility vehicles in both years, although the counts decreased in 2020, consistent with the overall trend of fewer accidents. The types of vehicles involved remained the same, but with lower frequencies.

The probabilities of certain accident types, given the involvement of specific vehicle types, show some changes. Notably, the confidence for accidents involving 'Pick-up Truck' leading to 'Sedan' increased in 2020, which could reflect a change in vehicle usage or reporting practices during the pandemic.

PLOT ON THE BASIS OF ACCIDENT COUNT (2019 vs 2020)

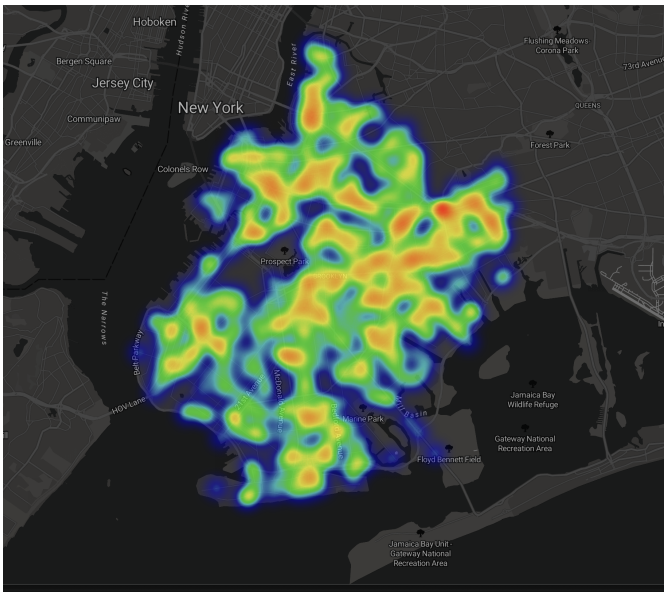


PLOT ON THE BASIS OF SEVERITY (2019 vs 2020)

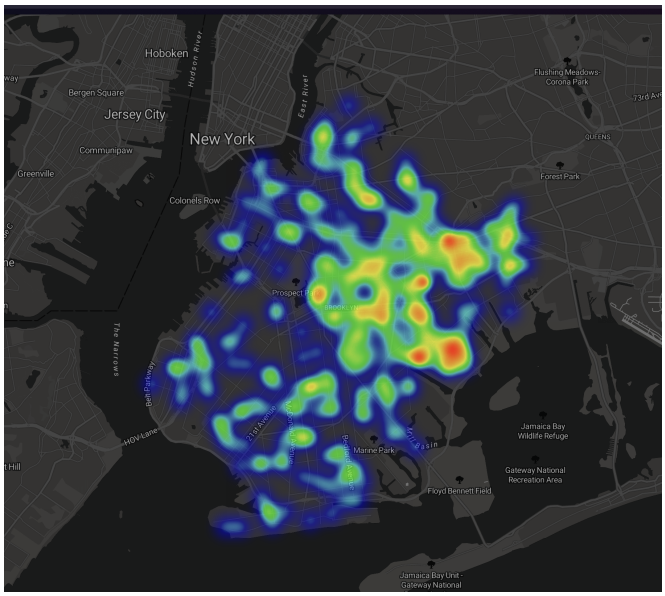
As seen in Question 1 we observe a similar trend between the hotspot for frequency of accidents and hotspots for severity of accidents.

We have also plotted the heatmap of where the most common type of accident happens (Sedan , Station Wagon) for both years 2019 and 2020

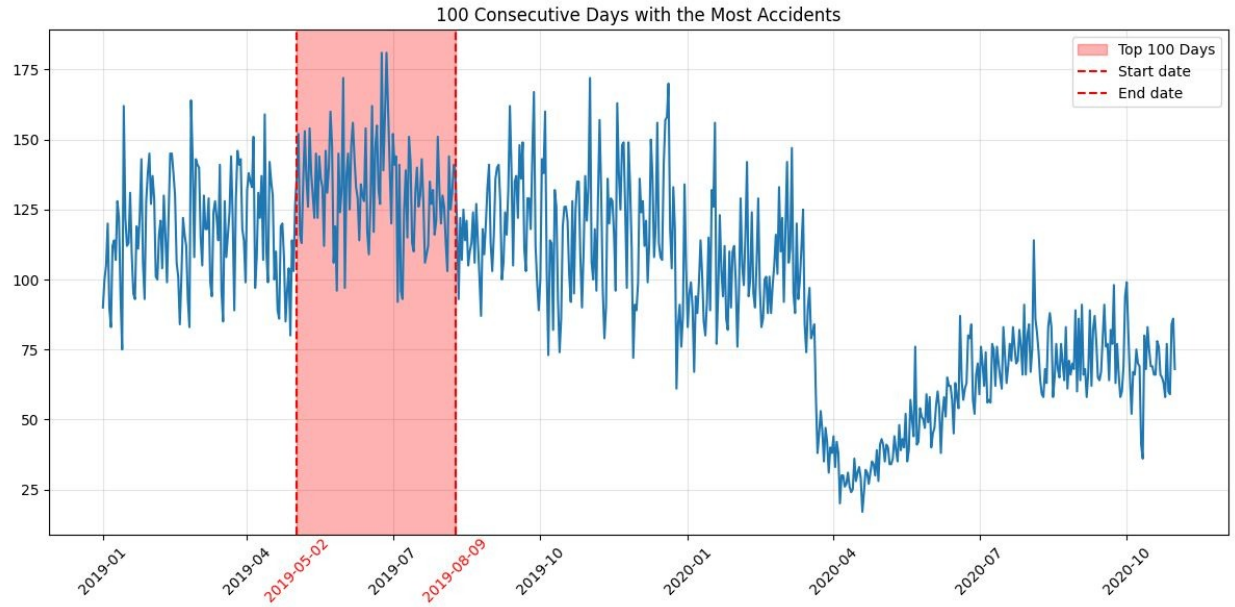
JULY 2019 (Sedan , Station Wagon)



JULY 2020 (Sedan , Station Wagon)



Q4 For the year of January 2019 to October of 2020, which 100 consecutive days had the most accidents?



The graph illustrates the number of accidents over the period from January 2019 to October 2020, with a highlighted section indicating the 100 consecutive days with the most accidents. Based on the data and the visualization:

- The 100 consecutive days with the most accidents fell within the highlighted timeframe. This period likely reflects specific underlying factors such as seasonal weather conditions, holiday-related traffic, or other temporal influences that could have increased accident rates.
- The trend seen in the graph, particularly the sharp drop-off toward the end of the highlighted period, might coincide with the beginning of the COVID-19 pandemic and subsequent lockdowns, which likely led to a decrease in traffic and accidents.

Q5 Which day of the week has the most accidents?

Day of the Week with most accidents: **Friday**

Day of the Week with most accidents considering the number of people killed or injured: **Saturday**

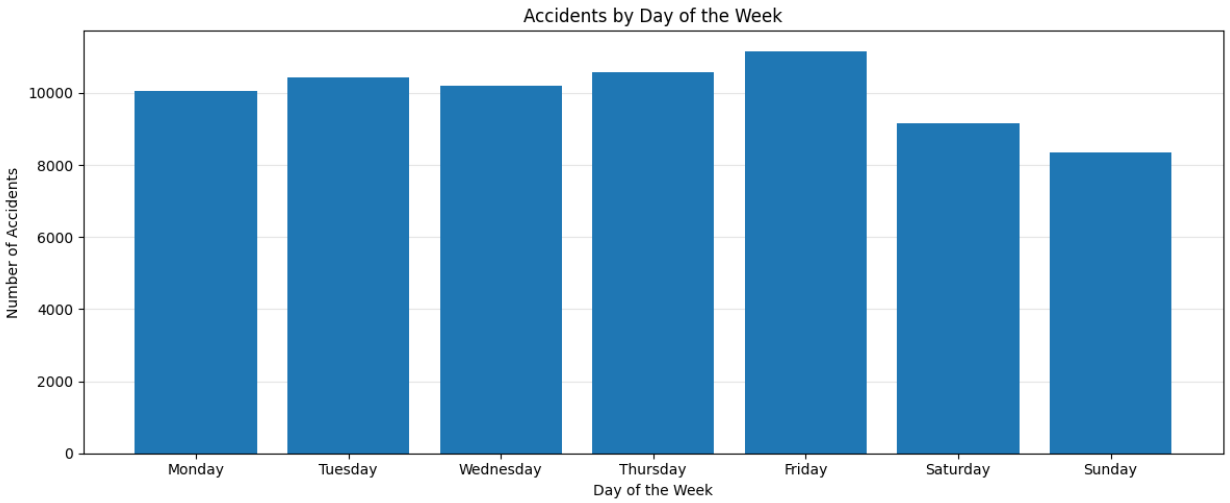
Week data:

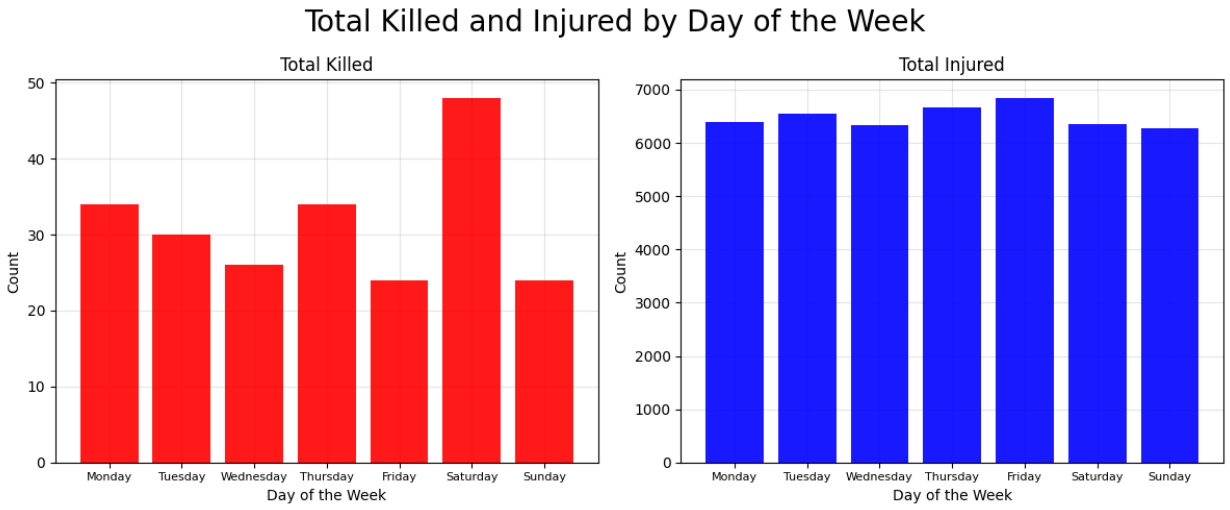
	DAY_OF_WEEK	ACCIDENT COUNT
0	Monday	10065
1	Tuesday	10426
2	Wednesday	10206
3	Thursday	10571

4 Friday 11164

5 Saturday 9159

6 Sunday 8360



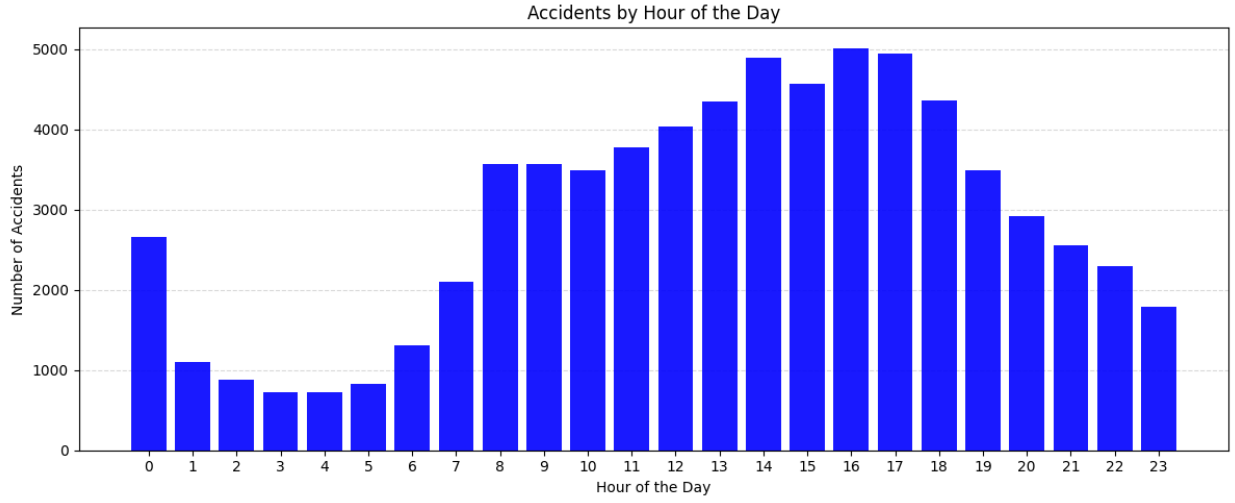


Based on the analysis, the following points can be inferred regarding accidents and days of the week:

- Most Accidents on Friday: The highest overall count of accidents occurred on Fridays. This could be due to increased traffic from weekend commutes, after-work social activities, and potentially more relaxed or distracted driving behaviors at the end of the workweek.
- Highest Injury and Fatality Rates on Saturday: Despite fewer total accidents, Saturdays have the most accidents considering the number of people killed or injured. This suggests that accidents on Saturdays might be more severe, potentially due to factors like higher speeds, increased recreational travel, or alcohol-related driving late at night.
- Accident Frequency Pattern Across the Week: The data shows a gradual increase in accidents from Monday to Friday, peaking on Friday. The decrease on Saturdays and Sundays might be due to lower commuter traffic, but the heightened severity on Saturdays is a notable exception. We can see from the graphs above that on the weekends the freq of accidents is low which can be because people prefer to stay at home and frelax during the weekends rather than going out causing less amount of accidents

Q6 Which hour of the day has the most accidents?

Hour of the day with most accidents: 4 PM(16:00)



The bar chart provided shows the distribution of accidents by hour of the day, with the highest number of accidents occurring at 4 PM. This observation can be attributed to several factors commonly associated with this time of day:

- Rush Hour Traffic: 4 PM is typically part of the afternoon rush hour when there is a significant increase in traffic as people leave work and schools. The higher volume of vehicles on the road increases the likelihood of collisions.
- Changing Road Conditions: As daylight begins to change, particularly in the fall and winter months, drivers may experience more difficulty adapting to the lighting conditions, potentially leading to more accidents.
- Driver Fatigue: By late afternoon, drivers may be more tired, especially after a long workday, which can lead to decreased alertness and slower reaction times.
- School Dismissal: In many areas, this time coincides with students leaving school, adding to traffic congestion and increasing the number of pedestrians, particularly children, who are more vulnerable to traffic accidents.
- Distractions: The end of the workday may also see more drivers using their phones to make calls or send messages as they leave work or run errands, leading to greater distraction-related accidents.
- Urgency to Reach Destinations: There may be a sense of urgency among drivers to get home or to other destinations, which can result in more aggressive driving behaviors such as speeding or unsafe lane changes.

Q7 In the year 2020, which 12 days had the most accidents?Can you speculate about why this is?

12 days that had the most accidents are:

	CRASH DATE	ACCIDENT COUNT
0	2020-01-18	156
1	2020-03-06	147
2	2020-02-07	142
3	2020-03-03	142
4	2020-02-27	133
5	2020-01-16	132
6	2020-02-03	129
7	2020-02-14	129
8	2020-01-17	126
9	2020-03-13	125
10	2020-02-10	124
11	2020-01-21	123

Analyzing the dates with the highest accident counts in 2020, several patterns and potential causes for the spikes in accidents on these specific days can be speculated:

- Winter Weather Conditions: Many of the dates with high accident counts are in January, February, and March, which are winter months in New York. Inclement weather, such as snow and ice, can lead to poor road conditions, reduced visibility, and an increase in accidents.
- Post-Holiday Traffic: The date of January 18th, which had the highest accident count, this can be because most of the universities in U.S start spring semester around this time and that's the reason why accidents are seen to be increased in these 7 days which can be seen in the data above
- Valentine's Day Traffic: February 14th is Valentine's Day, and the increase in accidents could be due to more people going out to celebrate, potentially leading to higher traffic volumes and possibly more incidents of driving under the influence.
- Beginning of the Month: The early days of the month, such as February 3rd and March 3rd, may see increased activity as people run errands or take care of tasks like paying bills, leading to more congestion and potential accidents.
- Fridays: Notably, March 6th and March 13th are Fridays, which often have higher traffic volumes due to weekend travel and social activities. Fridays can also see an increase in after-work socializing and drinking, which can contribute to accidents.
- COVID-19 Pandemic Onset: The dates in March coincide with the early stages of the COVID-19 pandemic in New York. The initial response to the pandemic may have affected driving patterns, with people potentially rushing to stock up on supplies or alter their routines.

Conclusion

Q1 What did you learn overall?

Throughout the analysis of the NYPD Motor Vehicle Collisions dataset, I learned that traffic accident trends are influenced by a complex interplay of factors, including temporal patterns like time of day and day of the week, as well as broader societal influences such as weather conditions and public events. Fridays showed the highest number of accidents, which suggests a combination of heavy traffic, leisure activities, and potentially riskier driving behaviors associated with weekends. The peak in accidents at 4 PM aligns with rush hour traffic, end-of-day fatigue, and varying light conditions.

The most accident-prone 100-day period identified within the 2019 to 2020 range, likely precedes the significant changes in traffic behavior due to the COVID-19 pandemic. The clustering of high-accident days during this period may suggest seasonal effects, such as winter weather conditions, or increased travel due to holidays, which historically correlate with higher accident rates. The data underscores the variability of accident patterns before the pandemic and can serve as a reference point for understanding the full impact of COVID-19 on road safety and traffic dynamics.

Overall, the analysis underscores the necessity of context when interpreting traffic data and the value of a multifaceted approach to improving road safety that accounts for these varied factors.

Q2 What went wrong, or what challenges did you face?

During the analysis, the main challenges faced were related to the quality and completeness of the data. There were instances of missing or incomplete information, such as absent street names and unspecified contributing factors for accidents, which made it difficult to pinpoint exact

locations or causes. The pandemic's unprecedented impact on traffic patterns in 2020 also introduced a variable that was not present in previous years, adding complexity to the year-over-year comparison.

Q3 What was interesting about this?

The most interesting aspect was seeing the clear impact of external factors, like the COVID-19 pandemic, on traffic accident patterns. As shown in the graph above we can clearly see a steep drop in the number of accidents right around when the lockdown started. It was intriguing to observe the shifts in peak times for accidents and how societal behavior, such as increased activity on Fridays or specific times like rush hour, directly correlates with accident frequency. This analysis highlighted the dynamic nature of urban traffic and the potential for data to inform safer road policies**.**

Q4 Which algorithm worked best?

The kernel density estimation, particularly the Parzen density algorithm, proved to be the most effective in visualizing and analyzing the changes in traffic accident patterns between the years of 2019 and 2020.

To highlight the differences:

- Visualizing Changes in Summer: We used Circle Marker to represent the number of accidents at each location. The area where multiple circle markers are overlapping is a potential hotspot for accidents. Thus we could visually identify changes in accident hotspots and the overall spread of incidents across Brooklyn.
- June 2019 vs. June 2020: To demonstrate the difference between these two months, I created two separate maps side by side, one for each year each containing Circle Markers.

The contrast between them showed a clear decrease in accident frequency during June 2020, likely due to the pandemic-related lockdowns.

- July 2019 vs. July 2020: Similarly, by comparing heat maps for July of each year, I could highlight the changes in traffic patterns. The July 2020 map showed not only a reduction in overall accidents but also a shift in where the most densely occurring accidents were located, which could be attributed to changes in traffic flow during the pandemic.

These visual comparisons provided a clear and immediate understanding of how accident patterns shifted between the two years, with the Parzen density algorithm effectively capturing the nuances in data distribution

Q5 What else would you like to share about the project?

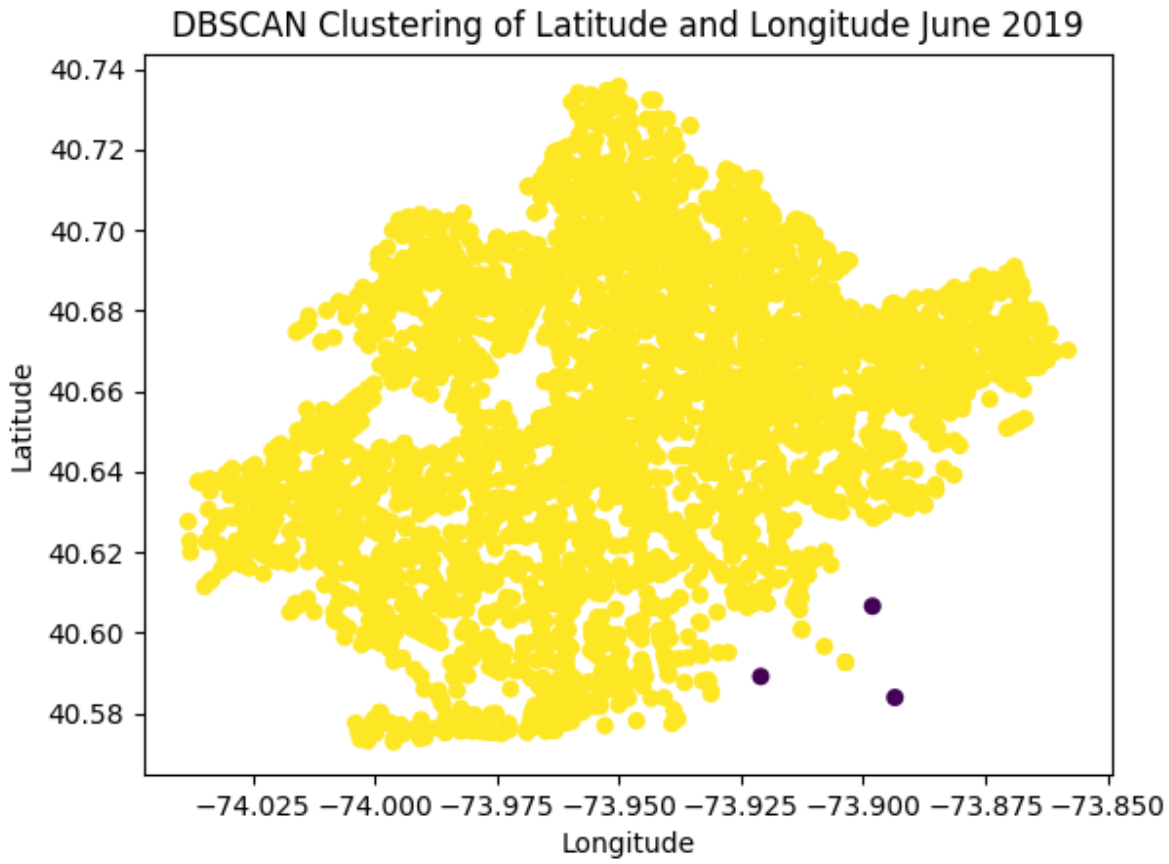
This project was a deep dive into the intricacies of urban traffic patterns and their susceptibility to a wide range of factors. The use of advanced data visualization techniques like heat maps offered a powerful tool for interpreting complex data in an intuitive manner. It also highlighted the importance of having robust and comprehensive data when conducting such analyses. Despite challenges with data completeness, the insights gained could be invaluable for policymakers and city planners focused on improving road safety. The project reinforced our appreciation for the potential of data analysis to make meaningful contributions to public health and safety.

Q6 Which algorithms did you finally use?

For finding the densities about the places with large number of accidents we used the Parzen Density Estimation as stated in the above question. The Parzen Density Estimation helped us in identifying the hotspots of accidents. We also use Apriori Algorithm to find the most common

types of vehicles involved in accidents and the Most Probable reason for accidents in the summer of 2019 and summer 2020. For both the summers we see that the most probable cause of accident is Driver Inattention/Distraction. We also notice that the most common type of accidents involved a Sedan, and a Station Wagon/Sport Utility Vehicle. We also applied association rules on the subsets we found from Apriori Algorithm. We found that a Pickup truck was most likely to have an accident with Sedan for both Summer 2019 and Summer 2020. Overall by applying Apriori and Association rules we see the most accidents involved a Sedan and a Station Wagon/Sport Utility Vehicle.

We also used clustering techniques like DBScan. But it didn't work for the Brooklyn Borough as all the points where accidents occurred were getting quantised into one cluster. Below is the result we got when applying DBScan to the Data



Q7 What went wrong, or what challenges did you face?

Challenges included handling incomplete or missing data, particularly in street names and accident details, which hindered precise location analysis. We were also not able to apply DBScan to cluster the data points. Instead we use Circle Markers to potentially identify the most common locations of Accidents for both years.

Q8 What was interesting about this?

The intersection of human behavior, environmental factors, and societal changes like the pandemic, and their tangible impact on traffic accident patterns, was particularly fascinating. It highlighted the dynamic nature of what might seem like random events but are actually deeply interconnected with daily life and larger societal shifts.

Q9 What else would you like to share about the project?

We would like to share that the project underscored the potential of data analysis in addressing real-world issues like road safety. Despite the data challenges, the insights gained were significant and could inform targeted, data-driven interventions to improve public safety.

Q10 What did you learn about data mining by doing this project?

Through this project, We learned the importance of thorough data preprocessing in data mining, especially dealing with real-world, imperfect datasets. It reinforced the value of using advanced visualization techniques to uncover hidden patterns and the necessity of contextual understanding for interpreting data mining results effectively.