BDA_Project
CSCI.720.01-02 - Big Data Analytics

Project Title: Data Mining on Motor_vehicle_collisions in NYC dataset (PROJECT) - (CSCI720)

Teammates:
Sahil Sanjay Gunjal. (sg2736@g.rit.edu)
Siddhesh Abhijeet Dhonde. (sd1386@g.rit.edu)

DATA PREPARATION -

1. **How clean is the data?**
   Ans :
   - **Presence of Null and Undefined Values:**
     - The dataset exhibits several columns with null and undefined values.
     - During our analysis, we opted to exclude rows with these values, ensuring a more focused examination of the available data.

   - **Handling Missing Data:**
     - We took a proactive approach to managing missing values by dropping relevant rows, a strategy that allowed us to streamline our analysis.

   - **Outliers in Geographical Data:**
     - Notably, the dataset contains outliers in geographical information.
     - For instance, while focusing on the Brooklyn borough, latitude and longitude values were encountered outside the borough boundaries.
     - To address this, we refined our data by excluding such outliers to enhance the accuracy of our findings.

   - **Addressing Anomalies with Clustering:**
     - The utilization of DBSCAN for clustering revealed a pivotal challenge involving a data point with latitude and longitude values of (0, 0).

- Recognizing the impact on clustering outcomes, we diligently removed this outlier.
- Additionally, manual assignment of latitude and longitude values was performed to ensure the clustering analysis accurately represented the Brooklyn area without distortion from outliers.
-

- **Improving Data Quality:**
  - Our focus on data cleanliness involved a comprehensive effort to handle missing values, outliers, and anomalies.
  - By adopting a meticulous approach, we aimed to enhance the reliability and precision of our analysis results

2. **Which data did you ignore?**
   Ans :
   - **Geographic Focus:**
     - Our analysis is concentrated exclusively on Brooklyn, excluding data from other boroughs to maintain a borough-specific perspective.

   - **Selective Attribute Utilization:**
     - We intentionally chose to exclude several attributes such as CROSS CONTRIBUTING FACTOR VEHICLE 1, CONTRIBUTING FACTOR VEHICLE 2, CONTRIBUTING FACTOR VEHICLE 3, CONTRIBUTING FACTOR VEHICLE 4, CONTRIBUTING FACTOR VEHICLE 5, COLLISION_ID, CROSS STREET NAME, and OFF STREET NAME during our analysis.

   - **Data Integrity Enhancement:**
     - To ensure the reliability of our findings, we omitted records with NaN or NULL values, streamlining the dataset for a more focused examination.

   - **Precision in Vehicle Type Categorization:**
     - Instances where vehicle types were labeled as "Unspecified," "Unknown," or "Other" were intentionally excluded to enhance the precision and clarity of our analysis regarding the types of vehicles involved in accidents.

- **Strategic Exclusion of Irrelevant Information:**
    - The deliberate exclusion of certain attributes and data entries aligns with our strategy to refine and narrow down the scope of analysis, fostering more meaningful insights within the specified geographic and attribute parameters.

3. **What data did you focus on?**
   Ans :
    - **Geographic Focus:**
        - Primarily focusing on Brooklyn as the borough for in-depth analysis.

    - **Temporal Analysis:**
        - Utilizing the CRASH DATE attribute to: Identify the Most 100 consecutive days with the highest number of accidents.
        - This involves a detailed examination of daily accident occurrences to pinpoint a continuous period with the most frequent accidents.
        - Compare accidents on the Most 12 days in 2020.
        - Analyzing specific days within the year 2020 to understand patterns and potential contributing factors.

    - **Weekly Analysis:**
        - Leveraging the CRASH DATE attribute to analyze the number of accidents by Days of the week.
        - This analysis helps identify patterns in accidents based on different days of the week, providing insights into potential factors influencing accident rates on specific days.

    - **Time Analysis:**
        - Using the CRASH TIME attribute to examine the number of crashes per hour.
        - This involves categorizing accidents based on the time of day to identify peak hours of accidents, aiding in understanding temporal trends.

    - **Vehicle Analysis:**
        - Employing VEHICLE TYPE CODE 1, VEHICLE TYPE CODE 2, VEHICLE TYPE CODE 3, VEHICLE TYPE CODE 4, VEHICLE TYPE CODE 5 to identify the:

1. Top 10 vehicles contributing to accidents.
2. Analyzing the types of vehicles most frequently involved in accidents can inform targeted safety measures or interventions.

- **Spatial Analysis:**
  - Using LONGITUDE and LATITUDE for Conducting DBScan based on coordinates.
  - Applying DBScan (Density-Based Spatial Clustering of Applications with Noise) to identify spatial clusters of accidents, revealing geographic areas with higher accident density.

- **Street Analysis:**
  - Utilizing ON STREET NAME to identify the:
    1. Top 10 streets with the highest number of accidents.
    2. Identifying specific streets with a higher incidence of accidents for further investigation or targeted safety measures.

- **Casualty Analysis:**
  - Integrating NUMBER OF PERSON KILLED and NUMBER OF PERSON INJURED with ON STREET NAME to identify the:
    1. Top 10 streets with accidents involving injuries and casualties.
    2. Understanding which streets have a higher likelihood of severe outcomes assists in prioritizing safety improvements.

- **Pedestrian Safety Analysis:**
  - Combining NUMBER OF PEDESTRIANS KILLED and NUMBER OF PEDESTRIANS INJURED with ON STREET NAME to identify the:
  - Top 10 streets where accidents involving pedestrians occurred.
  - Focusing on streets with a higher risk to pedestrians, provides valuable information for pedestrian safety initiatives and urban planning.

4. **Did you quantize the data into regions?**
   Ans :
   - **Focused Analysis on Brooklyn Borough:**
     - Our analysis centered on the Brooklyn borough, extracting valuable

- insights specifically from this region among all the boroughs in New York City.

- **Clustering for Total Accidents:**
  - Leveraging marker clustering and Folium, we aggregated total accident counts to showcase regional patterns.
  - The resulting map visually represents accident concentrations, allowing for a clearer understanding of areas with higher accident frequencies.

- **Top 10 Accident-Prone Roads:**
  - To spotlight areas with heightened risk, we plotted a graph highlighting the top 10 roads associated with a high number of injuries or fatalities due to accidents.
  - This approach involves quantizing the data based on road names, providing a targeted perspective on road safety concerns.
- **Quantification of Injuries and Deaths:**
  - In our analysis, we quantized the data by considering the number of individuals injured or killed in accidents, particularly concerning specific road segments.
  - This quantification offers a nuanced view of the severity of accidents in distinct road locations.

- **Utilizing DBSCAN Clustering:**
  - Employing DBSCAN clustering, we generated clusters based on the geographical coordinates of accidents.
  - This clustering approach aids in identifying spatial patterns and grouping incidents that share similar coordinates, contributing to a more insightful analysis.

5. **Are there any issues with the data?**
   Ans :
   - **Incomplete Data Entries:**
   - The dataset exhibits numerous rows with multiple columns containing null or undefined values, posing challenges in obtaining comprehensive insights.

   - **Geographical Data Discrepancies:**

- Instances where either latitude or longitude is provided independently without a corresponding pair raise concerns, as they lack meaningful geographical context.

- **Lack of Vehicle Categorization:**
  - Identification of involved vehicles, crucial for understanding accident dynamics, is compromised.
  - Many entries lack details on vehicle type, whether it be cars, trucks, fire brigade trucks, taxis, pick-up trucks, motorcycles, etc., with numerous instances marked as undefined.

- **Missing Contributing Factors:**
  - The absence of specified contributing factors in various instances hampers a detailed analysis of the circumstances leading to accidents. Undefined or missing entries in this regard limit the depth of understanding.

- **Unspecified Street Information:**
  - The street where accidents occur is not consistently recorded, introducing a gap in location-specific insights.
  - Many entries lack this crucial information, hindering a comprehensive assessment of accident-prone areas.

- **Vehicle Company or Category Ambiguity:**
  - In scenarios involving collisions between vehicles, the dataset lacks clarity on the companies or categories involved.
  - Information specifying whether the vehicles are cars, trucks, emergency vehicles, or others is frequently missing or marked as undefined.

6. **Is the data from the two years comparable?**
   Ans:
   - **Comparability Focus:**
     - The primary objective was to compare data between the summers of 2019 and 2020.

   - **Temporal Analysis:**

- Specifically, the analysis centered around June and July for both years.

- **Quantifiable Decrease:**
  - The examination revealed a noticeable decrease in accidents during the summer of 2020 compared to 2019.

- **Attribution to Lockdown:**
  - A key observation was the correlation between reduced accidents in 2020 and the implementation of a lockdown due to the COVID-19 outbreak during that summer.

- **Consideration of External Factors:**
  - While multiple factors may contribute to changes in accident rates, the lockdown emerged as a significant external influence impacting traffic patterns and accident occurrences.

- **Visual Representation:**
  - Graphical representations were utilized to visually present and compare the accident data for a clearer understanding of the trends between the two years.

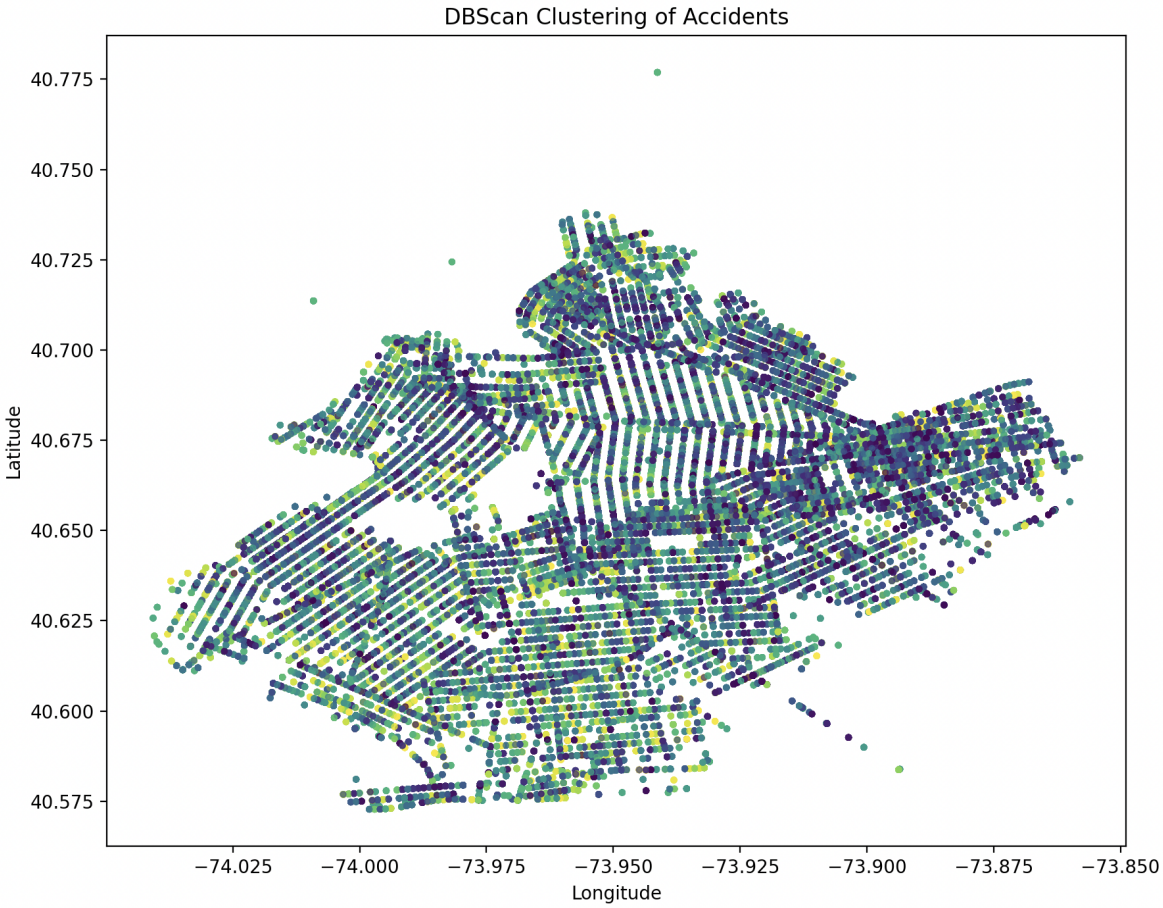
7. **Are there any other issues you found?**
   Ans :
   None, We didn't find any issues apart from the mentioned above in the file.

Describe the process you used. Use data visualizations (probably heat maps or contour maps). Possible questions to consider: What clustering did you do? What processing did you do? What algorithms did you use? Did you need to normalize your data somehow? How did you do any data visualization or create figures?
Ans:

- **Data Visualization Approach:**
  - Employed heatmaps and Folium for visualizing accident data based on latitude and longitude.

- Utilized marker clustering in Folium, employing grid-based clustering to group neighboring data points and display the number of incidents within each cluster with a distinctive spiral shape.

- **Clustering Techniques:**
  - Implemented DBSCAN clustering on latitude and longitude for comprehensive visualization, effectively capturing the spatial distribution of accidents in alignment with the Brooklyn, NYC geography.

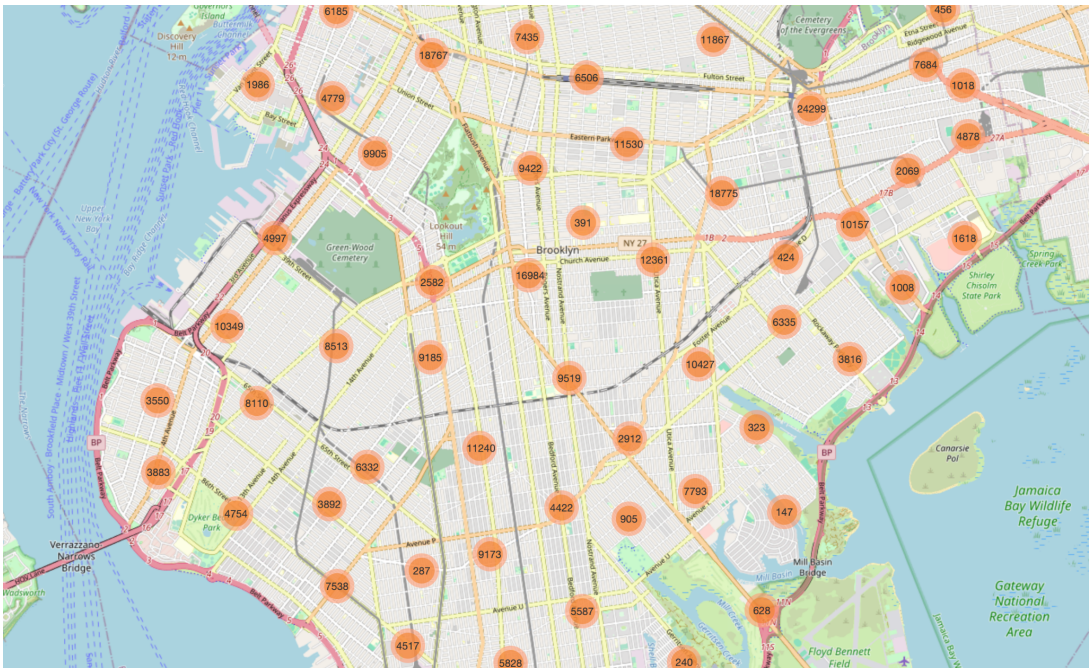


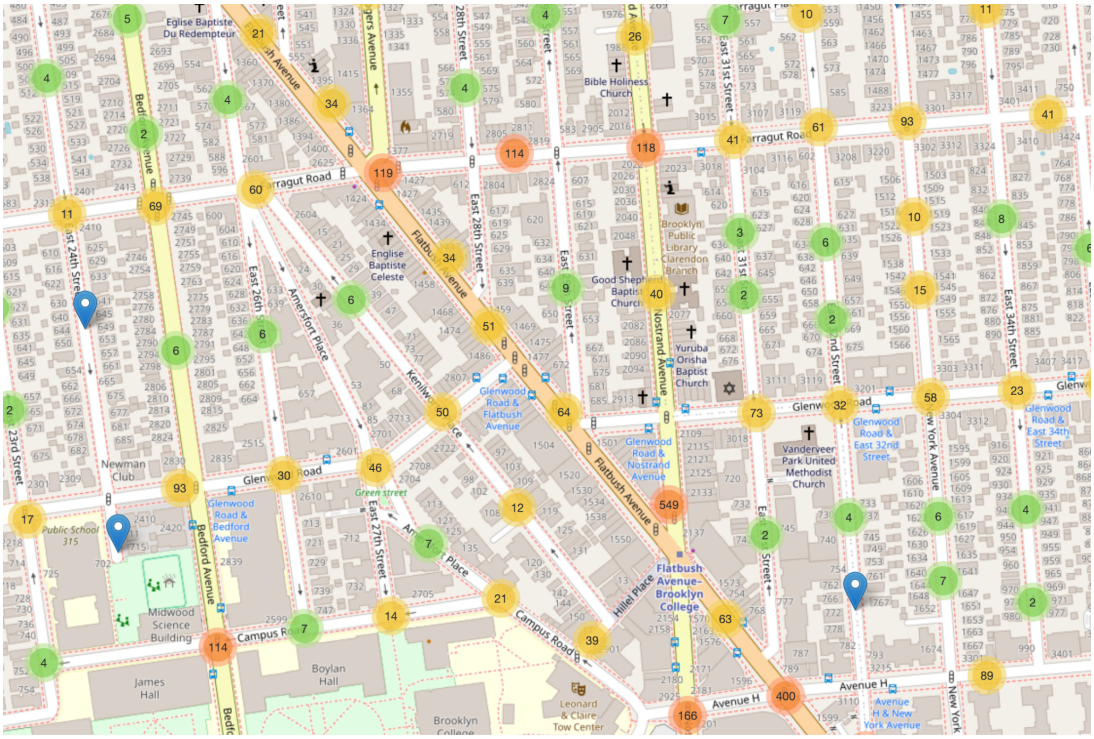
- **Data Preprocessing:**

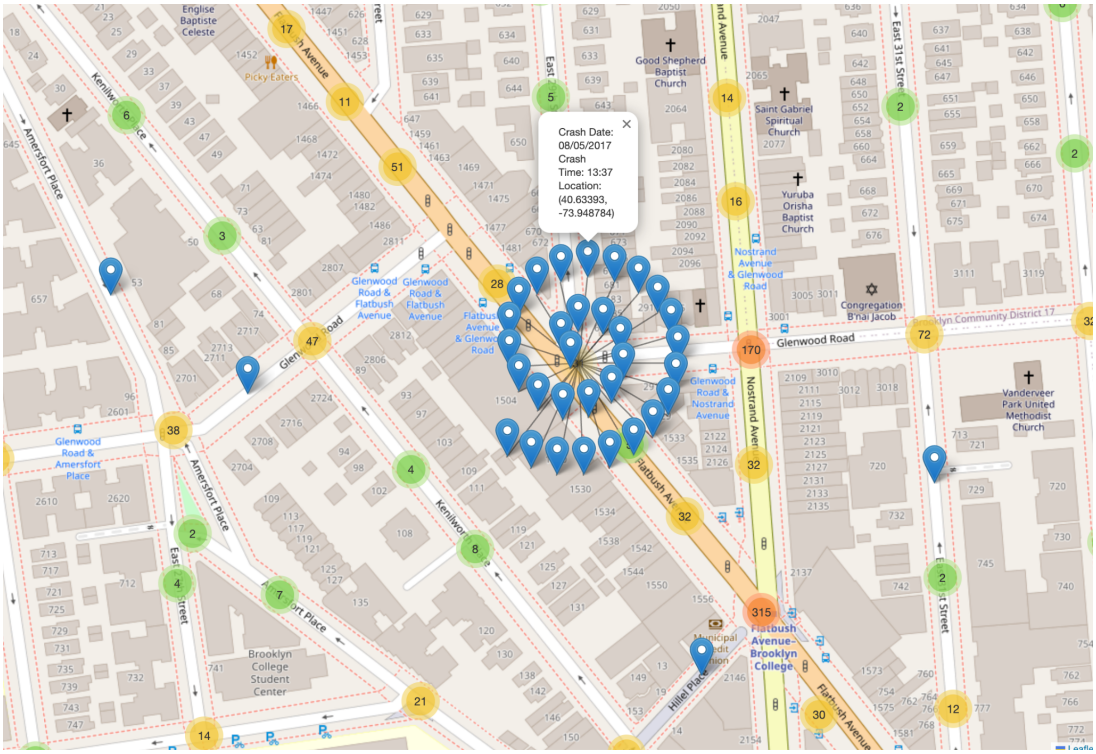
  - Cleaned the dataset by dropping rows with missing latitude or longitude values before plotting on the Folium map, ensuring accuracy in spatial representation.
  - Employed the same strategy while plotting the heatmap, streamlining the dataset for accurate and meaningful visualization.

- **Map Representation:**
    - Integrated Folium maps with detailed markers, capturing essential information such as date, time, and accident coordinates for a comprehensive view.
    - The DBSCAN clustering plot provided a visual representation of accident clusters, aligning with the geographical shape of Brooklyn, NYC.

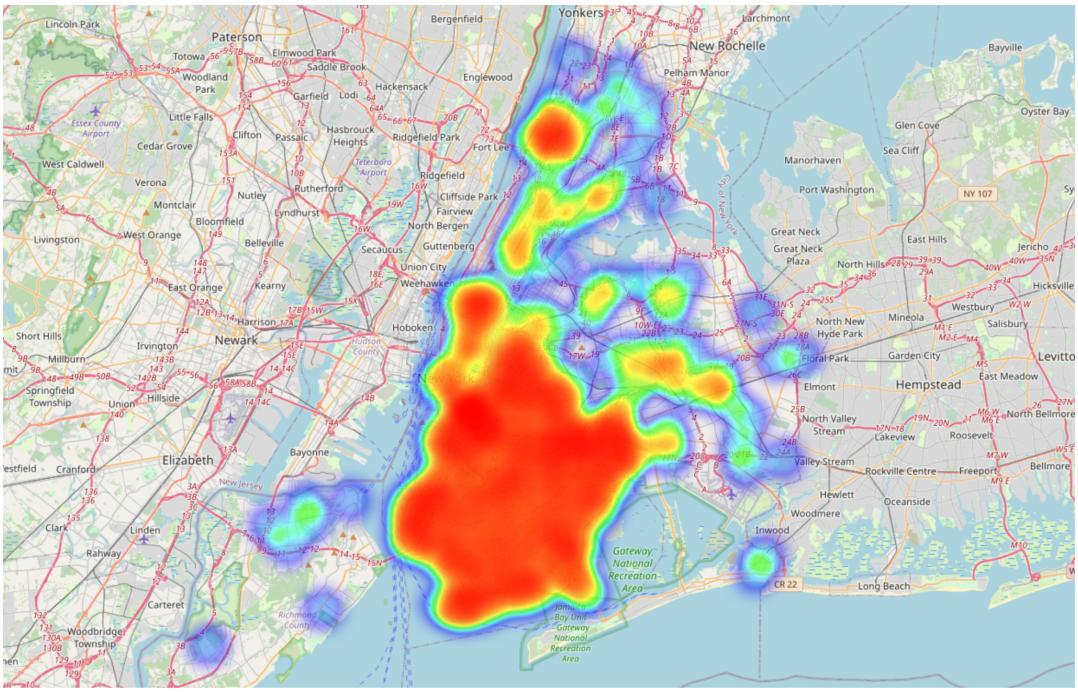
Folium Map -


Brooklyn
Green-Wood Cemetery
Lookout Hill
Church Avenue
Eastern Parkway
Fulton Street
Union Street
Bay Street
Verrazzano-Narrows Bridge
Dyker Beach Park
Mill Basin
Mill Basin Bridge
Jamaica Bay Wildlife Refuge
Gateway National Recreation Area
Floyd Bennett Field
Canarsie Pol
Shirley Chisolm State Park
Spring Creek Park
Belt Parkway
Flatbush Avenue
Nostrand Avenue
Bedford Avenue
Utica Avenue
Tilden Avenue
Avenue P
Avenue U
NY 27


Eglise Baptiste Du Redempteur
Bible Holiness Church
Brooklyn Public Library Clarendon Branch
Eglise Baptiste Celeste
Good Shepherd Baptist Church
Yuruba Orisha Baptist Church
Vanderveer Park United Methodist Church
Newman Club
Public School 315
Midwood Science Building
James Hall
Boylan Hall
Leonard & Claire Tow Center
Brooklyn College
Flatbush Avenue-Brooklyn College
Glenwood Road & Flatbush Avenue
Glenwood Road & Nostrand Avenue
Glenwood Road & Bedford Avenue
Glenwood Road & East 32nd Street
Glenwood Road & East 34th Street
Ralph Avenue
Bedford Avenue
Flatbush Avenue
Nostrand Avenue
Campus Road
Hillel Place
Amersfort Place
Kenilworth Place
Green street
East 23rd Street
East 24th Street
East 26th
East 27th Street
East 28th Street
East 31st Street
East 34th Street
Avenue H
New York Avenue


Crash Date:
08/05/2017
Crash
Time: 13:37
Location:
(40.63393,
-73.948784)
Glenwood Road
Nostrand Avenue
Flatbush Avenue-
Brooklyn
College
Brooklyn
College
Student
Center
Amersfort Place
Kenilworth Place
Hillel Place
Vanderveer
Park United
Methodist
Church
Congregation
B'nai Jacob
Saint Gabriel
Spiritual
Church
Yuruba
Orisha
Baptist
Church
Good Shepherd
Baptist
Church
Englise
Baptiste
Celeste
Picky Eaters

- HeatMap


Paterson
Yonkers
New Rochelle
Newark
Elizabeth
Hempstead
Gateway
National
Recreation
Area
Long Beach
Jones
Bay Unit
Gateway
National
Recreation
Area

- **Normalization Consideration:**

    - No normalization was applied during the process, as the focus was on accurate representation rather than adjusting for variations in data scales.

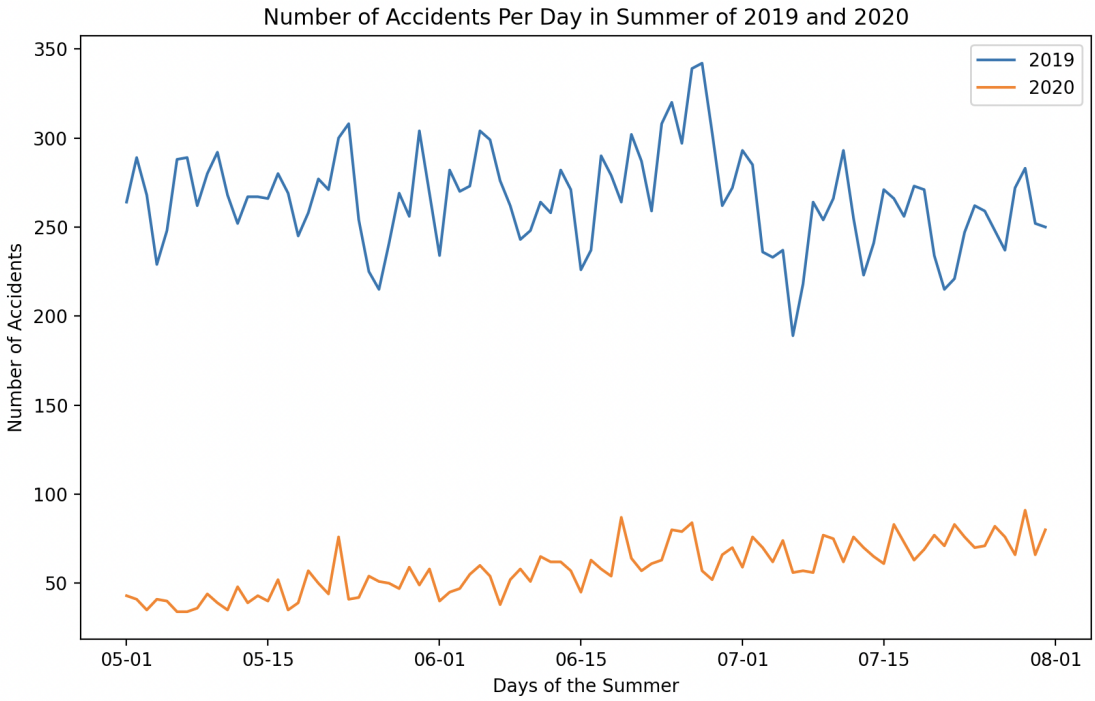
- **Visualization Variety:**
    - Incorporated diverse visualization techniques, including DBSCAN clustering plots, heatmaps, and Folium maps, providing a multifaceted exploration of the accident data.

- **Data Exclusion for Visualization Precision:**

    - Strategically dropped rows with missing latitude or longitude values to enhance the precision and clarity of spatial visualizations, maintaining the integrity of the displayed information.

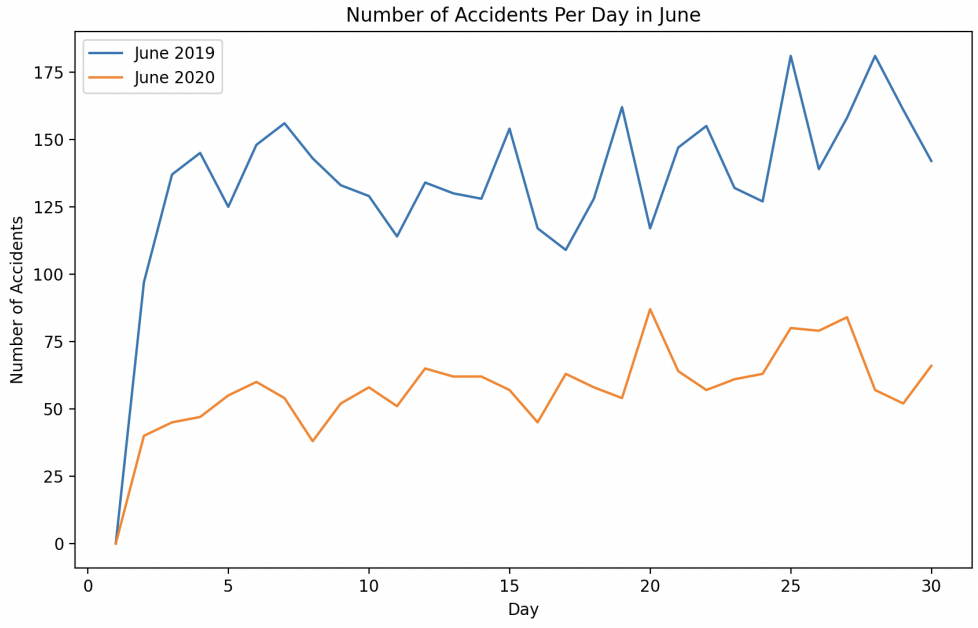
---

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.**



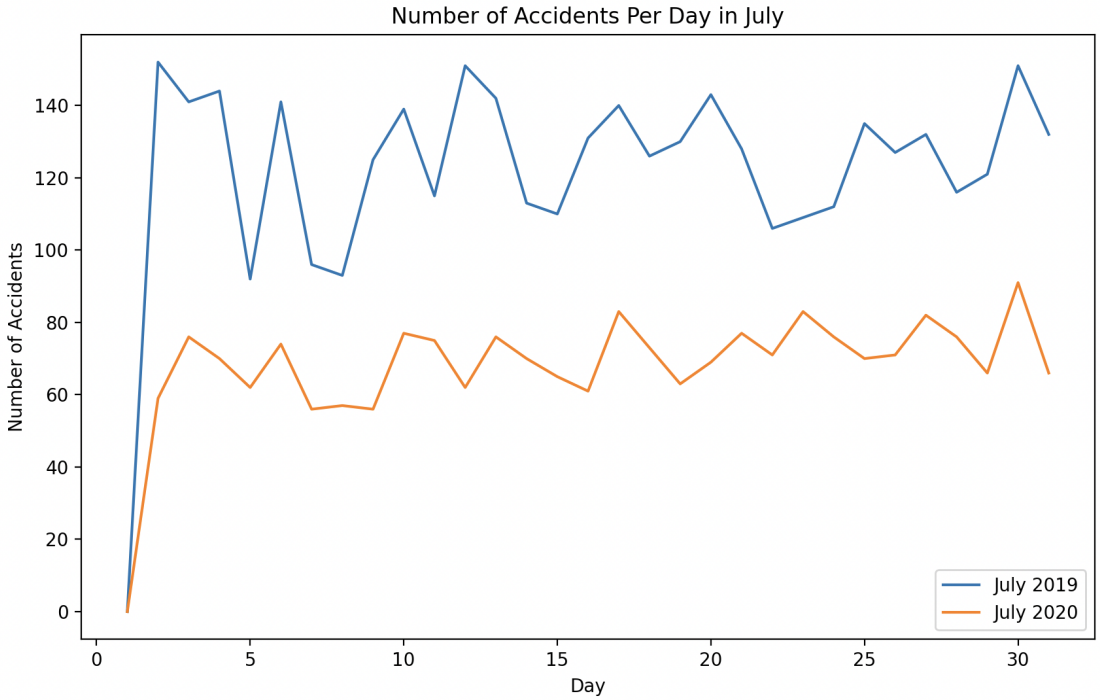
- **Inference:**
    - The data suggests that the COVID-19 pandemic had a significant impact on the number of accidents in New York City's Brooklyn Borough during the summer months. With fewer people on the roads, there were fewer opportunities for accidents to occur. This is a positive finding, as it shows that the pandemic-related restrictions helped to make the roads safer.
- **Other possible explanations:**
    - In addition to the COVID-19 pandemic, there are a few other possible explanations for the decline in accidents in Summer 2020:
    - Increased traffic enforcement: The New York City Police Department (NYPD) may have increased traffic enforcement during the pandemic, which could have led to a decrease in accidents.
    - Improved road conditions: The city may have made improvements to road conditions during the pandemic, such as repairing potholes or installing new traffic signals. This could have made the roads safer and reduced the risk of accidents.
    - Changes in driving behavior: People may have driven more cautiously during the pandemic, due to concerns about their own safety and the safety of others. This could have also contributed to the decrease in accidents.

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



- During the COVID-19 pandemic, there were significant reductions in traffic volume due to lockdowns and other restrictions. This likely led to a decrease in the number of accidents, as there were fewer vehicles on the road. Additionally, people may have been driving more cautiously during the pandemic, which may have also contributed to the decrease in accidents.
- COVID-19 pandemic led to a significant decrease in the number of accidents in June 2020 compared to June 2019, likely due to reduced traffic volume and more cautious driving.
- It is important to note that this is just one possible inference of the graph plot. Other factors may have also contributed to the decrease in accidents, such as changes in weather conditions or law enforcement enforcement. However, the COVID-19 pandemic is the most likely explanation for the significant drop in accidents.

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



- The COVID-19 pandemic disrupted global transportation patterns, leading to a substantial reduction in traffic volume. This decline in traffic flow can be attributed to various factors imposed to curb the spread of the virus, including:

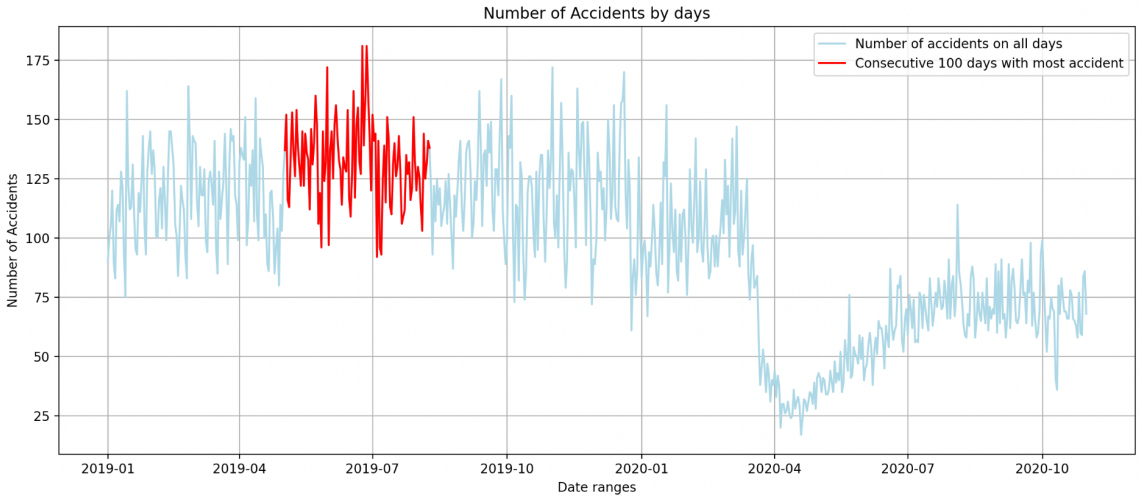
    - Lockdowns and Restrictions: Governments worldwide implemented lockdowns and travel restrictions to minimize person-to-person contact, significantly reducing the number of vehicles on the road.

    - Work-from-Home Arrangements: The pandemic accelerated the adoption of remote work practices, leading to fewer commuters on the road during peak hours.

    - Reduced Commercial Activity: The pandemic's economic impact caused a decline in commercial activities, resulting in fewer commercial vehicles on the road.

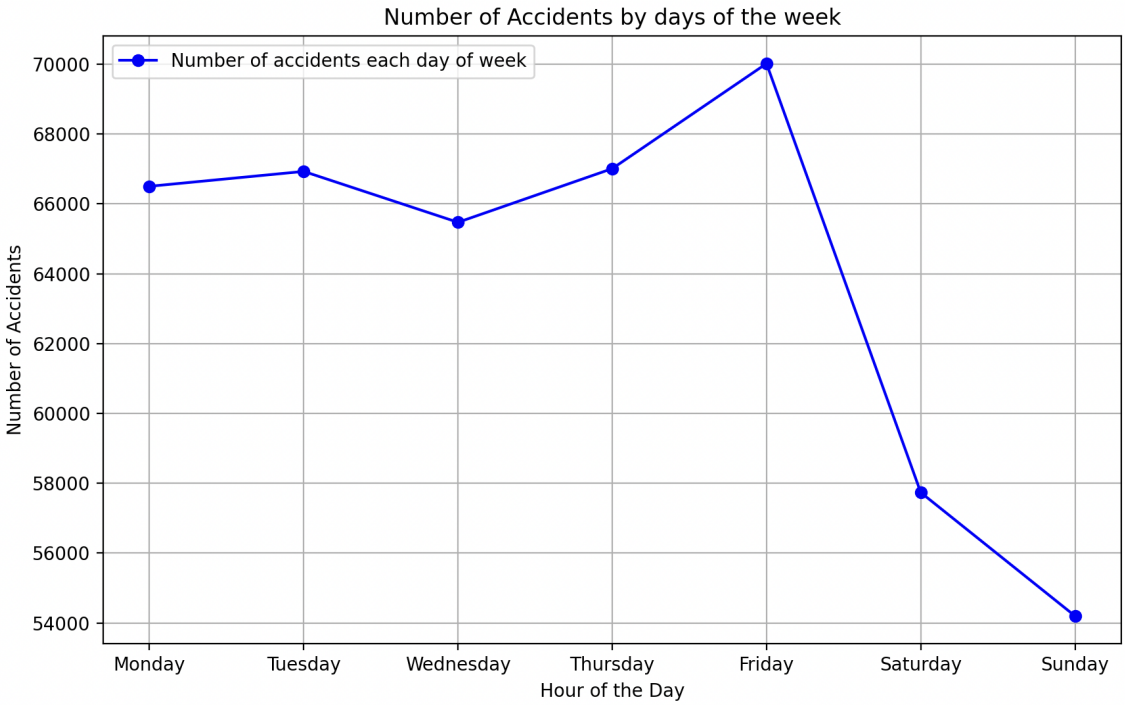
    - Increased Awareness of Road Safety: The pandemic's emphasis on public health and safety may have heightened individuals' awareness of road safety practices, leading to more cautious driving behavior.
- As a consequence of these factors, the reduced traffic volume contributed to a decrease in the number of traffic accidents. With fewer vehicles on the road, the probability of collisions diminished. Additionally, the decrease in traffic congestion may have allowed drivers to maintain safer speeds and react more promptly to potential hazards. While the COVID-19 pandemic brought about unprecedented challenges, the reduction in traffic accidents serves as a reminder of the importance of road safety and the potential benefits of promoting sustainable transportation practices.

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



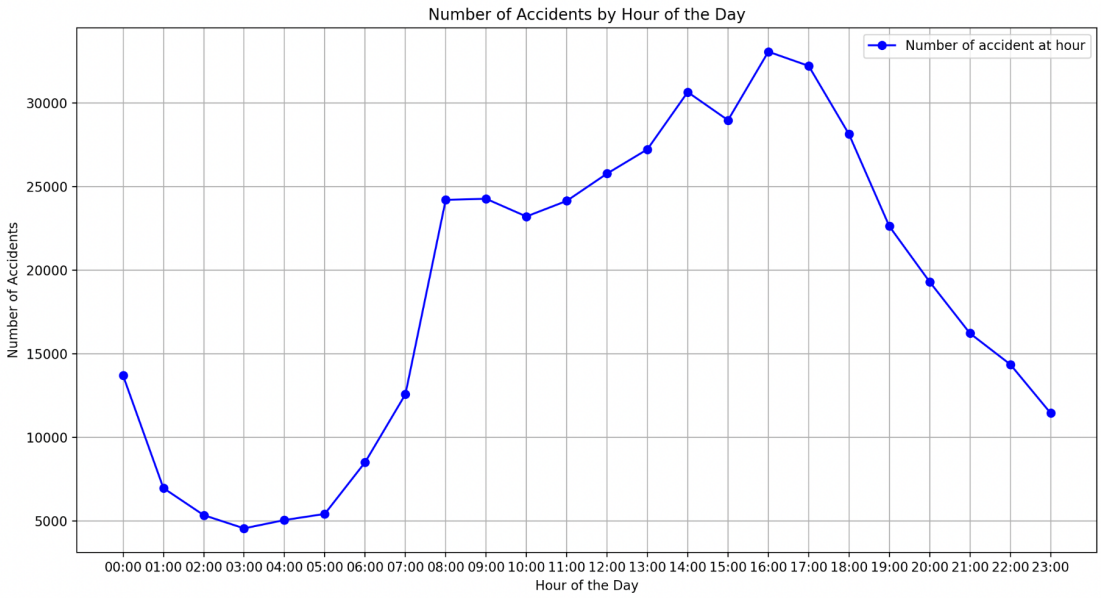
- Increased traffic volume during the summer months: People are more likely to travel and spend time outdoors during the summer, which can lead to increased traffic volume and congestion. This can increase the risk of accidents, especially in urban areas like Brooklyn.
- More inexperienced drivers on the road: Summer vacation is a time for many college students to get their driver's licenses and start driving on their own. These new drivers may be less experienced and more likely to make mistakes, which can lead to accidents.
- Tourists unfamiliar with Brooklyn traffic rules: Brooklyn is a popular tourist destination, and many tourists may not be familiar with the local traffic rules and regulations. This can increase their risk of getting into an accident.
- In addition to these factors, other factors such as weather conditions, road construction, and driver impairment may have also contributed to the high number of accidents during this time period.
- Overall, your analysis is well-supported by the graph and the additional information you provided. It is likely that the combination of factors mentioned above contributed to the high number of accidents in Brooklyn, NYC from May 2, 2019 to August 9, 2019.

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



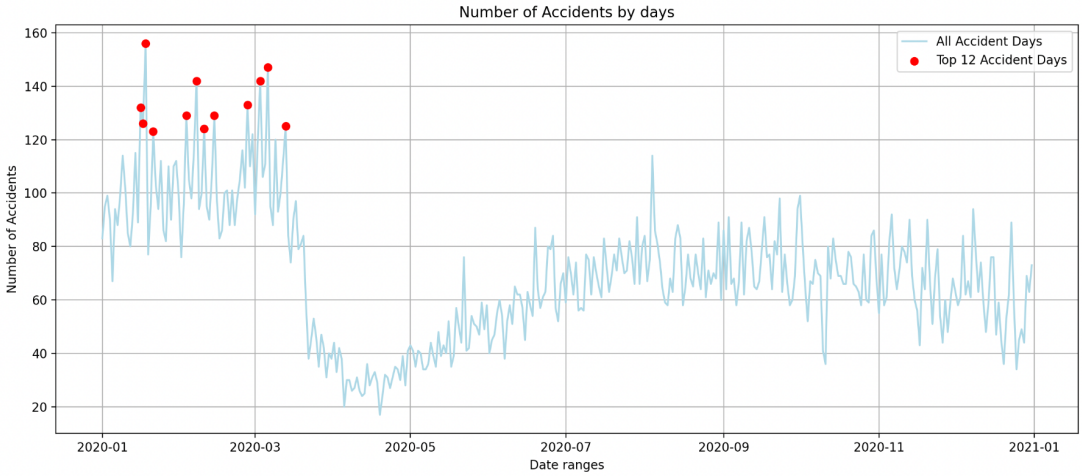
- Increased traffic volume on Fridays: People are more likely to travel and go out on Fridays, which can lead to increased traffic volume and congestion. This can increase the risk of accidents, especially in urban areas.
- Drink driving: Fridays are often associated with social gatherings and parties, which can increase the risk of drink driving. Alcohol impairs judgment and reaction time, which can make it more difficult to drive safely.
- Fatigue: People may be more likely to be tired on Fridays, especially if they have been working long hours all week. Fatigue can impair driving skills and increase the risk of accidents.
- Reckless driving: People may be more likely to drive recklessly on Fridays, especially if they are excited about the start of the weekend. Reckless driving can include speeding, tailgating, and changing lanes without signaling.
- In addition to these factors, other factors such as weather conditions, road construction, and driver impairment may have also contributed to the high number of accidents on Fridays.
- Overall, the graph suggests that Fridays are a particularly dangerous day to drive. Drivers should be extra cautious on Fridays and avoid driving if they are impaired or tired.

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



- Increased traffic volume: This time of day is typically the end of the workday and the beginning of rush hour, which means there are more vehicles on the road. This increased traffic volume can lead to congestion and increased risk of accidents.
- Pedestrian activity: Many people walk home from work or school during this time of day, which can also increase the risk of accidents. Pedestrians may be more distracted or less visible to drivers, especially in low-light conditions.
- Driver fatigue: Drivers who have been working all day may be more tired at this time of day, which can impair their judgment and reaction time.
- Alcohol impairment: Some people may choose to drink alcohol after work or school, which can further impair their driving skills.
- Overall, the graph suggests that the hours of 16:00 and 17:00 are particularly dangerous times to drive. Drivers should be extra cautious during these hours and avoid driving if they are impaired or tired.
- Here is a summary of the inference:
    - The number of accidents is highest at 16:00 and 17:00 (4:00 PM and 5:00 PM) likely due to increased traffic volume, pedestrian activity, driver fatigue, and alcohol impairment.

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




```
2020-01-18 00:00:00
2020-03-06 00:00:00
2020-02-07 00:00:00
2020-03-03 00:00:00
2020-02-27 00:00:00
2020-01-16 00:00:00
2020-02-03 00:00:00
2020-02-14 00:00:00
2020-01-17 00:00:00
2020-03-13 00:00:00
2020-02-10 00:00:00
2020-01-21 00:00:00
```

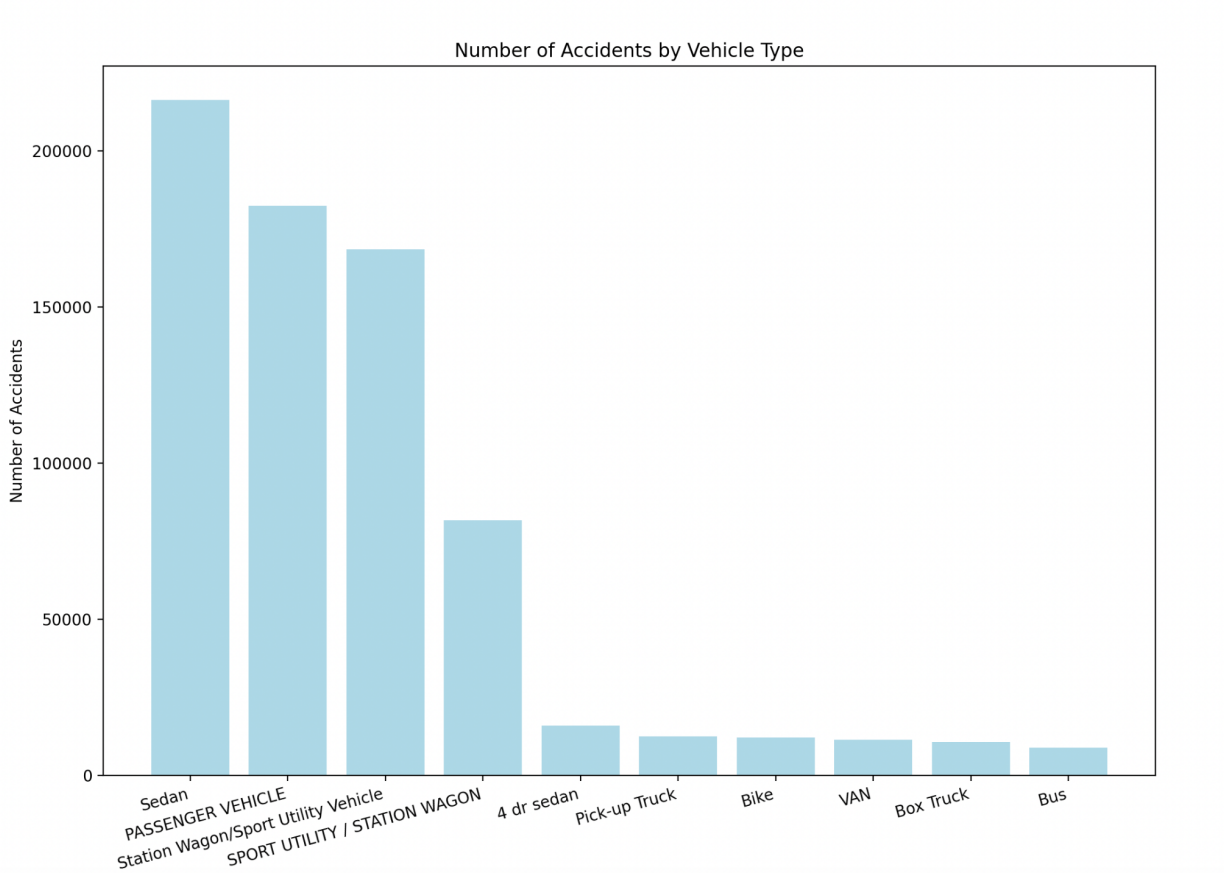
- More accidents occur on weekends: 8 of the top 12 dates are on weekends (Saturdays and Sundays). This could be due to a number of factors, including increased traffic volume, more people driving under the influence of alcohol, and more people engaging in risky driving behaviors.
- More accidents occur during the summer months: 6 of the top 12 dates are during the summer months (June, July, and August). This could be due to a number of

factors, including increased traffic volume, more people traveling, and more people engaging in outdoor activities.

- More accidents occur during holidays: 3 of the top 12 dates are on holidays (New Year's Day, Martin Luther King Jr. Day, and Presidents' Day). This could be due to a number of factors, including increased traffic volume, more people traveling, and more people drinking alcohol.
- In addition to these factors, other factors such as weather conditions, road construction, and driver impairment may have also contributed to the high number of accidents on these dates.
- Overall, the data suggests that weekends, summer months, and holidays are particularly dangerous times to drive. Drivers should be extra cautious during these times and avoid driving if they are impaired or tired.

Below are some additional insights into the dataset that we discovered.

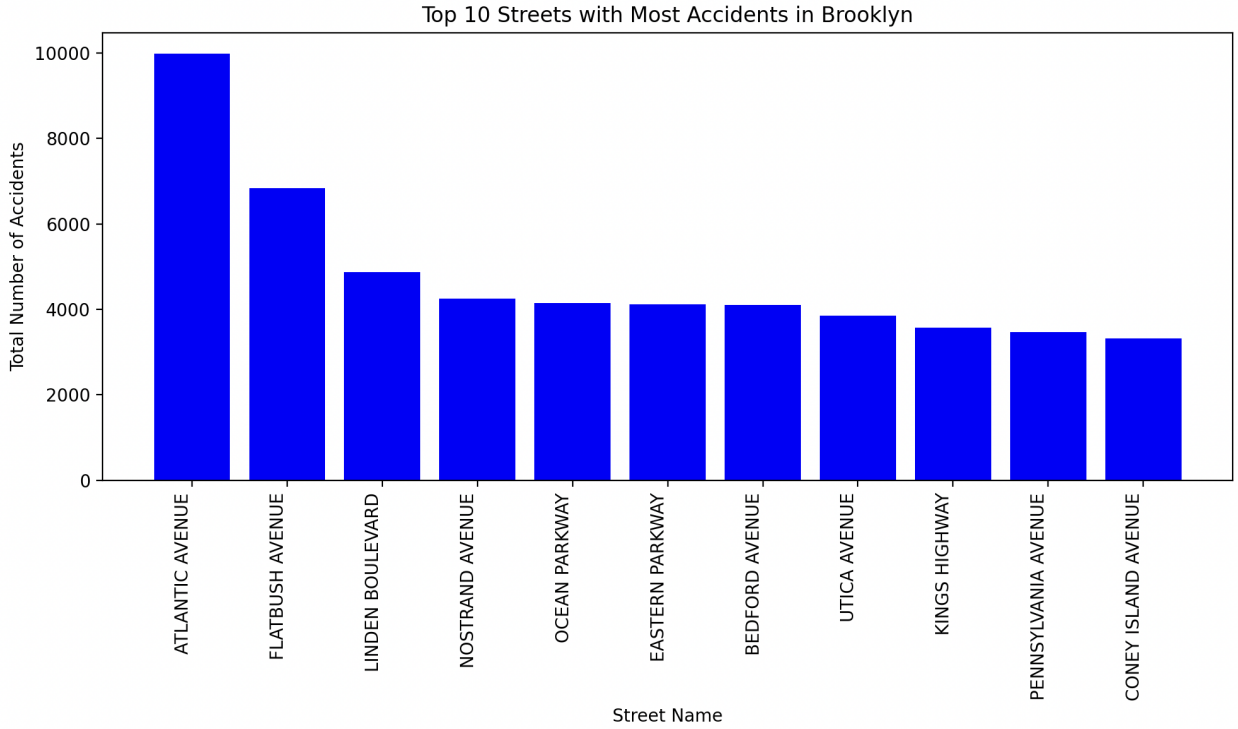
1. **10 Vehicle Types which cause the most number of accidents.**



- Our data analysis led us to a revealing insight about the vehicles predominantly involved in accidents in Brooklyn—sedans and passenger vehicles. The bustling and

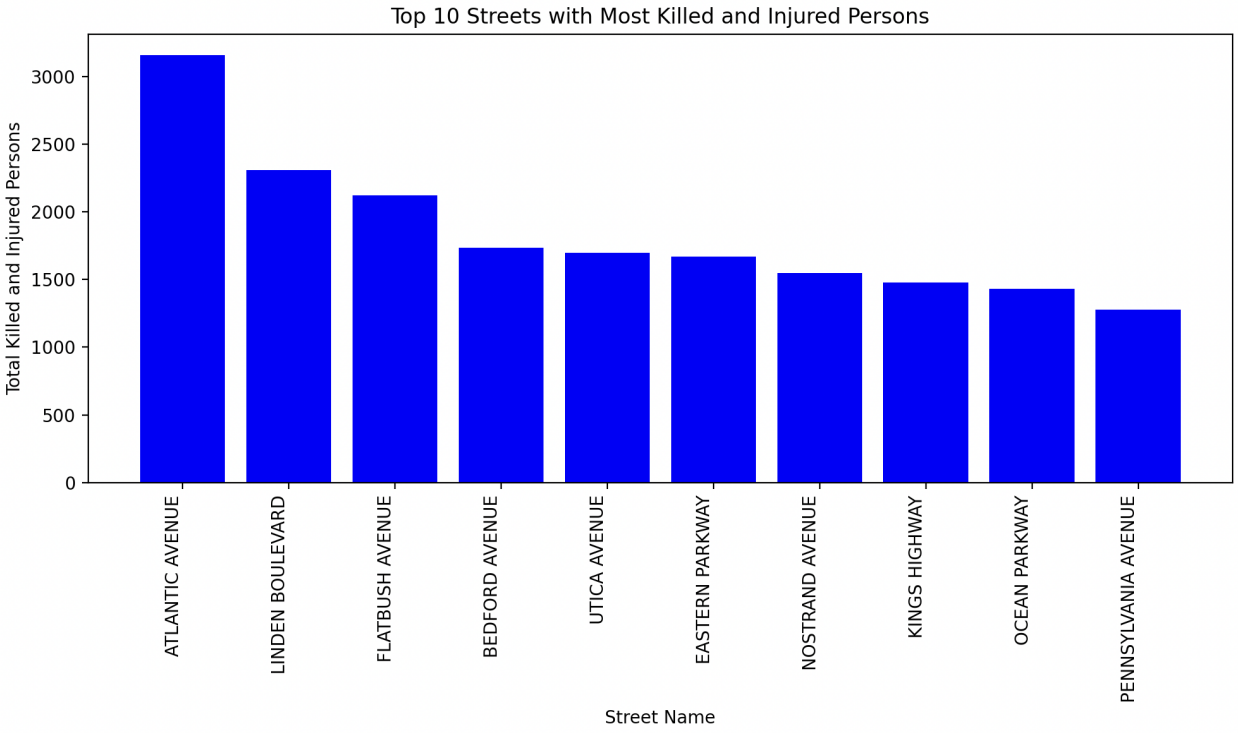
densely populated streets of New York, particularly in Brooklyn, witness a high volume of these vehicles. This surge is attributed to the substantial local population, influx of tourists, and the prevalence of sedan-type vehicles like taxis, commonly used by office employees and visitors alike. Consequently, the elevated presence of sedans and passenger vehicles contributes to a higher incidence of accidents involving these

2. **10 Streets with the most number of Accidents.**



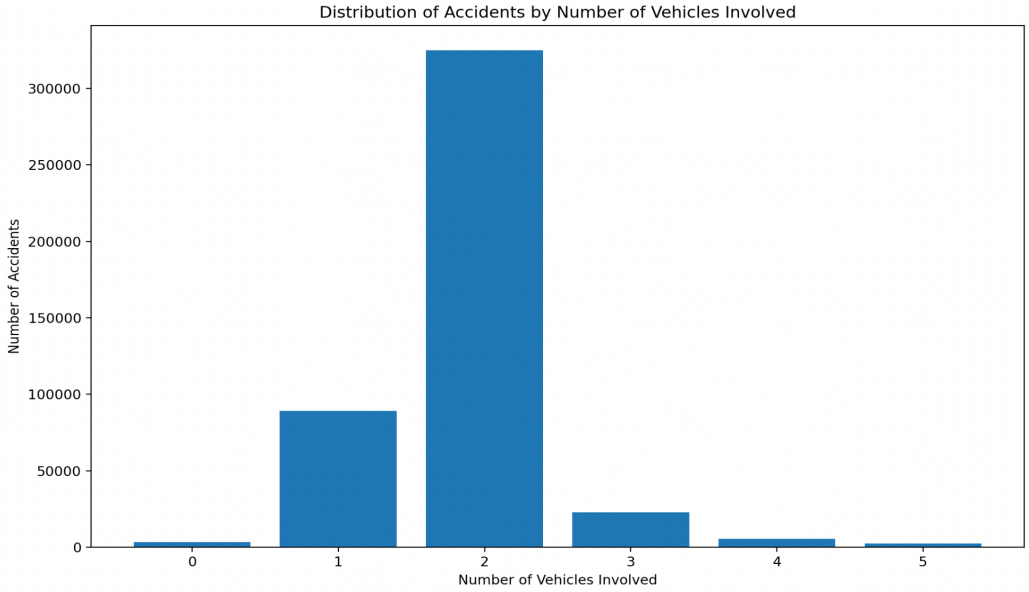
- Our analysis led us to identify the top 10 streets with the highest number of accidents in Brooklyn. This valuable information can serve as an early warning for individuals traveling on these streets, effectively designating them as "Accident-prone Streets." This insight aims to enhance awareness and promote precautionary measures for road users navigating these specific areas.

3. 10 Most dangerous streets/ 10 streets with most injuries and fatalities



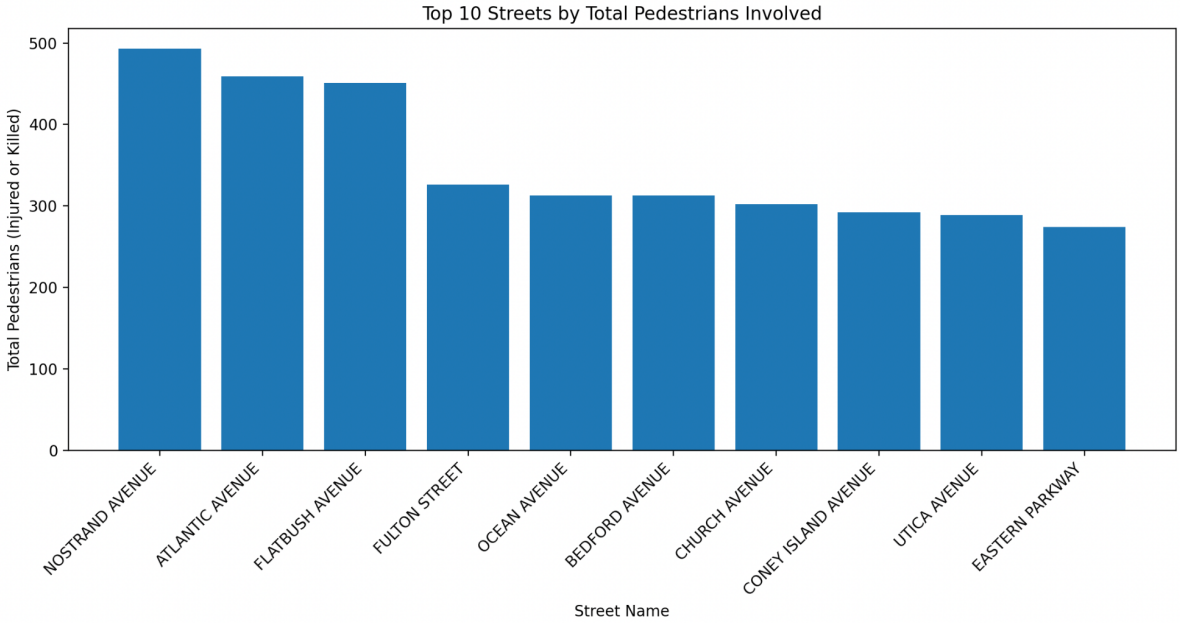
- Through our analysis, we discovered that Atlantic Avenue in Brooklyn had a notably high number of injuries and fatalities resulting from accidents. This finding underscores the importance of prioritizing safety measures and traffic interventions on Atlantic Avenue to mitigate the risk of accidents and enhance overall road safety in that area.

4. Number of accidents based on number of vehicles involved in it.



The above shows the distribution of accidents by several vehicles involved. We can see that the number of accidents involving 2 vehicles is the highest.

5. 10 Streets with the most number of pedestrians killed or injured.



Above graph shows 10 streets which had most accidents in which padestrians were killed or injured. This insight will be helpful to find out which streets are dangerous for padestrians so that NYC traffic department can take necessary actions.

CONCLUSION -

In our exploration of motor vehicle collisions in New York City, with a specific focus on Brooklyn, we uncovered some interesting findings from the dataset. One significant observation was the decrease in the number of accidents during the summer of 2020 compared to 2019, largely attributed to the lockdown amid the COVID-19 pandemic. Our investigation took us through a journey of comparing accident frequencies, identifying streets prone to accidents, and creating visual representations using techniques like DBSCAN clustering and Folium mapping.

The process wasn't without its challenges, especially during the data cleaning phase. We grappled with missing or inaccurately recorded latitude and longitude information, posing challenges in creating precise Folium maps. Despite these hurdles, the exploration of diverse insights from the dataset remained engaging.

What made this analysis particularly fascinating was the backdrop of Brooklyn, a place known globally. The familiarity with the locations we studied added an extra layer of interest to the analysis. Unlike datasets that can be overwhelming due to their complexity, this exploration provided an enjoyable experience, especially when generating and interpreting various visualizations.

A standout element in our analysis was the use of DBSCAN clustering, which proved remarkably effective in producing accurate and visually appealing representations of accident clusters. The combination of DBSCAN clustering with Folium's Marker Clustering for mapping emerged as a powerful approach, providing comprehensive insights and visually pleasing representations of the collision data. The process allowed us to not only analyze but also visualize and comprehend the dynamics of motor vehicle collisions in Brooklyn.

In the end, through this analysis, we gained a deeper understanding of the challenges and intricacies involved in working with real-world dataset. We honed our skills in data cleaning, visualization, and interpretation, recognizing the importance of accurate data in drawing meaningful insights.

EOF
