Data Mining Project – Fall 2023: BRONX BOROUGH

CSCI - 720

Team

Anurag Kallurwar (ak6491), Vishal Panchidi (vp8760)

Data Preparation

The data appears to have missing values in various columns such as "BOROUGH," "ZIP CODE," "LATITUDE," "LONGITUDE," and others. The missing data has been handled in the code by dropping the NAN and missing values. In terms of outliers, have checked and found out that there are a lot of error locations and wrong locations described as Bronxs, but we handled these outliers while processing the charts and ignoring the values that fall outside the region i.e. Bronx. Below is the data and its analysis shown. With the help of this data, we made decisions to ignore outliers and keep data which is important. Eg. pd.describe() gave the minimum, and maximum of the latitude and longitude, which helped us to understand the potential outliers in the data and the need to clean it. The problem of duplicate data is also present in the dataset, we found that 24 values had duplicate data which could be ignored as the data set is pretty large.

```
bronx_df.describe()
```

	LATITUDE	LONGITUDE
count	35227.000000	35227.000000
mean	40.787519	-73.772866
std	1.568478	2.836679
min	0.000000	-74.005790
25%	40.825912	-73.905594
50%	40.844420	-73.886940
75%	40.869001	-73.858770
max	40.912884	0.000000

```
checking_NaN(bronx_df)
```

```
CRASH DATE - 0
CRASH TIME - 0
BOROUGH - 0
ZIP CODE - 2
LATITUDE - 1516
LONGITUDE - 1516
LOCATION - 1516
ON STREET NAME - 14865
CROSS STREET NAME - 14868
OFF STREET NAME - 21878
NUMBER OF PERSONS INJURED - 0
NUMBER OF PERSONS KILLED - 0
NUMBER OF PEDESTRIANS INJURED - 0
NUMBER OF PEDESTRIANS KILLED - 0
NUMBER OF CYCLIST INJURED - 0
NUMBER OF CYCLIST KILLED - 0
NUMBER OF MOTORIST INJURED - 0
NUMBER OF MOTORIST KILLED - 0
CONTRIBUTING FACTOR VEHICLE 1 - 229
CONTRIBUTING FACTOR VEHICLE 2 - 8792
CONTRIBUTING FACTOR VEHICLE 3 - 34269
CONTRIBUTING FACTOR VEHICLE 4 - 36080
CONTRIBUTING FACTOR VEHICLE 5 - 36551
COLLISION_ID - 0
VEHICLE TYPE CODE 1 - 436
VEHICLE TYPE CODE 2 - 11135
VEHICLE TYPE CODE 3 - 34373
VEHICLE TYPE CODE 4 - 36099
VEHICLE TYPE CODE 5 - 36552
```

```
duplicates = bronx_df[bronx_df.duplicated(subset=bronx_df.columns.difference(["COLLISION_ID"]), keep=False)]
print("Number of Duplicate Accident Information: " + str(len(duplicates)))
```

```
Number of Duplicate Accident Information: 24
```

In this project, the focus is on change detection in traffic accidents within the Bronx borough during the summer months of June and July for the years 2019 and 2020. To achieve this, a subset of relevant columns has been selected, and specific data filtering has been applied. Data outside the specified months (June and July) and years (2019 and 2020) has been ignored for the analysis. All columns except 'CRASH DATE', 'CRASH TIME', 'BOROUGH', 'ZIP-CODE', 'LATITUDE', 'LONGITUDE', 'LOCATION', 'NUMBER OF PERSONS KILLED', 'NUMBER OF PEDESTRIANS INJURED', 'NUMBER OF PEDESTRIANS KILLED' has been ignored for analysis.

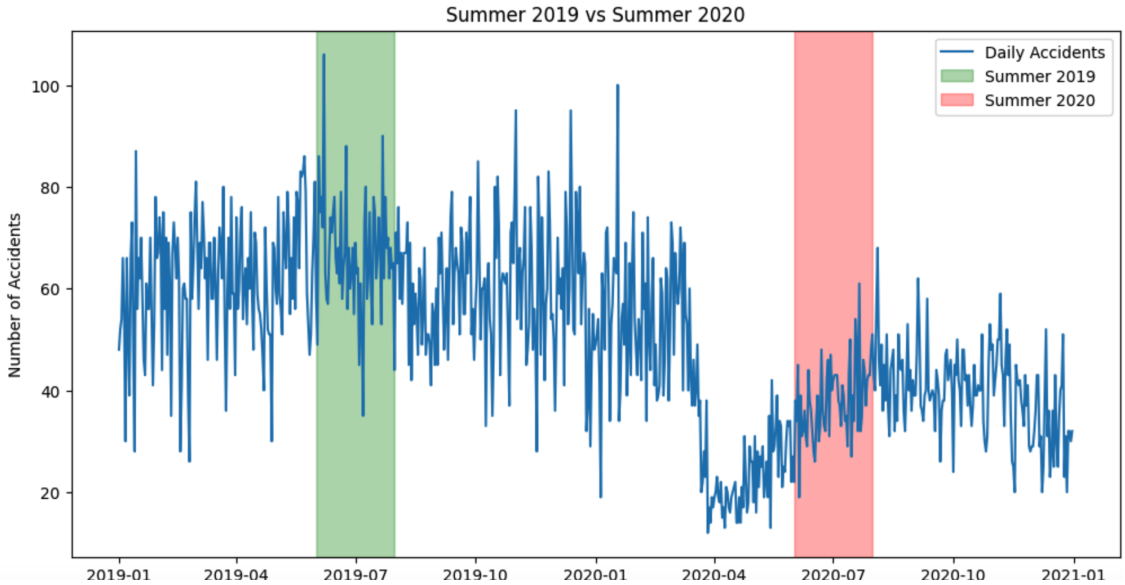
We looked at data for the Bronx during the summer months of June and July in 2019 and 2020, following Professor Dr. Kinsman's instructions. By concentrating on this specific time and place, we tried to understand how traffic accidents changed. We paid attention to things like injuries, fatalities, and the total number of accidents. This helped us see if there were any noticeable patterns in how traffic safety evolved in the Bronx during these summer months.

We did not quantize the data as we deemed it unnecessary for our analysis. Our dataset comprises individual accident records, each uniquely identified by a collision ID, which is represented by a continuous range of values. Each collision ID serves as a unique identifier for a specific accident, eliminating the need for quantization explicitly. This unique identification system enables us to maintain the granularity of individual accidents within the dataset, facilitating a detailed examination of each incident without the need for further discretization.

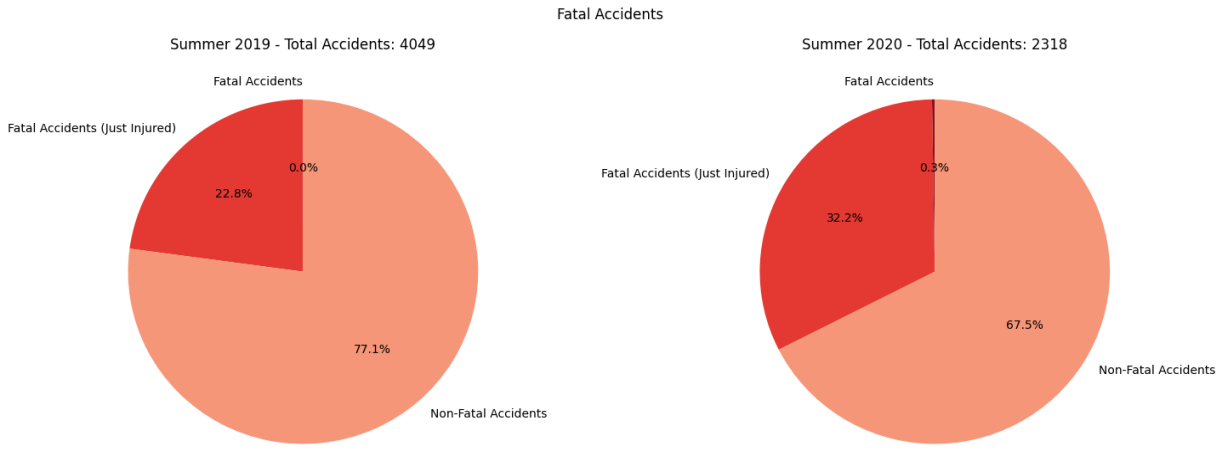
Data from 2019 and 2020 are comparable. The dataset includes essential information such as crash date, time, borough, latitude, longitude, and contributing factors consistently for both years. Ensure that the entire dataset for complete and consistent in key identifiers and data entries.

Data Analysis

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.

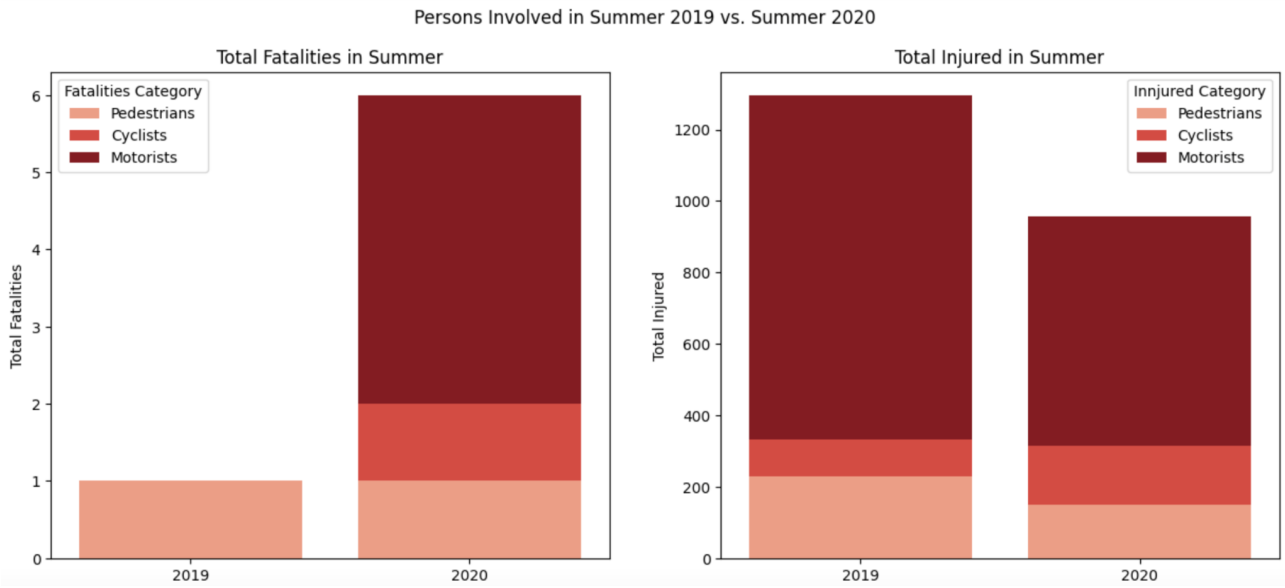


The analysis reveals a notable drop in traffic accidents during the summer months from 2019 to 2020 in the Bronx borough. While the total number of accidents decreased significantly from 4049 in the Summer of 2019 to 2318 in the Summer of 2020, there was an observed increase in the percentage of fatal accidents.



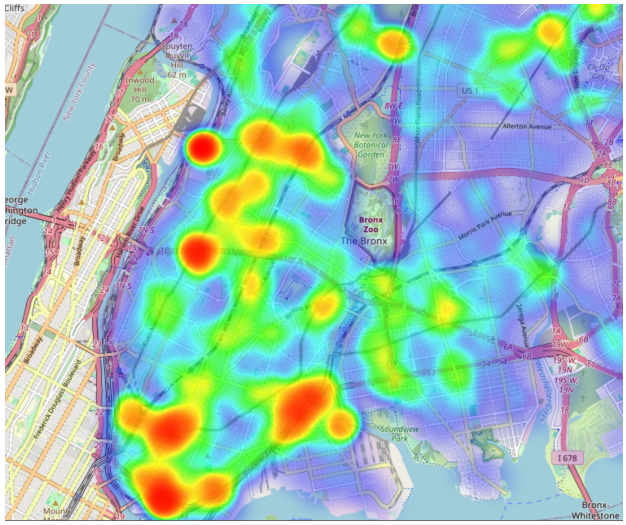
Specifically, the percentage of fatal accidents rose from 22.8% in the Summer of 2019 to 32.5% in the Summer of 2020. This shift indicates that, proportionally, a higher percentage of accidents in Summer 2020 resulted in fatalities or severe injuries compared to the previous year.

The analysis of pedestrian, cyclist, and motorist data in the Bronx area of New York reveals a notable trend. The data indicates that being a motorist comes with a higher risk, as the number of accidents involving motorists is comparatively higher. Additionally, concerning safety, there has been a concerning increase in deaths for both motorists and cyclists from 2019 to 2020.



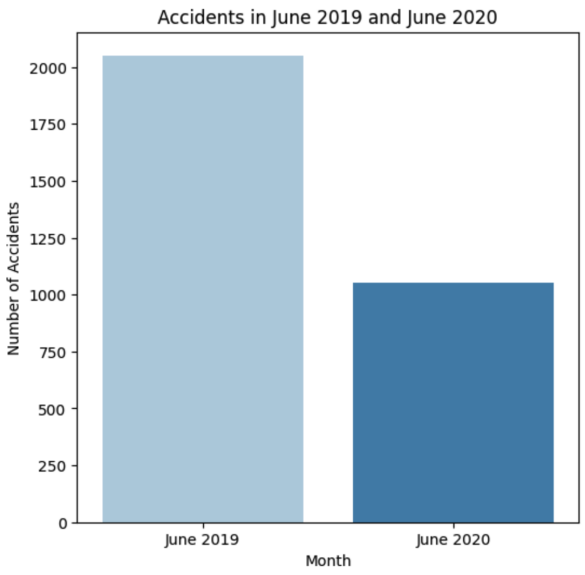
This insight suggests a heightened level of danger for individuals driving in the Bronx, emphasizing the need for enhanced safety measures and awareness, particularly for those commuting as motorists or cyclists.

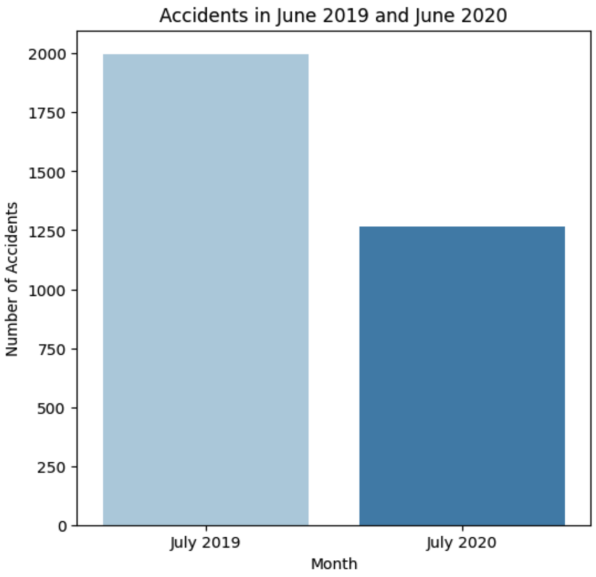
June 2019



June 2020

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





June 2019 June 2020

There were fewer accidents in June 2020 compared to June 2019, with a noticeable drop of over 50%. We used kernel density estimation plots to create heat maps, helping us analyze crash locations. It was observed that there was a significant increase in accidents around Webster Avenue and Tremont Avenue during the COVID-19 pandemic in June 2020.

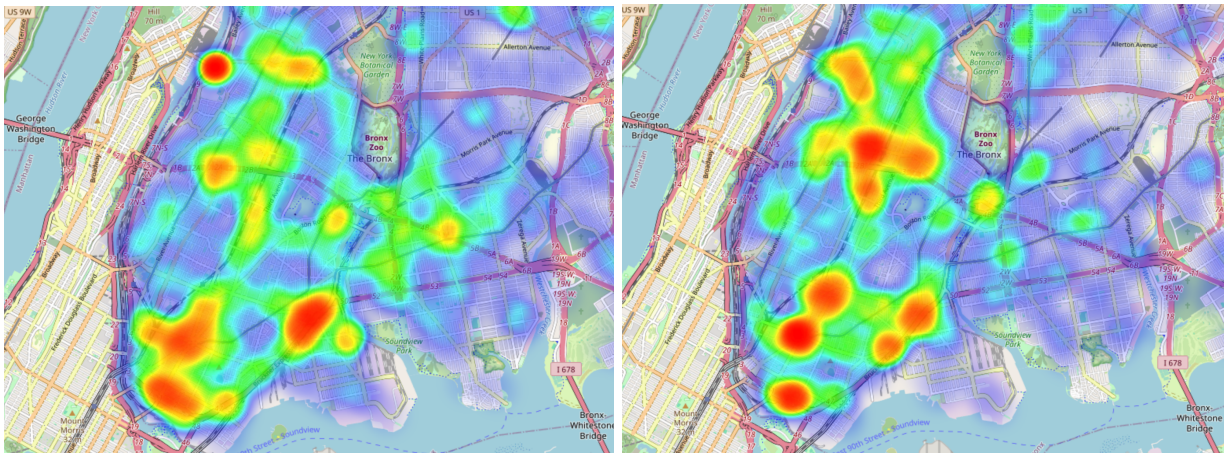
The Bronx Center, on the first floor of 1920 Webster Avenue in Tremont, meets the comprehensive needs of COVID-19 patients, providing them with pulmonary, cardiology, radiology and mental health services.

Dec 7, 2020

This surge in accidents can be attributed to the establishment of one of the first COVID-19 testing centres in that area, leading to increased traffic and activity around the testing facility.

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

July 2019 July 2020



In July 2020, similar to June, there were fewer accidents compared to July 2019, as seen in the bar plot from the previous answer. We utilized the same kernel density algorithm to map accident details. Notably, accidents were concentrated around key locations such as the transport station at 3 Av - 149 St in the Bronx and road circles near hospitals. Interestingly, office spaces and public parks continued to show either no or fewer accidents, resembling the trends observed in June 2020.

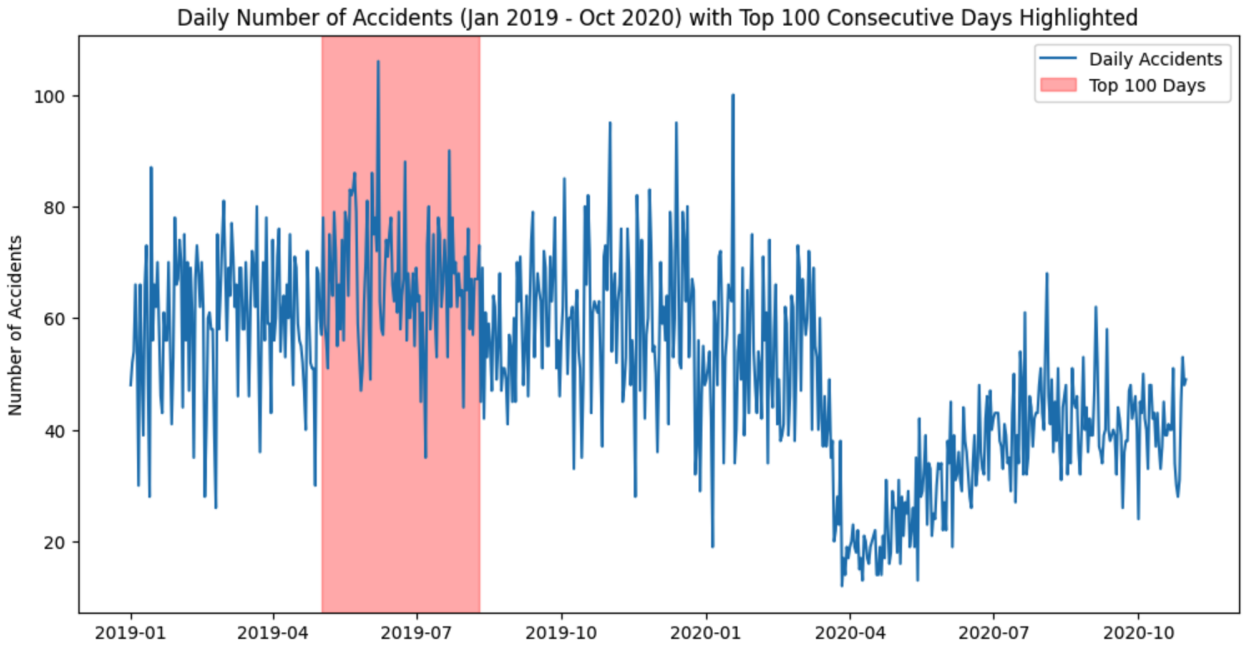
By J. David Goodman

Published June 7, 2020 Updated June 29, 2020

THE LATEST New York City has reopened, with as many as 400,000 people returning to work.

This could be attributed to reduced overall accidents but an increase in incidents around essential places, possibly due to eased travel restrictions during that period.

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



The 100 consecutive days with the most accidents

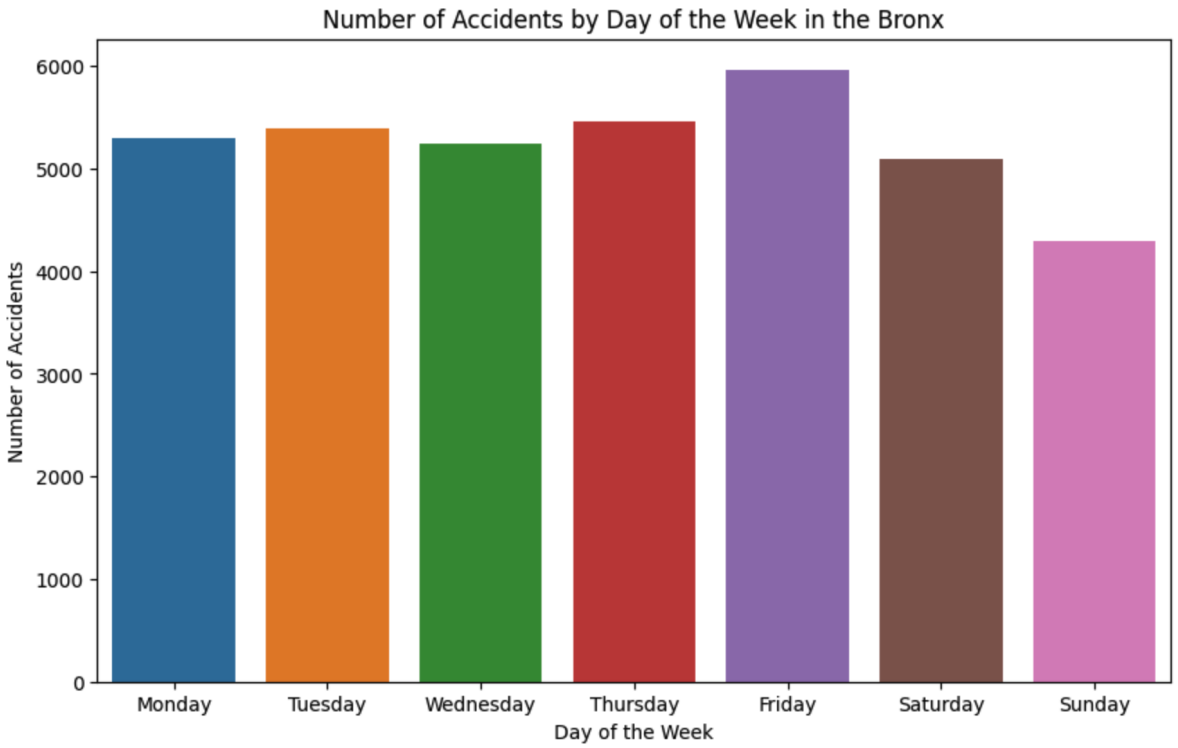
Start Date: 2019-05-02

End Date: 2019-08-10

The 100 consecutive days with the highest number of accidents in the Bronx, from May 2, 2019, to August 10, 2019, can be attributed to a combination of seasonal, environmental, and urban factors. During late spring and mid-summer, the city experiences increased outdoor activities, events, and tourism, contributing to elevated traffic volumes. The milder weather during these months encourages more people to travel, engage in outdoor pursuits, and attend events, potentially leading to a higher likelihood of accidents. Additionally, construction activities, common in urban areas during the summer, and the peak tourism season can impact traffic patterns.

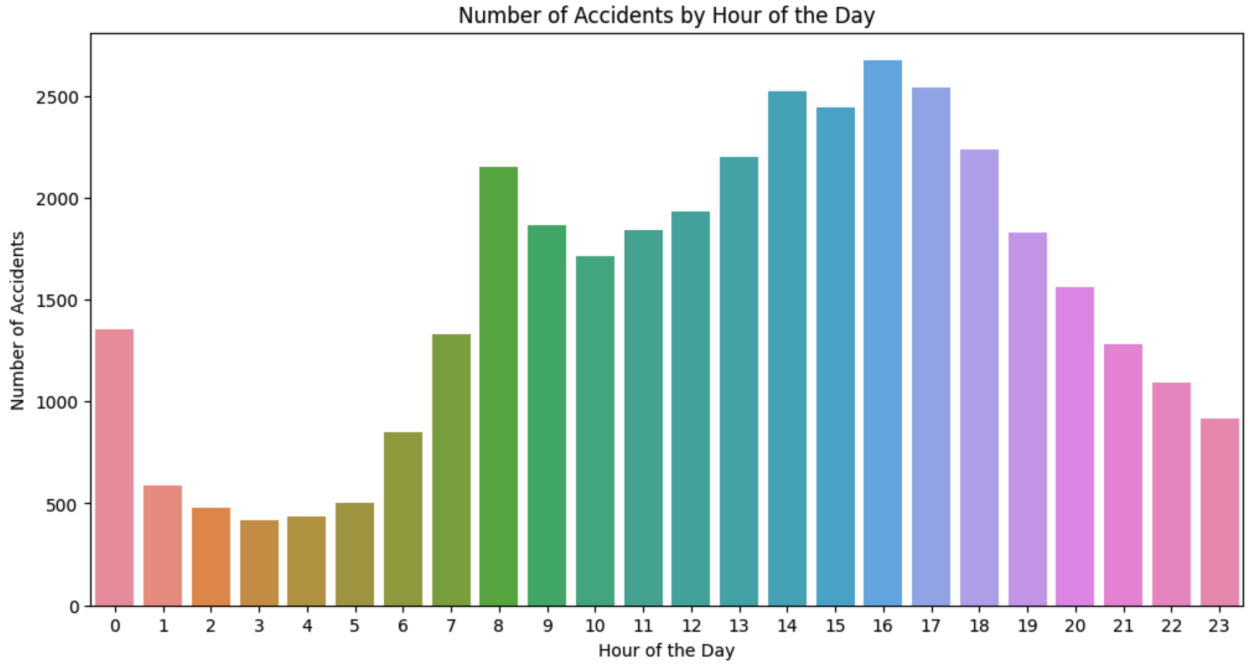
5. Which day of the week has the most accidents?

The day with the most accidents is Friday, tallying up to 5965 accidents.



The day with the most accidents is Friday, totalling 5965 accidents. This could be because Fridays are special – people are excited about the upcoming weekend, and they might be in a hurry to finish work and start their plans. In the Bronx, where there are lots of offices, the roads get busier during rush hours with people heading out for weekend activities. Sometimes, after a long week, people may feel tired or distracted, making accidents more likely. Friday nights often see more social and recreational activities, leading to increased traffic and unfortunately, more accidents. It's important for everyone, especially in areas with many offices like the Bronx, to be extra careful on Fridays and always prioritize safe driving habits.

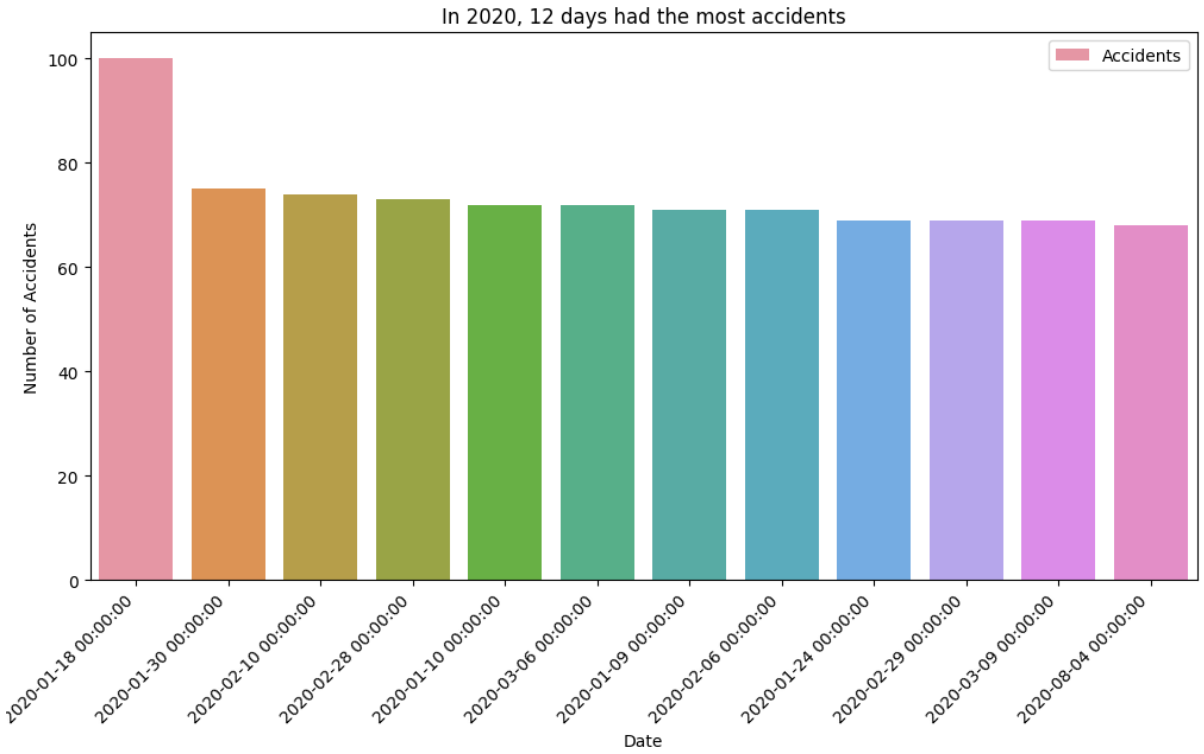
6. Which hour of the day has the most accidents?



The hours with the most accidents are 16:00 (4:00 PM), 17:00 (5:00 PM), 15:00 (3:00 PM), 14:00 (2:00 PM), 9:00 AM, and 8:00 AM. These particular hours see a higher concentration of accidents, with the frequency decreasing as you move away from the peak times. The reasons behind this pattern may include increased rush hour traffic during the late afternoon, mid-morning, and early afternoon hours, as well as potential distractions or fatigue during the morning hours. A spike in accidents at midnight can be because of cases like people drinking and driving during the night time or low visibility problems.

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

In the year 2020, the 12 days with the highest number of accidents were concentrated within the first two months and the last few weeks of the year. Several factors could contribute to this pattern. Adverse weather conditions, including snow, ice, and rain, often make driving more challenging and increase the likelihood of accidents. Poor road conditions, such as potholes and uneven pavement, may exacerbate the risk. Days with heavy traffic, possibly influenced by increased travel associated with holidays like New Year's Day, Christmas, and Thanksgiving, tend to see more accidents due to the higher volume of vehicles on the road.



Furthermore, the winter months with shorter daylight hours can contribute to reduced visibility and potentially contribute to accidents, particularly during nighttime driving. Overall, a combination of weather-related challenges, road conditions, heavy traffic, and the holiday season likely contributed to the higher accident rates on these specific days in 2020.

Fact checked: On January 18, 2020, there were a lot of accidents because of heavy snowfall. The roads were slippery, and it was hard to see clearly. To make things worse, the company forgot to put salt on one road, which usually helps melt the snow and ice. This made driving even more dangerous, leading to more accidents on that day.

Conclusion

This project has provided us with valuable insights into data preparation techniques, handling missing values, identifying and addressing outliers, and performing meaningful data analysis to derive actionable conclusions. We learnt about the patterns and trends of traffic accidents in the Bronx borough of New York City during the summer months of 2019 and 2020.

Firstly, the challenge we faced was the presence of outliers and incorrect data entries, particularly regarding location information. These outliers required careful handling to ensure accurate analysis. The challenges included handling outliers, suppressing warnings, and ensuring consistent interpretation of data within the team. Additionally, the need for domain knowledge to contextualize the data and understand the reasons behind observed patterns was evident. In this exploratory data mining project, the focus was on descriptive statistics and visualizations. However, kernel density estimation plots and geographical heatmaps proved effective for visualizing the spatial distribution of accidents to understand the latent patterns in the data.

Secondly, The identification of changing patterns in traffic accidents, especially the notable decrease in total accidents in 2020 but an increase in the percentage of fatal accidents, was particularly interesting. The impact of external factors such as the COVID-19 pandemic on traffic patterns highlighted the dynamic nature of accident data. Our analysis primarily involved descriptive statistics, kernel density estimation plots, and geographical heatmaps to visualize and interpret traffic accident data. Understanding the context of the data was crucial. For instance, recognizing the influence of external events like the establishment of COVID-19 testing centres on accident patterns provided a more comprehensive view.

In conclusion, we understood the importance of thorough data exploration, understanding the context of the data, and considering external factors that may influence patterns. This project highlighted the need for effective communication within the team to ensure consistent interpretation and decision-making. Additionally, it has significantly contributed to our professional development in the realm of exploratory data mining.