CS 732: Data Visualization Assignment 1 Report

M Srinivasan
IMT2021058
IIIT Bangalore
m.srinivasan@iiitb.ac.in

Siddharth Kothari
IMT2021019
IIIT Bangalore
siddharth.kothari@iiitb.ac.in

Sankalp Kothari
IMT2021028
IIIT Bangalore
sankalp.kothari@iiitb.ac.in

***Abstract*—Assignment 1 report made by group ”Are you able to visualize?” for the course CS-732 Data Visualization on the dataset ”Motor Vehicle Collisions-Crashes” [1].**
This report also mentions all the observations made through the visualization of the data - certain obvious, as well as some surprising, providing a comprehensive perspective of the data.

I. Introduction

The given dataset is about all the motor vehicle accidents in NYC from the year 2020-2023(July). The actual dataset is from 2012, and contains around 20,00,000 entries, which we had to cut down in order to be able to work with it. Some of the important data the dataset contains is listed below:

- Date of collision
- Time of collision
- Borough of New York City in which the accident had taken place.
- The latitude and longitude of the location where the accident took place.
- Number of people injured and killed (pedestrians / motorists / cyclists)
- Major contributing reason to the accident etc.

Using various visualizations, we have done a comprehensive analysis on various correlations between the following fields - time of the day, accident type, borough, number of people injured / killed.
We have analysed the data sets and we have come to some conclusions which shall be mentioned in detail in coming sections. We have used *numpy, pandas* for handling data and *plotly.express* and *matplotlib* libraries in Python to create the visualizations for this assignment.

II. Tasks

Through the visualisations, we have tried to showcase the trends between various columns of the dataset, and to try and show an overall analysis of how safe New York City and its various boroughs (Bronx, Brooklyn, Manhattan, Queens, Staten Island) are. The visualisations present the following -

1) **Task 1** - Analysis of the causes of accidents, their locations and frequencies in the boroughs of New York City.
2) **Task 2** - Analysis of the effect of the time of the day on the accident type, injuries and deaths
3) **Task 3** - Analysis of the deaths and injuries and their correlations with the causes of the accident and boroughs.

III. Accident Categories

There were a lot of causes of accidents, and highly disproportionate data per category (some had numbers which were skyrocketing, others only had a few dozens). Hence we divided the accident types into 7 broad categories for comprehensible representation -

1) Distractions - They represent various cases where the one of the parties was distracted due to something and caused the accident. They broadly include the following sub categories:
   - Driver Inattention/Distraction
   - Outside Car Distraction
   - Passenger Distraction
   - Glare
   - Cell Phone (Hand held or hands free)
   - Other Electronic Devices
2) Car Defects - They represent faults in machinery or equipment which caused the accidents. They broadly include the following:
   - Accelerator Defective
   - Brakes Defective
   - Headlights Defective
   - Steering Failure
   - Tire Failure/Inadequate
   - Tow Hitch Defective
   - Windshield Inadequate
   - Other Lighting Defects
3) Substance Abuse - These include the accidents caused by the parties being involved in consuming alcohol or illegal drugs. They include the following:
   - Alcohol Involvement
   - Illegal Drugs
   - Prescription Medication
4) Driver Not At Fault - These include the accidents caused by environmental or other factors which affected the car. They broadly include the following:
   - Animals Action
   - Lane Marking Improper/Inadequate
   - Obstruction/Debris
   - Other Vehicular
   - Pavement Defective/Slippery
   - Pedestrian Confusion
   - Reactions to Other Uninvolved Vehicles

- Shoulders Defective/Improper
- Traffic Control Device Not Working
- View Obstructed/Limited

5) Medical/Fatigue - These include the accidents which have been caused by the driver either falling asleep or some other medical problem. These include the following -
   - Fatigued/Drowsy
   - Fell Asleep
   - Illness
   - Lost Consciousness
   - Physical Disability
6) Traffic Rule Violations - These include accidents caused by a direct breach of traffic rules and/or accepted methods of driving safely. They include the following -
   - Aggressive Driving/Road Rage
   - Backing Unsafely
   - Failure to Keep Right
   - Failure to Yield Right of Way
   - Following Too Closely
   - Oversized Vehicle
   - Passing Too Closely
   - Passing or Lane Usage Improper
   - Traffic Control Disregarded
   - Turning Improperly
   - Unsafe Lane Changing
   - Unsafe Speed
7) Driver Inexperience - caused by the Driver being inexperienced in driving

IV. Visualizations

All the visualizations mentioned here are done using *matplotlib* and *plotly.express* as mentioned earlier.

1) *Analysis of number of accidents based on time* - We first segregated the data set we got into four parts on the basis of accident time as follows:
   - Midnight - after 22:00hrs till 6:00hrs next day.
   - Morning - from 6:00hrs till 12:00hrs.
   - Afternoon - from 12:00hrs till 16:30hrs.
   - Night - from 16:30hrs till 22:00hrs.

   We then plotted pie charts for the following against time of the day -
   a) Accident numbers (Figure 1)
   b) Deaths (Figure 2)
   c) Injuries (Figure 3)

   Based on the chart we can conclude that most of the accidents happen during the night and midnight. The borough wise versions of these files can be found in the images directory.

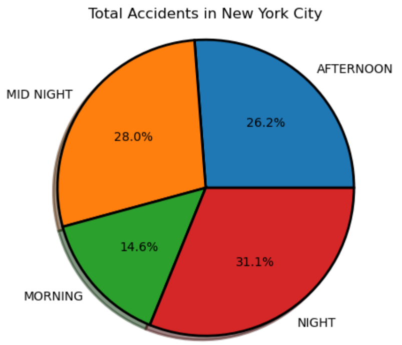

Fig. 1. Image showing accidents % in NYC by time

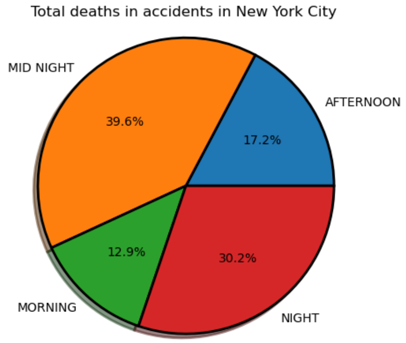

Fig. 2. Image showing deaths % in NYC by time

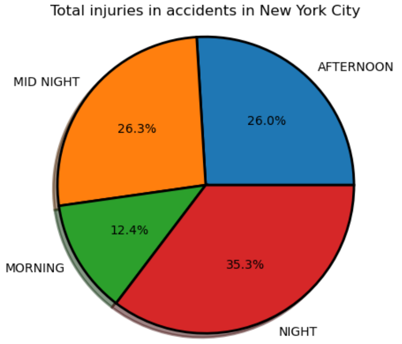

Fig. 3. Image showing injuries % in NYC by time

2) *Geographic locations of accidents in NYC* - We separated the accidents based on which boroughs they occurred in. The 5 boroughs considered for the same are:
   a) Bronx
   b) Manhattan
   c) Brooklyn
   d) Queens
   e) Staten Island

   Based on which borough the accidents happened in, we have used geographic scatter plots to denote the latitude and longitude (location on map) at which the

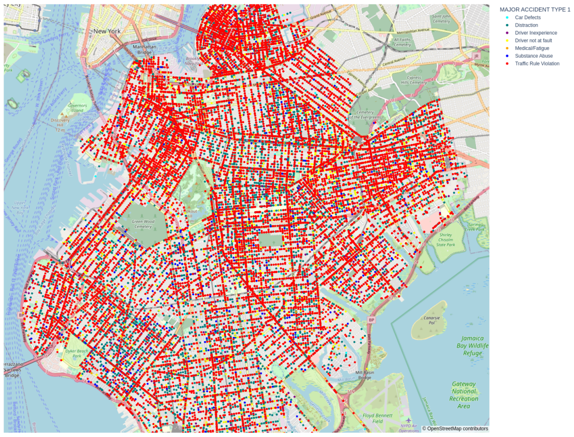

Fig. 4. Geographical scatter plot of accidents in Brooklyn

collision took place. We have further sub-grouped them based on the primary causes of the accident, i.e., what was the reason that the parties/vehicles got involved in the collision. The scatter plots for the other boroughs can be found in the images directory.

3) *Co-relation between deaths/injuries and cause of accidents -*

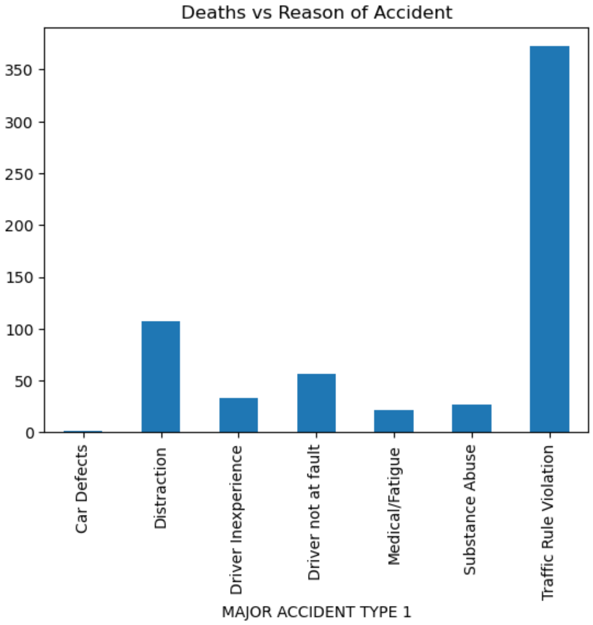

Fig. 5. Image showing deaths vs Cause of accidents.

We grouped the data into categories as mentioned above based on the accident types. Then for each of the 7 grouped categories we summed up the total injuries and death numbers separately and plotted the bar graphs for these two categories.(Figure 5, Figure 6).

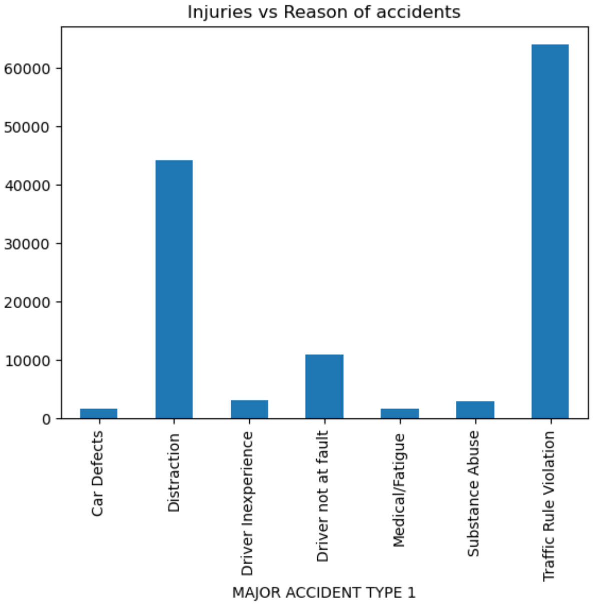

Fig. 6. Image showing injuries vs cause of accidents.

From the initial view on the graph we can observe that Traffic Violations lead to the highest number of fatalities and Traffic Violation and Distraction seem to be the major reasons for the injuries in the accidents.

4) *Co-relation between time of the day and the cause of the accident* - We grouped the data based on the accident types, and for each of the grouped data, we further grouped them based on the time of the accident and plotted pie charts showing the percentage of accidents during the 4 time divisions. (Figure 7)

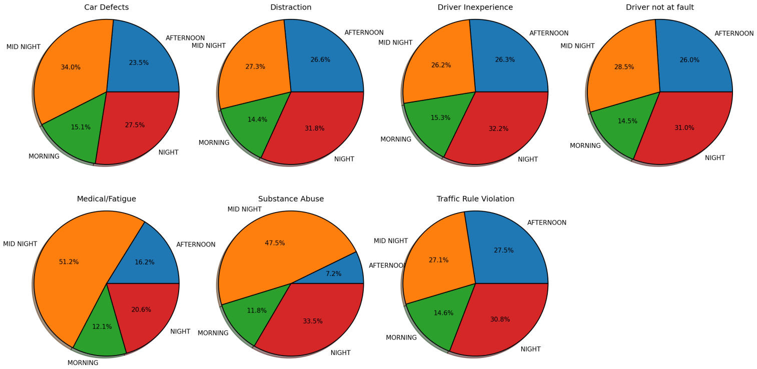

Fig. 7. Plots showing the % by time of day for each of the type of accidents

From initial view of the chart we observe that Medical/Fatigue, Substance Abuse are majorly taking place in the night and rest of them are almost evenly distributed throughout the day (Figure 7).

5) *Boroughs and Injuries/Deaths* - We counted the number of Pedestrian/Motorists/Cyclists fatalities and injuries (separately) for each of the Boroughs of NYC.

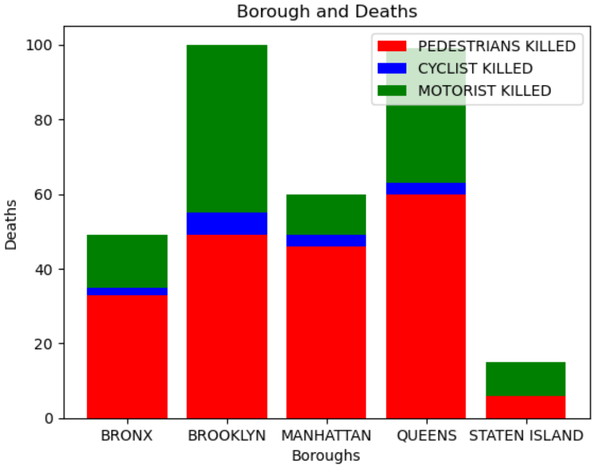

Fig. 8. Plot of the Deaths by Borough

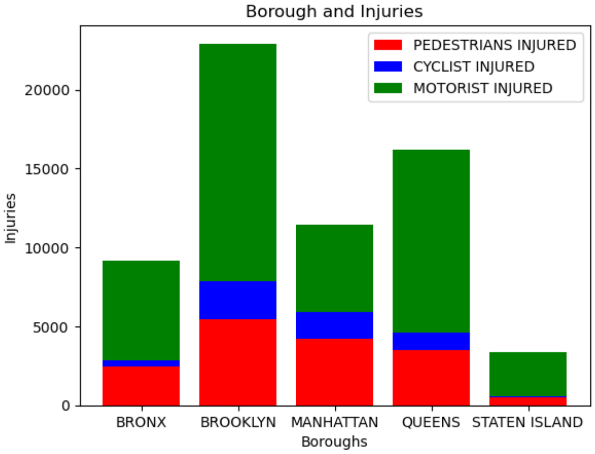

Fig. 9. Plot of the Injuries by Borough

We then plotted stacked bar charts for the same(Fig 8, Fig 9). From the initial view of the graph we observe the Staten Island has the least number of injuries and deaths, while Brooklyn and Queens have higher injuries as well as deaths.

6) *Accident Types in Boroughs in NYC* - We grouped the data on the basis of each of the boroughs of NYC and then for each group we counted the number of accidents by type and then plotted a bar graph for the same. (Figure 10).
An initial look at the charts, shows that most accidents are due to Distraction and Traffic Rule Violation, and all boroughs follow the same trend.

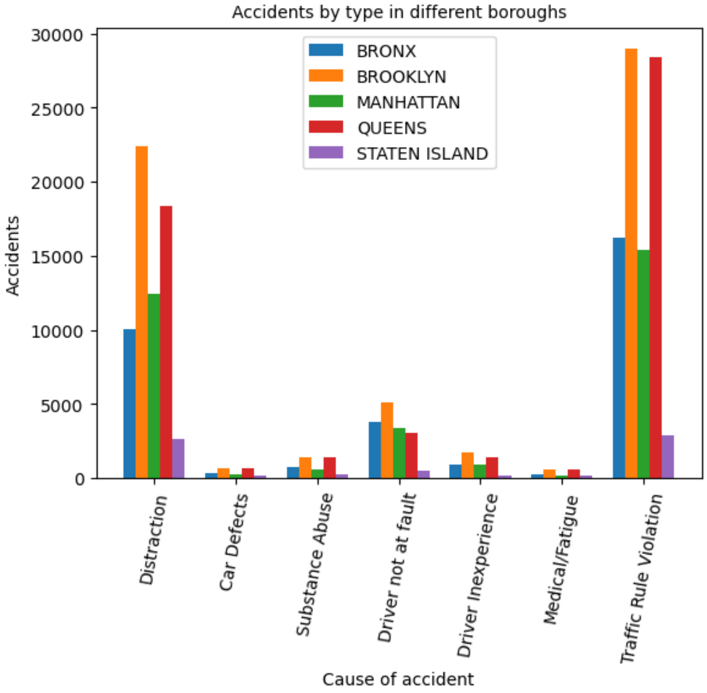

Fig. 10. Plot of the type of accident for each of the boroughs.

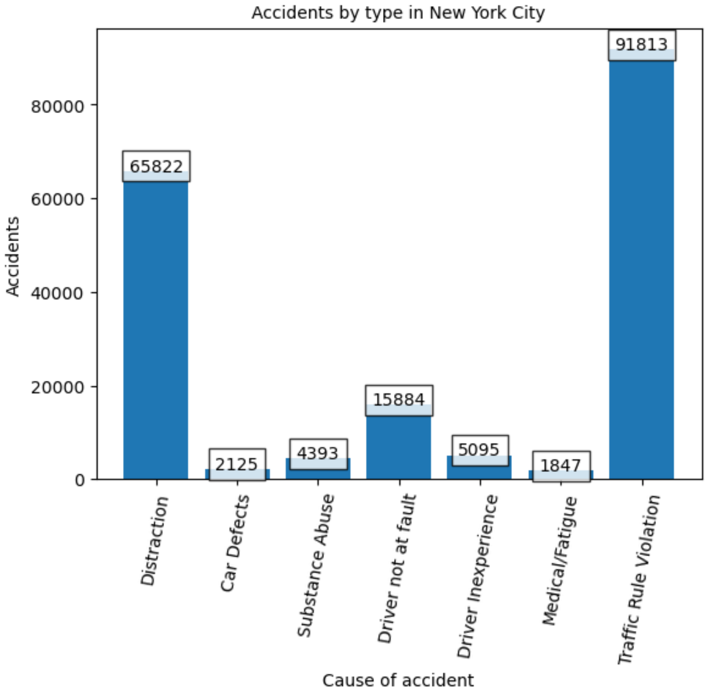

Fig. 11. Plot of the type of accident for the entire city.

V. Inferences

1) By looking into the pie charts plotted for the number of accidents in NYC, we observe that the majority of accidents are taking place in the night (evening included). This is usually the time when people get back from work or head outside for personal work. We also observe that the accidents that take place during the midnight are relatively more fatal compared to the ones that take place in the morning/afternoon. The plots shown here are for the entire city. Similarly we plotted them for each of the boroughs of NYC, and they have been added to the images directory in our submission. The inferences for the boroughs are similar.
2) Another inference worth mentioning here is that there is a very less percentage of accidents that take place in the mornings, given that is peak traffic time and number of vehicles on the road would be relatively higher. The accidents taking place in the mornings have neither high fatalities nor high injuries suggesting that most of them are minor accidents.
3) On further view on the borough wise geographical scatter plots (put in the images directory), we observe a lot of red points on each map compared to other colours, indicating that most accidents that take place are due to traffic rule violations. Other causes such as substance abuse etc. are relatively less in number.
4) By plotting the co-relation between deaths and causes of accident, we observe that the number of deaths corresponding to traffic violations are high and so is the number of injuries which was expected. But to our surprise we observe that the number of injuries are high for the distraction case, indicating that distraction accidents are mostly non-fatal or minor accidents.
5) Plotting the correlation between time of the accident and cause of the accident yields some interesting information. Usual accidents are somewhat evenly distributed across all times of day. However, accidents due to Medical Reasons/Fatigue and substance abuse occur very frequently in the midnight hours, which is expected. People who drive will usually be tired/sleepy during the late night hours which will make them loose control over driving. Most of the drug abuse, drink and drive issues happen only during the odd hours.
6) Plotting the number of accidents and injuries for each of the boroughs, we observe that the number of deaths, injuries and accidents are higher for Brooklyn and Queens. Staten Island has the least number of deaths as well as accidents among all the boroughs, which is expected given that it is a suburb and is away from the bustling city of NYC.
7) On analysis of the number of cyclists, pedestrians and motorists killed/injured, we realise that cyclists are relatively much safer in New York City than pedestrians and motorists.
8) There is a very low ratio of the number of deaths vs injuries. However, this may be due to the fact that all people injured, regardless of the severity of the injury, have been put into the same category.

VI. Work Distribution

The tasks for this assignment were initially brainstormed during a collaborative meeting, where we collectively built upon each other's ideas. Our team worked closely together, offering suggestions and making improvements to the visualization concepts. We opted not to divide tasks among ourselves because we believed in the value of a collaborative approach, where everyone's input and expertise contributed to creating a more compelling and informative presentation and analysis.

References

[1] US data.gov catalog - City of New York - Motor Vehicle Collisions-Crashes [Online] Available: https://catalog.data.gov/dataset/motor-vehicle-collisions-crashes

[2] Schulz, H. J., Nocke, T., Heitzler, M., & Schumann, H. (2013). A design space of visualization tasks. IEEE Transactions on Visualization and Computer Graphics, 19(12), 2366-2375. https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=6634156

[3] Matplotlib Documentation https://matplotlib.org/stable/index.html

[4] Plotly.express Documentation https://plotly.com/python/plotly-express/