Data Visualization
group project

Group number 7

Team members:

**Aditya Satpute, Yongjia Heng, Jaskeerat Brar,
Vivienne Xiang, Srinivas Abhilash Chintaluru**

About the dataset

Source: NYPD Motor Vehicle Collisions - Kaggel.com

Data dictionary:

Column Name	Description
COLLISION_ID	Unique record code generated by system
ACCIDENT_DATE	Occurrence date of collision
ACCIDENT_TIME	Occurrence time of collision
BOROUGH	Borough where collision occurred
ZIP CODE	Postal code of incident occurrence
LATITUDE	Latitude coordinate for Global Coordinate System, WGS 1984, decimal degrees (EPSG 4326)
LONGITUDE	Longitude coordinate for Global Coordinate System, WGS 1984, decimal degrees (EPSG 4326)
ON STREET NAME	Street on which the collision occurred
CROSS STREET NAME	Nearest cross street to the collision
NUMBER OF PERSONS INJURED	Number of persons injured
NUMBER OF PERSONS KILLED	Number of persons killed
NUMBER OF PEDESTRIANS INJURED	Number of pedestrians injured
NUMBER OF PEDESTRIANS KILLED	Number of pedestrians killed
NUMBER OF CYCLIST INJURED	Number of cyclists injured
NUMBER OF CYCLIST KILLED	Number of cyclists killed
NUMBER OF MOTORIST INJURED	Number of vehicle occupants injured
NUMBER OF MOTORIST KILLED	Number of vehicle occupants killed
CONTRIBUTING FACTOR VEHICLE 1	Factors contributing to the collision for designated vehicle
CONTRIBUTING FACTOR VEHICLE 2	Factors contributing to the collision for designated vehicle
VEHICLE TYPE CODE 1	Type of vehicle based on the selected vehicle category (ATV, bicycle, car/suv, ebike, e-scooter, truck/bus, motorcycle, other)
VEHICLE TYPE CODE 2	Type of vehicle based on the selected vehicle category (ATV, bicycle, car/suv, ebike, e-scooter, truck/bus, motorcycle, other)

The Hypotheses

The following hypotheses were made before conducting actual visualizations on the dataset

1. Certain locations have more accidents occurring compared to the other areas in New York City.

2. Frequency of crashes increases in the winter season compared to other seasons.

3. Driver Distraction is the leading cause of fatalities for pedestrians.

Analyses

Exploring Hypothesis 1

Hypothesis 1: Certain locations have more accidents occurring compared to the other areas in New York City

As we are trying to identify locations with a high number of accidents, we used the ‘Borough’ column which has the location data of the accident, and since it’s a categorical variable, we plotted a bar graph to visualize the frequency of accidents at different locations.

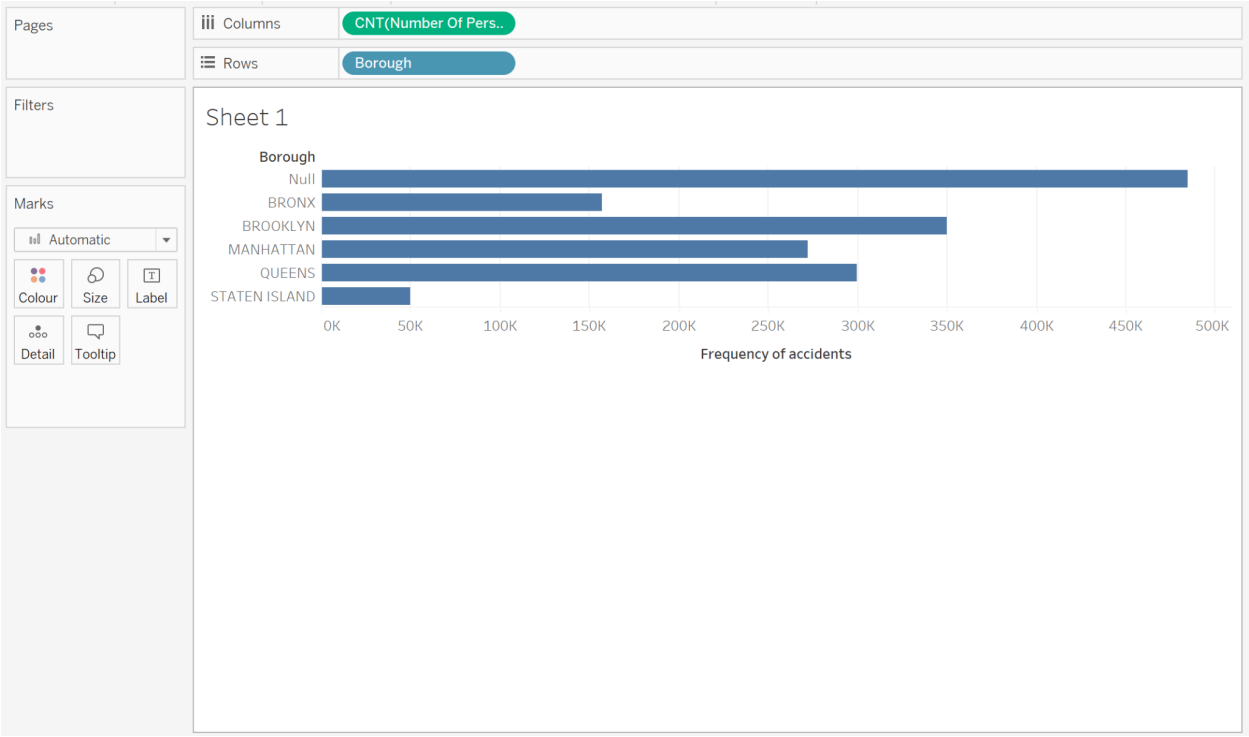
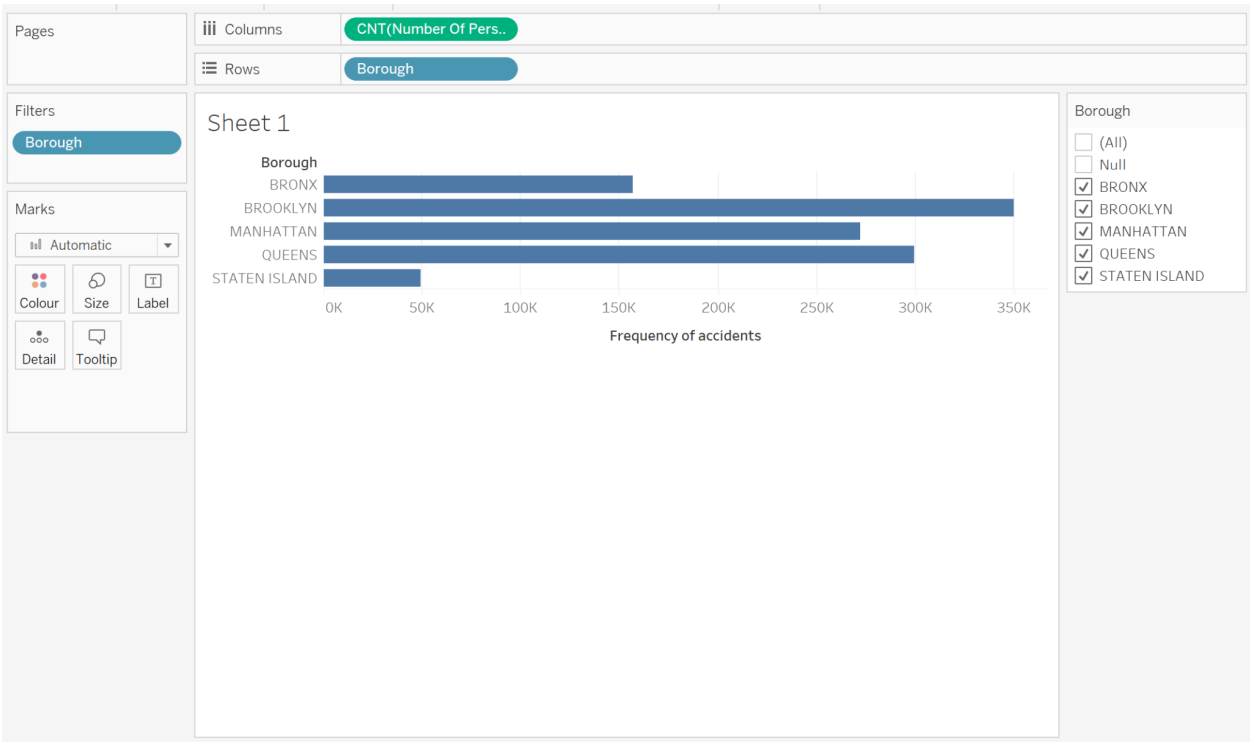
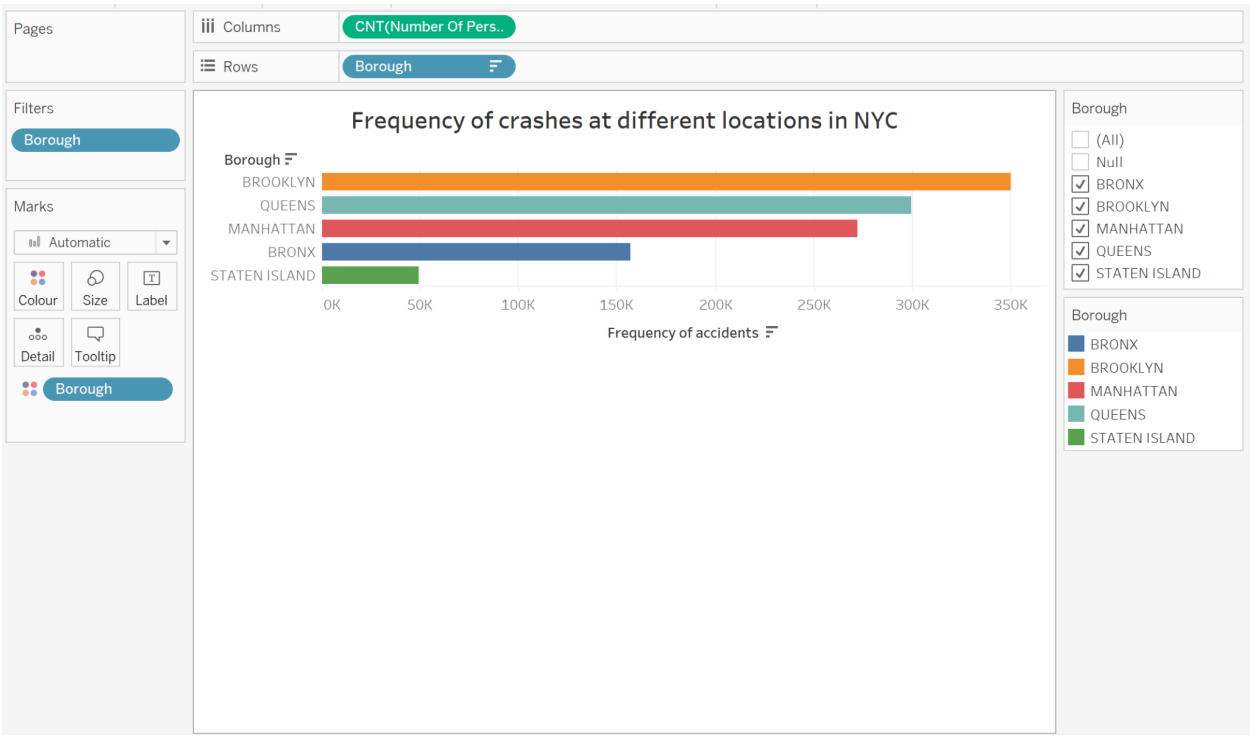


Fig. 1.1 Initial Visualization

As there are a lot of null values, for the accidents where the system was not able to capture the location, we placed a filter to remove the null values and visualize the frequency of the accidents at the rest of the locations where data is available.



The bar chart looks something like this after we remove the null values. To make the visualization more appealing, we sorted the data based on the number of accidents as well as colored the different region bars and added a title to the bar chart.



The visualization shows that the Brooklyn, Queens, and Manhattan areas have the highest crashes.

On further exploration of the data, we found that the pin code, latitude, and longitude data is available, which could eventually help us understand the crashes at a more local level rather than the broad region level.

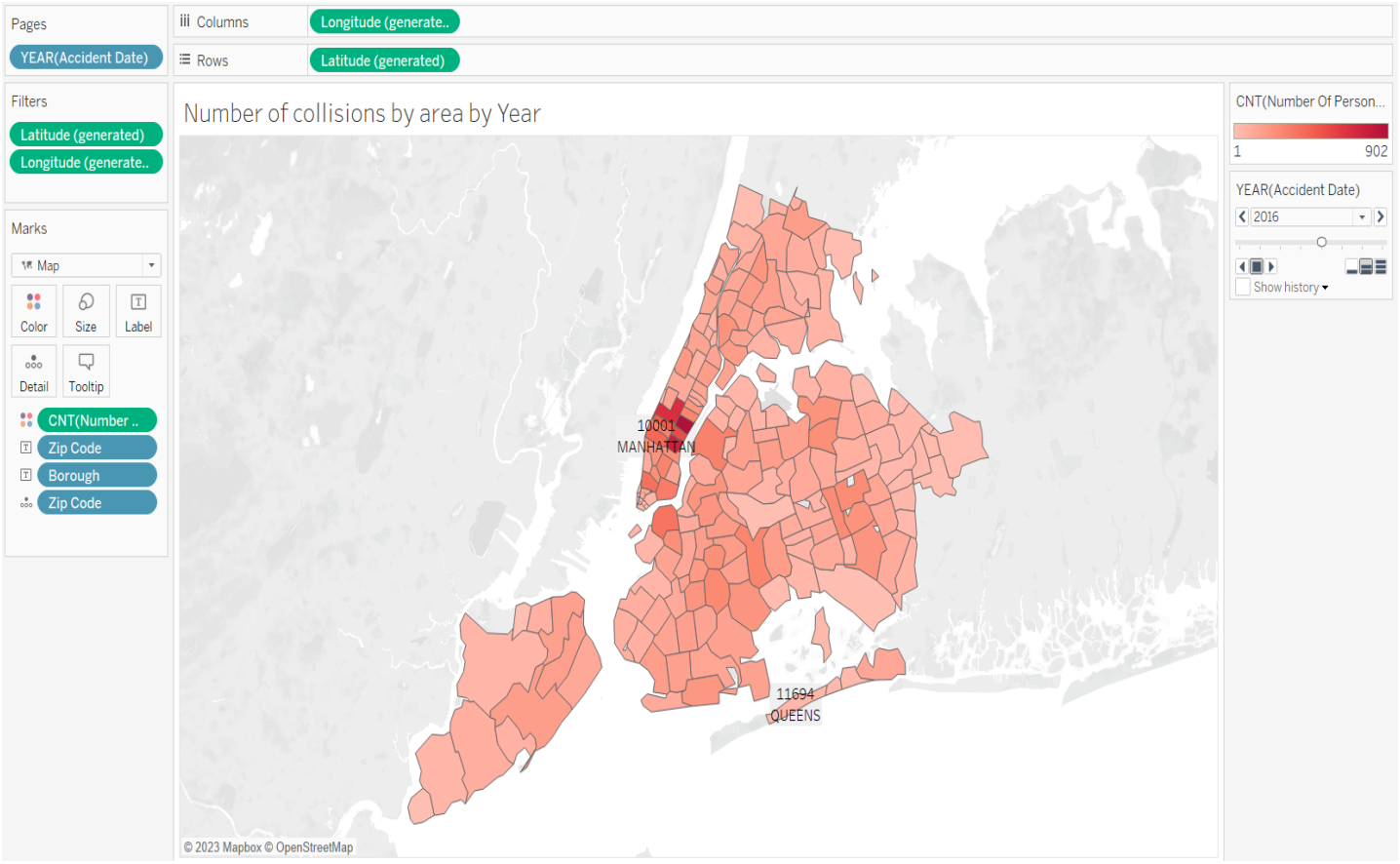


Fig. 1.4 Final Visualization

We build a heatmap on Tableau using the pin code, latitude, and longitude data.

The above graph represents the frequency of crashes occurring at different locations in New York City in a more detailed way as compared to the bar chart that we built at the beginning.

Exploring Hypothesis 2

By intuition, we believe that collisions occur frequently in winter due to the weather conditions in New York. For example, the road can be slippery when it's snowy, and the cold weather affects engines. Thus, we made a histogram to display the frequency of collisions by season.

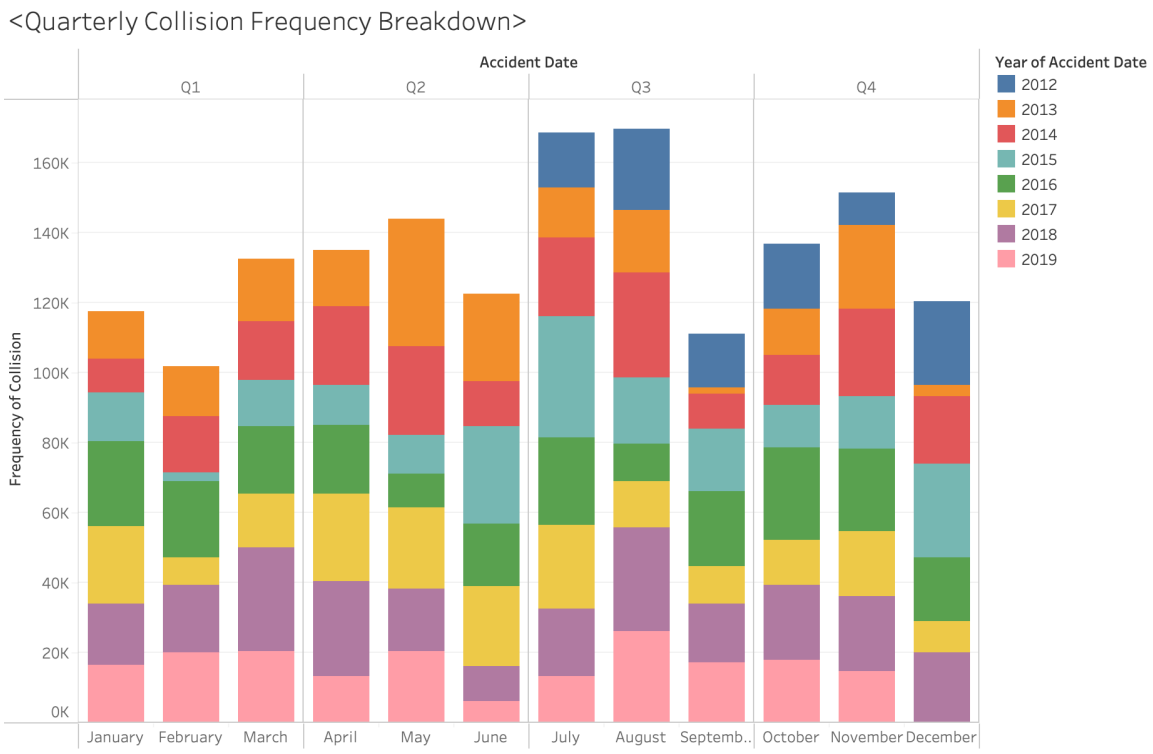


Fig. 2.1 Initial Visualization

Here we used the column "Accident Date" versus "Collision Id". To specify seasons, a calculation field is defined, where Spring contains March to May, Summer contains June to August, September to November is in Fall, and Winter includes December to February. As mentioned in the dataset description, each accident has a Collision ID, so a COUNT measure is used. Besides, to show the distinction between collisions occurring in winter and other seasons, a reference line of the average occurrence number is added.

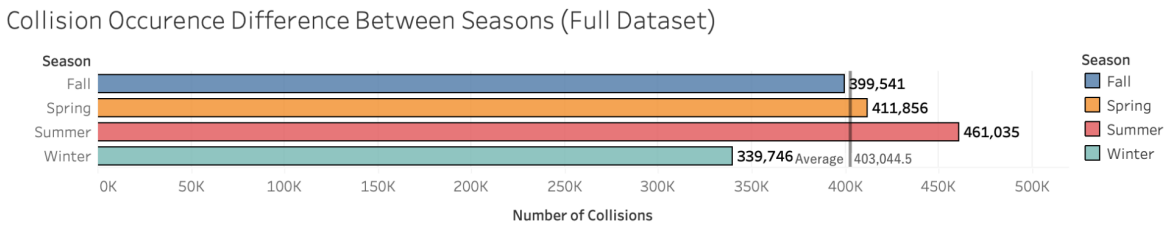
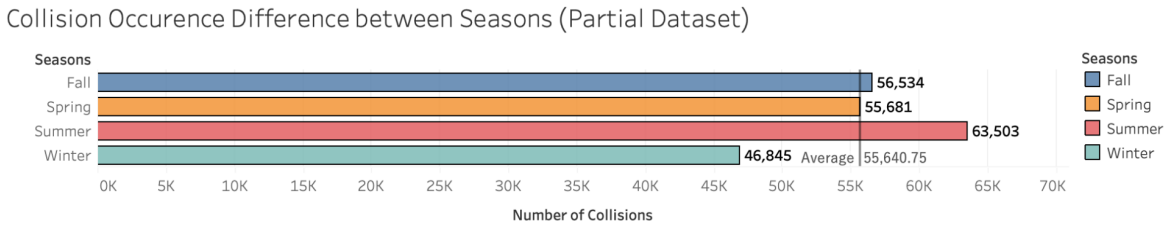


Fig. 2.2 Final Visualization

To our surprise, the total amount of collisions in winter is far below the average. Thus, we should reject our null hypothesis that the frequency of crashes increases in the winter season compared to other seasons. To be more accurate, a statistical test can be conducted by Python, which could give us more solid evidence to support the conclusion.

In addition, to figure out the main cause of fatality, we created a partial dataset in which unspecified collision causes are removed. To figure out whether the unspecified causes affect collision distribution among seasons, another visualization is made. It can be seen that the noise does not influence the frequency of collisions between seasons.



Exploring Hypothesis 3

Driver Distraction is the leading cause of fatalities for pedestrians.

We wanted to investigate what is the leading factor which causes the highest fatalities. To take a wild guess, lots of ideas came relating to driver fatigue / distraction.

'Unspecified' factor leads the way for contributing factors of fatalities. The NYSDOT data engineering team should focus on enhancing the classification of this factor to provide data analysts with a more meaningful framework. This improvement will enable us to extract optimal value and address specific business requirements more effectively.

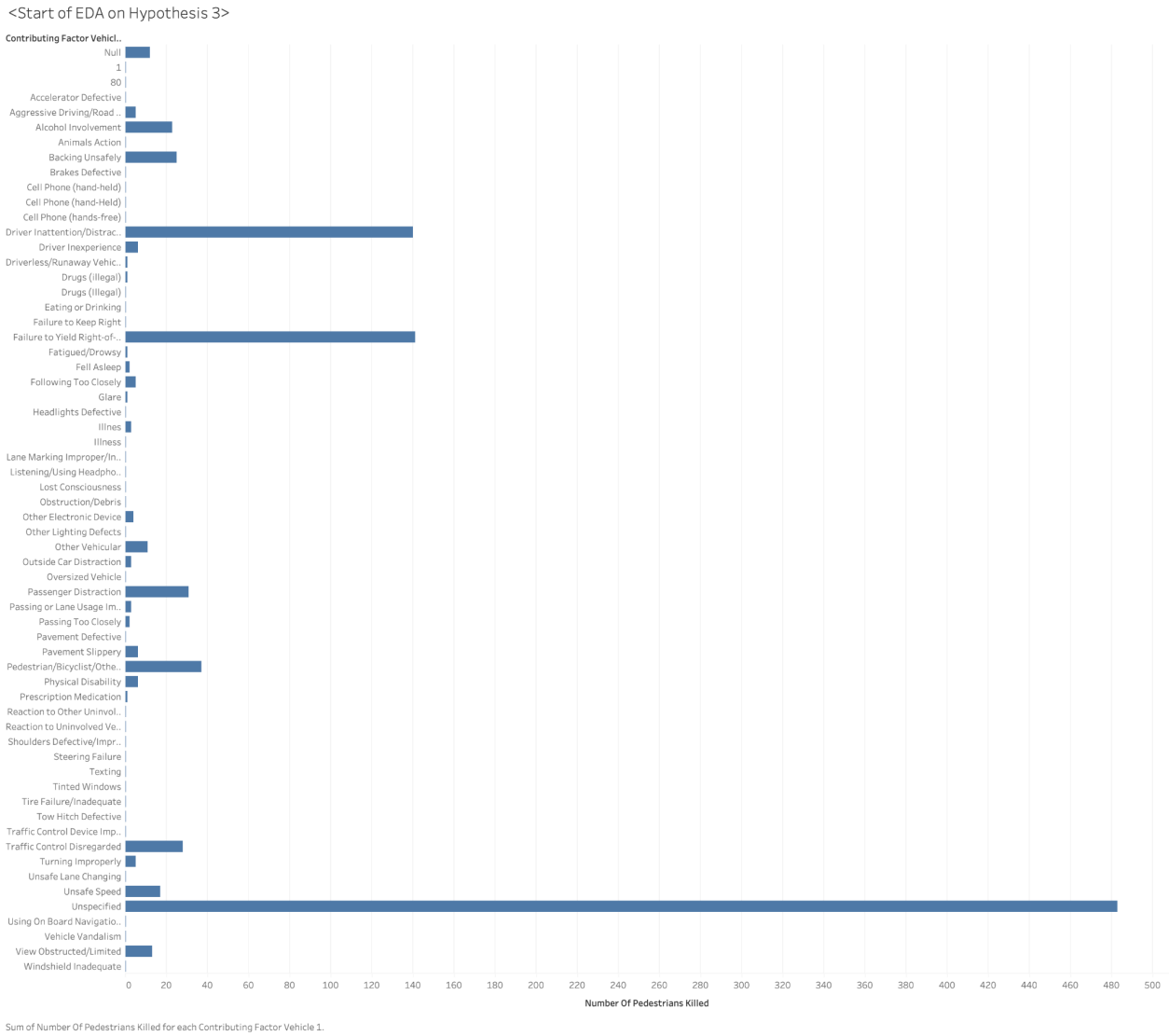


Fig. 3.1 Initial Visualization

We will have to possibly take this variable out to explore the other factors in greater granularity.

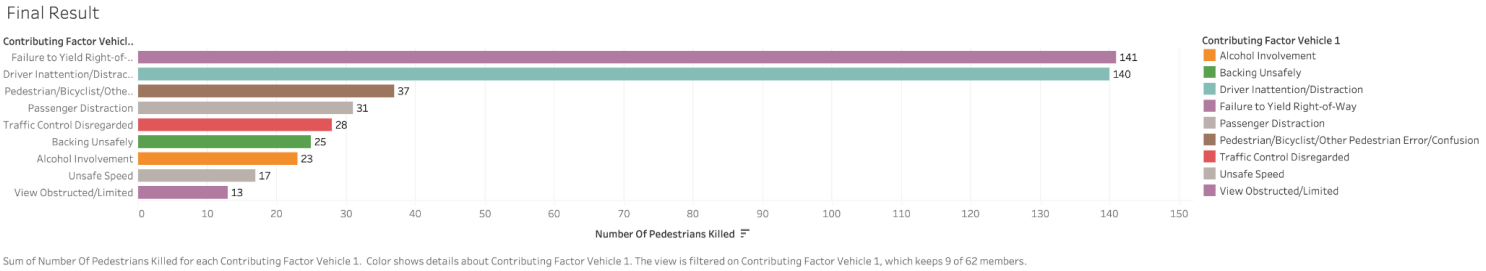


Fig. 3.2

After taking out null values and *Unspecified* factors, We were able to achieve great granularity in assessing the leading factors which cause fatalities to pedestrians.
In case of pedestrians, Driver Fatigue comes in a close second. *Failure to Yield Right of the way* beats it by a tiny margin!

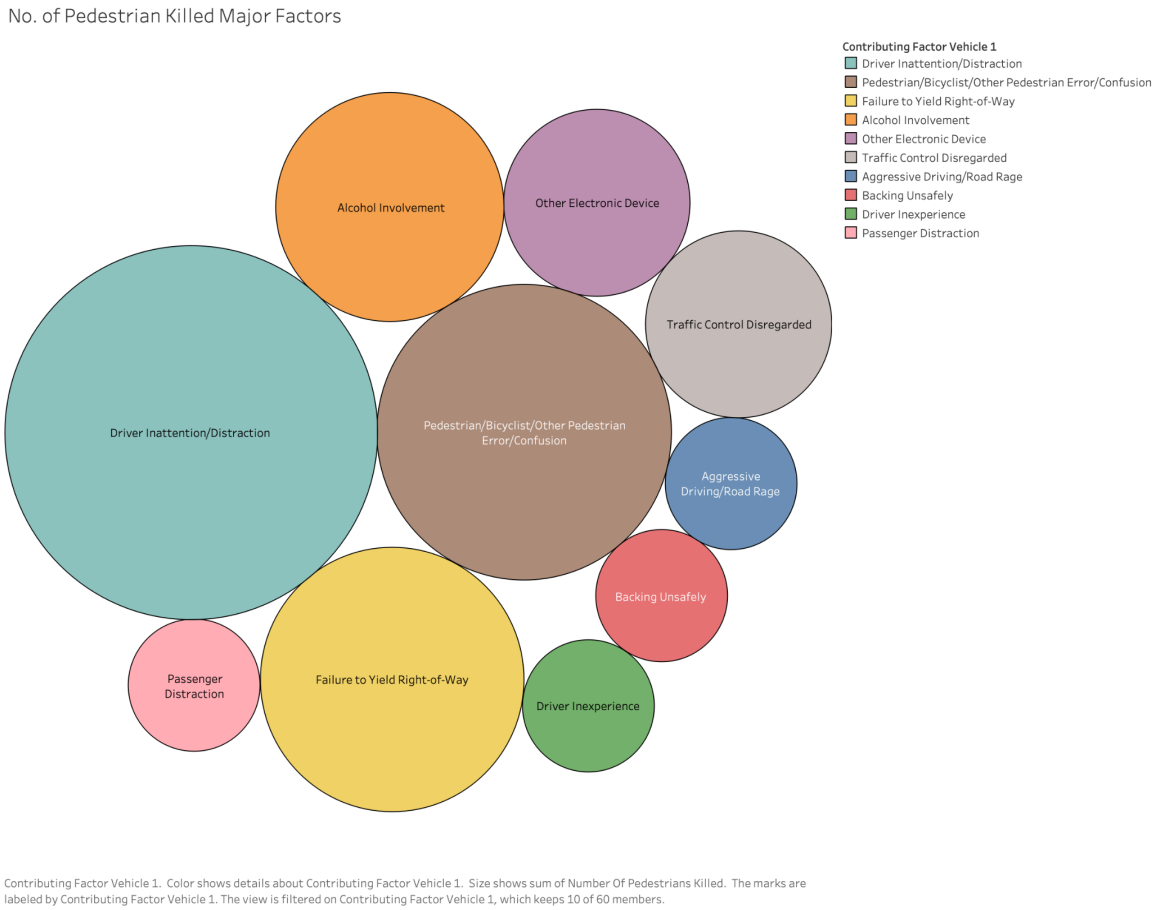


Fig. 3.3 Final Visualization

The Number of Pedestrians killed metric was calculated by using the function SUM.

Although we have close margins here, we will ***Reject our null hypothesis*.**

However, our analysis doesn't stop here. We set out to explore other fatalities under cyclists, persons and motorists.

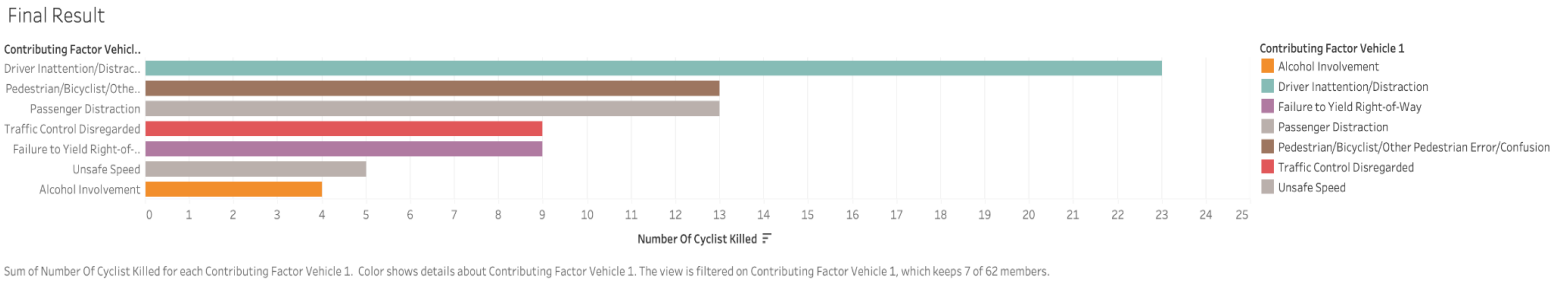
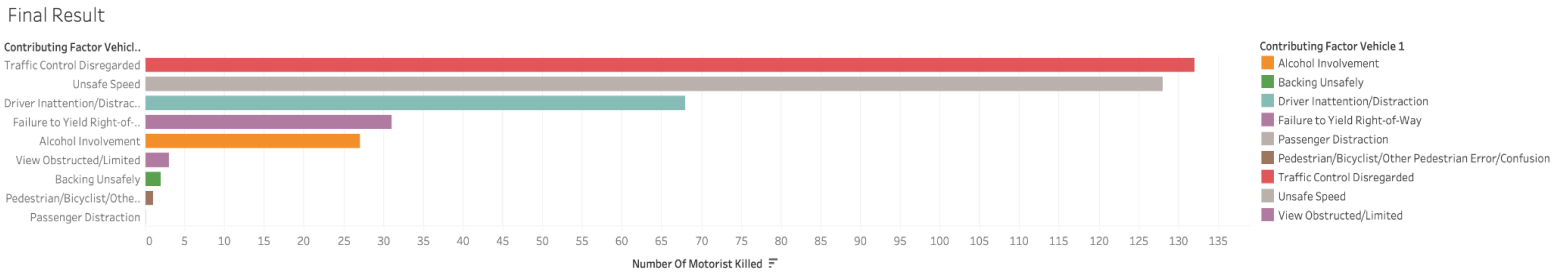
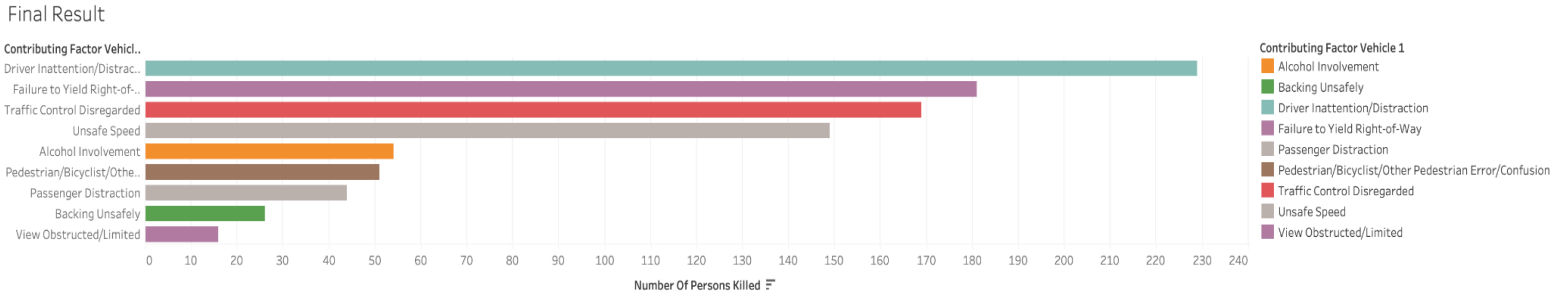


Fig. 3.4 Additional Visuals





Much to our surprise, *Driver Fatigue* is the leading cause of fatalities for cyclists and persons. However, for motorists, factors like *Vehicle Speed & Traffic Control Disregard* are the leading causes of fatalities.

Conclusions

Based on the key findings of our project, the following conclusions can be drawn:

Geographical Variation in Accident Frequency: The analysis indicates a significant geographical variation in accident frequency across New York City, with Brooklyn, Queens, and Manhattan exhibiting the highest number of crashes. This suggests a need for targeted traffic safety measures and infrastructure improvements in these areas.

Seasonal Impact on Accident Rates: Contrary to the initial hypothesis, the data does not support a significant increase in the number of collisions during winter months. This finding challenges common assumptions about seasonal effects on road safety and points towards the effectiveness of winter road maintenance and public awareness in mitigating accident risks during this period.

Primary Causes of Fatalities Differ Across Victims: The leading causes of fatalities vary among pedestrians, cyclists, and motorists. For pedestrians, "Failure to Yield Right of Way" and "Driver Fatigue" are predominant, whereas for cyclists and other persons, "Driver Fatigue" is a leading cause. For motorists, "Vehicle Speed" and "Traffic Control Disregard" are critical factors. This diversity in causation highlights the necessity for multifaceted road safety strategies that address specific risks faced by different road users.

In light of the findings from our project, specific recommendations are proposed for the New York Police Department (NYPD) and the New York Department of Transportation (NYDoT) to enhance road safety in New York City:

- Focusing on High-Risk Areas: The analysis identified Brooklyn, Queens, and Manhattan as regions with notably high accident frequencies. It's imperative to concentrate on safety measures in these areas. Strategies could include increasing traffic surveillance, enhancing street lighting, and improving road signage to ensure better visibility and awareness. Such targeted interventions are crucial for mitigating accident risks in these densely populated and high-traffic zones.

- Year-Round Safety Campaigns: The study refutes the assumption that accident rates spike during winter, highlighting the effectiveness of existing winter safety protocols. This insight advocates for the continuity of safety measures throughout the year. Regular road safety campaigns, focusing on driver education and enforcement of traffic rules, could be beneficial in maintaining a consistent level of road safety awareness among the public, irrespective of the season.

- Educational Initiatives for Specific Risks: Differentiating the primary causes of fatalities among pedestrians, cyclists, and motorists is essential for effective safety campaigns. For pedestrians, focusing on the dangers of "Failure to Yield Right of Way" and "Driver Fatigue" is crucial, while for motorists, addressing issues like "Vehicle Speed" and

"Traffic Control Disregard" is vital. Tailored educational programs can significantly reduce fatalities by targeting these specific risk factors.