*“More than 1 million people die in Vehicle Collision in the United States” What measures can we take to reduce & prevent these collisions.*



***Group 5***

*Archana Mahapatra*

*Kaumudi Kulkarni*

*Madhvika Sehgal*

Business Intelligence Project

Vehicle Collisions in New York City

CONTENTS

[I. BUSINESS PROBLEM 3](#_Toc40017889)

[II. CLIENT REQUIREMENTS 3](#_Toc40017890)

[III. DATA 4](#_Toc40017891)

[IV. PROJECT GOALS 7](#_Toc40017892)

[V. SOLUTIONS 8](#_Toc40017900)

[VI. RESULTS 9](#_Toc40017901)

[VII. CONCLUSION 9](#_Toc40017902)

[VIII. REFERENCES 10](#_Toc40017903)

[IX. APPENDIX 11](#_Toc40017914)

[Dashboard 1: Street Wise Collisions Based on Borough 11](#_Toc40017915)

[Dashboard 2: Heatmap of Collisions by each borough based on vehicle type 12](#_Toc40017916)

[Dashboard 3: Areas affected the most by intensity on the contributing factor 12](#_Toc40017917)

[COLLISIONS BY THE BOROUGH: 13](#_Toc40017918)

[COLLISIONS BY THE WEEK: 14](#_Toc40017919)

[HOUR OF THE MAJORITY OF COLLISIONS: 14](#_Toc40017920)

[VEHICLE COLLISION GROUPED BY CYCLISTS, MOTORIST & PEDESTRIANS: 15](#_Toc40017921)

[TWO-WHEELER COLLISIONS BY THE MONTH: 15](#_Toc40017922)

[ANOMALY IN NUMBER OF PEDESTRIANS INVOLVED IN MANHATTAN: 16](#_Toc40017923)

[PEDESTRIANS INJURED BY SEASON 16](#_Toc40017924)

[TOP STREETS WITNESSING COLLISIONS EACH YEAR BASED ON VICTIMS: 17](#_Toc40017925)

[FORECASTING COLLISIONS FOR 2020 & 2021: 17](#_Toc40017926)

[WEATHER CASE STUDY: 18](#_Toc40017927)

# BUSINESS PROBLEM

“Every year more than 1.35 million people die in road accidents worldwide every year resulting in about 3700 deaths a day”.

There is an increase in the number of motor vehicles per year due to which even the resulting deaths are increasing. Innumerous factors are contributing these rates including traffic laws (differs by state), weather, topography, speed limit, driving under influence, dangerous intersections, car conditions, distracted driving. Also, an important factor that contributes to accidents is the population density of the region.

In the US, there is a huge economic cost associated with the crashes as they would need medical care to attend both deaths and injuries. These collisions can be avoided with better safety measures and stricter laws.

There is a popular conception that most motor vehicle collisions that happen are primarily because of driving under the influence of alcohol. But we think, there are other factors affecting accidents bringing me to reason why we want to study and analyze the datasets further.

# CLIENT REQUIREMENTS

New York reports about at least 28 accidents take place per hour and one accident every two minutes.

NYPD publishes vehicle collision data being updated daily on the NYC Open Data. It consists of all the police reported records also the number of people injured and killed from 2013-2020 consisting of above 1 million collision records. NYC has the Vision Zero Program aiming at zero fatalities by 2024 created in 2014. We found it interesting to analyze the trend of the accidents during the Vision Zero implemented, whether there was an impact on the number of collisions.

New York City is primarily divided into five boroughs that are Manhattan, Bronx, Brooklyn, Queens, and Staten Island. Through this project, we want to analyze and visualize the discrepancies between the number of car accidents that are happening in each borough.

Our goal through this project is to be able to explore where accidents occur and the reason which will help in getting valuable information for reducing accidents and loss of life. This inference is especially important as it can lead policymakers to both understand what similarities there may be in these accidents as well as the locations where the most lives can be saved.

# DATA

For our Business Intelligence project, we have utilized the NYC open data set of vehicle collisions dated from 2013 to 2019.

The dataset has been obtained from the below link:

<https://data.ny.gov/Public-Safety/Motor-Vehicle-Collisions-Crashes/h9gi-nx95>

**DATASET DESCRIPTION:**

The dataset contains more than 1.1 million records and 29 variables of collisions taking place all across New York City. Each row indicates a collision record. These records are gathered from police-reported vehicle collisions in NYC.

The dataset contains the following attributes:

**UNIQUE KEY:** Each collision is given a unique identifier

**Date:** The date on which the collision occurred

**Time:** Time of the collision

**Borough:** 5 boroughs Manhattan, Queens, Brooklyn, Bronx, and Staten Island.

This signifies the borough in which the collision occurred.

**Zipcode:** The zip code within the borough where the collision occurred.

**Latitude:** Latitude of the collision.

**Longitude:** Longitude of the collision.

**Location:** Combination of latitude and longitude, signifying the exact location.

**On-street name:** Name of the street on which the collision occurred.

**Cross street name:** Nearest cross street to the collision

**Off-street name:** Nearest off-street to the collision location

The following attributes provide several persons, pedestrians, cyclists, and motorists injured or killed in each collision instance.

**Persons injured**

**Persons killed**

**Pedestrians injured**

**Pedestrians killed**

**Motorists injured**

**Motorists killed**

**Cyclists injured**

**Cyclists killed**

The following attributes describe the type of vehicle involved in the collision, for example, sedan, sports utility vehicle, wagon, etc. The minimum number of vehicles involved is 1 and the maximum is 5. Depending on the number of vehicles involved in each collision, each vehicle type is specified.

**Vehicle 1 type**

**Vehicle 2 type**

**Vehicle 3 type**

**Vehicle 4 type**

**Vehicle 5 type**

Similarly, the following 5 attributes specify the factor of the collision, for example, alcohol involvement, driver inattention, etc.

**Vehicle 1 factor**

**Vehicle 2 factor**

**Vehicle 3 factor**

**Vehicle 4 factor**

**Vehicle 5 factor**

**DATA PRE-PROCESSING:**

1. Cleaning missing/null values:

Attributes like “unique key” which had no significance to the analysis were dropped from the data.

Other attributes like location, borough, latitude, longitude had multiple missing or null values that were impacting the analysis results. These records were deleted to achieve better results.

Also, for this project vehicle collision records were considered from 2013 to 2019. This is because the records from 2012 have been included from July onwards, which does not help the analysis since complete yearly data is not available, especially for forecasting.

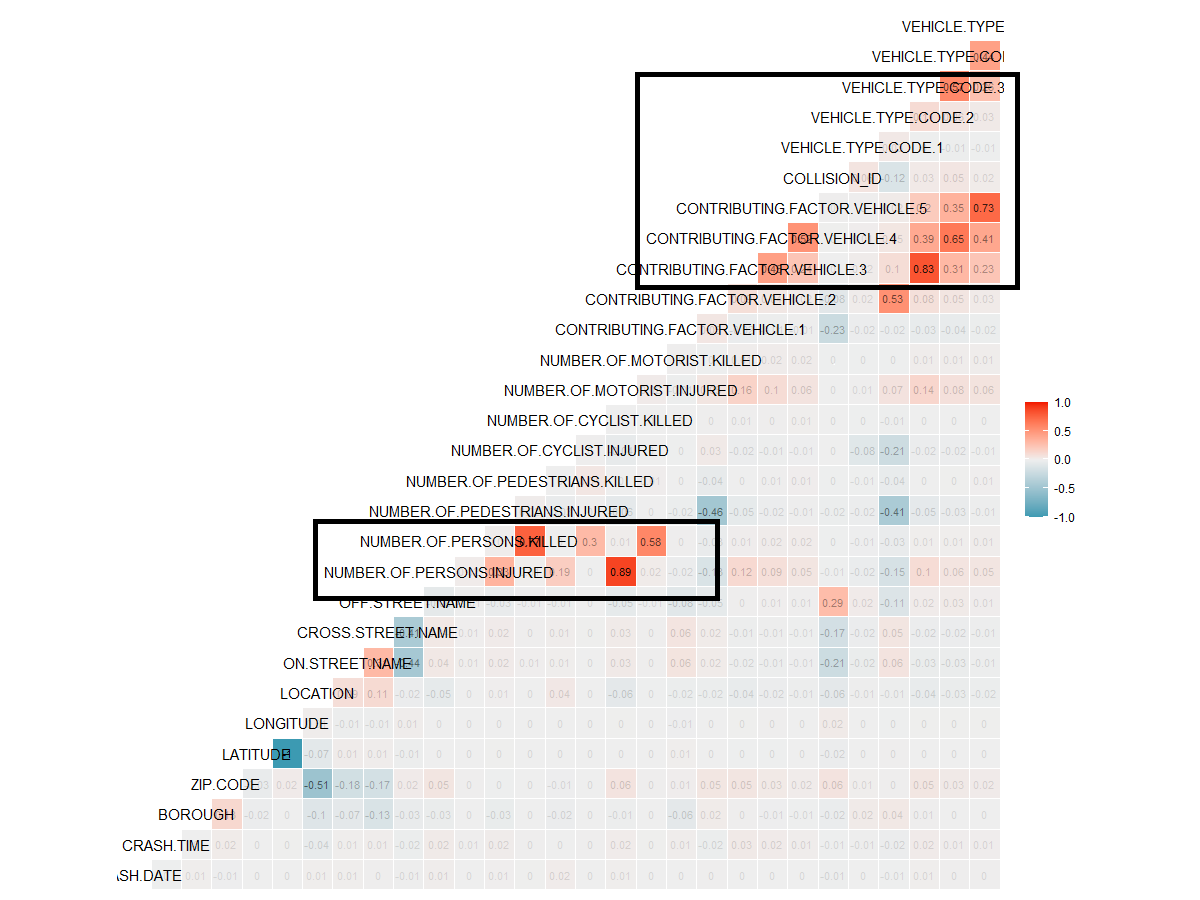
1. Renaming and grouping variables:

We also renamed and grouped certain variables like Motorists Involved, of Pedestrians involved, Cyclists involved by summing up injured and killed counts respectively. We also grouped variables for months based on seasons like Summer, Winter, Spring, and Fall.

1. Finding correlation:

We ran the code to find the correlation between the dimensions we obtained the following correlation matrix in the below Figure 1:

From the correlation matrix, we can see that there is no significant positive or negative correlation between the variables. For example, there is a strong relationship between the number of persons injured and the number of cyclists killed, but this does not represent any important relationship or trend that can impact the analysis.



*Figure 1: Correlation Matrix for Attributes*

**EXPLORATORY ANALYSIS:**

From the pre-processed data we derived some initial insights in the form of visualizations. We plotted a borough wise trend over the years in the form of “Crash by the time”, “crash by year”, “crash by months” and “persons killed and injured”.

The insights obtained were as follows:

For all boroughs, the number of collisions increases from 9 am to 6 pm, with a maximum number of collisions occurring between 5 pm and 6 pm.

There was a significant fall in the number of collisions from 2015 to 2016. This can be attributed to the Vision zero program adopted by the New York City mayor’s office to minimize the number of accidents to zero by 2024. However, the number of collisions is seen to have a minor increase after 2016.

The number of accidents drops significantly in April. This is because NYPD increases the number of traffic personnel during the start of the spring season which is in April.

The number of persons injured is significantly more than the number of persons killed consistently over the years.

*NOTE: The visualizations have been included in the appendix*.

**WEATHER CASE STUDY:**

As a case study, for further research and exploration purpose, we decided to conduct a case study analysis of the effect of weather on collisions in NYC. For this purpose, we considered a subset of the dataset, for the year 2016 and combined it with NYC weather data for the year 2016.

This weather data was obtained the below link.

<https://www.kaggle.com/mathijs/weather-data-in-new-york-city-2016>

This data included attributes like maximum, minimum, and average temperature, precipitation, snowfall, and depth of snow across different months of 2016.

This weather data was combined with the collision data. From the performed analysis and visualizations, the results obtained showed that the weather does not have a significant impact on the collisions.

*NOTE: The visualizations have been included in the appendix.*

# PROJECT GOALS

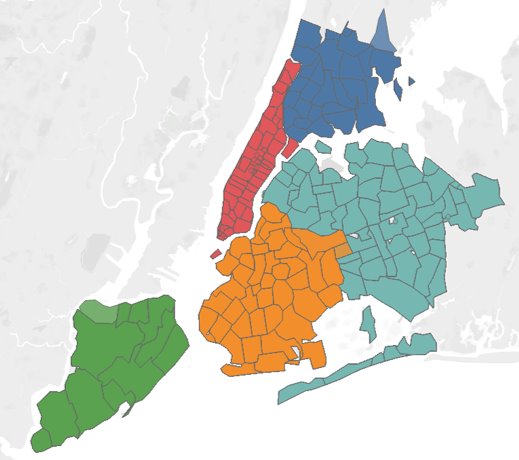
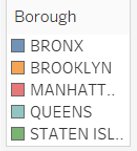
This project does not involve a live client. However, we believe that the analysis obtained from this research can be utilized by the New York City Police Department, NYC Traffic Department, NYC Department of Transportation, etc.

The main goals for this project are as follows:

1. To find patterns involved which might cause vehicular collisions, hence reduce the risk of collision.
2. Suggest recommendations to reduce city traffic fatalities.
3. Forecasting number of collisions for the upcoming years thus helping the authorities to better monitor the number of collision records.
4. Finding the impact of weather on collisions in NYC.
5. Develop suitable graduate quality visualizations to support the performed analysis.

# SOLUTIONS

After pre-processing and cleaning the data, we performed data analysis and data visualization to achieve our project goals and derive conclusions based on visualization and analysis which will help city architects, traffic management, and suburbanite commuters to make informed choices and reduce risk of collisions.



*Figure 2: Layout of the 5 Boroughs*

**DATA VISUALIZATION**

Tableau, being a powerful business intelligence tool for visualization was very helpful to get a picture of what is happening where and what conditions. Using Tableau, we created dashboards and sheets representing the below:

* Number of accidents by the year, month and seasons
* Number of people collisions and injuries each year and in each borough
* Weekly representation of the number of crashes by the day
* Crash time versus the number of crashes involved
* Type of victims involved in collisions like cyclists, motorcyclists, and pedestrians
* Streetwise count of collisions in each borough
* Heat map of collisions by the zip code and vehicle type
* Heat map of collisions by each street and even by the factors causing the collision
* Forecasting for years 2020 & 2021

**DATA ANALYSIS**

Although the correlation matrix did not show any major trends relevant to linear dependencies among the attributes used in the datasets, the visualizations helped in analyzing some other unknown facts and trends that ultimately lead to breaking down to the root of these collisions.

The visualizations mentioned above helped us drill down to the borough, vehicle type, streets, seasons, months, hours, and contributing factors that caused or had the greatest number of collisions.

We started by analyzing each borough leading us to the most dangerous intersections in those boroughs. Visualizations also helped us analyze the intersections that were most dangerous with the help of the heat map leading us to figure our which factors played a major role in contributing to the uncountable collisions in NYC every year.

With the help of Tableau, we also performed a predictive analysis to get a glimpse of how the situation might look like by the end of 2024 so that certain precautions can be taken proactively to avoid any injuries or loss of humans.

# RESULTS

After the visualization and analysis of the NYC collisions data set, where each visualization either represented a story or facts that helped us dig more to identify the problem, we made the below conclusions:

* The maximum number of collisions took place in Brooklyn that mainly involved passenger vehicles.
* The following intersections are the most dangerous in the respective boroughs:

1. Jerome Ave – East Tremont Ave (Bronx)
2. 3rd and 4th Ave (Brooklyn)
3. Broadway – Canal Street (Manhattan)
4. Jamaica Ave – Hillside Ave (Queens)
5. Bedford Ave – Hylan Boulevard (Staten Island)

* The majority of accidents take place on workdays (Mon -Fri) between 4 pm to 6 pm.
* A motorcyclist is the most affected victims in terms of injuries and deaths.
* Manhattan has higher pedestrians’ fatalities involved compared to the other boroughs​ considering it is the smallest borough in NYC.
* Winter season is the most dangerous of all the seasons.
* The small borough with cluttered streets could be the cause of more injuries on Manhattan streets.
* Aggressive driving is the most contributing factor throughout all boroughs.
* Forecasting estimates that by the end of 2020 NYC will witness 117,880 crashes whereas 2021 will see 120, 508 crashes if there is no improvement in city planning and traffic management.

# CONCLUSION

After the entire analysis, we conclude that the below suggestions will help contribute to the improvisation of the NYC traffic situation in the coming years:

* Increase the patrolling during peak hours​
* Make residents aware of dangerous intersections
* Be cautious of cyclists and pedestrians during specific periods​
* Increase the price of traffic tickets for certain traffic rules ​
* Improve roads and traffic lights near specific intersections​

All in all, we achieve the goal of the project to observe and anticipate the patterns in vehicle collisions of NYC and looking at the analysis certain timely measures can be executed to make New York City an accident-prone area which will in future lead other big cities in the world to follow similar guidelines and precautions to avoid human loss by vehicular collisions.

# REFERENCES

Djurkovic, Djurkovic, N., Djurkovic, N. D. N., & Medformis. (2020, April 17). How Many People Die in Car Accidents? (US & Global Stats). Retrieved from <https://policyadvice.net/car-insurance/insights/how-many-people-die-in-car-accidents/>

(NYPD), P. D. (2020, May 8). Motor Vehicle Collisions - Crashes: NYC Open Data. Retrieved from <https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Crashes/h9gi-nx95>

Traffic collisions. (n.d.). Retrieved from <http://people.ischool.berkeley.edu/~ellyrath/data.html>

Waegemakers, M. (2017, September 24). Weather data in New York City - 2016. Retrieved from <https://www.kaggle.com/mathijs/weather-data-in-new-york-city-2016>

Waegemakers, M. (2017, September 24). Weather data in New York City - 2016. Retrieved from <https://www.kaggle.com/mathijs/weather-data-in-new-york-city-2016>

Wilson, D. (2018, May 4). Using Machine Learning to Predict Car Accident Risk. Retrieved from <https://medium.com/geoai/using-machine-learning-to-predict-car-accident-risk-4d92c91a7d57>

Almohimeed, R. (2019, July 23). U.K. Traffic Accidents - Data Analysis (10 years). Retrieved from <https://medium.com/@rawanme/u-k-traffic-accidents-data-analysis-10-years-c81293180ee5>

Gallery: Tableau Public. (n.d.). Retrieved from <https://public.tableau.com/en-us/gallery/nyc-cyclist-collision-report?tab=featured&topic=all&type=featured>

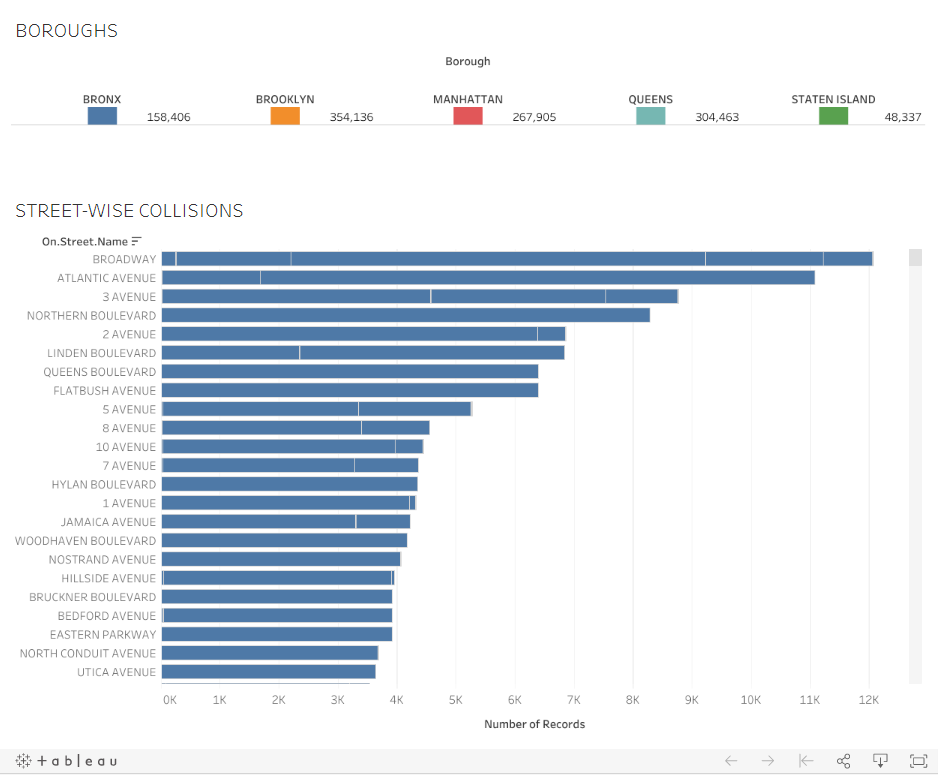
New York City Motor Vehicle Collision Data Visualization. (n.d.). Retrieved from <https://nycdatascience.com/blog/student-works/new-york-city-motor-vehicle-collision-data-visualization/>

Auygur. (n.d.). auygur/ExploratoryDataAnalysis-NyC\_Accidents. Retrieved from <https://github.com/auygur/ExploratoryDataAnalysis-NyC_Accidents>

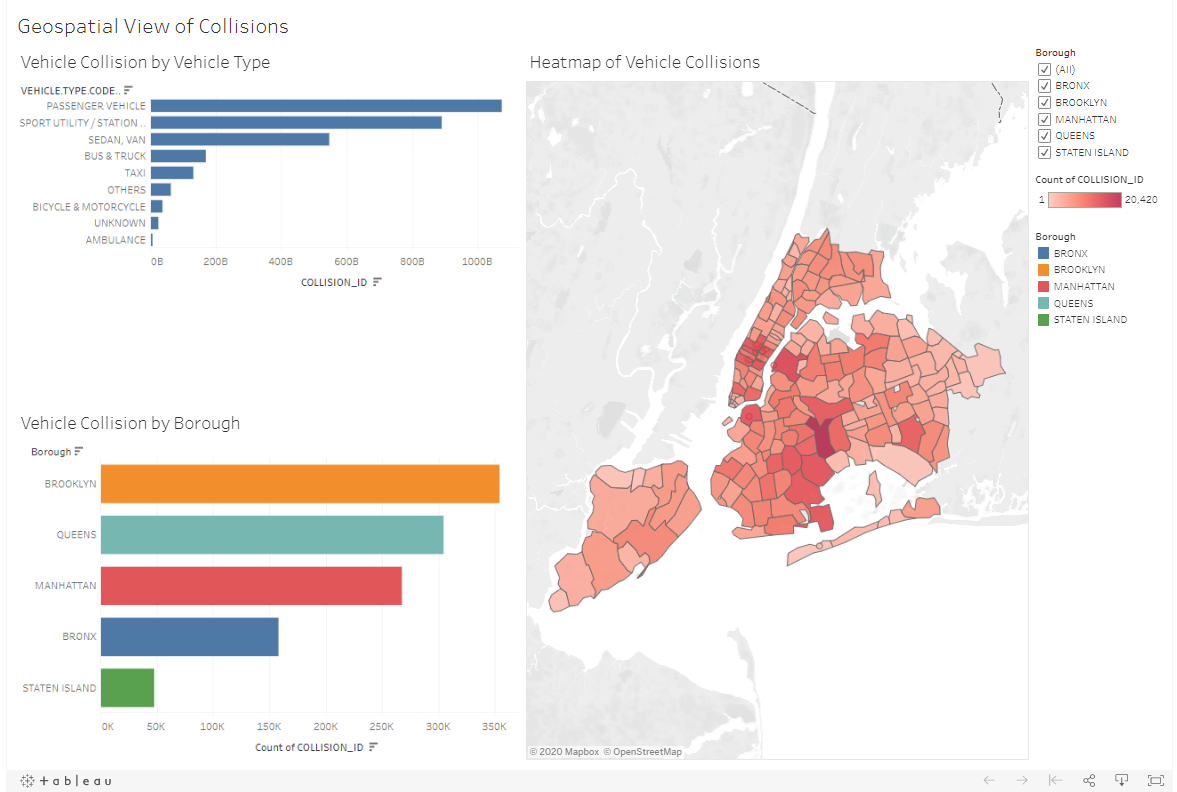
# APPENDIX

This section encloses the visualization and analysis done by our team for the NYC Collisions Dataset (2013-2019). The dashboards snapshots for our project is placed below respectively:

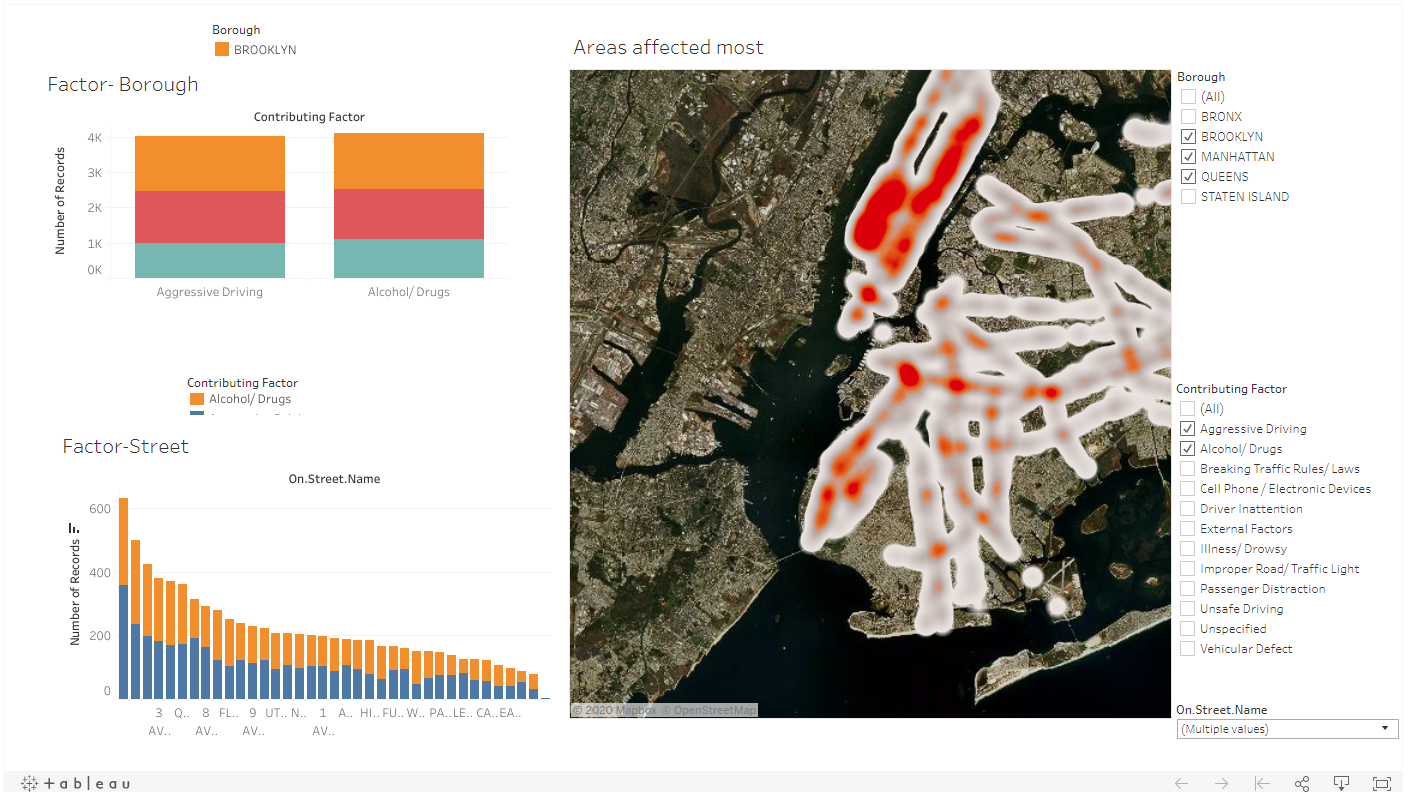
## Dashboard 1: Street Wise Collisions Based on Borough



## Dashboard 2: Heatmap of Collisions by each borough based on vehicle type



## Dashboard 3: Areas affected the most by intensity on the contributing factor

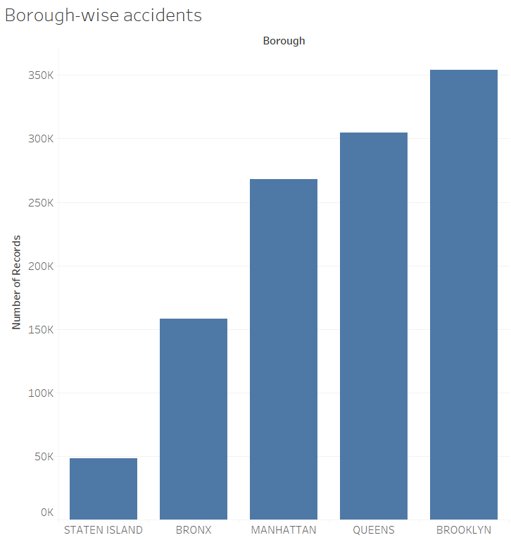


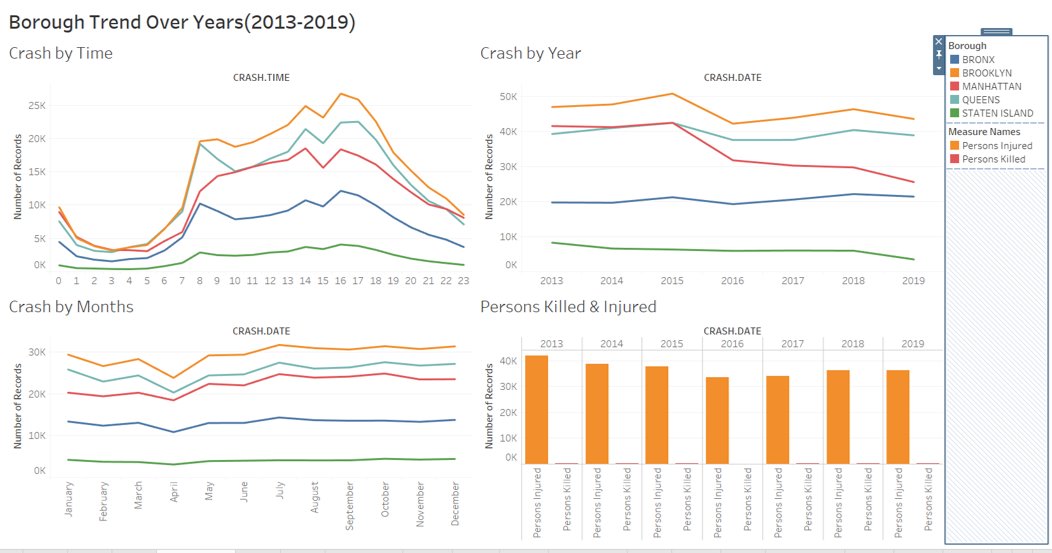
The Tableau Public link to the above three dashboards is listed below:

* <https://public.tableau.com/profile/kaumudi.kulkarni#!/vizhome/Book3-KAUMUDI/Dashboard1>
* <https://public.tableau.com/profile/archana.mahapatra#!/vizhome/VehicleCollisionsStatistics/Dashboard3>
* <https://public.tableau.com/profile/archana.mahapatra#!/vizhome/Madhvika_BI_DashboardNew/StreetIntersectionandFactors>​

Below are the snapshots of some critical visualizations that help form a story of the data that can be used in the future to make important decisions for and by the traffic police of NYC:

## COLLISIONS BY THE BOROUGH:

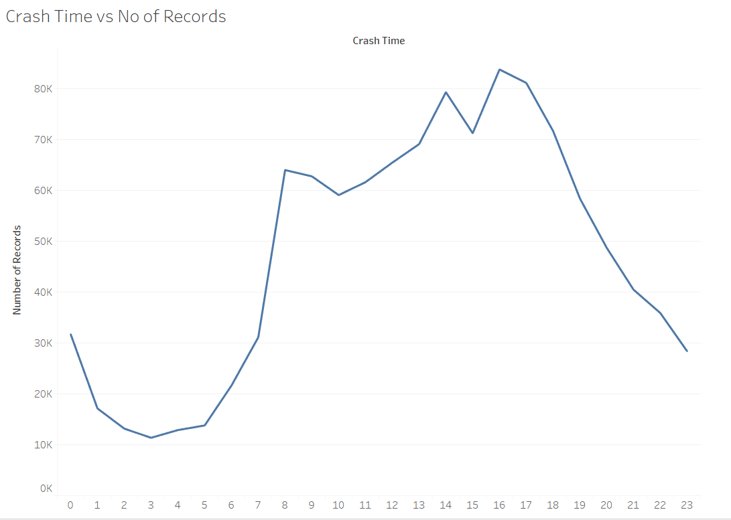




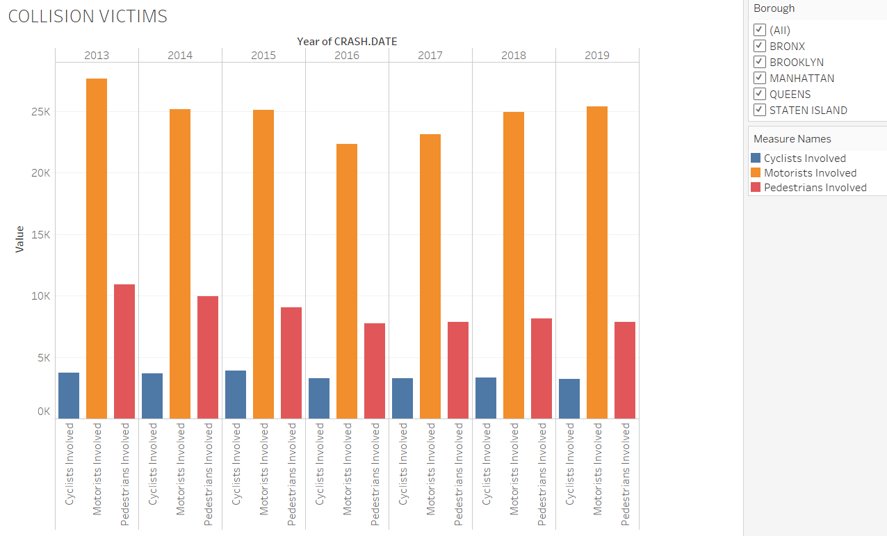
## COLLISIONS BY THE WEEK:



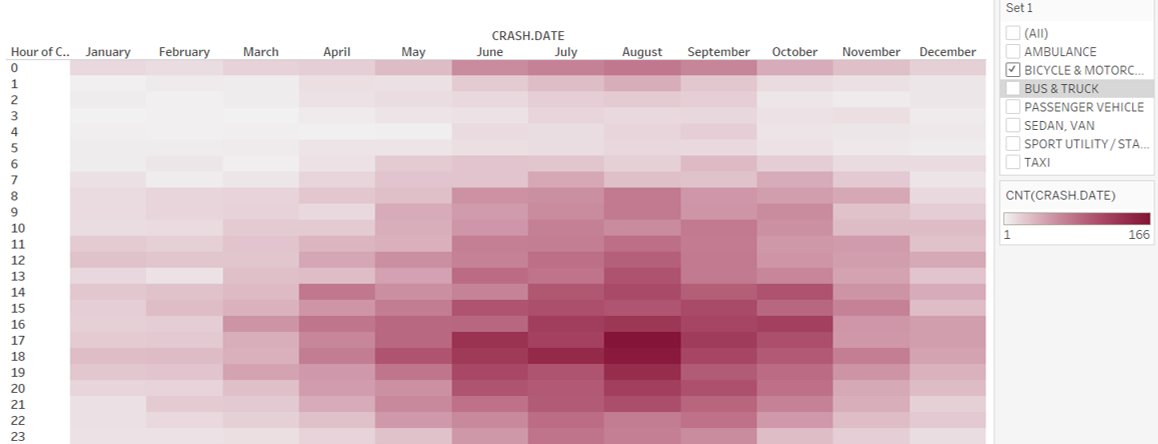
## HOUR OF THE MAJORITY OF COLLISIONS:



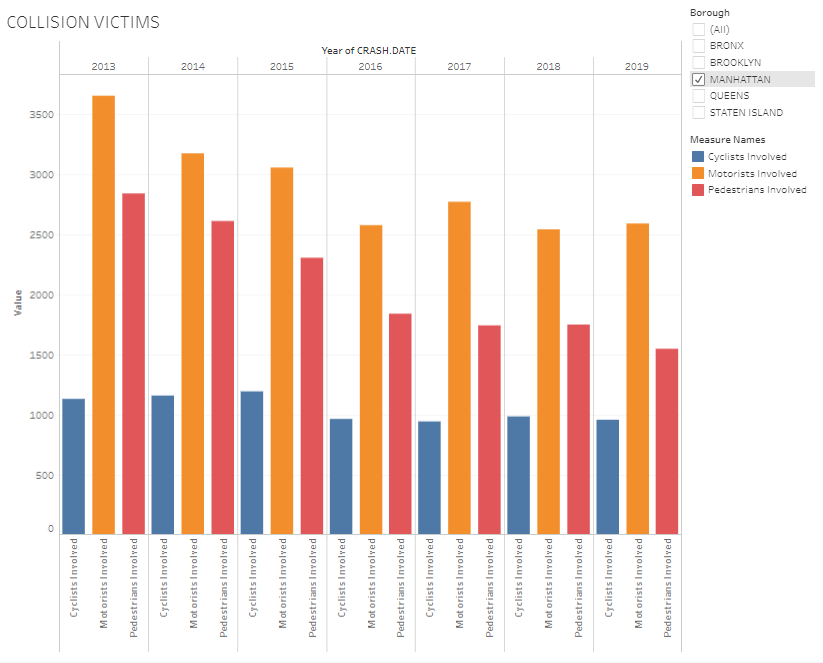
## VEHICLE COLLISION GROUPED BY CYCLISTS, MOTORIST & PEDESTRIANS:



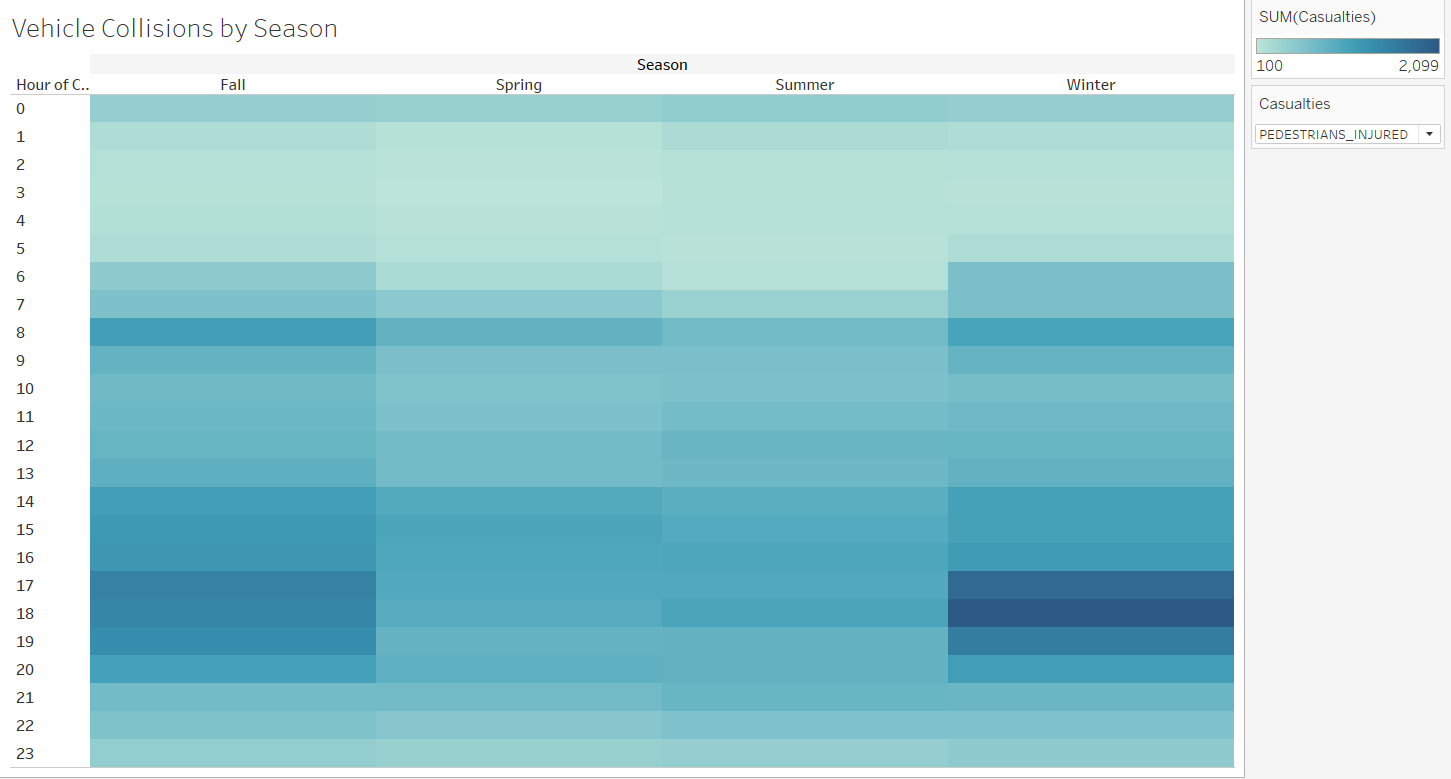
## TWO-WHEELER COLLISIONS BY THE MONTH:



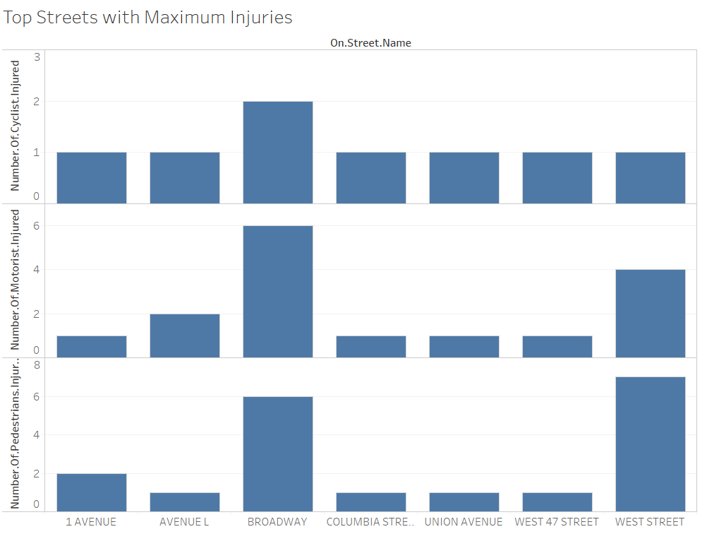
## ANOMALY IN NUMBER OF PEDESTRIANS INVOLVED IN MANHATTAN:



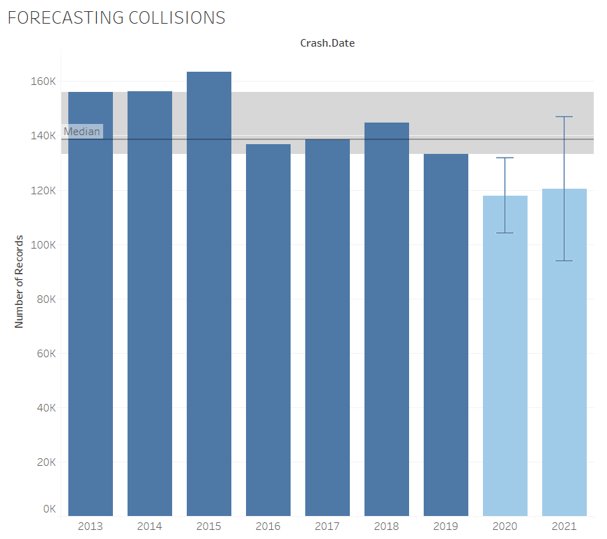
## PEDESTRIANS INJURED BY SEASON



## TOP STREETS WITNESSING COLLISIONS EACH YEAR BASED ON VICTIMS:



## FORECASTING COLLISIONS FOR 2020 & 2021:



## WEATHER CASE STUDY:

