Project Report

Divya, Nupur, Ranadheer, Varun

Data Description:

The dataset NYPD_Motor_Vehicle_Collisions_1000 has data structured into 29 columns with 1000 rows. It consists of data over 8 years beginning from 2012 to 2019. Data captures information regarding accidents activities in the United States considering location, year, town, street, vehicle type, contributing factor, unique key. Data is organized in the following manner:

Date: Specifying the date of accident.
Time: The time at which the accident took place.
Location: Latitudinal and Longitudinal Location of accident.
ZIP Code: Area's zip code where accident happened.
On Street Name: Street where accident happened.
Cross Street Name: The crossing street connected with on street.
Off Street Name: The off street connected with crossing street.
Number of Person Injured: Count of people injured.
Number of Person Killed: Count of people killed.
Number of Pedestrians Injured: Count of pedestrians injured.
Number of Pedestrians Killed: Count of pedestrians killed.
Contributing Factor Vehicle 1: The primary reason for accident.
Contributing Factor Vehicle 2: The secondary reason for accident.
Unique Key: A column uniquely identifying the all the rows.
Vehicle Type Code 1: Type of vehicle involved in the accident.

Data Cleaning-

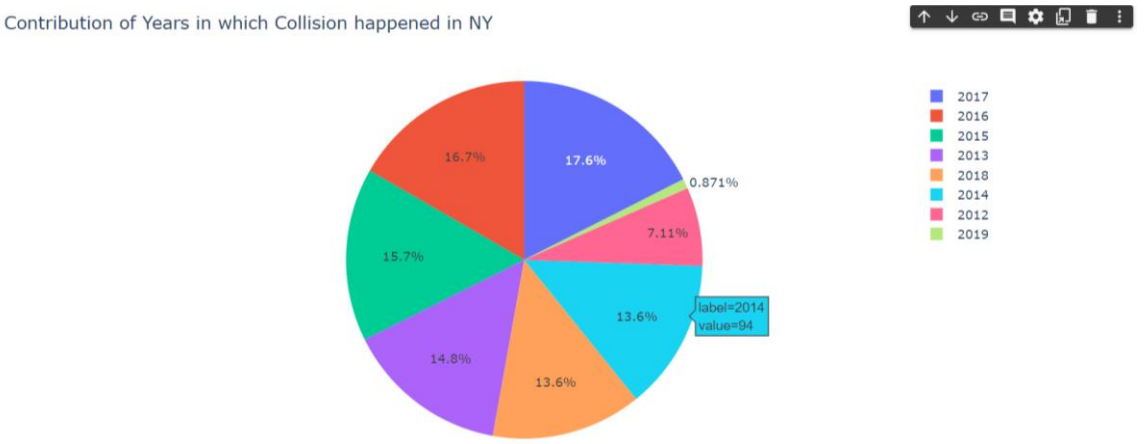
We have performed certain amount of data cleaning to scrutinize the best possible insights from the data set starting from the basic data manipulation of data year wise. This feature helps us extract and explore data from time series visualization and add valuable pinpoints to the analysis. The summation of people on the category of being pedestrians, motorists and cyclists annually has played key role in understanding the difference between no. of persons got injured and no. of persons got killed. Division of data set on interval of 0-4 people interval helped to realize how often group of people were endangered in fatalities.

Hypothesis-

While drawing insights and conclusion, broadly we have assumed a few points to be particularly being the possible shortcomings in the system. Few of dominant reason we assumed for mishaps was the weak infrastructure for transport activities, impoverished base, lack of awareness, disobeying the traffic rules, traffic ingestions, volume of people, lack of safety gears, missing the quality checks and maintenance activities at regular interval, poor technology, limited medical facilities etc. Also, we postulated environmental and social factor contributing to the accidents in particular region.

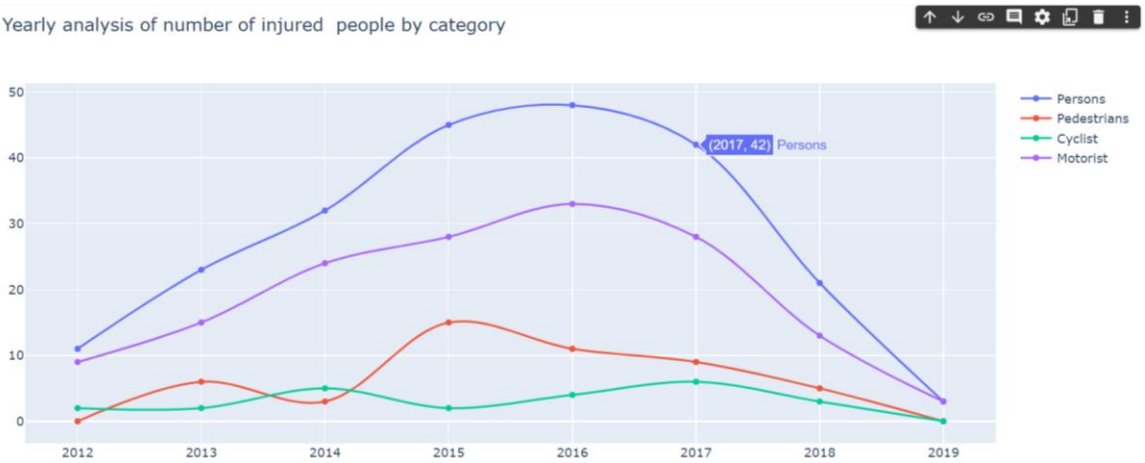
Understanding the Visualizations:

1. **Plot-1 Contribution of Year in which Collision happened in NY through *"Pie Chart"***



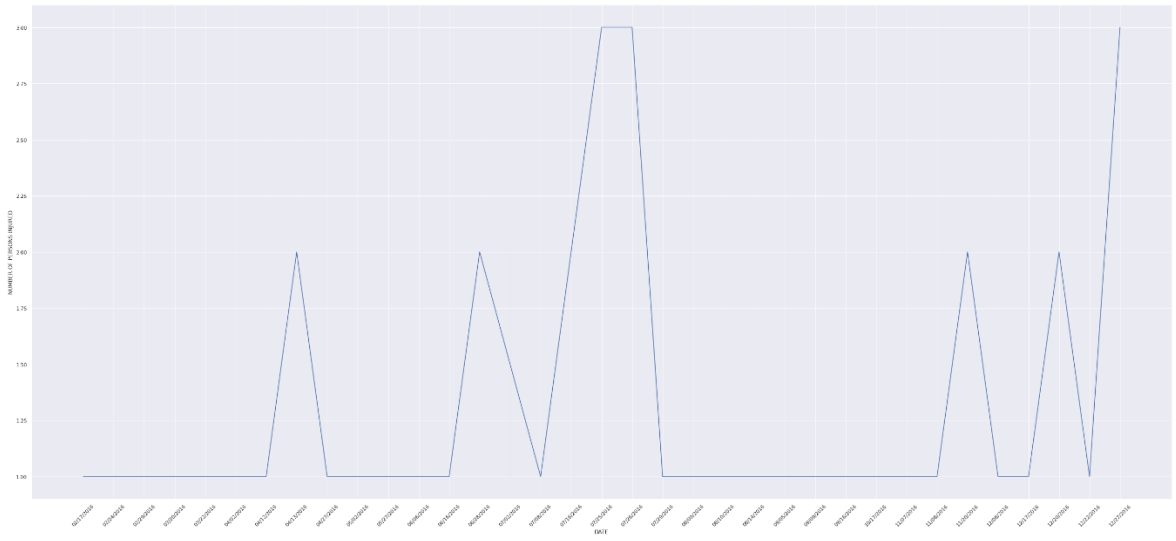
After performing data manipulation, New York emerged as the Hotspot of Collision. Analyzing the total accident count **year wise** from 2012-2019, we found that collisions regularly increased from year **2015-2017** with **1/6th** of total collisions encountered in **2017**. One significant insight from the chart is the reversal of trend from **2018-2019**, which implies the improvement in constructional, infrastructural activities and better traffic rules follow up.

2. **Plot-2 Yearly analysis of injured people by category on "Label Line Annotations"**



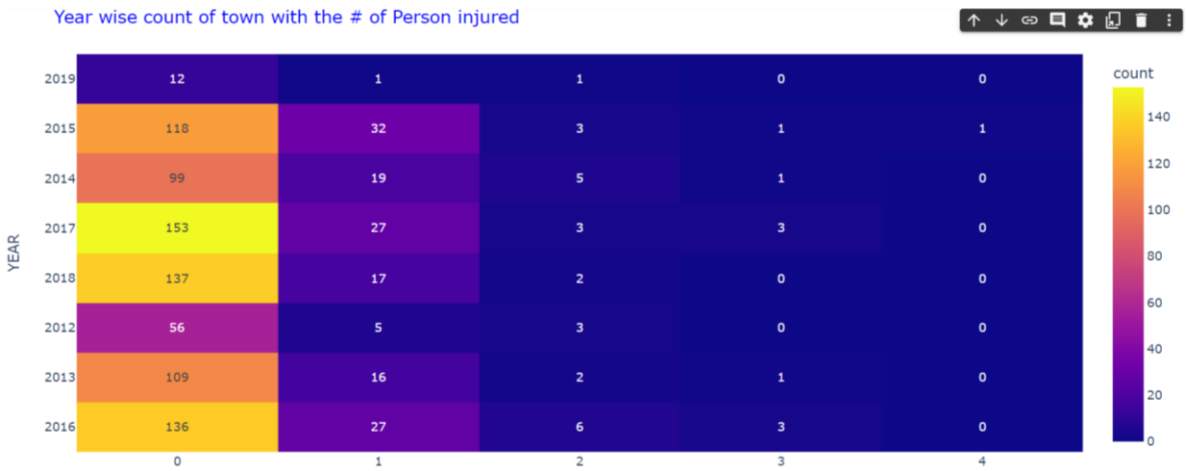
Talking broadly on injured people year-wise, the count has increased very evidently with **2016** at **top** of the chart. On scrutinizing categorically **Motorist** were most prone to injuries based on multiple factors like avoiding safety gears, over speeding vehicles etc. The number of injuries for pedestrians and cyclists always **remained under 10 until 2014**. The pattern changed **2016 onwards** with improvement of safety gears, introduction of automated cars, and sheer amount of vigilance.

3. Daily level of breakdown of Injured people for year 2016 through “Line Chart”



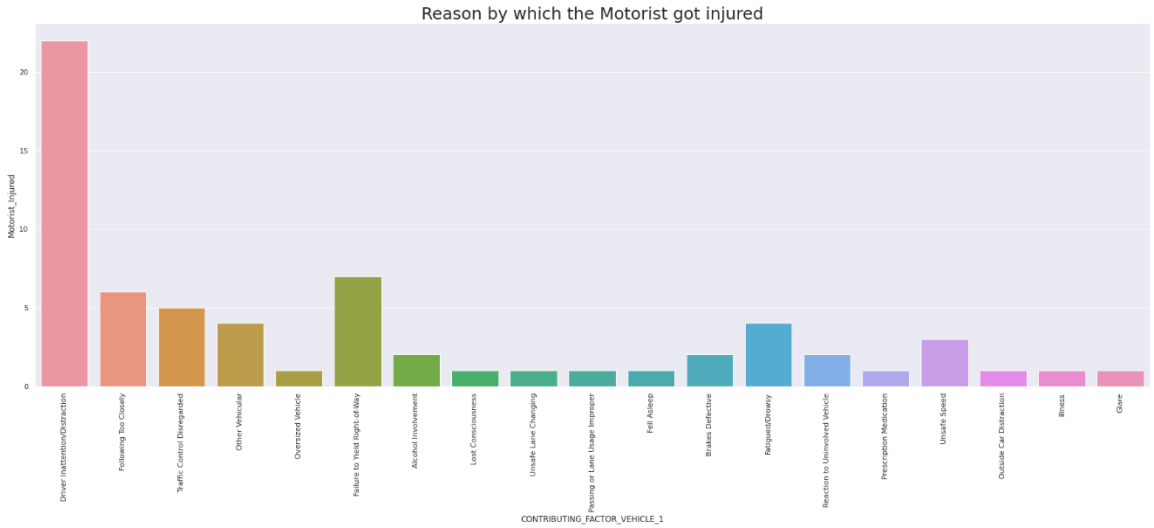
As 2016 had inflated injuries, so emphasizing **2016 monthly**, which is having **maximum** count of injuries overall, we found that in an interval of almost three-four months starting from **April**, then **July**, followed by **October** and **December** the injuries were more often. Holidays, the change of seasons, beginning of session etc. can be the contributing factors that made people travel during that time interval and some way or the other the poor infrastructural status made accidents and injuries frequent.

4. No. of people injured year wise in the town through “Heat Map”



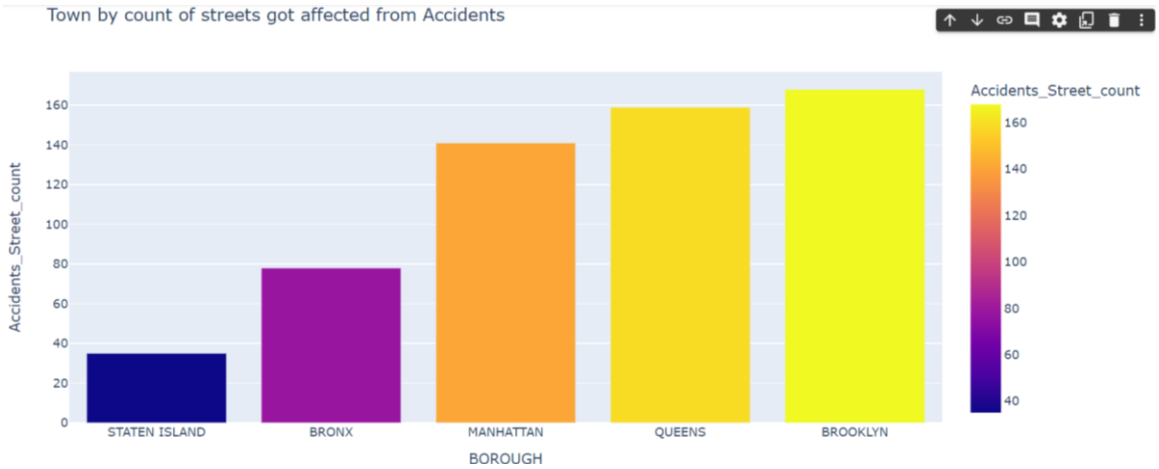
On a scale of **0-4 people** injured year-wise, the positive sign was the numerical data being in **single digit** having **2 or more people** under the radar of injuries. Also, the first column of heat map indicates that notable amount of people **didn’t get injured** even though they were involved in struggle.

5. Plot-5 Reason by which Motorist got injured through “Bar Chart”



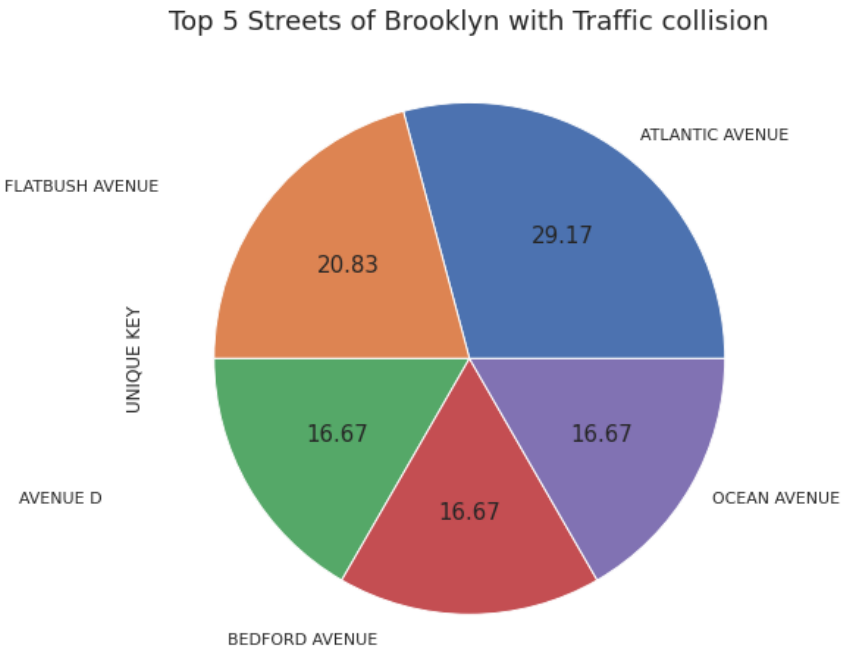
Taking injuries of **Motorist** into light, the possible reason of grapple comes out to be the Driver’s **lack of attention**/ any distraction, **Failure** to yield right of way, following too closely, **Traffic Control** disregarded etc. before the mishap.

6. Plot-6 Town by count of street got affected by Accidents through “Bar Graph”



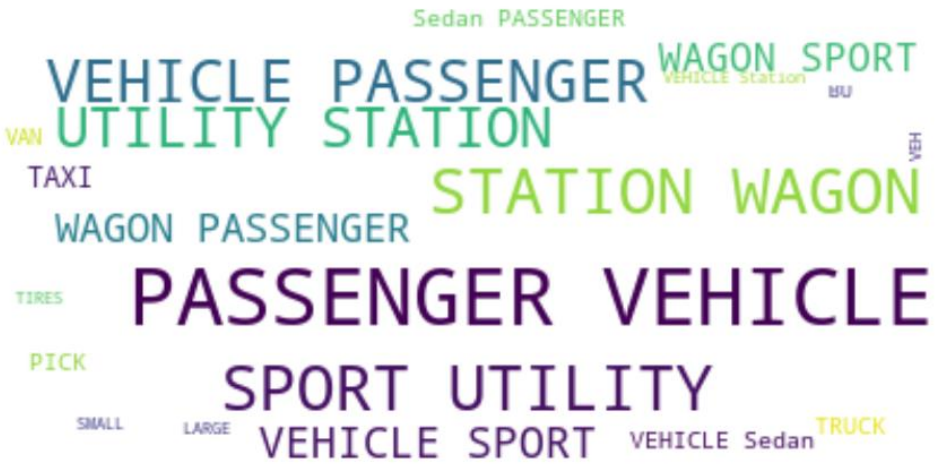
Narrowing our vision towards **Town**, we realized that **Brooklyn** and **Queens** had the most No. of accidents cumulating around **60% of total accidents** across all the towns. The traffic ingestion, lean roads, and volume of people at that town can be the dominant reason of tragedies there.

7. **Plot-7 Top 5 Street of Brooklyn with Traffic Collison through “Pie Chart”**



Since **Brooklyn** had the **most accident count**, we further investigated our dataset to find the demographic region through street to conclude accurate insights. **Atlantic Avenue** and **Flatbush Avenue** had **50% of fatalities** in Brooklyn, the area being less of developed, having narrow routes and high transportation activities are some valid shortcomings and contributing factors to accident.

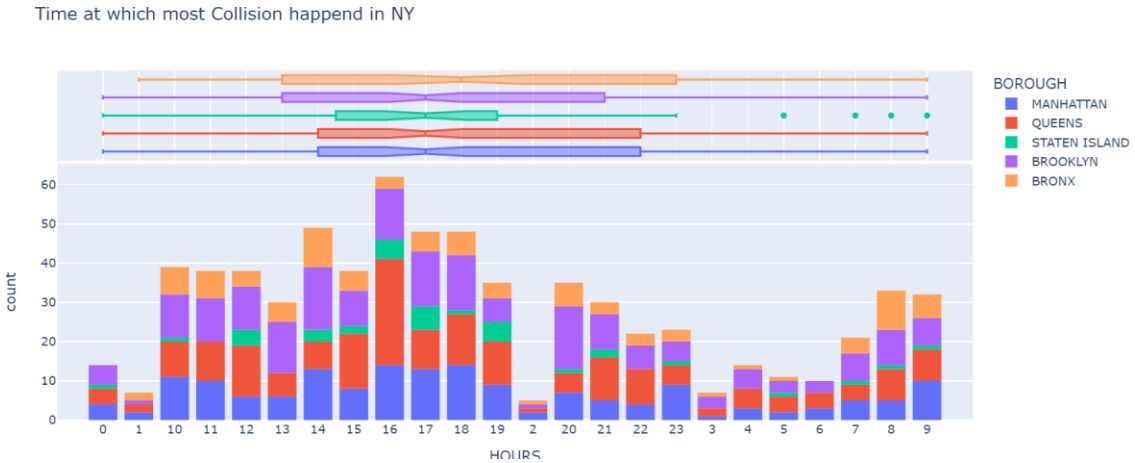
8. **Plot-8 Most frequent vehicle types responsible for collision through “Word Cloud”**



Clearly from the Word cloud, it is clear enough to identify that the **Passenger Vehicle**, **Station Wagon**, and **Sport Utility** are the most common vehicles that are frequently

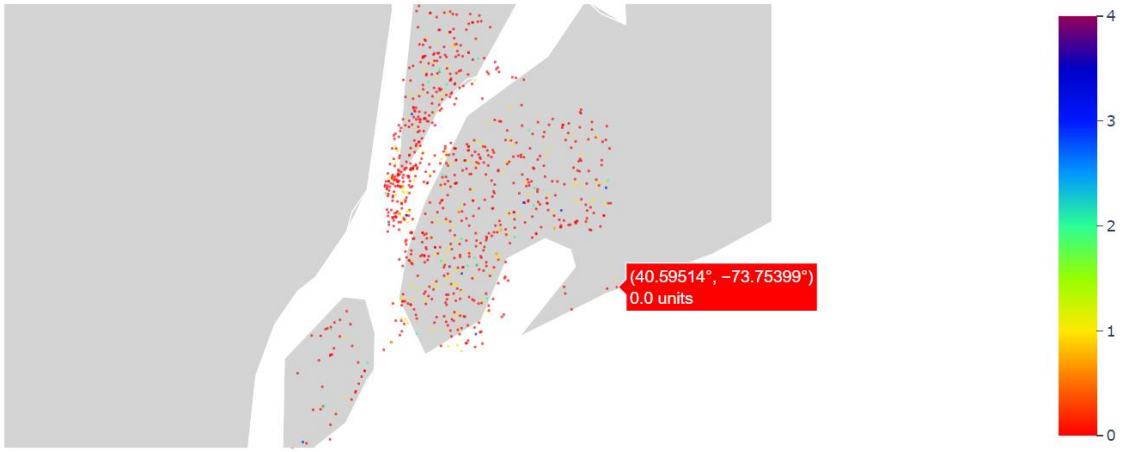
involved in accidents across the country. All these vehicles are **Mass vehicle**, which require proper maintenance and quality check at regular interval to safeguard life of the people.

9. Plot-9 Time at which Most Collision happened through “Stacked bar and Boxplot”



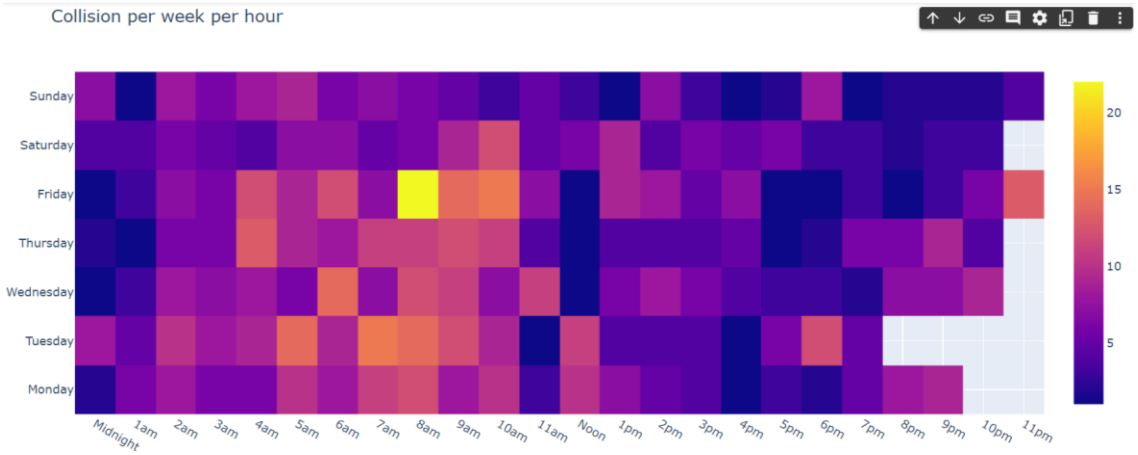
On a **24-hour format** when we analyze the no. of accident **town wise**, we identified an **exceptional insight**, the conclusion till now indicated that Brooklyn had maximum casualties but at **4 PM** when the no. of accidents peak, the count was maximum at **Queens**. The accidents were more often at **4 to 7 PM** in **evening**, the time when most of the people returned from office, shops and factories, students from school and college etc.

10. Plot-10 Locating Latitude and Longitude of Collision through “Choropleth”



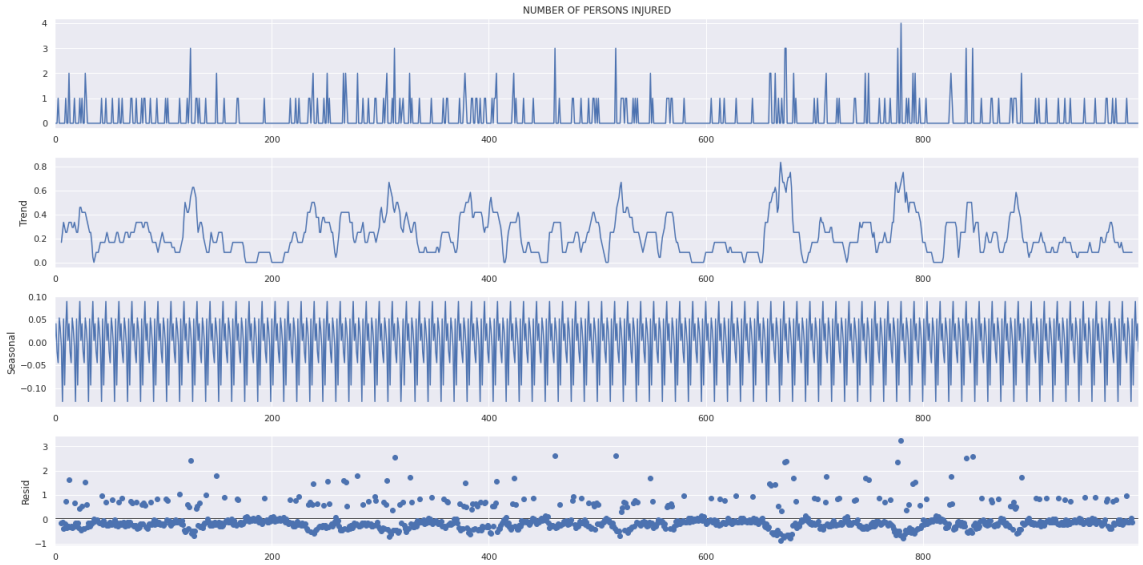
Locating **demographically** the **coordinates** through latitudes and longitudes of the collision in NY region. This helps to have an overview of the **most common area** that has been in the hotspot of accident. It provides valuable insights to improve the area and decrease the accident count.

11. Plot-11 Collision per week per hour through “Heatmap”



Analyzing the dataset **weekly** on a scale of **24-hour format**, collisions were **more frequent** in **morning hours** on **weekdays** around **8-10AM**. This time is usually the **rush hour** for reaching office, school, shops, industries etc. which can be the primary cause of casualties.

12. Plot-12 Trend and seasonal analysis of Vehicle Collison through “Decomposition”



Decomposing the Number of persons injured on the grounds of trend, seasonal, raised functionality. It shows a **multi-dimensional view** of the number of persons injured for getting profound knowledge of its trends.

Conclusion-

The NYPD_Motor_Vehicle_Collisions_1000 was used to investigate accidents across New York in the past 8 years ranging from 2012-2019. In this we have critiqued several data visualizations, learned common pitfalls, and presented solutions along the way. Various plots have been laid out systematically to analyze the trends year wise, location wise, on grounds of type of vehicle engaged etc. involved in the Accident count of people based on the number of parameters. Major trends were captured using graphical tools like Bar Graph, Choropleth, Word Cloud, Pie Chart and Label Line with annotations to name a few, these visualizations allowed us to synthesize and explore spatiotemporal conditions in New York more effectively.