Northeastern University

Final Project Report
on
Motor Vehicle Crash Data Analysis

Submitted by

Parth Dhameliya

Brijesh Savaj

Zeel Jodhani

Fenil Savani

Rishita Bhutani

Dr. Sivarit Sultornsanee

Subject

Computation and Visualization

Submitted to
Prof. Sivarit Sultornsanee

College of Engineering

Northeastern University, Boston

College of Engineering

Northeastern University

Dec – 14, 2022

INDEX

CHAPTER 1
INTRODUCTION

1.1 Introduction

The main goal of data visualization in this project is to communicate information clearly and effectively through graphical means. It doesn't mean that data visualization needs to look boring to be functional or extremely sophisticated to look beautiful. To convey ideas effectively, both aesthetic form and functionality need to go hand in hand, providing insights into a rather sparse and complex data set by communicating its key aspects in a more intuitive way.

90% of all data has been created in the last two years. With so much data, it's become increasingly difficult to manage and make sense of it all. It would be impossible for any single person to wade through data line-by-line and see distinct patterns and make observations. Data proliferation can be managed as part of the data science process, which includes data visualization.

1.2 Problem Introduction

Motor Vehicle Accidents are by large the most common causes of personal injury, and the leading non-natural killer in the world today.

MVA injuries accounts for nearly 25% of the 5 million people who die from injuries each year in the USA.

Almost as many people riding motor vehicles are treated by cities EMS and every year thousands of drivers are injured, put in danger, or delayed by collisions with other vehicles.

While only a handful of these crashes are fatal, every tragedy leaves a trail of grieving family and friends, and the despair of unfulfilled potential.

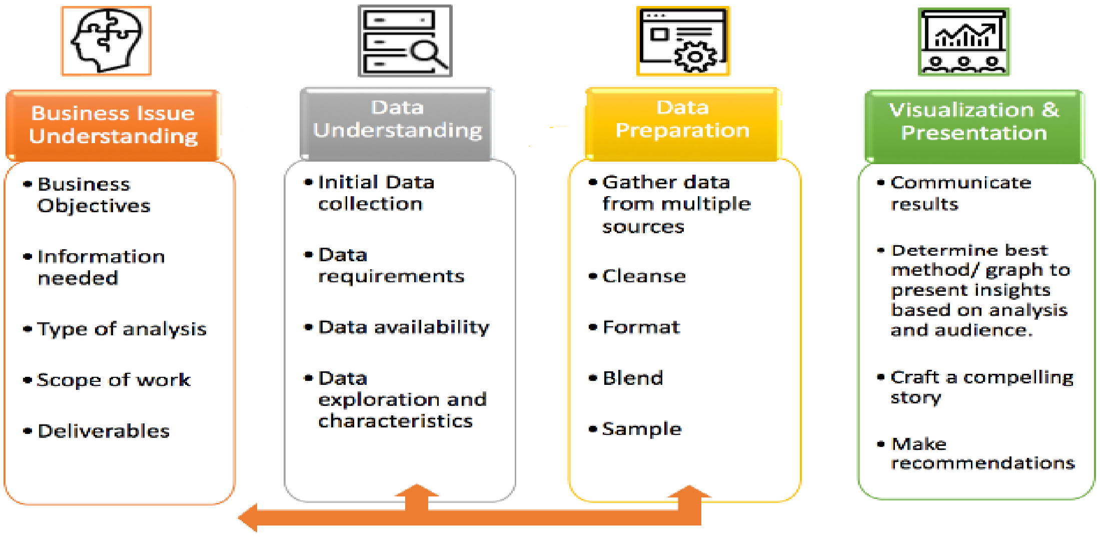
CHAPTER 2

DATA ANALYSIS

2.1 Data Source

The Motor Vehicle Collisions crash table contains details on the crash event. Each row represents a crash event. The Motor Vehicle Collisions data tables contain information from all police reported motor vehicle collisions in NYC. The police report (MV104-AN) is required to be filled out for collisions where someone is injured or killed, or where there is at least $1000 worth of damage. It should be noted that the data is preliminary and subject to change when the MV-104AN forms are amended based on revised crash details. For the most accurate, up to date statistics on traffic fatalities, please refer to the NYPD Motor Vehicle Collisions page (updated weekly) or Vision Zero View (updated monthly).

2.2 Cycle of Data Analysis


Business Issue Understanding
• Business Objectives
• Information needed
• Type of analysis
• Scope of work
• Deliverables
Data Understanding
• Initial Data collection
• Data requirements
• Data availability
• Data exploration and characteristics
Data Preparation
• Gather data from multiple sources
• Cleanse
• Format
• Blend
• Sample
Visualization & Presentation
• Communicate results
• Determine best method/ graph to present insights based on analysis and audience.
• Craft a compelling story
• Make recommendations

Step – 1) Understanding the Business Issue

At the start of the project, the focus is to get a clear understanding of the overall scope of the work, business objectives, information the stakeholders are seeking, the type of analysis they want you to use, and the key deliverables. Defining these elements prior to beginning the analysis is important, as it helps in delivering better insights. Also, it is important to get a clarity at the beginning as there may not be another opportunity to ask questions before the completion of the project.

Step – 2) Understanding Dataset

This phase starts with an initial data collection and proceeds with activities like data quality checks, data exploration to discover first insights into the data, or to detect interesting subsets to form hypotheses for hidden information. There are a variety of tools we can use to understand the data. Depending on the size of the dataset, we used Excel for manageable datasets, or used more rigid tools like Python, Tableau to explore and prepare the data for further analysis.

Key things to remember would be to identify key variables of interest to study the data, look for errors (omitted data, data that doesn't logically make sense, duplicate rows, or even spelling errors) or any missing variables that need to be amended so we can properly clean the data.

Step – 3) Data Preparation

Once the data has been organized and all the key variables have been identified, we can begin cleaning the dataset. Here, we will handle missing values (replace with means, drop the rows or replace with the most logical values), create new variables to help categorize the data, and remove duplicates. Data preparation tasks are likely to be performed multiple times, and not in any prescribed order. After this step, the final dataset is ready to be fed into a modelling tool for further analysis.

Throughout the data preparation process the need is to develop an ever-increasing understanding of the data's structure, content, relationships, and derivation rules. It is imperative to verify that the data exists in a usable state, and its flaws can be managed, and understand what it takes to convert it into a useful dataset for reporting and visualization. In such a scenario, leveraging Data profiling can help explore the actual content and relationships in the enterprise' source systems.

Step – 4) Visualization

Creation of the model is generally not the end of the project. Even if the purpose of the model is to increase knowledge of the data, the derived information will need to be organized and presented in a way that is useful to the customer. Depending on the requirements, this step can be as simple as generating a report or as complex as implementing a repeatable data scoring (e.g., segment allocation) or data mining process.

In many cases, data visualization will be crucial in communicating your findings to the client. Not all clients are data savvy, and interactive visualization tools like Tableau are tremendously useful in illustrating conclusion of this projects. Being able to tell a story with your data is essential. Telling a story will help explain to the client the value of your findings.

2.3 Data Processing

The first set of operations we performed with the uploaded data frame are as follows:

we have assigned a new Data Variable to store the same dataset while keeping the original data frame intact for visualization purposes. Here, we are replacing the unknown values in our DataFrame using the np.nan function. we are transforming the categorical variable into numeric variables using labelencoder () from the library sklearn. The variables on which we are performing this operation here. After changing the categorical variable to a numerical variable, we are removing the unwanted column which had more null values. To double-check null values, we have omitted all non-important object datatypes and selected only the float64 and int64 datatypes. We are using the isnull().sum() function through which we will check our dataframe. Then, we are trying to check for outliers using a predefined statistical function called find_outliers_IQR() which uses the concept of finding outliers using the Interquartile formula. Now, we are using the function on one features "Person Injured" to find the outliers.

2.4 Analytics Software

We are using Plotly libraries for defining all graphs in this Data Visualization task and after creating this all graph we used streamlit web development tools to create dashboard. This dashboard created through the visual studio editor to deploy chart in streamlit platform.

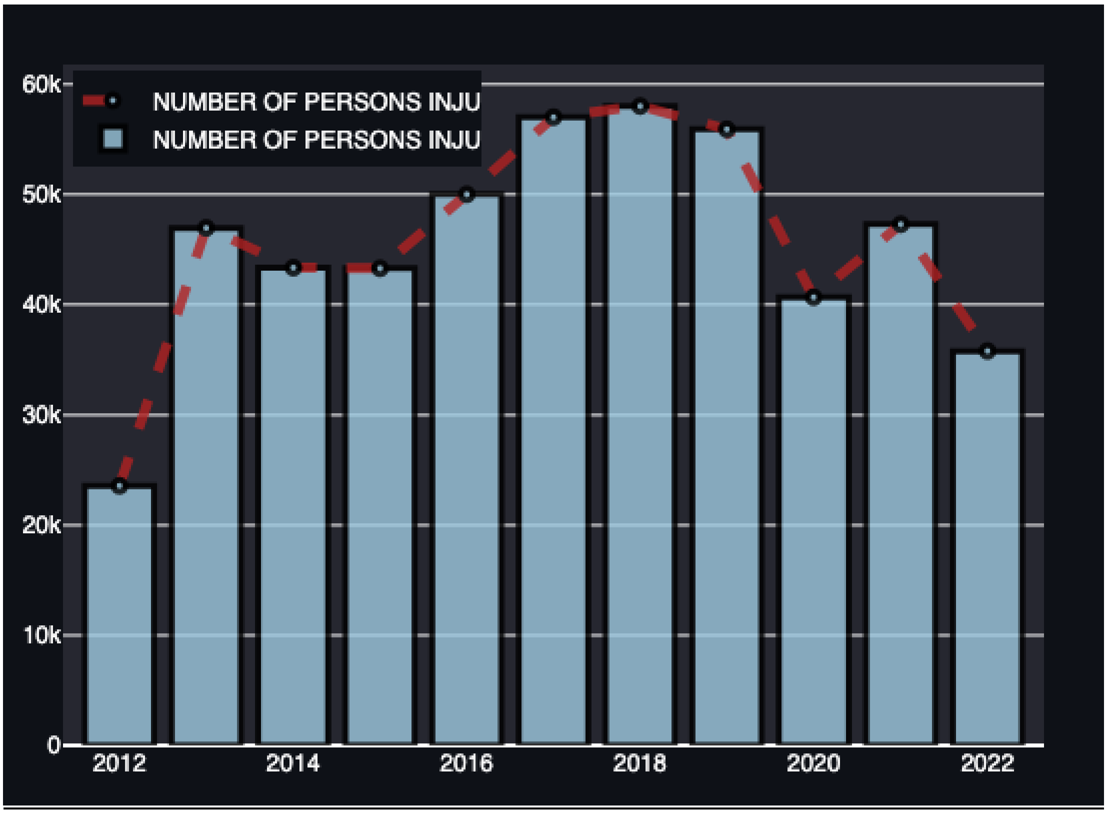
Tools: Streamlit, Visual Studio, Jupyter.

CHAPTER 3

DATA EXPLORATION

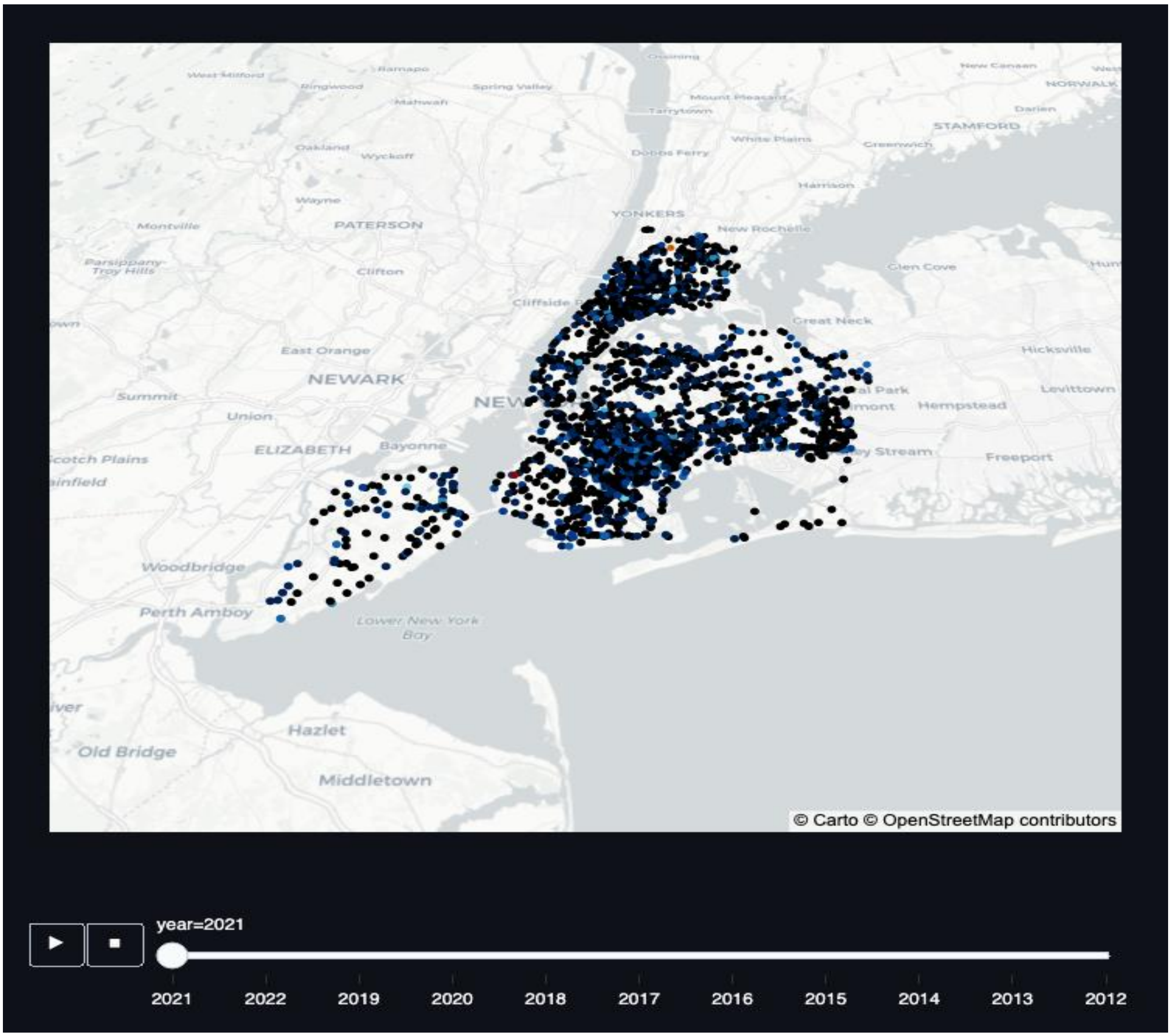
3.1 Visualization Chart

Chart – 1. Bar char with line Chart



Plotly figures are not constrained to representing a fixed set of "chart types" such as scatter plots only or bar charts only or line charts only: any subplot can contain multiple traces of different types. In this Plotly data visualization, we have plotted bar and line charts together which is usually called Plotly combination chart in python.The chart shows the number of people injured in particular years between 2012 to 2022. This data represents all areas in New York City.

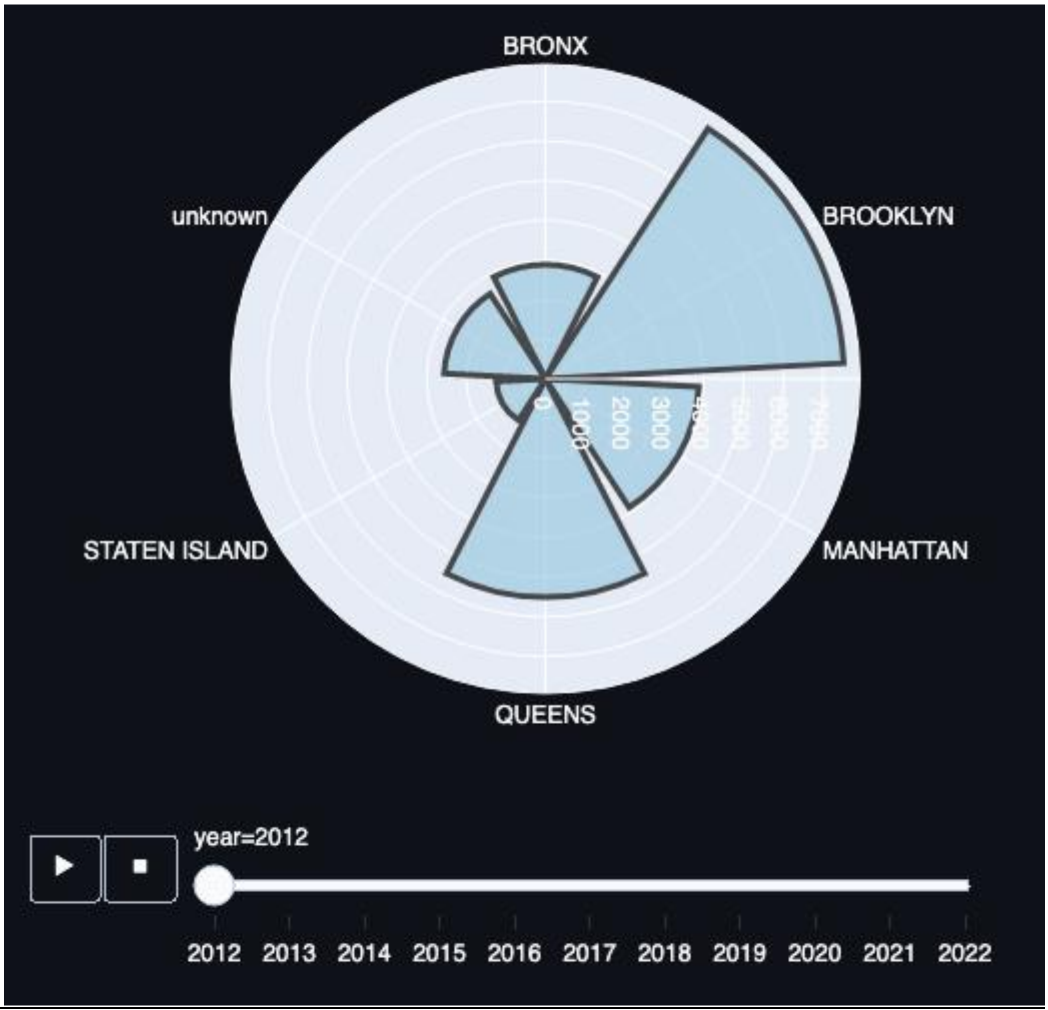
Chart – 2. Scatter Mapbox



Making interactive maps can be a challenge, but there are some great libraries and other cool stuff that can make things much easier by using Plotly is a really great library that allows you to do all kinds of maps (along with other graphs), and the open-street-maps feature, it allows us to zoom in and out at various heights above your target area on the earth and gives you the basic set of cities, roads, national parks, and other features to make a map complete. From these libraries, Mapbox maps are tile-based maps. mapbox object in the figure contains configuration information for the map itself.

Our project data includes the latitude and longitude of where each person was injured by a vehicle accident, so one of the things I tried was mapping them with Mapbox and plotly Express. To represent New York City we used centroid of longitude and latitude in this map to zoom in on specific boundaries of the geospatial map. This map shows the places of death of the people injured. We have assigned animation frames to the map between 2012 and 2022 in streamlit dashboard. the opacity of the dots' Color changes according to the density of injuries that happened in particular longitude and latitude area. We have applied a filter on the street area where people injured more than 3 where we displayed opacity in the map.

Chart – 3. Radar Chart



A Radar Chart (also known as a spider plot or star plot) displays multivariate data in the form of a two-dimensional chart of quantitative variables represented on axes originating from the center. The relative position and angle of the axes is typically uninformative. It is equivalent to a parallel coordinates plot with the axes arranged radially.

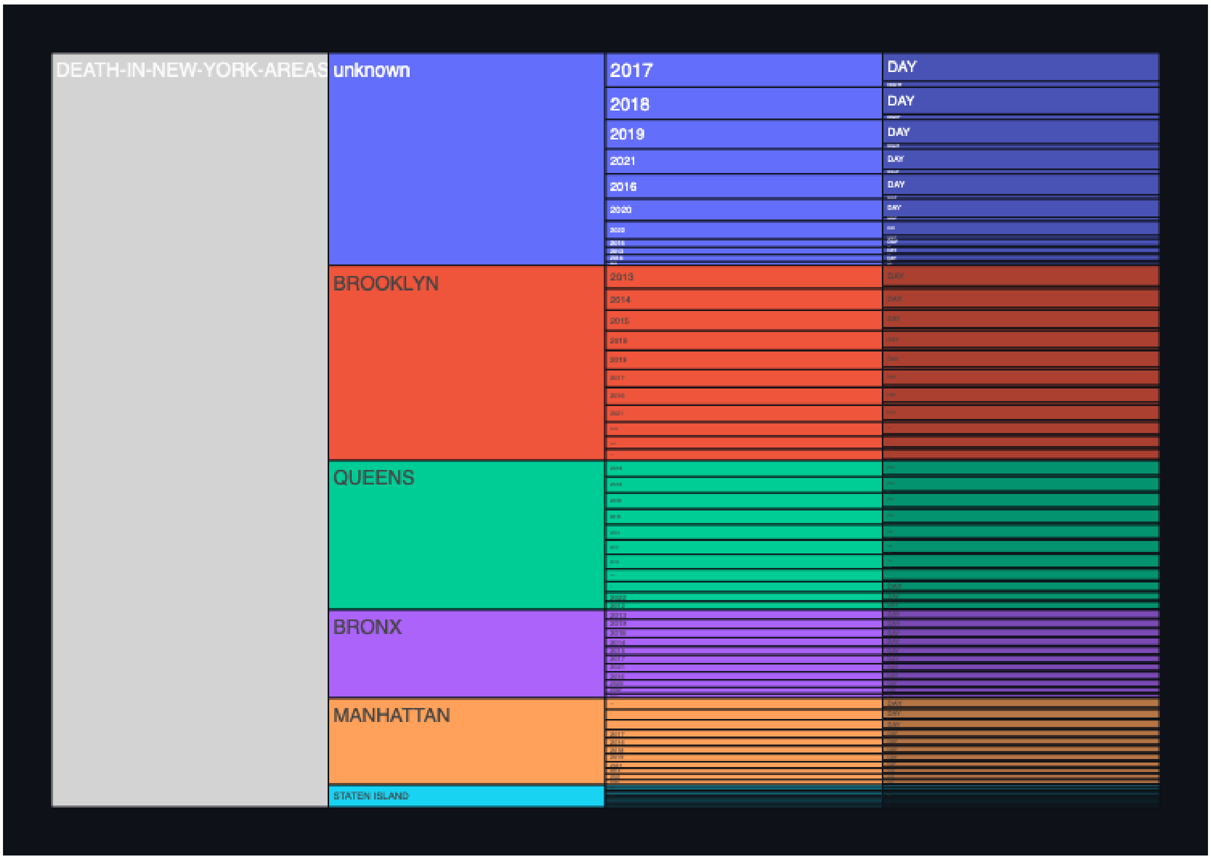
syntax of radar chart in plotly:

fig = px.line_polar(df, r='r', theta='theta', line_close=True)

here, 'r' stands for radius and 'theta' is category. In our chart 'r' represented the number of person injured, person killed, cyclists injured, cyclists killed, etc..

to show all years data we have added animation in bottom of chart. That include year of 2012 to 2022.This radar charts shows that how many person were killed or injured in which borough in perticular year. The lowest number of perosn injured in staten island while highest number of person injured in Brooklyn.

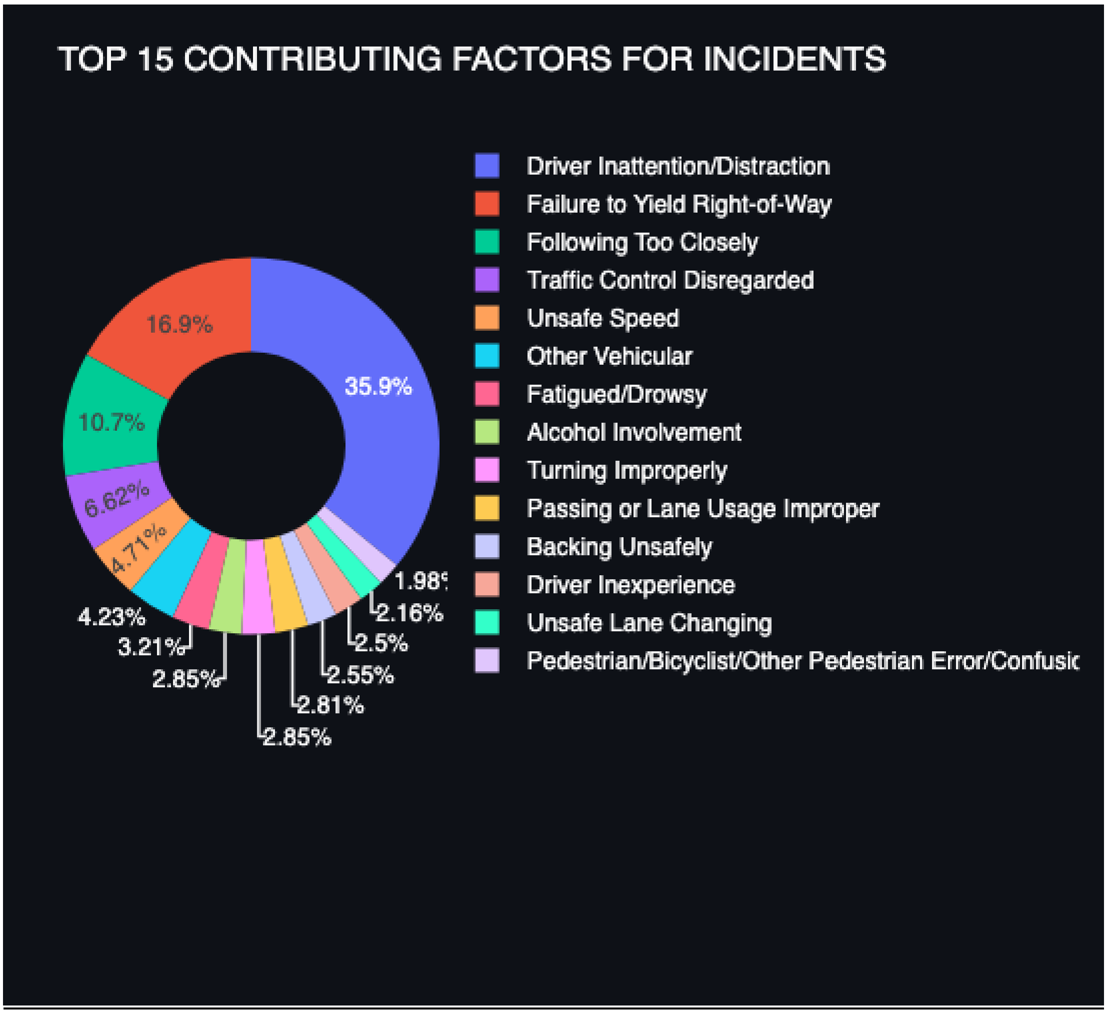
Chart – 4. Icicle Chart



Icicle charts visualize hierarchical data using rectangular sectors that cascade from root to leaves in one of four directions: up, down, left, or right. Similar to Sunburst charts and Treemaps charts, the hierarchy is defined by labels (names for px.icicle) and parents attributes. Click on one sector to zoom in/out, which also displays a pathbar on the top of your icicle. To zoom out, you can click the parent sector or click the pathbar as well.

This chart main root directory is Death in New York area and root node represent different Borough and sub nodes of this part is year range between 2012 to 2022. With different columns corresponding to different levels of the hierarchy. px.icicle can take a path parameter corresponding to a list of columns like day timing, and another features in columns.

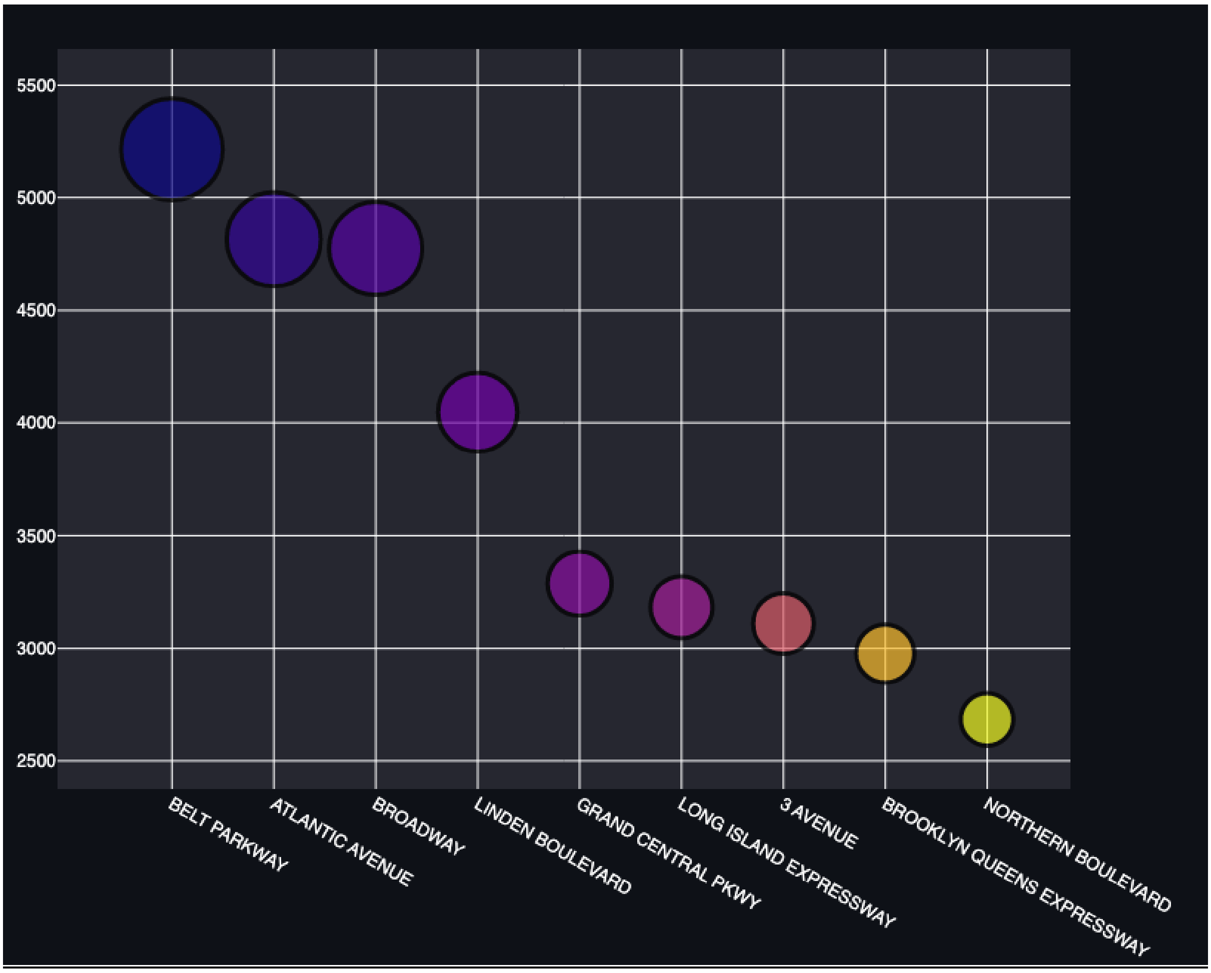
Chart – 5. Donut Chart



A Donut chart is a circular statistical chart, which is divided into sectors to illustrate numerical proportions. Here, we have created this plot with help of plotly express which is easy-to-use, high-level interface to Plotly, which operates on a variety of types of data and produces easy-to-style figures. In order to improve visualization and minimize the interface, we used a variant of pie chart known as a donut chart. Donut charts are made using the hole attribute. Our analysis of the leading contributing factors to motorcycle vehicle collisions led us to generate a total of 52 different types of contributing factors, but for better visualization we narrowed it down to the top 15 contributing factors chart, which is easier to visualize.

We can see from plotting top 15 contributing factors that derives attraction and distraction are the largest factor contributing to crashes, with 35.9%, followed by other factors.

Chart – 6. Bubble Chart



A bubble chart is a scatter plot in which a third dimension of the data is shown through the size of markers. The size of markers is set from the dataframe column given as the size parameter. Using our generated bubble, we are able to retrieve the top New York street that has had the most incidents. According to motor collision data, Belt Parkway street was most dangerous, while Northern Boulevard was least dangerous. By using radar chart, we can verify that the results of the bubble chart distribution are accurate. Use sizeref to scale a bubble's size. The following formula could be used to calculate a sizeref value: sizeref = 2. * max(array of size values) / (desired maximum marker size ** 2).

CHAPTER 4

CONCLUSION AND REFERENCES

4.1 Conclusion

Defining the project Behind this project is an organizational need. The need could be as simple as a weekly, monthly and annually visualization dashboard or a sophisticated predictive recommendation engine. Addressing these needs with concrete measurable objectives provide the right framework to deliver the right information in right fashion. The communication of Key Performance Indicators (KPIs) from the end product is very important for the consumers. To do this, you need to collect requirements, set design processes.

Emerging sources of intelligence, theoretical developments and advances in multidimensional imaging are reshaping the potential value that analytics and insights can provide, with visualization playing a key role. The principles of effective data visualization won't change. However, nextgen technologies and evolving cognitive frameworks are opening new horizons, moving data visualization from art to science. Looking back, much attention has been given to the principles of effective data visualization, such as substance, context and actionability.

More than anything else, data visualization should facilitate decision-making**,** a goal that is difficult to achieve for many. According to a recent [KPMG study - 2015], while data and analytics are deemed increasingly important to organizations, generating actionable insights remains a top challenge.

4.2 References

Data link:

https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Crashes/h9gi-nx95

Documentation:

https://mschermann.github.io/data_viz_reader/conclusion-1.html