PAIR Q

Declaration on Plagiarism

Name/s:	Teena Sharma and Teenu Prathyush
Student Number/s:	21261593 and 21262966
Programme:	MSc. in Computing (Data Analytics)
Module Code:	CA682
Assignment Title:	Data Visualisation
Submission Date:	26-11-2021
Module Coordinator:	Dr Suzanne Little

I/We declare that this material, which I/we now submit for assessment, is entirely my own work and has not been taken from the work of others, save and to the extent that such work has been cited and acknowledged within the text of my work. I/We understand that plagiarism, collusion, and copying are grave and serious offences in the university and accept the penalties that would be imposed should I engage in plagiarism, collusion, or copying. I/We have read and understood the Assignment Regulations. I/We have identified and included the source of all facts, ideas, opinions, and viewpoints of others in the assignment references. Direct quotations from books, journal articles, internet sources, module text, or any other source whatsoever are acknowledged and the sources cited are identified in the assignment references. This assignment, or any part of it, has not been previously submitted by me/us or any other person for assessment on this or any other course of study.

I/We have read and understood the referencing guidelines found at http://www.dcu.ie/info/regulations/plagiarism.shtml, https://www4.dcu.ie/students/az/plagiarism and/or recommended in the assignment guidelines.

Name: Teena Sharma Date: 26-11-2021

Name: Teenu Prathyush Date: 26-11-2021

DATA VISUALIZATION AND ANALYSIS OF MOTOR VEHICLE COLLISIONS IN NEW YORK

1. ABSTRACT:

All over the world, traffic accidents occur hundreds of thousands of times per year. The problems are so prevalent and frequent that motor vehicle safety is one of the primary concerns of all countries. In this project, we will be using New York vehicle collision data to generate useful insights. The details which we are using from the dataset include the following: Crash day, Crash date, Borough, Latitude, Longitude, Aftermath of Accident, Number of Persons Injured, Number of Persons Killed, Contributing factor, Collision ID, and Vehicle Type. Ultimately, we want to visualize accidents across each Borough in the state of New York. This will enable us to see where the most accidents are occurring and how many people suffer injuries, death, or no injuries. Additionally, we collected data about vehicles to see which vehicles are involved in the accidents. Using this data, we can get a better understanding of the areas that need further attention in terms of public safety.

2. DATASET:

Link to the datasets:

1. https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Crashes/h9gi-nx95
2. https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Vehicles/bm4k-52h4

Our visualizations are mainly achieved on Tableau using two datasets, Motor Vehicle Collisions – Crashes and Motor Vehicle Collisions – Vehicles, obtained from the NYC OpenData website, in CSV formats. Among the three characteristics of Big Data, two are present in our data — volume and variety. The NYC OpenData Website updates these two datasets regularly and we are provided with the latest information regarding accidents in New York.

The Motor Vehicle Collisions – Crashes dataset contains information on accidents happening in New York across each Borough. It has 1.84 million rows and 29 columns where each row is a motor vehicle collision. The Motor Vehicle Collisions – Vehicles dataset contains information on the type of vehicles involved in the collision. It has 3.86 million rows and 25 columns where each row is a motor vehicle involved in a crash. Both the datasets contain entries of mainly three data types - DATE, INT and CHARACTER types and the combined size of both datasets is 1.1 GB.

3. DATA EXPLORATION, PROCESSING, CLEANING:

The datasets contained a lot of null values and mismatching entries. We cannot open the complete dataset on Excel as it has a limit on the number of rows and columns. For this purpose, we used RStudio for the majority of the cleaning process and then Microsoft Excel is used for cleaning the subset of data that was required for our visualization. Out of the 29 columns present in the Crashes dataset, we used the following 8 columns for our visualization – Crash Date, Borough, Latitude, Longitude, Number of Persons Injured, Number of Persons Killed, Contributing Factor and Collision ID. And out of the 25 columns present in the Vehicles dataset, we used the following three columns – Collision ID, Crash Date and Vehicle Type.

- The Crash Date column was obtained by merging the original Crash Date and Crash Time columns. This column was then formatted to use POSIXct timestamp format.
- Additional columns were created for our visualisation in the Crashes dataset. These are Crash Day specifying the day of the crash and Aftermath of Accident which specifies whether any person was injured, killed or not injured in the accident.
- The Vehicle Type column in the Vehicles dataset contained a lot of mismatching entries so we categorically encoded each entry to contain one of the following values – Sedan, SUV, Two Wheeler, Heavy Vehicle, Taxi, Bus, Commercial Vehicle, Fire Engine, and Ambulance.
- The Contributing Factor in the Crashes dataset also contained a lot of mismatching entries so we cleaned the data to only have valid entries.
- We have removed all the entries containing null values using RStudio.
- Finally, for combining both datasets, we used Tableau to perform an inner join on the Collision ID column present in both datasets.

4. VISUALISATION:

In order to determine the patterns, trends, and relationships between big data, they need to be extracted and analysed by computers. A common objective is to reveal the hidden knowledge underlying these data sets. In order to extract hidden knowledge from data, data visualization can be presented graphically and dynamically [1].

I. DOT DISTRIBUTION MAP:

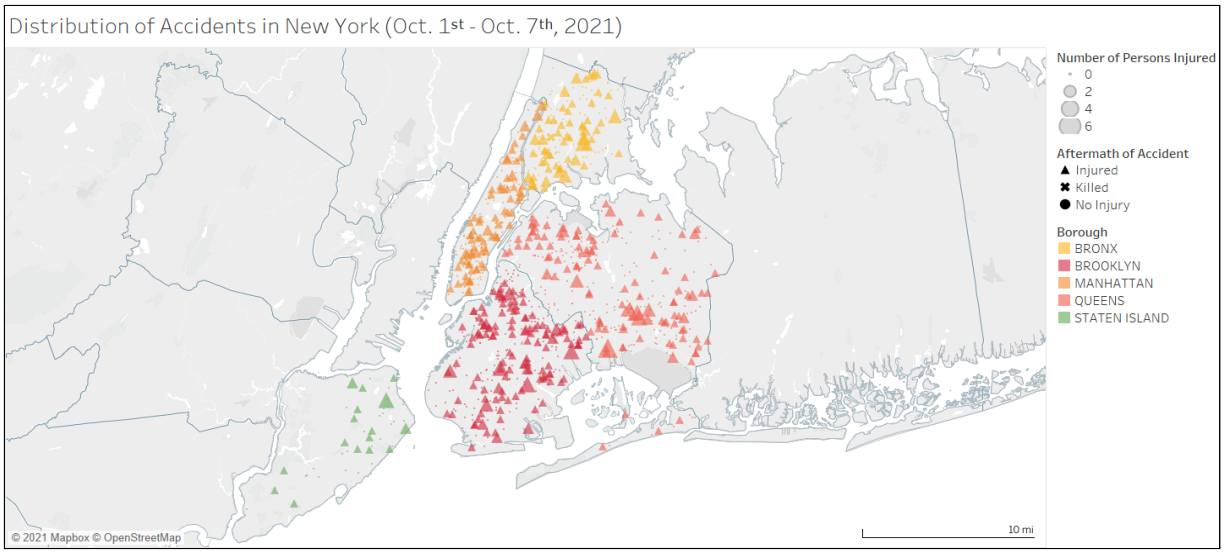

Figure 1: DOT DISTRIBUTION MAP

The dot distribution map displays the values of one or more fields by displaying dots or other symbols on the map. Each dot on a dot distribution map represents a certain amount of data [2]. By using a dot distribution map, we have visualized the number of persons injured, killed or not injured across each borough in New York for a period of one week, starting from October 1st 2021 to October 7th 2021.

Colour is encoded according to Boroughs in New York. We have used Dark Red for Brooklyn to signify the most number of accidents, followed by Queens which has a slightly lesser shade of Red to

signify the second most number of accidents, followed by Manhattan which has Orange as its colour to signify the third most number of accidents, followed by the Bronx which has Yellow to signify the next most number of accidents, and finally, Staten Island has the colour Green to signify the least number of accidents.

Three different shapes have been used to signify the aftermath of accidents, to visualize if any person was injured, killed or not injured. For No Injury, we have used filled circles as the shape. For any person having injuries, we have used filled triangles as the shape and finally, for any person killed in the accident we have used a filled cross mark as the shape.

We have also used the size feature to denote the number of injuries sustained in an accident. In our data, the most number of injuries sustained in an accident is six, so we have used a larger size to denote this accident in which six people were injured. Size decreases as the number of injuries decreases, in other words, size is directly proportional to the number of injuries in an accident.

We have also used an opacity of 60% to visualize any shapes which are overlapped by bigger shapes. The background of the map is chosen to be grey colour instead of satellite view or street view which makes the map much easier to visualize. County Borders have been used to differentiate between each borough. A scale according to the US metric system is displayed on the map for an accurate understanding of distance. The steps undertaken above will allow us to visualize where the accidents are happening, what the aftermath of the accidents is and how many persons are injured in an accident.

II. HEAT MAP:

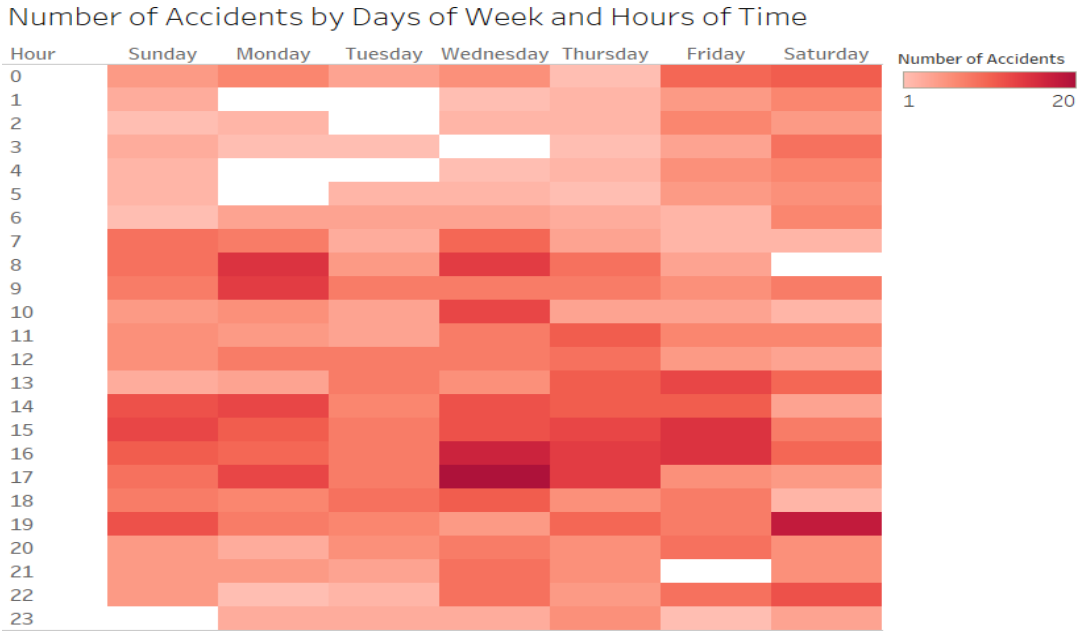

Figure 2: HEAT MAP

Heat maps are useful for communicating the highs and lows of a metric. Based on one week's accident data from October 1st 2021 to October 7th 2021, we used heat maps to visualize the number of accidents per hour. We have used Red as the colour of our choice as it denotes danger and it fits what we are trying to visualize, that is accidents occurring in New York. Colour saturation is taken into

account to visualize the number of accidents happening every hour. The more accidents there are, the darker the shade of red. The colour White denotes that there are no accidents in that particular hour.

III. 100% STACKED BAR CHART:

In a 100% stacked bar chart, a viewer can compare how each component contributes to the sum of comparable values across categories. Values on the X-axis range from 0 to 100%. In our visualization, we determined the percentage of accidents caused by a certain type of vehicle in a particular borough.

We used different colours for each vehicle type. The colour choices are mostly based on what colour vehicles we usually find on roads as well as personal preferences. For example, most people usually drive blue Sedan cars, SUVs are usually black, most taxis in New York are yellow, green is used for buses. A dark shade of White is used for ambulances and so on. We have generally tried to use light colours in our graph so that it is pleasing to the eye. For better visual aid, we have restrained from using percentage labels in each category as there are vehicle types like Bus and Ambulance which occupy less space in the stacked bar graph.

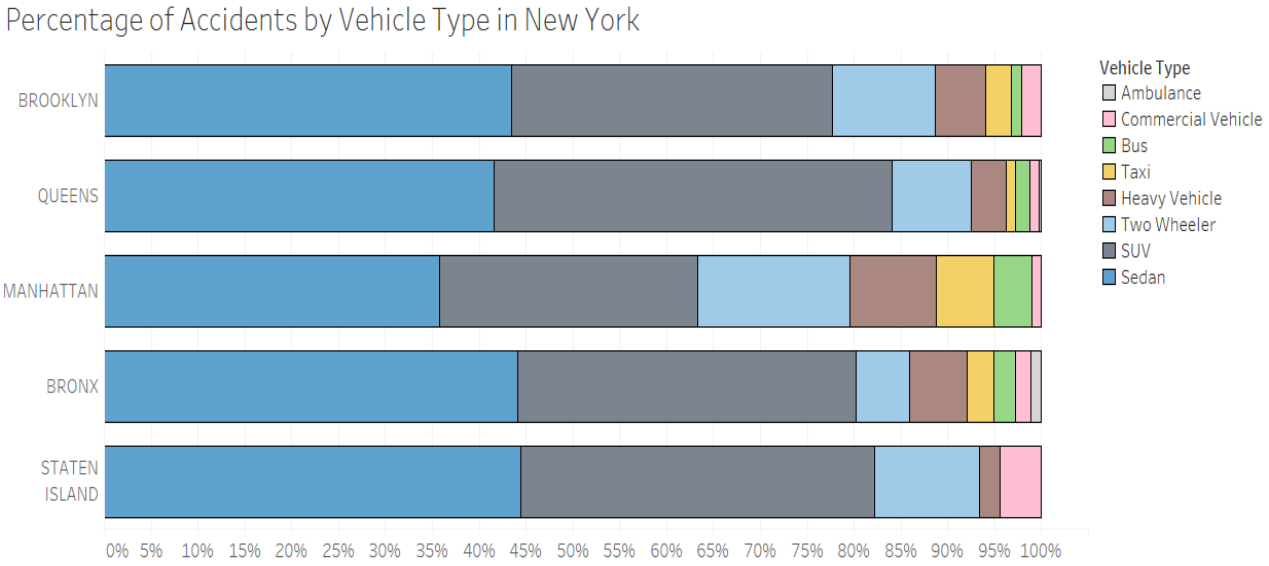

Figure 3: 100% STACKED BAR CHART

We have used Tableau for developing all of our visualizations as it is an interactive tool that offers a wide range of features such as visualizing data at the most detailed levels of granularity and other features such as filters and highlighters which makes it easy to develop quick and attractive visualizations. For all the charts displayed above, we have used legends on the right-hand side for easy movement of the eyes and interactivity is provided in the form of filters, highlighters and tooltips in Tableau. Filters can be used to view a select amount of data. For example, we can only visualize the number of persons killed while filtering out the persons who are injured and not injured. Highlighters can be used to highlight a particular category or point on the chart. For example, we can highlight a particular vehicle type (SUVs) to stand out while dimming the rest of the categories. Tooltips provide a description of the visualization when a user hovers the mouse over a particular point or category.

5. CONCLUSION

I. Observations from Dot Distribution Map

- We observe that most accidents are happening in Brooklyn and the least amount of accidents are happening in Staten Island. This could be due to the size of the population of the two boroughs.
- We observe that there are clusters within the map where accidents are frequent. Upon analysing in Tableau, we found that one cluster in Queens had the same contributing factor (Failure to yield right of way/Turning improperly) for almost all accidents. The authorities could look at this information and take necessary actions to prevent it from happening.

II. Observations from Heat Map

- A lot of accidents generally happen during the evening and on weekends as people like to go out for entertainment and this can be seen in our heat map. Many accidents are taking place during the evening and on weekends. Monday mornings (around 8 am) is another time when there are many accidents. This could be due to people going to work early in the morning after the weekend is over.
- Surprisingly, many accidents occurred from 17:00 - 18:00 hours on Wednesday. Upon filtering the data in Tableau, we could see that this seems to be an anomaly as the accidents are spread out and there is no evidence of them being related to each other.

III. Observations from 100% Stacked Bar Chart

- Unsurprisingly, Sedan and SUVs are the most common types of vehicles involved in accidents, followed by Two Wheelers, while the rest have a similar percentage of being involved in an accident.
- Around 25% of accidents in Manhattan are contributed by Two Wheeler vehicles. This is significantly more when compared to other Boroughs and it could also be the reason that people tend to use Two Wheeler vehicles more instead of cars due to the busy streets in Manhattan. Heavy vehicles are also involved in 10% of the total accidents in Manhattan which is significantly more when compared to other Boroughs. This could be due to goods vehicles carrying material to construction sites.

We wish to conclude by saying that this assignment was a great opportunity for us to learn about different visualization techniques using Tableau. We also learned how to clean Big Data using RStudio and Excel and overall it was a great learning experience. This assignment was divided equally among ourselves. The data collection and processing part was done by Teena Sharma. The visualisation part was done by Teenu Prathyush. Report writing was done by both team members.

6. REFERENCES

[1] Z. Idrus, Z. Idrus and S. N. Ismail, " StarVizAlgo: Visual Analytic Dot Mapping in Data Visualization," *IEEE,* 2020.
[2] "Caliper Corporation," [Online]. Available at: https://www.caliper.com/glossary/what-is-a-dot-density-map.htm.
[3] "AnyChart," [Online]. Percent Stacked Bar Chart, Available at: https://www.anychart.com/chartopedia/chart-type/percent-stacked-bar-chart/.
[4] Yeap Z. C., “Medium”, Dec 2020, [Online]. Available at: https://medium.com/analytics-vidhya/Visualising-uk-road-traffic-accidents-data-with-tableau-48ae78485807.