Analysis of vehicle collisions in New York City

Wassim BEN YOUSSEF, Visual Analytics coursework, City University of London

1 Motivation, data and research questions

Road accidents and vehicles collisions are the cause of approximately 1.3 millions of deaths and between 20 and 50 millions of injured each year in the world according to the World Health Organisation [1]. If anything is done, this number could double by 2030. The visualization of data about road accidents collected by authorities could be very helpful to give an overview of the situation. This might facilitate the choice of decisions and measures by authorities and give them directions of what should be done to improve safety on the road.

I chose to use a dataset of vehicle collisions in New York City between January 2012 and November 2016, provided by the Police Department (NYPD). The choice of this dataset seems to be relevant since a lot of information and details are given about each accident (the dataset contains 29 variables) and a large number of accidents are reported (more than 900,000). Moreover, several features are geospatial variables, including Latitude and Longitude, which can allow the visualization of maps on Tableau, a software that I appreciate and that I want to use during this study. The number of rows being very high, we can select the data which best suits us. To be able to use a map properly, we firstly choose only accidents which have a value for the variables « Location », « Latitude », « Longitude », « Borough » and « Zipcode », so we deleted all the observations with a missing value for one of these variables. Then, I selected only the 6500 first observations for the study. I also chose to add a column which gives the number of vehicles involved in the accident, called « Number_vehicles_involved ».

During this study, we will analyze characteristics of accidents in the different boroughs to see if there is a different kind of accidents in popular boroughs as the Bronx than in most busy and touristic borough as Manhattan. Furthermore, we will study the effect of the time to see the influence of the time on accidents. To sum up, the analysis will try to solve the following research questions:

- How accidents are geographically distributed?
- When accidents are the most frequent, and when are they the most dangerous?
- Which kind of vehicles are the most involved in the different boroughs?
- Which causes are the most frequent and which ones are the most dangerous?

2 Tasks and approach

The study was made using Tableau and R. During the study; we will mostly focus on three features and see the behavior of these features: the number of accidents reported, the number of people injured and the number of the vehicle involved.

2.1 Overview of accidents in the boroughs using geovisualization

The spatial distribution of accidents between the different boroughs is an important issue raised by this study. To analyze this question, we use a map showing the location of each accident (figure 1). The size of the point gives the number of vehicles involved and the color gives the borough in which happened the accident. This visualization allows using filters to have more accurate information, for instance, the repartition of accidents in a time interval or accidents involving only a type of vehicle. We can also compare the density of accidents and detect which areas show a particularly important number of accidents.

2.2 Visualization of temporal distribution using graphic plots

Now, we need to see the evolution of events during the time. These events can be the number of people injured, the number of accidents, the number of vehicles involved in an accident … Graphic plots are used for this visualization, permitting to see the evolution of the number of accidents or people injured regarding the hour of the day. As for the last task, we can use different filters and clusters, for example to see the evolution of records for each borough or each type of vehicle.

2.3 Barplots to visualize more accurately accidents characteristics

In order to compare more precisely the difference of involvement of different variables and parameters on the accidents, bar plots are good tools giving accurate results, allowing for instance to compare more precisely the number of accidents in the different boroughs. This visualization helps to see the precise number of events, which was not possible with map display.

2.4 Analysis of causes using heat maps

Another important aspect of collisions is their causes. To visualize the importance of each cause in each borough, we use heat maps. This tool shows the significance of each cause in the number of accidents using saturation. With that, we can quickly identify the most involved causes and compare the five boroughs. Moreover, we can use a time filter to see the impact of each cause in different periods of the day.

2.5 Computation of probability to be injured

The dataset gives the number of people injured for each observation. We then used these values to calculate the probability to be hurt when we have a vehicle collision for each borough for finally plotting these results in a bar plot. This probability was calculated by dividing the number of accidents in which there are people injured by the total number of accidents (figure 17). This visualization permits to see in which boroughs accidents often lead to injuries and so to find which boroughs have the most dangerous roads. The same thing was done for contributions.

3 Analytical steps

3.1 Distribution of accidents over boroughs

In Figure 1, we can see the location of each accident in the map of New York City. The accidents are clustered by borough identified with colors: navy blue is for the Bronx, orange for Brooklyn, red for Manhattan, light blue for Queens and green for Staten Island. Moreover, the size of the point represents the number of vehicles involved in the accident. This permits to visualize the importance of the observation. The overall visualization is not clear enough, so we did a zoom on each borough. As expected, Manhattan seems to concentrate an important number of accidents in almost all its surface. This can be explained by the high density of vehicles caused by congestion, which is an important problem in Manhattan, especially around the Central Business District [2].

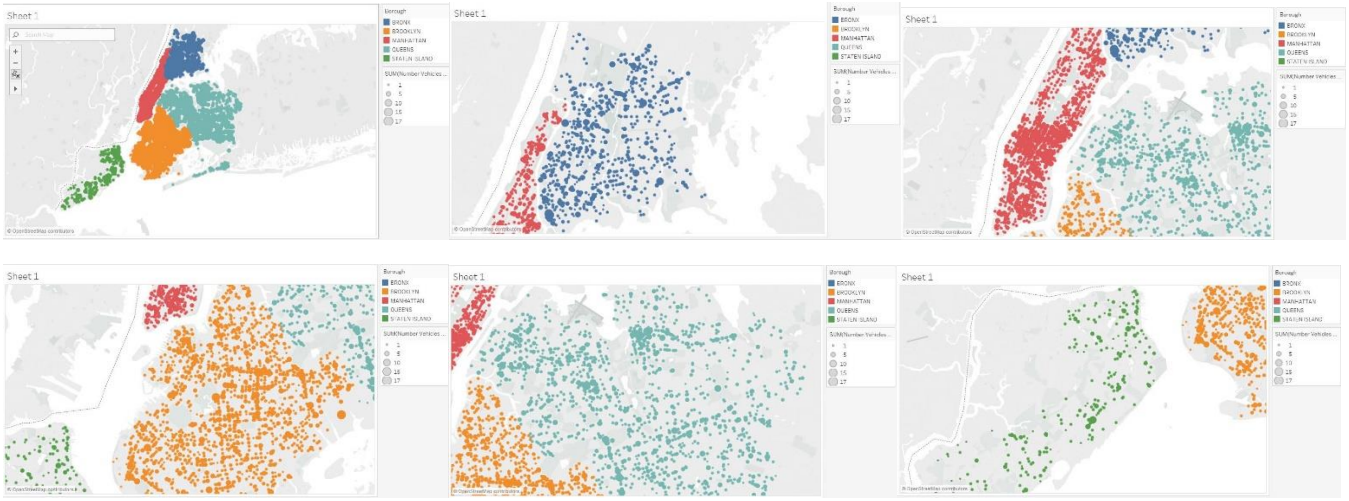

Figure 1: Maps of location of each accident clustered by borough with specification of the number of vehicles involved

This visualization allows us to use different filters. For instance, in Figure 2 we can see the repartition of accidents involving taxis. Given the fact that Manhattan is a touristic and business borough, it seems to account the most important number of records. For Brooklyn, we can observe that most of the accidents concerning taxis are concentrated in the North-West of the borough, a touristic area famous for its cultural history [3].

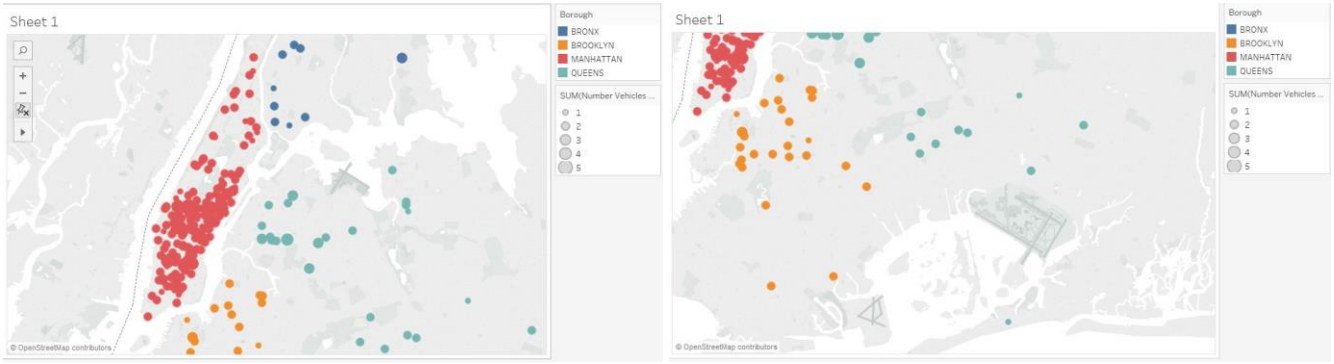

Figure 2: Maps of accidents involving taxis

3.2 Temporal variations

The map visualization can also be used to make a temporal analysis. Indeed, we can visualize the accidents for a specific time interval. We then divided the day into four intervals to see records for each interval. For instance, in Figure 3, we observe not so many collisions but most of the points are quite big which shows that the accidents often involve several vehicles.

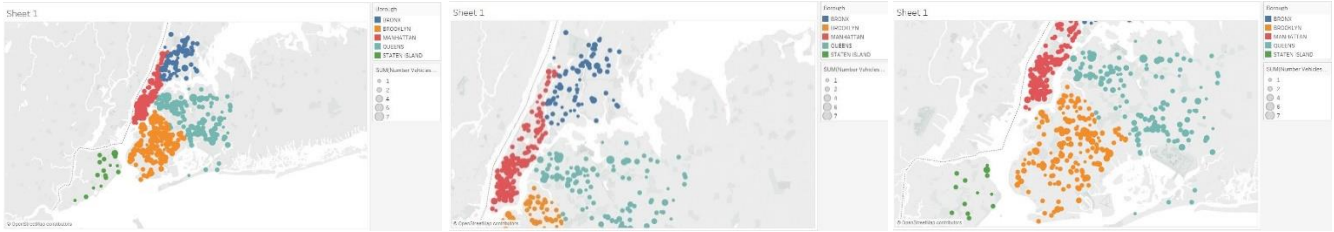

Figure 3: Maps showing accidents between 00:00h and 5:59h

Same kinds of analysis can be done for the others time intervals in Figure 4, Figure 5 and Figure 6. We can see nearly the same distribution between 6:00h and 11:59h and between 12:00h and 17:59h with numerous accidents distributed quite homogeneously in the space. Except for Staten Island, where observations are still located on the East coast.

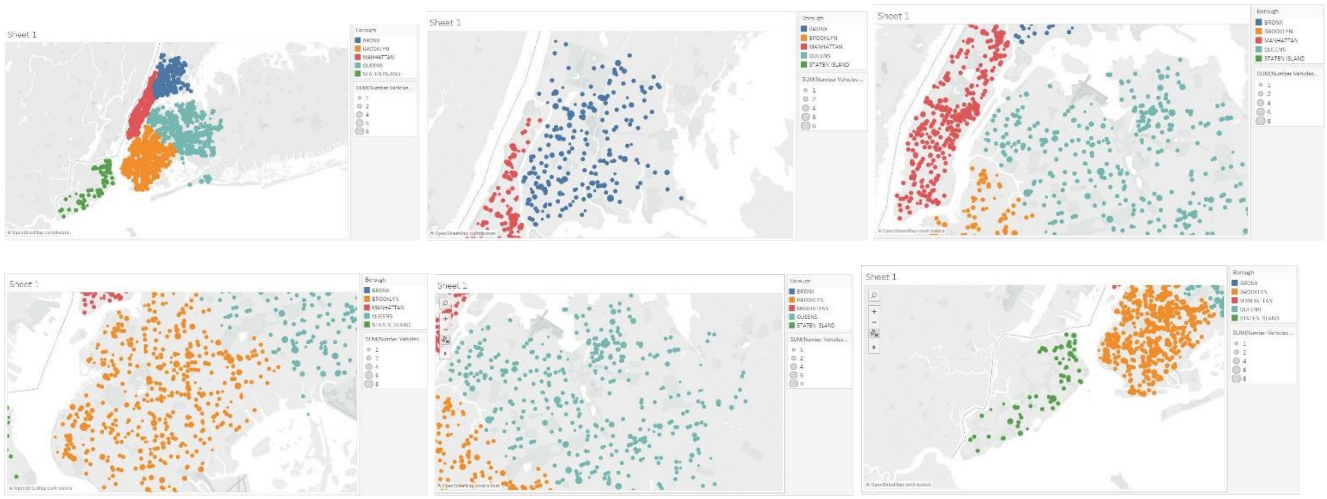

Figure 4: Maps showing accidents between 6:00h and 11:59h

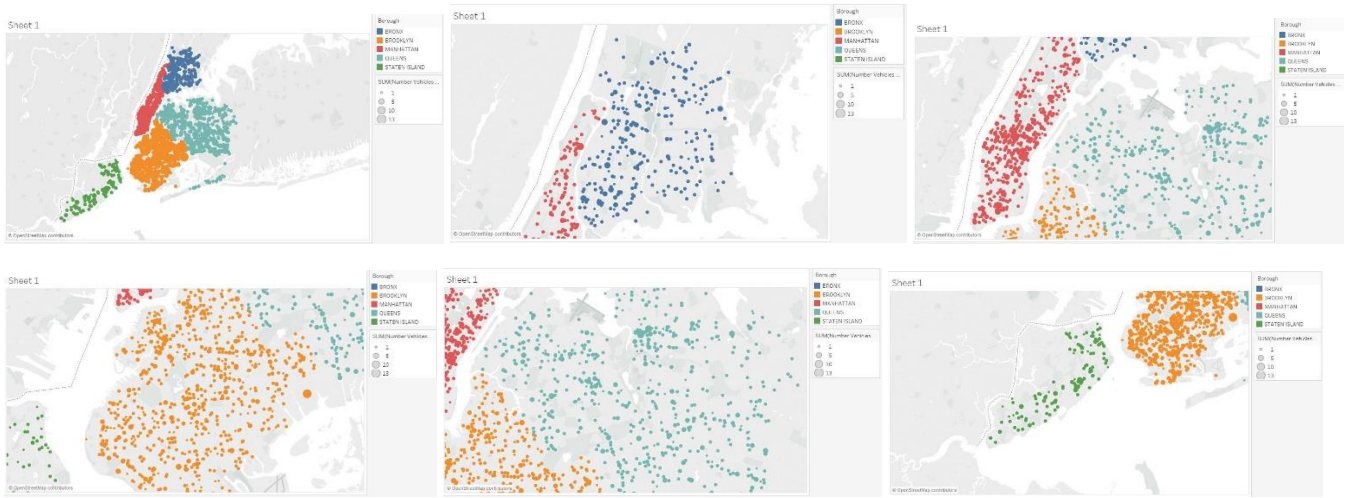

Figure 5: Maps showing accidents between 12:00h and 17:59h

Figure 6 shows that it is the time interval grouping the less number of accidents but with some very important ones involving up to 16 vehicles.

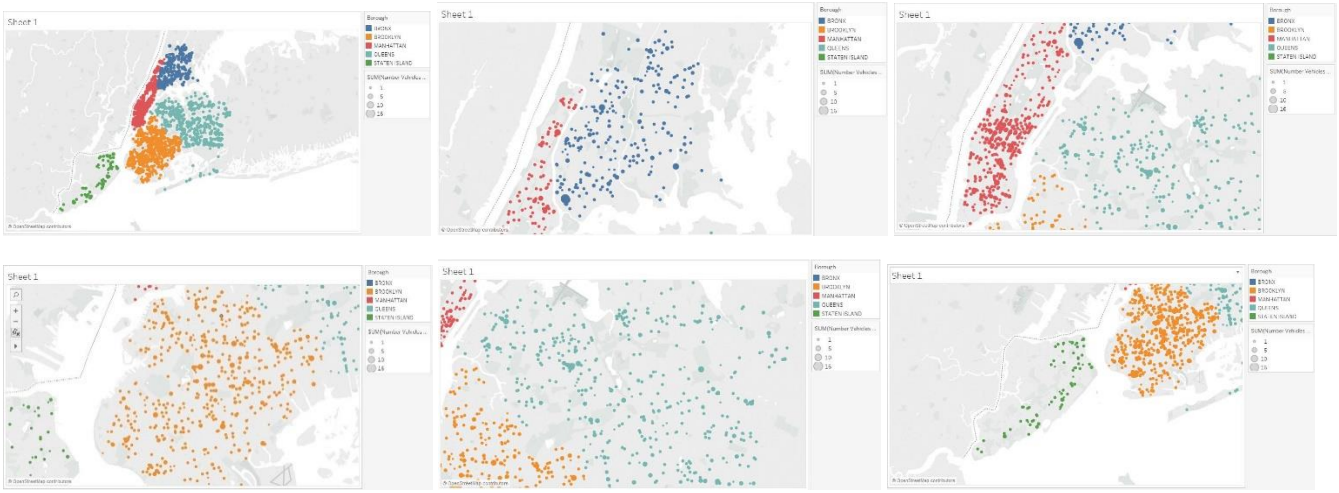

Figure 6: Maps showing accidents between 18:00h and 23:59h

Temporal graphic plots are others appropriate tools to see the evolution of the number of collisions during the day. With graphics in Figure 7, we can observe the number of accidents all along the day for all New York City in the first graph, for each kind of vehicle in the second graph and each borough in the third graph. As observed previously, we can see that the number of records is at its peak from 8:00h to 18:00h. These visualizations allow us to detect two prominent peaks at approximately 9:00h and 17:00h.

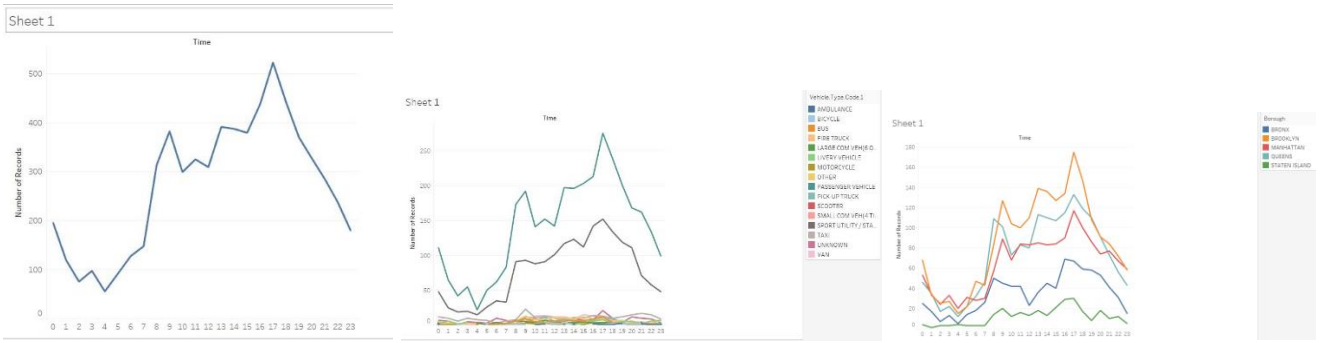

Figure 7: Evolution of number of records during the day for different parameters

We can also visualize the number of records for each day of the month and each month of the year (Figure 8), but this seems to be biased. Indeed, we took only the 6500 first observations of the original dataset in which observations were sorted in chronological order.

Other relevant features can be analyzed using this method. For instance, Figure 9 shows that the average number of vehicles involved in an accident observes a different evolution than the number of records, with an overall peak at 2:00h and very various variations between boroughs. Figure 10 shows another visualization for the number of injured over the time.

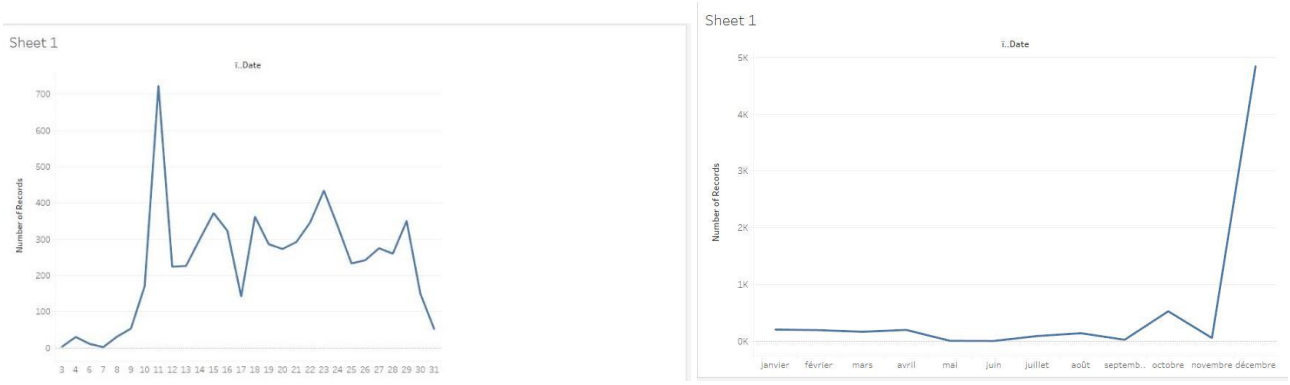

Figure 8: Evolution of records during the month and during the year

Figure 9: Evolution of Average number of vehicle involved in accidents

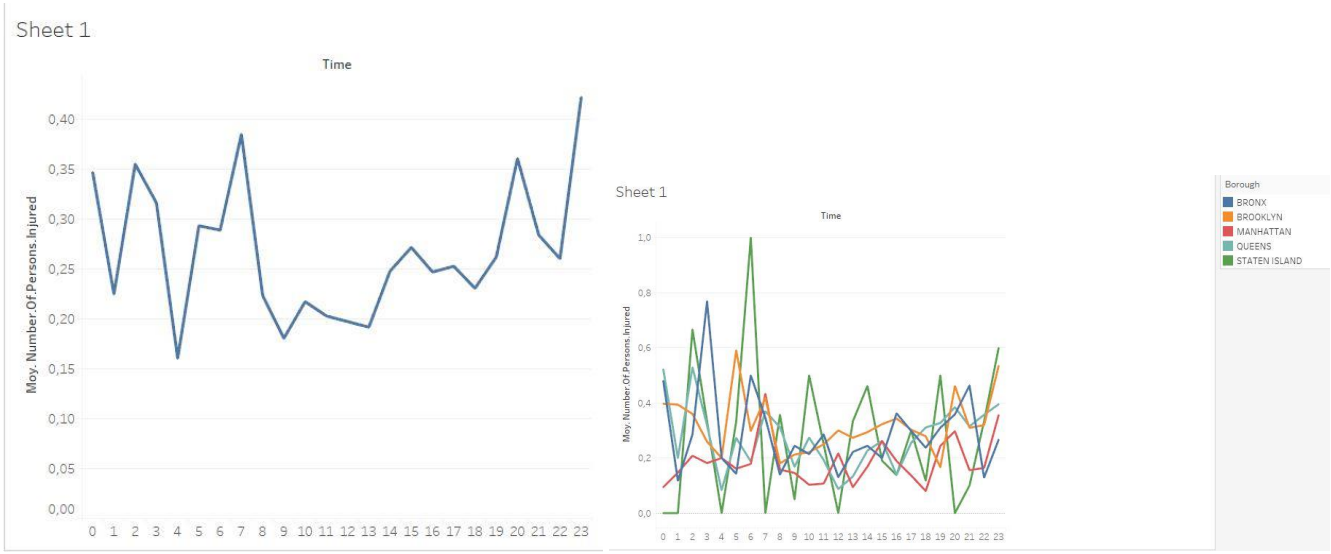

Figure 10: Evolution of Average number of people injured in accidents

3.3 Different comparisons using bar plots

The use of bar plots allows comparing more precisely different results. Indeed, with Figure 11 we can see that Brooklyn accounts the most of the observations far ahead Staten Island, the less concerned with road collisions. Moreover, Queens has the highest average number of vehicles involved. However, the difference of this value is not so different, varying only between 1.87 and 1.95.

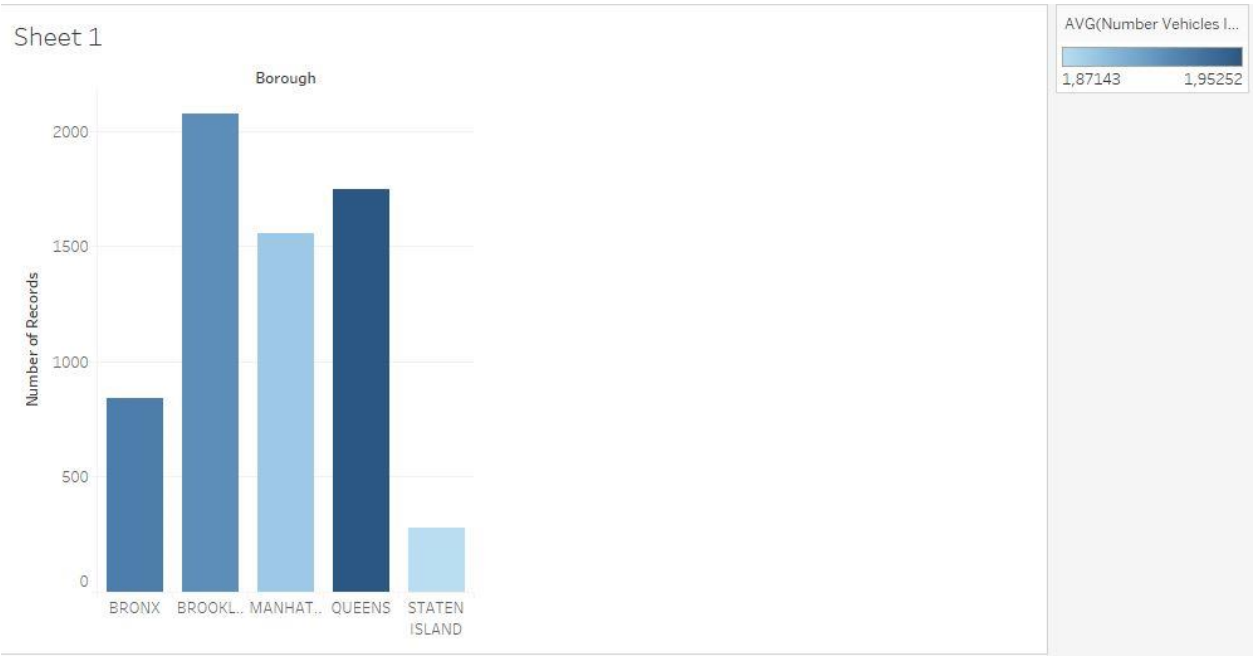

Figure 11: Number of records and average of vehicles involved for each borough

Figure 12 gives a more accurate visualization with the 20 most involved Zip codes.

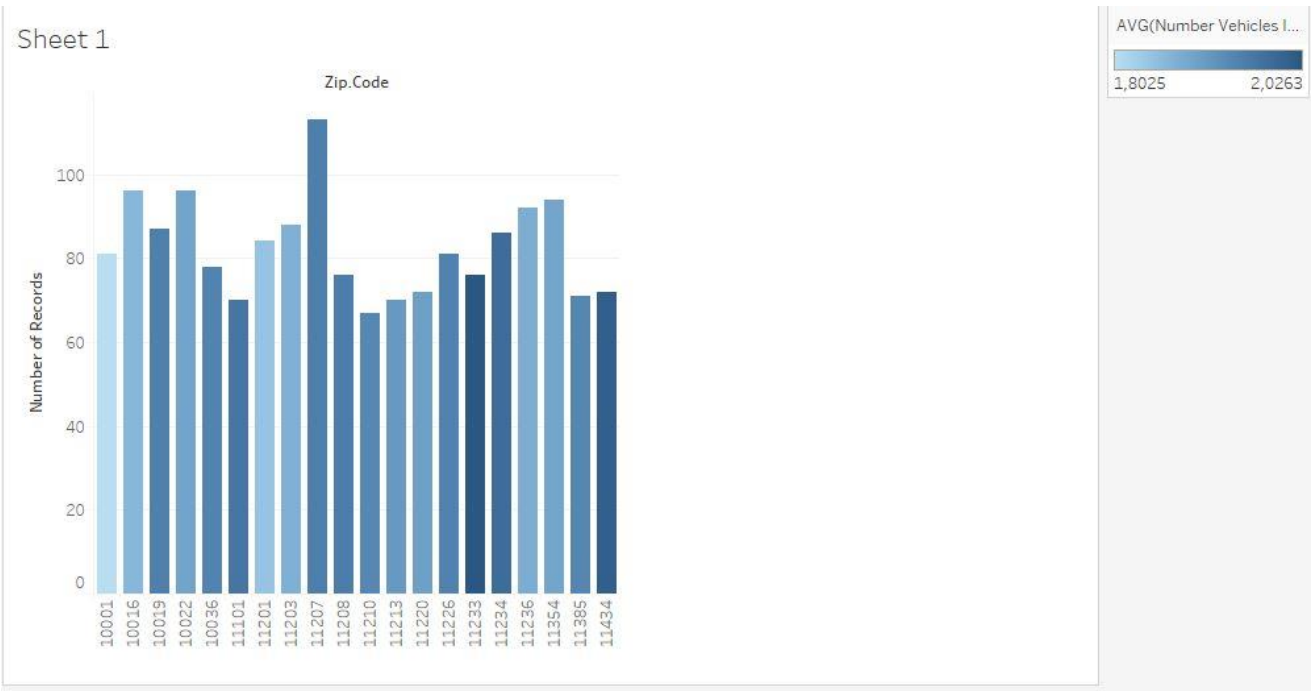

Figure 12: Number of records and average of vehicles involved for the 20 most common Zip codes

The type of the vehicle involved is also an important factor which can be studied using bar plots. Figure 13 is then relevant to see the involvement of each vehicle type. This visualization allows us to analyze the presence of these vehicle types in the boroughs. For instance, the Taxi bar confirms or previous map analysis, giving that most of the accidents involving taxis are located in Manhattan. Moreover, it is the third most involved vehicle type.

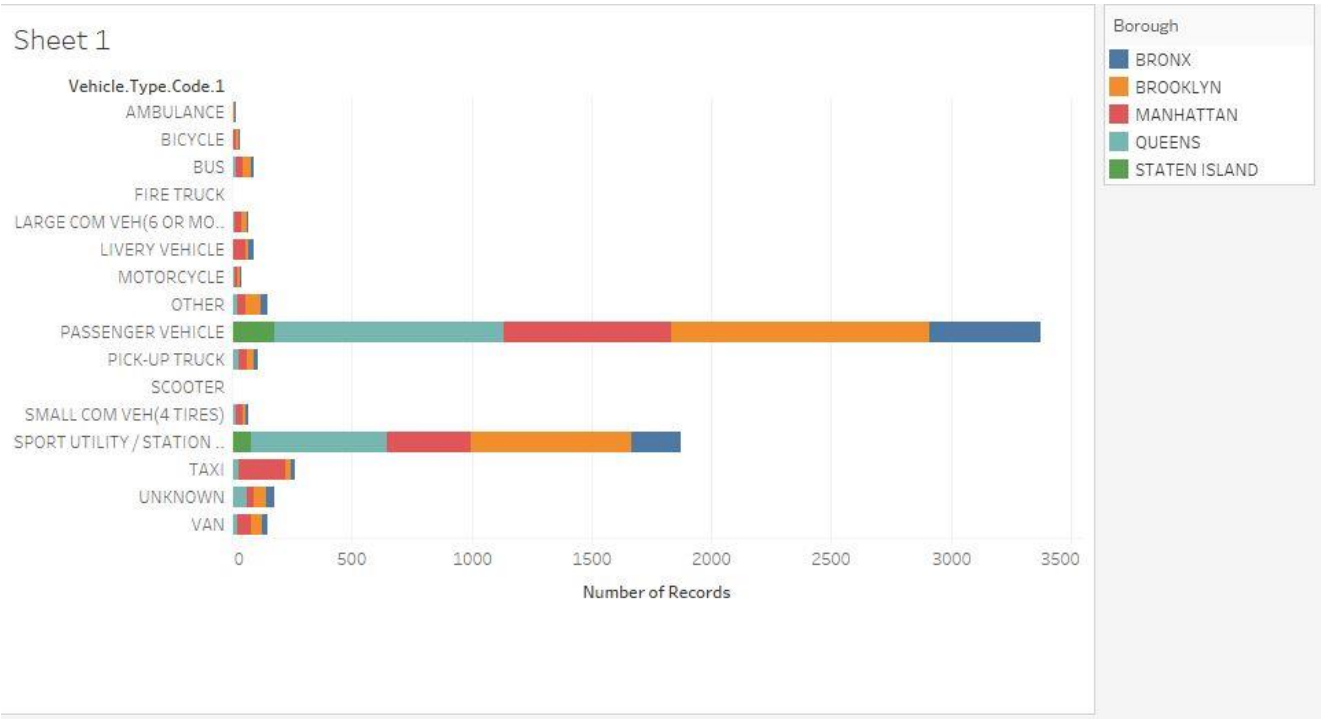

Figure 13: Number of accidents for each type of vehicle

3.4 Analysis of causes of accidents

Firstly, to see the importance of each cause of accidents, we use a heat map, the color hue giving the number of accidents. A first heatmap shows that a large number of observations have the value « Unspecified ». Then we could not differentiate between the other causes, so we just deleted it for the visualization (Figure 14).

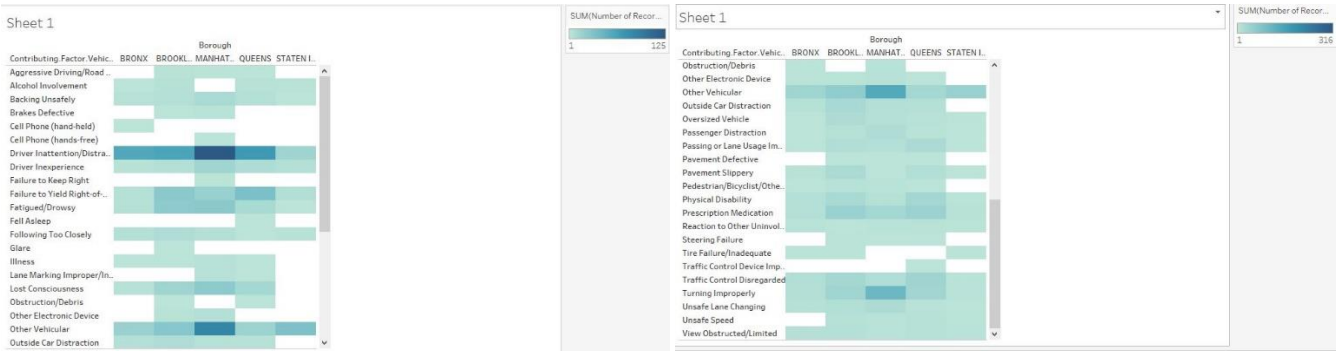

Figure 14: Heatmap of accidents contribution for each borough

We also have the possibility to use some filters for precise results. For instance, in Figure 15, we only have observations between 00:00h and 05:59h and Figure 16 uses only accidents involving bicycles, motorcycles and scooters.

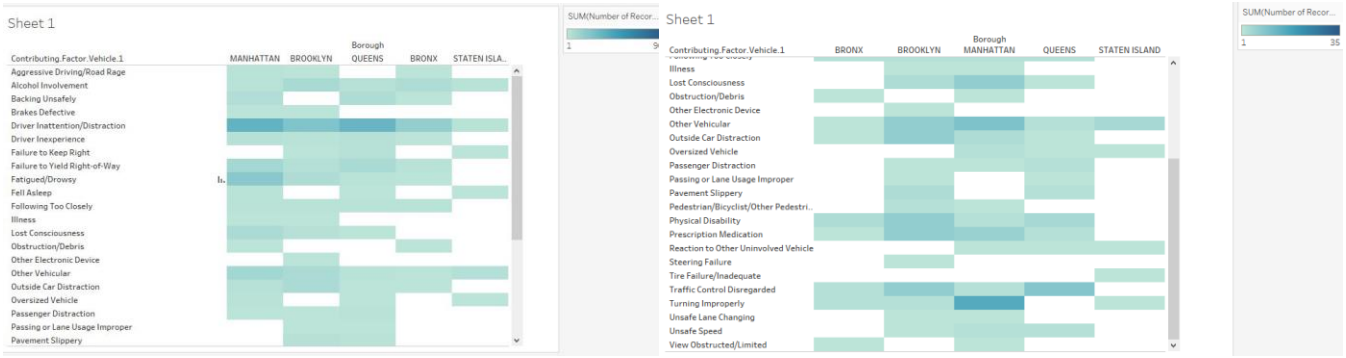

Figure 15: Heatmap of accidents contribution for each borough between 00:00h and 05:59h

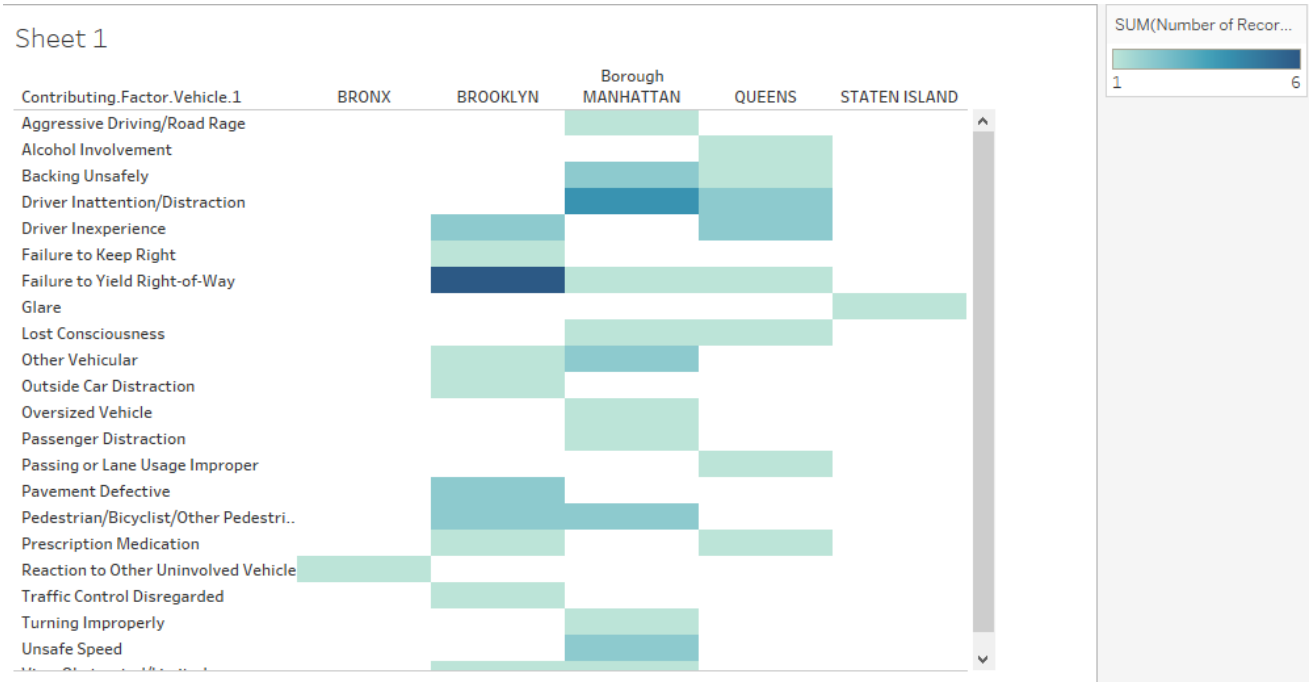

Figure 16: Heatmap of accidents contribution for each borough for bicycles, motorcycles and scooters

Another option would be to make the heatmap for vehicle types and then filter by borough.

3.5 Evaluate the risk to be injured

We thought it would be interesting to calculate the probability to be injured when we are victim of a vehicle collision for each borough. This shows in which boroughs accidents are the most dangerous. We also did the same for every causes listed. These probabilities were visualized in a bar plot using variation of hue color (Figure 18 and 19). We can see three contributions having a probability of 1 but that can be explained because there are only one or two observations with these values so these results are not relevant for these values. This shows another limit of the use of a too small sample.

	Borough	Number_of_accidents_with_injured	Proba_to_be_injured
1	BRONX	136	0.2020802
2	BROOKLYN	366	0.2319392
3	MANHATTAN	163	0.1419861
4	QUEENS	287	0.2070707
5	STATEN ISLAND	39	0.1813953

	Contribution	Number_of_accidents_with_injured	Proba_to_be_injured
1	Aggressive Driving/Road Rage	2	0.10000000
2	Alcohol Involvement	13	0.30952381
3	Backing Unsafely	12	0.07228916
4	Brakes Defective	6	0.31578947
5	Cell Phone (hands-free)	1	1.00000000
6	Driver Inattention/Distraction	192	0.21993127
7	Driver Inexperience	14	0.17948718
8	Drugs (illegal)	1	1.00000000
9	Failure to Keep Right	3	0.16666667
10	Failure to Yield Right-of-Way	140	0.37333333
11	Fatigued/Drowsy	30	0.13888889
12	Fell Asleep	4	0.50000000
13	Following Too Closely	25	0.25252525
14	Glare	4	0.33333333
15	Illness	5	0.26315789
16	Outside Car Distraction	8	0.03389831
17	Passing or Lane Usage Improper	14	0.51851852

Figure 17: Dataframes used to make the visualization of probability of being injured

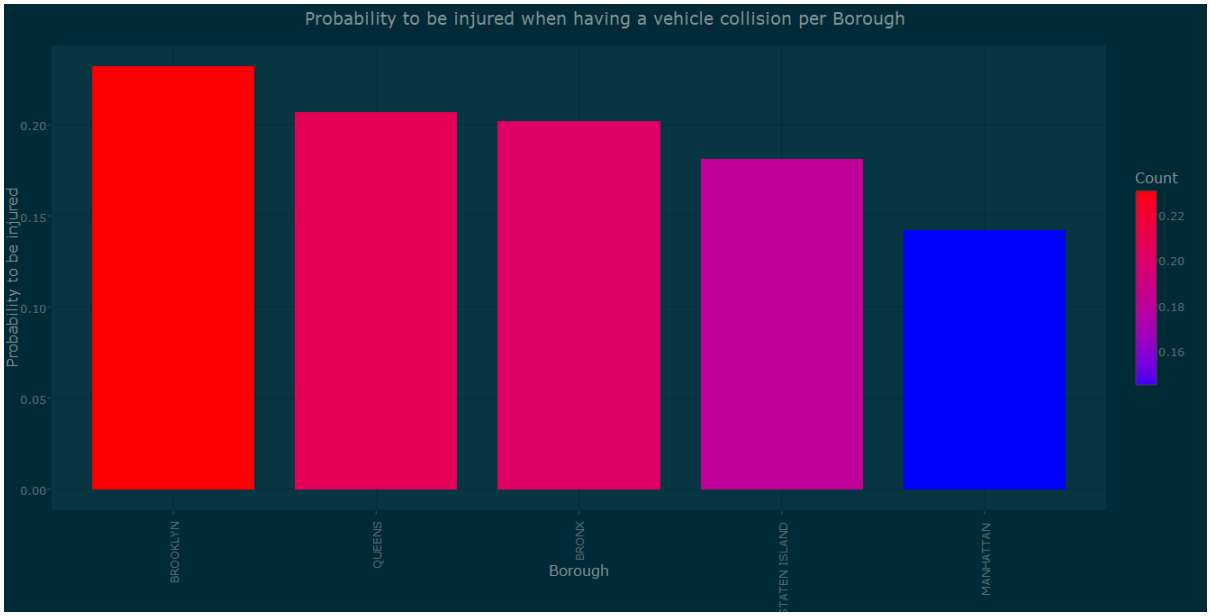

Figure 18: Risk to be injured for each borough

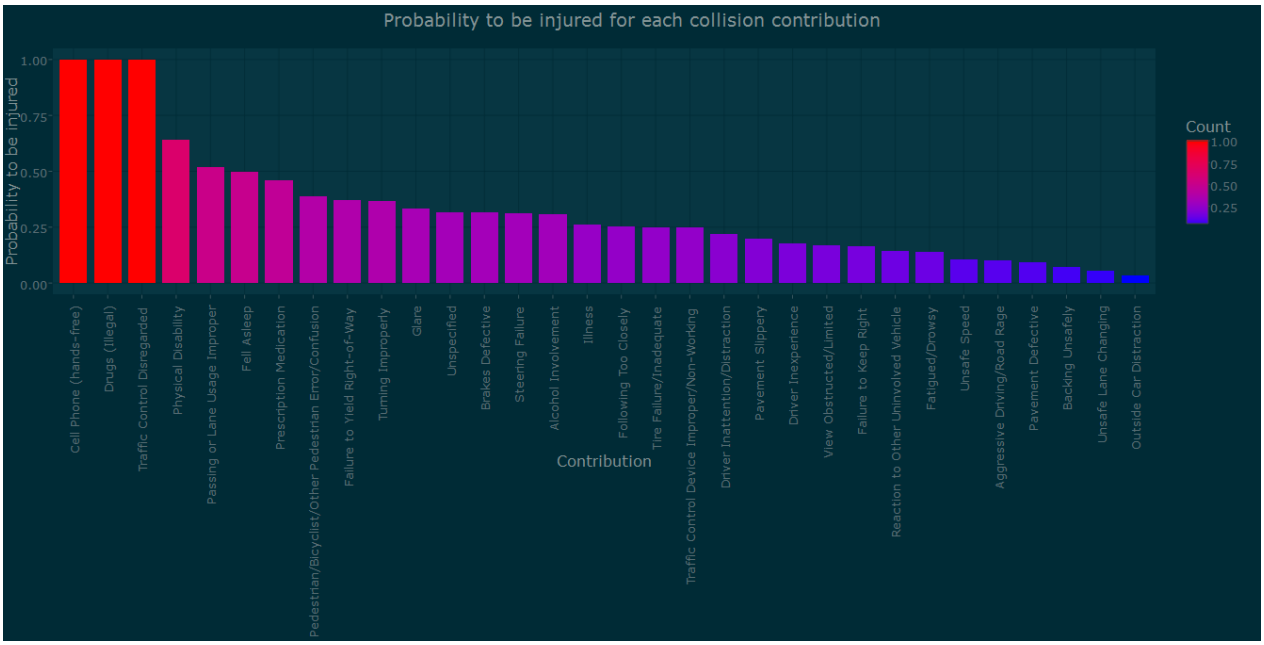

Figure 19: Risk to be injured for each contribution

4 Findings

This study deals with characteristics of road accidents and how the different factors of the accidents vary across New York City. Figure 1 shows that the density of accidents is important in Manhattan, while they are more scattered through all the borough for the Bronx, Brooklyn and Queens with major areas and road axes regrouping many accidents, especially the Zip Code 11207 in Brooklyn and the Zip Code 11417 in Queens (Figure 12 and 20).

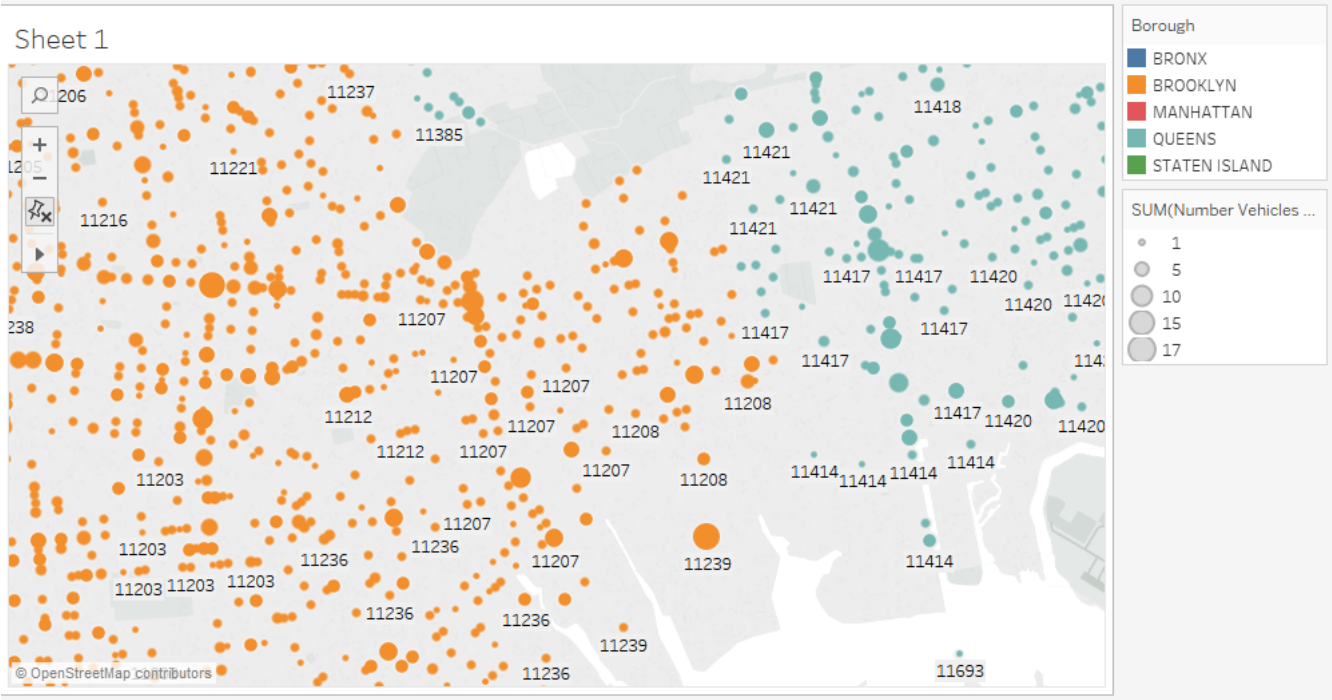

Figure 20: Zoom on the map visualizing accidents with the indication of Zip Codes location

For Staten Island, we can observe significantly fewer accidents than in other boroughs, located on the east coast. These results can be explained by the population, the density and the size of boroughs. Manhattan is the smallest of the five boroughs but has the highest density, while Staten Island population and density are very low comparing to the other boroughs [4]. Figure 18 shows that in Manhattan, accidents are often benign with less than 15% to be injured when having a collision while almost 1 on 4 causes injuries in Brooklyn (nearly 1 on 5 in Queens and the Bronx).

The temporal analysis shows two peaks of accidents similar to traditional traffic peak hours between 8:00h and 9:00h and between 17:00h and 18:00h (Figure 7). The time behavior is similar for the five boroughs, showing a slight decrease after the first-morning peak until the most important peak of the evening, followed by a decrease during the night. However, the average number of people injured and the average of vehicles involved do not follow the same evolution. According to figures 9 and 10, the average of vehicles involved has a peak at 2:00h and stays quite stable during all the day. Moreover, Manhattan often has the lowest average of vehicle involved. The average of people injured is much more important during the night. Conclusion: accidents are more dangerous during the night, especially during the intervals 22:00h-00:00h, 2:00h-3:00h and 5:00-7:00h.

According to Figure 13, passenger vehicle is the most involved kind of vehicle, followed by sport utility/station wagon and taxi. A remark about distribution over boroughs: livery vehicles, taxis and vans are the most often located in Manhattan.

Figure 14 shows that Driver Inattention and Failure to yield Righ-of-Way are the most involved causes. Physical disability, passing or lane usage improper and fell asleep are the the 3 contributions causing more than half the time injuries (Figure 19).

5 Critical reflexion

5.1 Implications of findings for domain

This study gives an overall view of the situation of road accidents in New York City. The results can be bounded and useful for institutions working for the regulation of the traffic with the objective of decrease the number of vehicle collisions. Indeed, findings show which areas are the most affected by overall accidents and by dangerous accidents reporting injuries for all the day and for time intervals. The study of injury risks also gives good indexes to see probabilities to have injuries for boroughs and accidents contributions. Although, we only have global results, giving numbers for large areas as boroughs or zip codes. So a much accurate study might be relevant to target precisely in which streets/avenues/bridges concrete measures must be taken.

Moreover, the results of major and most dangerous causes might be used to take specific measures as advertisements and awareness campaign to inform and warn the population. Combined with a geographical analysis, this can help to find where particular measure can be effective, for instance in which streets/avenues/bridges advertisement panels must be installed. The principal contribution of accidents, which is driver inattention/distraction, is one of the most concerned by these kinds of measures because other restriction rules or police surveillance might not be enough to encourage people not to drive when they are too tired and to keep focus on the road.

5.2 How well the data and visual analytics approaches enabled answers to research questions

The dataset is well adapted for a geographical analysis, giving that we have precise location points with Latitude and Longitude and also Zip Codes and Boroughs. That was enough for an overall study but could also be suitable for a more accurate study with smaller areas, and that could also be done using variables with the name of streets (Cross street name, Off street name and On street name). The dataset is also full of time information as the hour, the exact date, the month, the month day, the weekday, the year … That allows different kinds of time analysis, as the one done on the hour.

The type of vehicle involved and the contribution of the accidents are also two variables relevant for the study. They are well structured, especially because there are not so many unique values and the values are concise. So they did not need any transformation, instead of the dataset that I used for my project of Principles of Data Science. This dataset listing all the Shark Attacks (the Global Shark Attacks File [5]) has some features which needed to be modified to be useful. For instance, the feature giving the name of the species is composed of sentences as « Bullshark, 4' to 5' » or « Bullshark, 6’ ». So this is two different unique values. I then had to transform the data to recognise the shark species. Other features had to be handled and that took a lot of time. Indeed, according to a study by Kandel et al. [6], variables transformation is one of the most time-consuming challenges when dealing with a data analysis problem. The database used in this study did not need any changing in these features, except the creation of the new variable « Number_of_vehicles_involved », so that was an excellent advantage. However, the number of observations might not be enough for some aspect of the analysis. Indeed, the observations were not representative of what happen during all the year, most of the accidents happened during December 2016. We could use a larger dataset, but that would exclude the use of maps

representations made in this study (too many points would make the map illegible). However, a choropleth map might be a relevant approach.

5.3 Application of approach to other domains

This kind of approach can be generalized for the study of vehicle collisions in other bigger or smaller areas as States, regions or one of the five boroughs of New York City, or a borough of another city. This can also be applied to a whole country. But during this study, we saw different kind of analysis which can be applied to many other domains. For instance, the geographical approach can be used to analyze some dataset about real estate, like the one proposed by the Mayor of Paris about social housing [7] while the temporal analysis can be useful to study the evolution of Stock Exchange. Finally, the fact that we use standard visual tools like graphs, bar plots and maps commonly used in visual analysis means that this approach can easily be generalized to a multitude of domains and applications.

References

[1] World Health Organization. (2009). *Global status report on road safety: time for action*. World Health Organization.

[2] Hirschman, I., Mcknight, C., Pucher, J., Paaswell, R. E., & Berechman, J. (1995). Bridge and tunnel toll elasticities in New York. *Transportation*, *22*(2), 97-113.

[3] Haw, R. (2005). *The Brooklyn Bridge: a cultural history*. Rutgers University Press.

[4] LLC Books (2010). Boroughs of New York City: Manhattan, Queens, the Bronx, Brooklyn, Staten Island. General Books LLC

[5] Shark Research Institute (2005) http://www.sharkattackfile.net

[6] Kandel, S., Paepcke, A., Hellerstein, J. M., & Heer, J. (2012). Enterprise data analysis and visualization: An interview study. *IEEE Transactions on Visualization and Computer Graphics*, *18*(12), 2917-2926.

[7] Mayor of Paris https://opendata.paris.fr/explore/dataset/logements_sociaux_finances_a_paris/