NYC Traffic Accidents Visual Analysis

Alba García Ochoa
Visual Analytics course - Sapienza, Rome

Abstract
New York City's bustling environment presents significant challenges to public safety, with traffic accidents being a prominent concern. This paper investigates the underlying causes and conditions that contribute to motor vehicle crashes and collisions by analyzing open data provided by the New York City government. By designing an interactive visual analytics tool, the study enables comprehensive exploration and interpretation of extensive traffic datasets from the summer months of 2018. The tool integrates various data sources, including traffic incident reports and meteorological information, and employs data preprocessing and dimensionality reduction techniques to highlight key patterns and relationships. Through intuitive visualizations, users can easily navigate complex data, gaining insights into factors such as location, weather conditions, vehicle types, and contributing causes. This research demonstrates the potential of leveraging open data and visual analytics to enhance understanding of urban traffic dynamics and supports efforts to develop effective strategies for improving road safety.

Keywords: Big Data, Visual Analytics, Traffic Accidents, New York City, Data Visualization, Road Safety

1 Introduction

Big data has revolutionized the way issues regarding everyday challenges are faced, and when this problems concern citizens life and security, this data becomes of great value. Thus, data about traffic accidents and collisions is of great use in order to determine the causes and factors which are of influence and to determine the countermeasures to stop this kind of accidents from happening.

The publication of Open Data, such as the broad databases provided by New York City government [3], is a key resource for stimulating further research and engaging society to understand key issues and find solutions to improve social well being.

This paper provides an analysis of motor vehicle collisions and crashes in New York City during the summer months of 2018 through an interactive visualization, as well as the development of the analysis tool. This consists on an easy-to-use visual interface which allows a clear understanding of data and relations among different factors in order to draw conclusions. It also includes a more in-depth research using dimensionality reduction techniques which can serve for further analysis.

Furthermore, not only will this paper serve of interest to drivers for both personal and professional reasons, but also to government and authorities as they have the resources to improve road safety, yet need this kind of analytics to know where to invest their limited budget. Insurance companies and health care systems also benefit by this sort of analysis, for a larger control on the risks and factors that trigger them.

2 Background

Previous research related to the topic

Previous analysis on the topic is extense, as traffic accidents are a topic of great importance and Open Data encourages social interaction and study of this issues which benefits society at large. However, most of this research is more linked to manipulating and using data for a predictive serve, such as Shujie Zhang work[5] which proposes machine learning algorithms to predict NYC traffic accidents using weather data.

Moreover there are also some studies focused on the collisions allocation, as Benjamin Romano and Zhe Jiang paper, which focuses on spatial hotspot detection constrained to road networks, using therefore dimension and temporal dynamics[4]. This works made it clear that allocation must be taken into consideration when analyzing accidents, and influenced the use of geospatial point maps in the final visualization.

Finally, there is also some research on how to manage, store and mine data for analytics purposes, and create big data ecosystems, which can be understood through Eyad Abdullah and Ahmed Emam application, which uses Hadoop to store massive data and a parallel computing framework and a posterior Web services interface[1].

This paper completes previous research by the creation of a visual interface for analyzing data and drawing insights to reach conclusions and develop in posterior works countermeasures.

3 Methodology

Data processing, transformation and transformation pipeline

3.1 Data extraction

Two major datasets where extracted and stored locally for creating the application. First one included motor vehicle collisions and crashes in New York, filtered by year (2018) and month (June, July, August and September), and was extracted from New York's government's open source API[3]. Second one was used to enrich this first dataset, and contained weather conditions for each day of the month and belonged to Noaa's open source meteorological data API[2].

Nevertheless, data related to traffic collisions was collected from NYC Police Department preliminarily reports of crash events, therefore there is some missing data. To solve this issue a following preprocessing stage was conducted.

3.2 Preprocessing

As datasets are seldom complete and clean, and the one being used was no exception, a preprocessing stage had to be conducted. Main stages included the withdrawal of columns which were of no use in order to lower data load. In order to get meaningful insights and ease data management, both vehicle types and contributing factors were clustered in major groups, to reduce the number of categories.

As there was a large number of missing values, specially regarding NYC collisions' ubication (Borough and Zip Code columns), data was inferred through the longitude and latitude coordinates. Finally, as already stated, in order to enrich the dataset, weather information from Noaa agency was merged to the dataset,

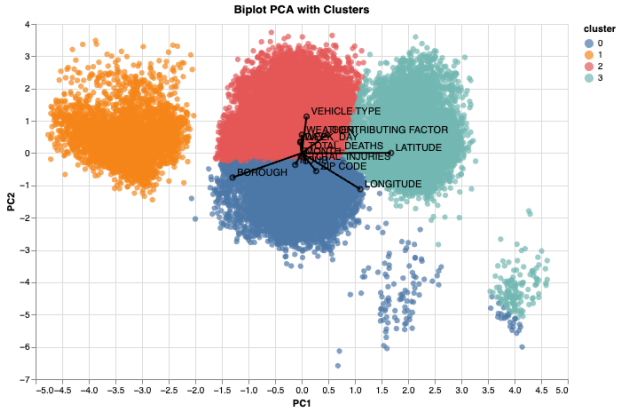
to get information on the weather condition for each collision.

3.3 Dimensionality Reduction

In pursuance of insights regarding relations among variables and observation, several dimensionality reduction techniques were applied, using scikit learn, Python, libraries.

First, a Principal Component Analysis was conducted, with a clear view of loading components and contributions of each variable. Results, on the one hand, were not very significative, as variance explained by the first principal components was not very significative. Only two first components would be used for the visualization, being the ones more meaningful.

On the other hand, however, results obtained through k-means clustering of observations for a biplot of first 2 principal components, showed some very interesting information. As it is shown in the chart below, there were 4 main clusters which precisely coincided with each borough, except for Manhattan and Brooklyn which belonged to the same cluster. Therefore, it can be stated that among other causes and conditions, there is a relation between accidents occurring in the same borough, and that accidents in Manhattan and Brooklyn are linked. Because of this, it would be interesting to consider a district management of traffic control and safety.



Other dimensionality reduction techniques, such as t-sne, were also tested, yet results were inconclusive and did no compare to discoveries made with the PCA data inspection.

3.4 Visualizations and design process

Several types of visualizations where tested using Python and Altair libraries, in order to get all meaningful insights data had to offer. It was clear that to see evolution over time, a dropdown for changing months was necessary, which was linked to each of the charts.

After testing the use of histograms, density plots and box plots for temporal evolution of collisions within each month, it was clear that the best way of laying data was by means of an histogram were rows represented months and columns were days, representing redness intensity the accidents' count. Evolution for the hours of the day was studied through a line chart, showing accumulation accidents per hour through the day.

For representing accidents' location, a cloropleth was discarded as the heavy data aggregation would lead to missing relevant ubication details, so instead a geospatial point map was use, where each point was an observation and the color of the observation encoded the accidents' severity, with red accounting for deaths, orange injuries and yellow no damage. In addition, a bar chart with total count for each Borough was included, to get an overview of collisions in each area.

To understand how weather conditions affected the number of accidents, an histogram was used, with a bar for each of the four types of weather (rain, cloudy, partially cloudy and sunny), and severity of accidents was encoded with same color as the ones used in the previous map.

Finally, to get a closer look on contributing factors, a horizontal bar chart proved very useful as differences among factors were clearly stated, including a color encoding for bars related to the type of vehicles.

3.5 Final Dashboard

Final dashboard is presented through Streamlit, as it eases the management of the large dataset used for this application. It includes the above mentioned charts, which are: a geospatial point map, a bar chart with accumulation of accidents per borough, a bar chart with contributing factors and vehicle types, a PCA scatter plot with clusters encoded through different colors, an histogram with weather conditions, a line chart for collisions evolution over hours of the day and a final heatmap for accidents count per month. Additionally, the applications' intuitive interface provides a direct data visualization, allowing the interaction among charts by selecting different conditions on demand, apart from the dropdown for time dimension. This way, if, for instance, user is interested on the peak collisions' hour on Manhattan in rainy conditions, by selecting in the corresponding charts these conditions, the result would be shown in the accidents x day-hour line chart. This tool will be therefore useful, not only to evaluate conditions on their own, but to understand relations among variables and the effect of factor combinations.

4 Analytics

NYC collisions' discovered insights

Brooklyn is the neighborhood to suffer more accidents overall, although if we analyze by month, we can see that for all months this rule remains except for September, when Queens suffers the largest amount of collisions, with 119 more than Brooklyn.

Another interesting issue is that when computing the dimensionality reduction through a Principal Component Analysis (PCA), we could see that there is a clear relation between the clusters find through a k-means algorithm and the boroughs. Thus, each cluster explained for a borough which except for Manhattan and Brooklyn which belong to the same cluster, so there might be relation between accidents and their causes happening in both neighborhoods. Nevertheless, explainability for the first two components, although being larger than for the other principal components, is significantly low so conclusions are not of a high importance.

Rainy weather is often linked to more traffic accidents, and data in this visualization confirms the theory. Interesting enough, although studying summer month (June, July, August and September) in NYC for which, as opposed to autumn or winter, warm temperatures and clear skies are the most common, the amount of accidents for rainy days is by far the largest for all months, so indeed we can conclude that more attention should be paid in case of precipitations and could be considered including more measures to regulate traffic in this kind of weather.

Moreover, accidents related to sunny weather decrease over the months, reaching its lowest on September, probably due to more sunny days on June and July than on August and September, as opposed to rain accidents which reach it's highest on September.

We can see that it is cars which by large suffer the most traffic accidents, regardless of other conditions (weather, contributing factor, borough, month...), and that the overall distribution of vehicle type has little changes for different conditions.

It can also be seen that, at night hours there are more taxi accidents, while at day hours the number of truck accidents increases (always being on second place to car vehicle type accidents'). This makes sense, since trucks are at most operated by daytime, vehicle lot of

taxis are used at nigh to return safely, which increase the possibility of having more accidents.

Another interesting insight is that is Manhattan where most of the Taxi accidents take place, which is a reasonable conclusion given is the a very touristic place worldwide and also compared to other NYC boroughs. Queens accounts for the larger number of car accidents and Brooklyn for truck accidents, yet both are write paired to other boroughs numbers.

We can see there is no big difference among the total number of accidents per month, nor any condensation of accidents for several days. We can see that for June there are several days with many accidents, probably days in which people go out on vacation, reaching its highest on the 29th with 772 collisions. For the beginning of September there are also some days with a large number of accidents (from the 4th to 6th), maybe the reciprocal to June, as people come back from vacation, although most of the days for this last month have little accidents.

We can see that worst hour for driving is around 4 pm, which remains unchanged regardless of any other condition, and that there are two local maxima at 9 am and 2 pm. In general, from 1 pm to 6 pm there is a spike on accidents, coinciding with rush hours (kids come back from school and adults from work), and the least amount of accidents are from 1 am to 6 am, maybe influenced by the fact that people at night pay more attention as it is often though to be more dangerous to drive at late hours.

When analyzing the causes for the traffic accidents, we can see that it is improper driving and traffic rules violation the most usual reason for NYC collisions, followed by driver distractions and vehicle interaction factors, accidents due to other vehicles, (leaving aside unspecified reasons).

5 Results and Conclusions

To sum up, worst conditions for motor vehicle collisions and crashes in New York City for the summer months of 2018 considered by their own (if added as combination most of them remain unchanged although there are some differences that can be discovered by the reader through the application) are as follows: in terms of location Brooklyn borough, June month around 4 pm, for time dimension, under rainy conditions, traveling by car and due to improper driving and traffic rules violation.

Furthermore, this study has only been conducted for the summer months of 2018, to ease the computations and focuses mainly on the data visualization and analysis side. However, with a proper data management pipeline, in a further study it could be extended to cover a wider range of time and include more years in order to get more insights. Approaches to include data prediction algorithms based on interaction with environmental factors such as the exposed in this paper, using the proposed application to analyze results could also be of interest to extend this research.

The moment data is available and has been proved to enhance public safety, there is a moral obligation to develop the necessary tools to use this data in such important aspects as road safeness.

References

[1] Eyad Abdullah and Ahmed Emam. Traffic accidents analyzer using big data. In *2015 International Conference on Computational Science and Computational Intelligence (CSCI)*, pages 392–397, 2015.

[2] National Centers for Environmental Information (NCEI). Climate data online search. https://www.ncei.noaa.gov/cdo-web/search, 2025.

[3] NYC Office of Technology and Innovation. Nyc open data. https://opendata.cityofnewyork.us, 2025.

[4] Benjamin Romano and Zhe Jiang. Visualizing traffic accident hotspots based on spatial-temporal network kernel density estimation. In *Proceedings of the 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, SIGSPATIAL '17, New York, NY, USA, 2017. Association for Computing Machinery.

[5] Shujie Zhang. Urban traffic accident prediction research based on meteorological data. In *2022 International Conference on Machine Learning and Knowledge Engineering (MLKE)*, pages 130–133, 2022.