Analyzing vehicle collisions and traffic violations in New York City

Omkar Patinge
New York University
New York, U.S.A.
opp212@nyu.edu

Shreya Kakkar
New York University
New York, U.S.A.
sk7748@nyu.edu

Urmi Jadhav
New York University
New York, U.S.A.
uj256@nyu.edu

***Abstract*—**

There are a huge number of motor vehicle collisions reported across the world, leading to major loss of precious life and property. Having said that, we thought it would be insightful to draw a relation between the petty offenses and the major one's. The idea we are trying to establish is that regions where the residents showcase a general indifference towards trivial traffic rules like parking violations, skipping a red light, etc are more susceptible to life threatening road accidents.

Keywords— big data, Impala, MapReduce, Apache Hadoop

I. INTRODUCTION

The main idea of the project is to find a link between the amount of traffic violations that happen across the different boroughs and the number of motor collisions. We wish to show that if the residents exhibit a general trend of carelessness concerning the trivial traffic rules, then it is more likely that those areas will be more prone to major road accidents like motor vehicle collisions. The violations taken into consideration are: illegal parking and moving violations (backing unsafely, use of the wrong lane, improper turns, not giving proper indicators, skipping traffic lights, etc.) The data will be grouped on the basis of boroughs and precincts and the different patterns will be recorded.

II. MOTIVATION

There are a huge number of motor vehicle collisions reported across the world, leading to major loss of precious life and property. Our work here finds a link between minor traffic offenses and the motor collisions. This analytic can be used by the traffic police department to encourage a positive attitude towards all traffic rules alike, among the residents and hence bring down the loss of life**.**

III. RELATED WORK

[1] The paper describes big data as a framework for processing collision data from California. The project uses DataBricks Spark over Hadoop and explains its advantages like its faster speed, ability to run as a standalone model or on top of HDFS, processing data in-memory as well as access to comfortable APIs for Java, Scala, Python and Spark SQL. The analysis concentrates on the SWITRS data from 2009-2013 and gives actionable insights like types of collisions, frequencies of collisions and timings of collisions. The paper provides a list of important factors involved in the analysis of the traffic data. This can be helpful to the law enforcement and traffic departments to better manage the traffic, prevent collisions as well as take measures to ensure safety of the public.

[2] The paper investigates the causes of accidents from various data sources and predicts whether a particular accident could be fatal or not. The datasets were filtered using Spark, Scala and Zeppelin and common variables were found out from both. Google maps were used for geocoding, to obtain the missing parts of the data. Data visualization was done by Zeppelin, HighChart, ArcGIS, and QGIS. Predictive modeling was done by Spark's machine learning libraries. The paper provides a list of important factors involved in the analysis of the traffic data. The machine learning algorithm used on data, random forest, can be used for our traffic dataset for classification of accidents. The method used to compare the cities can be used in our project to compare the boroughs.

[3] The paper debates on how Spark is a solution to most of Map Reduces' shortcomings. The main data structures/ techniques used for the analytic are Decision Tree(DT) and Apache Spark. The author throws some light on how the data is grouped into classes having some common features in the decision tree. ETL and profiling is done on the initial dataset to get rid of the spurious data, further Spark(c4.5 algorithm) and ML techniques are used to extract the decision making rules for the DT. The paper concludes with a correct prediction rate of: 91.12% and and an incorrect prediction of: 7.88% in the training data.

[4] Traffic accidents are increasing with urbanization and this paper demonstrates a software that tries to analyze the traffic data to derive insights into the problems causing the accident and look for a solution for this. The architecture proposed in the paper uses MySql to store the data. When data mining techniques are to be used, the data is loaded into Hadoop and Mahout for processing. The software can be used as a service by AJAX calls. The application provides an analysis of the huge traffic data set using Hadoop for more complex operations. The paper demonstrates various important functions which are essential in analyzing any kind of traffic data.

[5] This paper analyses the relation and pattern between moving violation and accident with some error margins. The study considers over 99 road surveys with 144 features

extracted. They take into consideration the similar variables in each of the studies and try to see the behavior of it on the results. The study uses meta-analysis to analyze the relation between crash and violations considering factors like age, sample size, time, regional factors and the type of data used. To keep consistency in the comparisons of the papers the factors like mean number of crashes, the mean number of traffic offenses, the timeframe for sampling crashes and the mean age of study participants were collected.

[6] The traffic accident data are analyzed, modeled, forecasted and verified. The accident data is merged with the geolocation data for the dataset to be completed. Heat map and a normal map with pointers are used for data visualization of the current dataset. Fisher Linear Discriminant is used to identify the patterns in the dataset. Random forest and decision trees are also used for classification of the dataset into 1(Minor accident), 2(General accident) and 3(Major Accident) type of accident. The paper provides 3 algorithms which can be used in our project for accident prediction types given the various factors in our dataset.

IV. Design and Implementation

A. Design Details

The unnecessary columns are pruned from the data sets and the zip codes are linked, based on the precinct numbers. All of this is done using MapReduce in Dumbo. Further, cleaned data is injected into impala and find aggregations on the counts of traffic violations and the motor collisions. We use statistical analysis tools to find a correlation between the minor and major offenses. In the end we plan on using tableau for representation of the data.

B. Description of Datasets

1. Motor Vehicle Collisions

The dataset contains a breakdown of every collision in NYC by location and injury. Each record represents a collision in NYC by city, borough, precinct and cross street.

2. Parking Violations Issued

The dataset contains a list of all the parking violations reported in NYC by location. Each record represents a parking violation in NYC along with the violation code, borough, time and street.

3. Traffic Tickets Issued Per Year

The dataset contains records of tickets on file with NYS DMV. The tickets were issued to motorists for violations pertaining to the involvement of a motor vehicle in acts of assault, homicide, manslaughter and criminal negligence resulting in injury or death.

4. Weather data in New York City

The dataset contains records with date, temperature, snow fall and rainfall information in NYC.

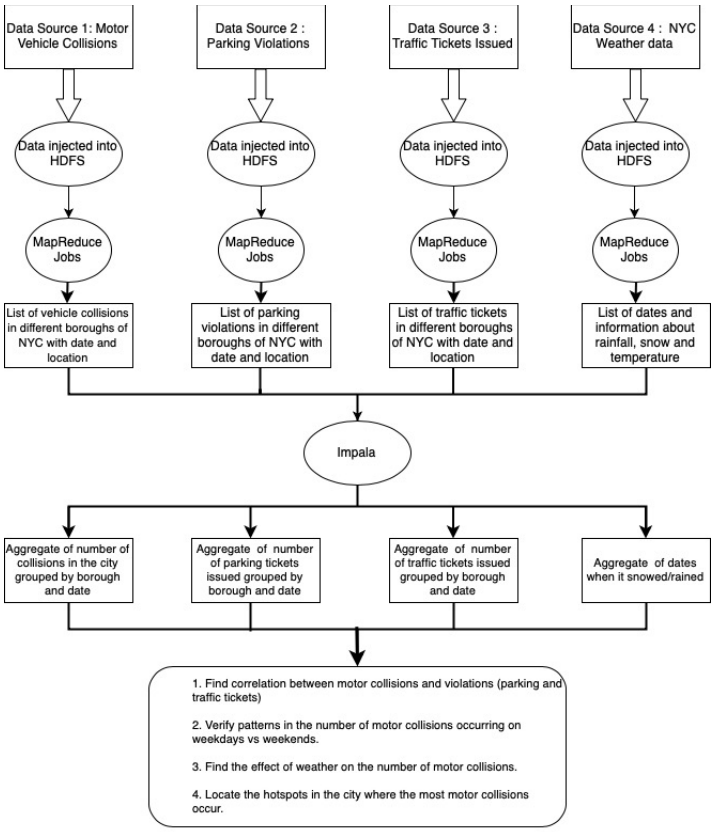


Fig.1. Motor Collisions and Traffic Violations Design Diagram

C. Data Cleaning and Profiling

The datasets are filtered to remove unnecessary columns and records using MapReduce. The data only includes the records for the years 2016 and 2017 for the analytic. Records with null values, state other than New York or inconsistent date formats were deleted. Other unnecessary columns are also filtered out.

D. Analytic

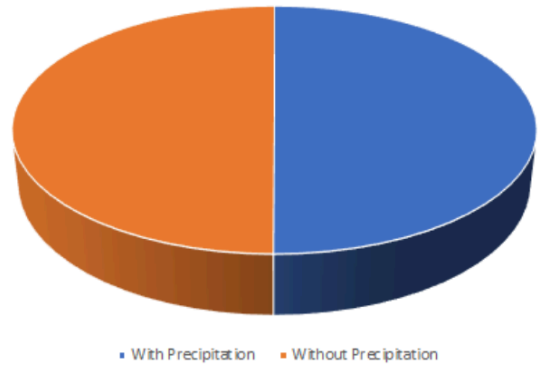
1. Correlating traffic offenses and vehicle collisions:

The hypothesis behind this analytic was that the areas where people have a general disregard for traffic rules, the chances of vehicle collisions happening in that area was higher. The traffic tickets dataset and non-moving traffic violations dataset were aggregated according to month and year borough-wise. A similar transformation was done in the motor collision dataset and a correlation coefficient of 0.88 was achieved. The graph shows the side by side comparison for the year 2016. The orange bar indicates vehicle collisions and blue bars indicate traffic violations.

2. Effect of weather on Motor Collisions

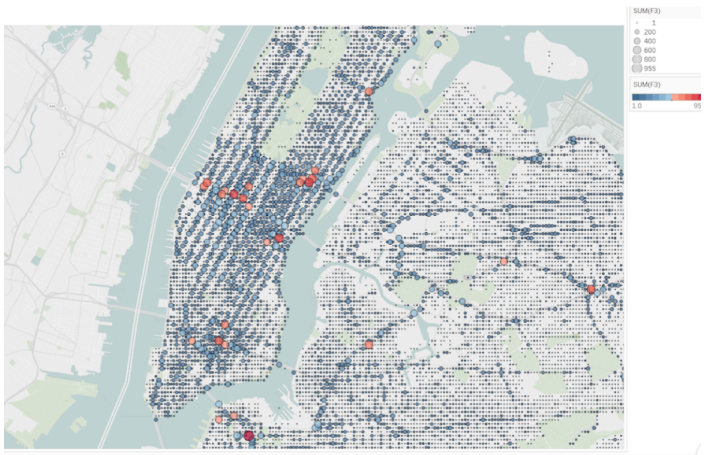
The weather data for 2016 for NYC was taken into consideration along with the Motor Collision dataset. The Motor Collision dataset was transformed for daily accidents counts for comparing it with the weather. Unfortunately, there wasn't any correlation between weather(precipitation) and motor vehicle collisions. This shocking result can be attributed to some missing data in the Motor Collision dataset. Also as the weather data was daily which could affect the result.

The following pie chart shows the accident ratio for days with precipitation and the ones without



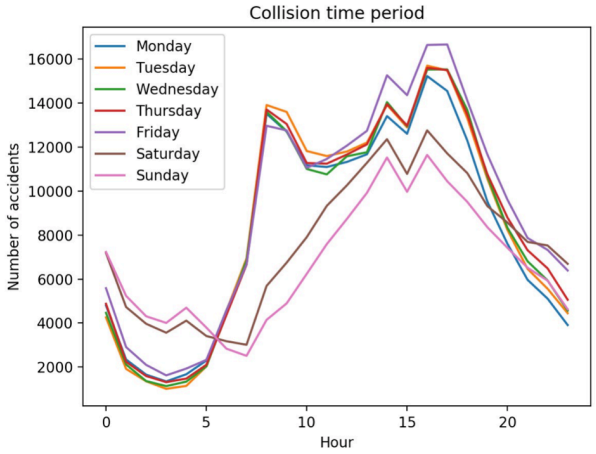
3. Collision Hotspots

The motor collisions dataset had GIS-based data of the accident areas. Hence using latitude and longitude(up to 3rd decimal place), the accident hotspots were identified up to a 100m accuracy. A lot of police reports online bolster our claim about the hotspots. The red bubbles indicate the most accident-prone areas.



4. Identifying the peak accident hours and days

The collision dataset can be transformed to give the day of the week and along with the time. Plotting the data shows that the accident occurs in the peak travel time of 8am-9am and 3.30pm-5pm on weekdays. The graphs clearly show that the accidents are very less on weekends.



V. RESULTS

The main analytic in our project shows that there is a strong correlation between the number of motor collisions and petty traffic violations. This is proved by the correlation coefficient which is 0.8874 in this case. We expected this value to be higher but since the data is not up to the mark(which can be seen by the exceptionally low values for the months of October-December) the correlation coefficient suffers.

The second analytic shows the hot spots for accidents in the boroughs with an accuracy up to 100m distance.

The third analytic shows that there is a similar trend for all the weekdays with more collisions happening during office hours (6-8am and 5-7pm). There are more collisions on weekdays as compared to weekends. This is in tandem with what we expected.

VI. FUTURE WORK

We can apply this analytic to other major cities like Chicago and Dallas. Two datasets lacked latitude and longitudes which affected the quality of our analytic. Therefore, a better weather dataset with hourly data and proper weather information and detailed information of the petty traffic offenses with latitude, longitude based data can be used for improved analytics.

VII. Conclusion

The correlation between the number of collisions and petty offenses was lower than we expected. This can be attributed to the fact that the data is inconsistent for some months. The trend for collisions for weekdays and weekends is in accordance with our expectations.

Acknowledgment

We would like to thank Professor Suzanne McIntosh for her constant support and guidance. We are also thankful to NYU HPC for providing access to Dumbo and other big data tools. Lastly, we would like to express our gratitude to the graders and fellow classmates for resolving our queries promptly.

References

1. Manik Katyal, Parag Chhadva, Shubhra Wahi & Jongwook Woo. Big Data Analysis using Spark for Collision Rate Near CalStateLA. Global Journals Inc. (USA) ISSN: 0975-4172
2. Rene Richard, Suprio Ray. A tale of two cities: Analyzing road accidents with big spatial data. IEEE International Conference on Big Data (Big Data), 2017
3. Addi Ait-Mlouk*, Fatima Gharnati, Tarik Agouti. Application of Big Data Analysis with Decision Tree for Road Accident. Indian Journal of Science and Technology, Vol 10(29), August 2017
4. Eyad Abdullah, Ahmed Emam. Traffic Accidents Analyzer Using Big Data. International Conference on Computational Science and Computational Intelligence (CSCI), 2015.
5. Peter Barraclough, Anders af Wåhlberg, James Freeman, Barry Watson,Angela Watson. Predicting Crashes Using Traffic Offences. A Meta-Analysis that Examines Potential Bias between Self-Report and Archival Data. PLOS one,Aus April 2016
6. Chen CHEN. Analysis and Forecast of Traffic Accident Big Data. ITM Web of Conferences 12, 04029 (2017)
7. T. White. Hadoop: The Definitive Guide. O'Reilly Media Inc., Sebastopol, CA, May 2012.
8. A. Gates. Programming Pig. O'Reilly Media Inc.,Sebastopol, CA, October 2011.
9. J. Dean and S. Ghemawat. MapReduce: Simplified data processing on large clusters. In proceedings of 6th Symposium on Operating Systems Design and Implemenation, 2004.
10. S. Ghemawat, H. Gobioff, S. T. Leung. The Google File System. In Proceedings of the nineteenth ACM Symposium on Operating Systems Principles – SOSP '03, 2003.