NYC SafeGo Application

Yue Luo
Dept. Computer Science
New York, USA
yl6127@nyu.edu

Cong Yu
Dept. Computer Science
New York, USA
cy1505@nyu.edu

Jiaqi Liu
Dept. Computer Science
New York, USA
jl8456@nyu.edu

***Abstract*—**

Walking around alone on the street can be terrifying for most of the people, especially in the middle of the night in New York City. It may not be ideal to find an accompany right away but would be more assuring to get a better idea of the safety index in the neighborhood. The objective of our project is to develop an application that can help individual identifies and avoids possible surrounding danger. The safety index comes from the analytics of four datasets, including 311 service request data, NYPD complaint data, NOAA weather data, and NYC street centerline data.

Keywords—Spark, Public Safety, Crime, 311 reports, Weather

I. Introduction

Criminal action is a threat to every good behaving citizens in the world. New York City, one of the major cities in the United States with low crime rates in today, used to have critical crime rates in the 80s and early 90s as the social consequences of crack epidemic. The New York Police Department, abbreviated as NYPD, with many other law enforcement agencies, has used a variety of tactics to maintain the order of this city by preventing crime from happening and bringing criminals to justice. Combating crime is a war for all, and we have conducted a big data analytics project focusing on crime predictions and to develop a safety route guidance application to contribute to the well-being of this city.

The objective of analytics is to develop a comprehensive understanding of crime patterns within Manhattan, New York, which enables us to flag potential crime hot zones at any given time, police arrest histories, weather, and the non-emergency requests. The outcome of this analytics can facilitate the law enforcement agencies to optimize its police deployment and patrolling strategy and to further reduce crime from happening.

The final route guidance application is built on top of the insights learned from the crime pattern analytics. Its target users are individuals traveling within the Manhattan island. With given input of starting location and final destination, the application is able to provide a relatively safe route choice with low probabilities of ending up as a victim in a crime.

Our datasets for the analytics are NYPD's Complaint data, 311 (the request through the city's non-emergency municipal services) data, NYC City Street Centerline data and NOAA weather data, obtained from NYC Open Data and NOAA. We have extracted the datasets from NYC Open Data and NOAA, ingested them into NYU Dumbo Clusters, and performed data cleaning and transforming using Spark. The raw dataset sizes of NYPD complaint data, 311 data, and NOAA weather data, NYC City Street Centerline respectively have roughly 2 Gigabyte, 12 Gigabytes, 1 Megabyte, and 44.8MB.

The technology used for analytics are Apache Spark framework, in particular its Spark SQL and Data Frames to perform data manipulation and MLLib to perform supervised gradient based machine learning, and the Hadoop Distributed File System on Dumbo machine from the Prince cluster of NYU High Performance Computing.

We organize our paper as follows: The next section presents the motivation for the reasons of performing this analytics; Section III presents the related work covering aspects from data filleting algorithms to data profiling algorithms such as using Spark MLlib to determine the likelihood of causing a crime for each factor in our analytics; It is followed by Section IV which includes the detailed schemas regarding our three datasets; In Section V, we formally introduce how we come to the conclusion via the data analytic; then, we demonstrate our big data application design UI in the Section VI; after that, we share the actuation responses designed in the application in Section VII; Section VIII will discuss the limitations and the problems we ran in both the analytics and during implementing the application; we conclude the analytics project and the development of our safety route application in Section IX. Finally, we will discuss all future works that can expand the project Section X.

II. Motivation

This application can be important to increase public safety awareness. For instance, individuals who will be walking on the street during night times; Government who wants to improve the security rating of a certain neighborhood, the machine learning produced by this project can also help them to determine where to allocate necessary resources.

III. Related Work

Several research works have been conducted in the crime analysis field. In this paper [1], it looked into the relationship between weather and aggressive crime, in particular, whether thermal conform is correlated with aggressive crimes. The author begins with 4 major theories on violent crime model: Negative Affect Escape Model, General Affective Aggression Model, Routine Activity Theory, and Social Escape/Avoidance

Model. Each theory contains its conjecture regarding the crime pattern. To verify the soundness of each model, the author analyzed the crime data between 1999 and 2004 in Cleveland, Ohio. The author filtered out all crime data other than the following six categories: domestic violence assault, non-aggravated assault, aggravated assault, robbery, rape, and homicide. The crime data are analyzed in a daily level, 3-hour interval level, spatial level. The author also analyzed weekends and weekdays crimes separately to better fit into some of the crime model mentioned above. The analytic result indicates that the crime counts follows a diurnal cycle that is similar to the social patterns of humans. There exists a general linear increase of all aggressive crime types. The summer has the highest aggressive crime count and the winter has the lowest crime counts. In addition, non-aggressive assaults and domestic violence assaults have the great response to temperature. In conclusion, the study strengthened the Routine Activity Theory and General Affective Aggression Model. The Social Escape/Avoidance Model is only partially supported by the analytics and there is little support to the Negative Affect Escape Model.

Some researchers are also trying to find ways to find solutions based on crime data. The app proposed by this work makes use of three data sources, including Bahía Blanca Municipal Open Data API, Google Maps API and Facebook API. Bahia Blanca Municipal Open Data API provides information about the 911 emergency service's events in real time, and information about the city crime data in the map. Google Maps API is used to display or create geographical information on a map. All 911 emergency services and crime events are marked on a Google Map. The event can be shown to users when they are available and get pressed. Users can report the issues through Facebook APIs to their friends lists. The inspiration from this paper comes from not only the concept- build an application that could improve citizen's safety, but also the idea that users can share the safety information among through the Facebook APIs. The application that is described in this paper also supports safety information sharing among users. They can provide tips and reported issues in the user's proximity, as a way for keeping users engaged. We are also thinking about including a similar feature into our application that enable users in the same neighborhood to find accompanies. But we are researching on this feature and deciding its feasibility at this moment. Also, the application gives us a fresh start to think about what we can bring to the table about citizen's safety. Similar to this research, we also have the crime data and geographic data, we can build a similar product that tell users about all danger events in their proximity based on our analytics and make this information available on the map when they pass by [2].

Weather is an important environmental factor regarding crime although some may post different opinion. According to the finding of Tompson et al [3] based on robbery data provided by Strathclyde police and detailed local weather data, mean temperature and humidity have significant influence on robbery. However, other seasonal factors such as fog, rain, snow and seasons themselves shows little correlation. A hypothesis for it is that people tend to reduce outdoors activity for weather in the short term while they keep norm activities in the average of the whole seasons. And other papers [4], [5] support the result using combined data source from tweet and weather. But they two vary on the choice of methods. Streetlight is another significant factor for crime. A research [6] using the crime data from Detroit shows that street light intensity is greatly correlated to all crime and three selected major crime (burglary, stealing vehicle, weapon offence). Not surprisingly, another finding from England [7] proves the same result using light condition (on or off) instead of light intensity. One limitation of their finding is that they only investigated 71 examples. But the two paper both look into robbery, vehicle and violence as important factors besides whole crime.

There are also study focuses on certain crime type. In this study [8], the authors obtained a collection of data of over 400 cases of convenience store robberies in metropolitan areas of Alexandria, Richmond, and Norfolk, Virginia between February 1, 1995 and September 30, 1996. In order to understand the significance of each factor that contributes to a robbery, the authors applied a case-control methodology by using conditional logistic regression. With this approach, the researchers has carefully assessed the environment of each store on many factors such as the existence of ATM machine in store. One of these factors that I am interested in, and may be beneficial to our crime analytics project is the factor that considers the visibility from the inside, outside, and within the store. This is, to some extent, similar to our analytics regarding whether dark or broken street light may induce more crimes to late night pedestrians. Many robbery cases indicated that visibility to the inside of the store was strongly associated with robbery, as stores with poor visibility to the inside of the store were at twice the odds of robbery. Therefore, In the end, the study concludes that there exists a significant correlation between brightness and robberies, and the result suggested that areas with low visibility can be one of the motives for crimes such as robbery since the likelihood of being witnessed is small.

Comprehensive studies about 311 calls across the country helps us to get a better understanding about the 311 dataset. As it stated [1], the 311 services not only alleviates the burden of 911 congestion but also brings a number of benefits to the community that empowers them to detect problems and provide feedback in time to the government. Using the 311 open data sources, scientists have developed an analysis portal for real-time data visualization to explore request trends at various time units and compare requests cross different cities. Besides these, they also did the privacy information-PII tests. Through their analytics, it is interesting to find out that the citizen engagement has been improved through the 311 services for cities such as NYC and Boston. However, for the PII, they found out that not all open dataset has removed the sensitive personal information, which would be desirable for data anonymization. Diving deeper into the datasets, the common types of 311 requests for each cities have also been plotted using streamgraphs, which could be a good demonstration on our insight presentation. As time progresses, the scientific community have shown a growing interest in making the environment smart. The research team has shown the interest in developing an application for noise detection using 311 call data [2]. Based on the noise analytics and 311 complaints, they proposed a Smart311 system that can detect the noise happening around

and classify and prioritize the noise using machine learning algorithms to interact with the 311 service server through a mobile client. The Smart311 system can report to 311 server without direct involvement of neighborhood. The machine learning algorithms used in this application are CNN, random forest, LSTM, SVM and decision tree. The experiments shows that the CNN achieves the best performance. Since machine learning is also a consideration in our development, this work serves as a guidance in understanding the 311 data and machine learning approach on the data set.

There are also studies that mainly discusses the inference of crime rate of regions in Chicago from three aspects: demography, geography, and transportation [11]. For demographic investigation, they collected: total population, population density, poverty, disadvantage index, residential stability, ethnic diversity, race distribution. Among all the subfactors, Poverty Index, Disadvantage Index, and race distribution are found to be strongly correlated to crime. The second aspect they inspect is the location data. The researchers use POI (points of interest) data from Foursquare as geographical information. Based on POI data, areas of the city are divided into 10 major categories such as food and education. However, only the "professional" category is found to be the crucial factor of crimes. Also, they depict how transportation would influence crime rate by tracking taxi trips. They conclude that overall the crime rate is positively correlated with the taxi flow. After doing feature selection and correlation analysis, scholars construct a linear regression model to inference crime rate based on the factors discussed above.

IV. DATASETS

A. NYPD Complaint Data

The link to the data is: https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Historic/qgea-i56i

This dataset includes all valid felony, misdemeanor, and violation crimes reported to the New York City Police Department (NYPD) from 2006 to the middle of year (2019). This dataset is released on NYC open data platform. This is a static dataset and we load it once. The size of the dataset is 2GB.

The schema of this data is as follows,

Variable Name	Variable Type	Description
Date	String	Date of occurrence for the reported event
Hour	integer	Representing the hour of the Exact time of occurrence for the reported event
Crime_Type	String	Description of offense corresponding with key code
Crime_Level	String	Description of offense in felony, misdemeanor, violation
Borough	String	The name of the borough in which the incident occurred
Latitude	double	latitude from 40.657273946 to 40.498905363
Longitude	double	longitude from 73.684788384 to 77.519206334

B. 311 Service Request from 2010 to Present

The link to the data is: https://data.cityofnewyork.us/Social-Services/311-Service-Requests-from-2010-to-Present/erm2-nwe9

NYC311's mission is to provide the public with quick, easy access to all New York City government services and information while offering the best customer service. Each row represents a 311 service requests. This dataset is released on NYC open data platform. This is a static dataset and we load it once. The size of the dataset is 12GB.

The schema of this data includes date, time, complaint type (drinking in public, street light, homeless encampment), latitude and longitude.

Variable Name	Variable Type	Description
Date	String	Date of occurrence for the reported event
Hour	integer	Between 0 to 23 representing the hour of the complaint
Type	String	Description of the complaint type
Latitude	double	latitude 40.498948846168354 to 40.912868795316655
Longitude	double	longitude from 74.25495171973925 to 73.70059684703173

C. NOOA Weather Data

The link to the data is https://www.ncdc.noaa.gov/data-access/land-based-station-data

Weather data collected from the National Weather Service. It contains for each day the minimum temperature, maximum temperature, average temperature, precipitation, snowfall, and current snow depth. The temperature is measured in Fahrenheit and the depth is measured in inches. This dataset is released on NYC open data platform. This is a static dataset and we load it once. The size of the dataset is 200KB.

The schema of this data includes date, wind speed, precipitation, Snow, max temperature, min temperature and weather type.

Variable Name	Variable Type	Description
Date	String	Date of each weather record
Wind_Speed	double	Wind speed (meters per second)
Precipitation	double	Precipitation (mm)
Max_Temperature	double	Temperature in celsius
Min_Temperature	double	Temperature in celsius
Weather_Type	integer	Denoting the type of weather

D. NYC City Street Centerline (CSCL)

The link to the data is: https://data.cityofnewyork.us/City-Government/NYC-Street-Centerline-CSCL-/exjm-f27b

A road-bed representation of New York City streets containing address ranges and other information such as traffic directions, road types, and segment types. The size of the data is 44.8MB.

The schema of this data is as follows,

Variable Name	Variable Type	Description
Physicalid	integer	Unique street ID
Full_Street	String	Street name
Shape_leng	double	Street length from 4.51 to 6974.75
Polygons	double	An array of string denoting the street latitude and longitude pair

V. Description of Analytic

A series of analysis have been conducted on these three datasets, which yield some interesting results to convey the story of crime occurrences. Our analytics can be categorized as three parts, first of all, the weather impact on crime rate; secondly, the occurrence of drinking and homeless events impact on crime rate; then finally, the occurrence of street light out events impact on crime rate. The following section will be focused on the results from these three analytics.

Since the crime data is the vital part in the scope of this project. There are several individual analytics towards the crime dataset (Figure 1.2. depicts crime events in Manhattan area). In order to understand better about the spread of crime occurrence, analytics about the relationship between locations and time of the day has been conducted on the crime data of past 9 years. The Figure 1.1. below shows the crime count in

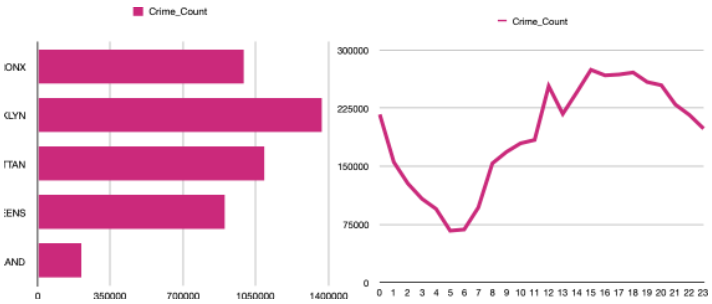

Figure 1.1 Crime count by Boroughs and by Hour of the day

terms of different boroughs in New York City and during different hours of the day. Among the five boroughs, Brooklyn has the highest crime count, which is 1369718, while Staten Island has the lowest number of crimes, which is 21057. Bronx, Manhattan and Queens seem to have similar amount of crime events. The safety level of different boroughs cannot be concluded right away because the population has not been taken into consideration. From the trend of crime count during different hours, crime occurrence tends to decrease during midnight and dawn time (hour 00 - 06) while comes to a surge during the daytime (hour 07 - 17). Even though the crime count in daytime tends to larger than that in the evening time, it is possible the result of increasing human activities in the daytime. Similarly, without the information about population of each hour, it cannot be concluded safely that dawn time is safer than any other time during the day.

Besides this, crime occurrences of different crime types and crime levels are also interesting topic to look at. There are in total 70 crime types in the crime dataset. The top 10 crime types are 'petit larceny', 'grand larceny', 'harassment', 'criminal mischief & related of', 'assault 3 & related offenses', 'off. angst pub ord sensibility', 'burglary', 'dangerous drugs', 'felony assault' and 'robbery'.

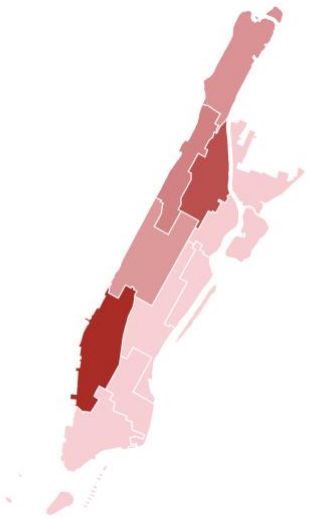

Figure 1.2 Crime count in Manhattan

A. Crime and Weather

Environmental factors can influence an individual's mindset, which may or may not push offenders commit certain crimes. From the available fields in weather dataset, the following analytics will be focused on whether temperature changes and precipitation will influence the occurrence of crimes.

1) Precipitation

Environmental factors can influence an individual's mindset, which leads to the following analytics set up about whether precipitation plays a role in determining crime count. Also, more specificity, whether precipitation acts different for each crime type or crime level. To make this happen, crime counts are aggregated by days, which are then joined with the weather data to get the crime counts for each precipitation level. This process has been conducted on the general crime, each crime level and top 10 crime types respectively. The tendency can be seen from Figure 2. Clearly, there is a negative correlation between precipitation and crime count, which means that the occurrence of crime shrinks as the level of precipitation increases. Furthermore, after comparing with different crime levels and crime types, this tendency seems to hold under these circumstances. There can be several reasons that crime count negatively correlated with precipitation level. First of all, precipitation can restrict the number of pedestrians, which in turns decrease the number of crimes. Furthermore, precipitation can restrict the selection of crime tools for offenders.

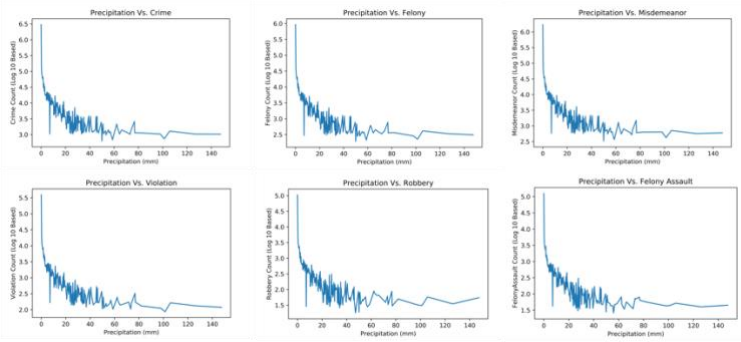

Figure 2. Precipitation verse Crime Count

2) Temperature

Temperature can be another factor that acts on the crime rate performance for the reason that temperature change can limit the pedestrian activity and alter the mindset of offenders. For example, higher temperature can make people more easily get furious, anxious and stressed than normal weather. With these assumptions in mind, the relationship between crime count and average temperature for each month in the past nine years is displayed in Figure 3. The average temperature for each month is calculated by average the Max_Temperature and Min_Temperature in the NOOA weather data. The crime count is presented by the number of crime events for each crime level. From the graph, there is no strong correlation between temperature and crime count. The number of Felony and Misdemeanor crime fluctuates as the temperature changes, while the number of Violation crime remains nearly the same in all months. Overall, there is no strong evidence to conclude that the crime rate is impacted by temperature.

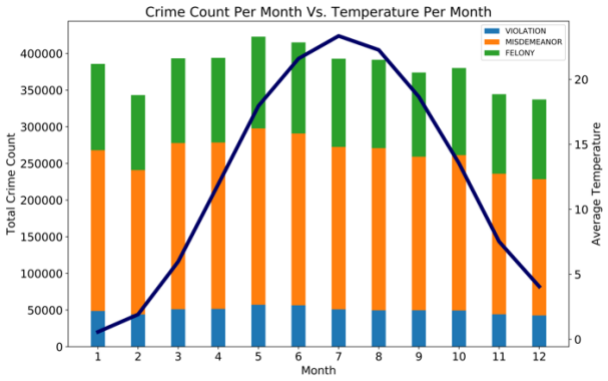

Figure 3. Crime count Per Month Verse Temperature Per Month

B. Crime and Drinking in Public and Homeless encampment events

As drinking in public and homeless encampment may create hidden danger to the surrounding environment, it is absolutely worth the interest to see if these two events have impacts on the crime rate in the neighborhood. To be more specific, the number of crimes a 311 complaint regarding homeless encampment within a block radius is collected. The block radius is 0.0018 for both latitude and longitude.

For drinking in public event, the following analytics is acted on the two datasets. The crime that occurs within 4 hours window after the 311-complaint call is considered here since we want to look at the impact after the event occurrence and this type of event cannot take effect permanently. The *Figure 4.1.* shows the spread of crime levels that happens under this set up. From this graph, Felony and Misdemeanor crimes together occupies the most part of crime occurrence. The *Figure 4.2.* displays the top 10 crime types happens near the drinking in public events. However, there is only 333 such events happen, which concludes that drinking in public event has no direct impact on crime rate in the neighborhood.

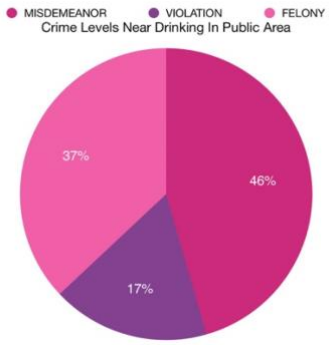

Figure 4.1. Crime Count Near Drinking in Public Area by Crime Levels

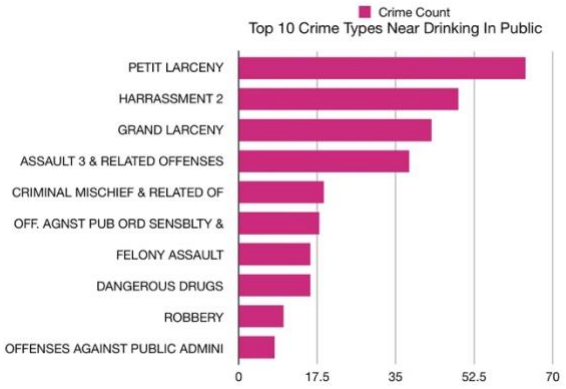

Figure 4.2 Crime Count of Top 10 Crime Types Near Drinking in Public Area by Crime Levels

For the homeless encampment event, the following analytics is acted on the two datasets. The crime that occurs within 6 hours window after the 311-complaint call is considered here since we want to look at the impact after the event occurrence and this type of event cannot take effect permanently. The *Figure 5.1* shows the spread of crime levels that happens under this set up. From this graph, Felony crimes occupies nearly half of crime occurrence. The *Figure 5.2.* displays the top 10 crime types happens near the homeless encampment events. However, there is 6222 such events happen, which concludes that homeless encampment event has some impact on crime rate in the neighborhood.

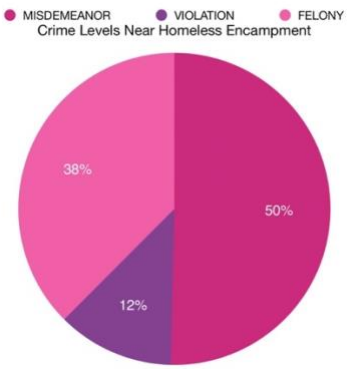

Figure 5.1 Crime Count Near Homeless Encampment *by Crime Levels*

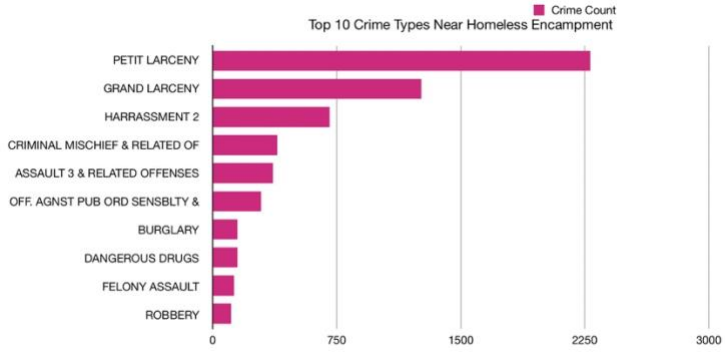

Figure 5.2 Crime Count of Top 10 Crime Types Near Homeless Encampment *by Crime Levels*

C. Crime and Street Light Out Events

Sudden darkness could also be a lead for crime occurrence. Broken or flashy Street light out events are recorded in the 311-compliant data and this complaint may take days to close, which provides possibilities for crime to occur. Under the nature of this type of event, the crime that occurs within 14 days window after the 311-complaint call is considered here since we want to look at the impact after the event occurrence and this type of event cannot take effect permanently. The Figure 6.1. shows the spread of crime levels that happens under this set up. From this graph, Felony and Misdemeanor crimes occupies nearly the same and half of crime occurrence. And there are few Violation crimes happen near this event. The Figure 6.2. displays the top 10 crime types happens near the homeless encampment events. However, there are 150804 such crime happens near this kind of event, which concludes that street light out event has an impact on crime rate in the neighborhood.

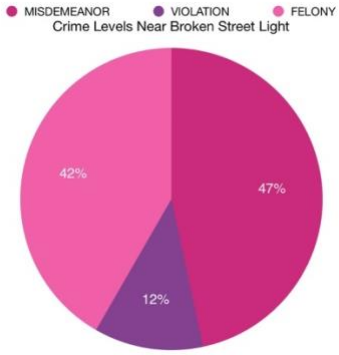

Figure 6.1 Crime Count Near Street Light Out *by Crime Levels*

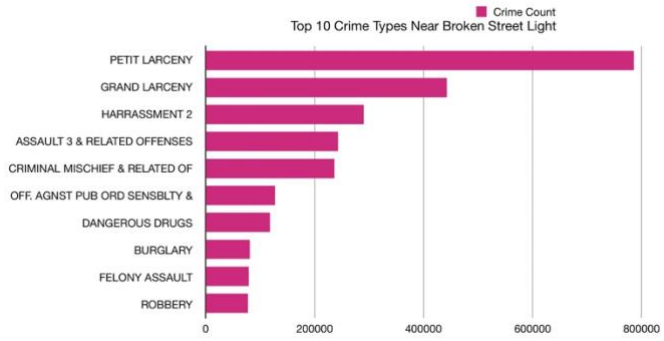

Figure 6.2 Crime Count of Top 10 Crime Types Near Street Light Out *by Crime Levels*

VI. Application Design

A. Design Details

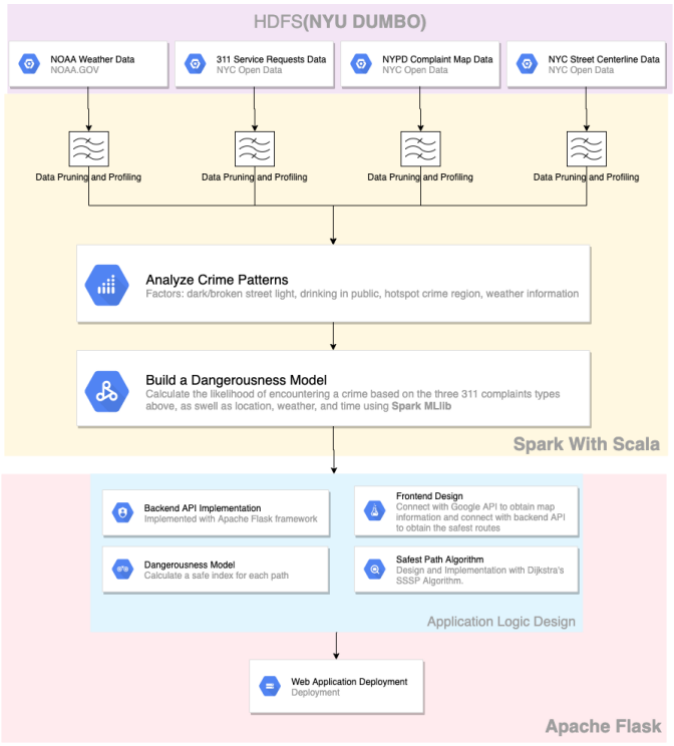

Figure 7. Data Flow Diagram

Figure 7 depicts our data flow paradigm. Crime data, 311 Service data, NOAA weather data, and NYC Street Centerline data are ingested into NYU Hadoop cluster from local file system. The data is stored in the Hadoop Distributed File System. Then, we profile, clean the data, and conduct the analysis with Spark. After the relevant data fields have been selected from analytics, the joint data set will be fed into Spark MLlib to build a dangerousness evaluation model to predict the likelihood of certain crime occurrence using logistic regression. For the web service, the front-end and back-end was implemented with Apache Flask framework. The front-end invokes the Google Map API to get the geological coordinates of the starting and ending points. Then the back-end is responsible for applying the Dijkstra's Single Source Shortest Path algorithm and the Dangerousness Model trained using Spark MLLib to output the safest and shortest path to the front-end. The front-end then, will render the path on the Google Map. In the end, this web application service was deployed at the website specified in the Actuation Section.

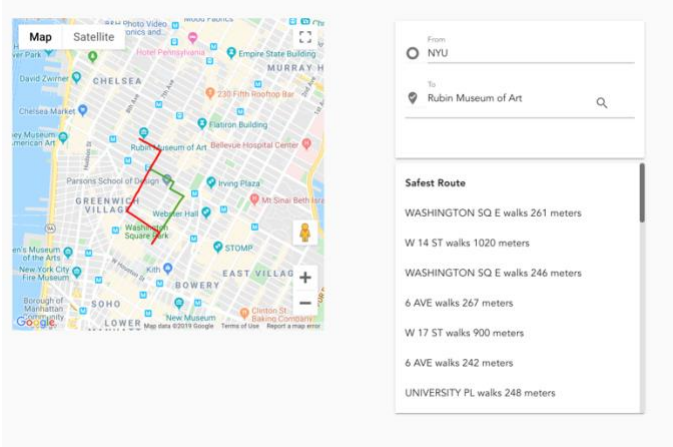

Figure 8. Application UI

Figure 8 depicts the user-interface of our web service; Users can enter their starting point and destination on the upper right side of the web page. Then the route results will be displayed on the lower right side. There are two kinds of route – safest route and shortest route. Both of the route results consist of the name of the street, walking distance and the safety results of this route choice. The left side part of the page is the google map that will show users their route results. Apart from displaying the route by drawing lines, colors will be assigned to lines according to street safety indexes. Under the scope of this project, only route search within Manhattan area is supported.

VII. Actuation or Remediation

Safety matters! The application we developed, named as SafeGo, was built on top of our analytics results.

The actuation derived can be summarized as following parts:

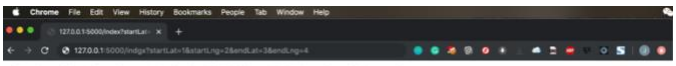

Figure 9. Application Backend Output

1.The core of this application is a backend api that calculates both the safest and the shortest path between two coordinates and returns the result to the visualization component. From a blackbox aspect (more details in the following paragraphs), this API takes four inputs: 1. start point latitude; 2. start point longitude; 3. end point latitude 4. end point longitude. The

API then produces outputs as a JSON object with two fields that describes the safest route and the shortest route between the starting point and end point. One field is named safest and another one is named shortest. Each field is an array of street objects This API has been deployed and can be accessed via this link: http://project.heliumyc.top.

The backend API, implemented using Apache Flask, performs input checking to provide stable and fast computations. It takes get requests from the following path:

```
/index?startLat=[a]&startLng=[b]&endLat=[c]&endLng=[d]
```

With the correct input, the caller will receive a 200 status code and the json objects that contains an array of Streets. Otherwise, the api will respond a 400 status code and a Bad Request string to incomplete inputs or invalid inputs.

Each Street object, similar to GeoJson object, is defined using the following schema:

- sid: Each street has a unique number id, this field is unique.
- streetLen: a number denoting the physical street length of each street
- safetyIndex: a number denoting the safety index of a street
- Polygons: an array of latitudes and longitudes numbers that represents the vertex of the street. In the format of `[{lat: Numbere , lng: Number}]`

Each street's information, such as the physical length, latitudes & longitudes of its vertices was obtained utilizing the cities street centerline information. With the geological coordinate of all NYC streets and their conjunction points, we used a very classic algorithm, Dijkstra's SSSP (Single Source Shortest Path), to compute both the physically shortest path and the safest path and the shortest path of the inquired route. In the safest path calculation, we modeled the graph with directed edges weighted by the safety index of getting from one vertex to another, apply the SSSP algorithm, and then backtrack the route.

2. The safety index of each street is calculated by performing machine learning on the analytics results. The machine learning model has been trained by ***Spark MLlib***. The model takes 8 factors into consideration, including latitude, longitude, average wind speed, precipitation, snow fall, minimum temperature, maximum temperature, weather type and complaints of homeless encampments, drinking in the streets and the damaged streetlights. Meanwhile, the goal of our model is to predict the crime level for every crime record (from 1 to 3). For simplicity, logistic regression with L2 regularization is chosen as our model. Since this is a multiclass classification task, the output layer of logistic regression is a softmax function. To choose the best parameters, we implemented grid search and cross validation. Those selected best parameters are used to train a final model with an accuracy of 60%. And at last, classification model is dumped for inference part. Despite low accuracy, we have several hypotheses for it and the result can be improved in future work. In the inference part, we loaded the model and calculated crime index by summing up different probability with their weights (1 for label 1, 2 for label 2 and 5 for label 3) for every street in Manhattan.

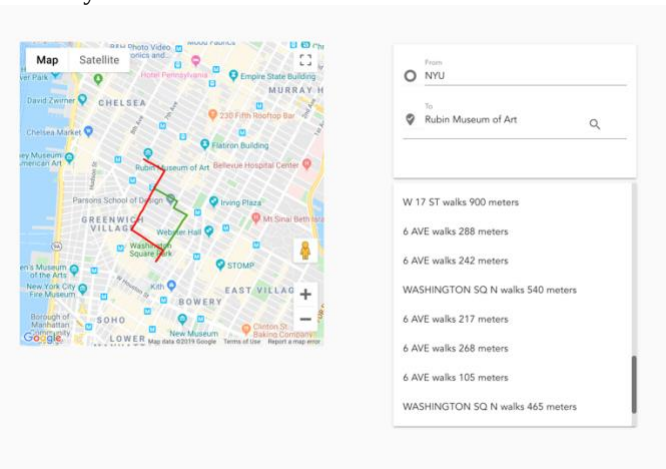

Figure 10. Application Frontend Example

3. The user are not expected to directly interact with the backend API directly, we have prepared a visualized interface serving as the front end of the overall application. The frontend is responsible for sending the request and for visualizing the safe routes. Based on the data fields that backend has sent back, the shortest path and the safest path displayed differently with distinct colors on a map in Manhattan area. In this way, users can get a better idea of the safety level of each street. Also, the specific details of the path such as every street names occurred in the path has also been displayed.

VIII. Analysis

The project is set up on Spark platform for analytics, Spark MLlib for Machine Learning, and Apache Flask for application development.

We encountered that NYU HPC Clusters are not designed for real-time analysis and so it does not have additional ports for serves from the outside to transmit data, which restrained for a real-time application. Also, since the NYC street data would contain ferry route, it is possible that some predicted route involve ferry route. This problem can be eliminated by filtering out the ferry route data.

We have developed a lot from this project in terms of skills to use Scala, Spark ecosystem, machine learning with Spark MLlib, backend developing with Apache Flask, data visualization and results documentation.

We would recommend others to understand better the dataset and also the platform when designing an application.

IX. Conclusion

In this project, we finish analytics on how crime count is related to factor of weather and 311 complaint events in New York City. The datasets we have used are 311 request data from 2010 to present, NOOA weather data and NYPD complaint data historic and current. For the analytics part, firstly, we conducted analytics on the relationship between precipitation with crime count, which we found out that it is a negative correlation. Followed by this, the relationship between temperature and crime count has also been calculated and concluded. Finally, among all the 311 complaint events, we selected the 'drinking in public', 'homeless encampment' and 'street light out' events and conducted the correlation analytics between these events and crime count in the defined neighborhood. The result turns out that the 'street light out' event has an impact on the crime count. Then an application built upon the analytics is discussed. With the application, route between any searched location with the safest index predicted can be displayed to users. In summary, we believe the insights we found from this project can be beneficial for economic and even political entities.

X. Future Work

Although we have achieved some interesting insights from the discussion above, there are more things to be finished in the future.

Firstly, for the crime analytics with boroughs and time of the day, there is room to improve in the future. If the population of the boroughs or the time of day can be obtained, the crime count per unit can be compared among boroughs or the time of the day. In this way, the relationship between crime count and boroughs and time of the day can be compared more accurate.

Secondly, for the analytics of crime count and 311 requests, the time impact of each event (4 hours window for drinking in public; 6 hours window for homeless encampment; 14 days window for street light out) as well as the block area is taken into account. Although conclusion has been made that street light out event has influences on crime count in the neighborhood while the rest of events does not, it is still possible that the results will be slightly different if the time window or the block area are adjusted.

Thirdly, the accuracy of the machine learning model can be improved if there are more relevant fields added on in the future.

Finally, the model inference is now based on manually update data from Spark processed data because the limitation of Dumbo. NYU HPC Clusters are not designed for real-time analysis and so it does not have additional ports for serves from the outside to transmit data. If the limitation can be removed in the future, the application can be achieved near real-time to update data for more smooth experience.

Acknowledgment

We would like to express our thanks to the Department of Computer Science at NYU Courant Institute of Mathematical Science and the NYU High Performance Computing group for supporting this project. We would also want to thank the New York Police Department for publishing NYPD complaint data, New York 311 Service for publishing the NYC 311 data and New York City Government for publishing the NYC Street Centerline (CSCL) data.

References

1. P. Butke and S. C. Sheridan, "An Analysis of the Relationship between Weather and Aggressive Crime in Cleveland, Ohio," Weather, Climate, and Society, vol. 2, no. 2, pp. 127–139, 2010.
2. A. G. Pereira, E. Estevez, and P. R. Fillottrani, "An innovative mobile app integrating relevant and crowdsourced information for improving citizens safety," Proceedings of the 11th International Conference on Theory and Practice of Electronic Governance - ICEGOV 18, 2018.
3. L. A. Tompson and K. J. Bowers, "Testing time-sensitive influences of weather on street robbery," Crime science, vol. 4, no. 1, p. 8, 2015.
4. Z. M. Wawrzyniak et al., "Relationships between crime and everyday factors," in 2018 IEEE 22nd International Conference on Intelligent Engineering Systems (INES), 2018, pp. 39–44.
5. X. Chen, Y. Cho, and S. Y. Jang, "Crime prediction using Twitter sentiment and weather," in 2015 Systems and Information Engineering Design Symposium, 2015, pp. 63–68.
6. Y. Xu, C. Fu, E. Kennedy, S. Jiang, and S. Owusu-Agyemang, "The impact of street lights on spatial-temporal patterns of crime in Detroit, Michigan," Cities, vol. 79, pp. 45–52, 2018.
7. R. Steinbach et al., "The effect of reduced street lighting on road casualties and crime in England and Wales: controlled interrupted time series analysis," J Epidemiol Community Health, vol. 69, no. 11, pp. 1118–1124, 2015.
8. B. H. Kim and D. Kim, "A Matched Case-Control Study of Potential Risk Factors for Convenience Store Robberies," Korea CPTED Association, vol. 8, no. 2, pp. 38–70, 2017.
9. B.-Y. Choi, M. K. Al-Mansoori, R. Zaman, and A. A. Albishri, "Understanding what residents ask cities: open data 311 call analysis and future directions," Proceedings of the Workshop Program of the 19th International Conference on Distributed Computing and Networking - Workshops ICDCN 18, 2018.
10. Z. Tariq, S. K. Shah, and Y. Lee, "Smart 311 Request System with Automatic Noise Detection for Safe Neighborhood," 2018 IEEE International Smart Cities Conference (ISC2), 2018.
11. H. Wang, D. Kifer, C. Graif, and Z. Li, "Crime Rate Inference with Big Data," Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 635–644, Aug. 2016.
12. Cohn, Ellen G. "The Prediction of Police Calls for Service: The Influence of Weather and Temporal Variables on Rape and Domestic Violence." Journal of Environmental Psychology, vol. 13, no. 1, 1993, pp. 71–83.
13. Gyimah-Brempong, Kwabena. "Alcohol Availability and Crime: Evidence from Census Tract Data." Southern Economic Journal, vol. 68, no. 1, 2001, p. 2.
14. Mares, Dennis M., and Kenneth W. Moffett. "Climate Change and Crime Revisited: An Exploration of Monthly Temperature Anomalies and UCR Crime Data." Environment and Behavior, vol. 51, no. 5, 2019, pp. 502–529.
15. Rotton, James, and Ellen G. Cohn. "Global Warming and U.S. Crime Rates: An Application of Routine Activity Theory." Environment and Behavior, vol. 35, no. 6, 2003, pp. 802–825.
16. Belesiotis, Alexandros, et al. "Analyzing and Predicting Spatial Crime Distribution Using Crowdsourced and Open Data." ACM Transactions on Spatial Algorithms and Systems, vol. 3, no. 4, 2018, pp. 1–31.
17. Ferrari, Alan, et al. "Can Smart Devices Protect Us from Violent Crime?" Proceedings of the 2015 Workshop on Wireless of the Students, by the Students, & for the Students - S3 15, 2015.