Traffic Collision Hotspots Project Update 2.0

Information Visualization Project Group 6
Bartosz Jerzy Bonczak (bjb417), Awais Malik (am5801) & Radu Stancut (rs1933)
April 15th, 2015

Introduction

Last year, 268 fatal traffic crashes occurred in New York City.[1] In order to achieve "Vision Zero" (the ideal scenario where NO fatal traffic accidents take place), the New York City Police Department (NYPD) will need to better understand city traffic patterns in order to prioritize allocation of resources to any collision "hotspots." The objective of this project is to provide the NYPD with the desired enhanced cognition of traffic collision patterns in the city by identifying these hotspots and correlating them with specific enforcement-related causes (such as DUIs and cell phone use).

Injuries and deaths related to motor vehicle accidents take a heavy toll on communities worldwide. According to the World Health Organization, 3,400 people die each day due to traffic collisions around the world.[2] The United States has experienced an annual death toll of 30,000 to 40,000 people over the past decade, with another two and a half million injured.[3] Additionally, road traffic injuries are a leading cause of deaths around the world for children and young adults.[4]

The crisis surrounding car accident deaths and injuries is a serious public health issue, one that all countries/cities must contend with (in fact, the name "Vision Zero" comes from a Swedish initiative started in the 1990's to deal with these issues). We view this project as a practical step towards better understanding the nature of NYC collisions by aiming to create an interactive visual platform that is meant to extract meaning and information not inherently accessible in its current state. We hope our visualization tool will allow the NYPD to aggregate and visualize the available data in a way that will be easier to understand and explore than before.

---

1 http://project.wnyc.org/traffic-deaths/
2 http://www.who.int/violence_injury_prevention/road_traffic/en/
3 http://www-nrd.nhtsa.dot.gov/Pubs/811856.pdf
4 http://www.who.int/violence_injury_prevention/publications/road_traffic/world_report/chapter1.pdf?ua=1

Related Work

The majority of identified papers using visualizations in their collision research leveraged figures, graphs, and maps for the sake of supporting, informing, and explaining their work and findings. The major focus was rarely to develop an information visualization tool, though there were exceptions (Yi, 2001). More often the spatial and visual elements were used to motivate and review statistical findings and models (Driss, 2015; Levine, 1995).

Before running down the different visualization uses below, a quick rundown of common approaches to mapping collisions and roadway safety included 1) a map identifying the area of study, 2) collision incidents plotted, 3) tables listing accidents broken out by features (speed, date, route, etc.), and 4) statistical formulas/models used. Beyond the aforementioned, below were common uses of visualizations.

Physical Form: Intersection or Motorway Focus – In a few instances the focus of study was drilled down to the street level and dealt with the geometry of the roads in question (Wang, 2009). Examples of this included diagram/figures portraying common road types (four-approach intersection) and a subsequent collision matrix with arrows meant to portray direction/movement of cars and point of accidents (Yi, 2001). This approach was interesting but too granular for our purposes, though there is a thematic overlap with collision points containing accident features (date, road condition) and chart summaries for intersections (location v. condition). The granular exploration of intersection collisions might be relevant for future study on identified New York City hot spots.

3D Rendering – Related to the focus on physical form was the use of three-dimensional (3D) tools. In one instance LiDAR images were overlaid on shapefiles to create a more robust representation of the physical environment, with photos for drivers' passing sight used to supplement 3D rendering (Khattak, 2005). Again, an approach not immediately applicable to our project purposes.

Loosely connected to this section was the placement of bars showcasing accidents and risks on a map projection to 3D. This was an idiom that was more visually appealing than informative. The map/chart proved difficult to read; a choropleth would have sufficed.

Collision Spatial Area & Density (Hot Spots) – There were different levels of sophistication in charting collisions on maps, from simple accident density (Soltani, 2014), to seasonal hot spots based on monsoons (Prasannakumar, 2011), to taking roadway congestion (Wang, 2009) or population density (Levine, 1995) into consideration, to finally intricate statistical means and models of grouping accidents (Erdogan, 2008); this latter example has to do with algorithm creation, beyond the scope of the project, but helps to motivate ideas about how hot spots are to be grouped generally.

Choropleth Mapping – Examples of choropleth, which we intend to use to some extent, highlighted accident fatalities and accident fatalities/capita (Aguero-Valverde, 2006), maps identifying sections of traffic casualties side-by-side with average traffic speeds (Quddus, 2008), and accident specific choropleths of pedestrian and bicycle crashes (Siddiquia, 2012).

Supportive Graphs & Plots – This section contains examples and concepts most applicable to our work. Visualizations include stacked bar charts of monthly accidents over 10 years, accidents as bar plot by day of week, and a pie chart for collisions over hourly range of day (Erdogan, 2008); currently we plan on combining day of week and hour of day into what we believe will be a more helpful vis tool, described in sketch section below.

Many of the remaining examples are variations of the same theme: line graph of average hourly traffic volume by time of day, weekdays and weekend superposed (Wang, 2009); side-by-side bar chart of road deaths per 100,000 population/10,000 vehicles in 2010 for various countries (Driss, 2015); crash risks by day of week, time of day, and road type/day of week (Li, 2007); and line graphs of hourly and monthly incidents, as well as a spider web chart plotting accidents by day of week, which was more visually appealing than readable for comparison.

Prediction/Modeling – From the granular aspect of investigating specific road sections we come to the global scope of predicting motor vehicle collisions. These projects used visualizations to highlight their models'/algorithms' findings and projections. Either a map was shaded to reflect model predictions against actual accidents (Levine, 1995) or stretches of roads were color-coded based on danger exposure (Driss, 2015). Though not tasked with providing a predictive element in our project, the concept of highlighting specific stretches of road or city cross sections as specifically dangerous is immediately applicable.

Data

The main dataset used for this project is NYPD Motor Vehicle Collisions available on the NYC Open Data platform[5] and can be downloaded directly from the NYC Open Data website in either CSV or JSON, among other types. The dataset comes from processed TrafficStat reports made by NYPD and was used inter alia as a basis for developing the Vision Zero View web application. It represents each collision reported to NYPD from 7/1/2012 onward. The dataset is updated daily with an one or two day delay, making it a dynamic dataset with more than 543,000 observations, as of 9th of March 2015. The data type is a flat table consisting of 29 attributes describing each collision, which can be grouped into 6 categories. Both categories and described attributes are listed in a table in the appendix (Table 1).

For the purpose of our project, the dataset will be treated as a static table limited to collisions from 07/01/2012 to 02/28/2015. The size of the generated file (CSV) is ~100MB. The dimensions of the filtered table is 541,764 rows by 29 columns. Because the project will be presenting spatial distribution, location information is crucial. Therefore we decided to focus only on the collisions that provided latitude and longitude. There are 455,575 such items which account for 84% of the total reported collisions over the selected time period.

Though location will be our primary spatial attribute, we anticipate using nearly all attributes in order to provide in-depth descriptive information and/or for aggregation purposes. Temporal features (DATE, TIME) will be used to show collisions in different time frames selected by users, while victim, contributing factor, and vehicle information will create additional layers to choose from.

In addition to the NYPD Motor Vehicle Collisions Data we will be using geometry type datasets like New York City's shapefiles including borough boundaries, zip-codes, police precincts, street network etc. available through BYTES of the BIG APPLE database.[6] They will be used mostly to create a base map for the visualization.

---

[5] https://data.cityofnewyork.us/Public-Safety/NYPD-Motor-Vehicle-Collisions/h9gi-nx95
[6] http://www.nyc.gov/html/dcp/html/bytes/applbyte.shtml

Further Data Transformation

For the purpose of the visualization with d3, police precincts shapefile (.shp) obtained from BYTES of the BIG APPLE database had to be converted into GeoJSON file. It was achieved by using ogr2ogr utility which is part of Geospatial Data Abstraction Library (GDAL)[7] package. The transformation created 4 MB JSON file that maps all 77 precincts of New York City that will serve as a framework for the general view map.

Despite the initial process of cleaning data, the data set requires additional steps in order to make it useful for the purpose of creating visualization. First of all each collision was mapped to the corresponding police precinct using ESRI ArcMap software. It resulted with additional attribute called "Precinct".

The next step was to extract date and time information for each collision. It was achieved using Python's Pandas package. Attributes "DATE" and "TIME" were converted from plain text into *datetime* objects what allowed to easily create new attributes for each item called "year", "month", "day", "weekday" and "hour".

Because of the fact that in original data set the information about both contributing factor and type of the vehicle involved in an accident was mapped to each vehicle. For the analytical purposes these attributes were deconstructed into new column for each category and calculated based on description for each vehicle involved in an accident. This operation created 46 new attributes for each contributing factor and 17 for each vehicle type. Additional column for number of vehicles involved in accident was added as well. This way prepared database was exported into the file of approximately 160 MB.

Completed database was grouped by precinct and exported to 77 JSON files, one for each precinct. This way, d3 script will require only small part of data for detailed view.

For the purpose of main choropleth, the data set required substantial aggregation. It was done by grouping all accidents by precinct and aggregating by given attribute like year, day of the week, hour, a sum of contributing factors and vehicle types as well as average number of vehicles involved. This way prepared data was added to properties of each police precinct in previously created GeoJSON file.

---

[7] http://www.gdal.org/

Visualization Questions

The following is a list of preliminary questions we hope to answer with the support of our visualization:

1. Are there areas of the city seeing new collision hotspots?
2. Are there local hot spots that correspond to increases in enforcement-related causes (such as cell phone use, illegal drug use, or alcohol involvement)?
3. Which areas should the NYPD prioritize its resources toward?
4. Which enforcement-related causes should the NYPD pay particular attention toward?

Visualization Sketches

Figure 1 (below) shows the first iteration sketch of our proposed visual solution: an interactive heat-map of the number of collisions in each locality in New York City (spatial options include police precincts, zip codes and census tracts). Temporal options include selecting figures for a particular day of the week, week of the month, month of the year, a particular year in the dataset, as well as by day and night. The timeline scroll bar will assist in any temporal analysis (by day, month or year). Once a particular region is clicked (or the mouse is hovered over it), that region's ID (borough name and zip code/police precinct/census tract) will be displayed as well as a bar chart of the major causes of collisions. At this point, the greatest anticipated challenge is allowing the visualization to surface hotspots without undue burden on user's time and effort. This meant certain revisions to the proposed visual solution had to be made.

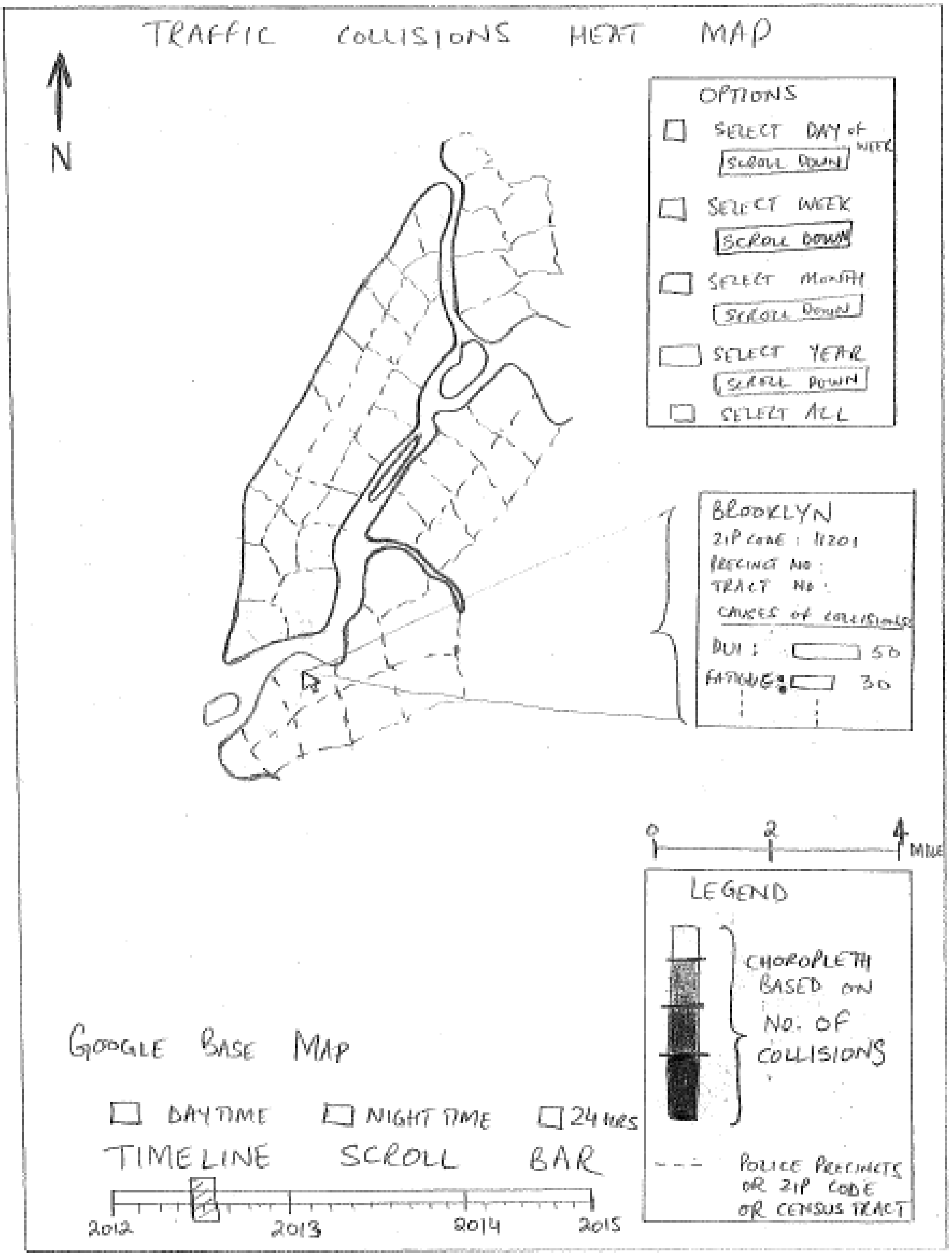

Figure 1: Visualization Sketch 1 (first iteration)

Figures 2 and 3 (below) show a revised version of the proposed visual solution. Based on user feedback, we have decided to make certain changes to our visualization in order to enhance cognition and make our solution more useful. Now the initial webpage will show a heat map of New York City Police Precincts based on collision density (number of traffic collisions per area).

The sections below the map will display information regarding hourly and monthly collision totals as well as the top 5-10 causes for the collisions sorted in descending order (based on number of collisions) for the city. Clicking on a certain month's bar chart will dynamically change the collisions totals for the causes and hourly sections to represent only collisions for that month. There will also be a scroll down menu which will allow the user to only see collisions for a certain year.

Hovering the mouse over a precinct will display the borough name, precinct number and the collision density for that precinct. Clicking on that precinct will make the map zoom in to that precinct as well as dynamically changing certain features of visualization (see Figure 3). In the zoomed-to-precinct level, the choropleth of precinct collision density will be replaced by actual collisions as points on a street view map (an opacity choropleth representing point density). The timeline and causal information displayed in the sections below will also change to show collision totals for that precinct only. The hourly collisions plot will feature the city average for comparison. The year selection option from the scroll-down menu will still be available, as will the option to click on a specific month in the bar plot and have the causal and hourly information update to only show collisions for the selected month. Multiple month and year selection is a desirable feature that we hope to have in our final visualization.

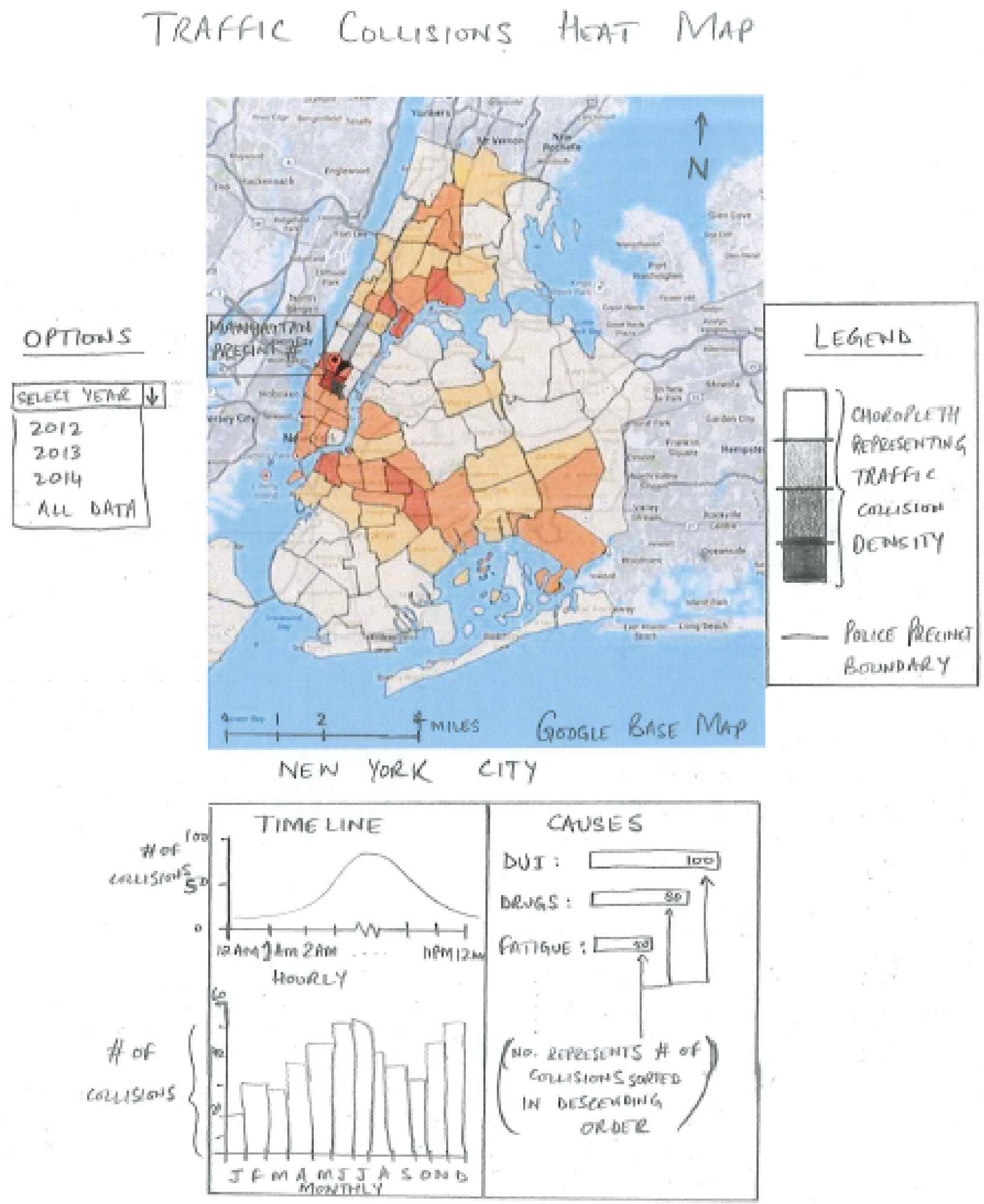

Figure 2: Visualization Sketch 2 (second iteration)

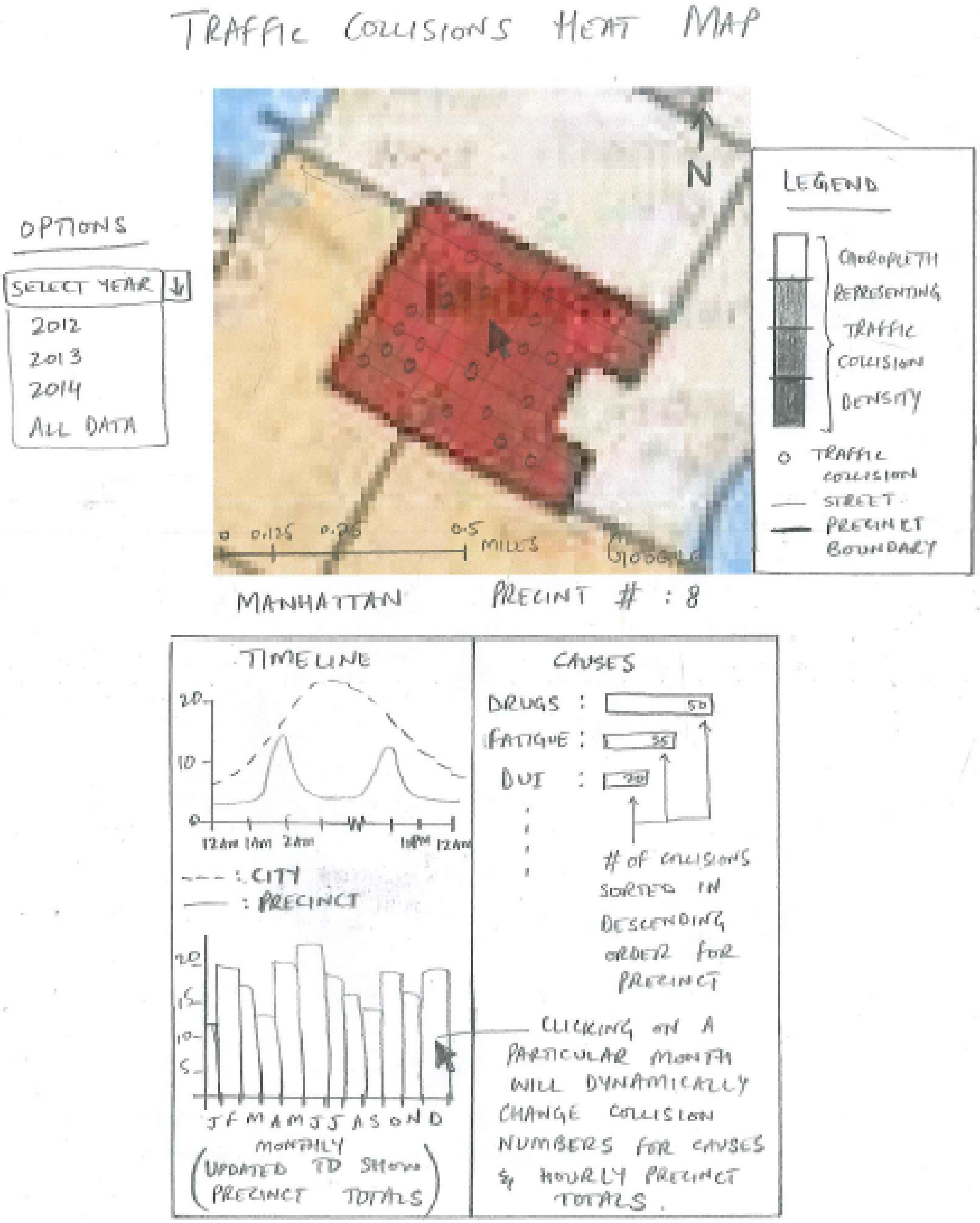

Figure 3: Visualization Sketch 2 (zoomed in to precinct)

Visualization Prototype 1

D3 Heat Map

Figure 4 (below) shows the first development of our proposed solution. Currently, the D3 visualization shows the police precinct heat map of total number of accidents with brighter red precincts depicting greater number of accidents. The mouse-over feature has a box that shows the Precinct Number and Accident totals for each year from 2012 to 2015 of the precinct on which the mouse is hovering. This is a very early rendition, and we still have to add several features to reach the envisioned visual solution shown in our sketches (see Table 1 below).

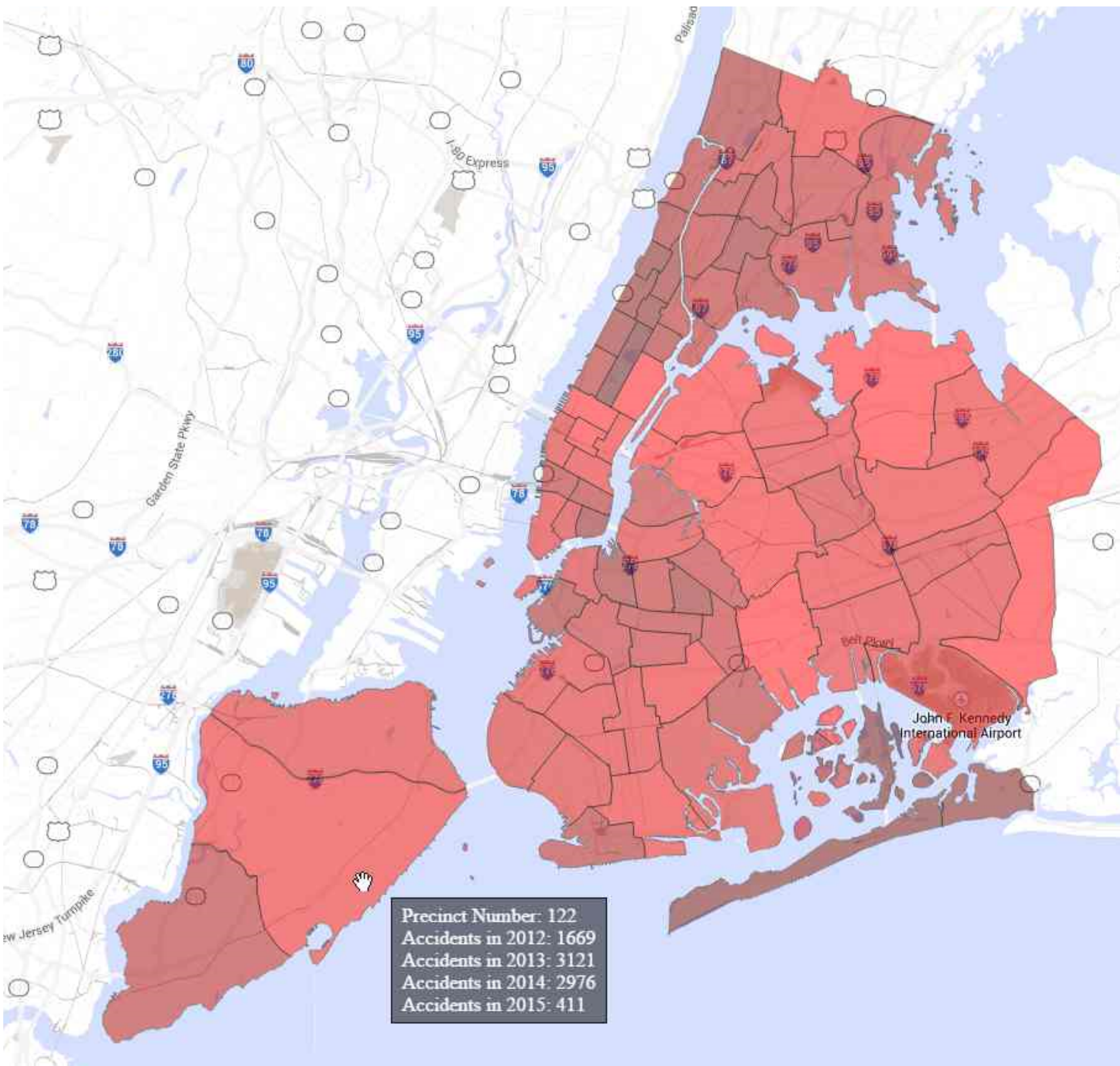

Figure 4: D3 Visualization Prototype 1.0 (Heatmap of 2014 Accidents)

Table 1: List of Attributes Available for D3 Visualization

Attributes	**Visual Channels**
Police Precincts	Shapes (polygons on map) Position (spatial location) Text (precinct number in mouseover box)
Total Accidents	Color (heat map)
Accidents by Year	Text (accident totals in mouseover box)
Accident Location	None (position & color to be included in zoomed-to-precinct feature)
Collision Causes	None (bar plot to be included)
Time (hourly, daily, weekday/weekend, monthly, yearly collision figures)	None (time series plots to be included)

D3 Time Series Plot

One of the features we have and intend to integrate into our map is a 24-hour time series plot that displays the average collision per hour. In Figure 5 below we are displaying the numbers for 2013 and 2014 in pink and green, respectively. Currently these numbers are aggregated across the entire city with the thought being to implement this on a precinct level. The plot would be interactive with respect to time and space, allowing the user to specify both the year and precinct of choice.

This user-generated focus will allow for the opportunity to identify how one precinct is doing relative to another, as well as to the city as a whole. For example, the data currently shows a city-wide pattern of collisions rising in the morning, remaining at a high level throughout the afternoon, and trailing off in the evening. Whether or not this profile holds for all precincts is a question we hope to allow the use to answer, among others.

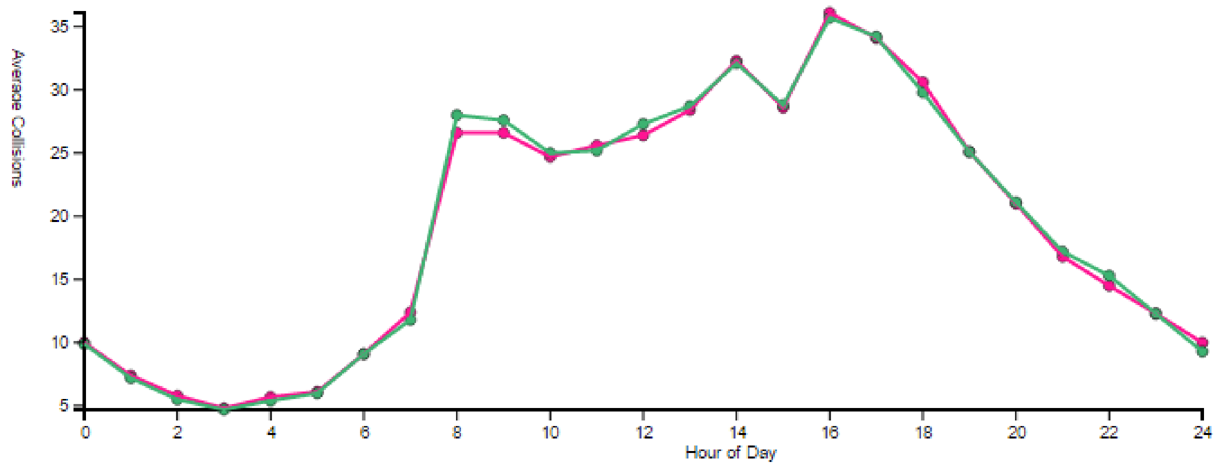

Figure 5: Twenty-four hour time series of collisions.

The above plot is one of several we are currently looking into implementing, all at the precinct level:

- collisions by month
- collisions by weekday versus weekend
- collision factors

Work Plan

The first priority identified by our team is to better understand the data and prepare it for use moving forward. As referenced above we will be focusing exclusively on collisions that provide latitude/longitude information. The task of isolating these records, as well as identifying and dealing with additional anomalies, will fall under data cleaning. Two other exercises will be going on concurrently as the data is being readied: spatial joining and D3 structural preparation. In anticipation of the finalized collision table, team members will work with representative data subsets to work through the process of assigning each collision to a police precinct via lat/long coordinates, which will be critical in aggregating collision data at the client's desired spatial level, and testing the importing of data into a first pass skeletal map using D3.

Lastly, before moving forward and testing different visual idioms with the client, each team member will be responsible for data exploration in a tool of their choice, as well as

Tableau. This exercise is meant to mirror the workflow of the sketches, where each team member worked independently and later came together to share ideas and insights.

All of the work tasks described are referenced in the table below, with the responsible team member identified. We stress that this is the preliminary work plan needed to better understand our data while we continue communicating with our client. It is believed that once the work above has been accomplished we will be in a stronger position to do the necessary data and task abstractions as well as any additional data transformations.

Table 2: Work Plan

Task	Awais	Bartosz	Radu	STATUS
Data Cleaning		x		Complete
Data/Spatial Joining			x	Proof of Concept
D3 (heat map)	x			Proof of Concept
Data Exploration (Tableau/R)	x	x	x	Ongoing
Police Precinct JSON File		x		Ongoing
GeoCode JSON w/Collision Data	x	x	x	Ongoing
D3 (collision plots)	x	x	x	To be broken out into discrete tasks.

Addressing Concerns in Project Update 1.0

We met with Professor Bertini to discuss his concerns regarding aggregating the data in the overview stage. The problem is our dataset includes approximately 450,000 data points on traffic collisions and showing all of them at once in D3 is not technically feasible. Therefore, we have decided to represent the detailed information individually for each precinct (once the precinct is clicked). This is also preferred for our end user, since Police officers are mainly interested in collision trends pertaining to their own precinct.

Table 3: Deliverable Timeline

Deliverable	**Proposed Date**	**Comments**
JSON: Precinct level collisions	Data: 4/20 JSON: 4/24	Aggregated by year, month, weekday v. weekend, & cause.
Graph: Collisions by 24-hour	Data: complete JSON: 4/24 Graph: 4/27	Precinct level.
Graph: Collisions by Month	Data: 4/20 JSON: 4/24 Graph: 4/27	City and precinct level.
Graph: Collisions by Weekend	Data: 4/20 JSON: 4/24 Graph: 4/27	City and precinct level.
Graph: Collisions by Cause	Data: 4/20 JSON: 4/24 Graph: 4/27	City [data complete] and precinct level.
Collision Plotting	Data: complete	Precinct level. Investigate options of color, opacity, and kernels.
Integrate map with graphs	4/30	

Appendix 1

Table 3: Attributes and their description

Data category	Name	Format	Type	Description
Temporal identification	DATE	date format (MM/DD/YYYY)	sequential	day when collision happened
	TIME	plain text in 24 hours format (HH:MM)	cyclic	time at what collision happened or was reported
Geographical reference	BOROUGH	plain text	categorical	name of the NYC borough for collision location
	ZIP CODE	plain text	categorical	zip-code for collision location
	LATITUDE	number	diverging	geographical latitude in decimal format of collision location
	LONGITUDE	number	diverging	geographical longitude in decimal format of collision location
	LOCATION	geographical location (degrees)	diverging	georeferenced attribute locating accident
	ON STREET NAME	plain text	categorical	name of the street where collision happened
	CROSS STREET NAME	plain text	categorical	name of the nearest cross street where collision happened
	OFF STREET NAME	plain text	categorical	address where collision happened

Victims	NUMBER OF PERSONS INJURED	number	quantitative	self-explanatory
	NUMBER OF PERSONS KILLED	number	quantitative	self-explanatory
	NUMBER OF PEDESTRIANS INJURED	number	quantitative	self-explanatory
	NUMBER OF PEDESTRIANS KILLED	number	quantitative	self-explanatory
	NUMBER OF CYCLIST INJURED	number	quantitative	self-explanatory
	NUMBER OF CYCLIST KILLED	number	quantitative	self-explanatory
	NUMBER OF MOTORIST INJURED	number	quantitative	self-explanatory
	NUMBER OF MOTORIST KILLED	number	quantitative	self-explanatory
Contributing factors	CONTRIBUTING FACTOR VEHICLE (1 to 5)	plain text	categorical	factor most contributing to collision of vehicle number 1 to 5 involved in an accident grouped into 46 different categories (appendix 1)
Identification	UNIQUE KEY	number	ordinal	unique key number given to each collision
Vehicle type	VEHICLE TYPE CODE (1 to 5)	plain text	categorical	vehicle type involved in a collision grouped into 17

				different categories (appendix 2)

source: https://data.cityofnewyork.us/Public-Safety/NYPD-Motor-Vehicle-Collisions/h9gi-nx95

Appendix 2

Table 4: List of categories for contributing factor

1.	Lost Consciousness
2.	Fatigued/Drowsy
3.	Pavement Slippery
4.	Driver Inattention/Distraction
5.	Unspecified
6.	Brakes Defective
7.	Glare
8.	Backing Unsafely
9.	Traffic Control Disregarded
10.	Physical Disability
11.	Failure to Keep Right
12.	Prescription Medication
13.	Failure to Yield Right-of-Way
14.	Other Vehicular
15.	Driver Inexperience
16.	Turning Improperly
17.	Outside Car Distraction
18.	Oversized Vehicle
19.	Illness
20.	View Obstructed/Limited

21.	Alcohol Involvement
22.	Passenger Distraction
23.	Reaction to Other Uninvolved Vehicle
24.	Steering Failure
25.	Other Electronic Device
26.	Unsafe Speed
27.	Fell Asleep
28.	Aggressive Driving/Road Rage
29.	Obstruction/Debris
30.	Unsafe Lane Changing
31.	Traffic Control Device Improper/Non-Working
32.	Windshield Inadequate
33.	Shoulders Defective/Improper
34.	Accelerator Defective
35.	Pavement Defective
36.	Passing or Lane Usage Improper
37.	Lane Marking Improper/Inadequate
38.	Cell Phone (hand-held)
39.	Tire Failure/Inadequate
40.	Drugs (Illegal)
41.	Other Lighting Defects
42.	Following Too Closely

43.	Animals Action
44.	Pedestrian/Bicyclist/Other Pedestrian Error/Confusion
45.	Headlights Defective
46.	Cell Phone (hands-free)

Appendix 3

Table 5: List of categories for involved vehicles

1.	BUS
2.	PASSENGER VEHICLE
3.	LARGE COM VEH(6 OR MORE TIRES)
4.	SPORT UTILITY / STATION WAGON
5.	TAXI
6.	UNKNOWN
7.	VAN
8.	OTHER
9.	SMALL COM VEH(4 TIRES)
10.	PICK-UP TRUCK
11.	FIRE TRUCK
12.	LIVERY VEHICLE
13.	AMBULANCE
14.	MOTORCYCLE
15.	BICYCLE
16.	SCOOTER
17.	PEDICAB

Bibliography

Aguero-Valverde, Jonathan and Paul P. Jovanis. “Spatial analysis of fatal and injury crashes in Pennsylvania.” *Accident Analysis and Prevention* 38 (2006): 618–625.

Driss, M. and K. Benabdeli, T. Saint-Gerand & M.A. Hamadouche. “Traffic safety prediction model for identifying spatial degrees of exposure to the risk of road accidents based on fuzzy logic approach.”
Geocarto International 30.3 (2015): 243-257.

Erdogan, Saffet and Ibrahim Yilmaz, Tamer Baybura, Mevlut Gullu. “Geographical information systems aided traffic accident analysis system case study: city of Afyonkarahisar.” *Accident Analysis and Prevention* 40 (2008): 174–181.

Khattak, Aemal J. and Huda Shamayleh. “Highway Safety Assessment through Geographic Information System-Based Data Visualization.” *JOURNAL OF COMPUTING IN CIVIL ENGINEERING* (Oct 2005)

LEVINE, NED and KARL E. KIM and LAWRENCE H. NITZ. “SPATIAL ANALYSIS OF HONOLULU MOTOR VEHICLE CRASHES: I. SPATIAL PATTERNS.” *Accident Analysis & Prevention* 27.5 (Feb 1995)

Li, Linhua and Li Zhu, Daniel Z. Sui. “A GIS-based Bayesian approach for analyzing spatial–temporal patterns of intra-city motor vehicle crashes.” *Journal of Transport Geography* 15 (2007): 274–285.

Prasannakumar, V. and H. Vijith, R. Charutha, N. Geetha. “Spatio-Temporal Clustering of Road Accidents: GIS Based Analysis and Assessment.” *Procedia Social and Behavioral Sciences* 21 (2011): 317–325.

Quddus, Mohammed A. “Modelling area-wide count outcomes with spatial correlation and heterogeneity: An analysis of London crash data.” *Accident Analysis and Prevention* 40 (2008): 1486–1497.

Siddiquia, Chowdhury and Mohamed Abdel-Atya, Keechoo Choi. “Macroscopic spatial analysis of pedestrian and bicycle crashes.” *Accident Analysis and Prevention* 45 (2012): 382– 391.

Soltani, Ali and Sajad Askari. “ANALYSIS OF INTRA-URBAN TRAFFIC ACCIDENTS USING SPATIOTEMPORAL VISUALIZATION TECHNIQUES.” *Transport and Telecommunication* 15.3 (2014): 227–232.

Wang, Chao and Mohammed A. Quddus, Stephen G. Ison. “Impact of traffic congestion on road accidents: A spatial analysis of the M25 motorway in England.” *Accident Analysis and Prevention* 41 (2009): 798–808.

Yi, Ping and Yingcai Xiao, Anthony Ciccolini, Greg Frommer, Dongning Zhang. “RULE-BASED MODEL FOR TRAFFIC ACCIDENT VISUALIZATION AND ANALYSIS.” *JOURNAL OF COMPUTING IN CIVIL ENGINEERING,* (April 2001)