Data Analysis of Crimes Reported in Philadelphia in 2012

Evans, David; Furman, Maxwell; Kellam, Jeffery; Smith, Joshua

Drexel University

Abstract – This project involved searching for and manipulating data through information visualizations to reveal what is not comprehensible when reading a large data set. The data focuses on Philadelphia Crime Statistics. The statistics were then analyzed for different possible trends and the visualizations of these trends are analyzed. The analysis and use of these crime statistics were also compared to current and proposed uses of similar data as part of predictive policing.

1 Introduction

Our group decided to focus on crime rates as reported by "Open Data Philly" for the city of Philadelphia. We sought to understand where in the city crime was most likely to be reported, the most common type of crime (violent vs. non-violent[1]), and if the time of day played a factor. The data was formatted and grouped using Microsoft Excel and Access. The results were then analyzed on the above 3 factors. This information could then be used as a model for predicting crime behavior. Proactive policing techniques such as increased patrols and community policing could then be focused on those areas at the highest risk of crime [1]. In addition, these efforts could then be focused at the time of day where they would be the most effective.

With the above information, we were able to show an example of how this data might be used by comparing it against two fast food chains in areas with the highest reported crime rates to determine if there appeared to be a correlation between reported crimes and one of the two food chains. This is just one example of using this data to help determine if outside factors, such as a restaurant are contributing to crime rates.

2 Approach

Our team's desire is to create multiple visualizations that could be plotted on a map, showing relevant information, or provide information that could be useful within geographically location related to our dataset. The dataset that we choose, crime reports in Philadelphia, was a perfect fit for our overall desire. Based on a recommendation from our professor we gave Google Fusion tables the benefit of a doubt and imported our data into the tool. We also wanted to create other none geographical related visualizations to provide information on types of crime. We also want to use visualizations to compare crime activity throughout the day to identify key hours. Our team wants to use this information to identify bad areas and dangerous times of day. Identify trends and provide prediction analysis that could be useful to law enforce agencies and administrations.

3 Solution

Google Fusion Tables offers a free cloud solution for structured data with multiple visualization options [2]. This online application is available to any user who has a google or gmail account. The main key feature that attracted us to Fusion Tables is that the tool could plot latitude and longitude data points, which were included in our dataset. Additional we required a system that supported online collaboration due to the fact our team was geographically separated. Again the professor also recommended that we check out this tool based on other student projects and his experience.

Along with the visualization option to map data on Google Maps, Tables allows users to create scatter plots, line charts, intensity maps, and network graphs. Given the attribute fields in the data we were able to use several of these visualizations. Other important options that Tables provided us are its filtering capabilities and data aggregation. This has proven very useful to consolidate or eliminate data out of the 89,000 entries of information.

There were some drawbacks related to using Fusion Tables that are worth mentioning for future researchers. Tables lack a needed visualization customization feature, such as applying your own color code to nodes on some of the visualizations. This was noticeable with the scatter chart diagrams when there was no possible way, in regards to color, to distinguish between crime categories that had been aggregated.

Our initial attempt to import data into Google Fusion Tables ended unsuccessfully. Our original raw dataset was 82 MBs and contained over 600,000 entries. It became apparent after many attempts to create visualizations and to change column properties the amount of information was overwhelming the system. The tool was slow to respond and was unable to properly geocode the latitude and longitude columns. Visualizations were also unable to handle the sheer amount of

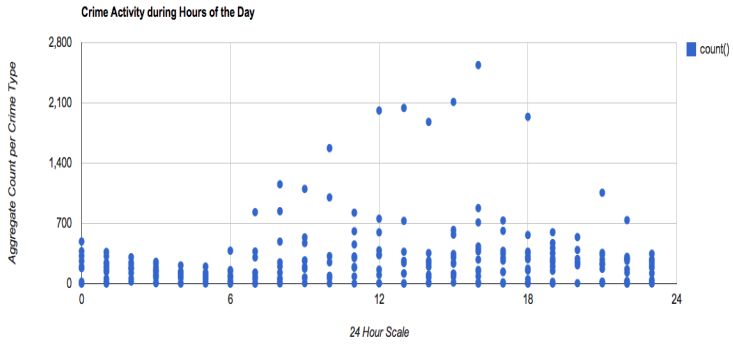
[1] FBI Standard were used;
http://www2.fbi.gov/ucr/cius2009/about/offense_definitions.html

node entries in their visualizations; the maximum amount noticed was a sample of 20,000 entries.
To correct these issues the data was modified in Microsoft Excel and then imported into Google Fusion Tables. Columns that were not pertinent to our information discovery were removed from the dataset such as internal identification numbers. Originally the latitude and longitude information was separate into two different columns, in order for Google Fusion Tables to properly process this information they had to be combined into one column. Lastly, the dataset rows had to be reduced from 600,000 to a more manageable size. The decision was made to only review information from Jan 1, 2012 to Feb 28, 2013. This reduced the row to 89,000 entries. After the information was cleaned up and reduced the table was uploaded into Fusion tables. We believe that the 13 months of sample data is sufficient to identify trends, potential hostile areas, safe areas, provide foundation for predictive analysis, and identify areas of interest that may need more research.

4 Data

We used crime data from the City of Philadelphia and mapping data from Google Maps [3] [4]. The crime data consisted of the date and time of a police report, the type of crime reported, and the location, and covers crimes committed in 2012. We believe this data is credible because it comes directly from the City of Philadelphia; it is the Philadelphia Police Department's own internal data and as such is the most reliable source for crime data within Philadelphia. This data was made available through the Open Data Philly project, an arm of the non-profit Philadelphia Public Interest Information Network that seeks to encourage active citizenship and government transparency through open access to information.

5 Visualizations and Major Conclusions



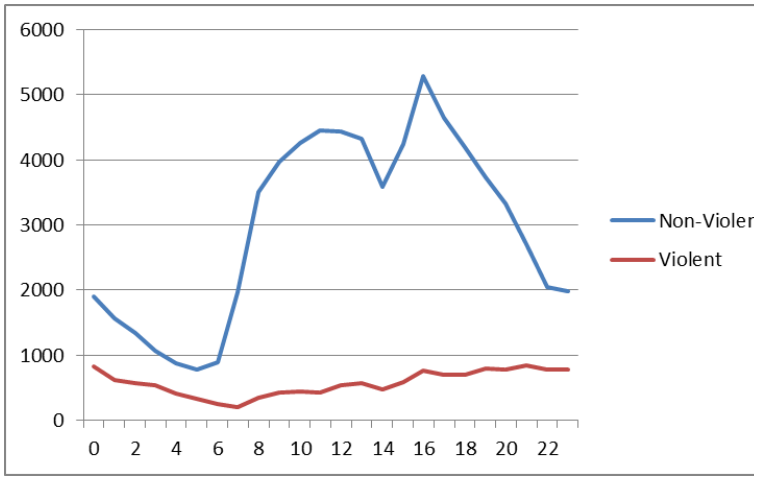
Visualization 1 – Produced and analyzed by Kellman

5.1 Description of the Visualization 1

The visualization focuses on crime reports at different times of the day for all areas and all types of crime. The X axis consists of the hours of the day on the 24 hour scale. To provide some clarity about the data, all crime reports land on the exact hour. One of the columns in the data, Hours, is the truncated hour of day that the report was called in. An example, any report called in between 12am to 12:59am are all logged under 0000 in the visualization, as midnight is 0000 on the 24 hour clock. Although the hours scale goes to 24 hours there will never be any entry on the 24th hour due to the way they categorized their information.

Visualization 1 Analysis

After reviewing the information there are apparent trends that should be noticeable to the reviewer. The first obvious trend is that most crime is reported in the afternoon between the hours of 1300 to 1800. This actually disproves our theory that most crime occurs during the night hours. We believe that this is a common theory that most people possess. This could be due to news media coverage or a reflection of how the entertainment industry portrays crimes. Looking at the scatter plot people can see that crime reports taper off to even levels through the night dropping to the lowest levels between the hours of 0300 and 0600. Other visualizations we provide will show more information about the amount of violent crime and non-crime but one can notice a significant difference in total reports during the peak hours of activity. The leading figures deal with thefts. We know this not because of the visualization but because the theft reports outnumber the other crimes by several fold and are the only category that can produce that amount of activity. What this information shows is that crime activity corresponds with human activity. As less people are active during the night less crime occurs and vice versa.
There are some interesting things to consider when reviewing this scatter plot. Just as this data shows that crime activity corresponds to human activity the reports are more than likely distorted due to human activity. The hours during the day when there is the least amount of activity, 0400 through 0600, are also the hours of the day that most people are going to be asleep. To take the thought further, this reduction in reported activity could actually be a result of the fact that people are no longer awake to report the incident. The greatest amount of crime activity is when most people are going to be awake and out and about. So there is a possibility that a portion of the crimes actually occur earlier during the day but they are only reported when people notice or have the time to report. Any other reporting discrepancies should be negated since police would not be called out to the location for most of the older crimes.

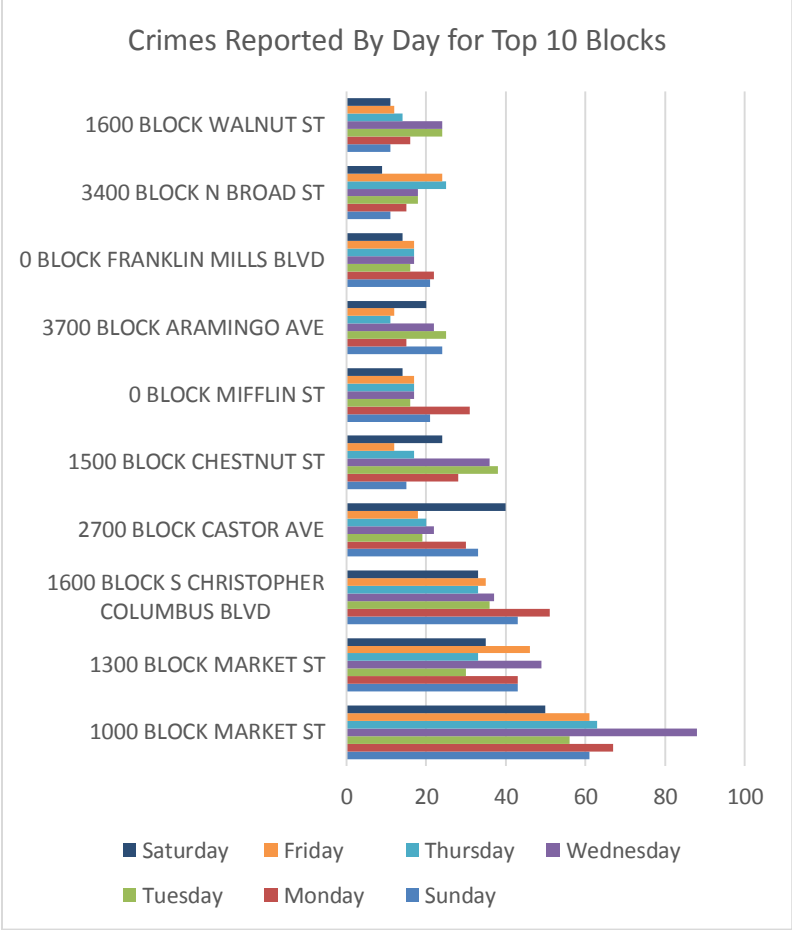


Visualization 2 – Produced and analyzed by Smith

5.1 Visualization

For this analysis of the crime data, we first separated out the violent and non-violent crimes as well as those listed as motor vehicle recoveries. We discarded the motor vehicle related incidents because we believed they were not important to the overall analysis to look at how crime develops. Having separated the non-violent and violent crimes out from the master data set we then organized both sets by the hour of the day during which they were reported starting with 0 for between midnight and 1 am and progressing throughout the day until reaching 11 pm at night. We had some presuppositions about the crime data before we mapped it and felt they were important compared to the results that the data showed. We knew that from glancing at the list that non-violent crime would be much greater than violent crime, but we also thought that both would follow similar trends in their rates by hour throughout the day. We also assumed that that crime would be at its highest during the late night (from 10pm till 2am) for both types of crime and that there would be a steady progression rising to a peak and then tapering off.

As we totaled the numbers by the hour we knew that our original estimates were not going to be correct, but the complete falseness of our assumptions were not completely clear until we charted the data. First, violent crime and non-violent crime fell and rose completely unrelated to each other. Violent crime matched up with my assumptions about a rise to a peak and then a drop off to a low, but we were also incorrect about the number of crimes committed in the late night time period (10pm to 2am) we expected versus earlier on in the evening for non-violent crime. Violent crime peaks earlier at midnight and then tapers off to a low at 7am whereas non-violent crime's lowest point is 5 am. It's possible to infer that non-violent crime and its reporting is driven by traditional work schedules as the crime rate rises at 6am with a large number of people waking up, commuting to work, and starting there day at the office. Non-violent crime then surges up in only an hour or two, reaching a plateau around 11 am. It then dips sharply at 2pm and quickly rises to its highest point at 4pm in the afternoon. From there it steadily trails downward till 10pm. From there it plateaus shortly till midnight then nose dives to the low point at 5 am. As a fellow researcher pointed out, overall crime reporting and responses would ultimately be higher for bother crime rates when more people are awake. And the shared low points in the morning coupled with the decline throughout late night into early morning would support this. Beyond that non-violent crime's tendency to move with the work day would make sense as all sorts of property crime would either depend on stores being open, people being away from home for work, or large groups of people to prey upon.



Visualization 3- Produced and analyzed by Evans

5.3 Visualization 3 Analysis

Microsoft Access was used to convert the dispatch date into the corresponding day of the week. For example, January 1, 2012 became "Sunday." When we compared all crimes reported to days of the week, Monday was the highest, followed by Friday. Sunday was last with the least number of crimes reported on that day. It was believed that Monday lead with the most number of crimes reported because non-violent crimes were skewing the results. (i.e. Business owners returning to work on a Monday to find a crime that happened

over the weekend) Surprisingly however, a second query removing all non-violent crimes kept the results in the exact order; Monday having the most crimes reported in 2012.

Our group had initially thought that crime would shift from weekdays when most businesses are open and where there are more people in the business district of Philadelphia. Our thoughts were that the crime for the weekends (Friday included) would be centered around areas with lots of bars and clubs (for example, Northern Liberties or Old City).

Once we had days of the week assigned to the crimes reported, we again used Access to group on day of the week and the street block the crime was reported on. The data showed that the most reported crimes on each day of the week corresponded with the data results from our overall crime data by location. The 1000 block on Market Street was previously shown to have the most crimes reported overall in the year 2012[2]. When comparing to days of the week, the 1000 block on Market Street held 7 of top 8 overall crime reported areas. The 7th overall spot (which, although a Monday, was the second Monday, and does not count for the "worse place to be on a Monday") is held by a city block that was the third on the overall most reported crimes list; the 1600 block of South Columbus Blvd[3].

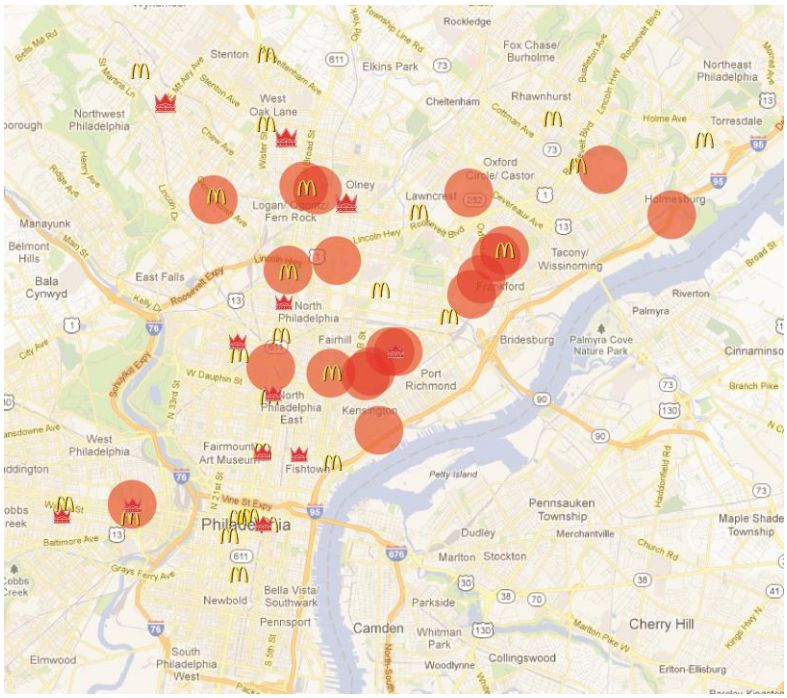
Having seen that the day of the week did not shift the highest crime areas, we decided to compare the number of crimes reported on each day of the week with the top 10 areas with the most reported crimes over all.

First we determined the 10 blocks with the most reported crimes over all in the 2012. This was done in Microsoft Access by grouping on the reported block location and using the "count" function to get the number of reports by block. What *Visualization 3* showed us what was to be expected; that there was not a single location skewing the data for crimes reported on Mondays. In all but two cases in each of the top 10 locations, Monday was in the top 3 days of crimes reported. Therefore having shown that the day with the highest reported amount of crimes lines up with the top 10 locations with the highest reported amount of crimes over all, we can generalize that the same would hold true as we venture outside of the top 10 locations.

This information could be invaluable to Police Officers on patrols not only in these areas, but during the "worst" days as well. Outside factors would need to be compared, such as sports games, festivals, and other large events to see if they are driving crime towards a certain day. This information can help police departments better their understanding of how and when to place patrols, plan events, and work with local businesses.

[2] See Appendix Table 1

[3] See Appendix Table 2



Visualization 4- Produced and analyzed by Furman

5.4 Visualization 4 Analysis

For our last visualization, we were inspired by the question, which is a more dangerous place to eat, McDonald's or Crown Fried Chicken? We selected these two chains because they were fairly common in the city of Philadelphia, and have locations in many different neighborhoods. We suspected they would make a good contrast as McDonald's is a huge international chain while Crown is found mainly in the largest cities of the Northeastern United States. To answer our question, we first used the crime data from Open Data Philly to determine the twenty blocks where the most violent crimes occurred. We visualized them on *Visualization 4* as red circles. Areas where the circles overlap have higher concentrations of violent crime. Then, using location data from Google Maps, we plotted the locations of the McDonald's and Crown Fried Chickens onto the same map [4]. We were looking to see which restaurant locations would be found in the most reported violent neighborhoods and from there which restaurant chain had the highest percentage of locations within those neighborhoods.

We went in without a strong feeling either way as to which one would be more dangerous, but were surprised to see McDonald's so clearly out in front. While only two Crown Fried locations are within red circles, and two boarding the edge of a circle, a whopping six McDonald's locations are in dangerous areas and four boarding an edge. One could make the argument that the reason that more McDonald's locations are in dangerous neighborhoods is that there are more McDonald's locations in general, however, out of 27 McDonald's on the map 6 are dangerous for an alarming rate of 22%. Meanwhile, two out of 12 Crowns on the map are in

high-crime areas for a rate of only 16%. That means that if we picked a McDonald's at random we would be more likely to be the victim of a crime than if we picked a random Crown Fried Chicken. Based on this data, we can only conclude that in Philadelphia it is much safer to eat at Crown Fried Chicken than at McDonald's.

6 Current Crime Mapping Projects

Overview

While knowing where crime is highest and what times most crimes occur could be helpful for buying a house, deciding when and where to go out for dinner and planning staffing levels for a police precinct, there are better, more proactive applications available. One such application is predictive policing. Predictive policing is a, "technique of integrating data analysis with professional law enforcement expertise to understand why a problem arises and how to avoid the next problem is called predictive policing. It builds on and melds pieces of community policing, intelligence-led policing and hot spots policing" [5]. Proactively policing an area known to have a high crime rate is not a new concept, but predictive policing takes this core concept of solid police work and applies to it the latest techniques and approaches in data analysis and data mining to process years of police data and trends and output patrol maps and highlight areas where crimes are most likely occur to within a few blocks. Predictive policing is already being tested and applied in US cities such as Los Angeles, Memphis, New York and Santa Cruz with results already being seen as Memphis has seen an overall decline in crime by 15% between 2008 and 2012 [6].

A key component of predictive policing

The statistics used to create the visualizations that have been featured are exactly the kind of statistics that were and are being used by predictive policing initiatives. The chart for violent versus non-violent crime could easily be applied to scheduling and patrol planning to decided what kind of officers should be staffed at different times of day. Regular uniform officers on foot and on bike could be placed in areas of high crime during the daytime hours to deal with the much larger incidence of non-violent crime whereas patrol cars with pairs of officers and a more duty load (shotguns and rifles in the car, wearing armor) could be dispatched during the peak violent crime hours in the late evening and nighttime hours. However, "Predictive-policing methods make use of far more variables than the times and locations of recent crimes" [7], meaning that simply using the crime stats analyzed above would not be nearly enough for an actual predictive policing program in real life. All kinds of data from housing values to school truancy rates to public transportation routes can factor into both how crimes occur in the real world and how an effective predictive policing program is.

7 Conclusion

Therefore through the use of Google fusion tables, information visualizations, and traditional analysis methods, the crime data was analyzed and trends not inherently apparent in the data set was brought to the fore through the use of information visualizations. These visualizations took presuppositions and tested them against what the data truly showed. From anecdotal beliefs about the most dangerous time of day to where the most crimes actually occurred within the city the data was able to reveal intriguing correlations. Beyond the data and the visualizations was also the growing field of predictive policing that exists solely to leverage similar results increase the effectiveness of police resources and drive crime rates. This real life application for both the data and the type of visualizations created is still in its infancy, only adopted by a few big city police departments across the US, but it is easy to how predictive policing will one day be standard practice. Nothing can replace pure police procedural and investigation skills, but predictive policing and the use of in depth data analysis can certainly aid it. This is merely one example of the many ways in which information visualizations can aid and improve efficiency and give rise to greater insight.

References:

[1] Miller, L., & Hess, K. (2005). Community Policing - Partnerships for Problem Solving. Belmont, CA: Thomson-Wadsworth.
[2] Google. (2013). Google Fusion Tables. Retrieved from Google.com: www.google.com/drive/start/apps.html
[3] Open Data Philly. (2011). Public Safety. Retrieved from Open Data Philly Project: http://opendataphilly.org/opendata/tag/9/?sort=name&dir=asc
[4] Google. (2013). Google Maps. Retrieved from Google.com: maps.google.com
[5]"Predictive Policing." *National Institute of Justice*. USA.gov, 6 Jan. 2012. Web. 11 Mar. 2013
[6] Greengard, Samuel. "Policing the Future." *Communications of the ACM* 55.3 (2012): 19-21. 3 Mar. 2012. Web. 11 Mar. 2013.
[7] Vlahos, James. "The Department of Pre-Crime." *Scientific American* 306 (Jan. 2012): 62-67. Web. 11 Mar. 2013.

Appendix:

Table 1

Count by Block	
COUNT	**LOCATION_BLOCK**
460	1000 BLOCK MARKET ST
283	1300 BLOCK MARKET ST
271	1600 BLOCK S CHRISTOPHER COLUMBUS BLVD
185	2700 BLOCK CASTOR AVE
173	1500 BLOCK CHESTNUT ST
143	0 BLOCK MIFFLIN ST
131	3700 BLOCK ARAMINGO AVE
126	0 BLOCK FRANKLIN MILLS BLVD
121	3400 BLOCK N BROAD ST
113	1600 BLOCK WALNUT ST

Table 1 Description: Top 10 blocks by number of crimes reported in 2012 (COUNT).

Table 2

Count by Day by Block		
COUNT	DISPATCH DAY	LOCATION_BLOCK
88	Wednesday	1000 BLOCK MARKET ST
67	Monday	1000 BLOCK MARKET ST
63	Thursday	1000 BLOCK MARKET ST
61	Friday	1000 BLOCK MARKET ST
61	Sunday	1000 BLOCK MARKET ST
56	Tuesday	1000 BLOCK MARKET ST
51	Monday	1600 BLOCK S CHRISTOPHER COLUMBUS BLVD
50	Saturday	1000 BLOCK MARKET ST

Table 2 Description: Most reported crimes by day by block in order of total number of crimes reported on that day (COUNT).

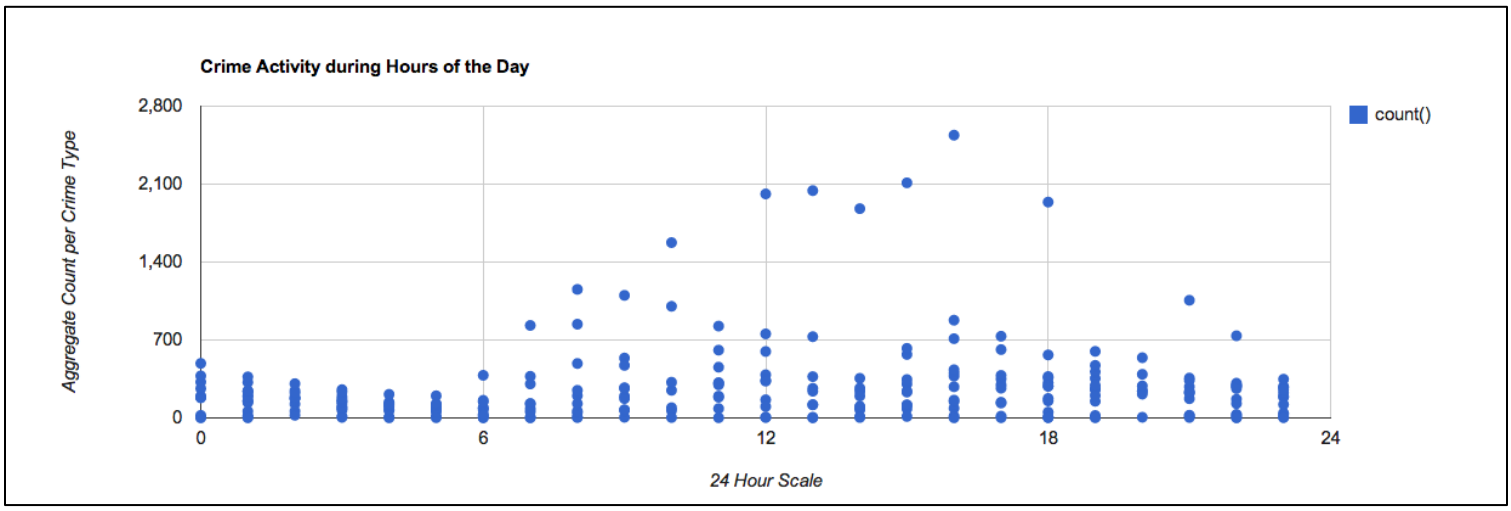


Figure 1 – Full size image of Visualization 1

Appendix Cont.:

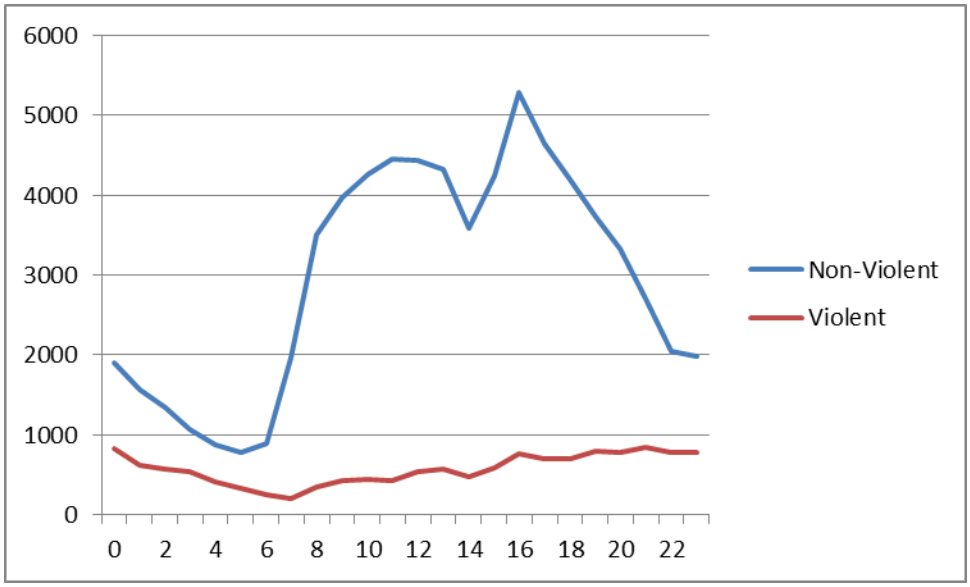


Figure 2 – Full size image of Visualization 2

Appendix Cont.:

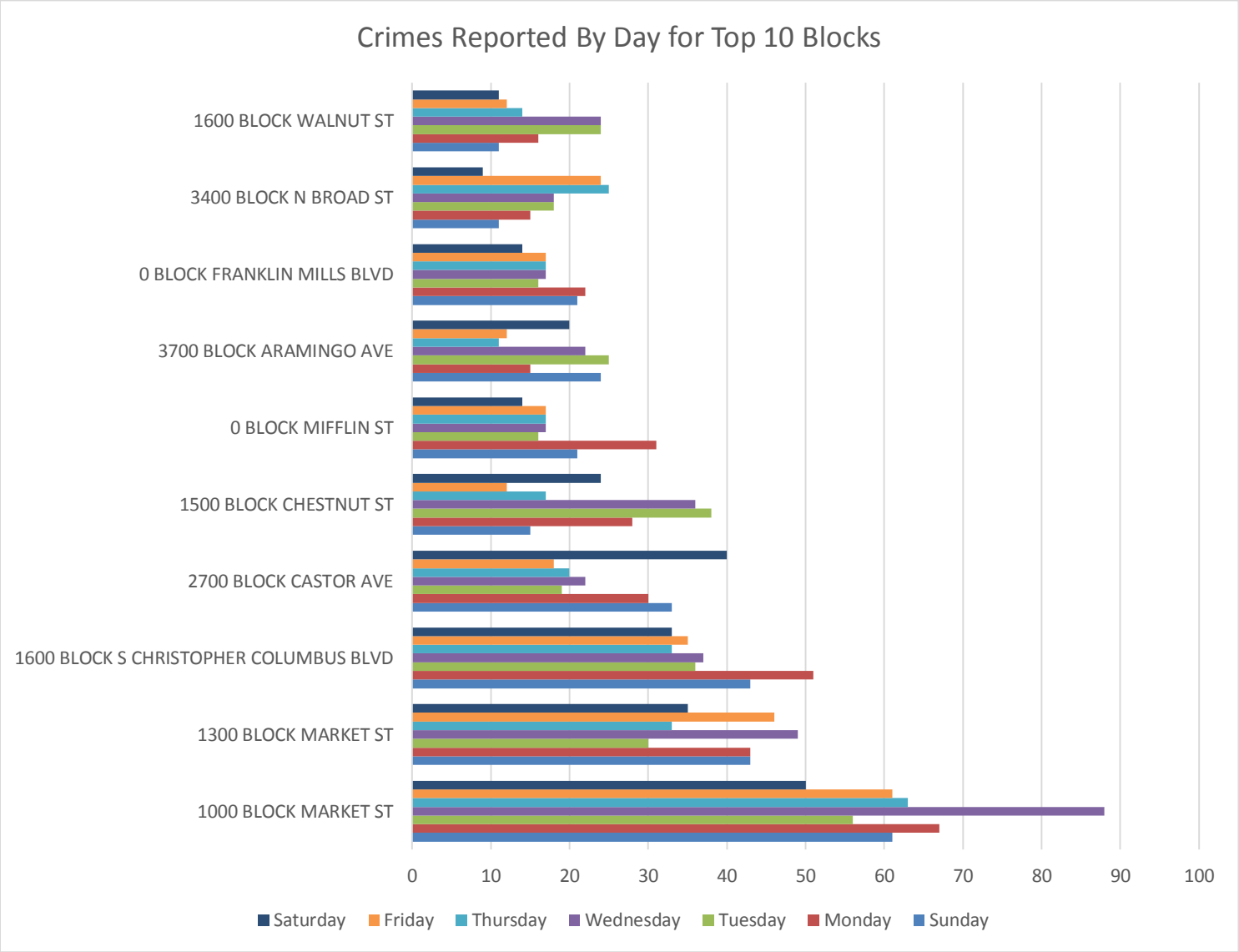


Figure 3 – Full size image of Visualization 3

Appendix Cont.:

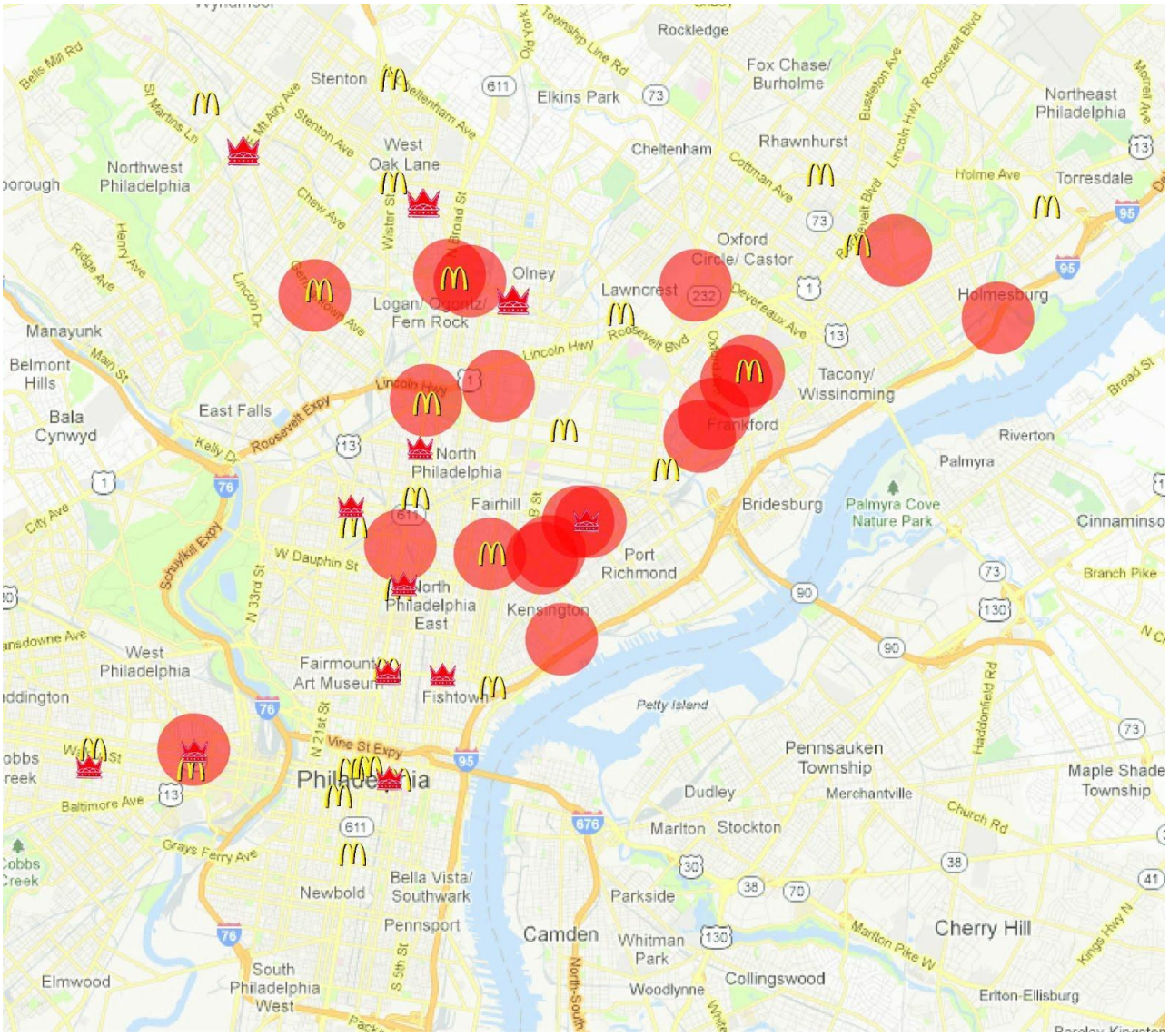


Figure 4 – Full size image of Visualization 4