4/29/2019

CRIMEWATCH: Analysis of Crime Data in Allegheny County

Final Project, CS 3551

Debarun Das, Nannan Wen
UNIVERSITY OF PITTSBURGH

Introduction

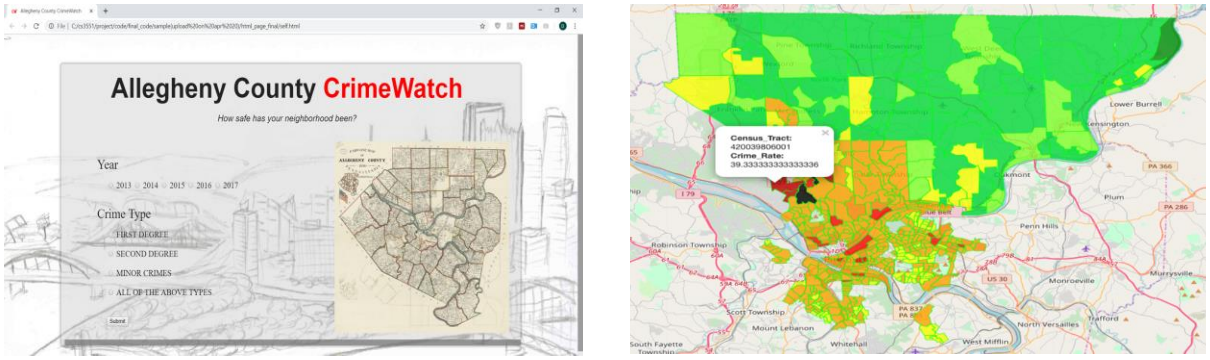
Safety of neighborhoods is a primary concern for any person looking to settle down in a new place. Pittsburgh is an evolving city with an increasing number of scholars who study in the numerous universities around the city. Moreover, with a string of new startups (like Duolingo) and giants (like Google) opening their offices here in the recent past, more people have flocked to the city in pursuit of the fresh opportunities. Awareness of the safety of neighborhoods helps people to choose places to stay. Moreover, this helps investors who want to invest in a new business to select safe neighborhoods for running their businesses as it is more likely to be successful in such neighborhoods. Finally, such knowledge would also help police to make decisions on deployment of force as more police officers can be deployed to unsafe neighborhoods to prevent more crimes. In this project, we study the distribution of crimes in Allegheny County over a period of five years (2013-17). The contribution of this project is twofold:

- Visualization of the intensity of crimes across Allegheny County based on:
    - Year of crime
    - Type of Crime
- Analysis of the crime dataset to establish potential trends across years, quarters, months and census tracts.

We initially discuss about the Datasets we use and the different ways in which we processed our data, followed by discussions on the implementation and results. Finally, we have the conclusion and future works.

Datasets and Data Pre-Processing

The primary datasets that we use are the Police Incident Blotter datasets of Allegheny county (2005-15, 2016-19) [1]. These datasets maintain crime incident data across Allegheny county and were published on a nightly basis. Initially, we isolate the data according to the year of occurrence of a crime incident by using the value of the *INCIDENTTIME* feature in the datasets. Furthermore, we divide the dataset according to the type of crime [2]. We classify crimes into three categories – First Degree (*Homicide*, *Rape*, *Kidnapping*, *Arson*, *Terrorist threats/activities*), Second Degree



Census_Tract:
420039806001
Crime_Rate:
39.333333333333336

Figure 1a

Figure 1b

Figure 1: a) The interactive interface for choosing the crime type and year, b) The visualization of normalized crime count of a specific type across the census tracts in the map of Allegheny county

(*Theft*, *Assault*, *Burglary*) and Minor Crimes which consist of every other remaining type of crime.

As we wanted to study the distribution and intensity of crimes across census tracts in Allegheny county, we transformed the values of the *X* and *Y* coordinates of an incident to their corresponding census tract. For this purpose, we used the “Allegheny County Zip Code Boundaries” dataset [3], “Allegheny County Census Block Groups 2016” [4] dataset and the “Zillow Neighborhoods PA” [5] dataset to use the *X* and *Y* coordinates in order to extract their corresponding twelve-digit Census Tract numbers, Zip Codes, Region Ids and Region Names. While browsing through the Police Blotter dataset, we observed that there are several entries with invalid (empty or zero) values of *X* and *Y* coordinates of a crime incident. We cleaned the dataset by removing such entries. In order to account for the variation in population across different census tracts, we normalize the crime count in each tract by their population. We were only able to access population data [6] of Allegheny county from 2013-17, so we used the Police Incident Blotter dataset from the years 2013-17 too. We tried to link the Allegheny County Property Assessment [7] dataset to the Police Blotter dataset. However, as the Property Assessment dataset lacks the *X* and *Y* coordinates feature, so we could not proceed to do so.

Implementation and Results

We implemented a simple interactive interface that allowed visualization of normalized crime count across the census tracts of the Allegheny County map, which was implemented using the Folium library in Python 3.7.2. The interface was created using HTML, CSS and JavaScript. As shown in figure 1a, the interface allows a user to visualize the normalized crime count based on the year and type of crime. In figure 1b, we see that the census tracts are colored ranging from Dark Green to Dark Red representing census tracts with lowest and highest crime rates, respectively. We chose to color some census tracts black to indicate tracts having a crime count greater than zero, despite having zero population.

Figures 2a and 2b show the crime count by year and type before and after data cleaning. With data cleaning, we remove the data entries with invalid (empty or zero) *X* and *Y* coordinates in the Police

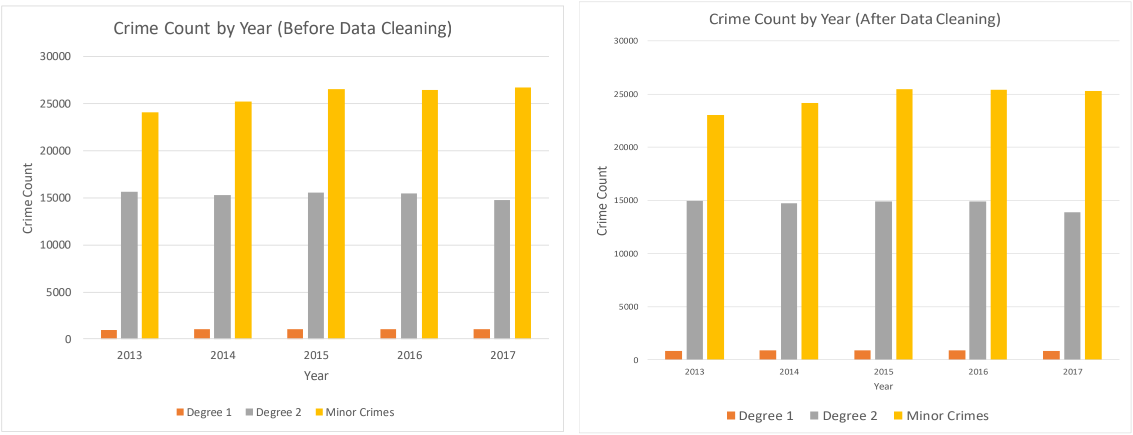


Figure 2a

Figure 2b

Figure 2: Crime count by year and type a) before and b) after data cleaning

Blotter dataset. For both the graphs, we see that the crime count of minor crimes is the highest, followed by second degree crimes and first degree crimes, respectively. There has been an overall increase in total number of crimes from 40,790 to 42,552 before data cleaning and from 38,839 to 40,002 after data cleaning, over the five years. However, the year 2015 has had the highest total crime count (of 43,117) among all the years before data cleaning and the year 2016 has had the highest total crime count (of 41,249) among all years after data cleaning. Hence, we see that even though the overall crime count has increased over the five years, yet there is no definite increasing trend in crime count from one year to the next. One of the interesting observations we made while browsing across the Police Blotter dataset, was that some specific types of crimes are almost always without a valid *X* and *Y* coordinate (like almost all of the entries for crimes like Rape and Sexual Assault had no valid *X* and *Y* coordinates). This can potentially be a reason why the trend in total crime count change over the five years is almost similar before and after data cleaning.

Figures 3a and 3b show the total crime count across different quarters and months over the five years, respectively. All the graphs in figure 3 include data before data cleaning because we wanted to present a complete representation of the crime counts over different months and quarters. In figure 3a, we see that the crime count is consistently higher in the second and third quarters than the first and fourth quarters. This is further confirmed by figure 3c where we see that the average crime count in the second and third quarters is higher than that of the first and fourth quarters, over the five years. Figure 3b shows a general trend of increase in crimes during the months in the

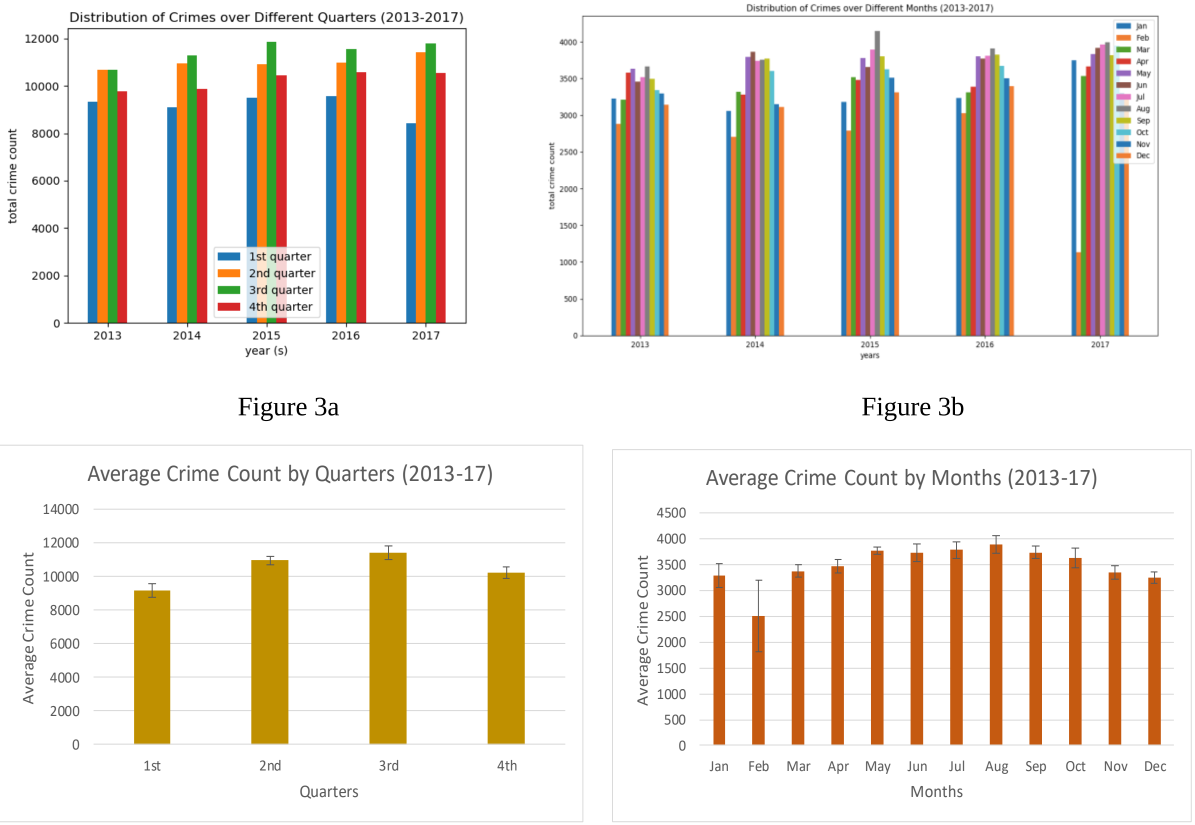


Figure 3a

Figure 3b

Figure 3c

Figure 3d

Figure 3: Distribution of crimes over a) quarters, b) months. Average crime count by c) quarters and d) months

middle of the year for all the five years, which is further confirmed by figure 3d. Over the five years, August has had the highest average crime count of 3895.2 with a standard deviation of 171.3 and February has had the lowest average crime count of 2505.8 with a standard deviation of 695.93. The high standard deviation in February is primarily attributed to the fact that in the year 2017, the crime count in February decreases drastically to 1130. This might be attributed to any popular event that might have taken place at that time or due to severe weather changes that might have happened in February that year or it may simply be due to a lapse in data entry during that time. Overall, we see that the warmer months tend to have more crimes than the colder months.

Figures 4 and 5 show the distribution of crimes across the census tracts in Allegheny county. For both the figures, we do not plot the outliers. Any census tract with normalized crime count greater than 1 is considered an outlier. Such a census tract has more crimes than the number of people staying there. Since the number of such census tracts is quite small (~10), so we discard them. Figure 4 shows the distribution of crimes across census tracts after the data has been sorted in descending order by the normalized crime count. We observe that there is a gradual decrease in crime count across the census tracts for all the years. While the highest normalized crime count changes for different years, the trend in decrease of normalized crime count across all the census tracts is similar for all the years. Since this does not gives us enough information about the variation in distribution of crimes across census tracts for different years, so we plot another set of graphs in figure 5. The graphs in figure 5 are plotted using data that is sorted by the census tract numbers in ascending order. We observe that the distributions of crimes over the different years vary and are not completely similar to each other. However, they do show some general trends with crimes happening in approximately the first 3/4th fraction of the census tracts followed by an almost barren patch, which is followed by some amount of crimes towards the end of the graph. The census tracts in the barren patch of the graphs in figure 5 have been relatively safe consistently over the five years and is possibly a good neighborhood to stay.

Conclusion and Future Works

In this project, we primarily focused on visualization of the intensity of crimes across the census tracts of Allegheny county. We developed a simple web based interactive interface for this purpose. In addition, we analyze the Police Blotter dataset of Allegheny county to study trends in crime count values for the years 2013-17 across different months, quarters and census tracts. We observed that there has been an overall increase in crime from 2013-17. Also, we see that more crimes occur in the warmer months than in the colder months. Finally, we establish some general trends in distribution of crimes across the census tracts in Allegheny county.

Potential future work for this project include extension of the analysis of crime data to more years (2005-19) to understand the trend in distribution of crime across the census tracts over the years more clearly. This data can also be used to predict future crime rates by using different time series algorithms. If the data is large enough, then we can potentially train LSTM and other deep learning models for prediction. This can be made more interesting by associating the crime data with other datasets like the Property Assessment dataset and the Transit dataset. In such a scenario, we can create a predictive model that can predict the crime rate of a neighborhood based on the average property valuation of the neighborhood or based on the average public transportation frequency of a neighborhood. We can further evaluate the performance of several regression models to determine the model that fits best for such prediction. Finally, the visualization interface can be improved and designed in the form of an Android/iOS app that makes it more user friendly and has more functionalities.

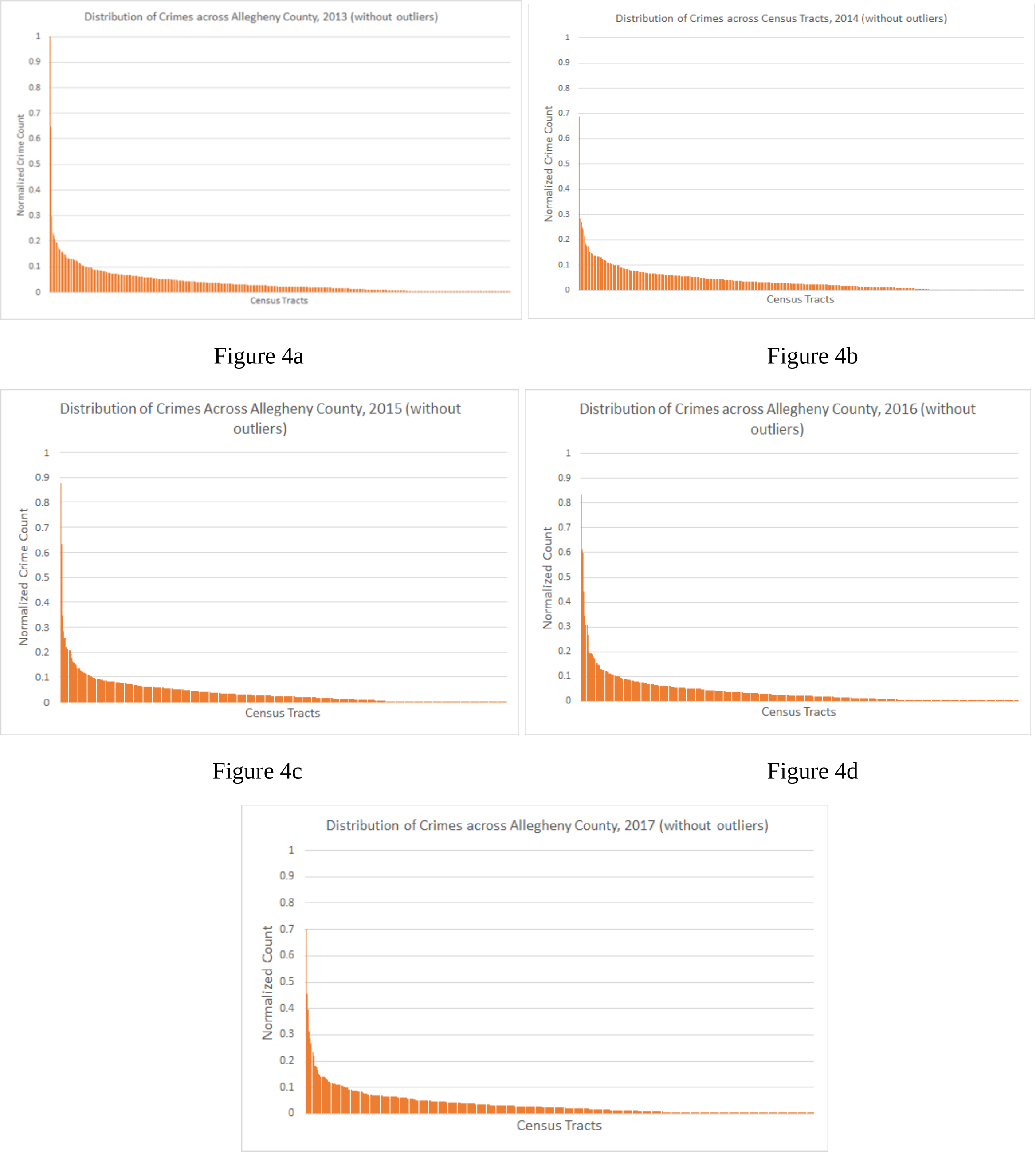


Figure 4a

Figure 4b

Figure 4c

Figure 4d

Figure 4e

Figure 4: Distribution of crimes across Allegheny County with sorted normalized crime counts in a) 2013, b) 2014, c) 2015, d) 2016, e) 2017

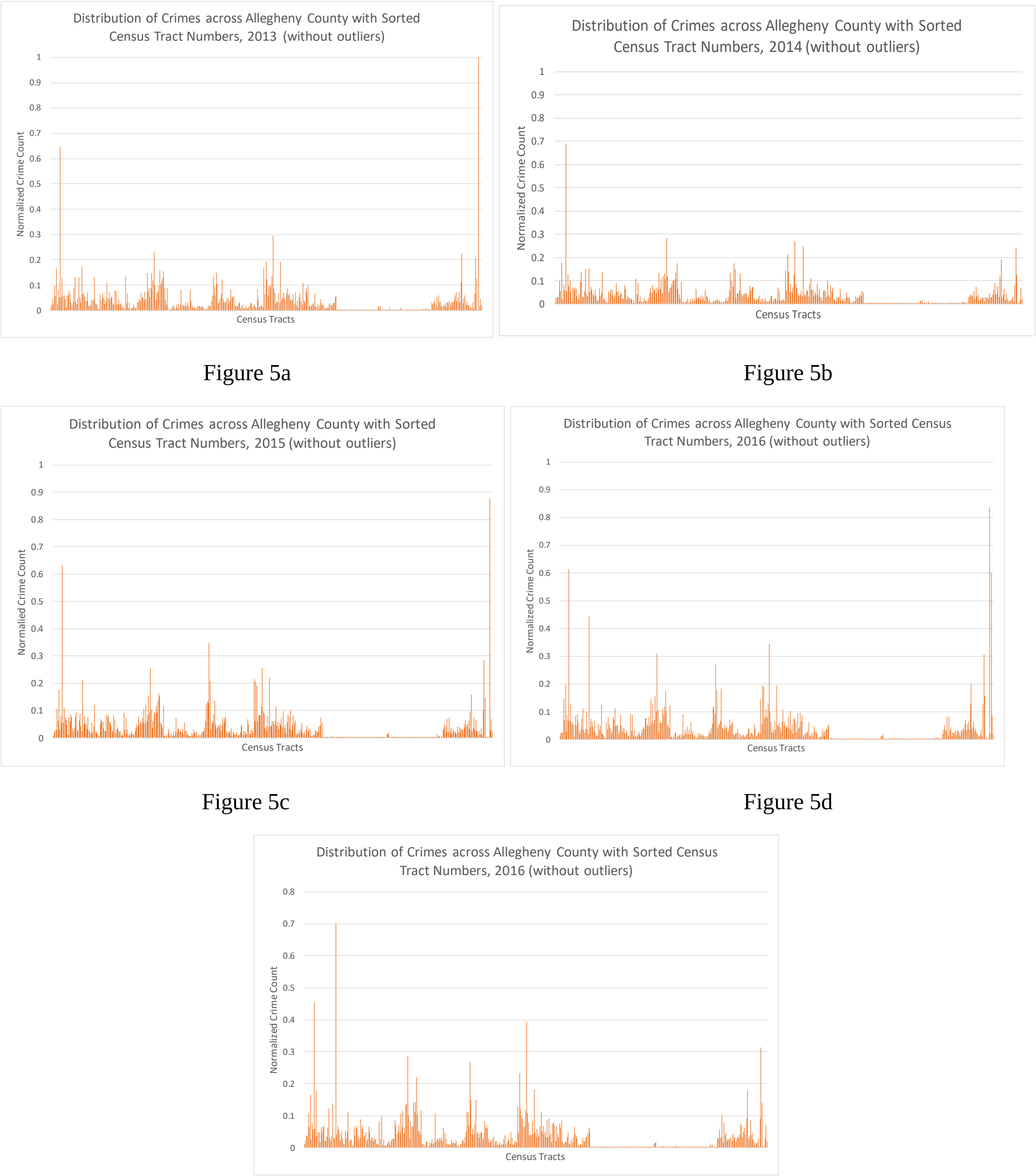


Figure 5a

Figure 5b

Figure 5c

Figure 5d

Figure 5e

Figure 5: Distribution of crimes across Allegheny County with sorted census tracts in a) 2013, b) 2014, c) 2015, d) 2016, e) 2017.

References

[1] WRPDC, “Police Incident Blotter (Archive),”. [Online]. Available: https://data.wprdc.org/dataset/uniform-crime-reporting-data.

[2] “Pennsylvania Crime Classification | David J. Cohen Law Firm, LLC.” [Online]. Available: https://www.davidcohenlawfirm.com/pennsylvania-crime-classification.

[3] “Allegheny County Zip Code Boundaries - Data.gov.” [Online]. Available: https://catalog.data.gov/dataset/allegheny-county-zip-code-boundaries-9a066.

[4] “Allegheny County Census Block Groups 2016 - Datasets - WPRDC.” [Online]. Available: https://data.wprdc.org/dataset/allegheny-county-census-block-groups-2016. [Accessed: 29-Apr-2019].

[5] “Zillow Neighborhood Boundaries | Zillow.” [Online]. Available: https://www.zillow.com/howto/api/neighborhood-boundaries.htm. [Accessed: 29-Apr-2019].

[6] U. S. C. Bureau, “American FactFinder - Search.” [Online]. Available: https://factfinder.census.gov/faces/nav/jsf/pages/searchresults.xhtml?refresh=t.

[7] “Allegheny County Property Assessments - Datasets - WPRDC.” [Online]. Available: https://data.wprdc.org/dataset/property-assessments.