Visualization of crime data in Baltimore

Process Book

Group No: 14

Name: Ravi Teja Kodali

Email: rkodali@g.clemson.edu

Clemson id: C10966668

Name: Venkata Revanth Naidu Danala

Email: vdanala@g.clemson.edu

Clemson id: C18459834

Repository link: VenkataRevanthNaidu/DatavizG14

1 Overview: The main overview of our project is to create a scatter plot based on probability and intensity. The scatter plot shows the crime with the highest intensity and the probability of crime that is likely to occur in that particular hotspot. After that step, we want to display the hotspots in the map of Baltimore city by using the coordinates, the latitude, and longitudes. Lastly, we want to create a dynamic interaction between the scatter plot, map and from the map to line chart to see the overall trends.

1.1 Motivation: Now a days, all the countries and their governments in the world are stepping towards public security to control the crimes that are being occurred. Billions of dollars are being spent each year to stop criminal activity. We can see that the crime rate in many countries gradually reduces year by year. In countries like the United States of America, the police department keeps an eye on the criminal activity that occurred with their data from past crime history.

After seeing all this, the police departments in the USA are speedy and tactical in finding out the crimes, solving the crimes. This is one of the main reasons to decrease criminal activity in the USA. This is our main motive behind selecting this field. By studying the crime data, we can create many Visualizations, which will be helpful to understand the data quickly, and we will implement Viz techniques and their tools to a particular crime dataset to get better outcomes. Using the results of our visualization, the police department can quickly get an idea of the crime data by looking at the visualizations. We started searching in online for the problem we were looking then we found the journal from IEEE "CriPAV: Street-Level Crime Patterns Analysis and Visualization." We as a graduate student the goal is to replicate the designs and techniques from the reference paper to our project.

2 Related Work: One of the research papers we took as the reference is "CriPAV: Street-Level Crime Patterns Analysis and Visualization." The overview of the journal is, they created a new tool called CriPAV, used to get the street-level view of crime patterns. This research paper mainly focuses on analyzing the crimes based on Probability vs Intensity. We found this paper interesting because they classified hotspots based on probability vs intensity, which is a different approach. After categorizing the hot spots, they have presented that on a Geospatial map where these visualizations may be helpful to solve real-world problems. They further used different techniques in deep learning, Python, and image processing. But Dr. Fedrico suggested us to focus on calculating probability and intensity to create a scatter plot and then to represent the data in the street-level map. Finally, to show the overall trends. Taking this as inspiration, we want to use these concepts for our data set and analyze Baltimore City's crimes from 2012 to 2017. The visualizations we discussed in the class and the assignments helped us immensely. We took the scatter plot, line chart, interaction techniques, and creation of the maps as a reference while doing our project.

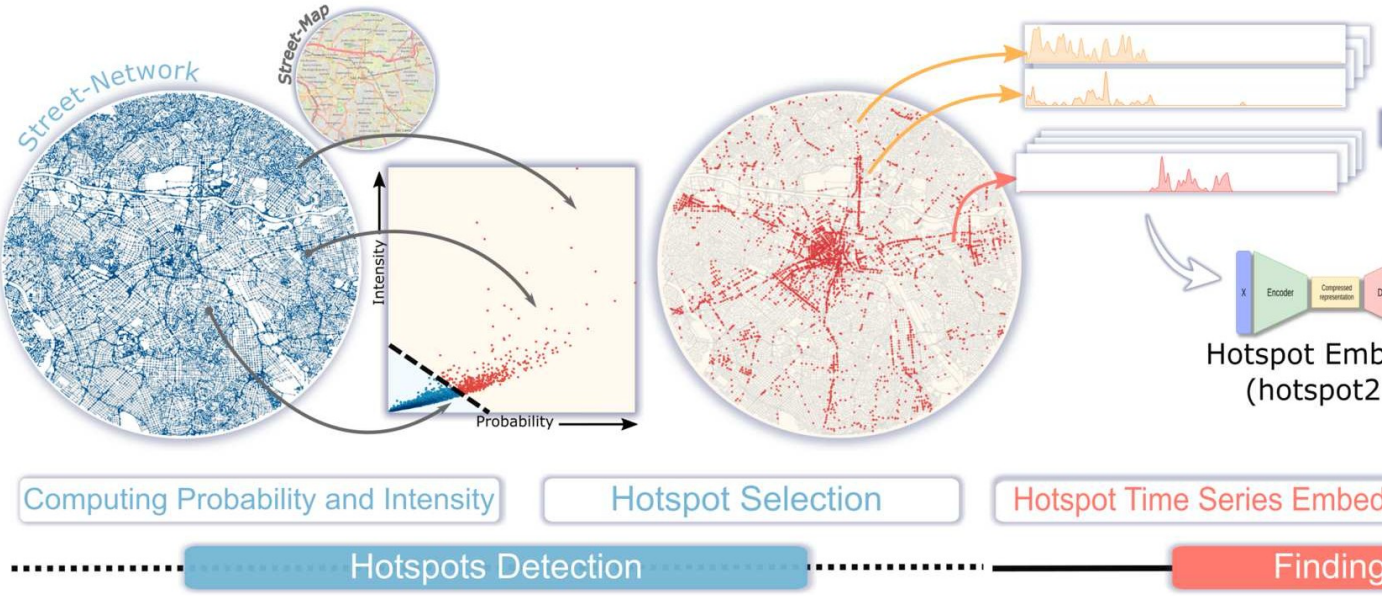

Figure 1 shows the pipeline that we should replicate from our reference paper to our project.

2.1 Pipeline: we have stuck to a pipeline from the paper we have taken as a reference. The pipeline in the paper is to create a probability vs scatter plot to identify the hotspots and then represent the hot spots in the street level map. Lastly, presenting the overall trends that are obtained from the above two steps. So, we wanted to do the same as they did in their research paper. We want to create interactions between the scatter plot, map and from map to line chart to see the overall trends from 2012 to 2017.

3 Questions:

1. What are the hotspots in Baltimore city based on the probability of the crimes? Scatter plot
2. What is the crime that has the highest probability and intensity of occurring in a neighborhood?Geospatial graph
3. What is the overall trend of crimes over the years?

All these questions evolved gradually and at the initial stage of the project, it seemed to be easy, but as we started developing the visualizations, we must do many things with our data. We have done a lot of data preprocessing and used several libraries in JavaScript. In the end, we got the desired output.

4 Data: We have collected the data from the government site of Maryland, USA. The dataset consists of the crimes that have been happened in the Baltimore city from 2012 to 2017. It has 2.7 million rows of data. Initially, the dataset has 13 columns. They are Crime Date, Crime Time, Crime Code, Description, In/Outside, Weapon, Post, District, Neighborhood, Longitude, Latitude location of the crime that happened, Premise, Total Incidents.

After implementing the data in order to visualize, we found some difficulties. So that we did some data cleaning. We have removed the attributes that are not required for our visualization part; they are Crime Time, In/Outside, Post, Premise, Total Incidents. Then, we did the data preprocessing for the dataset. There are some null values in the dataset. We removed all the null values by using python. To reach the goals of our project, we should have the probability and intensity of the crimes that we did not have in the dataset. Using PANDAS, we calculated the Probability and Intensity of the crimes in each location. The formula for calculating the probability is max(crime)/total number of crimes at a specific location. We have taken the crime, which is the highest in every location, probability of the crime that is likely to occur, and the intensity is nothing, but the total number of crimes that happened at a specific location. Also, we have used Date Time to remove the day and month from year to present the data year-wise. By using pandas, we have created six CSV files from 2012 to 2017. After computing the probability and intensity we have merged all six files into one. Lastly, we must do a little bit of data preprocessing for our final interaction. First, we took a geojson file from the web and then created a similar dataset with our data. As per the geojson file we have, we combined 2-3 locations and aggregated total crimes in those locations for six years of data.

5 Exploratory Data Analysis:

At the starting stage of the project, we could not replicate the probability*intensity scatter plot. Then Dr. Fedrico suggested to calculate both probability and intensity. We started the project with aim to replicate the techniques used in research journal to our project. As we are doing our project on crime hotspots, the intention is to create the scatter plot and then represent the hotspots in the map. Then Dr. Federico allowed us to use any JavaScript libraries to create a map. Then we started searching for the best library to create the street level map. We saw the leaflet a JavaScript library and found that good to create a street level map. It allowed us to create the Baltimore city map with the help of coordinates. Then, we created duplicate data to check whether we were getting the desired visualizations and started the Visualization part. Firstly, we have created a duplicate CSV file with probability and intensity and created a scatter plot. It has obtained the desired visualizations. Then we have created a duplicate geojson file and used leaflet, an open-source JavaScript library. Using the coordinates of Baltimore and the help of the latitudes and longitudes in our dataset, we plotted the crime locations in the map. Then we worked the same with our dataset, created the map, and plotted the actual locations in the map. We also used the Observable site as a reference to check whether all our visualizations are coming in the right way or not. We also saw some geospatial visualizations in the observable site and then created the map using JavaScript. All our results came in the right way, and we gained all the insights we required. These insights are used until the end of our project and help us get our desired outputs.

6 Design Evolution:

After picking the topic and dataset, we thought of doing the bar chart, Sankey graph, and Line chart as our visualizations to present the data. Later Dr. Fedrico suggested to pick a research paper and asked to replicate the same visualizations in that paper. We took an IEEE journal as our reference paper. The main goal of our project is to replicate the techniques used in the paper that we took at the start of the semester. After reading the journal, we learned that we need to do interactions to visualize our data so that it will be easy for the user to understand the data. Then we stopped doing the above visualizations. After understanding our reference paper, we started computing probability for all the locations from 2012 to 2017. We took the crime which

has occurred more number of times in a specific location, and then we thought a scatter plot would be the ideal visualization to present the data by choosing the dependent variable "Intensity." Using the scatter plot for this visualization, we can easily plot the data in a scatter plot. We started using the observable site and then created the scatter plot with dropdown. With the help of leaflet, we created the map of Baltimore city and then plotted the hotspots in the map. Then for selecting the years we kept the dropdown option to select a year. Both our Scatter plot and the geomap looks great at that time. Dr. Fedrico suggested us not use the observable site. He also said to use the observable site just as a reference and asked us to write the code in JavaScript file. Then we started did the scatter plot in d3.

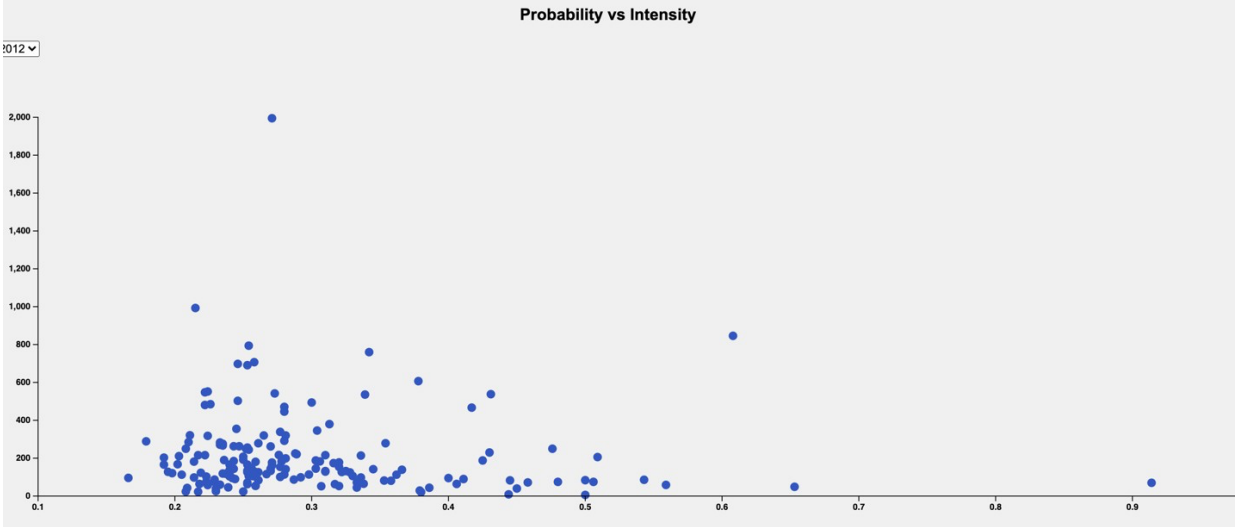

Fig 2. The scatter plot above shows the probability vs intensity.

As our project is based on crimes in Baltimore city, displaying the hotspots in the map would be great because visualizing the data on the map can be visible clearly to everyone, and the design of the map and the data can be easily understandable. Now we focused on how to utilize the leaflet library. We first tried to create a map of Baltimore. For obtaining a map we placed our dataset in the ArcGIS website to read the data via latitudes and longitudes. Then we created the street level map of Baltimore. We worked on plotting the locations on the map using their latitudes and longitudes of every hotspot. We displayed the crime which has occurred more number of times in the location and the location name to make it easier for the user to understand.

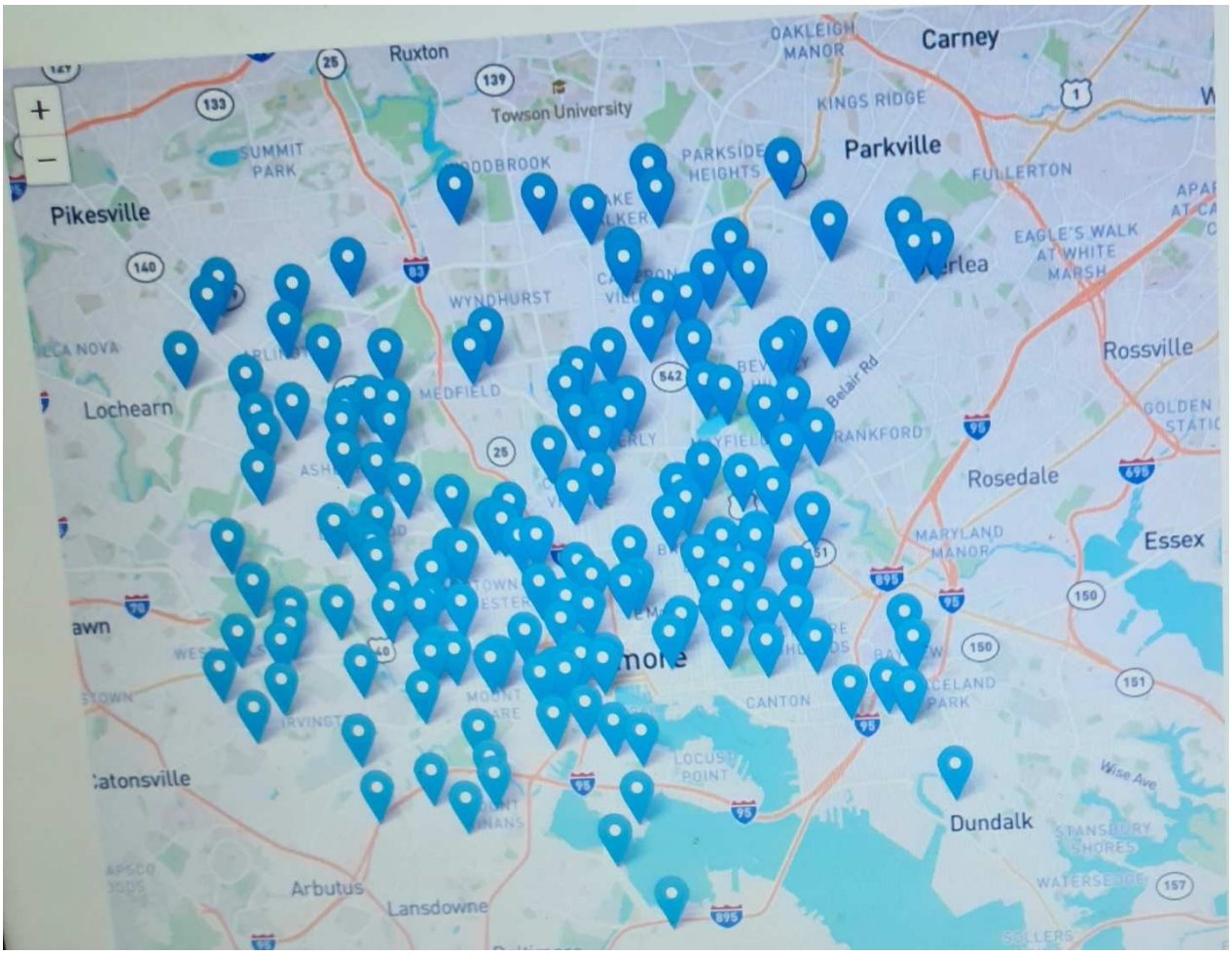

Figure 2.1 shows the initial street level map

From the feedback the feedback given during our prototype we added the tooltips for all our visualizations.

After designing all the three visualizations, Dr. Fedrico suggested us to create an interaction between the visualizations. We started working on implementing an interaction between the Scatterplot and Geomap. Firstly, we used some libraries in JavaScript to create dynamic interactions. We used the same map for the dynamic interaction. Then we choose the similar probability vs intensity scatter plot using chart.js, it is also used for the visualizations in JavaScript. With the help of leaflet, we used Api link and created the street level of Baltimore city and then we written a code of scatter plot in the leaflet code. Using some functions, we performed dynamic interaction between map and the scatter plot. We also kept an option, like if we zoom the map closely, we can see those particular locations on the map. We also kept impact and intensity on tooltip for the scatter plot. It allows us to see both impact and intensity of that specific location when we place the cursor.

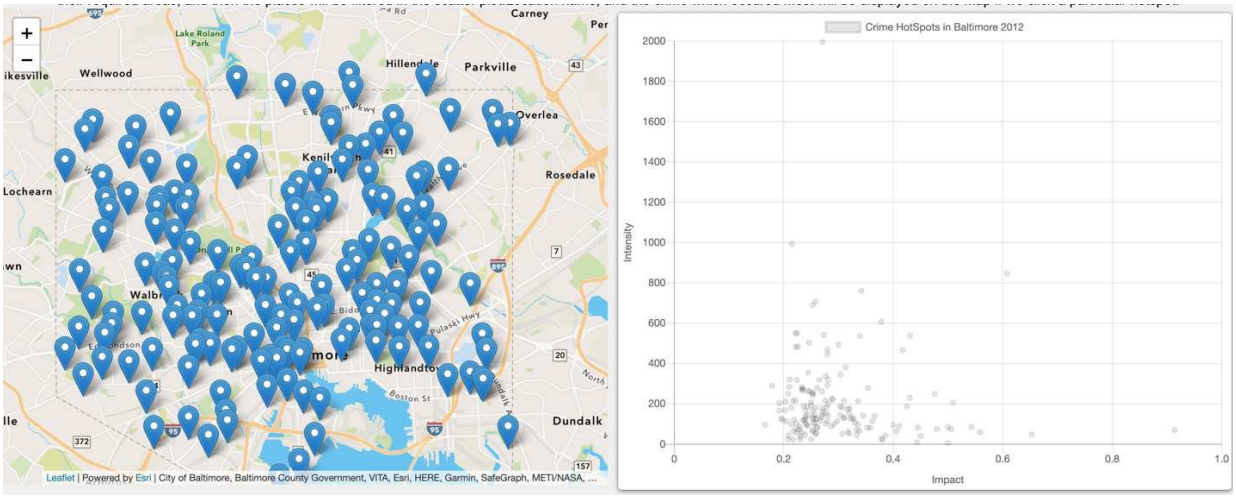

Figure 2.2 Shows the hotspots in the map and the scatter plot shows the hotspots in Baltimore.

Coming to the final interaction, we have combined intensity of all crimes that have been occurred in a neighborhood. we created the Baltimore map by using d3.geo equirectangular() function. We used the community_statistical_Areas_CSAs_Reference_Boundaries geojson file to create the boundaries of Baltimore city. Then we used some d3 functions to create the map. To display the ranges of the crimes in different hotspots we applied the color hue and given the range. Then we developed the code for the line chart to display the overall trends of the crimes in every location in Baltimore. We have chosen the years as X-axis and the count of the crime on Y-axis. Now, we have inserted the data into the map and then written the code for line chart.

Figure 2.3 shows the map of Baltimore with color hue and the line chart shows the trends of the crimes from 2012 to 2017

Until the completion of our project we were in the pipeline and made one change from the professor has asked. Dr. Fedrico asked to make interactions for all the 3 visualizations at a time. But unfortunately, we made interaction for scatter plot and Street level map and then we created new map based on boundaries and then created an interaction between the map and line graph. The only difference between the pipeline in our reference paper and our project is the authors created interaction from scatter plot to map and then to the line chart. But we used a scatter plot, two maps and then a line chart. (Scatter plot => Map & Map => Line chart)

7 Implementation:

Probability vs Intensity scatter plot

This visualization shows the hotspots in Baltimore city from 2012 to 2017 based on probability and intensity. The circles represent location in Baltimore city.

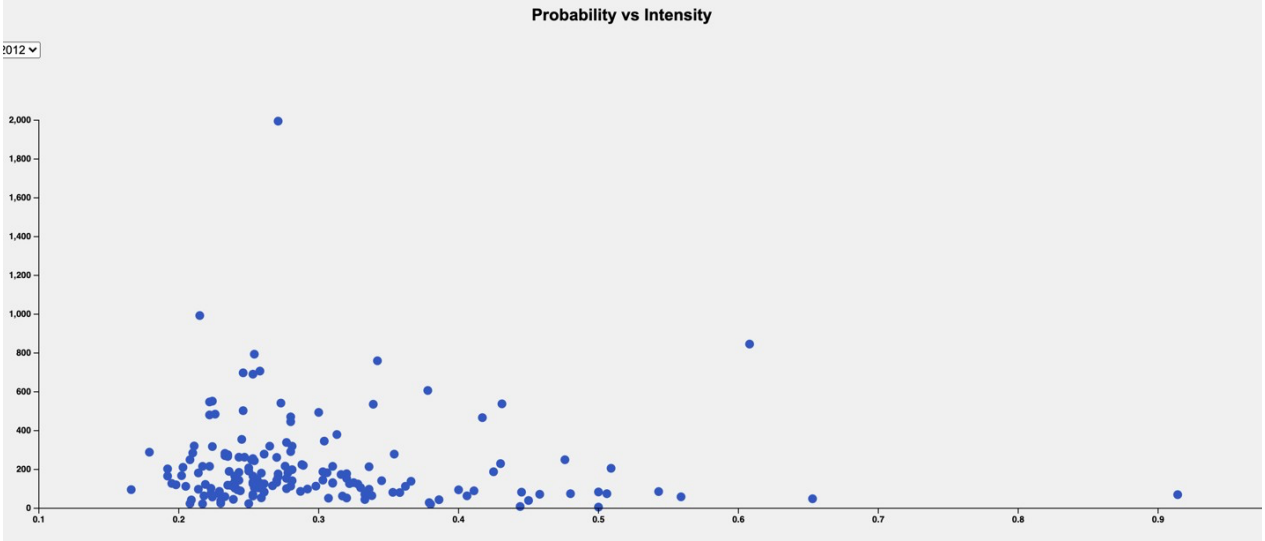

Figure 3 shows the scatter plot based on probability and intensity

We placed the option to choose the year from 2012 to 2017. By selecting the year, the graph gets updated and show the hotspots on that particular year.

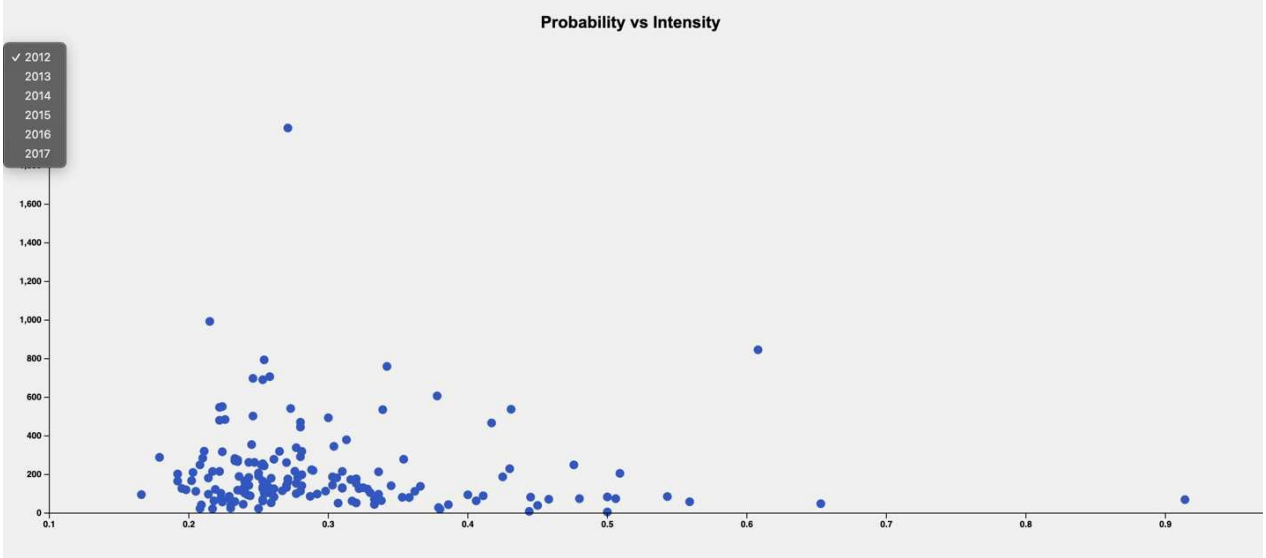

Figure 3.1 shows the drop-down option to select the years from 2012 to 2017

We have also mentioned both impact and intensity of every location in the scatter plot by using the tooltip. We can see both impact and intensity below the drop-down option.

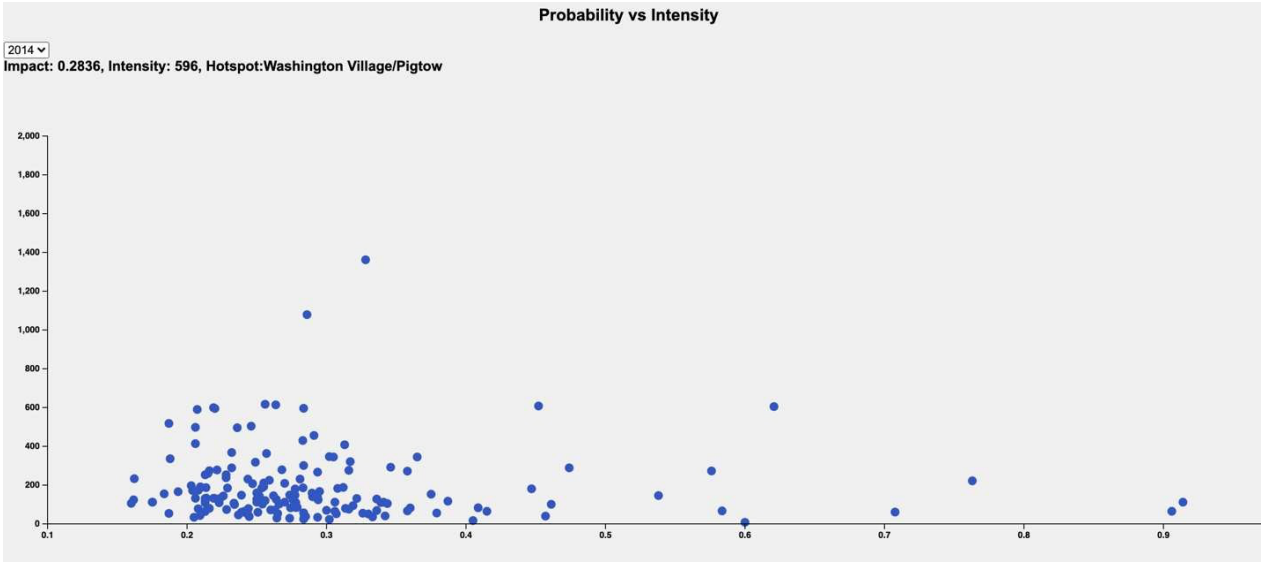

Figure 3.2 shows Impact, Intensity and the location name by using tooltip

Main Visualization

In the main visualization, we used the same map we created by leaflet and then choose the similar probability vs intensity scatter plot. Where we kept the tool tip on scatter plot to display the intensity and probability. We have also kept the option that if we zoom the map the hotspots in the scatter plot will be filtered.

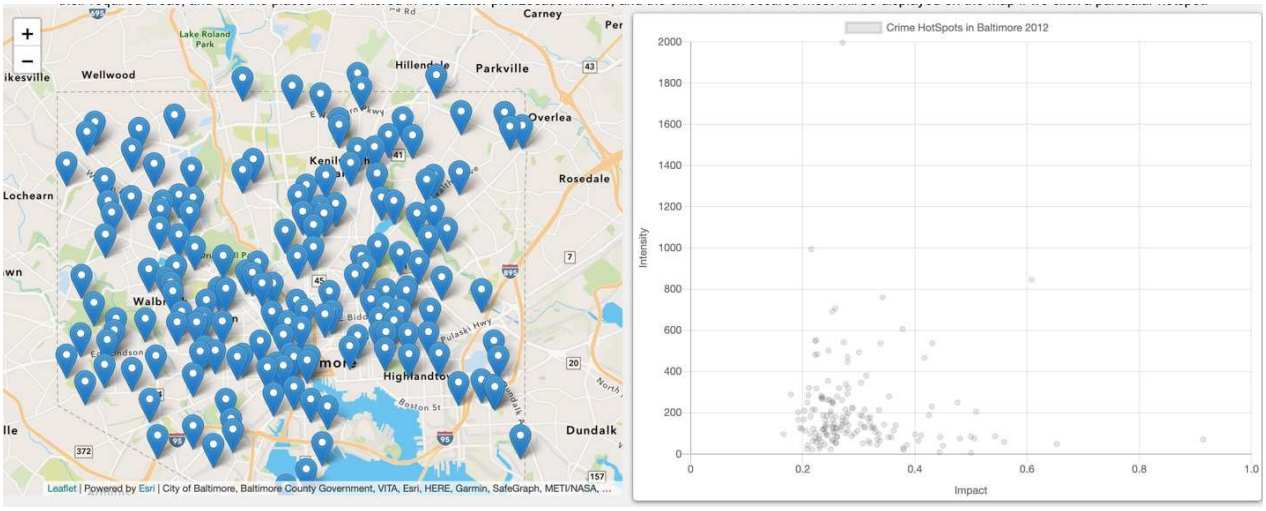

Figure 4 Shows the hotspots in the map and the scatter plot shows the hotspots in Baltimore.

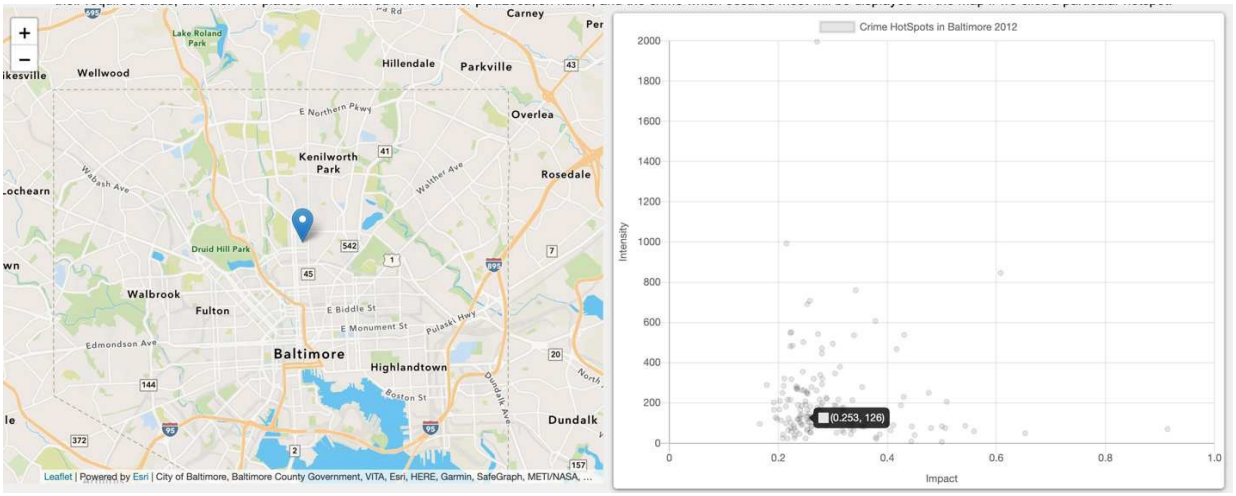

Fig 4.1 shows the dynamic interaction created between scatter plot and map. The hotspot selectedin the scatter plot shows the impact and intensity and that location is showed in the map.

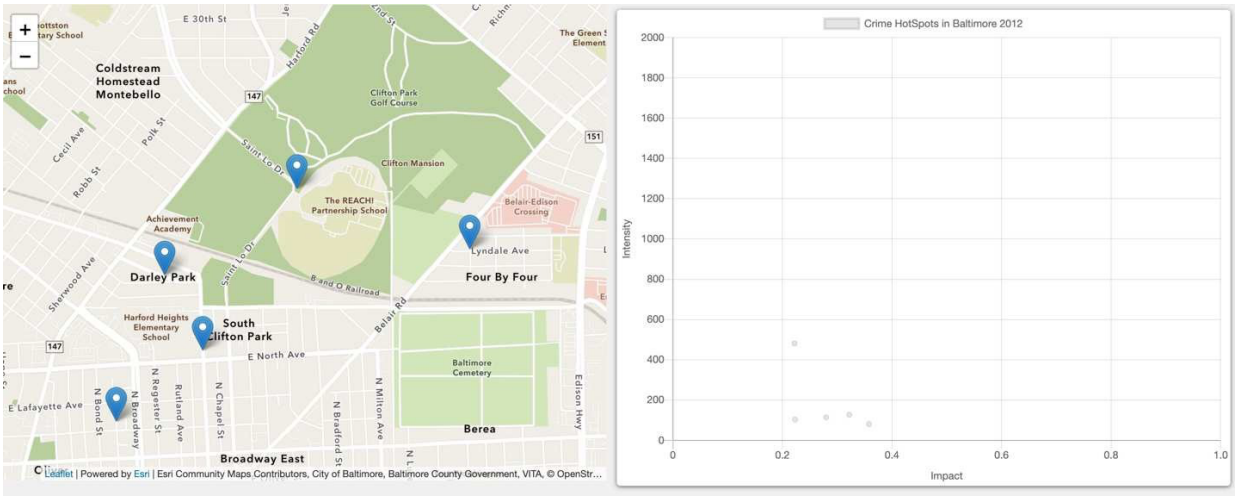

Figure4.2 shows the filtered locations in the scatter plot when we zoom the map.

In our final interaction we used the map that is created by the boundaries of Baltimore and with the help of d3 functions and then interacted that with the line chart.

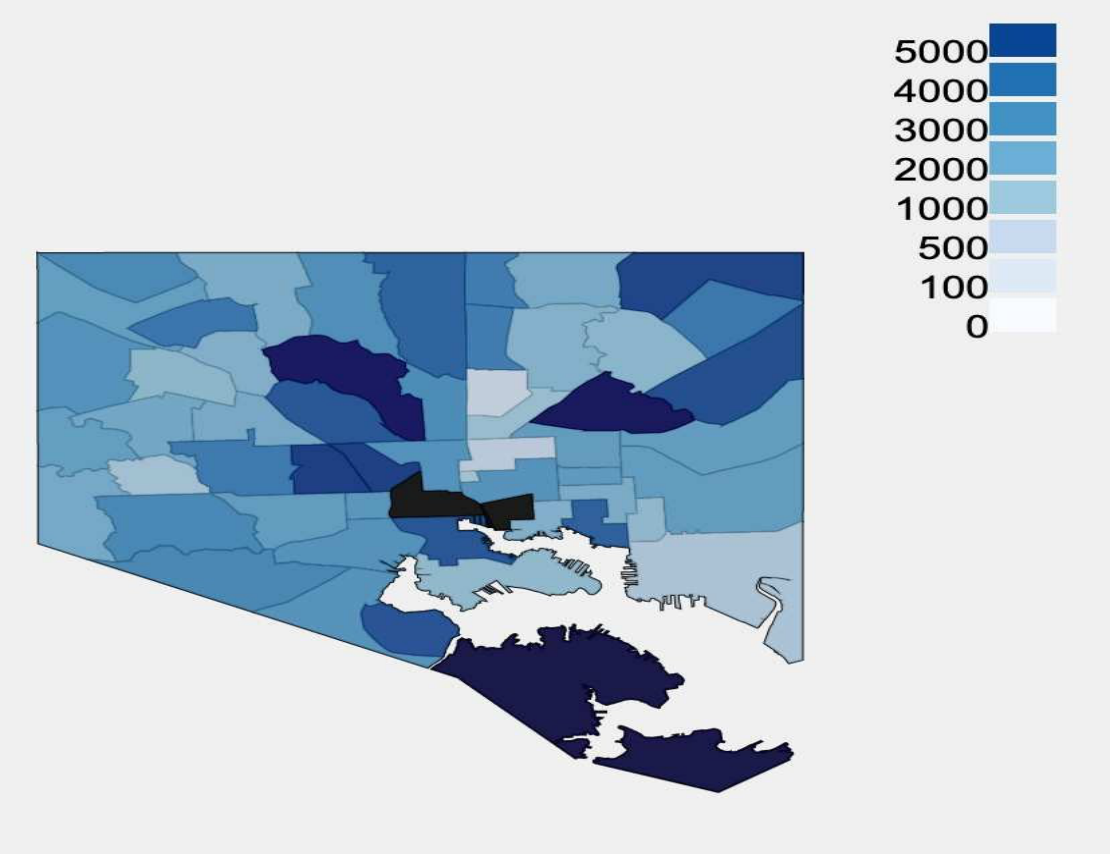

Figure 5 shows the map of neighborhoods in Baltimore with color hue applied

Figure 5.1 shows the overall trend of the crimes from 2012 to 2017 by using the line chart

Evaluation:

By performing these visualizations there are many insights gathered.

- Firstly, the main highlight of the project is "**finding the hotspots no on the total count of the crime but, by the probability of the crime that is likely to occur.**" We have found the hotspots in Baltimore city.
- Secondly, we can see the range of the crimes occurring in Baltimore.

- We can also see the exact location of the crime that have been occurred. Let's say that the crime has occur on the street, grocery store etc.
- Lastly, we can see the overall trends of the crimes over the years.

In the probability vs intensity scatter plot, we can see that many locations have the probability of occurring crimes is around 0.2 to 0.3 crimes. There is one location which is named as Downtown has registered a greater number of crimes from 2012 to 2017. Another insight we found is, that there is one common crime; Larceny, which has been occurred more number of times in majority of the locations from 2012 to 2017.

Overall, our visualizations answered all the questions mentioned above.

- **What are the hotspots in Baltimore city based on the probability of the crimes?**
- **What is the crime that has the highest probability and intensity of occurring in a neighborhood?**
- **What is the overall trend of crimes over the years?**

Our first visualization allowed us to answer for the first question by creating a scatter plot, we got the hotspots in the Baltimore city. We used intensity of the location as the dependent variable.

By creating the street level map, it allowed us to answer for our second question. We have plotted the location in the map and displayed the crime that have occurred more number of times. By including interaction and tooltips between street level map and the scatter plot we can have a clear idea where exactly the crime has been happened and which crime has the highest intensity.

By implementing the Line chart, we have answered to the final question by comparing the trends in all neighborhoods at Baltimore. By implementing the interaction, it is so clear that we can easily see the trends changing over the years.

One thing would like to improve is, to bring the Map and line chart side by side.

Here by, we are confident that we have answered and implemented the pipeline without missing and accomplished the project successfully. We would like to say that user can understand our data easily and will play the interactions that we have created.