Exploratory Data Analysis and Spatial Visualization of Crime Data in New York City

Rohith Reddy Depa
2295660
Dept. of Computer Science
rdepa@cougarnet.uh.edu

Yashwanth Chowdary Pavuluri
2183774
Dept. of Computer Science
ypavulur@cougarnet.uh.edu

Naga Badra Kali Mylavarapu
2183705
Dept. of Computer Science
nmylavar@cougarnet.uh.edu

***Abstract*— Crime data in New York is significant importance for various reasons. It plays a crucial role in understanding and addressing public safety concerns. This geospatial analysis project aims to address the challenge of presenting crime data effectively by employing advanced analytical techniques and visualization methodologies within the context of New York City. Utilizing a comprehensive dataset sourced from New York City's crime records and operating with GeoPandas, Folium, and Plotly, this study explores the spatial distribution of criminal incidents across the boroughs interactively. This implementation emphasizes the broader implications of geospatial data analysis, extending beyond law enforcement to demonstrate its potential applications in traffic management, logistics, and educational accessibility.**

Keywords: - Crime Data, New York, Boroughs, Geospatial Data Analysis, and Visualization.

1. INTRODUCTION

The objective of this project is to conduct a comprehensive analysis and visualization of the geospatial distribution of crimes across New York City's boroughs, utilizing the rich dataset provided by the New York City crime/offense records. The primary focus lies in leveraging GeoPandas to create informative geoplots, illuminating the spatial patterns of criminal incidents across different boroughs. Beyond visualization, this project endeavors to delve deeper, aiming to identify and address specific challenges and issues within the dataset.

Understanding the geographical distribution of crimes is pivotal in shaping effective law enforcement strategies and community-oriented interventions. By harnessing the power of geospatial analysis, this study aims to unravel spatial nuances, identify crime hotspots, and discern temporal trends within the diverse neighborhoods of New York City. The visual representation of crime incidents through geoplots offers a dynamic and intuitive means to comprehend the geographic disparities in criminal activities, providing actionable insights for law enforcement and urban planners alike.

The amalgamation of geospatial visualization techniques, critical analysis, and data refinement pursuits in this project aims not only to uncover spatial crime patterns but also to contribute towards creating more robust and reliable crime analysis frameworks beneficial for policymakers, law enforcement agencies, and community stakeholders.

2. RELATED WORK

GeoPandas serves as a dynamic bridge between pandas' data manipulation prowess and geospatial data handling, enabling seamless operations on shapefiles, *GeoJSON* files, and more, with an intuitive pandas-like syntax. Its extensions to pandas, like *GeoDataFrame.area*, offer robust spatial analytics by computing geometry areas, *GeoDataFrame.boundary* unveils polygon perimeters, while *gdf.centroid* precisely pinpoints central locations. These features are foundational, empowering users to gauge spatial relationships, extract boundary insights, and quantify geographical areas, fueling detailed analyses and visualization within Python.

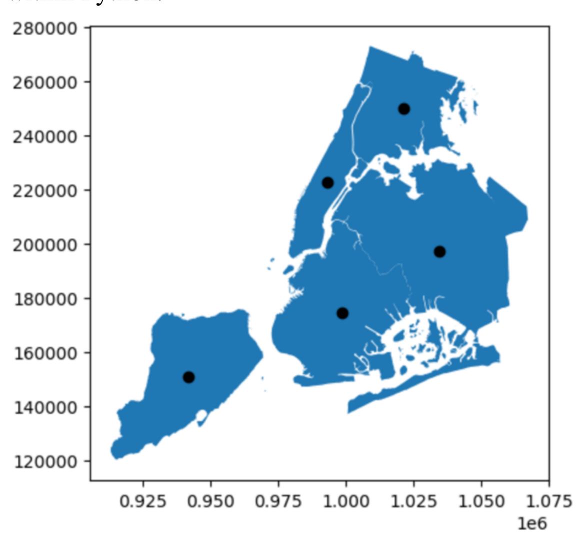

Figure 2.1 Geometric Centroid Of New York Broughs

The process of spatial exploration initiates by computing centroids' distances within a *GeoDataFrame*. Leveraging GeoPandas distance method, it calculates and stores distances between the initial centroid and others, shedding light on spatial distances across boroughs. This seamless integration of spatial attributes within a familiar pandas framework unlocks statistical insights, exemplified by the *mean()* method effortlessly deriving the average distance between centroids, enriching spatial comprehension within the dataset.

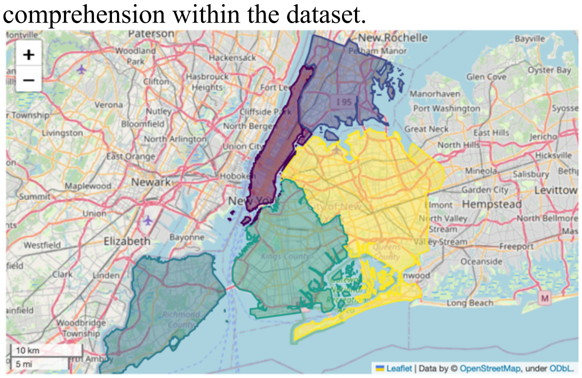

Figure 2.2 Interactive plot of broughs

GeoPandas visualization capabilities are pivotal. Using *gdf.plot("area", legend=True)* vividly portrays geometries color-coded by their respective area values, facilitating quick comprehension of variations. Employing *gdf.explore("area", legend=False)* generates interactive maps, enabling dynamic exploration of spatial data. Shifting to centroid representations *(gdf = gdf.set_geometry("centroid"))* brings a new perspective, transforming data into point geometries and allowing different insights into spatial distributions.

Delving deeper, GeoPandas excels in spatial analyses. It effortlessly evaluates intersections between Brooklyn's geometry and buffered boroughs, revealing proximities within a 10,000 feet radius. Additionally, it discerns which buffered centroids fall entirely within their original boroughs' boundaries. This analytical prowess, culminating in visual representations confirming the spatial relationships, underscores GeoPandas' capacity for intricate spatial analyses, facilitating comprehensive geographical assessments.

3. MOTIVATION

The fundamental drive behind this project is to unravel the sophisticated geo-communication framework utilized by law enforcement agencies, particularly in comprehending their swift mobilization strategies following offenses in specific locations. By dissecting the intricate methods employed by police departments when responding to incidents based on geographic cues, this study seeks to unveil the operational intricacies and efficiency of their tactical approaches.

Moreover, transcending its immediate implications for law enforcement, the insights derived from this analysis possess extensive applicability. These spatial analysis and geo-communication strategies find resonance in tackling challenges beyond crime, playing pivotal roles in mitigating traffic congestion and optimizing logistical routes within E-commerce delivery systems. Understanding and leveraging these methods have the potential to markedly enhance logistical efficiency and resource allocation across diverse industries.

Furthermore, this project endeavors to extend its impact to the educational sphere. By harnessing spatial analysis techniques, it aims to pinpoint and support hard-to-reach students within educational institutions. Identifying underserved areas or student populations can facilitate targeted interventions, ultimately enhancing educational accessibility and fostering equity within the educational landscape.

In essence, this project's motivation encompasses a broader exploration of the geo-communication systems employed by law enforcement agencies, seeking to extrapolate its insights for optimized responses, logistical efficiency improvements, and equitable educational outreach strategies spanning multiple sectors.

4. PROPOSED SYSTEM

After acquiring the dataset from the New York Police Department, our focus shifts to conducting insightful visualizations that will serve as pivotal outcomes for this project. Upon extraction, our initial step involves analyzing the dataset to discern key factors essential for our objectives. To achieve this, we systematically list and display the dataset items, meticulously examining the contents within the table. Given the nature of our work with crime data, a primary aim is to identify prevalent occurrences within New York State. Leveraging the `value_counts` function becomes instrumental in this process, allowing us to swiftly pinpoint and quantify the frequency of various case occurrences across the state. This initial exploration lays the groundwork for a more detailed and targeted analysis of crime patterns within New York City's boroughs.

Going forward in the model we will, we will implement the concept of GEOPANDAS which is the very core part of this project. We create map of New York City boroughs using geographical data. By loading datasets of world countries and NYC boroughs. It uses Geoplots to visually display the boroughs on a map, employing the *Albers Equal Area** projection to accurately represent their sizes. The Albers Equal Area projection is a type of map projection that aims to preserve the relative sizes of areas on the map. Albers Equal Area projection seeks to maintain accurate proportions of areas across the entire map. This makes it suitable for thematic mapping where preserving the accurate representation of geographic areas is important.

We use *Plotly* Express to create an interactive *choropleth* map of New York City boroughs. It starts by reading datasets for world countries and NYC boroughs. The *Plotly Express* library is then used to generate the *choropleth* map. This concept takes the boroughs' *GeoDataFrame* and specifies parameters such as the geometry, title, locations, and color scale for the map.

The *choropleth* map is configured to display the spatial distribution of boroughs based on an index, using the '*Viridis*' color scale. The map is centered on New York City with a specified latitude and longitude. By using this map, we get a brief understanding by hovering over boroughs to view additional information, and gain insights into the geographical distribution of data across New York City.

After involvingthe fundamental data analysis on the data provided, we observed the *age_value* data is a significant factor impacting crime the data and this needs a careful observation to locate the important locations for surveillance.

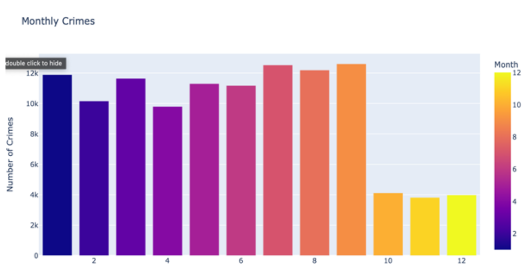

Figure 3.1: Monthly Crimes in New York

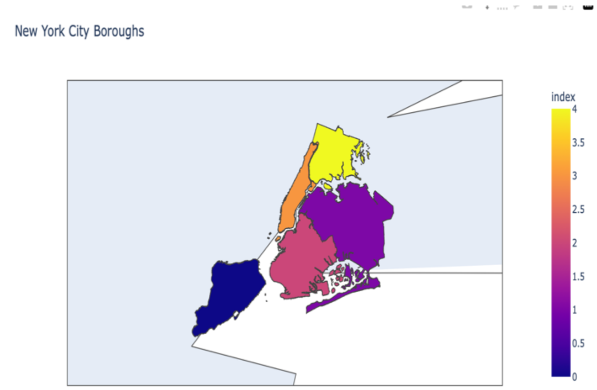

Figure 3.2: Different Boroughs In New York

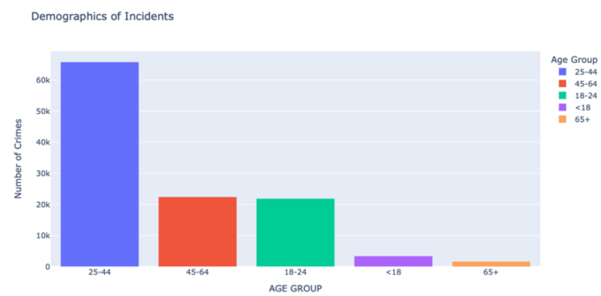

Figure 3.3: Demographics Of Incidents Based on Age Group

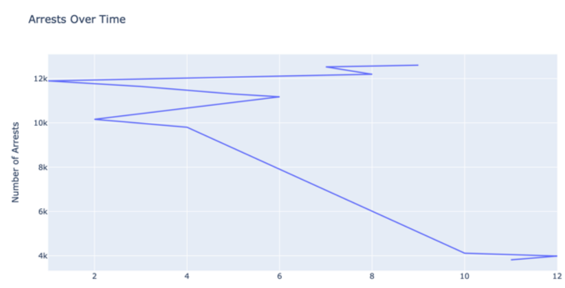

Figure 3.4: Arrests Over Time

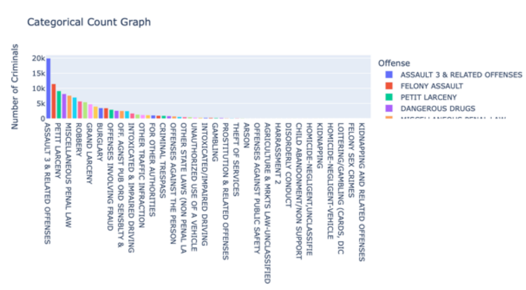

Figure 3.5: Categorical Count Graph

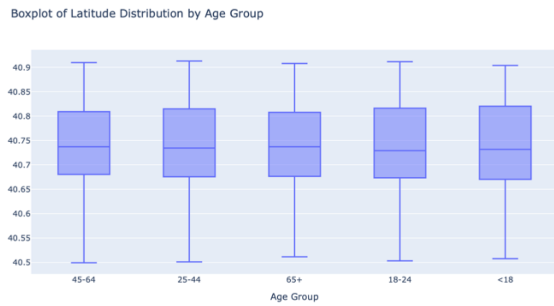

Figure 3.6: Boxplot Of Latitude Distribution by Age Group

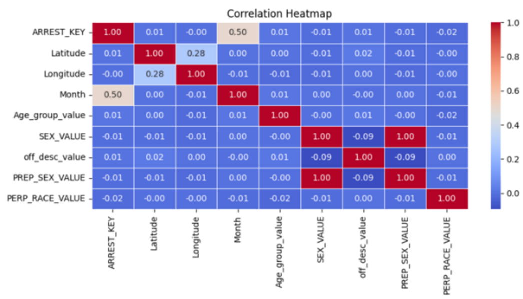

Figure 3.7: Correlation Heatmap

5. RESULTS

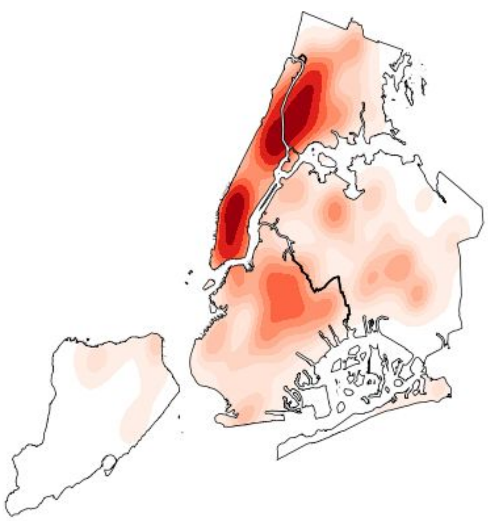

Figure 5.1: Static Geo-Plot Based on Gender

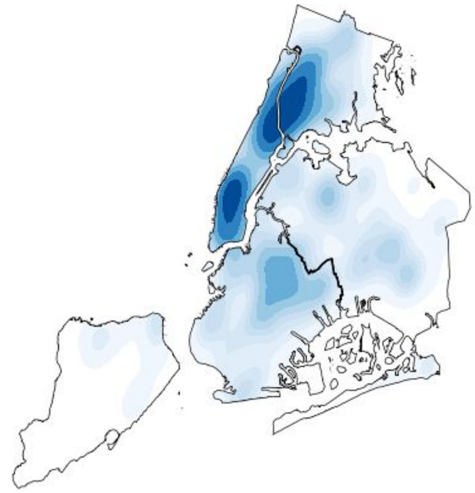

Figure 5.2: Static Geo-Plot Based on Age

Figure 5.3: Interactive Map Displaying Crime Data using Co-ordinates.

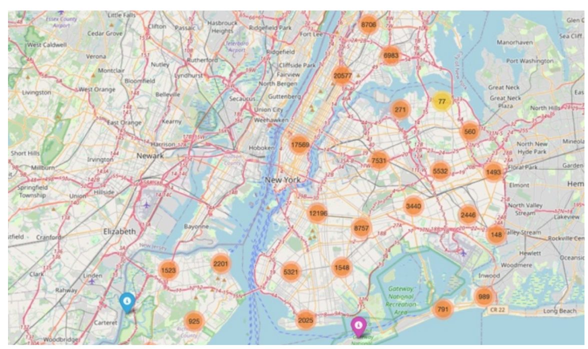

Figure 5.4.a: Interactive Map using Folium Representing Crime Date Based on Age Group

Figure 5.4.b: As we zoom in, we can observe the incident performed by specific group.

6. LIMITATIONS AND FUTHER GOALS

6.1 Limitations

The limitations of the geospatial analysis project include potential biases and inaccuracies in the dataset affecting crime pattern interpretations, granularity constraints limiting fine-scale spatial analysis, temporal restrictions hindering comprehensive trend exploration, a narrow focus on crime data overlooking broader contextual factors, and a heat-map attribute limitation to gender. Mostly these practices will take part using static images, we overcame this by using different libraries like plotly, shapely and Folium. We also important concepts of *chloropleth* as part of plotly which helped in converging the data points projection while displaying interactive maps.

6.2 Future Work

As we implemented the specific location based on *age_group*, we intend to develop this by displaying the individual details and records present under their history. This feature helps the police or other investigators to get the detailed knowledge on the incident and leads to predictive modeling or develop expected outcomes.

Extending geospatial analysis methodologies beyond crime to explore applications in disaster management, or environmental planning for broader societal benefits.

7. TEAM MEMBERS CONTRIBUTION

As we are a team of three, we divided the teaks accordingly. During initial analysis we observed *GeoSpatial Analysis* is a unique way of representing the data using visualization methods.

Badri played a huge role in data gathering and processing for this project. Her efforts in sourcing and organizing the crime dataset from various sources provided the foundational data for our analysis.

Yaswanth expertise and decision-making skills were instrumental in determining the appropriate observations and visualizations to effectively showcase our findings. His insights guided the selection of visual representations using Folium, Plotly and Shapely.

Rohith took charge of the consolidating the complete efforts of visualization aspect in our project. His skillful execution in creating informative and visually compelling representations facilitated a better understanding of our geospatial analysis.

8. CONCLUSION

In culmination, this project embarked exploration of the geospatial distribution of crime incidents across New York City, leveraging analytical methodologies and visualization techniques. Through meticulous analysis, the study illuminated spatial nuances, identified crime hotspots, and discerned temporal trends within the diverse boroughs. The visualization of crime patterns through Geoplots provided a dynamic depiction of geographical disparities in criminal activities, offering valuable insights for law enforcement strategies and protection in important locations. Primary takeaway from this project is, how the police department communication system works when alert pop-ups in particular location. This also helps them strategically to place their sentry, patrol positions and offices effectively. Beyond its application in law enforcement, the versatility of geospatial analysis emerged, showcasing potential applications in traffic management, logistics, and educational accessibility. The project's broader impact lies in its potential to inform evidence-based decision-making, improve resource allocation, and foster proactive measures across various sectors, transcending the realm of crime analysis.

9. REFERENCES

[1] Anselin, Luc, and John O'Loughlin. 1992. "Geography of international conflict and cooperation: spatial dependence and regional context in Africa." *The New Geopolitics*, 39–75.

[2] Baddeley, Adrian, Ege Rubak, and Rolf Turner. 2016. *Spatial Point Patterns, Methodology and Applications with r*. Florida: CRC Press.

[3] Bailey, Trevor C., and Anthony C. Gatrell. 1995. *Interactive Spatial Data Analysis*. England: Prentice Hall.

[4] Dykes, J. A., and D. J. Unwin. 2001. "Maps of the Census: A Rough Guide." *In Case Studies of Visualization in the Social Sciences: Technical Report* 43 (43): 29–54. http://www.agocg.ac.uk/reports/visual/casestud/dykes/dykes.pdf.

[5] O'Sullivan, David, and David Unwin. 2010. *Geographic Information Analysis*. New Jersey, USA: Wiley.

[6] Pebesma, Edzer. 2018. "Simple Features for R: Standardized Support for Spatial Vector Data." *The R Journal* 10 (1): 439–46. https://doi.org/10.32614/RJ-2018-009.

[7] https://educationcommission.org/updates/five-ways-geospatial-analysis-can-help-visualize-and-solve-some-of-educations-biggest-challenges/

[8] https://geopandas.org/en/stable/docs/user_guide/projections.html

[9] https://www.google.com/search?q=WGS84&rlz=1C1GCEA_enUS1081US1082&oq=WGS84&gs_lcrp=EgZjaHJvbWUyBggAEEUYOdIBBzM2NGowajmoAgCwAgA&sourceid=chrome&ie=UTF-8

[10] https://geopandas.org/en/v0.10.2/gettingstarted/introduction.html#Concepts

[11] https://towardsdatascience.com/interactive-geographical-maps-with-geopandas-4586a9d7cc10