CS 732: Data Visualization Assignment 3 Report

Munagala Kalyan Ram
IMT2021023
IIIT Bangalore
Kalyanram.Munagala@iiitb.ac.in

Vikas Kalyanapuram
IMT2021040
IIIT Bangalore
Vikas.Kalyanapuram@iiitb.ac.in

Ramsai Koushik
IMT2021072
IIIT Bangalore
Ram.Polisetti@iiitb.ac.in

Please find the link [8] to the A1 report attached in the references section.

I. DATASET

In this assignment, we worked on the following 2 datasets:

1) Crime Data of Los Angeles [1], which was the same dataset as used in A1. The given dataset has details of crimes that have taken place in Los Angeles from the January 2020 to beginning of December 2023. There are a total of 28 data fields where each field covers different aspects of the crime that took place. The major data fields that were used for the visualizations are [2]:
   - Date OCC : The date of occurrence of the crime.
   - Area Name : The 21 Geographic areas of Los Angeles.
   - LAT, LON : Latitude and Longitude values of the crime.
2) Arrest Data of Los Angeles [3], which supplemented the first dataset. This dataset had details on the arrests that had taken place in Los Angeles from January 2020 to beginning of December 2023. There are a total of 25 data fields where each field covers different aspects of the arrest. The major data fields that were used for the visualizations are: [4]
   - Booking Date: The date on which the person was booked in detention facility
   - Area Name : The 21 Geographic areas of Los Angeles.

II. DATA PRE-PROCESSING

The dataset was loaded into a python notebook using the pandas library. The preprocessing that was done in A1 was kept for this assignment. In addition to that, we also carried out another step of pre-processing since our requirements were different in this assignment: We dropped all rows that had 0° longitude and 0° latitude. The dataset description had mentioned that this implies that the location of the crime was unknown. Since these rows were few in number and since our analysis this time focused on the location of the crime, we decided to drop these rows. How the data was transformed during the feedback loop will be explained in following sections.

III. REQUIREMENT OF DATA ANALYSIS

We want to confirm that the location (within Los Angeles) and time of the year both play a very important role in determing the count of crimes committed. In order to confirm this, we followed the following series of steps:

1) We initially filtered crimes recorded within each year (2020-2023).
2) Subsequently, we determined the number of crimes occurring in each area of Los Angeles for each of the 12 months in the respective year.
3) Hence, we are left with the following: For each year, for each area within Los Angeles, the count of crimes during the 12 months of the year.
4) Finally we plotted box plots [Fig 1] for each area for each individual year:

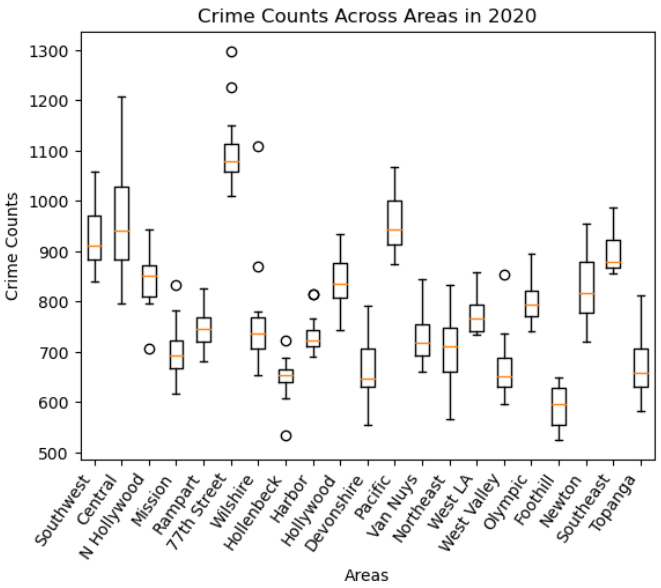

Fig. 1. Box Plot for Count of crimes across months of the year, for 2020

Note that we have attached the box plots only for the year 2020 and 2021. You can find the box plots for the other years here : (Box Plots Images Folder)
Here as well as in the box plots of all the years, we can see that there are several areas which have a discrepant number of crimes around the median. Namely in some areas, most of the data is lying above the median. Hence,

we can conclude that in a given year, different areas have a different number of crimes happen at different times of the year.

IV. TASKS

Through the visualizations, the users must be able to perform the following tasks:

- **T1** : Identify temporal periods when crime counts surge in different Los Angeles areas. Furthermore, in case of a notable crime spike in a specific area during a specific time frame, we aim to investigate the spatial distribution of crimes within that area.
- **T2**: How did the surge in criminal activities across the regions identified impact the quantity of arrests within those specific areas?

V. ANALYTICS WORKFLOW

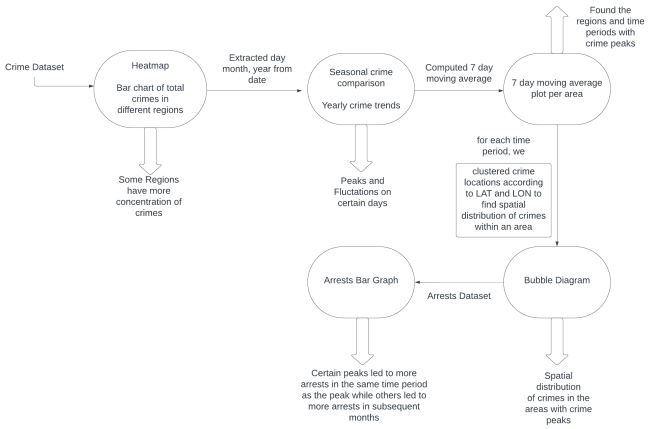

Fig. 2. Analytics Workflow

The [Fig 2] represents the analytics workflow that was used. Each circle represents the set of visualizations that were made in that iteration and the large arrow out of each circle denotes the inferences that were made. The thin arrow out of each circle denotes the data transformation that was performed during the feedback loop and a box denotes that an ML algorithm was used.

VI. VISUALIZATIONS

All the visualizations have been created using Python libraries, namely Matplotlib, Seaborn, and Folium except for the final visualization, which was created using tableau.

Our analysis of temporal crime patterns across different areas began with a series of crucial visualizations. We initiated our exploration by plotting a bar chart illustrating the distribution of crime counts across various areas.

A. Bar Chart - Distribution of Crime Counts Across Areas

This bar chart[Fig 3] depicts the total number of reported crimes across different areas. Each bar represents a distinct area, showcasing the count of reported crimes in that specific area. The height of each bar corresponds to the total count of crimes, providing a visual comparison of crime frequency among various areas.

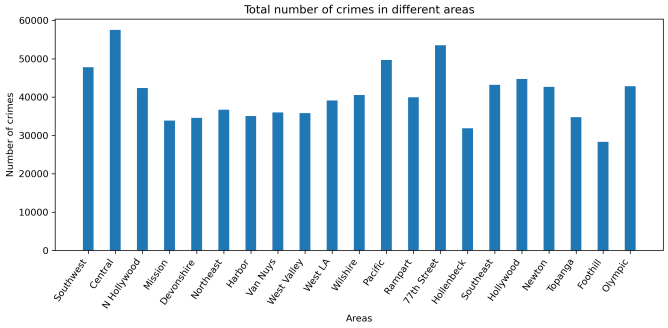

Fig. 3. Bar Chart (Total number of crimes in different areas)

B. Heatmap of Crime Incidents Across Los Angeles

This heatmap[Fig 4] visualizes the spatial distribution and intensity of reported crime incidents across different areas. It showcases the geographical concentration of crime incidents, offering insights into crime hotspots or patterns based on location coordinates.

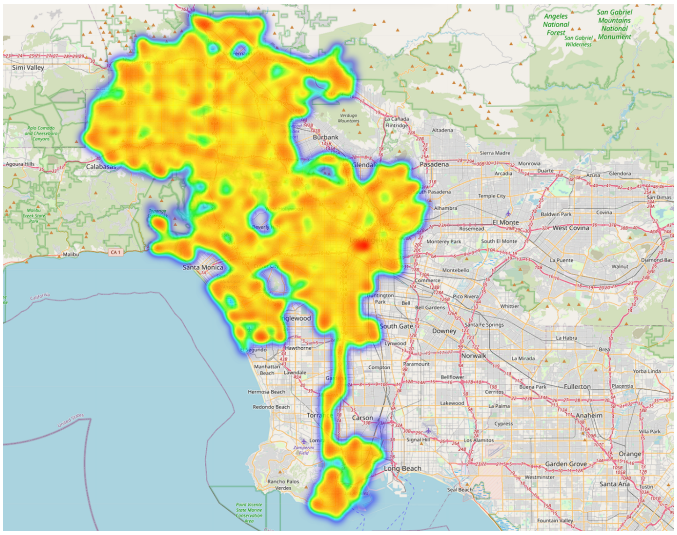

Fig. 4. Heatmap of Crime Incidents Across Los Angeles

Feedback Loop 1: The heatmap[Fig 4] analysis across different areas in Los Angeles reveals contrasts in reported crime incidents. While certain regions exhibit minimal or no reported crimes, others distinctly show a higher concentration of incidents. However, this analysis primarily focuses on spatial distribution, lacking insights into the temporal aspect of these occurrences within specific areas. This emphasizes the need for a more detailed temporal breakdown to comprehend

when these incidents transpired in the identified high-density crime regions. Incorporating temporal data into the analysis will enhance our understanding of the time-specific trends within these areas, fostering a more comprehensive evaluation of crime patterns. To achieve this, we transformed the dataset by adding three essential columns—month, year, and day (extracted from 'DATE OCC' column).Additionally, we added a 'season' column(extracted from 'month' column), categorizing incidents into Fall, Spring, Summer, and Winter.

C. Seasonal Crime Comparison Across Areas

This visualization presents small multiples[Fig. 5] showcasing the comparison of seasonal crime incidents across different areas in Los Angeles. The plot comprises four subplots, each subplot represents a specific season—Fall, Spring, Summer, and Winter—depicting the count of reported crimes within various areas during these seasons across all years. The y-axis represents the number of crimes reported, while the x-axis displays distinct areas. By visualizing crime occurrences across seasons for each area, it allows for an assessment of seasonal variations in crime density across different areas of the city.

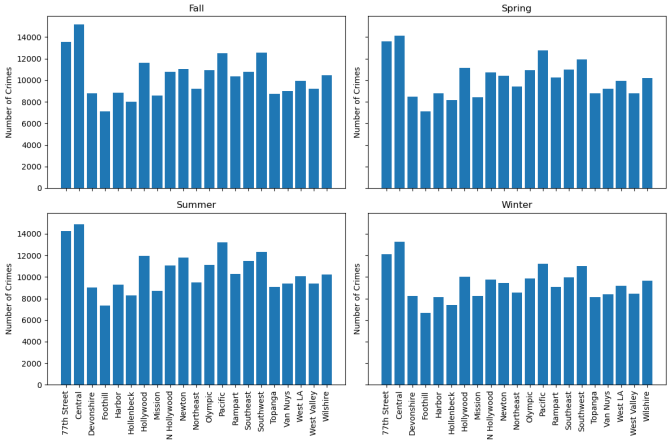

Fig. 5. Seasonal Crime Comparison Across Areas

D. Yearly Crime Trends Over Days

This visualization[Fig 6] presents the yearly trends of reported crime incidents from 2020 to 2023, demonstrating how the count of reported crimes fluctuates over the days of the respective years. The plot comprises four subplots, each representing a specific year, showcasing the variation in daily crime occurrences. The x-axis represents the days of the year, the y-axis denotes the count of reported crimes, and each subplot's title specifies the examined year.

Feedback Loop 2: "Seasonal Crime Trends" conveys comparative crime rates for each season across different areas. Similarly, "Daily Crime Counts" conveys the frequency and fluctuations in crime occurrences over time. However, due to clustered representations and erratic spikes in the visualizations, inferring clear patterns becomes challenging. To

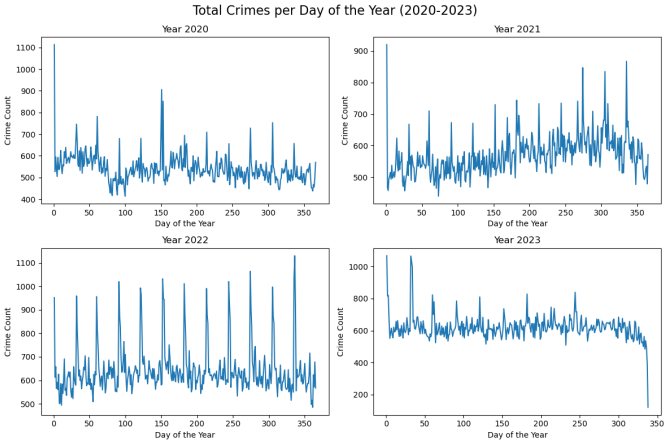

Fig. 6. Yearly Crime Trends Over Days

overcome these issues and enhance interpretability, we introduce the utilization of moving averages. This method helps in smoothing the data, minimizing irregularities and facilitating a better understanding of crime trends over time. To visualize the moving averages, we transformed the data by adding a column for featuring a 7-day moving average for crime incidents.

E. Moving Average - Small Multiples of Crimes Across Months Per Area

The Moving Average plots visualizes the 7 day average of crimes (Temporal Transformation of data) in each area for each year as a line chart where the X-axis represents the months and the Y-axis represents the number of crimes. The number of crimes on each day is also plotted as a scatter plot, i.e both the scatter plot and the line chart are in the same visualization. The plots in [Fig 7] represent the crimes in the year 2020. The moving average plots for the year 2021, 2022 and 2023 can be found in the Moving Averages Image folder . Each small multiple corresponds to the number of crimes in an "Area Name".

Feedback Loop 3 : The visualization in [Fig 7] helps in better understanding the trends of the crimes where it smooths out short-term fluctuations. We were able to identify the time periods where the number of crimes suddenly increased in each "Area Name". However, this visualization does not provide details on the regions in the given "Area Name" that caused the increase in the crimes, i.e the regions with a greater concentration of crimes. To accomplish this, we need to incorporate clustering techniques based on the Geo-spatial data (LAT, LON) of the crimes within the areas with fluctuations.

F. Bubble Map within each Area per Year

In this visualization, we performed a spatial transformation of the data by finding clusters in a given "Area Name" using the **K-Means** unsupervised clustering algorithm on the Geo-spatial data. Each "Area Name" corresponds to an anomaly identified for a time period of a given year and it is represented by a different color (channel). All the clusters of an "Area Name" have the same color. Each cluster within the clusters of

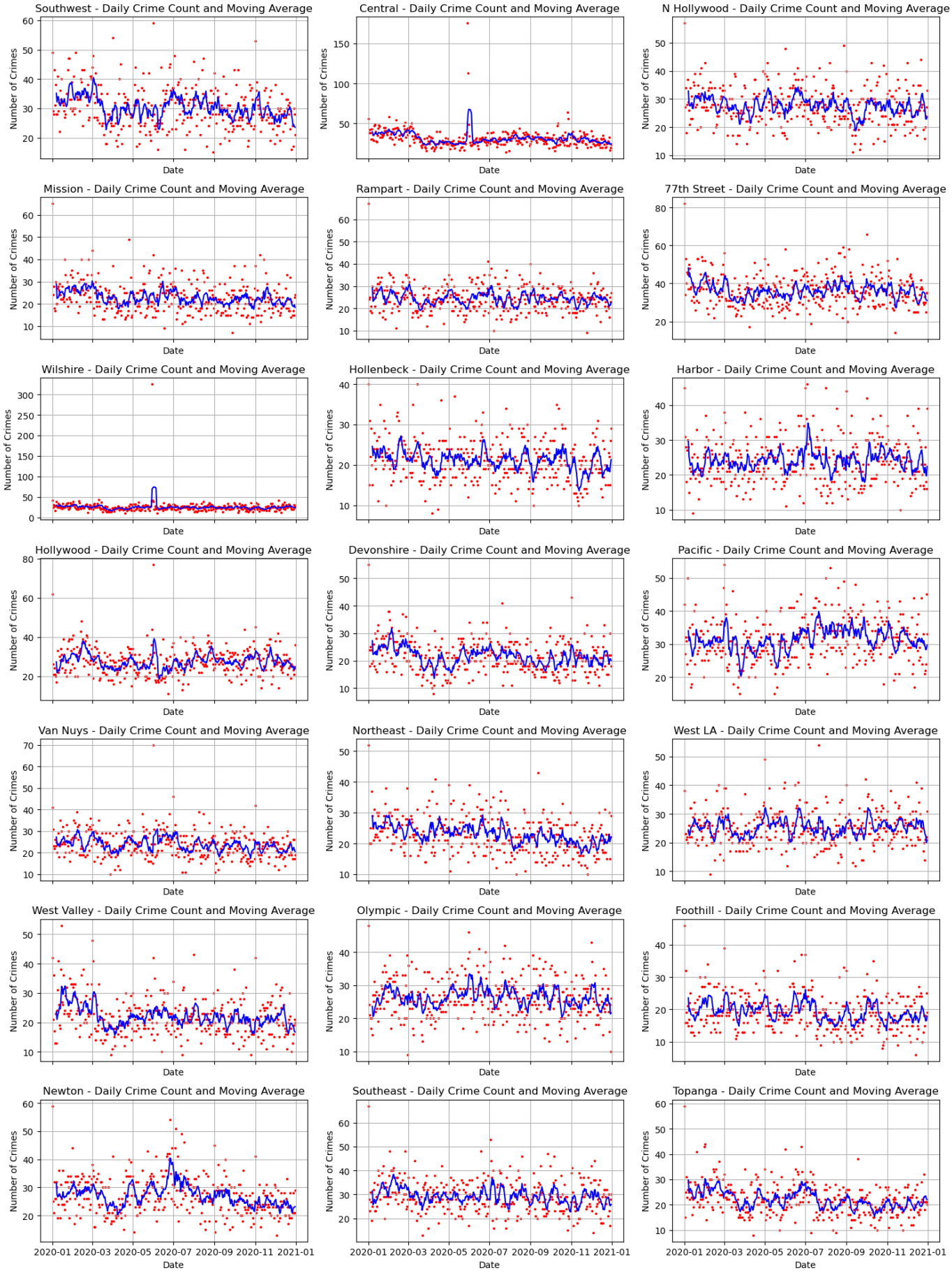

Fig. 7. Moving Average Plot Per Area in Year 2020

an ”Area Name” represents the concentration of crimes which, is represented by a circular mark whose size is proportional to the number of crimes (channel). The marks are plotted on the Los Angeles map using the LAT and LON values of the centroid of each cluster. The plots [Fig 8, Fig 9, Fig 10, Fig 11] represents the clusters computed for the anomalies detected in the year 2020, 2021, 2022, 2023 respectively.

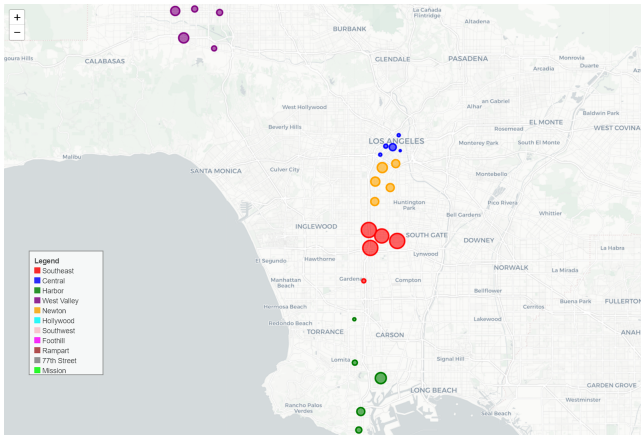

Fig. 8. Clusters of the Area Names for Year 2020

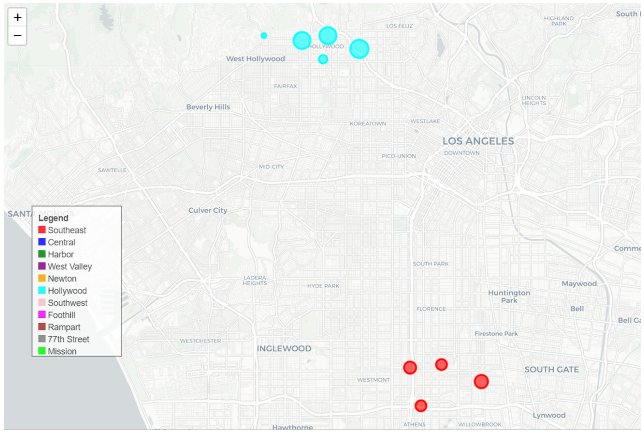

Fig. 9. Clusters of the Area Names for Year 2021

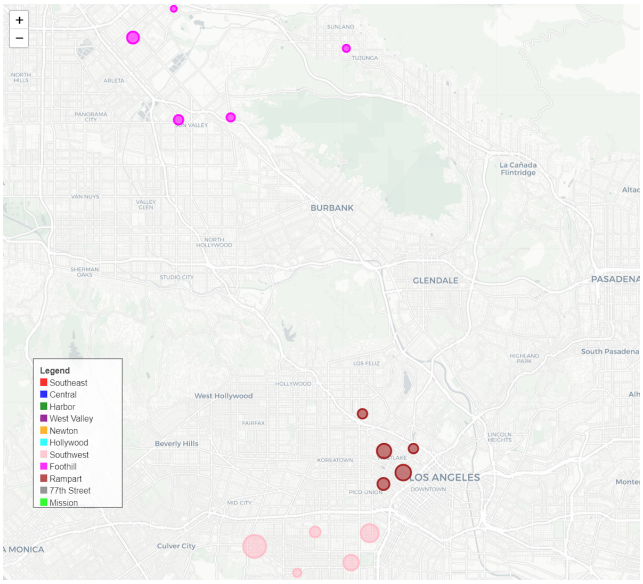

Fig. 10. Clusters of the Area Names for Year 2022

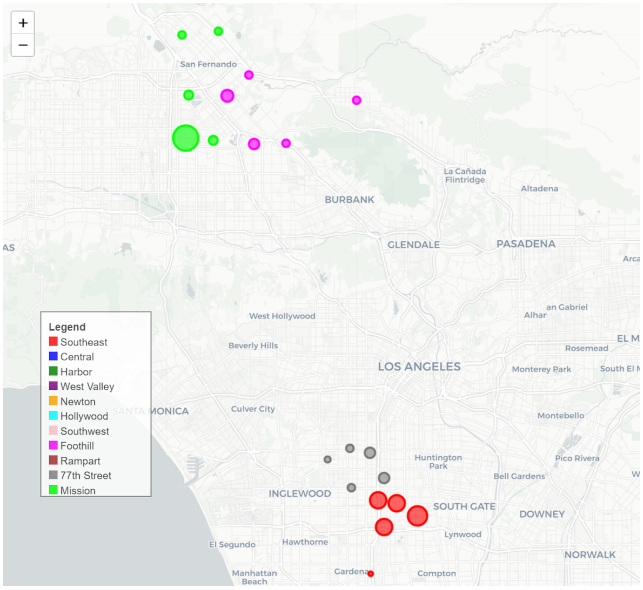

Fig. 11. Clusters of the Area Names for Year 2023

Feedback Loop 4 : We now know the regions where there is a greater concentration of crimes (clusters). We now find the number of arrests that have occurred in these regions for the corresponding time frames by utilizing the Arrest dataset [4].

G. Bar Graph of Arrests in the Detected Area Name

The visualizations in [Fig 12], [Fig 13] are bar graphs where the X-axis is the months of the corresponding year and the Y-axis is the number of arrests that happened in the month.

Note : The visualizations created in python are made as functions that we defined where the user only needs to pass the

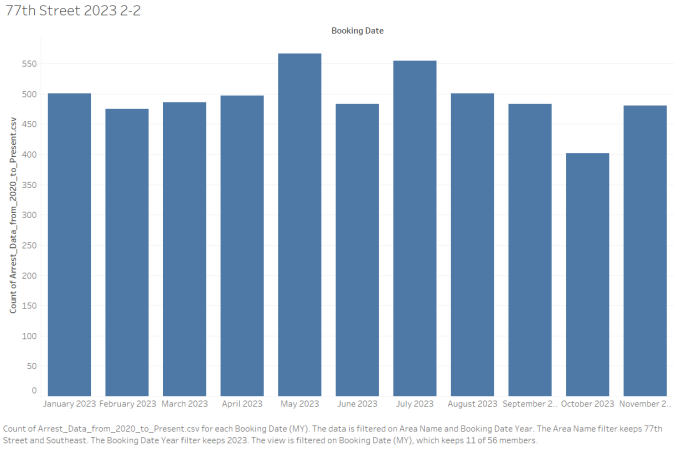

Fig. 12. Arrests in 77th Street in year 2023 February. Generated using Tableau

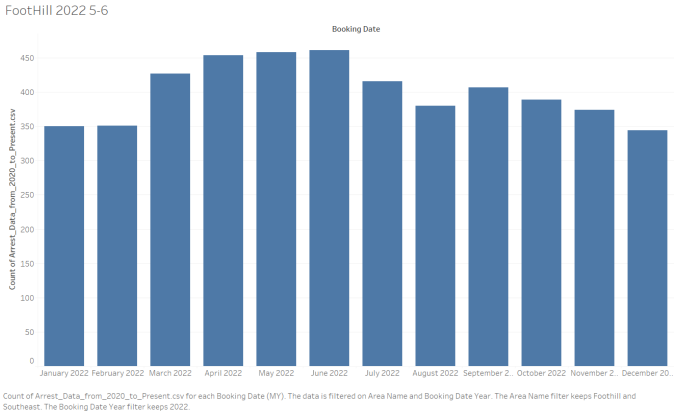

Fig. 13. Arrests in FootHill 2022 May - June. Generated using Tableau

required input parameters to get the expected output (EX : The Moving Average Plot and the Bubble Map can be computed by passing any valid Area Name and Time period). Hence, they can be made user interactive if needed.

VII. Inferences

The following inferences can be made from the visual analytics workflow:

1) From the heatmap [Fig 4] and total crimes bar chart [Fig 3], we can see that the "Area Name" Central has the most count of total crimes as this region is colored red(T1).
2) From the 7-day moving average plots [Fig 7] and the bubble map plots [Fig 8], [Fig 9], [Fig 10], [Fig 11], we can see that the following places at the following periods of time have a peak in the count of crimes(T1):
   a) South-West region of the area "West Valley" during the months January to February(2020).
   b) Central region of the area "Central" during the month of June(2020).
   c) North-West region of the area "Newton" during the months June to July(2020).
   d) East region of the area "Harbor" during the months July to August(2020).
   e) Throughout the "Southeast" area during the months January to March(2020).
   f) Central region of the area "Hollywood" in the month of November(2021).
   g) North-East region of the area "Southeast" in the month of December(2021).
   h) West region of the area "Foothill" during the months May to June(2022).
   i) South-East region of the area "Rampart" during the months March to April(2022).
   j) West region of the area "Southwest" during the months September to October(2022).
   k) South region of the area "Mission" during the months January to February(2023).
   l) West region of the area "Foothill" during the months January to February(2023).
   m) East region of the area "Southeast" during the months January to February(2023).
   n) South-East and North-East region of the area "77th Street" in the month of February(2023).
3) We will now look at the trends observed when comparing the number of crimes with the number of arrests at the peaks (T2) mentioned above :
   a) From the bar graph in [Fig 12] , we see that the number of arrests are less compared to the other months even though the number of crimes are more in this period(February 2023) [Fig 11]. Here we can conclude that the arrests for these crimes have occurred later , i.e in May to July of 2023 as there is an increase in the number of arrests, but we know that there are relatively fewer crimes in this period.
   b) From the bar graph in [Fig 13], we see that the months May to June in the year 2022 have the most number of crimes [Fig 10] and also have the most number of arrests in this time period. Here we can say that the arrests for most of crimes happened in the same time period as the number of arrests reduce in the following months.
   c) From the above 2 inferences, we can conclude that the number of crimes and the number of arrests don't necessary correlate to each other.

VIII. Work Distribution

The tasks detailed in this report were discussed among the team members. All team members contributed equally to deciding the tasks and contributing to the analytics workflow to accomplish the tasks.

References

[1] Crime Data 2020 to Present in Los Angeles: https://catalog.data.gov/dataset/crime-data-from-2020-to-present
[2] Crime Dataset Description : https://data.lacity.org/Public-Safety/Crime-Data-from-2020-to-Present/2nrs-mtv8
[3] Arrest Data 2020 to Present in Los Angeles : https://catalog.data.gov/dataset/arrest-data-from-2020-to-present

[4] Arrest Dataset Description : https://data.lacity.org/Public-Safety/Arrest-Data-from-2020-to-Present/amvf-fr72/about_data
[5] Matplotlib Tutorial : https://matplotlib.org/stable/tutorials/pyplot.html
[6] Seaborn Turtorial : https://seaborn.pydata.org/tutorial.html
[7] Folium Tutorial: https://python-visualization.github.io/folium/latest/getting_started.html
[8] A1 Report : https://iiitbac-my.sharepoint.com/:b:/g/personal/kalyanram_munagala_iiitb_ac_in/ETSo_xeBDB9ImJysPQMVnyABuH0ohHUDJ6R1PhWnHGIFCw?e=CsLAnq