IE 6400 Foundations for Data Analytics Engineering

Project 1 Report

Cleaning and Analyzing Crime Data

Group 1

Aliya

Qiuyi Chen

Siyuan Gao

Xinyue Niu

Yiqian Ning

Introduction

Los Angeles, situated at the heart of California, USA, is a bustling metropolis with over 4 million residents. However, crime remains one of the biggest concerns for both residents and visitors of LA. This project aims to leverage data analytics to uncover insights into crime patterns and trends in Los Angeles.

Our analysis focuses on data inspection, cleaning, exploratory data analysis, and time series forecasting to examine crime rates over time, across different regions, crime types, and in relation to major events and demographic factors. We employ statistical techniques and time series forecasting models to analyze historical crime data and predict future crime rates.

The core dataset for this project spans from January 2020 to September 2023 for the city of Los Angeles, sourced from Data.gov. A comprehensive dataset and description are available at this link: https://catalog.data.gov/dataset/crime-data-from-2020-to-present.

Python is utilized as the primary analytics tool, leveraging libraries and packages such as *Pandas* and *Numpy* for data processing, *Matplotlib*, *Seaborn*, and *Folium* for visualization, and *Statsmodels*, *pmdarima*, *scikit-learn*, and *Prophet* for time series analysis. Through exploratory data analysis and predictive modeling, this project aims to extract meaningful insights from crime data to inform law enforcement policies and public safety strategies for Los Angeles. The results may also be applicable to other major metropolitan areas grappling with crime management.

Data Collection

Upon obtaining the most updated dataset from Data.gov, we proceeded to import it into our preferred data analysis tool, Python with Pandas, for further analysis. This initial step enabled us to prepare the dataset for thorough examination.

To gain insights into the dataset's structure and content, we began by displaying the first five rows. This visual representation allowed us to identify any immediate irregularities or inconsistencies in the data. We conducted a data type check for each column using the 'dtypes' attribute, ensuring that the assigned data types aligned with our expectations. Additionally, we

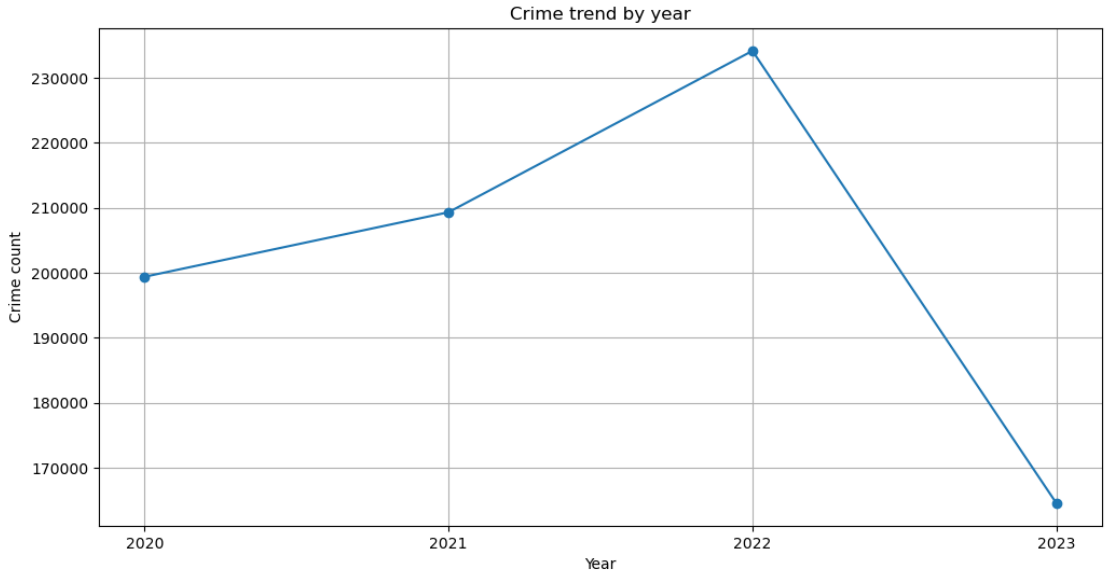
referred to dataset descriptions [Here] to gain a comprehensive understanding of the column meanings, facilitating our data interpretation. Notably, the first column, labeled as 'DR_NO,' was identified as representing unique crime identification numbers.

Our data cleaning process involved identifying missing values, with a record of the total count of missing data in each column. A thorough check for duplicate entries within the 'DR_NO' column revealed the absence of duplicate rows in the dataset. Furthermore, we addressed datetime-related issues by converting the 'DATE OCC' and 'Date Rptd' columns, originally labeled as 'object' data types with datetime values in the format "01/08/2020 12:00:00 AM," into a standardized datetime format while removing the time component. Additionally, we standardized the 'TIME OCC' column, which contained non-standard time formats such as '2230' or '330,' by utilizing the '*pd.to_datetime*' method, resulting in a consistent 24-hour format for time entries. After all, the dataset is stored in to '*clean_data.csv*' file for furfure analysis.

Exploratory Data Analysis

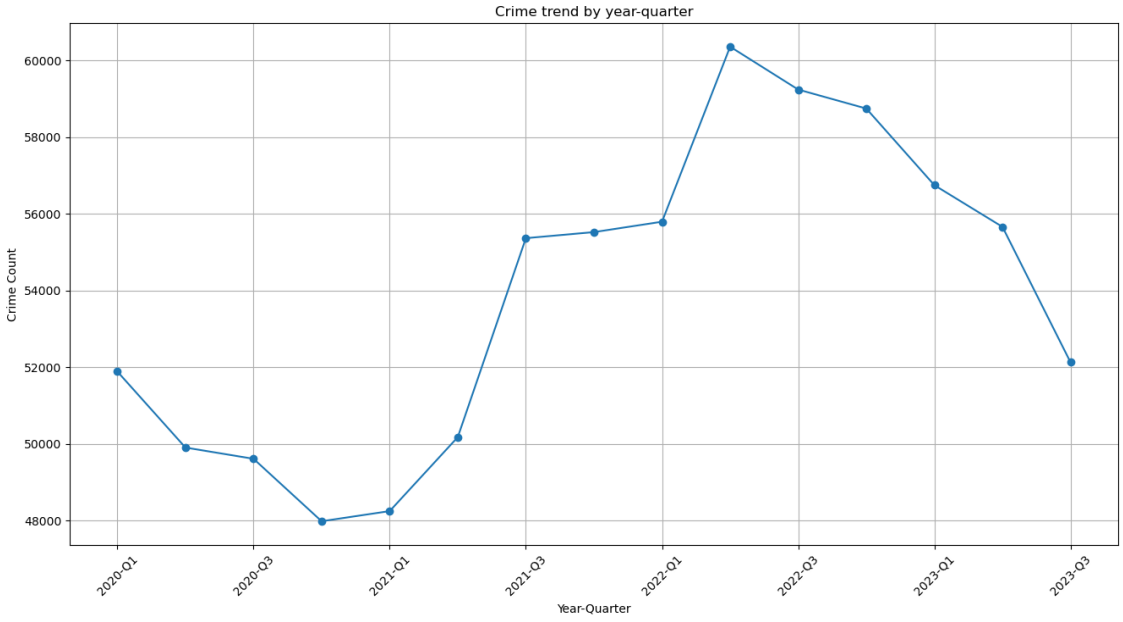
Now, let's delve into the dataset to unveil its insights. Our initial step involves visualizing the overarching crime trends spanning from 2020 to the present. This visual exploration will enable us to observe temporal patterns at various time granularities, ranging from the broader yearly perspective down to quarters and months.

Upon examining the Yearly Trend chart, we observe a notable rise in crime rates from 2020 to 2022, succeeded by a decline in 2023. It is essential to consider that the decrease in 2023 may be linked to incomplete data, which could affect the accuracy of this decline. However, it is essential to highlight the broader trend in crime rates, which signifies an overall upward trajectory over the span of 2020 to 2022. The figures for reported cases further underscore this trend: 2020 saw 199,384 cases, 2021 reported 209,313 cases, 2022 documented 234,144 cases, and 2023 accounted for 164,536 cases.



(Year trend chart)

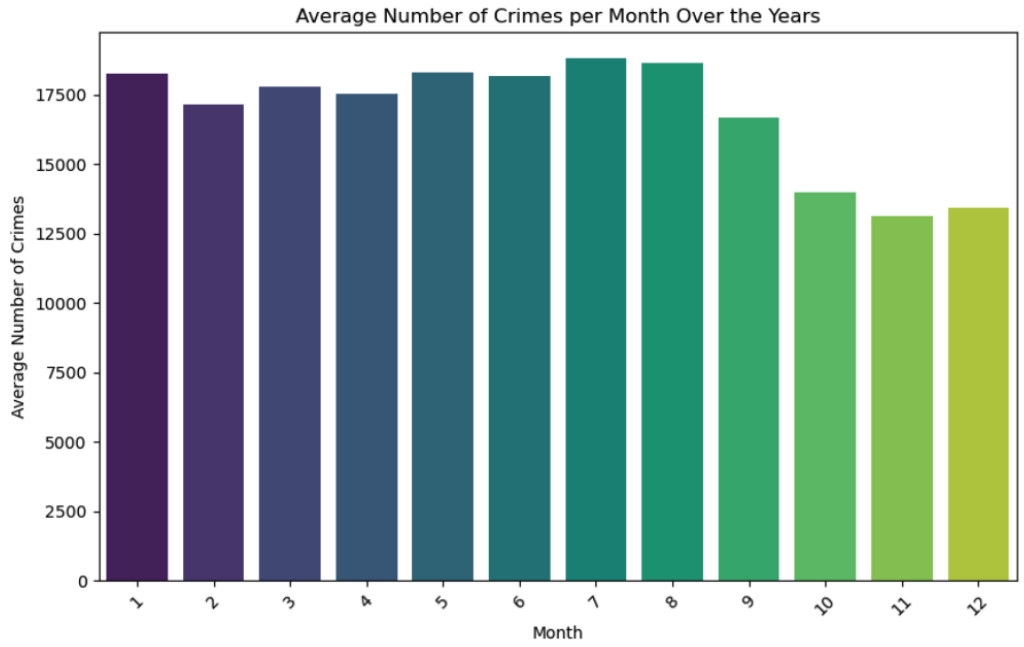
From the Year-Quarter trend chart, we can observe from the chart that the crime rate has decreased significantly in the four quarters of 2020 Q1-2020 Q4, and the crime rate has increased significantly in the five quarters of 2021 Q1-2022 Q2.



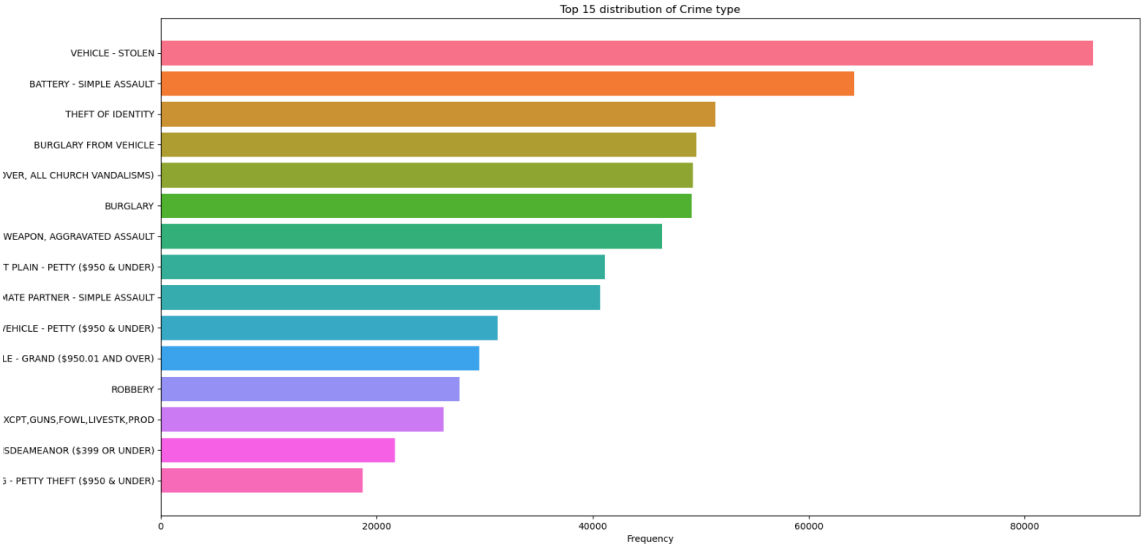
(Year-Quarter trend chart)

By grouping the dataset by month and calculating the average number of crimes per month over the years, we can see the results reveal intriguing insights into the temporal patterns of crime rates. On average, the dataset records the highest number of crimes in July (18,793 cases) and

August (18,617 cases), suggesting a peak in crime during the summer months. Conversely, the lowest average number of crimes is observed in November (13,134 cases) and October (13,996 cases). These findings provide valuable information about the seasonality and variation in crime rates, enabling us to better understand and potentially address the underlying factors contributing to these patterns.

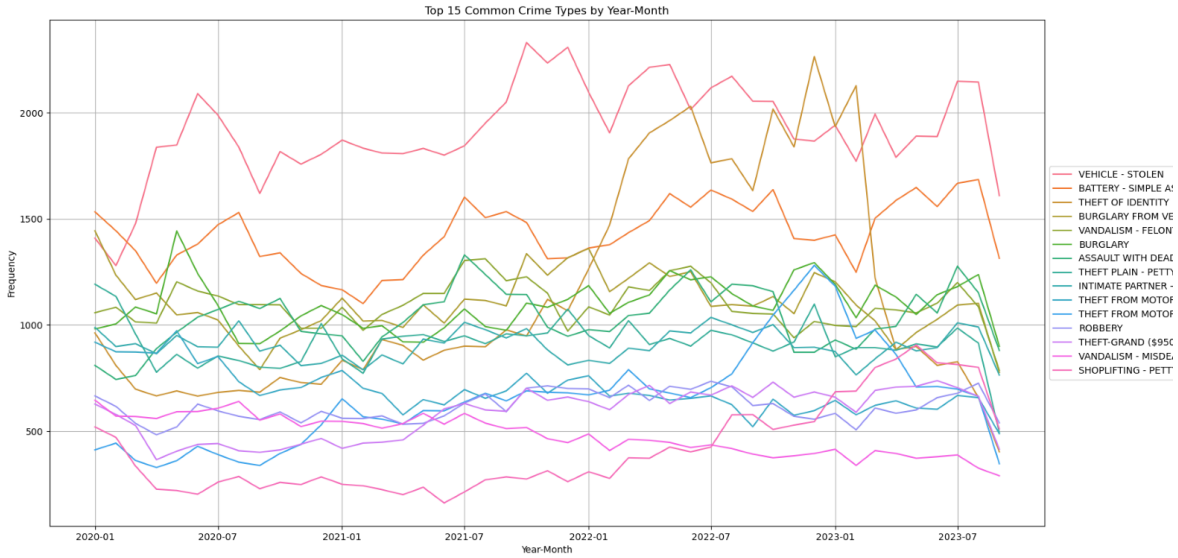


Next, our objective was to identify the most prevalent crime types in the dataset. Recognizing that there was a myriad of crime types, we decided to focus on showcasing the top 15, as displaying them all could result in clutter. To achieve this, we applied the '*.value_counts().head(15)*' method, which enabled us to reveal the fifteen most frequently occurring crime types. It is noteworthy that "Vehicle stolen" emerged as the unequivocal leader in terms of frequency, with a staggering count exceeding 80,000 cases.



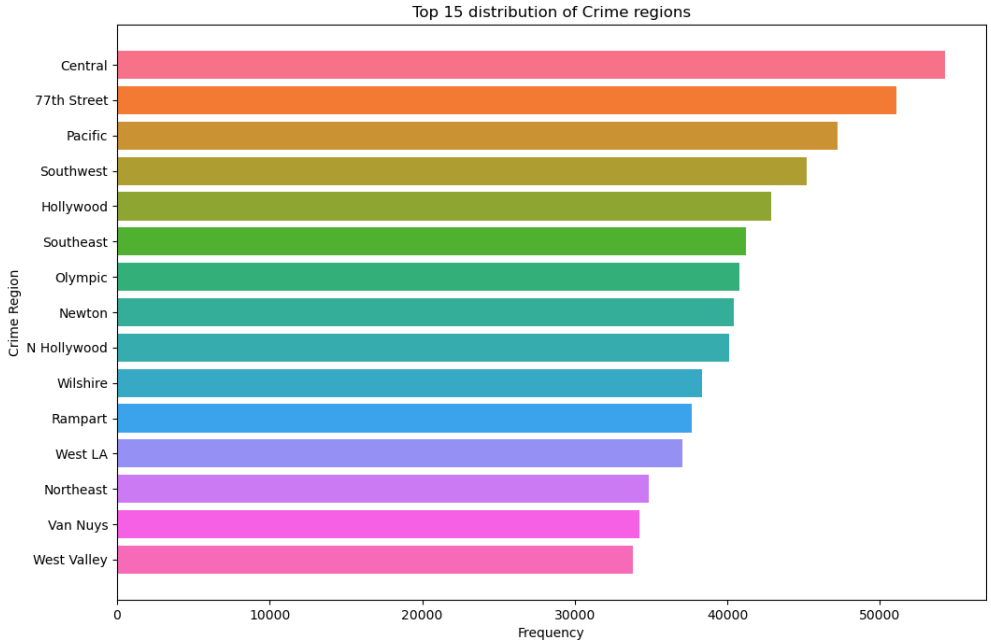
In addition, we also crafted two distinct charts representing the trends of the top 15 common crimes. This allowed us to conduct a comparative analysis of the patterns exhibited by different crime types. Our analysis revealed that not all crime types follow a uniform pattern. While some crimes exhibit a declining trend, such as "Vandalism-Misdemeanor", others demonstrate significant increases at certain points in time. For instance, "Theft of Identity" started to surge notably from January 2022, and "Theft from Motor Vehicle" and "Shoplifting" experienced substantial upswings starting in July 2022. This variation in trends among different crime types underscores the importance of understanding the specific dynamics within each category for effective analysis and response.

(Top 15 common crime type trend by year-month)



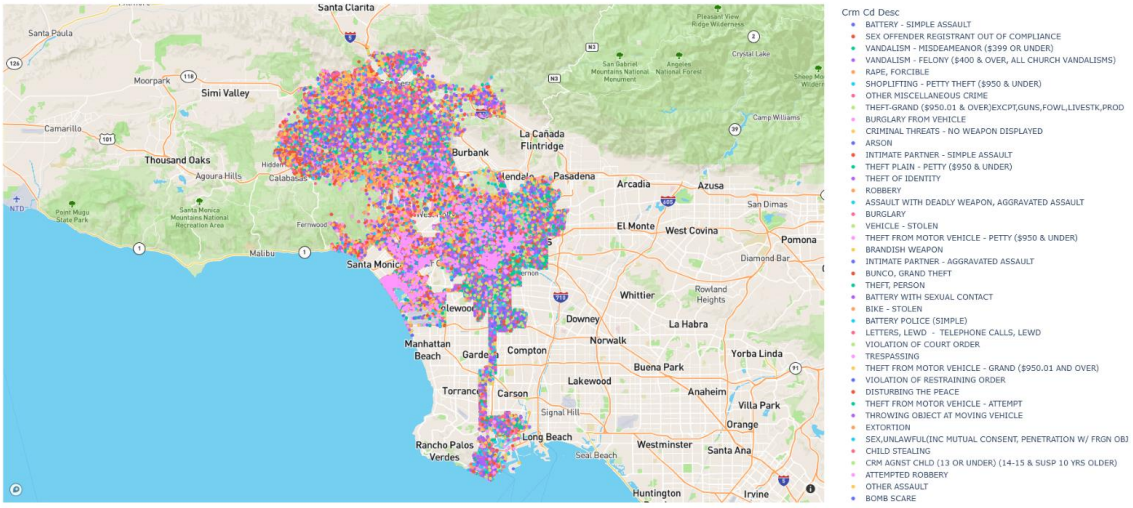
(Top 15 common crime type trend by year-month in one picture)

Our analysis of the top 15 crime regions revealed the Central region as the leading location for reported crimes. However, we observed a relatively even distribution of crimes across different regions rather than an isolated concentration. To directly visualize the crime distribution, we generated heatmaps showing all crimes and specifically "VEHICLE-STOLEN" incidents.

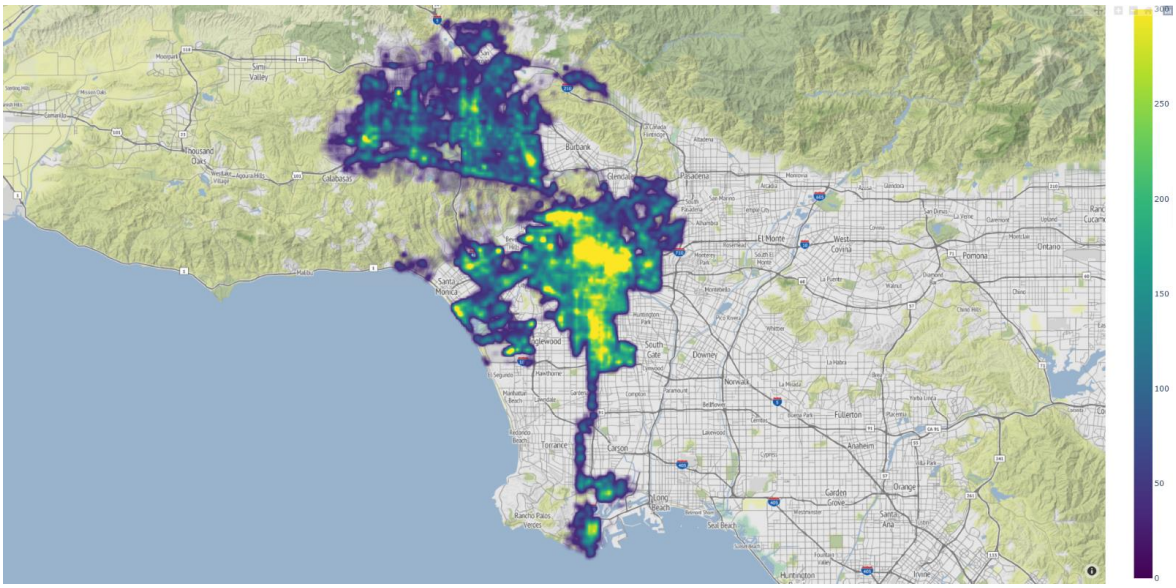


(Top 15 crime regions bar chart)

The heatmaps highlighted a striking "7" shape concentration of crimes in the city center and Central region. This focal point indicates the Central region's significance for multiple crime types, not just an isolated category. In summary, while the Central region leads in total crimes, our analysis found a more widespread distribution rather than a single crime epicenter. The heatmaps provided vital spatial insights into the city-wide spread of criminal activity centered around the vital Central region.



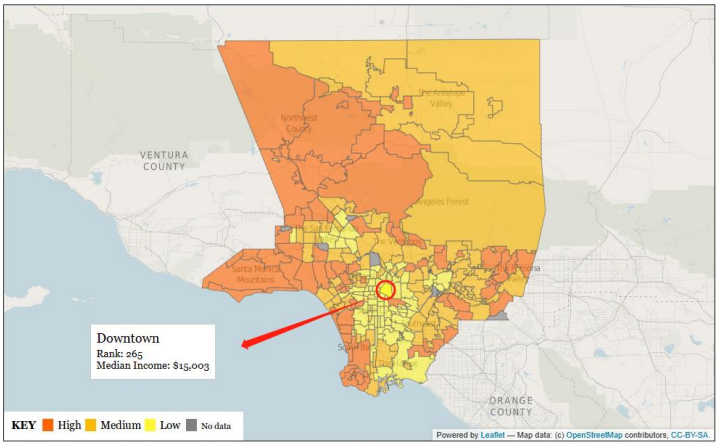
(All crime--splattering)



(All crime--heatmap)

Upon closer examination of the heatmap, we pinpoint areas with the highest incidence of crime, marked by distinct red circles. Specifically, when we narrow down our focus on the heatmap, it becomes evident that the red-circled region corresponds to downtown Los Angeles.

Now that we know the regional distribution of crime rates, we need regional economic data if we want to determine the correlation between crime rates and economic factors. We found the data of the median income of each district in LA to judge the correlation between the crime rate and economic factors. Notably, the downtown area exhibited a ranking of 265 in terms of median income within the broader Los Angeles area. This observation leads us to infer a potential negative correlation between the regional economy and crime rates, as lower median incomes in the downtown region coincide with higher crime frequencies.

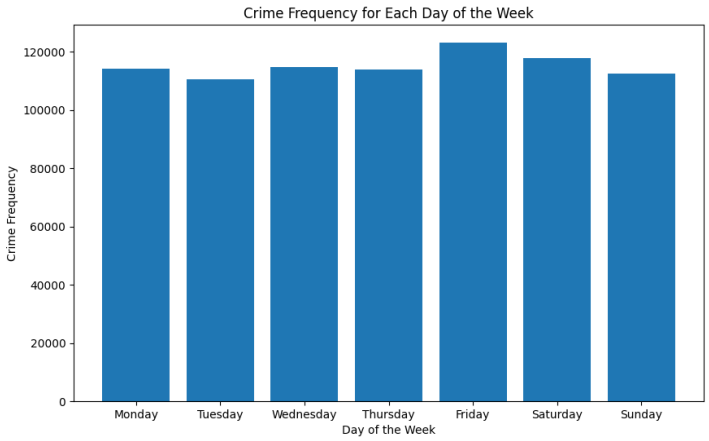


(Data source: https://maps.latimes.com/neighborhoods/income/median/neighborhood/list/)

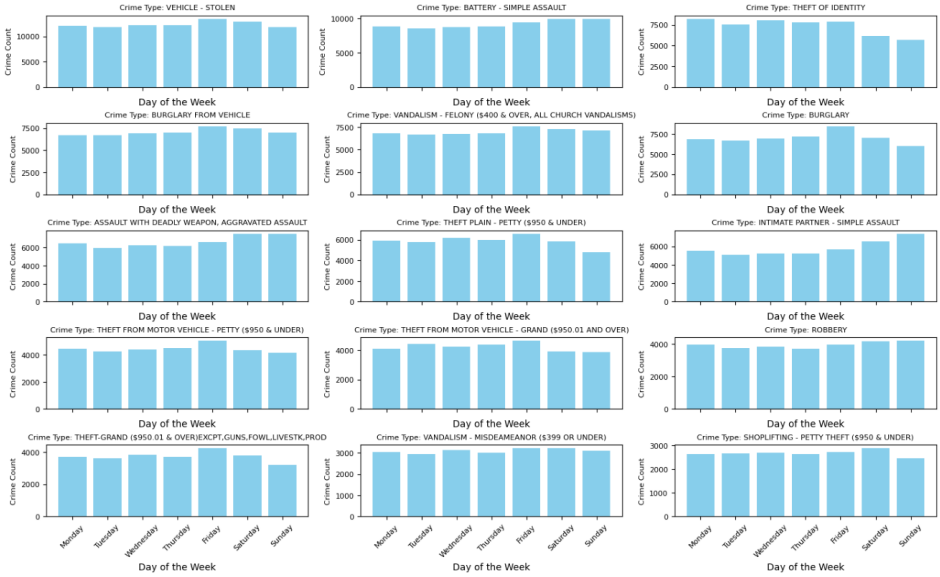
We then conducted an analysis of crime frequencies by grouping the data according to the days of the week. The results, presented in a bar chart, shed light on the following weekly crime patterns:

Mondays and Fridays exhibit comparatively higher crime frequencies. Tuesdays, Wednesdays, and Thursdays are associated with relatively lower crime frequencies. Saturdays and Sundays fall in the middle range in terms of crime frequencies.

It suggests that there is no significant fluctuation in crime rates throughout the week. However, we do observe a more noticeable uptick in crime frequencies on Mondays and Fridays, while midweek days (Tuesday through Thursday) show lower crime frequencies.

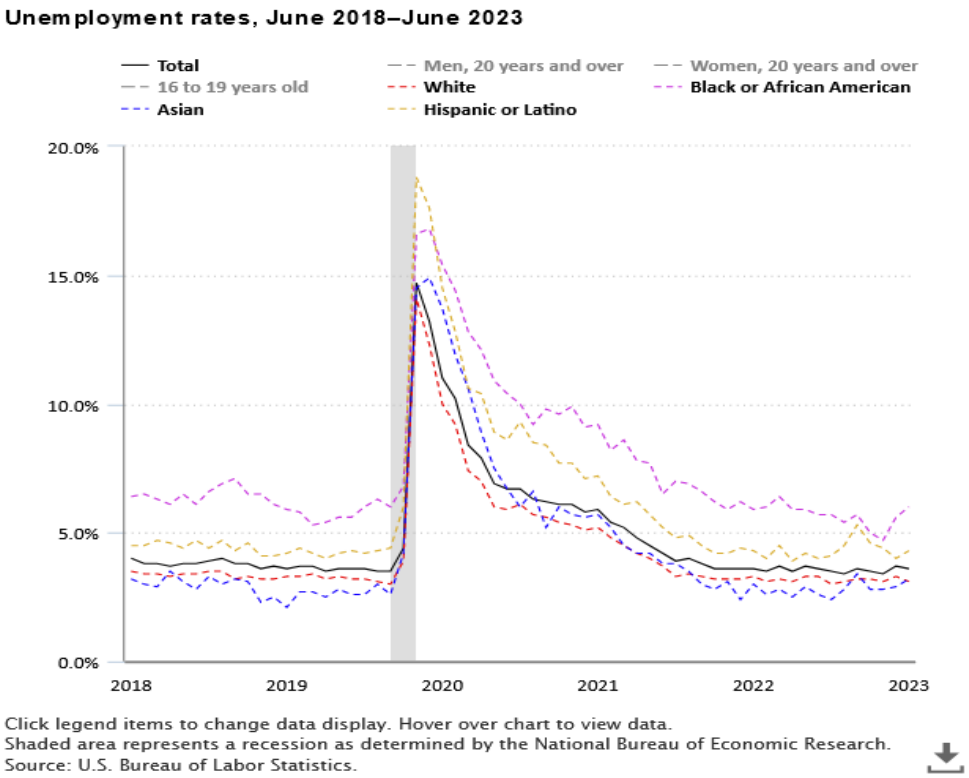


Our more granular examination of the top 15 specific crime types revealed that their distribution remains relatively stable across all days of the week. This suggests these crimes tend to occur with similar frequencies regardless of the day. Among the top types, "VEHICLE - STOLEN", "BATTERY - SIMPLE ASSAULT", "THEFT OF IDENTITY", and "VANDALISM - FELONY ($400 & OVER, ALL CHURCH VANDALISMS)" are particularly notable. While the distributions were largely consistent, we did observe subtle variations for certain crimes. For instance, "BURGLARY" saw a slight decrease on particular days compared to others. Overall though, our analysis found that the patterns for most top crime types show consistency throughout the week, with only minor fluctuations. The broader takeaway is that day of week does not seem to be a major factor driving variability in these common crime categories. But the analysis provides a useful baseline of their weekly distribution and highlights some types that stand out in magnitude.



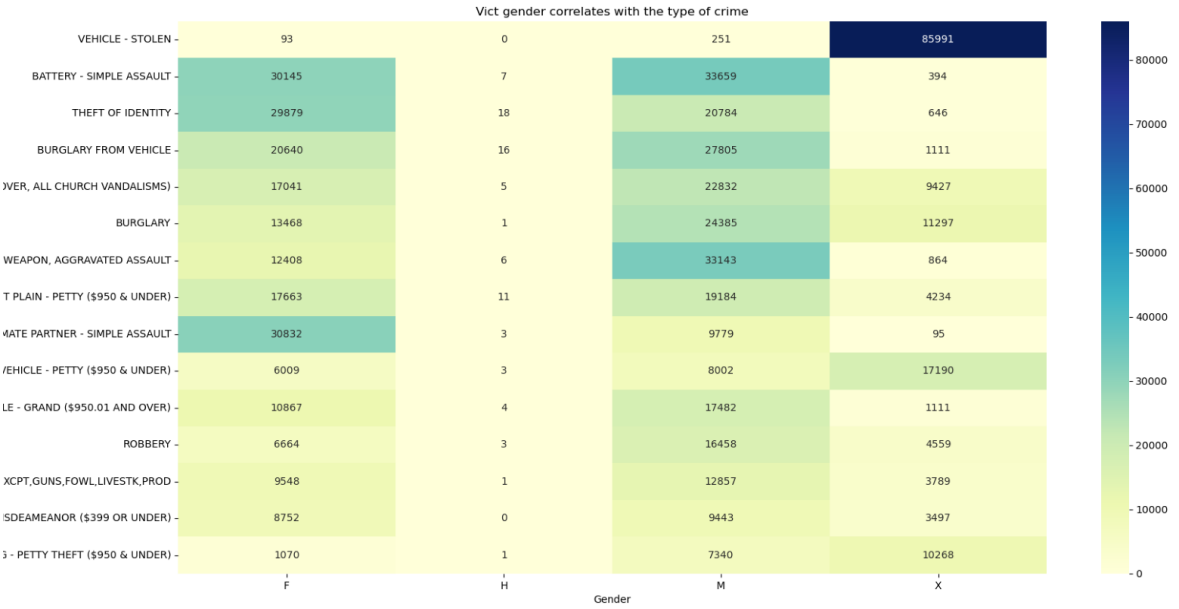
Analyzing the variations in crime rates within the dataset period reveals noteworthy factors contributing to these changes. In 2020, the COVID-19 pandemic played a pivotal role in driving crime rates down as the health crisis led to public health restrictions and reduced opportunities for criminal activities. Additionally, the Federal Reserve's actions, such as raising interest rates and implementing monetary measures, aimed at aiding unemployed citizens during the pandemic, may have influenced crime rates.

Conversely, the subsequent increase in crime in 2021 and 2022 can be attributed to distinct factors. In 2021, the unemployment rate failed to reach anticipated levels, and people grappled with the ramifications of the interest rate hike in 2020. As a result, inflation surged, driving up prices and creating economic challenges. These economic pressures likely contributed to the notable upswing in crime rates during this period.

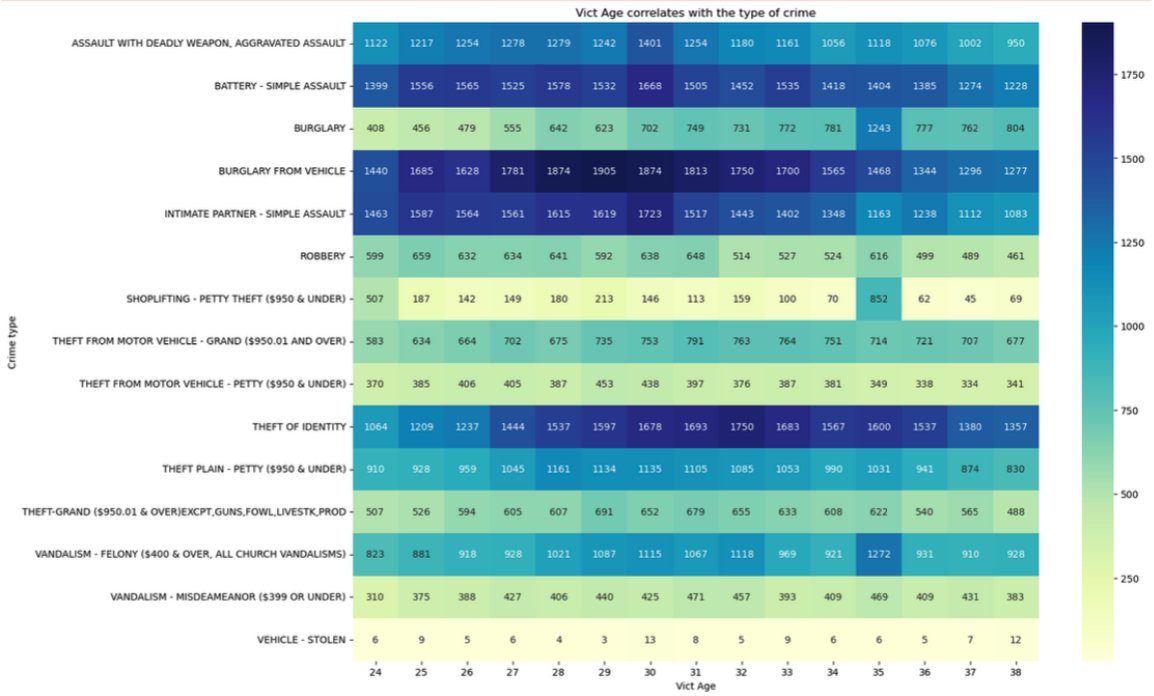


(Source: https://www.bls.gov/charts/employment-situation/civilian-unemployment-rate.htm)

Then, we sought to uncover patterns and correlations between demographic factors, such as age and gender, and the occurrence of specific crime types. To explore these connections, we constructed a matrix heat map that focuses on the gender of victims and the top 15 crime types. The analysis revealed distinctive gender distributions among different crime types. Particularly, some crime categories, like "Intimate partner-simple assault," exhibit a significant prevalence of female victims, greatly surpassing the number of male victims. This emphasizes the need for a gender-specific approach when addressing and preventing such crimes.

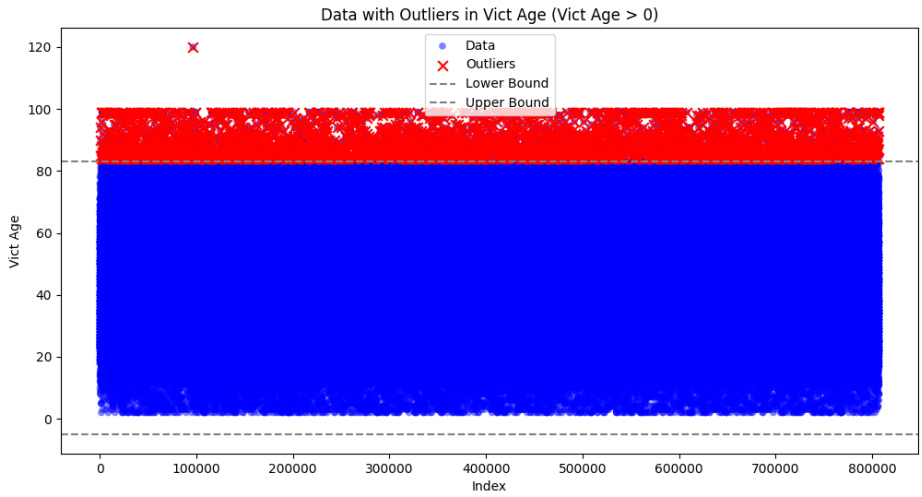


The 15 age groups with the most extensive distribution of crime victims, extracting and sorting this data. Subsequently, we generated a matrix heat map featuring the top 15 crime types and their age distribution. The results revealed that the majority of crime victims fell within the age range of 26 to 33 years old, shedding light on a notable age demographic among crime victims.

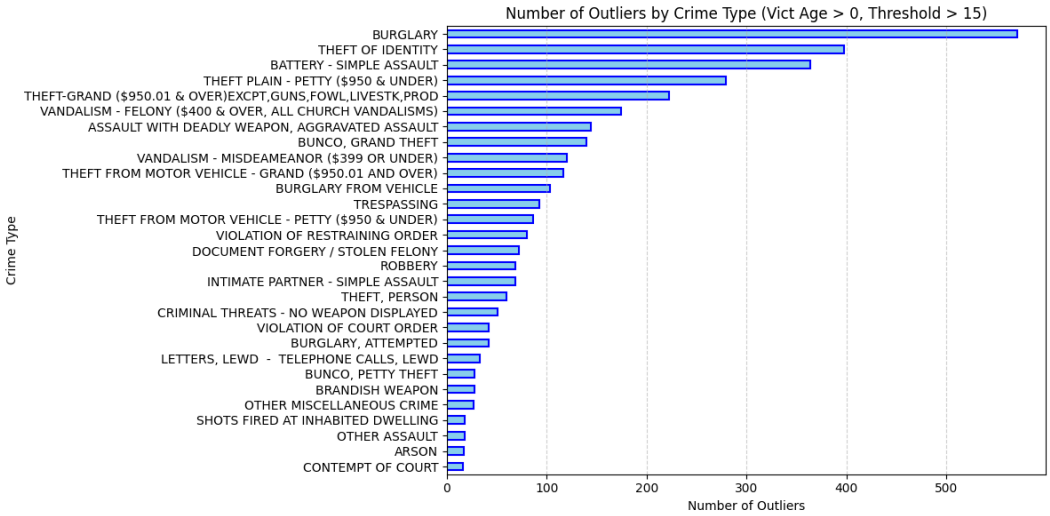


Outliers and Anomalies

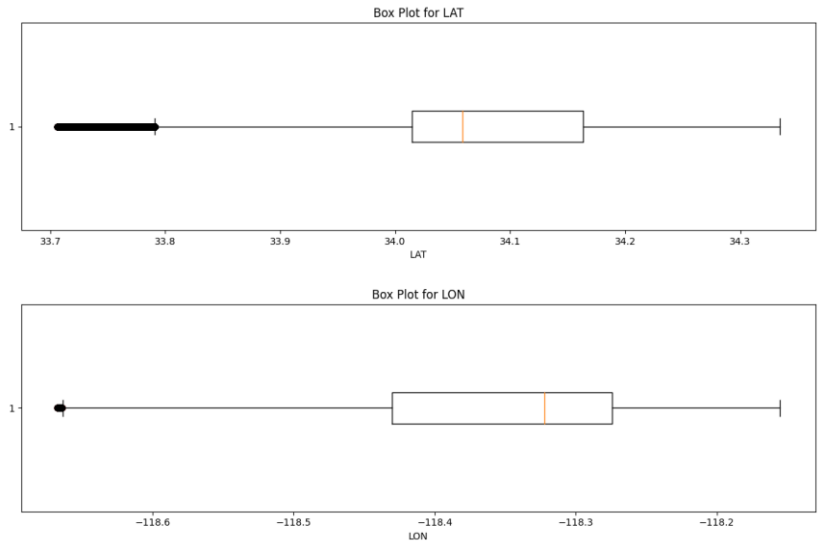
Our analysis aimed to identify outliers within the dataset and investigate any unusual patterns. The scatter plot titled "Data with Outliers in Vict Age (Vict Age > 0)" portrays "Vict Age" values on the y-axis against their corresponding indices on the x-axis. Notably, the majority of data points cluster in the lower age range, with a concentrated grouping of blue points representing ages below 20. A horizontal distribution of data is observed in the age range of approximately 20 to 80. The plot includes dashed lines denoting the upper and lower bounds for outlier detection, and data points lying beyond these boundaries are marked with red 'x' symbols, indicating their status as outliers. Also, there are discernible outliers exceeding the age of 80, with a few scattered data points hovering around the age of 100.



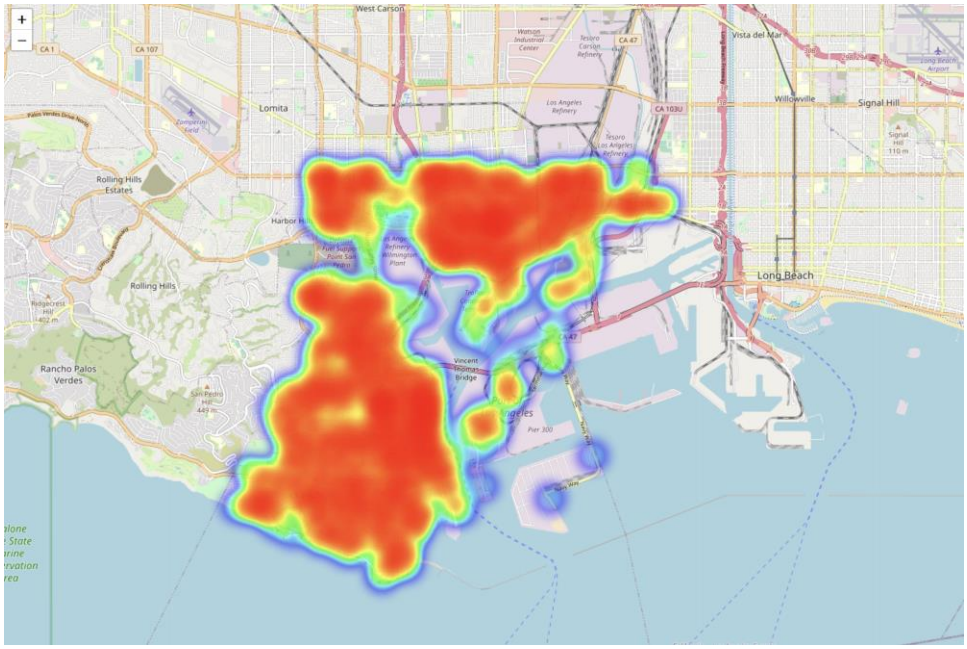
"Number of Outliers by Crime Type (Vict Age > 0, Threshold > 15)," which plots different crime types on the y-axis and their respective outlier counts on the x-axis. The chart highlights variations in the number of outliers across crime types, with "Burglary," "Theft of Identity," and "Battery - Simple Assault" standing out as categories with a notably higher incidence of outliers. This visual representation not only underscores the disparity in outlier occurrences but also provides a foundation for more comprehensive investigation, potentially revealing insights into the underlying reasons why certain crime types exhibit a greater number of outliers.



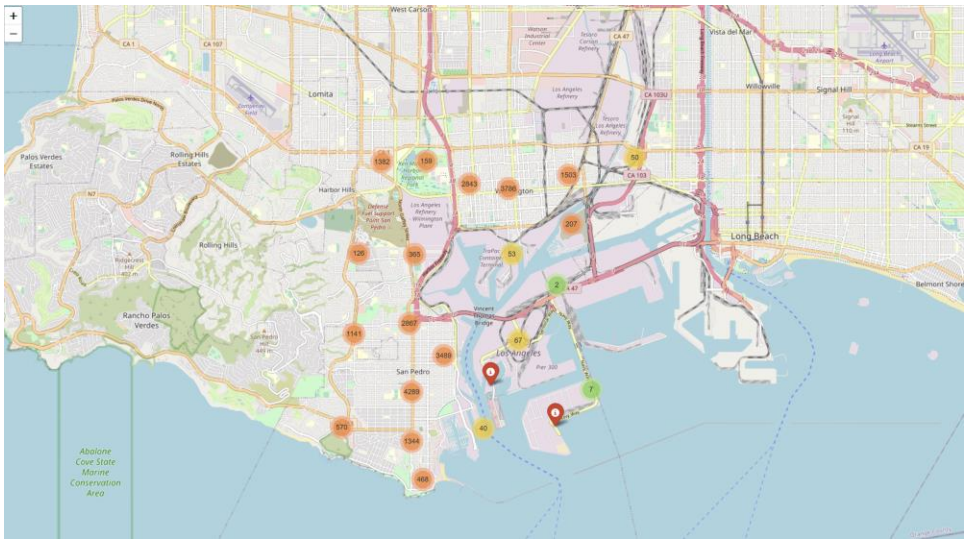
The box plots for latitude (LAT) and longitude (LON) illustrate that most data points are concentrated within narrow ranges, with a few distinct outliers. For LAT, values span approximately 33.7 to 34.3, but the box shows the bulk of data packed close to 34.1. A prominent outlier on the far left sits away from this central range. Similarly for LON, values range from -118.6 to -118.2, but the box reveals the majority of data clustered near -118.3. An outlier is again visible on the left side, distant from the main data concentration. In summary, the visualizations demonstrate dense clustering of data around specific latitude and longitude values, indicating the spatial concentration of crimes within defined location bounds. The few outliers likely represent incidents outside the high-density zones. Overall, the tight clusters highlight the geographic specificity of crimes within expected LAT and LON ranges in the city.



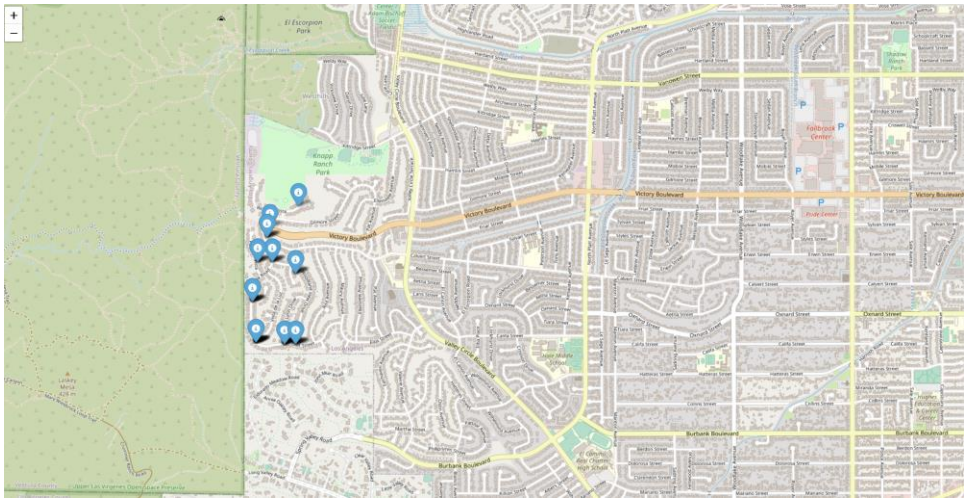
The heatmap visually captures concentrated outlier zones across Los Angeles. By mapping color intensity from blue to red based on outlier density, it immediately highlights areas of highest concentration in red. These red zones pinpoint specific locations with abnormally high outlier occurrences that may merit further investigation or priority. Overall, the heatmap concisely synthesizes the geographical distribution of outliers, using color coding to allow quick identification of outlier epicenters that could represent anomalies or risks. The visualization transforms the abstract concept of statistical outliers into an actionable spatial map of concentrations demanding attention.



The map visualization depicts geographical outlier clusters across Los Angeles, with circles denoting the count of markers within each cluster. These clustered circles synthesize the outliers into condensed representations of high-density zones. When zoomed, the clusters decompose to reveal the individual outlier markers comprising them. This interactive format enhances understanding of both macro-outlier distributions and micro-level spatial patterns. For example, "San Pedro" contains a substantially larger cluster circle of 11078 outliers versus sparser regions. Overall, the map concisely transforms granular outlier markers into digestible macro-level clusters, enabling quick identification of high outlier density areas for further investigation while still retaining access to the underlying granular data.

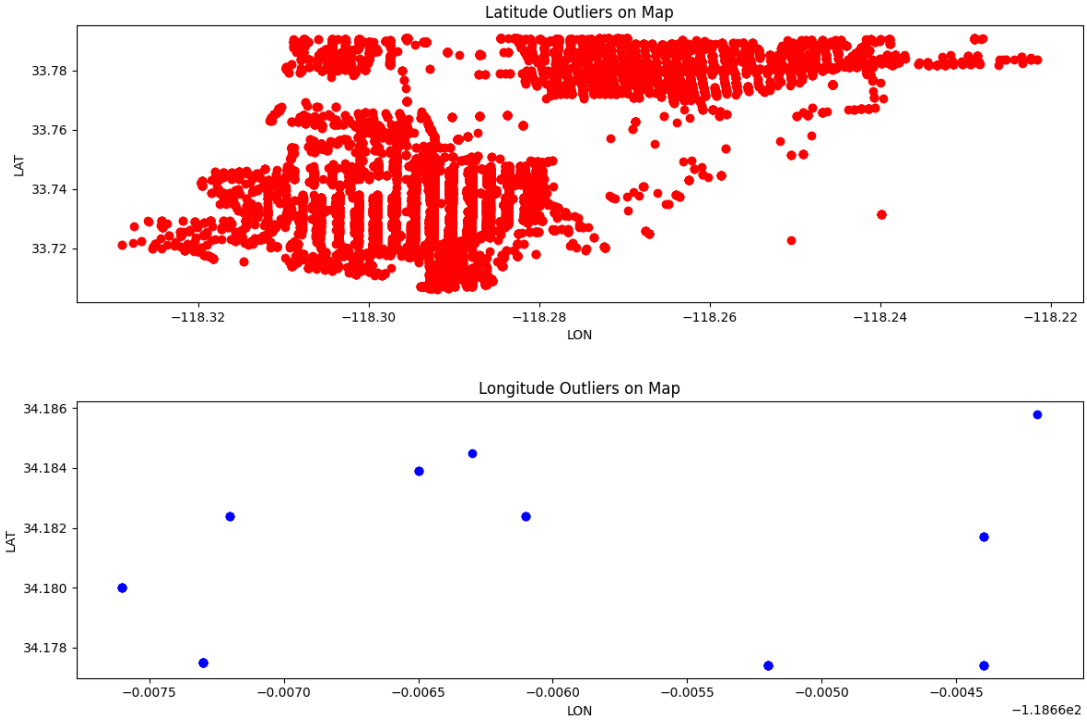


The image presents a map of a region near Los Angeles, with specific locations marked by blue pins. These pins represent the longitude outliers, as indicated by the outliers_lon dataset. The map centers on West Hills and its nearby areas, with multiple blue markers clustering around the southwestern edge of West Hills, near the "Upper Las Virgenes Open Space Preserve." These markers provide insights into the spatial distribution of longitude outliers within this region. The dense cluster suggests a significant concentration of these outliers in a relatively confined area.

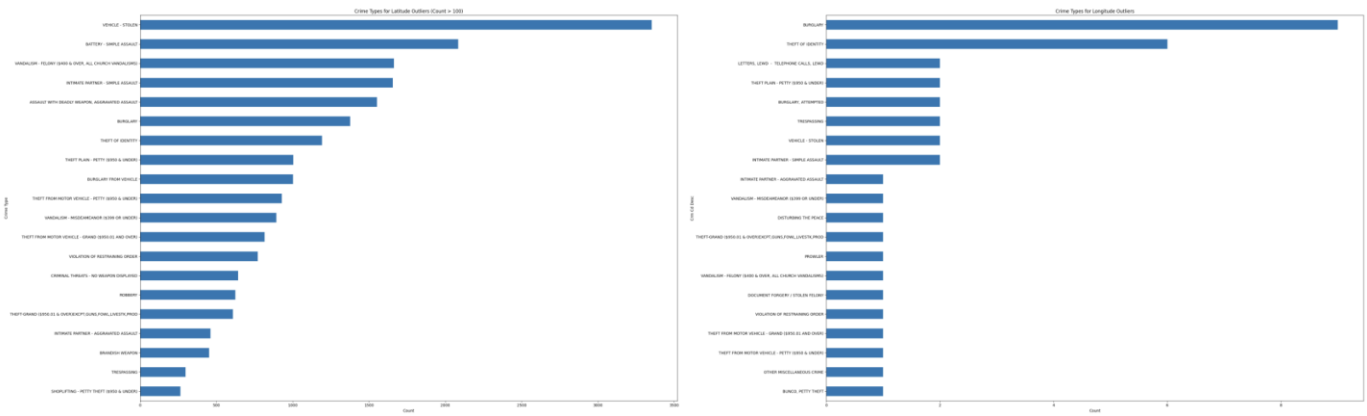


The visualizations effectively represent latitude and longitude outliers in separate subplots for clear distinction. The red dots depicting latitude outliers exhibit clustering around defined areas, highlighting potential regions of interest or data inconsistencies for further investigation. In contrast, the sparse and dispersed blue longitude outliers suggest more sporadic anomalies in that dimension. The separation into two visuals allows quick assessment of the

different outlier characteristics. Overall, these geographic data visualizations enable rapid identification of spatial anomalies and inconsistencies through the outlier distributions. The clustered latitude outliers point to specific zones that may warrant attention, while the sporadic longitude outliers indicate more isolated anomalies.



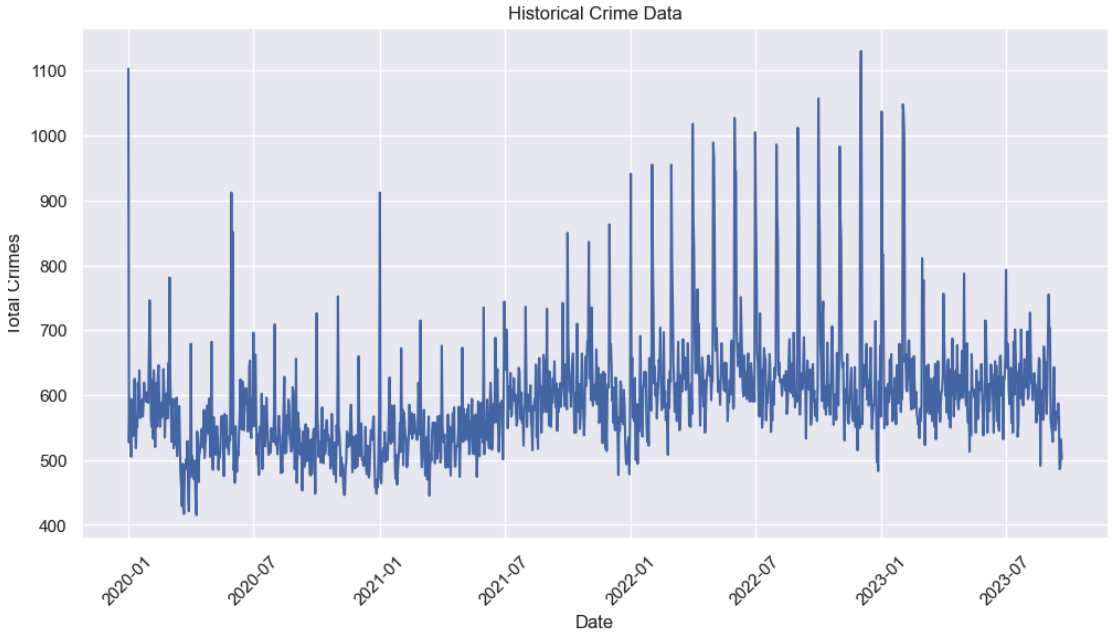
The latitude outliers analysis chart reveals "Vehicle - Stolen" as the most frequent crime type associated with latitude anomalies, followed by "Battery - Simple Assault" and "Vandalism - Felony ($400 & over, All Other Vandalism)". For longitude, the analysis shows "Burglary" as the predominant outlier-related crime, with a markedly higher count than other types. Taken together, these visualizations identify the specific crimes that repeatedly occur in areas exhibiting geographic data inconsistencies or anomalies. The latitude analysis highlights vehicle theft, assault, and vandalism as crimes meriting attention in the clustered outlier zones. Meanwhile, the longitude analysis points to burglary as the key crime type arising in areas with sporadic geographic anomalies. By associating crimes with spatial outliers, these charts provide focus for investigative efforts and data verification in affected high-crime areas.



*All html files can be download [Here]

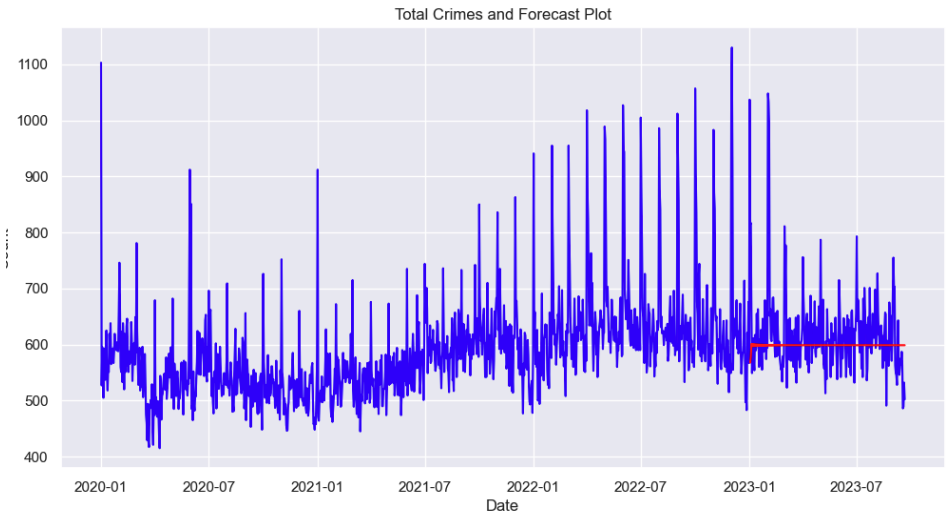
Data Modeling and Algorithms

In our time series forecasting analysis, we employed three distinct prediction models: ARIMA, SARIMA, and Prophet. To initiate the time series forecasting process, we began by extracting the total daily crime data and aggregating it by date. We then saved this processed data in a 'daily_crime.csv' file for convenient access throughout the analysis. Following this, we visualized the historical crime data to discern any observable trends, which aided us in determining the most suitable model for this dataset.

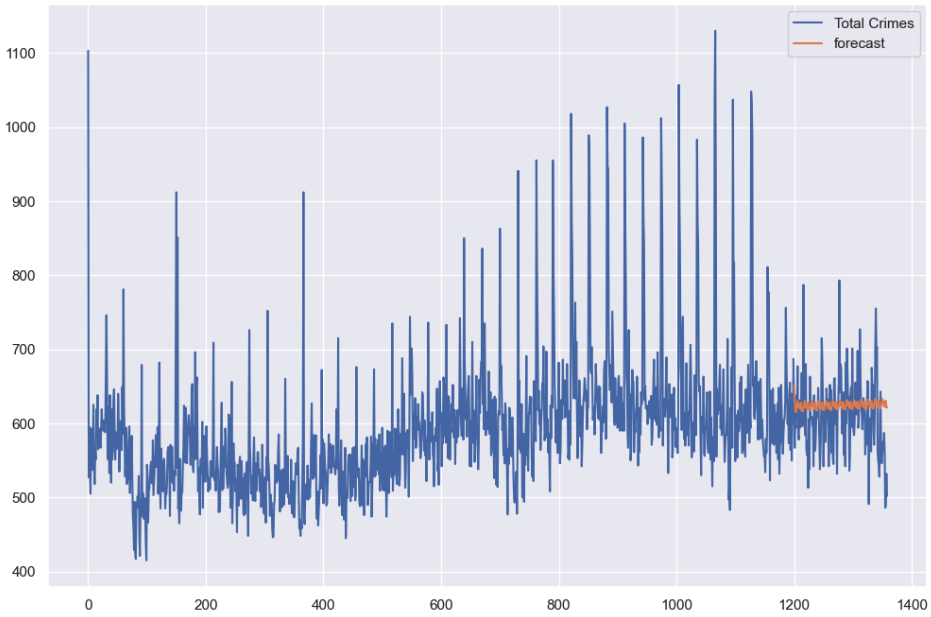


Next, we divided the data into two subsets: a training dataset containing observations up to January 1st, 2023, and a testing dataset comprising data from January 1st, 2023, onwards.

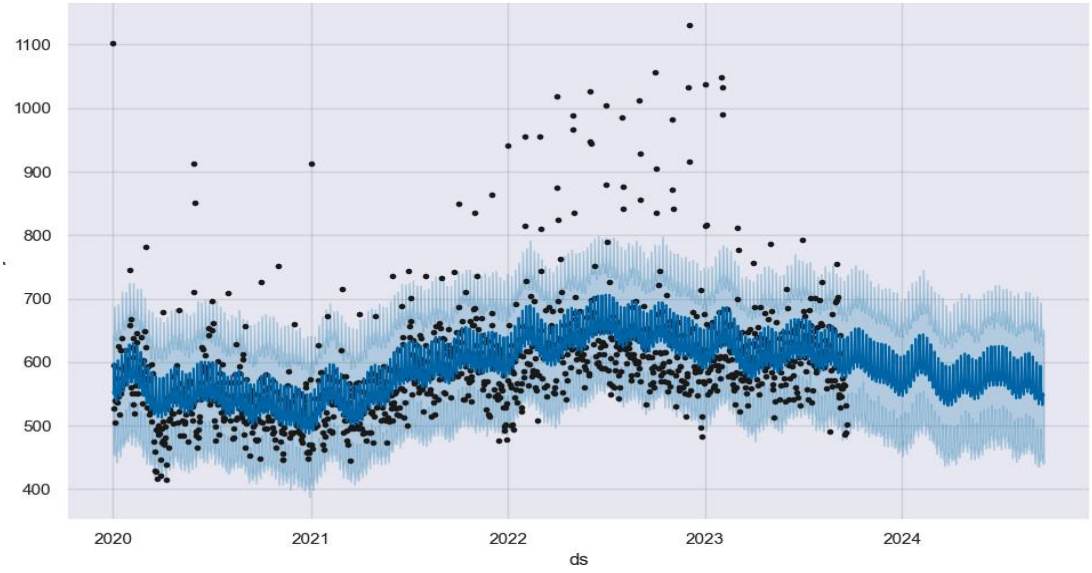
Our first modeling approach involved utilizing ARIMA (Autoregressive Integrated Moving Average). To identify the appropriate ARIMA parameters (p, d, q), we began by generating autocorrelation and partial autocorrelation plots (ACF and PACF) and conducted an Augmented Dickey-Fuller (ADF) test. The ADF test indicated a p-value below 0.05, allowing us to reject the null hypothesis, signifying the presence of a stationary pattern. As a result, we determined that the value of d should be set to 1. Then we employed the *'pmdarima'* library, which is equivalent to R's '*auto*.arima' functionality, and it yielded an ARIMA(1,1,4) model as the optimal choice. We further assessed the model's performance by examining residuals, their density, and the ACF and PACF plots of the residuals. The white noise appearance of the residuals indicated the model's adequacy. Additionally, the residuals show a normal distribution with a mean centered at 0, while the ACF and PACF plots of the residuals showed minimal significant spikes. These assessments confirmed the suitability of the (1,1,4) parameters for making predictions, and the resulting forecast is presented below.



The second model applied in our analysis is SARIMA, an extension of the ARIMA model specifically designed for time series data with a seasonal component. We augmented the SARIMA model with seasonal order (1, 1, 1, 12) to account for the seasonality, and the resulting forecast is displayed below.



Our final model, Prophet, is a powerful tool for time series forecasting. It employs an additive model that captures non-linear trends, yearly, weekly, and daily seasonality, and holiday effects. This model excels when dealing with time series data exhibiting strong seasonal patterns and multiple seasons of historical data. Prophet exhibits robustness in handling missing data and trend shifts and is typically effective at accommodating outliers. With Prophet, we were able to make predictions regarding the next 356 days crime rates.



In the context of our time series analysis, the ARIMA model's performance metrics reveal important insights. The Mean Absolute Error (MAE) for the ARIMA model is computed to be 45.29. This value signifies that, on average, the absolute disparity between the actual and predicted values is approximately 45.29. Additionally, the Mean Absolute Percentage Error (MAPE) is calculated at 0.06887, indicating that, on average, the absolute percentage discrepancy between actual and predicted values amounts to approximately 6.89%. Furthermore, the Root Mean Square Error (RMSE) is quantified at 72.32, which implies that, on average, the squared variance between actual and predicted values is approximately 72.32. It's important to note that lower values for MAE, MAPE, and RMSE suggest superior model performance, while higher values indicate more substantial errors. In this assessment, the ARIMA model exhibits subpar performance.

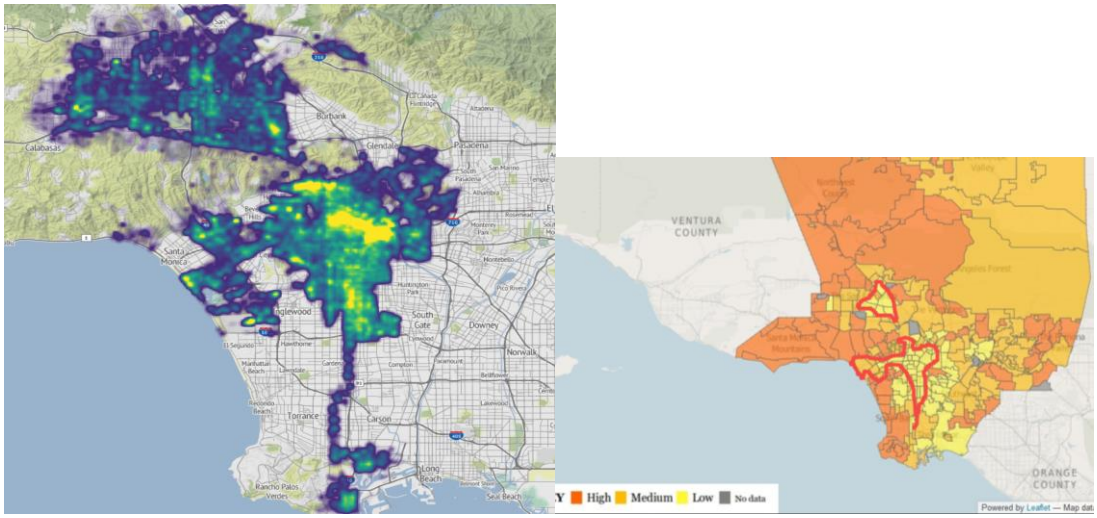
Comparatively, the SARIMA model demonstrates improved performance with MAE at 42.093, MAPE at 0.0708, and RMSE at 54.17. These metrics indicate that the SARIMA model outperforms the ARIMA model, presenting lower error measures.

Lastly, in our evaluation of the Prophet model, we uncover a discernible seasonal pattern characterized by an annual recurrence. This pattern manifests as a sharp rise in crime rates at the onset of each year, peaking around January and February, aligning with the typical post-holiday spikes. Afterwards, crime levels decline through the spring, reaching a nadir around May to July, coinciding with the influence of warmer weather on reducing certain types of crimes. Late summer sees another increase, peaking around August and September, likely due to schools reopening and increased opportunities. Towards the end of the year, crime rates surge again, likely due to the holiday season, with the cycle repeating annually.

The seasonal fluctuations encompass a variation of roughly +/- 100 crimes, contributing to approximately 10-15% of the overall variance around the baseline trend. It's worth noting that the shape of this annual pattern, with its distinctive peaks, troughs, and slopes, may have also been influenced by external factors such as the COVID-19 pandemic. Notably, we observe a distinct peak following lockdown periods and a slowdown during quarantine measures, indicating the impact of these external events on the pattern.

Conclusion

By mapping the crime heat map to the median income gradient map, it is easy to find that most crime is concentrated in light-colored areas, namely low-income areas. The darker areas in the income map, known as high-income areas, have little or no crime. So, we can draw a conclusion that **the crime rate is negatively correlated with the regional economy.**



Upon examining the spatial distribution of crime across different crime types, a prominent trend emerges, with many crimes converging in the central "7" shape of the city's Central region. Consequently, it is prudent to allocate additional police resources to this area on Friday, Saturday and holiday season to address the concentrated criminal activities effectively.

Furthermore, when considering the day of the week and its relationship with crime types, it is evident that a notable spike in criminal incidents occurs on Friday, Saturday, and Sunday. Consequently, optimizing the police presence during these specific days, when crime frequencies are at their highest, is advisable.

In our analysis of the correlation between gender and crime types using a matrix heat map, a noteworthy insight surface. Specifically, the number of female victims in the "Intimate partner-simple assault" crime category significantly surpasses that of male victims. This finding underscores the importance of raising awareness among women in Los Angeles about the risk of assault from intimate partners, highlighting the need for targeted prevention and support efforts.

Our analysis primarily focused on the temporal aspects of crime rates. While Prophet recognized the impact of COVID-19, a more comprehensive analysis could benefit from considering additional external factors. The accuracy of our analysis is dependent on the quality and completeness of the dataset. Incomplete or inaccurate data can lead to suboptimal model performance.

Enhancing the feature engineering process by incorporating additional relevant variables can improve model accuracy. Exploring more advanced time series models, including machine learning-based models or hybrid approaches, can further enhance predictive capabilities. We also suggest developing real-time forecasting capabilities for crime rates to assist law enforcement agencies in proactive planning and resource allocation. Our time series analysis provided valuable insights into crime rate forecasting. While the models exhibited varying levels of performance, there is ample opportunity for improvement through further refinement of modeling techniques and the incorporation of additional factors. This analysis represents a solid foundation for future research in the field of crime rate forecasting and its broader applications.

References

https://data.lacity.org/Public-Safety/Crime-Data-from-2020-to-Present/2nrs-mtv8

https://www.usatoday.com/in-depth/money/2020/05/12/coronavirushow-u-s-printing-dollars-save-economy-during-crisis-fed/3038117001/

https://www.usbank.com/investing/financial-perspectives/market-news/federal-reserve-tapering-asset-purchases.html

https://www.depledgeswm.com/depledge/the-us-printed-more-than-3-trillion-in-2020-alone-heres-why-it-matters-today/

https://medium.com/well-red/outlier-and-anomaly-detection-using-facebook-prophet-in-python-3a83d58b1bdf

https://machinelearningmastery.com/time-series-forecasting-with-prophet-in-python/

https://medium.com/@josemarcialportilla/using-python-and-auto-arima-to-forecast-seasonal-time-series-90877adff03c