LA Crime Data Analysis Report

Shihan Wang

0. Dataset

The dataset I analyzed is LA Crime data. This dataset has 2.12 million rows and 28 columns, and each row reflect incidents of crime in the City of Los Angeles from 2010 to 2019. The main goal of this project is to provide safety guidance to visitors who would like to visit LA and give advice to the police department on efficiently distributing the police.

1. OLAP

Before diving into the analysis, I retrieved the data from the LA government website and saved as a DataFrame in Spark. I used both Spark DataFrame and Spark SQL to do data cleaning and data preprocessing work. And utilized DataBricks built-in tools for data visualization.

First, I used OLAP to explore the general patterns of LA crime incidents. For example, the 5 most frequently occurring crime categories in LA are: Battery - Misdemeanor, Burg from Vehicle, Stolen Vehicle, Theft - $950 & Under, and Burglary. And the 5 most dangerous (with the largest number of crimes) areas in LA are 77th Street, Southwest, N Hollywood, Pacific, and Southeast.

Then I would like to know the safety trend of the most popular area for visitors, LA downtown. I defined LA downtown via the range of spatial location and visualized the number of crimes on Sunday at LA downtown. The plot shows seasonality but also indicates that the safety is getting worse at LA downtown on Sunday within 10 years.

As the last plot shows that safety is getting worse these years, I filtered the data and only keep the records between 2015 and 2019 to test this statement. Generally, 2015 is the safest year with the lowest number of crimes every month. However, the plot also shows as the number of crimes reached a peak in 2017, safety is getting better since then, which is against my previous statement. Looking into each month, I highly recommend the safety-concerned visitors to plan their trip in February.

To provide the visitor with more detailed guidance on safety, I analyzed the number of crimes of each hour through a single day. Taking 2019/12/15 as an example, 8 pm is the time when most crimes occur while 4 am is the time associated with the lowest number of crimes. Generally, morning hours are safer than evening hours. Visitors should avoid walking alone during the evening hours.
Moving forward to the age, sex, and ethnicity patterns of victims, I found most of the victims are around 19 to 35 years old. And in general, younger females are more likely to

be victims than younger males. Conversely, for people over 35 years old, the male is more likely to be the victim. Overall, Hispanic/Latin/Mexican, White, and Black are 3 groups with the most victims. But when I narrowed down to consider sex in these 3 ethnic groups, things are different. The white male is more likely to become the victim than the white female, while the black female is more likely to become the victim than the black male. Meanwhile, sex didn't play an important role in the Hispanic victim group.

My last part in OLAP is to analyze the percentage of victim ethnicity of major crime categories. As the pie chart shows, the Hispanic group is more likely to suffer from the stolen vehicle while the black group is more likely to become the victims of battery misdemeanor.

2. K-means Clustering

   To better understand this dataset, I leveraged the Spark ML clustering algorithm to cluster the spatial data, and then visualize the clustering result. By evaluating the squared Euclidean distance, I found the optimal number of centroids is 7 in this dataset.

3. Time Series Analysis

   Inspired by the trend between the number of crime and time variables, I wanted to identify trends and seasonality using time series analysis. I first visualized the time series with the seasonal decomposition tool from Statsmodels and found a strong cyclical pattern in the seasonal plot. Then, I tried to stationarize the data with different lags and found combining first difference and seasonal difference outperformed. I plot ACF and PACF and tried to find out the optimal parameter of ARIMA models from these two plots. I tried several combinations of parameters and kept the set with the lowest AIC score and make predictions based on this model. As shown in my prediction plot, I successfully capture the seasonality effect but failed to explain the drop of the number of crimes since 2017.

4. Conclusion
   In conclusion, I think we could leverage this dataset to build a dashboard-based website or app to provide all visitors a personalized guide for a safe trip. For example, users could type in their age, sex, race group, and locate the safest place to stay and to hang out.