BIG DATA ANALYTICS PROJECT

Report

Submitted by:
Team #5 (Gamal, Amr, Ahmed, Ahmed)

Problem & Motivation

Crime in Chicago has been tracked by the Chicago Police Department's Bureau of Records since the beginning of the 20th century. The city's overall crime rate, especially the violent crime rate, is substantially higher than the US average. Chicago was responsible for nearly half of 2016's increase in homicides in the US, though national crime rates stayed near historic lows.

As of 2017, Chicago's homicide rate is significantly higher when compared to the larger American cities of New York and Los Angeles, but lower when compared to smaller American cities. The reason for the violence which is localized to some areas of the city, including change in police tactics or increase in gang rivalry, remain unclear.

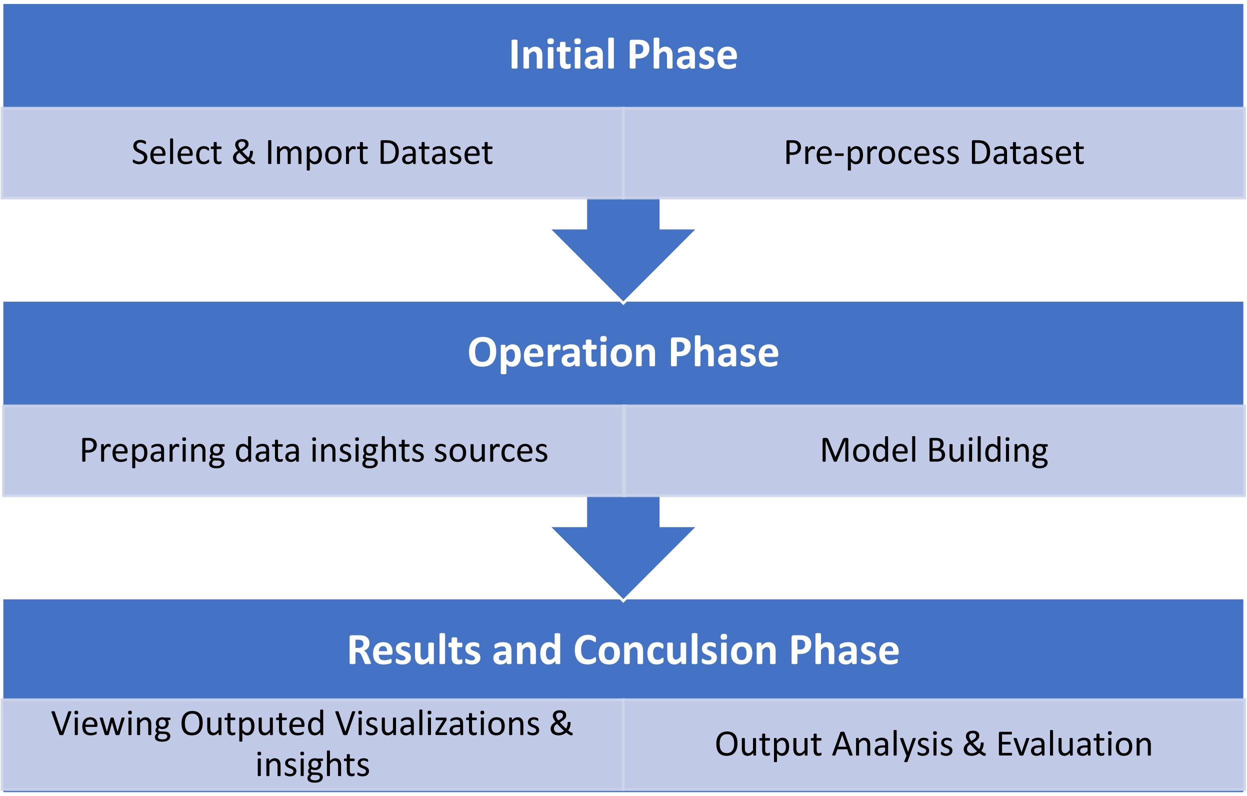
<u>So, what's the motivation for harvesting such data (i.e. business problem)?</u>

- Chicago Crime Numbers have been in the media for all the wrong reasons.
- US President widely used these numbers during his presidential campaign and He still talks about it!
- These numbers are impacting the overall numbers of the country.
- Do homicides rates are really getting worse in Chicago in 2016?

<u>Thus, lots of work can be done to study and answer such important questions:</u>

- How has crime changed over the years?
- Is it possible to predict where or when a crime will be committed?
- Which areas of the city have evolved over this time span?

Project Pipeline


Initial Phase
Select & Import Dataset
Pre-process Dataset
Operation Phase
Preparing data insights sources
Model Building
Results and Conculsion Phase
Viewing Outputed Visualizations & insights
Output Analysis & Evaluation

Solution & Analysis

The solution works on 2 main tracks:

- Exploring & Visualizing previous years crimes recorded in the dataset starting from 2005 to 2016
- Forecasting future crimes in the next few years till 2019

I. Data Preprocessing:

- Dataset was cleaned from NA-valued records
- "Year, Month, Day" columns are added to each record, derived from the existing date column. They are added to facilitate later insights

II. Data visualization:

- In the Exploration part, useful insights and nice informative plots were developed, including Heatmaps, Time series analysis diagrams and Bar charts.
- Visualizations were mainly on crimes/arrests time series analysis, crimes types, locations and time.
- There were visualizations too for time series forecasting of crimes including the trend, weekly and yearly components
- We use High charter and ggplot libraries

III. Model Preparation

- In the forecast part, crime time series forecast analysis diagram was developed showing the future crime predictions.
- Model was prepared from the dataset records by extracting the Id and the data columns and modifying the date format to be suitable for our use.
- We used the "Prophet" time series forecasting package. It is based on an additive model where non-linear trends are fit with yearly and weekly seasonality, plus holidays.
  Prophet is robust to missing data, shifts in the trend, and large outliers.

IV. Model fitting & usage

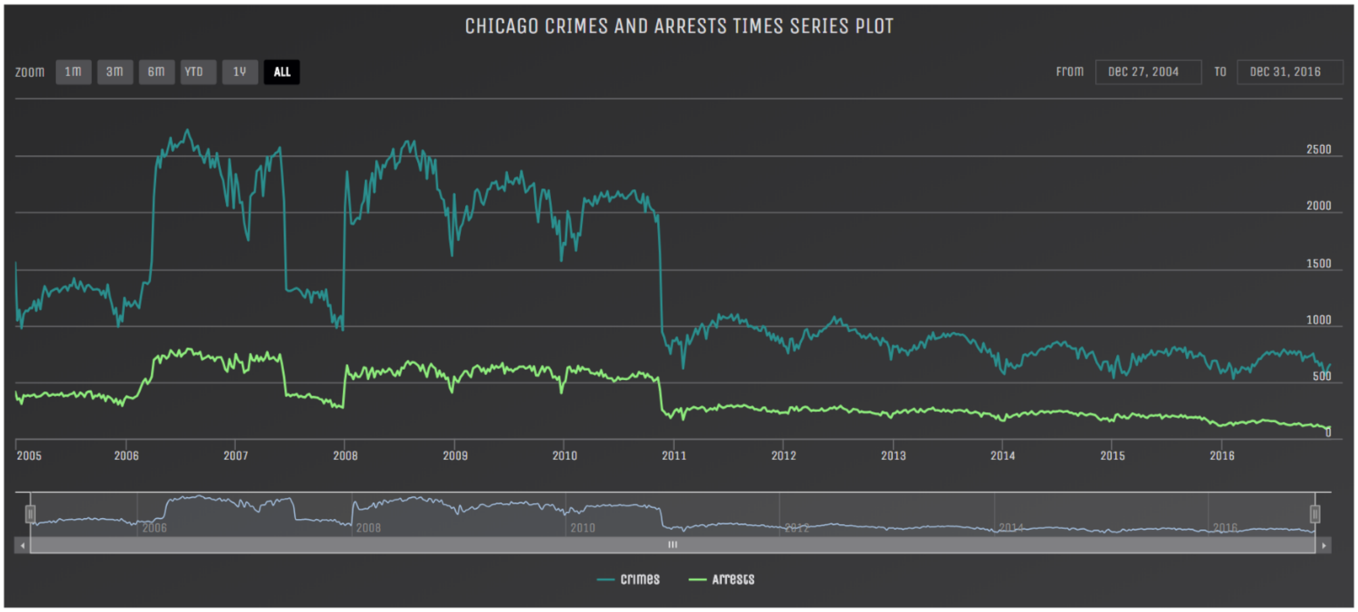
- Model was fitted using the previous year dataframe prepared
- A future dates dataframe was created containing the future dates to predict
- The future dataframe is supplied to the predict function of Prophet and will provide a future value prediction based on the model and future dataframe.
- Prophet provides a min, max and average values for each future prediction
- Moreover, we can plot the time series analysis decomposition into Trend, weekly and yearly analysis.

Results

- All generated plots can be found and viewed in the attached file.

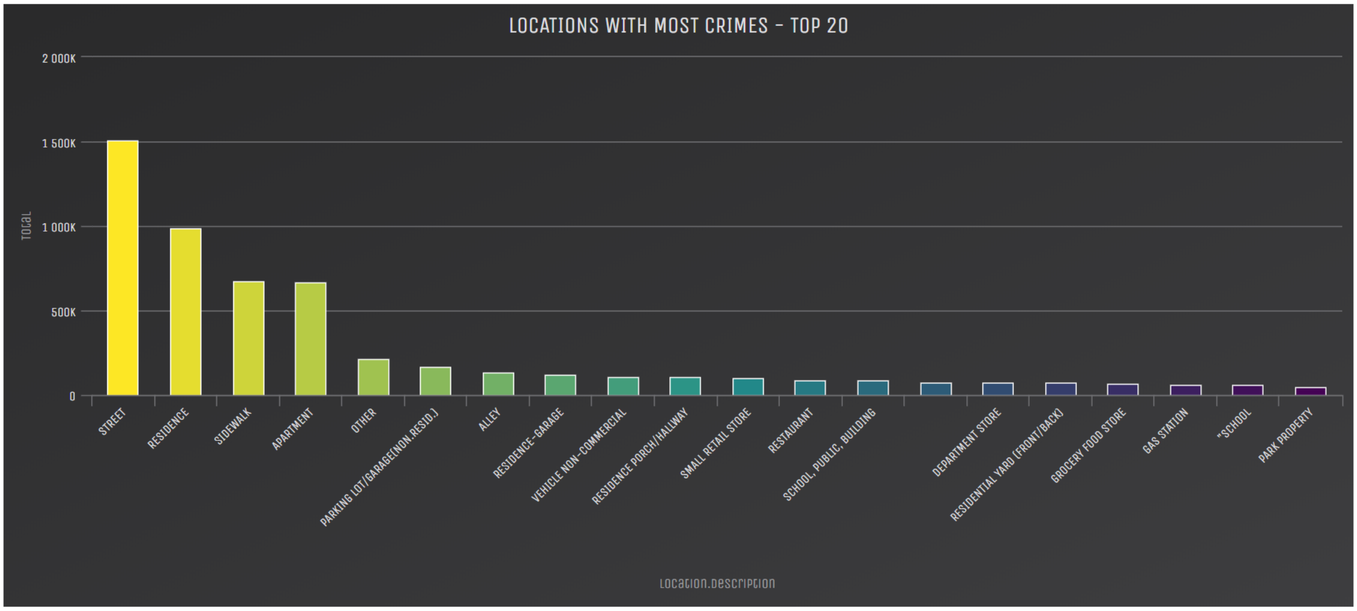
Sample Plots:

Plot #1:



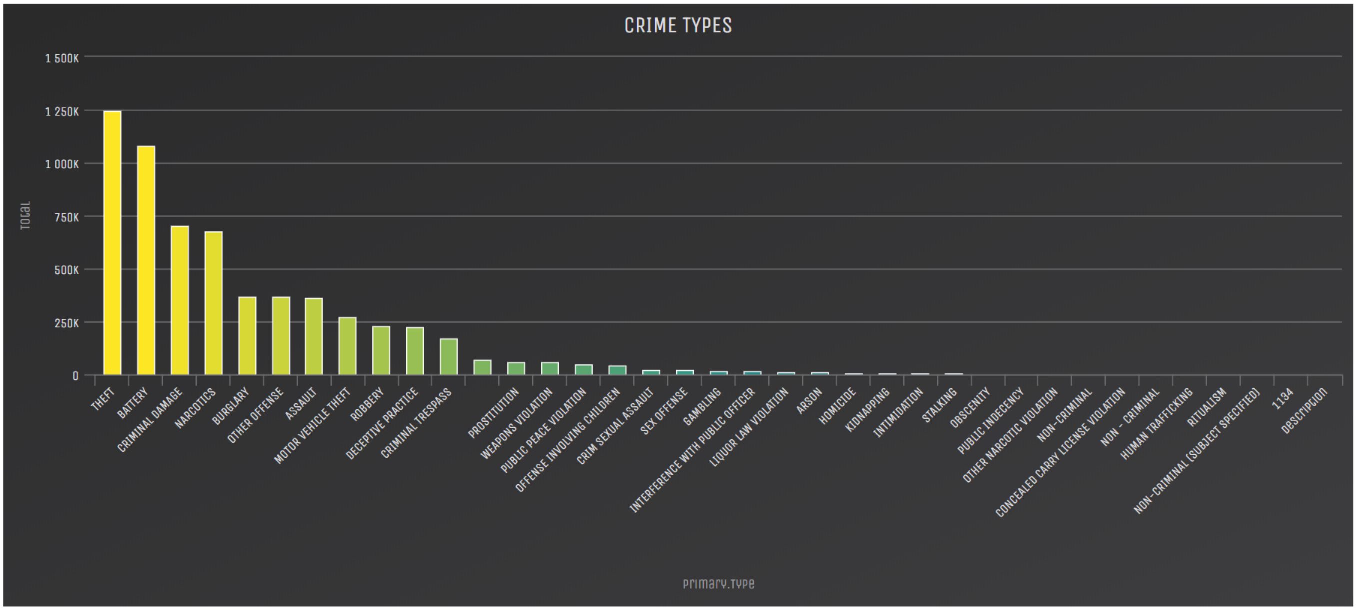
We can see that crimes and arrests have decreased starting from 2011 till 2016 and the peaks of crime numbers is during the middle of each year in summer!

Plot #5:



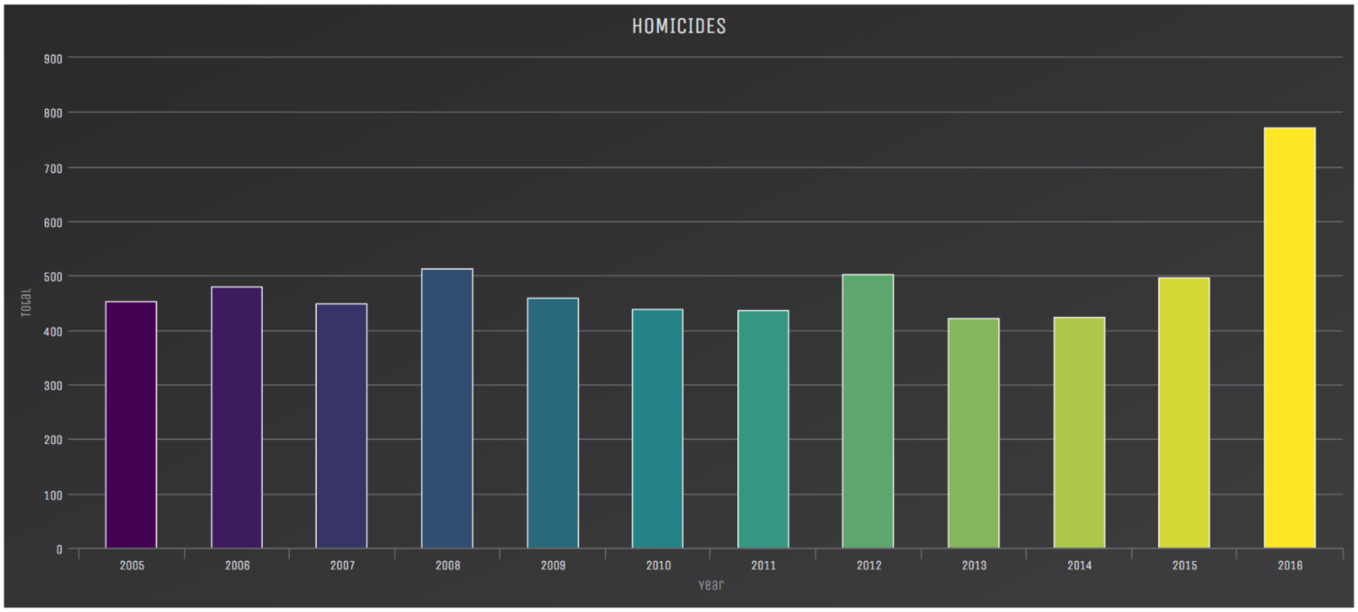
We obviously see that crimes occur most at streets, residence, sidewalk and apartments

Plot #7:



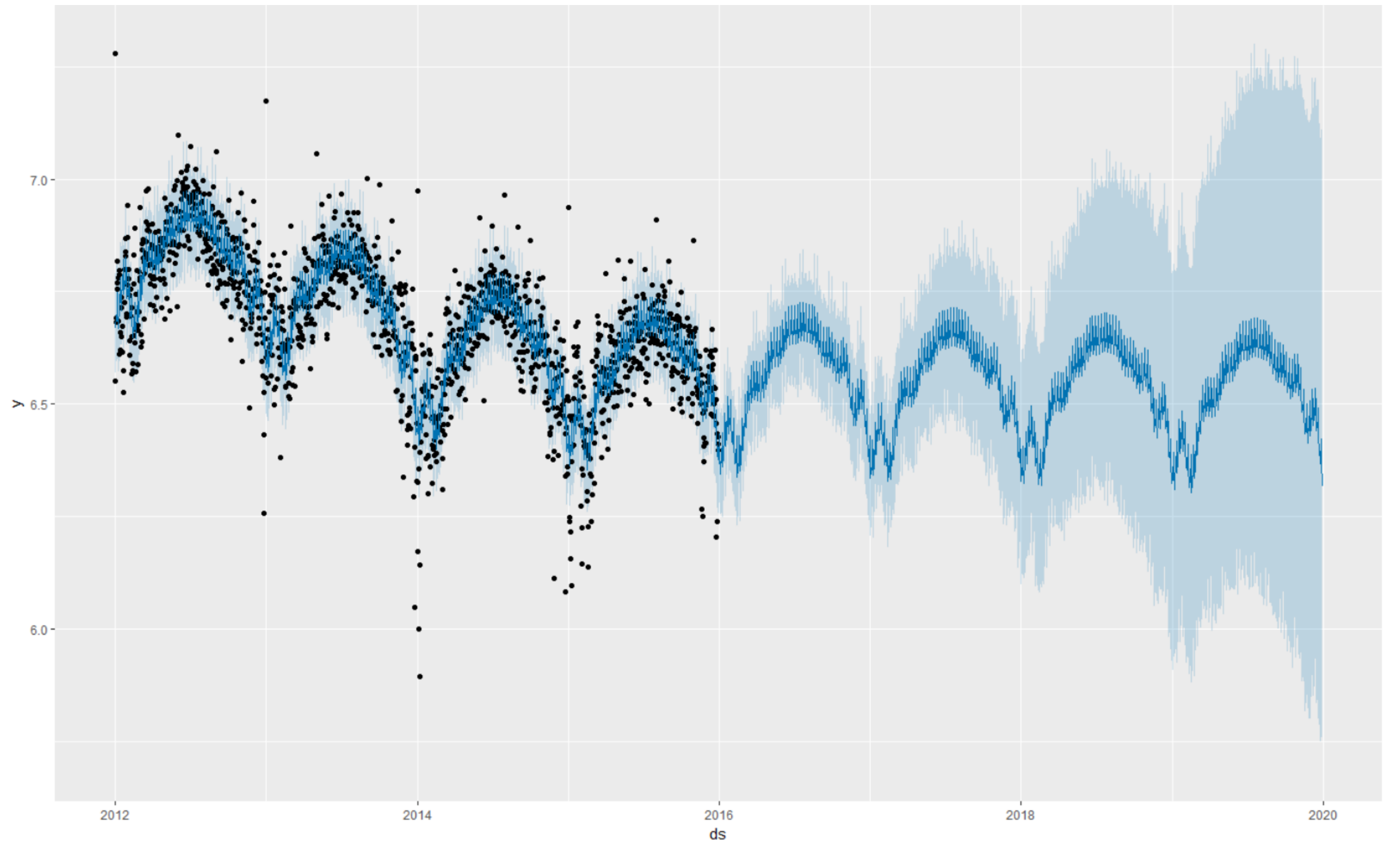
We can obviously see that Theft, Battery, criminal damage and Narcotics are top crime types.

Plot #9:

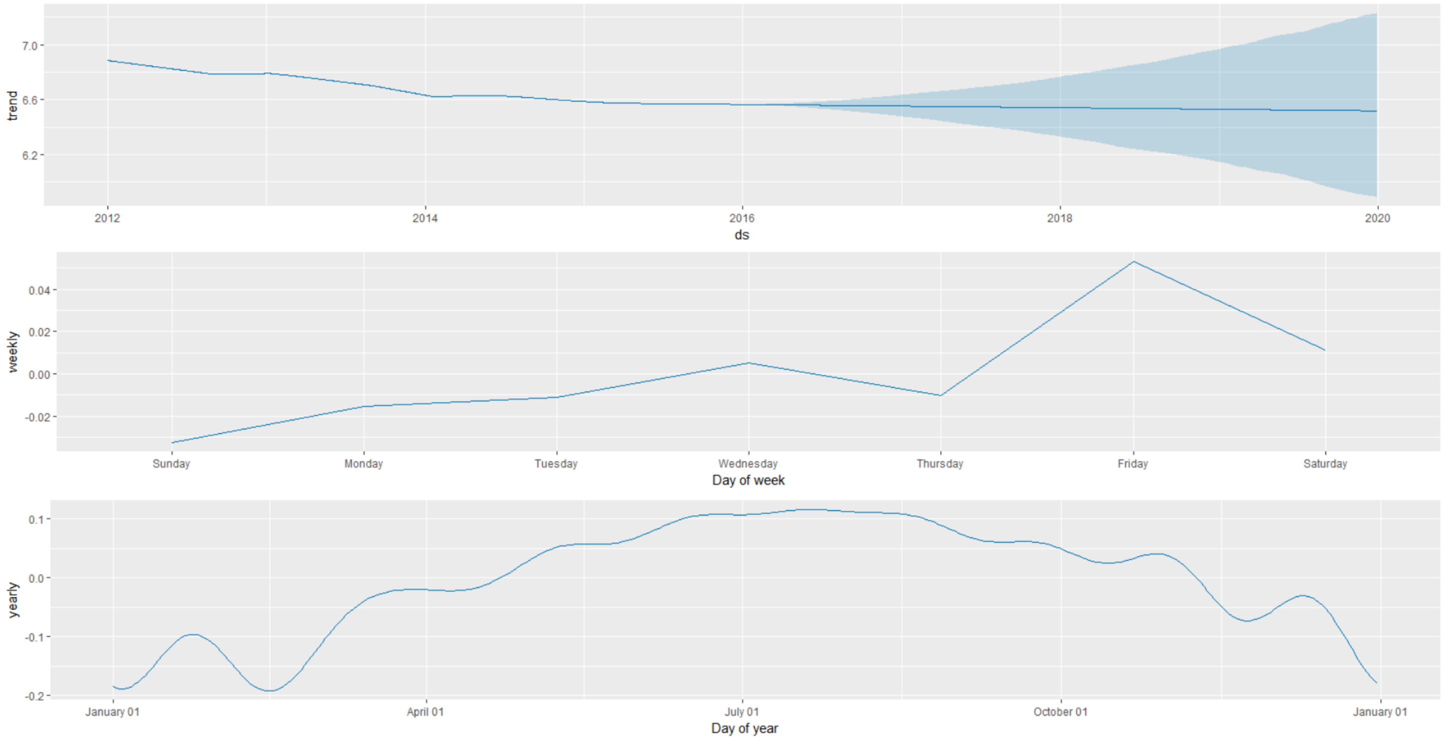


Surprisingly, as we can see that 2016 has the greatest number of Homicides than ever!

Forecasting Plot:



Forecast decomposition to trend, weekly and yearly



We can see that crime numbers will continue to grow again significantly!

Prediction Model Evaluation

- To measure mode accuracy, we used the year 2016 existing records as the cross-validation data and compared it to the predicted values got by the model.
- We used the “Root mean square error” and the “Mean absolute difference” formulas to measure error percentage.
- RSME % ~= 8%, MAPE % ~= 6%
- So, we can say the model is accurate by around 93%!

Future work enhancements

Other useful visualizations could be added like plotting the crime location on maps and indicating safe/un-safe points and warning against common crimes in each location plotted on map.
Moreover, we can add also a nice year timeline recommending safe days of the week. It’s really a promising dataset!