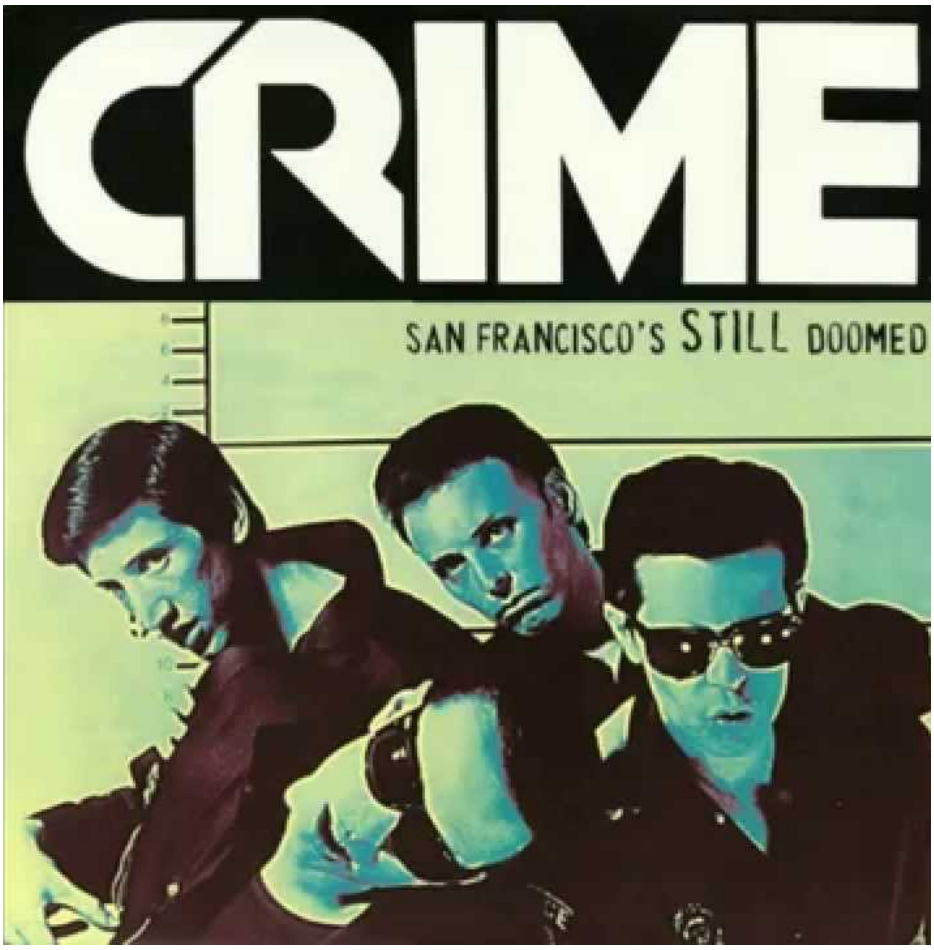
Crime in San Francisco

– A PySpark Analysis



By
Jonas Hellevang

What: Background description

In this assignment I have analyzed a dataset found on a governmental webpage about crime in San Francisco. I originally found a dataset with about 150.000 rows and 13 columns on Kaggle, but soon later found an updated dataset from the same category while searching on google. Why did I change from the "2016" to the "2018 until now" dataset? As you can imagine, I thought it was more interesting to work on newer data, as well as there is data from more than one year in the new dataset. The data was also continuously updated on the webpage, so I updated it last 17. February 2020 to include the freshest data at the end, although that did not change any of my results. The new dataset had also more than twice as many rows as the previous one and was supposed to have 26 columns according to where I collected the data, but when I did my analysis it turned out to be 36.

Why did I choose such a dataset? First of all, it was very important to me to work on an interesting dataset to boost my motivation for what could potentially be a steep-learning-curved assignment (which it was). Second, I have been to San Francisco and it is one of my favorite cities in the world. I find it especially interesting to work on something that is more familiar than a place I haven't visited yet such as Washington D.C. Third and last, crime is something I have always found interesting, watching every tv-show about it I can (until they made too many of them). Sit back and enjoy the ride, because now I will take you through a journey of details and insight!

Why: Goal of the Analysis

Although I wouldn't say that San Francisco is known for its crime, rather for its bridge and steep hills and houses. Still crime is everywhere, and it is interesting to see how this can be reduced. So how can we help the police of San Francisco, or SFPD, with its crime fighting? I have focused the analysis on location, time and severity of crime, and by doing so we can provide SFPD with facts such as in what areas there is most crime, what day crime happens the most and where.

How: Analysis deep dive

San Francisco Crime Analysis (2018 until today)

San Francisco Crime Analysis is going to be performed as follows:

1. PySpark **environment setup**
2. Data source and **Spark data abstraction** (DataFrame) **set up**
3. Data set **metadata analysis**:
   A. Display **schema and size** of the DataFrame
   B. Changing the **names of**, and **deleting** Columns
   C. Changing **DayOfWeek** from string to integer
   D. Changing **time related** columns
   E. Get three **random samples** from the data set to better understand what the data is all about
   F. Identify **data entities**, **metrics** and **dimensions**
   G. **Column/field categorization**
4. Columns groups **basic profiling** to better understand our data set:
   A. **Timing related** columns basic profiling
   B. **Crime related** columns basic profiling
   C. **Location related** columns basic profiling
5. **Answer to business questions:** How can SFPD make the city of San Francisco safer?
   A. **Ratio of crime based on severity**
   B. **Severity of crimes and their location** by level of severity sorted by the most severe crimes (9)
      b. Severity of crimes and their location
      c. Severity of crimes and their location including most dangerous days of the week
   C. **Top 10 dangerous intersections with severe crime**
      c. Top 10 dangerous intersections with any severe crime
      d. Top 10 dangerous intersections with any severe crime including days of the week

1. PySpark environmental setup is all about setting up the environment to run Spark on Python smoothly. Here we set up the spark session and spark context.
2. Our second step is to import our data into the PySpark environment and saving it into a data frame.
3. Data set metadata analysis:
   a. Display schema and size of the data frame: Here is where you can see all columns and rows in the dataset, and I included a count of columns just because I thought it looked like more than 26 columns, which I was right about because it was 36. Count of rows also turned out to be 318,743.
   b. In this step I changed the name of columns and deleted a few rows such as Latitude, Longitude, Point and Analysis Neighborhood as they would not be useful for my analysis. I also renamed almost every column to make it easier to maneuver through and understand in the code.

c. Changing DayOfWeek column to integer was an important step to do summary analysis. It is also very intuitive to understand that the value 1 = Monday and so forth.
d. I had to change the time related columns for later analyzes I was going to make. For instance, it was very important to be able to tell which hour crime happens in, not in which exact minute of the day, because there is a lot of distinct values there! 1440 to be exact! That's why it is easier to see within hours of the day when most crimes are happening. Also, I created columns for DayOfMonth and Month, as these were not included in the original dataset. Changing the date and time columns were one of the most challenging parts of this assignment, and it took a long time to grasp how this actually works with Spark, but with the help of internet and slides from the professor I was able to create new features of value for the analysis.
e. Here I took 3 samples of the dataset to see example values of the different columns. Why 3? To me it made sense to have a couple more than one in case it would give some interesting different results.
f. Before digging into the real analysis, it is important to understand what kind of data you are dealing with, therefore I classified different columns as entities, metrics and dimensions. The result can be viewed here:

Entities: IncidentID, PoliceDistrict, Neighborhood

Metrics: DateAndTime, Date, Month, DayOfMonth, Time, Hour, Year, DayOfWeek, ReportDateAndTime

Dimensions: RowID, IncidentNumber, CADNumber, ReportTypeCode, ReportTypeDescription, Online, IncidentCode, Category, Subcategory, Description, Resolution, Intersection, IntersectionID, PoliceDistrict, Neighborhood, SupervisorDistrict, NeighborhoodID, PoliceDistrictID, CurrentSupervisorDistrict, HSOCZones, OWEDPublicSpaces, CentralMarket/TenderloinBoundaryPolygon, ParksAllianceCPSI, ESNCAG-BoundaryFile and AreasOfVulnerability

g. After looking deeper into the identity of each column we have to see if we can further group them together for our analysis. In this dataset I have located 3 different categories, timing, crime and location related columns:

Timing related columns: DateAndTime, Date, Month, DayOfMonth, Time, Hour, Year, DayOfWeek and ReportDateAndTime
Crime related columns: IncidentID, CADNumber, ReportTypeCode, ReportTypeDescription, Online, IncidentCode, Category, Subcategory, Description and Resolution
Location related columns: Neighborhood, Intersection, IntersectionID, PoliceDistrict, Neighborhood, SupervisorDistrict, NeighborhoodID, PoliceDistrictID, CurrentSupervisorDistrict, HSOCZones, OWEDPublicSpaces, CentralMarket/TenderloinBoundaryPolygon, ParksAllianceCPSI, ESNCAG-BoundaryFile and AreasOfVulnerability

4. Our columns groups basic profiling is used to better understand our data by looking at statistics and checking for null values. Here we analyze based on our categories from step 3g.
    a. Timing related columns basic profiling: Here I used Date, Month, DayOfMonth, Time and Hour to get more insight into the data and found out that every hour and all other values in timing features are included, as well as there is no null-values present. We can also see that there is slightly more crime in the last six months of the year than the first, there is more crime from 12pm to 12am than the first part of the day, and the exact hour of most crime is 12pm. We can also see that the most frequent day of the week is 5, which can be nice to remember for later exploration of the dataset.
    b. Crime related columns basic profiling: Here I explored directly crime related columns and found out CADNumber and Online had many missing values, but that can easily be explained by CADNumbers not always being present on purpose, and that the columns that aren't filed online is marked as the opposite. We can also see that there is 20 rows does not have a category of crime, and that there is 75 different Subcategory values. We can also see that

the most frequent crime is larceny theft, and to be even more specific, theft from a vehicle. Maybe people should try to hide their valuables better in cars?

c. Location related columns basic profiling: To be able to answer the business questions we not only need to know somethings about the time and crime, we also need to know a little bit more about the areas. We can see that there is around 17,000 missing values in most categories, there are 11 police districts, and most importantly for our further analysis we have

5. Over to the big question; **How can SFPD make the city of San Francisco safer**? Well, by identifying where the most dangerous crimes occur, when they typically happen and the distribution of severity, we can find patterns and we get a couple nice findings.

   a. Ratio of crime based on severity: First of all, there is no column that is called Severity. That tells us that we have to make it our self. How did I do that? By creating a new column in the existing data frame where all 75 different subcategories of crime are split into 10 different levels where 0 is not criminal, 1 is very low crime, and 9 being life sentence. To be able to categorize these crimes I used a document I found from a US website with crime related to prison sentence (https://pap.georgia.gov/sites/pap.georgia.gov/files/CSL-s_Post_1-1-2006_considerations.pdf). We can see that 372 of the cases can lead to life sentence in prison, 50,502 of the cases can lead to almost no punishment and even 83,123 of the cases are not even criminal.

The next two categories are divided in two of the sole purpose to investigate location and severity of crime on one side, and then repeat the process only this time including DayOfWeek.

   b. Severity of crimes is sorted by highest value of severity for distinct intersections around town. From there you can see that where the most severe crimes happen is not necessarily where the other types of crimes are happening and vice versa. From there on including the day of week it looks like Thursday, Friday and Saturday are typical days where the most severe crimes happen at those districts.

   c. To find the top 10 dangerous intersections with severe crime all degrees of severity was included from 6 to 9, where they combined and formed the intersections where most crime happened in general. From the results you can

see that you should most definitely steer away from Boardman PL \ Bryant St! When including the day of the week as well you can tell that no matter what day of the week, you do not voluntarily go there unless you have to.

If the assignment had been of an even bigger scale it would be interesting to look into questions such as what police district has the highest workload, how should their resources be divided between districts and more, but what we really are interested in knowing now is what insight did I obtain from this analysis?

Insights: Conclusion

First of all, not all intersections are dangerous even though crime might happen there, and you should also think about what crimes are not too scary for you. For example, in my analysis of the severity of the crimes I did not include those crimes categorized under level 6. This is due to the fact that those crimes are not the ones you actually should be afraid of. What you should be more concerned about is the ones values 6, 7, 8 and 9. In my analysis you can see that these crimes happen many places, and with a few more tools you will be able to see actually how close from each other some of these occurrences are. What if all of them were focused in the same area, which most likely many of them are? This opens a question that could be interesting for someone to further investigate. SFPD needs to be where the crimes happen more so than others to prevent them for happening. By knowing the time, quantity, frequency and location of these crimes SFPD can really plan well their police schedules, as well as taking security measures around those intersections.

Link to dataset: https://data.sfgov.org/Public-Safety/Police-Department-Incident-Reports-2018-to-Present/wg3w-h783