White Paper

By Achraf Safsafi

Topic:

Crime Analysis using Unsupervised Machine Learning.

Business Problem:

When we need to visit a city or want to go to a particular location or move to live in a specific place, we must have prior knowledge of the intended places, such as the level of safety, especially if we are unfamiliar with those areas. Therefore, crime analysis is a technical method that can benefit us in that. It can identify criminal patterns and trends, and thus we would be aware of the level of safety in those places we want to visit. So dividing the city zones into some clusters similar to each other in terms of safety level could help avoid some of the risks and crimes that could be occurring in our society.

Background/History:

Our societies have known crimes since we have lived on the earth. Predicting crimes has always been based on traditional methods such as random patrols; Therefore, crime avoidance was limited because it relied only on help calls or human intuition. However, nowadays, with our ability to collect and analyze data, we can use historical crime data to recognize criminal patterns and trends then avoid crime before happening.

Implementation Plan:

1) Problem understanding.
2) Data understanding.
    - Describing data.
    - Exploring data.
3) Data preparation:
    - Cleaning data.
    - Selecting data.
4) Modeling
    - K-means clustering.
5) Model evaluation

Data Explanation:

In this project, the dataset is obtained from data.sfgov.org. The data is published by the City and County of San Francisco. This dataset includes historical incident reports from 2003 to May 2018 across all San Francisco's neighborhoods with 14 columns and around 2.1 M rows, where each row is an incident report. Only historical data from 2014 to 2017 (518160 rows) is used in this project.

The following table shows the columns description:

Column	Description
PdId	Unique Identifier for use in update and insert operations
IncidntNum	Incident number
Incident Code	Incident code
Category	Category of the crime incident
Descript	Description of the crime incident
DayOfWeek	Day of week
Date	The date of the crime incident
Time	Time of the crime incident
PdDistrict	Name of the Police Department District
Resolution	Resolution of the crime incident
Address	Street address of the crime incident
X	Longitude
Y	Latitude
location	Location of the crime incident

Figure 1 shows the distribution of the crime categories that occurred from 2014 to 2017 across all San Francisco's neighborhoods. The "NON-CRIMINAL" category is removed from the analysis because only criminal incident types would be analyzed. Larceny/Theft crime is considered the most crime type in that time, followed by crimes classified under the Other Offenses category then assault.

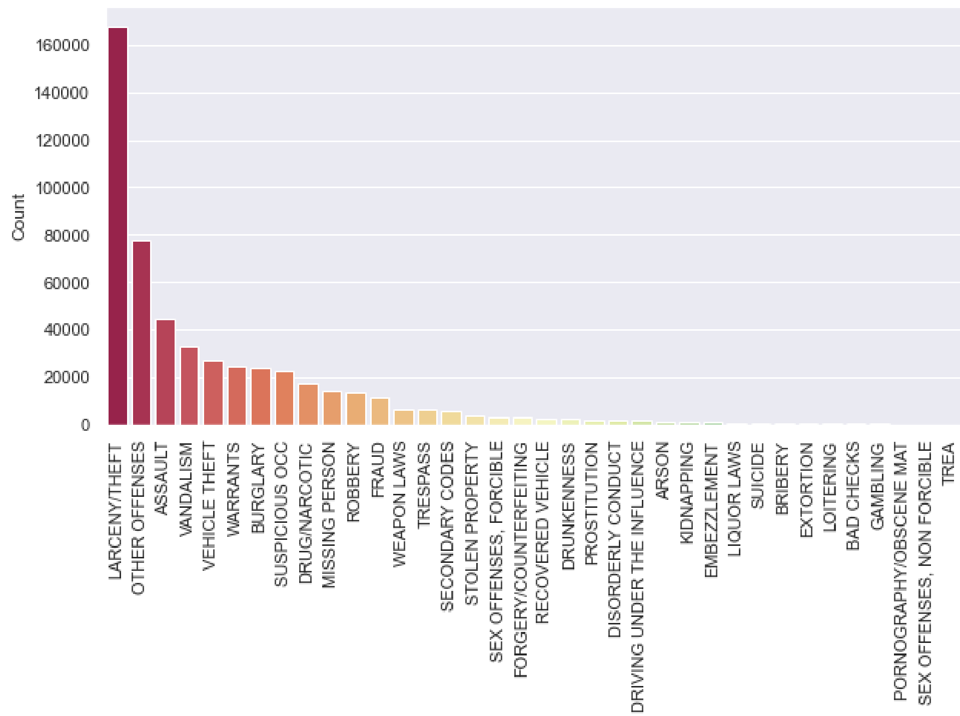

Figure 1 : Crime categories distribution in SF between 2014-2017.

Figure 2 indicates that crimes occur throughout the week, but crime incidents are the highest Friday and Saturday.

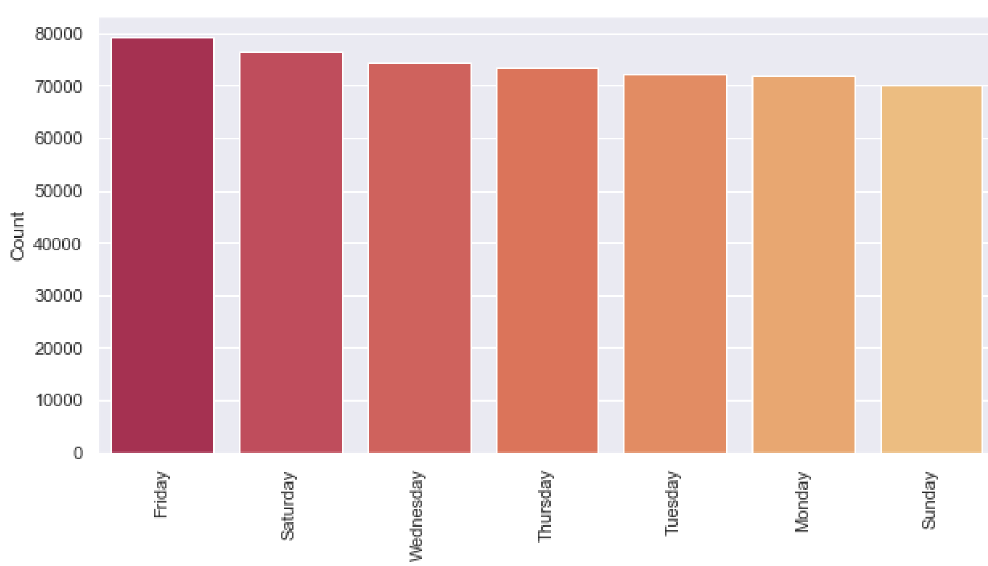

Figure 2: Total number of crimes by weekday in San Francisco County between 2014-2017.

Figure 3 shows that the Southern District has the highest number of crime incidents between 2014 and 2017. However, Park District has the lowest number of crime incidents in the same period.

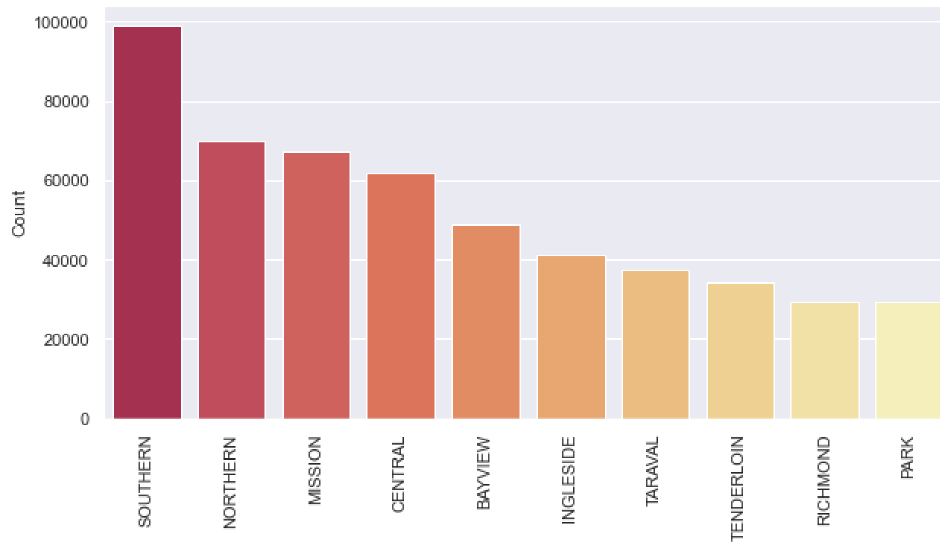

Figure 3: Total number of crimes by districts in San Francisco between 2014-2017.

Methods:

After preparing data, various visualization techniques are applied to assist crime analysis. Cluster analysis is performed on the dataset to cluster San Francisco areas into groups with similar characteristics. Since the features used in this analysis are all numerical (number of crime incidents of each category), the K-means clustering algorithm is applied. The method needs the number of clusters to be picked before starting the analysis. The Elbow technique is used to help find the optimal number of clusters in the dataset.

Analysis:

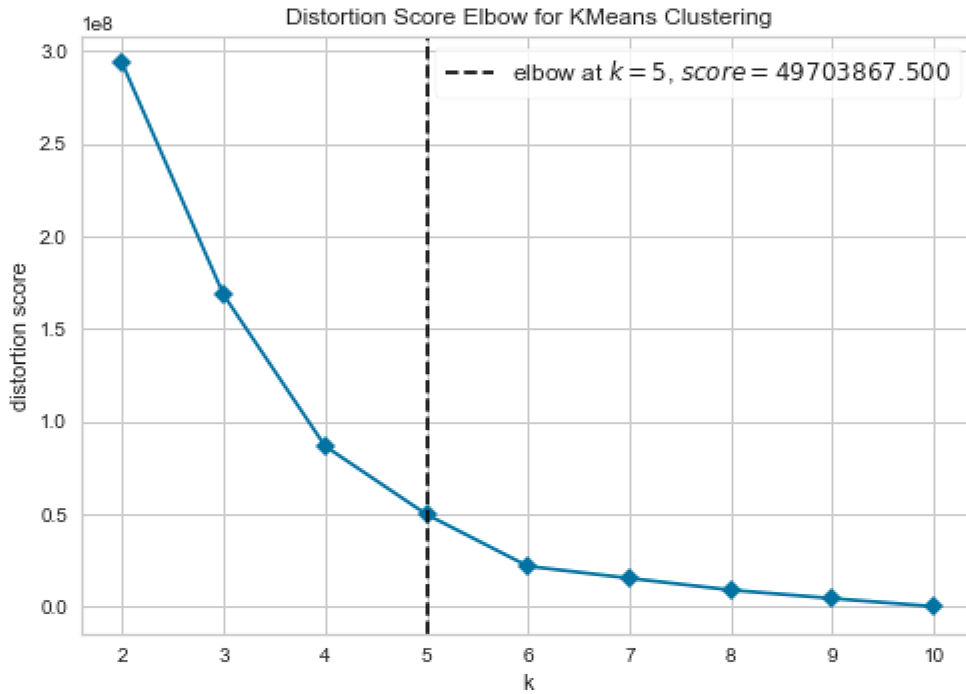

Figure 4: The Elbow method using distortion score.

Figure 4 shows the plot of the Elbow method using distortion score. This elbow method runs K-means clustering on the dataset for a range of values from 1 to 10. The optimal number of clusters K at the point where the "elbow" is seen. This is the same point where the distortion score starts dropping linearly. Based on the plot, the optimal number of clusters K is 5.

Figure 5 shows the size of the clusters. Cluster 0 contains four districts: Park, Richmond, Taraval, and Tenderloin District. Cluster 1 has only one district, Southern District. Cluster 2 includes two districts, Central and Northern District. Cluster 3 has one district, Mission District.

Cluster 4 contains two districts, Bayview and Ingleside District. Table 1 gives more details about those 5 clusters.

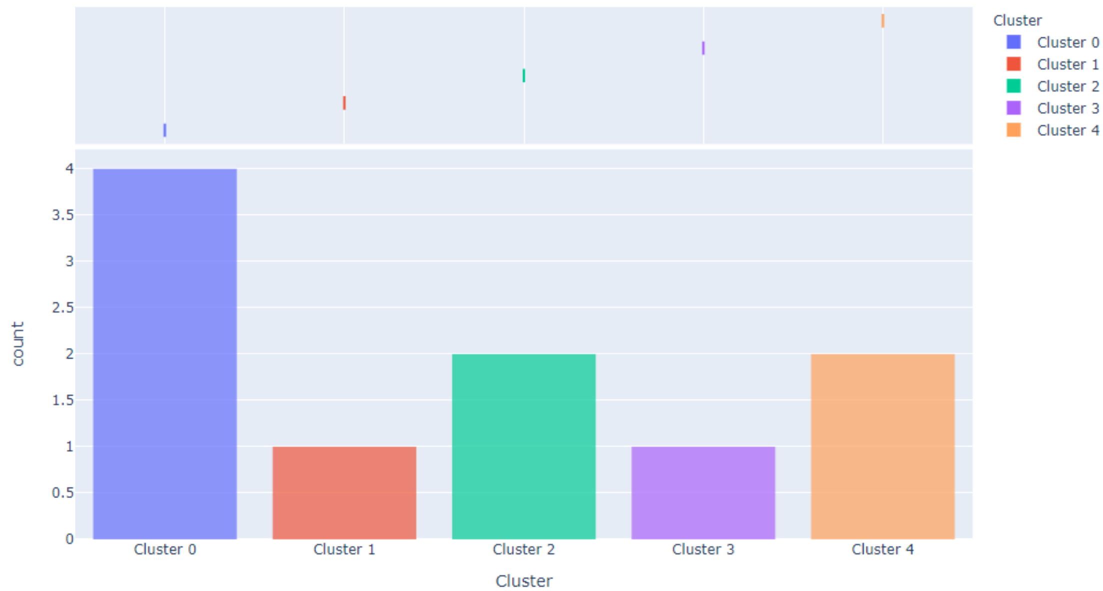

Figure 5: The size of clusters

Table 1 shows the total crime incident number by the district. Districts of cluster 0 have the lowest number of crimes, followed by districts of cluster 4, then districts of cluster 2. Districts of cluster 1 have the highest number of crimes in the period from 2014 to 2017 compared to all San Francisco's neighborhoods. Districts of cluster 0 have lower crime rates in all crime categories, except drug/narcotic, extortion, pornography/obscene mat, and sex offenses (non forcible) which recorded

District	Cluster	Total crime incident number by district
Park	Cluster 0	29255
Richmond	Cluster 0	29340
Taraval	Cluster 0	37222
Tenderloin	Cluster 0	34123
Southern	Cluster 1	99124
Central	Cluster 2	61890
Northern	Cluster 2	70002
Mission	Cluster 3	67301
Bayview	Cluster 4	48654
Ingleside	Cluster 4	41248

Table 1: Total crime incident number by SF districts (2014-2017)

the highest rate in Tenderloin, Taraval, Richmond, and Park, respectively. Cluster 1, Southern District, has the highest number of crimes for about 40% of the crime types. These crimes include assault, bad checks, embezzlement, forgery/counterfeiting, fraud, burglary, larceny/theft, missing person, other offenses, robbery, stolen property, suspicious occ, trespass, vandalism, and warrants. In the same period, cluster 2 districts reported crimes with significant numbers, but these numbers are not the highest within any of the numbers reported across all San Francisco areas. Cluster 3, Mission District, has the highest number of crimes for disorderly conduct, driving under the influence, drunkenness, kidnapping, liquor laws, loitering, prostitution, and sex offenses (forcible). Districts of Cluster 4 recorded high crime rates in other crime types. The most significant number of gambling, suicide and vehicle theft incidents were recorded in Ingleside District. Furthermore, The largest number of arson, bribery, pornography/obscene mat, recovered vehicle, secondary codes, trea, and weapon laws incidents were recorded in Bayview District. For more details, see the table in Appendix A.

Conclusion:

By applying K-means clustering, San Francisco city is divided into five similar areas based on the crime category and crime rates recorded between 2014 to 2017. Such analysis will help identify the hot spots in terms of crime rates.

References:

DataSF. (n.d.). Police department incident reports: Historical 2003 to May 2018. Retrieved from https://data.sfgov.org/Public-Safety/Police-Department-Incident-Reports-Historical-2003/tmnf-yvry

Brownlee, J. (2020, August 20). 10 clustering algorithms with Python. Retrieved from https://machinelearningmastery.com/clustering-algorithms-with-python/

Mahmud, S., Nuha, M., & Sattar, A. (2021, January 1). (PDF) Crime rate prediction using machine learning and data mining. Retrieved from https://www.researchgate.net/publication/347219439_Crime_Rate_Prediction_Using_Machine_Learning_and_Data_Mining

Appendix A

The Highlight of the maximum and the minimum values of the crime incidents per crime category in SF Districts between 2014 – 2017

Category/PdDistrict	BAYVIEW	CENTRAL	INGLESIDE	MISSION	NORTHERN	PARK	RICHMOND	SOUTHERN	TARAVAL	TENDERLOIN
ARSON	257	102	93	184	120	51	50	170	86	56
ASSAULT	5354	4255	4408	7331	4925	1848	1549	8033	2722	4222
BAD CHECKS	10	12	9	18	23	2	11	25	18	2
BRIBERY	58	16	35	50	25	5	6	36	15	13
BURGLARY	2188	2989	2144	2878	3440	1712	1574	3535	2400	695
DISORDERLY CONDUCT	169	160	47	535	253	79	36	383	67	213
DRIVING UNDER THE INFLUENCE	132	97	132	265	129	96	211	220	161	40
DRUG/NARCOTIC	1384	911	854	2519	2147	1009	395	3315	465	3986
DRUNKENNESS	123	237	90	466	183	91	96	372	165	172
EMBEZZLEMENT	80	105	34	72	60	22	22	177	59	54
EXTORTION	21	23	22	21	17	8	19	26	27	9
FORGERY/COUNTERFEITING	149	379	177	363	384	116	154	571	174	178
FRAUD	554	1919	664	1356	1521	598	781	2029	1156	709
GAMBLING	16	5	18	9	3	0	0	5	2	12
KIDNAPPING	149	90	158	165	105	33	39	138	77	96
LARCENY/THEFT	9321	27576	7889	14494	29881	9227	11663	40063	10467	7337
LIQUOR LAWS	50	37	21	92	36	31	6	54	16	59
LOITERING	6	21	7	33	22	1	3	25	7	11
MISSING PERSON	1339	1061	1450	1954	1209	2044	754	2261	1560	649
OTHER OFFENSES	9813	6441	7899	11736	8290	4345	4375	12672	6902	5131
PORNOGRAPHY/OBSCENE MAT	4	1	2	0	1	0	4	0	2	2
PROSTITUTION	19	378	28	525	223	6	24	443	266	65
RECOVERED VEHICLE	546	143	368	196	249	96	106	329	170	157
ROBBERY	1264	1573	1394	2326	1518	480	458	2430	695	1398
SECONDARY CODES	966	446	764	811	569	260	315	798	549	345
SEX OFFENSES, FORCIBLE	195	333	267	691	355	147	131	527	221	206
SEX OFFENSES, NON FORCIBLE	0	1	0	1	2	4	2	0	0	0
STOLEN PROPERTY	273	599	246	510	558	158	168	724	186	206
SUICIDE	15	42	48	34	40	26	12	30	35	21
SUSPICIOUS OCC	2425	2240	1981	3346	2292	1089	1448	3836	1802	1926
TREA	2	0	1	1	0	0	0	1	1	0
TRESPASS	428	754	274	1171	787	244	203	1204	375	480
VANDALISM	3845	4432	3118	3904	4337	1807	1890	5797	2572	1312
VEHICLE THEFT	3919	2061	4427	3955	3177	2037	1922	2644	2501	515
WARRANTS	2328	2025	1538	4145	2588	1370	746	5223	988	3260
WEAPON LAWS	1252	426	641	1144	533	213	167	1028	313	586
Cluster	Cluster 4	Cluster 2	Cluster 4	Cluster 3	Cluster 2	Cluster	Cluster 0	Cluster 1	Cluster 0	Cluster 0

: Min

: Max

Appendix B

Cluster PCA Plot