San Francisco City Crime Data Analysis

– **Special Report on Kidnapping Incidents**

Capstone project by: Jun Liu, March 2019

I. Problem and background

There have been 257 kidnapping incidents occurring in city of San Francisco in 2016. Given the limited police force resource, the police department has contacted the data scientist to address the following questions:

- *Where have most kidnapping incidents occurred and is there a seasonal pattern (weekday, time) for example?*
- *Is there a pattern in terms of criminal choice of favourite location (eg. common venue) for these crimes and is there a rational explanation for such pattern?*
- *Could we get some insight about the time vs. the location of kidnapping in order to improve police's time allocation more efficiently for crime prevention (i.e. patrolling certain type of location at certain time of the day?)*
- *Can we derive information about the neighbourhood without prior knowledge of the incident area?*

II. The data

The San Francisco City Police Department has provided the data scientist raw crime information from their database. It consists of a csv file containing over 150,000 crimes of various types recorded in 2016. Each crime incident has its detailed date, time, location, geographic coordinate information and the data scientist will clean & normalize the data if needed, filter out relevant crime type for this analysis, get the venue information (via Four Square API) for each crime location and conduct detailed analysis to derive business insight and answer the police department's questions.

III. Methodology

1. *Overall Crime information in San Francisco*

We imported the data into Jupyter Notebook and started with slicing & dicing the relevant data by groups and presenting the results graphically to understand the data better. For example, we have done some exploratory analysis to look at the overall crime statistics:

Figure 1(A) % of Crime Incidents by Crime Types (Top 10)

	Category	SeriousCrimeTag	Count	pct%
0	LARCENY/THEFT	N	40409	40.0
1	ASSAULT	Y	13577	13.4
2	VANDALISM	N	8589	8.5
3	VEHICLE THEFT	N	6419	6.4
4	WARRANTS	N	5914	5.9
5	BURGLARY	N	5802	5.7
6	DRUG/NARCOTIC	N	4243	4.2
7	ROBBERY	Y	3299	3.3
8	FRAUD	N	2635	2.6
9	TRESPASS	N	1812	1.8

There are 34 labelled crime types in total.

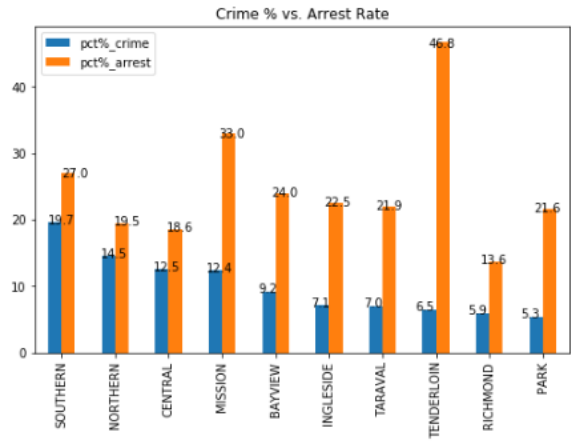
Top 5 crimes types covered 74.2% of all crime incidents and Herfindahl index value is 2005 which confirms moderate concentration of the crimes.

Most of the top 10 crime types (covered over 90% of the crimes incidents) are non-serious offenses.

Note: We cleaned the original crime dataset by removing incidents labelled as "non-crime" (such as lost item report) or "unspecified crime".

It's also interesting to group crime incidents by Police District to see how it is compared with actually arrest rate. Would we observe some anomaly like high crime rate coupled with low arrest rate or vice versa?

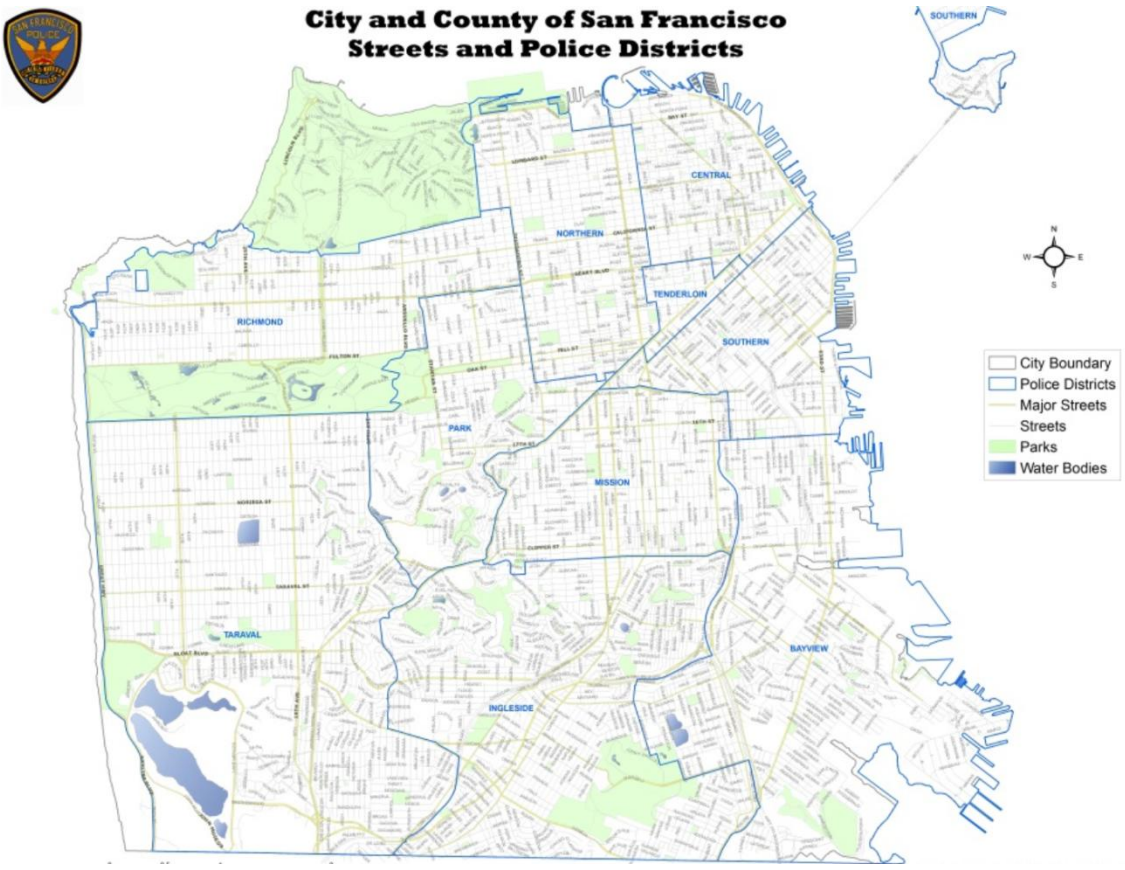
Figure 1(B) Crime & Arrest Rate by Police District



There are 10 police districts in San Francisco with Southern district having the highest crime incident rate 19.7%. Without any prior information about San Francisco, we could guess it's a troubled neighbourhood linked with poverty and high unemployment.

Tenderloin has the highest arrest rate made despite having relative small crime incidents proportion. One might conclude that the police force is particularly effective in that district. But it's probably simply due to fact that Tenderloin is one of the smallest police district in San Francisco, which made police work easier.

Figure 1(C) San Francisco Police District Map



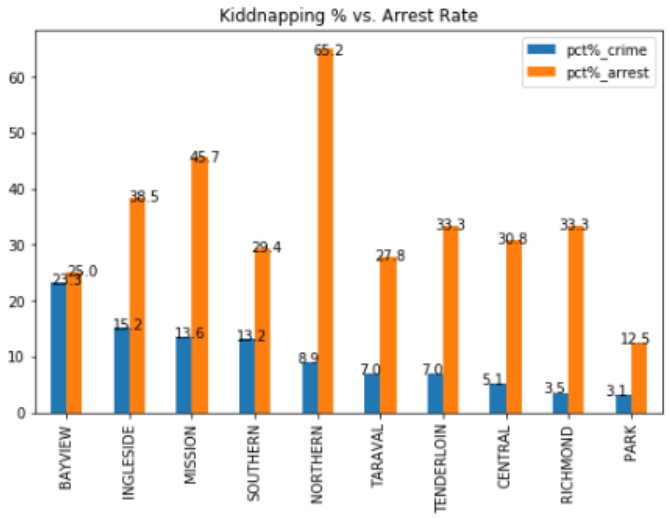
Source: http://sanfranciscopolice.org/police-district-maps

2. *Diving into the Kidnapping section of the crime data*

OVERVIEW

Next we start to look into the key topic of this study, i.e. the crime type "Kidnapping". We first apply the same grouping analysis like above to see which police district has the highest incidents of kidnapping and how it is compared with the actual arrest rate

Figure 2(A) Kidnapping & Arrest Rate by Police District



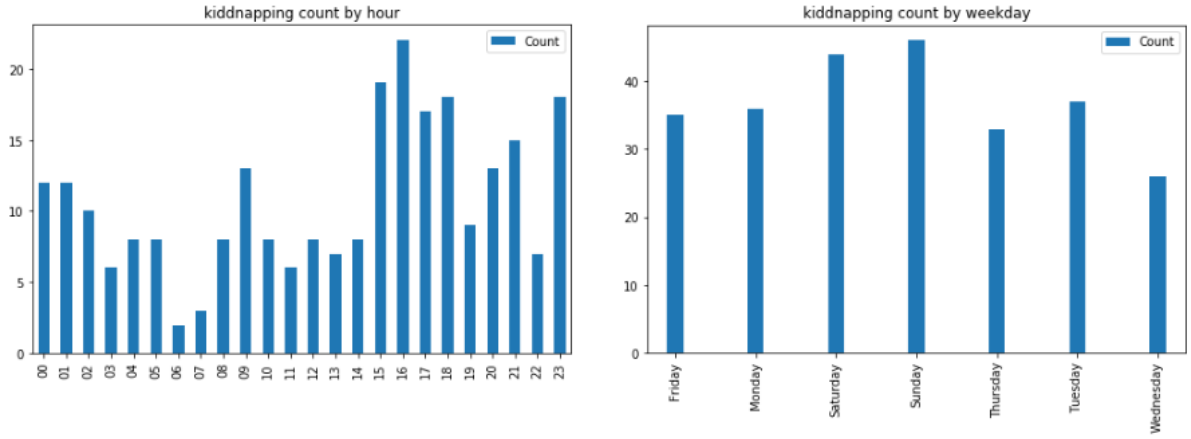
Bayview has the highest % of kidnapping incidents. Bayview is located in Easter coast of San Francisco city. Even without prior information about the region, the common sense tells us that it should be an expensive area to live. So it might make sense to see high kidnapping rate (presumably for ransom money) as this could be a popular target area. But could there be other reason why Bayview stands out with high kidnapping rate vs. other rich neighbourhood?

We observe impressive arrest rate related to kidnapping case in Nothern police district. It might be helpful to conduct additional interview to understand how that was achieved and whether it could be applied to other police districts.

SEASONALITY

As the next step we looked at the crime selection of kidnapping category only and to see whether there is any seasonality related to kidnapping – eg. Are kidnapping more likely to happen over weekend? Do kidnapping tend to occur at certain time of the day?

Figure 2(B) Seasonality (weekday, hour of the day) about kidnapping Incidents

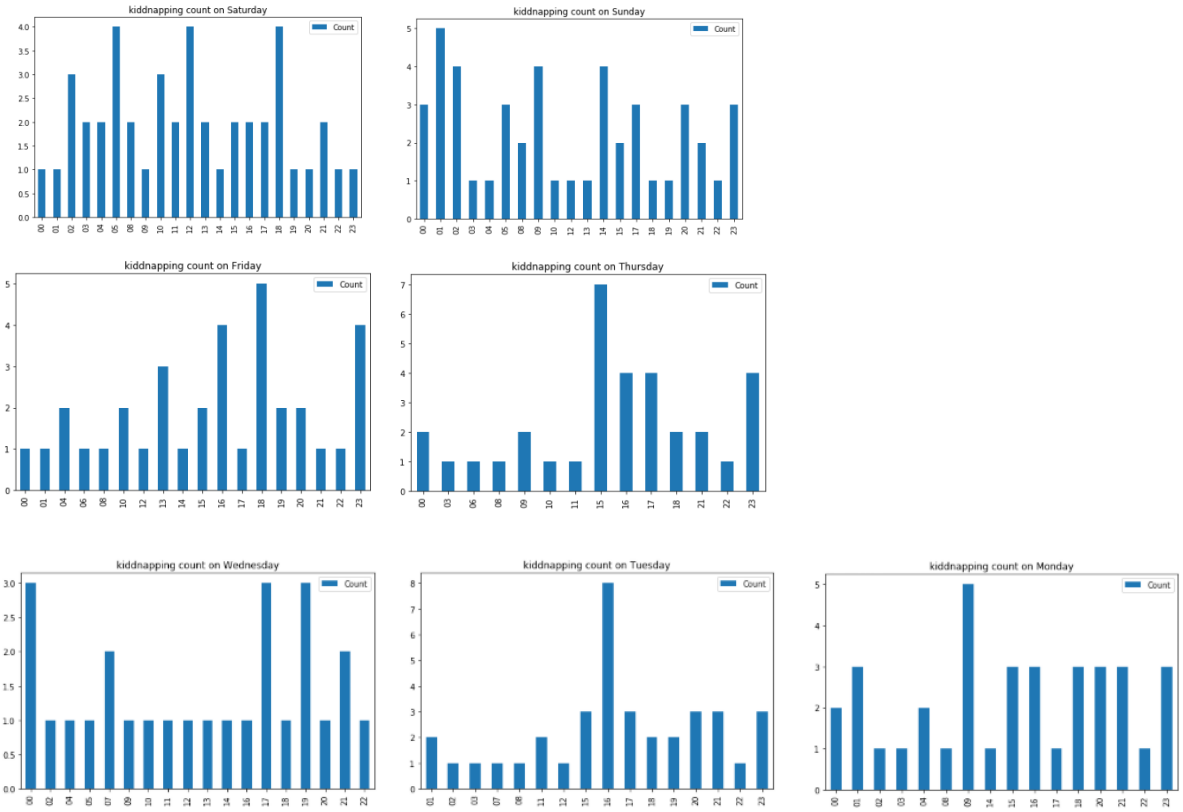


In terms of the hour of kidnapping incident, we notice a high density of incidents occurring between 15:00 – 18:00. It means that kidnapper's favourite choice of is the time window between finishing school/work and going home. Prime evening time 20:00 -21:00 when the targets probably have just finished a dinner/drink night out or eve midnight when there is likely to be fewer witness is also a popular choice. There doesn't seem to be a strong

seasonality around weekday although weekends do seem to have higher chance of kidnapping incidents occurring.

We also try to find any seasonality around incidents hours by weekday in order to see whether there are any cluster in incidents taken place in particular hour on a particular weekday/weekend. The results are shown in figure 2(c).

Figure 2(C) Seasonality of Kidnapping Incidents by particular Hour on a particular Day



Although there are indeed spikes around certain hours by different weekday, we don't have enough data set to make any meaningful statistical inferences such as "kidnapping is more likely to occur around 9am on Monday" or "kidnapping is more likely to occur around 1-2 am on Sunday". However, the kidnapping incidents do tend to occur more during at least one of the two occasions:

- **When victims are in transition between homes and work/school or vice versa**
- **When there are less risk of detection/witness/intervention (for example, midnight or early morning)**

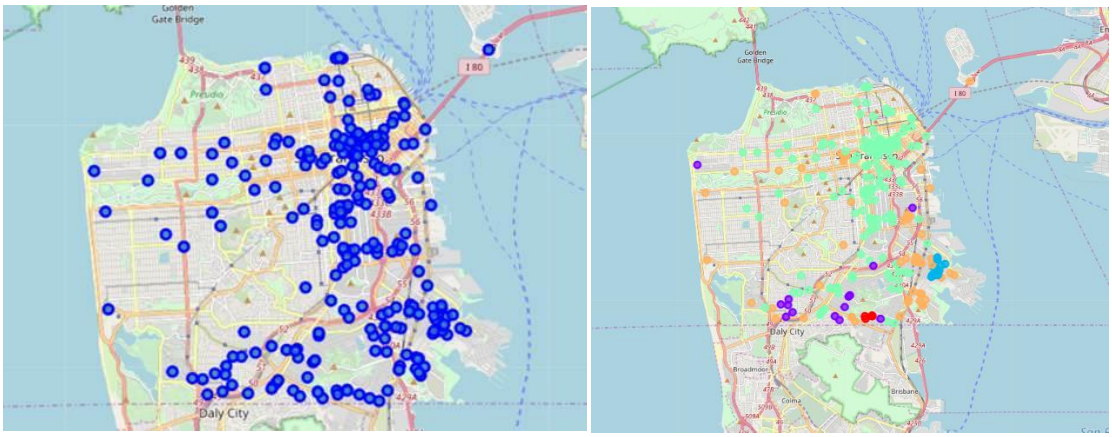
LOCATION DATA

Next we focus on the location data by downloading relevant venue information via Four Square API sorted by relative frequency. We record the top 10 common venues for each incident, and then run a K-mean clustering analysis on the data. With K-mean clustering, we hope to derive insight on whether there are any common venue features that make a location particularly attractive for kidnapper within the respective location cluster.

Figure 2(D) Data sample after getting location data via Four Square API

	PdDistrict	Latitude	Longitude	Neighbourhood	DayOfWeek	Hour	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	INGLESIDE	37.712200	-122.420864	INGLESIDE / 100 Block of BROOKDALE AV	Friday	08	4	Pool	Playground	Basketball Court	Bike Trail	Mexican Restaurant
1	SOUTHERN	37.784189	-122.407634	SOUTHERN / 800 Block of MARKET ST	Friday	16	2	Women's Store	Coffee Shop	Clothing Store	Toy / Game Store	Cosmetics Shop
2	INGLESIDE	37.723986	-122.435408	INGLESIDE / 4600 Block of MISSION ST	Friday	23	2	Chinese Restaurant	Mexican Restaurant	Bakery	Latin American Restaurant	Liquor Store
3	BAYVIEW	37.719033	-122.398004	BAYVIEW / 1000 Block of LECONTE AV	Saturday	12	3	Breakfast Spot	Historic Site	Mountain	Burger Joint	Park
4	BAYVIEW	37.729203	-122.374019	BAYVIEW / 700 Block of KIRKWOOD AV	Saturday	18	3	Harbor / Marina	Construction & Landscaping	Spa	Business Service	Zoo
5	CENTRAL	37.796903	-122.406832	CENTRAL / PACIFIC AV / GRANT AV	Saturday	05	2	Chinese Restaurant	Italian Restaurant	Dive Bar	Coffee Shop	Cocktail Bar
6	RICHMOND	37.780285	-122.477772	RICHMOND / 5400 Block of GEARY BL	Wednesday	11	2	Chinese Restaurant	Grocery Store	Sushi Restaurant	Mexican Restaurant	Café

Figure 2(E) Kidnapping incident location (left) vs. Kidnapping incident location in five clusters (Right)



3. A curious analysis inside the Clusters

Finally, we want to look into each of the k-mean cluster and see whether we can derive any insight without prior knowledge of the relevant area. We then check whether our findings reflect the reality by cross-checking with facts. The k-mean clustering divided the kidnapping location into five clusters. The clusters that are in particular interests are cluster 3 (light green dotted area in figure 2 (E)) and Cluster 4 (light orange dotted areas in figure 2 (E)) as both clusters covered 87.5% of all the kidnapping incidents.

Cluster 3 Analysis

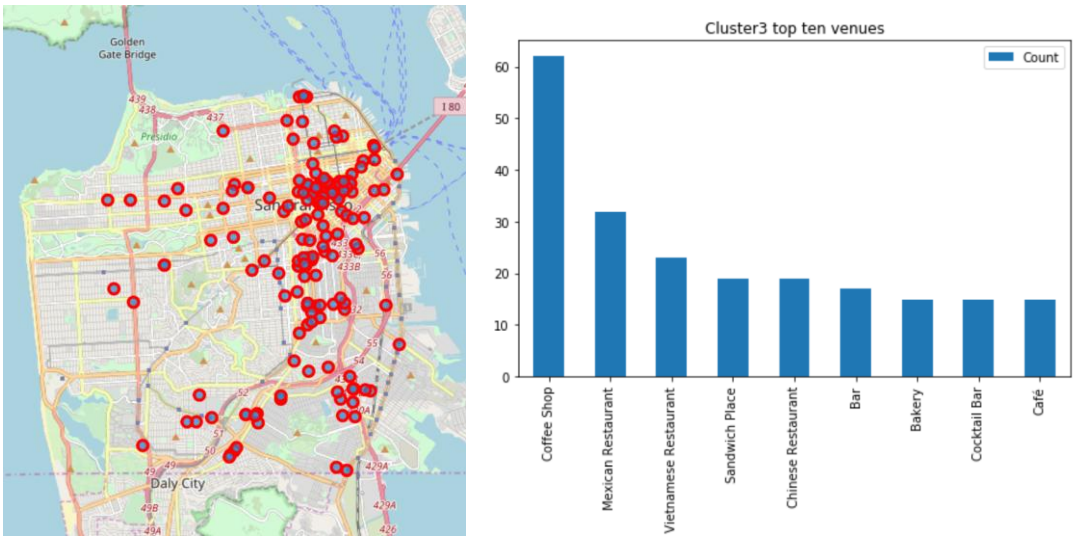
(167 incidents, note in below table it shows Cluster Labels 2 due to indexing starting from 0)

Below is a table of cluster 3 location feature data

	PdDistrict	DayOfWeek	Hour	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
1	SOUTHERN	Friday	16	2	Women's Store	Coffee Shop	Clothing Store	Toy / Game Store	Cosmetics Shop	Food Truck	Thai Restaurant	Marijuana Dispensary	Bubble Tea Shop	Department Store
2	INGLESIDE	Friday	23	2	Chinese Restaurant	Mexican Restaurant	Bakery	Latin American Restaurant	Liquor Store	Grocery Store	Sandwich Place	Vietnamese Restaurant	Japanese Restaurant	Pharmacy
5	CENTRAL	Saturday	5	2	Chinese Restaurant	Italian Restaurant	Dive Bar	Coffee Shop	Cocktail Bar	Bakery	Tea Room	Vietnamese Restaurant	Asian Restaurant	Szechuan Restaurant
6	RICHMOND	Wednesday	11	2	Chinese Restaurant	Grocery Store	Sushi Restaurant	Mexican Restaurant	Café	Dim Sum Restaurant	Vietnamese Restaurant	Korean Restaurant	Bubble Tea Shop	Bakery
7	BAYVIEW	Tuesday	20	2	Chinese Restaurant	Vietnamese Restaurant	Bakery	Grocery Store	Coffee Shop	Bubble Tea Shop	Bus Station	Sandwich Place	Dim Sum Restaurant	Recreation Center
8	SOUTHERN	Tuesday	16	2	Coffee Shop	Cocktail Bar	Beer Bar	Theater	Bakery	Marijuana Dispensary	Gym	Café	Brewery	Taco Place
9	BAYVIEW	Wednesday	22	2	Furniture / Home Store	Coffee Shop	American Restaurant	Art Gallery	Brewery	Gym / Fitness Center	Burger Joint	Café	Nightclub	Massage Studio
10	MISSION	Thursday	23	2	Mexican Restaurant	Coffee Shop	Bakery	Bookstore	Fish Market	Italian Restaurant	Latin American Restaurant	Performing Arts Venue	Public Art	South American Restaurant
15	NORTHERN	Sunday	15	2	Spa	Gym / Fitness Center	Bar	Café	Wine Shop	Park	Italian Restaurant	Mediterranean Restaurant	Liquor Store	French Restaurant
16	BAYVIEW	Saturday	12	2	Chinese Restaurant	Grocery Store	Pizza Place	Vietnamese Restaurant	Intersection	Pharmacy	Rental Car Location	Dim Sum Restaurant	Diner	Park

We plot the cluster 3 incident locations on the San-Francisco city map and also aggregate top three venue features by each location in cluster 3 into a common feature vector in order to see its distribution.

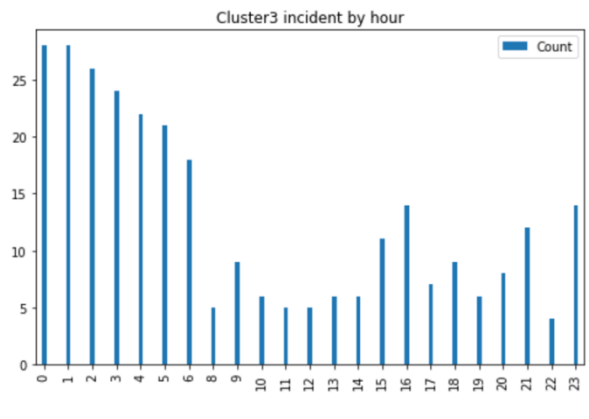
Figure 3(A) Kidnapping incident location (left) and most common venue features in cluster 3 (Right)

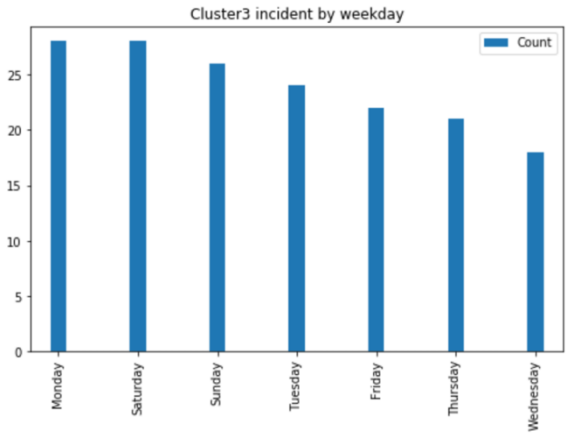


Most common venues within cluster 3 are coffee shops, Mexican/Vietnamese/Chinese restaurants & Sandwich bars, there doesn't seem to be a causality explanation why these locations would make it preferred kidnapping choices. However, these venue features do reflect something about the incidents areas. Given that these are mostly venues like fast food type of restaurants, it might be a sign of low income, high immigrant density (for example Asian or Hispanic population). These areas are commonly prone to crimes due to poverty. It's also plausible to assume that kidnapper might specifically target this area as low income area might have low media coverage and less police resource allocation.

Again we want to check the presence of seasonality in this cluster. **Interestingly there seems to be a high proportion of incidents occurring in early morning (midnight to 6 am). It might be effective to increase police patrol at those hours in these cluster areas.**

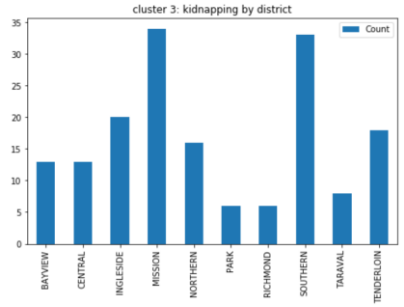
Figure 3(B) Kidnapping incident seasonality analysis (cluster 3)

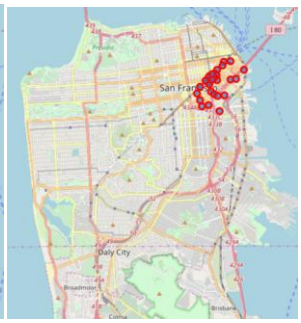




Finally, we take a look at the districts with most kidnapping incidents within this cluster.

Figure 3(C) Cluster 3 kidnapping incidents by district & selective mapping of district with high incidents

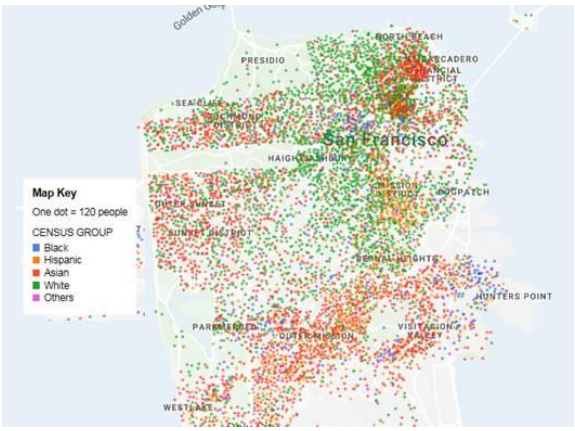


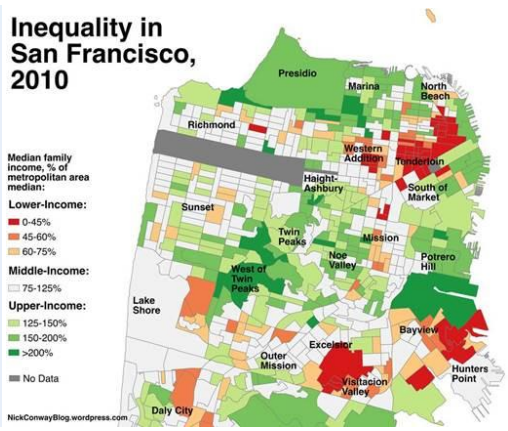


P.District: ***MISSION*** *P.District:* ***Southern***

Our hypothesis from analysing cluster 3 data has shown strong inference to potential poor economic condition in these areas. Hence we take a look at the income map of San Francisco to see whether this is the case using police district Mission and Southern as example. Indeed Southern and Mission district are known as low-income areas and are generally perceived to be less safe. In addition, there is indeed a higher population density for Hispanic and Asian residents in these areas. Another hypothesis we could pose here is that it's plausible to assume the purpose of kidnapping in this area might have to do with other crime related to human trafficking, drug etc.

Figure 3(D) Census Group Density (Left) and Income Inequality Map (right)





Cluster 4 Analysis

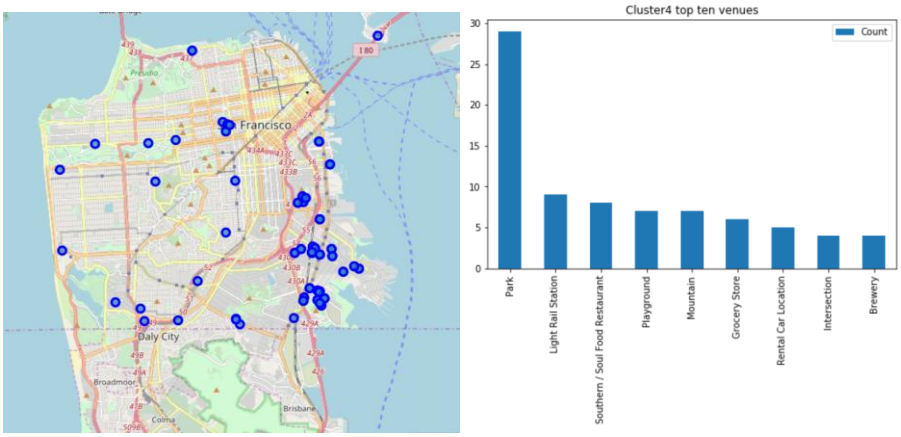
(58 incidents, note in below table it shows Cluster Labels 3 due to indexing starting from 0)

Below is a table of cluster 4 location feature data

	PdDistrict	DayOfWeek	Hour	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
3	BAYVIEW	Saturday	12	3	Breakfast Spot	Historic Site	Mountain	Burger Joint	Park	Bike Rental / Bike Share	Martial Arts Dojo	Farm	Event Space	Exhibit
4	BAYVIEW	Saturday	18	3	Harbor / Marina	Construction & Landscaping	Spa	Business Service	Zoo	Farm	English Restaurant	Ethiopian Restaurant	Event Space	Exhibit
12	BAYVIEW	Saturday	20	3	Southern / Soul Food Restaurant	Light Rail Station	Fried Chicken Joint	Bakery	Theater	BBQ Joint	Park	Market	Grocery Store	African Restaurant
13	PARK	Thursday	0	3	Park	Café	Deli / Bodega	Coffee Shop	Garden	Middle Eastern Restaurant	Sushi Restaurant	Bus Station	College Gym	Rental Car Location
29	TARAVAL	Sunday	21	3	Rental Car Location	Light Rail Station	Park	Burger Joint	Thai Restaurant	Gas Station	Garden	Laundromat	Gym	Mexican Restaurant
33	INGLESIDE	Sunday	18	3	Trail	Park	Dog Run	Grocery Store	Playground	Café	Scenic Lookout	Chinese Restaurant	Exhibit	Elementary School
35	BAYVIEW	Thursday	17	3	Grocery Store	Light Rail Station	Vietnamese Restaurant	Spa	BBQ Joint	Latin American Restaurant	Park	Business Service	Mexican Restaurant	Distillery
37	INGLESIDE	Sunday	19	3	Light Rail Station	Convenience Store	Vietnamese Restaurant	Café	Sandwich Place	Donut Shop	Park	Train Station	Coffee Shop	Breakfast Spot
41	NORTHERN	Sunday	14	3	Park	Indian Restaurant	Liquor Store	Playground	Record Shop	Coffee Shop	Roller Rink	Sandwich Place	Pakistani Restaurant	Dog Run
44	NORTHERN	Sunday	14	3	Park	Indian Restaurant	Liquor Store	Playground	Record Shop	Coffee Shop	Roller Rink	Sandwich Place	Pakistani Restaurant	Dog Run

Again we plot the cluster 4 incident locations on San-Francisco city map and also aggregate top three venue features by each location in cluster 3 into a common feature vector in order to see its distribution.

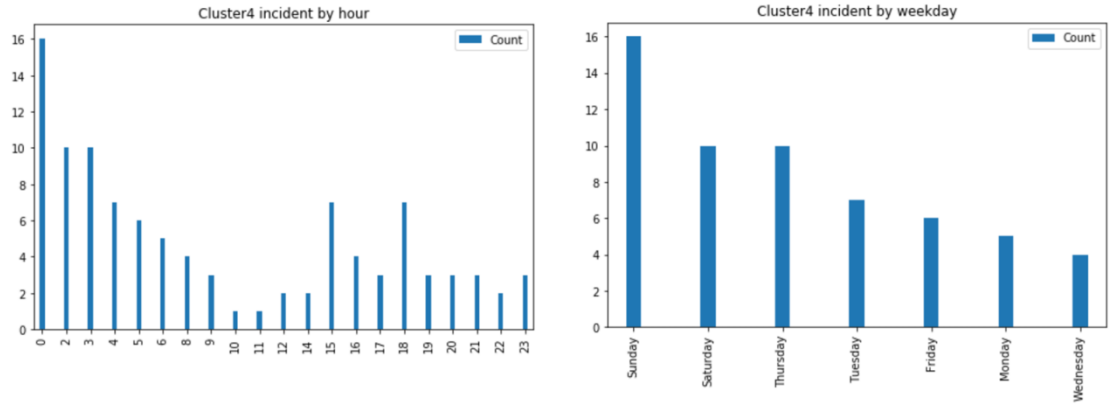
Figure 3(E) Kidnapping incident location (left) and most common venue features in cluster 4 (Right)



Most common venue within cluster 4 is Park! This is interesting. Indeed if we look at other common features like rail station, playground, mountain … we can conclude that this location cluster is marked with many outdoor areas which is typically large and less population density. This makes sense as Kidnapper would prefer to reduce the risk of being witnessed or interrupted with their attack.

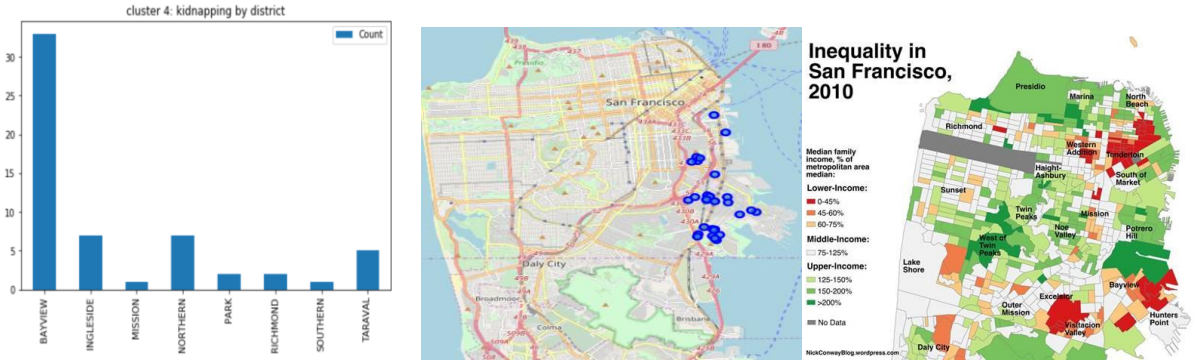
Further we want to check the presence of seasonality in this cluster. **Interestingly there seems to be a high proportion of incidents occurring in early morning (midnight – 6 am) as well as around afternoon. We will need more data to examine this as from rational point of view, for example it's puzzling why a potential victim would find him/herself in a park at these hours. It might be effective to increase police patrol at those hours in these cluster areas. There is also a higher tendency of incidents occurring on Sunday, presumably because more people tend to spend time in outdoor activities during weekend.**

Figure 3(F) Kidnapping incident seasonality analysis (cluster 4)



Finally, we take a look at the districts with most kidnapping incidents within this cluster.

Figure 3(G) Cluster 4 kidnapping incident by district, BAYVIEW map and Income Inequality Map



P.District: ***BAYVIEW***

Among all the locations, Bayview stands with highest incidents, potentially because of the presence of many outdoor places in the area as well as Bayview is typically populated with Upper-income household which make them more prone to be potential kidnapping target for ransom purpose.

IV. Results section

Based the data analysis conducted in the previous section, we can now address the client's questions:

a) ***Where have most kidnapping incidents occurred and is there a seasonal pattern (weekday, time) for example?***

About a quarter of the kidnapping incidents occurred in district BAYVIEW. The fact that it has many outdoor areas with less population density and is also an upper-class neighbourhood makes it an attractive kidnapping location presumably aimed at ransoms. The incidents seem to occur more often on Sunday and between midnight & dawn although we need more data to make a statistically significant conclusion about seasonality.

b) ***Is there a pattern in terms of criminal choice of favourite location (eg. common venue) for these crimes and is there a rational explanation for such pattern?***

In general we think there is a pattern, but it is highly dependent on the motives of the kidnapping. For example, ransom driven kidnapping is most likely to target upper-income neighbourhood with lots of outdoor area (less population density). This was proven in the clustering analysis using cluster 4 with district Bayview as example. Kidnapping linked to other crimes such as human-trafficking is more likely to occur in low-income neighbourhood and the common venue in those places are probably fast food type of restaurant, coffee, bar although the location venue information alone might not be sufficient without further details. Kidnapping driven by other motives such as murder, sexual assault might appear random in terms location venue features.

c) Could we get some insight about the time vs. the location of kidnapping in order to improve police's time allocation more efficiently for crime prevention (i.e. patrolling certain type of location at certain time of the day?)

Yes, in general the kidnapper would target when victims are in transition between home and work/school or vice versa or when there is less risk of detection/witness/intervention (for example, midnight or early morning).

So depending on the area type: In Bayview district, police force should consider patrolling more in the park in the midnight/early morning and on weekends. In areas like Southern or Mission, police force should focus on time that potential vulnerability time slot: for example, morning or late afternoon in certain radius around school area.

d) Can we derive information about the neighbourhood without prior knowledge of the incident area?

Possibly, for example we have shown in both cluster 3 and 4 that we managed to get more useful information about the incident area without prior information. However, as mentioned above it is highly dependent on motives of the kidnapping. Kidnapping driven by other motives such as murder, sexual assault might appear random in terms location venue features and hence we might not be able to derive meaningful insight about the incident area.

V. Discussion

In general, location venue data is useful to help police department generate more insight about the crimes as well as about the neighbourhood itself. K-mean clustering analysis is a powerful way to conduct unlabelled cluster analysis. One thing police should include in their future database is the victim information, for example, male/female, age group, ethnic group, profession etc. Combined with location data, this should help to generate even more insight, especially about potential motives or link seemingly unrelated kidnapping incidents into the same group.

Another observation is the arrest rate related to kidnapping cases. Some police district like Northern has impressive arrest rate related to kidnapping case. There should be some cross-communication between police district to share their experience/method in tracking down kidnapper in timely fashion.

VI. Conclusion.

In this report we have used crime data from San Francisco to analyze in particular kidnapping related crimes. We used explorative data analysis to understand the crime picture in general in San Francisco and then focused on crime type "Kidnapping". We combined police report information about day & time of the incident, combined with location venue feature data via Four Square API to generate a large feature set for each incident location. We then run K-mean analysis to cluster these incidents into five clusters in order to derive useful insight about the cluster area, crime pattern, seasonality, motive and suggestions to optimize police patrolling activities. We have shown location data is a powerful addition to the traditional police data collection and can really help the police division to generate new insight into the crime in the era of big data.