Question 1: Introduction where you discuss the business problem and who would be interested in this project.

Answer: Discussed different use case with colleagues who are working on ML and giving such exam and I found interesting to solve this use case where an investor wants to open a new restaurants in San Francisco ,California.

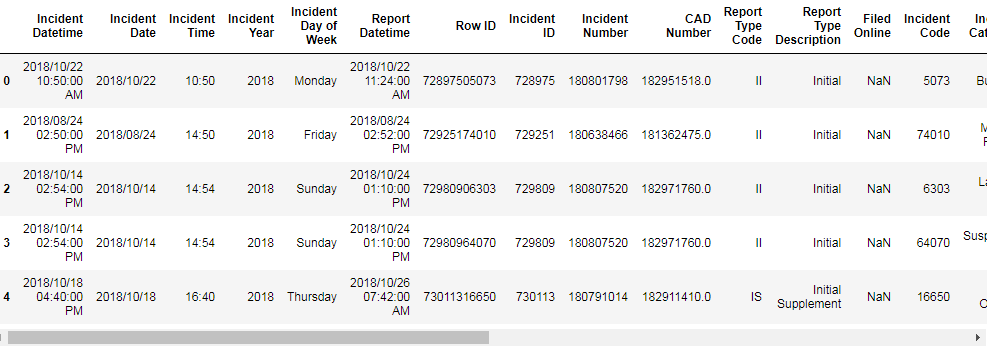
An investor is looking to open a new restaurant in San Francisco, but he is not sure about the best location for his new venue. So you got a call from him asking your input to help him choose the location. San Francisco is a very busy city, best known for tourist attractions and business innovation. Strolling around the city blocks, it pretty easy to notice that the city already has a lot of restaurants in town. How should we proceed and decide the location?

Following this line of thinking, the basic strategy here is then to know what are the most critical factors that contribute to the restaurant’s profitability. According to a report written by Tom Larkin published in the FSR magazine, these components stand out as the most important ones: visibility, parking, space size, crime rates, surrounding businesses and competitor analysis, accessibility, affordability, and safety. Using public datasets, we could actually address some of these considerations pretty straightforwardly.

Question 2:Data where you describe the data that will be used to solve the problem and the source of the data.

Answer: I have downloaded different data from Kaggle for San Franciso as mentioned below :

df\_crime = pd.read\_csv("./data1/Police\_Department\_Incident\_Reports\_\_2018\_to\_Present.csv",na\_values = missing\_values)



CSV data find the exact location where crime is less and other factors to used by pandas (extnsion with .shp or .prj) which includes SF location dimensions

GEOJSON data

Based on definition of our problem, factors that will influence our decission are:

\* number of existing restaurants in the neighborhood (any type of restaurant)

\* number of and distance to restaurants in the neighborhood, if any

\* distance of neighborhood from city center

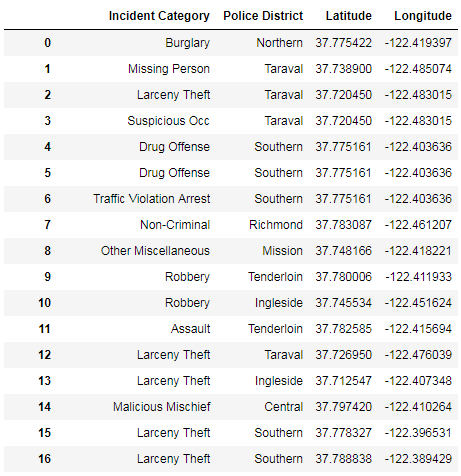
We decided to use regularly spaced grid of locations, centered around city center, to define our neighborhoods.

Filtered and Clean data output

Explanation for dropping Columns :

There are lots of columns. But for our purpose, we only need ‘Incident Category,’ ‘Police District,’ ‘Latitude,’ and ‘Longitude.’ One notices that multiple incidents might have the same Incident ID. This is because a report (corresponding a row in the dataframe) can have multiple incidents associated. After dropping unwanted columns and cleaning up NaNs, the number of crimes can be plotted by categories:

df\_crime.drop(['Incident Datetime','Incident Date','Incident Time','Incident Year','Incident Day of Week','Report Datetime','Row ID', 'Incident ID','Incident Number','CAD Number','Report Type Code','Report Type Description','Filed Online','Incident Code','Incident Subcategory','Incident Description','Resolution','Intersection','CNN','Analysis Neighborhood','Supervisor District','point'],axis=1)



df\_crime.shape

(178109, 26)

Question 3:Methodology section which represents the main component of the report where you discuss and describe any exploratory data analysis that you did, any inferential statistical testing that you performed, and what machine learning’s were used and why.

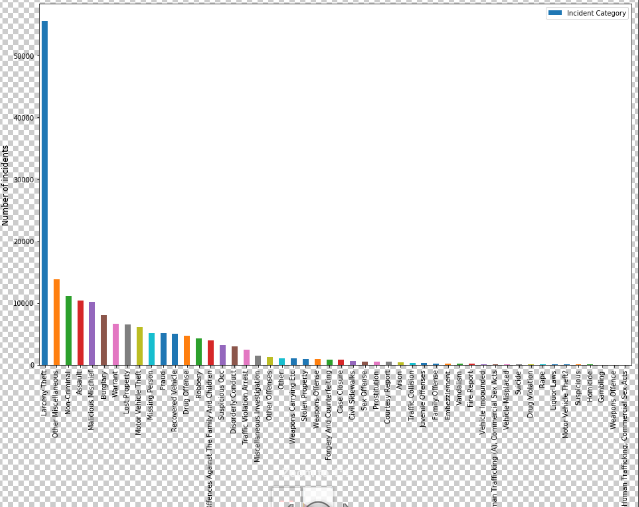
Answer: We used different charts like Bar Chart :

Bar charts and Horizontal Bar charts are useful when comparing different categories, it would be nice if we could put these data on a map using their GPS coordinates. This could be achieved through using the GeoPandas and the folium package. We first form a Shapely geometry object by combining ‘Latitude’ and ‘Longitude’ columns and merge it with the original dataframe to form a GeoDataFrame.

use Foursquare API to get info on restaurants in each neighborhood.

we will use \*\*heatmaps\*\* to identify a few promising areas close to center with low number of restaurants in general

**Comparison of crime number over incidents Bar chart**

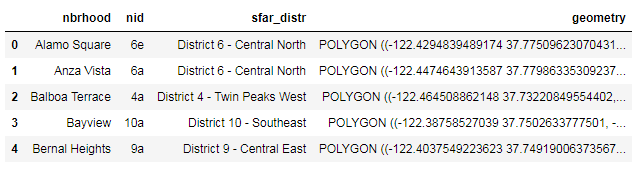


Question 4:Results section where you discuss the results.

Answer : Let's create latitude & longitude coordinates for centroids of our candidate neighborhoods. We will create a grid of cells covering our area of interest which is aprox. 12x12 killometers centered around San Francisco city center.

Let's first find the latitude & longitude of Berlin city center, using specific, well known address and Google Maps geocoding API

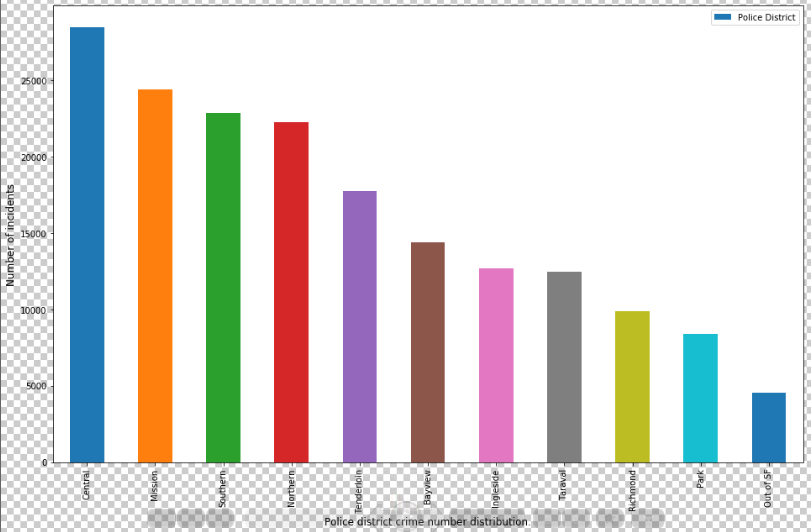
nbrhoods = gpd.read\_file("./data1/geo\_export\_0347b3c0-4ff8-483e-9b1c-327c9a2be580.shp")



Apparently, the number one category is larceny theft. A bit more than 30k incidents in the span of approximately 9 months (up to the time when the dataset was downloaded). That is about 4,000 incidents per month across the entire city. The second most category is assault, followed by burglary (we ignore ‘Other Miscellaneous’ and ‘Miscellaneous Mischief’). But the frequency of these two categories is far less than theft. Alright, now we know the type of crime committed most often in the city. What about its distribution? We could take advantage of the ‘Police District’ column. Using the value\_counts function, the total number of incidents grouped by Police District is plotted as follows:

ax = df\_crime['Police District'].value\_counts().plot.bar(legend=True,stacked=True, figsize=[16,10])

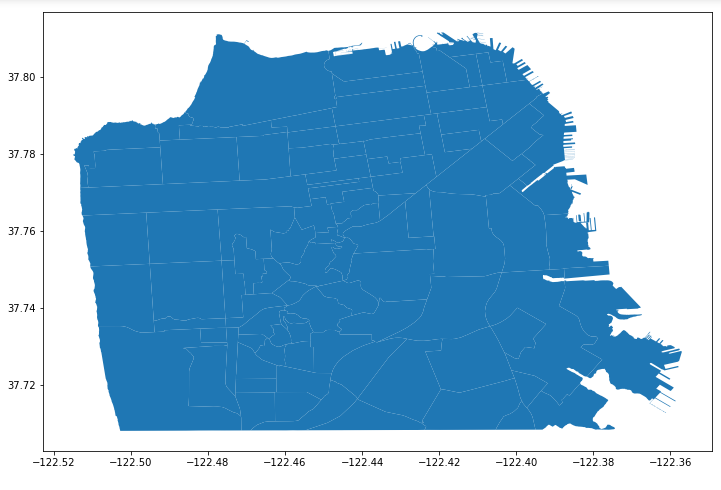
**Police district crime number distribution**

****

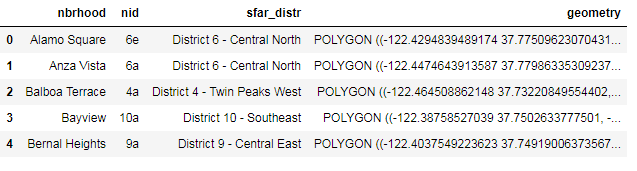
It looks like the Central, Mission, Northern, Southern, and Tenderloin Districts combined contribute the majority number of incidents to the report. This is perhaps not surprising at all because these areas are the most busy districts in San Francisco. For example, the Central District covers Fisherman’s Wharf and the Embarcardero which has many waterfront attractions.

While the above bar charts are useful when comparing different categories, it would be nice if we could put these data on a map using their GPS coordinates. This could be achieved through using the GeoPandas and the folium package. We first form a Shapely geometry object by combining ‘Latitude’ and ‘Longitude’ columns and merge it with the original dataframe to form a GeoDataFrame:

Each row is a neighborhood defined by a shapely polygon object. We can plot the shape file easily using the object’s plot method:



How do we bin the crime reports based on neighborhoods? Similar to SQL’s join function that allows us to selectively combine two databases, GeoPandas’ sjoin provides an approach to spatial join of two GeoDataFrames. Because we want to aggregate the number for each neighborhood, we set the parameter op=’intersects.’ when calling sjoin. The results are grouped by neighborhoods:



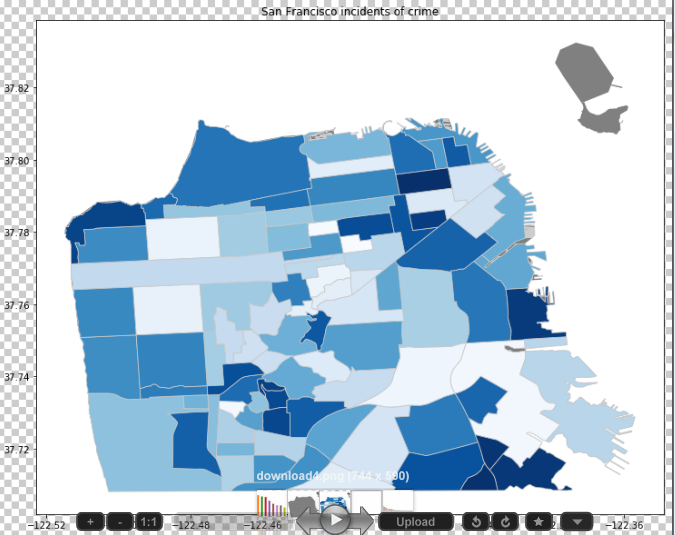
GeoJSON data of San Francisco

These are saved as Polygons rather than Points, and draw out the boundary lines for each of the neighborhoods.

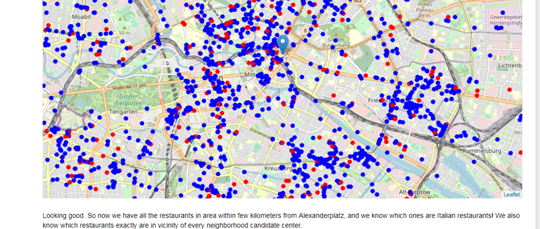


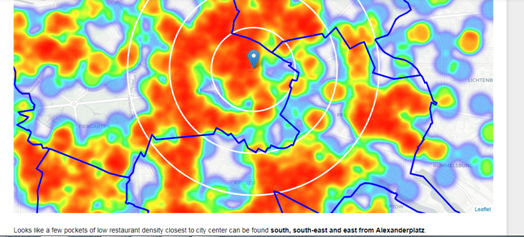
Overlaying Points on top of Polygons

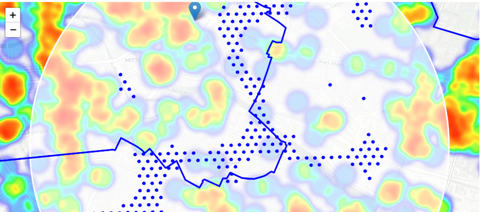
E relationship between diffe



\







use Google Maps API to get approximate addresses of those locations.

**RESULTS AND DISCUSSIONS**

**Our analysis shows that although there is a great number of restaurants in Berlin (~2000 in our initial area of interest which was 12x12km around Alexanderplatz), there are pockets of low restaurant density fairly close to city center. Highest concentration of restaurants was detected north and west from Alexanderplatz, so we focused our attention to areas south, south-east and east, corresponding to boroughs Kreuzberg, Friedrichshain and south-east corner of central Mitte borough. Another borough was identified as potentially interesting (Prenzlauer Berg, north-east from Alexanderplatz), but our attention was focused on Kreuzberg and Friedrichshain which offer a combination of popularity among tourists, closeness to city center, strong socio-economic dynamics \*and\* a number of pockets of low restaurant density.**

After directing our attention to this more narrow area of interest (covering approx. 5x5km south-east from Alexanderplatz) we first created a dense grid of location candidates (spaced 100m appart); those locations were then filtered so that those with more than two restaurants in radius of 250m and those with an Italian restaurant closer than 400m were removed.

Those location candidates were then clustered to create zones of interest which contain greatest number of location candidates. Addresses of centers of those zones were also generated using reverse geocoding to be used as markers/starting points for more detailed local analysis based on other factors.

Result of all this is 15 zones containing largest number of potential new restaurant locations based on number of and distance to existing venues - both restaurants in general and Italian restaurants particularly. This, of course, does not imply that those zones are actually optimal locations for a new restaurant! Purpose of this analysis was to only provide info on areas close to Berlin center but not crowded with existing restaurants (particularly Italian) - it is entirely possible that there is a very good reason for small number of restaurants in any of those areas, reasons which would make them unsuitable for a new restaurant regardless of lack of competition in the area. Recommended zones should therefore be considered only as a starting point for more detailed analysis which could eventually result in location which has not only no nearby competition but also other factors taken into account and all other relevant conditions met.

**Discussion section where you discuss any observations you noted and any recommendations you can make based on the results.**

Performance of different models as mentioned above

Different charts for identification of location as mentioned in charts

**Conclusion section where you conclude the report.**

I analyzed the relationship between different factors to advice for opening a new restaurant.I build different models whether and used different charts to find out the exact location.

Future Predictions :Models in this study mainly focused on individual features and there is scope of improvement considering more factors and depends on investor investment cost and other factors.

Purpose of this project was to identify San Franciso areas close to center with low number of restaurants (particularly Italian restaurants) in order to aid stakeholders in narrowing down the search for optimal location for a new Italian restaurant. By calculating restaurant density distribution from Foursquare data we have first identified general boroughs that justify further analysis , and then generated extensive collection of locations which satisfy some basic requirements regarding existing nearby restaurants. Clustering of those locations was then performed in order to create major zones of interest (containing greatest number of potential locations) and addresses of those zone centers were created to be used as starting points for final exploration by stakeholders.

Final decission on optimal restaurant location will be made by stakeholders based on specific characteristics of neighborhoods and locations in every recommended zone, taking into consideration additional factors like attractiveness of each location (proximity to park or water), levels of noise / proximity to major roads, real estate availability, prices, social and economic dynamics of every neighborhood etc.