1. Introduction - Business Case

For prospect businessmen who can afford setting up a restaurant or club etc. in the San Francisco area a preliminary research might be interesting. This research project aims to provide insights in the neighborhoods of the largest Colleges and Universities in San Francisco. The assumption is that neighborhoods around the educational institutes have lots of footprint and students have sufficient means for buying coffee and visiting restaurants, besides, the usual supermarket visits and they also tend to have parties and fundraisers. Presumably sponsored by their parents, scholarship or an unhealthy study loan. What might be interesting is the crime rate of the several serious recorded offences in the neighborhoods when selecting an appropriate place to open a new venue given that the subjects of your choice are available at multiple locations. Therefore, this project serves two purposes:

- Selecting a nice neighborhood for opening a new venue.
- For the personal security conscious, also insights in the crime rate that can be factored in your choice.

1.1 Discussion of the business case

Wanting to know the nice location is obvious, these can also be clustered with a machine learning algorithm. The crime rate might also be interesting because when your choice of location is only available at limited number, you know beforehand which areas to avoid and at what periods of the day. Possibly it might also be interesting to explore a few points within walking distance of each education location.

2. Discussion of data sources and usage

The folium package will be used to show OpenStreetMap data.
The Foursquare API provides nearby venues with some additional data.
The https://data.sfgov.org provides the crime rates of 2018.
The http://www.city-data.com/city/San-Francisco-California.html provides the list of Universities and Colleges that will be scraped with BeautifulSoup4 into a Pandas Data Frame. These locations provide the basis for the neighborhood analysis of shops and crime rate.

The crimes will be grouped in the vicinity of these beforementioned locations with a Latitude and Longitude band width, furthermore there will be a selection of the more violent and invasive crimes that have a personal experience potential. This means that non-criminal, fraud, found license plates and so on, will be excluded from the further analysis. Then these crimes could even be clustered with the K-means unsupervised clustering algorithm to provide a 'to be labeled' crime profile. Distinctions could be made between working hours, evening and nighttime events to provide further differentiation between crimes. It might also be possible that there is a correlation between time of events and shop-venue clusters, therefore this will be researched. It might well be that this possible correlation is quite insignificant or non-existent at all. Though the presumption is that in an area with many bars and cafes the assault rate would be higher.

3. Methodology section

Pandas will be used for data mangling, thus cleaning, reading, grouping and aggregating data in a table formatted style. The beforementioned data sources will be filtered on:

- the vicinity of the educational locations
- shops and venues around these locations
- crimes by these neighborhoods, crime type and timeliness of occurrences

3.1 Exploratory analysis

- describe the Latitude and Longitude bandwidths of the selected locations for crime and shop-venue selections
- plot the crimes in total and per location with a histogram with time-period on the horizontal axis and count on the vertical axis.
- plot these crimes also per selected main category.

For visualization purposes also provide a detailed map with aggregated crimes around each location, with markers that provide details. Also plot the found venues in the areas around each location for generating an idea about the neighborhoods.

3.2 Machine Learning: Clustering

For creating insights into the neighborhood, clustering is an appropriate algorithm that can classify similar 'themed' shops and venues in the vicinity. This clustering could be tweaked manually by performing different selections. This selection could theme the area quite significantly.
The crimes will be clustered with K-Means and Density-Based Spatial Clustering of Applications with Noise (DBSCAN).

4. Results

This results section will discuss the gathered Education Locations, Foursquare venue data, crime data, feature selection, applied algorithms, the non-successful search for correlations, and lastly the most successful combination of clustering techniques.

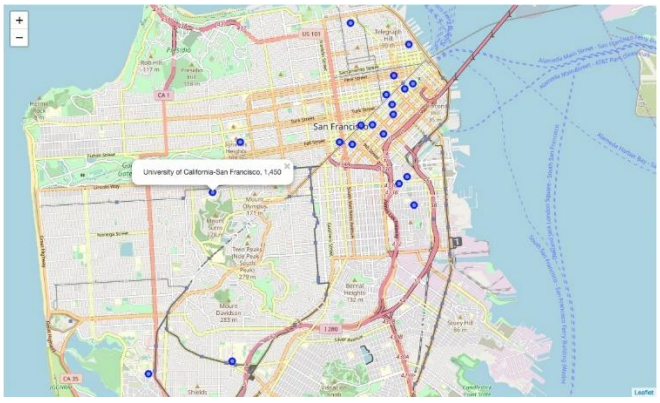

Figure 1:Education Institutes

Figure 2:Venues

Figure 3: Crime Trend analysis

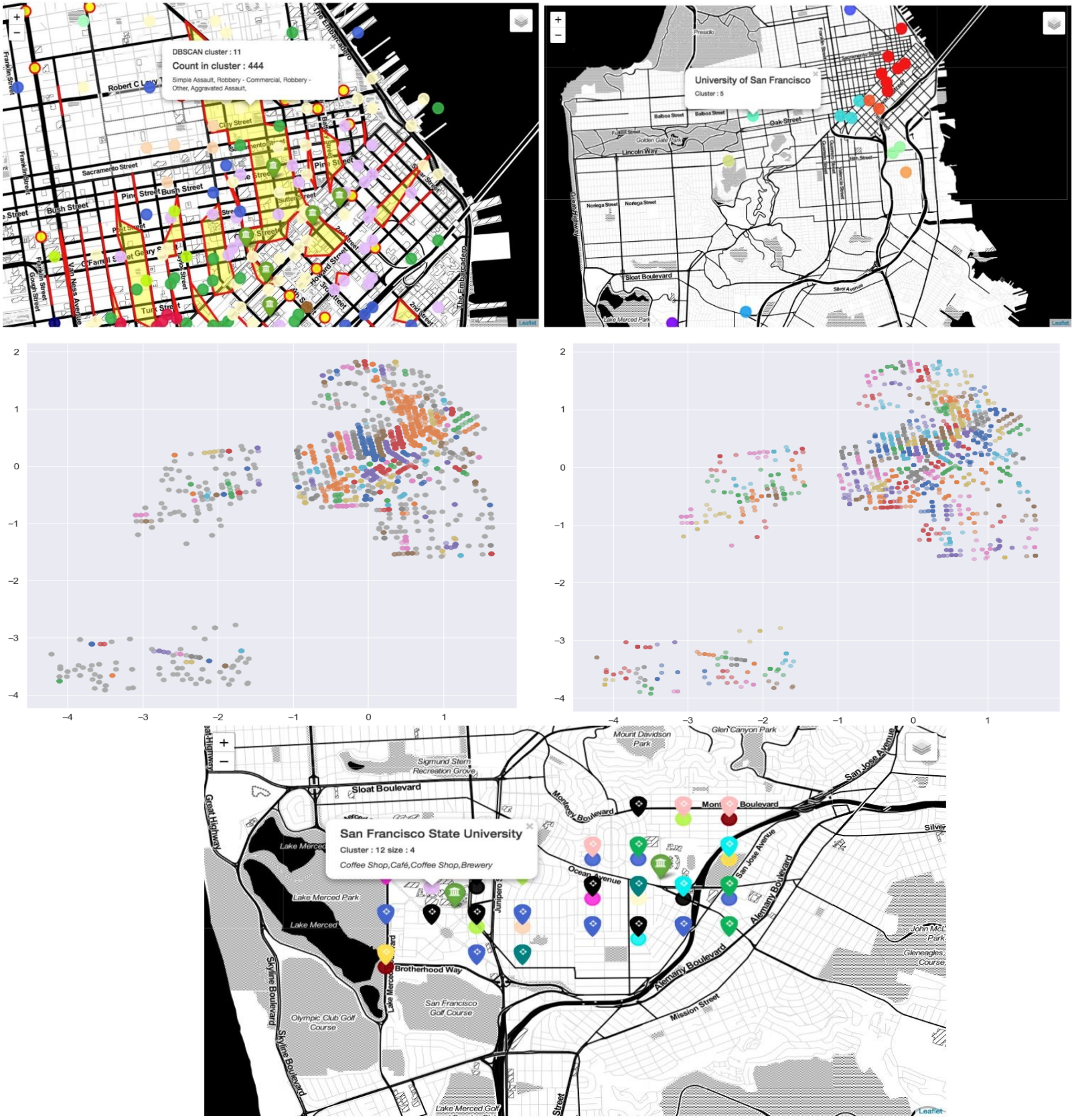

Figure 4: Clustering using K means and DBSCAN

4. Discussion

Aside from the non-found correlated expected effects, this project was able to produce quite insighful perspective with the explorative goals in mind, of neighbourhood exploration based on two factors, crime-incident and venues.
Most interestingly, a few applied algorithms with daytime-only crime-incidents produces quite detailed results when all incidents were not quadrantised, thus without predetermined neighborhood sizes.
The DBSCAN algorithm produces results that were easily interpretable for individuals without specific domain-knowledge available. The cluster naming produced an quantifiable neighborhood impression, with displayed points per cluster and cluster name.
These points allow an individual to make sense of the density of each cluster, and especially in case of crime, correlate this to the *actual* chance that one could be confronted with an event that could have a traumatizing effect, either emotionally of physically.
The venue clusters appear to clutter the view, this a known effect of the quadrant neighborhood determination. This intentionally cluttered view disappears when the Folium Layer Controls are used. Each education location can be removed from the display. This can be done individually in all versions of the Folium crime-venue compound maps. The crimes can be removed only all at once. This is also intentional since the high amount of unsupervised-non-quadrantised clusters have little or no locational-related relation with any of the education locations.

One of the main recommendations is when an individual explores any version of the map, is to consciously be aware that there are a few important things to keep in mind at all times :

- All crime related incidents are anonymized, many could happen indoors and would not be noticeable for pedestrians on the sidewalk of a street.
- Incidents happened in the past, do not have to occur again at any time in the future with any guarantee. The crime related data is *historic* data and give an impression of the incidents last year. This reasoning is also one of the *better* arguments for choosing DBSCAN, since that algorithm is designed to cluster inherent noisy data. Combined with the choice to minimize the hyperparameters to minimize the noise count, the effect could be that spurious occurrences, like homicide, are simply not visually clustered. This has the accompanying effect to *not* over-frighten an individual inspecting these clusters.

5. Conclusion

The fulfilling of the project was quite educational, to observe the actual effects of the applied algorithms, tuning hyperparameters and the effects in a numerical sense and visually. To explore different manners of tackling programmatic challenges and observed malfunctioning of implemented packages, finding satisfiable solutions to those.
The gaining of more practical experience in the complete process of Data Science, from research question formulation, gathering and processing data to presenting the results within the limits of the *supplied* tools like *free-version-API-limitations* was a quite interesting process.

Concluding that the research question is successfully answered, a presentable solution has been developed to provide insight into the venue and crime climates in the neighborhoods of educational locations for investors in San Francisco, who can explore the neighborhoods visually and interpret the

provided labelled clusters with quantified venue and crime incident counts. This should help a security conscious individual to make an educated choice, or to possibly avoid some streets.

In the current representations a user-selectable crime-layer is optimal, because of the different clustering characteristics and a user can decide which clustering method is personally preferable. A default clustering method is suggested and is the DBSCAN clustering since this algorithm ignores noise in the visualizations and produces the least unintended effect of crime-climate over appreciation.