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.