Where to build a pizza place in San Francisco?

A coursera capstone project by Wolfgang Hamer

1. Introduction

Pizza places enjoy great popularity. For the customer, they provide a well-known and well-liked product, and for the owner, they offer good added value and correspondingly high profit margins. In order to maximize the profit, however, the optimal location is crucial. In this project we try to find the optimal position for a pizza place in San Francisco. It would seem logical at first to look for a district in which there is not yet a pizza place, as it would obviously serve a gap in the market there. But following Hotelling's spatial competition (also known as Hotelling's law) a new pizza place fits nice next to another one. If there is no or only one pizza place in a district, few customers will go to that district to eat a pizza. However, if the district is known for pizza, our client's new pizza place can catch and win over the customers of other pizzerias by offering better quality, a wider selection and more exotic creations. In addition to the number of pizzerias per district, the crime rate in that district is also important to our customers. Let us assume that he has closed his last pizza place due to frequent vandalism and is therefore looking for a district with a low crime rate. The goal of this project is to cluster the districts of San Francisco according to their equipment of shops and restaurants and the criminality occurring in them and recommend suitable districts for our client.

2. Data

The analysis is based upon three datasets:

1. Geometries of the districts of San Francisco which are downloaded from the official website data.sfgov.org
2. Foursquare data for each district which are downloaded via the API
3. Police Department Incident Reports which are also downloaded from the official website data.sfgov.org

The Geometries of the districts of San Francisco are downloaded from the official website data.sfgov.org and imported into python via the geopandas package. The geopandas dataframe contains the geometries itself and the numbers and names of the districts. In the course of this project, especially the names will continue to be used (Fig. 1).

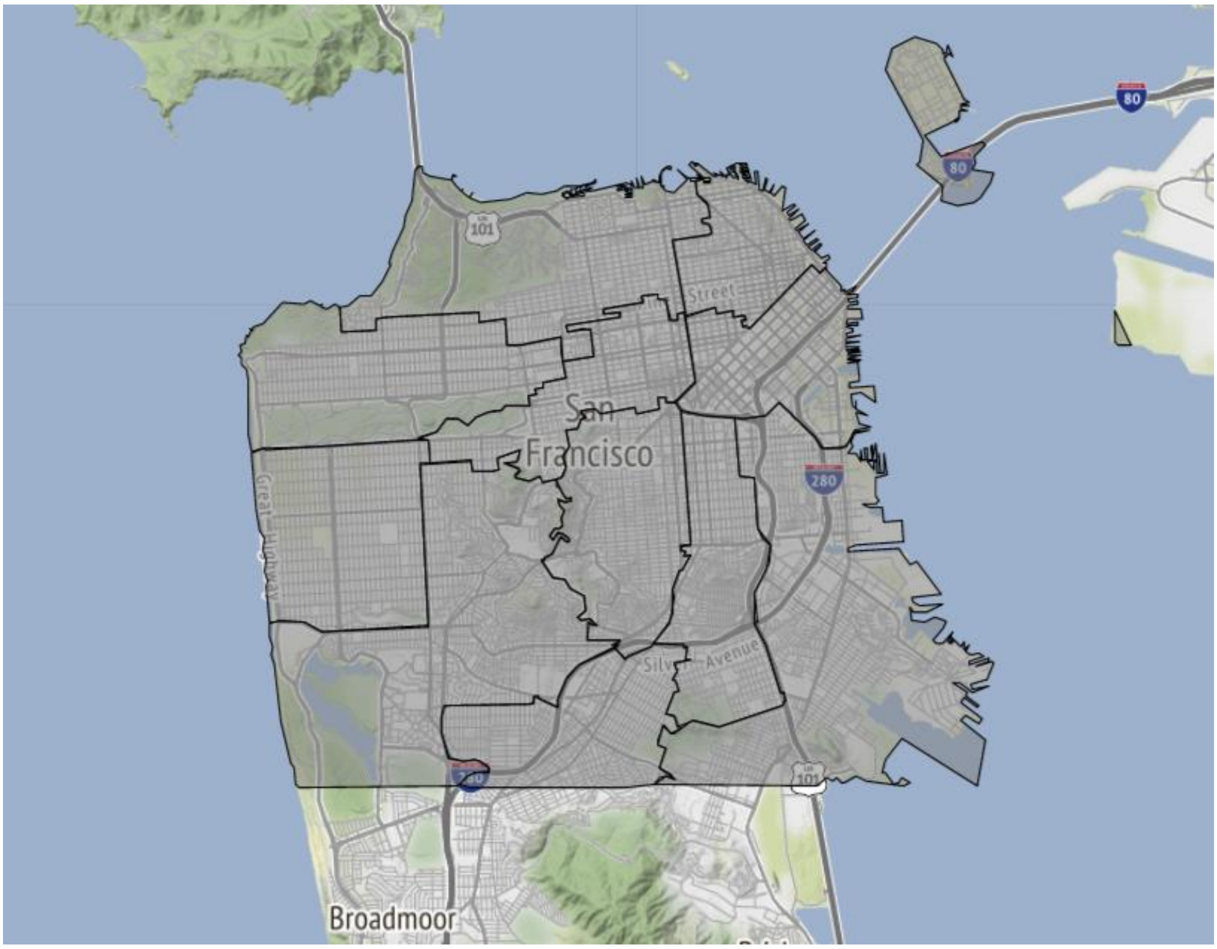
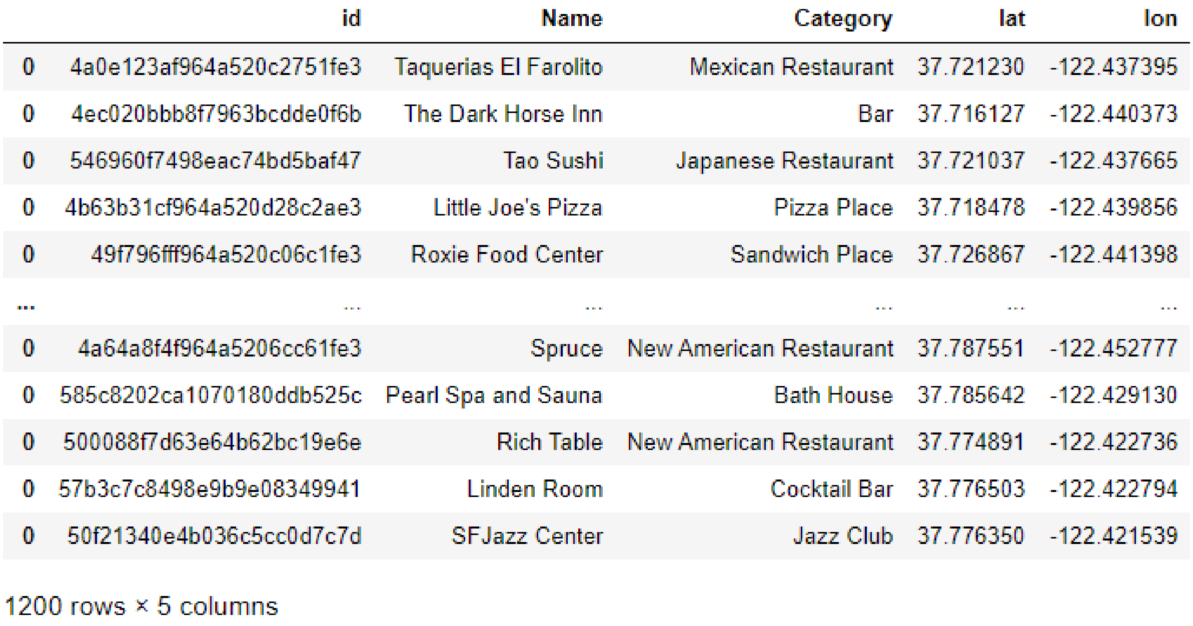


Figure 1: Spatial representation of districts (zoomable and clickable in the Jupyter Notebook)

The Foursquare data for each district are downloaded via the API. First the credentials to download from Foursquare are inserted and the limit of locations and the maximum distance are set to 3000 m. The Foursquare data are downloaded and transferred to a pandas dataframe containing the locations id, the Name, the Category and the coordinates (Fig. 2). For this project the Category as well as the location are of specific interest, since the Category contains the information about pre-existing pizza places and Italian restaurants.

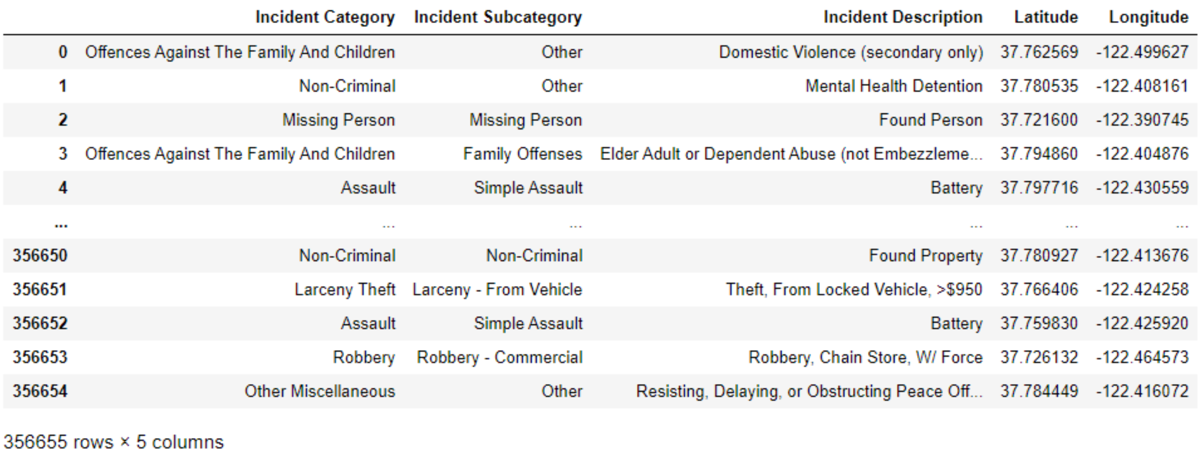


	id	Name	Category	lat	lon
0	4a0e123af964a520c2751fe3	Taquerias El Farolito	Mexican Restaurant	37.721230	-122.437395
0	4ec020bbb8f7963bcdde0f6b	The Dark Horse Inn	Bar	37.716127	-122.440373
0	546960f7498eac74bd5baf47	Tao Sushi	Japanese Restaurant	37.721037	-122.437665
0	4b63b31cf964a520d28c2ae3	Little Joe's Pizza	Pizza Place	37.718478	-122.439856
0	49f796fff964a520c06c1fe3	Roxie Food Center	Sandwich Place	37.726867	-122.441398
...	...	...	...	...	...
0	4a64a8f4f964a5206cc61fe3	Spruce	New American Restaurant	37.787551	-122.452777
0	585c8202ca1070180ddb525c	Pearl Spa and Sauna	Bath House	37.785642	-122.429130
0	500088f7d63e64b62bc19e6e	Rich Table	New American Restaurant	37.774891	-122.422736
0	57b3c7c8498e9b9e08349941	Linden Room	Cocktail Bar	37.776503	-122.422794
0	50f21340e4b036c5cc0d7c7d	SFJazz Center	Jazz Club	37.776350	-122.421539

1200 rows × 5 columns

Figure 2: Foursquare data

Since our customer is afraid of malicious mischief, robbery, burglary and vandalism, we want to know in which district he could get affected by it. Therefore, Police Department Incident Reports are downloaded from the official website (https://data.sfgov.org/Public-Safety/Police-Department-Incident-Reports-2018-to-Present/wg3w-h783). The dataset includes many attributes, but for our analysis we are specifically interested in the incident category and subcategory, since here the information about vandalism and robbery is contained, as well as in the location information (Fig. 3).



	Incident Category	Incident Subcategory	Incident Description	Latitude	Longitude
0	Offences Against The Family And Children	Other	Domestic Violence (secondary only)	37.762569	-122.499627
1	Non-Criminal	Other	Mental Health Detention	37.780535	-122.408161
2	Missing Person	Missing Person	Found Person	37.721600	-122.390745
3	Offences Against The Family And Children	Family Offenses	Elder Adult or Dependent Abuse (not Embezzleme...	37.794860	-122.404876
4	Assault	Simple Assault	Battery	37.797716	-122.430559
...	...	...	...	...	...
356650	Non-Criminal	Non-Criminal	Found Property	37.780927	-122.413676
356651	Larceny Theft	Larceny - From Vehicle	Theft, From Locked Vehicle, >$950	37.766406	-122.424258
356652	Assault	Simple Assault	Battery	37.759830	-122.425920
356653	Robbery	Robbery - Commercial	Robbery, Chain Store, W/ Force	37.726132	-122.464573
356654	Other Miscellaneous	Other	Resisting, Delaying, or Obstructing Peace Off...	37.784449	-122.416072

356655 rows × 5 columns

Figure 3: Crime data about San Francisco

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, if any, and what machine learnings were used and why.

The methodology is split in three parts:

1. Combination of Districts and Foursquare data
2. Combination of the foregone datasets with the crime data
3. Use of the unsupervised kmeans cluster machine learning process to determine the optimal districts for the safe location of a pizza place

3.1. Combination of Districts and Foursquare data

First the duplicates are removed from the Foursquare data. These are created because the API is used to retrieve all localities within a fixed radius of the district centre. Especially for smaller districts, it is possible to retrieve locations of neighbouring districts, which are then duplicated. This results in a dataset of 849 observations. The next goal is to assign a district to the Foursquare data according to their geographical location. To do this, the Foursquare data is first transferred to a spatial object of the class geopandas.GeoDataFrame. Then the `within` method is used, to select only the elements which are geographically within the specific district. This results in Foursquare data points relative to the location in the cities districts (Fig. 4).

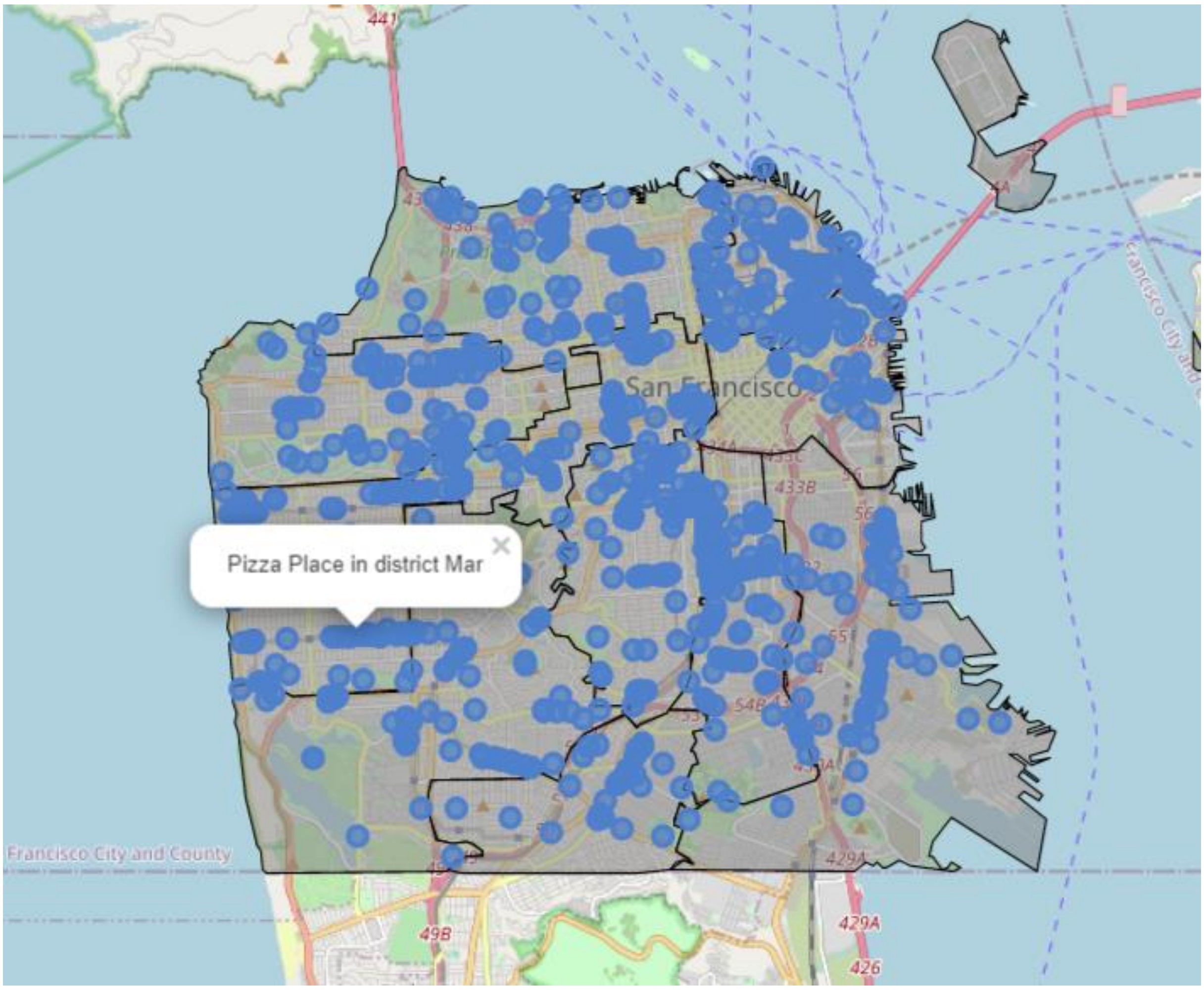


Figure 4: Foursquare data in San Francisco (zoomable and clickable in the Jupyter Notebook)

The next step is designed to show the number of Foursquare categories per district. Afterwards, the categories are manually classified to identify, how many pizza places, restaurants, etc. a district has. This has to be done manually, since a first attempt, which automatically classified for similarity, could only combine places that explicitly contained "restaurant" or "store". Figure 5 clearly shows, that restaurants are dominant in the studied districts. Only in Stefani, the district with the most outdoor and fitness activities, there are more outdoor places referred to in the foursquare database. The most pizza places are located in Peskin followed by Mandelman and Yee.

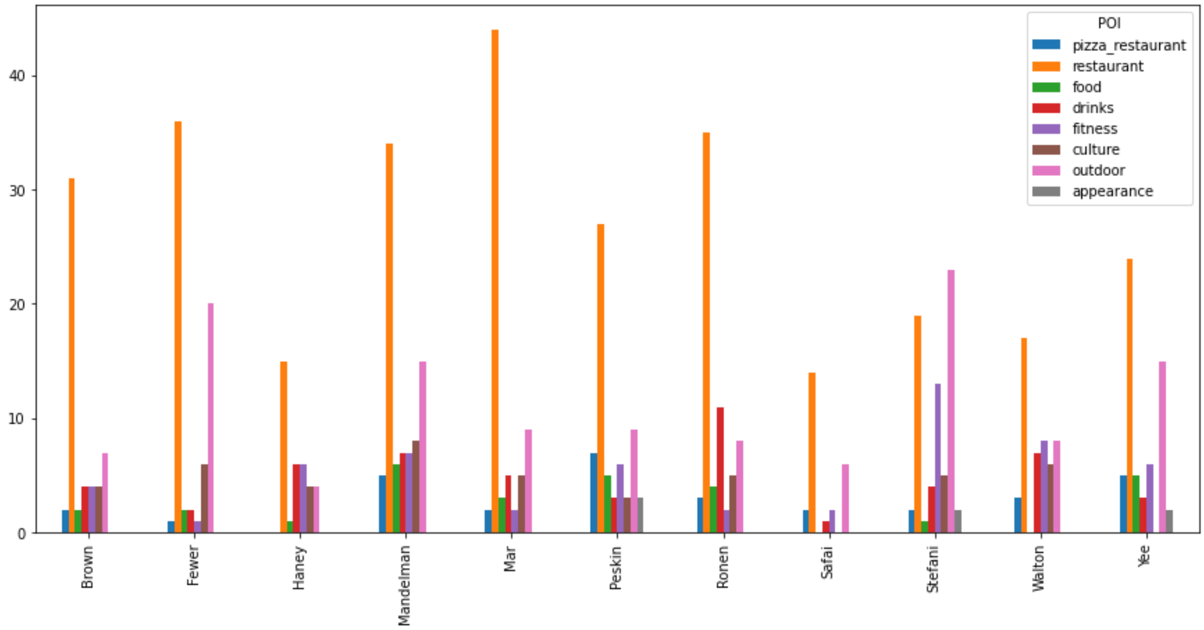


Figure 5: POIs in each district

Since the districts are not of equal size, the foursquare points have to be set relative to the districts size (Fig. 6). The dominance of Peskin in relation to pizza restaurants has become even greater due to the relative size of the area. Mandelman still follows in 2nd place, but Yee now seems to be on par with Brown.

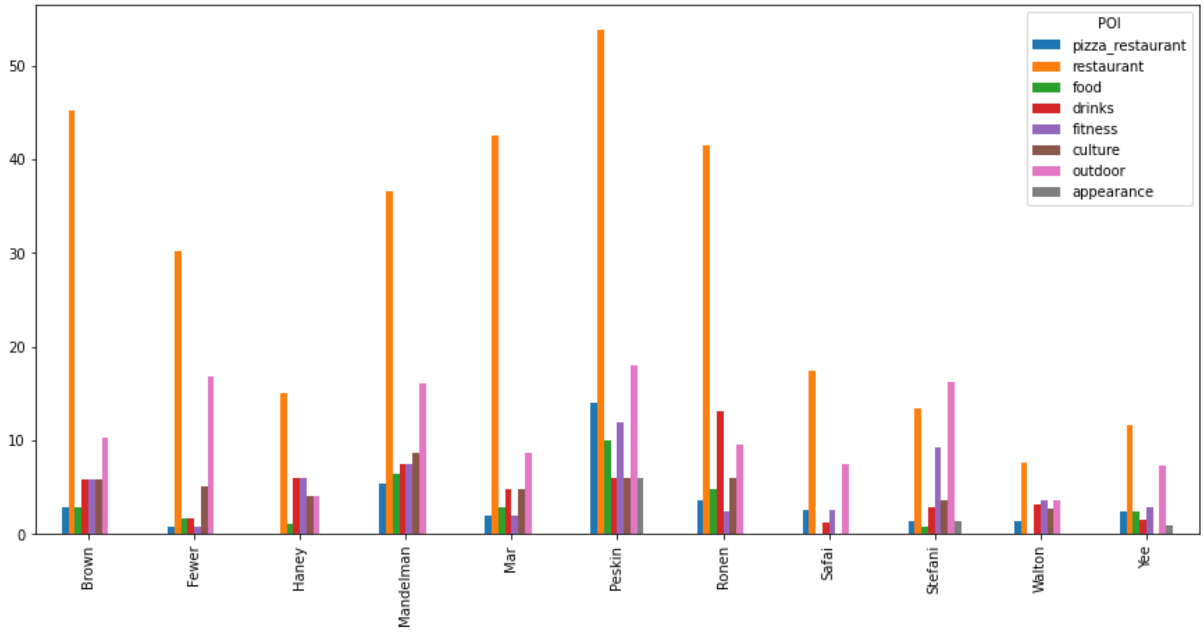
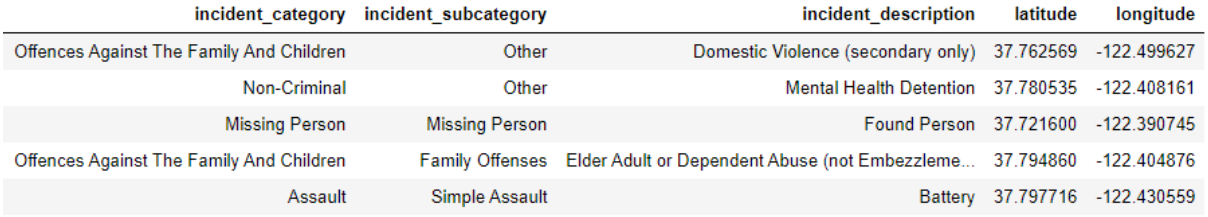


Figure 6: POIs in each district relative to the districs size

3.2. Combination of the foregone datasets with the crime data

First the total data set is reduced to the parameters relevant for this study and the NA values are removed resulting in the dataset represented in Figure 7.



incident_category	incident_subcategory	incident_description	latitude	longitude
Offences Against The Family And Children	Other	Domestic Violence (secondary only)	37.762569	-122.499627
Non-Criminal	Other	Mental Health Detention	37.780535	-122.408161
Missing Person	Missing Person	Found Person	37.721600	-122.390745
Offences Against The Family And Children	Family Offenses	Elder Adult or Dependent Abuse (not Embezzleme...	37.794860	-122.404876
Assault	Simple Assault	Battery	37.797716	-122.430559

Figure 7: Crimes dataset consisting of 337,759 observations

A look at the crimes (Fig. 8) shows that many categories are included which are not relevant to the concerns of the pizza entrepreneur. He is mainly interested in malicious mischief, robbery, burglary and vandalism. Similar to the classification of the Foursquare categories, the author also made a manual selection based on the incident categories (Fig 9).

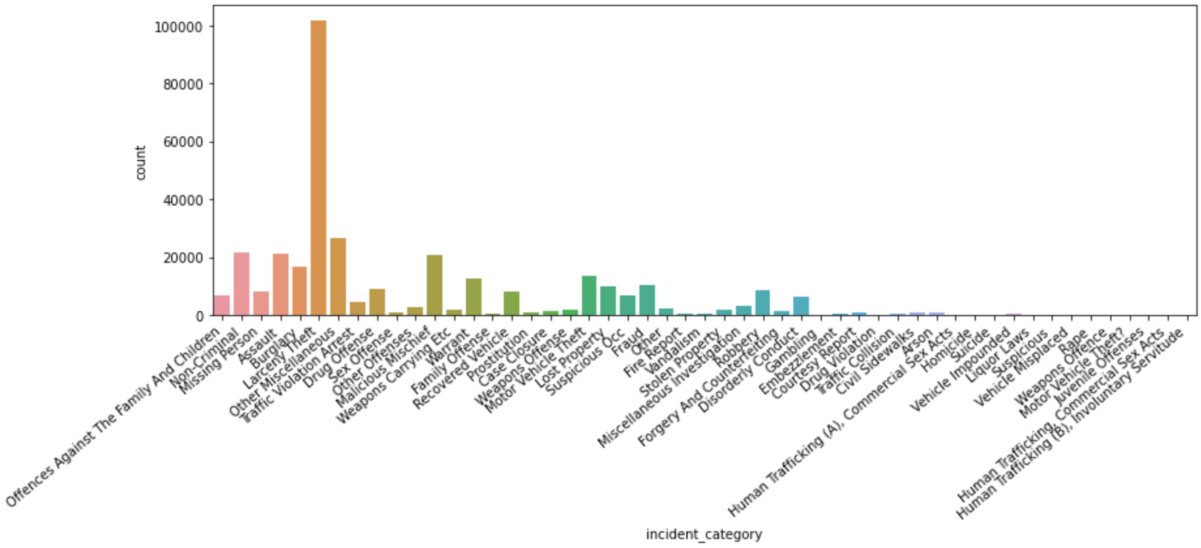


Figure 8: Observed crimes in San Francisco

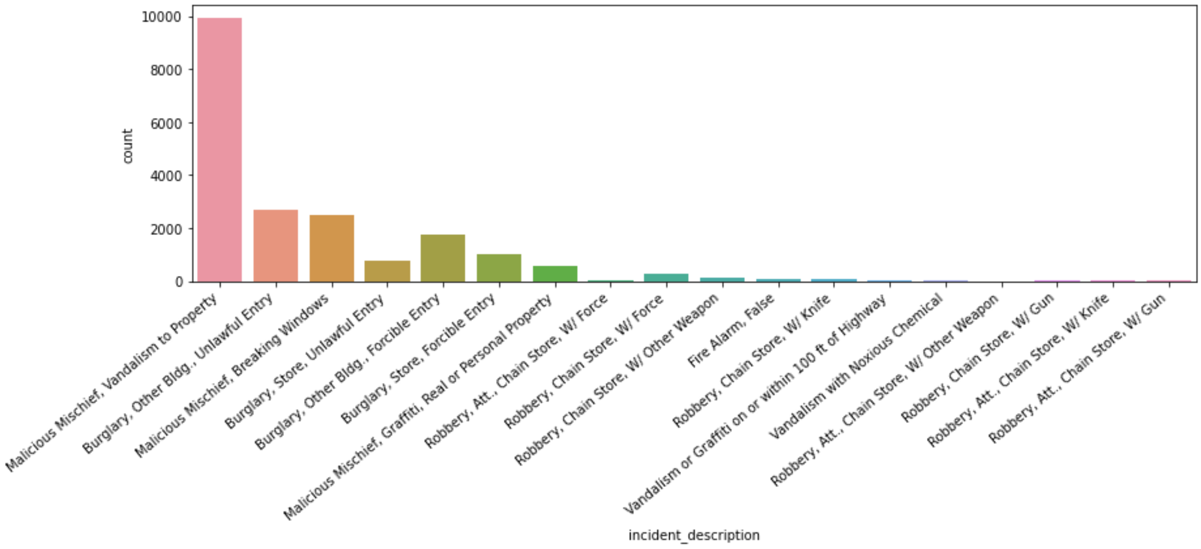


Figure 9: Selected observed crimes in San Francisco

As with the Foursquare data, the aim now is to determine in which district the crimes were committed. For this purpose these are also transferred to a spatial object of the class geopandas.GeoDataFrame and then selected by the `within` method. In this case, the problem arises that the official record does not list any crimes in the Safai district. This can have two reasons. Firstly, there may not actually be any crime in the district. Then they could be assigned the value 0 in the data record. The author considers this to be rather unlikely due to the size of the district. The more probable case is that no data for this district are contained in the data set, although crimes do occur. Therefore this district is excluded in the following. For the other districts it seems as if Haney has the most crimes out client wants to avoid Fig 10). But similar to the foursquare data, the number of crimes should be normalise by the area of the districts (Fig. 11). Are the crimes considered relative to the area of the district, Peskin takes the lead followed by Haney and Brown.

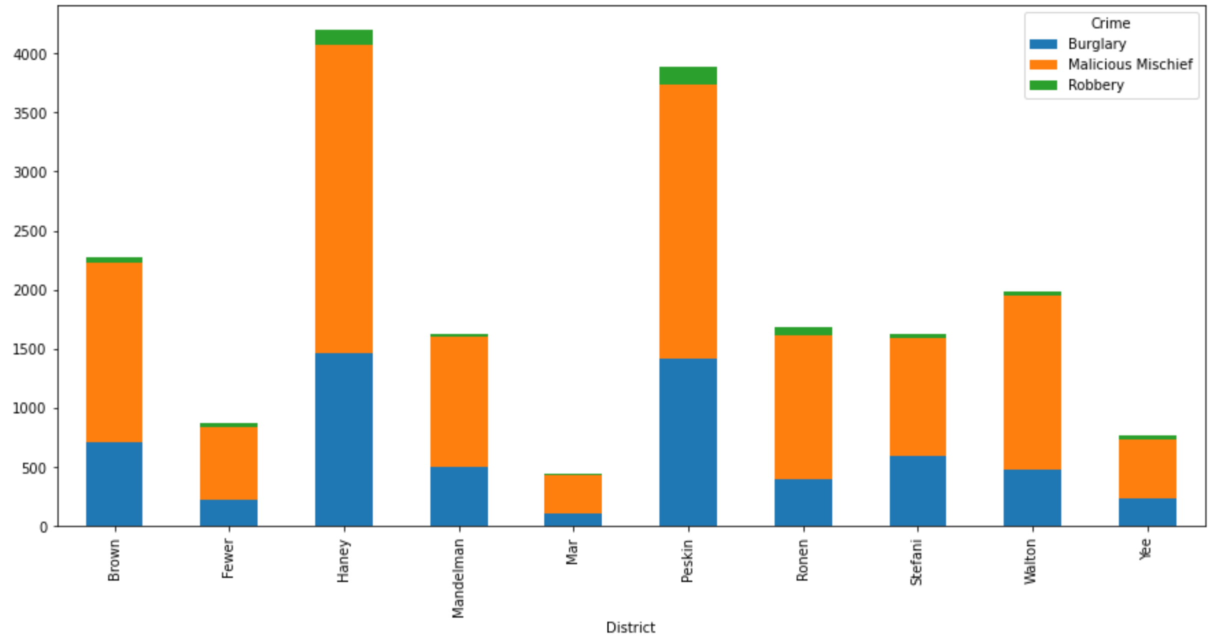


Figure 10: Crimes in each district

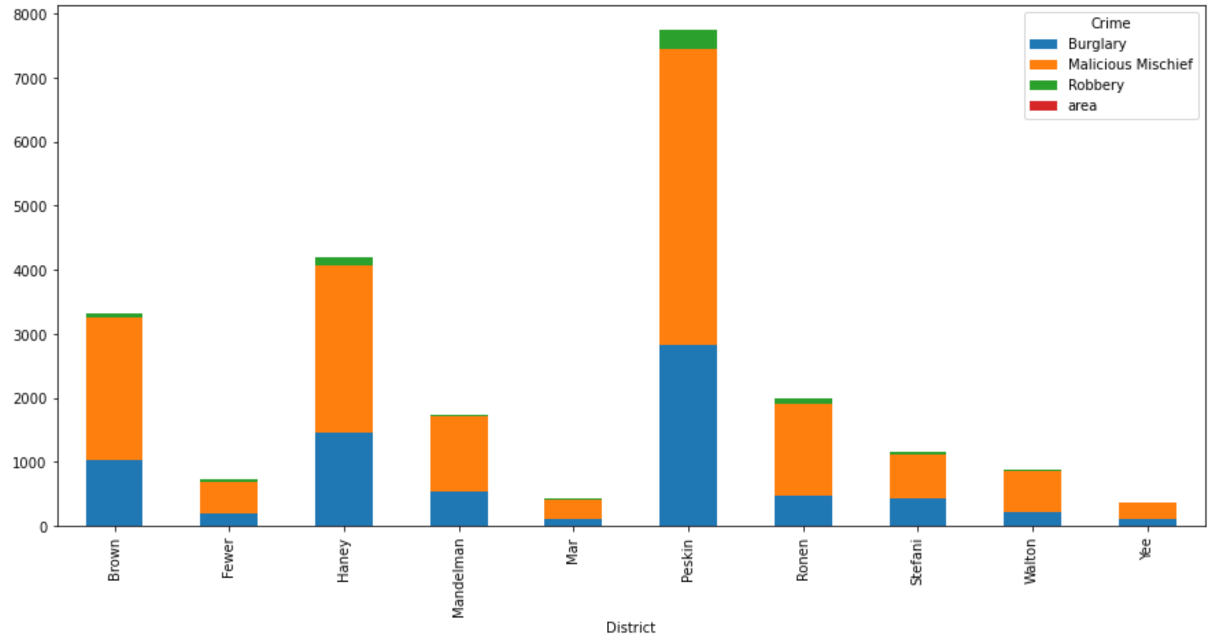


Figure 11: Crimes in each district relative to the districs size

3.3. Use of the unsupervised kmeans cluster machine learning process to determine the optimal districts for the safe location of a pizza place

Since these data shall be studied by clustering, first the values are normalized for each column. Then the kmeans cluster algorithm can be applied. This is based on the basic assumption that the districts provide different services, which can be identified with the foursquare data and the crimes committed in the districts. Thereby, similar districts are assumed to have a similar suitability for the opening of a new pizza place.

4. Results

Initially, the characteristics of the districts will be compared (Fig. 12). The machine learning process has divided the districts into 5 clusters. Two clusters (2 and 4) consist of only one district. Cluster 0 and 3 each consist of three districts and Cluster 1 consists of two districts. Cluster 2, which consists solely of the district of Peskin, has the highest number of pizza places and other restaurants relative to the area. This is also shown by the comparison with the corresponding figure in the Methodology chapter. However, the crime rate is also significantly higher than in the other districts. The high crime rate as well as the wide range of restaurants and other activities indicate that this cluster is the pulsating heart of the city. Cluster 4, which consists exclusively of the Mandelman district, has the second highest number of pizza places relative to the area. The number of other restaurants is comparable to clusters 1 and 3, with crime in the lower midfield. The large cultural offer as well as the outdoor activities that are possible here distinguishes this cluster from clusters 0, 1 and 3. The districts of Cluster 1 and 3 are relatively similar. Both have significantly more restaurants than pizza places. However, a clear differentiation can be made on the basis of crime. Here the districts of Cluster 1 are clearly more dominant than the districts of Cluster 3, and in addition Cluster 1 has more fitness facilities than Cluster 3. Cluster 0 differs from the other clusters by generally lower values. As a result, the crime value is also the lowest. It is interesting to note that this is the only cluster with more pizza places than other restaurants.

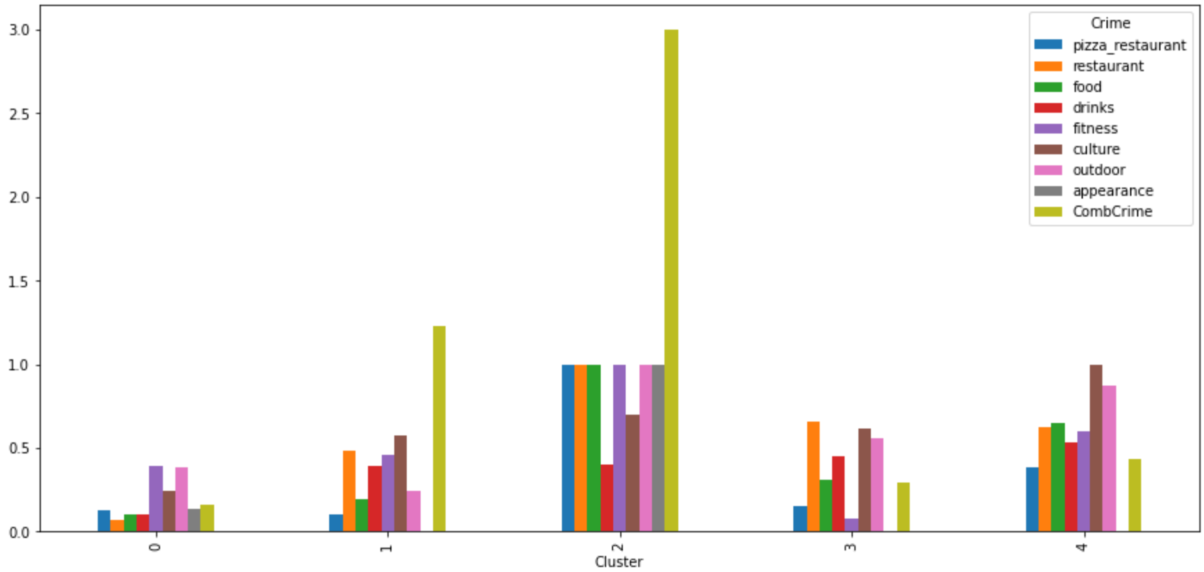


Figure 12: Characteristics of the districts

Next, the geographical location of the clusters will be described (Fig. 13). Cluster 2 (orange), the district of Peskin, is located in the north between the two bridges over which San Francisco can be reached. Cluster 4 (blue), the Mandelman District, is situated in the centre of the districts of San Francisco. The districts of cluster 0 (green) enclose San Francisco in the north, south and west. Two of the three districts of cluster 3 (yellow) are located in the west of the city. The third district is situated in the centre of the districts close to cluster 4. The districts of cluster 1 (violet) build a line from the northeast to the centre of the city and thereby separate cluster 2 from cluster 4.

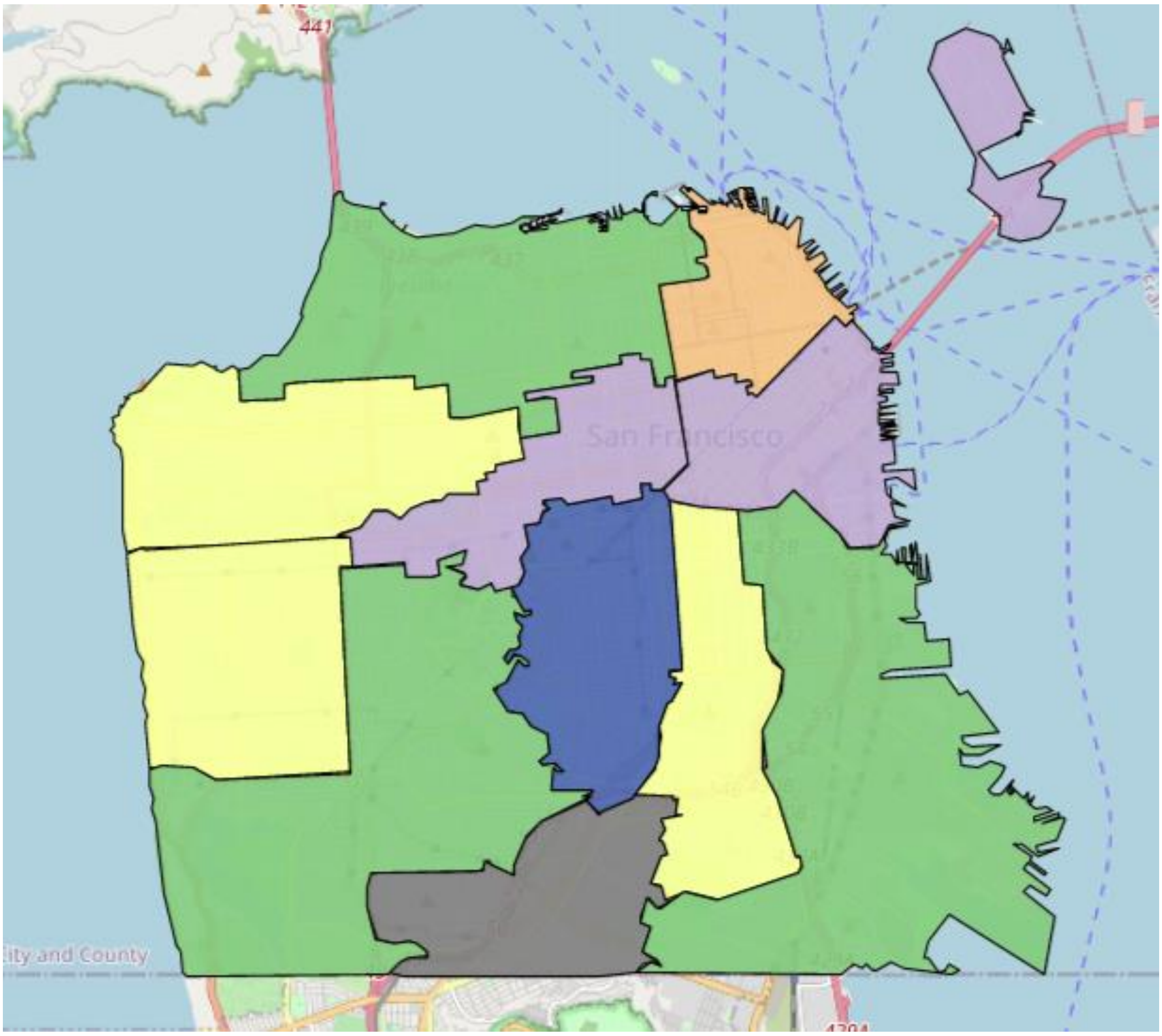


Figure 13: Geographical location of the clusters (0: „green“, 1: “violet”, 2: “orange”, 3: “yellow”, 4: “blue”)

5. Discussion

The district of cluster 4 (Mandelman) seems to be the most appropriate locations for a pizza restaurant (Fig. 14). In these district on average the second most pizza places are located and Following Hotelling's spatial competition a new pizza place fits nice next to another one. In addition, crime is comparatively low here. The districts of Cluster 0 could also be interesting, as they have a high number of pizza places in relation to other restaurants and the crime rate is even lower. Cluster 2, which has more pizza places than the other clusters also has a very high crime rate. Therefore cluster 2 is not considered. Cluster 1 has also a high crime rate, and therefore is not considered either. Cluster 3 has much more differing restaurants but not that much pizza restaurants. Therefore, it can be assumed, that people would go there to search for something else than pizza to eat. This cluster therefore would be considered after cluster 4 and 0.

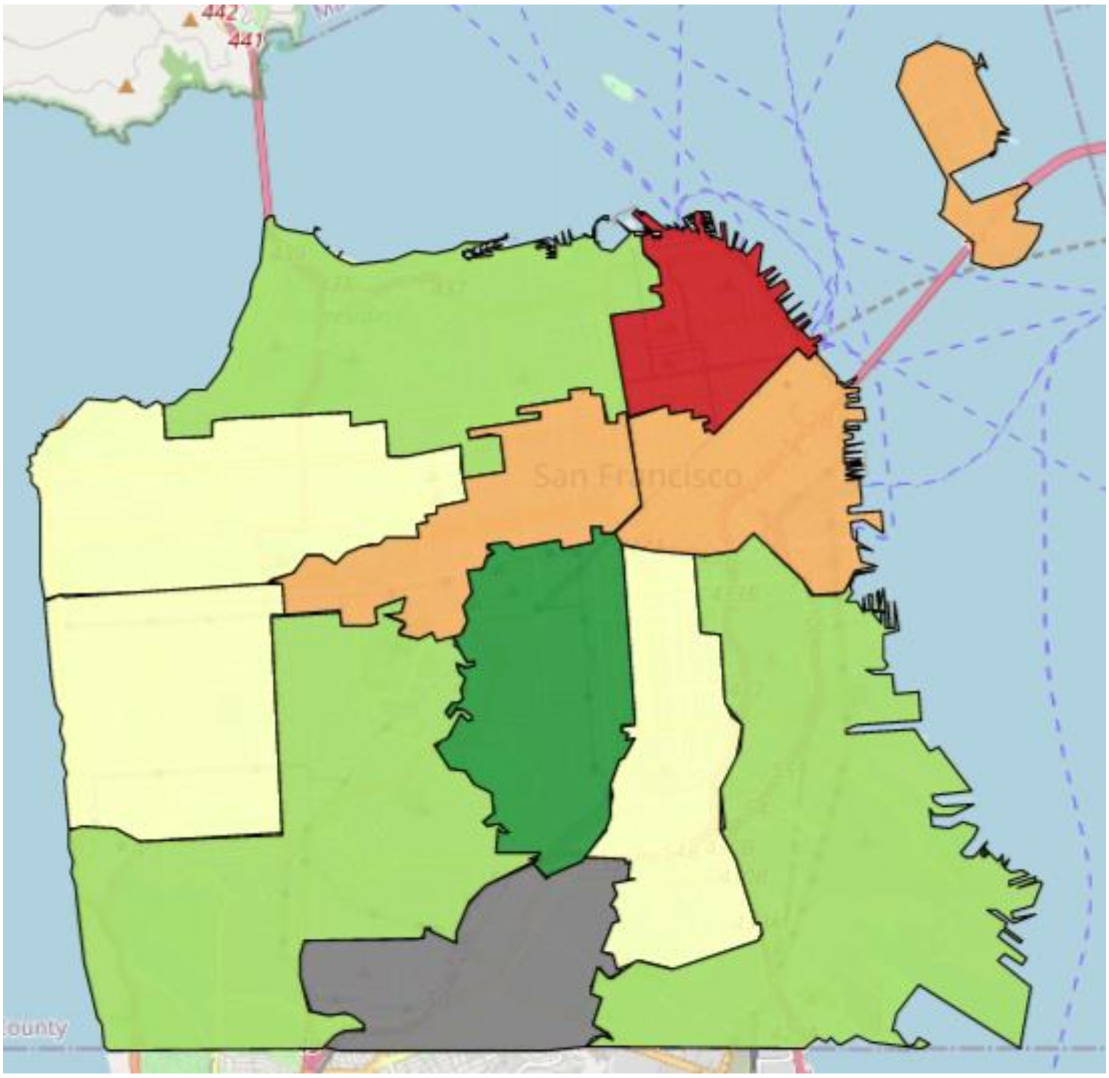


Figure 14: Geographical location of the clusters ranked by colour (First: „green“, Second: “lightgreen”, Third: “yellow”, Fourth: “orange”, Fifth: “red”)

6. Conclusion

The analysis shows a clear recommendation of the Mandelman district for the customer. The district of Peskin seems to be the least suitable because of the high crime rate. Apart from the crime rate, however, there would be a lot to recommend this district. If the client had further information, such as the targeted shop rent or if the pizza place has the aim to deliver pizza or which income class is expected among the customers further data would be required and further analysis could be carried out. For example. ff the pizza place has the aim to deliver pizza, a good connection close to fast streets would be important. If the pizzas are based on cheep ingredients, a district with younger inhabitants would be better for the pizza place than a district with older and more wealthy inhabitants. Without further information, however, the recommendation for the district of Mendelman remains.