Capstone project

The Battle of Neighborhoods

Riccardo Angelo Giro

April 2020

1. Introduction

1.1. Description of the problem and discussion of the background

The purpose of this project is to determine the best district in San Francisco for a person that may want to live in such city. The evaluation metrics for the problem considered are:

1. Safety, which is related to the total number of crimes corresponding to each district;
2. Attractiveness, which is assessed by listing the ten most common venues in each neighborhood.

The above features have been chosen because they have a relevant impact in determining the choice of a new location for a given individual.

Lastly, each district will be clustered using a procedure based on **HDBSCAN**, such that neighborhoods with similar characteristics (from a safety and attractiveness point of view) are assigned the same label.

To briefly summarize:

1. **Business problem**: determining the best district to live in San Francisco;
2. **Target audience**: individuals willing to relocate to San Francisco;
3. **Methods**: unsupervised clustering of data using the HDBSCAN algorithm.

2. Data

2.1. Description of the data that will be used to solve the problem and source of such data

The crime dataset used in this work corresponds to all the crimes that occurred in San Francisco in 2016. It can be freely downloaded from Kaggle website at the following link: https://www.kaggle.com/roshansharma/sanfranciso-crime-dataset.

The crime dataset, displayed in detail in Section **3.1**, is structured in a table. The column labels have the following meaning:

1. **IncidntNum**: incident number;
2. **Category**: category of crime;
3. **Descript**: description of the crime;
4. **DayOfWeek**: day of the week in which the crime occurred;
5. **Date**: date in which the crime occurred;
6. **Time**: time in which the crime occurred;
7. **PdDistrict**: Police department district;
8. **Resolution**: kind of punishment given to the criminal to resolve the case;
9. **Address**: address where the crime scene happened;
10. **X**: longitude of the crime location;
11. **Y**: latitude of the crime location;
12. **Location**: Exact location (latitude, longitude);
13. **PdId**: Pd ID.

The venues related to each district in San Francisco are retrieved from Foursquare API, using the same procedure displayed in the assignments of the previous weeks.

San Francisco Police Department addresses are instead retrieved here (https://sfgov.org/policecommission/police-district-maps), while the geographical boundaries of each district is available at the following website (https://data.sfgov.org/Public-Safety/Current-Police-Districts/wkhw-cjsf).

3. Methodology

3.1. Preprocessing crime data

San Francisco crime dataset is structured in a table, whose columns have been described in Section 2.1. The first operation consists in removing faulty data points, such as rows containing missing or NaN values; Successively, the columns of interest are extracted from such table, which are the ones related to: district name in which the crime occurred; type of crime; geographical coordinates of each event. The result of such operation is displayed in the figure below, where the last five rows of the dataframe are represented.

	PdDistrict	Category	X	Y
150494	TENDERLOIN	WEAPON LAWS	-122.409661	37.786439
150495	TENDERLOIN	WEAPON LAWS	-122.411966	37.784914
150496	TENDERLOIN	WEAPON LAWS	-122.412054	37.781614
150497	TENDERLOIN	WEAPON LAWS	-122.416711	37.783357
150498	TENDERLOIN	WEAPON LAWS	-122.416711	37.783357

Successively, the total number of crimes per district must be extracted. This operation can simply be performed by counting the number of rows having the same *PdDistrict* field. The result is displayed below.

```
----------------------------------------
District      Number of crimes

SOUTHERN      28445
NORTHERN      20100
MISSION       19503
CENTRAL       17666
BAYVIEW       14303
INGLESIDE     11594
TARAVAL       11325
TENDERLOIN     9942
RICHMOND       8922
PARK           8699
Name: PdDistrict, dtype: int64
----------------------------------------
```

3.2. Location data retrieval from Foursquare API

Information about the local venues pertaining to each district can be easily retrieved through API calls using Foursquare, as shown in the assignments of the previous weeks. For each district, it is possible to retrieve all the venues located within a certain radius from the district center. Considering that San Francisco districts have a relatively large extension, it was decided to extract venues data over a radius of 3500 m from the district center. The result of this operation reveals the presence of a total number of 1000 (this small number is a limitation of the Foursquare “free” account, there more

than 1000 venues in the whole city of San Francisco!) venues, which belong to 181 different categories. Each venue is then associated to its district through the evaluation of its geographical coordinates. Successively, one-hot encoding of the 181 categories is performed.

	PdDistrict	Accessories Store	Adult Boutique	African Restaurant	American Restaurant	Aquarium	Arcade	Argentinian Restaurant	Art Gallery	Art Museum	...	Trail	Turkish Restaurant	Udon Restaurant	Vietn Res
0	BAYVIEW	0	0	1	1	0	0	0	1	0	...	0	1	0	
1	CENTRAL	0	0	0	0	0	0	0	0	3	...	2	0	0	
2	INGLESIDE	0	0	0	0	0	0	0	0	0	...	1	1	0	
3	MISSION	0	1	0	0	0	0	1	0	0	...	0	0	0	
4	NORTHERN	0	1	0	1	0	0	0	0	1	...	1	0	0	
5	PARK	1	0	0	0	0	1	0	1	1	...	1	0	0	
6	RICHMOND	1	0	0	0	1	1	0	0	1	...	1	0	0	
7	SOUTHERN	0	0	0	0	0	0	0	1	2	...	0	0	0	
8	TARAVAL	0	0	0	0	0	0	0	0	0	...	2	0	1	
9	TENDERLOIN	0	0	0	1	0	0	0	0	3	...	0	0	0	

10 rows × 182 columns

3.3. Construction of the final dataset

Similar venues (e.g., restaurants, fast food places, pizza places, etc.) are then grouped together and are assigned a categorical label. Lastly, the total number of crimes dataset is merged with the one just discussed. The final dataset is displayed in the figure below.

	PdDistrict	Latitude	Longitude	Number of Crimes	Restaurants	Stores/Shops	Groceries	Sports Facilities	Entertainment/Culture	Landscape	Other
0	BAYVIEW	37.729978	-122.398246	14303.0	30	20	8	12	4	10	2
1	CENTRAL	37.798769	-122.409932	17666.0	24	25	5	8	8	10	4
2	INGLESIDE	37.726698	-122.446569	11594.0	36	24	9	8	0	15	2
3	MISSION	37.762997	-122.421984	19503.0	21	32	8	19	7	3	1
4	NORTHERN	37.780146	-122.432471	20100.0	20	24	11	15	9	5	2
5	PARK	37.767771	-122.455166	8699.0	19	26	9	8	8	21	2
6	RICHMOND	37.760460	-122.462860	8922.0	18	23	7	9	5	26	1
7	SOUTHERN	37.772236	-122.389044	28445.0	16	29	8	20	9	4	4
8	TARAVAL	37.743731	-122.481459	11325.0	31	27	11	10	1	14	1
9	TENDERLOIN	37.783675	-122.412919	9942.0	21	28	5	13	12	5	3

3.4. Unsupervised clustering using HDBSCAN algorithm

Before performing unsupervised clustering through HDBSCAN (a detailed explanation of such algorithm can be found at https://hdbscan.readthedocs.io/en/latest/how_hdbscan_works.html), it is necessary to carry out standardization of the selected features. To this purpose, Z-Score normalization is performed to the columns of the table displayed above, except for the first three columns, which are not given as input to the clustering algorithm.

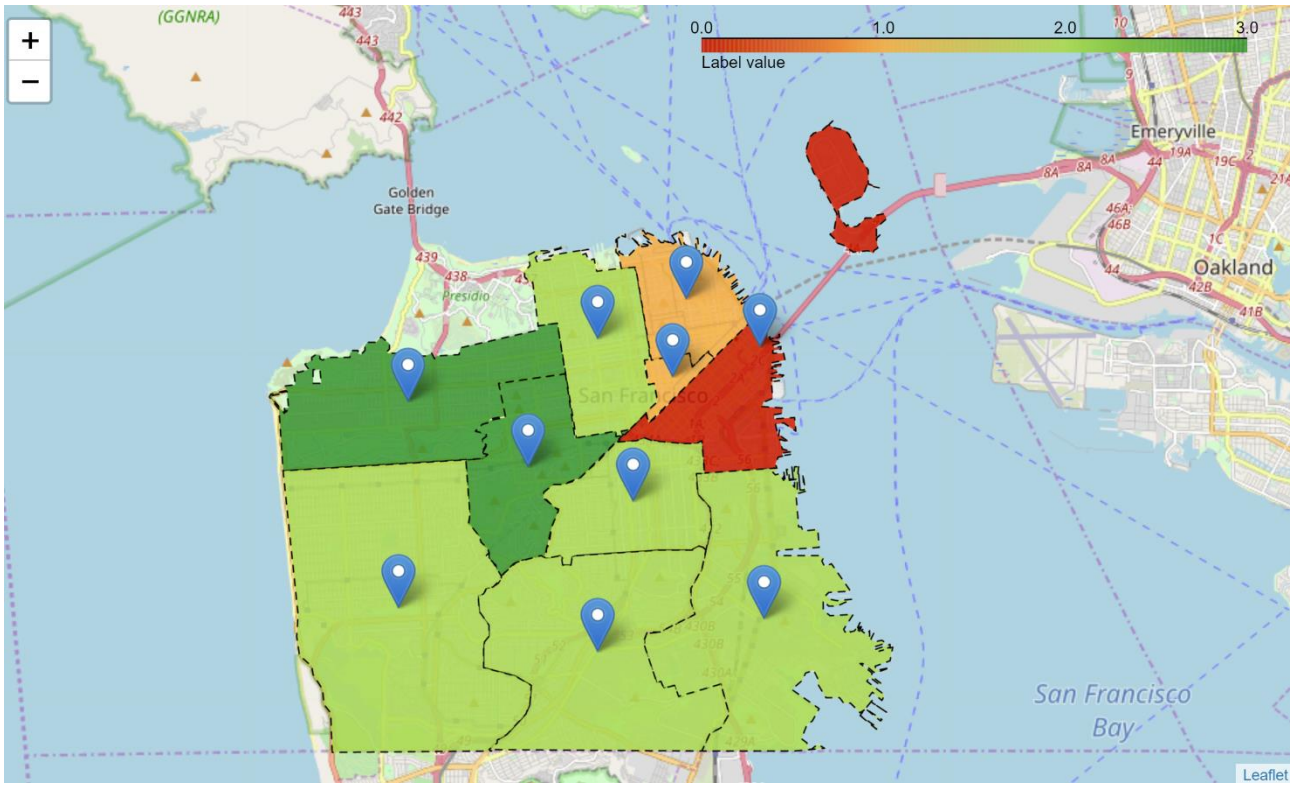
4. Results and discussion

HDBSCAN is able to automatically detect four different clusters: therefore, the districts of San Francisco can be assigned to four different tiers (respectively labeled with 0, 1,…, 3). The values of the tags can be interpreted as to what degree the different districts are safe and attractive: in fact, areas characterized by a large number of crimes (e.g., Southern) are classified as “not desirable” (label value = 0), whereas the most desirable districts achieved a score of 3. Moreover, The presence of landscape elements (e.g., parks, gardens, etc.) positively impacts the desirability of a given area. The most desirable districts (Park and Richmond) not only present the lowest number of crimes, but also have the highest number of landscape elements.

The output of the clustering algorithm (set of labels) is appended as a column to the final dataset, as shown in the figure below.

	PdDistrict	Latitude	Longitude	Number of Crimes	Restaurants	Stores/Shops	Groceries	Sports Facilities	Entertainment/Culture	Landscape	Other	Labels
0	BAYVIEW	37.729978	-122.398246	14303.0	30	20	8	12	4	10	2	2
1	CENTRAL	37.798769	-122.409932	17666.0	24	25	5	8	8	10	4	1
2	INGLESIDE	37.726698	-122.446569	11594.0	36	24	9	8	0	15	2	2
3	MISSION	37.762997	-122.421984	19503.0	21	32	8	19	7	3	1	2
4	NORTHERN	37.780146	-122.432471	20100.0	20	24	11	15	9	5	2	2
5	PARK	37.767771	-122.455166	8699.0	19	26	9	8	8	21	2	3
6	RICHMOND	37.760460	-122.462860	8922.0	18	23	7	9	5	26	1	3
7	SOUTHERN	37.772236	-122.389044	28445.0	16	29	8	20	9	4	4	0
8	TARAVAL	37.743731	-122.481459	11325.0	31	27	11	10	1	14	1	2
9	TENDERLOIN	37.783675	-122.412919	9942.0	21	28	5	13	12	5	3	1

Lastly, the result of the clustering operation is graphically displayed on an interactive map, where the less attractive areas are colored in orange/red, while the most attractive districts are colored in green.



5. Conclusions and future work

The work presented here shows how the relocation process in a new city can be aided through data-driven approaches. In particular, the proposed method can help a person get a better understanding of the individual districts of a target city, allowing to determine the best ones, according to certain evaluation metrics (e.g., crime levels, attractiveness of each district, etc.). The results obtained so far reveal that Park and Richmond districts are the most desirable areas to live in San Francisco. Future work may include testing the performance of the algorithm on other urban environments (other cities, small towns, etc.), but also the introduction of additional features to enrich the model.