Capstone project
The Battle of Neighborhoods

Riccardo Angelo Giro

April 2020

1. Introduction

- **Background**: relocation to a new city is not a simple task, especially in big metropolis such as San Francisco. Several factors must be considered, such as local crime rates and attractiveness factor of different districts
- **Problem statement**: is it possible to determine the best district in San Francisco from a safety and attractiveness point of view?
- **Target audience**: any individual who is interested in relocating to San Francisco
- **Methods**: unsupervised clustering of San Francisco districts, using a procedure based on HBSCAN algorithm

2. Data acquisition

- Sources:
  1. San Francisco crime dataset (2016), available at: https://www.kaggle.com/roshansharma/sanfranciso-crime-dataset;
  2. Foursquare local venues data, obtained through API calls
  3. San Francisco Police Department addresses, available at: https://sfgov.org/policecommission/police-district-maps;
  4. Geographical boundaries of San Francisco districts, available at: https://data.sfgov.org/Public-Safety/Current-Police-Districts/wkhw-cjsf

3.1. Methodology – data processing

- San Francisco crime dataset (structured in a table):
  - Removal of faulty data points;
  - Selection of columns of interest;
  - Evaluation of the total number of crimes per district.
- Foursquare local venues data:
  - Association of venues to the corresponding district;
  - One-hot encoding of each venue type;
  - Grouping similar venues to a common category.
- Final merging of such processed datasets into one, named `df_final`

3.1. Methodology – data processing

```
In [20]: df_final
```

Out[20]:

	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	9941.0	21	28	5	13	12	5	3

3.2. Methodology – data clustering

- Unsupervised clustering of San Francisco districts using the HDBSCAN algorithm
- The columns of the dataframe `df_final` (except *PdDistrict*, *Latitude* and *Longitude*) are given as input to the algorithm, after having performed z-score normalization of the data

4. Results and discussion

- HDBSCAN automatically splits the districts of San Francisco into four different clusters, respectively labelled from 0 to 3 (rightmost column)

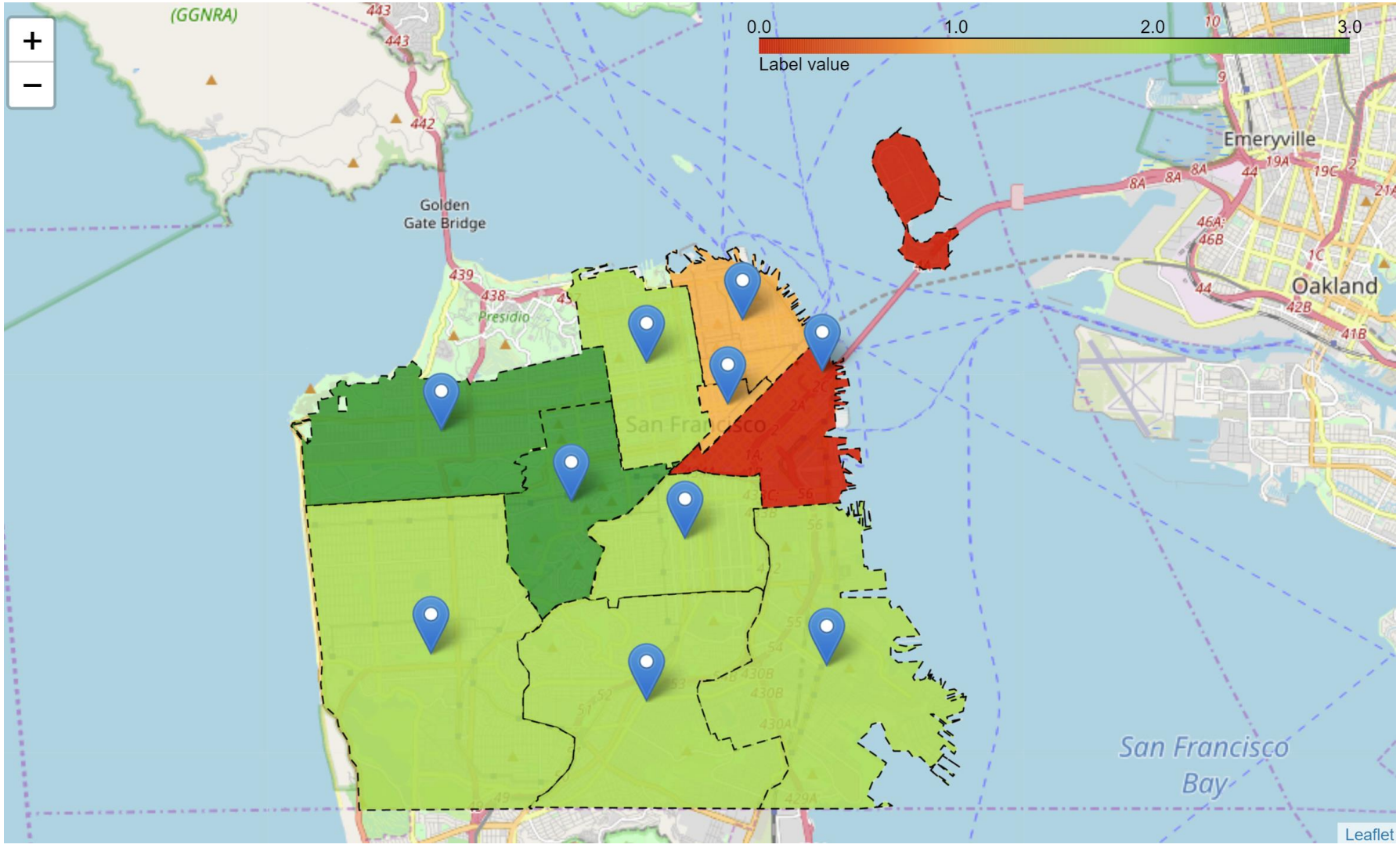
Out[33]:

	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

4. Results and discussion

- Interpretation of the output labels: higher values correspond the safest and most attractive districts
- As expected, areas characterized by a large number of crimes (e.g., *Southern*) are classified as “not desirable” (label value = 0)
- 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

4. Results and discussion (labelled map)



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.)
- Concerning the case of San Francisco, *Park* and *Richmond* districts are identified as the best ones
- **Future work**: testing on other cities; evaluation of additional features (e.g., rent cost of each district, etc.)