Airbnb in San Francisco

Stanislas Hascoet - Coursera Capstone project

April 2019

Introduction

- Airbnb price will most likely depends on its location

→Customers are likely to be interested to know which neighborhood is the most suitable for his wallet.

- What are the keywords that lead to high prices?

- Missing nearest restaurant on Airbnb website

Data

- Airbnb release some the collected data and are free to use.
- Data to be used:

name/description
-price
-neighborhood
-number of reviews
-Reviews per month
-Comments
-Foursquare API
-Data cleaning and preparation:

Data acquisition and cleaning

- Airbnb open data are available as .csv files. Panda will be used to convert this file to a data frame (cropped here for better visibility:

	id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimum_nights	number_of_reviews	last_review	reviews_per_month
0	958	Bright, Modern Garden Unit - 1BR/1B	1169	Holly	NaN	Western Addition	37.76931	-122.43386	Entire home/apt	170	1	180	2019-02-17	1.54
1	5858	Creative Sanctuary	8904	Philip And Tania	NaN	Bernal Heights	37.74511	-122.42102	Entire home/apt	235	30	111	2017-08-06	0.93
2	7918	A Friendly Room - UCSF/USF - San Francisco	21994	Aaron	NaN	Haight Ashbury	37.76669	-122.45250	Private room	65	32	17	2016-11-21	0.15
3	8142	Friendly Room Apt. Style -UCSF/USF - San Franc...	21994	Aaron	NaN	Haight Ashbury	37.76487	-122.45183	Private room	65	32	8	2018-09-12	0.15
4	8339	Historic Alamo Square Victorian	24215	Rosy	NaN	Western Addition	37.77525	-122.43637	Entire home/apt	785	7	27	2018-08-11	0.23

```
id                                0
name                              0
host_id                           0
host_name                         0
neighbourhood_group            7151
neighbourhood                     0
latitude                          0
longitude                         0
room_type                         0
price                             0
minimum_nights                    0
number_of_reviews                 0
last_review                    1377
reviews_per_month              1377
calculated_host_listings_count    0
availability_365                  0
dtype: int64
```

NaN removal

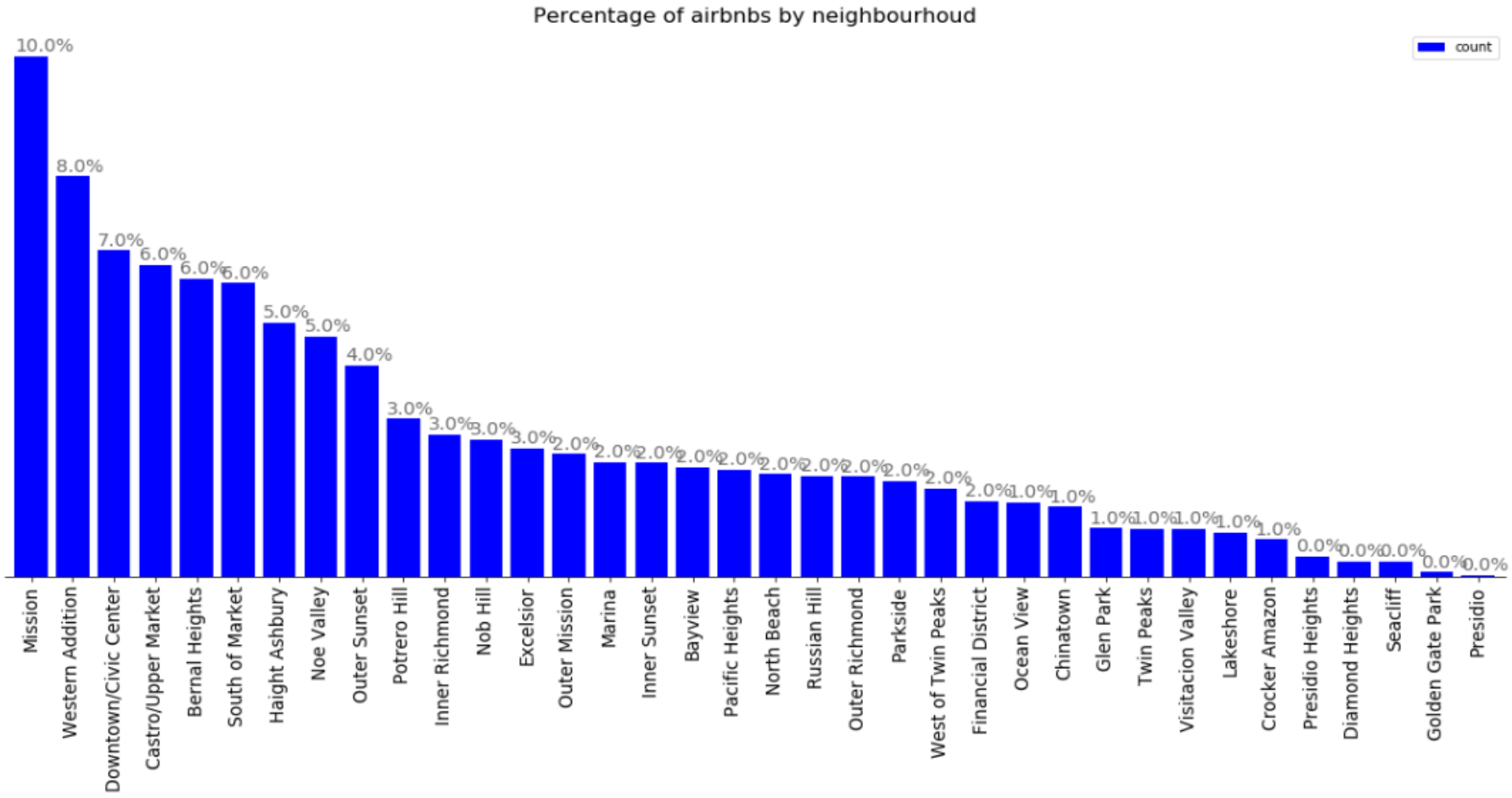
```
id                                0
name                              0
host_id                           0
host_name                         0
neighbourhood                     0
latitude                          0
longitude                         0
room_type                         0
price                             0
minimum_nights                    0
number_of_reviews                 0
last_review                       0
reviews_per_month                 0
calculated_host_listings_count    0
availability_365                  0
dtype: int64
```

Method

- Histogram (matplolib) will be used to plot the bar chart of the count of airbnb as a function of neighbourhoods
- Chroropleth mapis also used to show the neighborhoods with highest Airbnb count+ Geocoder from geopy is used to center the map to San Francisco.
- Scatter plot (matplotlib) will then be used to evaluate correlation between different data
- Word Cloud (WordCloud) is used to evaluate the most frequent word in the name/reviews of the airbnbs (“San” and “Francisco” and “Home” added to the stop words)

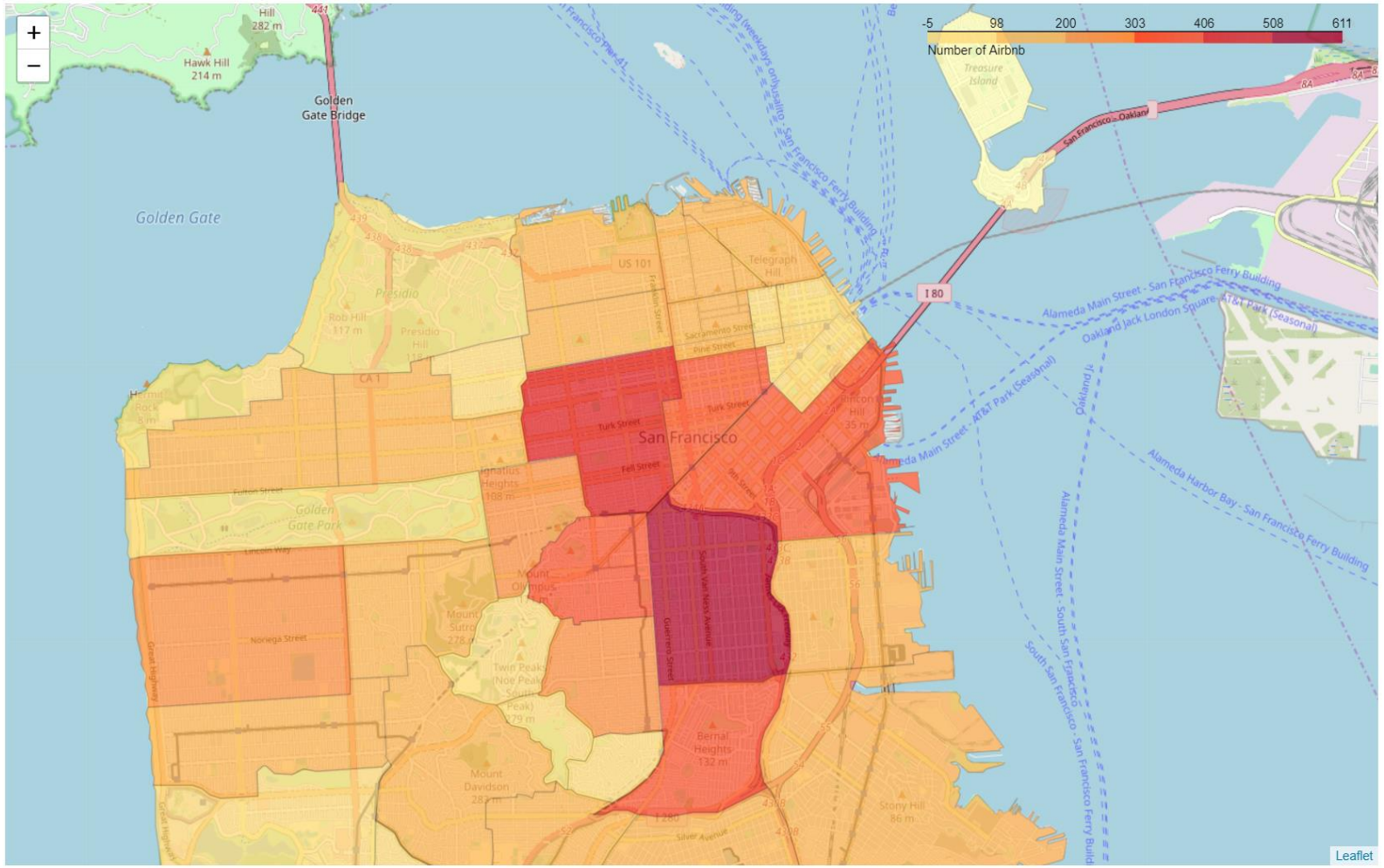
- Use FourSquare API to get close point of interest (limit set to 100 and radius to 500)
- Filter json file from API and convert to database
- Filter by restaurant, food, pizza
- Calculate distance using haversine formula and apply to database
- Plot on the map the location of the Airbnb and the closest restaurant (distance readable in the tag)

Result: Airbnb count



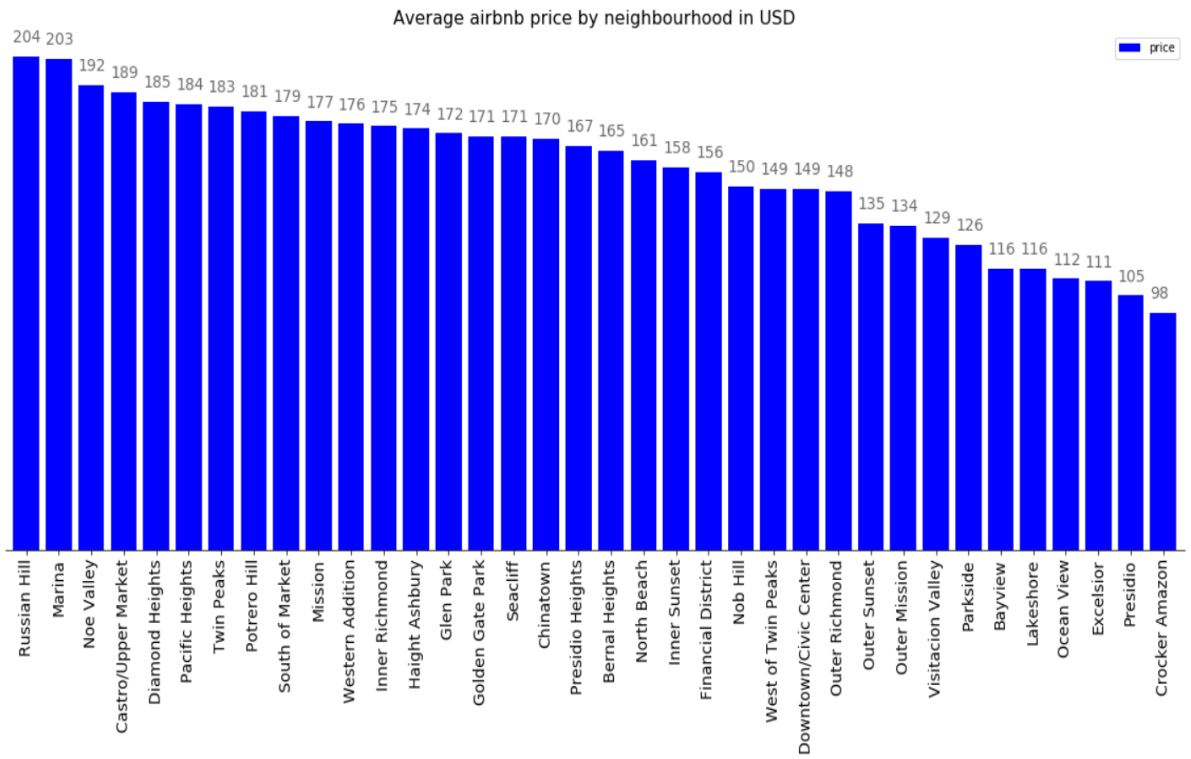
From the above histogram it can be deducted that nine neighborhood share a major part of the total Airbnb.

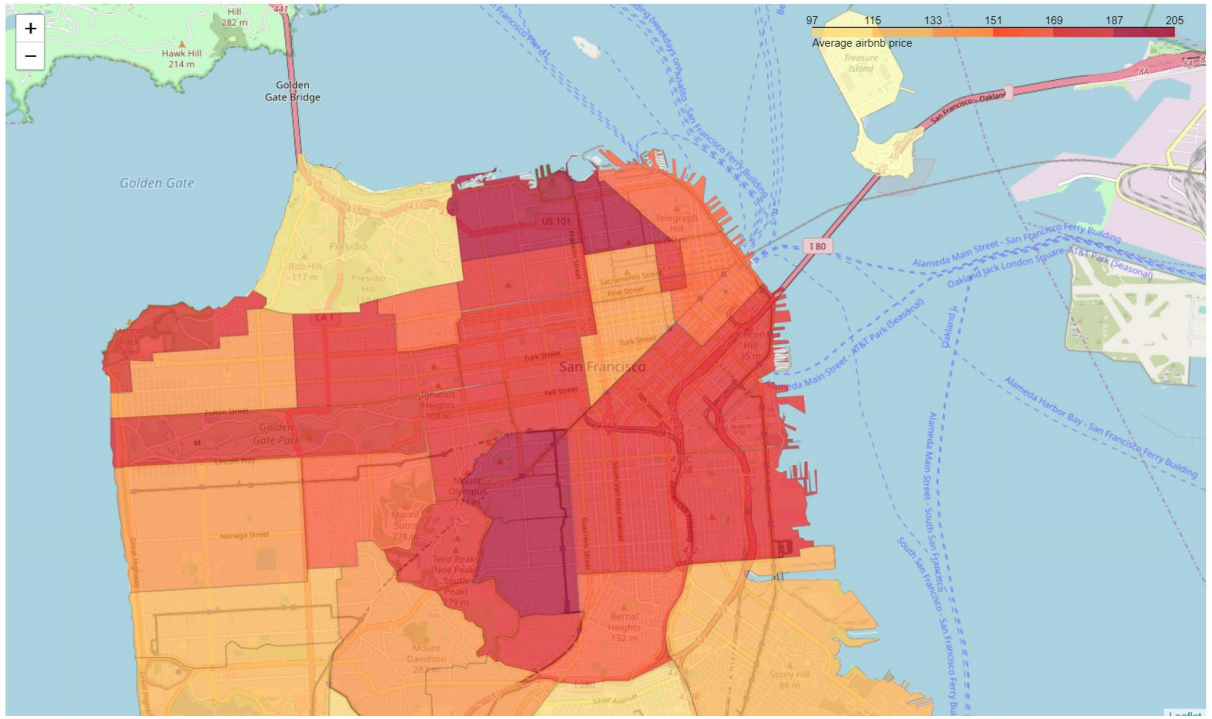
Result: Airbnb count



From the map, the Airbnb are located mainly around the neighborhood Mission

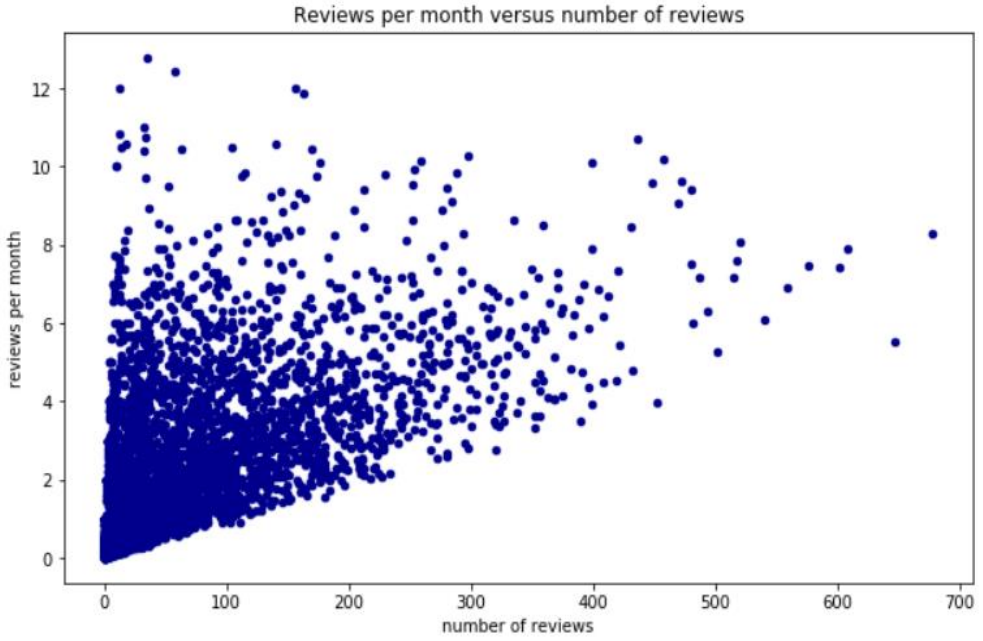
Result: Airbnb price





- The two most expensive neighborhoods for Airbnb are Russian Hill and Marina with respectively an average price of 204 and 203 USD per night. Russian Hill offers seems to offers great view on the San Francisco bay but customers will have to pay for it. As a comparison, the average price for the cheapest neighborhoods is about 100 USD.

Result: correlation



- From graph on the left, there is no obvious correlation. It still can also be supposed that most popular airbnb (more number of reviews) are in the 100 to 200 USD price range. Number of customer would be needed to normalize the numbers of reviews before exploring correlation possibility more in detail. Ratings would also be a great parameters to look for correlation but moreover for customer numbers/frequency.
- From the graphic on the right, there is a more obvious correlation. It is more likely that an Airbnb get a high number of reviews only if it has a decent number of reviews per month.

Result: Airbnb description/name and price

Full dataset

“Private” is the most frequent keyword. This means that most of the host consider that a private room of flat is appealing.
“Mission” is also very frequent which seems logical as “Mission” the neighborhood that offer the highest number of Airbnb

Result: Airbnb description/name and price

Price below **150** USD

"No big difference are visible compared to word cloud with full dataset but it still can be noticed that luxury disappeared. "Victorian" is also much smaller than on the word cloud for the full dataset.

Result: Airbnb description/name and price

Price above **1000** USD

"No big difference are visible compared to word cloud with full dataset but it still can be noticed that luxury disappeared. "Victorian" is also much smaller than on the word cloud for the full dataset.

Result: Airbnb description/name and price

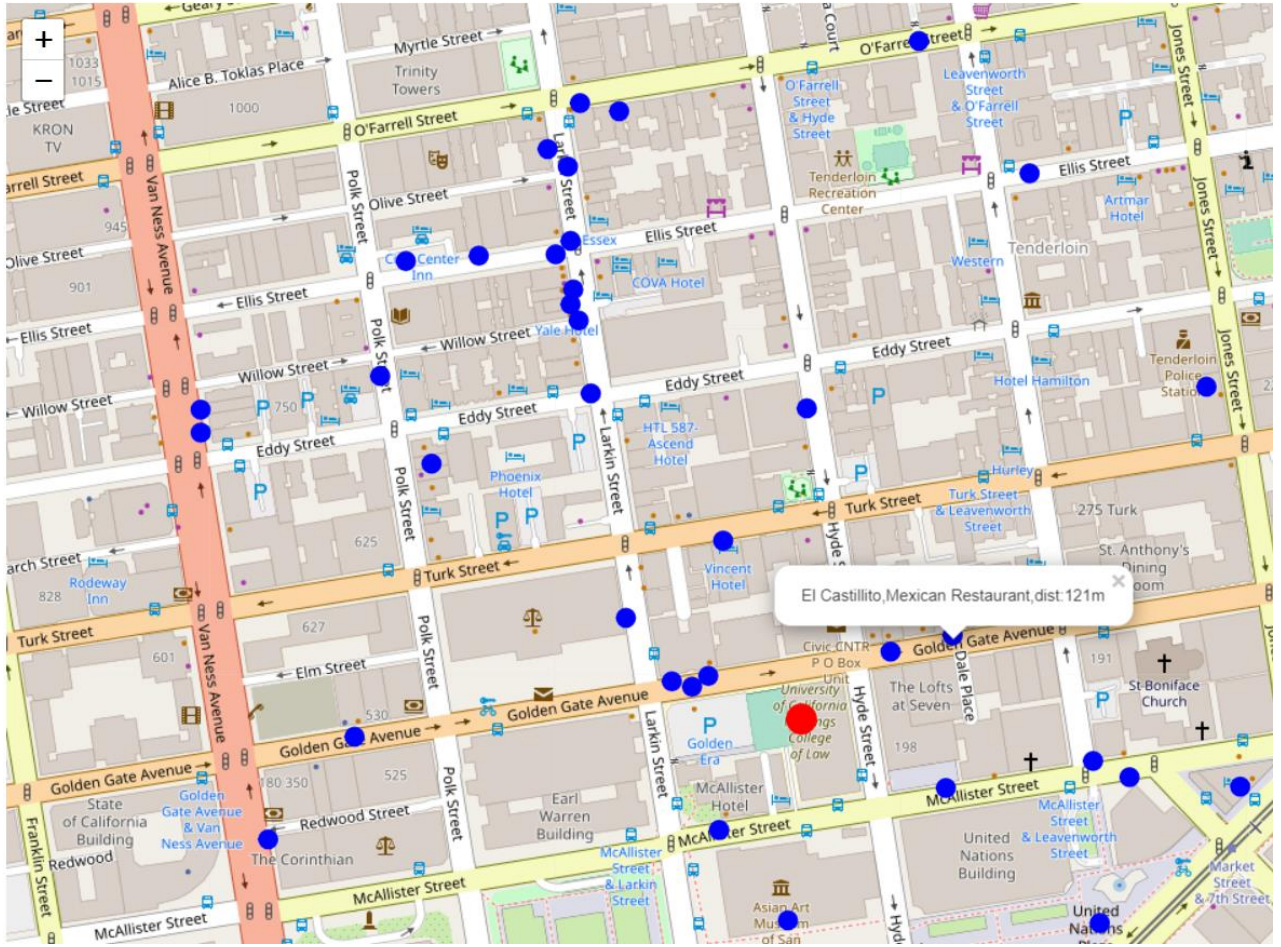
Victorian style

Source: https://www.tripsavvy.com/

This is Victorian style, unlikely to access an Airbnb with Victorian sytle for less than 150USD per night…

Looking at the word cloud with “great” for comments it seems that staying in an Airbnb in San Francisco is a good experience! But be careful of cancellation!

Result: Airbnb and nearest restaurants



Check nearest restaurant when selecting the Airbnb for your future trip sound good! It also show the distance to the restaurant: might impact the choice… Try it in the notebook ;)

Conclusion

- Through this project, repartition of Airbnb in San Francisco has been discussed. Price dependency to location has also been demonstrated. Form the available dataset for Airbnb, some important data are missing: the ratings and number of customers but also the revenue of the Airbnb (visible on Airbnb website but no accessible).
- Next steps: explore frequency of keyword frequency versus the ratings. Machine learning could be used to predict the rating from the comments! For instance with the “word” great in the review, a good rating can be expected while with “cancelled” it might be lower rating.