Airbnb in San Francisco

1 Introduction:

Airbnb has become very famous since its creation back in 2008. There is now a bunch of offers with very different prices. Airbnb price will most likely depends on its location so customers are likely to be interested to know which neighborhood is the most suitable for his wallet. There is also chances that some keyword in the name of the Airbnb will lead to higher prices. So customer will be interested to know which characteristic of the rents that impact its price.

Currently on Airbnb website there is no restaurant proposal from a selected Airbnb while the most frequent question to the host is where can we have dinner? This project aim at answering those question.

2 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

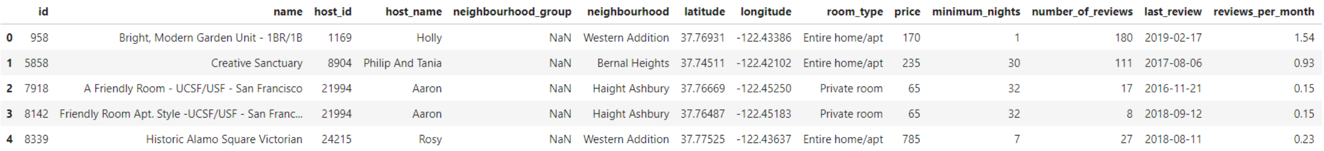
-Comments

-Foursquare API

Data cleaning and preparation:

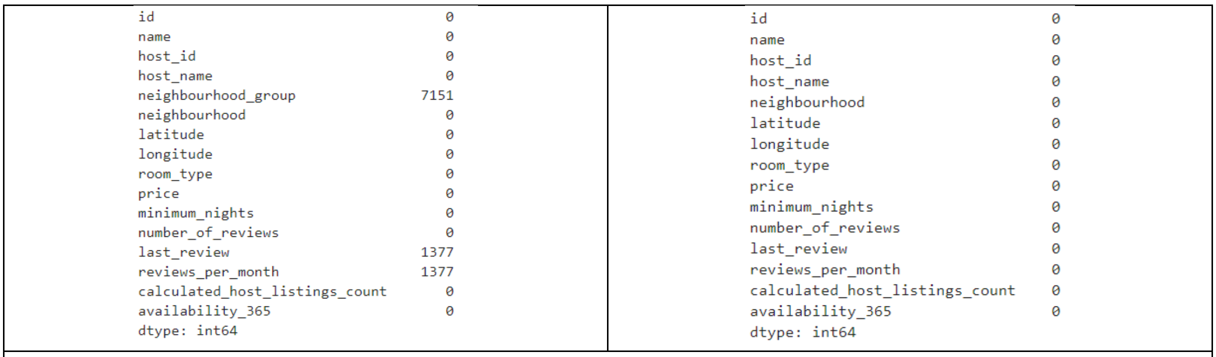
Data acquisition and cleaning:

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

Data cleaning is necessary to get rid of the NaN:



```
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
```

```
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
```

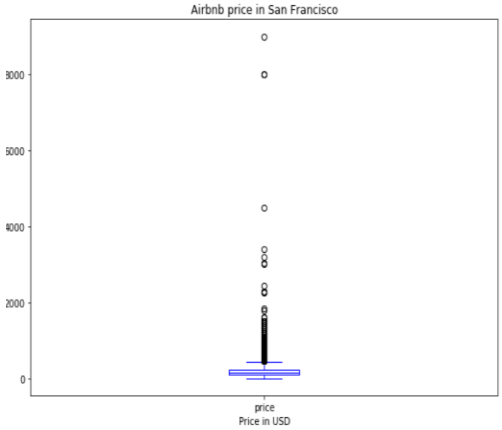
Dataframe summary after NaN removal

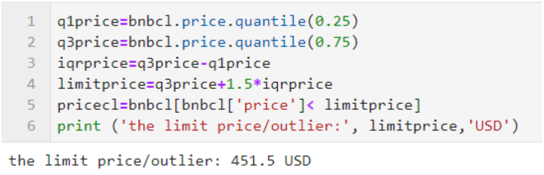
3 Method:

For this project, choropleth (Folium) map will be used to evaluate the concentration of Airbnb in San Francisco as a function of the neighborhoods.

Airbnb price influence study:

- Histogram (matplolib) will be used to plot the bar chart of the count of airbnb as a function of neighbourhoods. Chroropleth map is also used to show the neighbourhoods with higher Airbnb count. Geocoder from geopy is used to center the map to San Francisco.
- To study the price, average price is calculated as a function of the neighborhood. Then a box plot will be used first to evaluate and then eliminate the outliers. Then matplotlib and folium will be use to plit the bar chart and choropleth.





```
q1price=bnbcl.price.quantile(0.25)
q3price=bnbcl.price.quantile(0.75)
iqrprice=q3price-q1price
limitprice=q3price+1.5*iqrprice
pricecl=bnbcl[bnbcl['price']< limitprice]
print ('the limit price/outlier:', limitprice,'USD')
```

```
the limit price/outlier: 451.5 USD
```

Figure 1: Box plot and calculation to remove outliers

- Scatter plot (matplotlib) will then be used to evaluate correlation between different data:

  -Price versus number of reviews
  -Price versus reviews per month
  -Number of reviews per month versus number of reviews

- Word Cloud (WordCloud) is used to evaluate the most frequent word in the name/descriptions of the airbnbs
  -Full data base will be used first
  -Data based will be split into two different categories of prices

- Word Cloud (WordCloud) is used to evaluate the most frequent word in the comments of the airbnbs
  -“San” and “Francisco” added to the stop words
  -it is for information, as the ratings are not available for machine learning

Airbnb closest restaurant suggestions:

- Random Airbnb will is selected from the dataframe :extract latitude and longitude
- 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)

4 Results:

4.1.1 Airbnb count and price

Neighborhoods concentration of Airbnb in shown on below histogram

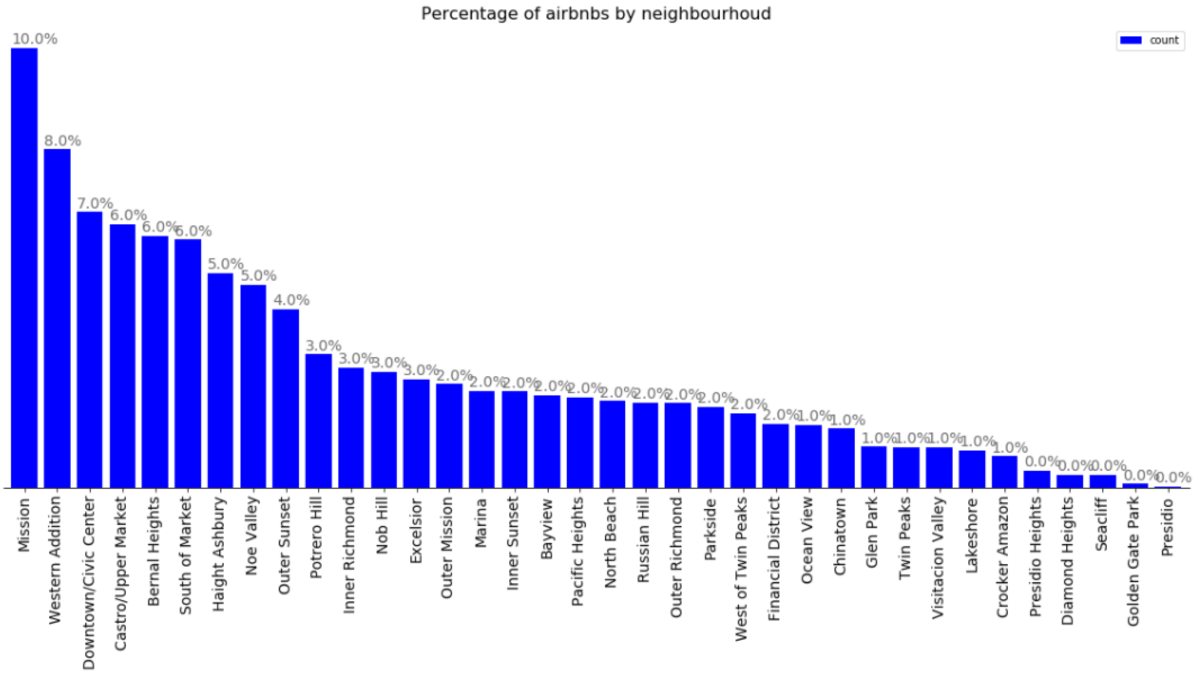


Figure 2: Histogram for Airbnb concentration by neighborhood

From the above histogram it can be deducted that nine neighborhood share a major part of the total Airbnb. This is also visible on the map:

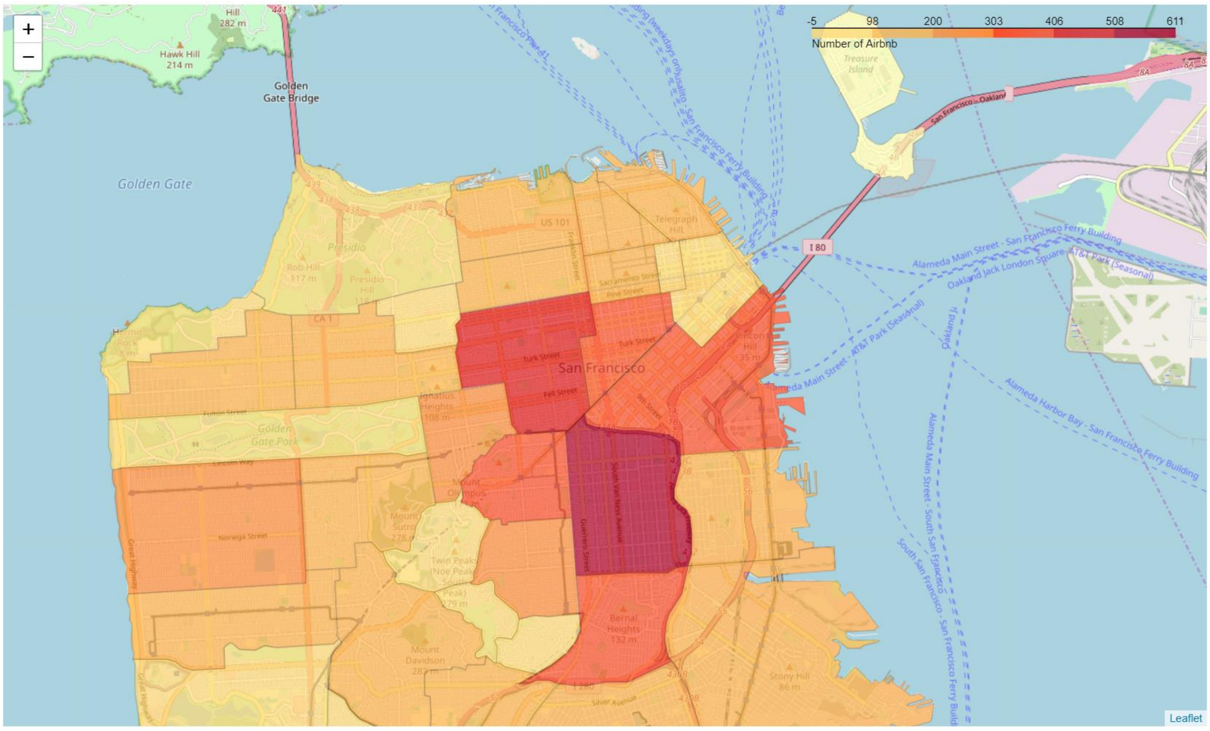


Figure 3: Chroropleth for airbnb concentration

From the map, the Airbnb are located mainly around the neighborhood Mission. It is interesting to compare Airbnb concentration to price:

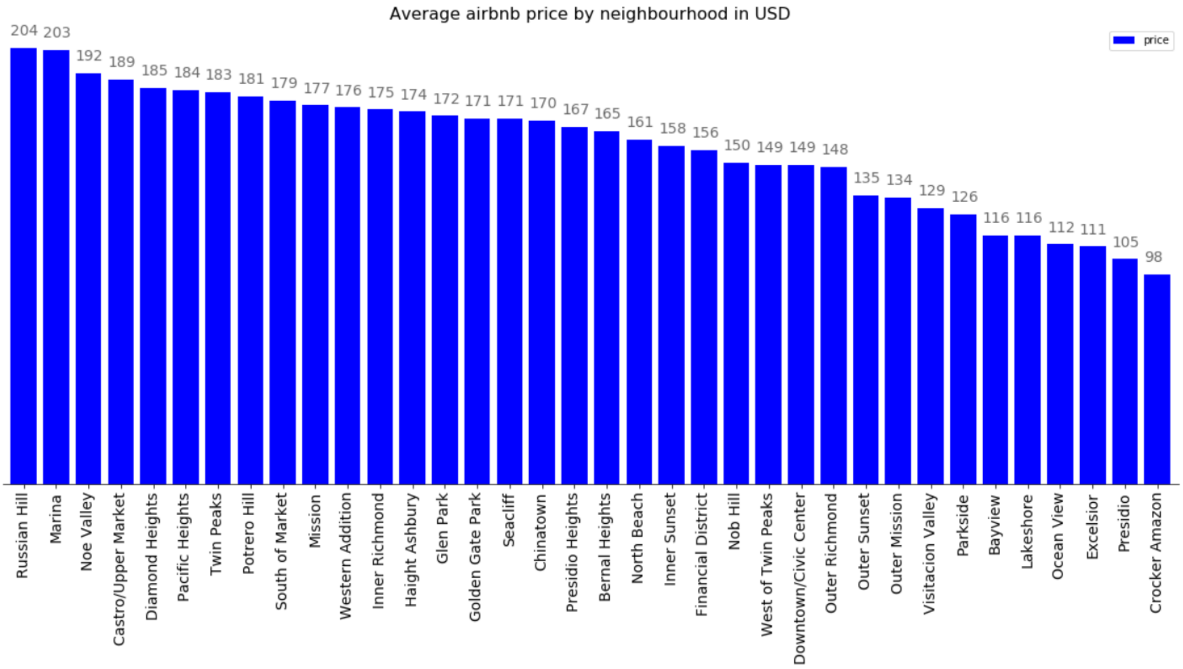


Figure 4: Histogram price repartition by neighborhood

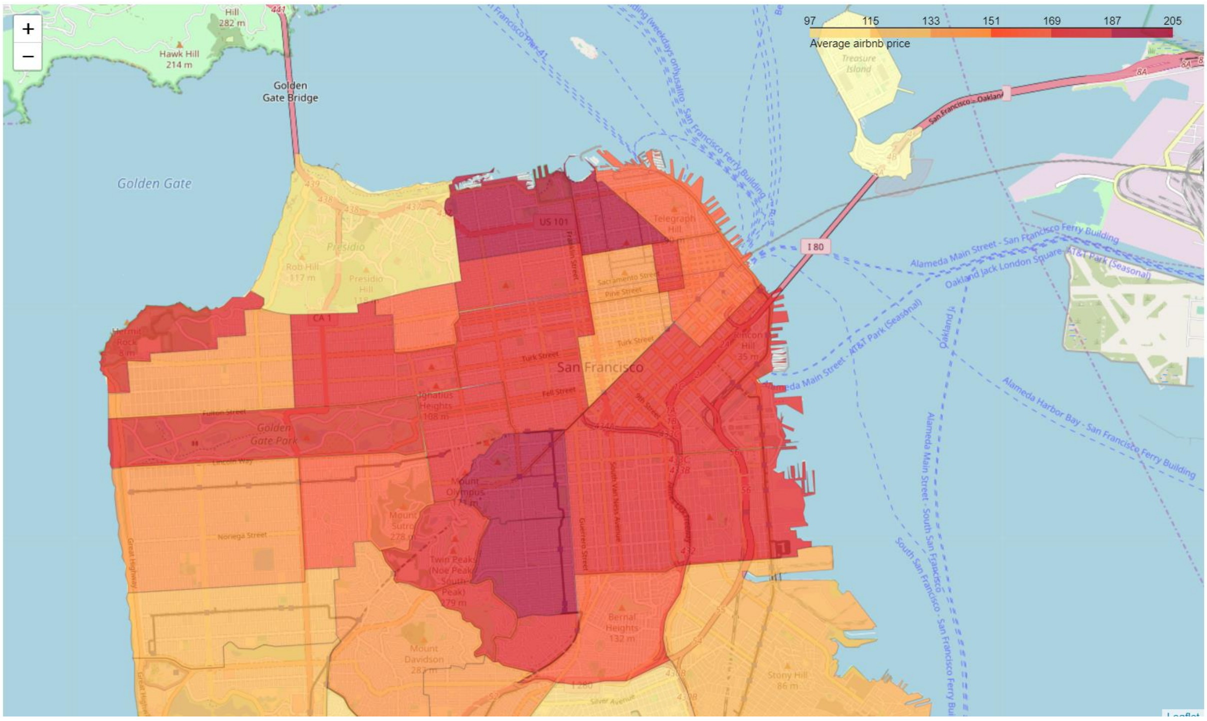


Figure 5: Choropleth map for price concentration

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.

4.1.2 Correlation:

Available data have been used to evaluate potential correlation:

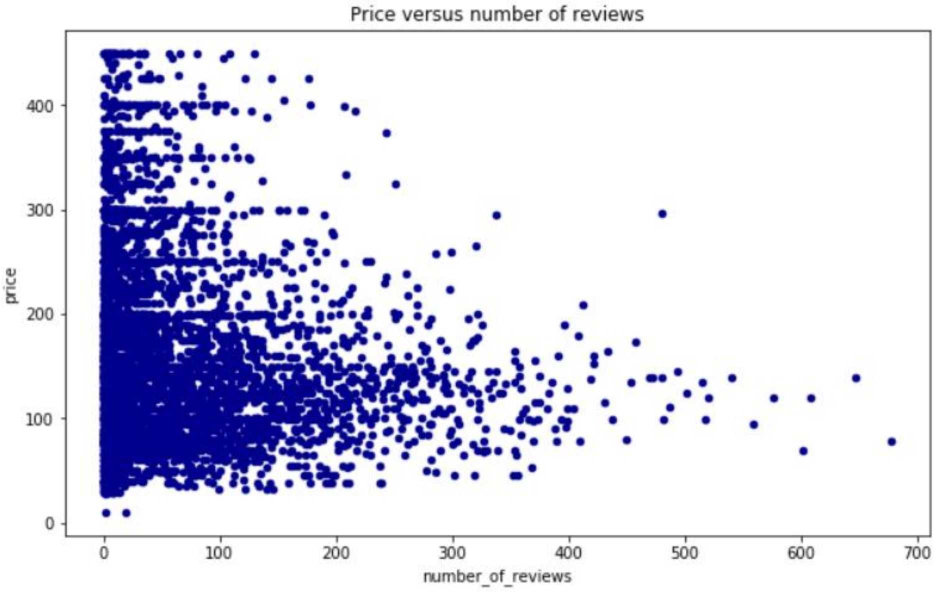


Figure 6: Scatter Plot number of reviews versus price

Even if a trend can be seen, there is no obvious correlation. It 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.

Remark: Review per month versus price and is very similar to the above scatter plot.

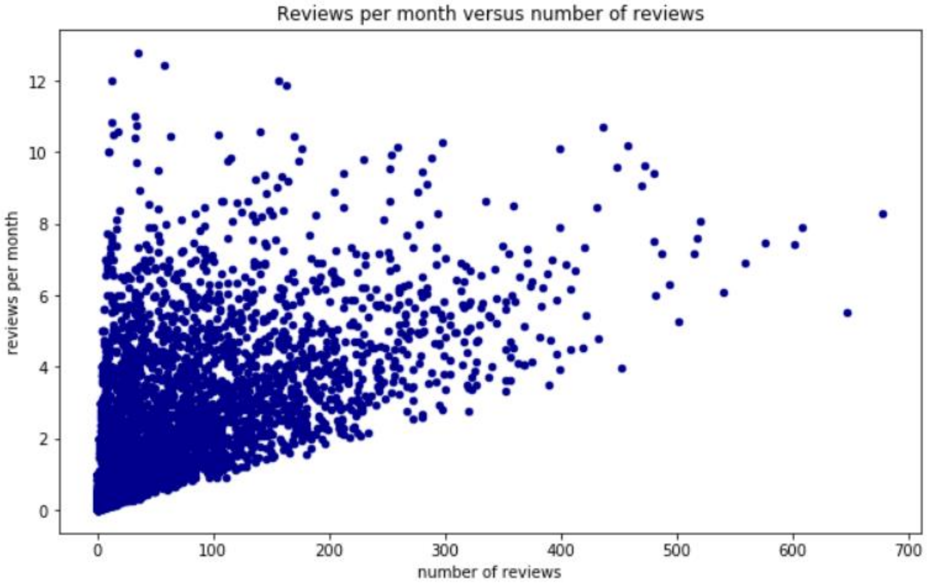


Figure 7: Scatter plot for number of reviews versus reviews per month

From the above graphic, 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.

4.1.3 Airbnb description/name and price

Most frequent word used in the description/name for Airbnb in San Francisco can be visualized using a word cloud:

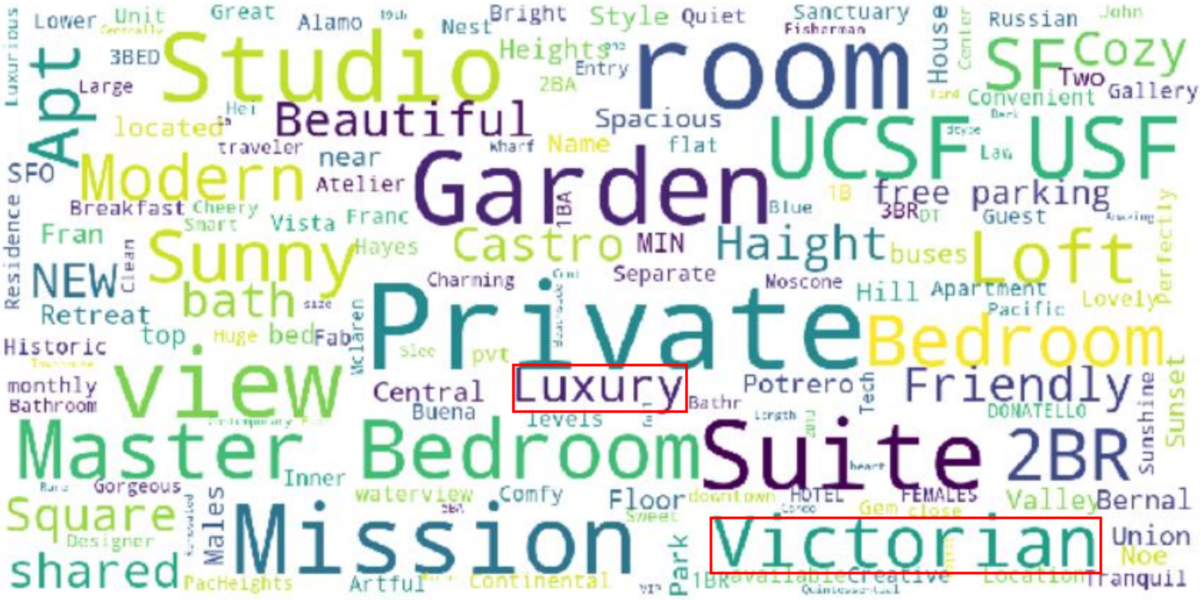


Figure 8: Word cloud for full database

It is interesting to see that “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 (10% see Figure 2).

Same exercise can be done with two different price range: below 150 USD per night and above 1000 USD per night.

<u>Below 150 USD:</u>

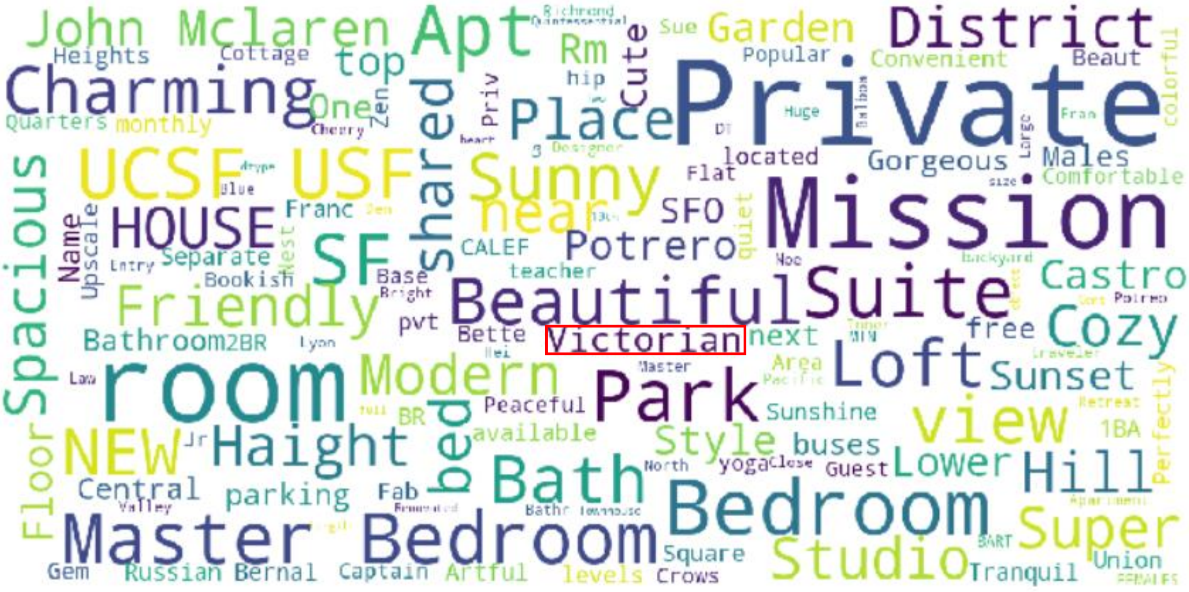


Figure 9: Word cloud for price below 150 USD

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

Above 1000 USD:

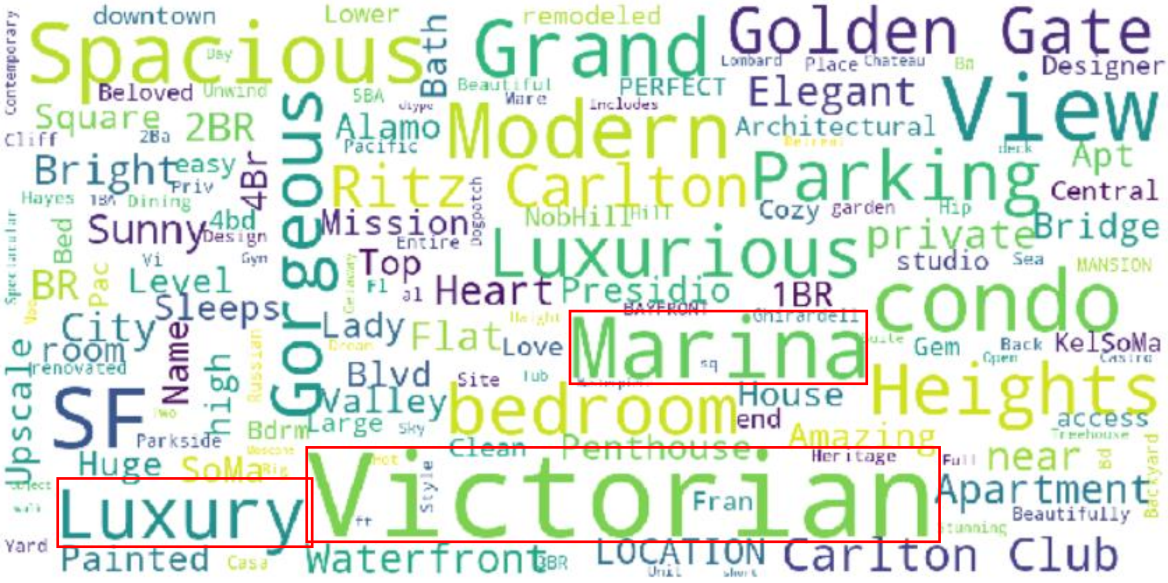


Figure 10: Word cloud for price above 1000 USD

From above word cloud, “Victorian” is the most frequent word for Airbnb name above 1000 USD per night. Luxury is also more frequent as well as Marina which is the second most expensive neighborhoods. Also here is more frequent word about the view and the size of the Airbnb while private disappeared (private is a minimum at 1000 USD per night, isn’t it?).

This is Victorian style… Looking at the word cloud with “great” for comments it seems that staying in an Airbnb in San Francisco is a good experience! (be careful about cancellations!)

Figure 11: Victorian style (source : https://www.tripsavvy.com/)

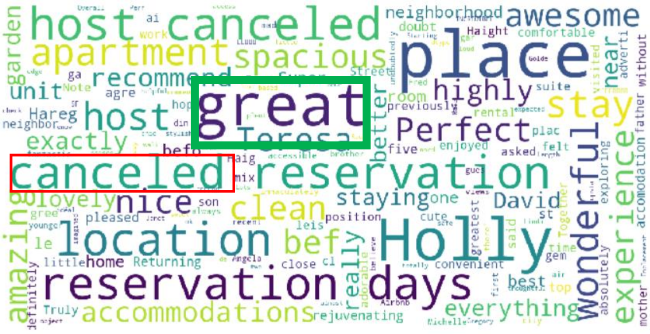


Figure 12: Word Cloud for reviews

4.1.4 Airbnb and nearest restaurants:

Now customers where to find an Airbnb suitable for their wallet, they also know what are the keyword they should look for (or not) in the description depending on their budget. Last thing is to find where to eat when you arrive (red dot is the Airbnb location):

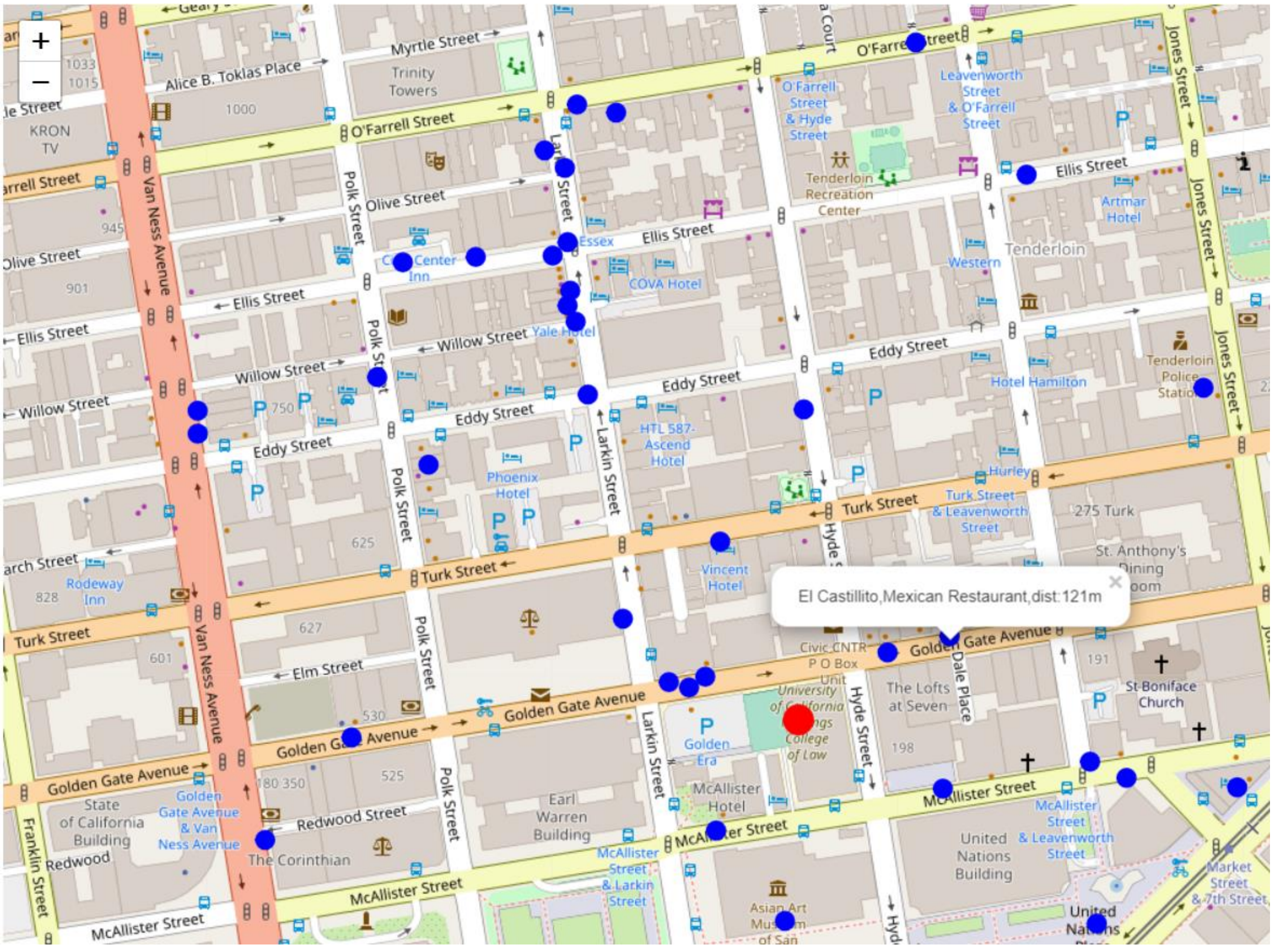


Figure 13: Nearest restaurant

Access the interactive map here:

https://github.com/stany298792/Coursera_Capstone/blob/master/SF_Airbnb_capstone%20project.html

Number of displayed restaurant depends on the Airbnb location, and if the customer is tired she/he can check the distance.

5 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.