“Chicago’s Airbnb price predictor”

Rabab Mohamed & Gabriel Palacios

DS-320, “Data Analysis & Visualization”

Project Proposal

01//20/22

DS-320 Project Proposals

Project title: "Chicago's Airbnb price predictor"

Team members: Rabab Mohamed & Gabriel Palacios

Motivation

As people try to get into new businesses to increase their income after a hard year, we found that there is no way to predict how much someone should charge for an Airbnb listing. To make that transition easier we want to create a model that can predict the price of the market for a new place taking into account its characteristics. This model then could be used to create web applications or mobile apps and help people to get into the Airbnb business.

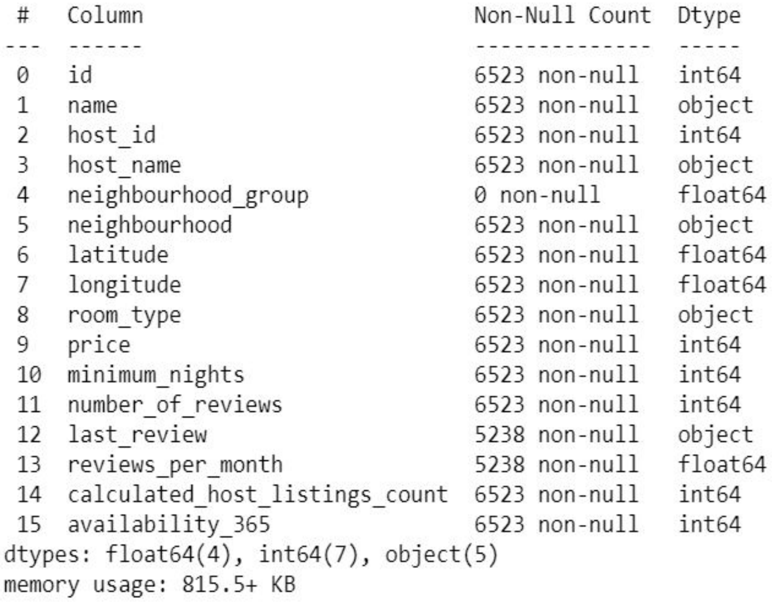
Data

Source: This dataset belongs to *Inside Airbnb*. Specifically, we will use Chicago's dataset posted on December 20th, 2020. It can be found at http://insideairbnb.com/chicago/.

Description: As will use a raw dataset from *Inside Airbnb*, then we will have to spend a lot of time cleaning it. The data consists of 16 columns and 6523 rows.

Attributes information: Id, name, host_id, host_name, neighbourhood_group, neighbourhood, latitude,longitude, room_type, price,minimum_nights, number_of_reviews, last_review, reviews_per_month, calculated_host_listings_count, and availability_365.

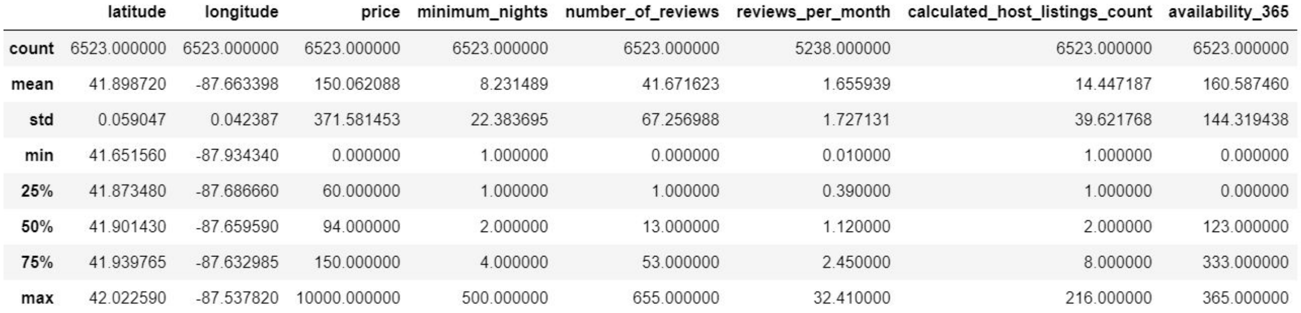
Dimensions: On figure 1, we can observe that the dataset has 16 columns and 6523 rows. There is an empty column neighbourhood_group and 1285 missing values from last_review and review_per_month.



```
 #   Column                          Non-Null Count  Dtype  
---  ------                          --------------  -----  
 0   id                              6523 non-null   int64  
 1   name                            6523 non-null   object 
 2   host_id                         6523 non-null   int64  
 3   host_name                       6523 non-null   object 
 4   neighbourhood_group             0 non-null      float64
 5   neighbourhood                   6523 non-null   object 
 6   latitude                        6523 non-null   float64
 7   longitude                       6523 non-null   float64
 8   room_type                       6523 non-null   object 
 9   price                           6523 non-null   int64  
 10  minimum_nights                  6523 non-null   int64  
 11  number_of_reviews               6523 non-null   int64  
 12  last_review                     5238 non-null   object 
 13  reviews_per_month               5238 non-null   float64
 14  calculated_host_listings_count  6523 non-null   int64  
 15  availability_365                6523 non-null   int64  
dtypes: float64(4), int64(7), object(5)
memory usage: 815.5+ KB
```

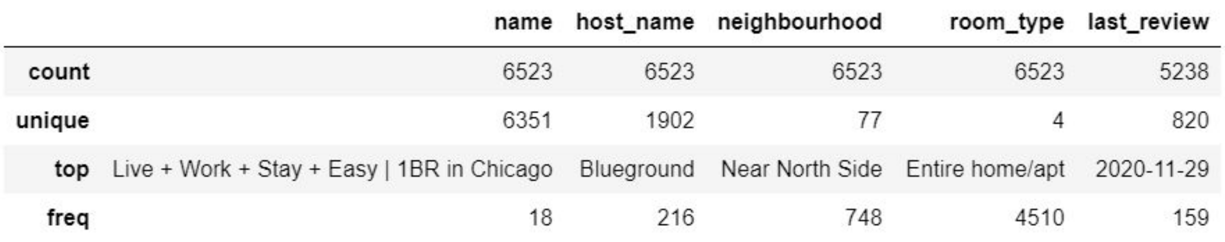
Figure 1 - Columns, entries and data types.

Discussion: The data is not pre-processed. It has some missing values and five categorical attributes (see figure 3). We would need to do some processing as one-hot encoded, labeling and data cleaning.



	latitude	longitude	price	minimum_nights	number_of_reviews	reviews_per_month	calculated_host_listings_count	availability_365
count	6523.000000	6523.000000	6523.000000	6523.000000	6523.000000	5238.000000	6523.000000	6523.000000
mean	41.898720	-87.663398	150.062088	8.231489	41.671623	1.655939	14.447187	160.587460
std	0.059047	0.042387	371.581453	22.383695	67.256988	1.727131	39.621768	144.319438
min	41.651560	-87.934340	0.000000	1.000000	0.000000	0.010000	1.000000	0.000000
25%	41.873480	-87.686660	60.000000	1.000000	1.000000	0.390000	1.000000	0.000000
50%	41.901430	-87.659590	94.000000	2.000000	13.000000	1.120000	2.000000	123.000000
75%	41.939765	-87.632985	150.000000	4.000000	53.000000	2.450000	8.000000	333.000000
max	42.022590	-87.537820	10000.000000	500.000000	655.000000	32.410000	216.000000	365.000000

Figure 2 - Statistics of numerical columns.



	name	host_name	neighbourhood	room_type	last_review
count	6523	6523	6523	6523	5238
unique	6351	1902	77	4	820
top	Live + Work + Stay + Easy | 1BR in Chicago	Blueground	Near North Side	Entire home/apt	2020-11-29
freq	18	216	748	4510	159

Figure 3 - Statistics of categorical columns.

Visualizations of some Airbnb Characteristics

The pie Chart in Figure 4 shows percentage of each room type in the data set. And as observed we can see that 69.14 % Entire home, 1.44% Shared room, 1.09 % Hotel room and 28.33 % Private room.

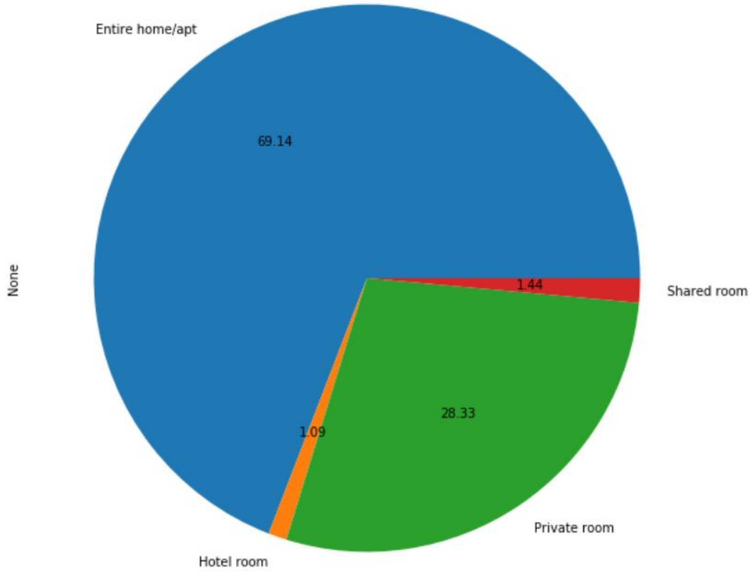


Figure 4 -Pie Chart of room type

The box plot in Figure 5 is a visualization of number_of_reviews values, and the data is right-skewed with some outliers.

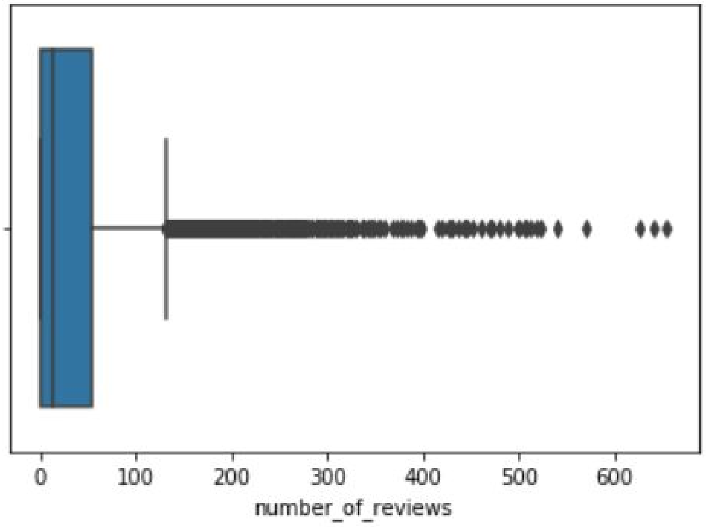


Figure 5 -Box Plot of number_of_reviews

Milestones

We will be working virtually, and we will use Github to collaborate on this project. The following table shows our meeting dates and tasks.

Date & Time	Tasks
01/21/2021 at 1 p.m	**Data:** You need to attach all the data pre-processing, quality and relevance work from the proposal. We specifically studied outlier detection and removal from data so you need to deal with these. Use visualization to support your arguments for outlier detection and use appropriate methods to deal with them. Additionally, you will discuss how you choose sample size? How you defined Training and Test sets? Attribute selection process (You must have to use at least 3 different Visualization techniques and statistical or descriptive summaries to support your arguments). * Make sure your Project 2 has at least 2-3 visualization techniques we studied in the second part of the course. Boxplots, Regression plot, Waffle charts, etc.
01/22/2020 at 1 p.m	**Model/Algorithms:** This portion will discuss what model you used (We studied Regression and Clustering)? Why did you choose a particular model? Why is it suitable? What is efficiency and accuracy of a model? Comparison with other techniques/models? Show predictions on test cases? You need to provide some decent model evaluation metrics to justify your results.
12/23/2020 at 1 p.m	**Significant findings and contributions:** You need to state significant findings and novel contributions to the problem.
01/24/202 at 1 p.m	**Finish all comments on Jupyter notebook and create a PowerPoint presentation & poster.**
01/26/2021 at 1 p.m	**Presentation practice.**
01/27/2021	**Final presentation**