The Battle of Neighborhoods: Airbnb in Mexico City

Guillermo Velazquez
September 5, 2020

1. Introduction

1.1 Business Understanding/Problem Description

Airbnb accommodation is a booming industry with the latest rise in tourism worldwide. This means the demand for Airbnb accommodation is only going to grow further and more people would want to be in the game.

- How should a new business person decide where to open a Airbnb?
- What factors should he look at before investing?
- Which neighborhood venues affect a user's rating for *location* of an Airbnb?

At the same time, it is difficult for a traveller, especially first-timers, to select an Airbnb from among many options. Airbnb reviews are subjective and differ from person-to-person and one cannot solely depend on them to make a decision. It is especially important to consider other aspects like price and neighborhood, which can greatly influence one's experience of the city/country. I will try to answer the following questions

- How does price vary with location?
- How does proximity to transportation affect airbnb rating?
- Suggest similar airbnb but cheaper.

For this project, we will be looking at Airbnbs in Mexico, in particular, Mexico City.

1.2 Target Audience

1. **Traveller:** Help them make an informed decision while choosing an Airbnb by providing an in-depth analysis of Airbnbs and their neighborhood.
2. **Business Person:** Provide useful information and models which can help them where to open their first/next Airbnb.

2. Data

Following are the datasets used in the project along with the reasons for choosing them:

1. **Mexico City Airbnb's Dataset**: The data behind the Inside Airbnb site is sourced from publicly available information from the Airbnb site. The data has been analyzed, cleansed and aggregated where appropriate to faciliate public discussion.
2. **Foursquare API**: This API will help me get the venues around the Airbnbs which I will use for EDA and clustering.
3. **Mexico City Land Price**: I will scrape this website to get land prices of various neighborhoods in Mexico City.

Firstly, I will use the list of Airbnbs from *Mexico City Airbnb's dataset* and use *Foursquare API* to get venues around the Airbnbs. I will then use EDA to explore the neighborhood and how it affects the price of the airbnb. I will also use the combined dataset to cluster similar airbnbs as per pricing and neighborhood.

Second, I will combine the above data with the land price for the area in which the Airbnb is situated and them develop clustering and regression models to predict where a new Airbnb should be opened and how much should it be priced at.

2.1 Mexico City Airbnb Dataset

The original dataset on Insideairbnb has the following columns:

- 'id',
- 'listing_url',
- 'scrape_id',
- 'last_scraped',
- 'name',
- 'summary',
- 'space',
- 'description',
- 'experiences_offered',
- 'neighborhood_overview',
- 'notes',
- 'transit',
- 'access',
- 'interaction',
- 'house_rules',
- 'thumbnail_url',
- 'medium_url',
- 'picture_url',
- 'xl_picture_url',
- 'host_id',
- 'host_url',
- 'host_name',
- 'host_since',
- 'host_location',
- 'host_about',
- 'host_response_time',
- 'host_response_rate',
- 'host_acceptance_rate',
- 'host_is_superhost',
- 'host_thumbnail_url',
- 'host_picture_url',
- 'host_neighbourhood',
- 'host_listings_count',
- 'host_total_listings_count',
- 'host_verifications',
- 'host_has_profile_pic',
- 'host_identity_verified',
- 'street',
- 'neighbourhood',
- 'neighbourhood_cleansed',
- 'neighbourhood_group_cleansed',
- 'city',
- 'state',
- 'zipcode',

- 'market',
- 'smart_location',
- 'country_code',
- 'country',
- 'latitude',
- 'longitude',
- 'is_location_exact',
- 'property_type',
- 'room_type',
- 'accommodates',
- 'bathrooms',
- 'bedrooms',
- 'beds',
- 'bed_type',
- 'amenities',
- 'square_feet',
- 'price',
- 'weekly_price',
- 'monthly_price',
- 'security_deposit',
- 'cleaning_fee',
- 'guests_included',
- 'extra_people',
- 'minimum_nights',
- 'maximum_nights',
- 'minimum_minimum_nights',
- 'maximum_minimum_nights',
- 'minimum_maximum_nights',
- 'maximum_maximum_nights',
- 'minimum_nights_avg_ntm',
- 'maximum_nights_avg_ntm',
- 'calendar_updated',
- 'has_availability',
- 'availability_30',
- 'availability_60',
- 'availability_90',
- 'availability_365',
- 'calendar_last_scraped',
- 'number_of_reviews',
- 'number_of_reviews_ltm',
- 'first_review',
- 'last_review',
- 'review_scores_rating',
- 'review_scores_accuracy',
- 'review_scores_cleanliness',
- 'review_scores_checkin',
- 'review_scores_communication',
- 'review_scores_location',
- 'review_scores_value',
- 'requires_license',
- 'license',
- 'jurisdiction_names',
- 'instant_bookable',
- 'is_business_travel_ready',
- 'cancellation_policy',
- 'require_guest_profile_picture',
- 'require_guest_phone_verification',
- 'calculated_host_listings_count',
- 'calculated_host_listings_count_entire_homes',
- 'calculated_host_listings_count_private_rooms',
- 'calculated_host_listings_count_shared_rooms',
- ‘reviews_per_month'

2.2 Airbnb borough:

This dataset contains all the neighborhoods or venues within 500m radius of a airbnb. It has the following columns:

- ✦ ‘AirbnbName’: Name of the airbnb
- ✦ ‘VenueName’: Name of the venue
- ✦ ‘Category’: It is the primary category of the venue
- ✦ ‘VenueLatitude’, ’VenueLongitude’: Coordinates of the venue

Below is a snapshot of the dataset:

	AirbnbName	VenueName	Category	VenueLatitude	VenueLongitude
0	CHIC apartment for 1 person monthly stay	Parque Arboledas	Park	19.379073	-99.162159
1	CHIC apartment for 1 person monthly stay	Starbucks	Coffee Shop	19.376799	-99.162081
2	CHIC apartment for 1 person monthly stay	Il Forno by 50 Friends	Italian Restaurant	19.373420	-99.162346
3	CHIC apartment for 1 person monthly stay	Farmacia San Pablo	Pharmacy	19.375991	-99.161727
4	CHIC apartment for 1 person monthly stay	Área de Perros Parque Arboledas	Dog Run	19.377361	-99.162016

2.3 Mexico City Land Price:

This dataset contains the locality name and the average price of the land per square meter. Below is a snapshot:

	Alcaldia	Metraje	Terrenos	Precio m2
0	Álvaro Obregón	695025	288	20714
1	Azcapotzalco	682413	109	17255
2	Benito Juárez	226107	487	44167
3	Coyoacán	164882	129	27000
4	Cuajimalpa de Morelos	394801	149	13725
5	Cuauhtémoc	234167	368	45037
6	Gustavo A. Madero	373860	146	14545
7	Iztacalco	79590	64	16949
8	Iztapalapa	367266	148	9960
9	La Magdalena Contreras	99293	84	12678
10	Miguel Hidalgo	308232	341	46193
11	Tláhuac	301334	81	7000
12	Tlalpan	1655199	364	6667
13	Venustiano Carranza	32615	59	18251
14	Xochimilco	467972	128	6278

3. Methodology

3.1 Data Collection:

- The Mexico City Airbnb dataset is freely available on InsideAirbnb and was built by scraping Airbnb.com website.
- We used Foursquare API to get the venues around the hostel.
- We extract the land prices of various boroughs in Mexico City from a web article, https://www.entrepreneur.com/article/342506.

3.2 Analytic Approach:

I took two approaches in the project.

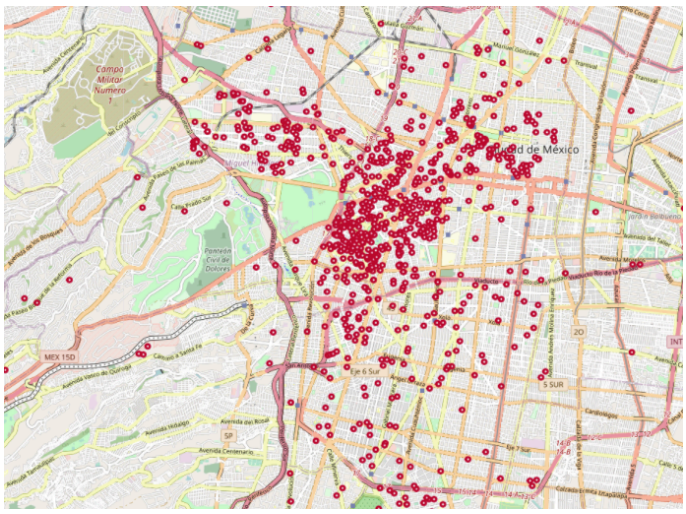
Firstly, I used exploratory data analysis (EDA) to uncover hidden properties of the data and provide useful insights to the reader, both future traveller and investor. I used the list of airbnbs form Airbnb dataset and use Foursquare API to get venues around the Hostel. I will then use EDA to explore the boroughs and how it affects the price of the airbnb. Also I will use the combined dataset to cluster similar hostels as per pricing and neighborhood.

Secondly, I used perspective analytics to help a business person decide a location for new airbnb. I will use clustering (K-Means). I combined the above data with the land price for the area in which the airbnb is situated and then develop clustering models to predict where a new hostel should be opened.

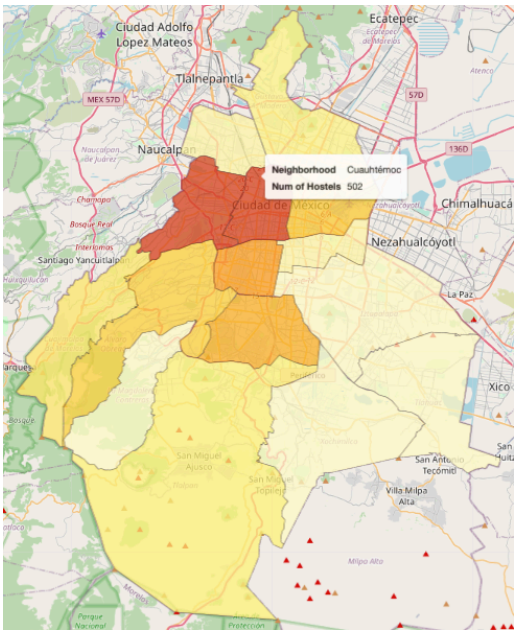
4. Analysis

4.1 Exploratory Data Analysis

We visualized all the airbnbs on a map using Folium and Open Street Maps. Below is the geo-visualization of the airbnbs in Mexico City. As we can see, many airbnbs area located near the centre with density reducing as we move away from it. In the report we have visualized the same map based on multiple criteria and clusters.



For a person interested in opening a new airbnb, it is essential to identify the right area to do so. The main factors to consider while making such decision is demand, supply and cost. Demand and supply usually go hand-in-hand in such cases i.e. we can say



	Neighborhood	Count	Metraje	Terrenos	Precio m2
14	Xochimilco	1	467972	128	6278
12	Tláhuac	1	301334	81	7000
11	Iztapalapa	4	367266	148	9960
10	La Magdalena Contreras	4	99293	84	12678
7	Iztacalco	8	79590	64	16949
6	Azcapotzalco	8	682413	109	17255
8	Gustavo A. Madero	10	373860	146	14545
13	Tlalpan	11	1655199	364	6667
9	Cuajimalpa de Morelos	13	394801	149	13725
4	Álvaro Obregón	19	695025	288	20714
5	Venustiano Carranza	21	32615	59	18251
3	Coyoacán	77	164882	129	27000
2	Benito Juárez	119	226107	487	44167
0	Miguel Hidalgo	148	308232	341	46193
1	Cuauhtémoc	502	234167	368	45037

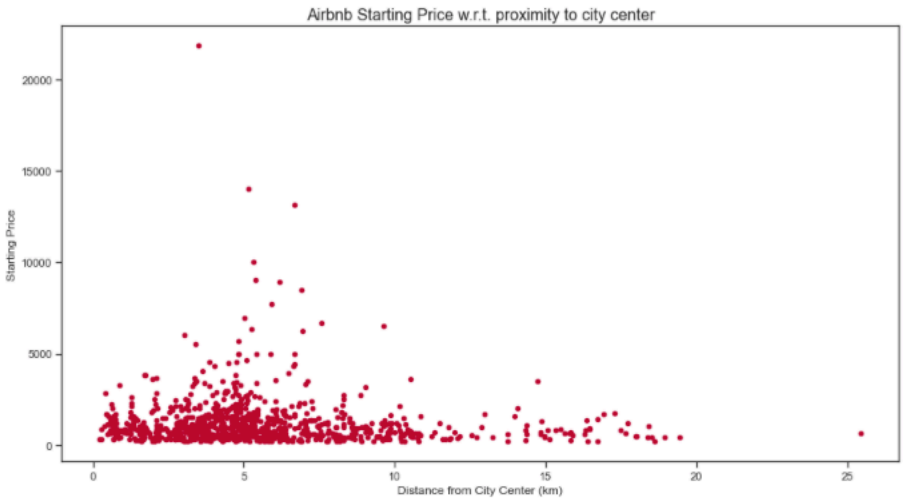
that the more the airbnbs in a region, the higher is the demand. It is evident from the above map that Cuauhtémoc and Miguel Hidalgo are host to many airbnbs. Airbnb density is highest in the boroughs. However, if we compare the land prices in these regions, the area among the most costly localities to buy a property.

Coyoacan seems to be popular since it ranks 4th in the list of number of airbnbs and at the same time, it is not that expensive like the first ones.

The price of property (per sq. meter) in Coyoacan is almost 67% less than Cuauhtémoc which combined with the fact that it is close and has fewer airbnbs than Cuauhtémoc, makes it an exciting prospect for new investors.

Next, we checked if the distance of a hostel from the city centre has any effect on the price.

We can see a very weak negative correlation between the price and distance of the



airbnb form the city center. Our intuition is supported mathematically by the Pearson coefficient which turned out to be -0.067. The correlation coefficient is very less to make any strong inferences.

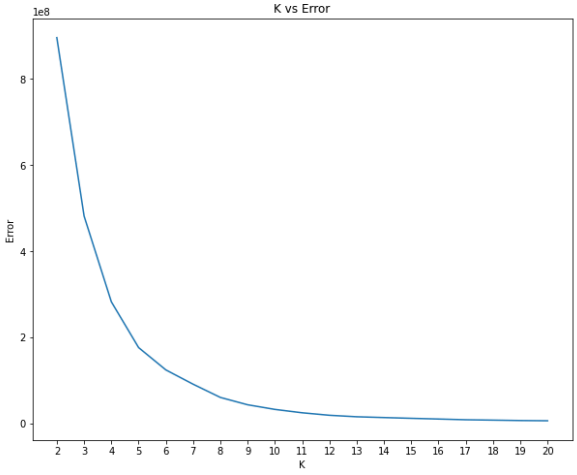
Next, we tried to see how does the neighborhood of a airbnb affect its overall rating. Of course, this alone is not a strong predictor of rating, we aimed to find out some rough patterns.

Following are the things we found out:

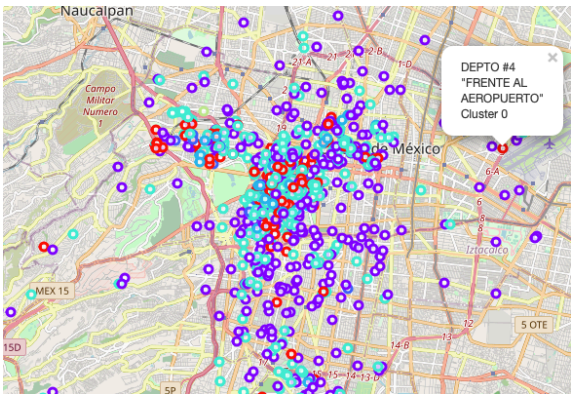
- Proximity to a mode of transportation seemed to affect ratings, since almost 20% of the airbnbs near to a bus/metro station were rated with 9 of 10.
- Proximity to convenience store did not have any effect
- Airbnbs near museums and historic sites had a lower rating than the overall average.

4.2 Clustering

We performed clustering twice based on different set of parameters. First, we clustered using the different rating scores, distance from city centre, and starting price. We used K-Means clustering algorithm and found out the K by using the elbow method. The K on our case is 7, since the error doesn't decrease much after this point.



Let's see a geo-visualization of the clusters and also examine the proprieties of each cluster.



•Cluster 0: Medium High Cost. Good score rating. Very good location and cleanliness rating. Close to city centre.

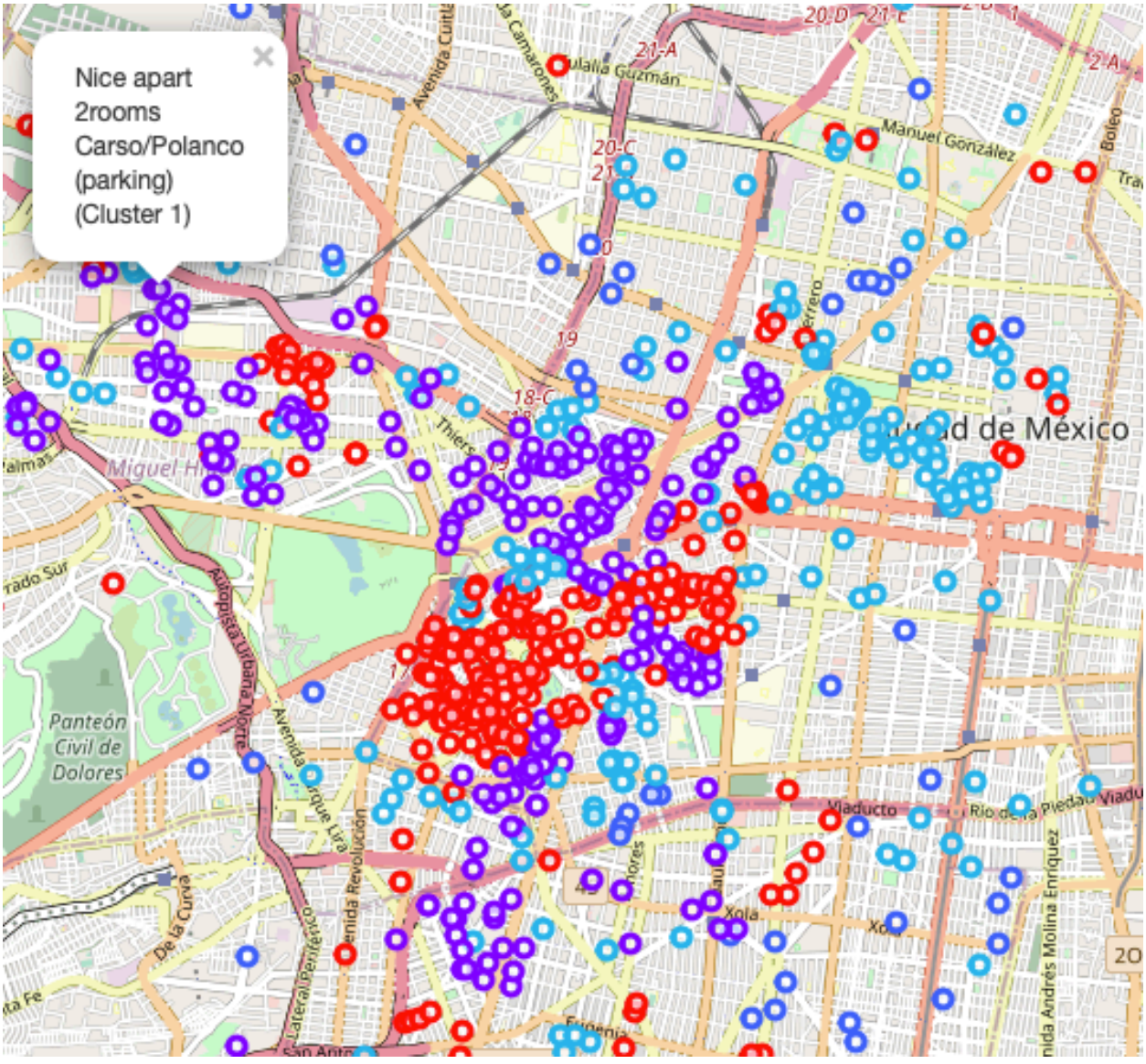
•Cluster 1: Low price. Not far from city center. Very good cleanliness and location rating. Very good score rating.

•Cluster 2: High price. Not far from city center. Very good cleanliness and location rating. Very good score rating.

•Cluster 3: High price. Very good score rating. Very good location and cleanliness ratings. Close to city center.

- Cluster 4: Medium Cost. Not so far from city center. Very high cleanliness and location rating. High score rating.
- Cluster 5: Very High Price. Very good score rating. Very good location and cleanliness. Close from city center
- Cluster 6: Very High Price. Very good score rating. Very good location and cleanliness. Not so far from city center

Second we clustered the airbnbs based on the venues in its vicinity. This time, we fixed K in our K-Means algorithm to be 4 since otherwise each airbnb would be assigned a unique cluster which defeats the purpose.



We can examined each cluster and determine the discriminating venue categories that distinguish each cluster. Based on the defining categories, we can then assign a name to each cluster.

Cluster 0: Variety of places. Coffee Shops, Restaurants, Ice Cream Shops and Bakerys

Cluster 1: Coffee Shops and Mexican Restaurants.

Cluster 2: Taco places and Mexican Restaurant.

Cluster 3: Mexican Restaurants.

5. Result

- We got a glimpse of the airbnb scene in Tokyo and were able to find out some interesting insights which might be useful to traveller as well as people with business interests. Let's summarize our findings:
- Most hostels are located in Miguel Hidalgo, Cuauhtemoc and Benito Juarez.
- Coyoacan seem to be an intersting locality since it is close to Cuauhtemoc and 68% cheaper than it.
- The starting price of hostels doesn't vary much depending on its distance from the city centre.
- Proximity to mode of transportation affects the airbnb rating.

6. Discussion

According to the above analysis, opening an Airbnb in Coyoacan is the most appropriate option. Miguel Hidalgo and Cuauhtemoc contain the highest number of airbnbs from out dataset. This means that these wards must attract lot of tourists, no other reason appropriately justifies why they house so many airbnbs inspite of being very costly. Coyoacan is a borough of Mexico city and is almost 68% cheaper than Cuauhtemoc. It also contain decent amount of airbnbs but not so many as Cuauhtmoc and Miguel Hidalgo, which leaves space for new opportunities.

The clusters will help tourists to identify alternate airbnbs in case their airbnb is not available for some reason. For example, suppose that I want to stay at "**Zen Loft**". However, when I go to book, it shows that it doesn't have any spots available/ Fortunately, I can now use the cluster result to find a airbnb similar to **Zen Loft**.

Some drawbacks of the analysis are that our prescription to new business persons for opening a new airbnb is solely based on neighborhood and land price and not on other factors like how much the business is willing to invest, what facilities will they provide, how will they price the airbnb. Without this data, it is difficult to predict the success of the investment. However, in our analysis, we have ignored this since we don't hace such data and it would be difficult to farm it for small exploratory study like ours.

Hence, our analysis only helps a business person to identify a region to open an airbnb but doesn't guarantee its success.

Also, it would have been beneficial if we had additional features such as crime rate in the locality, and average number of tourists in the locality, This would give us a more complete picture of the neighborhood of an airbnb resulting in better analysis.

7. Conclusion

In the above study, we explored and analyzed various aspects of the airbnb scene in Mexico City, Mexico using data science. We used an existing dataset and combined it with data collected from Foursquare API as well as data extracted from a website. We performed EDA and clustering on the datasets in our pursuit of solutions. We were able to find satisfactory answers to the questions we posed before the study.

The study is based on limited data, bu is is nevertheless a significant step in shedding light on the airbnb scene in Mexico City. This study can be repeated easily for other cities of the world.