Identification of factors affecting rental pricing on Airbnb

Zhang Jia Jun

December 31, 2019

1. Introduction

1.1 Background

Airbnb is an online marketplace for arranging lodging or tourism experiences. Airbnb is established in 2008 and currently it has 150 million people using their website. (Airbnb by the Numbers: Usage, Demographics, and Revenue Growth, n.d.) As many transactions flows through the website each day, it is interesting for us to find out how to price our property rent better.

1.2 Problem

This project aims to analyse the factors affecting the pricing of housing rental on Airbnb website. There will be a greater focus on how different neighborhoods in Madrid affect the prices of the Airbnb in this project.

1.3 Interest

The findings from this report may interest current and potential home lenders by advising them on pricing of their homes. It can also provide information to potential tenant to see if the listed property's rental is overpriced or a good deal.

2. Data Acquisition and Cleaning

2.1 Data Sources

In this project we will be analysing one of Kaggle's dataset titled " Madrid Airbnb Data". This dataset contains information on Airbnb listings in Madrid, Spain. The data is downloaded from: https://www.kaggle.com/rusiano/madrid-airbnb-data#reviews_detailed.csv

In order to supplement this dataset, I will also be using Foursquare API to find out the amenities which are located near respective properties.

2.2 Data Cleaning

Generally the data provided is clean. As such we only need to do some simple cleaning and also formatting of data to get what we want.

As the column labelled "square_feet" refers to the area of the apartment. However, there are some entries which are labelled as 0. We assume that they are labelled wrong and will be replaced with null value.

Column labelled as "prices" contains a currency before the value and are removed to allow for better manipulation of data.

Columns labelled "hosue_rules", "license" and 'instant_bookable" are initially in text format. They are converted into binary option to answer question on whether the presence of "house_rules"/"license"/instant_bookable" will affect apartment prices.

The dataframe which will be used for modelling will have to have null values removed. As such all null values are dropped for dataframe which would be used for modelling in the later section.
Futhermore, columns labelled “property_type”, “room_type”, “cancellation_policy” and “neighbourhood” have to be converted to dummy variables since they contain non-numerical data.

2.3 Feature Selection

From the website there is a total of 6 CSV files and 1 GeoJSON file. “neighbourhoods.geojson” is used to find the coordinate of respective neighborhoods. This is because finding all the amenities of each listing is time consuming and can be approximated by amenities of each neighborhoods instead.

As “listings_detailed.csv” contains most of the information that is required, we will only be using “listings_detailed.csv” as the dataset for data manipulation and modelling. The dataset contains 106 features, of which only 18 features are selected. The rationale for selecting and excluding of the features are tabulated as follows.

Selected Features

Name of feature	Reason for Selection
id	Easy identification of rows and allows for joining of datasets if need so
latitude, longitude	Used for geopositioning of apartment. Will be used in conjunction with Foursquare API to determine the amenities nearby.
property_type, room_type	To see if type of apartment and rooms will affect how people is willing to pay for the room.
bathrooms, bedrooms, beds	To see how large a magnitude the number of bathrooms, bedrooms and beds will affect the prices.
square_feet	To see how large a magnitude the size of apartment affects the price.
minimum_nights, maximum_nights, house_rules, cancellation_policy, instant_bookable	To see if restriction on bookings will affect the price
review_scores_rating	To see how big of a magnitude will the ratings affect the price of apartment
license	To see if a licensed listing will receive a higher price
neighbourhood	To see if different neighborhoods will affect prices.
price	The dependent variable for our project.

Excluded Features

Name of feature	Reason for Exclusion

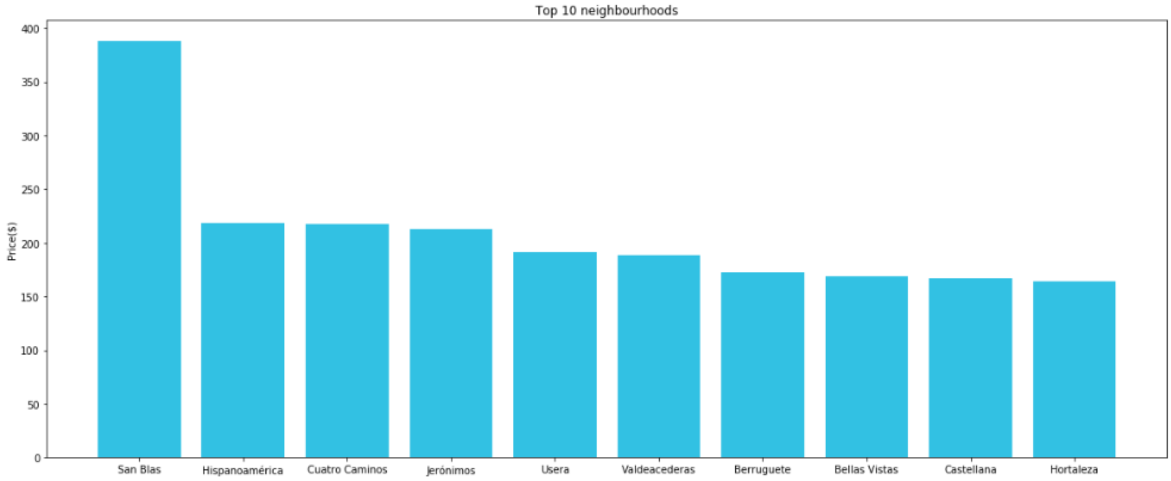
listing_url, scrape_rd, last-scraped, name, neighborhood_overview, thumbnail_url, medium_url, picture_url, xl_picture_url, host_id, host_url, host_name, host_since, host_location, host_about, host_response_time, host_response_rate, host_acceptance_rate, host_is_superhost, host_thumbnail_url, host_picture_url, host_neighbourhood, host_total_listing_count, host_verifications, host_has_profile_pic, host_identity_verified, street, neighbourhood_cleansed, neighbourhood_group_cleansed, city, state, zipcode, market, smart_location, country_code, country, is_location_exact, bed_type, security_deposit, cleaning_fee, guests_included, extra_people, minimum_minimum_nights, maximum_minimum_nights, minimum_maximum_nights, maximum_maximum_nights, minimum_nights_avg_ntm, maximum_nights_avg_ntm, calendar_updated, has_availability, hosting_listings_count, host_id, calendar_last_scraped, number_of_reviews, number_of_reviews_ltm, first_review, last_review, review_scores_accuracy, review_scores_cleanliness, review_scores_checkin, review_scores_communication, review_scores_location, review_scores_value, requires_license, jurisdiction_names, accommodates	Irrelevant to our research.
summary, space, description, experiences_offered, notes, transit, access, interaction, amenties	May contain useful information which can be extracted however difficult to analyse text data. Some of these information will be complemented with Foursquare API to determine the amenities nearby.
weekly_price, monthly_price	Similar to price.
availability_30, availability_60, availability_90	Similar to availability_365.
is_business_travel_ready	Difficult to define what is considered as business ready.

3. Exploratory Analysis

In this section we explore some of the variables affecting prices.

<u>3.1 Graphical analysis of different neighborhood on prices</u>

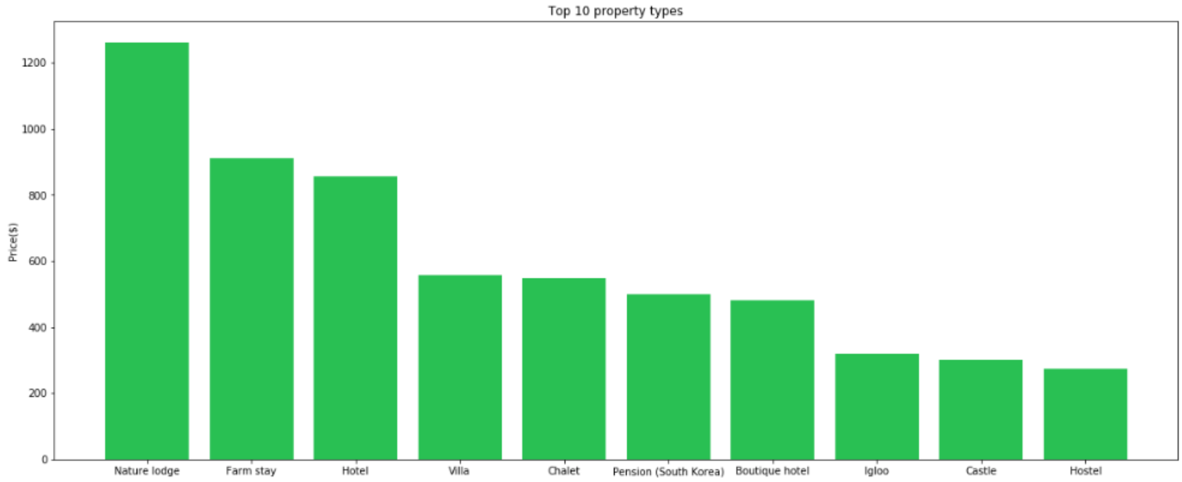
As the focus of this project is on neighborhood on apartment prices, we shall start off with this variable.



When we plot the top 10 neighborhood in terms of how expensive they are, we realise that San Blas tops the chart.

<u>3.2 Graphical analysis of property type on prices</u>

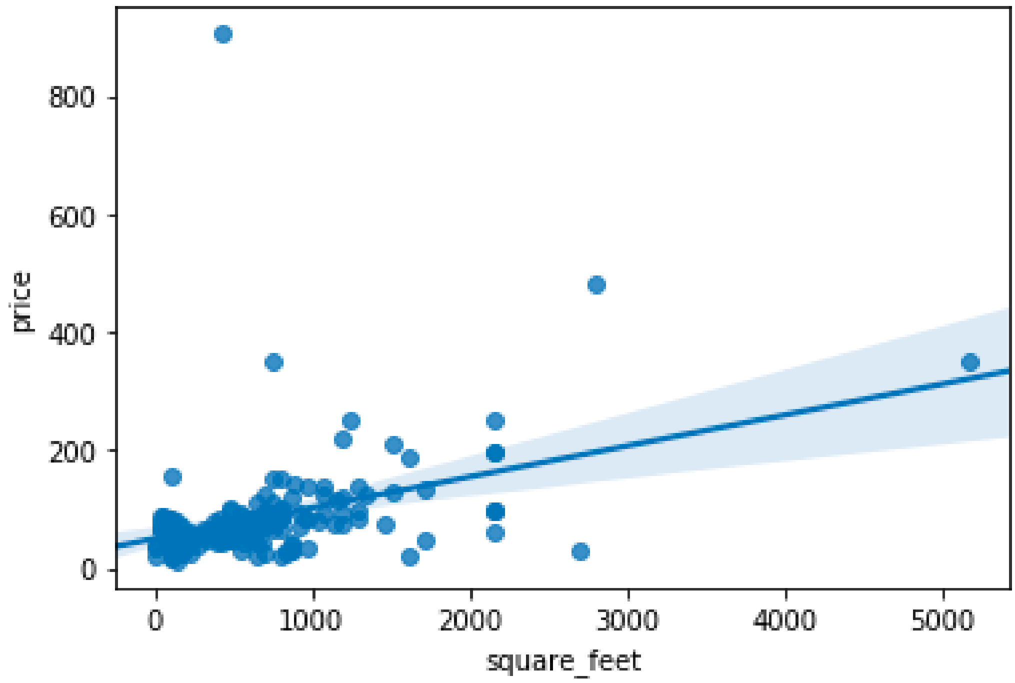
To answer the question on whether property type of will affect the property prices, we have plot the top 10 property type against their prices.



We realise that Nature lodge tops the chart, followed by Farm stay then Hotel. This may be due to their small quantity of such property type. For example, there is only 1 listing of Nature lodge amongst the 20837 entries.

3.3 Graphical analysis of how area of property affects prices

To see if the size of apartment whether affects the price greatly, we have plot the area to the prices of the apartment.



As can be seen from the graph, generally there is a positive relationship between area and prices of apartments.

4. Modelling

We will generally be using multi-linear model to explain the factors affecting Airbnb apartment prices. This is because we are predicting prices which is a continuous variable.

4.1 Multi-Linear model of all relevant factors on prices

The first multi-linear model will be using all relevant factors to predict the price of apartment. As there are numerous factors in this model, we have decided to cover only the top 10 factors which affects the price based on magnitude.

Out of the 10 factors, 7 of them represent neighborhoods. This tells us that neighborhood is a huge factor in determining the rental prices. As such I have created a multi-linear model and will further examine the effect of neighborhood on prices in the next section.

The other 3 factors pertains to the type of property, whereby hostels are expected to decrease the price by $187, Townhouse to increase the price by $98, and Bed and Breakfast to increase the price by $63.

4.2 Multi-Linear model of neighborhood on prices

As our project focuses on neighborhood as a factor and as aforementioned, neighborhood seems to be an important factor, we will be creating a model for neighborhood specifically. After modelling neighborhood to the apartment prices, we realise the neighborhood which increase the price the most is San Blas by $229. On the other hand the neighborhood decrease the price the most is Almenara by $101. We are curious to find out what is the reason which differentiates between the neighborhoods.

Taking a look at the amenities present in the neighborhoods of the top 10 neighborhoods presented by FourSquare API, we realise that the differentiating factor as to why there is this disparity in prices is due to the larger presence of restaurants in the top neighborhoods which allows them to be priced higher. More specifically, they are Spanish restaurants. This can be explained since most of rental of Airbnb are likely to be tourists and they would likely want to try out the local cuisine.

5. Conclusion

In this study, we have analysed the relationship of various factors which influences the apartment rental prices on Airbnb. We have taken a more in depth look at neighborhoods effect on the prices and have determined that the higher prices is mainly due to the presence of more restaurant, or more specifically Spanish cuisine restaurant. This research in general benefits various stakeholders who have a stake in Madrid market of Airbnb.

6. Future Directions

It would be best if this topic is looked into again with better datasets. As current dataset only contains listing prices, it does not tell us accurately whether such apartments will be taken up by tenants. It would, however, be interesting if we have information on transactions which went through, as this tell us that guests have indeed accepted such prices. It would thus be more meaningful to extract information from such dataset, then to derive insights on listings which may or may not be taken up by tenants. However, this project dataset still does provide meaningful insights on how people price their apartments.

7. References

Airbnb by the Numbers: Usage, Demographics, and Revenue Growth. (n.d.). Retrieved from muchneeded: https://muchneeded.com/airbnb-statistics/