KSCHOOL DATA SCIENCE MASTER PROJECT

GONZALO RODRÍGUEZ

13th NOVEMBER 2017

CONTENT

INTRODUCTION

Nowadays landlords of any mid size city in the western world renting their properties have to decide what type of rental is more suitable for them. Vacation rental or residential rental.

The aim of the project is to find out what kind of rental is more profitable in Madrid.

Since 2010 online vacation platforms, like Airbnb or Homeaway, have experienced an explosive growth worldwide. In less than five years, they have become a serious threat to hotels, a new alternative for travelers seeking accommodation and lastly but no less important, a new real state business.

DATA DESCRIPTION

Vacation rental market is analyzed through data of insideAirbnb website. A table containing all properties listed in Airbnb Madrid.

It is made up of 13500 rows and 95 columns. Each column is a variable describing a property of the room or apartment.

FEATURE	DESCRIPTION
Id	Numeric identification of each listing
name	Header of the listing
Summary	Brief description of the property
Space	Description of area
Description	Detailed description of the house
Neighbourhood_overview	Description of the neighbourhood
Notes	Check in conditions. Timetables
Transit	Transportation facilities near thehouse
Access	Floor, access description to the house

interaction	Message of host to guests. Contact
house_rules	Conditions rules guests must follow during stay. Noise policy
Host_id	Id code of host.
host_since	Date of first listing
host_location	Host city
host_about	Brief introduction to the host
Host_response_time	Time
host_response_rate	Percentage of questions replied

host_is_superhost	Boolean, Only hosts with very good reviews
Host_neighbourhood	Area where the host lives
Host_total_listings_count	Number of listings managed by the host
Host_has_profile_pic	Boolean
Host_identity_verified	Boolean
Street	Address of the house
Neighbourhood	Popular name location
Neighbourhood_cleansed	Official administrative neighbourhood “Barrio”
Neighbourhood_group_cleansed	Official district
Zipcode	Zipcode of listing
Smart_location	City
Latitude	Geo coordinates of listing location
Longitude	Geo coordinates of listing location
Is_location_exact	Boolean. Exact address provided

property_type	House, apartment… type of home
room_type	Category. 3 types
Accommodates	Number of people that can stay
Bathrooms	Quantity
Bedrooms	Quantity
Beds	Quantity
bed_type	Size of the bed
amenities	Devices available in the house
price	USD price per night
weekly_price	USD charge for 7 days stay
Security_deposit	USD amount payable if rented
Cleaning_fee	Extra amount charged to guests
Guests_included	Number of people included in the price
Extra_people	Usd charge for additional guest
Minimum_nights	Minimum stay. Days
availailability_30	Number of days the property is available in the next 30 days. **(key variable)**
availailability_60	Number of days the property is available in the next 60 days
availailability_90	Number of days the property is available in the next 90 days
Review_scores_rating	General score
Review_scores_cleanliness	Notation of house cleaning
Review_scores_value	Importance of previous scores
instant_bookable	Boolean. Whether host acceptance is required
cancellation_policy	Condtions to cancel reservation

Files used in the project:

- Listings.csv.gz : airbnb data
- airbnb.csv: airbnb data cleaned
- aibsuma.csv: airbnb dataset for visualization
- Idealista1.csv: idealista dataset
- Idealista2.csv: idealista dataset
- RLM.csv: dataset for linear regression and machine learning.
- center.csv: dataset for visualization.
- Barrios.shp: file for choropleth map.

METHOD

The project is divided in 5 stages:

1) Filtering and cleaning of raw data.
2) Model building for availability prediction.
3) Retrieval of Idealista data and comparison to Airbnb.
4) Visualization of results.
5) Findings

1) FILTERING AND CLEANING (Data Cleaning github folder).

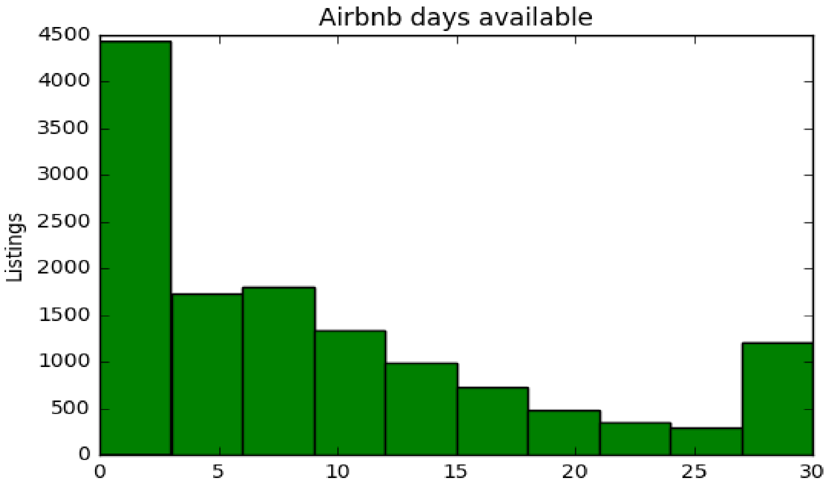
The dataset containing Airbnb listings is not from Airbnb. It is available in an independent web that posts twice a year datasets of major cities of the world.

The problem addressed in this part is the identification and removal of fake and temporary listings. Approximately 25% of listings are not real. Many listings are not available for rent, so in order to have an accurate sample of vacation rental market we need to get rid of those listings.

This is a crucial step in the project since either occupancy rate forecast or estimated gross revenue are based on one variable: availability_30.

Availability_30 shows the number of days each property is available for rent in a month or in the following 30 days. As far as this project is concerned, it is the occupancy measure.

It is essential to work with listings whose availability is real and accurate. Fake listings and inactive listings show availability_30 equal to 0 and therefore distort the accuracy of the future predictions.



As seen in the graph above, listings fully booked or with availability below 2 days are numerous. The reason for such a weird distribution relies on all those inactive listings that are disrupting the real picture.

2) *MODEL BUILDING (Data Analysis github folder).*

Given that the scope of this project is gross revenue estimation, two variables become the focus of the study:

- Price
- Availability

Price is a variable given by the market. For every neighbourhood and house size there is a price interval where demand meets offer. It is a known feature.

Availability_30 is then the key variable to forecast. Hence, the second stage of the project is an attempt to build a model through machine learning algorithms to predict occupancy (inverse availability_30) and what is more important, the variables affecting availability of a property offered in the vacation rental market.

The final objective of the predictive models implemented is to let landlords know what are the key variables that affect occupancy the most and provide guidelines to increase occupancy and therefore revenue.

3) *RETRIEVAL OF IDEALISTA DATA AND COMPARISON TO AIRBNB.*

In order to compare rents of residential and vacation rentals we first need to have a sample of the residential market of Madrid.

We turned our attention to Idealista.com, the biggest real state website in Spain. All listings for rent in Madrid were withdrawn via API to get a grasp of the city residential market.

To run the script of the Idealista API, an APIkey and secret are required:

APIKEY: hds4ce6tz7wgnkm7z3pgtsiu0ua1ppq2

Secret: WvB72IPBDAfn

Next, we merge Airbnb dataset with Idealista dataset to compare prices, rents and offer size.

Properties size comparison with unknown square meters (Airbnb does not provide this data) is an issue. To sort this out, we take the number of rooms as an approach of property size.

Finally, a summary dataset show average price, availability, premium, rents and so on. All figures a landlord needs to decide what rental suits his/her interests.

4) *VISUALIZATION OF RESULTS (Visualization folder)*

Dashboard.

A Dashboard has been developed to visualize key figures of both markets and compare estimated average rents in Airbnb and Idealista (vacation vs residential rental markets).

Choose the “barrio” where the house/apartment is located and the number of rooms it has. The chart displays the average rent a landlord receives in the residential market (Idealista) and in the vacation market (Airbnb).

A premium chart is displayed to understand how much more a landlord makes on average if his/her property is listed on Airbnb.

Finally a graph with number of listings in each neighbourhood in both platforms.

Easy to see how rental offer has pivoted from the residential market to the vacation market in lots of areas.

Map

A choropleth map of Madrid neighbourhoods is created to spot the most popular “barrios” among travelers. The colour scale is linked to the average availability, so the more intense the red is the more popular the neighbourhood is.

Note: figures of neighbourhoods outside M-30 are extracted from few listings so they are not reliable.

5) *FINDINGS*

Machine Learning (see R files in Data Analysis folder).

Madrid Rental Markets:

According to the data, and with few exceptions, vacation rental is more profitable than residential rental.

Although out of the scope of this study, the massive imbalance in the offer of rental market (see dashboard chart) may explain (along with other factors like credit conditions, monetary policy or unemployment rate) the recent rise of residential rents in Madrid. Furthermore, it confirms the disruptive effect of vacation websites in the urban landscape. Local tenants have been replaced by tourists in central and trendy neighbourhoods.

UNKNOWNS.

Price:

Airbnb price used across study is unique. However the reality is that every listing may have several prices: weekly prices, weekend prices, holiday prices and so on. The dataset did not contain all these prices so only one has been implemented. Considering that weekend prices and holiday prices (always higher) are unknown and not taken into account, rent estimations and average prices are probably lower than in reality.

Availability:

Unlike occupancy, availability is not fixed since it has not happened yet. It is just the number of bookings, and bookings may be canceled or increased during the month so it is important to be aware that all figures shown are estimations.

Risks:

There is no information about airbnb robberies, damages, fines…that may translate into costs to the host. Moreover, future regulations to limit holiday rentals are a threat difficult to evaluate.

Tax & Airbnb fee:

Fee charged by Airbnb and taxes due are unknown. That is why all income estimations are gross.

REFERENCES

An introduction to Statistical Learning. Springer Texts in Statistics.

http://www.um.es/ae/FEIR/40/#regresion-lineal-multiple

http://r-statistics.co/Linear-Regression.html

http://r-statistics.co/Top50-Ggplot2-Visualizations-MasterList-R-Code.html

https://www.datanalytics.com/libro_r/

http://kevinfw.com/tags/r/

Python for Data Analysis. O`Reilly

Python Data science Handbook. O`Reilly

http://darribas.org/gds_scipy16/

http://pythonforengineers.com/

https://plot.ly/dash/gallery

https://github.com/urschrei/Geopython