Project C4: KAGGLE-NYC-AIRBNB

Predictions on Airbnb 2019 NYC listings

CRISP-DM REPORT

Anton Slavin (Group 6)
Elen Liivapuu (Group 6)
github.com/tonysln/ids2020-project

Business Understanding

Background

Each year, thousands of flats and homes in New York are either reconstructed or specifically made for use on Airbnb. Airbnb is an online marketplace for renting flats and houses, mostly for tourists or anyone looking for an alternative to hotels and motels.

The service of Airbnb is quite beneficial to both hosts and guests. As a guest, you often have a much larger choice of different locations and price ranges compared to hotels, as well as different types of rooms and amenities. As a host, you are able to start earning income from Airbnb and potentially make it your primary business.

When it comes to renting apartments, the key aspects include its' location (neighbourhood), room type, price, reviews, and rating of the host (previous reviews). For most users, the most important aspect is the price. This is why it is very important to choose a right price based no the other parameters (aspects). A price set too high will make most guests choose the competitors' apartments, and a price too low may not bring a profit.

The main goal would be to create a model, that, given some parameters (e.g. neighbourhood, room type), would give a predicted price for the apartment. This can be used, then, to decide on the locations of new apartments and their profitability. This will help hosts avoid creating unprofitable apartments.

Business goals

- Give aid to hosts on choosing new locations and apartment types.
- Help hosts predict the price of their competitors' newly created apartments in any areas.
- Discover profitable neighbourhoods in New York for Airbnb apartments.
- Analyze the relationships between different parameters for Airbnb listings.

Business success criteria

The success of this project will be defined by the amount of correctly (close enough) predicted prices for listings based on different parameters for a part of the data. The predicted prices must be close to the prices of real listings with similar parameters. Using the predicted info must result in more success (profitability) than randomly chosen locations for new apartments.

Inventory of resources

People:
- Two computer science students

Data:
- Data from Airbnb (on Kaggle) for all listings in NYC in 2019.

Hardware:
- HP Elitebook Laptops
- Apple Macbook Pro
- Possible cloud infastructure from Google and Microsoft (if needed)

Software:
- Python 3.8
- Jupyter Notebook
- Tableau
- Microsoft Excel 2016

Other:
- Coffee

Requirements, assumptions, and constraints

Requirements:
- Access to the Internet
- Access to Kaggle (account created)
- Access to hardware and software as defined in the Inventory of resources

Constraints:
- Time limit (2-3 weeks)
- Possibly limited number of features present in the dataset

Risks and contingencies

- All of our hardware may break for some reason.
- Long-term Internet outage.
- Technical difficulties with software.
- We might encounter last-minute problems without the required time and/or skills to fix.
- We might not be able to create a successful model and will lack the skills to understand why and how to fix it.
- We might come to incorrect conclusions during the analysis and model design due to fallacies of some kind or just incorrect reasoning.

Terminology

- Model - a system of rules, a program of sorts that has been trained on data to make a prediction based on given parameters.
- Training a model - giving a model data so it can find the rules and patterns to learn.
- MSE - mean squared error, the average of the squares of the errors for values (difference between the actual and predicted values).

Costs and benefits

Not relevant to our project.

Data-mining goals

- Process the dataset (AB_NYC_2019.csv).
- Find (and present) any possible correlations and/or clusters between different parameters.
- Create a model for predicting a price of a listing based on various parameters.
- Create a model for predicting some parameters based on the given price & other parameters.
- Create a presentation for our project.
- Create a video to introduce and showcase our project.

Data-mining success criteria

- The MSE of the created model must be as close to zero as possible.
- The predicted prices must not be too different from the prices of real listing with similar parameters.
- The whole dataset must be analyzed and cleaned up, if needed.
- At least 3 visualizations must be created on the found correlations.
- The completed presentation must cover the whole work.
- The completed video must fit in 3 minutes and cover the whole work.

Data Understanding

Gathering data

Data requirements

For each listing, our project requires:

- Listing information
- Host information
- Neighbourhood description
- Precise location information
- Type of listing/room/apartment
- Price of the listing

Data availability

All required data (the dataset) is available on Kaggle, provided by the user Dgomonov. The data originates from Inside Airbnb.
A local copy of the data has been made and uploaded to Github.

Selection criteria

Required dataset: AB_NYC_2019.csv in .csv format, using the fields: *id, name, host_id, host_name, neighbourhood_group, neighbourhood, latitude, longitude, room_type, price, minimum_nights, number_of_reviews, last_review*.
The fields have been selected based on the Data requirements for the project.

Describing data

Filename: AB_NYC_2019
File format: csv
File size: 6.75 MB

The following analysis was made using Python 3.8 and Jupyter Notebook, using the pandas library and optionally matplotlib.

Rows in the dataset: 48895.
Number of columns (fields): 15 (+ index), including: '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', 'availability_365'.

All of the required fields (as described in Data requirements) are present in the dataset, along with multiple extra fields that we have decided not to use. Following is the description of each field present in the dataset:

- Name - Name of the listing, set by the host.
- Host_id - Unique ID of the host, can be used for grouping by hosts and looking at all of their listings.
- Host_name - Name of the host, may not be unique.
- Neighbourhood_group - A less precise grouped neighbourhood location for the listing.
- Neighbourhood - The neighbourhood of the listing (more on neighbourhoods in NYC).
- Latitude and Longitude - Exact coordinates of the listing.
- Room_type - Type of the listing: private room, entire apartment, shared room etc.
- Price - Price for the listing as set by the host, dollars, per night.
- Minimum_nights - The minimum amount of nights to stay in the apartment, defined by the host.
- Number_of_reviews - The amount of reviews left for this listing.
- Last_review - The date of the last review for this listing.
- Reviews_per_month - The amount of reviews left for this listing each month.
- Calculated_host_listings_count - The amount of listing each host has.
- Availability_365 - Number of days in the year when the listing is available for rent.

Exploring data

We used Python 3.8, the Pandas library and Jupyter Notebook for data exploration.

Different values for *Room type*:

Entire home/apt, Private room, Shared room.

Different values for *Neighbourhood group*:

Manhattan, Brooklyn, Queens, Bronx, Staten Island.

Values for *Neighbourhood*:

set of 221 different values, most common: Williamsburg.

Values for *Price*:

average = 152, min = 0, max = 10000

Values for *Minimum nights*:

average = 7, min = 1, max = 1250

Values for *Number of reviews*:

average = 23-24, min = 0, max = 629

While exploring the data, we found a lot of missing values for the Last_review and Reviews_per_month fields, as well as some for the name of the listing and host name. Yet, all of the other ~50000 fields are present and are mostly valid (for some exceptions that we found, described in the Verifying data quality section).

Verifying data quality

We found missing values in the following fields (columns):

- Name - 16 missing values.
- Host_name - 21 missing values.
- Last_review - 10052 missing values.
- Reviews_per_month - 10052 missing values.

We found some possibly invalid values, such as the ridiculously high price of 10000$ per night and the required 1250 minimum number of nights to stay. Also, there were multiple listings with the price set to 0$, which are also most likely invalid or placeholder values.

The rows with 1-2 missing values will be removed, as well as the columns with most missing values, such as Last_review and Reviews_per_month.

Overall, the quality of the data seems to be very good.

Plan

#	Task Name	Methods	Tools	Effort (person-hours)	Comments
1	Explore the data	Using the pandas library for exploring	Python, MS Excel	5	The first step, started while creating this report.
2	Clean up the data	Using Python and the pandas library to replace missing values	Python	5	Continuation of task #1, using Python to clean up the data and fix any issues with its’

					quality.
3	Search for correlations	Using Python pandas, matplotlib and scipy libraries	Python, MS Excel	10	Search for any correlations and clusters between parameters.
4	Create visualizations	Using matplotlib, seaborn Python libraries	Python	8	Creating the visualizations that will be used in the presentation.
5	Build model	Using the scipy & pandas Python libraries, different combinations of their functions	Python	12	The main step of building a model for making predictions.
6	Test & improve model	Using scipy & pandas Python libraries, previous knowledge	Python	10	Continuation of step #5 (possibly repeated multiple times), fixing the model and searching for ways to improve it.
7	Create documentation & presentation & video	Reviewing all results, combining them for presentation & overview	Google Docs, Slides, possibly Latex	10	The final step, creating the presentation, video and any documentation for code and data.