Elis Vingisar, Group 8

Rahel Pettai, Group 8

KAGGLE – Airbnb prices in European cities

Project number: H8

Setting up

Our repository is hosted in GitHub and can be found at this link: https://github.com/ElisVingisar/IDS-project

Business understanding

Identifying the business goals

Background:

In the hospitality industry, Airbnb has become a prominent player, offering a diverse range of accommodations across European cities. For the hosts or customers themselves, it's crucial to understand the factors that influence Airbnb prices. The 20 datasets that we have been provided with, representing 10 European cities with weekday and weekend pricing details, provide a great source of information for comprehensive analysis.

Business goals:

1. Price analysis across cities:
   Our first goal is to understand and analyze how the prices between the 10 given European cities vary**. We aim to identify the overall prices in the 10 cities under observation, which could be useful for future travelers in deciding where to go depending on their budget.**
2. Weekday vs. weekend price disparities:
   The second goal is to examine the variations in Airbnb prices between weekdays and weekends for each of the cities. **This will help us identify the city with the biggest price difference, depending on whether it is a weekday or weekend, of the 10 European cities.**
3. Price prediction model:
   The final goal of the project is to predict one of the cities accommodation's price based on it's attributes. **A reliable pricing prediction can benefit hosts by setting competitive rates and even assisting travelers in budgeting for their trips.**

Business success criteria

The project is considered successful when

- we have identified the cities with highest and lowest housing prices,
- we have gained insight on which of the cities housing is most affected by the time when making a housing reservation through Airbnb.

Assessing our situation

Inventory of Resources

1. Data:

amsterdam_weekdays.csv

amsterdam_weekends.csv

athens_weekdays.csv

athens_weekends.csv

barcelona_weekdays.csv

barcelona_weekends.csv

berlin_weekdays.csv

berlin_weekends.csv

budapest_weekdays.csv

budapest_weekends.csv

lisbon_weekdays.csv

lisbon_weekends.csv

london_weekdays.csv

london_weekends.csv

paris_weekdays.csv

paris_weekends.csv

rome_weekdays.csv

rome_weekends.csv

vienna_weekdays.csv

vienna_weekends.csv

2. Software: We are going to use Python, Jupyter Notebook and different Python libraries, for example Pandas and NumPy, and Google.

3. Hardware: We are going to use our HP laptops from the Institute of Computer Science.

4. Other: Our main contact in case of any questions or problems would be our lab supervisor Carel Kuusk.

Requirements, assumptions and constraints

1. Requirements: Most importantly we need access to Airbnb pricing data and computational resources for model development. Planning related requirement is to submit the project before the deadline – 11th December 2023.
2. Assumptions: The dataset is representative of Airbnb listings in the specified 10 cities. Prices are influenced by a combination of factors as outlined in the dataset.
3. Constraints: Limited to no ability to factor in external events affecting prices.

Risks and Contingencies:

1. Risk: Incomplete or inaccurate data in the dataset and model overfitting.
2. Contingency: Implementing rigorous data cleaning processes.

Terminology

Superhost – In Airbnb, a host is someone who is providing the housing. Superhosts on the other hand tend to have a higher occupancy rate (and more potential earnings) on average because they've met the Airbnb criteria of becoming a Superhost.

Costs and benefits

1. Cost: Time invested into completing the project
2. Benefit: Applying oneself, wider knowledge, grade based on the project outcome. Results of the project can be useful for future travelers choosing housing through Airbnb or for the hosts to estimate the total price per night of the housing.

Other than that, we do not need to assess any more costs and benefits.

Defining data-mining goals

Data-mining goals

Our goals are the following:

- to develop a predictive model to estimate accommodation prices (mainly focused on the location),
- to deliver pricing differences on various graphs and therefore find out more patterns.

Data-mining success criteria

The data mining is considered successful when

- we can make accurate predictions based on the accommodations factors,
- we can report on the cities overall accommodation's pricings.

Data understanding

Gathering Data

Data requirements outline

The data required for our analysis primarily revolves around the specifics of Airbnb accommodations. This includes the total cost of the listing, the type of room being offered, whether the room is private or shared, and the maximum number of people that can stay in the room. In addition, we need geographical data such as the distance from the city centre and the nearest metro station, as well as the exact location of the listing. We also require information about the host, specifically whether they are a superhost, and customer satisfaction ratings. Lastly, we need to know if the listing is for multiple rooms or for business purposes.

Data availability verification

The data we have acquired is from Kaggle* and has been verified.

* https://www.kaggle.com/datasets/thedevastator/airbnb-prices-in-european-cities/data

Selection criteria

The data we are using is in CSV file format. We have 20 CSV files which contain information about 10 cities accommodations pricings. For each city there is two datasets – weekday prices and weekend prices. For example, 1 of the 10 European cities is Amsterdam. The according datasets are:

amsterdam_weekdays.csv

amsterdam_weekends.csv

All of the datasets contain the same information:

Column name	Description
realSum	The total price of the Airbnb listing.
room_type	The type of room being offered (e.g. private, shared, etc.).
room_shared	Whether the room is shared or not.

room_private	Whether the room is private or not.
person_capacity	The maximum number of people that can stay in the room.
host_is_superhost	Whether the host is a superhost or not.
multi	Whether the listing is for multiple rooms or not.
biz	Whether the listing is for business purposes or not.
cleanliness_rating	The cleanliness rating of the listing.
guest_satisfaction_overall	The overall guest satisfaction rating of the listing.
bedrooms	The number of bedrooms in the listing.
dist	The distance from the city centre.
metro_dist	The distance from the nearest metro station.
lng	The longitude of the listing.
lat	The latitude of the listing.

Describing data

The data is from Kaggle and it seems to have been gathered since the beginning of 2023 (no exact date range stated in Kaggle).

Column name	Format
realSum	Numeric
room_type	Categorical – Entrie home/apt, Private room, Other
room_shared	Boolean – true, false
room_private	Boolean – true, false

person_capacity	Numeric
host_is_superhost	Boolean – true, false
multi	Numeric – 1, 0
biz	Numeric – 1, 0
cleanliness_rating	Numeric (max 10.0)
guest_satisfaction_overall	Numeric (max 100.0)
bedrooms	Numeric
dist	Numeric
metro_dist	Numeric
lng	Numeric
lat	Numeric

Exploring data

All of the datasets have been looked into and are clean, meaning that we don't have a need for any data cleaning processes. We can narrow down the selection of features that we are using, if there is need for it – for example there are columns attr_index, attr_index_norm, rest_index and rest_index_norm which meaning we don't know as they are neither intuitive nor explained in Kaggle.

As mentioned above, the data is clean. The housings represented in data are rather well rated – over half of the ratings are 70.0 and higher. The room person capacity is from 2 to maximum 6 people.

Data quality verification

The data we have aqcuired should be suitable for our project and we haven't found any data errors yet.

Project plan

Task	Tools	Elis	Rahel
Homework 10	Google Docs	1h	4h
Research about previous work done in that area and exploring more ideas for presenting the outcome	Google, Kaggle, Microsoft Excel	3h	2h
Decide on final approaches	Google, Microsoft Excel	2h	0h
Data exploration or cleaning	Jupyter Notebook, Microsoft Excel	4h	3h
Data visualisation	Jupyter Notebook	6h	6h
Data training for predictions	Jupyter Notebook	8h or more	8h or more
Results analysis	Microsoft Excel, Jupyter Notebook	4h or more	4h or more
Poster	Canva, Google'i joonised	3h or more	3h or more
Present the project	Poster	2h or more	2h or more