Carolina Martín

Task 6.1 Data Immersion

Data Source

The information is external and gathered by Inside Airbnb. 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 to facilitate public discussion. This data is licensed under a Creative Commons Attribution 4.0 International License. Therefore the information is **trustworthy**.

Data Interest: The analysis of Airbnb listings and reviews is current to today's real accommodation matters in the city I live in. Additionally, the data sets contain qualitative data (from teh reviews), as well as geospatial (from the districts listings locations) on top of the quantitative data which I find interesting to manage.

Data Collection: information compiled from the Airbnb web-site including the availability calendar for 365 days in the future, and the reviews for each listing. The collection is done automatically (Web Scraping)

Data content: The Dataset gathers information from the listings (+26K listings) and reviews (+1M reviews) from Airbnb accommodations in Madrid (Spain) for the past 12 months. (Explore)

File Name	**Description**
listings.csv.gz	Detailed Listings data
calendar.csv.gz	Detailed Calendar Data
reviews.csv.gz	Detailed Review Data
listings.csv	Summary information and metrics for listings in Madrid (good for visualizations).
reviews.csv	Summary Review data and Listing ID (to facilitate time based analytics and visualisations linked to a listing).
neighbourhoods.csv	Neighbourhood list for geo filter. Sourced from city or open source GIS files.
neighbourhoods.geojson	Neighbourhood list for geo filter. Sourced from city or open source GIS files.

Data Relevancy: All the data sets are relevant for the analysis on how the proliferation of Airbnb accommodations are shifting the neighborhoods in Madrid. Madrid Airbnbs are mostly controlled by real estate agents, specialized companies, large and small owners who are dedicated to extracting housing from the residential rental market to introduce it on Airbnb.The **Central district of Madrid** is a particular case that requires special attention, as its number of apartments per km2 it is 10 times higher than that of any other district in Madrid.

Additionally, traditional residential districts such as Arganzuela, Chamartín, Chamberí, Moncloa-Aravaca, Retiro, Salamanca and Tetuán excessive tourist overcrowding is forcing locals as prices increase. .

Data Assumptions:
Data Assumptions | Inside Airbnb

Data Dictionary:
Inside Airbnb Data Dictionary - Google Sheets

Data Cleaning

Conduct some basic data cleaning and consistency checks in Jupyter to ensure your data is ready for further analysis.

Listing.csv:

- Data cleaning completed via Jupyter Notebook.
- The df_listing_st shows 5304 price values edited as NAN and 24,182 listings with no License.
- A tourist license is required in Spain allowing you to rent out the property as a tourist accommodation. However Airbnb is a platform that allows hosts and guests to connect. It seems not to be a requirement form the platform to request the license number as 89.8% of the listings don´t count with a registered license.*

- **FIX** on the price I will calculate the average price and edit the missing figures. Otherwise the EDA will present outliers on 0 values, which is not realistic.

- **FIX** on the license number they will remain as they are. It is already an interesting insight as it is.

- **FIX** df_listing.st had 'id' as 'listing_id'. Changing column name to match with the rest of the dataset columns

Neighbourhoods.csv.

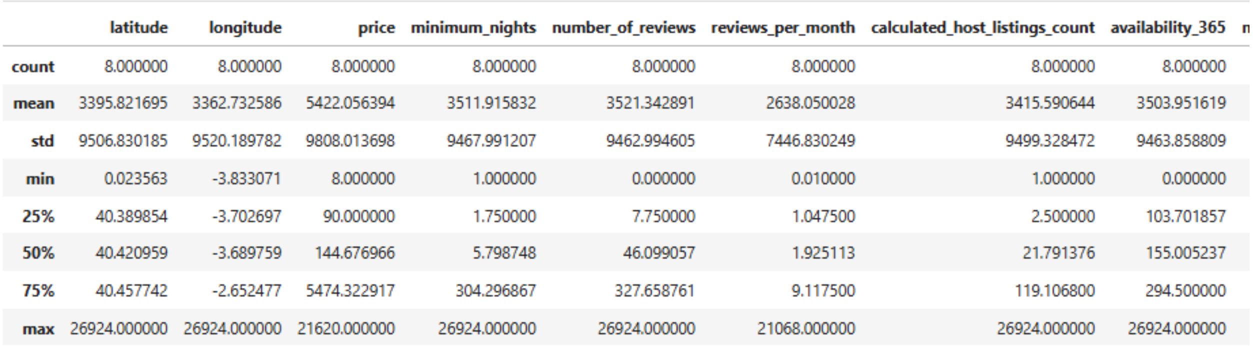
- Data cleaning completed directly on the csv
- It contained a total of 21 districts and 129 neighborhoods
- Both districts and neighborhoods have typos with accents presented and special characters such as 'ñ'(ex: ChamartÃn,ChamberÃ, TetuÃ¡n) . **Solution**:correcting the typos.

reviews.csv.gz

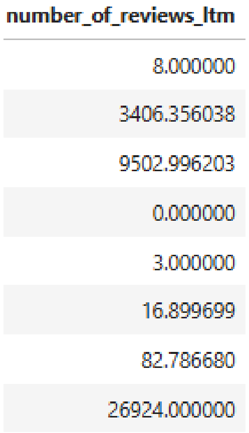
- Data cleaning completed via Jupyter Notebook.
- There were 93 reviews missing comments and no duplicates. No further action taken

Understand your data

Develop a basic understanding of your data set by reviewing the variables and performing basic descriptive statistical analysis

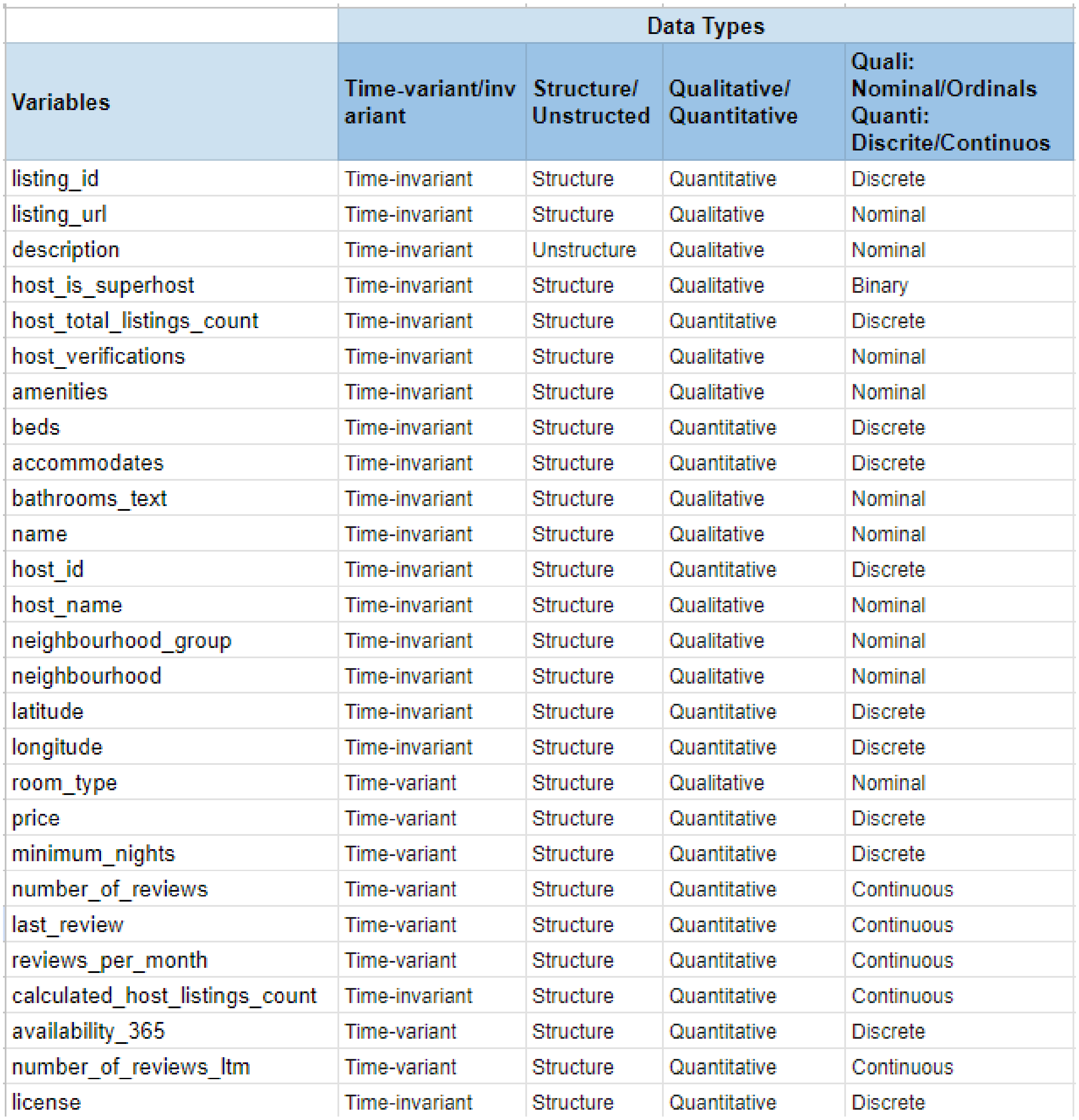


	latitude	longitude	price	minimum_nights	number_of_reviews	reviews_per_month	calculated_host_listings_count	availability_365	n
count	8.000000	8.000000	8.000000	8.000000	8.000000	8.000000	8.000000	8.000000	
mean	3395.821695	3362.732586	5422.056394	3511.915832	3521.342891	2638.050028	3415.590644	3503.951619	
std	9506.830185	9520.189782	9808.013698	9467.991207	9462.994605	7446.830249	9499.328472	9463.858809	
min	0.023563	-3.833071	8.000000	1.000000	0.000000	0.010000	1.000000	0.000000	
25%	40.389854	-3.702697	90.000000	1.750000	7.750000	1.047500	2.500000	103.701857	
50%	40.420959	-3.689759	144.676966	5.798748	46.099057	1.925113	21.791376	155.005237	
75%	40.457742	-2.652477	5474.322917	304.296867	327.658761	9.117500	119.106800	294.500000	
max	26924.000000	26924.000000	21620.000000	26924.000000	26924.000000	21068.000000	26924.000000	26924.000000	



number_of_reviews_ltm
8.000000
3406.356038
9502.996203
0.000000
3.000000
16.899699
82.786680
26924.000000

- 50% of the prices are around 144 EUR a night.
- Lat and longitude shows that Percentiles shows that most accommodations are located very close to one another.
- Median minimum number of booked nights are 5. with a 75% percentile of 304 days (long term bookings)
- Hosts have a median of 21 listings. Meaning most booking are done through agencies rather than individuals owners.



	Data Types			
Variables	Time-variant/invariant	Structure/ Unstructed	Qualitative/ Quantitative	Quali: Nominal/Ordinals Quanti: Discrete/Continuos
listing_id	Time-invariant	Structure	Quantitative	Discrete
listing_url	Time-invariant	Structure	Qualitative	Nominal
description	Time-invariant	Unstructure	Qualitative	Nominal
host_is_superhost	Time-invariant	Structure	Qualitative	Binary
host_total_listings_count	Time-invariant	Structure	Quantitative	Discrete
host_verifications	Time-invariant	Structure	Qualitative	Nominal
amenities	Time-invariant	Structure	Qualitative	Nominal
beds	Time-invariant	Structure	Quantitative	Discrete
accommodates	Time-invariant	Structure	Quantitative	Discrete
bathrooms_text	Time-invariant	Structure	Qualitative	Nominal
name	Time-invariant	Structure	Qualitative	Nominal
host_id	Time-invariant	Structure	Quantitative	Discrete
host_name	Time-invariant	Structure	Qualitative	Nominal
neighbourhood_group	Time-invariant	Structure	Qualitative	Nominal
neighbourhood	Time-invariant	Structure	Qualitative	Nominal
latitude	Time-invariant	Structure	Quantitative	Discrete
longitude	Time-invariant	Structure	Quantitative	Discrete
room_type	Time-variant	Structure	Qualitative	Nominal
price	Time-variant	Structure	Quantitative	Discrete
minimum_nights	Time-variant	Structure	Quantitative	Discrete
number_of_reviews	Time-variant	Structure	Quantitative	Continuous
last_review	Time-variant	Structure	Quantitative	Continuous
reviews_per_month	Time-variant	Structure	Quantitative	Continuous
calculated_host_listings_count	Time-invariant	Structure	Quantitative	Continuous
availability_365	Time-variant	Structure	Quantitative	Discrete
number_of_reviews_ltm	Time-variant	Structure	Quantitative	Continuous
license	Time-invariant	Structure	Quantitative	Discrete

Limitations & Ethics.

Outline any limitations and ethical considerations presented by the content of your data, its source, and/or how it was collected

Data Limitations: As the data set belongs to the past 12 months a review to uncover historic trends might be difficult to analyze. However it is updated every quarter, therefore the data is rather current reducing any bias.

Data privacy: No "private" information is being used. Names, photographs, listings and review details are all publicly displayed on the Airbnb site. Location information for listings are anonymized by Airbnb.

Define questions to explore

In a third section of your project document, define a list of questions to explore with your analysis

Madrid, as many other cities in the world, is experiencing a shift in their traditional districts upon the proliferation of short term accommodation facilities such as Airbnb, that provides solutions for tourism mostly.This is based on the high demand of this type of services that is not matching the reality of the offer. In consequence the number of peer to peer (P2P) solutions keeps growing. Making local residents leave the city or the city center to the outskirts while local shops/markets suffer upon the increase in prices.

The analysis focuses on Airbnb activity in Madrid as of today:

- Most booked type of property?
- Min and Max stay?
- What price does it have?
- Are there any amenities most in demand? Least?
- Do bookings present any seasonality? Is it reflected on the type of properties? Number of accommodates? Amenities requested?

To understand what makes a host or listing successful.

While we have a look into the districts activity :

- What is the distribution/density of listings per district? What are the most packed and least?
- Are there any pricing differences between them?
- What are the most expensive listings and where are they located?
- How many listings does a host have? In order to know if the booking is done to an individual or a company.

Is this growth pace sustainables in time?
What measure can be introduced to allow a more balanced growth within the touristic sector?

Bibliography:

Cerdá-Mansilla, E. (2022, junio 17). Airbnb y la turistificación en Madrid. Universidad Autónoma de Madrid. Retrieved from https://www.uam.es/uam/investigacion/cultura-cientifica/articulos/airbnb-turistificacion-madrid

Cerdá-Mansilla, E., Rubio, N., García-Henche, B., & Campo, S. (2022). Airbnb y la turistificación de los barrios en las ciudades: Un análisis de segmentación por barrios del alojamiento extrahotelero en Madrid. Revista Investigaciones Turísticas, (23), 210-238. Retrieved from https://rua.ua.es/dspace/bitstream/10045/121251/6/Investigaciones-Turisticas_23_10.pdf

My Lawyer in Spain. (2024). Explained: Spain's Touristic Licence. My Lawyer in Spain. Retrieved from https://www.mylawyerinspain.com/blog/explained-spains-touristic-licence/

Airbnb. (2024). What legal and regulatory issues should I consider before hosting on Airbnb? Airbnb. Retrieved from https://www.airbnb.es/help/article/961