**Airbnb Europe Exercise**

Link to data:

<https://docs.google.com/spreadsheets/d/1ecopK6oyyb4d_7-QLrCr8YlgFrCetHU7-VQfnYej7JY/edit?usp=drive_link>

Link to research paper:

<https://www.sciencedirect.com/science/article/pii/S0261517721000388>

**Instructions**

Create an Exploratory Data Analysis (EDA) Report in markdown and submit it with the respective notebook to GitHub. The activity is individual, but you can work in groups. Submit by next Tuesday.

***→ Practice your data wrangling***

-Import the data directly from google drive instead of saving it locally.

-Merge all the tables into a single dataframe

-Create a column for city, country, and weekday/weekend

-Change the name of realSum to Price

-Remove first column

***For advanced:***

-Instead of having two columns of room\_shared and room\_private, create one with the respective categories.

-Instead of dummy variables, create a column where one can know if the host of the listing has only one, two, four, and more than four listings.

→ ***Exploratory questions (suggestions for beginners)***

In addition to the initial inspection of your data using pandas functions (for examples, df.info(), df.describe, etc.), answer the following questions

1. How many listings are in each city in total and also per type of day?

2. Which city has the biggest proportion of superhosts?

3. Which cities have listings with more than four rooms?

4. Which city has the most entire home/apt type listings?

5. Are ratings typically high across listings, or is there a wide variation?

6. How does person\_capacity vary across listings? What is the most common capacity of listings?

7. Plot the distribution of realSum for both weekday and weekend offers. Is it normally distributed, skewed, or multimodal? If skewed, consider using transformations (e.g., log transformation) to normalize it.

8. Plot the distribution of guest\_satisfaction\_overall ratings.

9. Examine the distribution of cleanliness\_rating.

*For advanced students, come up with your questions, hypotheses, answers, and tests. Give us a surprise*

**→ Suggested questions for testing (for beginners)**

1. Is there a difference in price between two cities? Choose at least three pairs of cities during weekdays.
2. Does the price tend to be higher on weekends?
3. Are listings of superhosts more expensive than those of normal hosts?
4. Are superhosts closer to the city center and metro station than normal hosts?
5. Are superhosts cleaner than normal hosts?
6. Is there a pattern between room types and superhost status?
7. Is it more expensive to rent an entire home/apt than a private room? Does that depend on the city?  
     
   **For advanced**
8. Create a map for each city where the top 10 most expensive Airbnbs are shown and the top 10 cheapest Airbnbs are shown.

**For the testing questions, create your hypothesis and then test them. Create the respective plots and apply the respective test. Mind the assumptions behind the application of simple tests, such as normality, sample size, etc.**

1. Airbnb plays a big role in the phenomenon known as gentrification. Based on the data, what can you tell about this phenomenon, could you see a pattern? Is it different in each city? As a tip, consider 1) how much would it cost to rent each listing for a month, as if paying the Airbnb for a long-term stay (subselect only those listings whose price is below 300), and 2) how many listings belong to owners with more than property and that are closer to the city center.

2. Which information is missing in order to conduct a better analysis and to understand the gentrification phenomenon?