PYTHON GROUP ASSIGNMENT – AIRBNB

2019-2020

# Practicalities

## Form groups

* Form groups of 5 people. Subscribe as a team to a group on Minerva.

## Presentation

* At the end of the assignment each group will present their findings within 20 min.
  + 15 min presentation
  + 5 min Q&A
* Create a **professional** presentation
  + Presentation should be directed at a non-technical audience, however, questions might be technical as well.
  + Present as if you would present to real investors who are interested in starting an Airbnb business. Try to interpret your findings and communicate a message that is relevant to them.
  + Make it visually appealing
* There will be a strong time pressure:
  + Knowledge to do the analyses is learned during classes
  + Limited presentation time, so it is necessary to select the **most relevant results**, without losing information
    - E.g. Often different plots can be combined into a single plot, saving space and presentation time and making comparison easier
    - Having backup slides with additional information might be a good idea…
* Do not limit yourselves to the obvious outcomes. If you know additional analyses that might be relevant and that are not mentioned in the goals below, feel free to do it. Moreover, go beyond simply reporting results, interpret them as well and search for links.
* Try to come with a clear message during your presentation.

# Assignment

## Data preparation

1. Upload all provided datasets.

## Analysis

1. Calculate the average listing price per neighborhood.
2. Plot how the average price evolves through the year across New York.
3. Identify which neighborhood has the largest price fluctuations across the year. Plot the fluctuations for this neighborhood.
4. In marketing, there is a phenomenon known as ‘the long tail’ (Hint: look it up). This also translates to the number of reviews. Plot this on an intuitive graph.
5. Run a regression to explain the price per listing. (Hint: location, reviews, etc. may all explain this).
6. Find additional data sources to explain the average listing price per neighborhood. (Hint: think demographics).
7. Plot how the average prices differ across New York using a color-coded heat map of New York neighborhoods.
8. The latitude of Statue of Liberty National Monument, New York, USA is 40.689247, and the longitude is -74.044502. This monument is one the most popular tourist places in New York. Statistically test whether a distance smaller than 2 miles to the monument increases average listing price.
9. Create a timeline and plot for each year the highest, Q1, the median, Q3 and lowest price on one graph. Do this for each neighborhood group as well as for the entire city. Determine which neighborhood group stands out the most and create a comparative graph of this neighborhood with all other groups.
10. Plot the number of rooms per host in function of the number of reviews per host.
11. Are there a lot of hosts having multiple locations? Do most people just rent their own place? Is there a ‘host long tail’? Make a comprehensive plot.
12. Do hosts with multiple locations stay within the same neighborhood? (Hint: use subset).
13. What are the 5 most used words in reviews that are no stop words? (e.g. the, or, etc. Python can filter these automatically using packages such as NLTK).
14. Do these most frequent words differ across neighborhoods? What are the ‘most different’ areas? What distinguishes them? Interpret.
15. Plot the amount of reviews across time.
16. Is there a link between availability (days per year) with the price? Determine both graphically and statistically.
17. Is there a link between how many times the word ‘great’ appears in a review and the listing price? Determine both graphically and statistically.
18. Plot how the number of Airbnb locations are distributed across the city on a map. Plot the number of locations per neighborhood and color code according to neighborhood group.
19. Williamsburg is a ‘hip’ area in in Brooklyn with a lot of Airbnb locations on offer. Explore how this area differs from other locations and visualize. You may also use external data sources.
20. Create a stacked bar chart of the distribution of room type per neighborhood group. Statistically test whether these differences are significant.
21. Color-coded plot the most popular room type per neighborhood on a city map.

Overall, your presentation should make a comprehensive summary of the requested items. If you feel that one of the points requested above does not fit in your presentation, you can leave them out (or add it as extra slides). However, **you need to write code for all of the questions above**. There are a lot of extra aspects that can be looked at. You can include these in your presentation or coding if you think they may be of particular relevance or contain some unexpected results.

**NOTE**: We will also ask you to hand in your Python code, please structure it in a readable format using comments and numbering as below for each exercise. Any additional code you write, should be indicated as “extra”.

# Hints and guidelines

* New functions that have not been seen in class might be necessary, so use the Internet wisely.
  + Google
  + Stack Overflow
  + Quora
  + Etc.
* Handle missing values (delete them or find a way to integrate them in your analyses).
* Use plots and make your slides visually appealing.
* Other information that you can find on the Internet can be useful to provide more detailed information.