AIRBNB LISTING DATA TASK

The Data

The data is available on the web, at http://tomslee.net/airbnb-data-collection-get-the data

This website contains scraped data of AirBnb listings from different cities/regions all over the world and in different times. A listing is a property listed on AirBnb and it is identified by a unique ID.

There is data granular by city and referring to different days when the website has been scraped. Information on each listing includes the price, the review score received by guests, the number of bedrooms. Please read the specifics of each attribute and note that different files may not always contain all attributes.

Which data to consider -

- **Consider data from Europe alone, to avoid adding in other layers of data complexity**
- **Consider cities only**
- **Choose capitals and / or important cities only**

mallzee

The Task

We'd want to analyse this data to spot insights. It is real-world data and chances are, it won't contain (unlike other). The whole task is quite open-ended so we don't already have a list of specific questions we'd like to answer.

On a general note, we'd be interested in seeing what the data says about the price and quality of AirBnb offerings in European cities and whether there's been changes in time.
Note - that the service has been, or will soon be, banned in some places.

The task we propose is divided into two sections to be tackled in succession: the first is strictly focussed on a statistical analysis of the data, aimed at extracting information from the data we have and the subsequent one which is about applying some Machine Learning to the predict where would some new data sit.

We're interested in seeing how you get along with the proposed problem, how you tackle it and reach conclusions or learn from approaches which prove not to work.

Please feel free to send us all the work you've done, included what may not have worked, as it informs us of your reasoning and problem-solving attitude. It is not important to solve the problem, what is important are the strategies implemented.

The Analysis Part

This part is a classical data analysis task. We want to understand what information is there in the whole mole of data we have. The data is already quite clean per se, so we can directly switch on doing some statistics. Examples of questions (these are just examples) we might want to try to answer are:

- **Did each city see an upward / downward trend in prices?**
- **Did the same properties show interesting trends?**
- **Did the number of listings change in time per city?**

We're interested in how you'd tackle problems similar to these ones, how you get creative to ask yourself new or different ones and what you achieve. An important components of this is evaluating how good the data is to reach these insights and how robust the results appear to be.

The Prediction Part

This part is conceptually meant to follow the previous one.

The general idea is to see whether we can build a predictive model of the price of a listing based on where it is, which period we are interested in and what type of accommodation it is.

The preliminary part would be a feature engineering one, where features are built from the data you are given here, but potentially also on data you can add using other sources.

If a model is finally achieved, we are also interested in evaluating its performance to assess how good it is supposed to be on new data.

How To Submit Your Work

The ideal way to submit us your work is sending us a notebook we can run.

We will then provide feedback in any case.