

**HackWagon Academy Project:**

**AirBnB Data Analysis Project**

1. **Brief Description**

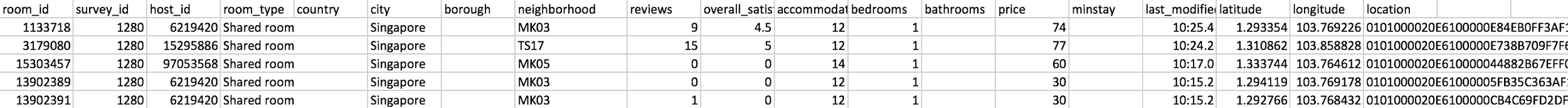
Airbnb is an online marketplace and hospitality service, enabling people to lease or rent short-term lodging including vacation rentals, apartment rentals, homestays, hostel beds, or hotel rooms. The company does not own any lodging; it is merely a broker and receives percentage service fees (commissions) from both guests and hosts for every booking made on their site. In this project, we aim to use algorithms and libraries to mine the reviews people have submitted on Singapore AirBnB rentals in order to provide descriptive analytics.

This course project which will **span over 2 weeks** aims to achieve the following learning outcomes:

1. Students should be able to translate business requirements into detailed technical analysis documents
2. Students should be able to write their own methods to retrieve data from a CSV file and store the data in a suitable python data structure.
3. Students should be able to make use of standard Python libraries, as well as language features to perform descriptive analytics on a limited dataset.

1. **Programming Portion of the Project**

You’re provided with a CSV file of actual rental data from AirBnB, and you are supposed to use the data to perform some simple analysis. A sample of the data looks like the following:



**Function 1: CSV File Reader**

Before you can even begin to perform any data analysis, the first thing you have to do is to read the data from CSV and store the data in a suitable python data structure. Make use of the methods learnt from the previous class to retrieve the data from the CSV file, and thereafter, store all the data in a **list of dictionary**. Basically, each data row should be stored as a dictionary, and look like the following:

{'room\_id': '1133718', 'survey\_id': '1280', 'host\_id': '6219420', 'room\_type': 'Shared room', 'city': 'Singapore', 'neighborhood': 'MK03', 'reviews': '9', 'overall\_satisfaction': '4.5', 'accommodates': '12', 'bedrooms': '1.0', 'price': '74.0', 'minstay': '', 'last\_modified': '2017-05-17 09:10:25.431659', 'latitude': '1.293354', 'longitude': '103.769226', 'location': '0101000020E6100000E84EB0FF3AF159409C69C2F693B1F43F'}

And each row of data that has been stored as a dictionary should be appended to a list.

**Function 2: Descriptive Analytics Report Generation**

The data team at AirBnB wishes to find out the answers to a few simple questions on the existing rooms in Singapore. Your goal is manipulate the data you have stored in the list of dictionary to generate a report for the following questions:

1. What are the top 10 priciest rooms in Singapore? List the **room\_id** as well as **price**
2. Which hosts in Singapore have got more than 30 rooms? How many of such hosts do we have? (Note: Each host can have more than one property listed, which in turn has varying number of rooms)
3. What are the different room types, and how many of each room types are there in Singapore?
4. What are the top 10 most reviewed rooms?
5. What are the top 10 rooms with the highest overall satisfaction?

**Function 3: Satisfaction Map**

The data team at AirBnB wishes to find out which region of Singapore has rooms that people are generally more satisfied with. Any room which has a rating of >4 would indicate a highly satisfactory room, and >2.5 would indicate an average room. Any rating <2.5 indicates an unsatisfactory room. The AirBnB team hypothesised that the North-East region of Singapore has more rooms with higher satisfaction rating. As the data team, your goal is to find out whether the hypothesis stands true by providing descriptive analytics to allow the AirBnB team to better visualise the satisfaction levels for the whole of Singapore.

Given the dataset provided to us, each room has a geographical location specified by its latitude and longitude. Latitudes and longitudes can be considered to be merely a set of X and Y coordinates on an axis, where the format follows such a manner e.g. (1.293354, 103.76922). What this means for us is that, geographical locations on a map can be represented easily on a scatter plot. This allows us to make use of our pre-existing knowledge on the construction of scatter plots to build a simple heatmap to visualise the satisfaction levels of different rooms in Singapore. An example of such a simple heatmap is shown below.

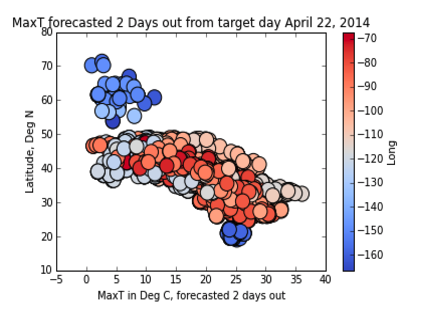


Figure 1: Example of a heatmap for weather forecasting

**Consider doing the following:**

1. Choose proper colors for your heatmap that allows you to easily distinct between the 3 levels of satisfaction
2. Should there be a need to consult existing heatmap-through-scatterplot examples, consider the following resource <http://geodesygina.com/matplotlib.html>

**Function 4: Basic Machine Learning Models**

The data science team at AirBNB wishes to make use of some simple machine learning models to: (1) learn more about their customers, and (2) predict the choice of customers. Using the data provided, do the following:

1. Using any method you prefer, find out which variable provides the greatest explanation for the overall satisfaction of customers.
2. Using your own judgment, make use of an appropriate machine learning model to predict the satisfaction level of the 30 customers in the test.csv file