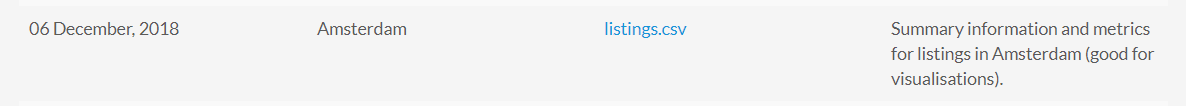
1. Selecting data set:

<http://insideairbnb.com/get-the-data.html>

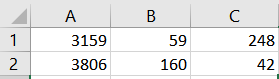




This dataset is provided by Airbnb company, it includes Airbnb id, Airbnb name, host\_id, host\_name, neighbourhood, latitude, longitude, room\_type, price, minimum\_nights, number\_of\_reviews, last\_review, reviews\_per\_month, calculated\_host\_listings\_count, availability\_365.

2. Data Cleasing:

Among all these factors, I select a dataset only includes hostid, price, numberofreviews. The first row should leave it blank.



3. Define Problem and Provide Solution

- Business Problem:

To determine the Total Number of Reviews received by each host, and then to sort Host ID descending by the Total Number of Reviews received by each host.

- Solution:

To use Python to create a MapReduce program (chaining two MapReduce jobs together).

- Findings: The total sum of reviews received by each host ranges from "00000" to "01772".

