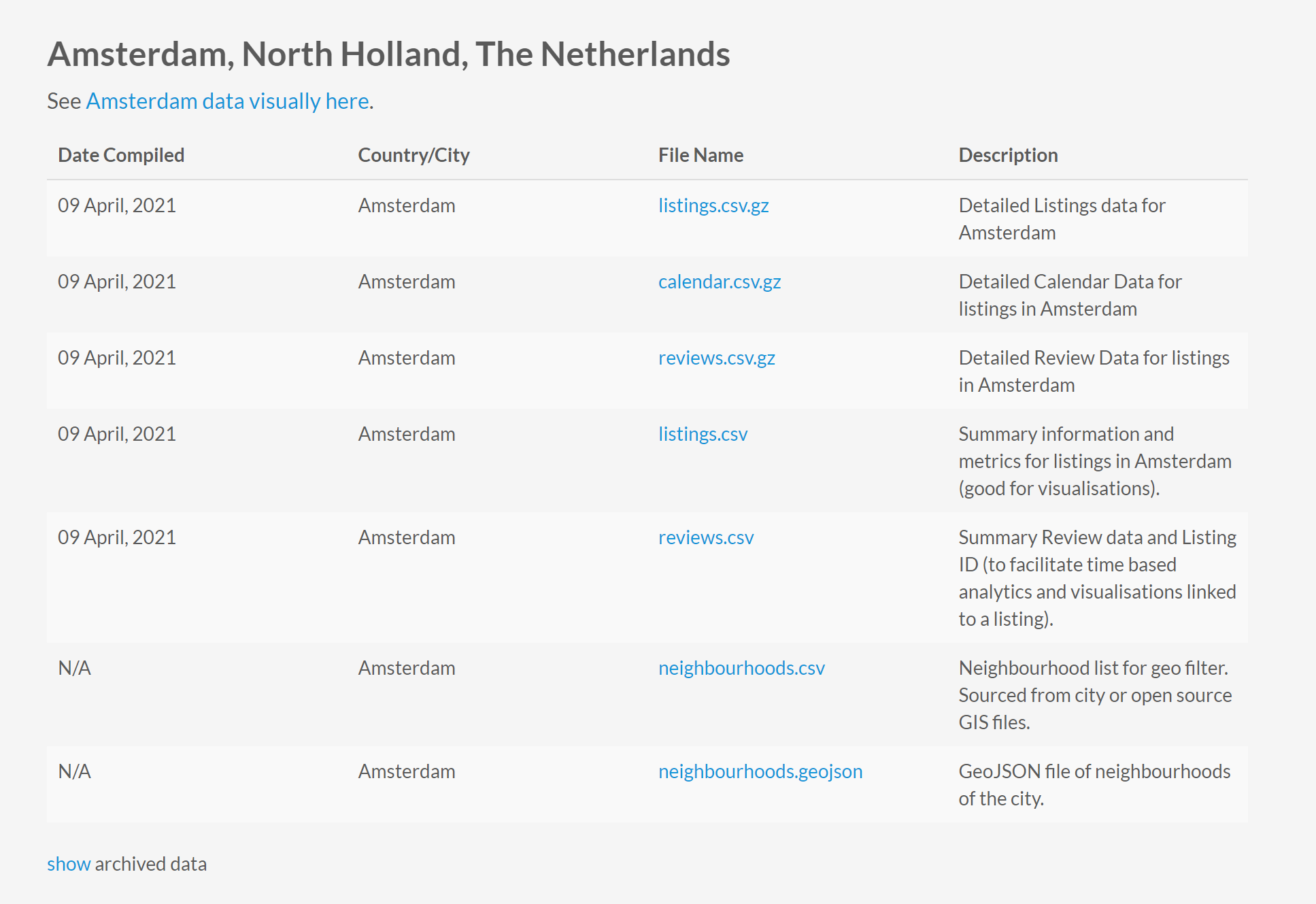
数据来源：<http://insideairbnb.com/get-the-data.html>

需要的数据列表：



每个城市都会有多个可下载的数据，我们重点focus在红色框里文件，也就是.gz结尾的文件，是最全的.

灰色框里的三个文件是.gz文件的简略版本，会少很多列variables，所以应该我们不需要。

**Neighbourhoods.geojson**文件我理解应该是绘制地图可视化用的，这里我们还应该也不需要，只需要下载InsideAirbnb选择城市列表excel文件里的19个欧洲城市里含.gz结尾的3个文件：

* **Listing.csv.gz:** airbnb上该城市的房源列表，里面有以房源id为主体对应的相关信息，例如房东姓名，房东是不是优秀房东（super\_host）, 是否支持instant booking，房源描述，房源所在neighbourhood描述等等一系列variables。
* **Reviews.csv.gz:** 房源id下对应的用户评论，每个review都有一个id，以及对应reviewer\_id, 评论日期以及评论内容，这个文件的columns列相对listing.csv.gz要少很多。
* **Calendar.csv.gz: 里面含有每个房源id在某个日期下对应的，是否available，每晚价格，maximum\_nights，minimum\_nights，这个文件可以用来**calculate price and occupancy rate dependence on previous month reviews，可以evaluate the time-based relationship between review satisfaction VS price and review satisfaction VS occupancy rate

**注意：**

1. insideairbnb这个网站会定期更新数据文件的version，目前reviews.csv.gz文件从2015至今一直是V1，calendar.csv.gz文件从2019年开始至今一直是V2, listing.csv.gz文件目前最新的是V4，2020年8月更新的，2019年开始一直是V3，V3版本到V4版本是删除掉了很多变量也增加了2个变量，具体可以查看Inside Airbnb Data Dictionary这个excel文件。如有需要，黄色框里面“show archived data” 是可以查看例是版本的数据的。
2. 因为还需要欧洲的疫情相关数据结合InsideAirbnb数据一起分析，可能需要注意一下InsideAirbnb里面城市以及国家的命名方式，比如荷兰Netherlands，有可能是The-Netherlands，或the\_netherlands类似这种命名需要注意，就是和其他数据结合的时候要注意下命名，有的国家叫法不一样。