

本次实验的数据是旅行者评价数据集，每行为一条数据，不同列之间使用"|"符号分隔，从前到后，每列的含义及数据类型为：

1. string review\_id
2. double longitude
3. double latitude
4. double altitude
5. string review\_date
6. string temperature
7. double rating
8. string user\_id
9. string user\_birthday
10. string user\_nationality
11. category user\_career
12. double user\_income

具体的，我们对本次实验内容所使用的属性如下：

1. 3.1节中的分层抽样所使用的属性A为user\_career属性；
2. 3.2节中的奇异值过滤所使用的属性B为两个属性，分别是longitude和latitude。为了避免同学们使用额外的流程计算奇异值边界，给定longitude的有效范围为[8.1461259, 11.1993265]，latitude的有效范围为[56.5824856, 57.750511]，可以在代码中直接使用；
3. 3.3中，数据格式属性涉及到属性user\_birthday和review\_date，这些日期字段可能使用2018-03-21、2018/03/21、March 21, 2019这些格式，转换为哪种格式取决于同学们自己；temperature有华氏与摄氏两种，同样的，目标格式取决于同学们；需要归一化的属性则是rating；
4. 3.4中存在缺失值的属性为rating和user\_income，根据先验知识，rating近似依赖于user\_income、longitude、latitude和altitude，user\_income近似依赖于user\_nationality和user\_career。对rating和user\_income的填充可以利用这些依赖关系；