《R 语言数据分析方法与实验》个人作业

高宝俊

2021 年 4 月 28 日

1 教学目标

本作业通过解决一个数据清洗、整形、探索性分析与可视化的实际问题，旨在达到以下知识目标和能力目标。

知识目标： R tidyverse 体系下的各种工具如 readr, dplyr, tidyr, stringr, ggplot2, 常见的 R 数据结构与编程，处理日期时间的 lubridate 包（可选）等。

能力目标：数据质量与数据意识，解决复杂实际问题的能力，在实践中学习新知识的能力。

2 数据集介绍

本次作业的数据集为：tripadvisor_content.csv。

这一数据集是从 TripAdvisor（https://www.tripadvisor.com/）的网站上爬取下来的，TripAdvisor 是全球最大的旅游点评社区，这一数据集被大量的研究所采用。该数据集中主要包含评论和酒店两方面的信息。爬取下来的数据需要做很多清洗、变形才可以用于进一步的数据分析。

数据集字段：

ReviewID,RatingDate,ReviewTitle,ReviewText,NumHelpful,AvgRatingStarsThisUser,StayDate_TravelType,Via_Mobile,StarClass,PriceRange,Services,HotelURL

数据集中字段的含义如下：

ReviewID：评论的 ID，是每一条评论的唯一标识，出现在评论的 URL 中。

RatingDate：写评论的日期。

ReviewTitle：评论的标题。

ReviewText：评论文本的正文。

NumHelpful：评论所获得的点赞数。

AvgRatingStarsThisUser：当前用户对酒店服务的评分。

StayDate_TravelType：一个字符串，阐明当前用户住宿的时间（月份），以及旅行方式。从中可以提取出停留日期和旅行方式，StayDate 和 TravelType。

Via_Mobile：用户是否通过移动设备发布这一评论，1 为是，0 为否。

StarClass：酒店的星级

PriceRange：酒店房间的价格区间

Services：酒店提供的服务种类，为一个字符串，从中可以识别中酒店是否提供某种服务。

HotelURL：酒店评论页面的 URL，从中可以提取出城市和酒店的唯一标识，CityID 和 HotelID。

图 1 -图 4 展示了各个字段在网页上对应的内容。访问酒店和评论页面，进一步观察数据的特征是理解数据的最好的途径。

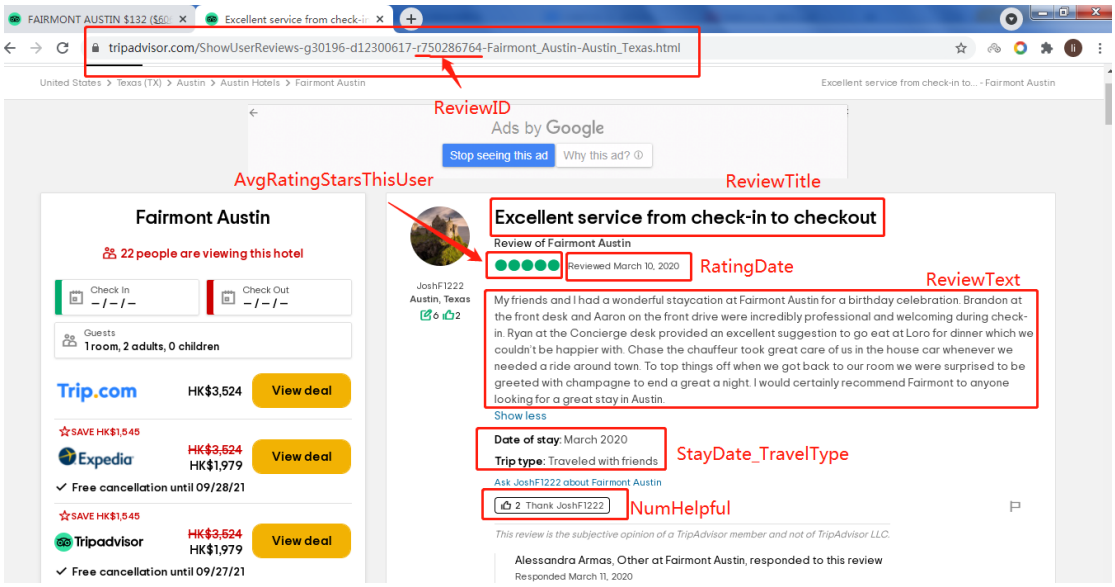

图 1 评论具体内容

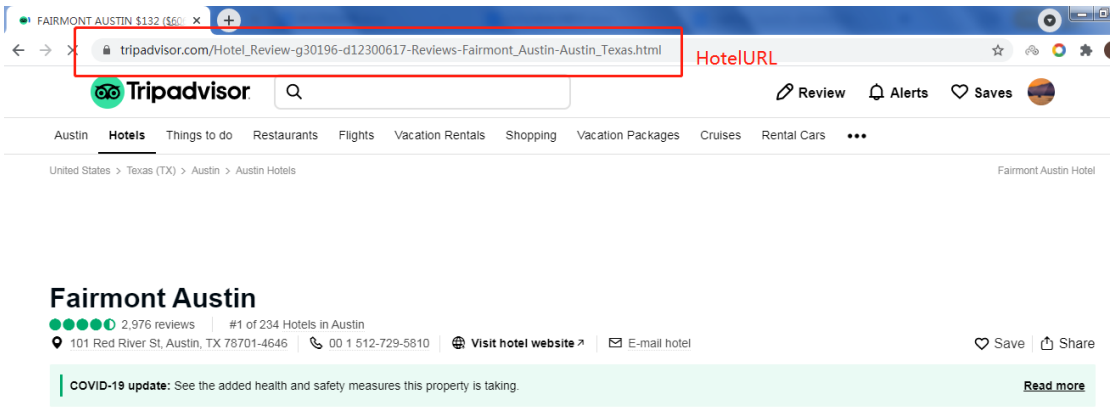

图 2 酒店 URL

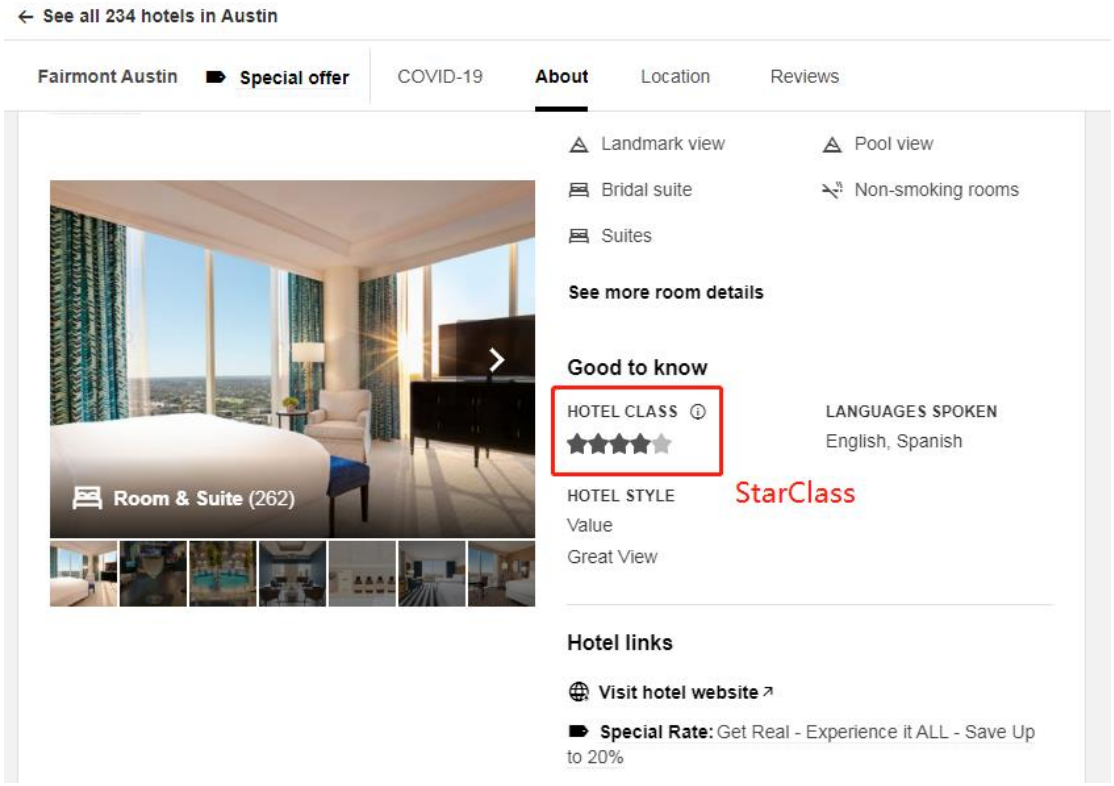

图 3 酒店 StarClass

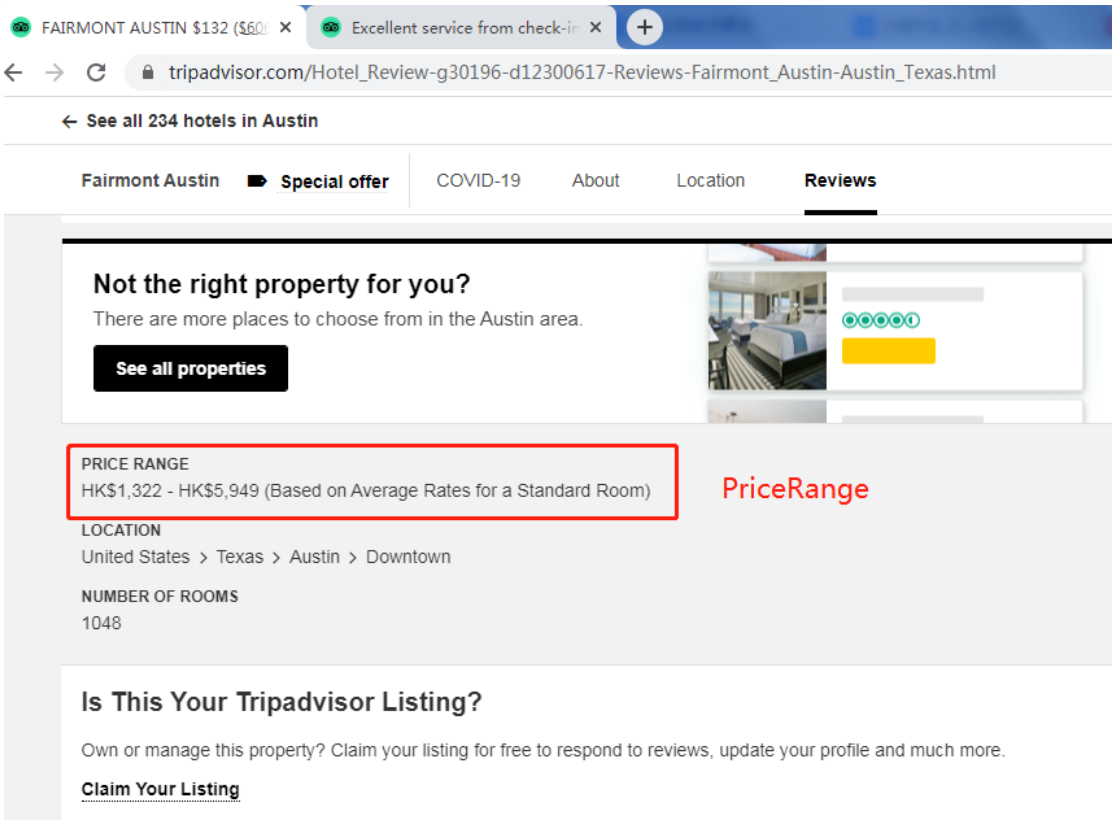

图 4 酒店 PriceRange

3 研究问题

问题 1：数据提取

从 HotelURL 中提取出 CityID 和 HotelID，将其转换为整数类型，并将其插入到 HotelURL 之前。

提示：在 HotelURL 字符串中，字母 g 和 d 后面的数字分别表示 TripAdvisor 中城市和 Hotel 的唯一标识，即 CityID 和 HotelID。

示例：HotelURL 为 https://www.tripadvisor.com/Hotel_Review-g60956-d73236-Reviews-Days_Inn_San_Antonio_Near_Lackland_AFB-San_Antonio_Texas.html，那么 CityID 为 60956，HotelID 为 73236。

问题 2：数据提取与转换

问题描述：StayDate_TravelType：一个字符串，阐明当前用户住宿的时间（月份），以及旅行方式。从中提取出停留日期和旅行方式，StayDate 和 TravelType。

将这两个字段插入到表中 StayDate_TravelType 之后，其它字段之前的位置。

将 Stay_Date，RatingDate 转换为日期型。检查 Stay_Date 和 RatingDate 的年和月份上是否一致，并在合适的位置增加一列 CheckDate 存储比较的结果。

将 TravelType 转换为 factor 类型，并将数据集中出现次数最多的 TravelType 类型设置为因子的基准水平。

问题 3：规范化小表的生成与表连接

HotelID（HotelURL 中生成）之后的所有字段都是对 Hotel 的描述，之前的字段则是对 Review 的描述。同一个 Hotel 的所有 reviews 中，有关 Hotel 的描述部分都是相同的，因而这部分数据是冗余的，不满足第三范式（3NF）。

通过下列操作，将一张表拆分成两张表，使得结果满足第三范式。步骤如下：

首先，将 HotelID 及其之后的所有字段提取出来作为一张表，命名为 Hotel，这张表的主键是 HotelID。然后，将原表中 HotelID 之后的所有字段都去掉，仅仅保留 HotelID 之前的字段和 HotelID，将其命名为 Review；这张表的主键为 ReviewID，外键为 HotelID。两张表可以通过 HotelID 进行连接。

新生成的 Hotel 表中有很多重复的行，请去掉重复的行，使得每个 hotel 在表中有且仅有一行。

最后，再将 Review 和 Hotel 两张表连接起来，观察与原始数据表有没有区别？此处连接时内连接和左连接得到的结果是否相同？为什么？

问题 4：因子识别与长宽表转换

字段 Services 列出了每个酒店提供的服务。通过观察数据可以发现，各种服务通过“-”连接成了一个较长字符串。

将这个字符串分隔开，就可以知道每个酒店分别提供哪些服务。

识别出数据集中所有酒店总共有哪些种类型的服务（假设为 N 种类），在数据集 Services 之后其他字段之前增加 N 列，分别命名为“is_服务的名称”，该字段类型为逻辑型变量，若酒店提供该种服务取 TRUE，否则取 FALSE。

提示：要知识点为 stringr，tidyr, 可能用到 list 和循环。为了方便操作，可以在 Hotel 表中进行操作，或者只提取出 HotelID 和 Services 两个字段生成一张小表，得到结果之后再与已有的数据表进行连接。

问题 5：数据聚合与面板数据生成

理论研究和实践表明，在线评论可以产生巨大的经济价值——可以影响产品销售，并提升公司的资本市场价值。为了研究在线评论对酒店销售的影响，我们就需要将数据聚合到酒店层面。

评论量（Volume）、平均评分（Avg_Rating）和评分标准差(Std_Rating)是最常用的三个评论指标。请计算下列 6 个酒店评论的指标：每个酒店每个月份的评论量（Volume）、平均评分（Avg_Rating）和评分标准差(Std_Rating)的当期值和累计值。当期值是基于发生在该月份的评论数据的计算结果；而累计值是基于从某个酒店的第一条评论开始，截至到该月月末的数据得到的计算结果。

上述六个指标和 HotelID，年月生成一张表，这张表的主键是 HotelID + 年月。这就是所谓的面板数据。

从理论上讲，每家酒店从开始有第一条评论的月份开始，每个月份的数据都应该存在。但是，在你生成的数据集中，如果某家酒店在某个月份没有评论的话，该酒店在该月份的观测值可能就没有计算出来，因而存在数据缺失。

请采取合适的方法补齐缺失数据。如果某家酒店在某个月份没有评论，那么三个当期值指标就取 NA；而累计值指标为与上一个月的累计值指标相同。

提示：可以先生成一个 HotelID + 年月的完整的数据表（注意每家 hotel 开始的时间不一样），然后再将该表与前面生成的不完整的表进行连接。

问题 6：探索性数据分析与数据可视化

到现在为止，我们有评论层面的数据（评论+酒店）和酒店层面的面板数据。请基于以上两个数据集，提出一些你认为有价值的研究问题，通过探索性数据分析与 ggplot2 数据可视化来回答这些问题。

问题的质量，用到知识点的广度和深度是该题的评分标准。

4 提交内容

本次作业需要用 Markdown 撰写，作业中加入必要的描述性文字。提交 Markdown 文件和编译好的 html 文件，以及生成的 RData 文件。