

**大数据分析**

数据分析报告

|  |  |
| --- | --- |
| 题 目 | 酒店预订数据分析与可视化 |
| 学生姓名 | 彭德龙、仪超 |
| 专业班级 | 软件工程 |
| 指导老师 | 孙娜 |
| 学 号 | 22301992、22301996 |
| 院 （系） | 信息与工程学院 |

|  |
| --- |
|  |

目 录

[1 环境背景 2](#_Toc132742299)

[2 基本技术框架 2](#_Toc132742300)

[3 数据分析 2](#_Toc132742301)

[3.1 数据内容 3](#_Toc132742302)

[3.2 建库建表、加载数据 4](#_Toc132742303)

[3.3 ETL数据清洗 5](#_Toc132742304)

[3.4 分析与可视化 5](#_Toc132742305)

# 环境背景

Expedia是一家全球知名的在线旅游公司，提供酒店预订、机票预订、租车预订等服务。Expedia通过自己的网站和移动应用程序向消费者提供这些服务，同时也为其他在线旅游代理商提供后端技术支持。

随着旅游业的发展和互联网技术的普及，越来越多的人开始使用在线平台进行旅游预订。这导致了大量的酒店预订数据的产生。这些数据包括预订时间、酒店地点、房间类型、价格等信息。这些数据对于Expedia来说非常有价值，可以帮助该公司更好地了解用户需求和市场趋势，优化产品设计和营销策略。

因此，可以利用大数据分析技术对这些数据进行分析，以便更好地了解市场和客户需求，提高业务效率和盈利能力。通过分析这些数据，可以获得更好的洞察，以指导其产品开发和市场营销策略，并最终提高公司的业务水平和竞争力。

# 基本技术框架



数据分析基本技术框架

本次数据分析基本流程和所需技术如图1-1所示，原始数据存储空间为1.92GB，本次数据分析，首先将原始数据切割为存储空间为300M的txt文本，数据信息为1628588条；其次将切割后的文本数据上传到Hadoop、Hive集群，基于集群使用SQL语言对数据进行粗处理；然后使用Python语言使用Pandas对粗处理后的数据进行细处理，并使用Pyecharts对细处理的数据进行数据可视化。

# 数据分析

## 数据内容

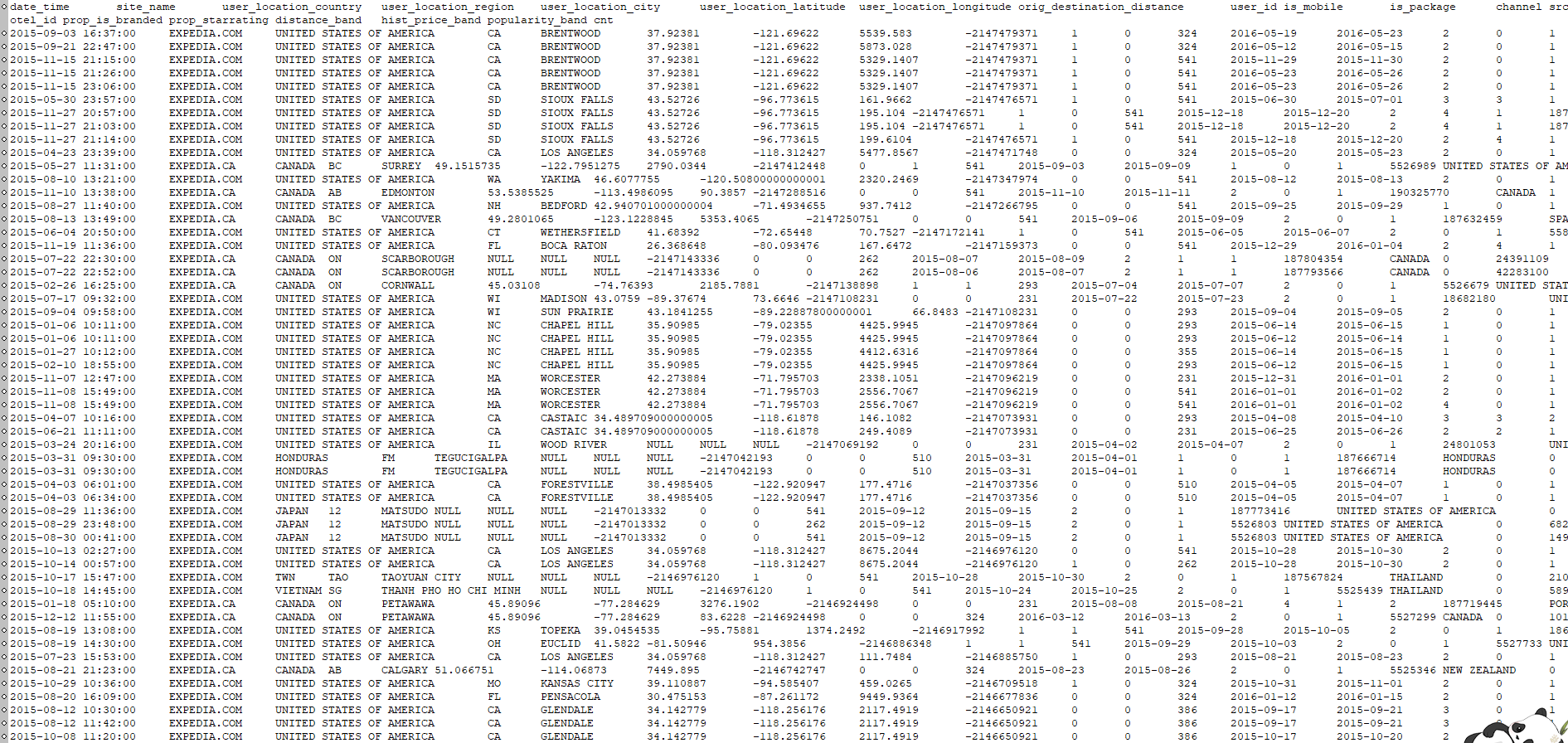
数据大小：存储空间为300M，1628588条数据。

列分隔符：制表符\t。

数据字典及样例数据：

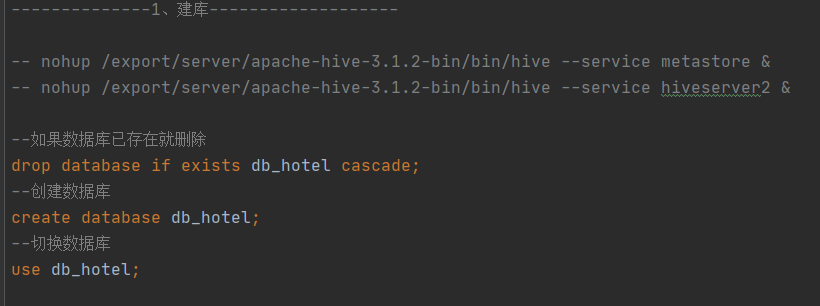
样例数据信息

|  |  |
| --- | --- |
| 数据名称 | 样例数据信息 |
| date\_time | 2015-03-26 22:08:00 |
| site\_name | EXPEDIA.COM |
| user\_location\_country | UNITED STATES OF AMERICA |
| user\_location\_region | CO |
| user\_location\_city | EAGLE |
| user\_location\_latitude | 39.660264999999995 |
| user\_location\_longitude | -106.8289 |
| orig\_destination\_distance | 2378.5325 |
| user\_id | 1227447697 |
| is\_mobile | 0 |
| is\_package | 0 |
| channel | 510 |
| srch\_ci | 2015-05-09 |
| srch\_co | 2015-05-10 |
| srch\_adults\_cnt | 1 |
| srch\_children\_cnt | 0 |
| srch\_rm\_cnt | 1 |
| srch\_destination\_id | 186006912 |
| hotel\_country | UNITED STATES OF AMERICA |
| is\_booking | 1 |
| hotel\_id | 48623065 |
| prop\_is\_branded | 1 |
| prop\_starrating | 3.0 |
| distance\_band | F |
| hist\_price\_band | L |
| popularity\_band | VH |
| cnt | 1 |

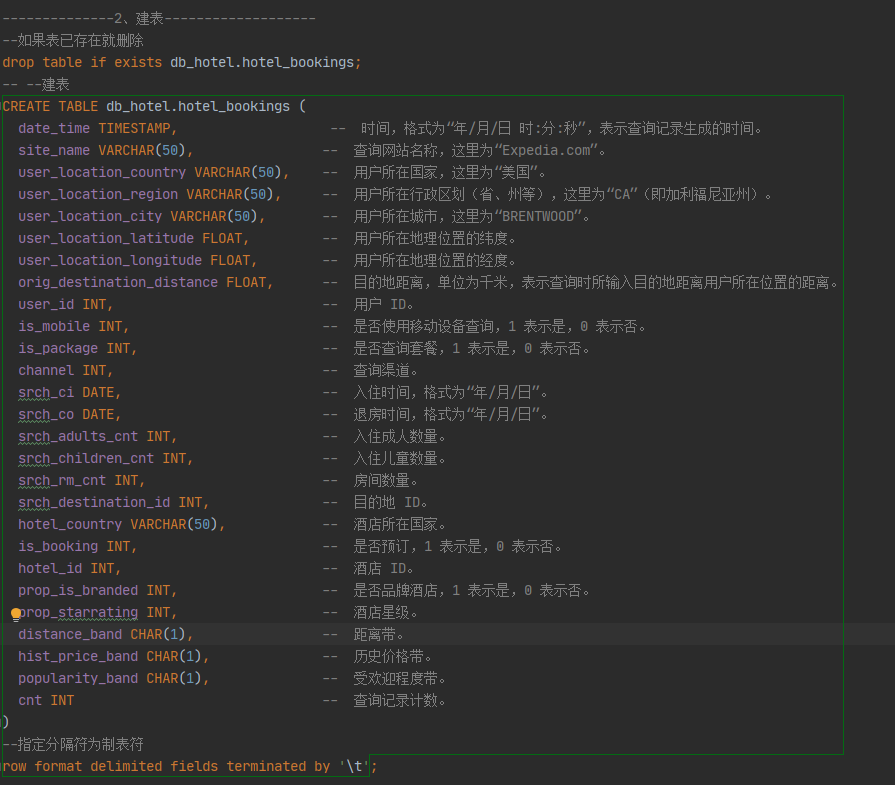


原始数据样例

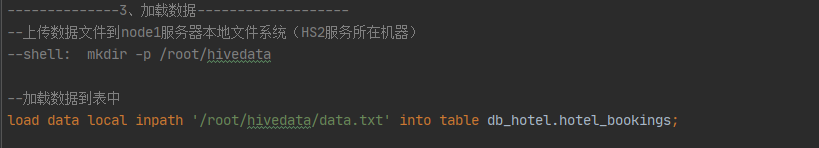
## 建库建表、加载数据



基于Hive集群建库、建表



数据分析并建表



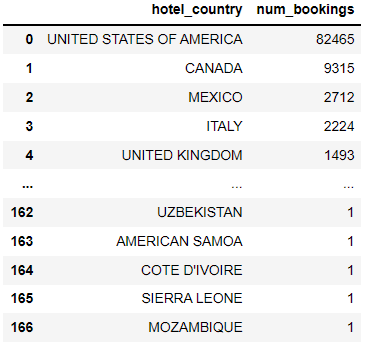
加载数据到集群服务器

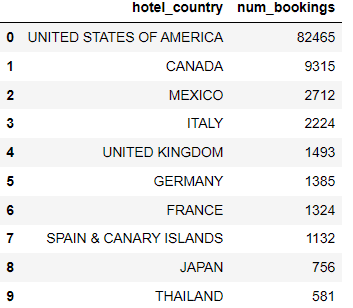
## ETL数据清洗

数据内容问题：部分数据信息中user\_location\_latitude、user\_location\_longitude 、orig\_destination\_distance 为NULL，清洗这部分为空的数据，清洗后剩余1291375条数据。

## 分析与可视化

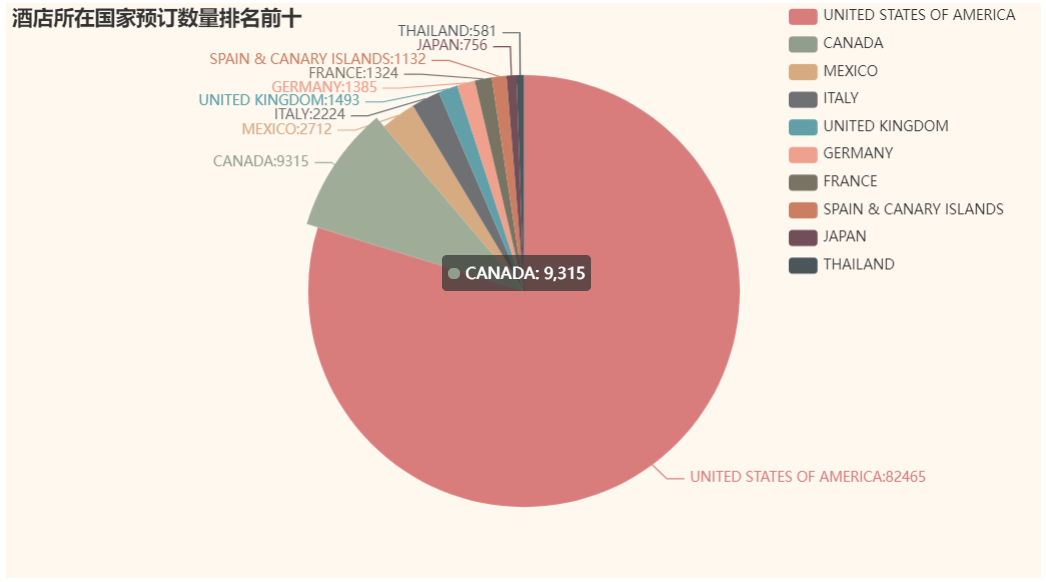
(1) 根据酒店所在国家计算预订的数量，并按预订数量从高到低排序。



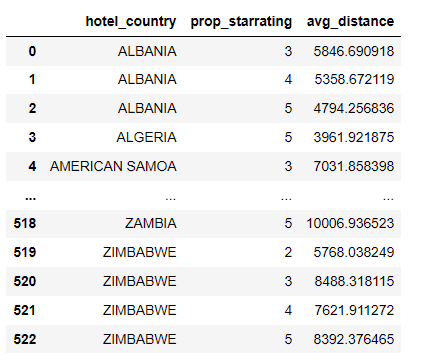


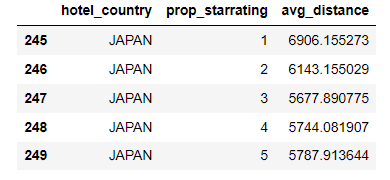
分析结果：本次分析计算酒店所在国家预订的数量，数据涉及166个国家及地区，其中预订次数最高的国家还是与美国周边国家以及主要语言为英语的国家，前十的国家只有日本和泰国为亚洲国家，可以分析出美国心仪的亚洲旅游国家为日本和泰国。

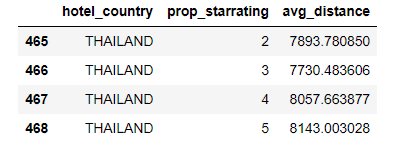
可视化：按预订数量对前十国家可视化。

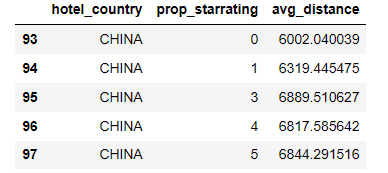


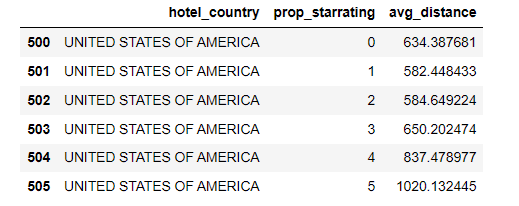
(2) 计算预订酒店所在国家和星级的平均目的地距离。





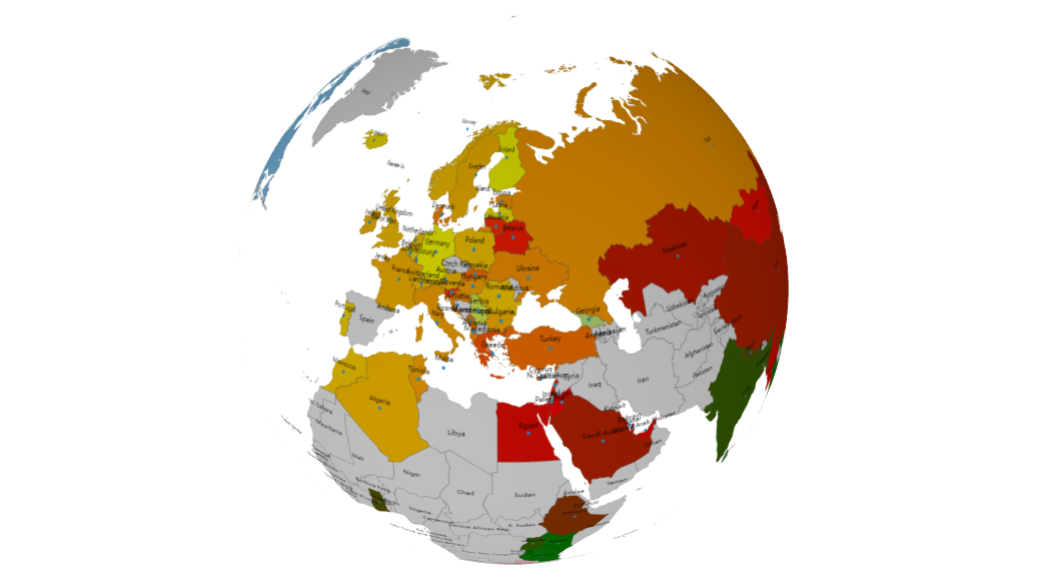




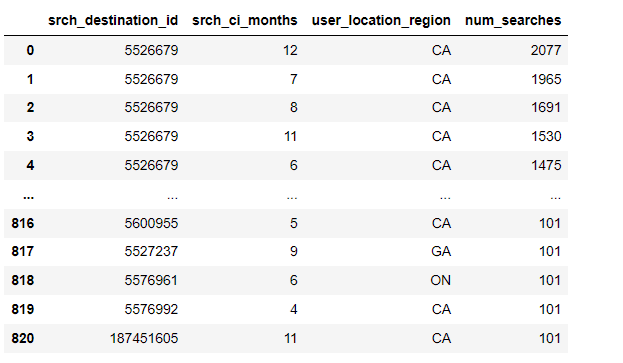


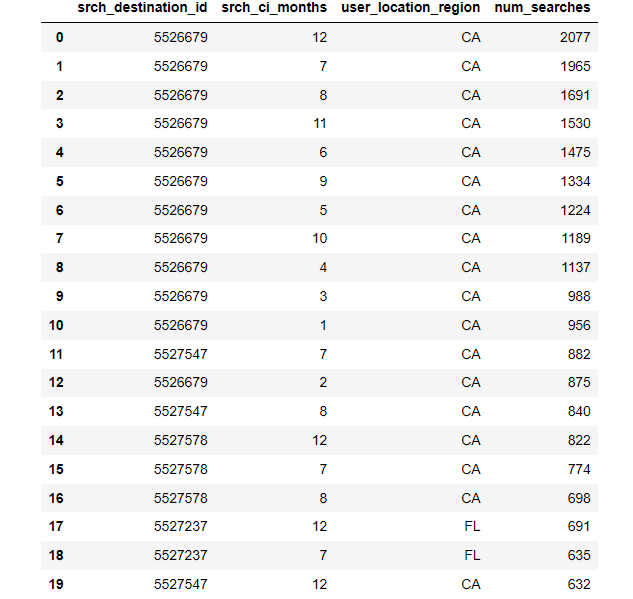
分析结果：本次分析结果统计了各个国家不同星级酒店的平均目的地距离，以亚洲国家中国、日本、泰国及数据本土美国为例，可以看出不同星级酒店平均距离还是保持在一定范围内并没太大差距，而美国四星级和五星级酒店的平均距离要远高于其它星级，可以分析出预订高星级的酒店人数较少，而亚洲国家没有这个情况，可以看出货币购买力在亚洲国家高，使得用户不必考虑酒店档次。

可视化：对预订酒店所在国家和五星级酒店的平均目的地距离可视化。



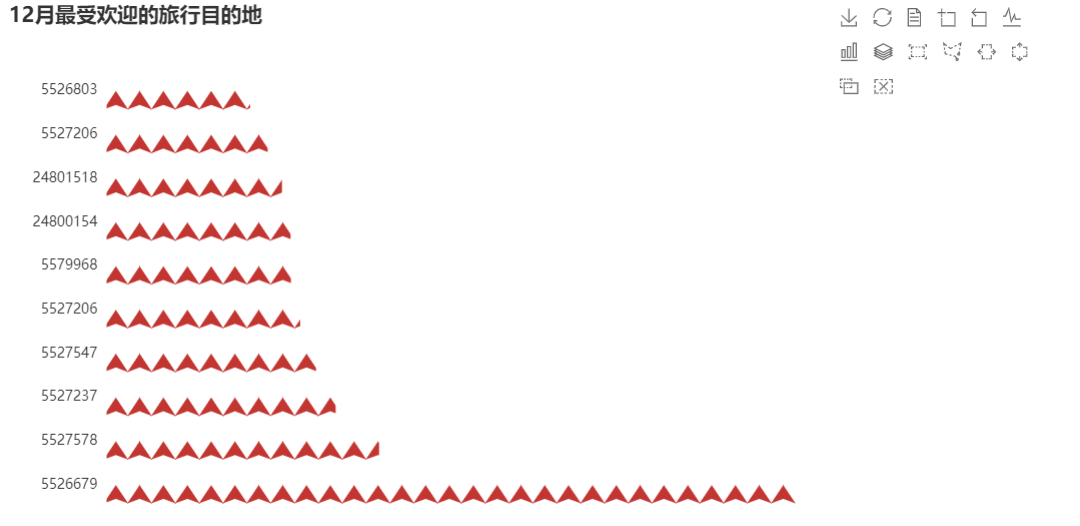
(3) 根据搜索的目的地、月份和用户所在行政区划计算搜索次数，并筛选出搜索次数大于 100 的记录，以便确定最受欢迎的旅行目的地。



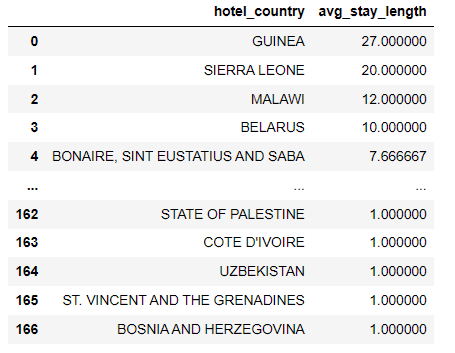


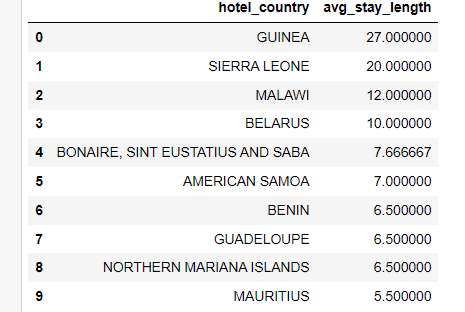
分析结果：可以看出数据前20条结果中前10条中最受欢迎目的地都是5526679，但该目的地相比于其它月份，2月份该地区是较为不受欢迎的。

可视化：筛选出搜索次数大于 100 的记录，确定12月份最受欢迎排名前十的旅行目的地。



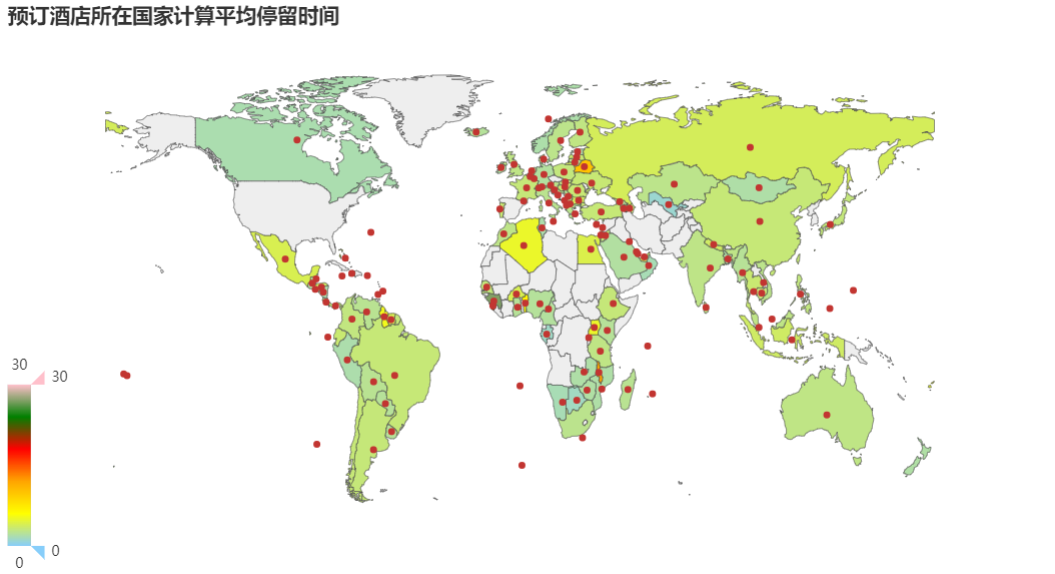
(4) 根据预订酒店所在国家计算平均停留时间，并按平均停留时间从高到低排序。



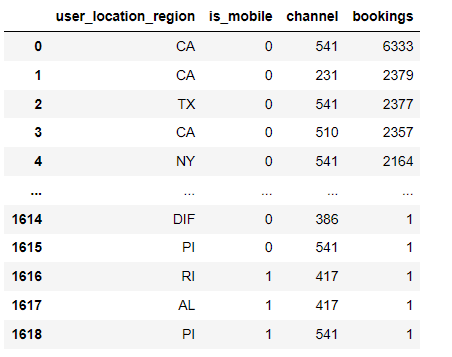


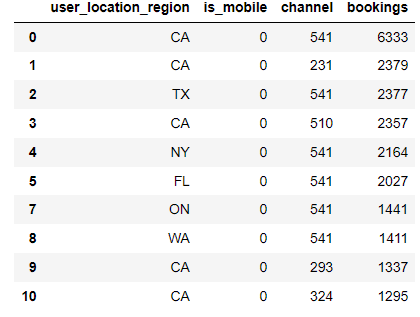
分析结果：根据不同国家的平均停留时长，可以分析出大部分国家停留时长保持在1-7天左右，而几内亚、塞拉利昂这些国家平均停留时长竟达到了20天以上。

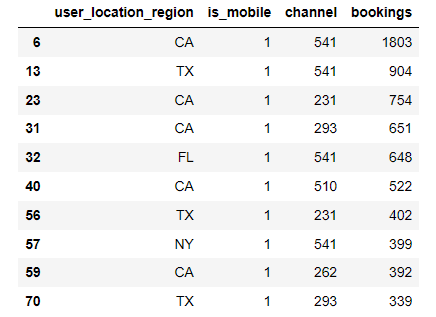
可视化：对预订酒店所在国家计算平均停留时间。



(5) 根据用户所在行政区划、设备类型和预订渠道计算预订次数，并筛选出最受欢迎的预订渠道。

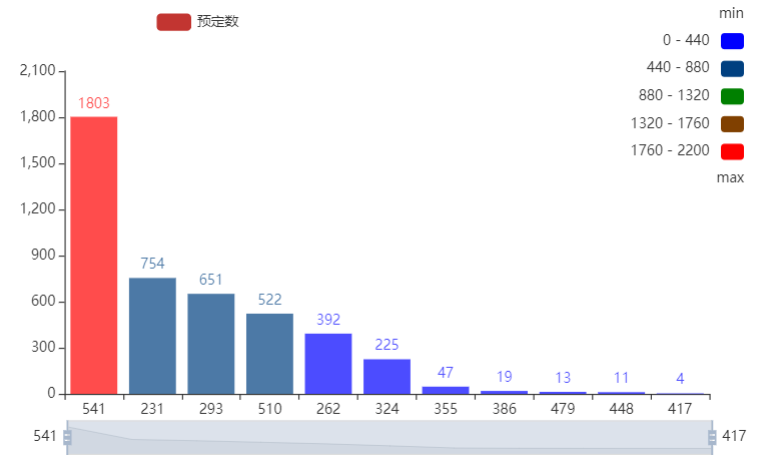




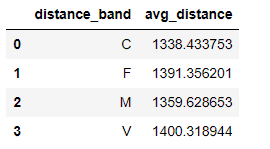


分析结果：根据预订次数可以分析出使用非移动设备最受欢迎的预订频道是541频道，使用移动设备最受欢迎的预订频道也是541频道。

可视化：对加利福尼亚地区使用移动设备的预订渠道的预订次数，并筛选出最受欢迎的预订渠道可视化。

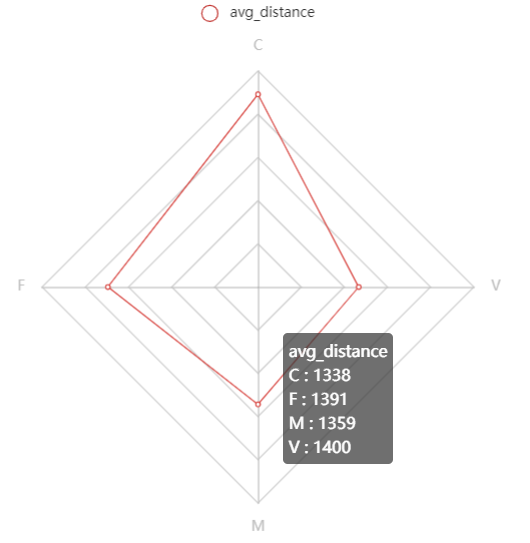


(6) 根据酒店距离带计算平均距离。

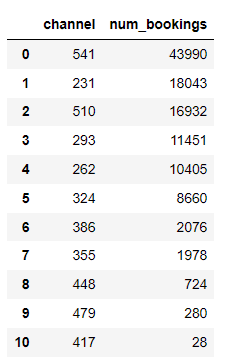


分析结果：酒店距离带共包含四个属性值，分别为C、F、M、V，计算不同距离带的平均距离，可以分析出它们之间差值并不明显。

可视化：

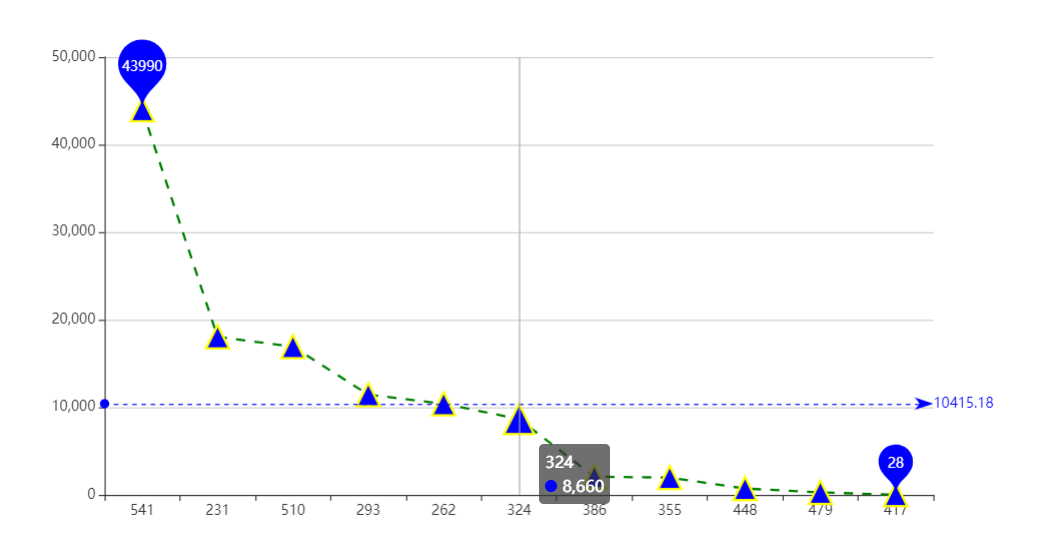


(7) 计算每个预订渠道的总预订次数,筛选出最受欢迎的预定渠道。

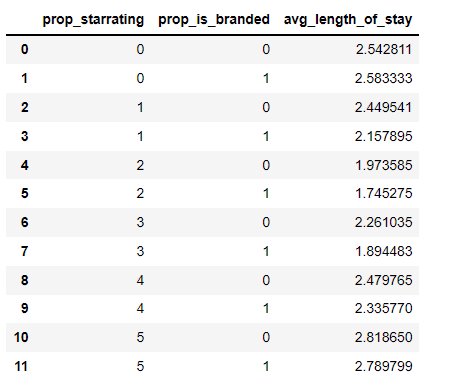


分析结果：共统计了10个频道，其中全国所有地区最受欢迎的频道为541。

可视化：

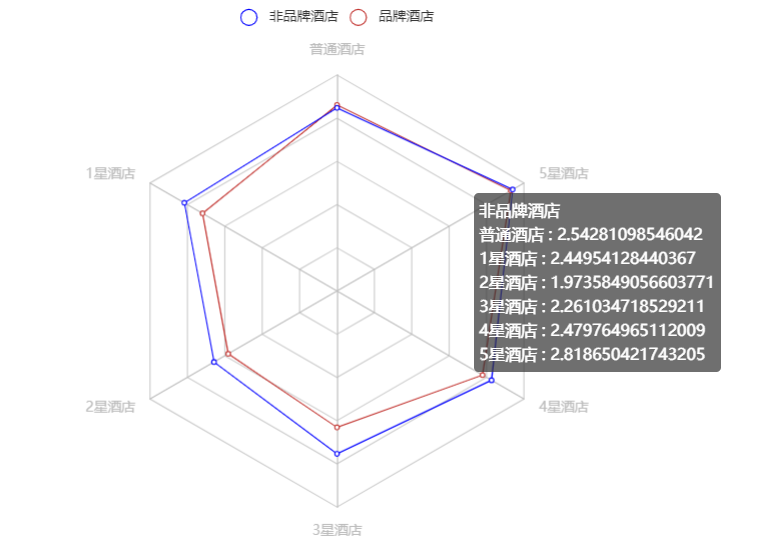


(8) 根据酒店的星级和品牌信息计算预订的平均停留天数。

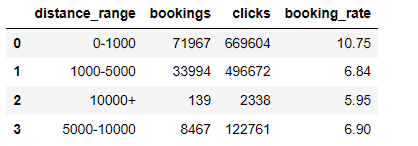


分析结果：从数据中可得出，除了一星级酒店，其它星级酒店品牌酒店比非品牌酒店的平均停留时间都略低，而一星级品牌酒店高于非品牌酒店。

可视化：



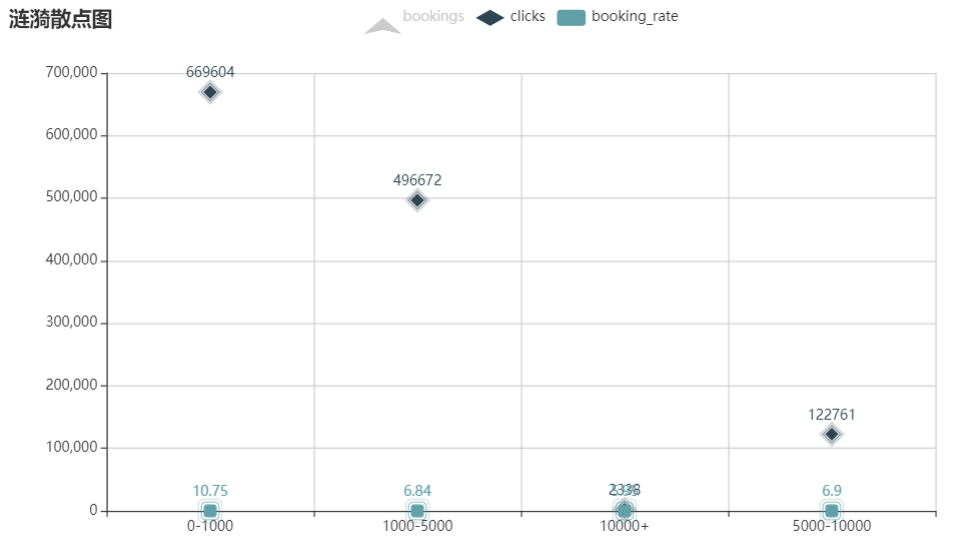
(9) 根据原始目的地距离段计算预订率。



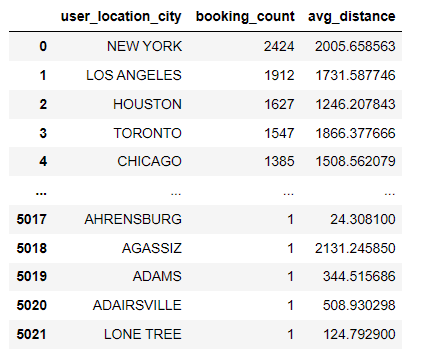
分析结果：将用户与目的地之间的距离划分为四阶段，统计各阶段的预订次数，非预订计算为点击数，统计其预订率。根据数据可得出，0-1000（km）之间的预订率是最高的，1000-5000、5000-1000（km）阶段的预订率保持在6.90%附近，在10000+（km）阶段的无论是预订率还是其次数都最低，可以得出人们更乐意去距离5000（km）以下的目的地旅游。

可视化：

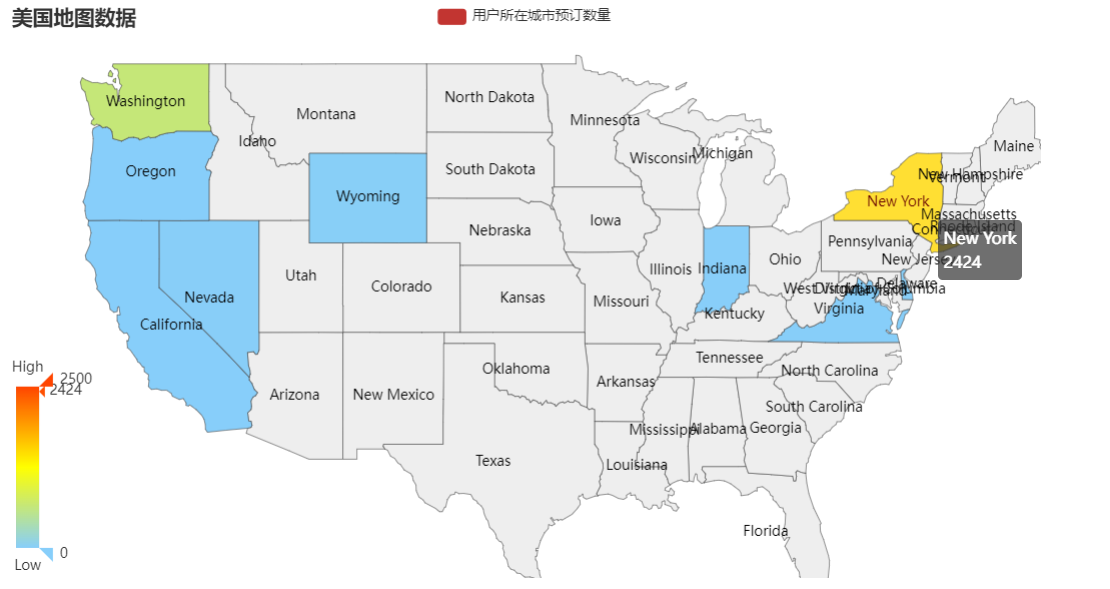


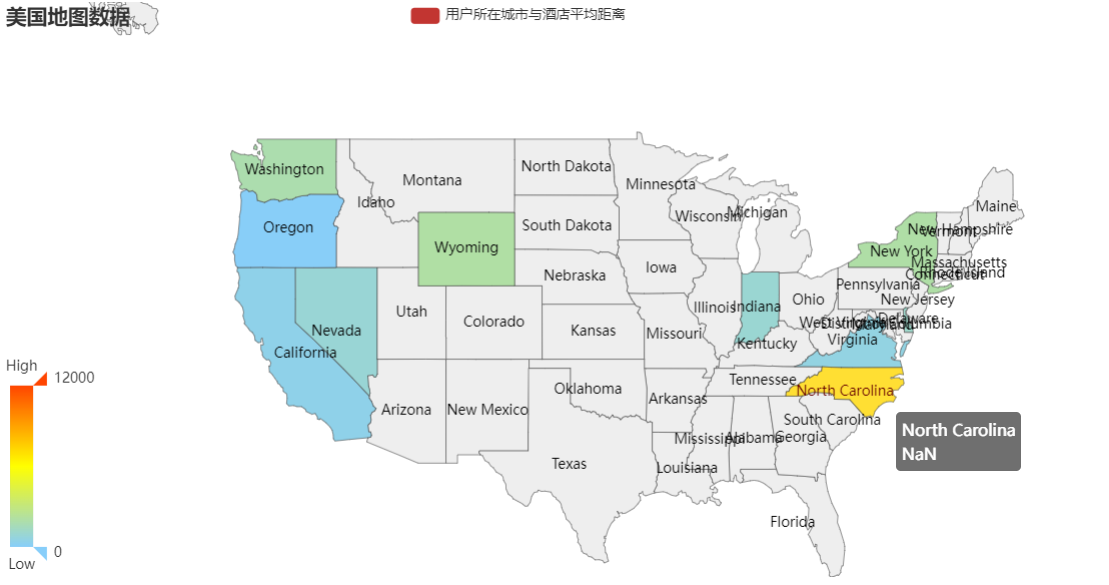


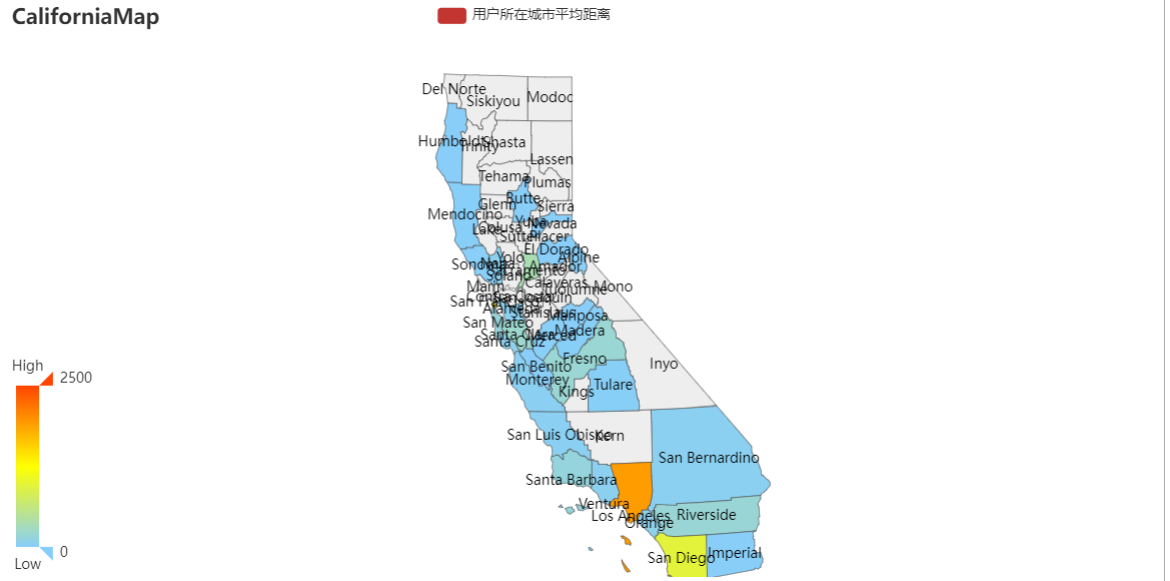
(10) 统计每个用户位置城市的酒店预订数量和平均酒店距离，并按预订数量降序排列。

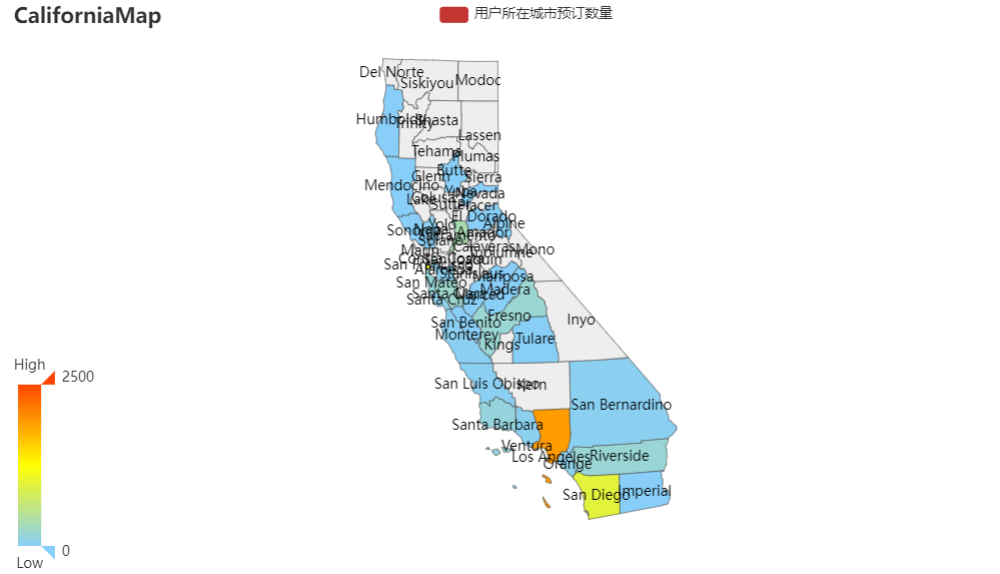


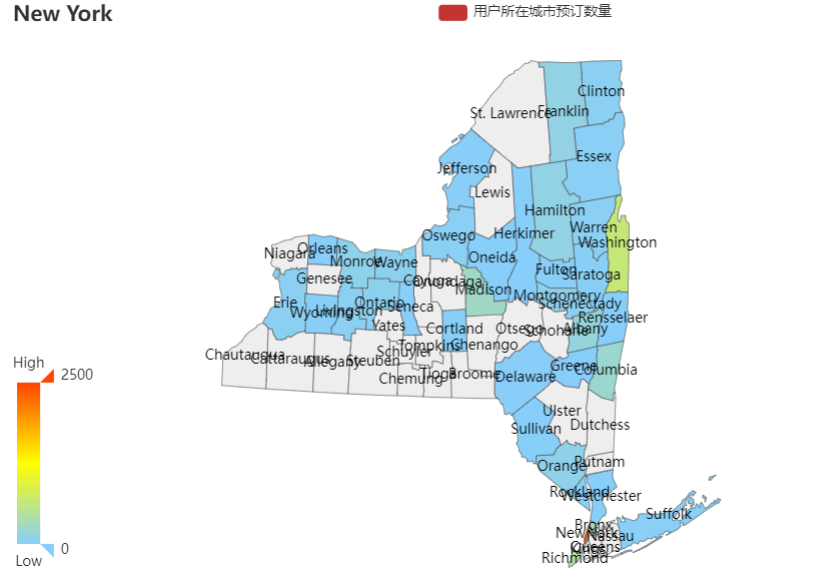
可视化：对用户所在城市预订数量和用户所在城市与酒店平均距离对美国、加利福尼亚、纽约等州分别进行可视化。

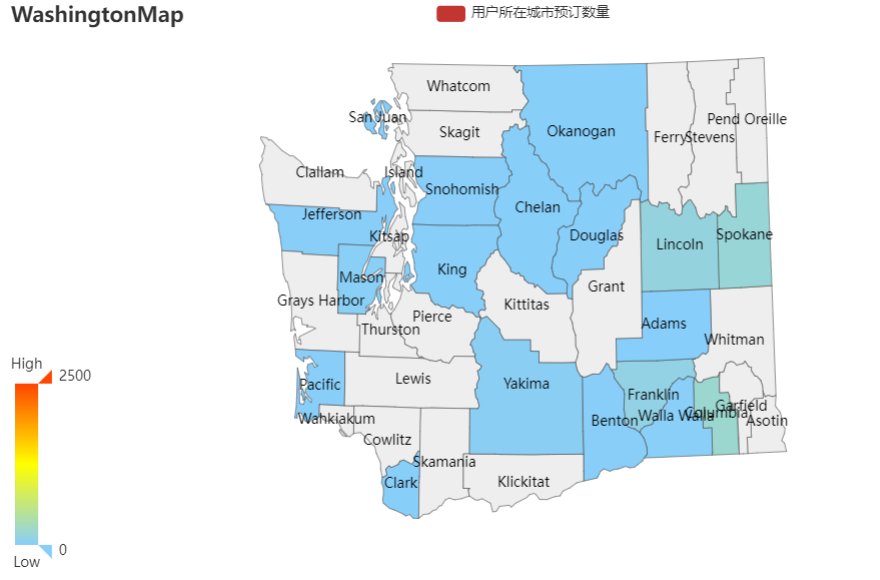


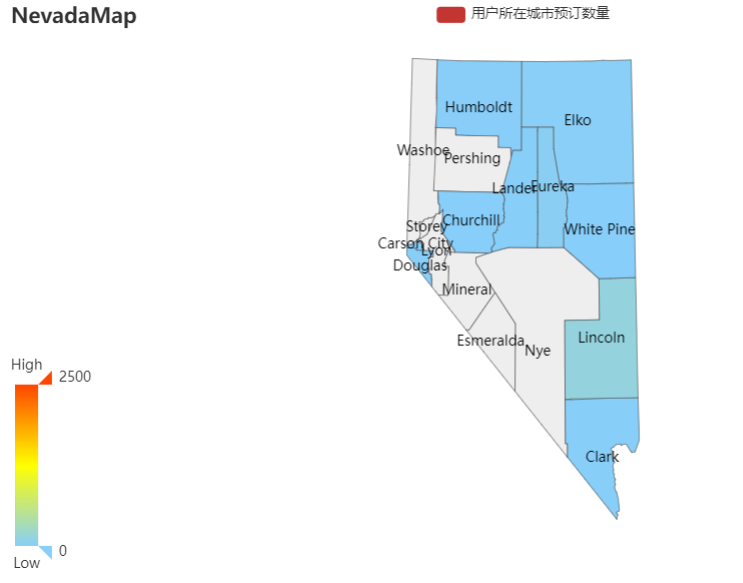




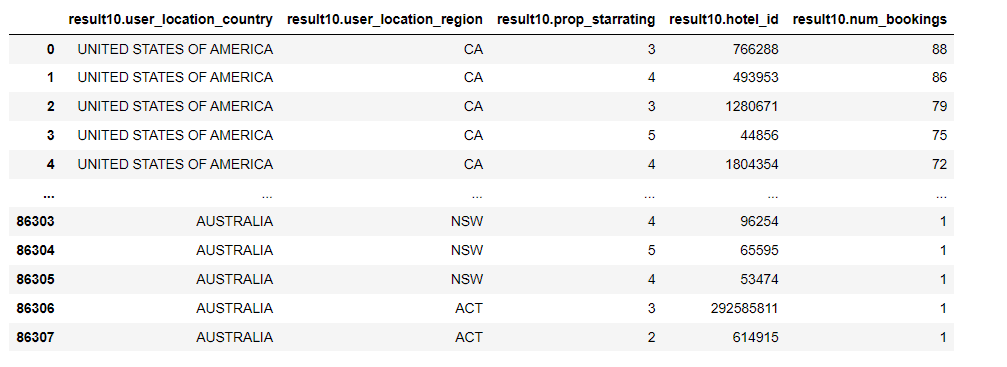




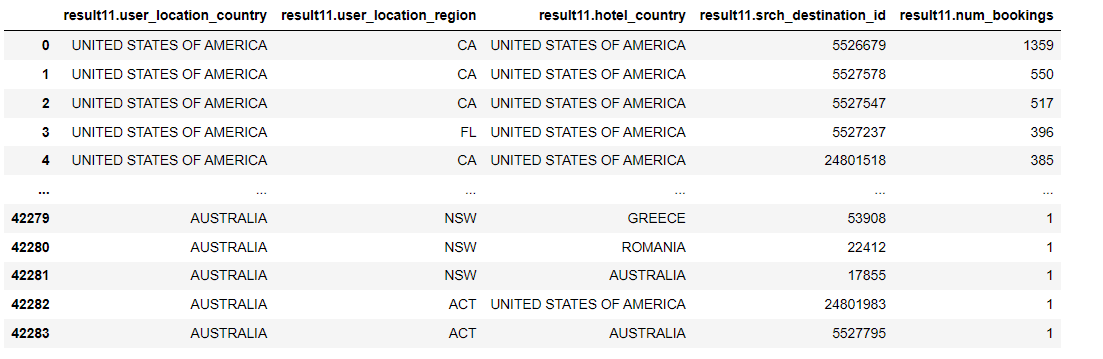




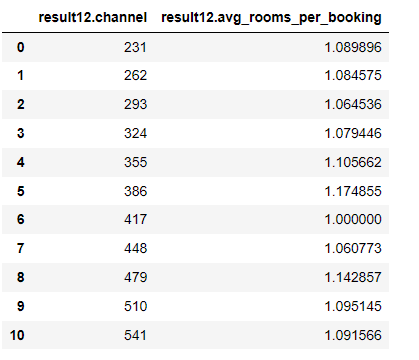
(11) 根据用户所在国家、行政区划和酒店星级计算预订次数，并筛选出最受欢迎的酒店。



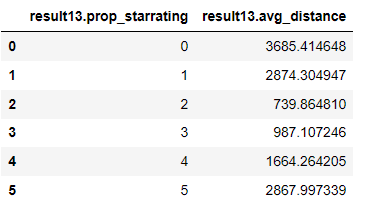
(12) 根据用户所在国家、行政区划和预订酒店所在国家,目的地ID计算预订次数，并筛选出最受欢迎的目的地。



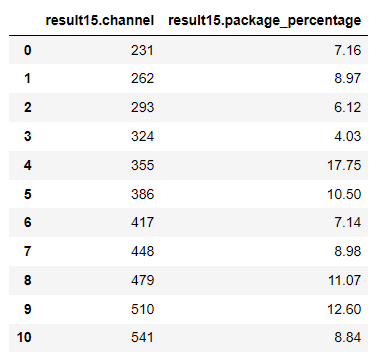
(13) 根据预订渠道计算平均预订房间数。



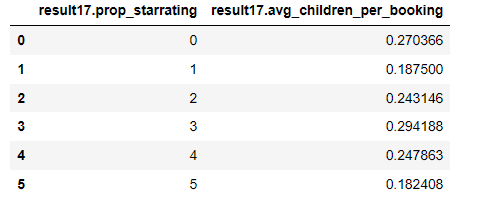
(14) 根据酒店的星级计算预订酒店与用户位置的平均距离。



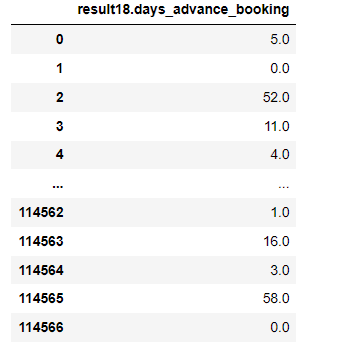
(15) 根据预订渠道计算打包销售的预订比例。

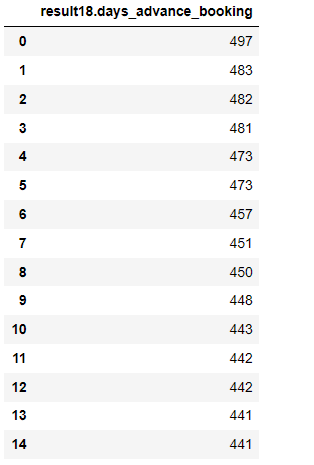


(16) 将根据酒店的星级计算平均每次预订的儿童数量。



(17) 将计算所有预订的提前天数，即预订日期和入住日期之间的天数。

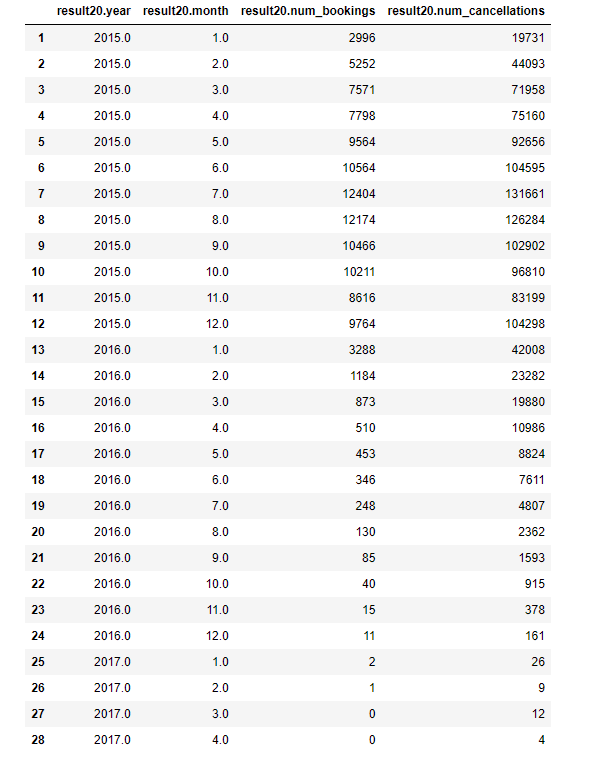




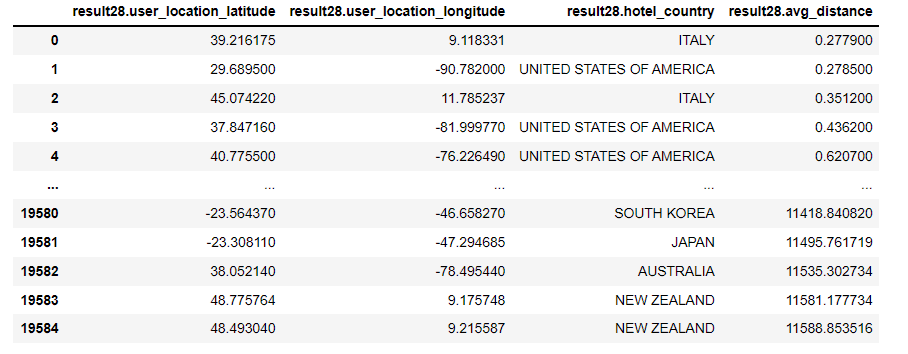
(18) 将计算所有预订的平均提前天数，即预订日期和入住日期之间的天数。



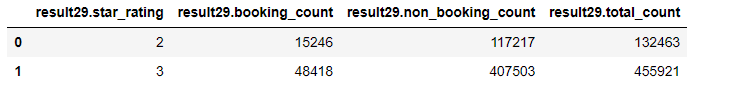
(19) 将统计每个月的预订数和未预定数，并按年份和月份分组。



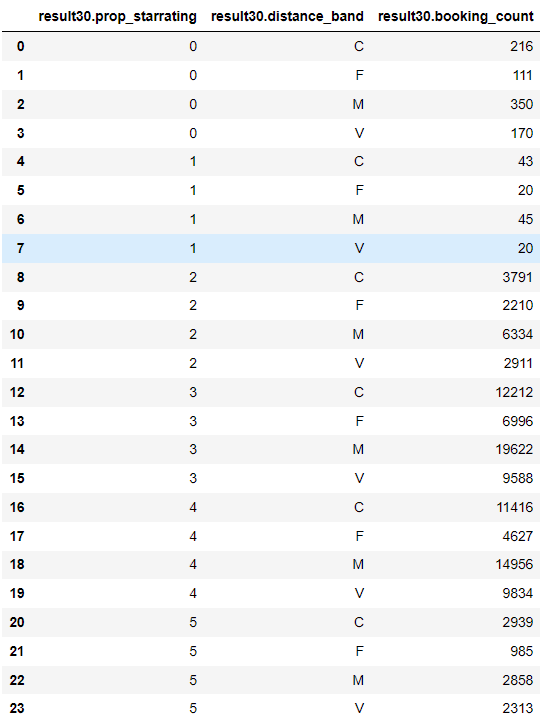
(20) 查询所有预订成功的用户位置和酒店位置之间的距离，并按距离升序排列。



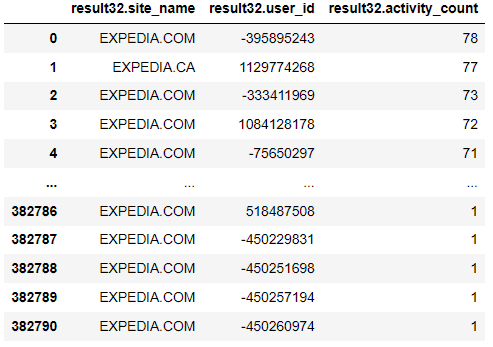
(21) 按照酒店星级分组，统计每个星级的酒店预订数和非预订数，并按照预订率从高到低排序，只显示预订率大于10%的星级。



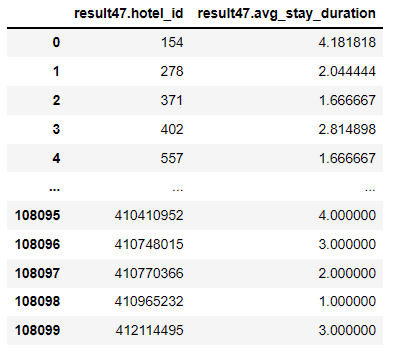
(22) 计算每个星级和价格带的预订次数。

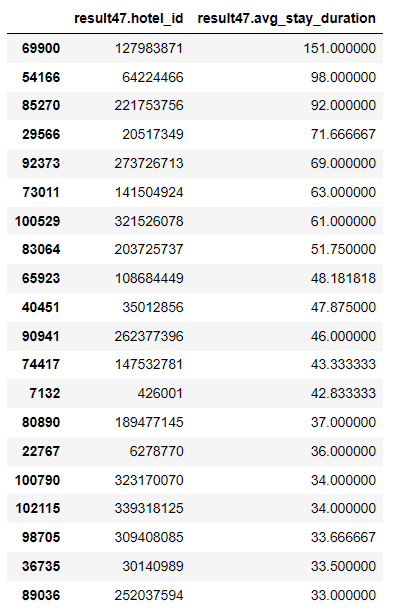


(23) 统计每个用户在该网站上的活跃度并按照活跃度从高到低排序。



(24) 统计每个酒店的平均住宿天数。





(25) 统计各个州/省的预订数量。

