数据科学导论 Capstone Project

数据可视化及 Shiny App 的制作

无 47　　刘前*　　2014011216

2017 年 1 月 5 日

1　基本说明

本人在此次 Project 中主要负责数据的可视化, 并且基于数据可视化使用 R 语言中的 shiny package 设计制作了一款 Shiny App。其中，数据可视化部分没有涉及太深入的机器学习或者数据挖掘，只是对数据进行了较为浅显的可视化操作，得出一些直观可见的结论。由于 Shiny App 的制作需要耗用较长的时间，因而较有深度的数据分析和挖掘由另外组员完成。

2　数据获取及说明

本人在做基本的可视化展示时，大部分内容以 2016 年的数据[1]作为数据源，只有在进行时间的纵向对比时，才用到前两年的数据。2016 年的数据主要包含以下文件:

表2和表3是对各文件内数据的具体说明[2]:

2.0.1　STATION INFORMATION

如表2所示。

*清华大学电子工程系 (E-mail: liuqian14@mails.tsinghua.edu.cn)

[1]实际的起止时间分别为 2015 年 9 月 1 日至 2016 年 8 月 31 日

[2]使用英文对数据进行说明

文件名	说明
201608_station_data.csv	公共自行车各站点基本信息
201608_status_trip_data.csv	2 使用公共自行车的旅行的基本数据信息
201608_weather_data.csv	2015-9-1 至 2016-8-31 之间各天的天气信息

表 1: 原始数据

Data	Explanation
station_id	station ID number
name	name of station
lat	latitude
long	longitude
dockcount	number of total docks at station
landmark	city
installation	original date that station was installed.

表 2: Station Information

2.0.2 TRIP DATA

如表3所示。

Data	Explanation
Trip ID	numeric ID of bike trip
Duration	time of trip in seconds
Start Date	start date of trip with date and time
Start Station	station name of start station
Start Terminal	numeric reference for start station
End Date	end date of trip with date and time
End Station	station name for end station
End Terminal	numeric reference for end station
Bike	ID of bike used
Subscription Type	Subscriber and Customer
Zip Code	Home zip code of subscriber

表 3: Station Information

2.0.3 WEATHER DATA

Daily weather information per service area, provided from Weather Underground in PST. Weather is listed from north to south (San Francisco, Palo Alto, Mountain View, San Jose).

3 数据的处理

在获取数据之后,重要的步骤是对数据进行清洗。数据清洗 (Data Cleaning) 虽然不是本次 Project 的重点，但是对于获取的数据仍然需要一些基础的清洗，以便于后续数据分析和可视化工作的开展。数据清洗，首先从定义上是指：发现并纠正数据文件中可识别的错误的最后一道程序，包括检查数据一致性，处理无效值和缺失值等。

本次 Project 获取的数据已经是较为规范的数据格式，但是仍有一些需要修正之处。本人重点关注了自己负责部分的数据清洗工作。数据清洗操作在数据可视化操作之前，为了不影响报告的安排，将在各数据可视化之前对必要的数据清洗操作进行说明。

同时，数据分析和可视化之前还需要数据的预处理工作，也在具体的数据可视化中一并给出。

4 需求分析与问题整理

由于选取数据是公共自行车的相关情况，所以数据的分析和可视化主要可以从两个角度展开：企业角度和用户角度。

企业和用户对数据的关注点明显是不同的。企业角度是指公共自行车的管理者或运营者的角度，比如国内的摩拜单车，此时的企业角度即指摩拜科技有限公司对数据的审视角度。数据所在的海湾地区的 Bike Share 项目也可能是企业或者政府来承办。而用户角度则主要是从公共自行车使用者 (上班族或者游客) 的角度分析这些数据，期望能够对他们的使用带来一些便利。本人从这两个角度对数据进行了分析与可视化。

4.1 企业角度

根据企业的自身利益，企业更关注一些较为全局的方面。比如公共自行车是否够用，或者是否存在使用者太少、造成资源浪费的车站等问题。这些问题直接关乎企业的经济效益，对企业的发展规划和决策起到了重要的作用。根据已获取数据所包含的信息，该部分主要重点关注了以下问题：

4.1.1 公共自行车使用情况与时间的关系

i. 按照日期分析自行车一年内整体的使用情况，便于把握数据的全貌；

ii. 按照季节分析自行车的使用情况，探究自行车的使用是否与季节的不同有关；

iii. 将数据按照一周时间进行分析，探究一周各天 (尤其是对比工作日与休息日) 的自行车使用情况；

iv. 将数据按照一天的时间进行分析，探究一天 24 小时使用自行车的分布情况，探究每天使用公共自行车的规律。

4.1.2 公共自行车骑行的路线情况

i. 从所有站点构成的骑行路线中，分析挑选出骑行人次最多 (最受欢迎) 的若干路线，从而作为向使用者推荐路线的参考；

ii. 在所有骑行路线的起点中，找出出发人次最多的若干站点，可以认为是整个线路图中骑行出发的枢纽站点。

iii. 在所有骑行路线的终点中，找出到达人次最多的若干站点，可以认为是整个线路图中的“热门”终点。

4.1.3 不同城市的自行车使用情况

按照不同的城市，分析自行车的使用情况并加以比较，作为企业在不同城市的规划部署的参考，分析的内容可以从前述分析中任选。

4.2 用户角度

从用户的角度来说，用户并不会关注每个车站的经济效益如何或者自行车是否得到了充分的利用。作为用户，更多的关注点在于实用功能，经过分析，用户期望的功能主要包含以下几个:

4.2.1 路线查询功能

假如用户在某一个地点，想要到达一个目的地，因而需要查询附近有哪些站点，以及自己的目的地大致在哪个方位，怎样才能到达目的地。不难看出，这一功能主要由地图来实现，在地图上明确标注出在运行的车站即可。

4.2.2 车站查询功能

用户确定了路线之后，需要选择从哪一站点开始出发，因而需要了解附近的站点是否有空闲的可使用的自行车。由于我们所获取的数据是往年的非

实时的数据，所以只能根据往年数据的一般情况进行可视化，然后由用户参考往年的结果来决策将哪一站点作为起始站点。

4.2.3 路线推荐功能

自行车的使用者也可能是游客，他们可能并不确定自己想要去哪些地方游览，因而需要一些参考的结果，来帮助他们了解从当前车站出发有哪些车站可以作为骑行的好去处。这些参考结果也是从往年旅行者的数据中分析并进行的数据可视化。

4.2.4 其他功能

其他功能主要包含一些更细致的查询结果，比如周末和工作日的分开查看等。

基于以上需求分析，本人认为最好的数据可视化方法是制作一个 Shiny App，将所有结果呈现在 Shiny App 中，为自行车使用者提供便利。具体的 Shiny App 的设计及最终结果详见详细分析及可视化结果请见**数据可视化**的**用户角度**部分。

5 数据可视化

由之前**需求分析与问题整理**部分的分析，本部分对数据可视化结果进行呈现。

5.1 企业角度

5.1.1 公共自行车使用情况与时间的关系

i. 按照日期分析

结果分析：

图1使用柱状图展示了一年时间内的所有数据，可以掌握数据的大致情况，但是不能从中看中什么确切的规律，唯一能够看出来的，是图中 (除中间部分之外) 定期出现的下降，感觉像是骑行数据固定地每周会有一次下降。

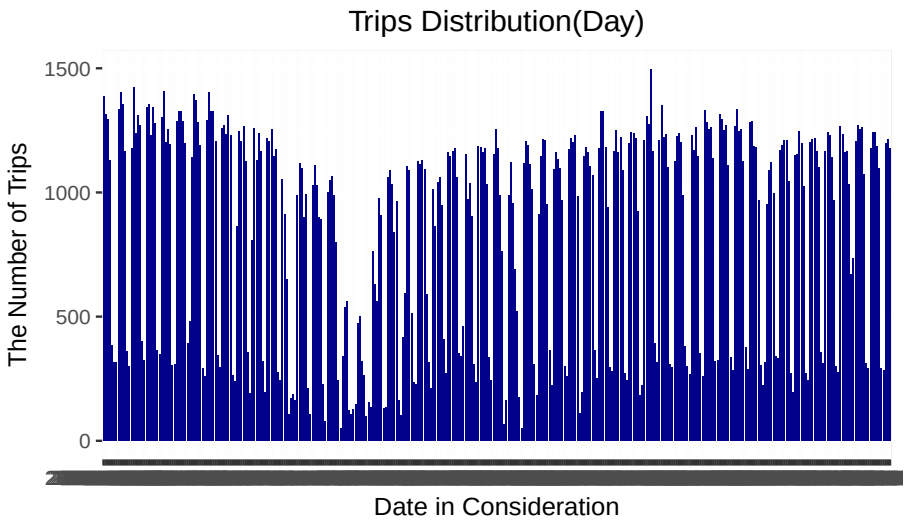

图 1: 所有骑行数据的全貌 (柱形图)

2用点图也体现了自行车的使用会定期的减少。至于更细致的分析，将在下面逐步展开。

ii. 一周各天的自行车使用情况

为了更清楚地研究之前发现的问题，我们将数据整合到一周内, 分析一周时间中骑行人次的分布情况。3非常直观地体现了工作日与周六、日的区别。周末的使用明显比工作日使用多出很多。这是因为，对于上班族来说，工作日使用公共自行车可能是由于距离公交车站或者地铁站有一定的距离，因而骑行过去比较快捷方便。工作日上班的人数减少，自行车使用量因而明显减少。

iii. 将数据按照一天的时间进行分析。

进一步，本人将用户的骑行数据按照一天 24 小时进行分析，这样基本可以分析出用户的生活规律。图4可以看出，自行车使用数每天会出现两个峰值，分别在 9 点和下午 17 点处。可以猜测，上午 9 点是早高峰时间，晚上 17 点是晚高峰时间，可见大部分上班族是“朝九晚五”型的工作。

iv. 按照季节分析一天 24 小时自行车的使用情况。

本人又将用户骑行数据按照不同季节进行分析，图5中体现出由总体得

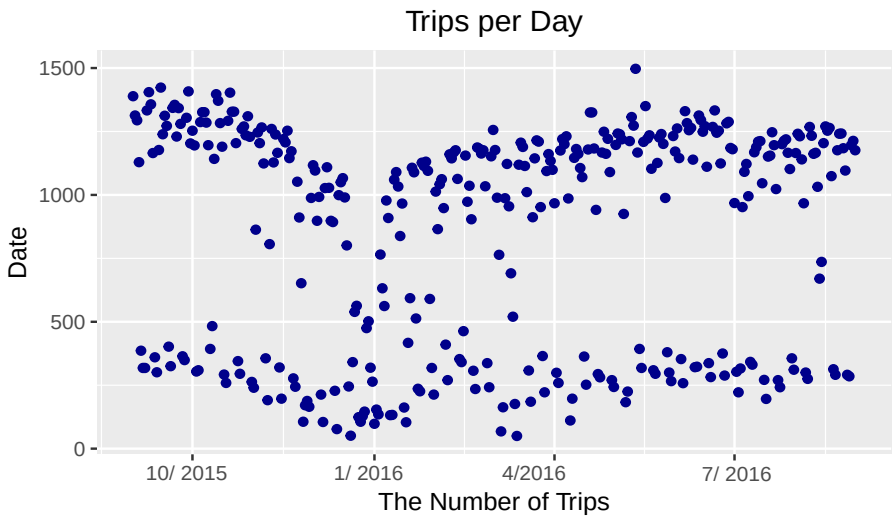

图 2: 所有骑行数据的全貌 (点图)

出的规律在不同季节仍然适用，只不过在春冬季节骑行的人次要略小于夏秋季节。

5.1.2 公共自行车骑行的路线情况

i. 从所有站点构成的骑行路线中，分析挑选出骑行人次最多 (最受欢迎) 的前 10 名路线，进行数据可视化。

图6展示了 10 条人流最多的路线，这对企业来说可以作为未来经济发展重点关注的路线和区域。

ii. 在所有骑行路线的起点中，找出出发人次最多的前 10 个站点，作为整个线路图中骑行出发的枢纽站点，进行可视化。

iii. 在所有骑行路线的终点中，找出到达人次最多的前 10 个站点，作为是整个线路图中的“热门”终点。

图7展示了 10 个出发人次最多的站点；而图8展示了 10 个到达人次最多的站点。企业可以在这些站点适当地增加车位数。

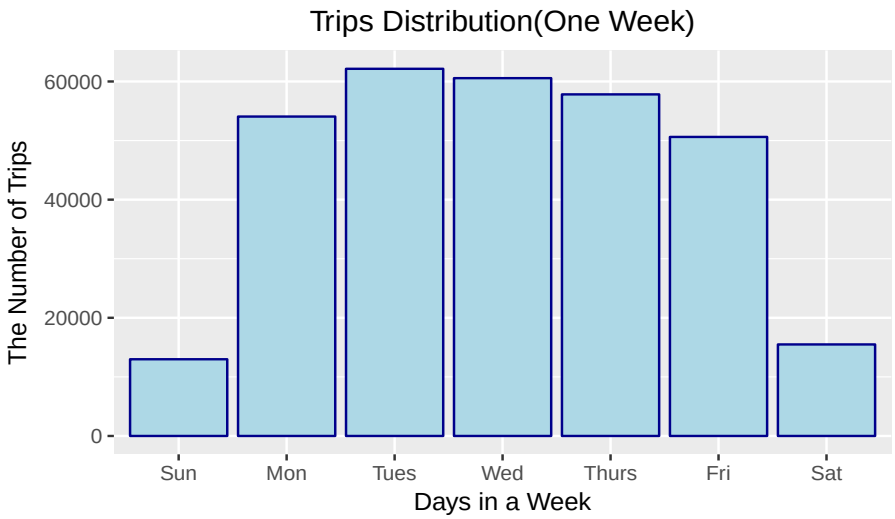

图 3: 一周各天自行车使用情况

5.2 用户角度

用户角度的数据可视化通过 Shiny App 进行了集成，主要使用了 shiny、ggplot2 和 leaflet 三个 package。shiny 实现了与使用者的交互；ggplot2 在界面中作出一些可视化的图，重点说明一下 leaflet package。这个 package 主要实现地图功能，在这方面其地图效果和功能都明显优于尝试的其他 package。leaflet 与 ggplot2 的语法十分相似，通过增加函数可以将地图界面做的非常优美。

下面将对本人设计制作的 Shiny App 做一个简要的使用说明。

5.2.1 用户界面

Shiny App 运行之后，主界面主要分为三部分。图9是主界面，图10是输入面板，图11是图形的输出面板。

输入面板主要包含四个输入：地图类型、城市、起始站点、终点站。

第一步输入 “Base Map”，其中提供了 5 种不同的地图类型，默认是 OpenStreetMap 地图 (见图12)。

第二个输入是 “City”，将自行车站点按照城市分类，选定一个城市，会

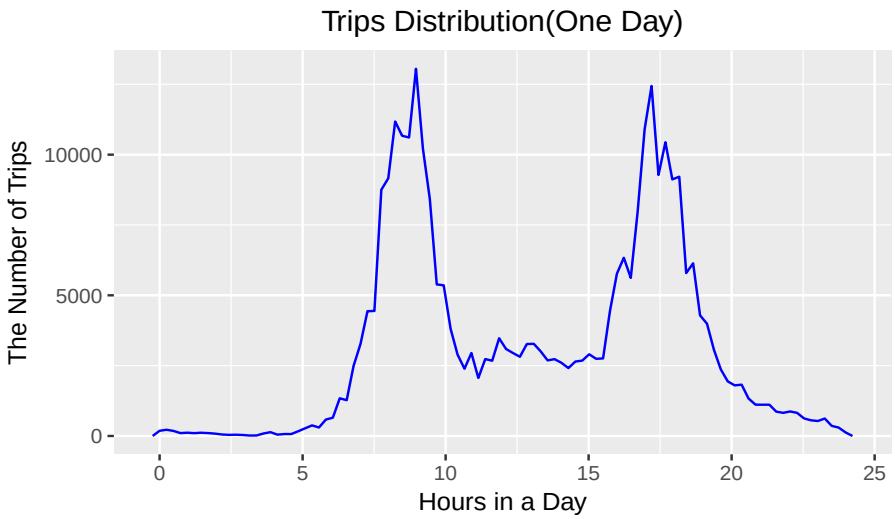

图 4: 一天 24 小时自行车使用情况

自动将地图放大到选定的城市 (以旧金山为例，见图13)，便于查询站点及线路。点击要查询的站点，会弹出该站点的所有相关信息 (如14所示)。

第三个输入是起始车站，包含车站的序号和车站名，选中起始地点之后，图11中的第 1 个图会展示出以该站点为起点的所有路线的分布情况，可以作为游客去处的参考，相当于向使用者进行了推荐。

第四个输入是终止车站，同样包含车站的序号和站名。选定起始车站和终止车站之后，图11中的第 2 张图会按照工作日和休息日展示出哪天骑行的人多，方便使用者 (游客) 规划路线和日程。第 3 张图则展示了往年数据中在不同时间出发的旅行数目，以此作为参考让用户选择合适的时间出行。

由于 Shiny App 的设计工作比较复杂，工作量过大，因而实现了上述较为基本的功能。总之，该 Shiny App 集路线查询、车站查询、路线推荐等功能于一体，基本能够满足使用者的需求。

6 个人工作小结

本人在本课程的 Capstone Project 主要完成了基本的数据可视化和 Shiny App 的制作。对 R 语言中 dplyr、lubridate 等包进行数据处理、leaflet

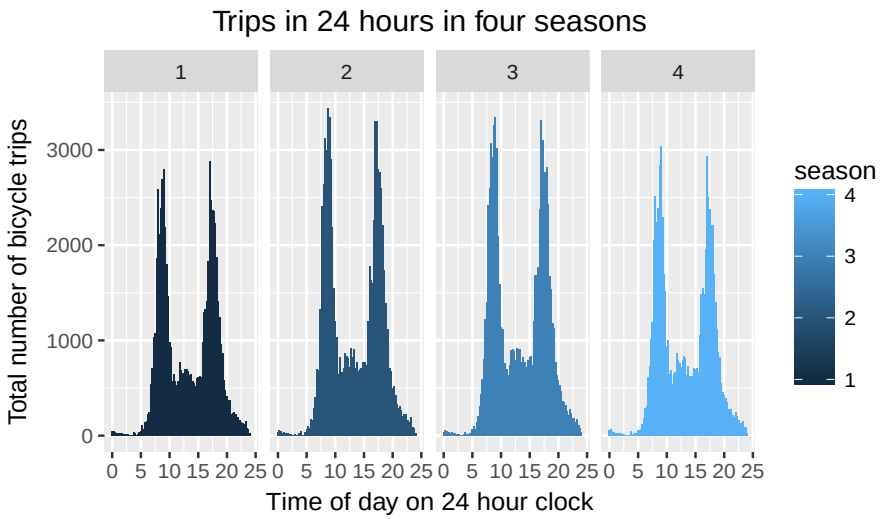

图 5: 不同季节一天 24 小时自行车的使用

与 shiny 将地图与 Shiny App 与地图相结合，再加上 ggplot2 在 App 中进行作图，最终的效果还是比较令人满意的。

总之，本次大作业充分利用了在课堂上学习的数据处理和可视化的相关知识，并且自己查阅了不少资料，对 Shiny App 和地图的使用进行了尝试，对整个学期的课程做了较为圆满的总结。最后一定要向俞声老师和助教一学期的努力与付出表示感谢!

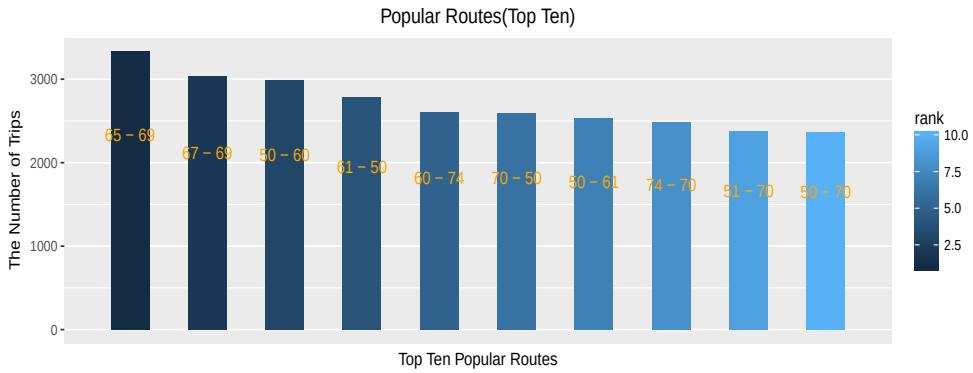

图 6: 骑行人次最多的前 10 条路线

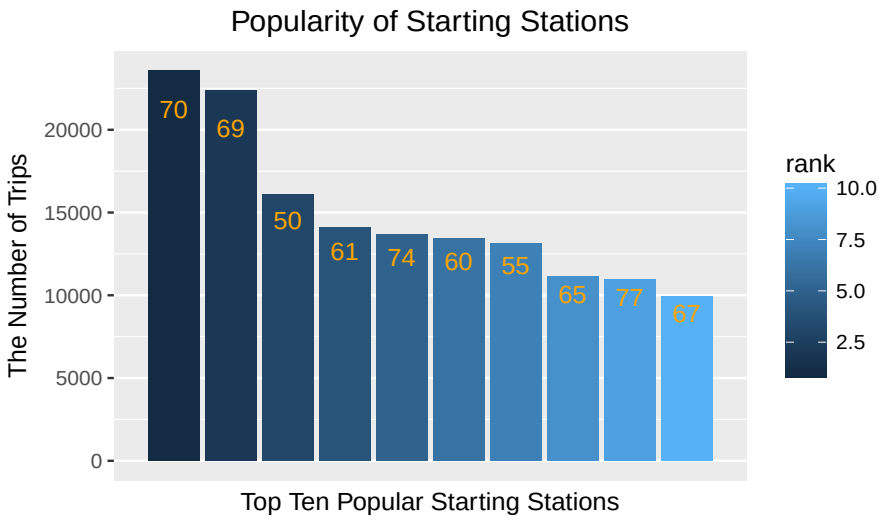

图 7: 出发人次最多的前 10 个站点

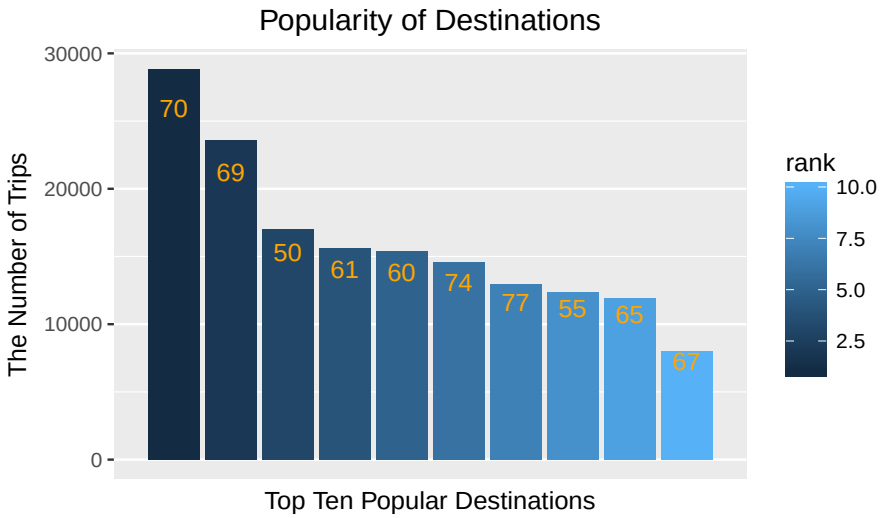

图 8: 到达人次最多的前 10 个站点

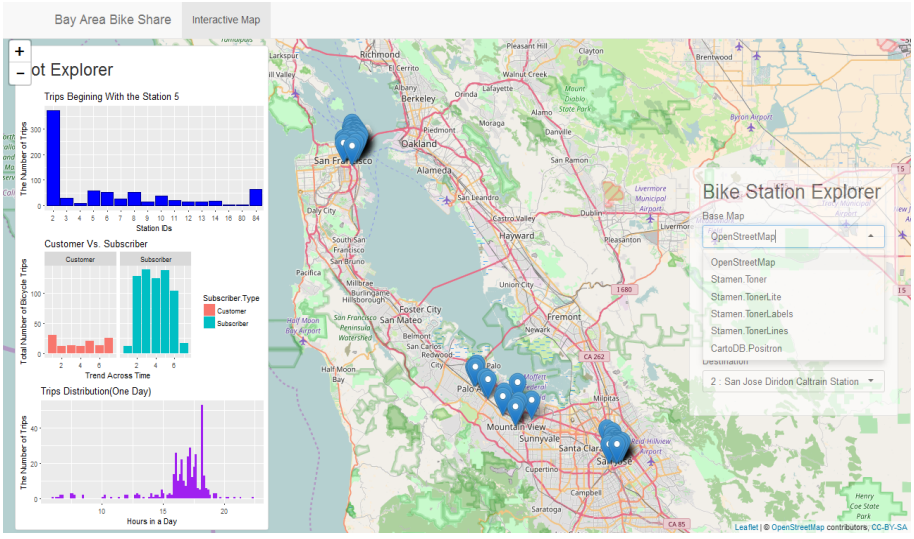

图 9: Shiny App 主界面

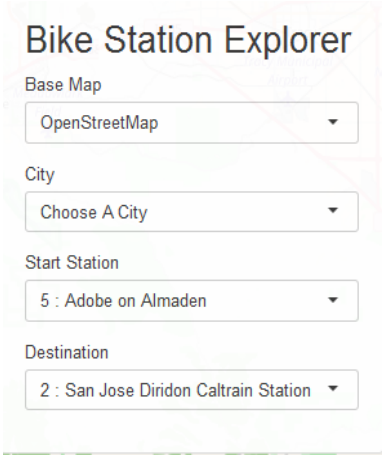

图 10: Shiny App 输入面板

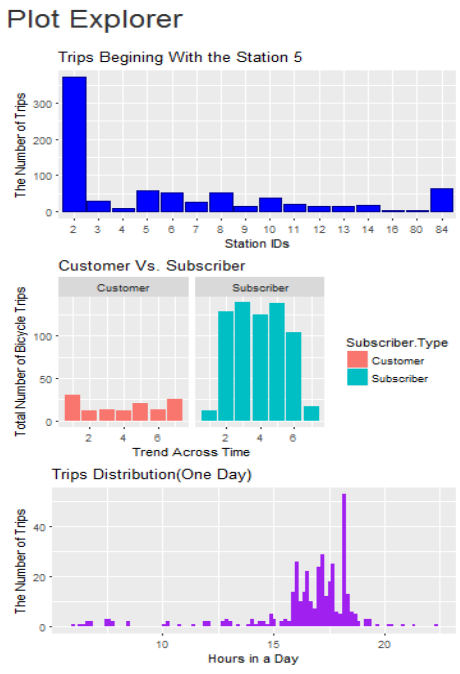

图 11: Shiny App 作图输出面板

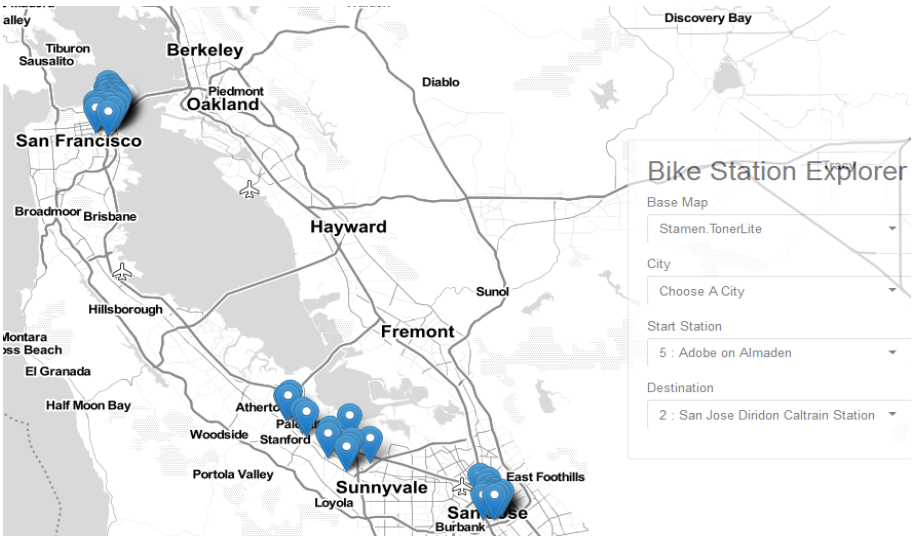

图 12: 选择不同的地图

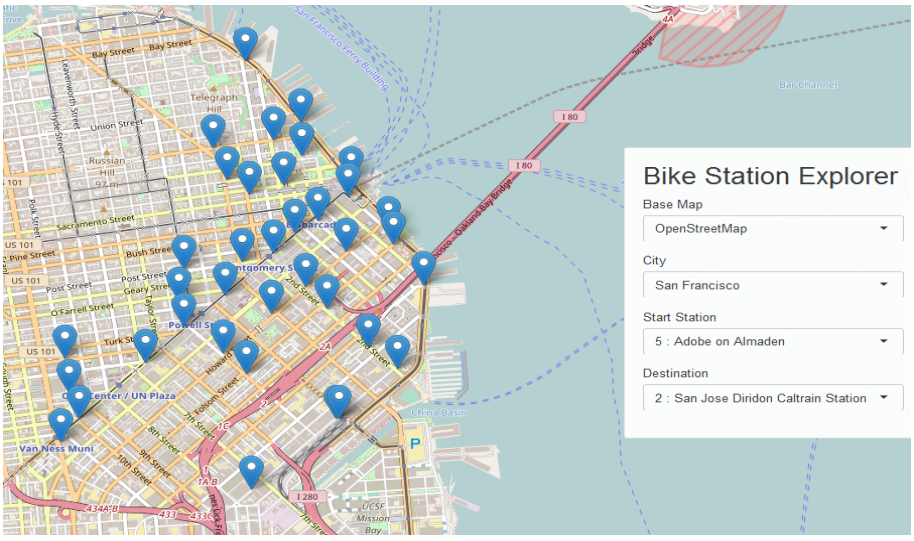

图 13: 放大所选择的城市地图

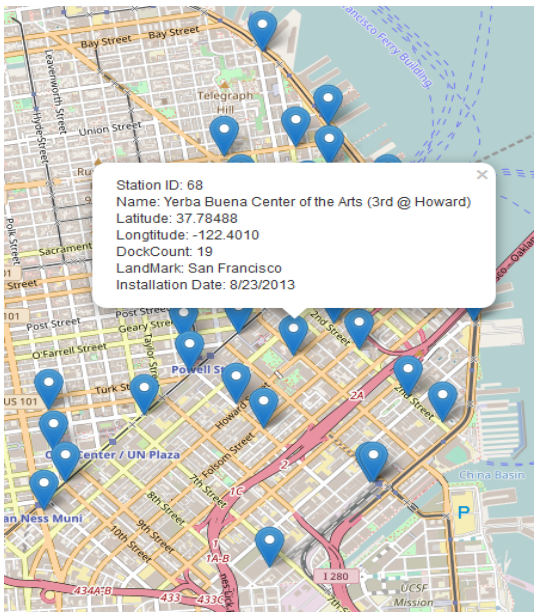

图 14: 显示车站详细信息

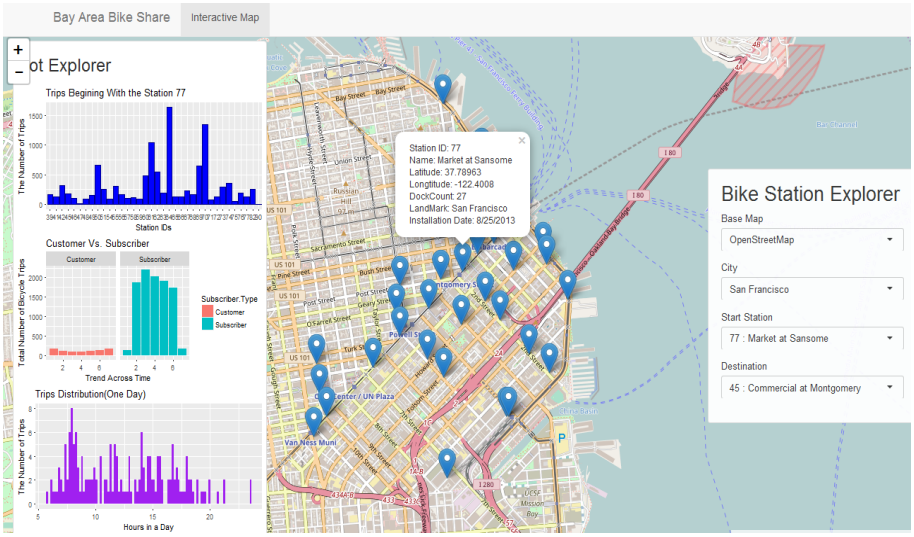

图 15: 选择起始站点和终点之后的界面

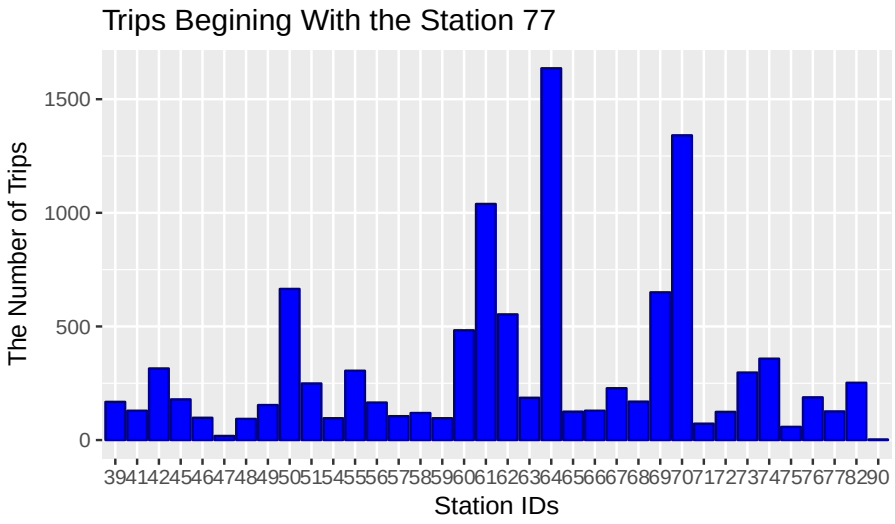

图 16: 查询结果 2

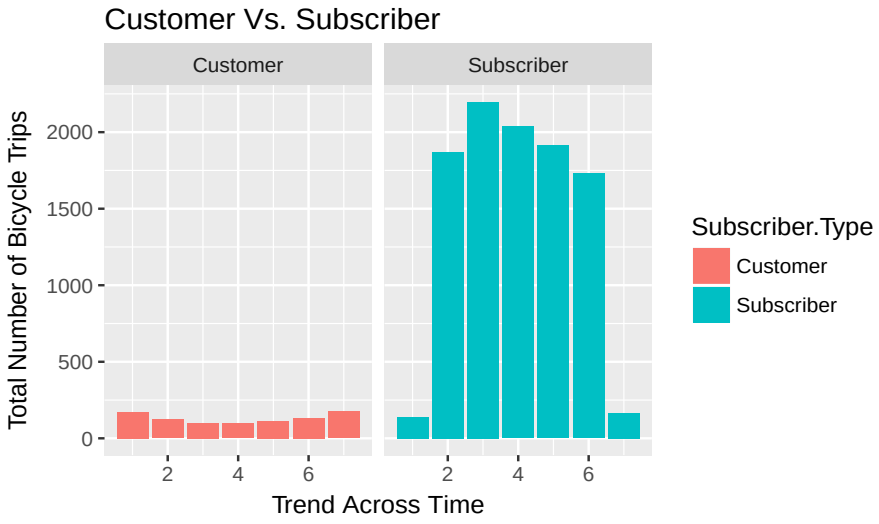

图 17: 查询结果 1

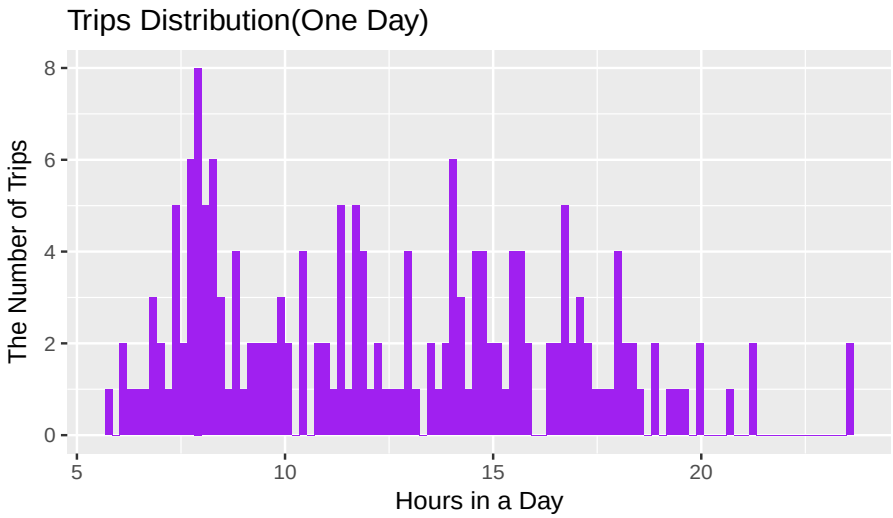
Trips Distribution(One Day)
The Number of Trips
0
2
4
6
8
5
10
15
20
Hours in a Day

图 18: 查询结果 2