基于 Tableau 的摩拜单车数据可视化

DAND 进阶课 P4 项目

何伟健 2019.5.:6

项目概述

本项目基于 2016 年摩拜共享单车在上海区的用户使用数据，使用 Tableau 对数据进行可视化，目的是通过一系列图表发现并挖掘一些有价值的规律和现象。

数据介绍

本项目的研究对象为 2016 年摩拜共享单车上海区中随机抽样的 100 万条的用户使用数据，该数据集来源于上海 SODA 比赛[1]，包含如下内容：

字段名	说明	数据类型
Orderid	订单编号	数值 – 连续型
Bikeid	车辆编号	数值 – 连续型
Userid	用户编号	数值 – 连续型
Start_time	订单开始时间	日期和时间值
Start_location_x	起始位置经度	数值 – 连续型
Start_location_y	起始位置维度	数值 – 连续型
End_time	订单结束时间	日期和时间值
End_location_x	终点位置经度	数值 – 连续型
End_location_y	终点位置纬度	数值 – 连续型
track	骑行轨迹点	文本值

作品链接

版本 1：
https://public.tableau.com/profile/kelvin.he6404#!/vizhome/Mobike_Shanghai_Data_Tableau_V1/Story

版本 2：

https://public.tableau.com/profile/kelvin.he6404#!/vizhome/Mobike_Shanghai_Data_Tableau_V2/Story

项目总结

在本项目中我所创建的可视化，主要想展示在上海地区摩拜单车用户使用特点及单车的利用情况，可视化图表涵盖了单车在不同时段的使用特点、骑行的时长分布及直线距离统计、单车被使用的频率、单车使用区域分布等。从图中我主要发现了用户一般骑行 7min 左右的时间以及不超过 1 公里的距离；总体上周三使用单车的时间最多；单车使用的频率一般是 1 个月 1-2 次；单车主要在市区内使用为主等规律。

项目设计

1. 订单的趋势和分布（Page 1 和 Page 2）

最开始我希望先了解各订单随着时间的变化趋势，以及在不同周期的分布。我选择折线来展现订单在 1 个月内每天的趋势，因为折线图适合表现时间序列数据。同时我以用户使用次数 10 次作为分界点来分组观察，并用蓝色和橙色区分，希望能看到不同的用户群体对订单增长的影响。

接下来我选择了按周（7 天）和按天（24 小时）这两个周期分别观察订单分布，以此来研究用户使用单车的行为。这里我采用直方图的形式，可以方便地对不同时间段的订单量进行对比。接下来我再把这两个周期结合起来一起观察，把每天 24 小时的订单分布拆分成一周 7 天进行纵向对比。

结论：从“ 单车 8 月份订单变化趋势图 ”可以看出，订单量呈现明显的增长趋势，而且对于使用次数 10 次以内的用户群，订单的增长幅度更大。从订单时间分布的几个图可以看出，周一至周三订单量逐步增长至峰值，同时在工作日有两个单车使用的峰值时段，分别是早上 8 点和下午 6 点。而在周末只有下午 6 点左右单车使用量稍微多一点。

2. 单车使用时长及骑行直线距离（Page 3）

接下来我想观察每个订单的车辆使用情况，我将分别从每次骑行的时长，以及骑行的起点和终点直线距离长短来分析。首先初始数据集的字段中并没有骑行的时长，因此我通过 end_time – start_time 来生成一个计算字段，命名为 using time，单位为分钟数。我仍以直方图的形式绘制单车的使用时间分布图，同时加入时间分组（从 0 点开始，以 6 个小时为一组）以供对比。另外，由于数据集中可能存在一些异常值，出线了一些几千分钟的值，使图形呈现出极端的长尾分布。因此我把横坐标的上限调整为 50（数量已经很少了），这样图形展现得更合理和美观了。

对于骑行距离，原始数据集也是不包含这个字段的。数据集可提供的有起点和终点的经纬度，以及路程当中一些的点的经纬度。由于路程中的点在数据集中的排列并非有序的，难以还原其真正的骑行路线。因此我采用起点和终点间的直线距离作为一个参考距离来分析，从网上找到了计算方法，并在 excel 表中计算。从直线距离的统计分布来看用饼状图来展示比较合适，同时我将直线距离划分成 6 个组来显示。

两个图我采用的是上下并列来对比查看。

结论：骑车时间在 7 分钟左右的情况是最多的，而且很少有超过 40 分钟的。而骑行的起点到终点的直线距离在 1 公里以内的情况占了 2/3，而且几乎没有超过 5 公里的情况。可见用户对共享单车的使用都是以短途代步为主，解决了最后的 1 公里。

3. 用户骑行频率/单车使用频率 （Page 4）

从使用频率的角度来继续深入研究用户的行为和共享单车的利用率。首先我希望了解到每个用户在一个月内使用共享单车的次数，这里我取字段 userid，计算每个 userid 的出现的次数，再进行统计。同样我也想知道每辆单车在一个月内被使用的次数，取字段 bikeid 并计算出现的次数。我将采用直方图展示这两项统计，上下并列显示。

结论：在一个月内大部分用户使用 5 次单车，很少有用户会使用超过 15 次，也就是说共享单车并没有成为大众的日常交通工具，只有在某些情况偶尔用一下。而对于单车的使用，一辆单车基本上只被使用 1-2 次，利用率非常低，也一定程度反映出供给大于需求。

4. 单车使用的区域分布 （Page 5）

最后我从单车使用的位置分布来了解单车在哪些地方使用得更频繁，获取这个信息有助于摩拜公司更合理地安放单车，减少资源浪费。从字段中我选择了起始点的经度和维度字段，采用的是地图的形式显示每个点在上海市各个位置的分布情况。我还添加了两个过滤器可以互动使用，第一项是时间的分组，可以分别查看在 4 个时段中的订单使用分布，用颜色来区分。第二项是按周的分组，用户可分别查看一周 7 天的订单使用分布。考虑到显示点的很多，我将点的尺寸适当缩小，这样看得相对舒服一些。

结论：从图上可以看出大部分的订单集中在黄浦江以北，以市中心的使用为主。

项目反馈

版本 V1：

反馈 1：

1. 第一页 weekday 的 1-7 最好改为星期一到星期天或者英文显示，否则可能有误解，有些地方的 1 表示星期天，并非星期一。
2. 第一页“ 订单每天的分布情况 ”纵坐标刻度线相隔太远，建议增加次级刻度线更容易观察具体数值。
3. 第四页“ 每辆单车使用次数统计图 ”更改图标题为“ Number of Bike Use ”

反馈 2：

1. 在可视化中的主要收获

最大的感觉是共享单车利用率不高，造成一定程度的资源浪费。还有就是哪些时段的单车使用最频繁。

2. 对数据的疑问

骑行距离的统计以直线距离作为参考，感觉不是很准确。实际的骑行距离比直线距离可能大得多，容易造成误导。

参考

[1] 上海 SODA 比赛：http://shanghai.sodachallenges.com/

[2] 通过经纬度计算两点间的直线距离：
https://community.tableau.com/thread/216882