**《交易数据统计报告》**

2017年6月1日星期四

北京邮电大学

# 需求说明

## 1.1异常数据分析

* + 1. 检查订单数和支付数是否一致
    2. 检查购票数和取票数是否一致
    3. 检查站点编号和站点名称是否一致

## 1.2广州地铁购票数据分析

1. 以天为单位统计机场南，广州东站进站人数变化规律
2. 以天为单位统计各大机场，火车站为终点的客流统计
3. 统计每个周六日（节假日做调整），以机场南，广州东站为起点的乘客终点分布情况
4. 统计每个周六日（节假日做调整），以各大机场，火车站为终点的乘客起点分布情况
5. 统计热门站点（热门起点站和热门终点站，给出统计指标），可以暂定为：一段时间内进站（或出站）乘客数目大于所有统计到的进站（出站）乘客均值
6. 分析各大站点每日进站高峰期（出站无法统计），只能统计进站客流，可以先从热门进站站点开始
7. 热点线路统计：以天为单位（以周为单位），统计起点站和终点站都相同的购票数量，给出排名
8. 统计购票数量前五（或者前十）的站每日购票乘客中不同购买方式的占比，而且给出每个站点的走势情况
9. 以站点为单位统计购票订单和出票转化率
10. 分析工作日上班高峰期（上午7点到9点），下班高峰期（下午17点到19点）每条路段的客流及拥堵情况。
11. 根据每个乘客历史购票记录分析并展示该用户的出行轨迹
12. 定义乘客画像，并将有相同习惯的乘客分类

## 1.3长沙充值记录分析

1. 长沙充值用户中，每日充值金额，支付金额，转化订单比例，转化金额比例
2. 长沙充值用户的用卡类型比例（这个注意是按照用户ID区分，不是单纯的订单中某一种卡所占的比例）
3. 按照充值频率和充值数额分别对设备ID排序

# 使用数据

1. 3月1日~3月31日广州地铁单程票购票记录
2. 2016年11月10日~2017年4月17日长沙充值记录

# 方法说明

## 3.1异常数据分析方法

* 1. 对比购票数量和实际取票数量，付款订单数量和未付款订单数量；
  2. 对比站点名称和站点编号，确认站点名称和编号是否一一对应；将付款乘客作为下一步分析对象

## 3.2广州地铁购票分析方法

1. 找到所有和机场，火车站相关的地铁站；获取以这些站为进站的购票记录；以一天为单位进行整合，并统计每天的购票记录总数
2. 找到所有和机场，火车站相关的地铁站；统计这些站以这些站为终点站的购票记录；以一天为单位进行整合，并统计每天的购票记录总数
3. 取出机场南、广州东站在周末的购票记录；对每一天的记录按照终点站分类，统计每一类个数；
4. 找到所有和机场、火车站相关的地铁站；取出每个站点周末购票记录；对每个站点每一天的终点站进行分类，统计每一类个数
5. 一个月里计算购票量超过均值的站点为热门站点，并给出
6. 将热门站点每天固定时间段的数据加和起来，形成一个1小时为单位一个月总购票量
7. 按照日期、起点站、终点站对购票记录进行三层分类；每一类统计购票数量
8. 获取每一天的购票记录；统计每天购票记录中不同支付方式的购票数量；按照站点对数据分类，并统计每个站点每天不同购票方式的走势
9. 以天为单位统计所有订单数量；以天为单位，统计所有订单中取票数量和购票数量相同的订单；计算每天付款订单比例
10. 从购票记录中筛选出某一工作日上班高峰期和下班高峰期的购票记录，按照起始站和终点站进行归类并计算票数。利用Logit算法对每个站点的进站客流计算路线选择概率，并根据概率值分配客流，计算每段线路在上下班高峰期的客流情况。将客流情况展示在广州地铁线路网中。
11. 从购票记录中获取一个乘客每天的购票记录，并且将一天的出行路线在轨道交通图上展示出来
12. 定义用户画像，根据用户画像对不同乘客进行聚类，找到出行规律相似的乘客。将每一类乘客每天的出行活动在轨道交通图上展示出来，并对乘客的群体特征进行分析

## 3.2长沙充值记录分析方法

1. 以天为单位计算长沙充值用户中，充值金额，支付金额，转化订单比例，转化金额比例等
2. 统计每个设备ID（不是订单ID）的充值记录中涉及的卡类型，并进行汇总
3. 对每个设备ID的充值频率，充值金额进行倒序排序

# 展示结果及分析

## 4.1异常分析结果展示

1. 统计订单数和取票数；

下表是订单，取票统计信息：

|  |  |  |  |
| --- | --- | --- | --- |
| 订单未支付数量 | 订单总数 | 支付订单的取票数 | 总票数 |
| 68459 | 3448294 | 4173872 | 4257227 |

1. 检查地铁站和地铁站编号是否一一对应；

下表是站点编号和站点名称数量：

|  |  |  |  |
| --- | --- | --- | --- |
| 起始站站点名称数量 | 起始站站点编号数量 | 终点站站点名称数量 | 终点站站点编号数量 |
| 137 | 154 | 159 | 177 |

从上表可以看出站点和其编号并不是一一对应的，如下站点对应两个站点编号：

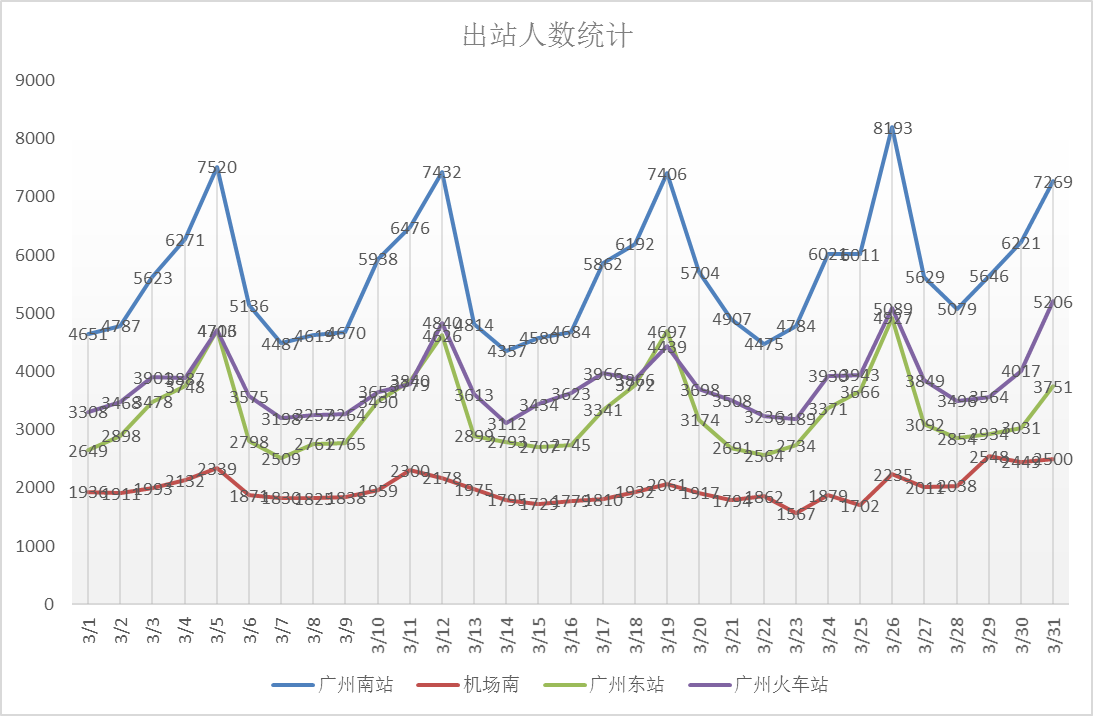
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 站点名称 | 广州火车站 | 客村 | 体育西路 | 大学城南 | 广州南站 |
| 对应编号 | [506 215] | [306 821] | [303 114] | [709 411] | [249 701] |
| 站点名称 | 广州东站 | 天河客运站 | 海珠广场 | 黄沙 | 汉溪长隆 |
| 对应编号 | [116 301] | [314 621] | [211 610] | [105 607] | [311 705] |
| 站点名称 | 坦尾 | 公园前 | 东山口 | 区庄 | 石壁 |
| 对应编号 | [502 605] | [109 212] | [112 614] | [615 509] | [248 702] |
| 站点名称 | 杨箕 | 珠江新城 | 西朗 |  |  |
| 对应编号 | [511 113] | [304 513] | [6014 101] |  |  |

其中不存在一个编号对应两个站点的情况

## 4.2广州地铁购票记录分析结果

1. 以天为单位计算机场南，广州东站进站人数，并展示其变化情况

从图上可以看出，广州东站具有“周末高平时低”的特点，这是传统的出行特点，但是并没有很强的周期规律。而机场南则比较平稳，没有周末等出行高峰造成的客流增长。原因可能有两点：1.机场南的客流本身就不具有周期高峰这样的性质，毕竟飞机成本要比火车成本高出很多；2. 也有可能是两个站点乘客的购票习惯不同，火车站出来的乘客转化为地铁乘客的概率要比出机场转化为地铁乘客的概率，即火车站的地铁乘客转化率要高于飞机场

1. 以天为单位计算以各大机场，火车站为终点的客流，并展示其变化情况

这张图就能看出三个火车站有比进站数量更强的周期性，例如广州南站的图中，最高峰的日期一致相差7天，同样其他时候的客流也存在很强的以7天为单位的周期性。这就说明广州本地居民出行的规律更符合周期性，也就是传统意义上理解的客流存在的周期性可能对常住居民，本地居民更加适用。因为这组数据较上一个excel来说最大的区别就是：这次涉及的乘客很大一部分是本地居民，而上一个excel涉及的乘客更多的是外来人口

1. 统计每个周六日（节假日做调整），以机场南，广州东站为起点的乘客终点分布情况

25日是周六，周六从机场南出发到广州各站的乘客分布情况图，可以明显看出人和，广州南站，体育西路，广州东站的人数总和占比较高，这说明这几站是下飞机的乘客选择概率较高的站点，像是广州东站和广州南，则是进行换乘火车，如果是人和和体育西路则需要根据这两地的商业或者是用地情况对乘客的目的地进行初步的判断

广州东站周末出站的乘客目的地分布就比机场南乘客分布稍微平均一些，尽管体育中心，体育西路，广州南站是他们选择率较高的车站，但是后续的像人和，公园前，烈士陵园等分布较为平均，而且数量也不少。从这里不难发现，最高的几个站点中，除了交通站点外，体育西路这个站是机场乘客和广州东站乘客共同偏好的站点，可以研究该站点的地理位置含义，以及它吸引乘客的原因。

下面是周末整体的目的地分布情况：

1. 统计热门站点（热门起点站和热门终点站，给出统计指标）

“热门”定义为：一段时间内进站（或出站）乘客数目大于所有统计到的进站（出站）乘客均值

一月购票数量超过均值的站点包括：广州南站、广州塔、体育西路、广州火车站、客村、广州东站、公园前、人和、岗顶

其中广州南站每天购票数量走势图如下：

从走势图上看并没有特别明显的周期性趋势，但是四个周末5日、12日、19日、26日为四个明显的极值点，这四个点在其附近一周的购票量中是最高的，所以可以判断一周末的购票数量一般是本周的高峰时期。

相比于广州南站，广州塔购票量就显得平缓很多。尽管周末时段（4日、12日、19日、26日）仍然是一周的极值点，但是整体的波动性，特别是在13日到21日一周的波动性相对平缓。

1. 分析各大站点每日进站高峰期（出站无法统计），重点针对热门进站客流

如果按照24小时划分，将一个月里固定一小时的购票数量加和起来，就形成了如下折线图，该折线图是广州南站以1小时为单位的整月购票量统计图：

从图中可以看出从整个月的角度来看，广州南站的客流高峰期处于中午11点至下午6点之间，这可能和进入广州南站的火车时间，乘客作息时间有关。而且从11点之后没有特别明显的突变点，说明客流在11点之后比较稳定，而且数量较大。

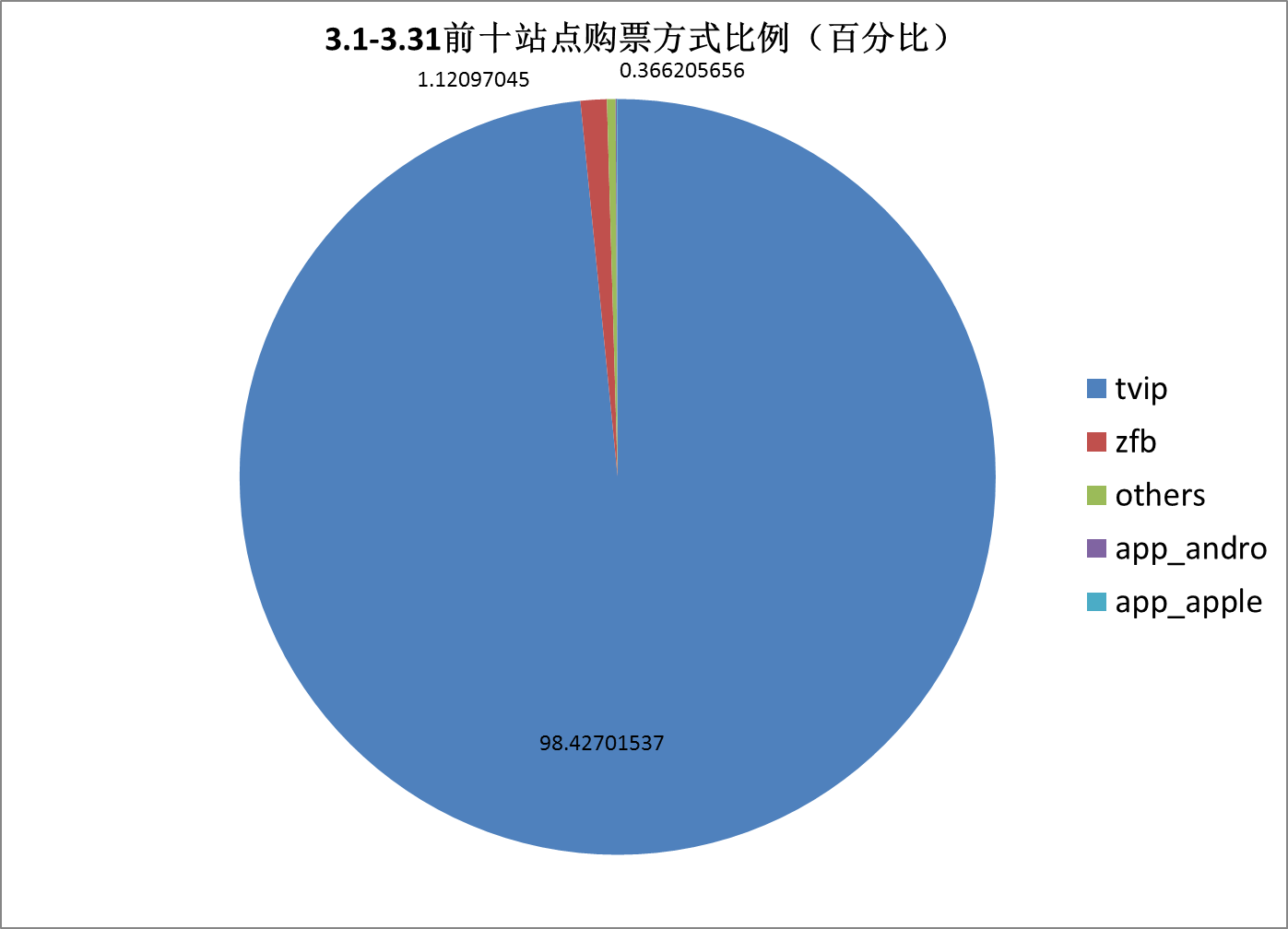
从整个月的角度看，广州塔一天的购票高峰期就比广州南站明显很多，从图上可以看出，广州塔很明显是晚上9点左右达到进站高峰，说明广州塔地铁站的乘客多为晚上返程的乘客。

1. 热点线路统计：以天为单位（以周为单位），统计起点站和终点站都相同的购票数量，给出排名

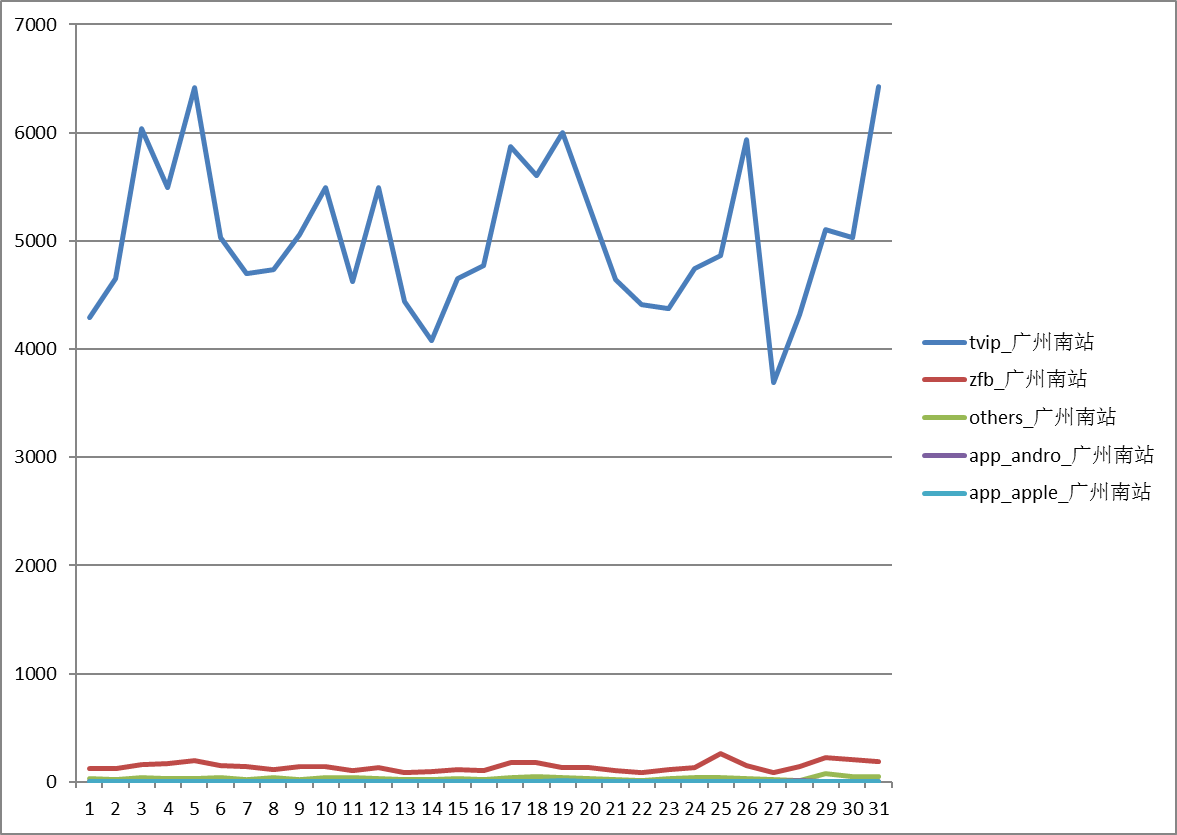
下表是广州3月14日购票前十的热门线路

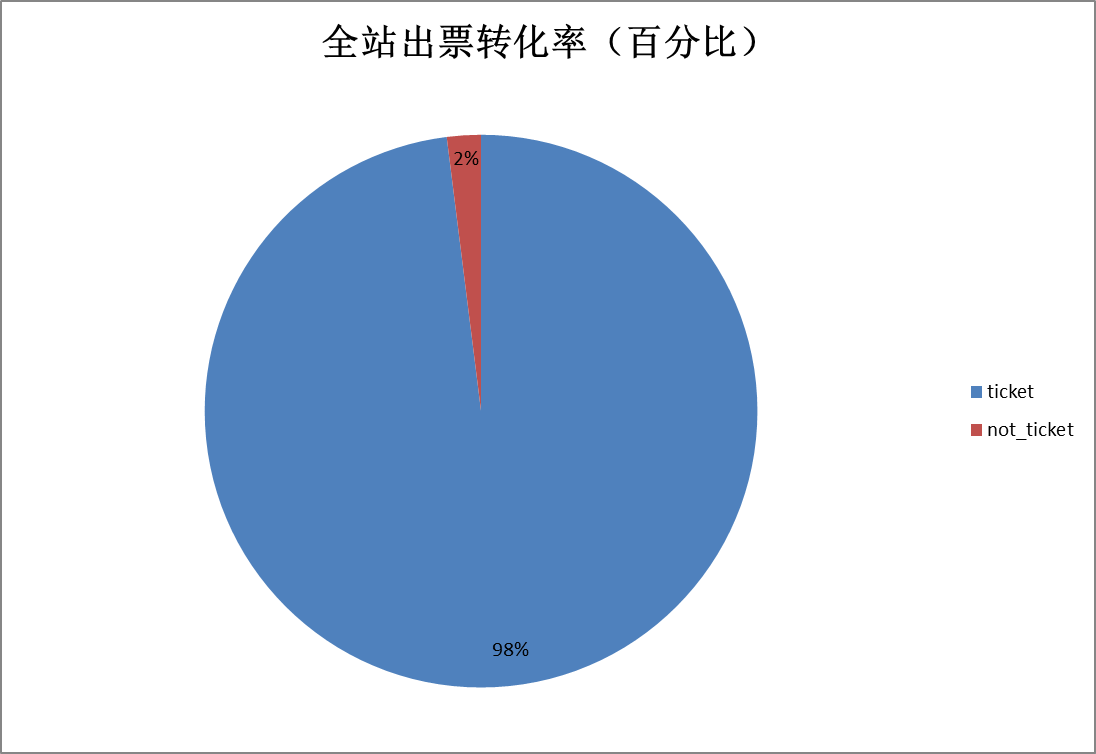
|  |  |  |
| --- | --- | --- |
| **广州南站** | **广州火车站** | 295 |
| **广州火车站** | **广州南站** | 286 |
| **广州南站** | **嘉禾望岗** | 182 |
| **大石** | **市桥** | 161 |
| **岗顶** | **天河客运站** | 150 |
| **广州火车站** | **三元里** | 148 |
| **五山** | **岗顶** | 146 |
| **广州南站** | **广州东站** | 141 |
| **广州塔** | **广州南站** | 141 |
| **体育西路** | 140 |

总的来说并没有很强的偏向性，单路线最多的购票数量占全天的1%左右，线路整体购票数量分布算是比较平稳的。从表中很明显看出，广州南站和广州火车站的来往客流的数量比较多，尽管在全天的购票量中占比不大，但是也能很明显看出有很大一部分乘客是将广州作为换乘城市的乘客。除此之外，广州南站到嘉禾望岗之间，大石到市桥等路线也能算作是乘客乘坐较多的路线。

1. 统计购票数量前五（或者前十）的站每日购票乘客中不同购买方式的占比，而且给出每个站点的走势情况

图上明显可以看出，tvip（不知道是什么设备）的购票数量远超其他所有购票类型。



1. 以站点为单位统计购票订单和出票转化率

全网的订单票数与实体票的转化率大约在98%左右，最低的几个站如下：

|  |  |  |  |
| --- | --- | --- | --- |
| **start\_station\_name** | **ticket\_count\_pick\_1** | **ticket\_count\_pick\_2** | **rate** |
| **万胜围** | 5172 | 4877 | 0.942962 |
| **谢村** | 561 | 538 | 0.959002 |
| **广州东站** | 73798 | 70826 | 0.959728 |
| **嘉禾望岗** | 50820 | 48959 | 0.963381 |
| **蕉门** | 1987 | 1917 | 0.964771 |

转化率作为乘客购买之后不退票的一种检验手段，可以看出在线购票的乘客大部分还是愿意再花一点时间取票的，转化率较低的站点可能机器数量较少导致乘客会出现购票之后反而不愿取票而选择退票的情况。

1. 展示某一乘客在地铁线路中的运行情况

下图展示的是一个乘客从广州赤岗站到万胜围站的运动情况



利用百度地图接口，并根据某一乘客购票记录的起点站和终点站画出其在地铁线路中的运行轨迹。该轨迹以实时动态展示形式，直观的反映出一个乘客比较真实的运动情况。

1. 分别展示每个站点工作日期间，周末以及三月份每小时购票数量

由于站点数量过多，下文将针对两个站点“东圃”和“广州火车站”进行比较详细的分析，其余站点的数据可以参照附录，附录中计算了所有站点三月份每天以小时为单位的购票数量，并画出了热门的10个站（**广州南站，广州塔，体育西路，广州火车站，客村，广州东站，公园前，任何，岗顶，珠江新城**）的工作日，双休日，一个月以小时为单位的购票趋势图。

下图为东圃站工作日期间（3月20日~3月24日）每天以小时为单位客流变化趋势：

可以看出再工作日期间每天早中晚分别有一个时间段可能出现客流高峰，其中早晚高峰情况比较明显，上午的高峰期推测与居民上班有关，晚上的进站高峰期与下班或者居民出行吃饭有关，所以可以判断该站周边以居民区或者工作区为主。

下图为东圃站工作周末（3月2日~3月26日）每天以小时为单位客流变化趋势：

周末的客流特征与工作日有很大区别，首先是高峰时段向中午靠拢，可以看出部分居民再周末的出行时间偏向中午或者是下午。

下图为东圃站3月份（3月1日~3月31日）每天以小时为单位客流变化趋势：

下图为广州火车站工作日期间（3月20日~3月24日）每天以小时为单位客流变化趋势：

从购票总量来说，广州火车站的购票量要远远大于东圃，波形比较平缓，而且购票时间更多趋近于下午时段，没有一般工作日的早晚高峰特点。

下图为广州火车站周末（3月25日~3月26日）每天以小时为单位客流变化趋势：

广州火车站周末的客流特点和工作日区别不大，可见乘坐火车的地铁乘客不存在特别明显的高峰时间，推测与广州火车站的火车到站时间有关。

下图为广州火车站3月份（3月1日~3月31日）每天以小时为单位客流变化趋势：

1. 计算3月30日上午上班高峰期（7点~9点）每段线路上乘客数量，并通过列表和地铁图的方式展示。乘客数量指的是该线路上所有经过的乘客的数量，不论乘客的起点和终点是否在路段的两段，而且计算结果区分上下行方向。

---拥堵级别的划分和地图上颜色的展示

目前的拥堵级别划分标准是从客流最小值到最大值区间等比例划分五个级别，级别的划分也可以根据需求进行调整；

---各个级别线路拥堵情况和颜色对照表

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 拥堵等级 | 0 | 1 | 2 | 3 | 4 |
| 颜色 | 绿色 | 蓝色 | 橙色 | 红色 | 紫色 |

---上午上班高峰期线路最拥挤路段排名及乘客数

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 站点1 | 站点2 | 线路名称 | 客流数量 |
| 1 | 员村 | 潭村 | Line5 | 1538 |
| 2 | 潭村 | 猎德 | Line5 | 1532 |
| 3 | 磨碟沙 | 新港东 | Line8 | 1532 |
| 4 | 猎德 | 珠江新城 | Line5 | 1511 |
| 5 | 广州塔 | 珠江新城 | Line3 | 1503 |
| 6 | 赤岗 | 磨碟沙 | Line8 | 1483 |
| 7 | 梅花园 | 燕塘 | Line3b | 1419 |
| 8 | 科韵路 | 员村 | Line5 | 1415 |
| 9 | 三元里 | 广州火车站 | Line2 | 1390 |
| 10 | 大塘 | 客村 | Line3 | 1381 |
| 11 | 京溪南方医院 | 梅花园 | Line3b | 1379 |
| 12 | 广州东站 | 林和西 | Line3b | 1348 |
| 13 | 越秀公园 | 纪念堂 | Line2 | 1319 |
| 14 | 江南西 | 昌岗 | Line2 | 1309 |
| 15 | 市二宫 | 江南西 | Line2 | 1299 |

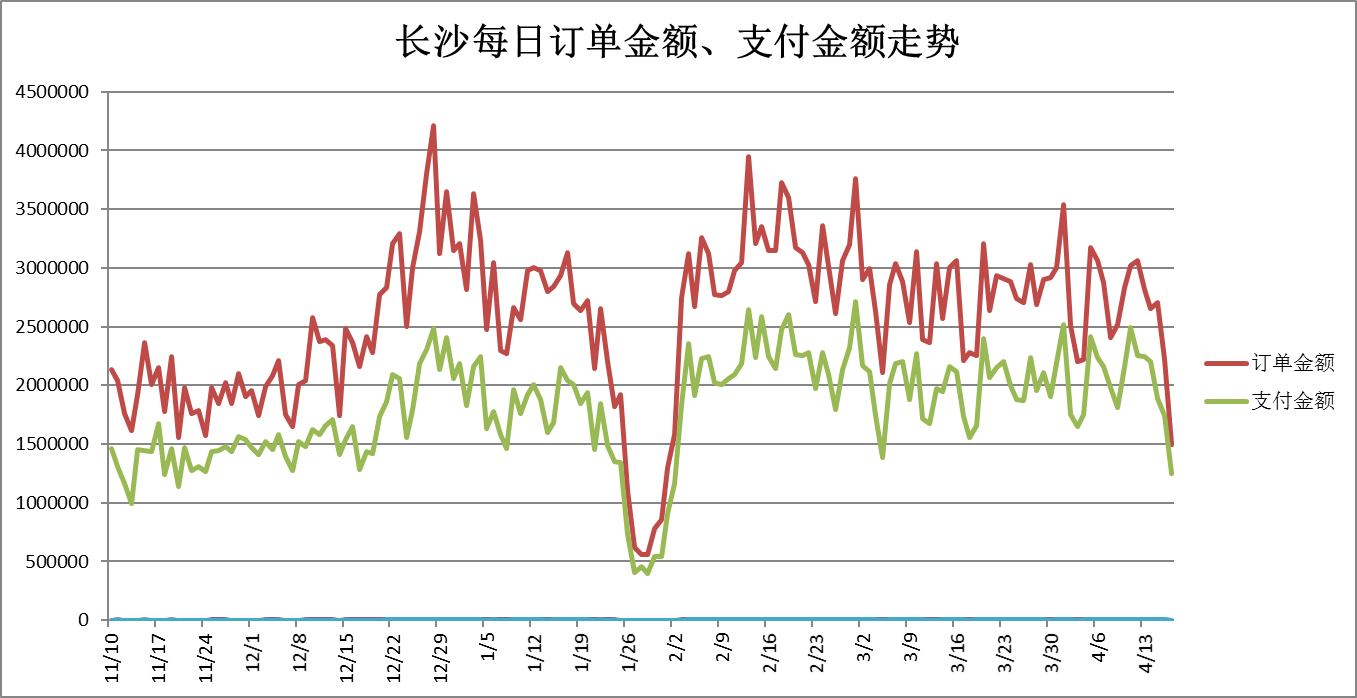
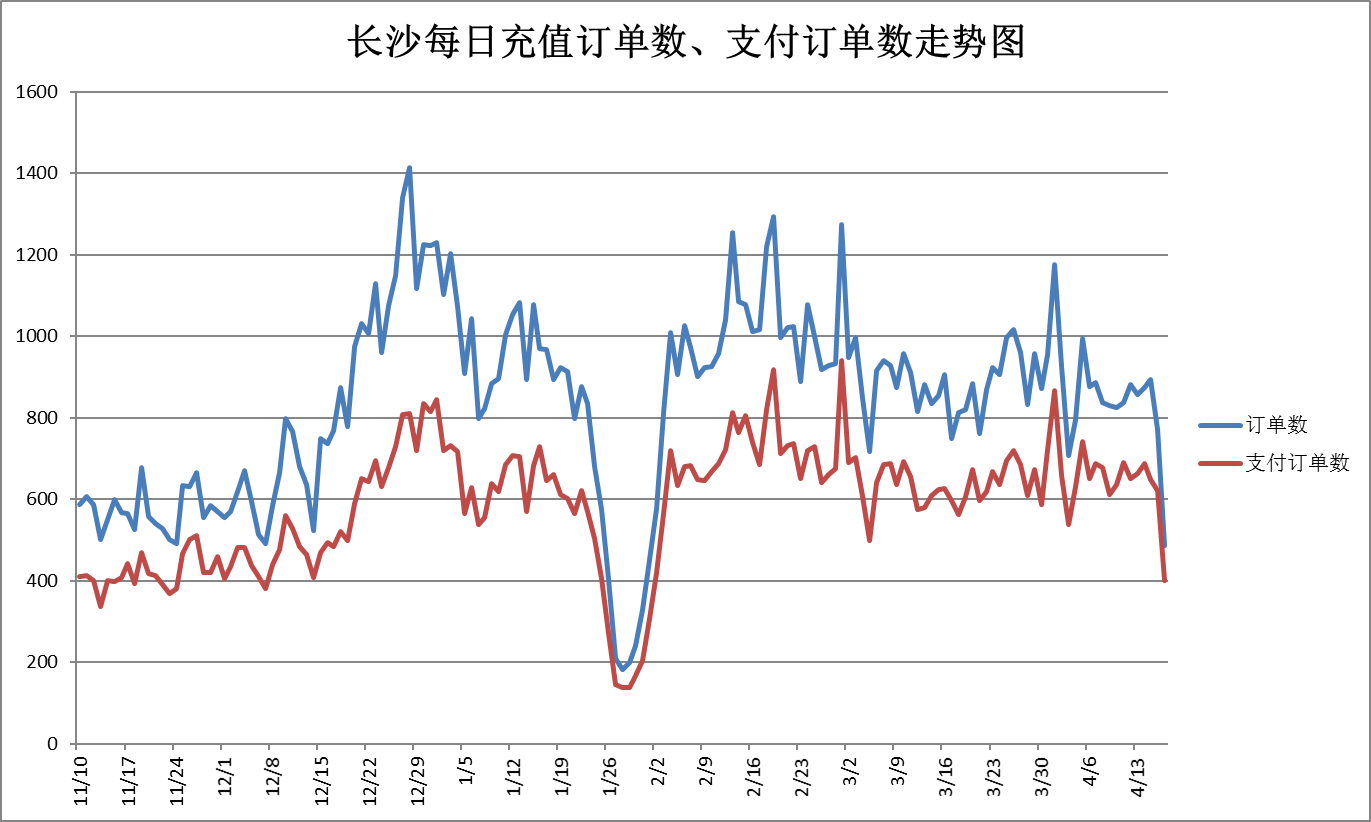
从图上可以看出，相比于其他路段，公园前站到越秀公园站路段在上午7点到9点期间的拥堵等级最高，属于最拥堵路段，其中纪念堂到公园前路段在广州全网路段中拥挤程度排13名；珠海广场站以南路段处于次拥堵路段；长寿路站以南路段相对最为空闲；除此以外其余路段拥堵程度均为正常。

1. 计算3月30日下午下班高峰期（17点~19点）每段线路上乘客数量，并通过列表和地铁图的方式展示乘客数量指的是该线路上所有经过的乘客的数量，不论乘客的起点和终点是否在路段的两段，而且计算结果区分上下行方向。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 站点1 | 站点2 | 线路名称 | 客流数量 |
| 1 | 广州塔 | 珠江新城 | Line3 | 2616 |
| 2 | 广州塔 | 客村 | Line3 | 2257 |
| 3 | 市二宫 | 海珠广场 | Line2 | 1961 |
| 4 | 江南西 | 市二宫 | Line2 | 1943 |
| 5 | 越秀公园 | 广州火车站 | Line2 | 1931 |
| 6 | 纪念堂 | 越秀公园 | Line2 | 1900 |
| 7 | 市二宫 | 江南西 | Line2 | 1771 |
| 8 | 大塘 | 沥滘 | Line3 | 1743 |
| 9 | 沥滘 | 厦滘 | Line3 | 1695 |
| 10 | 三元里 | 飞翔公园 | Line2 | 1693 |
| 11 | 江南西 | 昌岗 | Line2 | 1684 |
| 12 | 大塘 | 客村 | Line3 | 1658 |
| 13 | 梅花园 | 京溪南方医院 | Line3b | 1609 |
| 14 | 五羊邨 | 珠江新城 | Line5 | 1607 |
| 15 | 飞翔公园 | 白云公园 | Line2 | 1596 |

图上可以看出越秀公园站到纪念堂站的客流在17点到19点之间比较拥堵，属于比较拥堵路段，而且在广州全网路段中拥挤程度排第6名；公园前和珠海广场路段最为空闲；其余所有路段拥堵程度均为正常。

## 4.3长沙充值记录分析结果

1. 长沙充值用户中，统计每日充值金额，支付金额，转化订单比例，转化金额比例

支付总金额和订单总金额的变化幅度大致相同，总订单数和支付订单数的趋势也大致相同，只是他们之间都存在一个稳定的间隔，说明在充值用户中，并不是绝大部分用户下单后确认支付，下单和真正充值还是有所区别的。

同时还可以看出12月15日附近有大幅度上升，过年期间有大幅度下降，过年会降低乘客的出行频率。

1. 每个设备ID的充值记录中涉及的卡类型，统计结果如下

长沙充值用户的用卡类型比例（这个注意是按照用户ID区分，不是单纯的订单中某一种卡所占的比例）

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| IOS\_NFC全卡充值 | 长沙手环(IOS) | 长沙地铁充值 | 长沙地铁实体卡 | 长沙手环(安卓) |
| 3532 | 21 | 17877 | 2776 | 34 |

被充值的有五类产品，其中IOS\_NFC全卡充值、长沙地铁充值、长沙地铁实体卡充值占主导地位，其他两类也存在充值记录

1. 按照充值频率和充值数额分别对设备ID排序（单位：分）

|  |  |  |
| --- | --- | --- |
| **device\_id** | **price** | **count** |
| **22673562117667** | **424000** | **107** |
| **19589529435459** | **5983** | **99** |
| **18647215736283** | **98000** | **98** |
| **22302229615401** | **150000** | **95** |
| **23166049869043** | **111000** | **92** |
| **16865554846601** | **85000** | **85** |
| **16935266687853** | **175000** | **84** |
| **22551181111677** | **82000** | **82** |
| **19550110913361** | **80000** | **78** |
| **23415274893759** | **108000** | **77** |
| **device\_id** | **price** | **count** |
| **22673562117667** | **424000** | **107** |
| **18642432475867** | **318000** | **44** |
| **16791640844171** | **306000** | **44** |
| **16973700487823** | **285000** | **29** |
| **19003514858321** | **266000** | **22** |
| **17122847126573** | **265000** | **17** |
| **17119516316241** | **254000** | **33** |
| **16828658991937** | **254000** | **29** |
| **22915814409341** | **253000** | **18** |
| **23039711845847** | **238000** | **46** |

# 附件

1. 各站点每天以1小时为单位的购票数统计结果



1. 3月30日上午7点到9点各路段客流计算结果



1. 3月30日下午17点到19点各路段客流计算结果

