全国旅客出行特征及车流量分析

——基于2020年1月-2021年5月全国旅客出行数据

一、数据简介 [3](#_Toc69808039)

[二、数据预处理 3](#_Toc69808039)

[(一)缺失值处理 4](#_Toc69808040)

[(二)班次代码预处理 4](#_Toc69808044)

[(三)发车日期预处理 4](#_Toc69808068)

[(四)发车时间预处理 5](#_Toc69808072)

[(五)车站预处理 5](#_Toc69808075)

[1.乘车站预处理 5](#_Toc69808051)

[2.到达站预处理 6](#_Toc69808052)

[三、描述性分析 7](#_Toc69808049)

[(一)发车量最多的乘车站 23](#_Toc69808050)

[(二)发车量最多的时间 43](#_Toc69808053)

[(三)发车量最多的日期 60](#_Toc69808068)

[四、可视化分析 50](#_Toc69808057)

[(一)车站站点信息 50](#_Toc69808058)

[(二)跨省车流量统计图 54](#_Toc69808062)

[(三)不同日期的乘车站车流量统计 60](#_Toc69808068)

[1.热力日历图 23](#_Toc69808051)

[2.时序图 36](#_Toc69808052)

[(四)座位类型可视化 61](#_Toc69808072)

[(五)省内车流量情况（以广东省为例） 62](#_Toc69808075)

[五、结论 59](#_Toc69808065)

1. **数据简介**

本次数据分析所用数据为全国各个城市的车站发车记录，数据包含8个变量，其中舍去两个变量不做分析，其余6个变量分别为班次代码、发车日期、发车时间、乘车站名称、到达站名名称和座位类型。变量类型均为字符串型，共有107226143条数据记录。

1. **数据预处理**

**（一）缺失值处理**

首先去除第四、五列变量，并对其余各列重新命名，读取前10行数据如下：

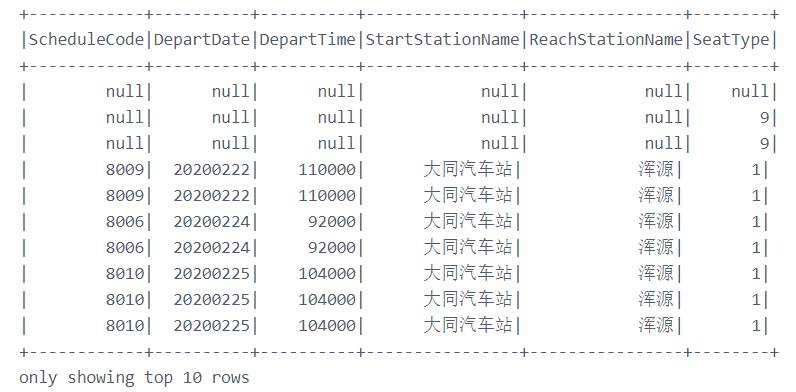


图1 数据预览

求得各列缺失值占比如下：

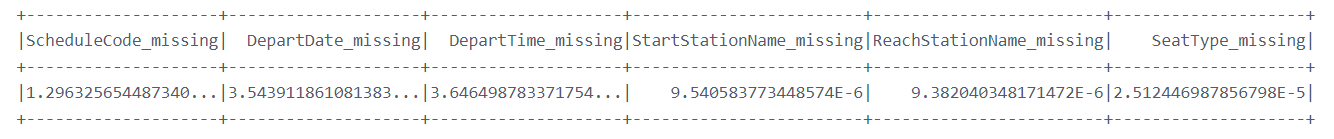


图2 数据缺失值比例

由于各列都有一定的数据缺失，且占比极小，删除后对整体数据的影响很小，故删去含有缺失值的行记录。

删除缺失值后，读取前10行数据如下：

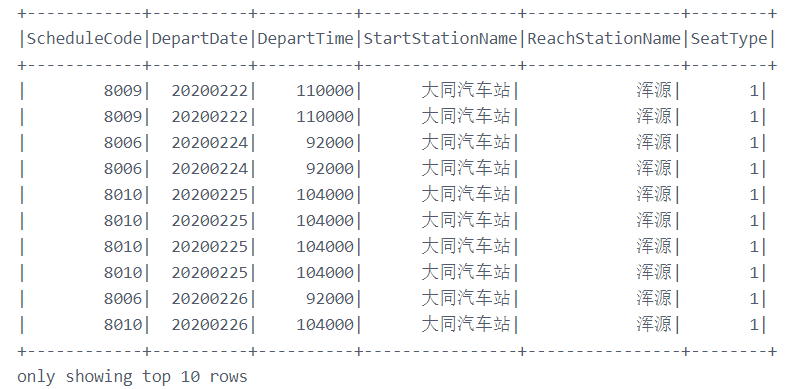


图3 缺失值处理后数据预览

**（二）班次代码预处理**

读取班次代码列以及分组计数列的前10行数据，发现无异常值，且在后续数据分析中，班次代码仅为计数，故不做过多处理。

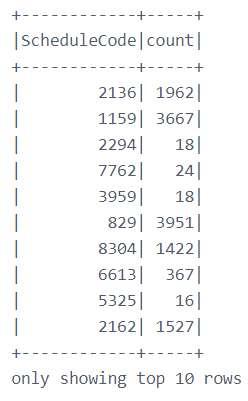


图4 班次代码及分组计数列展示

**（三）发车日期预处理**

对离站日期进行针对性处理，首先对离站日期进行分组统计，结果显示出“刘文东”等异常值及“2020-02-26”等具有不规范格式的日期数据。



图5 离站日期异常值①

为方便后续处理，我们删除或修改离站日期列的异常记录，并对留下的日期数据进行规范化处理。由于我们希望所有的日期都呈现“”的标准格式，首先将所有不满足8位数字格式的日期补全，并将所有超过8位数字的日期统一以字符串“”代替并删除。对补位操作后的离站日期进行分组统计如下：

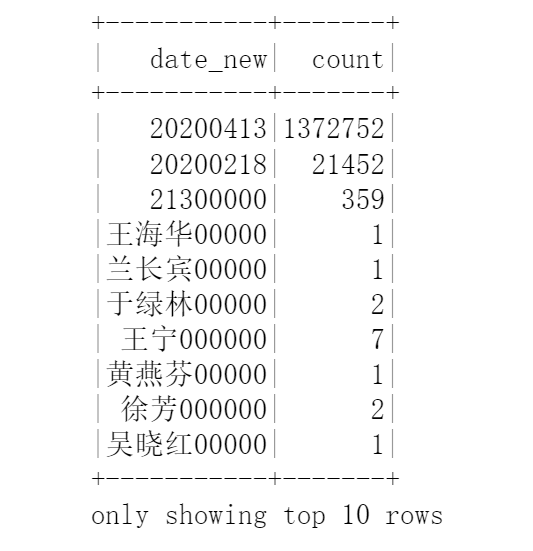


图6 离站日期异常值②

从上述数据截图不难看出，仍存在一些以字符形式存在的离站日期取值，因此通过操作将所有以非数字格式存在的日期记录删除。

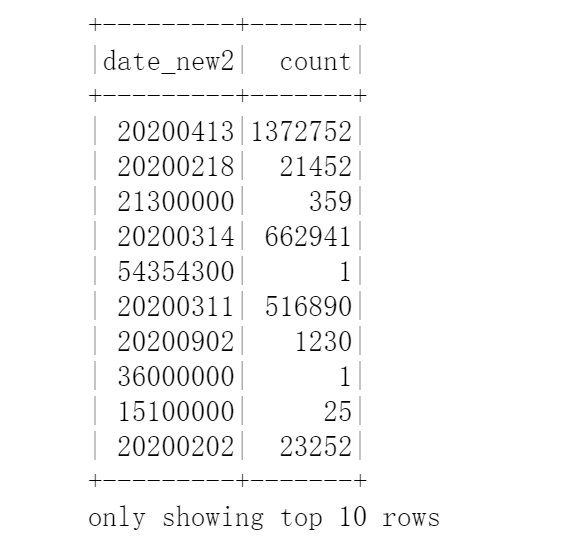


图7 格式化后离站日期

为了从离站日期中获取更多更有效、易分析的信息，依次提取其中的“年份”、“月份”和“日”的数据重新形成三列，并在每次提取过程中都进行有效范围的筛选，如剔除非“”形式的年份所在行。在月份处理中，为了保证位数一致，不妨用“\*”补齐“1～9”的首数位。经信息提取后的日期相关数据集展示如下：

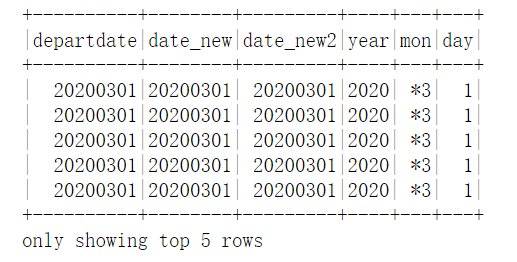


图8 提取信息后的离站日期（年、月、日）

**（四）发车时间预处理**

对发车时间进行预处理，为方便后续处理提取出发车时间的小时数。数据中有3位、5位和6位数字的记录，因此将对格式不同的数据进行分析修改。

对于只有3位数字的记录，无法判断具体时间，但占比很小，所以直接删除。

对于5位数字的记录，观察可知是小时数小于10的记录省掉了首位的0，不做处理。

对于6位数字的记录，认为是形式为hhmmss的标准格式，不做处理。

最后提取出表示小时的数据（DepartHour）。

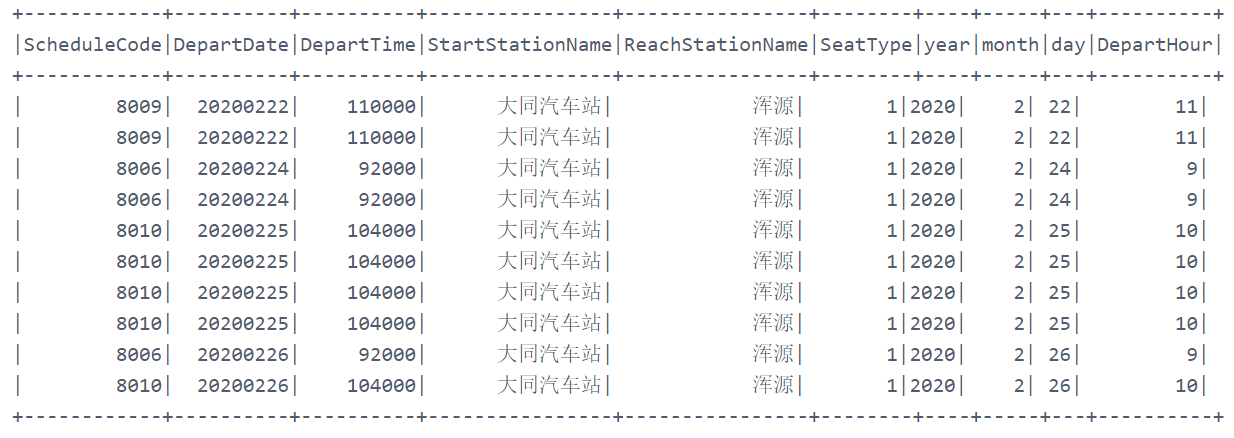


图9 提取信息后的发车时间（时）（最右一列）

**（五）车站预处理**

数据中出发车站和到达车站两个变量，能够反映车辆的流入量和流出量。

**（1）乘车站预处理**

“StartStationName”中为各乘车车站的名称，不能直接反映地理位置信息。我们通过函数提取出车站名称中包含的所在区/县、省、市的特征，其中一些只包含区/县名称，我们根据区/县名得到地级市名，最后根据地级市得到所在省名，还有一些车站直接包含地级市名，直接提取地级市，再推算出省即可，还有少数车站不包含地理信息（例如“万达广场”，全国各地有很多车站名为“万达广场”), 对于此类难以定位的车站，直接删除该行。

为了形成车站与其所在省、地级市和区/县的映射关系，我们从“中华人民共和国民政部”的官网上整理了一份精确到县级的行政区划表（见“行政区划.csv”）。但还存在一个问题：行政区划表包含的是各省市县的全称，而“StartStationName”中往往包含的并不是全称，而是两个字的简称。所以我们截取省\市\县全称的前两个字作为缩写，将三个变量命名为’province\_short’, ’city\_short','county\_short',用于与车站名的映射和匹配。

具体匹配方法是，对每一个车站名，遍历'county\_short'，如果某一个'county\_short'（县名）包含在车站名中，则认为该车站位于该县，然后根据县推算得到地级市、省。对于未能匹配成功的车站，再遍历‘city\_short'(因为某些车站中包含的不是县级而是地级市名)，同理匹配得到该车站所在地级市，再推得省。最后剩下的未能匹配成功的车站，将对应观测直接删除。

经过上述处理，我们从车站名衍生出了县级、地级市、省级三个层次的地理信息，观察发现，有很多观测的县级信息还是缺失的，再加上我们后续研究主要精确到地级市，所以把县级的地理变量删去。最后保留地级市、省级两个层次。经过处理后的前几行数据如下图所示：



图10 乘车车站及所在的县级、地级市、省级区域

统计得出，乘车车站共有4677个，能精确到地级市、省级地理信息的有3794个，其中比例最高的是广州省的乘车站。

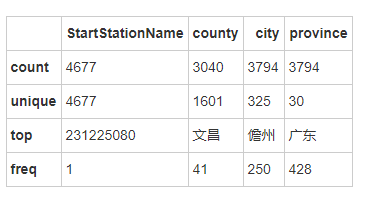


图11 乘车车站及地理信息描述

**（2）到达站预处理**

“ReachStationName”中主要为到达车站的地级市或县级名称，为了定位地级市、省级地理信息，我们用同样的方法将其与行政区划表进行匹配，得到处理后的数据。统计得到，到达车站共有51071个，其中有17599个能精确到地级市、省级地理信息，占比最大的同样是广东省到达车站。

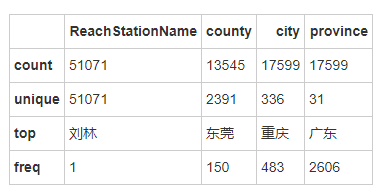


图12 到达车站及地理信息描述

1. **描述性分析**
2. **发车量最多的乘车站**

对处理后的乘车车站进行车流量计数，结果图如下，可以看出发车量最多的前七个乘车站。

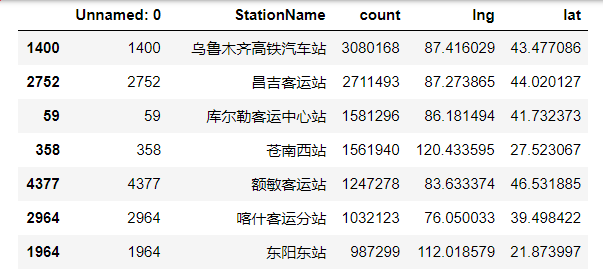


图13 发车量最多的乘车站

1. **发车量最多的时间**

对处理后的发车时间（即：发车所在时刻）分组统计，进行车流量计数。离站车流量最大的十个时刻及对应流量，如下图所示。

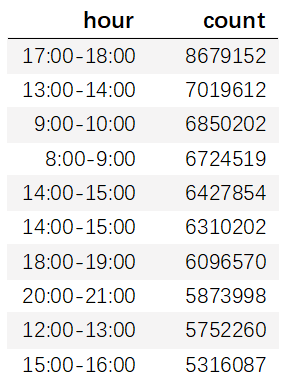


图14 发车量最多的时间

1. **发车量最多的日期**

对数据清洗后的数据，依据离站年份和月份分组对车流量大小进行统计，得到如下图所示的车流量TOP7对应月份。

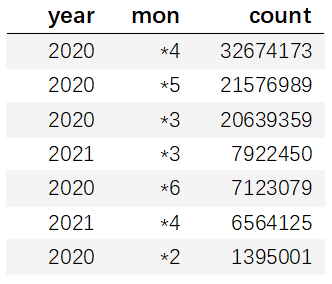


图15 发车量最多的月份

对数据清洗后的离站日期进行车流量分组统计，结果如下，不难发现发车量最高的日期集中在2020年4月。

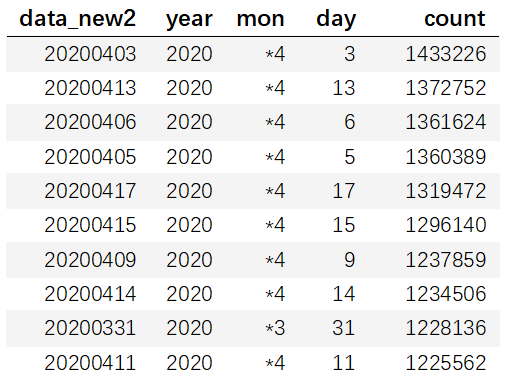


图16 发车量最多的日期

1. **可视化分析**

**（一）车站站点信息**

乘车车站经过划分区/县、市、省后，将这四列数据存下，借助百度地图，利用python定位车站地理坐标，并对乘车车站进行车流量计数，由于一些乘车站没有能够通过名称提取出所在县/区，将经纬度设为0，以便于后续删除，前几行数据如下：



图17 乘车车站经纬度信息

由此，绘制乘车站在各地的分布图如下：

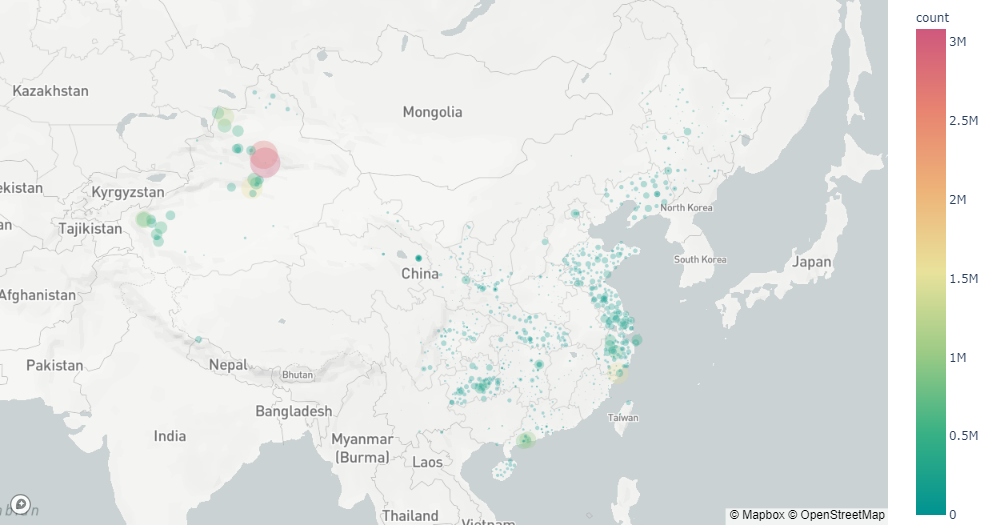


图18 乘车车站在各地的分布图

其中，以颜色深浅和点的大小代表车流量大小，点越大，颜色越深，表示车流量越大，其中乌鲁木齐高铁汽车站、昌吉客运站、库尔勒客运中心站、苍南西站和额敏客运站车流量前五。

**（二）跨省车流量轨迹图**

根据预处理得到的出发车站所在省份和到达车站所在省份，可以得到“出发地-到达地”的路线，选择出其中跨省的记录，根据某条路线在所有跨省记录中的占比得到权重，可以画出全国跨省的车流量轨迹图（选取权重大于0.5的前44条路线）。



图19 车流量最大的五条路线

图中，箭头移动的方向即为出发地至到达地的方向，线的粗细表示车流量大小，线条越粗，车流量越大，其中陕西-黑龙江、贵州-四川、四川-重庆、江苏-上海和吉林-辽宁车流量前五。（轨迹图在html文件中）

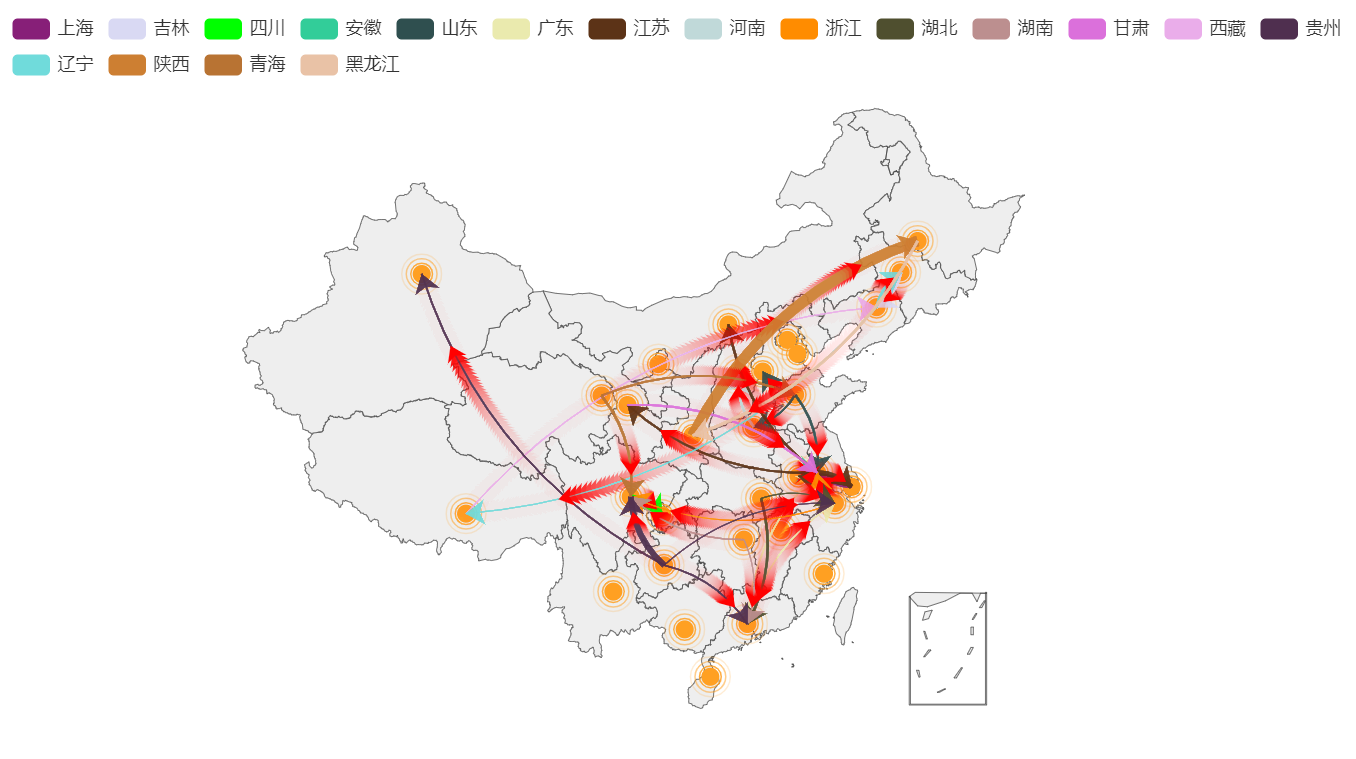


图20 跨省车流量轨迹图（动图）

**（三）不同日期的乘车站车流量统计**

**（1）热力日历图**

根据预处理中提取的离站时间信息，分组统计计算各月份车流量，并以热力日历图的形式进行可视化。

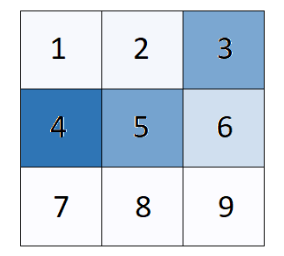
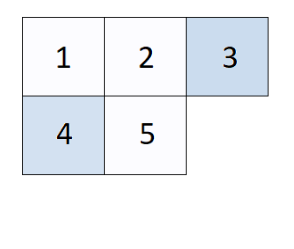
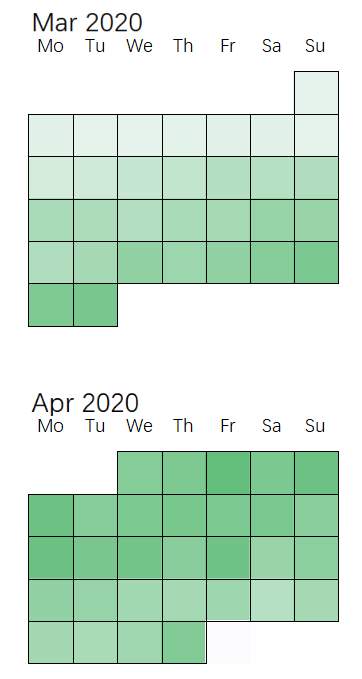
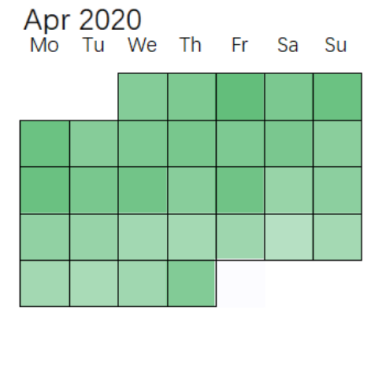
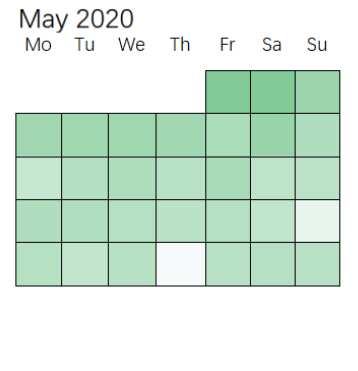
 

图21-1 2020年1-9月车流量 图21-2 2021年1-5月车流量

从上述热力日历图结果可以看出，在2020年3、4和5月的发车数量最多，其次是2020年6月、2021年3月和4月。这样的结论与我们预期中寒暑假假期车流量较多的假设有所矛盾，对各月份的发车记录登记形式和日期进行分批次查看，发现其原因可能在于某些月份的数据记录格式不符合常规，在对数据进行统一处理时被迫丢失。

在上述月度车流量热力日历图的基础上，我们发现2020年3、4、5和6月以及2021年3、4月的车流量远远高于其他月份，且对于这些月份而言，几乎每天都存在发车记录，具有较为良好的分析价值。因此，对上述筛选出来的月份的具体发车数据进行可视化分析。

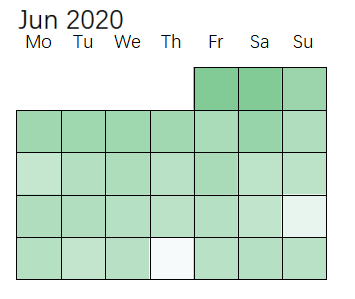
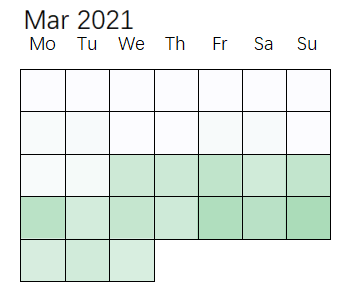
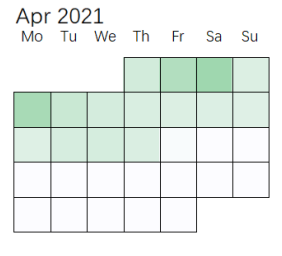
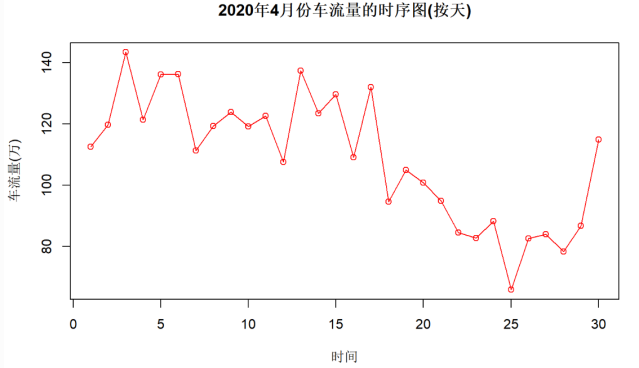
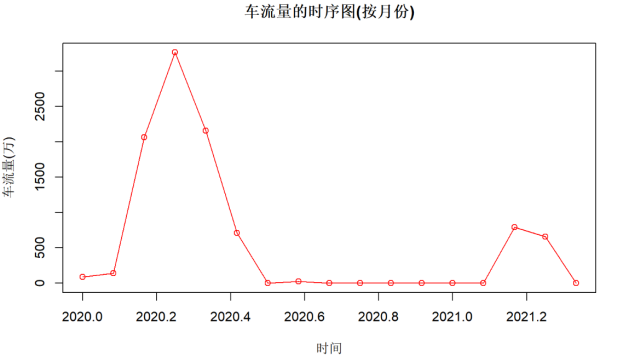
  

图22 车流量日历图（部分）

根据车流量日历图，我们不难发现位于周末的格子颜色整体深于位于工作日的格子，说明在这几个具有代表性的月份中，周末出行的人数普遍多余工作日，这一可视化数据图相对来说也比较符合现实情况。

**（2）时序图**

根据月度车流量的统计，我们绘制了以下四张时序图，分别描述了所有清洗后数据的月度序列以及其中月份（2020年4月、2020年5月和2021年3月）每天的离站车流量序列的走势。据下述时序图显示，车流量的序列不具有平稳性的性质，在2020年4月和5月有较为明显的下降趋势，而在2021年3月有较为明显的上升趋势。



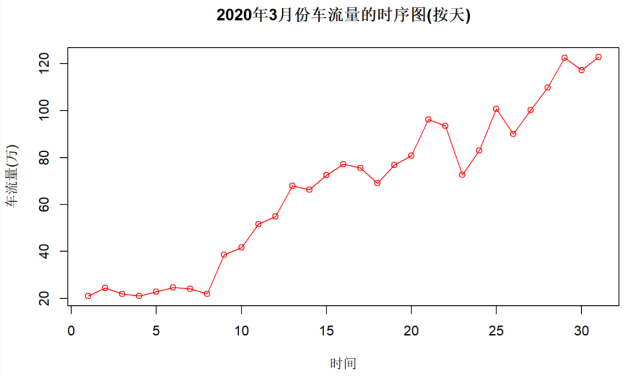
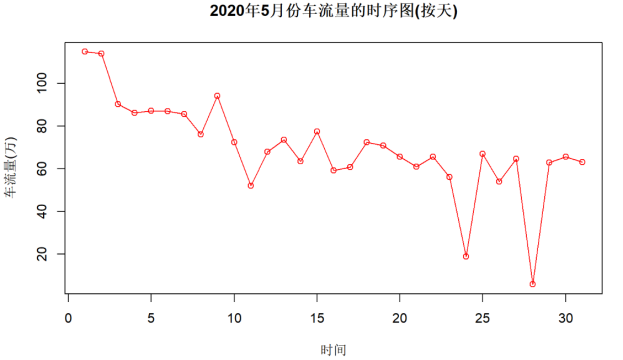


图23 车流量日历图（部分）

**（3）饼图（时间）**

根据不同时间段离站车流量的统计，我们绘制了如下饼图。为使得图形可观化，我们以某个数字代替一个时间段，如以“1”代表“1:00-2:00”。由饼图，我们发现离站时间集中在8:00-18:00，说明此数据集的记录对象大多选择在白天的工作时间进行外出。

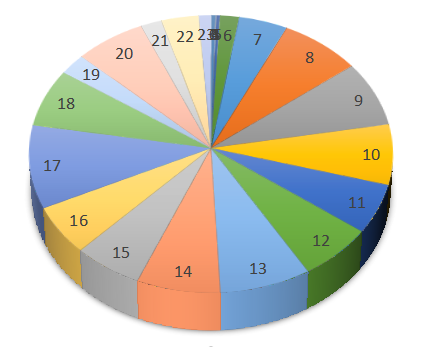


图24 不同时间段车流量饼图

**（四）座位类型可视化**

座位类型变量包括1、2、3、4、9，分别对应普通座、商务座、上铺、下铺和其他。

通过计数得到五种座位的数量，绘制柱状图如下，从图中可以看出，座位类型绝大多数为普通座和其他类型，几乎没有商务座和上下铺。

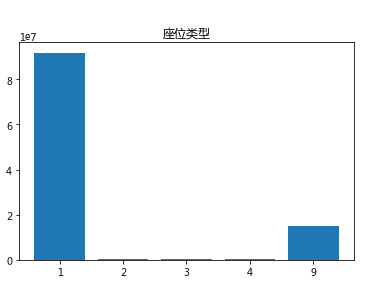


图25 座位类型柱状图

**（五）省内车流量情况（以广东省为例）**

由于全国各个省份数据过多，乘车与到达车站轨迹复杂，绘制流向图不够清晰明了，考虑到在乘车站与到达站中广东省均占比最大，所以选取广东省车辆流向进行分析。

筛选”StartProvince”和”ReachProvince”均为“广东”的记录，共得到1931385条数据。广东省共有20个市，对各市间的车辆流向进行可视化，绘制桑基图如下：

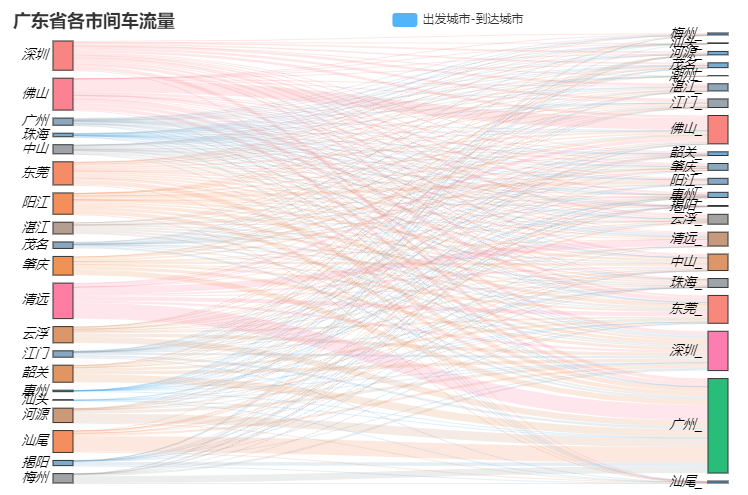


图26 广东省省内车流量桑基图

从图中可以看到，车流量较大的城市主要为深圳、佛山、东莞和广州，这些城市均为经济较为发达的一、二线城市。还可以看出，流向总体趋势是由小城市流向大城市，尤其是广州，流入量远远大于流出量。

1. **结论**

从上述数据分析及可视化可以得出，全国车流量最多的时间段位于17:00-18:00、13:00-14:00和9:00-10:00，具备一定的规律性，是现实生活中的常见车辆发车时间段，符合实际情况。

2020年1月到2021年5月，全国车流量最多的月份是2020年3、4、5月，由于疫情的影响，2020年1-2月份全国车流量很少，随着国内疫情得到有效控制，且复工复产稳定推进，3月份与4月份车流量逐渐增加，在五一小长假时达到出行小高峰。

从数据处理结果看，全国乘车站发车辆较多的站点有乌鲁木齐高铁站、昌吉客运站、库尔勒客运中心站等，同时，沿海城市也具备较大的车流量，如上海市、广州市等。丛生分之间车流量轨迹图可以看出：陕西-黑龙江、贵州-四川、四川-重庆、江苏-上海和吉林-辽宁之间的车流量最多。