基于深圳北站月旅客发送信息的数据分析

何俊锋 2021113362 交运茅班

摘要：针对深圳北站 30 日旅客发送信息进行了数据分析。首先对数据内容和结构进行初步分析，然后定义了本文要解决的三个问题：深圳北站发往其他车站的流量流向、各车次的供给与需求关系以及列车开点与上座率的关系。使用 Python/Numpy/Pandas 对数据进行清洗和预处理，然后使用 pyecharts 绘制了流量流向示意图，使用 seaborn 和 matplotlib 对供给需求关系以及列车开点与上座率的关系进行可视化描述，并对相应结果进行了分析。文章最后总结了本次数据分析值得注意的结果。

关键词：数据分析；深圳北站；旅客运输

Data analysis based on monthly passenger information sent by Shenzhen North Railway Station

Abstract: This paper analyzes the data of the passenger information sent by Shenzhen North Railway Station in a month. First of all, the data content and structure are preliminarily analyzed, and then three problems to be solved in this paper are defined: the flow direction of Shenzhen North Station to other stations, the relationship between the supply and demand of each train number, and the relationship between the train departure point and the occupancy rate. Python/numpy/pandas was used to clean and preprocess the data, and then pyechards was used to draw the flow direction diagram. Seaborn and matplotlib were used to visually describe the relationship between supply and demand, as well as the relationship between train departure and occupancy, and the corresponding results were analyzed. Finally, the paper summarizes the results of this data analysis.

Key words: data analysis; shenzhen north railway station; passenger transportation

1 数据内容分析及问题定义

数据分析首先需要对数据内容进行预分析，了解各数据的结构、数据类型等基本信息。从深圳北站该月的旅客发送信息来看，数据主要包括：1）从深圳北出发各列车的车次、开点、到站、总运能、总旅客发送量以及上座率，其中总运能和总旅客发送量按一日、一旬为周期分别进行了统计；2）以一旬和一月为统计周期，从深圳北站到达其他各站的总运能、总旅客发送量以及上座率；3）以一日、一旬和一月为统计周期，深圳北站向其他各站的流量流向。这些数据分别放在同一个 excel 文件下的五个工作表中，命名为发送人数上旬、发送人数中旬、发送人数下旬、发送人数合计及流量流向。

但是注意到，文件中存在许多缺失值(NaN)，这对数据分析工作产生了一定影响，故需要进行数据清洗。总的来说，缺失值主要是由于某些车次(如 G6204)在某旬内无旅客发送人数等数据或只有某几天有数据。考虑到该部分的数据相较于整个数据内容来说占比较小，本文采取直接删除(dropna)该车次数据的方法。

基于此，本文着重探讨三个方面的问题，分别为深圳北站发往其他车站的流量流向、各车次的供给与需求关系以及列车开点与上座率的关系。流量流向主要

根据每日发往其他各站的旅客人数进行分析，供给与需求关系主要考虑每列车的定员与发送人数之间的关系，列车开点与上座率源于现实经历，个人经验来讲晚上的上座率通常不如白天的上座率。

2 数据处理与探索

2.1 深圳北站发往其他车站的流量流向

流量流向的可视化见图 1。地图展示了 30 天内，各个站点\市的流量情况，流量范围从最低的 59 人次到最高的 900596 人次，使用不同的颜色表示不同的流量区间。较高的流量用红色和深红色表示，较低的流量用蓝色和浅蓝色表示。首先创建一个字典，映射每个车站与其所在的市。然后将 30 天到达各市的旅客人数进行汇总，最终形成图 1 的 30 天深圳北站总流量流向示意图。

从地图上看，东南沿海地区（尤其是广东、浙江和福建）的流量明显较高，这可能与这些地区的城市化水平、经济活动及人口密集度有关。特别是在广东省，特别是深圳附近，流量最为集中。随着流量区间的变化，从西北向东南，流量逐渐增加，表明东南沿海地区的交通流动性更强。北部地区（如内蒙古、新疆等）和西南偏远地区相对流量较低，这可能是由于距离较远，而高铁晚上需要检修无法在一天内到达这些地区导致的。

但是也有一些出乎意料的现象。虽然吉安市、赣州市、河源市等市距离深圳很近，但是没有流量流向这些区域。这是由于这些地区之间可达性较低，这也许是未来广东地区交通发展的一个方向。同时，虽然总的来说距离深圳较远的地方流量越低，但是流向郑州、西安、石家庄等市的流量仍然很高，尤其是郑州市和北京市，远超其他同等距离的城市的流量。原因可能郑州、西安和石家庄都是重要的经济中心，郑州作为中原地区的交通枢纽和商业中心，西安是西北地区的政治、经济、文化中心，石家庄作为河北省的省会，均具有良好的经济基础。

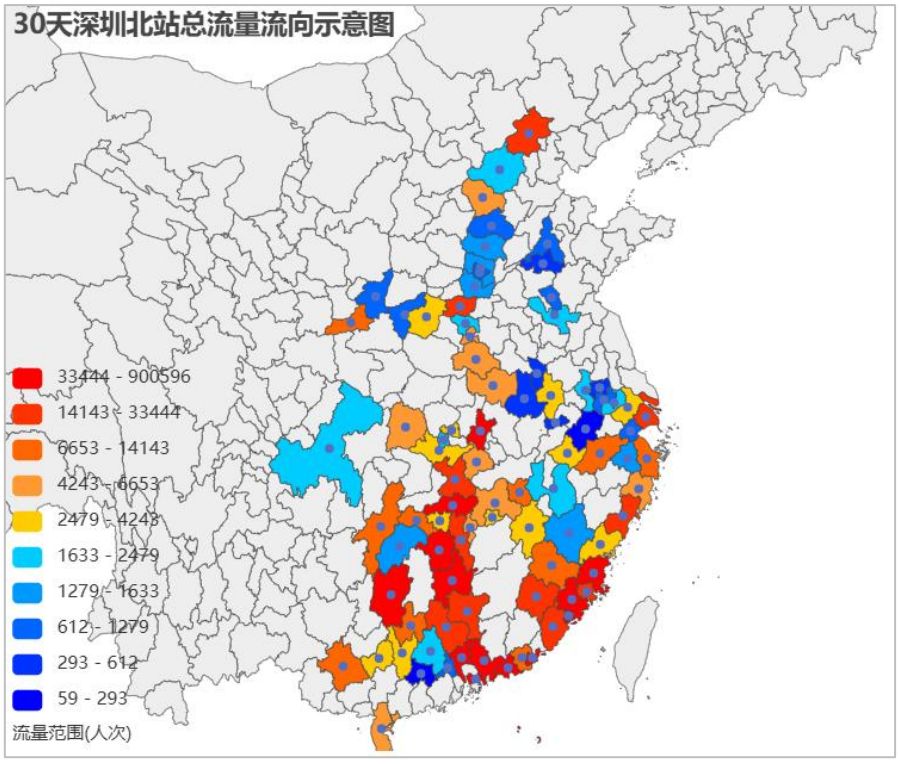

图 1 深圳北站发往其他车站的流量流向

2.2 各车次的供给与需求关系示意图

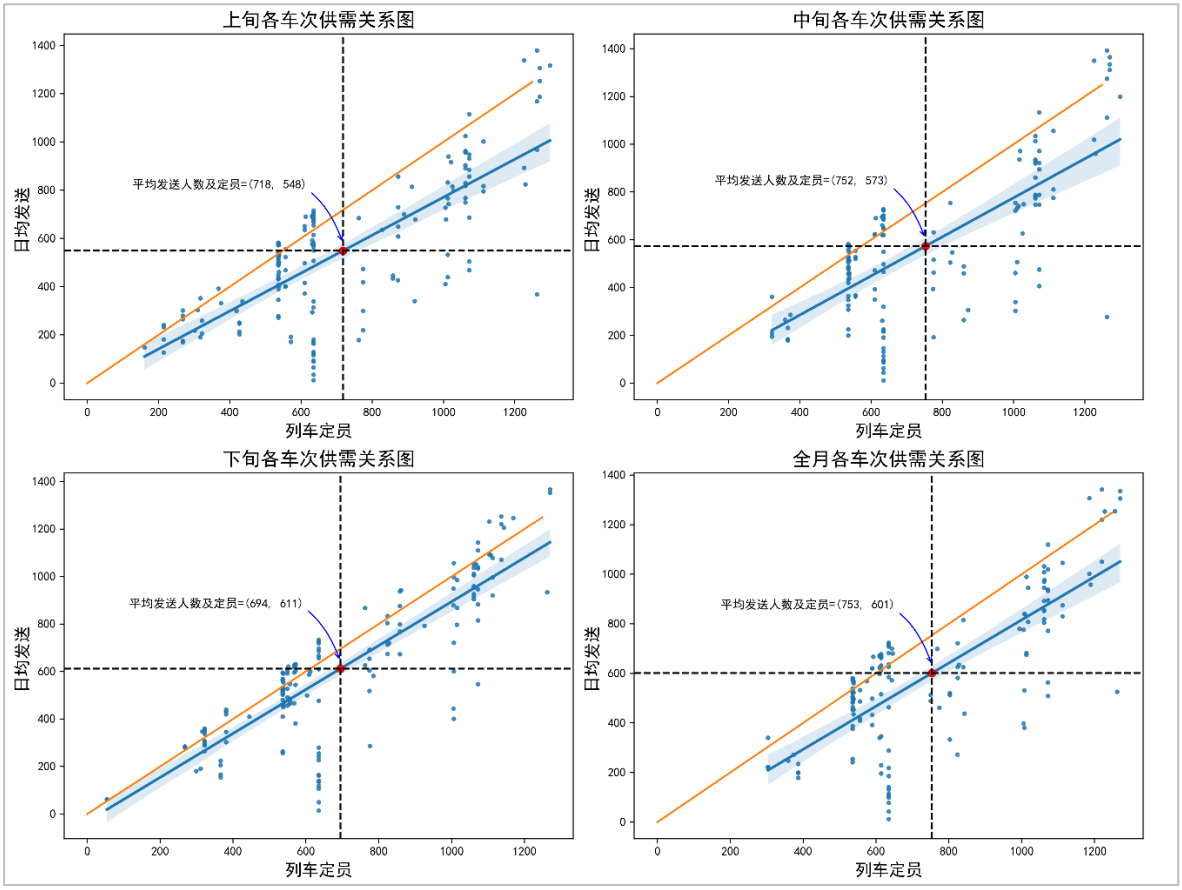

图 2 各车次的供给与需求关系示意图

各车次的供给与需求关系示意图见图 2。所有图表都展示了旅客需求与列车定员之间的正相关关系，说明列车定员越高，旅客的需求量也越大。每个图上都有显示所有列车平均定员和出行人数的红色点（如 `710, 548`），橙色线为列车定员数等于日均发送旅客人数的情况。总的来说，大部分列车的供给都能满足需求，但是仍然有少部分列车位于橙色线上方和红色点的左上方，表明这些列车的定员无法满足乘客的需求，相关部门应该针对这些车次采取一些措施，如加开列车，采用重联列车来满足对应的需求。针对位于橙色线下方，红色点右下角的列车，这些列车的定员高于对应的出现需求量，从经济层面考虑，可以减少这些列车的开行。

上、中、下旬和 30 日平均数据差距不大，波动性不太高，表明该地区的交通需求量较为稳定，这对铁路运营来说是一定优势。

2.3 列车开点与上座率的关系示意图

列车开点与上座率的关系示意图见图 3。总的来说，从早上 6 点开始到早上 10 点上座率逐渐上升，然后一直维持较高的上座率(0.9 左右)一直到下午 5 点。下午 5 点过后上座率开始急剧降低，到晚上 9 点上座率只有大概 0.4 左右，这也许和出行者的出现偏好有关。但是值得注意的是，晚上 10 点上座率普遍迎来了一次大幅上升，从 0.4 左右上升至 0.8 左右。这可能是商务会议和社会活动通常在晚上结束，乘客通常会选择在这些活动结束后立即乘坐高铁返回，或者从事商务活动的人常常选择夜间出发，以便在第二天更好地安排日程，导致晚上 10 点的上座率上升。

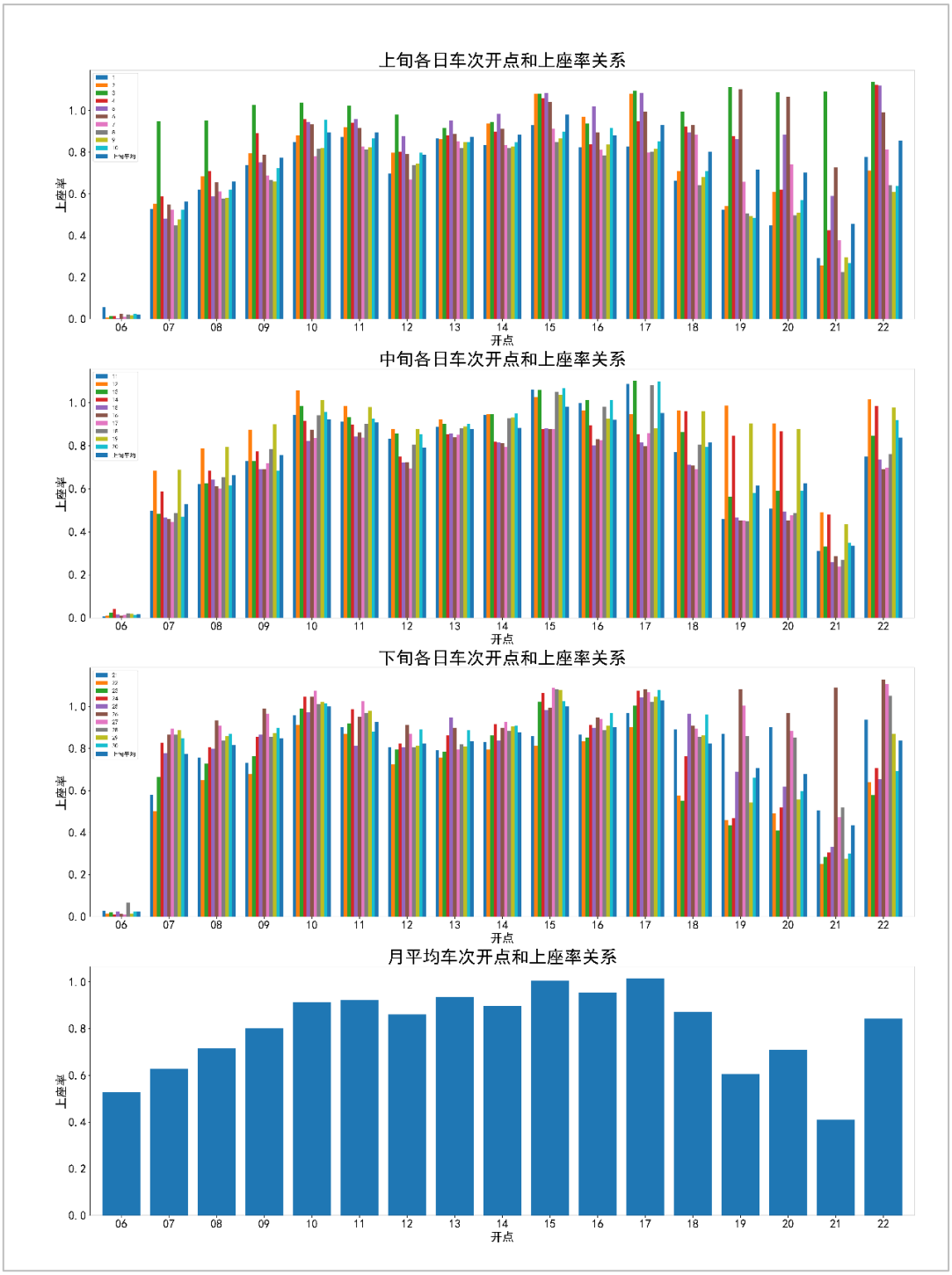

图 3 列车开点与上座率的关系示意图

3 总结

本文针对深圳北站旅客发送人数进行了数据分析，有几个有趣的地方：1）夜间上座率较低，但 22 点产生了上座率激增现象；2）部分车次存在供需不平衡现象；3）该站与西部地区可达性较低。这些现象对未来交通布局有一定启示作用。

4 附录

代码地址：https://github.com/SWJTUHJF/Miscel，以防涉密，原始数据未上传。