实验报告：城市轨道交通通勤数据分析

小组成员：曹紫昱 祁辉昶 苏正一 贺耀扬 张含玉

一、数据清洗

- **进出站点应符合线路特征**

考虑到本数据仅为某城市在2022年10月31日至2022年11月18日中，共15个工作日（不含周六、周日）某轨道交通通勤走廊线路下行方向的早高峰智能卡刷卡数据，及该轨道交通所有列车时刻表数据。由于为下行方向，则所有数据都应该满足出站站口在入站站口后方，即有：

$$Station_{ori} > Station_{des}$$

则不符合该条件的数据均为错误数据，都应删除。

- **进站时间应早于出站时间**

由于该城市轨道交通数据为单行方向，则出站时间应晚于进站时间。即所有的数据都应该满足：

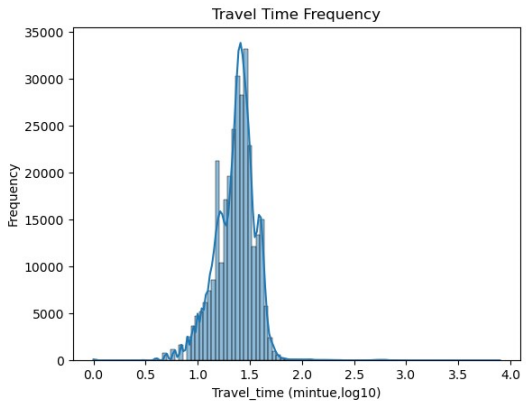
$$Time_{del} > Time_{ent}$$

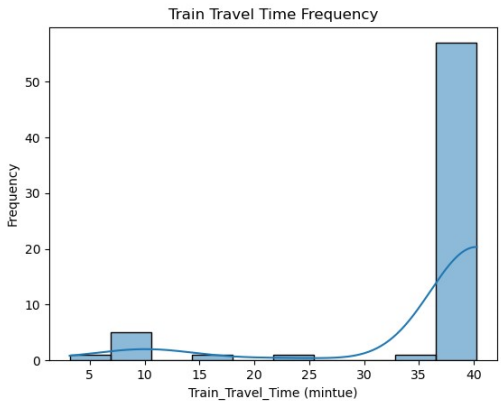
案例中的部分数据出现出站时间大于或等于进站时间，不符合实际，应当删除。

- **出行时长应符合规律**

数据中有部分乘客记录的出进站时差过长，这与实际情况不相符，应当结合地铁的运行的时差从而删除不符合实际的数据。

我们分别画出了所有人出行的时间频次图，和地铁运行时间频次图：





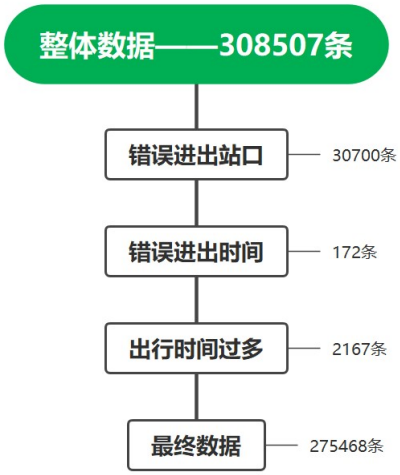
其中由于出行时间范围较大，对时间取对数10处理。结合图示，发现绝大部分出行数据都分布在$0.5 \sim 2.0$之间，且单程列车运行时间为40min，则依照实际情况剔除出行时间大于100min的数据。

- **出行数据应满足有相应列车匹配该数据**

该部分在后续部分予以讨论。

- **结果展示**

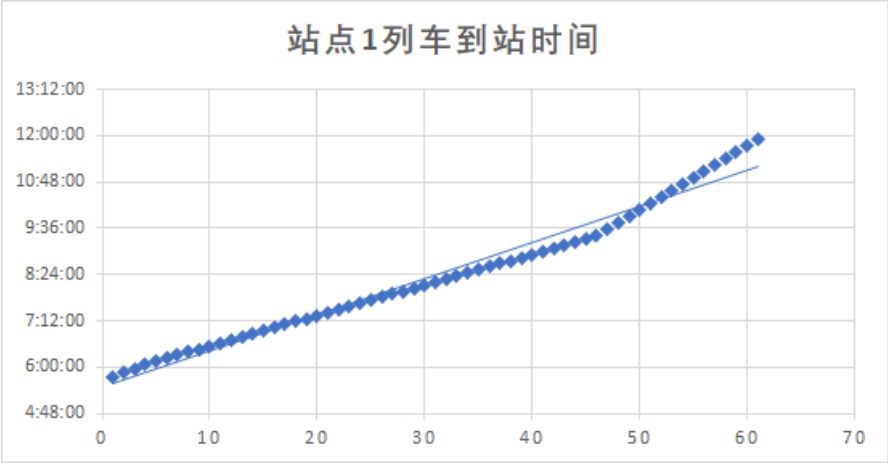
经过上述的数据清洗，得到结果如下：



二、整体数据分析

- **列车分析**

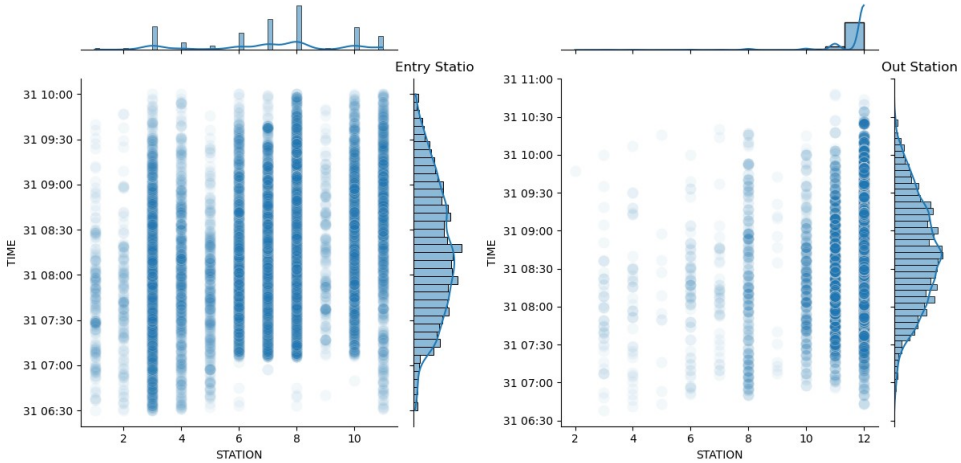
  考察所有列车随时间分布，由于所有列车在站点以及站点至站点的行为一致，则选取站点1的进站时间说明列车随时间的分布。



  可以的出列车在8:30之后车次较多，而在此之前车次较少。

- **进出站时间流量分布**

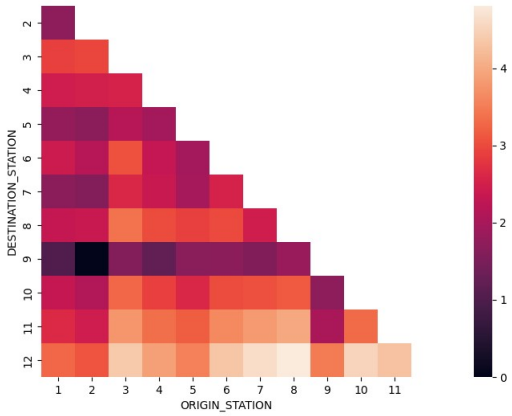
对清理的数据做具体分析，作出不同站点在不同时间段内的进、出站人数的流量分布图



可以看出对于进站，站点3、6、7、8、10为热门的进站点，而绝大多数出站口为站点11、12。出行时间近似正态分布，其平均数在早8:30前后。

- **热门路线分析**

以入站点和出站点为横纵坐标，分析所有OD对的频数，绘制相应的热力图。由于出行人数频数较大，且数据差异性较大，则对频数进行对数处理。热力图如下所示，

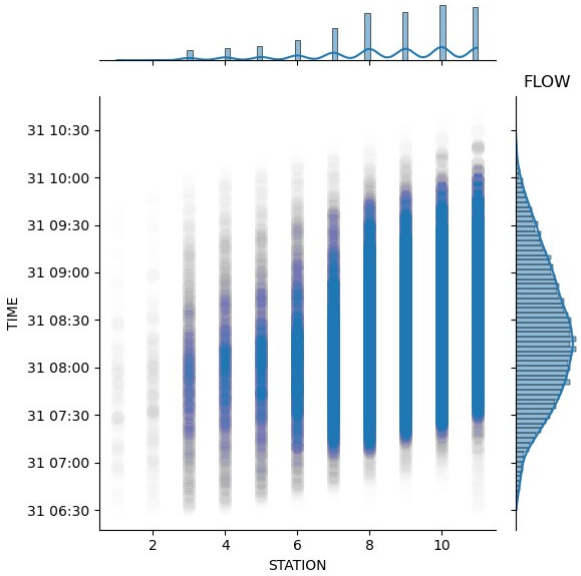


其中，颜色较浅为出行频数较多，颜色较深为频数较少。有图可得，较为热门的出行对为8-12、7-12、10-12、3-12。

- **路段流量**

根据上述的图示可以明显发现出行人数与时间和出站口、入站口关系。由于出行仍存在途径站，则可对路段进行分析。假设出行为从入站口9-12，则其对9-10，10-11，11-12路段均有贡献，并假设其时间根据入站、出站时间均匀分布。

对站点1-12中共11个路段进行分析，考虑其路段时间流量，绘制如下直方图



由图可以直观的看出各个时间段的人流量大致符合正态分布,高峰期为8：00左右，路段7-8、8-9、9-10、10-11、11-12路段出行人数较多，路段较为拥挤。

三、出行匹配相应列车

- **乘车分析**

分别考虑列车时间对某站点的进站时间和出站时间， 出行人进入出发站的时间以及离开目标站的时间，若满足列车在初始点列车的出站时间，晚于出行数据中进入该站的时间：

$$des_{train} \geq ori_{time}$$

同时对于出行的终点站，列车的进站时间早于出行数据中离开该站时间：

$$ori_{train} \leq des_{time}$$

即列车在出行人到达出发站时尚未发车，出行人离开目标站前已经的到达该站，说明该列车可以满足出行人的出行需求。若存在过多车辆，则依据现实假设出行人发现列车中乘客较多，出行人未选择前次列车，则选取最后一辆车为搭乘车辆。

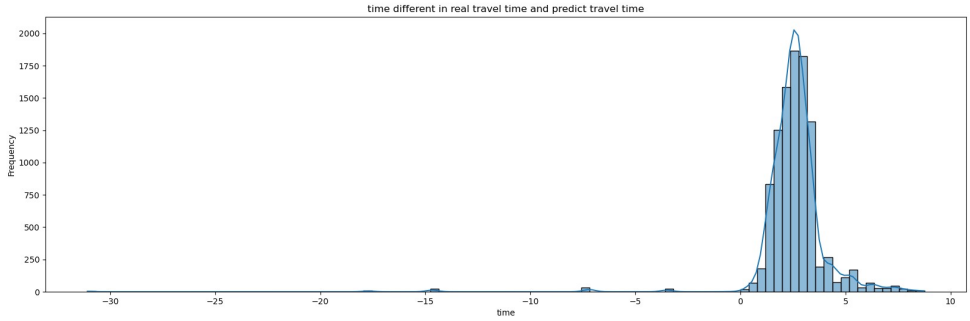
- **出行匹配**

根据前面分析的数据结果，有256760个乘客有车辆选择，剩余18708出行人无车辆匹配。考虑到列车的进站时间与出站时间精确至秒，而出行数据的进站、出站时间仅精确到分，且依据现实，绝大多数站点 从刷卡进站到乘车距离较短，可以在数秒内到达），考虑刷卡时间的误差，将出行人前后放宽30s（例如出站时间为8:10:29s，则仍会统计为8:10，我们假设对于无匹配的出行数据均出现过大的取舍问题，将数据恢复到精确至秒的时间），再进行测试算。此时有600人出现车辆选择。最终结果为：

有265360出行数据匹配的车辆选择，10108条数据无匹配列车。

- **原因分析**

由于车辆运行时间固定，且对出行人员时间已经进行较大范围的估计，则仍有未匹配数据不符合行里。由于所有列车的运行时间固定（在所有站点等待时间固定，且从站点至站点的运行时间固定），则考察人员出行时间（即离开目标站的时间与进入初始站的差值）与列车对应该路线的运行时间（目标站到站时间与初始站出站时间的差值）的差值的频数，直方图如下图所示，

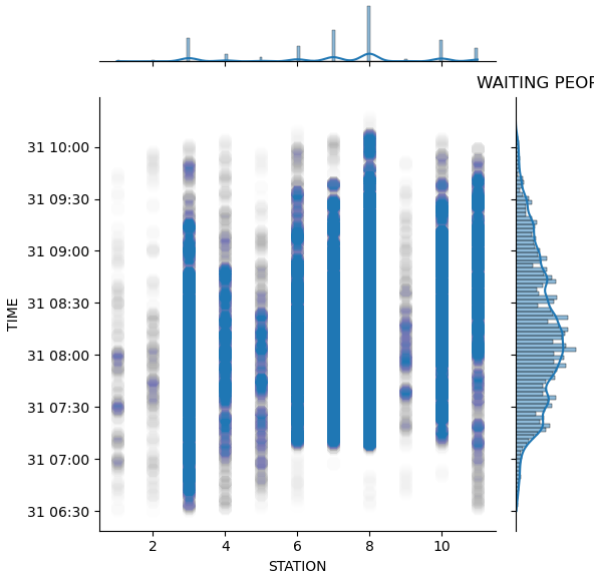


由图可知部分数据为无效数据（差值小于0min），不存在出行时间小于列车运行时间的情况。而绝大多数在$0 - -5min$，考虑为当日列车到站时间存在一定误差，与列车表数据并不完全相符，从而致使这些数据无车辆匹配。

- **所有站点等待人数**

（由于该分析为对所有数据进行分析，故假设此时已得出出行数据匹配的相应列车，并将离开站点时间假设为为列车离开站点的时间，则可以分析站点等待人员数量。）

进行仿真处理，对出行数据进入站点的时间进行模拟，从进站时间至离开该站点时间，每30秒进行打点，则可以得出不同站点进站人员随时间的等待情况，如图所示。

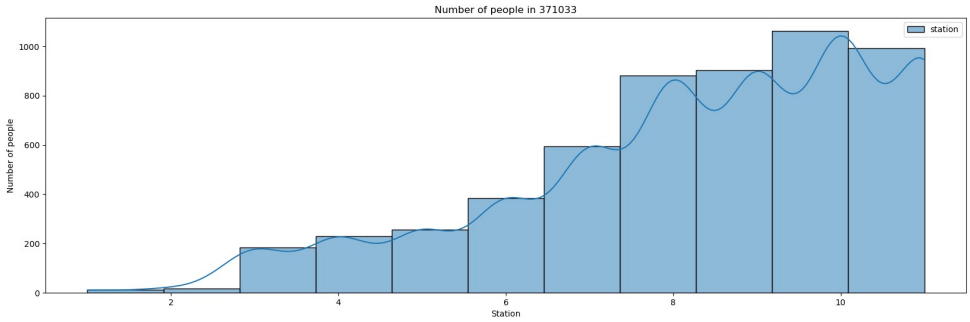


由图可得出站点3、7、8的等待人数较多，站台处于较为拥挤的状态。同时人数的高峰期为8：00左右，这个时间的站点等待人数最多，最为拥挤，出现不方便。而当提前或延后出行时，等待人数都会减少很多，出行较为便利。因此要想减少等待时间，尽量在7:00前或9点后出行。

四、单一车辆分析

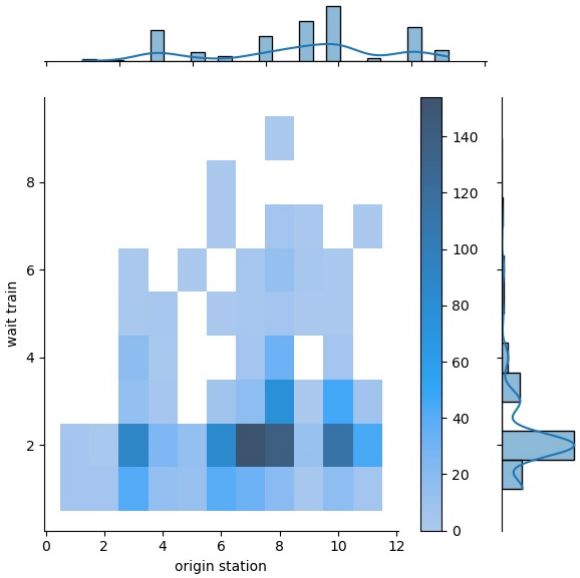
- **特例选取**

选取到10月31日这一天的数据，选取行人数最多的列车，其列车编号为371033。则对于371033次列车上人数和站点关系分析，结果如图片所示



我们发现8、9、10、11站点的车上人数较多。站点1、2时车上人数很少，站点3、6、7、8车上人数增加较多，站点8车上人数基本达到满员状态，站点10车上人数达到最多，这时车上较为拥挤。

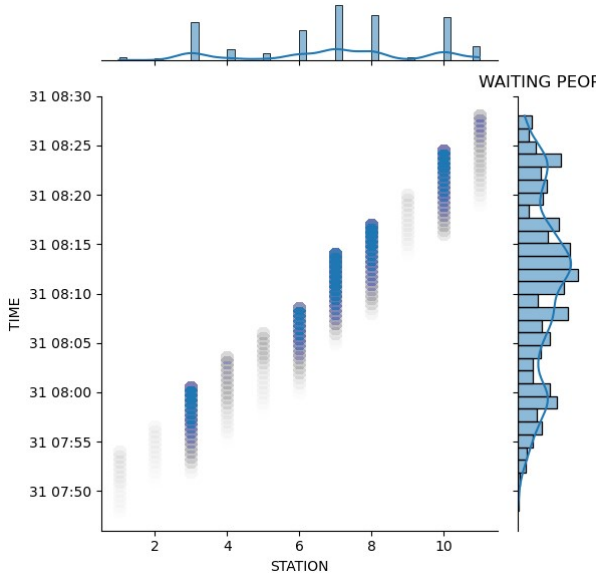
继续考察371033等待车辆个数和初始站的关系（对于部分人无车辆是否是车厢人多导致的不愿意上车）



发现有一部分人等待车辆个数较多，可能原因有当前站台等待人数过多，一辆车无法满足车站的所有人上车，于是一部分人选择等待下一辆车，例如站点7、8；

而对于站点3，也可能是当前站台的车上人数过多，没有座位，于是一部分人选择等待下一辆车。同时考虑到OD对3-12较多，对于该站上车乘客旅程较长，等待有座位的列车符合常理。

- **结果分析**
  - 站次3与6、7、8倾向于等待下一辆车
  - 站台3车内人数急增，一次性上车人数较多，部分人可能是因为没有座位或过于拥挤而选择不上车去等待下一辆车，等待可能使得有座位或不太拥挤
  - 8、10因为当前车上人数过多，无法继续上车，一部分人会选择不上车，继续等待更多车次。即当前车次无法满足上车需求
  - 站点9由于上车人数极少（见进站时间流量图）8站台到9站台车上人数增加很少，导致等车人数较少，于是9站台的等待车辆个数较少。(补充：考虑实际情况，例如站点8已经无法继续上车，在列车运行时间车内人员会继续调整站位，从列车门附近往列车中部走，使得下一站仍可继续满足乘客的上车需求，也可对该情况进行说明)
- **站点等待人数（站点内的拥挤情况）**



站点1、2的等待人数较少，这是因为上车人数少，车上人的数量较少，所以可以快速上车；站点3的等待人数很多，于是导致站点3的车上人数激增；站台6、7、8的等待人数较多，车上的人数也在较快的增加，基本达到了满员状态；站台9的等待人数很少，上车人数少，于是导致车上的人数基本没有变化；站台10的等待人数较多，上车人数较多，车上人数进一步增加，这时车站较为拥挤。

该情况与上述所有站点等待人数相符合，故不过多分析。

五、通勤用户识别及其出行规律分析

- **筛选通勤用户**

根据智能卡卡号的历史出行数据，筛选出平均每周在早高峰期间出行3次以上的用户数据。选择进站站点为7，出站站台为12，选择出行时间为$7:00-10:00$

- **聚类分析**

利用spss软件使用K均值动态聚类发对通勤用户的进站/出站时刻、出行频次数据特征进行聚类分析，结果如下图所示：

最终聚类中心

	聚类		
	1	2	3
ENTRY_TIME_TIME_STAMP	9.14	8.32	7.57
DEAL_TIME_TIME_STAMP	9.57	8.78	8.03

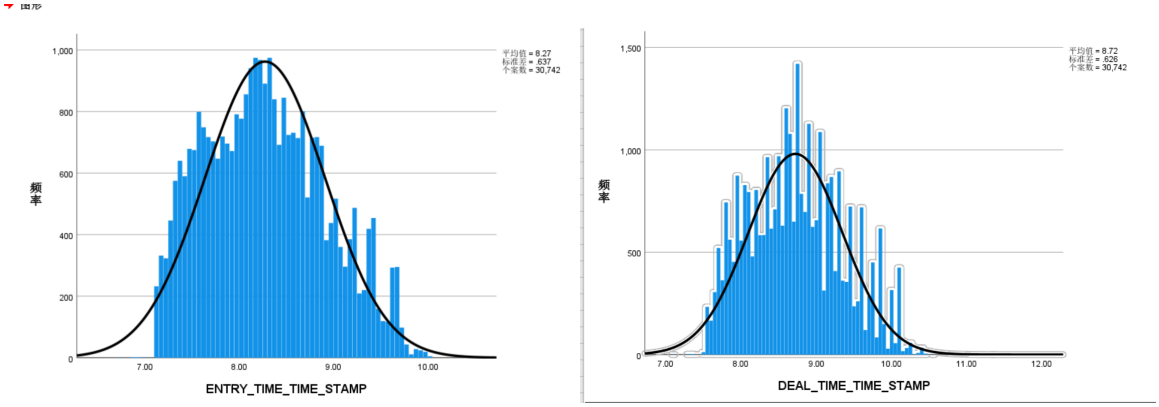
每个聚类中的个案数目

聚类	1	7495.000
	2	13070.000
	3	10177.000
有效		30742.000
缺失		.000

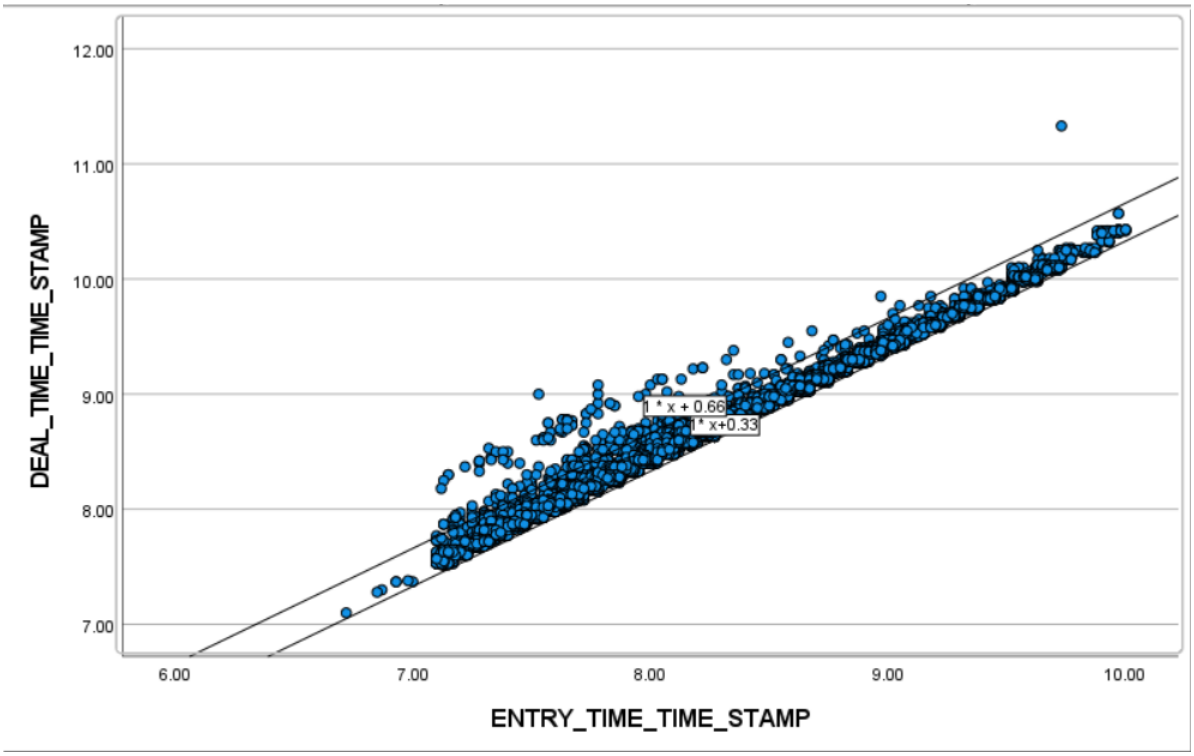
从图中可以看出分析结果聚三类，三个聚类中心对应的时刻分别为（9:08/9:34)、（8:19/8:52)、(7:34/8:02)。

- **出行规律分析**

下面两个图分别为进站时间和出站时间的频数分布直方图



所绘制的两个频数分布直方图大致拟合上述K-means动态聚类分析的结果，这三个聚类时间为行人出的高峰期，可考虑为为避免碰到过于拥挤的情况，部分人可能会选择提前或者是延后出行，从而出现另外两个次高峰。



上图两条直线的解析式分别为$Y_1 = x + 0.66$,$Y_2 = x + 0.33$。从图中可以只管的看出绝大多素的点都均匀的分布在这两条直线之间，由此可以看出绝大多数的出行耗时都在$19.8min - -39.6min$之间。

结论

通过对城市轨道交通通勤数据的分析，我们清洗了数据，匹配了流量，识别了通勤用户，并分析了他们的出行规律。

整体数据分析揭示了城市轨道交通系统的一些重要特征。我们能够看到不同站点在不同时间段的进站和出站人数，以及站点之间的流量分布。这些信息可以帮助交通部门更好地了解乘客的出行习惯，以便进行更好的资源分配和调度。

出行匹配分析,虽然大多数乘客能够成功匹配到列车，但仍然存在一些未匹配的情况，这可能是由于列车时刻表与实际情况之间存在一定差距。这个分析可以为改进列车时刻表和乘客信息提供有用的反馈。

单一车辆分析帮助我们了解不同站点上车人数的变化，以及乘客在车站之间的选择和等待行为。这对于车站布局和列车运营的优化至关重要，以满足高峰期的需求并提供更好的出行体验。

通过通勤用户的识别和出行规律分析，我们能够更深入地了解特定乘客群体的行为。这对于交通规划和服务改进非常重要。我们发现通勤用户在特定时间段内出行频繁，可能会选择避开高峰期，以获得更好的出行体验。

这些分析可以帮助城市交通部门更好地理解通勤客流，改善服务，并优化运营计划。