

# 20211202 作业

1、从以下链接下载数据：

北大网盘：<https://disk.pku.edu.cn:443/link/13D288DB58ED524005AD641F425087BB> 数据是北京地铁刷卡数据。第一行是数据标题，其含义如下：

- a) carType: SUB 是唯一取值，表明数据是地铁，本次作业可以不用；
- b) iclD: 交通卡编号，每个编号对应一张卡，已经消密脱敏；
- c) cardType: 卡类型
- d) tradeType: 交易类型
- e) UpLine: 上车线路编号
- f) UpTime: 上车时间
- g) UpStation: 上车站编号
- h) DownLine: 下车线路编号
- i) DownTime: 下车时间
- j) DownStation: 下车站编号
- k) State: 弃用

2、压缩包内的 PDF 文档是地铁线路编号和车站编号，供大家参考。

3、为便于程序调试，数据中有 Subway\_20190301\_top100000.txt(实际上是

Subway\_20180301\_top100000.txt，误为 2019)，是 Subway\_20180301.txt 的前 100000 行数据，可以用于初步调试程序，待调试完毕后再用完整数据跑，节省时间。

4、【60 分】计算乘车耗时时段内的人数。每条记录是一次乘车行为，下车时间减去上车时间，即为耗时，题目要求统计耗时 10 分钟有多少人，耗时 15 分钟有多少人。以分钟为单位，计算出每个时间段内的人数。如 1 分钟：10 人，2 分钟 20 人，……，10 分钟 500 人，30 分钟 5000 人。计算耗时，用 round()函数四舍五入到分中。对下车时间早于或等于上车时间的数据予以清除，对超过 120 分钟的数据视为 120 分钟，对于不是 20180301 的数据清除。输出时，按耗时升序排序输出到

PeopleInSubwayTime.txt，并在 Excel 中打开该文件，形成条形图。

5、【40 分】所有地铁内人数时间分布。以 10 分钟为间隔，统计地铁内人数多少。进站相当于人数增加，出站相当于人数减少。以凌晨 00:00 开始计数，此刻人数为 0。统计[00:00-00:10],[00:10-00:20]，以此类推，要求标记为：[00:00-00:10],[00:10-00:20]（以下同）。如果上下车时间不在 20180301 范围内，予以清除。按时间自然升序输出到 PeopleInSubwayCount.txt，形成折线图。

6、注意：数据处理时，要清除掉第一行，此为标题行。本数据仅用于完成作业，不得对外发布。