20181212 作业

这是本学期最后一次作业，请大家认真完成。

1、 请从百度云盘上下载数据。链接：

https://pan.baidu.com/s/1iTLXBi6SBaZsfYBY_8Vq0Q 提取码：aynp。

这是北京郊区区域内公交刷卡数据。该数据的特征是：上车产生一条数据，下车产生一条数据。交易日期和交易时间是刷卡的具体时间，上车时间实际上是根据交易日期和交易时间产生，上下车刷卡，上车时间保持一致，以判断两次刷卡属于一次交易。另外，个别卡主可能下车不刷卡，或者一次上车时产生两条数据【补刷上一次不完整交易，然后本次上车数据】，然后下车正常刷卡，也就是一次乘车行为可能产生一条数据（下车忘记刷卡）、两条数据（正常乘车）、三条数据（补一次，本次乘车两次）或者是上述乘车的中组合；

2、 根据上述数据进行清洗，留下正常数据，即清洗掉异常交易数据。根据正常数据，将一次乘车行为上下车数据合并成一条数据。包括：卡号、上车线路、上车车站、上车时间、下车车站、下车时间。合并时，用上车的线路作为上车线路、下车的上车车站和下车车站作为上下车车站，下车的交易日期和时间作为下车时间；

3、 按 15 分钟分段，统计出乘车时长，即每个卡号的乘车时长分布。比如：0-15 分钟有多少人，15-30 分钟有多少人？以此类推；

4、 从 0 时起，按 15 分钟分段，统计出每个时段内上车人数分布，下车人数分布，即 6:00-6:15 有多少人上车，有多少人下车？以此类推；

5、 第 3、第 4 小题形成的数据放在一个 Excel 之中，Excel 文件名为 Stat.xlsx，并形成柱状统计图；

6、 由于数据集较大（实际上不大），可以编写一个程序浏览前 10000 行，以了解数据集；

7、 提交内容包括：python 代码（一个或多个文件）、Excel 文件、readme.txt(对 Python 代码文件的简单说明，确保看到该说明既能了解 Python 程序文件的功能)。将上述内容压缩

成“学号_20181212.zip”的文件；

8、 本数据集还能产生诸多应用，个别同学可考虑升级为大作业，如果需要更多数据，可以和任课老师联系；

9、 作业提交截止时间：2018-12-31 23:59 前。