# 第一次实验报告

## 建立数据集

由于数据集要求10到15G大小，自己生成或者采集比较麻烦，所以选择去网上下载。在网上公开数据集中发现了一个2013年纽约出租车行车记录的数据集，下载后通过scp命令上传到了学校服务器上面。

数据集dataset.csv大概11GB大小，这个数据集是由2013年1，2，8，9，10月份的纽约出租车行车记录组成。

每行记录各个值的含义是

medallion,hack\_license,vendor\_id,rate\_code,store\_and\_fwd\_flag,pickup\_datetime,dropoff\_datetime,passenger\_count,trip\_time\_in\_secs,trip\_distance,pickup\_longitude,pickup\_latitude,dropoff\_longitude,dropoff\_latitude

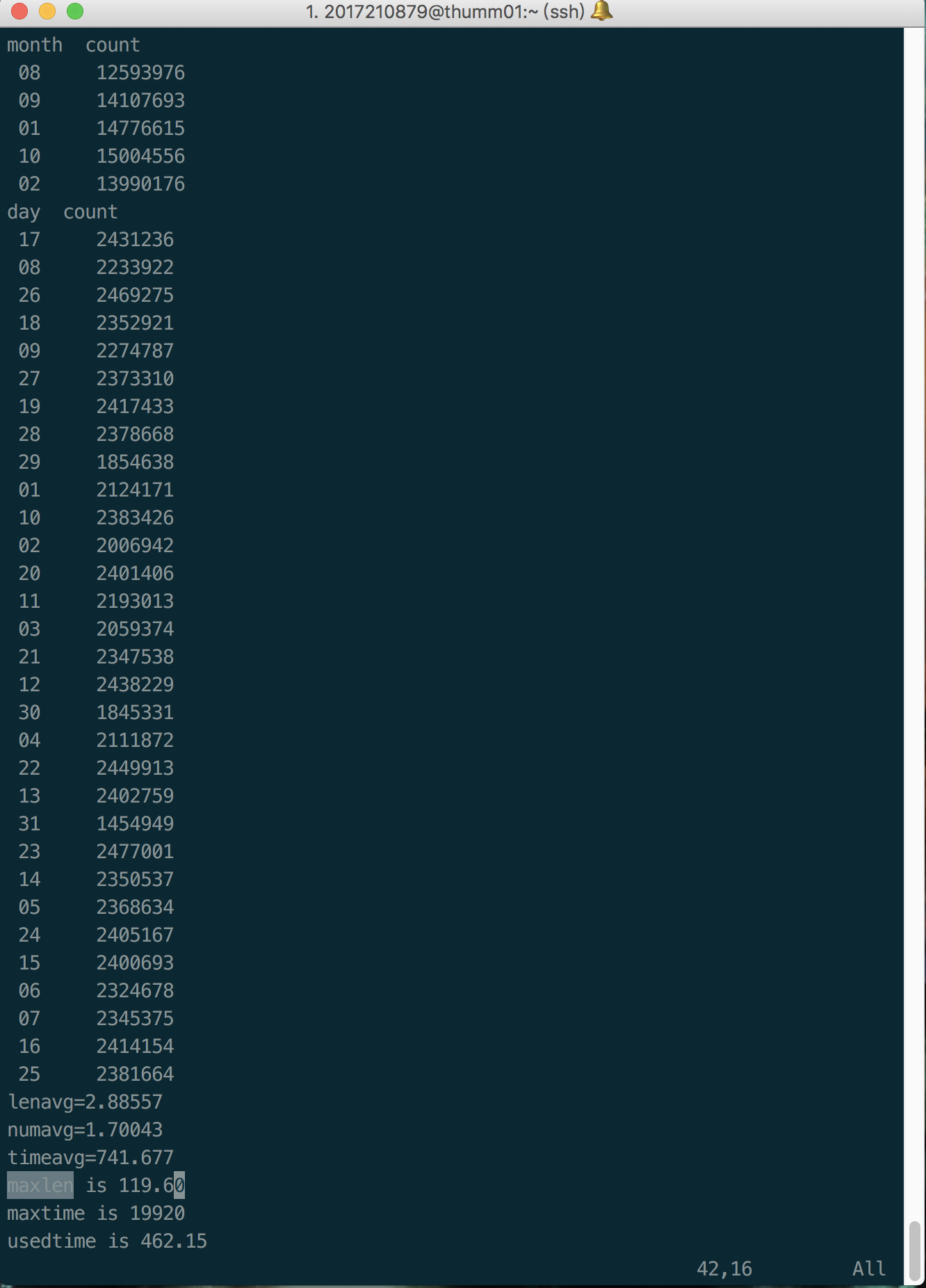
## 单线程统计

单线程测试用到了整个数据集即dataset.csv这个文件，统计是用shell脚本来完成,主要命令为awk，分别统计了最长路程、最长行驶时间、平均行驶时间、平均路程、平均乘坐人数,还有每月累计行车记录数，每日累计行车记录数和程序运行用时。

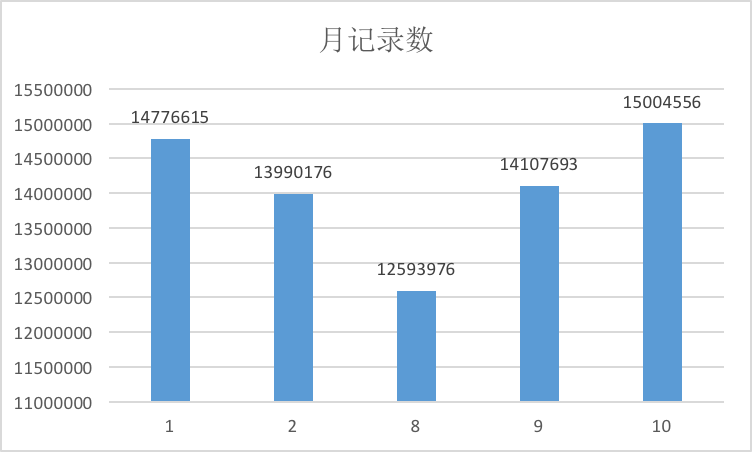
Shell脚本为附件中的work.sh。内容截图为：

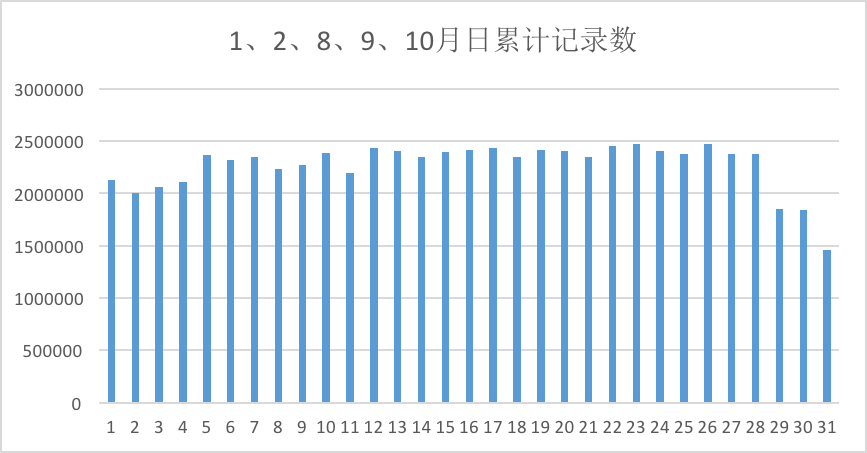


结果输出到result.txt文件中，内容截图为：



根据运行结果分别对日、月记录数用excel画图如下：





由于只统计了1，2，8，9，10这5个月的数据，并看不出一年中每个月的行车记录分布，但是由这5个月的日累计记录可以看出月末的行车记录相对较少，可能是因为某些月没有30号，31号导致的，同时单节点统计用时462.15秒。

## 多线程模拟多节点统

将dataset.csv数据集按照月份分为5个子数据集data1.csv、data2.csv、data3.csv、data4.csv、data5.csv，同时将单线程的代码改为5个线程同时统计5个子数据集，分别结果输出到result1.txt、result2.txt、result3.txt、result4.txt、result5.txt，等5个线程都运行完之后再分别读取result文件，整合出最大值，并且根据每个月的记录数占总记录数的比重来计算总体的平均值，最终结果输出到final\_result.txt。

这里由于我不是很熟悉shell代码，而且加上shell对浮点数的计算和判断比较不是很好，中途碰到很多错误，在查阅了很多资料和改代码后，在5个均值根据权重求总体均值这一直报错，所以最后结果整合的时候注释掉了这方面统计，结果整合的计算量很小不会影响到总体运算时间。

shell脚本文件为附件中的multiwork.sh。程序大体框架为：

for((i=1;i<=5;i++))

do{

分别统计每个子数据集并将结果分别输出到文件

}&

done

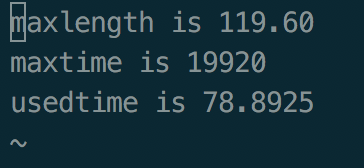
wait

{

对5个文件中的结果整合

}

查看最终结果文件final\_result.txt，内容如下：



可见结果与单线程一致，用时仅为78.8925秒，速度提升了很多。