Part1：数据介绍

本项目的数据有两部分，第一部分是纽约市自行车的交易流水表，第二部分是纽约市的天气数据。

1、 交易流水表

交易流水表指的是用户借还车的记录。以 2014 年 9 月份的数据集为例，此数据集包含 953887 条记录，15 个变量，变量说明如下：

变量名	变量含义	变量取值及说明
tripduration	旅行时长	骑行时间，数值型
starttime	出发时间	借车时间，字符串，m/d/YYYY HH:MM:SS
stoptime	结束时间	还车时间，字符串，m/d/YYYY HH:MM:SS
start station id	出发站点编号	定性变量，站点唯一编号
start station name	出发站点名称	字符串
start station latitude	出发站点纬度	数值型
start station longitude	出发站点经度	数值型
end station id	结束站点编号	定性变量，站点唯一编号
end station name	结束站点名称	字符串
end station latitude	结束站点纬度	数值型
end station longitude	结束站点经度	数值型
bikeid	自行车编号	定性变量，自行车唯一编号
usertype	用户类型	
birth year	出生年份	仅有此列存在缺失值
gender	性别	

下面对表中的部分变量进行简单描述，以形成初步认识。

1.1、 start station id，end station id 和 bikeid

这三个变量分别代表借车站点编号、还车站点编号和自行车编号。纽约市一共有 328 个站点，站点的编号多数为 3 位数字，也有少量 2 位数字和 4 位数字，分布如下：

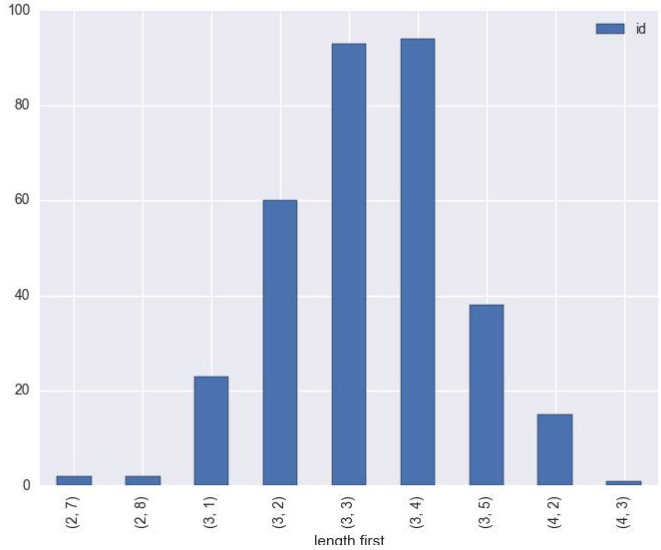

图 1 站点编号情况

纽约市公共自行车在 2014 年 9 月份的数量是 5888 辆，编号 bikeid 均为 5 位数，以 1 开头的有 4560 辆，以 2 开头的有 1328 辆。

1.2、 start station name、start station latitude、start station longitude、end station name、end station latitude 和 end station longitude

这些都是关于站点的基本信息，包括站点名称和经纬度，变量介绍已经放在了表格中。

1.3、 tripduration、starttime、stoptime

这三个变量分别是骑行时间、借车时间和还车时间。根据借还车时间我们可以得到 2014 年 9 月份纽约市一天中 24 小时的借还车分布如下，可以看出明显的早高峰和晚高峰。

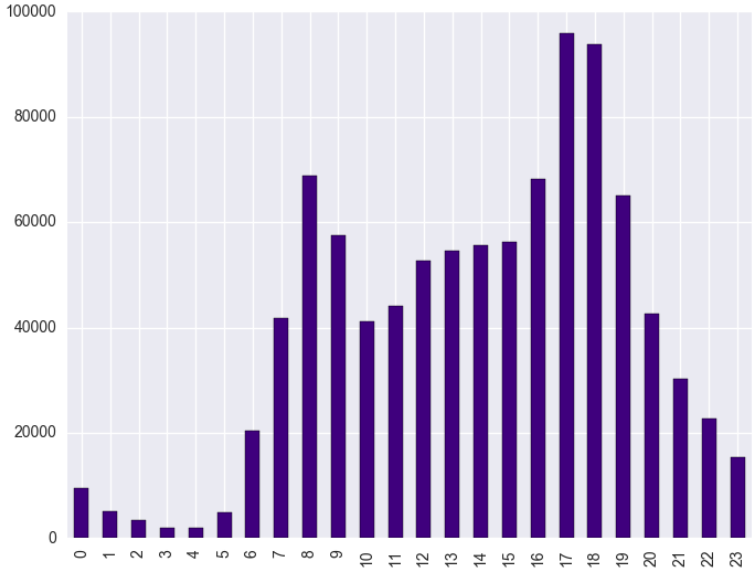

图 2 借车分布

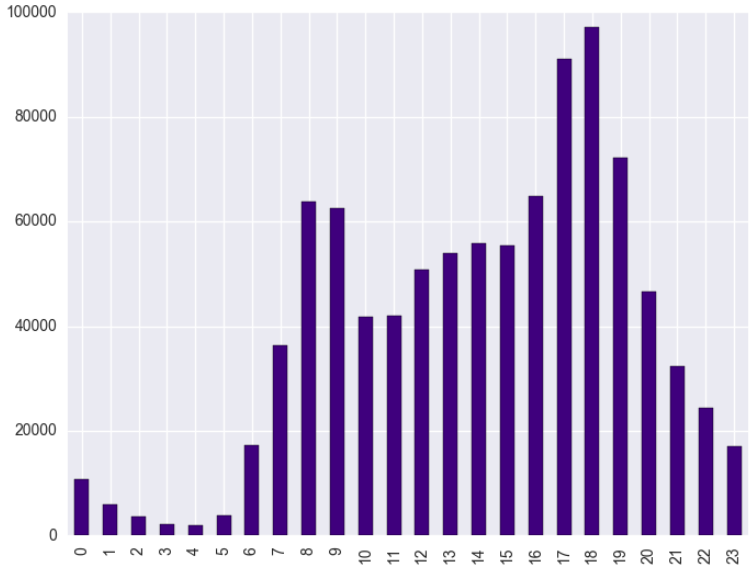

图 3 还车分布

tripduration 取完对数以后的密度估计图如下：

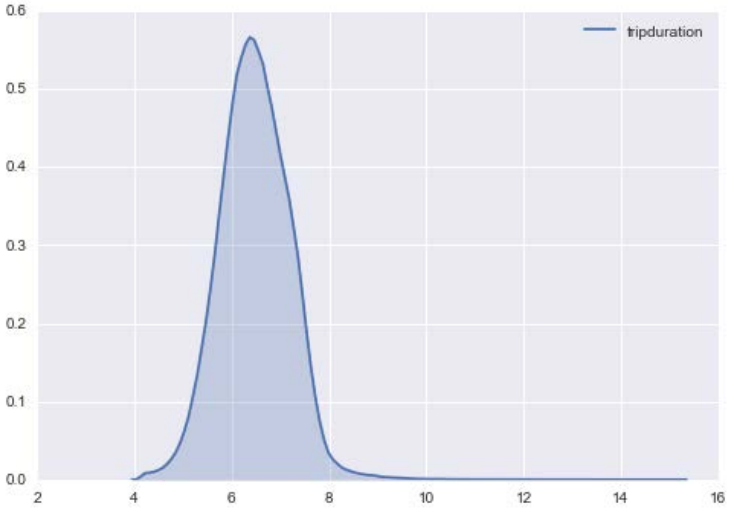

图 4：log(tripduration)的分布

可以看到，除了右边存在大量极端值以外，对数 tripduration 近似服从对称分布。

（未完）

Part 2：聚类分析

利用每小时各站点借换车数量差的数据进行聚类分析，每一个站点对应一个 24 维向量（c1,c2…c24）,ci 表示 i-1 时~i 时，该站点平均每天借还车数（还车数-借车数）的差值。对所有站点进行 kmeans 聚类聚成 4 类。

第 0 类的站点 24 小时的还车数减去借车数的情况如下。可以看到，这一类站点早高峰还车多，晚高峰借车数多。

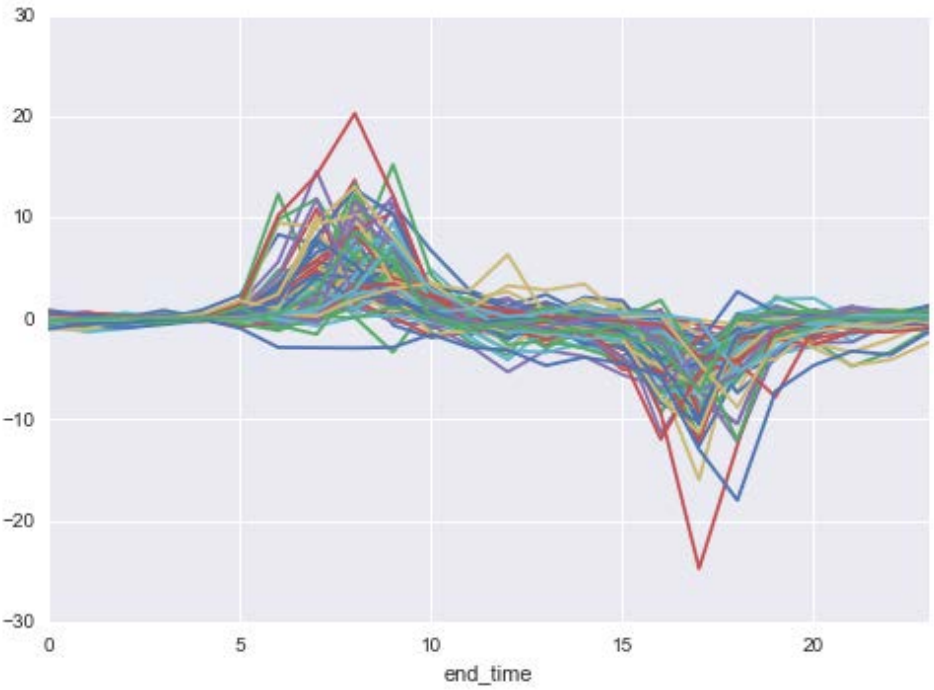

图 5：第 0 类

第 1 类的站点 24 小时的还车数减去借车数的情况如下。这一类站点恰与上相反，早高峰借车多，晚高峰还车多，应该多数位于住宅区。

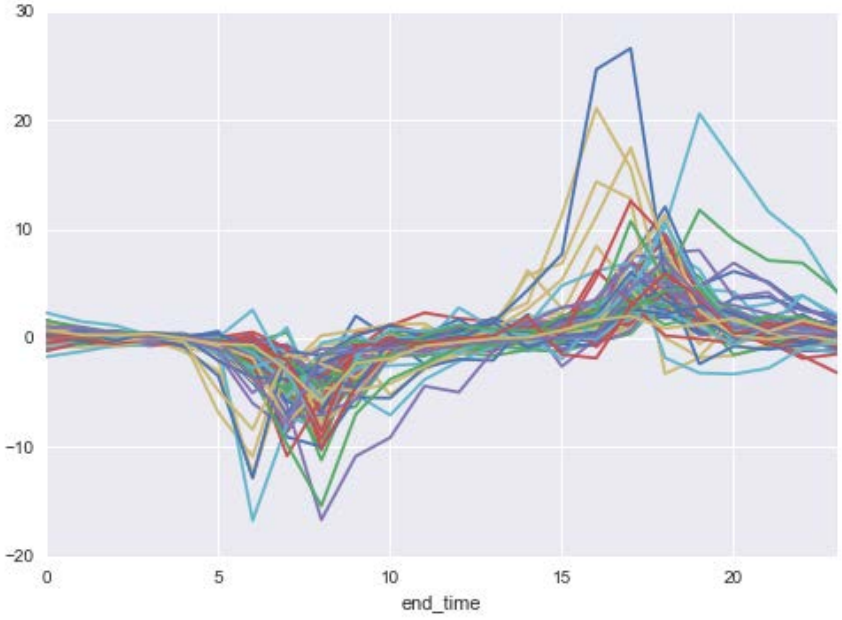

图 6：第 1 类

第 2 类的站点 24 小时的还车数减去借车数的情况如下。它没有上述两类站点那样明显的借还特征，且峰值普遍小于上两类站点，应该是借还车动态平衡的站点或者没

有什么明显规律的站点。

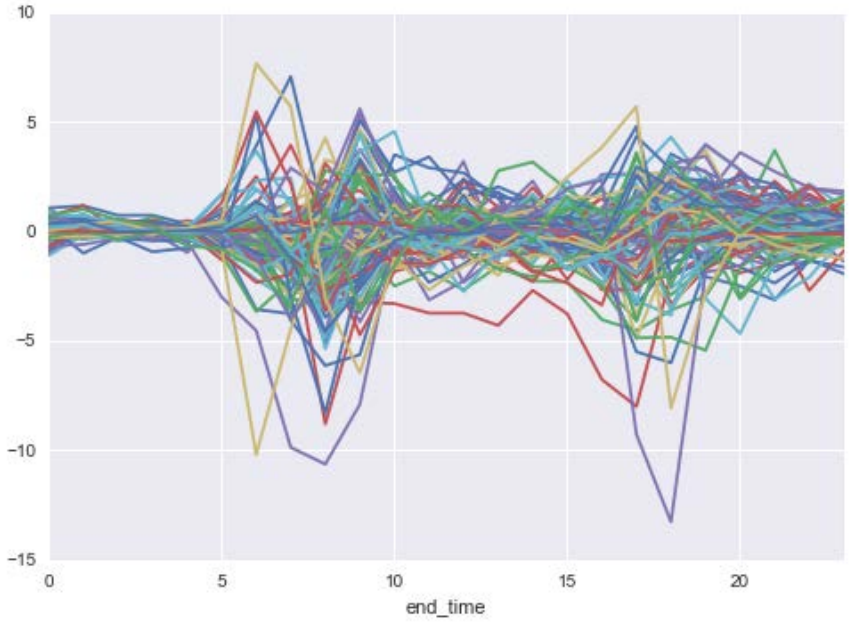

图 7：第 2 类

第 3 类的站点 24 小时的还车数减去借车数的情况如下。这一类只有一个站点，是 521 号站点。它其实也应该是属于第 1 类早高峰借车多晚高峰还车多的站点，但是由于它的峰值远远大于其他站点，所以聚类的时候把它分离出来了。我们以后可以将这个站点作为重要站点进行研究。