1. 主要工作。

我们做的主要工作有三个。

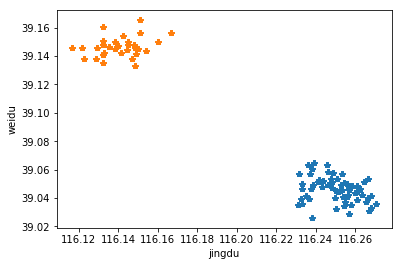
一是通过描出了k-means聚类后各被使用的共享单车的分布情况，我们将所有单车划分为两个位置。

二是分析了不同区域 共享单车的使用在一些特征上存在的差异，并用lda进行了有监督学习，验证我们的划分位置的正确性。也说明了根据不同区域人们的使用特点来进行共享单车投放的正确性。

三是通过分析 单车投放数量、每辆单车平均每天使用时间之间的关系确定了最优的单车投放数量。

1. 区域位置聚类

2.1、使用K-means算法，根据开始经度、开始维度、结束经度、结束纬度四个特征进行聚类，得到下面两个区域，分别记为area0 和 area1.



注Kmeans聚类介绍：

kmeans的计算方法如下：

1 随机选取k个中心点

2 遍历所有数据，将每个数据划分到最近的中心点中

3 计算每个聚类的平均值，并作为新的中心点

4 重复2-3，直到这k个中线点不再变化（收敛了），或执行了足够多的迭代

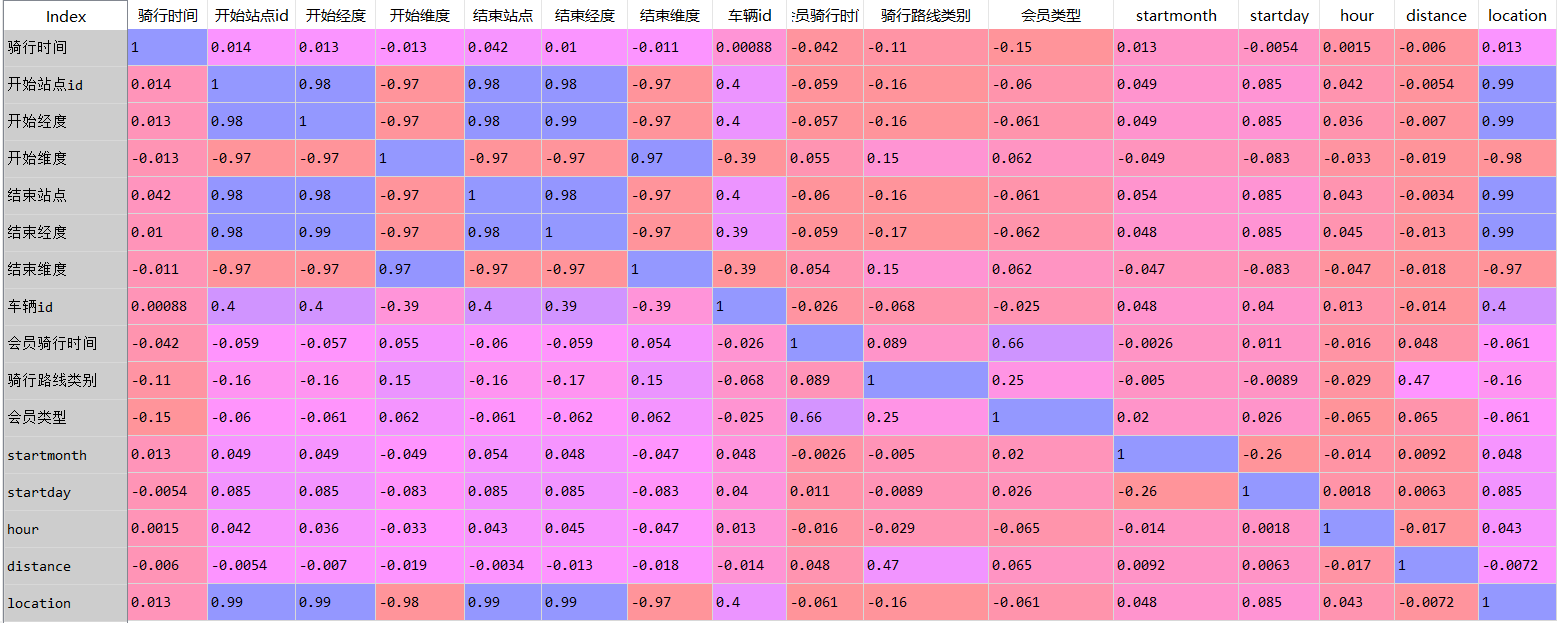
时间复杂度：O(I\*n\*k\*m)

空间复杂度：O(n\*m)

其中m为每个元素字段个数，n为数据量，I为跌打个数。一般I,k,m均可认为是常量，所以时间和空间复杂度可以简化为O(n)，即线性的。

2.2 通过函数我们将根据经度和维度计算出骑行距离这一特征，根据这一特征把骑行距离大于10KM的样本删除。

2.3各原始指标相关系数矩阵。

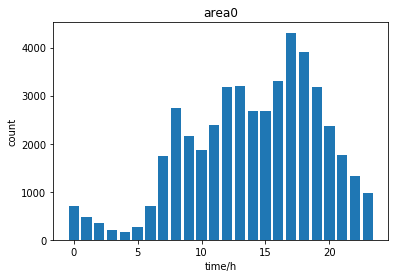


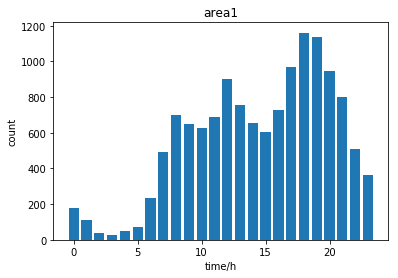
我们从中挑取与Location相关系数高的指标进行继续分析。

1. 不同区域差异性验证以及LDA分类模型

我们先分析了area0和area1两个区域的在汽车次数、会员类型、路线类型、骑行距离、平均骑行时间等特征上的差异。随后采用LDA分类模型，将上述变量作为特征输入进模型，验证了两个区域的差异性。

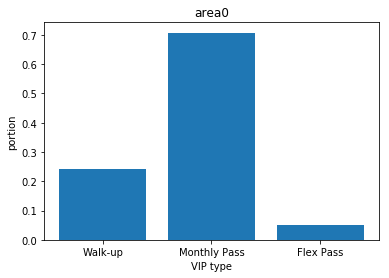
3.1两区域 骑车次数随时间变化。

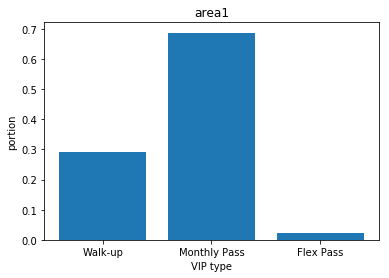




结论：共享单车使用量的高峰分别出现在早上七点、中午十二点、和下午五六点。主要差异在area1的峰值在每天下午6点，而area0的峰值在每天下午5点。并且area0的单车使用次数明显更多，所以投放时应该根据地区进行区别，并且在area0投放更多单车。

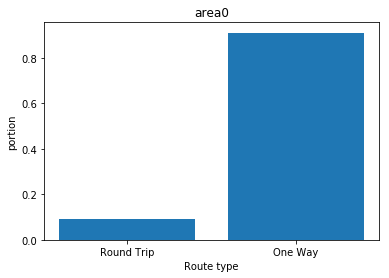
3.2两区域 会员类型差异。

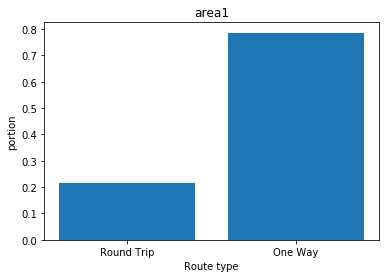




Area0 的 Flex Pass更多。

3.3两区域 路线类型差异。

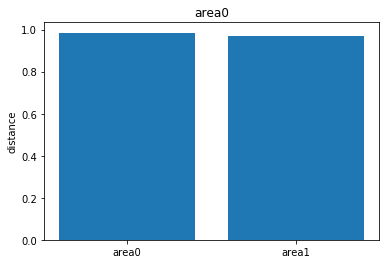




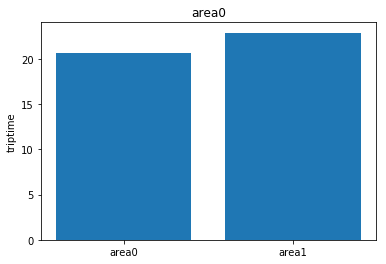
Area1骑行路线Round trip更多。

3.4两区域 骑行距离

差异不大，都接近一公里。



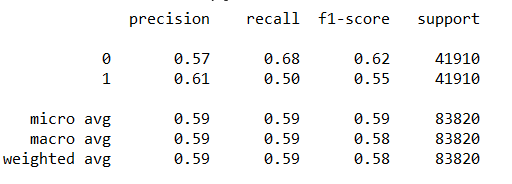
3.5、两地区骑行时间。



Area1平均骑行时间更高。

总结：所有的差异都是在投放共享单车时需要考虑的因素，接下来将采用LDA线性判别模型，根据这些特征判断区域位置。

3.6、LDA分类下召回率(recall)与精确度(precision)结果。可见效果接近60%。说明我们建立的模型对位置投放的选择是有效的。



1. 最优投放比例判定。

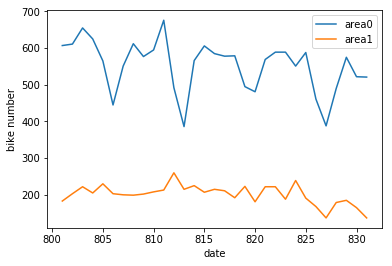
4.1 假设。

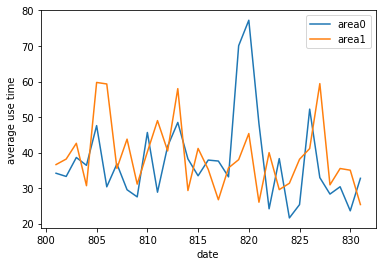
我们要最大化的目标是每天被使用的共享单车的总时间。

我们假设关系：总时间=每天的共享单车使用数量\*共享单车每天的使用时间 成立，并且这一关系是稳定的。

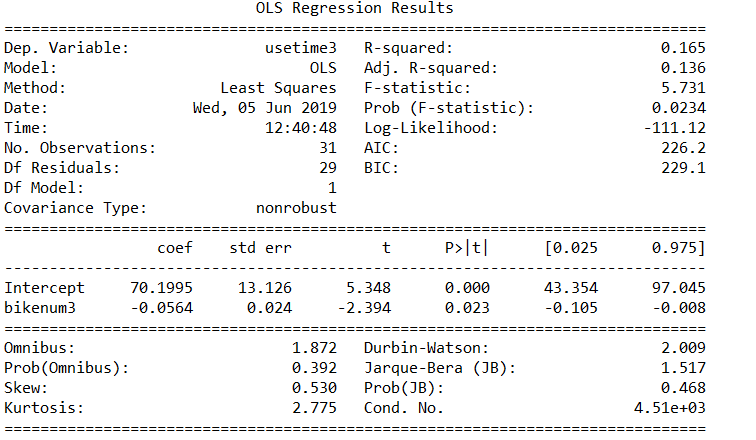
假设共享单车使用数量与共享单车每天的使用时间之间的负相关关系是线性的。

4.2 数量与使用时间随时间变化关系。





我们选取了一个月的时间，从第一幅图可以到area0每天的自行车使用量稳定在600辆左右，area1稳定在200辆左右。结合两幅图可以看到，当自行车每天的使用增加时，每辆自行车每天的使用时间在减少。我们分别对两个变量计算相关系数，发现area0自行车使用量与每辆自行车每天使用时间的相关系数为-0.41，area1自行车使用量与每辆自行车每天使用时间的相关系数为-0.29。因此两个区域都呈负相关关系。我们将通过线性回归刻画精准的模型。以area0为例，回归结果如下图所示：



建立的线性方程为

每天每辆自行车使用时间（设为y）=-0.0564\*每天自行车的使用辆数（x）+70.1995，这个方程是在5%的置信水平下显著的。

因为每天的自行车使用总时间U=x\*y，因此U=-0.0564x^2+70.1995x。

简单求解可知当x=622时，U取最大值，因此area0的最优投放量为622。

同理我们得到 area1的线性方程为:

y=-0.0493x+27.6613

area1的最优投放量为280。

目前area0的平均车辆数为552，area1的平均车辆数为200。因此我们的建议是两个地区都应该多增加车辆。