HW02

张侨伦 5140809064

代码说明：

第一题：1Wine.R

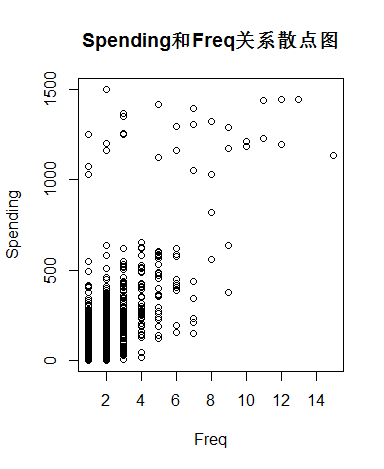
第二题：2Tayko.R

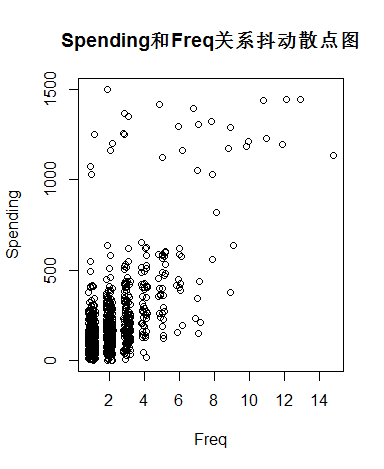
第三题：3fgl.R

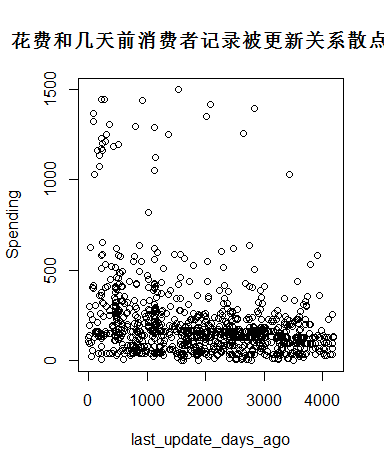
1. 主成分分析
   1. 如果不对数据进行标准化处理，从结果看，1个主成分即可捕捉到99.81%的总方差。
   2. 从结果看，要用8个主成分才能捕捉到90%以上的的总方差。（8个变量捕捉到0.92018%的总方差）
   3. 进行标准化处理比较好。在PCA中进行标准化处理，可以移除量纲对结果的影响。否则某些量纲会导致其取值范围超过其他变量，对结果造成误差。
2. 线性回归
   1. 在软件目录 source\_a 购买了软件的消费者平均消费金额是247.8985

在软件目录 source\_b 购买了软件的消费者平均消费金额是197.6111

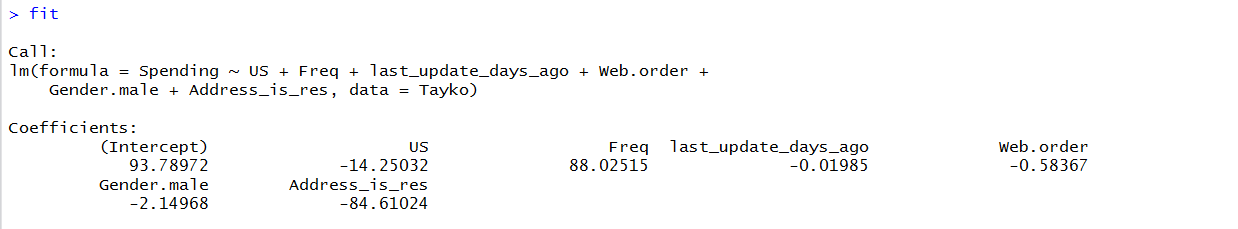
* 1. 图如下



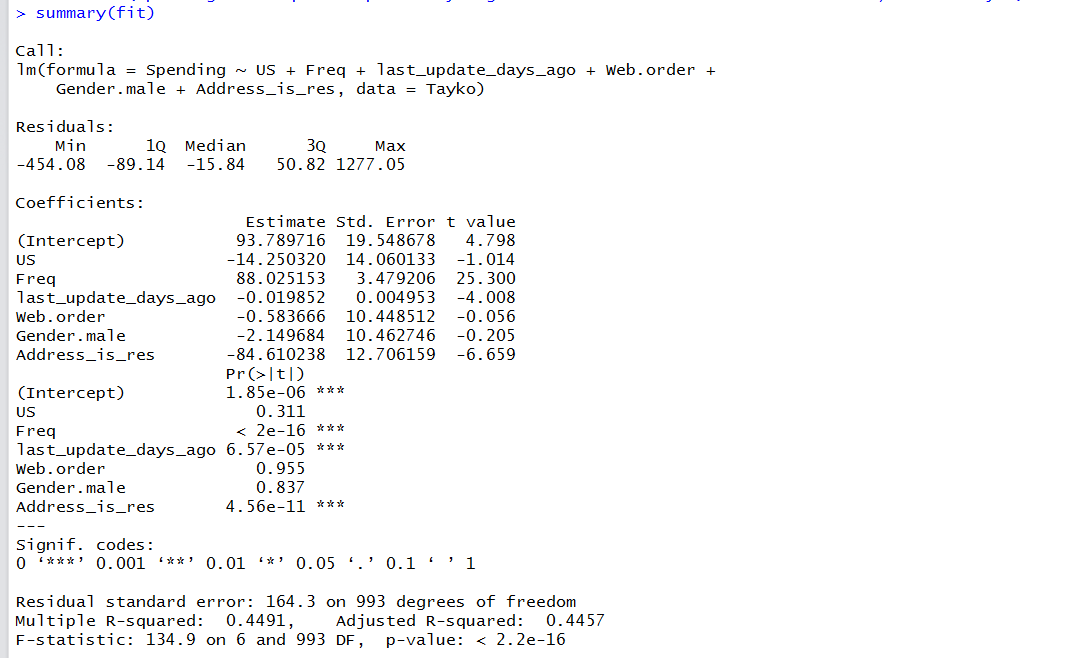




* 1. 线性回归的结果



也就是：Spending = 93.78972 - 14.25032\*US + 88.02515\*Freq - 0.01985\*last\_update\_days\_ago - 0.58367\*Web.order – 2.14968\*Gender.male – 84.61024\*Address\_is\_res。



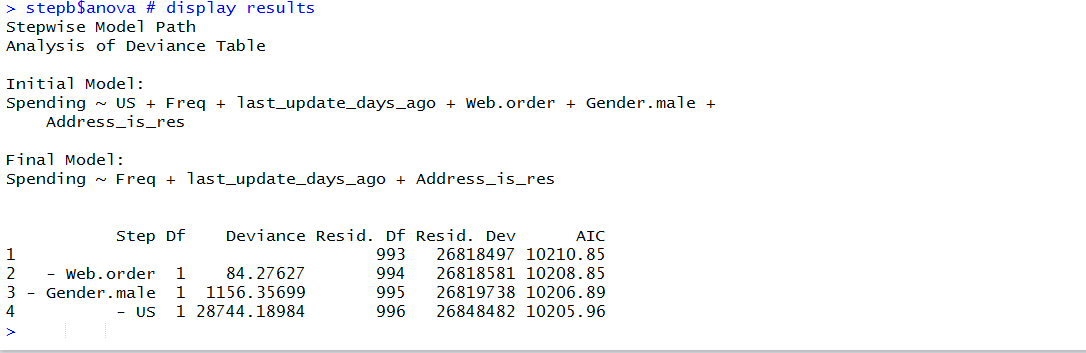
使用summary（fit）进行分析，可知Freq，last\_update\_days\_ago和Address\_is\_res这三个变量在5%的显著性水平上显著。

* 1. 根据输出结果可得最后被选择的变量为Freq，last\_update\_days\_ago和Address\_is\_res，模型的形式如下

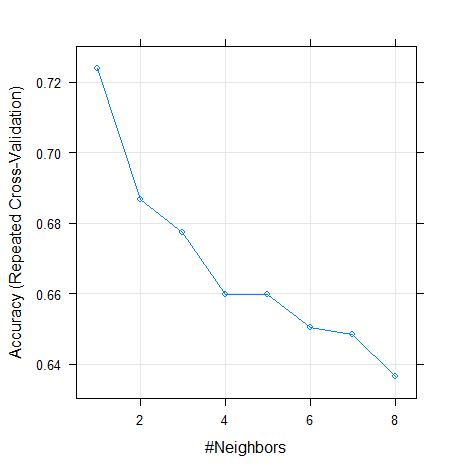
Final Model:

Spending ~ Freq + last\_update\_days\_ago + Address\_is\_res

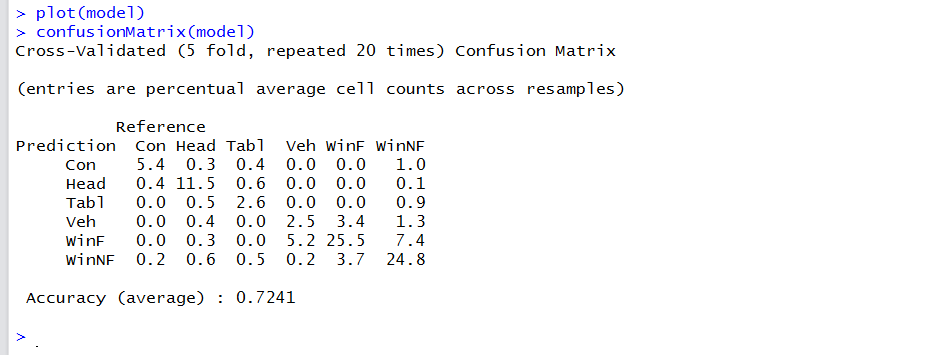
截图如下：



1. KNN
   1. 训练的结果最佳k是1



混淆矩阵如下



* 1. 可以对数据进行分段，给出在不同数据段里面的不同玻璃类型的数量进行分类。

1. 选择地址需要考虑两个因素：
2. 如果精确到小区，那么小区内数据的数量是否足够，如果小区内数据数量不够，那么选择精确到小区会使某些量的比例变为100%，反而会使结果不准确。
3. 以后推广的时候是否是要进入小区推广。如果推广的区域可以缩小到这种程度，那么选择精确到小区是有意义的。

综合上述两点考虑：如果是以后推广要精确到小区，并且数据数量充足，那么选择精确到小区。否则选择精确到县市一级。