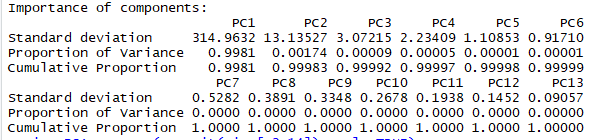
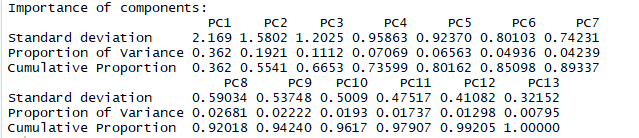
**Q1**

（1）如果不对数据进行标准化处理，由R脚本BA\_HW03\_Q1可得



故只需一个主成分变量即可捕捉到90%以上总方差。

（2）如果对数据进行标准化处理，由R脚本BA\_HW03\_Q1可得



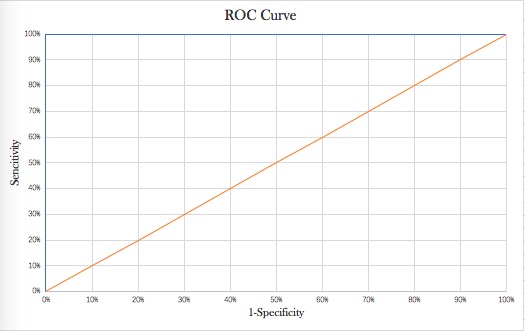
故需要8个主成分变量才能捕捉到90%以上总方差。

（3）第二种方法好，方法二消除了量纲的影响，是表达信息量的有效方式，避免了某一变量量纲选择过小，而成为主要因素。

**Q2**

****

****

****

**Q3**

此题的答案已在BA\_HW03\_Q3的R文件代码中以注释方式给出，但为了更清晰地呈现此题的文字性答案，我们将其在此文件中进行了总结。

1. 得出的两个数据框中，source\_a为1时，Spending对应的数值为在软件目录source\_a购买了软件的消费者平均消费金额，即247.8985；source\_b为1时，Spending对应的数值为在软件目录source\_b购买了软件的消费者平均消费金额，即197.6111

b) Freq, last\_update\_days\_ago, address\_is\_res在显著性水平5%上显著（至少得到一颗星）

1. 最终模型表达式为Spending ~ Freq + Address\_is\_res + last\_update\_days\_ago + source\_r + source\_a + source\_u + X1st\_update\_days\_ago
2. 最终模型表达式为：Spending ~ Freq + Address\_is\_res + last\_update\_days\_ago

**Q4**

1. 最佳k值为1。不同k值下KNN模型的预测准确率的图以及最终的模型的混淆矩阵在运行R语言脚本BA\_HW03\_Q4后可以得到。
2. 若想使用Naïve Bayes方法处理相同的任务，需要将fgl.csv中用来判断玻璃类型的9个特征数据（RI, Na, Mg, Al, Si, K, Ca, Ba, Fe）各自按一定的步长划分成几个区间，再将这些数值型变量转换成类别型变量。以Mg元素为例，可以将Mg元素含量以0.5的步长划分成0~0.5, 0.5~1.0, 1.0~1.5, 1.5~2.0, 2.0~2.5, 2.5~3.0, 3.0~3.5, 3.5~4.0, 4.0~4.5 这九个区间，然后将其转化成类别型变量。同理，其他特征也需经过相同的处理，但步长不一定要是0.5, 要根据每种特征的数值区间做出调整，例如，此文件即fgl.csv中所统计的所有玻璃的Fe元素含量的数值范围为[0, 0.51], 此时若仍将步长还设置为0.5, 则不够恰当。因为此时Fe元素含量只能分成两个区间，即[0, 0.5]、[0.5, 0.51]，并且通过文件得知在[0.5, 0.51]区间上的Fe元素含量数据只有一个，如此一来很难较为准确地得出Fe元素含量对玻璃类型的影响。