数据分析与可视化 第三-四次作业

说明：

1. 作业上交日期为 2022 年 12 月 22 日中午 12 点，将作业发给助教邮箱，逾期将无法提交，视为放弃此次作业。

2. 请将作业保存成 pdf（不要用 word）上传；文件名为 XXXX_YY_Ass5.pdf，其中 XXXX 为你的学号名，YY 为你的名字。

3. 作业中的每个问题，涉及到代码问题的，都需要在该题目位置附上相应的 R（Rstudio）源代码，不要用截屏记录代码，否则视为没有作答（助教将复制并运行你的源代码）。

4. 作业中涉及到简答问题的，请给出你的答案和理由。

5. 涉及到生成随机数的问题，请设定“种子”；其中 seed 的值请设定为你的 9 位数学号，请在合适的位置用如下代码： set.seed(XXXXXXXX)

辛苦大家在圣诞和新年期间完成此次作业，预祝大家新年快乐！期末取得好成绩！

通过如下命令，加载数据集 wine：

```
wineUrl <- 'http://archive.ics.uci.edu/ml/machine-learning-
databases/wine/wine.data'
wine <- read.table(wineUrl, header=FALSE, sep=',',
                   stringsAsFactors=FALSE,
                   col.names=c('Cultivar', 'Alcohol', 'Malic.acid', 'Ash',
                           'Alcalinity.of.ash','Magnesium','Total.phenols',
                           'Flavanoids', 'Nonflavanoid.phenols',
                           'Proanthocyanin', 'Color.intensity',
                           'Hue', 'OD280.OD315.of.diluted.wines', 'Proline'))

dim(wine)
wineTrain <- wine[, which(names(wine) != "Cultivar")]
```

该数据集有 14 个变量，178 条关于酒的记录；其中，第一列 Cultivar 为一个多分类指标的标签。（该数据集是一个开源数据，有兴趣的同学可以通过数据网站查看每个变量的具体含义）。进一步，通过如下命令，生成去标签后的训练样本集 wineTrain（注意，这里我们没有选 random sample 来做后续模型估计）。

(a) 针对变量 Alcohol, Malic. Acid, Magnesium, Total.phenols, Color.intensity, Hue，进行描述性统计分析。请用一幅图内展示每个变量在标准化之后的箱型图，选用适当的颜色以及图片的主标题和横纵坐标的标题。从图中，有显示出可能的异常值吗？如果存在，请找出其在原始数据集中的行数。

(b) 请选用 ggplot2 中适当的图表类型，展示每个变量的样本分布是否有偏，以及相关图标的格式，如颜色，标题，图例等等。

(c) 请选用合适的方式，计算并展示 wineTrain 数据集中所有变量的两两相关性。你哪些变量之间的相关性比较高？

(d) 设定随机数种子为你的学号，通过 k-means 方式进行聚类，其中，中心的个数定为 3 个。请通过合适的图表(建议 ggplot2 相关图表)，展示你的聚类效果。你认为 kmeans 的聚类效果如何？

(e) 请通过 silhouette 统计量和 gap 统计量，分别决定 cluster 组的个数的最优值，并将你得到的结果进行展示。两种方法给出的最优组数是否相同？如果不同，你觉得哪个更合理。其中 nstart 设定为 25.此时，组的个数与原始数据集中 wine 中的变量 Cultivar 的可能取值相比，是否相同？

(f) 设定随机数种子为你的学号，通过 k-means 方式进行聚类，其中，中心的个数定为 3 个。根据每个个体的分组情况，与其对应的标签相比，吻合情况如何？你可以展示一下 confusion matrix。

(g) 请展示通过层次聚类 hclust 函数进行聚类的结果，并通过合适的可视化方式进行展示。该方法与 k-means 相比，效果如何？

(h) 请通过任何一种你学过的分类方法，将 wine 进行分类，其中 Cultivar 作为响应变量，得到每个样本点的分类的预测值。对比 k-means 的 k 取 3 的时候的聚类效果，你认为通过 kmeans 方法聚类后用来做标签的预测效果怎么样？哪个更精准？你觉得可能的原因有哪些？