STA 135 Final project

Due on **June 10th** 11:59 pm PST. Please upload your report to Canvas.

1 Data sets

Three data sets are as follows.

Data set one: Glass Identification Database

The data set can be obtained from the R package `mlbench`. More specifically, after installing the package, you can obtain the data sets by running

```
library(mlbench)
data(Glass)
```

For this data set, people are usually interested in predict the glass type from chemical properties.

Data set two: Pima Indians Diabetes Database

The data set can be obtained from the R package `mlbench`. More specifically, after installing the package, you can obtain the data sets by running

```
library(mlbench)
data(PimaIndiansDiabetes)
```

For this data set, people are usually interested in predicting the onset of diabetes in female Pima Indians from medical record data.

Data set three: Air pollution in US cities

The data set can be obtained from the R package `HSAUR2`. More specifically, after installing the package, you can obtain the data sets by running

```
library(HSAUR2)
data(USairpollution)
```

For this data set, people are usually interested in understanding which of the climate and human ecology variables are the best predictors of the degree of the air pollution in a city as measured by the sulphur dioxide content of the air. Some people are even interested in ranking the performance of the cities.

2 Some requirements

You will analyze a data set of your choosing from one of the above three data sets, using some techniques from multivariate data analysis such as LDA, QDA, PCA or factor analysis. You should write a concise report summarizing your analysis. The report **should be no longer than four (typed) pages**, not counting any R output, graphs, etc., which you may wish to include as support or illustration for your analysis. The style of the report is up to you, but the best reports will address many of the questions and details studied in class when we discussed the relevant type of analysis.

Some things to include should be:

1. A description of the data set, with some background. Here I only provide the data. You may try to Google some related background on the data set you picked and summarize some useful information.

2. Visualization of the data sets before applying the techniques.

3. Applying some techniques discussed in class to analyze the data set. In your report, you need to provide detailed analysis based on the R outputs. For example, if you are using PCA, what are the PCs? How many PCs you want to choose? Can I interpret the PCs and use them for ranking? For another example, if you are using LDA/QDA, which method do you choose? What is the decision boundary? What is the confusion matrix? Does the splitting of training and test data sets affect my results? **Do not only enclose the R outputs, you need to focus on the analysis. Using words, numbers, equations and formulas to explain your results!**

4. Your overall conclusions about the data, based on your analysis.

5. In the end of the report, please specify the contribution of the group members.

I will not provide any proposal questions for you to answer. You need to deliver some messages regarding the data set to me based on your analysis. A key trick to write a good report: Image that you are writing something to your manager/boss/supervisor/client who does not know too much about the data sets and multivariate analysis.

3 Submission

The project is due on **June 10th** 11:59 pm PST. Please upload your report to Canvas. Everyone needs to submit the report to Canvas (of course if you are working as a group, all groups members have one report but each of you still needs to submit it.).