# Exercise on Cluster Analysis with R (Rmd). AMS 580

# Name:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_SBU ID:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Dear all, please study the lecture notes and lab notes, and learn how to do cluster analysis and principal component analysis (PCA) with R (Rmd) though this exercise. It is due Monday (4/24) before class. Please upload your solutions to Brightspace. Please also refer to the following website for reference: <http://www.sthda.com/english/articles/25-clusteranalysis-in-r-practical-guide/>

1. The pimaindiansdiabetes2 data contain information about diabetes of the Pima Indians (<https://en.wikipedia.org/wiki/Pima_people>). The data set contains 768 observations on 9 variables:

* pregnant -- Number of times pregnant (numeric);
* glucose – 2-Hour Plasma glucose concentration (numeric);
* pressure -- Diastolic blood pressure (mm Hg) (numeric);
* triceps -- Triceps skin fold thickness (mm) (numeric);
* insulin -- 2-Hour Serum insulin (mu U/ml)(numeric);
* mass -- Body mass index (weight in kg/(height in m)^2) (numeric);
* pedigree -- Diabetes pedigree function (numeric);
* age -- Age (years) (numeric);
* diabetes -- class variable: diabetic or not (factor with 2 levels: neg and pos).

One can load this data set as follows:

install.packages("mlbench")

data("PimaIndiansDiabetes2", package = "mlbench")

pima <- na.omit(PimaIndiansDiabetes2)

head(pima)

Please write up the entire Rmd code necessary to answer the following questions.

You may refer to the following website for the plain or fancy R procedures (either one is acceptable):

<https://rstudio-pubs-static.s3.amazonaws.com/33876_1d7794d9a86647ca90c4f182df93f0e8.html>

<https://www.datanovia.com/en/blog/types-of-clustering-methods-overview-and-quick-start-r-code/>

Our goal is to predict the diabetes diagnostics (although it has already been given) based on the data given. Subsequently we wish to compare our data-driven clustering to that of the true diabetes diagnoses. For each of the following six clustering methods (K-mean; Hierarchical: Ward, Single-linkage, Complete-linkage, Average-linkage, Centroid), we need you to:

1. Perform the clustering analysis;
2. Draw scree-plot to see whether two clusters is reasonable or not;
3. Show the 2D representation of the Cluster solution;
4. For K-means clustering, build a confusion matrix to evaluate the clustering performance;

For Hierarchical clustering, draw the dendrogram showing the 2 clusters obtained;

1. Finally, make a comparison of all six clustering methods, and rank their performance for the given problem.
2. Please write up the entire Rmd code necessary to answer the following questions.

Since this is the first time you run the PCA analysis, you may simply follow the following website for the step-by-step R procedures – and use the same data set ‘mtcars’:

<https://www.datacamp.com/community/tutorials/pca-analysis-r>

(\* Note: Once you have finished running procedures in the above designated website, please do study the following website for some alternative procedures and more details: <http://www.sthda.com/english/articles/31-principal-component-methods-in-r-practical-guide/118-principal-component-analysis-in-r-prcomp-vs-princomp/>)

You will need to install the following packages for the code to work in the designated website:

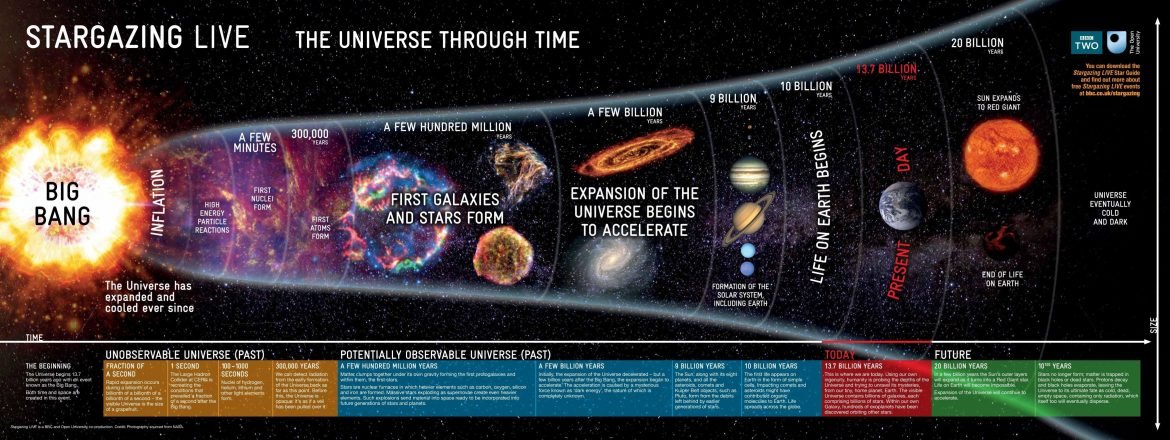
install.packages("devtools")

library(devtools)

install\_github("vqv/ggbiplot")

library(ggbiplot)

1. Please compute the Principal Components (PC) and print out a summary of the analysis – and in particular, please point out what percentage of the variations each PC would explain.
2. Please make a biplot, which includes both the position of each sample in terms of PC1 and PC2 and also will show you how the initial variables map onto this.
3. Now you will add the origin of each of the cars by putting them into one of three groups: US, Japanese and European cars. You shall then visualize the biplot by setting the ellipse argument to be TRUE, which will draw an ellipse around each group.
4. Last but not the least, we shall print out the PC1 as linear combinations of the original variables.



<https://steemit.com/space/@asaad/how-big-is-our-universe>