# Homework 3. Clustering practice

Your Name

2022-09-21

## Part 1. USArrests Dataset and Hierarchical Clustering (20 Points)

Consider the "USArrests" data. It is a built-in dataset you may directly get in RStudio. Perform hierarchical clustering on the observations (states) and answer the following questions.

```
head(USArrests)
```

```
##            Murder Assault UrbanPop Rape
## Alabama      13.2     236       58 21.2
## Alaska       10.0     263       48 44.5
## Arizona       8.1     294       80 31.0
## Arkansas      8.8     190       50 19.5
## California    9.0     276       91 40.6
## Colorado      7.9     204       78 38.7
```

**Q1.1.** Using hierarchical clustering with complete linkage and Euclidean distance, cluster the states. (5 points)

**Q1.2.** Cut the dendrogram at a height that results in three distinct clusters. Interpret the clusters. Which states belong to which clusters? (5 points)

**Q1.3** Hierarchically cluster the states using complete linkage and Euclidean distance, after scaling the variables to have standard deviation one. (5 points)

**Q1.4** What effect does scaling the variables have on the hierarchical clustering obtained? In your opinion, should the variables be scaled before the inter-observation dissimilarities are computed? Provide a justification for your answer. *(5 points)*

*Answer:...*

## Part 2. Market Segmentation (80 Points)

An advertisement division of large club store needs to perform customer analysis the store customers in order to create a segmentation for more targeted marketing campaign

You task is to identify similar customers and characterize them (at least some of them). In other word perform clustering and identify customers segmentation.

This data-set is derived from https://www.kaggle.com/imakash3011/customer-personality-analysis

```
Colomns description:
People
  ID: Customer's unique identifier
  Year_Birth: Customer's birth year
  Education: Customer's education level
  Marital_Status: Customer's marital status
```

```
  Income: Customer's yearly household income
  Kidhome: Number of children in customer's household
  Teenhome: Number of teenagers in customer's household
  Dt_Customer: Date of customer's enrollment with the company
  Recency: Number of days since customer's last purchase
  Complain: 1 if the customer complained in the last 2 years, 0 otherwise

Products

  MntWines: Amount spent on wine in last 2 years
  MntFruits: Amount spent on fruits in last 2 years
  MntMeatProducts: Amount spent on meat in last 2 years
  MntFishProducts: Amount spent on fish in last 2 years
  MntSweetProducts: Amount spent on sweets in last 2 years
  MntGoldProds: Amount spent on gold in last 2 years

Place
  NumWebPurchases: Number of purchases made through the company's website
  NumStorePurchases: Number of purchases made directly in stores
```

Assume that data was current on 2014-07-01

**Q2.1.** Read Dataset and Data Conversion to Proper Data Format *(12 points)*

Read "m_marketing_campaign.csv" using `data.table::fread` command, examine the data.

```
# fread m_marketing_campaign.csv and save it as df
```

```
# Convert Year_Birth to Age (assume that current date is 2014-07-01)



# Dt_Customer is a date (it is still character), convert it to membership days (name it MembershipDays)
# hint: note European date format, use as.Date with proper format argument
```

```
# Summarize Education column (use table function)


# Lets treat Education column as ordinal categories and use simple levels for distance calculations
# Assuming following order of degrees:
#    HighSchool, Associate, Bachelor, Master, PhD
# factorize Education column (hint: use factor function with above levels)
```

```
# Summarize Education column (use table function)


# Lets convert single Marital_Status categories for 5 separate binary categories
# Divorced, Married, Single, Together and Widow, the value will be 1 if customer
# is in that category and 0 if customer is not
# hint: use dummyVars from caret package, model.matrix or simple comparison (there are only 5 groups)
```

```
# lets remove columns which we will no longer use:
# remove ID, Year_Birth, Dt_Customer, Marital_Status
# and save it as df_sel


# Convert Education to integers
```

```
# hint: use as.integer function, if you use factor function earlier
# properly then HighSchool will be 1, Associate will be 2 and so on)
```

```
# lets scale
# run scale function on df_sel and save it as df_scale
# that will be our scaled values which we will use for analysis
```

## PCA

**Q2.2.** Run PCA *(12 points)*

```
# Run PCA on df_scale, make biplot and scree plot/percentage variance explained plot
# save as pc_out, we will use pc_out$x[,1] and pc_out$x[,2] later for plotting
```

**Q2.3** Comment on observation (any visible distinct clusters?) *(4 points)*

1st PC has highest variance more than two times larger than second, however it is still around 25%. Second and higher PC has PVE less than 10%. Slow grow in cumulative PVE shows that large number of features is needed.

May be two clusters?

## Cluster with K-Means

In questions Q2.4 to Q2.9 use K-Means method for clustering

### Selecting Number of Clusters

**Q2.4** Select optimal number of clusters using elbow method. *(6 points)*

**Q2.5** Select optimal number of clusters using Gap Statistic. *(6 points)*

**Q2.6** Select optimal number of clusters using Silhouette method. *(6 points)*

**Q2.7** Which k will you choose based on elbow, gap statistics and silhuettes as well as clustering task (market segmentation for advertisement purposes)? *(4 points)*

## Clusters Visulalization

**Q2.8** Make k-Means clusters with selected k_kmeans (store result as km_out). Plot your k_kmeans clusters on biplot (just PC1 vs PC2) by coloring points by their cluster id. *(4 points)*

**Q2.9** Do you see any grouping? Comment on you observation. *(4 points)*

*Answer...*

## Characterizing Cluster

**Q2.10** Perform descriptive statistics analysis on obtained cluster. Based on that does one or more group have a distinct characteristics? *(10 points)* Hint: add cluster column to original df dataframe

## Cluster with Hierarchical Clustering

**Q2.11** Perform clustering with Hierarchical method. Try complete, single and average linkage. Plot dendagram, based on it choose linkage and number of clusters, if possible, explain your choice. *(10 points)*

# Additional grading criteria:

**G3.1** Was all random methods properly seeded? *(2 points)*