HMW 1: due Feb 23, 4:30 pm in pidgeonhole 'Data Science Methods' of Prisma Building, second floor.

- if you cannot print color, find other ways to highlight your results
- you may work in groups but you should write your unique answers
- include below each subquestion of Question 3 the code you ran (if you use R Markdown it will be easy to do so)
- do not use code for Question 1, do it by hand.

Question 1

Suppose that you have four observations, for which we compute a dissimilarity matrix, given by:

$$\begin{bmatrix} & 0.5 & 0.9 & 0.7 \\ 0.5 & & 0.8 & 0.1 \\ 0.9 & 0.8 & & 0.3 \\ 0.7 & 0.1 & 0.3 & \end{bmatrix}$$

a) On the basis of this matrix, do hierarchical clustering using complete linkage, showing in detail each step.

b) Suppose we pick two clusters. Which observations are in each? Carefully draw the dendogram to show this.

Question 2

Suppose the true model is $y_i = f(x_i) + \epsilon_i$, where $\epsilon_i \sim iid(0,1)$ (for all i). Suppose we take a random sample of n observations $\{y_j, x_j\}$, and we estimate a regression function, i.e. $\hat{y}_j = \hat{f}(x_j)$. Possibly imposing sufficient but reasonable assumptions, prove that for a new observation $\{x_0, y_0\}$, where $\hat{y}_0 = \hat{f}(x_0)$ and $y_0 = f(x_0) + \epsilon_0$, we have that: $E(y_0 - \hat{f}(x_0))^2 = Var(\hat{f}(x_0)) + Bias(\hat{f}(x_0)) + 1$. If you make further assumptions, clearly state them along with all the derivations that lead to the result above.

Question 3

You are interesting in finding patterns in house prices. The data is taken from www.kaggle.com and the description of the variables observed (in the Data Dictionary) is self explanatory.

a) Load data and data dictionary, and make sure that the dates and region names are loaded in a date (**as.Date**), respectively a factor format (**as.Factor**).

b) Select variables 6 to 20 in the loaded dataset, for the date 30-11-2017.

c) Input data with 6 PC using the EM algorithm and ridge regularization of the variance-covariance matrix. The package you need to install is **missMDA** and the command is below:
c2<-imputePCA(c1, ncp = 6, center=TRUE, scale = TRUE, method = c(”Regularized”,”EM”), row.w = NULL, coeff.ridge = 1, threshold = 1e-06, seed = NULL, nb.init = 1, maxiter = 1000)

d) Run a PCA on the standardized observations using **princomp()**, and display the first 2 PC loadings up to two digits. Which variables have loadings below 0.01 on the first two components, and what do we conclude for these variables?

e) Plot the first two PC loadings and interpret the plot.

f) Plot the PC scores and interpret the plot.

g) Display a screeplot of the eigenvalues ordered from largest to smallest. Explain how many PC you would choose based on this screeplot.

h) Calculate from the eigenvalues the cumulative variance explained by the first 1,2,3 up to 15 PC, and plot it against the number of PC considered, in ascending order. Based on this plot, do you arrive at the same conclusion as in g)? Why/why not?

i) Take the first two PC from d). These are computed for different areas where homes are sold. You are interested in clustering these regions. Perform a k-means cluster analysis with 2 clusters, 20 random starts, and interpret the results.

j) Explain how you could construct from the initial dataset an index of home values in all available regions using PC. Explain one difficulty you foresee in the computation of such an index.