Universidad
Carlos III de Madrid
www.uc3m.es

ARCOS

TECHNOLOGICAL FUNDAMENTALS IN THE BIG DATA WORLD

CLUSTERING WITH KMEANS

Lab 2

Description

- To create a program to cluster data in a file with kmeans using Python.
  - File: computers.csv

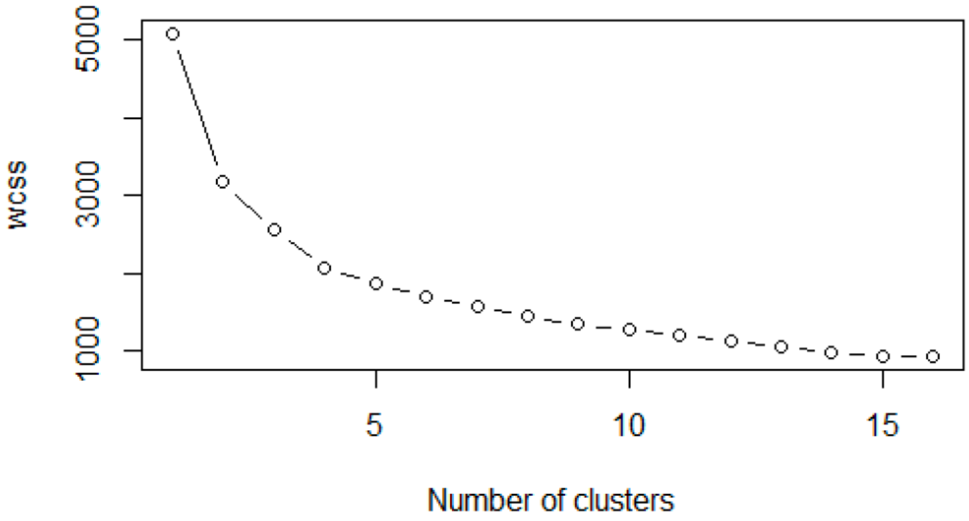
- You are asked to:
  - Write the serial version of the program in Python
  - Write the parallel version of the program in Python with multiprocessing and threads.
  - Write a memory explaining your results (maximum 12 pages)

- Notice:
  - In the dataset you have 3 fields that are not numerical:
    - cd multi premium
  - As they have only two values, you can substitute them with 0 (no) and 1 (yes) to normalize de data.

Serial program

- Make a program in that:

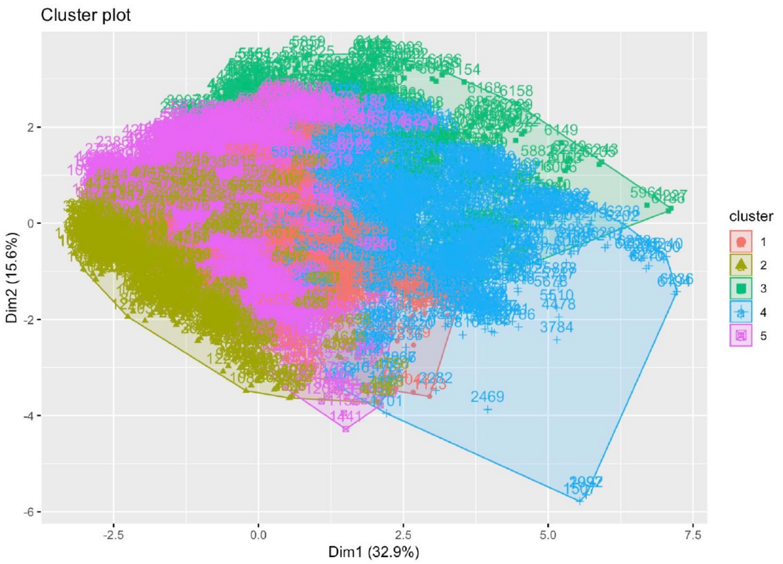
1. Constructs the elbow graph and find the optimal clusters number (k).
2. Plots the results of the elbow graph. Choose optimum.



Serial program

3. Clusters the data using the optimum value using k-means.

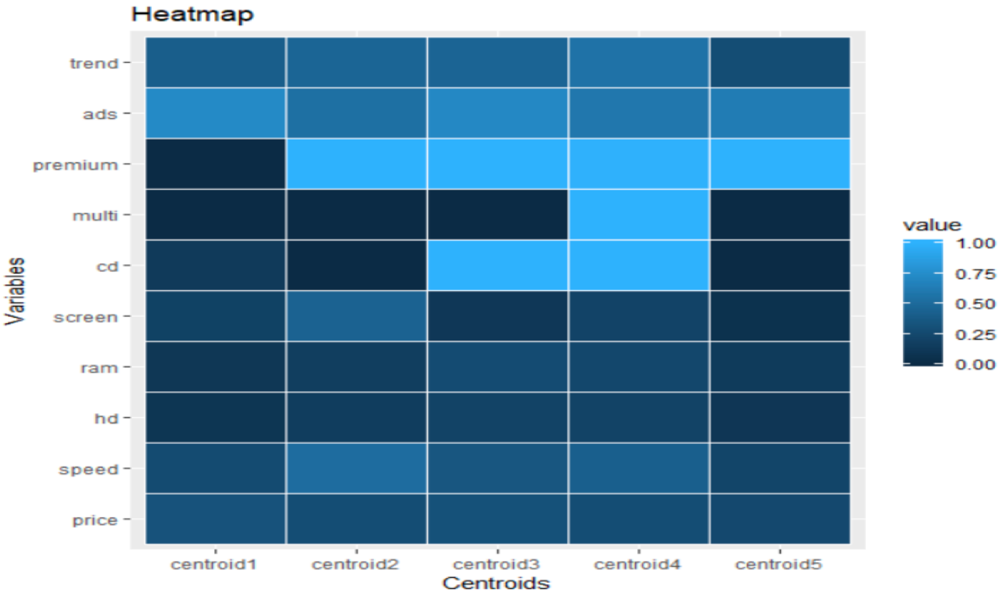
4. Plot the first 2 dimensions of the cluster



Serial program

5. Finds and print the cluster with the highest price average.

6. Prints a heat map with matplotlib.pyplot for the clusters centroids.



Parallel Python programs

- Multiprocessing version
  1. Write a parallel version of you program using multiprocessing
  2. Measure the time and optimize the program to get the fastest version you can.

- Threaded version
  1. Write a parallel version of you program using threads
  2. Measure the time and optimize the program to get the fastest version you can.

Rules

□ Make the lab in groups maximum 3

□ Do not use Jupyter Notebook

□ To deliver:

1. Programs in Python: serial and parallel versions
2. Written report

□ Deadline: **October 23rd 2022. 23:30 hours.**