Documentation for R assignment

B5

Abdelrhman mohamed elsayed – 20198113

Amira kadry – 20198014

Dina Elsayed – 20198029

Mahmoud Samy rezk - 20198075

- **The selected data:**

The selected data is Mall Customer Segmentation Data from Kaggle website

And that is its link [Mall Customer Segmentation Data | Kaggle]

- Sample of the data

CustomerID	Gender	Age	Annual Income (k$)	Spending Score (1-100)
1	Male	19	15	39
2	Male	21	15	81
3	Female	20	16	6
4	Female	23	16	77
5	Female	31	17	40
6	Female	22	17	76
7	Female	35	18	6
8	Female	23	18	94
9	Male	64	19	3
10	Female	30	19	72
11	Male	67	19	14
12	Female	35	19	99
13	Female	58	20	15
14	Female	24	20	77
15	Male	37	20	13
16	Male	22	20	79

- # The models used in the assignment

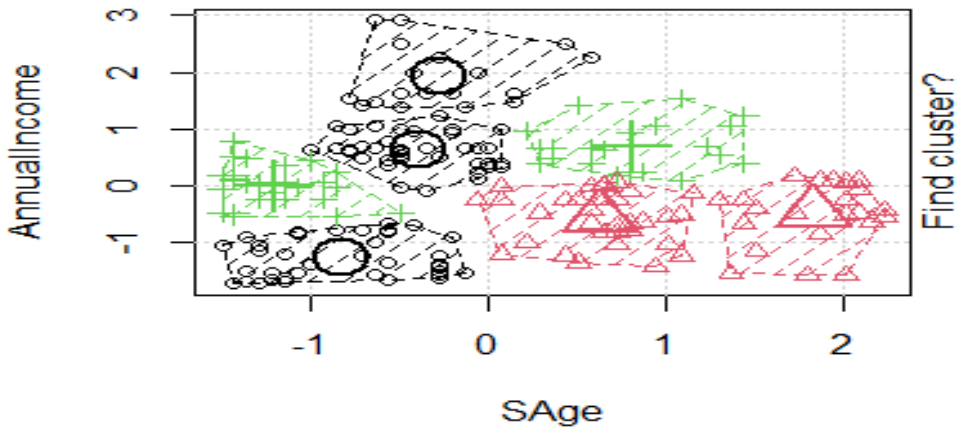
  There are two models used in the assignment

  - First one: trying to find relations (clusters) between the age of mall customers and the annual income
  - Second one: trying to find relations between the age of mall customers and spending score

- # First model:
  - **The model:**
    In this model we chose the optimal k is 7 and we use kmeans function which by default uses Euclidian as error function
    So the model used (k = 7, error function = Euclidian)

  - **The graph of clustered data:**

    the plotting here uses the animation library

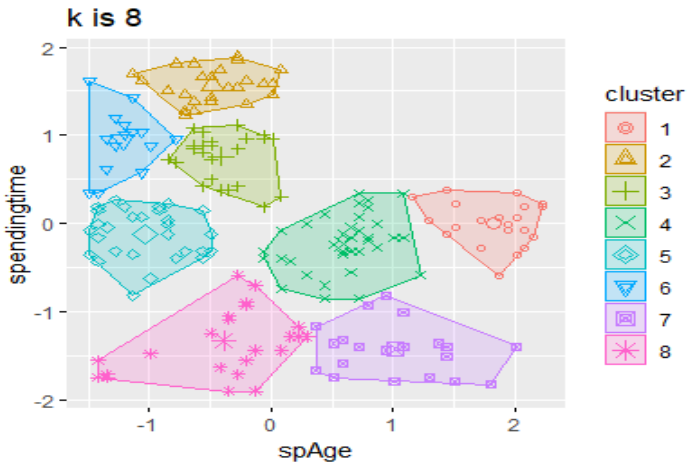


  - **Interpretation (Justification) of the data:**
    the clusters here show that the people almost (30 – 50) have higher annual than the other ages

- **second model:**
  - **The model:**
    In this model we chose the optimal k is 8 and we use kmeans function from **amap** library to manage changing the function error to Manhattan ,So the model used (k = 8, error function = Manhattan)

  - **The graph of clustered data:**

    the plotting here uses the factoextra library to manage change error function



  - **Interpretation (Justification) of the data:**
    the clusters here show that the people almost (less than 20 "teenagers") spend more time in the mall than other who are in older ages