HOMEWORK 8

- Anirudh Narayanan (an9425)

README:

Run command on terminal to create a database named clustering with collection name movies .

mongoimport --db HW4 --collection movies --file "/Users/anirudhramesh/Desktop/INTRO_TO_BIGDATA/ASSN4/mem_j2.json" --drop

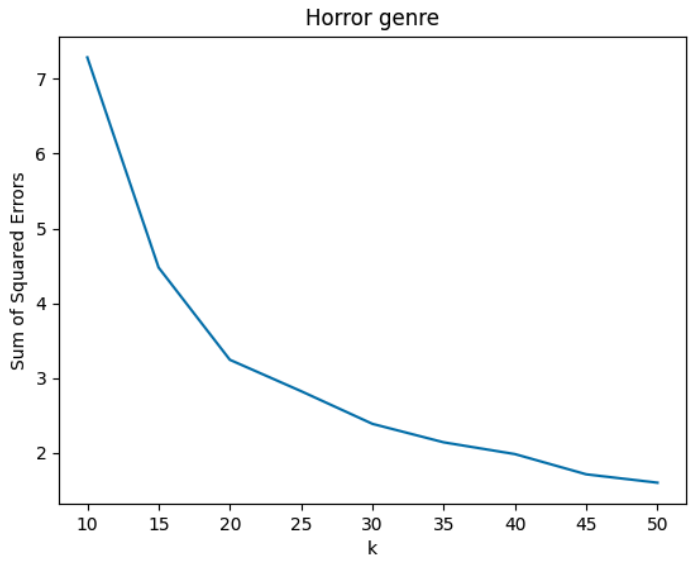
Q1 - Run Q1.py
Q2. - Run Q2.py (test case: genre = ‘Action’, k = 5)
Q3. Run Q3.py (test case : genre = ‘ Action’)
Code executes for initial centroid initialization and cluster assignment.
Q4, Q5. Run Q4_5.py. Code will save 5 graphs in the local directory and 5 scatter plots.

OUTPUTS:
Q4, Q5
For Q4, we will consider the elbow point as the best value of k.

For Q5, based on the scatter plot, we can see that data with close normalized values are grouped together which translates to movies with closer start year and average rating are grouped together.

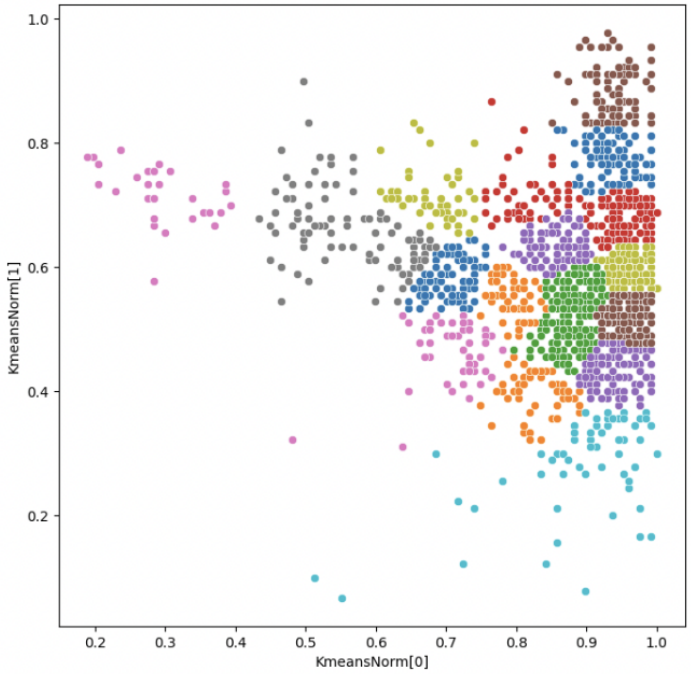
Outputs from next page.

HORROR:

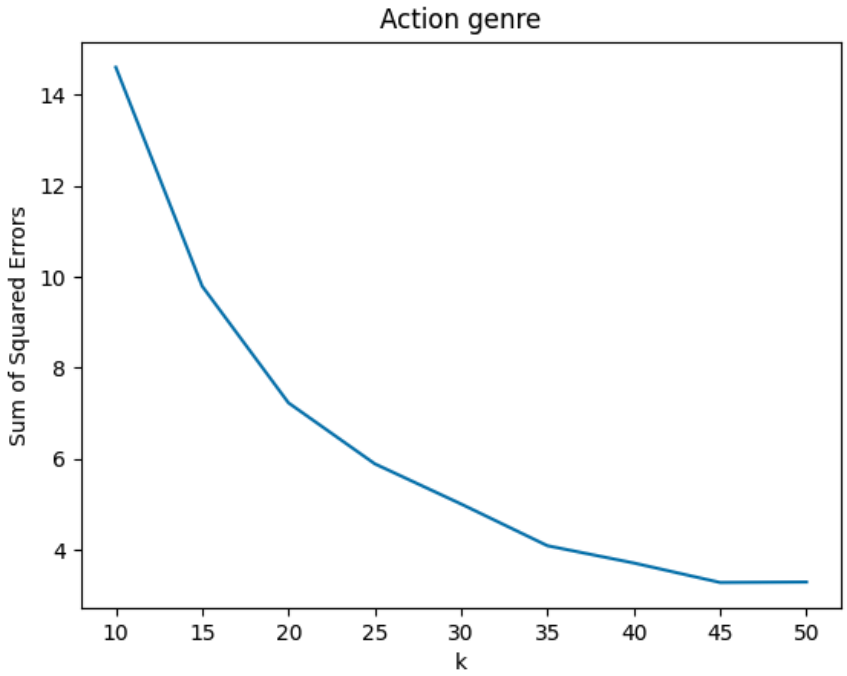


20 best value of k based on the graph above

Corresponding Scatter Plot for Q5:

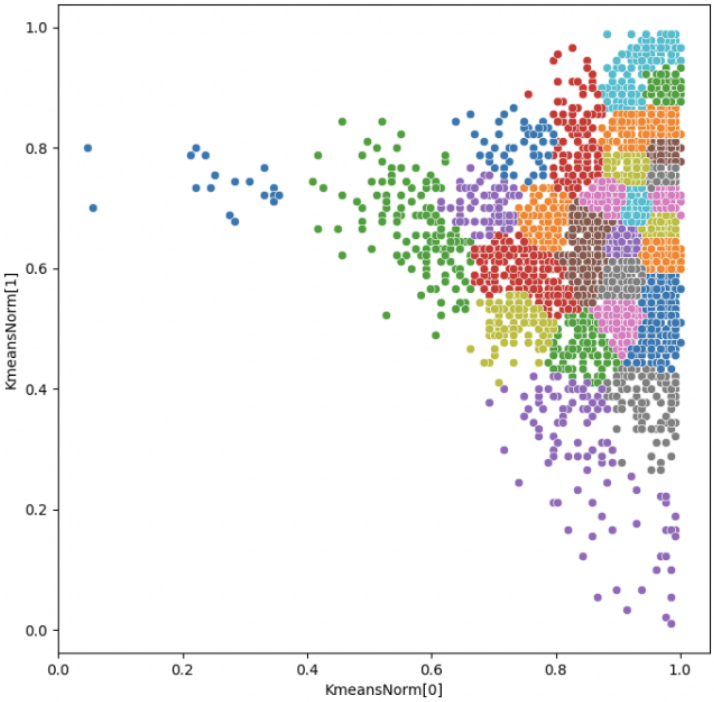


ACTION:

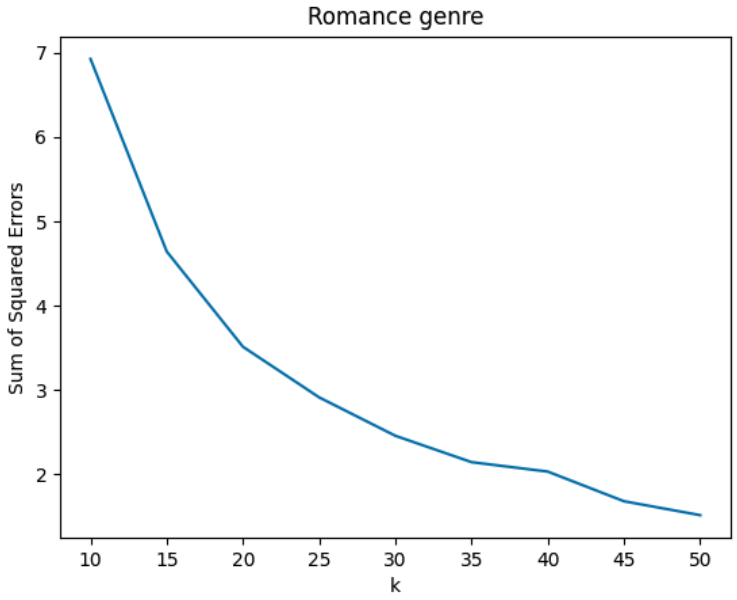


35 best value of k based on the graph above

Corresponding Scatter Plot for Q5:



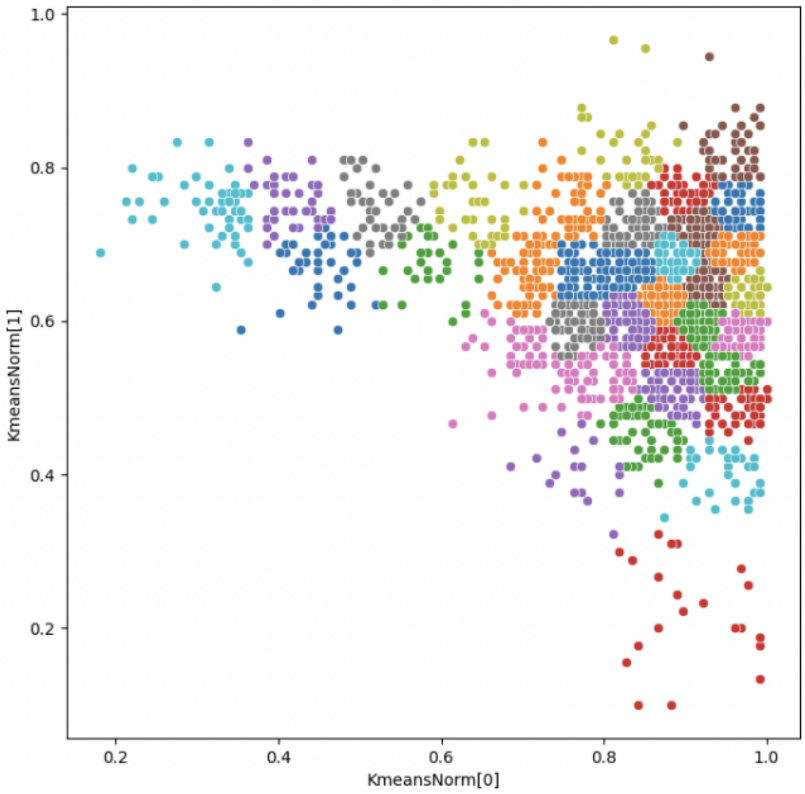
ROMANCE:



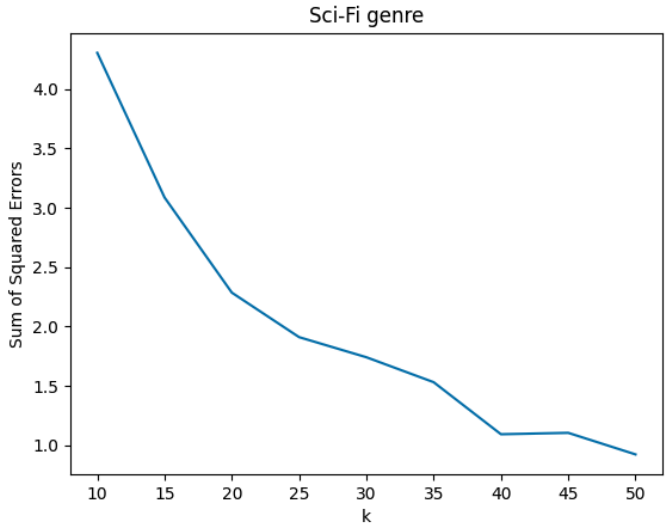
---

35 best value of k based on the graph above

Corresponding Scatter Plot for Q5:

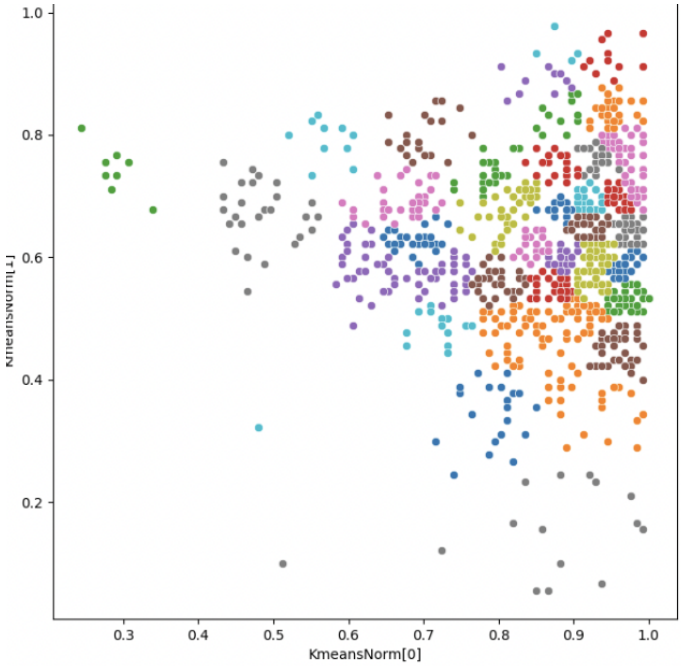


SCI-FI:

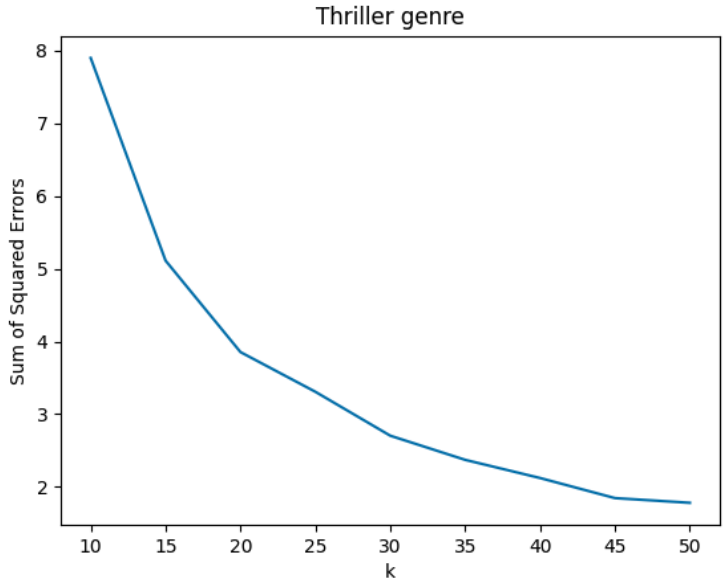


40 best value of k based on the graph above

Corresponding Scatter Plot for Q5:



THRILLER:


Thriller genre
Sum of Squared Errors
8
7
6
5
4
3
2
10
15
20
25
30
35
40
45
50
k

30 best value of k based on the graph above