CSCI.620.01/03

Homework 8 Report (ss7495)Suraj Sureshkumar

To consider movies (documents whose type=’movie’) with a number of votes greater than 10,000 for all questions in this assignment and where both startYear and avgRating exist implemented a mongodb db and created a new collection. Below is the query:

```
db.movie.aggregate([
    {
        $match:{
            'titleType': 'movie',
            'numVotes':{$gt:10000},
            'startYear':{$ne : null},
            'averageRating':{$ne : null}
        }
    },
    {
        $out : 'movie_cluster'
    }
])
```

Task 1:

```
db.movie.aggregate([
{"$setWindowFields": {
   "output": {
     "max_year": {$max: "$startYear"},
     "min_year": {$min: "$startYear"},
     "max_rating": {$max: "$averageRating"},
     "min_rating": {$min: "$averageRating"}
   }
 }
},
{$out : movie_cluster}])

db.movie_cluster.aggregate([
{$addFields : {
      'kmeansNorm': [{
         $divide :[
         {$subtract : ['$startYear', '$minYear']},
         {$subtract :['$maxYear', '$minYear']}
         ]
      },{
         $divide :[
         {$subtract : ['$averageRating', '$minRating']},
         {$subtract :['$maxRating', '$minRating']}
         ]
      }]

   }
},
{$out : 'movies_norm'}])
```

As you can see below the kmeansNorm field has been created.

_id	titleTyp	originalTitle	startYear	runtimeMinute	averageRating	numVotes	genre	directors	producer	actors	writers	kmeansNorm
17...	m...	The G...	1926.0	67.0	8.1	92514.0	[3 elements]	[2 elements]	[2 elements]	[5 elements]	[1 elements]	[2 elements]
19...	m...	Steam...	1928.0	70.0	7.8	15044.0	[3 elements]	[2 elements]	[2 elements]	[5 elements]	[1 elements]	[2 elements]
15...	m...	Sherlo...	1924.0	45.0	8.2	51065.0	[3 elements]	[2 elements]	[2 elements]	[5 elements]	[1 elements]	[2 elements]
15...	m...	The N...	1924.0	59.0	7.6	10243.0	[3 elements]	[2 elements]	[2 elements]	[5 elements]	[1 elements]	[2 elements]
14...	m...	Safety...	1923.0	74.0	8.1	21227.0	[3 elements]	[2 elements]	[2 elements]	[5 elements]	[1 elements]	[2 elements]
21...	m...	Little ...	1931.0	79.0	7.2	13642.0	[3 elements]	[2 elements]	[2 elements]	[5 elements]	[1 elements]	[2 elements]
23...	m...	Scarfa...	1932.0	93.0	7.7	28664.0	[3 elements]	[2 elements]	[2 elements]	[5 elements]	[1 elements]	[2 elements]
18...	m...	Wings	1927.0	144.0	7.6	13337.0	[3 elements]	[2 elements]	[2 elements]	[5 elements]	[1 elements]	[2 elements]
15...	m...	The G...	1925.0	95.0	8.1	112440.0	[3 elements]	[2 elements]	[2 elements]	[5 elements]	[1 elements]	[2 elements]
19...	m...	La pas...	1928.0	110.0	8.2	56076.0	[3 elements]	[2 elements]	[2 elements]	[5 elements]	[1 elements]	[2 elements]
21...	m...	L'âge ...	1930.0	60.0	7.2	14203.0	[3 elements]	[2 elements]	[2 elements]	[5 elements]	[1 elements]	[2 elements]
12...	m...	The Kid	1921.0	68.0	8.3	127612.0	[3 elements]	[2 elements]	[2 elements]	[5 elements]	[1 elements]	[2 elements]
18...	m...	The C...	1928.0	76.0	8.0	12112.0	[3 elements]	[2 elements]	[2 elements]	[5 elements]	[1 elements]	[2 elements]
21...	m...	City Li...	1931.0	87.0	8.5	186224.0	[3 elements]	[2 elements]	[2 elements]	[5 elements]	[1 elements]	[2 elements]
20...	m...	Anim...	1930.0	97.0	7.4	14732.0	[3 elements]	[2 elements]	[2 elements]	[5 elements]	[1 elements]	[2 elements]
22...	m...	Monk...	1931.0	77.0	7.4	13788.0	[3 elements]	[2 elements]	[2 elements]	[5 elements]	[1 elements]	[2 elements]
23...	m...	Horse...	1932.0	68.0	7.5	12900.0	[3 elements]	[2 elements]	[2 elements]	[5 elements]	[1 elements]	[2 elements]
18...	m...	The Ci...	1928.0	72.0	8.1	34412.0	[3 elements]	[2 elements]	[2 elements]	[5 elements]	[1 elements]	[2 elements]
16	m...	Seven	1925.0	56.0	7.8	10731.0	[3 elements]	[2 elements]	[2 elements]	[5 elements]	[1 elements]	[2 elements]

```
{
```

JSON OUTPUT:


```
"_id" : NumberInt(4972),
  "titleType" : "movie",
  "originalTitle" : "The Birth of a Nation",
  "startYear" : 1915.0,
  "runtimeMinutes" : 195.0,
  "genres" : "Drama,War",
  "averageRating" : 6.2,
  "numVotes" : 25191.0,
  "genre" : [
    "Drama",
    "War",
    null
  ],
  "directors" : [
    "428",
    null
  ],
  "producer" : [
    null,
    null
  ],
  "actors" : [
    {
      "actor" : "1273",
      "roles" : "[\"Elsie - Stoneman's Daughter\"]"
    },
    {
      "actor" : "178270",
      "roles" : "[\"Margaret Cameron - The Elder Sister\"]"
    },
    {
      "actor" : "550615",
      "roles" : "[\"Flora Cameron - The Pet Sister\"]"
    },
    {
      "actor" : "910400",
      "roles" : "[\"Col. Ben Cameron aka The Little Colonel\"]"
    },
    {
      "actor" : null
    }
  ],
  "writers" : [
```

"228746"
],
"kmeansNorm" : [
0.0,
0.6046511627906977
]
}

Task 2:

```
var g = 'Action';
var k = 20
var i = 1;
db.movies_norm.aggregate([
   {$unwind:"$genre"},
   {$match:{genre : g,genre:{$ne : null}}},
   {$sample : {size: k}},
   {$project: {"kmeansNorm":1}}
]).forEach(function(doc) {
   db.centroid.insertOne({ID:i, kmeansNorm:doc.kmeansNorm});
   i = i+1;
});



db.createCollection("centroid");


db.movies_norm.aggregate([
   {$unwind:"$genre"},
   {$match:{genre : g,genre:{$ne : null}}}]).forEach(function(doc) {

   });
```

Centroid collection output:

_id	ID	kmeansNorm
643904acba80a1...	1	[2 elements]
643904acba80a1...	2	[2 elements]
643904acba80a1...	3	[2 elements]
643904acba80a1...	4	[2 elements]
643904acba80a1...	5	[2 elements]
643904acba80a1...	6	[2 elements]
643904acba80a1...	7	[2 elements]
643904acba80a1...	8	[2 elements]
643904acba80a1...	9	[2 elements]
643904acba80a1...	10	[2 elements]
643904acba80a1...	11	[2 elements]
643904acba80a1...	12	[2 elements]
643904acba80a1...	13	[2 elements]
643904acba80a1...	14	[2 elements]

The json output:


```
{
   "_id" : ObjectId("643904acba80a144e24f8466"),
   "ID" : NumberInt(1),
   "kmeansNorm" : [
      0.8684957964099068,
      0.633471251248395
   ]
}
```

kmeansNorm collections

_id	ID	kmeansNorm
643904acba80a1...	1	[2 elements]
643904acba80a1...	2	[2 elements]
643904acba80a1...	3	[2 elements]
643904acba80a1...	4	[2 elements]
643904acba80a1...	5	[2 elements]
643904acba80a1...	6	[2 elements]
643904acba80a1...	7	[2 elements]
643904acba80a1...	8	[2 elements]
643904acba80a1...	9	[2 elements]
643904acba80a1...	10	[2 elements]
643904acba80a1...	11	[2 elements]
643904acba80a1...	12	[2 elements]
643904acba80a1...	13	[2 elements]

Task 3:

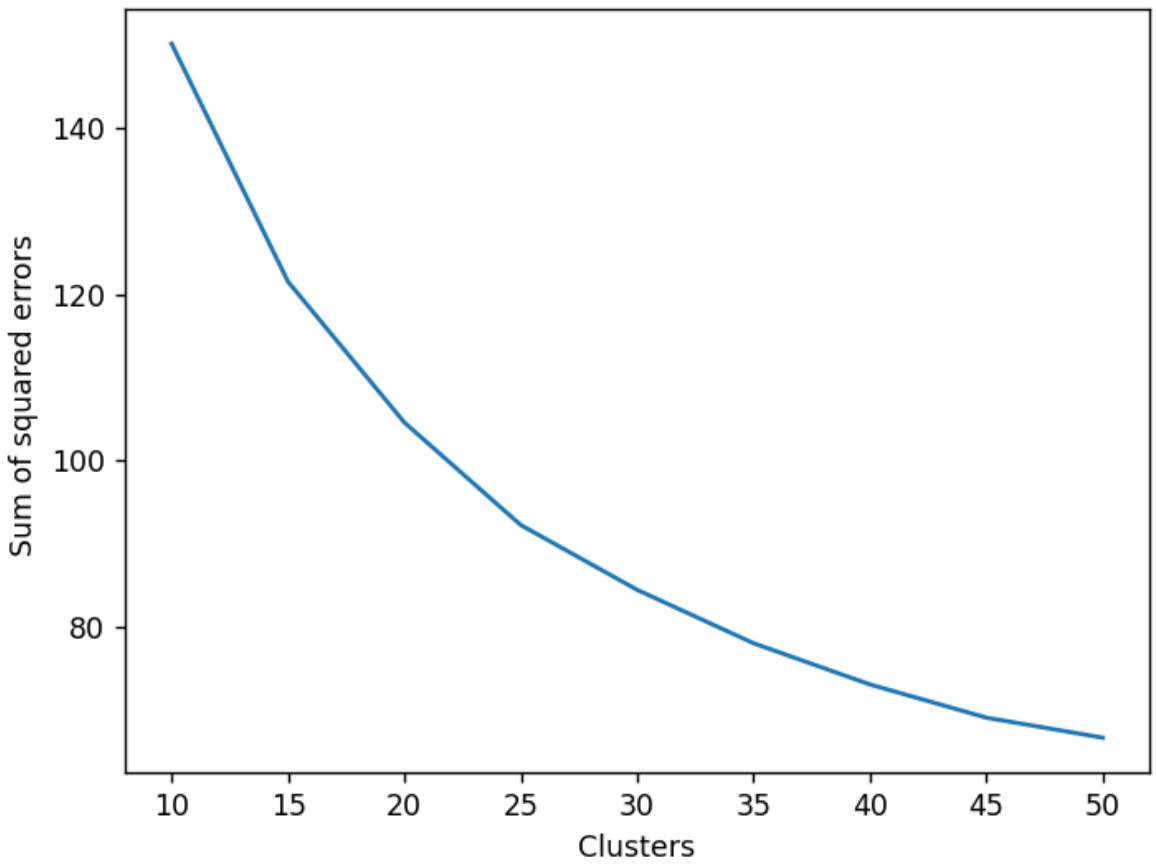
This was implemented in python and the code will attached in the sourcecode.zip and here too:

```
import math
from statistics import mean
import numpy as np
import pandas as pd
import pymongo

# mongo connection
client = pymongo.MongoClient("mongodb://localhost:27017")
db = client["hw4"]
movies_norm_collection = db["movies_norm"]
centroid_collection = db["centroid"]

g = "Action"

pipeline = [
    {
        "$unwind": "$genre"
    },
    {
        "$match": {
            "genre": g
        }
    }
]

movie_norm_list = list(movies_norm_collection.aggregate(pipeline))
centroid_list = list(centroid_collection.find({}))

movie_norm_df = pd.DataFrame(movie_norm_list)
centroid_df = pd.DataFrame(centroid_list)

movie_norm_df['cluster'] = np.nan

for index, row in movie_norm_df.iterrows():
    min_distance = float('inf')
    cluster_id = 0
    for index2, centroid in centroid_df.iterrows():
        eDistance = math.dist(row['kmeansNorm'], centroid['kmeansNorm'])
```

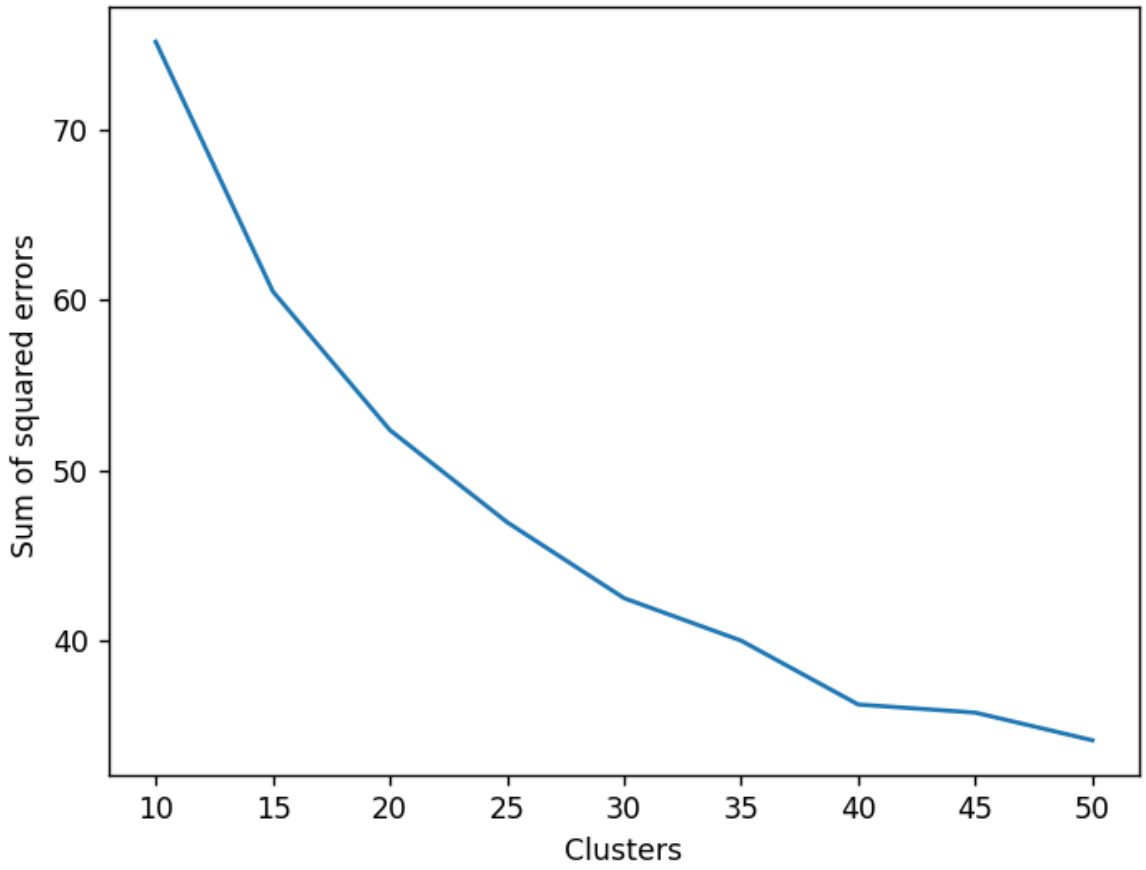
```
        if eDistance < min_distance:
            min_distance = eDistance
            cluster_id = centroid['ID']
    movie_norm_df.loc[index, 'cluster'] = cluster_id

for index2, centroid in centroid_df.iterrows():
    temp_df = movie_norm_df[movie_norm_df['cluster'] == centroid['ID']].copy()
    temp_df[['x', 'y']] = temp_df.kmeansNorm.tolist()
    x = mean(temp_df.x.tolist())
    y = mean(temp_df.y.tolist())
    centroid_df.at[index2, 'kmeansNorm'] = [x, y]
    centroid_collection.update_one({'ID': centroid['ID']}, {"$set": {"kmeansNorm": [x, y]}})
```

Task 4:

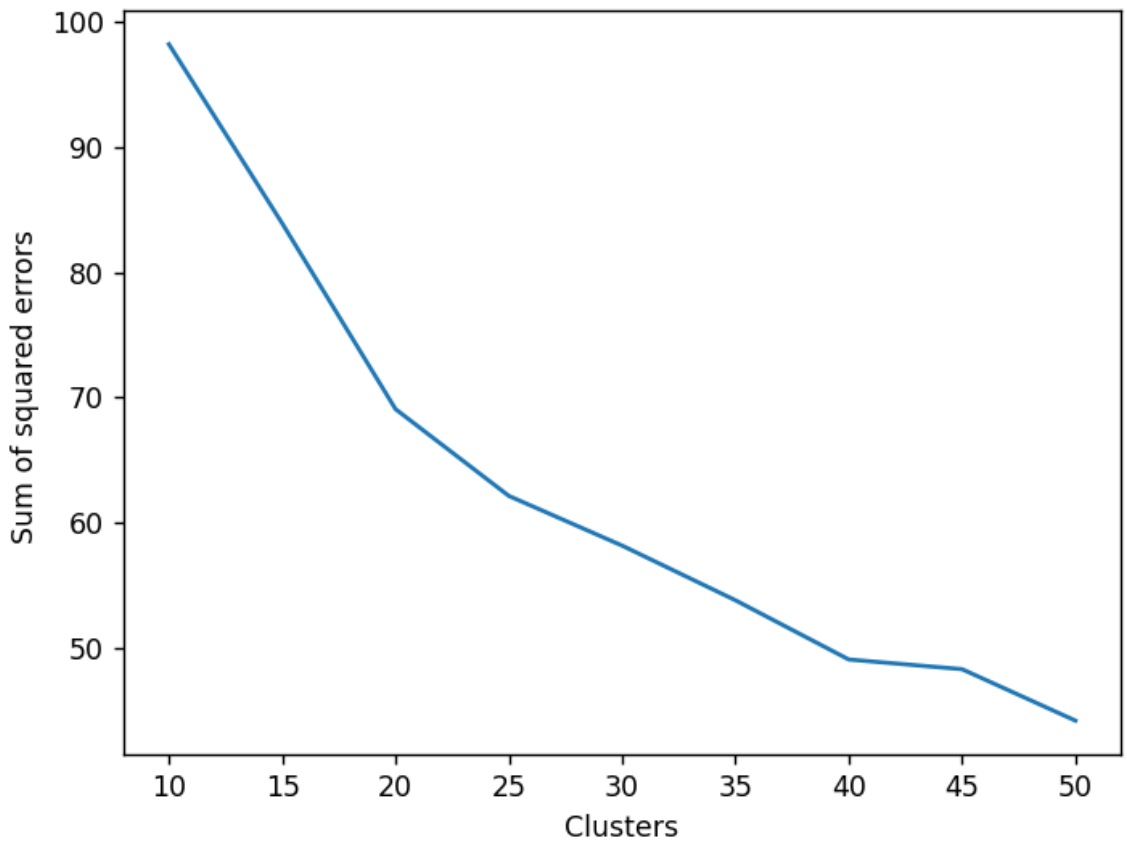
Plot for Action



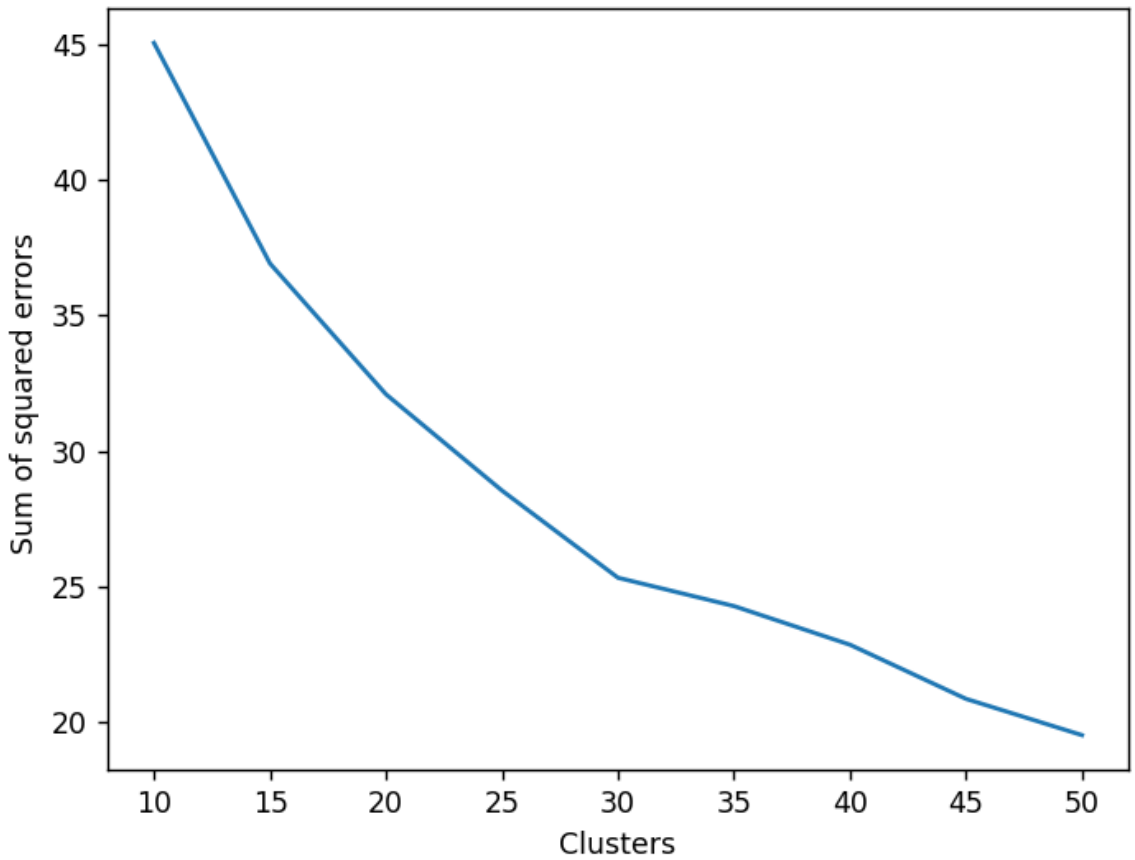
Plot for Horror



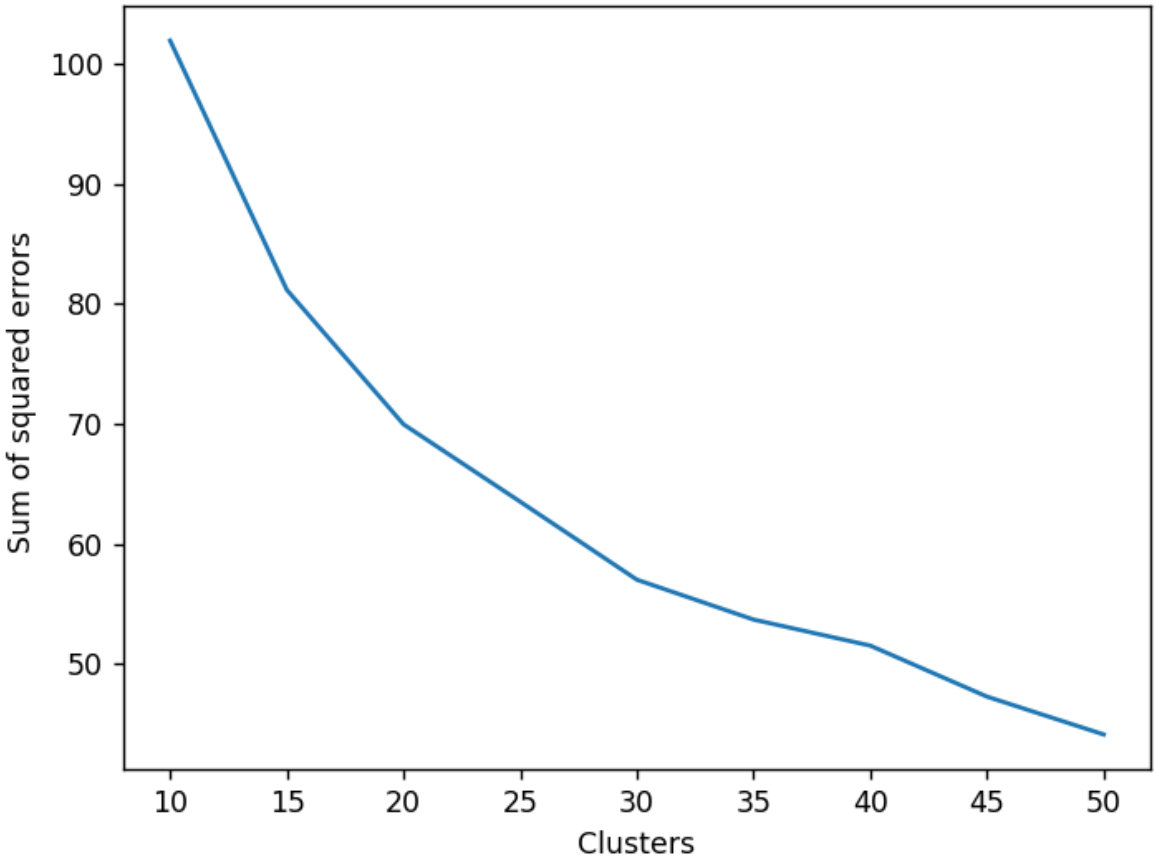
Plot for Romance



Plot for Sci-Fi



Plot for Thriller



Task 5

Dataframe output

```
0       14429     movie  ...  [0.07407407407407407, 0.8255813953488372]      1.0
1       15324     movie  ...  [0.08333333333333333, 0.8372093023255813]      1.0
2       15163     movie  ...  [0.08333333333333333, 0.7674418604651163]      1.0
3       18578     movie  ...   [0.1111111111111111, 0.7674418604651163]      1.0
4       17925     movie  ...  [0.10185185185185185, 0.8255813953488372]      1.0
```

Centroid output

```
[[0.31502525252525254, 0.7682346723044398],
```

The variants of the data assigned to this cluster with respect to the centroid are less.
Similarly the variants of the data assigned to 0 are close to the

```
[0.07407407407407407, 0.8255813953488372]
```