CIS/CSE 400, CIS 600, CSE 691: Assignment 6

Instruction: This is an individual **in-class** assignment. You are welcome to look anything up online but do NOT collaborate with your classmates or get help from anyone outside. You need to enter the responses to the questions that are given online on Blackboard (but we are also listing them below so you know the whole context). In addition to responding to the online questions on Blackboard, upload a PDF of the python code which includes the run outputs.

The dataset we will analyze is from Kaggle where a contributor used IMDB to get the top 1000 movies. The dataset has been cleaned up a little from the Kaggle version, as a result there are a fewer number of movies than 1000 in the list. The main column names and descriptions are:
`Series_Title` - Name of the movie (list includes international)
`Released_Year` - Year movie was released
`Certificate` - Certificate earned by that movie
`Runtime` - Total runtime of the movie (all in minutes)
`Genre` - Genre of the movie
`IMDB_Rating` - Rating of the movie in the IMDB site
`Overview` - Mini story/ summary
`Meta_score` - Score earned by the movie
`Director` - Name of the Director
`Star1, Star2, Star3, Star4` - Name of the main actors
`No_of_Votes` - Total number of votes in IMDB
`Gross` - Money earned by that movie

1. **Get data and pre-process**: Create a single code for all the components, so that you can upload a single PDF in the end. You may have to use aspects from one or more of the 3 codes we saw in class on clustering. Read the CSV file `imdb_top_1000_cleaned.csv` onto your code. In your code output make sure you print the data frame header.

   (a) Write a line in the code to check if any columns have NaN or empty values. In which features/columns there are NaN or empty cells (possibly more than 1 column)?
   (a) Certificate, (b) Runtime, (c) Meta_score, and (d) Gross

   (b) Use `dropna()` to drop all the rows with NaN or empty cells. Then reset your index so that the index numbers are continuous after dropping the NaN. Make sure you also drop the new column called "index" as well as the column "Overview". Now, how many rows are in the data frame?
   (a) 1000, (b) 831, (c) 714, (d) 127

   (c) In the column "Gross", you may find commas and so python may read it as a string. If your data frame is called `df` then make the following edit:
   `df['Gross'] = df['Gross'].str.replace(',', '').astype(int)`
   Likewise the column "Runtime" has the term minutes in it. So make the following edit:

```
df[’Runtime’] = df[’Runtime’].str.replace(’ min’, ’’).astype(int)
```

Now when you type `df.dtypes`, which of the following columns (there may be more than 1) are of the type `object`?
(a) Runtime, (b) Genre, (c) Meta_score, (d) Gross

(d) Run a correlation heat map between the features (you will get 6 numerical features for this). Take a moment to see if the correlation values make sense (some may not). With which variable does ‘Gross’ correlate most?
(a) Released_Year, (b) IMDB_Rating, (c) Meta_score, (d) No_of_Votes?

(e) Create two separate lists of columns: the names of the 6 numeric columns, and the names of the 2 rating columns (`IMDB_Rating` and `Meta_score`). Use a standard scaler to scale the 6 numeric columns (since the two rating columns are a subset, they will also get scaled). For the movie *The Shawshank Redemption*, what is the scaled value for "Gross"?
(a) 1.24, (b) 5.62, (c) -0.09, (d) -0.44

2. **K-Means Clustering with 2 Features**: The two features we will use here are the 2 rating columns (`IMDB_Rating` and `Meta_score`). The IMDB rating is based on average score of audience member users from 1 to 10, while metacritic rating (`Meta_score` value) averages the score of prestigious critics from 1 to 100 (you can think of these as the experts). Before you proceed, take a look at their correlation.

(a) Use $K = 6$ clusters and perform a K-Means clustering. Be sure to use ‘k-means++’ for init, n_init = 10, max_iter = 500, and random_state = 50 (do not use `tol`). Once you create the clusters, merge the cluster data with the unscaled data frame so you have the actual rating and other categories. Display the header of the resulting data frame. Which of the following statements (there could be more than 1) are TRUE?
(a) All movies with Meta score of 60 or lower are in the same cluster
(b) The movies *Rear Window* and *Citizen Kane*, both ranked 100 in Meta score, are part of the same cluster
(c) All movies with IMDB rating of 8.7 or higher are part of the same cluster
(d) The number of movies in each cluster is the same

(b) If you have not done it yet, graph the clusters on a graph with X-axis as the IMDB rating and Y-axis as the Meta score. There are two clusters with very small widths on the X-axis, what range of values in IMDB ratings do they take?
(a) 7.6 to 8.1, (b) 7.9 to 8.3, (c) 8.5 to 8.8, (d) 7.6 to 7.9

(c) For the same parameters for K-Means we had earlier, vary the number of clusters from 1 to 30. Draw a graph of the distortions using the two rating columns as features. What is the distortion approximately for $K = 5$ clusters?

(a) 850, (b) 350, (c) 2250, (d) 25

(d) Instead of 2 features, now use all 6 numerical features to create a graph between distortions and number of clusters (use the same parameters as before for the K-Means clustering). When we go from 1 to 10 clusters, by how much (approximately) does distortions reduce?
(a) 100, (b) 1000, (c) 2000, (d) 3000

3. Let us now try Agglomerative Clustering. But let us not specify the number of clusters. Instead, let us say the `distance_threshold` is 8.0, affinity is euclidean, and linkage is complete. Use all the 6 numerical features. Merge with the unscaled data frame.

(a) To answer this you can either use a command in python or just plot a graph (that is the next question anyway). How many clusters are obtained?
(a) 2, (b) 3, (c) 4, (d) 5

(b) Although we used all 6 features, we can still plot the clusters on a graph with X-axis as the IMDB rating and Y-axis as the Meta score. Which of the following statements is False?
(a) One cluster had all IMDB ratings higher than 8.5
(b) One cluster has all IMDB rating lower than 8.5
(c) One cluster did not have any movie with Meta Score rating of 100
(d) One cluster has all Meta Score rating higher than 74

(c) Only one of the clusters in the previous question had 6 movies. Which of the following does not belong to that cluster?
(a) The Matrix
(b) Avatar
(c) Star Wars: Episode VII - The Force Awakens
(d) Titanic

4. Now let us go back to K-Means clustering with all 6 features. And we also go back to the original data frame before scaling. Create a new data frame with movies with greater than 8.3 (that means 8.4 and higher) IMDB rating. To this top-movie data frame, perform a standard scaler.

(a) How many movies are in this top-movie data frame?
(a) 54, (b) 68, (c) 78, (d) 82

(b) Use 6 clusters and all 6 features to perform a K-Means clustering. Use the same values as before for other parameters. Which of the following were not part of the same cluster?
(a) The Departed

(b) Star Wars
(c) Coco
(d) Alien

(c) Continuing with the previous question set up, which of the following statements is False?
(a) There is a cluster with just 3 movies in them
(b) Interstellar, Fight Club, The Prestige and Joker are are in the same cluster
(c) Both movies directed by Alfred Hitchcock in the data frame are in the same cluster
(d) All movies with Tom Hanks in Star1 role are in the same cluster

5. Now let us go back to the original data frame before scaling. Create a new data frame with movies that fall under 'Genre' of Drama (just Drama, do not use rows where drama is combined with other genres). Add a seventh feature of "Certificate" but this is a categorical variable. So do a `pd.get_dummies()` and obtain a drama-movie data frame with numerical features. To this drama-movie data frame, perform a standard scaler.

   (a) How many movies are in this drama-movie data frame?
   (a) 58, (b) 68, (c) 78, (d) 88

   (b) Use 3 clusters and all features to perform a K-Means clustering. Use the same values as before for other parameters. Which of the following were not part of the same cluster?
   (a) The Color Purple
   (b) The Help
   (c) Rain Man
   (d) Fight Club

   (c) Continuing with the previous question set up, which of the following statements is True?
   (a) Movies were clustered by Certificate where each cluster has 2 certificate types
   (b) One cluster had all the Meta score of 80-100, one with 70-79, and one below 70
   (c) IMDB scores of 9s were clustered together, 8s were clustered together, and below 8 were the third cluster
   (d) Dramas made after 2000 were clustered together, those made in 1990s were together, and those before 1990 were in the same cluster

6. Upload a single PDF document with your code.