CSE 572
Assignment 2

Group Name: Aditi_13

Members:

Name	ASU ID	Email
Aditi Baraskar	1213175832	anbarask@asu.edu
James Smith	1208109080	jsmit106@asu.edu
Moumita Laskar	1204363181	mlaskar@asu.edu
Tejas Ruikar	1215161649	truikar@asu.edu

Results

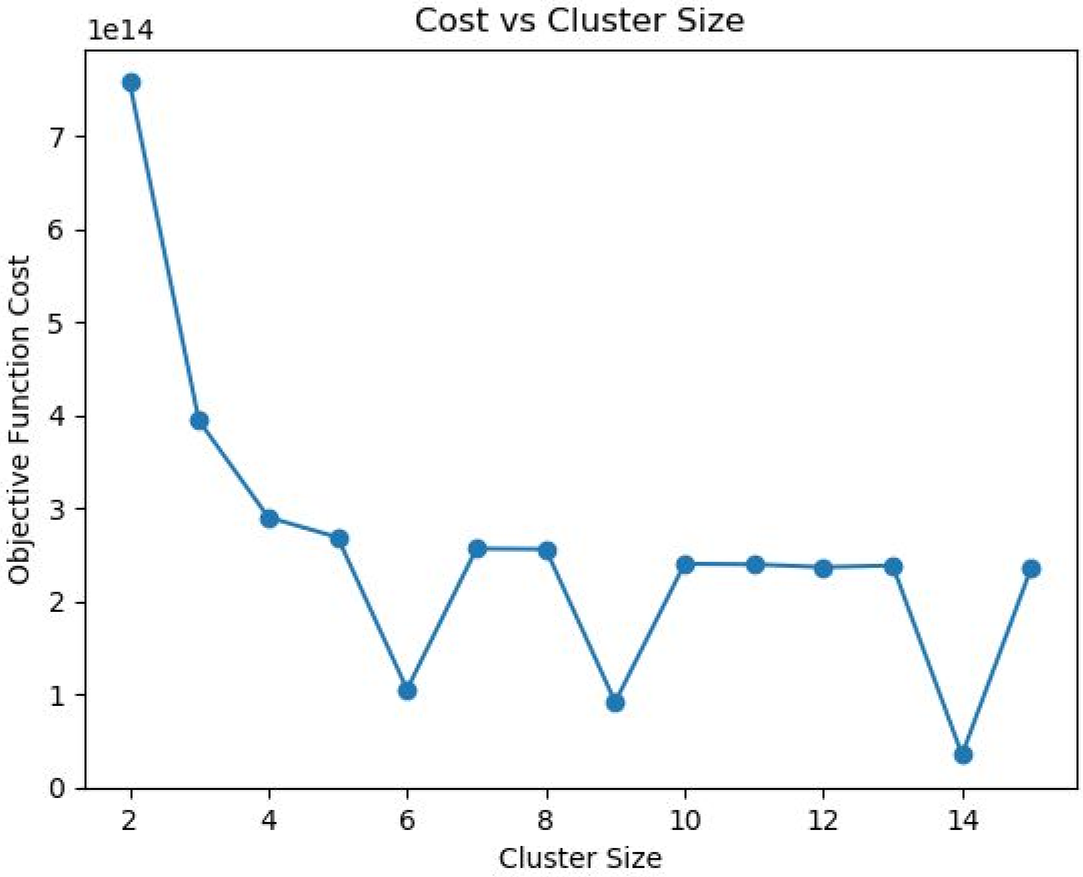
Task 1:

- Task1_data.xlsx has the extracted columns Population and Deaths for each of the 50 states from the overdoses.csv
- Task1_k5_table.xlsx has the clustering information for when k = 5. The first column is the index of the row from the data, and the second column is the index of the cluster assigned to it, 0 to 5.
- Assignment2.py has the code for k-means clustering and the calculation for the cost using the objective function.

$$J=\sum_{j=1}^{k}\sum_{i=1}^{N} m_{i,j}(x_i-C_j)^2$$

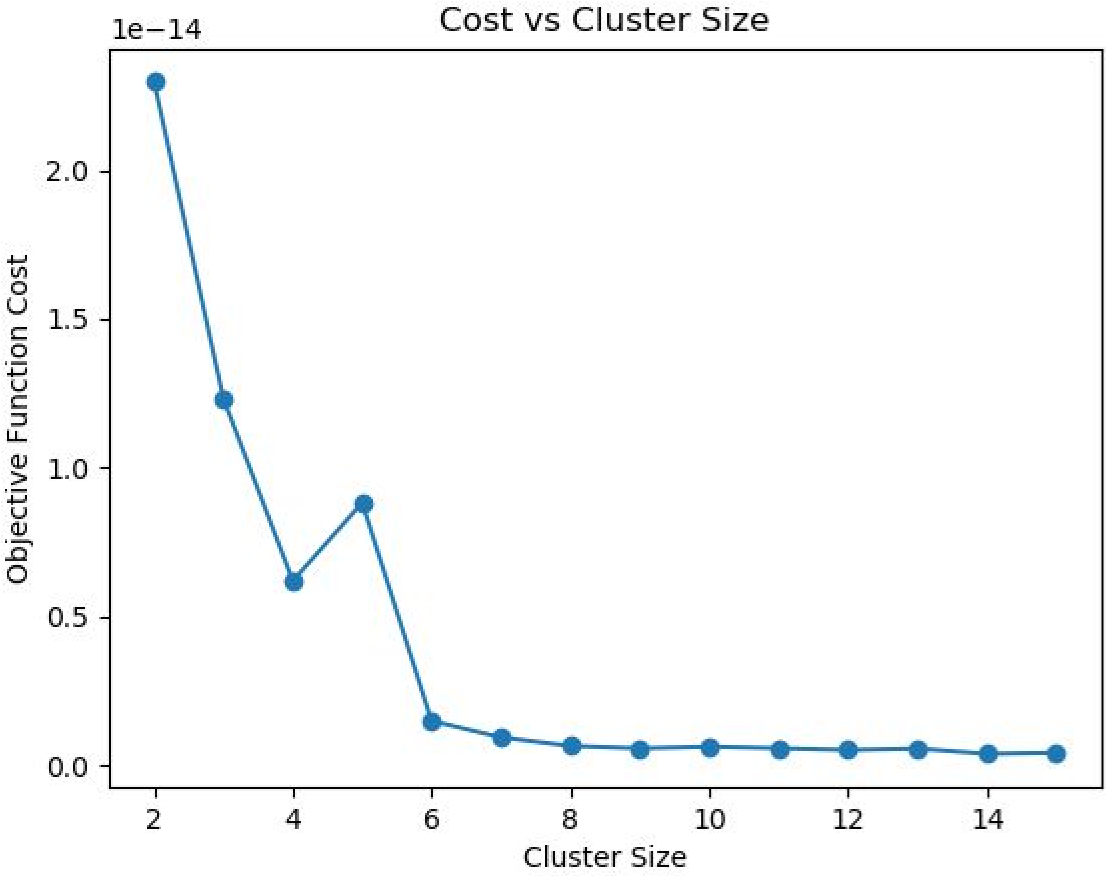
k = number of clusters
N = Total number of data points
$m_{i,j}$ = The cluster membership
C_j = jth cluster center

- The graph for the objective function value vs the number of clusters for the size from 2 to 15 is shown below.



Task 2:

- Task2_sim_matrix.xlsx contains the cosine similarity matrix based on the population and deaths of each state.
- The graph of the objective function value vs the number of clusters for size from 2 to 15 is mentioned below.



Task 3:

For the given dataset, using the cosine similarity metric would be better for grouping similar literary items together based on their topic. This is because cosine similarity ignores magnitude, unlike the euclidean distance. For example, the two articles are on separate topics, but due to their page size they are limited on words. This would lead euclidean distance to find the 2 articles more similar than the books. This contradicts what we want in grouping similar items by topic.