School of Information Technologies
Faculty of Engineering & IT

ASSIGNMENT/PROJECT COVERSHEET - GROUP ASSESSMENT

Unit of Study: **COMP5318 Machine Learning and Data Mining**

Assignment name: Assignment 2: Clustering

Tutorial time:____________________**Tutor name:**____________________________

DECLARATION
We the undersigned declare that we have read and understood the *University of Sydney Academic Dishonesty and Plagiarism in Coursework Policy*, an, and except where specifically acknowledged, the work contained in this assignment/project is our own work, and has not been copied from other sources or been previously submitted for award or assessment.

We understand that failure to comply with the *Academic Dishonesty and Plagiarism in Coursework Policy* can lead to severe penalties as outlined under Chapter 8 of the *University of Sydney By-Law 1999* (as amended). These penalties may be imposed in cases where any significant portion of my submitted work has been copied without proper acknowledgement from other sources, including published works, the internet, existing programs, the work of other students, or work previously submitted for other awards or assessments.

We realise that we may be asked to identify those portions of the work contributed by each of us and required to demonstrate our individual knowledge of the relevant material by answering oral questions or by undertaking supplementary work, either written or in the laboratory, in order to arrive at the final assessment mark.

Project team members				
Student name	**Student ID**	**Participated**	**Agree to share**	**Signature**
1. **Dayun Liu**	**490536519**	Yes	Yes	**Dayun Liu**
2. **Xiaochuan Xia**	**490353617**	Yes	Yes	**Xiaochuan Xia**
3.		Yes	Yes	
4.		Yes / No	Yes / No	
5.		Yes / No	Yes / No	
6.		Yes / No	Yes / No	
7.		Yes / No	Yes / No	
8.		Yes / No	Yes / No	
9.		Yes / No	Yes / No	
10.		Yes / No	Yes / No	

Assignment 2 report

0. Data Setup (Load and Preprocessing)

a. Load the data:

Before the start of our project, split the one dataset file into two CSV files and uploaded to google drive. Then download the files separately and use pandas to read both files.

b. Preprocessing:

For dataset 1, A PCA algorithm is used to decrease the dimension of data to 2 dimension which is easier to draw a diagram and can do better in cluster algorithms.

For dataset 2, We use TfidfVectorizer as lab 9. We select only ‘author_keywords’ and ‘abstract’ for the features. First, we combine these two columns and make it to one string for each row. The change into NumPy array and use theTfidfVectorizer covert them to TF-IDF features.

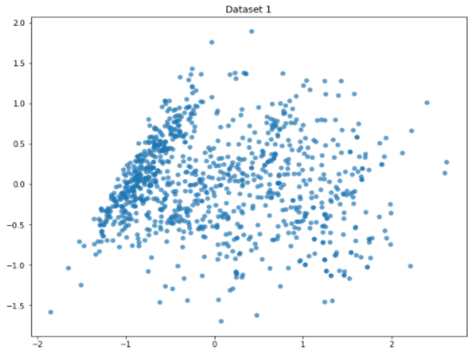


Figure 1

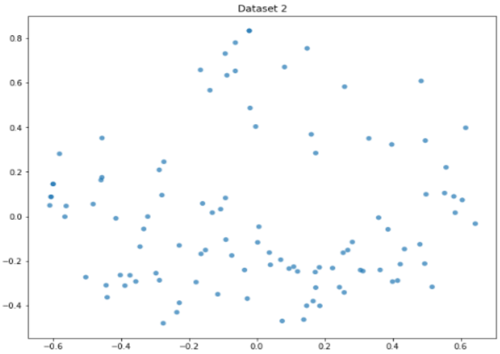


Figure 2

1.**K-means Clustering (2 marks)**

1.1. Travel Reviews Dataset

a. Similarity measure selection:

In sklearn’s Kmeans package, it is not possible to alter the distance metrics and in order to make the Euclidean distance not inflated, the PCA algorithm is needed to reduce the dimension of the data.

b. Evaluations:

(1) The elbow method

The first evaluation used is the elbow method. Increase the K will cause improve fit and eventually to overfit, the elbow method gave a direct change from underfitting to overfit, which can be used to determine the optimal K. (see appendix 1)

(2) Silhouette coefficient

The silhouette score presents the quality directly by a score. The score is a number between -1 and 1, which the closer to 1 means better clustering, -1 means incorrect clustering and 0 means overlapping cluster. By sklearn.metric’s silhouette_score function we can calculate the score directly. As the graph shows, 5 is the optimal value for K. (See appendix2)

c. Result and visualization:

The optimal K is 5 based on both evaluation methods, so we can draw the cluster diagram as K=5. In this diagram, different colors mean different clusters and the grep point means the centroid of each cluster.

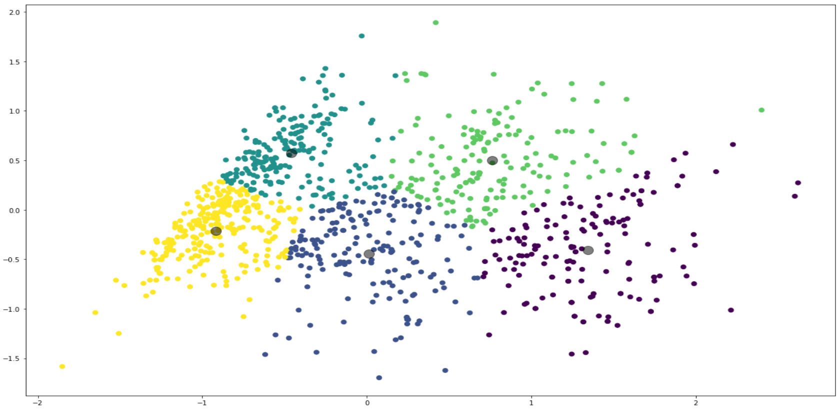


Figure 3

1.2 CMLA 2014 Accepted Papers Dataset

a. Similarity measure selection:

Because k-means package cannot change the distance measure, for this dataset, Euclidean distance is used.

b. Evaluations:

(1) F measure:

The advantage of f-measure is it considered both precision and recall. The measure outputs a score between 0 and 1. The cluster is best at 1and worst at 0. According to the f-measure, 24 is the optimal value for K. (See appendix 3)

(2) Adjusted rand index:

The adjusted rand index method has some advantages. First, it's better for random labels and the score is bounded from -1 to 1. Adjusted rand index method also does no assumption on cluster structure, which is good for our dataset. According to Adjusted rand scores, 16 or 21 is the optimal value for K. (see appendix 4)

c. Result and visualization:

Dataset 2 visualization with optimal K =24

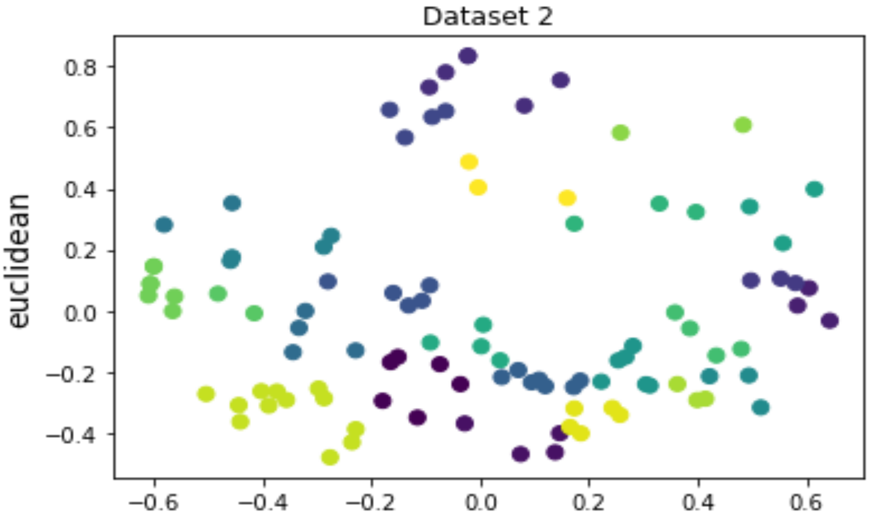


Figure 4

2. Hierarchical Clustering (2 marks)

2.1. *Travel Reviews Dataset*

a. Evaluation and Similarity measure selection:

Euclidean distance was chosen for this dataset. By evaluating all Euclidean, Manhattan and haversine distance in DBSCAN part, and find Euclidean is most suitable for this dataset.

(1) Silhouette score:
By using the silhouette score to interpret which linkage should be the best for this dataset, it turns out for ward linkage, the optimal K is 5 and for other linkages the optimal K value is 2. So, it is decided to use ward. (see appendix 5-8)

(2) Calinski-Harabasz Index
The evaluation used the linkage we selected in the previous part and use Calinski-Harabasz Index as the second evaluation. The score is higher for a standard cluster and the computation is really fast. As the graph shows, the optimal K value is 5. (see appendix 9)

b. Result and visualization:
The cluster results with K=5, linkage=" ward".

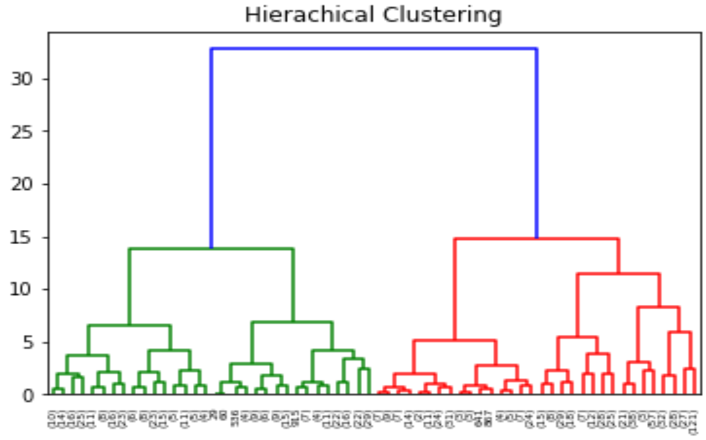


Figure 5

2.2. ICMLA 2014 Accepted Papers Dataset

a. Evaluation and Similarity measure selection:
Manhattan distance was chosen for this dataset, after tried 3 different distance metrics in the DBSCAN model and only find Manhattan performs best on this dataset. The same method as dataset 1 was applied to determine which linkage, for this dataset, and use F-measure to find the best linkage.

(1) F-measure
According to the result (see appendix 10-13), ward linkage will be used.
(2) Adjusted rand index
According to the results (see appendix 14). The optimal K is 5 or 23.

b. Result and visualization:

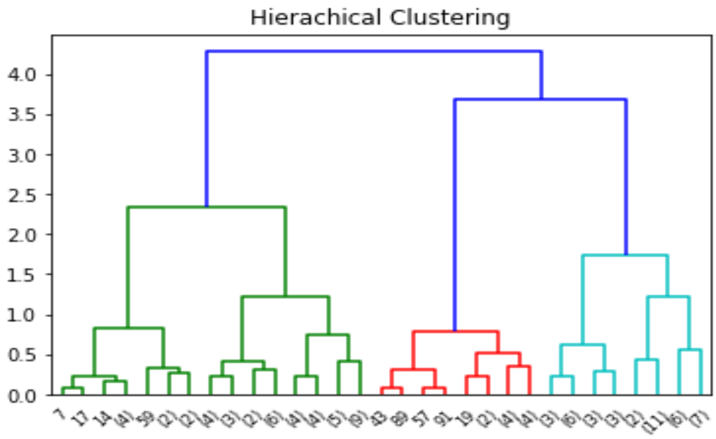


Figure 6

3. DBSCAN Clustering (2 marks)

3.1. Travel Reviews Dataset

a. The optimal eps and min_sample:

As used the method kth nearest neighbor, we found that the sharp change always occurs when then eps is relatively large which will cause all data into one cluster.

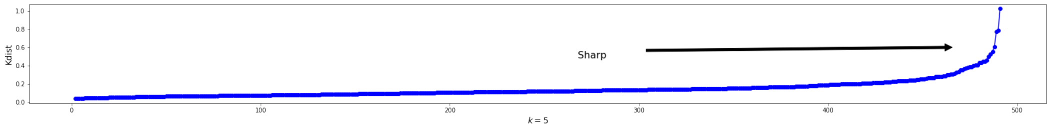


Figure 7

b. First evaluation method and Similarity measure selection:

The silhouette score was used to determine the best eps, min samples and optimal similarity measure. After the experiment, outcomes the best distance metric is Euclidean and the best eps and min_sample is 0.19 and 7. (see appendix 15-17)

c. Second Evaluation methods

As we know from b, Euclidean distance has a higher score, so we use Euclidean metric and use the Calinski-Harabasz Index to measure the model. we can see the Calinski-Harabasz index showed 0.12, 9 are the best parameters. (see appendix 18)

d. Result and Visualization

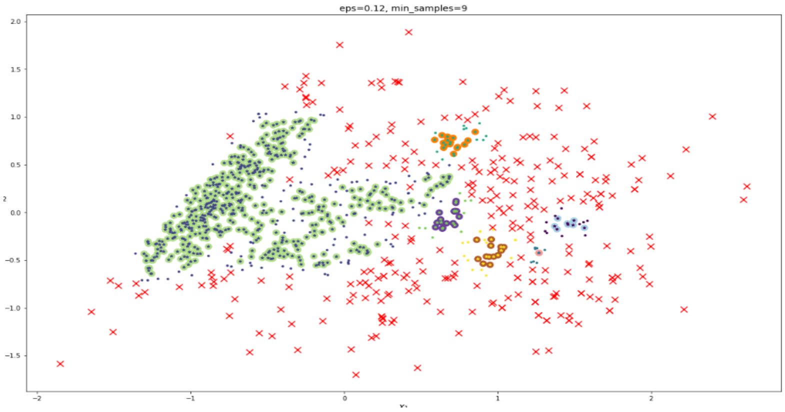


Figure 8

3.2 ***ICMLA 2014 Accepted Papers Dataset***

a. First evaluation method and Similarity measure selection:

The same methodology as the first dataset was applied to determine the best parameters and using the adjusted rand index for this data set. After experiment, the distance metric is Manhattan, the best parameters are 0.09, 2. (see appendix 19-21)

b. Second evaluation:
For the second evaluation, f-measure was used. The best parameters are 0.08,2. (see appendix 22)

c. Result and visualization
The distance metric is Manhattan.

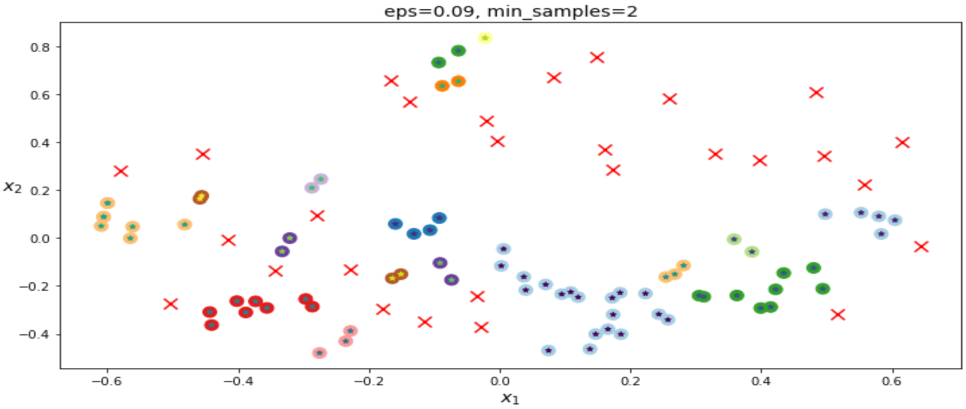


Figure 9

4. The Best Model (2 marks)

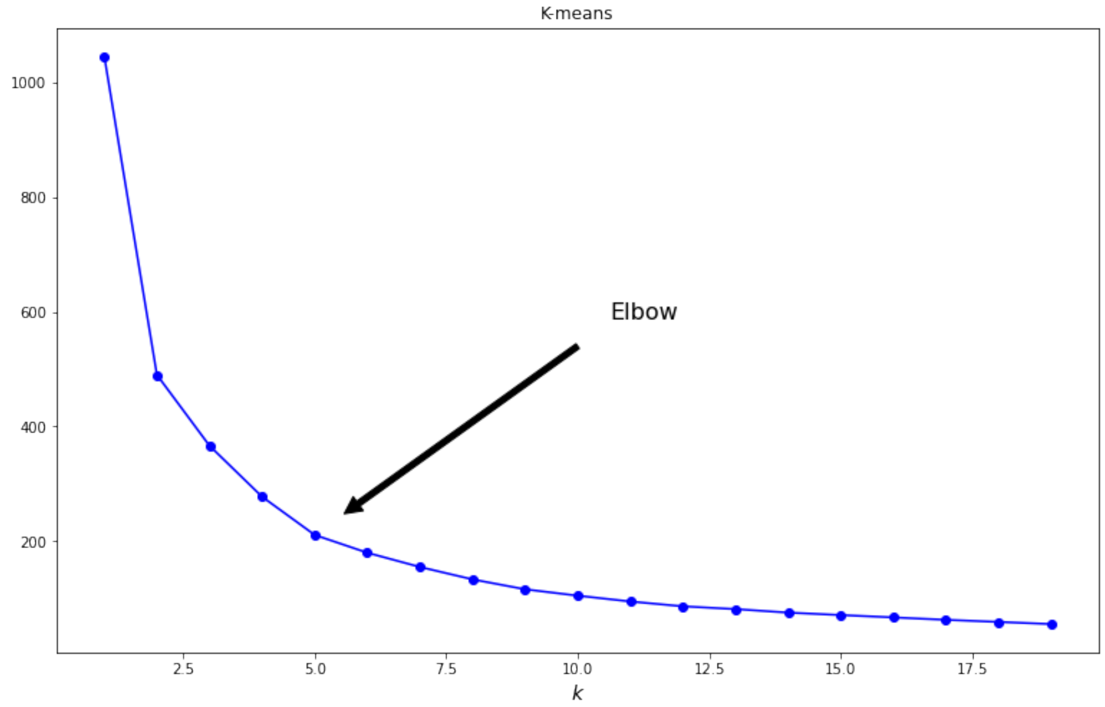
4.1 *Travel Reviews Dataset*

We think the best model is the K-means. Since we used PCA to reduce the dimension, we think this only can refer to some characteristics of the data set. We did the silhouette coefficient method to all 3 models, K-means has the highest score and according to the graph, it is the easiest one for observing clusters. And for this dataset we think it is not suitable for DBSCAN.

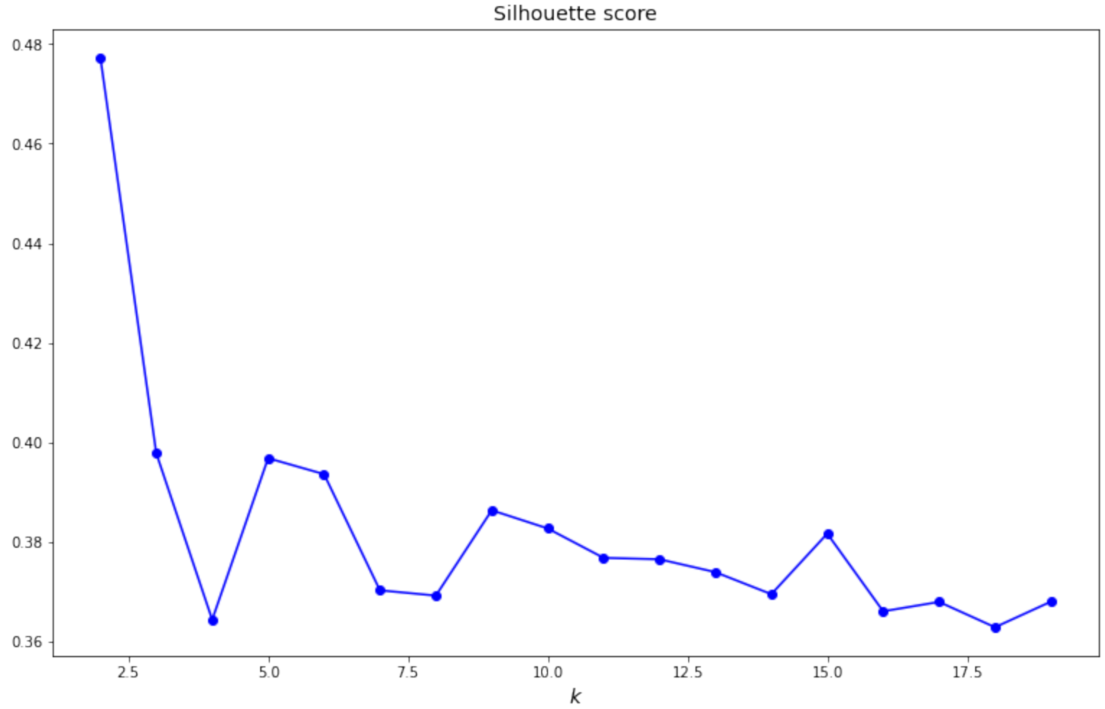
4.2 ICMLA 2014 Accepted Papers Dataset

We think the best model for this dataset is hierarchical clustering. Because K-means has a low adjusted rand score and hierarchical doesn’t need a particular K and easy to visualize in this case. By k-means and DBSCAN it did not have a nice graph to show the relationship of each paper, however, hierarchical can do a quite good tree to present the relationship between each book.

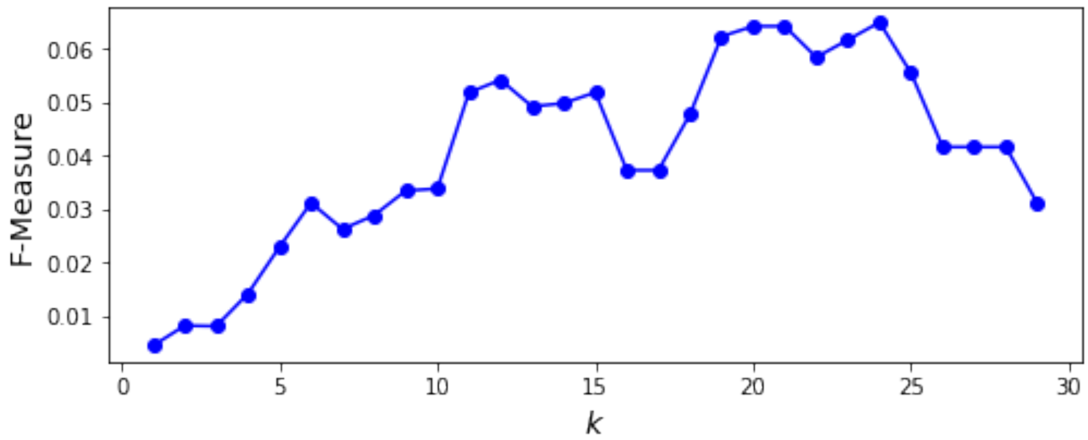
Appendix 1: Dataset1 Elbow method for K-means:



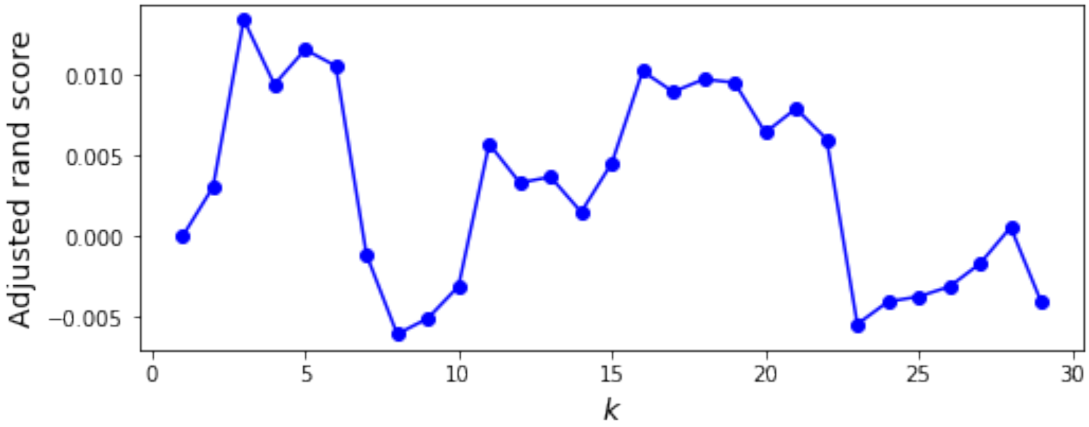
Appendix 2: Dataset1Silhouette coefficient for K-means:



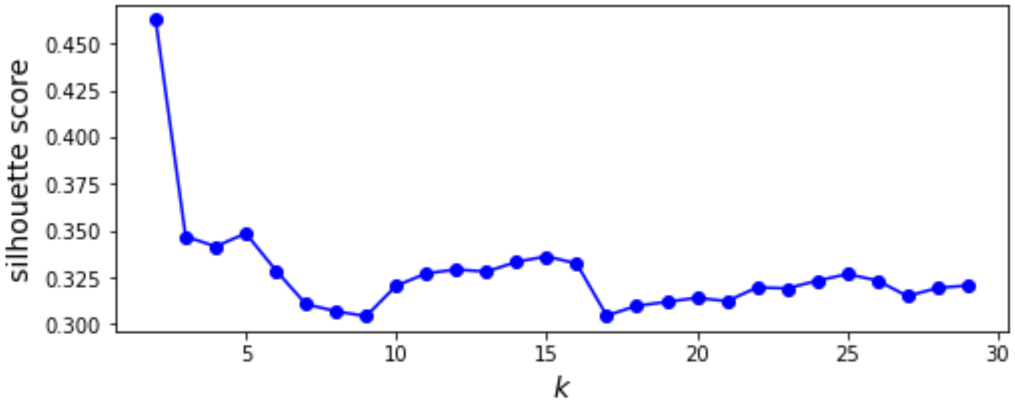
Appendix 3: Dataset 2 F-measure for K-means:



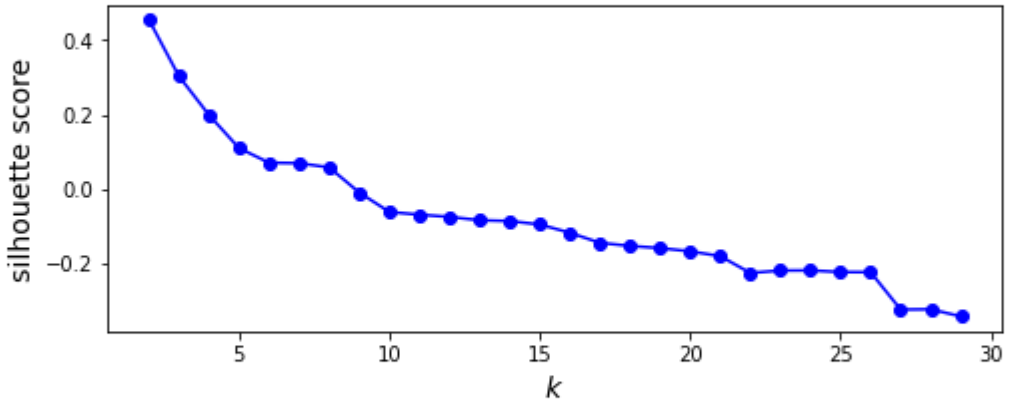
Appendix 4: Dataset 2 adjusted rand index for K-means:



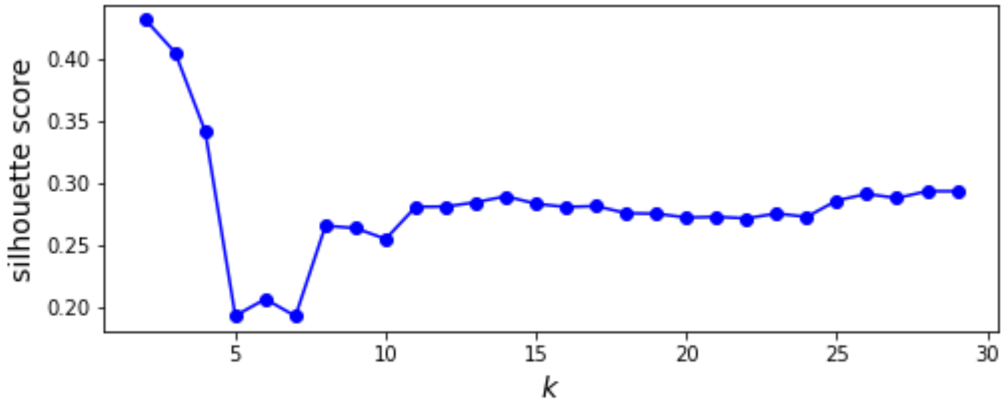
Appendix 5: Dataset 1 Silhouette score for hierarchical (ward linkage):



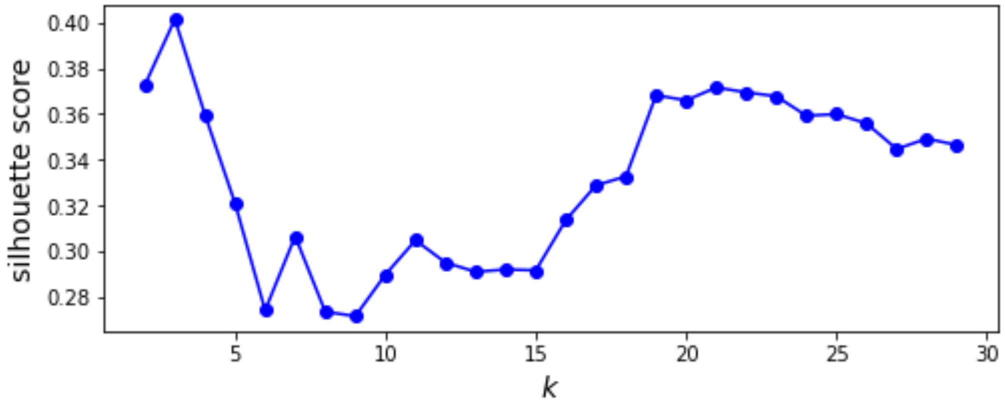
Appendix 6: Dataset 1 Silhouette score for hierarchical (complete linkage):



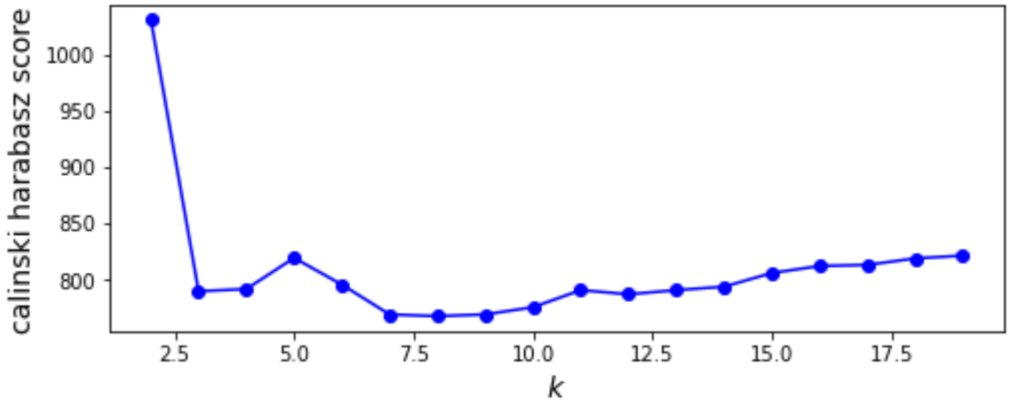
Appendix 7: Dataset 1 Silhouette score for hierarchical (average linkage):



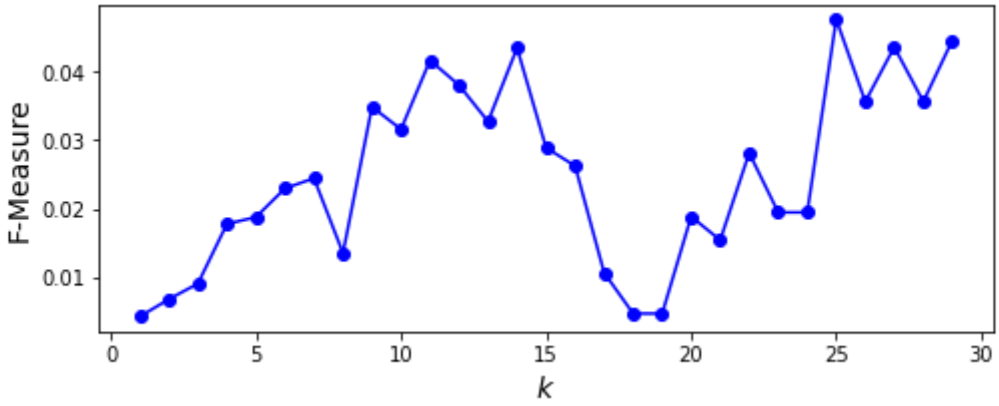
Appendix 8: Dataset 1 Silhouette score for hierarchical (single linkage):



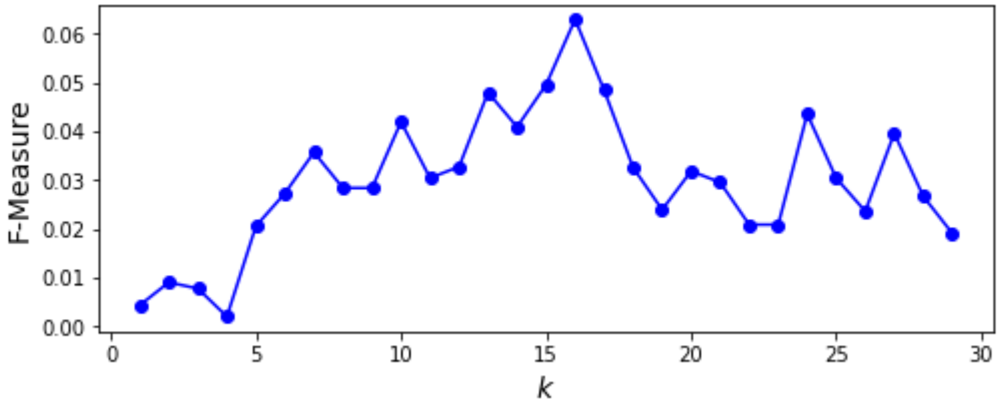
Appendix 9: Dataset1 Calinski-Harabasz index for heraichical:



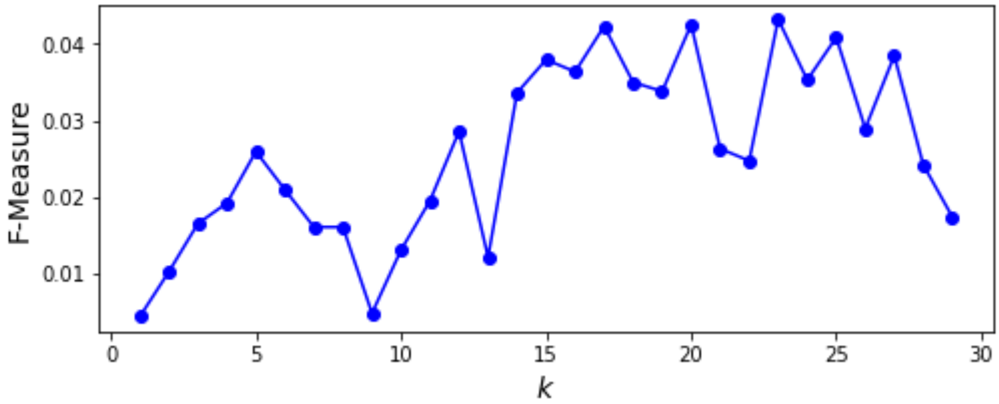
Appendix 10: Dataset 2 Silhouette score for hierarchical (ward linkage)



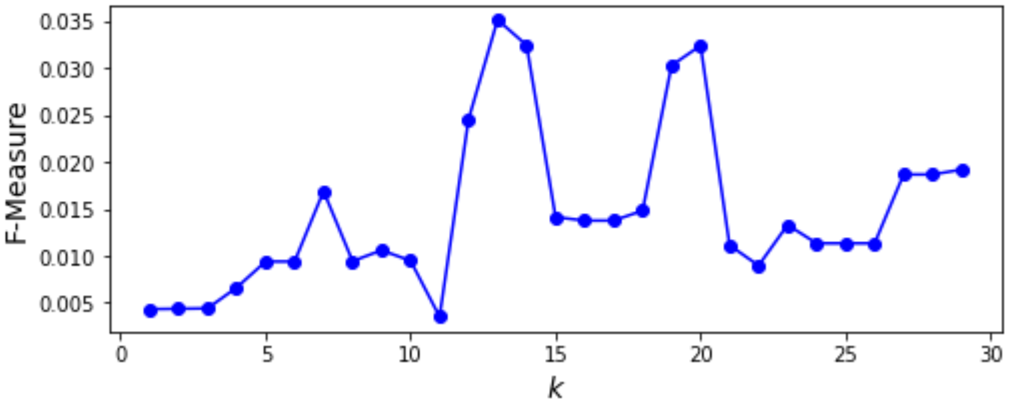
Appendix 11: Dataset 2 Silhouette score for hierarchical (complete linkage):



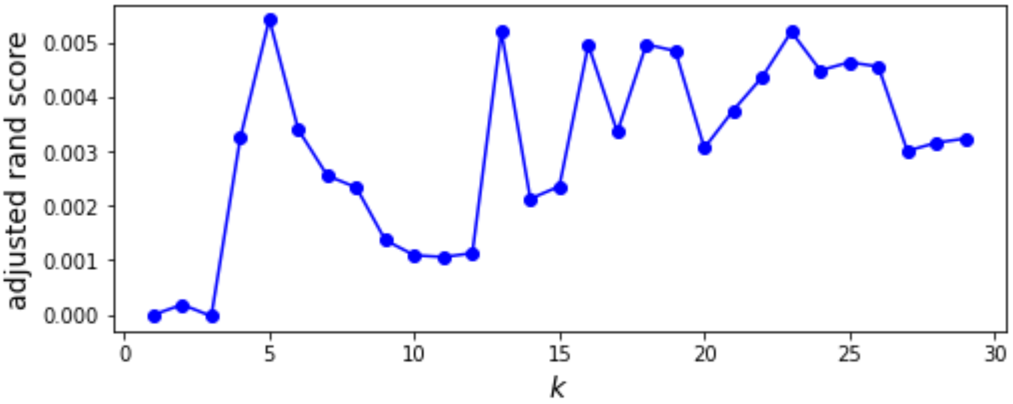
Appendix 12: Dataset 2 Silhouette score for hierarchical (average linkage):



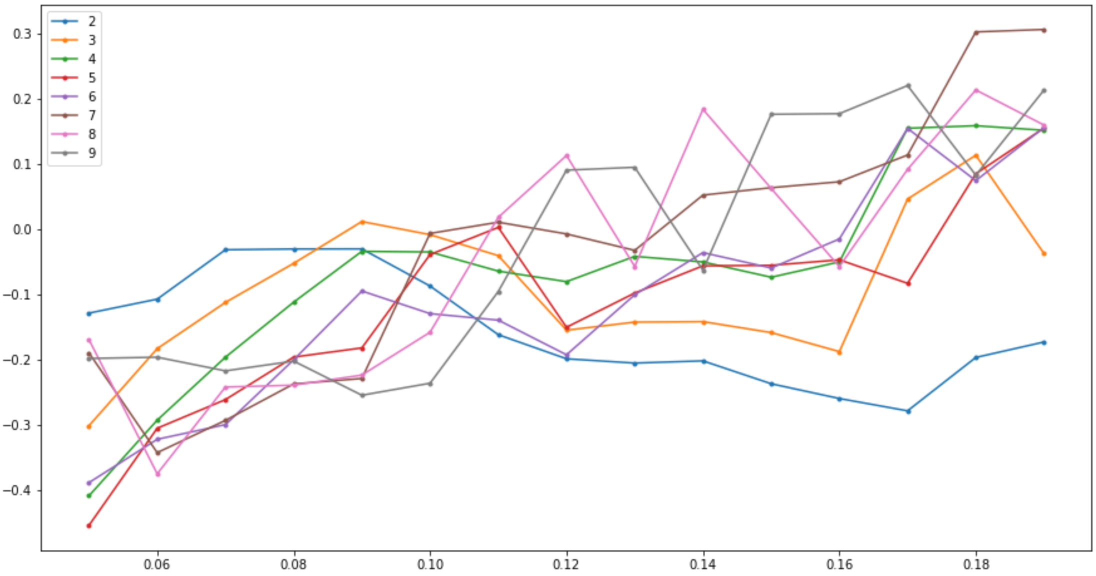
Appendix 13: Dataset 2 Silhouette score for hierarchical (single linkage):



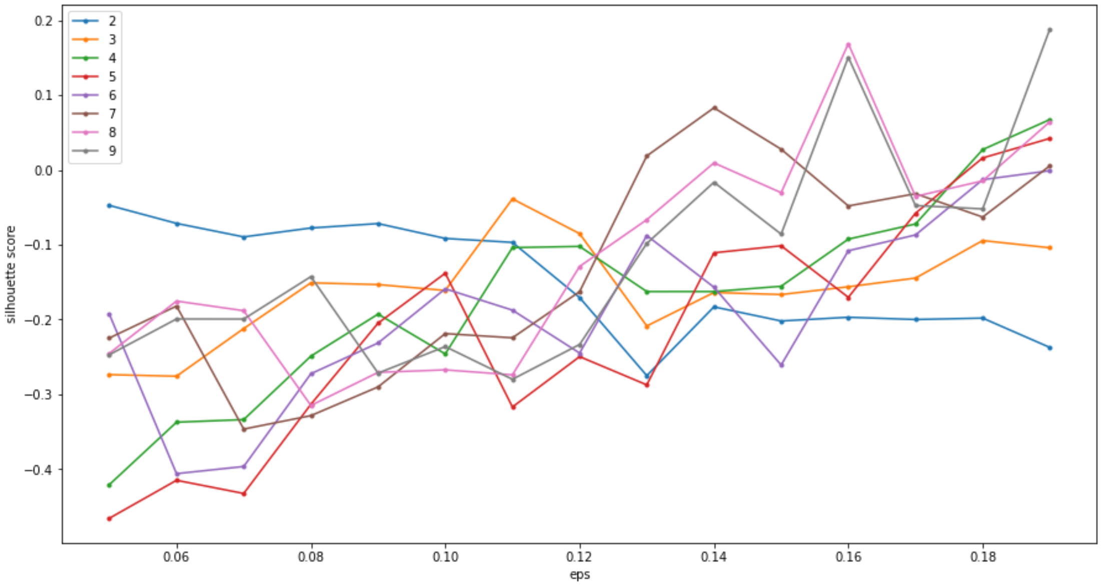
Appendix 14: Dataset 2 adjusted rand index score for hierarchical:



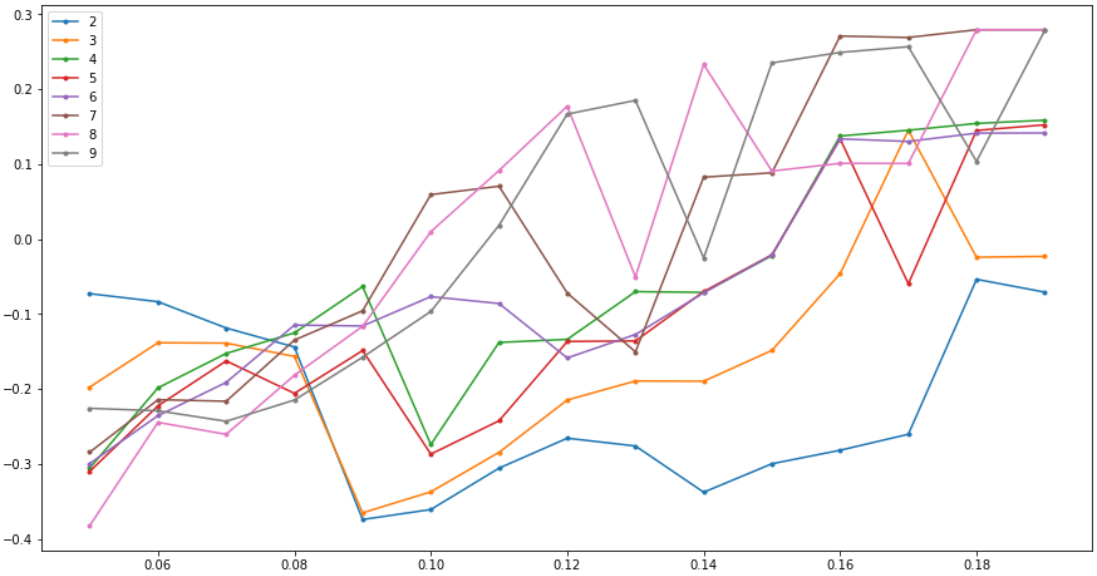
Appendix 15: Dataset 1 Euclidean distance silhouette score for DBSCAN:



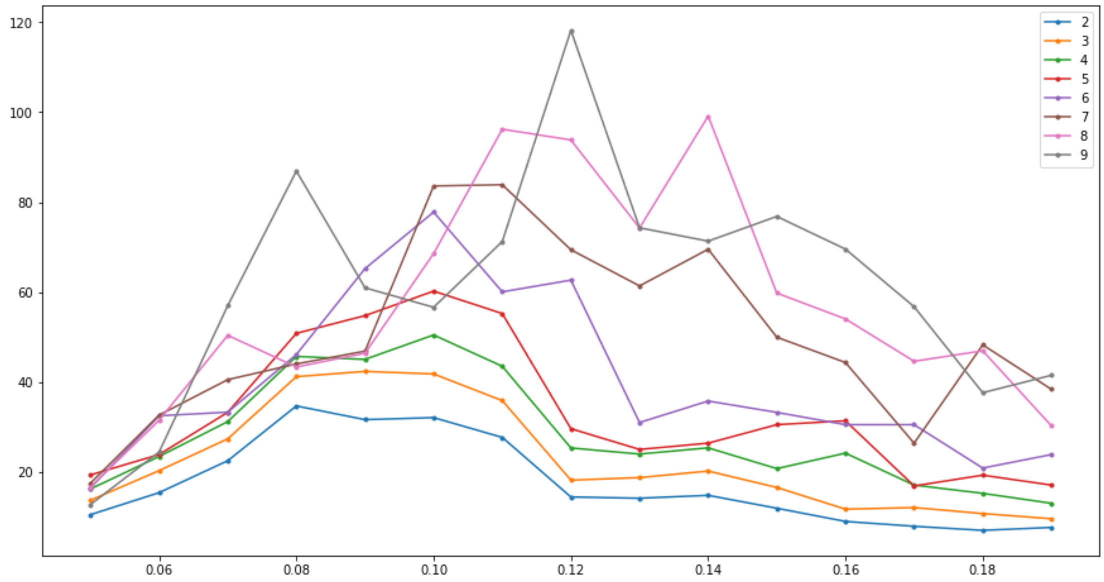
Appendix 16: Dataset 1 Manhattan distance silhouette score for DBSCAN:



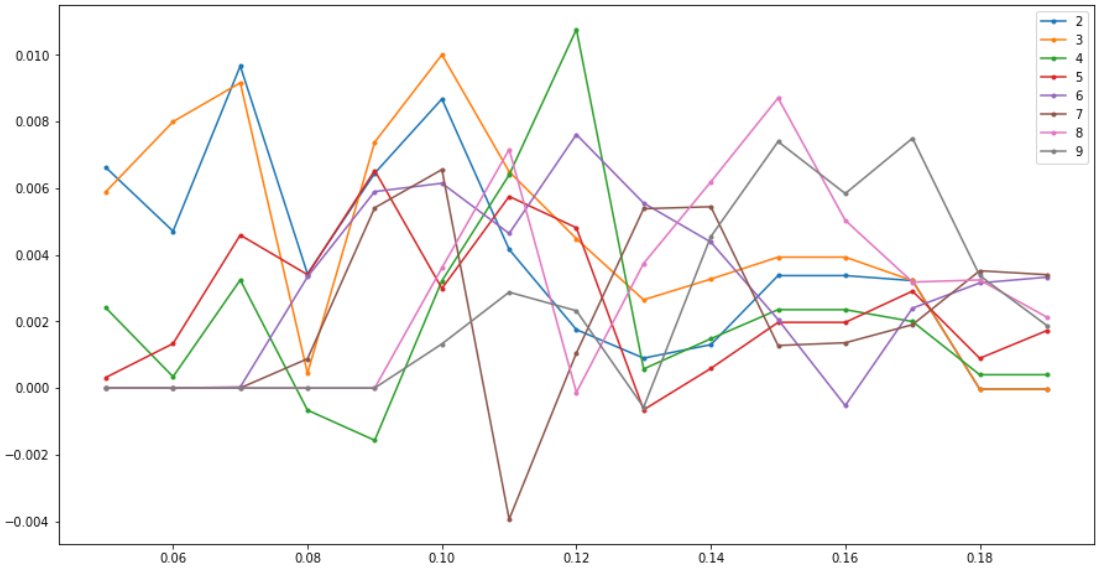
Appendix 17: Dataset 1 haversine distance silhouette score for DBSCAN:



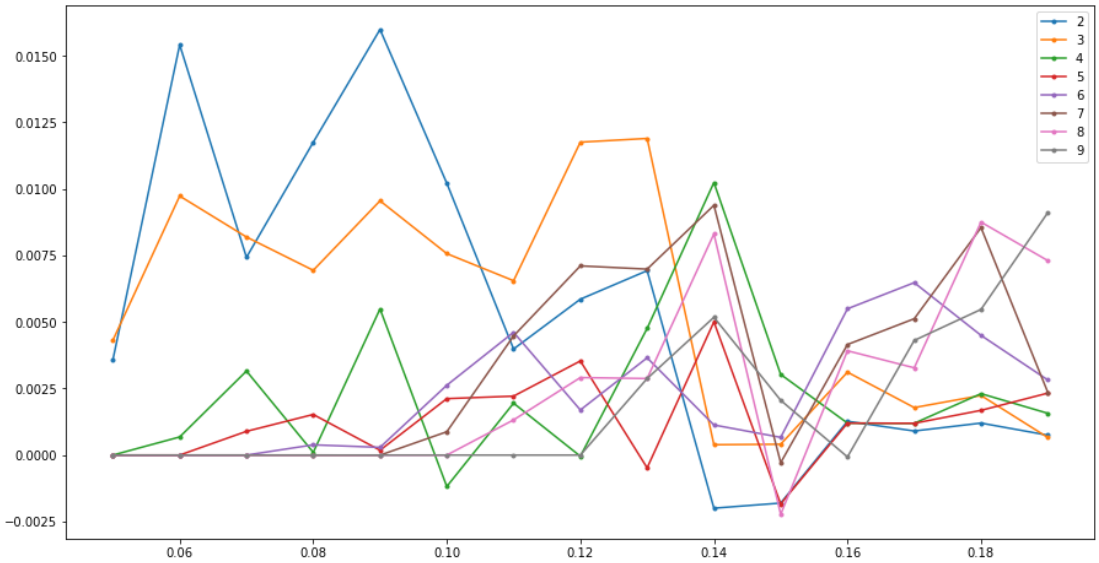
Appendix 18: Dataset1 Calinski-Harabasz index:



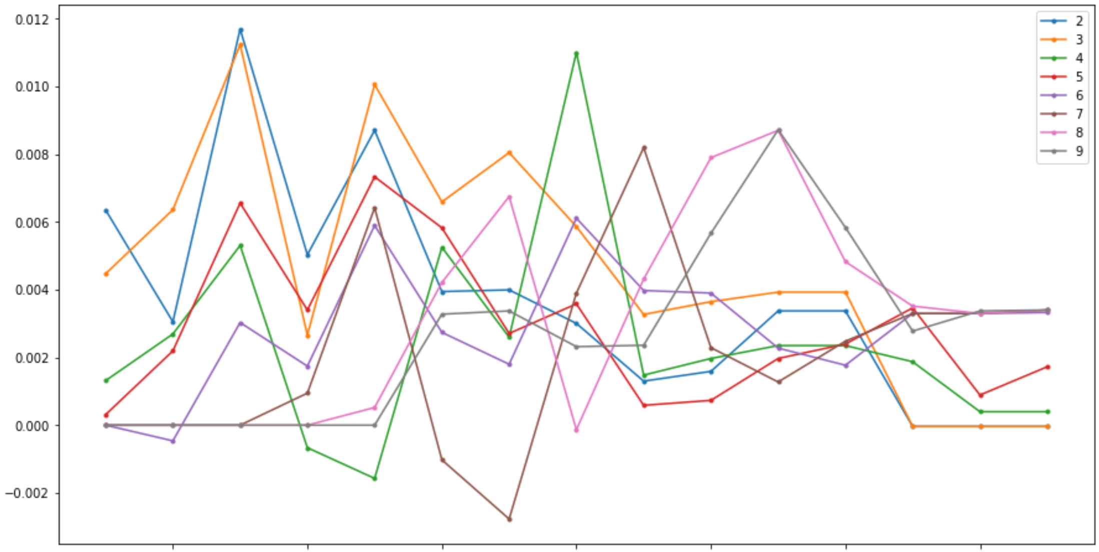
Appendix 19: Dataset 2 Euclidean distance silhouette score for DBSCAN:



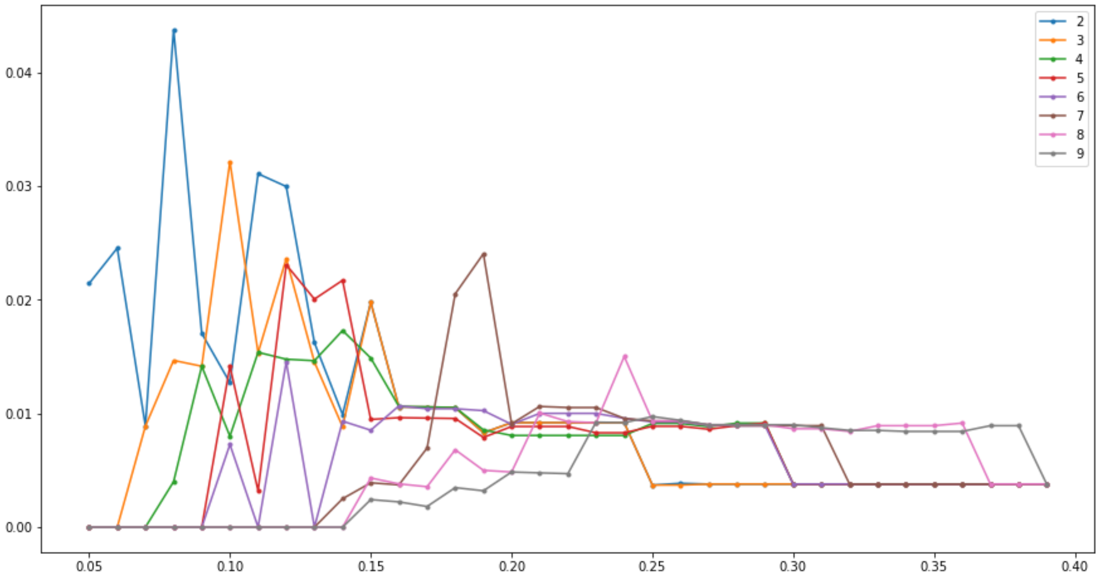
Appendix 20: Dataset 2 Manhattan distance silhouette score for DBSCAN:



Appendix 21: Dataset 2 haversine distance silhouette score for DBSCAN:



Appendix 22: Dataset 2 f-measure score for DBSCAN


0.04
0.03
0.02
0.01
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
2
3
4
5
6
7
8
9