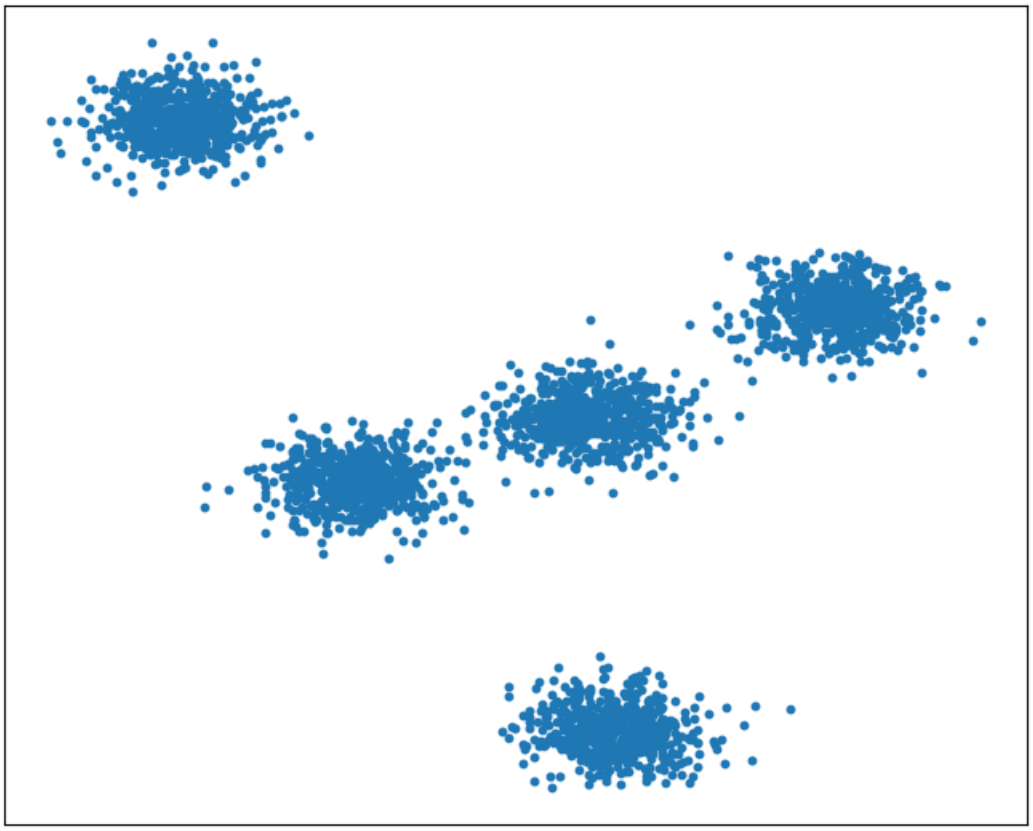
ASSIGNMENT 4

Christos Eleftheriou

1009537

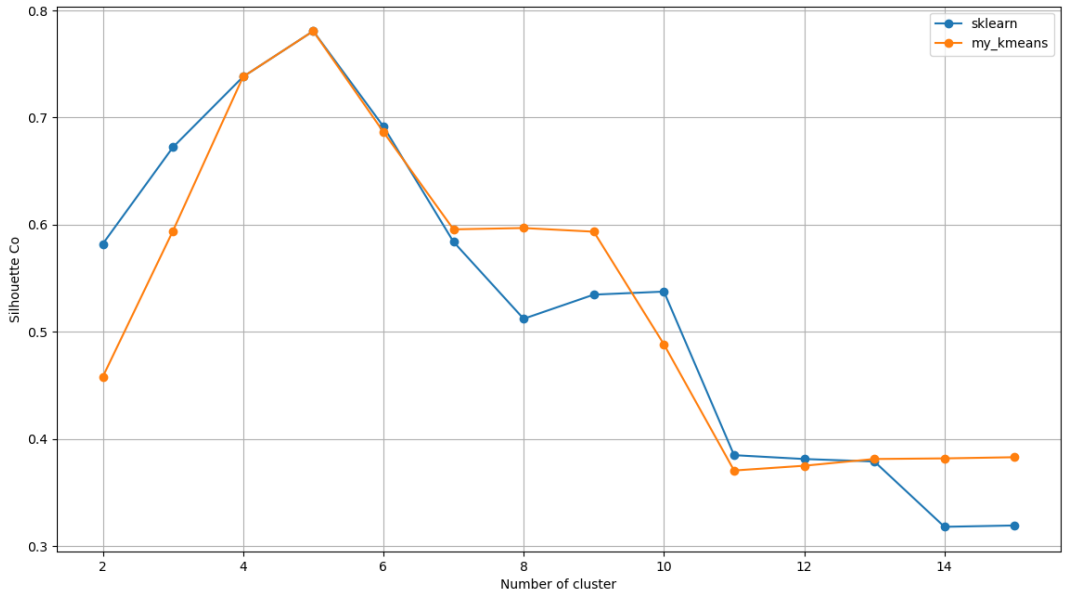
Part A – Clustering

1.



3.

a.

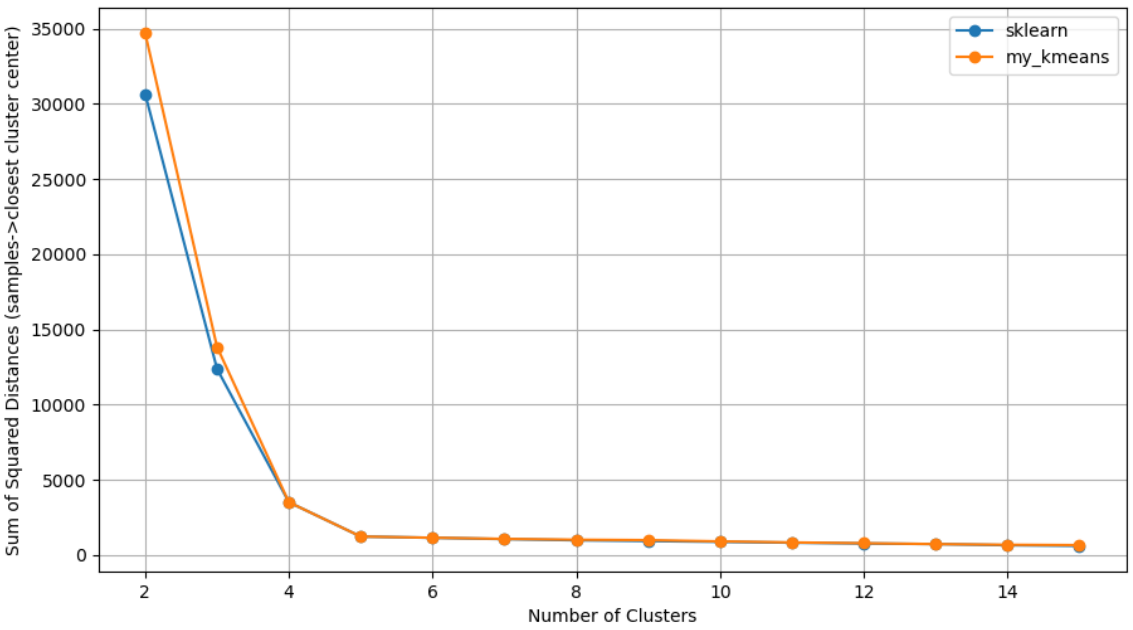


b.

To compare the silhouette scores of MyKMeans and Scikit-learn's kmeans, we can check the number of clusters that corresponds to the highest silhouette score for each of them. We can see that for both of them, the highest silhouette score occurs for 5 clusters.

4.

a.

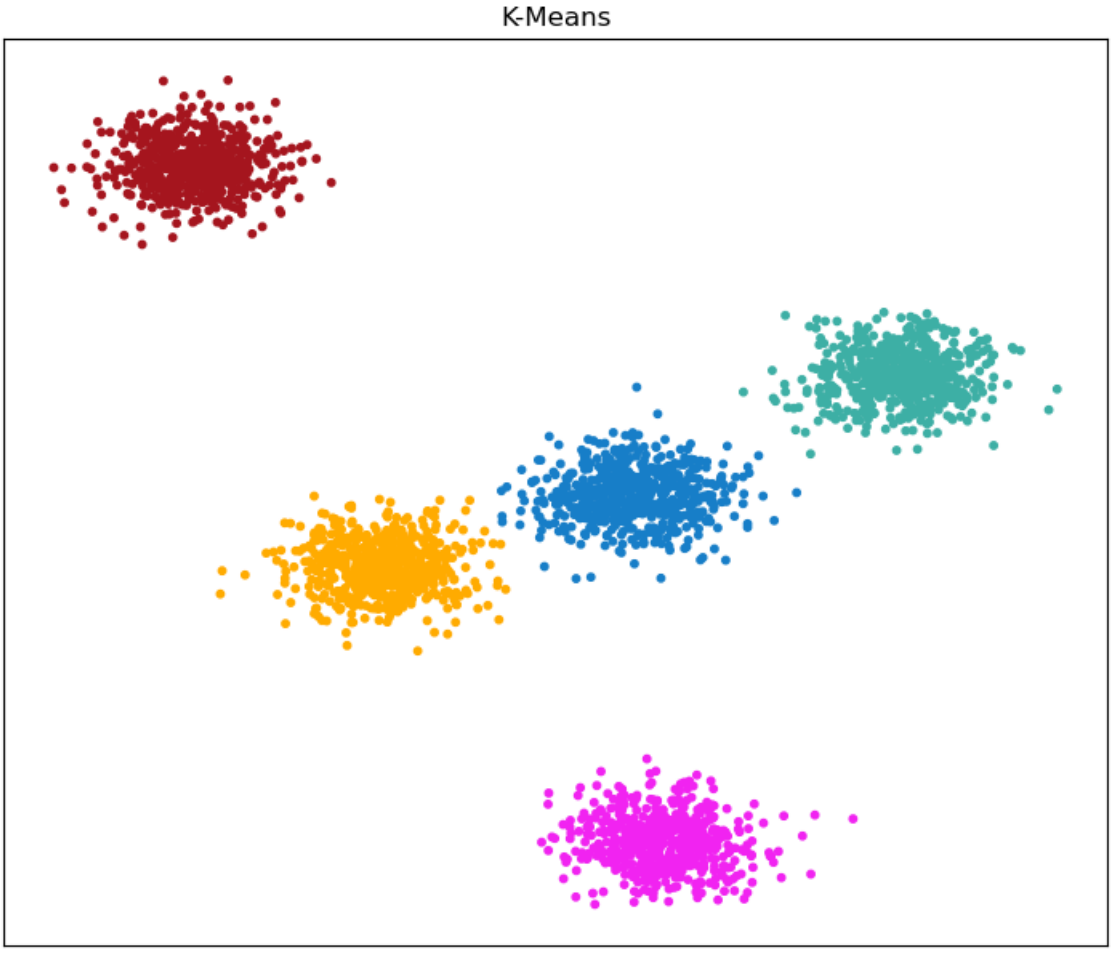


b.

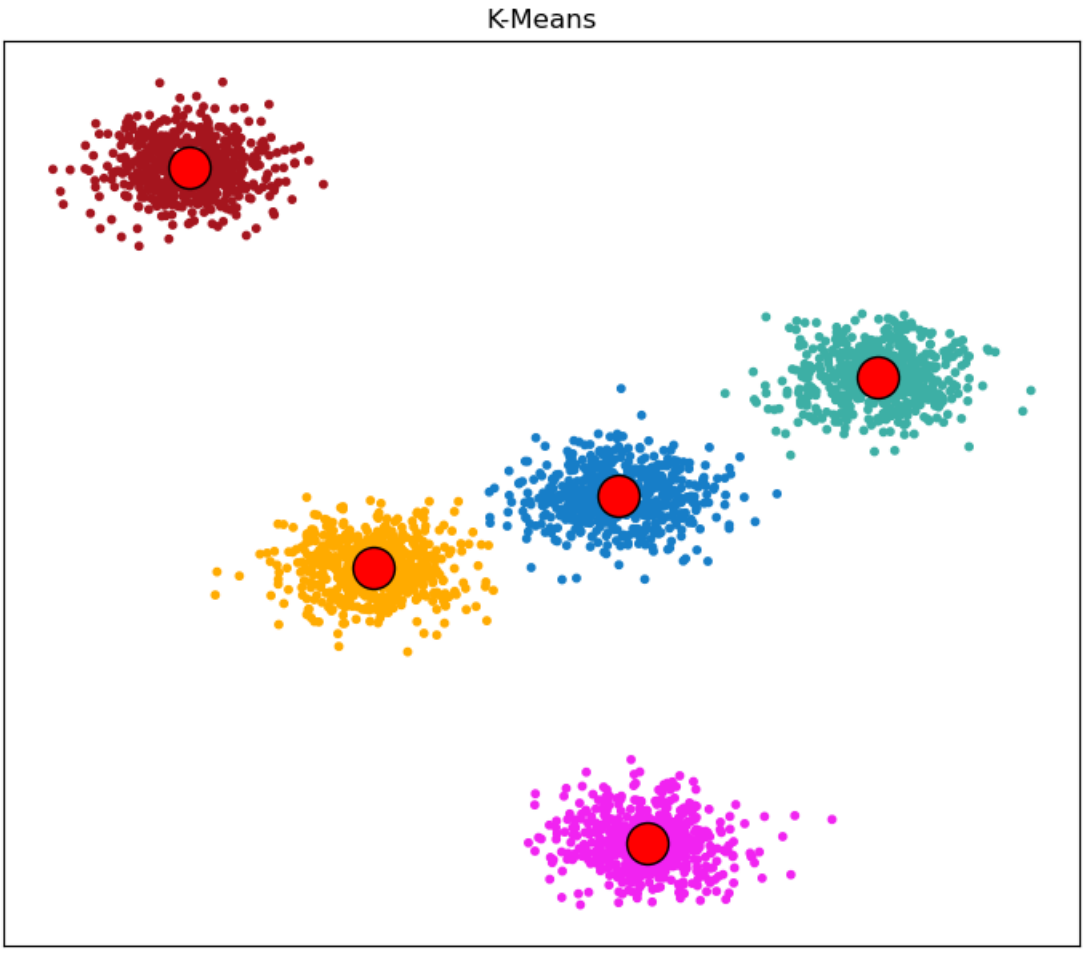
To compare the inertia plots of MyKMeans and Scikit-learn's kmeans, we can check the number of clusters for which the rate of inertia slows down identifying the elbow, which is the optimal number of clusters. We can see that that point is at 5 clusters for both of them.

5.

a.



b.



Part B – Anomaly Detection

1.

We should use outlier detection because outliers already exist in the training data, and we try to find them.

2.

```
Fraudulent transactions in original dataset: 487
Legitimate transactions in original dataset: 28423
```

3.

```
Percentage of fraudulent transactions in original dataset: 1.6845382220684884
```

4.

```
Fraudulent transactions captured by IsolationForest: 486
Fraudulent transactions captured by OneClassSVM: 4930
```

5.

```
Legitimate transactions incorrectly classified as fraudulent by IsolationForest: 224
Legitimate transactions incorrectly classified as fraudulent by OneClassSVM: 4681
```

6.

```
Percentage of frauds in the original dataset that were detected by IsolationForest: 53.798767967145785
Percentage of frauds in the original dataset that were detected by OneClassSVM: 51.129363449691986
```

The metric you are referring to is known as recall in machine learning. It represents the percentage of actual positive instances (in this case, frauds) that are correctly identified by a model.

7.

```
Time to train IsolationForest: 1.523932933807373
Time to train OneClassSVM: 88.4242057800293
```

The time to train the Isolation Forest is much smaller than the training time of One Class SVM. This is happening because of the nature of the algorithms, where Isolation Forest is a method based on decision trees, whereas One Class SVM involves solving a quadratic optimization problem. This makes its training process more computationally expensive.

8.

Isolation Forest

The Isolation Forest tries to isolate observations. It creates random forests by randomly selecting features and then selecting their corresponding split value. The number of splits required to isolate a sample is equal to the path length from the root to the leaf. This path length is used as a measure for normality. Anomalies tend to produce noticeably shorter paths, therefore when random forests produce shorter path lengths for particular samples, they are highly likely to be anomalies.

One Class SVM

One-class SVM uses a hypersphere to encompass all instances of the known class (normality). It tries to create the smallest possible hypersphere. Everything outside the hypersphere is considered an anomaly.

9.

An anomaly that I can spot is that Isolation Forest captured 486 outliers while One Class SVM captured 4930. This big difference between the two is unusual, and can attributed to the difference in the nature of each of them.

10.

A.

This anomaly detection is called Novelty Detection because we try to detect outliers in the new sample.

B.

```
Number of anomalies IsolationForest detected: 2
Number of anomalies OneClassSVM detected: 6
```