Experiment:K Means Clustering

Name:	Vivek Rajput
UID:	2019110046
Class:	BE ETRX
Batch:	A
Subject:	Data Analytics Lab

Objective: The focus of this lab is k-means clustering. We will look at the vanilla algorithm, its performance, and some better variants. Finally, we will use clustering for classifying the MNIST data set.

System Requirements: Python 3.9 or visual studio code

Code:

```
from math import *
import random
from copy import deepcopy
import numpy as np

def argmin(values):
    return min(enumerate(values), key=lambda x: x[1])[0]
def avg(values):
    return float(sum(values))/len(values)

def readfile(filename):
    '''
    File format: Each line contains a comma separated list of real
numbers, representing a single point.
    Returns a list of N points, where each point is a d-tuple.
    '''
    data = []
    with open(filename, 'r') as f:
        data = f.readlines()
    data = [tuple(map(float, line.split(','))) for line in data]
    return data

def writefile(filename, means):
    '''
    means: list of tuples
```

```
    Writes the means, one per line, into the file.
    '''
    if filename == None: return
    with open(filename, 'w') as f:
        for m in means:
            f.write(','.join(map(str, m)) + '\n')
    print('Written means to file ' + filename)


def distance_euclidean(p1, p2):
    '''
    p1: tuple: 1st point
    p2: tuple: 2nd point
    Returns the Euclidean distance b/w the two points.
    '''

    distance = None

    # TODO [task1]:
    # Your function must work for all sized tuples.
    dist = [(x1-x2)**2 for x1, x2 in zip(p1, p2)]
    distance = sqrt(sum(dist))
    ########################################
    return distance

def distance_manhattan(p1, p2):
    '''
    p1: tuple: 1st point
    p2: tuple: 2nd point
    Returns the Manhattan distance b/w the two points.
    '''

    # k-means uses the Euclidean distance.
    # Changing the distant metric leads to variants which can be more/less
robust to outliers,
    # and have different cluster densities. Doing this however, can
sometimes lead to divergence!

    distance = None
```

```
    # TODO [task1]:
    # Your function must work for all sized tuples.
    distance = sum([abs(x1-x2) for x1, x2 in zip(p1, p2)])
    ########################################
    return distance

def initialization_forgy(data, k):
    '''
    data: list of tuples: the list of data points
    k: int: the number of cluster means to return
    Returns a list of tuples, representing the cluster means
    '''


    means = []
    means = random.sample(data,k)

    # TODO [task1]:
    # Use the Forgy algorithm to initialize k cluster means.

    ########################################
    assert len(means) == k
    return means


def initialization_kmeansplusplus(data, distance, k):
    '''
    data: list of tuples: the list of data points
    distance: callable: a function implementing the distance metric to use
    k: int: the number of cluster means to return
    Returns a list of tuples, representing the cluster means
    '''


    means = []

    # TODO [task3]:
    # Use the kmeans++ algorithm to initialize k cluster means.
    # Make sure you use the distance function given as parameter.

    # NOTE: Provide extensive comments with your code.
```

```
    # The first centroid is randomly selected
    means.append(random.sample(data, 1)[0])
    # Current minimum distance squared from any centroid
    # The probability with which to select a point will be proportional to
this
    min_dist = [float('Inf')] * len(data)

    # Going over means; adding 1 new mean in each loop
    for i in range(k-1):
        # Updating minimum for all points
        for j in range(len(data)):
            # Calculate minimum distance
            d = distance(means[i], data[j])
            # Squaring the distance since the probability will be
proportional to this.
            d = d*d
            if d < min_dist[j]:
                min_dist[j] = d

        # Normalising min_dist array to get an array of probabilities
        s = sum(min_dist)
        prob = [i/s for i in min_dist]
        # Choose a random index with probabilities proportional to
min_dist
        idx = np.random.choice(len(data), 1, p=prob)[0]
        # Append the chosen point to means
        means.append(data[idx])
        # Note that in the next iteration, the min_dist for points in
`means` will become 0
        # So they will not be chosen again

    ########################################
    assert len(means) == k
    return means

def initialization_randompartition(data, distance, k):
    '''
    data: list of tuples: the list of data points
    distance: callable: a function implementing the distance metric to use
    k: int: the number of cluster means to return
```

```
    Returns a list of tuples, representing the cluster means
    '''

    means = []
    means.append(random.sample(data, 1)[0])
    # TODO [task3]:
    # Use the kmeans++ algorithm to initialize k cluster means.
    # Make sure you use the distance function given as parameter.

    # NOTE: Provide extensive comments with your code.
    for i in range(k-1):
        # Updating minimum for all points
        # Choose a random index with probabilities proportional to
min_dist
        idx = np.random.choice(len(data), 1)[0]
        # Append the chosen point to means
        means.append(data[idx])
    # The first centroid is randomly selected
    #means.append(random.sample(data, 1)[0])
    # Current minimum distance squared from any centroid
    # The probability with which to select a point will be proportional to
this
    #min_dist = [float('Inf')] * len(data)
    #indices = np.random.choice(range(0, k), replace = True, size =
data.shape[0])
    # Going over means; adding 1 new mean in each loop
    #for i in range(k):
    #    means.append(data[indices == i].mean(axis=0))

    ########################################
    return means

def iteration_one(data, means, distance):
    '''
    data: list of tuples: the list of data points
    means: list of tuples: the current cluster centers
    distance: callable: function implementing the distance metric to use
    Returns a list of tuples, representing the new cluster means after 1
iteration of k-means clustering algorithm.
    '''
```

```
    new_means = []
    k = len(means)
    dimension = len(data[0])

    # TODO [task1]:
    # You must find the new cluster means.
    # Perform just 1 iteration (assignment+updation)
    new_means = [tuple(0 for i in range(dimension))] * k
    counts = [0.0] * k
    for point in data:
        closest = 0
        min_dist = float('Inf')
        for i in range(k):
            d = distance(point, means[i])
            if d < min_dist:
                min_dist = d
                closest = i
        new_means[closest] = tuple([sum(x) for x in
zip(new_means[closest], point)])
        counts[closest] += 1

    for i in range(k):
        # import pdb; pdb.set_trace()
        if counts[i] == 0:
            new_means[i] = means[i]
        else:
            new_means[i] = tuple(t/counts[i] for t in new_means[i])
    ########################################
    return new_means

def hasconverged(old_means, new_means, epsilon=1e-1):
    '''
    old_means: list of tuples: The cluster means found by the previous
iteration
    new_means: list of tuples: The cluster means found by the current
iteration
    Returns true iff no cluster center moved more than epsilon distance.
    '''
```

```
    converged = False

    # TODO [task1]:
    # Use Euclidean distance to measure centroid displacements.
    for i in range(len(old_means)):
        p = [abs(x1-x2) > epsilon for x1, x2 in zip(old_means[i], 
new_means[i])]
        if True in p:
            return False
    converged = True
    ########################################
    return converged



def iteration_many(data, means, distance, maxiter, epsilon=1e-1):
    '''
    maxiter: int: Number of iterations to perform
    Uses the iteration_one function.
    Performs maxiter iterations of the k-means clustering algorithm, and 
saves the cluster means of all iterations.
    Stops if convergence is reached earlier.
    Returns:
    all_means: list of (list of tuples): Each element of all_means is a 
list of the cluster means found by that iteration.
    '''


    all_means = []
    all_means.append(means)

    # TODO [task1]:
    # Make sure you've implemented the iteration_one, hasconverged 
functions.
    # Perform iterations by calling the iteration_one function multiple 
times.
    # Stop only if convergence is reached, or if max iterations have been 
exhausted.
    # Save the results of each iteration in all_means.
    # Tip: use deepcopy() if you run into weirdness.
    means_copy = deepcopy(means)
```

```
    for i in range(maxiter):
        new_means = iteration_one(data, means_copy, distance)
        all_means.append(new_means)
        if hasconverged(means_copy, new_means, epsilon):
            break
        means_copy = new_means
    ########################################

    return all_means



def performance_SSE(data, means, distance):

    '''
    data: list of tuples: the list of data points
    means: list of tuples: representing the cluster means
    Returns: The Sum Squared Error of the clustering represented by means,
on the data.
    '''

    sse = 0

    # TODO [task1]:
    # Calculate the Sum Squared Error of the clustering represented by
means, on the data.
    # Make sure to use the distance metric provided.
    for point in data:
        min_dist = float('Inf')
        for i in range(len(means)):
            d = distance(point, means[i])
            if d < min_dist:
                min_dist = d
        sse += min_dist*min_dist
    ########################################
    return sse



import sys
```

```
import argparse
import matplotlib.pyplot as plt
from itertools import cycle
from pprint import pprint as pprint


def parse():
    parser = argparse.ArgumentParser()
    parser.add_argument('-i', '--input', dest='input', type=str,
help='Required. Dataset filename')
    parser.add_argument('-o', '--output', dest='output', type=str,
help='Output filename')
    parser.add_argument('-iter', '--iter', '--maxiter', dest='maxiter',
type=int, default=10000,
                        help='Maximum number of iterations of the k-means
algorithm to perform. (may stop earlier if convergence is achieved)')
    parser.add_argument('-e', '--eps', '--epsilon', dest='epsilon',
type=float, default=1e-1,
                        help='Minimum distance the cluster centroids move
b/w two consecutive iterations for the algorithm to continue.')
    parser.add_argument('-init', '--init', '--initialization',
dest='init', type=str, default='forgy',
                        help='The initialization algorithm to be used.
{forgy, randompartition, kmeans++}')
    parser.add_argument('-dist', '--dist', '--distance', dest='dist',
type=str, default='euclidean',
                        help='The distance metric to be used. {euclidean,
manhattan}')
    parser.add_argument('-k', '--k', dest='k', type=int, default=5,
help='The number of clusters to use.')
    parser.add_argument('-verbose', '--verbose', dest='verbose',
type=bool, default=False, help='Turn on/off verbose.')
    parser.add_argument('-seed', '--seed', dest='seed', type=int,
default=0, help='The RNG seed.')
    parser.add_argument('-numexperiments', '--numexperiments',
dest='numexperiments', type=int, default=1,
                        help='The number of experiments to run.')
    _a = parser.parse_args()

    if _a.input is None:
```

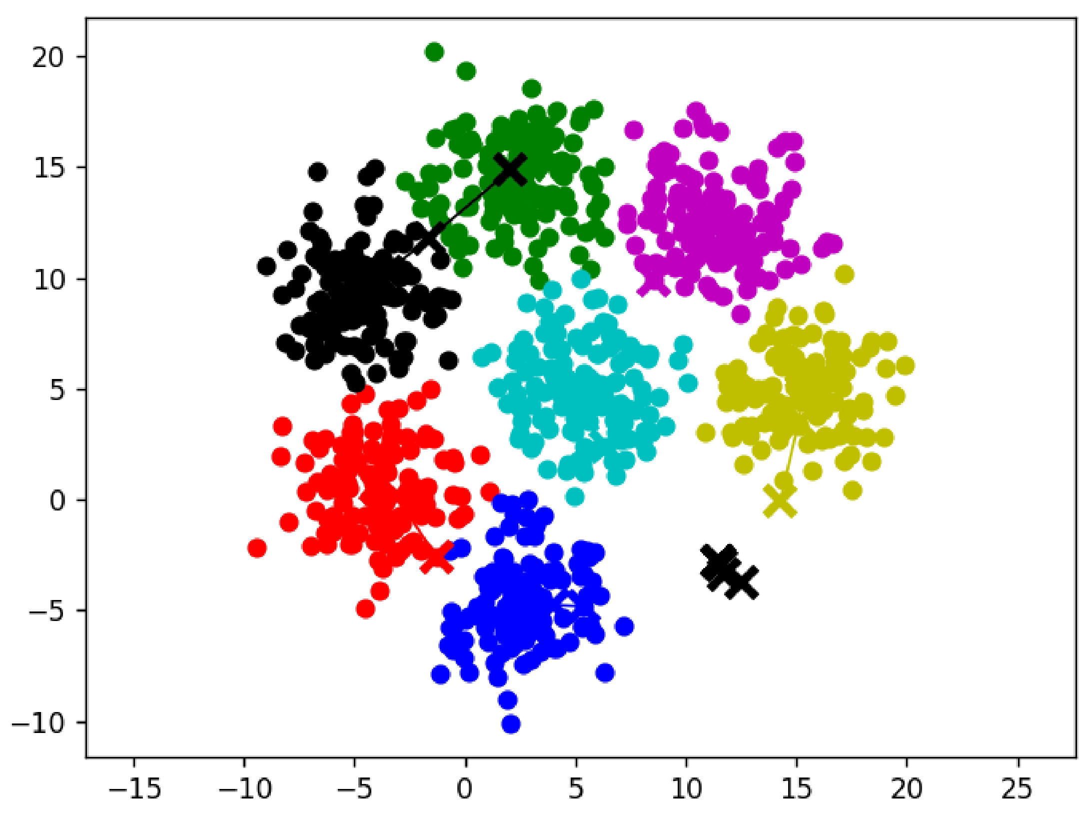
```
        print('Input filename required.\n')
        parser.print_help()
        sys.exit(1)

    args = {}
    for a in vars(_a):
        args[a] = getattr(_a, a)

    if _a.init.lower() in ['random', 'randompartition']:
        args['init'] = initialization_randompartition
    elif _a.init.lower() in ['k++', 'kplusplus', 'kmeans++', 'kmeans',
'kmeansplusplus']:
        args['init'] = initialization_kmeansplusplus
    elif _a.init.lower() in ['forgy', 'frogy']:
        args['init'] = initialization_forgy
    else:
        print('Unavailable initialization function.\n')
        parser.print_help()
        sys.exit(1)

    if _a.dist.lower() in ['manhattan', 'l1', 'median']:
        args['dist'] = distance_manhattan
    elif _a.dist.lower() in ['euclidean', 'euclid', 'l2']:
        args['dist'] = distance_euclidean
    else:
        print('Unavailable distance metric.\n')
        parser.print_help()
        sys.exit(1)

    print('-' * 40 + '\n')
    print('Arguments:')
    pprint(args)
    print('-' * 40 + '\n')
    return args


def visualize_data(data, all_means, args):
    print
    'Visualizing...'
    means = all_means[-1]
```

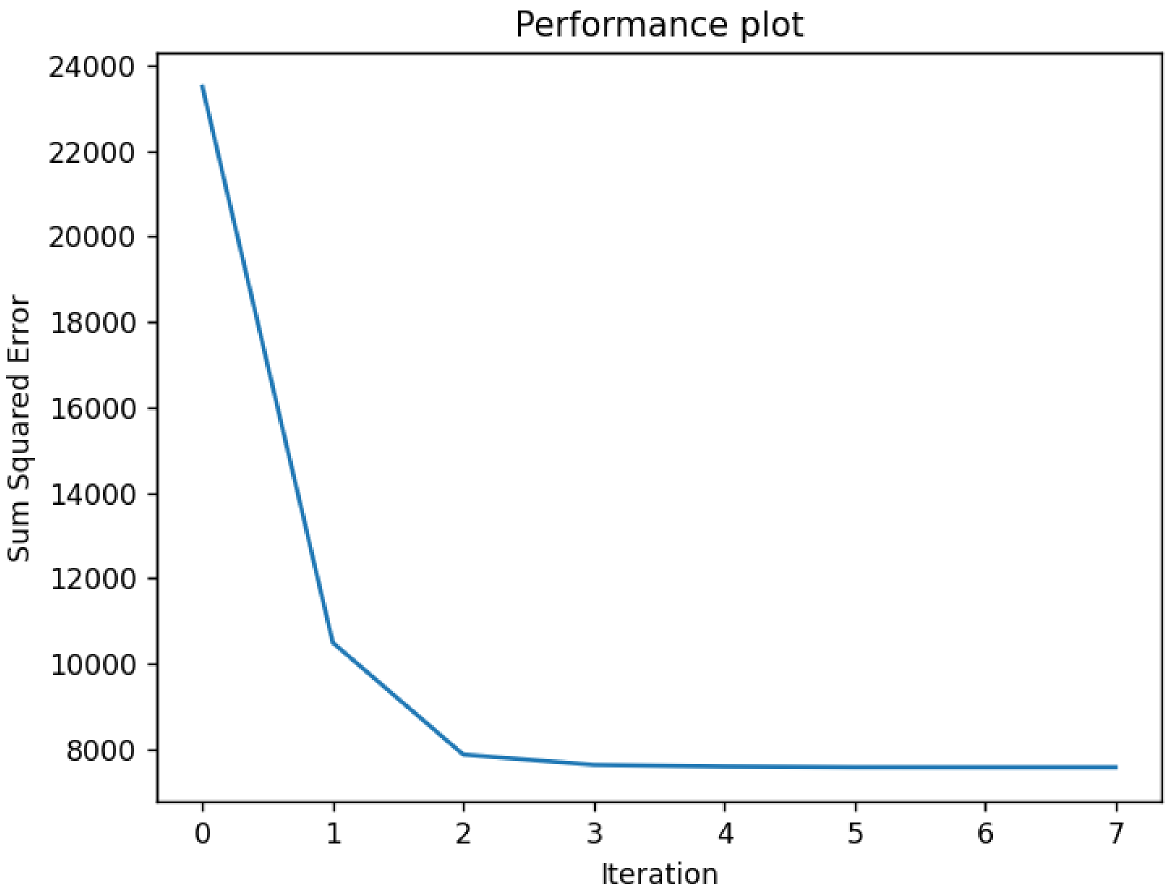
```
    k = args['k']
    distance = args['dist']
    clusters = [[] for _ in range(k)]
    for point in data:
        dlist = [distance(point, center) for center in means]
        clusters[argmin(dlist)].append(point)

    # plot each point of each cluster
    colors = cycle('rgbwkcmy')

    for c, points in zip(colors, clusters):
        x = [p[0] for p in points]
        y = [p[1] for p in points]
        plt.scatter(x, y, c=c)

    # plot each cluster centroid
    colors = cycle('krrkgkgr')
    colors = cycle('rgbkkcmy')

    for c, clusterindex in zip(colors, range(k)):
        x = [iteration[clusterindex][0] for iteration in all_means]
        y = [iteration[clusterindex][1] for iteration in all_means]
        plt.plot(x, y, '-x', c=c, linewidth='1', mew=15, ms=2)
    plt.axis('equal')
    plt.show()


def visualize_performance(data, all_means, distance):
    errors = [performance_SSE(data, means, distance) for means in
all_means]
    plt.plot(range(len(all_means)), errors)
    plt.title('Performance plot')
    plt.xlabel('Iteration')
    plt.ylabel('Sum Squared Error')
    plt.show()


if __name__ == '__main__':

    args = parse()
```

```
    # Read data
    data = readfile(args['input'])
    print('Number of points in input data: {}\n'.format(len(data)))
    verbose = args['verbose']

    totalSSE = 0
    totaliter = 0

    for experiment in range(args['numexperiments']):
        print('Experiment: {}'.format(experiment + 1))
        random.seed(args['seed'] + experiment)
        print('Seed: {}'.format(args['seed'] + experiment))

        # Initialize means
        means = []
        if args['init'] == initialization_forgy:
            means = args['init'](data, args['k'])  # Forgy doesn't need
distance metric
        else:
            means = args['init'](data, args['dist'], args['k'])

        if verbose:
            print('Means initialized to:')
            print(means)
            print('')

        # Run k-means clustering
        all_means = iteration_many(data, means, args['dist'],
args['maxiter'], args['epsilon'])

        SSE = performance_SSE(data, all_means[-1], args['dist'])
        totalSSE += SSE
        totaliter += len(all_means) - 1
        print('Sum Squared Error: {}'.format(SSE))
        print('Number of iterations till termination:
{}'.format(len(all_means) - 1))

        print('Convergence achieved: {}'.format(hasconverged(all_means[-
1], all_means[-2])))
```

```
        if verbose:
            print('\nFinal means:')
            print(all_means[-1])
            print('')

    print('\n\nAverage SSE: {}'.format(float(totalSSE) /
args['numexperiments']))
    print('Average number of iterations: {}'.format(float(totaliter) /
args['numexperiments']))

    if args['numexperiments'] == 1:
        # save the result
        writefile(args['output'], all_means[-1])

        # If the data is 2-d and small, visualize it.
        if len(data) < 5000 and len(data[0]) == 2:
            visualize_data(data, all_means, args)

        visualize_performance(data, all_means, args['dist'])
```

Results:
Task1:





Task2:

1. Run your code on datasets/garden.csv, with different values of k. Looking at the performance plots, does the SSE of k-means algorithm ever increase as the iterations are made? (1 mark)

Answer: In the k-means method, the SSE never rises. It can even be demonstrated that it will never rise. The first time the drop occurs is when we rename the points using the centroid that is closest to them. This indicates that the distance to the centroid closest to

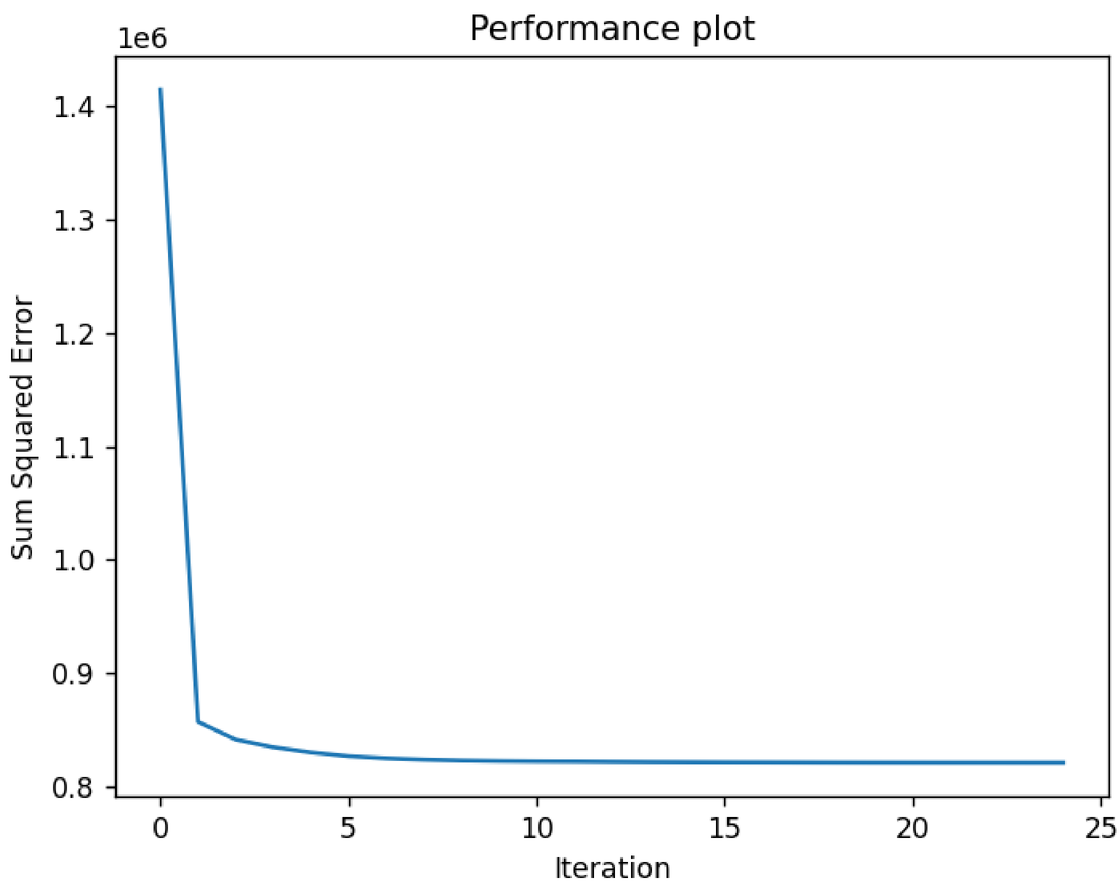
Every point that has been relabelled has dropped, hence the SSE must drop overall. As the centroid of the present clusters, we now move on to step two, where we obtain new means. The SSE will again drop since we know that the sum of squared distance is the least from the mean (centroid). And the cycle continues. All of this was mathematically demonstrated in class as well.

2. Look at the files 3lines.png and mouse.png. Manually draw cluster boundaries around the 3 clusters visible in each file (no need to submit the hand drawn clusters). Test the k-means algorithm on the datasets datasets/3lines.csv and datasets/mouse.csv. How does the algorithm's clustering compare with the clustering you would do by hand? Why do you think this happens? (1 mark)

Answer: I naturally group the three lines in the dataset together to form three oblong clusters. On the other side, the algorithm provided a very different response. If the cluster centroids are in the middle of each of the three lines, separation should be easy since the perpendicular bisectors would identify the dividing zone.

But given that the SSE of such a scenario is higher, this is not where we are convergent in this case. Additionally, as we can see, the centroid is in close proximity in this instance, which is something we strive to avoid when using kmeans++, leading to a higher SSE. The SSE is greater because the distances between the lines' endpoints and middles are greater. Since we're measuring euclidean distances, it's best if the points are arranged in a circle around the centroids, putting the majority of the points in each cluster as close as they can be to the centroid.
I instinctively grouped the face and each ear in one cluster for the mouse dataset. Even though all three of these clusters are circular, the algorithm once more cannot match it. This time, we can see that the face extends into the ear clusters in certain places. This occurs as a result of the mouse's face's enormous circumference and the presence of the ears at the face's edge. The centroid for the face cluster would be near to the face's centre given the face's geometry. The ears are the same way. Due to the enormous radius, the spots on the face near the ears are now closer to the centroid of the ear than the centroid of the face.

Task4



1. For each dataset, with kmeansplusplus initialization algorithm, report "average SSE" and "average iterations". (1 mark)
Answer:

Dataset	Initialization	Average SSE	Average Iterations
100.csv	forgy	8472	2.64
100.csv	kmeans++	8472	2.01
1000.csv	forgy	21337462	3.4
1000.csv	kmeans++	19400000	3.25
10000.csv	forgy	169946236	22.1
10000.csv	kmeans++	20783467	6.01

In every instance, the average SSE and the number of iterations are both less than forgy. This is as a result of our decision to use improved initializations for kmeans++. In k-means, the initial centroids are chosen to be further apart from one another, making it more likely that every point will find at least one close-by centroid; in other words, for the majority of the points, the minimum distance from the initial centroids will be smaller than in the case of forgy. Therefore, the SSE is low right away. Faster convergence results from this. Additionally, better initializations increase the likelihood

Conclusion:

- K means may also be used to classify data from the MNIST dataset. However, we discovered that it was not very accurate and frequently misclassified photos.
- When there are clusters with different densities and sizes, k-means have problems clustering the data. You must generalize k-means to cluster such data.
- Before applying k means, data normalization is a crucial preprocessing step that makes sure
- Outliers may drag centroids, or they may form their own cluster in place of being ignored. K means are therefore not resistant to outliers.