Introduction to Data Science

Data Exploration & Clustering

Assignment 4

Submitted by

Andras Csepreghy

xgj708

2nd of April, 2019

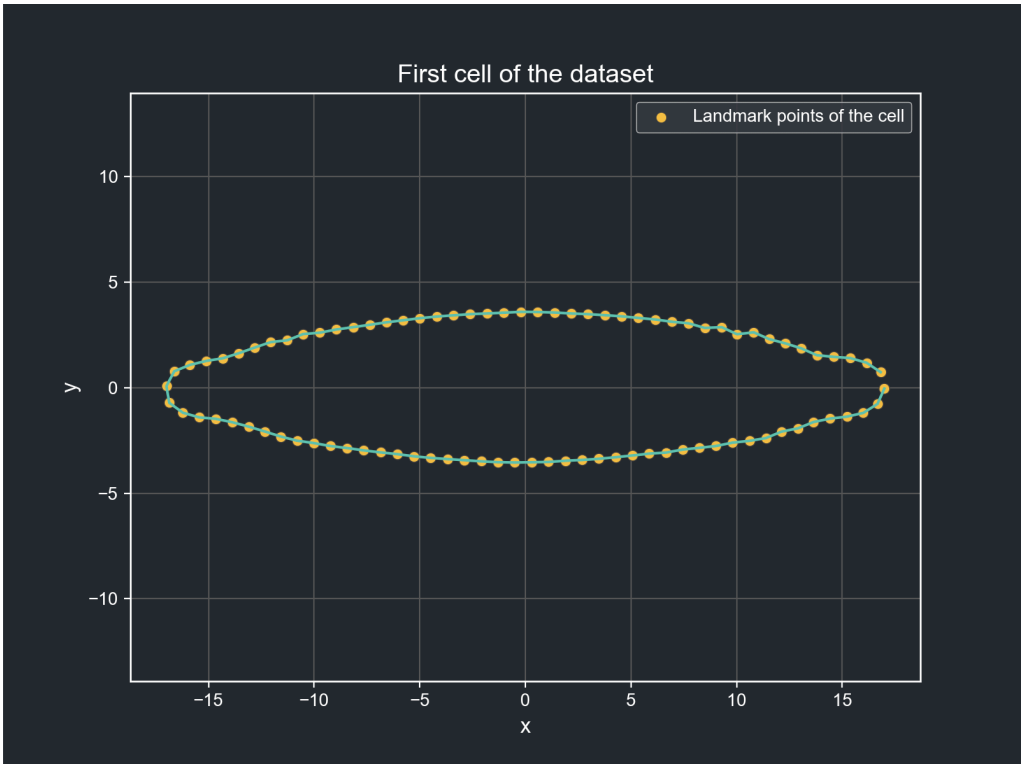
UNIVERSITY OF COPENHAGEN

Introduction

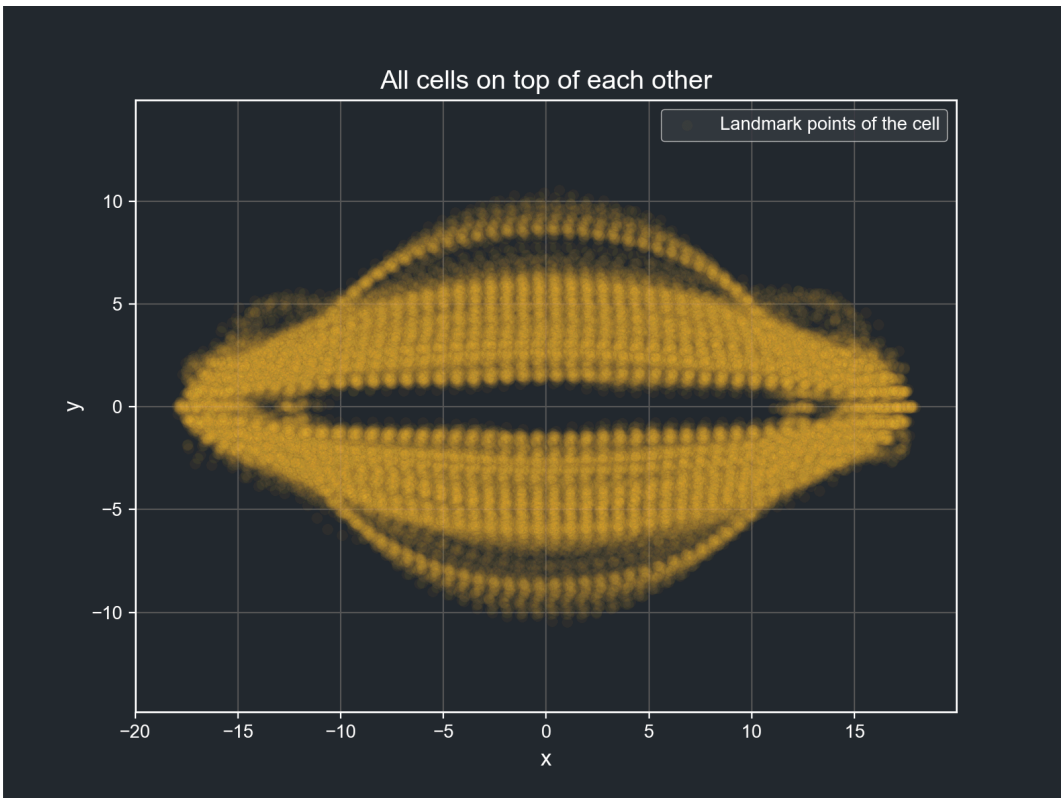
In this assignment we were asked to perform data exploration with PCA as well as some K-Means clustering.

Exercise 1

In exercise 1 I plotted the landmark points of the first cell in the dataset. The landmark points are orange and the line that connects them is blue for contrast. I used Plotify, a light matplotlib skin I created a few assignments back to give plots some unique look. To plot the points correctly I followed the structure given by the assignment, x and y coordinates are next to each other in a one dimensional array.

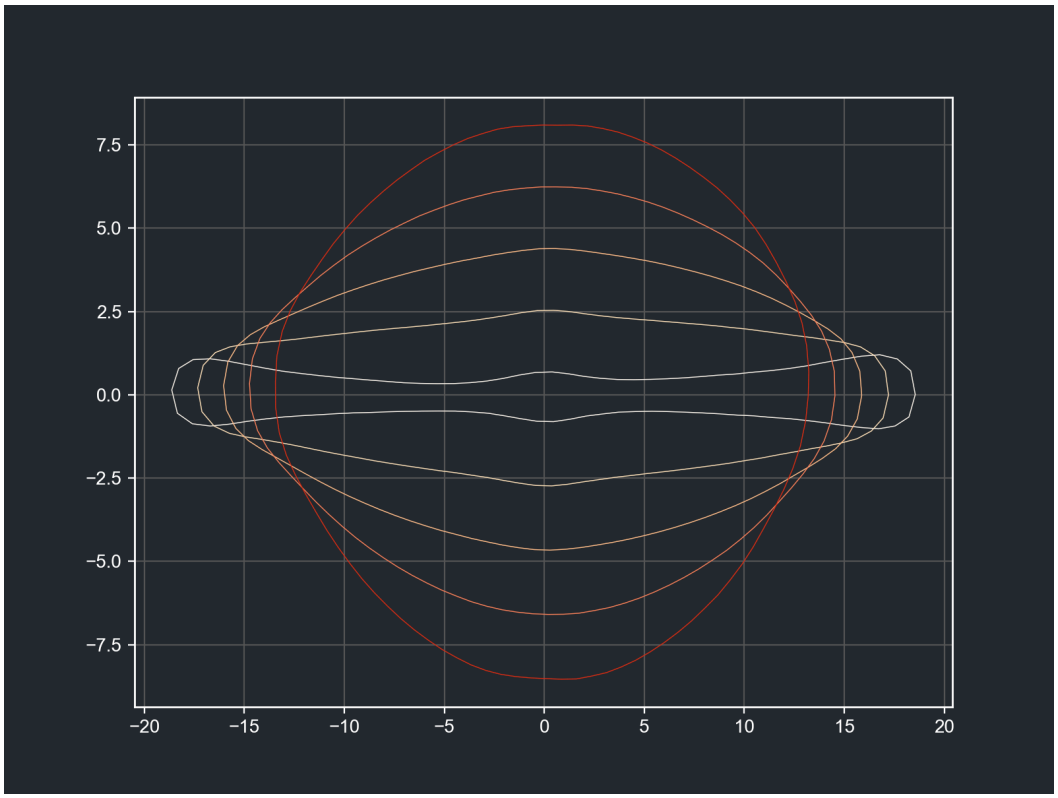


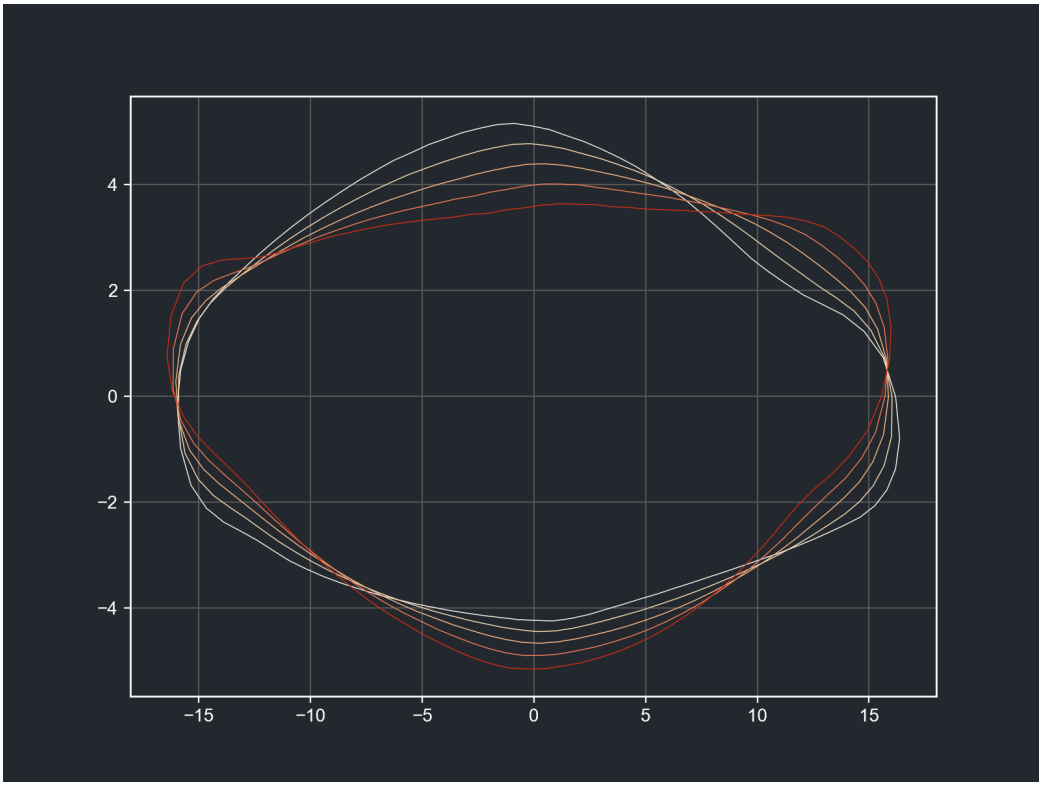
Next the plot that displays all the cells. Here I reduced opacity (alpha) to make the emerging shapes more apparent. Here we can notice some generic cell shapes and that the cells have rather similar outlines without many outliers. The shapes also reveal some of the difference between cell types.

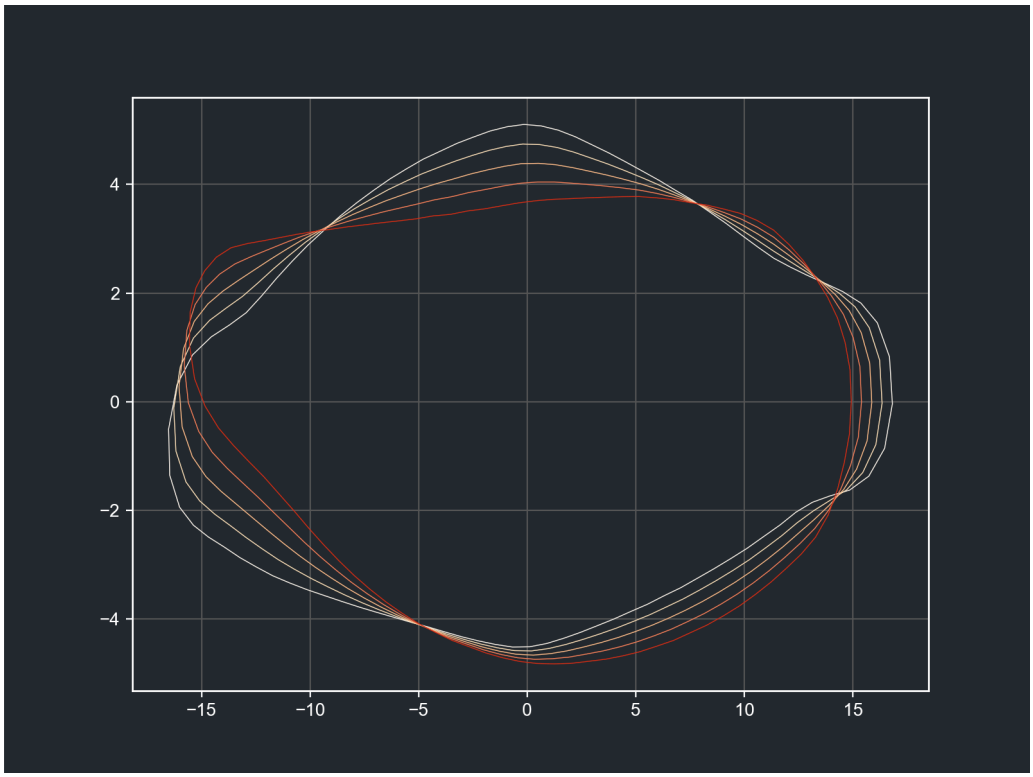


Exercise 2

Here I performed PCA on the diatoms dataset, which captured the variance among cell shapes. The first 3 components can be seen as 3 cell types. By multiplying the mean of the PC with the eigenvectors and eigenvalues we can show the temporal development. Using different coloring it becomes more apparent. The images are in this order show PC1, PC2 and PC3.




4
2
0
−2
−4
−15
−10
−5
0
5
10
15


4
2
0
−2
−4
−15
−10
−5
0
5
10
15

Exercise 3

a)

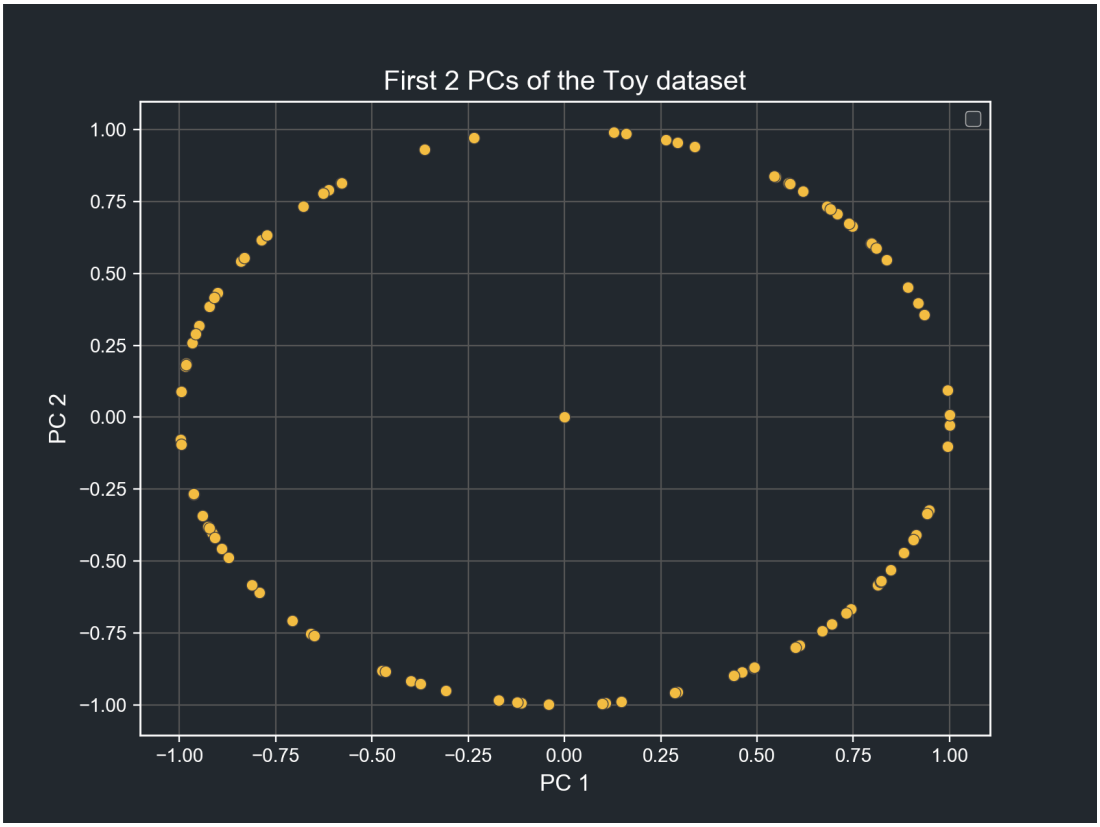
1: Centering: Centering is not necessarily required because the covariance matrix will be the same regardless of the data being centered. However it is generally a good idea to center our data before doing the PCA because it will be easier to work with it later on, for example drawing eigenvectors or doing dimensionality reduction

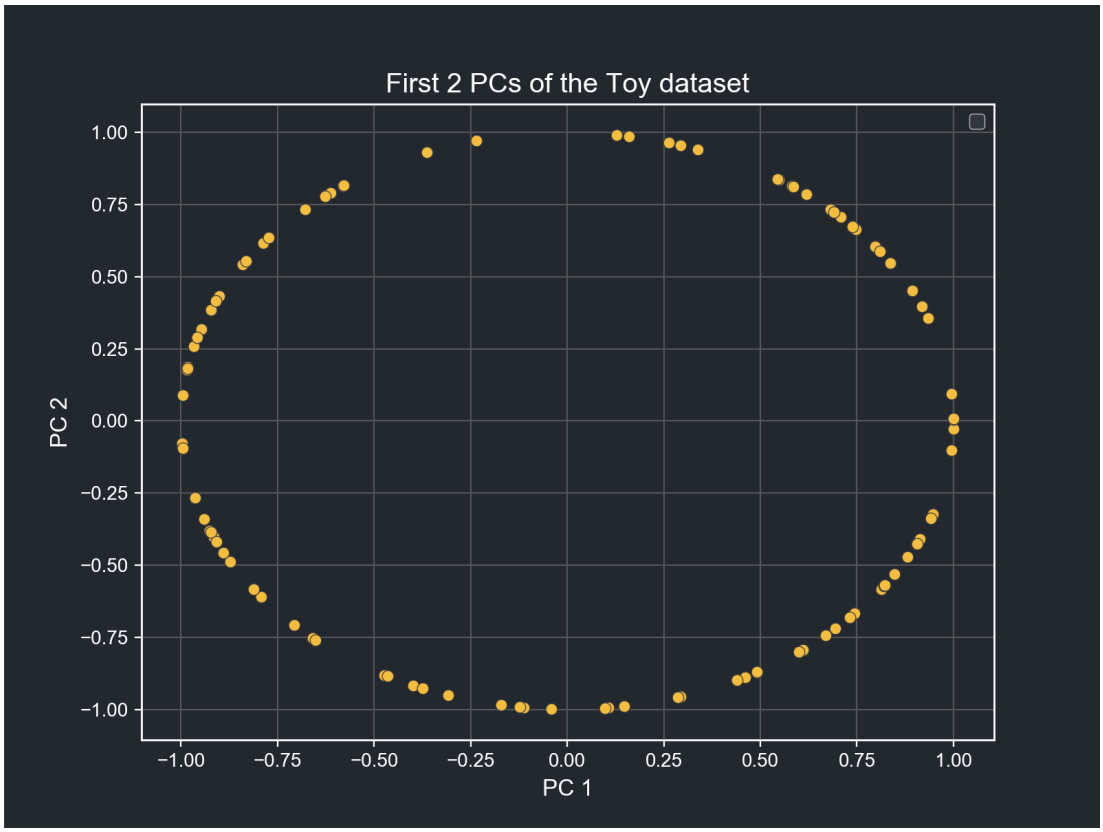
2: **Standardization**: This is entirely dependent on the data. There are some cases where it isn't needed, however as a rule of thumb standardizing data is important. In the previous assignment I did PCA both with and without standardization and argued that from the assignment description it wasn't clear if the data needs scaling. The result of the PCA will be different depending whether one has standardized data because it removes some of the variance from it. A good example when one must standardize data is when there are data points with different unit scales. For example age and salary where age ranges about 0-110 where salary is 30,000 - 400,000 USD. (Ranges not entirely accurate, only examples)

3: **Whitening**: Whitening data removes underlying correlation in the data completely, defeating the purpose of PCA. It decorrelates the the components and makes the variance 1. However it could be applied after PCA, when the eigenvectors and eigenvalues already captured the variance.

b)

I ran PCA on the toy dataset which captured the only two components that had numbers in them, while leaving out the two column filled with zeroes. Leaving out the last two datapoints removed the dots from the middle, which didn't seem a natural continuation of the data in the first place. It also removed variation from the data (what removing the outliers/noise does). First plot without removing any data, second without the last 2 points.





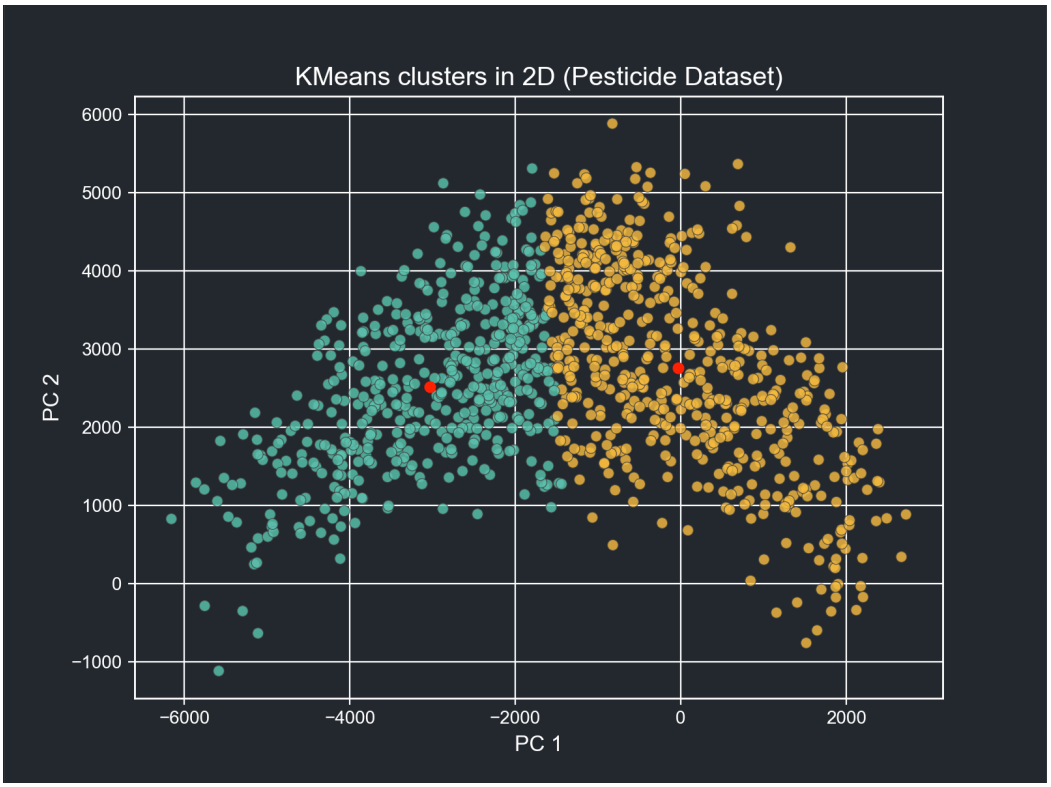
Exercise 4

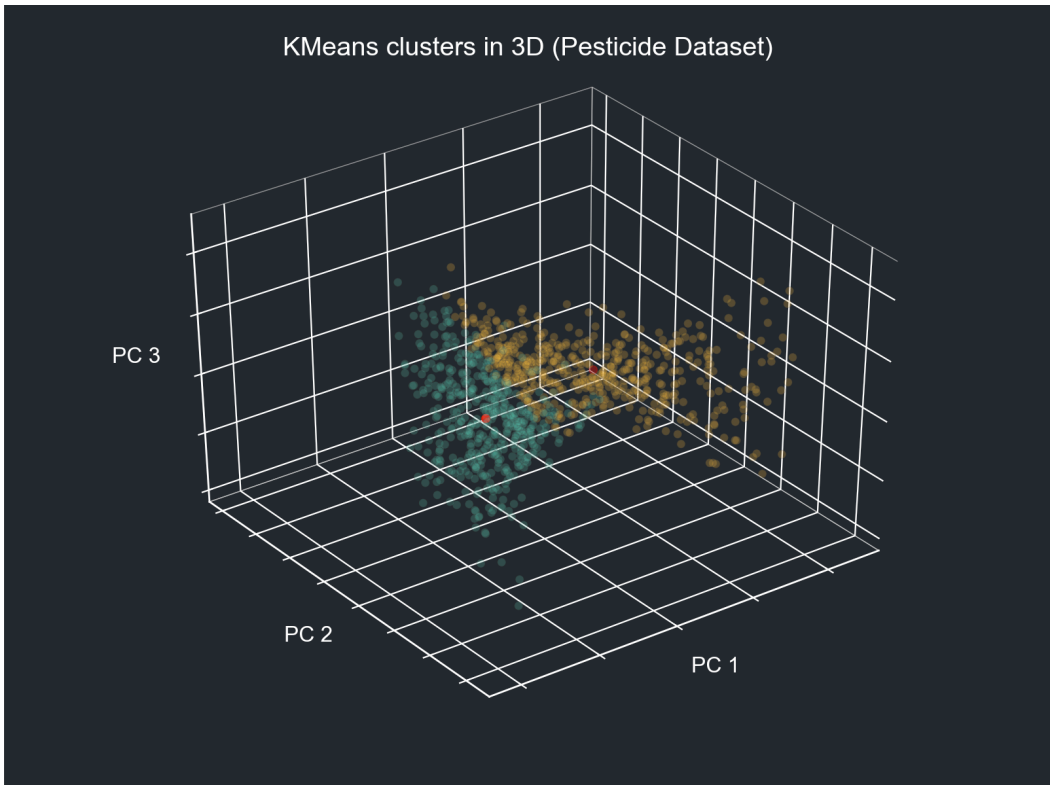
K-Means is a simple unsupervised learning algorithm that creates k number of clusters. K-Means is agnostic about the number of dimensions our data has, however in this exercise for the purpose of removing noise and to make it possible to visualize the data in 2D or 3D we performed PCA on the dataset first and projected the data down to the first principal components.

K-Means works by initializing cluster centers (in our case it is the first data points, which would be considered bad practice, but easier for grading) and then iteratively allocating datapoints to the cluster centers they have the least euclidean (in case of linear problems) distance to, and then moving the cluster centers to the mean of their clusters. This iterative process happens until the cluster centers no longer move and all data points are assigned.

After PCA and dimensionality reduction I ran the K-Means on the data and colored the two clusters in the plots. First in 2D, then 3D. In the 3D image it is a little harder to see the red dots (centroids).

The clusters seem somewhat meaningful, however when visualizing the data in 2 and 3 dimensions it looks more continuous than clustered. This can be seen by observing how close the data points are to each other that are not in the same cluster. However predicting performance and accuracy is very good when using K-Means as well as K-Nearest Neighbors.


KMeans clusters in 2D (Pesticide Dataset)
6000
5000
4000
3000
2000
1000
0
−1000
PC 2
−6000
−4000
−2000
0
2000
PC 1


KMeans clusters in 3D (Pesticide Dataset)
PC 3
PC 2
PC 1