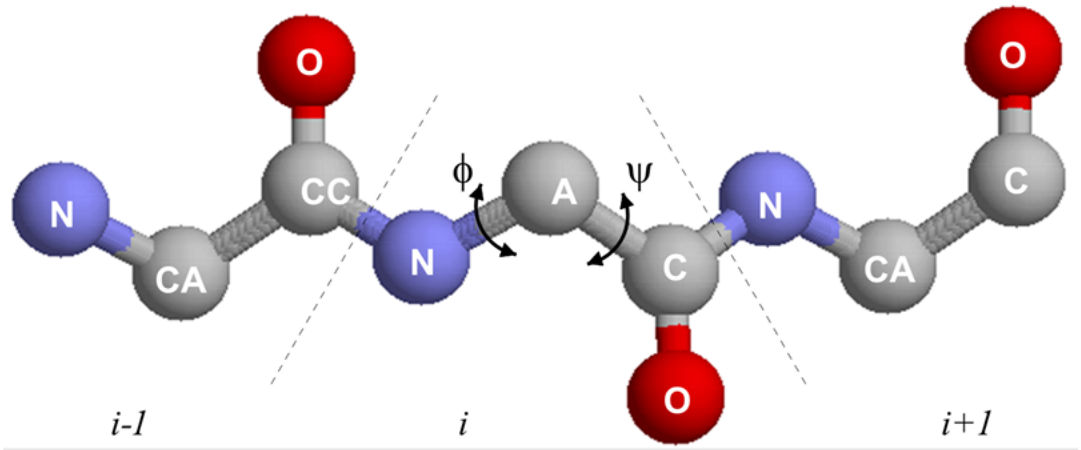
Introduction to Data Science and AI, Sofia, 2023

Assignment 3: Clustering

Pavel Petkov

Introduction

The assignment is about implementing Python code that clusters data using K-means and DBSCAN methods. The data is about phi and psi angles of protein channels.

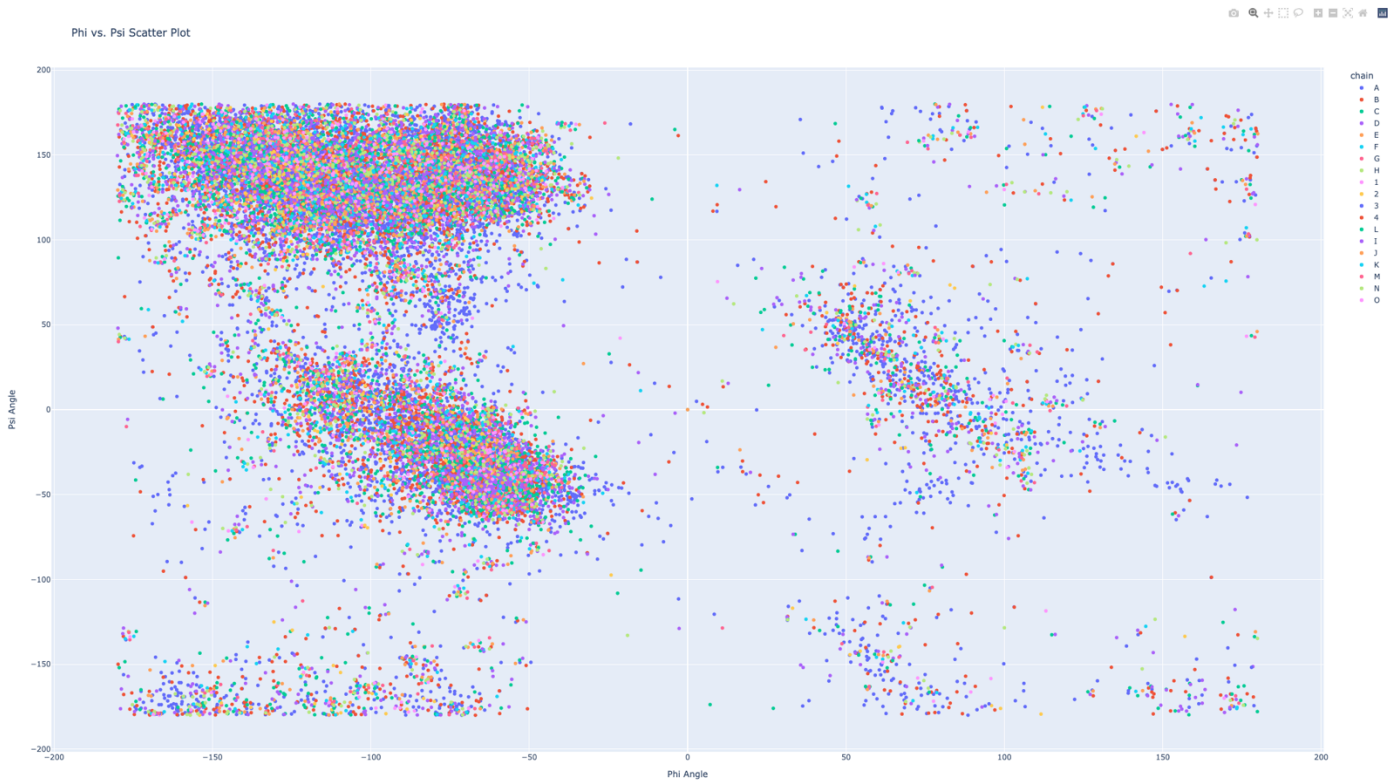


Example of the CSV with the data:

```
residue name,position,chain,phi,psi
LYS,10,A,-149.312855,142.657714
PRO,11,A,-44.283210,136.002076
LYS,12,A,-119.972621,-168.705263
```

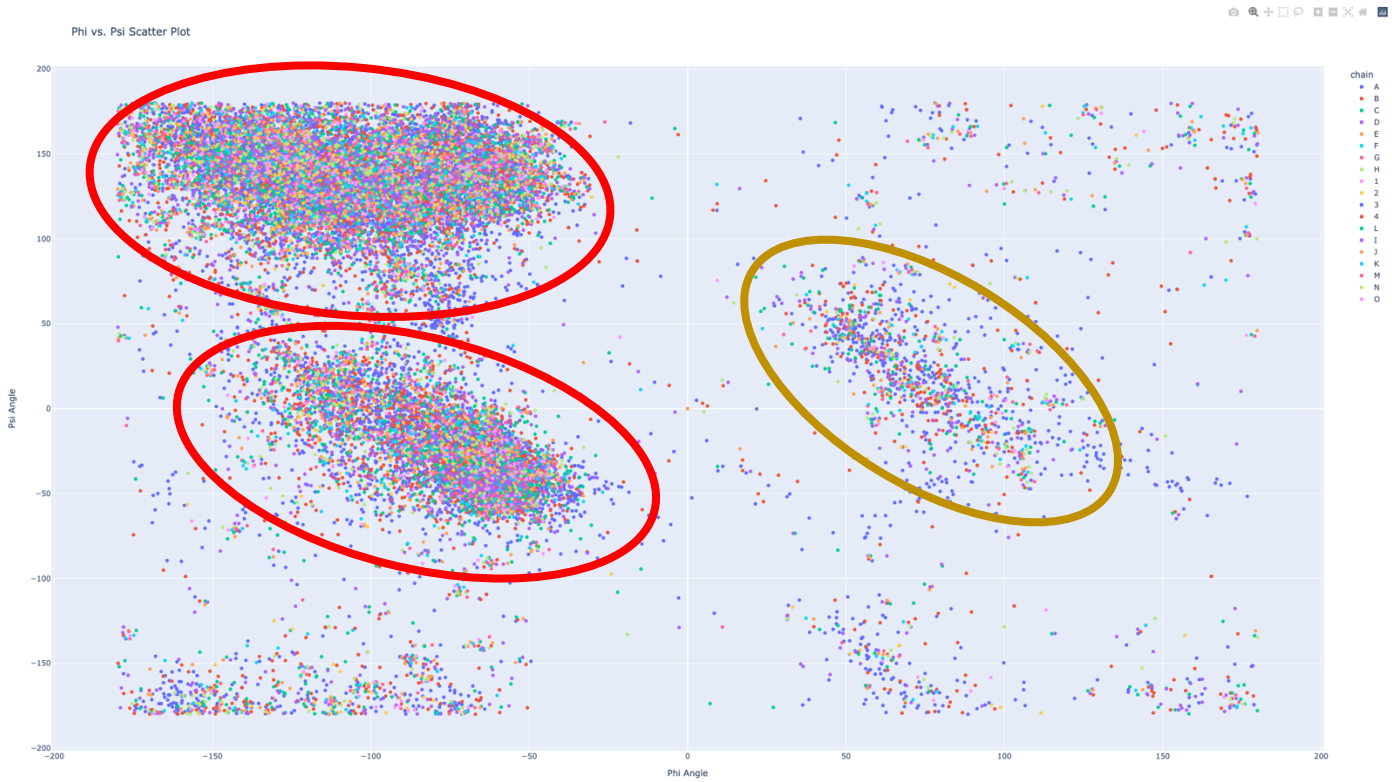
Plotting the data

First, we draw a scatter plot and heatmap to be able to visually see if any clusters are formed.

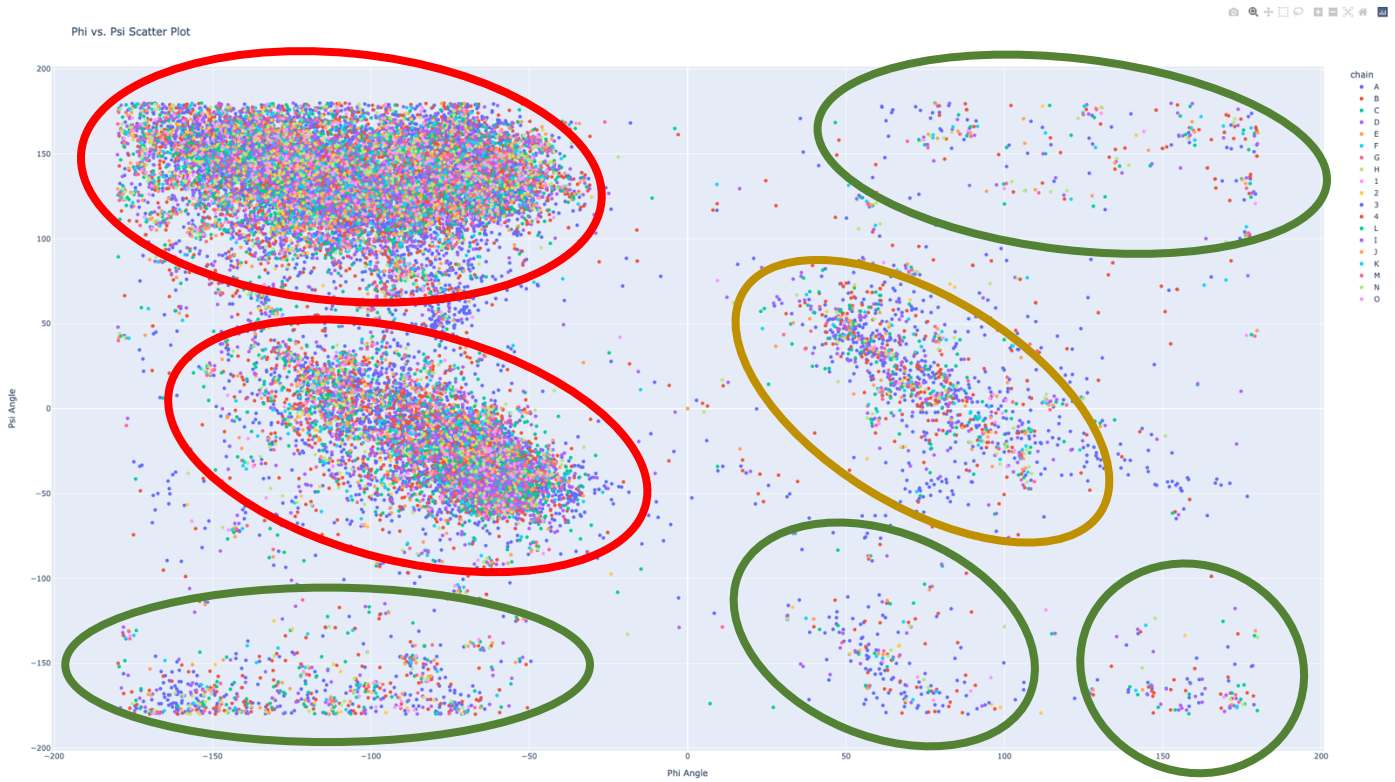


Eyeballing the scatter plot we can see that there are 2 clusters that are very dense and easily distinguishable (in the red circles.

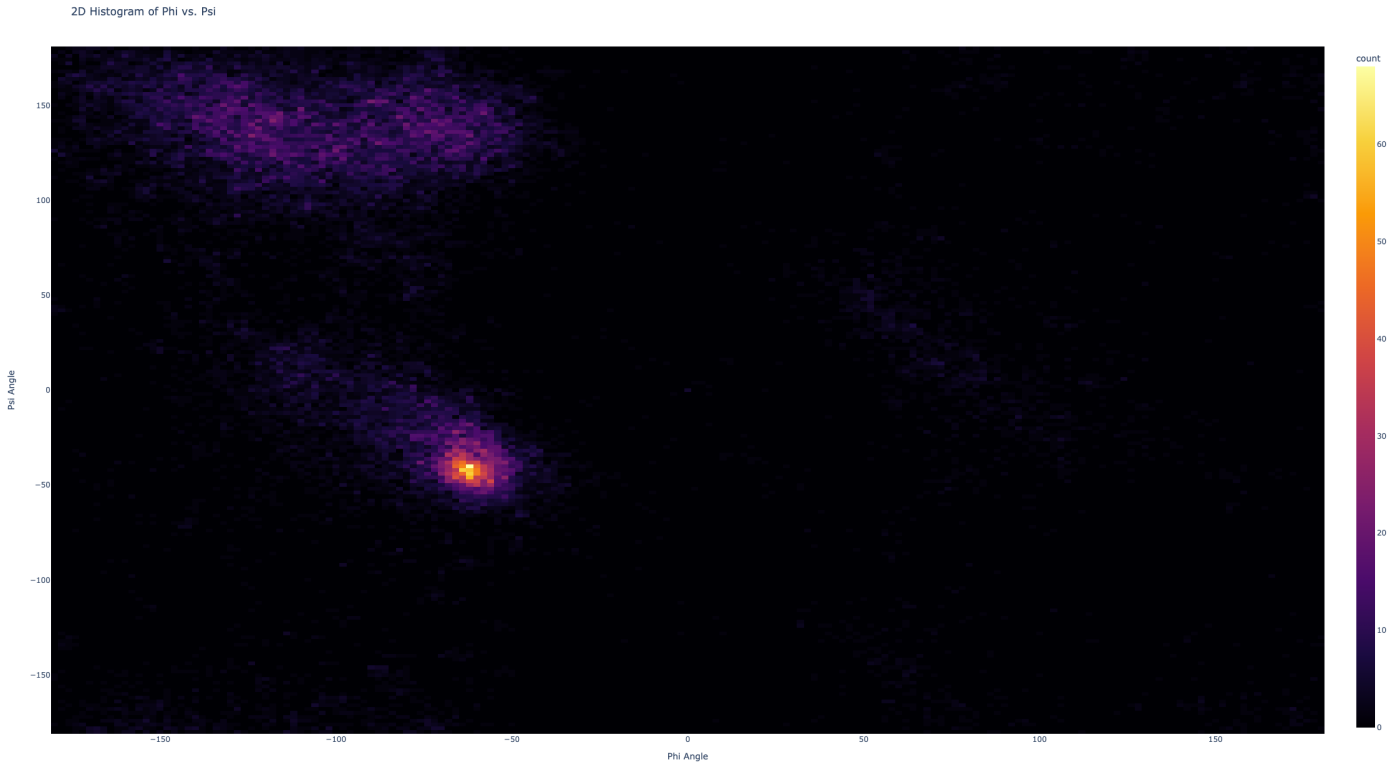
There is a third cluster that is not as dense as the first two, but still can be noticed (yellow circle).



We may also argue that the 4 more clusters are visible (green circles). However, they are not as dense as the first 3 and can as well be considered outliers. We will elaborate on this when we compare the K-means and the DBSCAN methods.



The above assumptions are also visible from the heatmap:

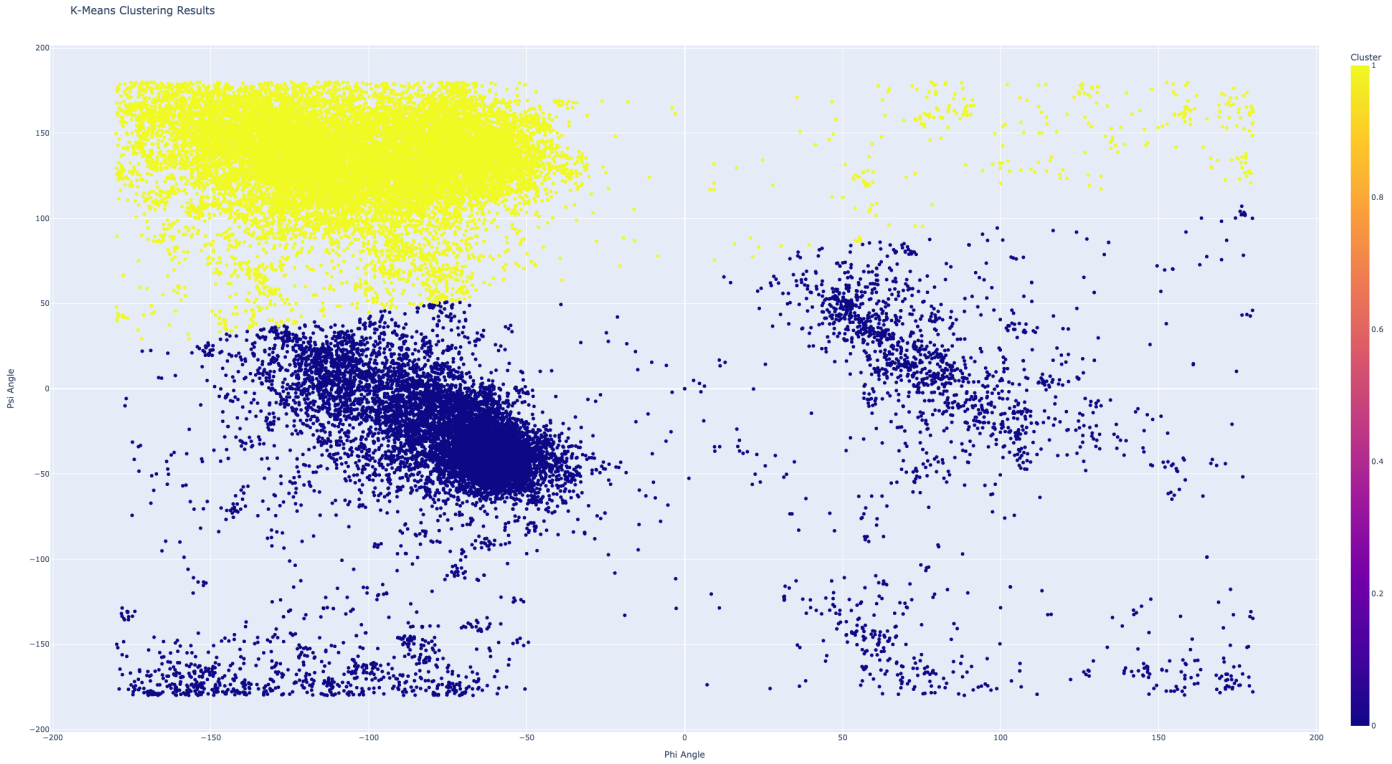


It clearly shows the 2 main clusters and the 3rd is barely visible. The other 4 can not be distinguished, so according to the heatmap, the points in them should be considered outliers.

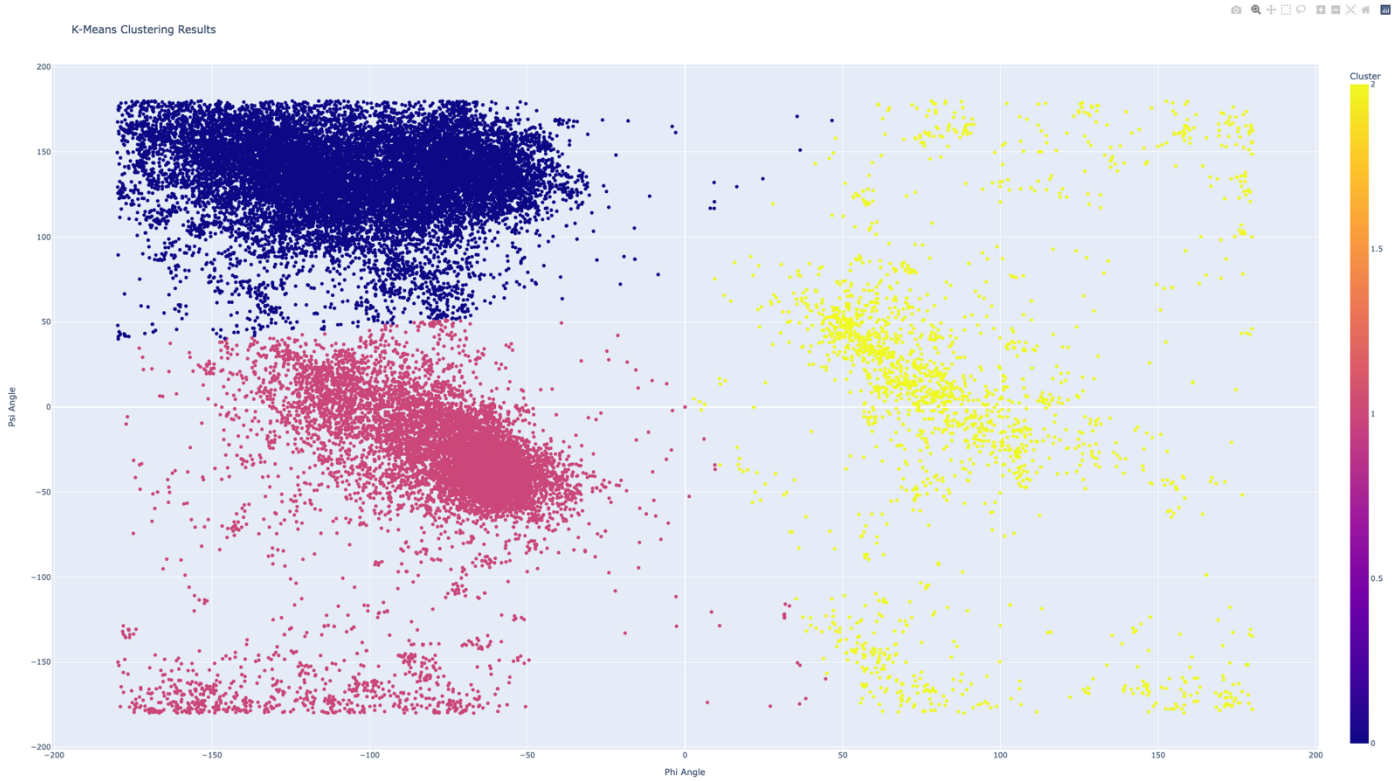
K-means

We cluster the data using the K-means method with K values 2, 3, 4, and 7 and random centroid initialization.

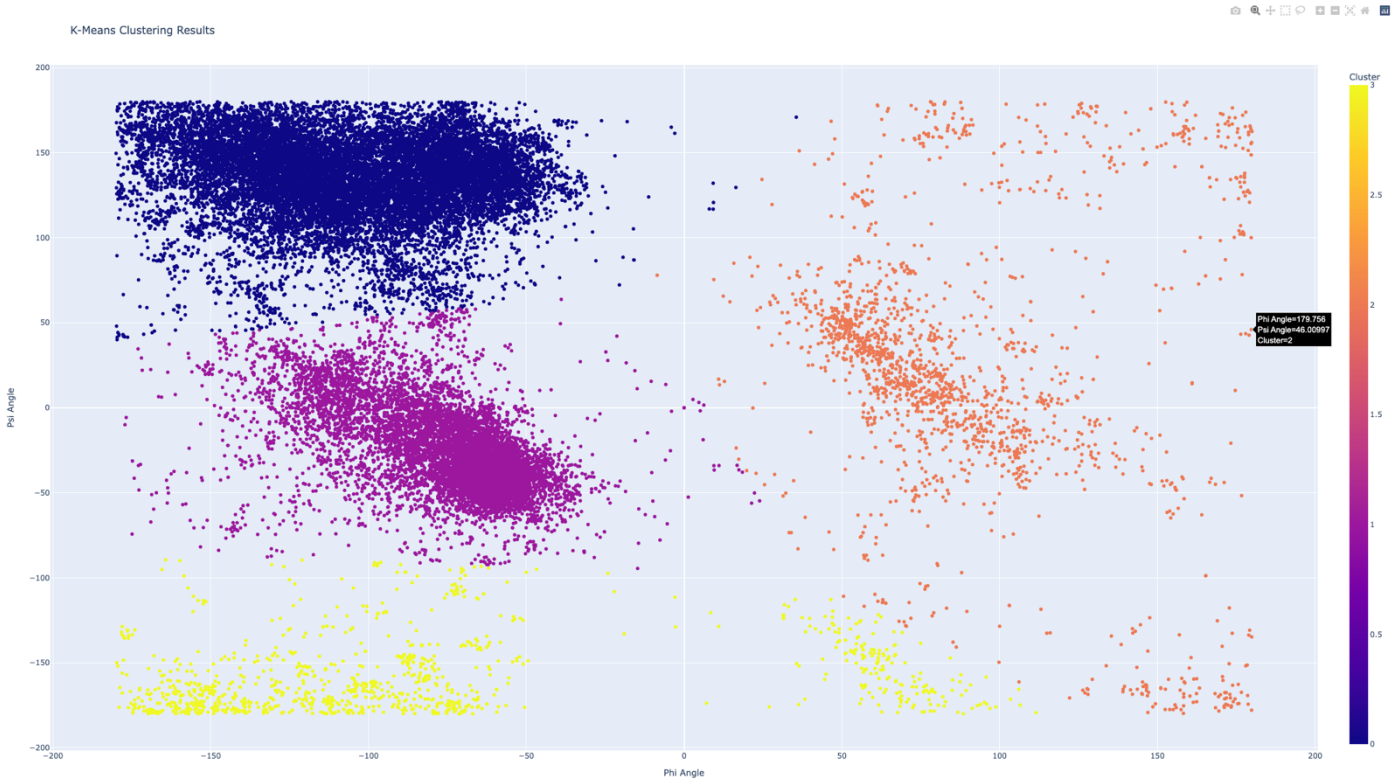
K = 2



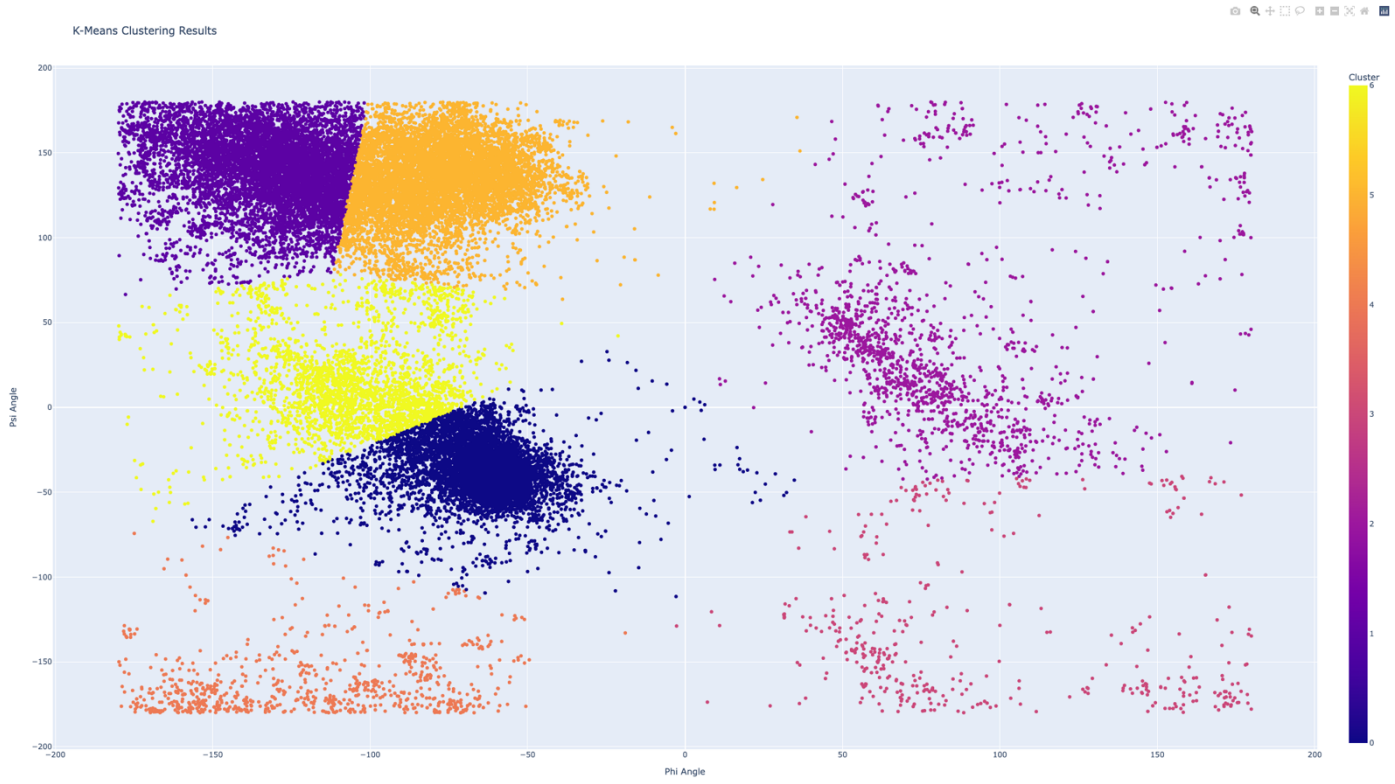
K = 3



K = 4



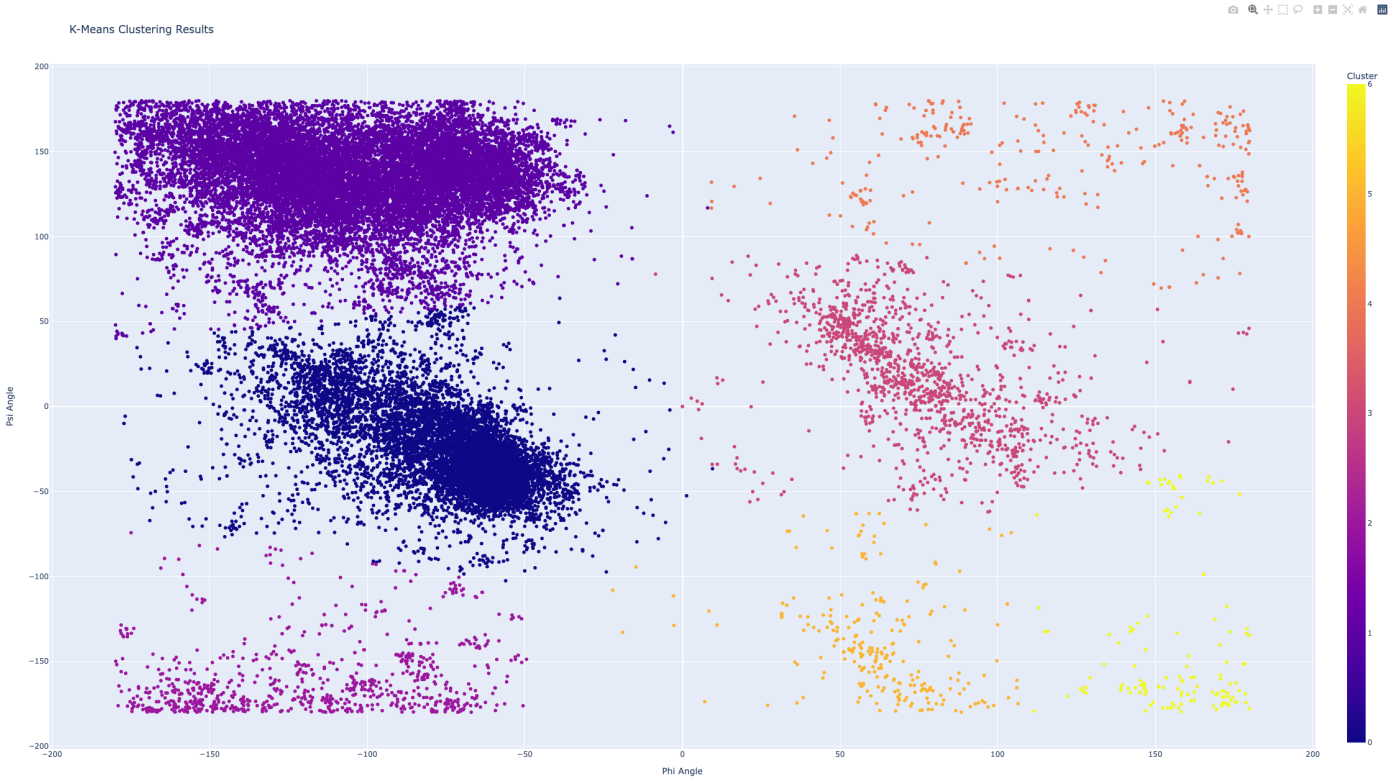
K = 7



We can see that with K values 2, 3, and 4 the results look reasonable and the clusters are formed as we expected. But with K = 7, the results are not as expected.

With a little help from a person, placing the 7 centroids around the centers of the 7 clusters that we distinguish and then using the K-means method, the 7 clusters are formed. We have to keep in mind that 4 of the clusters have significantly less data than the others and it might not be a good idea to use them to make informed decisions.

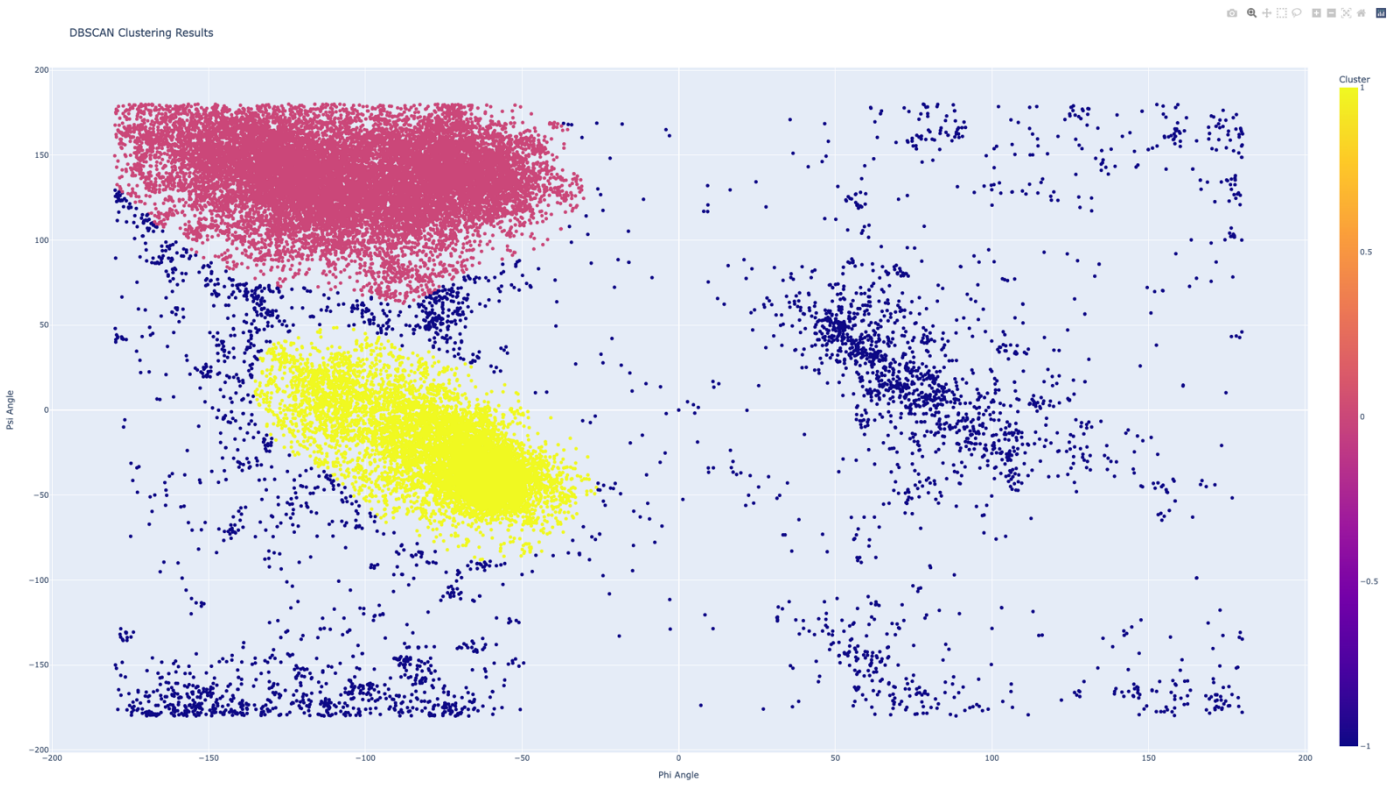
K = 7, with manual centroid initialization



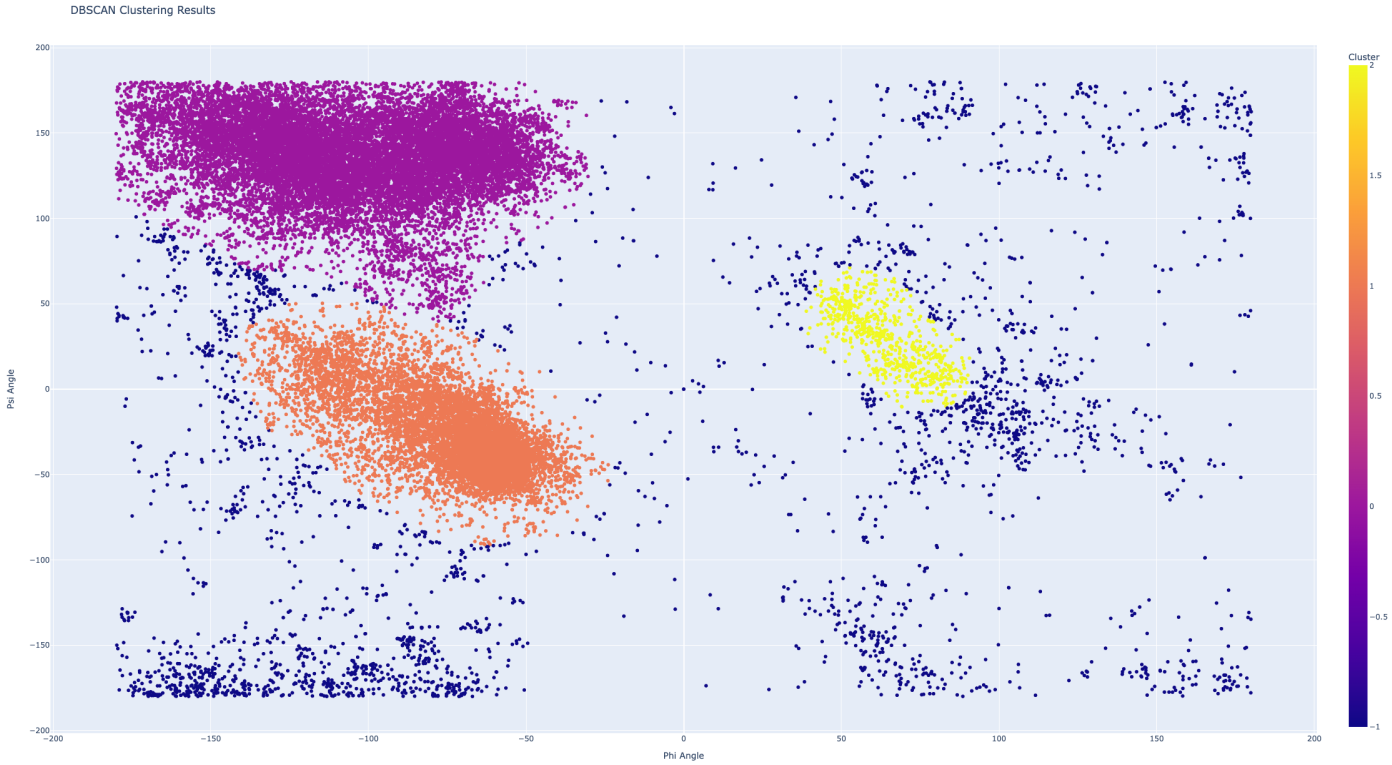
DBSCAN

Using the DBSCAN methods we also come up with 2 graphics that look as expected.

epsilon = 0.2, minimum number of samples = 300



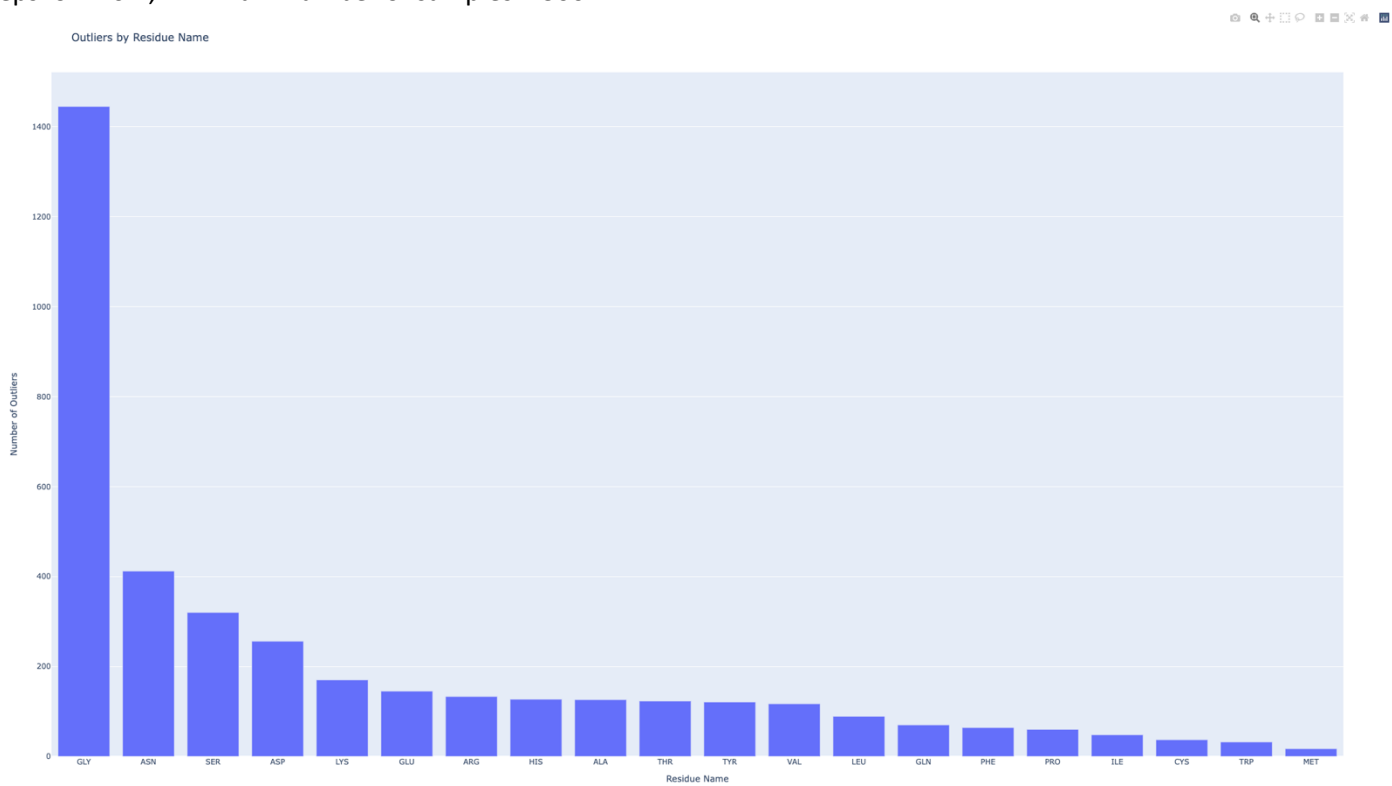
epsilon = 0.2, minimum number of samples = 200



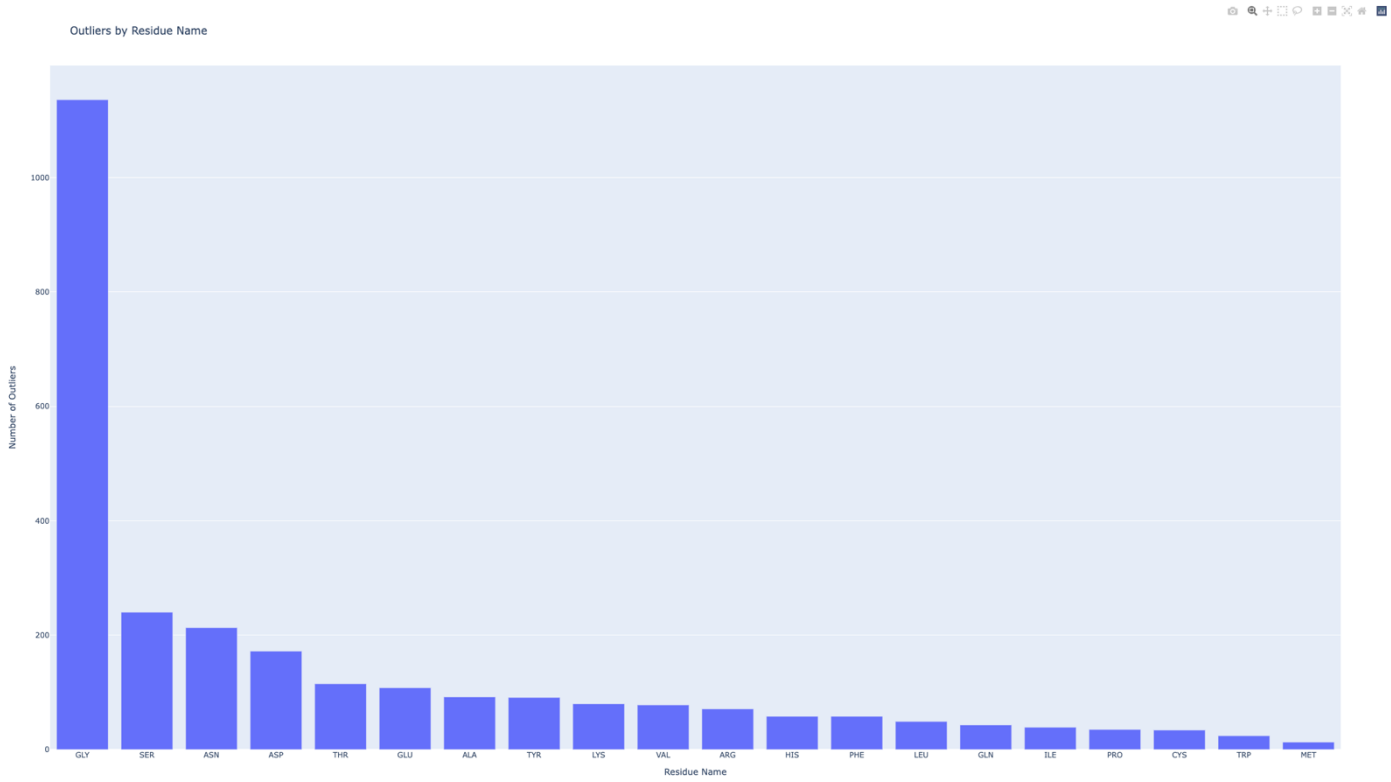
Using the DBSCAN method we are able to find 2 or 3 clusters and the rest is considered outliers. This confirms that even if we are able to find 7 clusters with the K-means method at least 4 of them would be irrelevant. I also tried different values for the epsilon and the minimum number of samples but each time the results were not reasonable.

The program also draws bar plot with the outliers separated by residue type.

epsilon = 0.2, minimum number of samples = 300

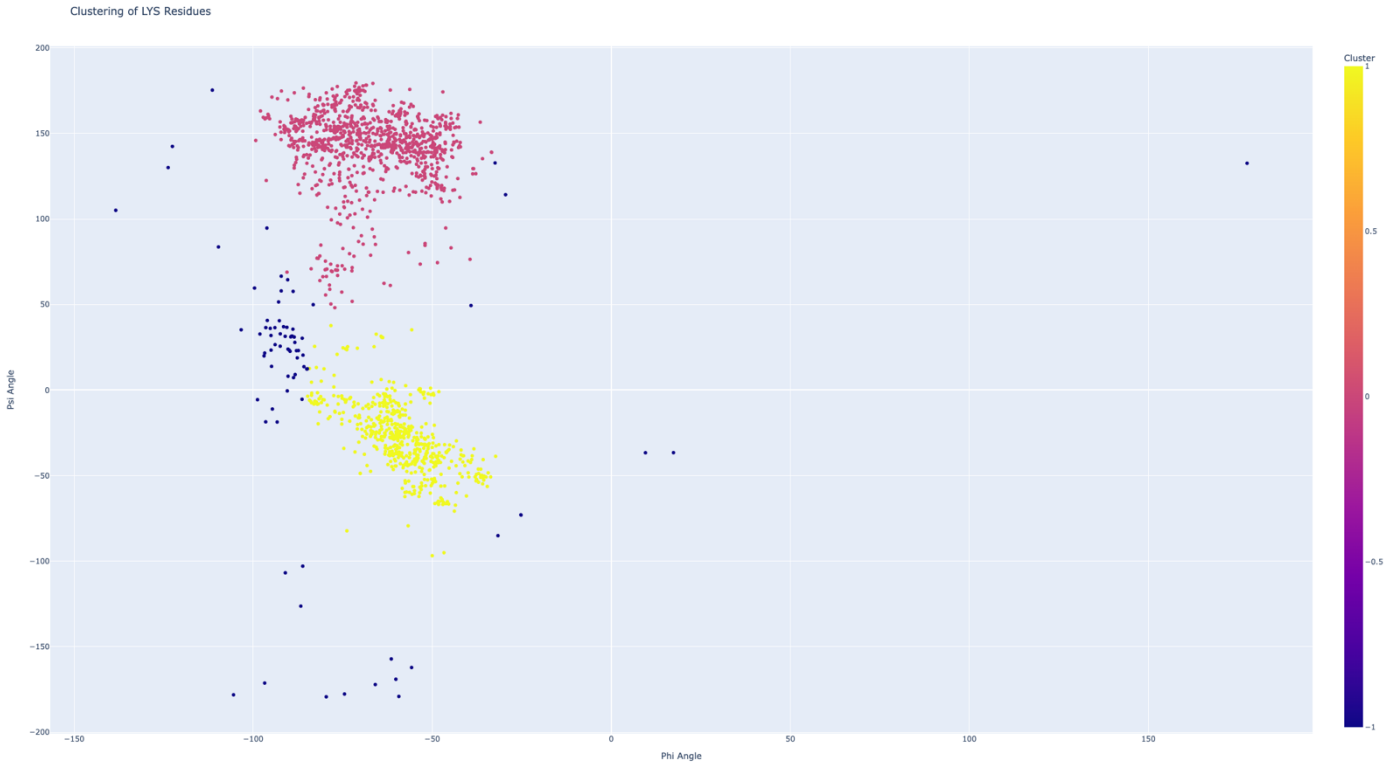


epsilon = 0.2, minimum number of samples = 200



In both cases, we can see that there are significantly more outliers of residue type GLY than of any other type.

Clustering the residues of LYS type, using the DBSCAN methods reveals the following graphic:



We can see that the 2 clusters that form strongly correlate with the clusters that are formed when the whole data.

Bonus

Mapping 2D Ramachandran plot onto a torus