Hypothesis:

a. Plot the dataset as a scatter plot with the labels.

b. Which clustering algorithm do you think will work the best for this dataset? Why?

Tha Kmeans and the Gaussian distributions will likely be tied for the most accurate clustering algorithms for the dataset because both algorithms are most effective for clusters that take on an evenly circular shape with varying densities and have distinguishable spacing between them.



Analysis:

K-Means, Agglomerative, DBSCAN and Gaussian Mixture model clustering algorithms

e. Answer these questions: (1 pt)

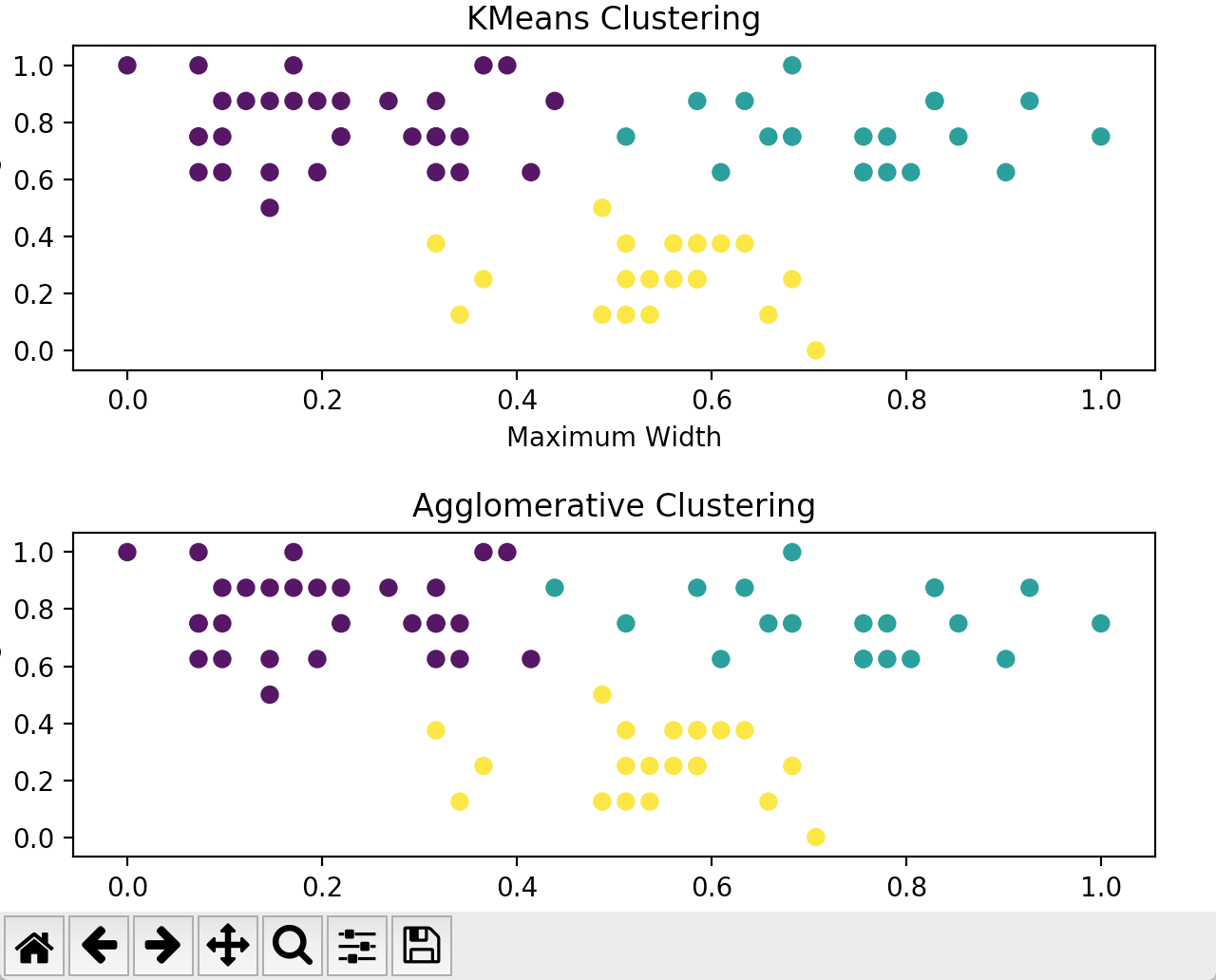
i. What parameters did you choose for each of the four algorithms? Why?

ii. How accurate was each algorithm?

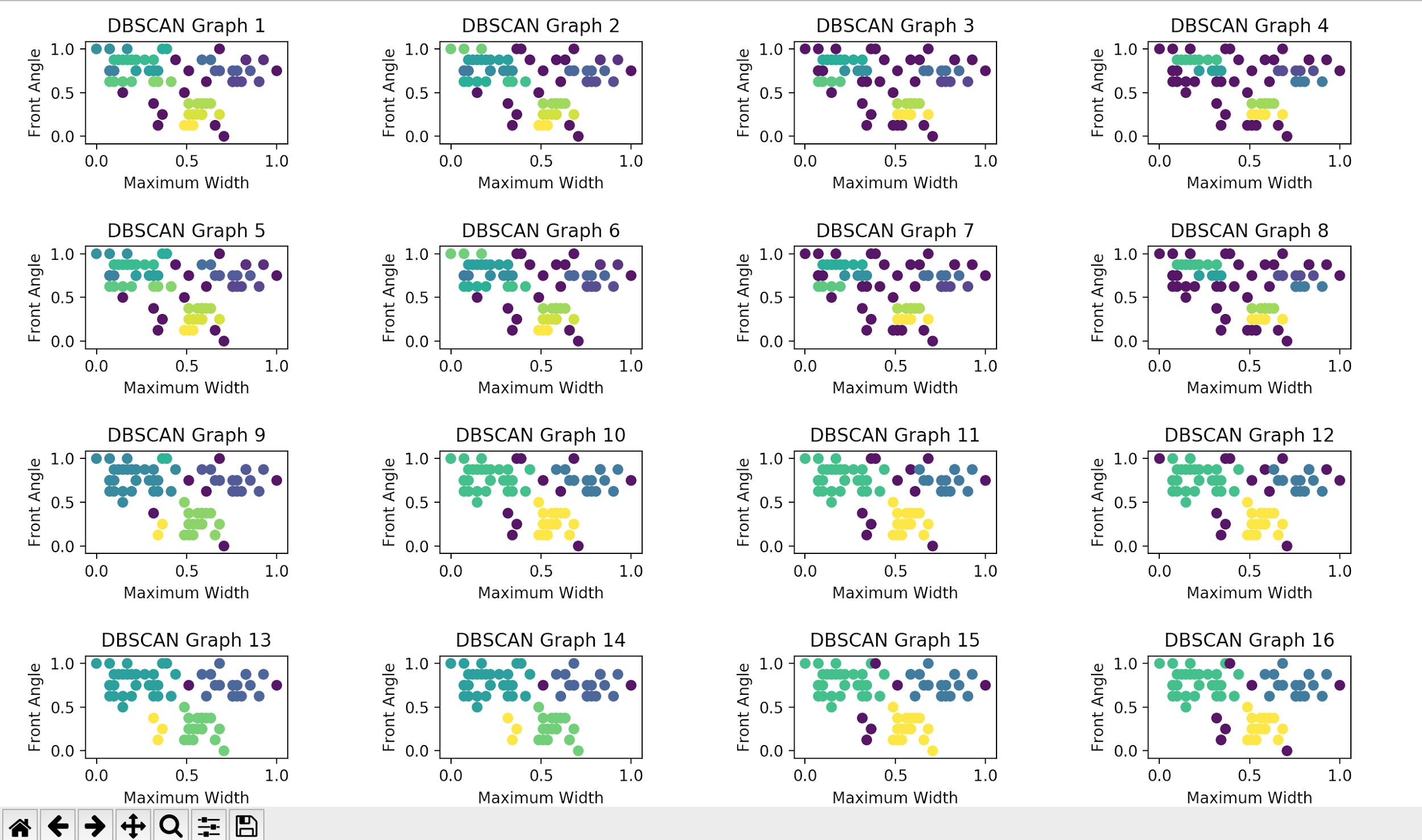
(# of correctly assigned points) / (# of points)

For the KMeans algorithm, I chose to set the number of user defined clusters at 3, because there were clearly 3 categories of data, so I set up the guidelines for the number of clusters to reflect my desired result. I also set the random state at 170, because I wanted consistent results every time, and not fluctuating results with switching points. Since some of the points were placed in a location where they had the potential of belonging to multiple clusters, I set the random state so the groups would be split the same way without fail each time. Only one point ended up being grouped into the wrong cluster.

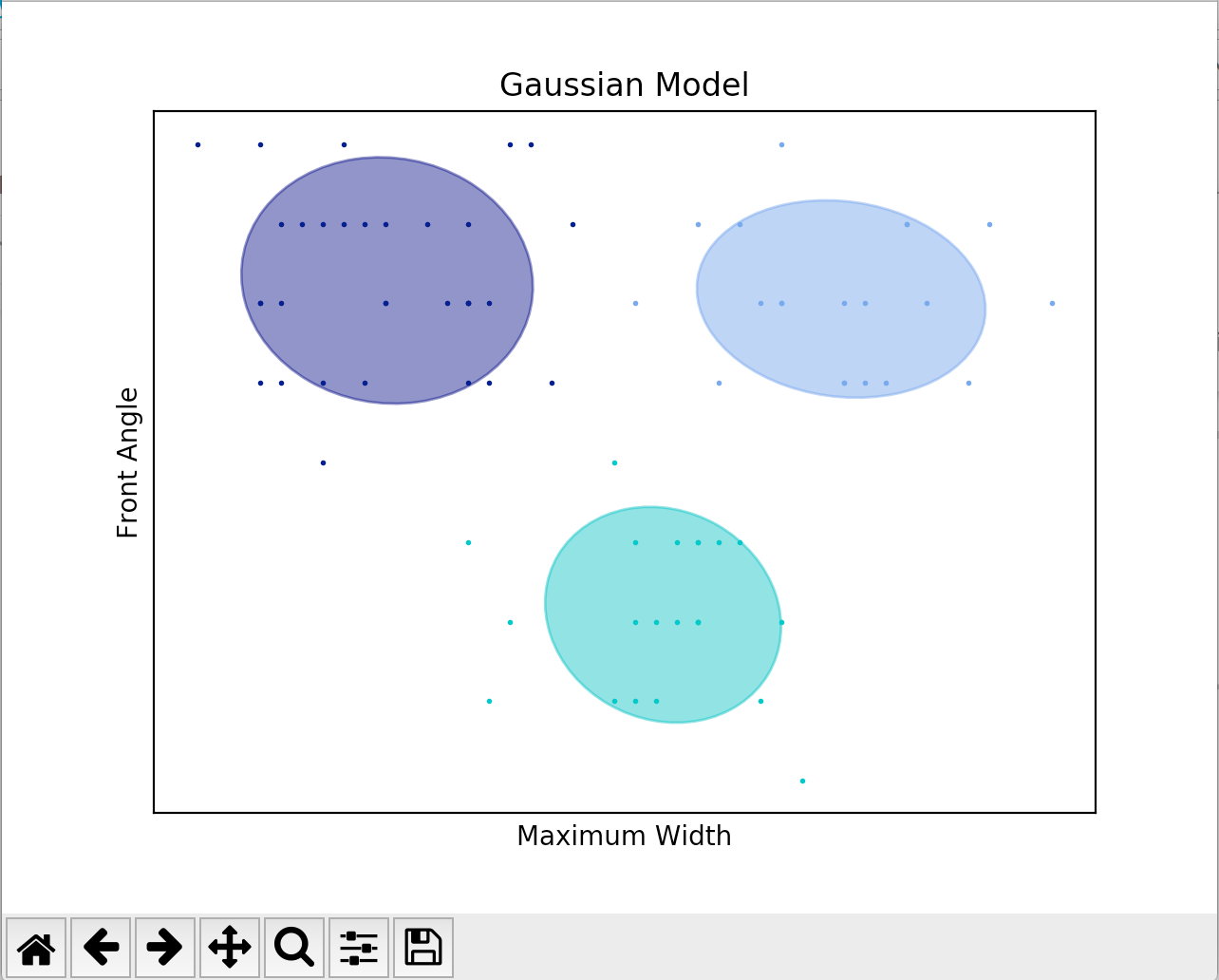
For the Agglomerative Clustering algorithm, I set the number of user defined clusters at 3, because I could clearly identify 3 groups of data from the graph. The rest of the algorithm is automatic, so the number of clusters defined only identifies when the algorithm should stop. All points were grouped correctly.



For the Density-Based clustering, I set the parameters of epsilon, the radius, at values between 0.11-0.14, and the minimum number of points needed in the circle of the specified radius to assign it to a group between 2 and 5. The various epsilon values I tested were to identify how selective I had to be with the circle to obtain the best results. I set the minimum number of points at various values also as a way to test and identify specificity when grouping points to find the best results. None of the parameter combinations produced the right results, ranging from getting about 10 of the points wrong to almost half of the points.



For the Gaussian Distribution method, I first set the user-defined number of distributions to be 3, as that was how the data was obviously organized and structured from what I saw. Unfortunately, the Gaussian distribution got one point wrong in the dataset.



Conclusion:

Which clustering algorithm gives the best results for this data? Why?

My hypothesis was wrong, because I had predicted that KMeans and Distribution would perform the best, but it turned out that the Agglomerative Clustering algorithm was the most accurate with this dataset. This is because even though KMeans and Distribution, compared to Agglomerative Clustering, were only one point off, that one point looked like it belonged in a different group because it was spaced a little further from its correct group and looked like it was part of another cluster assuming the groups of data were in circular shapes. Because KMeans and Distribution look at clusters that are more round in shape, that one point got classified wrongly. Meanwhile, Agglomerative Clustering just looks at the distance between data points to identify the groups, so it is able to deal with irregular shapes better.