**CIS 6930: Introduction to Data Mining**

**Individual Project 3 – Report**

In this project, the clustering techniques from the categories listed below are applied to the given datasets:

1. Distance-based clustering
2. Density-based clustering
3. Graph-based clustering

Please refer to the README file for the running of the program.

The observations for the various parts of the project are given below:

**PART-I:**

For the first dataset (*dataset1.csv*) I applied the following techniques based on the categories of clustering given:

1. K-Means Clustering for distance-based
2. DBSCAN for density-based
3. MCL (Markov Chain Clustering) for graph-based

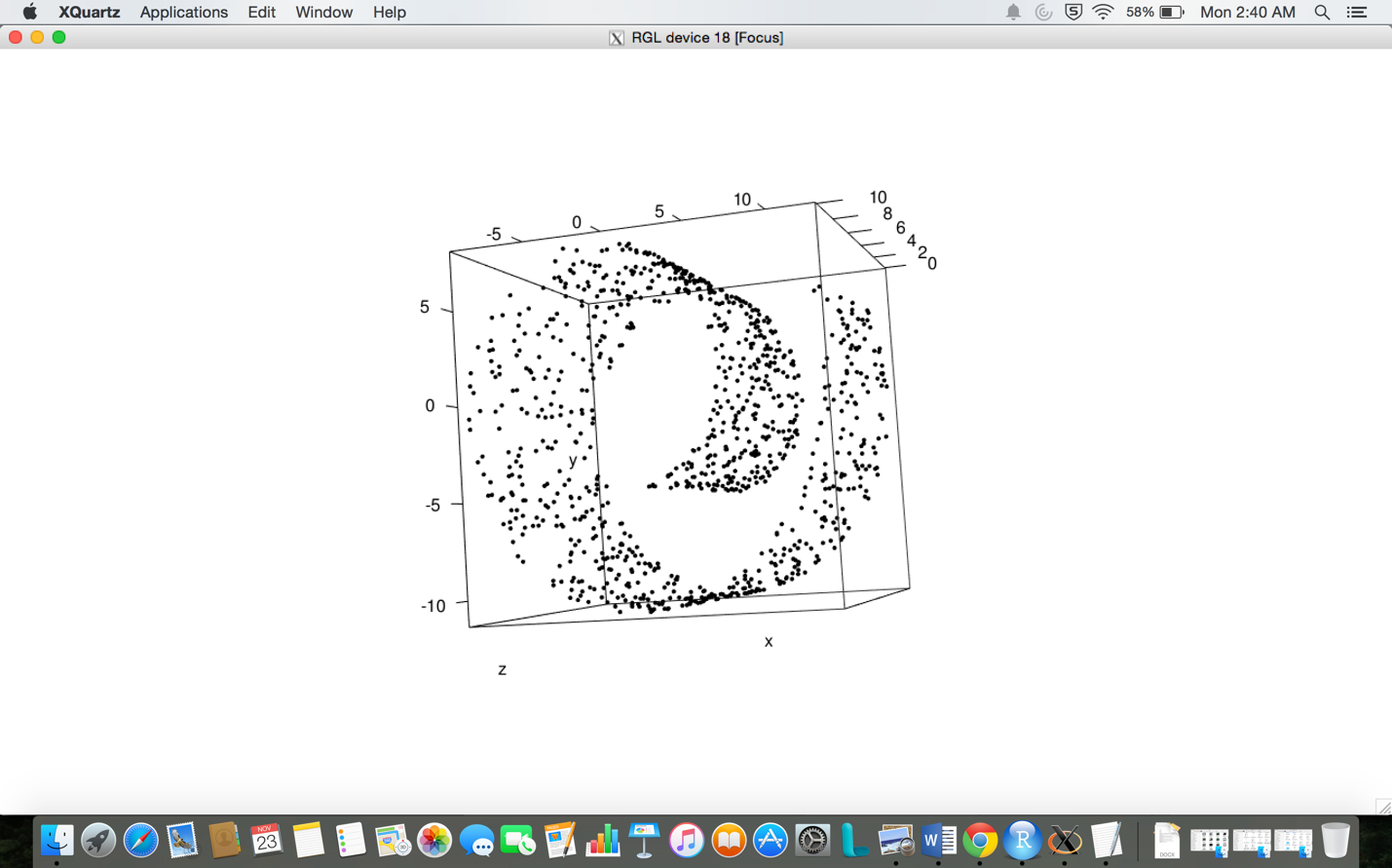
***Observations:***

For the K-Means clustering the output graph is as expected for the dataset given as I have used the k values as 10 in the code written and I have observed that the output changes for different values of k. I have chosen 10 as the k value as it gives the better output plot compared to others.

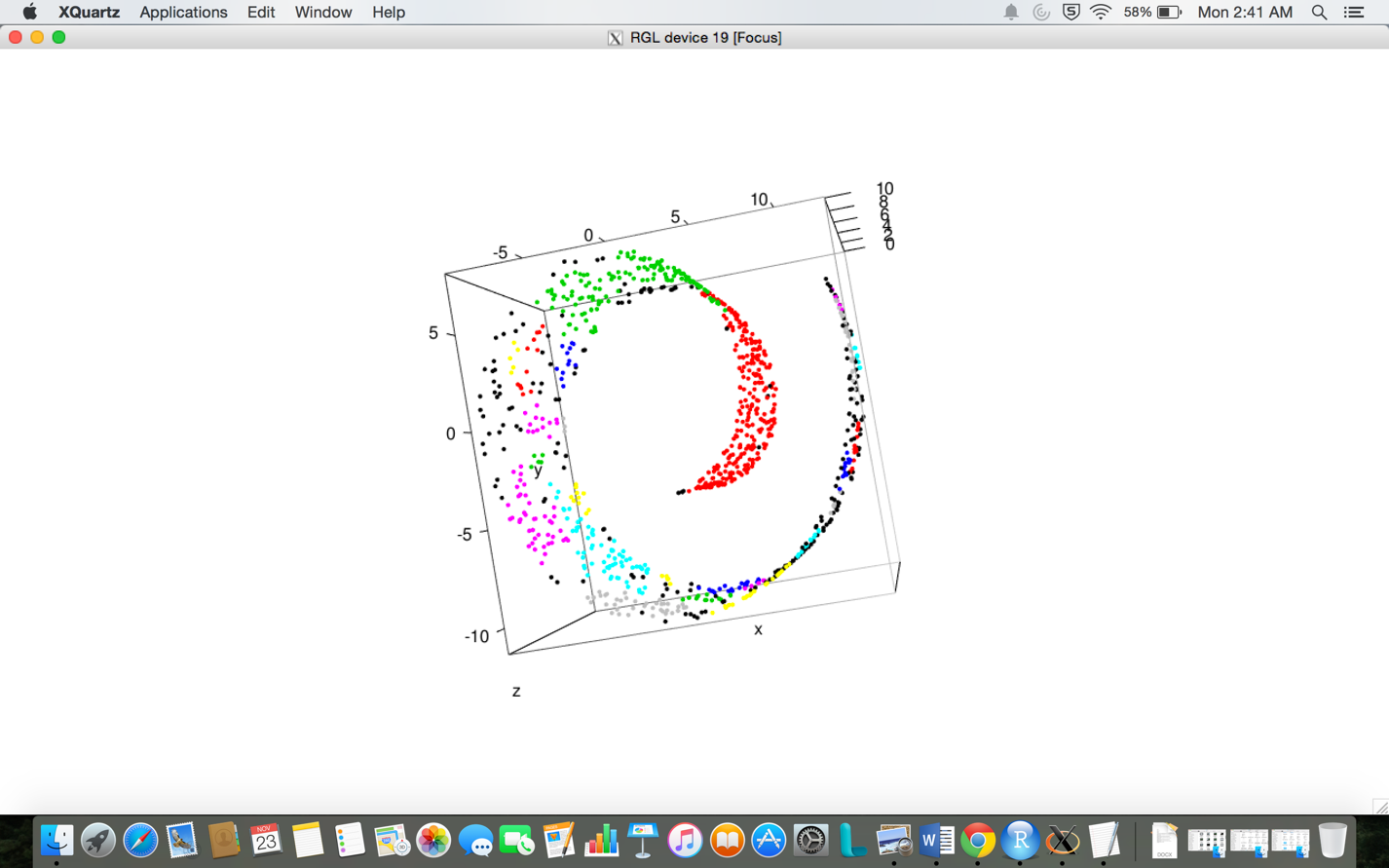
For the DBSCAN method, the output is as expected and you can see that by using the DBSCAN method it automatically assigns the number of clusters and since I have set the showplot=1, you can see the way in which the dbscan method clusters the data based on the density.

For the graph based clustering, I have used the Markov Chain Clustering method for the graph-based clustering method. Here the output of the plot shows the clustering of the data but initially I had to arrange the data in the matrix format as the MCL method only accepts matrix input for its calculation. The output graph shows us the expected outcome for the dataset used.

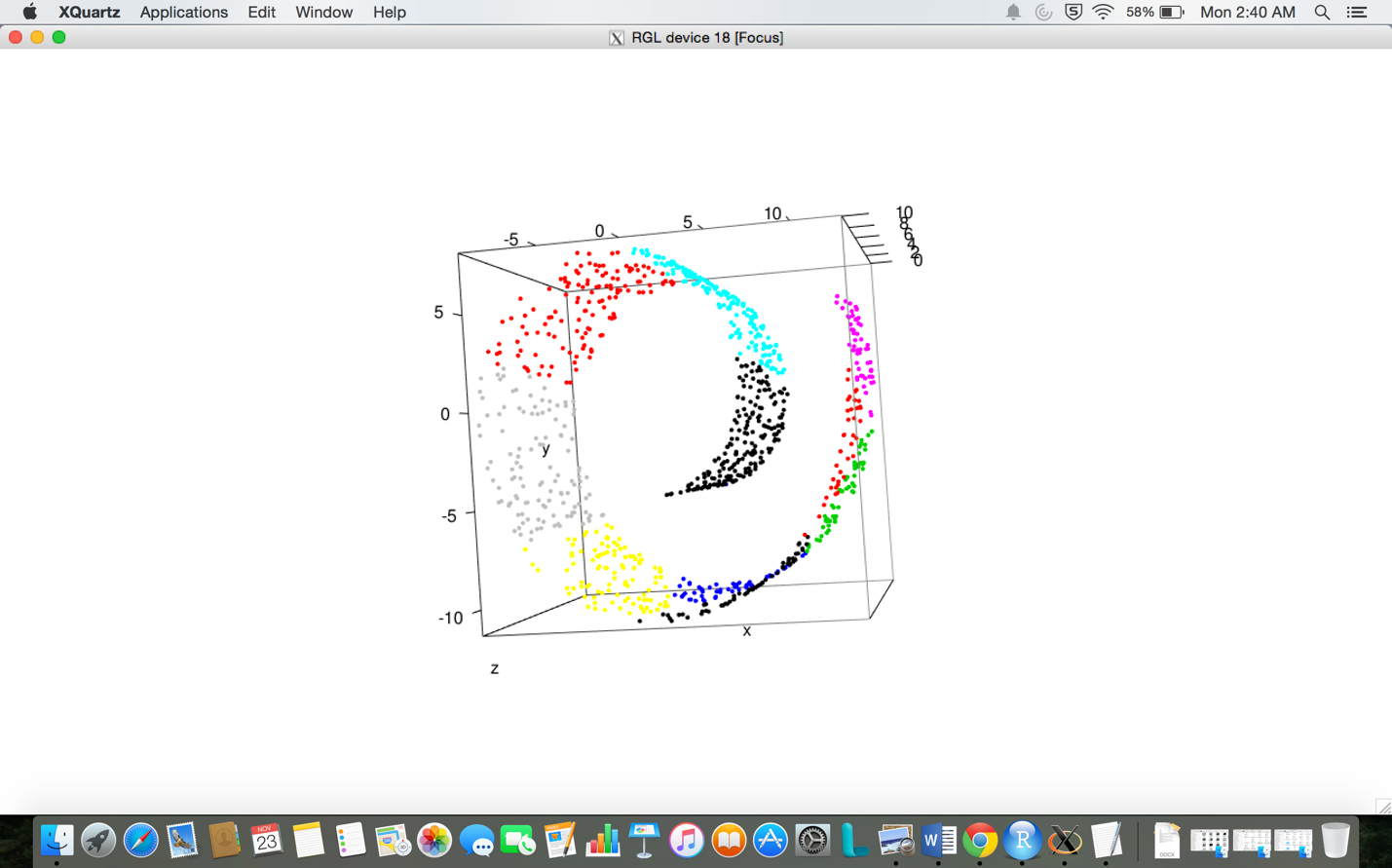
The output plots and graphs in 3D for the three techniques and for the actual data are given below:



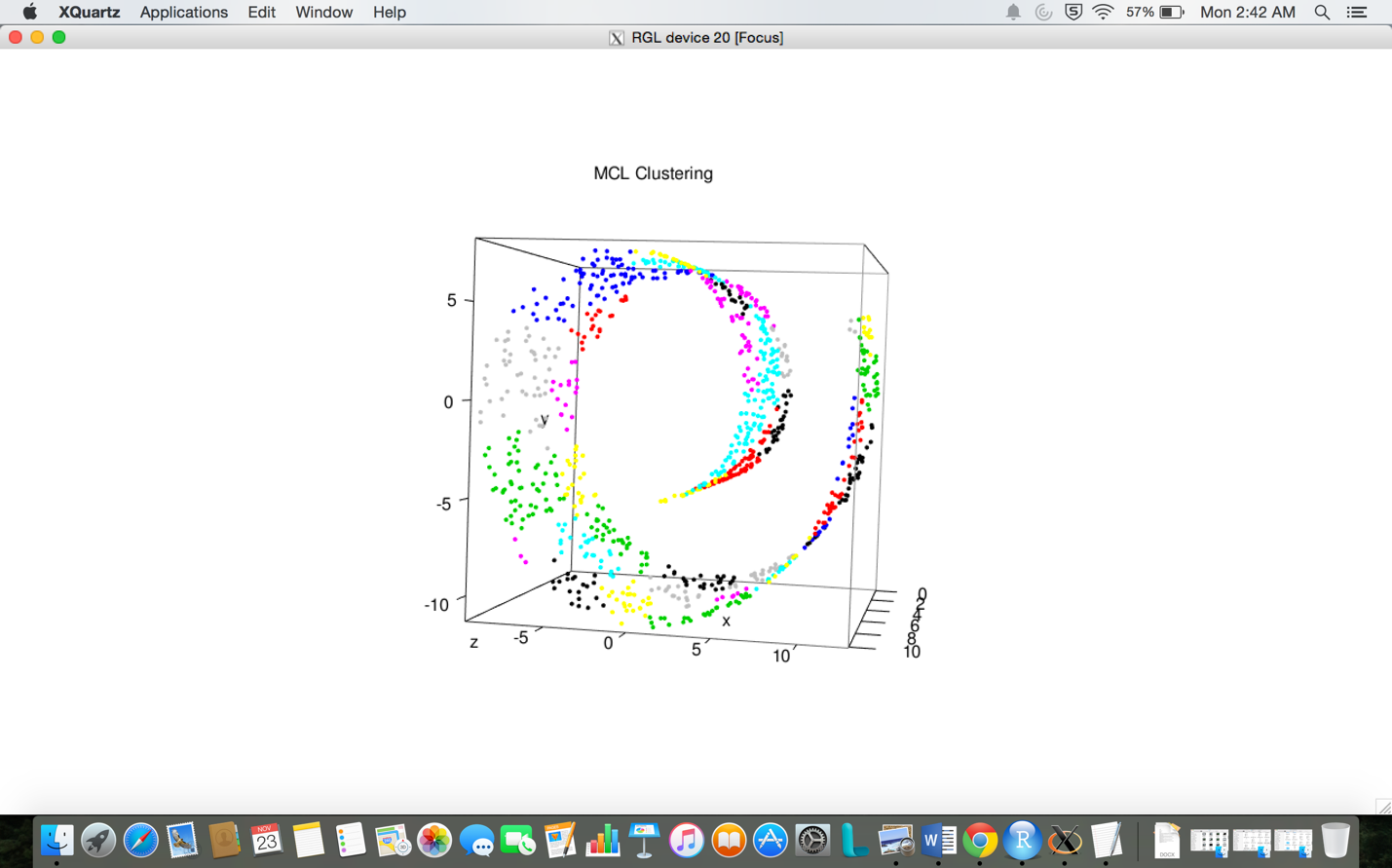
**Fig. Actual Data**



**Fig. K-Means Clustering**



**Fig. DBSCAN**



**Fig. MCL Clustering**

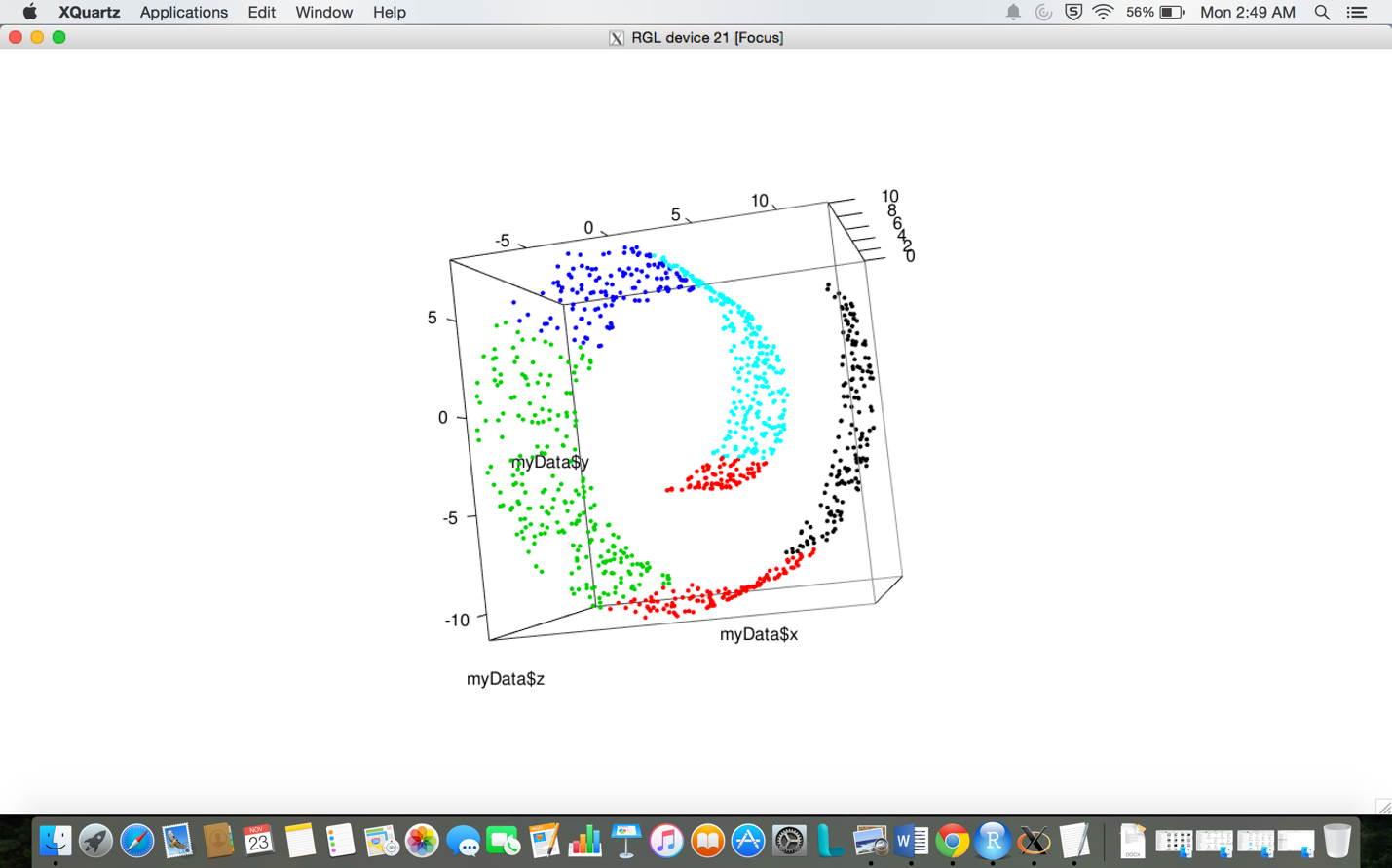
**PART-II**

***Observations:***

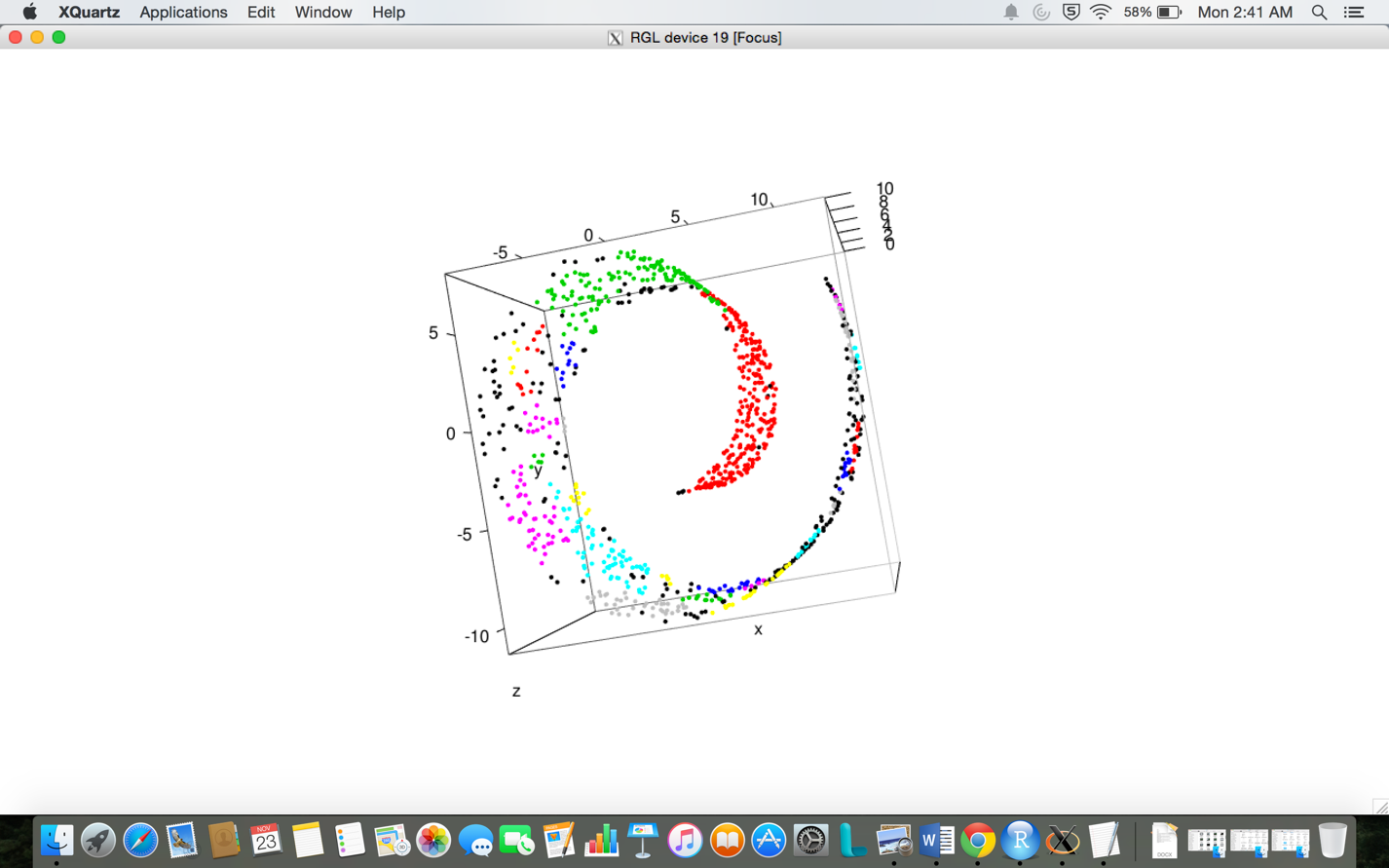
In the second part of the assignment, we were asked to write our own distance-based clustering method such that it assigns weight to *x* attribute 4 times, and *y* attribute 2 times and *z* attribute is treated normally, with weight 1. We were also asked to compare our method with *K-Means* clustering method.

After writing the code for the above mentioned clustering method, I found that the output does not vary that much from that of K-Means clustering method. In fact, our method does not improve the result when compared with that of K-Means method by much. I would say that the K-Means clustering method is a much more effective method for the given dataset.

The output plots in 3D for that of the K-Means and for the new distance-based clustering is given below:



**Fig. New Clustering Algorithm**



**Fig. K-Means Clustering**

**PART-III**

***Observations:***

For the third part of the assignment, we were asked to do the clustering for the second dataset (*dataset2.csv*). Since the dataset given is too large for us to apply different clustering methods and compare them, I have decided to use the K-Means clustering method (Distance-Based) for the clustering of the second dataset.

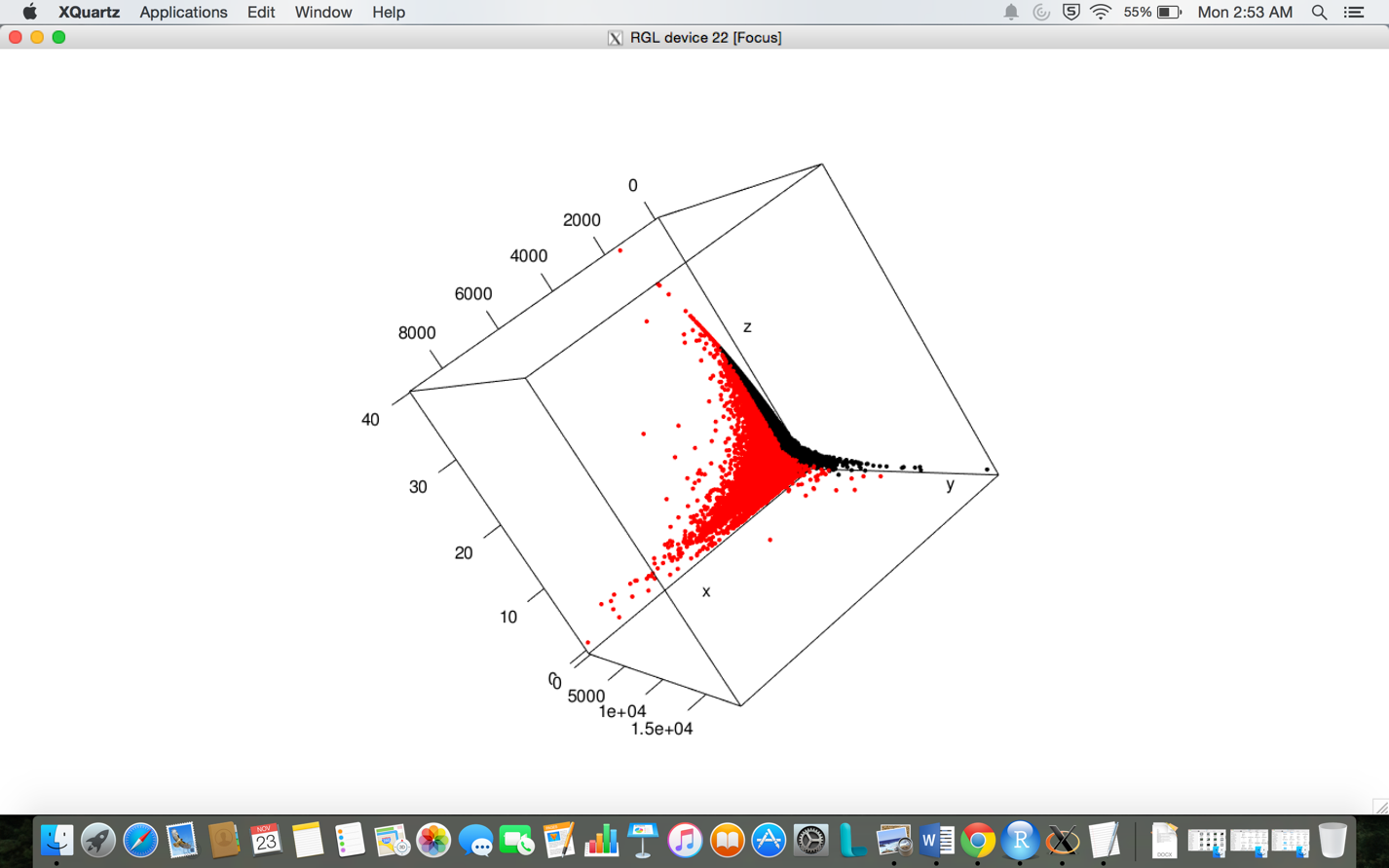
I use the K-Means clustering method because it is one of the most basic and widely used clustering algorithms. Also it is easy to specify the number of centroids in this algorithms. This algorithm also does not take much computation time when used in R for big datasets such as the one which we are using now.

I have chosen the number of clusters to be 2 as the final column of the dataset only contains two outputs- TRUE or FALSE. This has given an output graph with about 75% accuracy when used with K-Means. I also plotted a graph to find the clusters in the dataset in R and found that the number of clusters can be 2 to 4. I have also found that when I changed the number of centroids, the computational time increased and also the accuracy of the output seemed less than that of the previous one.

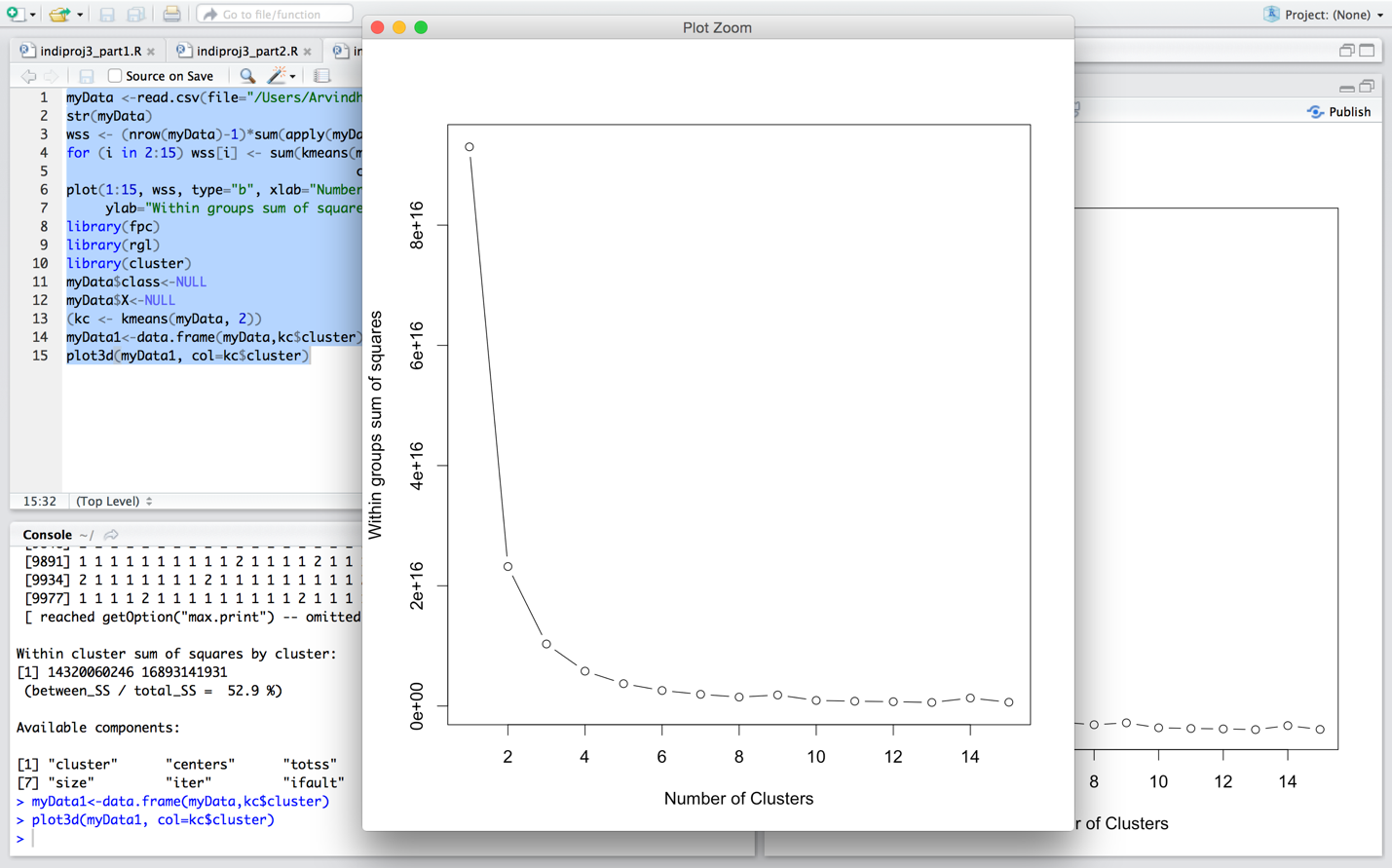
As for the challenges faced while doing this part, I would say that the the size of the dataset hindered the use of other clustering algorithms such as DBSCAN and hence it was not possible to compare the various algorithms based on their computational time and also that of their accuracy for the given dataset.

A better solution for this would have been to take some random number of data from the dataset, say 10000, and used the algorithms mentioned above to compare and deduce which of the algorithms was better suited for this dataset.

The output plots and graphs in 3D for this part are given below:



**Fig. K-Means Clustering for dataset2.csv**



**Fig. Finding the number of clusters in dataset2.csv in R**