# CENG 464

# PROJECT REPORT

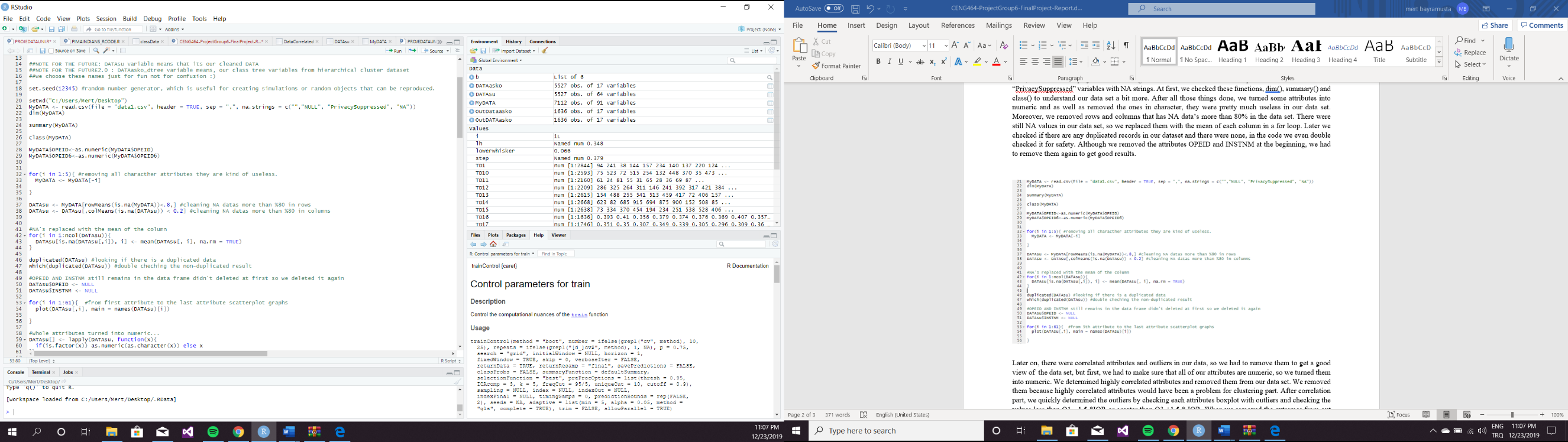
## PROJECT GROUP 6

**25/12/19**

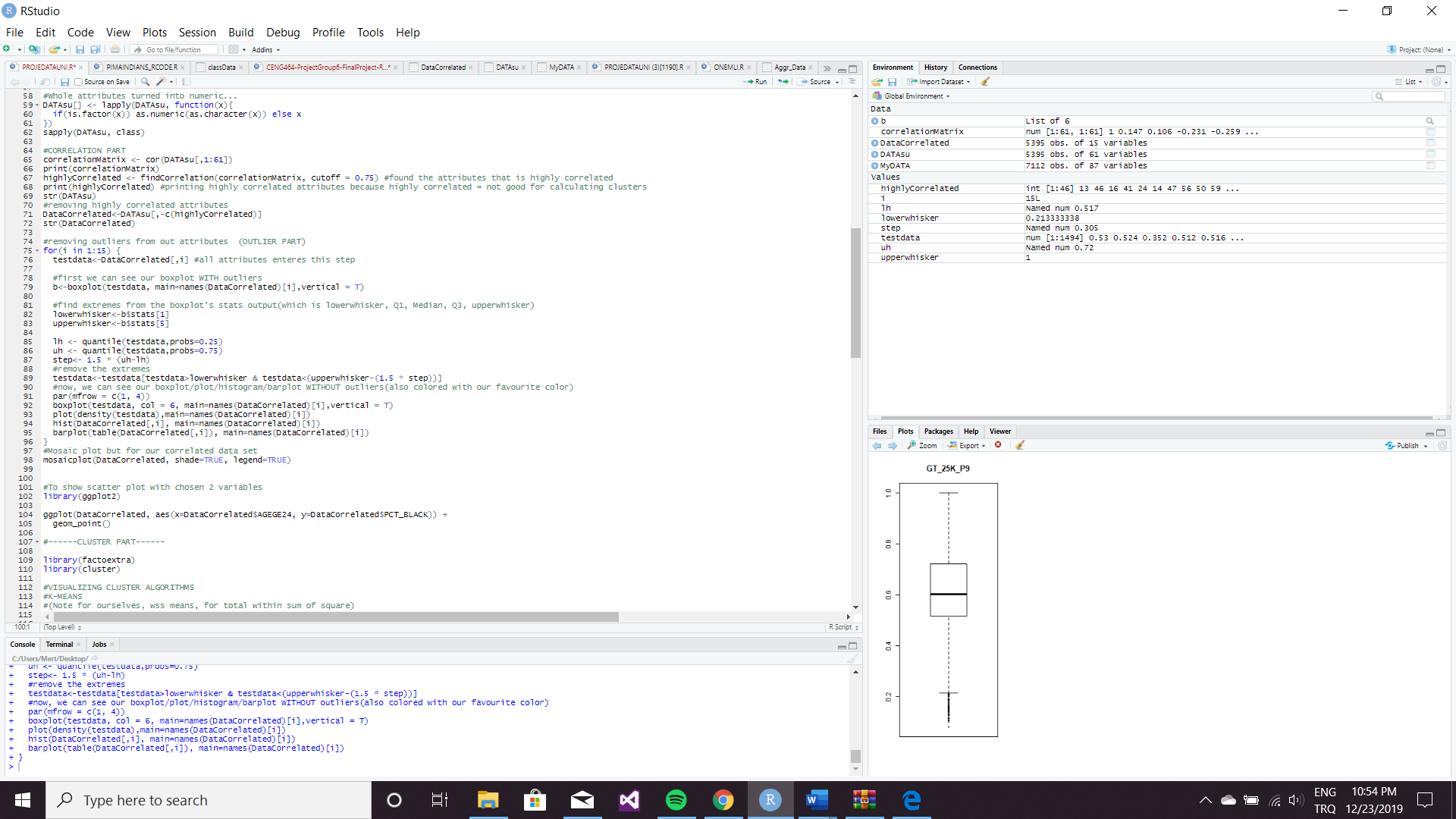
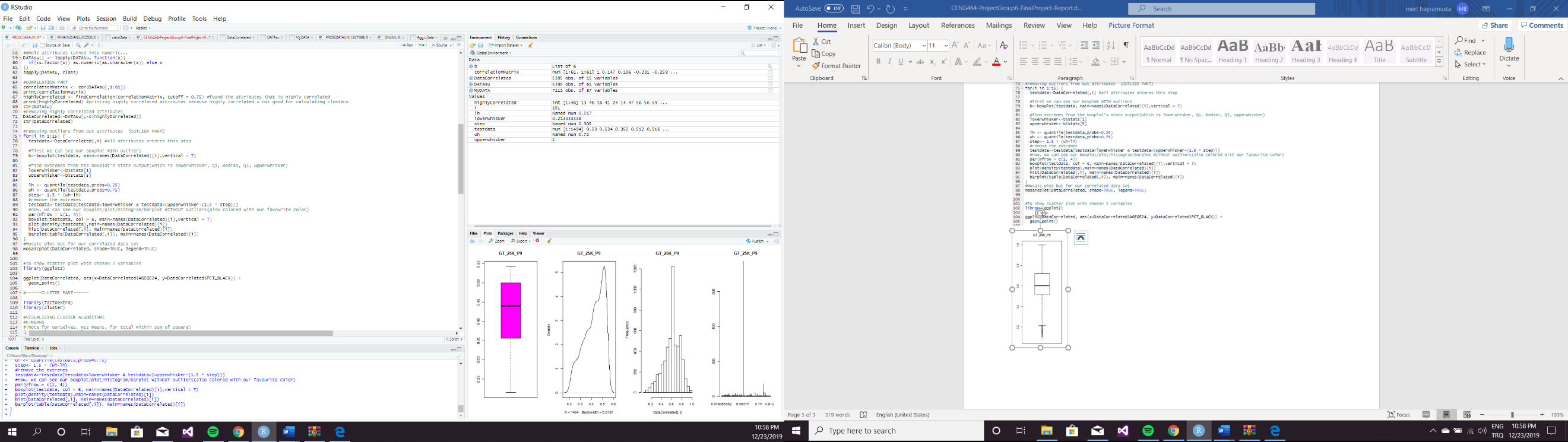
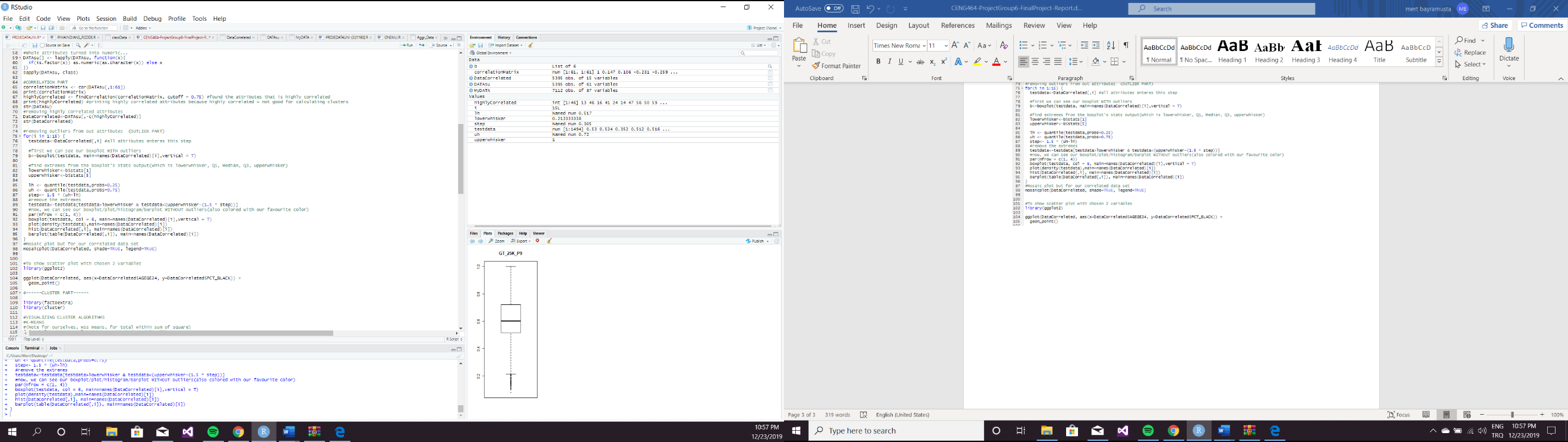
### 201711006 Mert BAYRAMUSTA

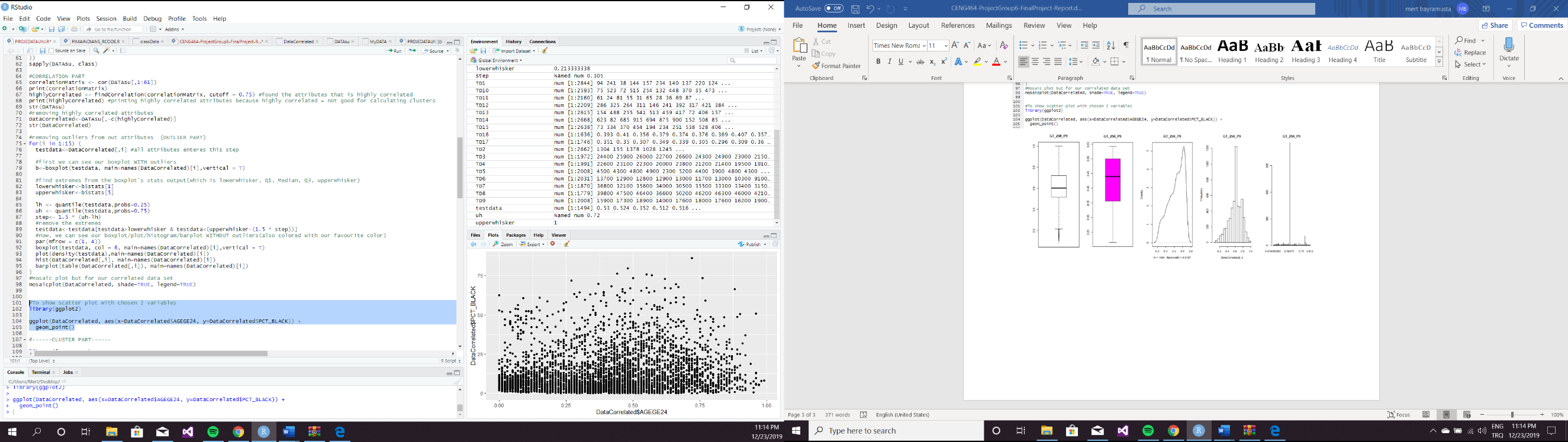
**201611047 İrem ÖZTÜRK**

First, we started our project with cleaning the useless data variables. We replaced all the “NULL” and “PrivacySuppressed” variables with NA strings. At first, we checked these functions, dim(), summary() and class() to understand our data set a bit more. After all those things done, we turned some attributes into numeric and as well as removed the ones in character, they were pretty much useless in our data set. Moreover, we removed rows and columns that has NA data’s more than 80% in the data set. There were still NA values in our data set, so we replaced them with the mean of each column in a for loop. Later we checked if there are any duplicated records in our dataset and there were none, in the code we even double checked it for safety. Although we removed the attributes OPEID and INSTNM at the beginning, we had to remove them again to get good results.

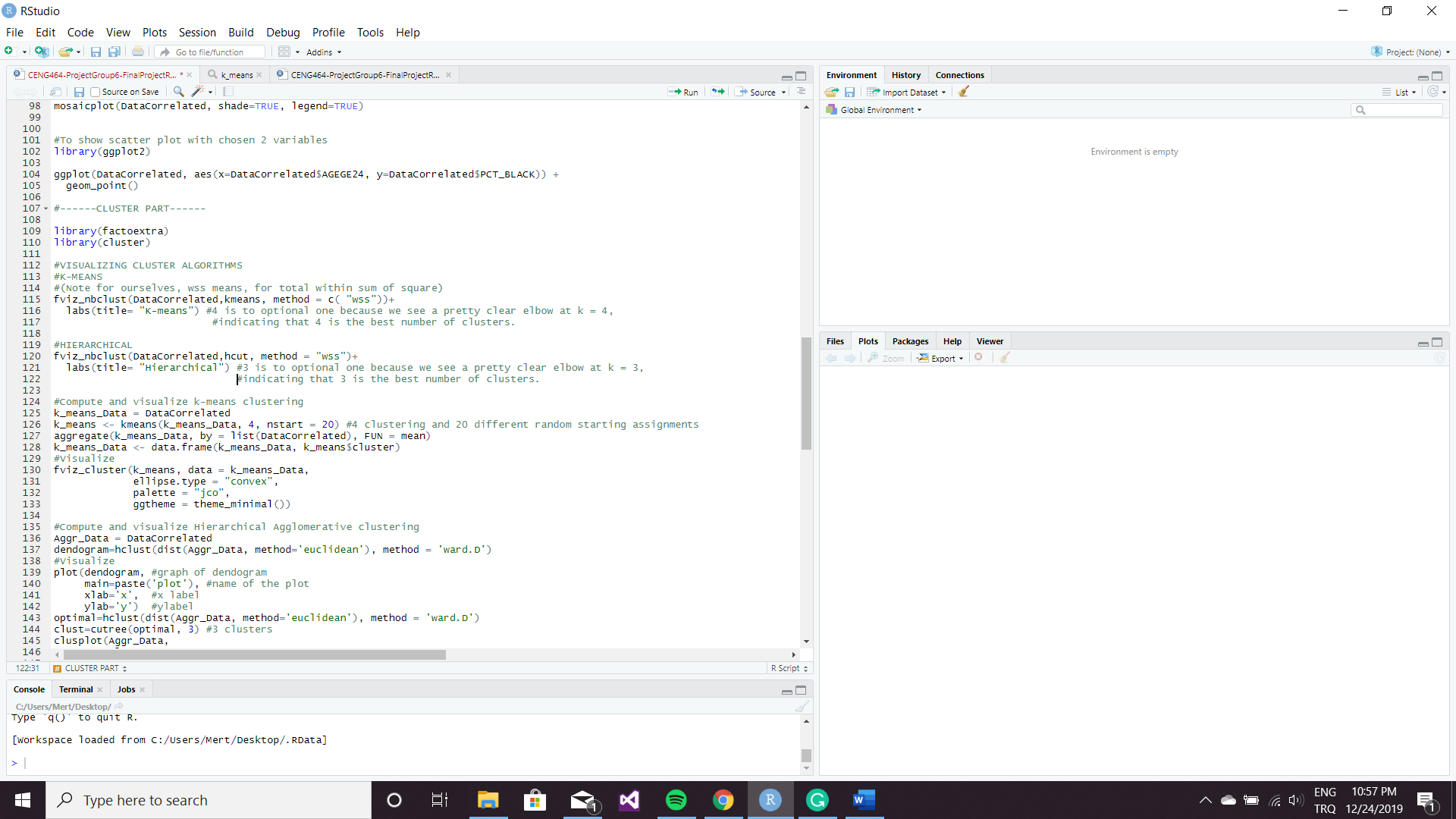


Later on, there were correlated attributes and outliers in our data, so we had to remove them to get a good view of the data set, but first, we had to make sure that all of our attributes are numeric, so we turned them into numeric. We determined highly correlated attributes and removed them from our data set. We removed them because highly correlated attributes would have been a problem for clustering part. After correlation part, we quickly determined the outliers by checking each attributes boxplot with outliers and checking the values less than Q1 - 1.5 \*IQR or greater than Q3 +1.5 \* IQR. When we removed the extremes from out data set, we checked boxplot, plot, histogram and barplot of the attributes, we colored the outlier-free boxplot with purple to see the difference. (For quality purposes only one of the attributes which is GT\_25K\_P9 has been displayed here.) After removing the correlations and outliers from our data set, we checked the mosaic plot and the scatter plot, but because our data still has so many dimensions, mosaic plot turned out to be unreadible, and with scatter plot we had to check with chosen 2 variables from the data set.





When we finished to preprocessing part, we got more understandable data frame. This is the type of data frame that we want because, in that way, we could find some valuable pattern to visualize. We started to do clustering as a first step because grouping data with similar properties will help us to create a new attribute that has information about each data’s group. With the help of this feature, it will be much easier and meaningful for us to classify our data set. Firstly, we use the k-means clustering method. With the help of this feature, it will be much easier and meaningful for us to classify our data set.



harita, metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

To reading that table we can easily say that 4 is to optional one because we see a pretty clear elbow at

k = 4, indicating that 4 is the best number of clusters.

As a second clustering method, we choose Hierarchical Agglomerative clustering method.

harita, metin içeren bir resim

Açıklama otomatik olarak oluşturulduHowever, when our code shows us the Hierarchical Agglomerative clustering method’s table, we decide that 3 is to optional one because we see a pretty clear elbow at k = 3, indicating that 3 is the best number of clusters. After to get information from those tables, we started to visualize k-means clustering and Hierarchical Agglomerative method.

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

metin, harita, tablo içeren bir resim

Açıklama otomatik olarak oluşturuldu

Image of k-means clustering Image of Hierarchical Agglomerative clustering

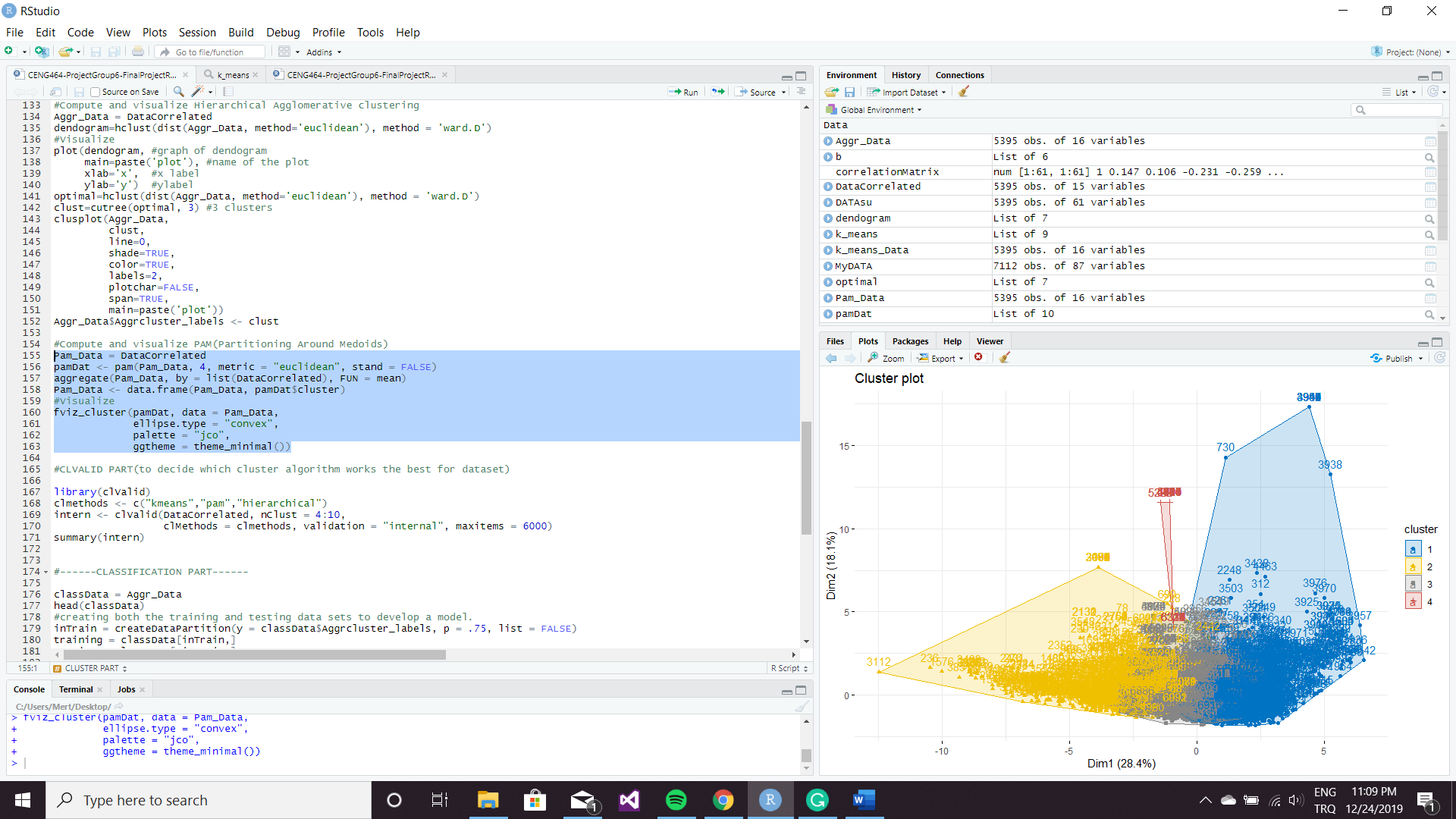
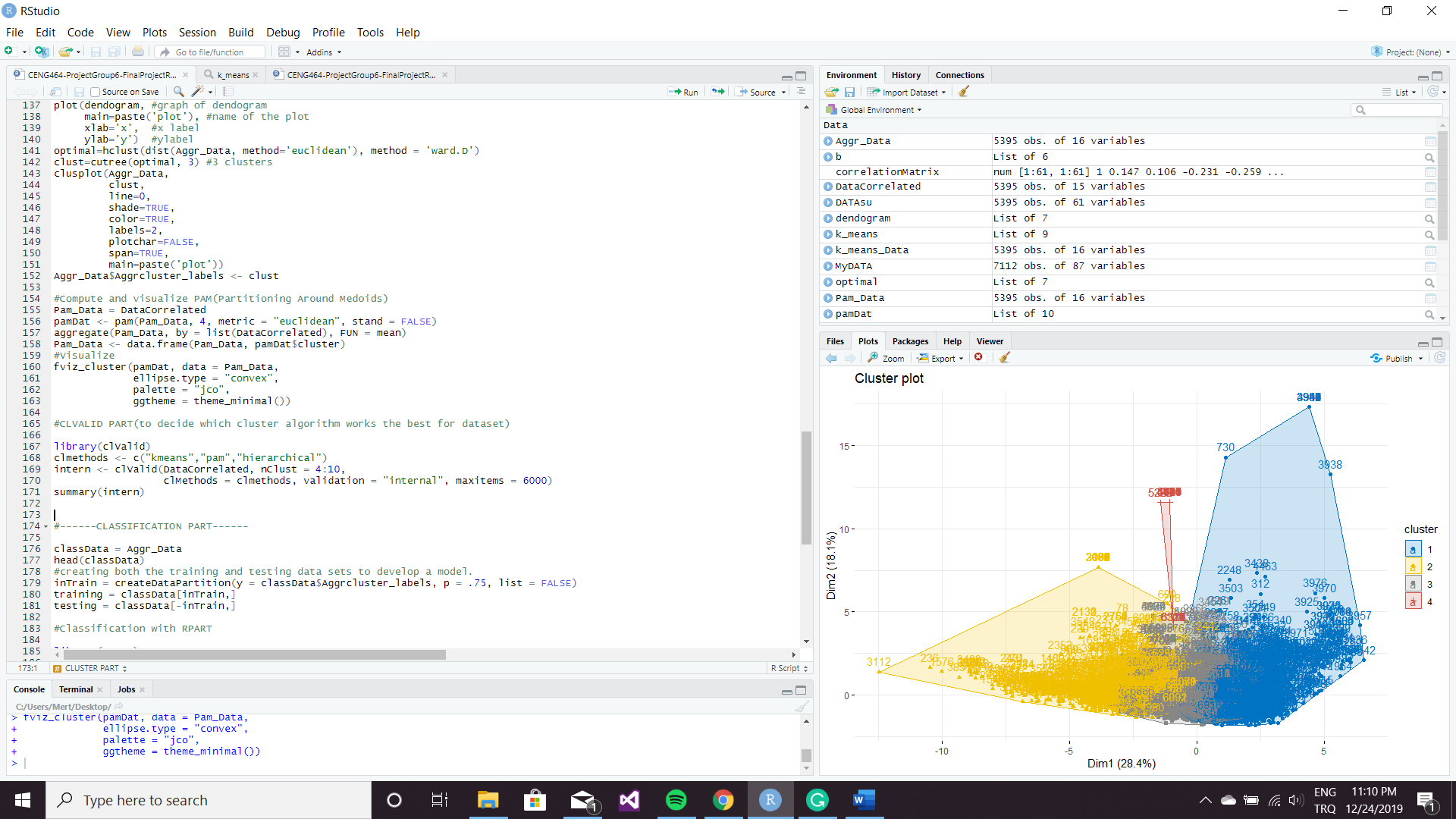
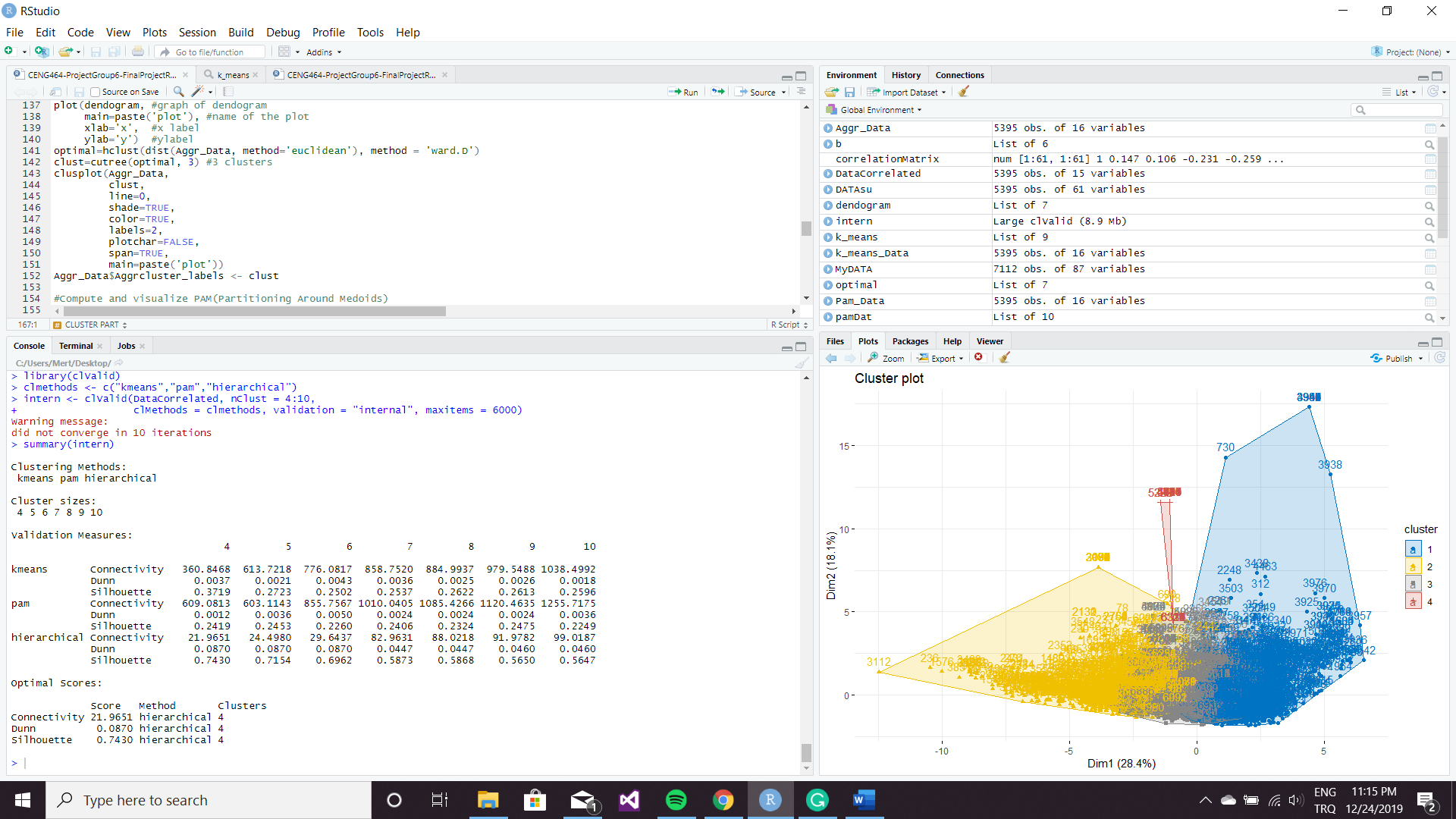
Our third clustering method is PAM (Partitioning Around Medoids). According to, information that we collect about PAM we thought that to using k=4 value for this method, would be better than using 3 because PAM is a part of k-means therefore, it should use k=4 value.

Image of PAM

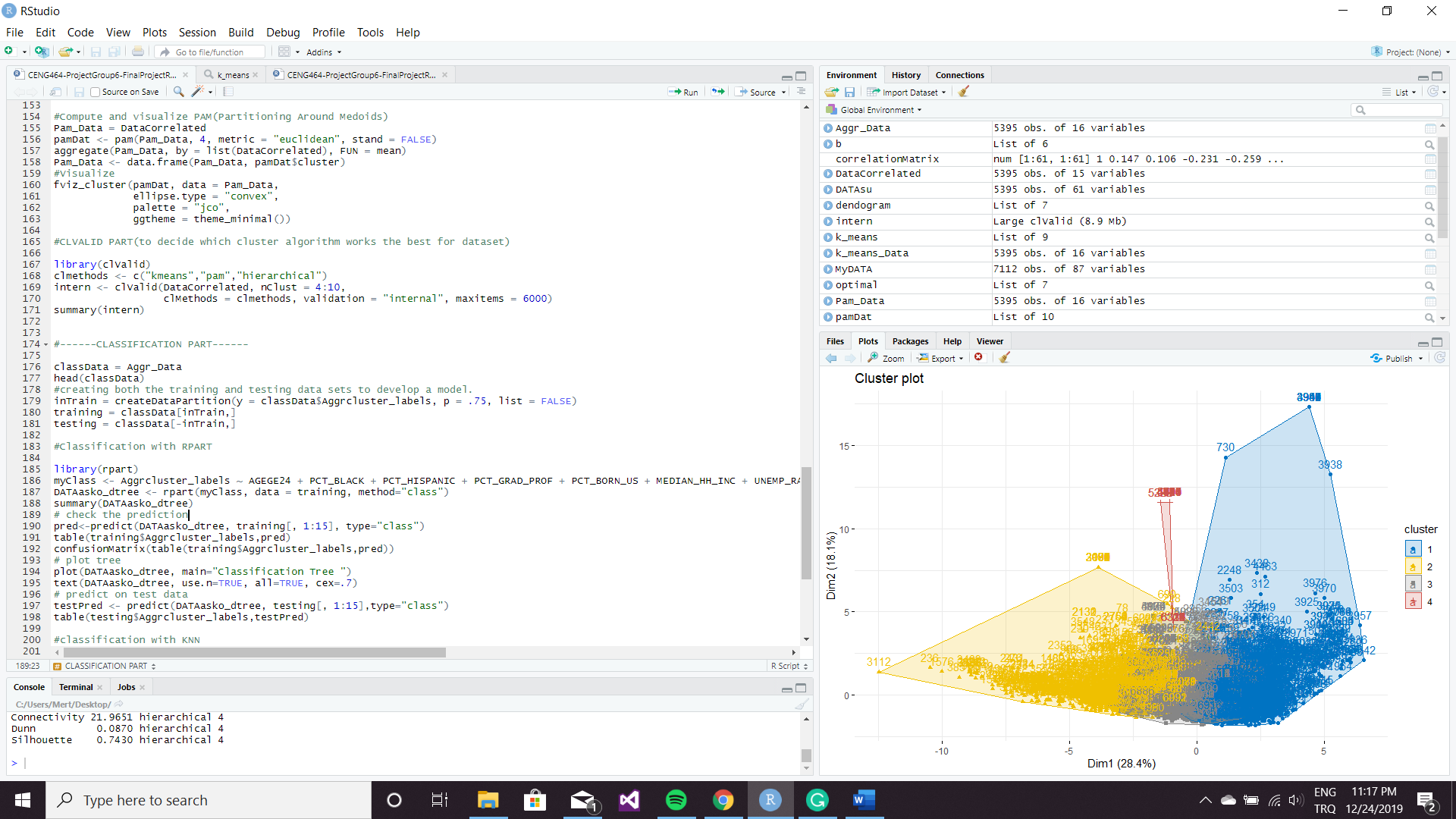
When we collected all three clustering methods’ information, we added a function that can help us to decide which clustering method is better than the others. After a search session on the net, we found the “clValid” library for it.



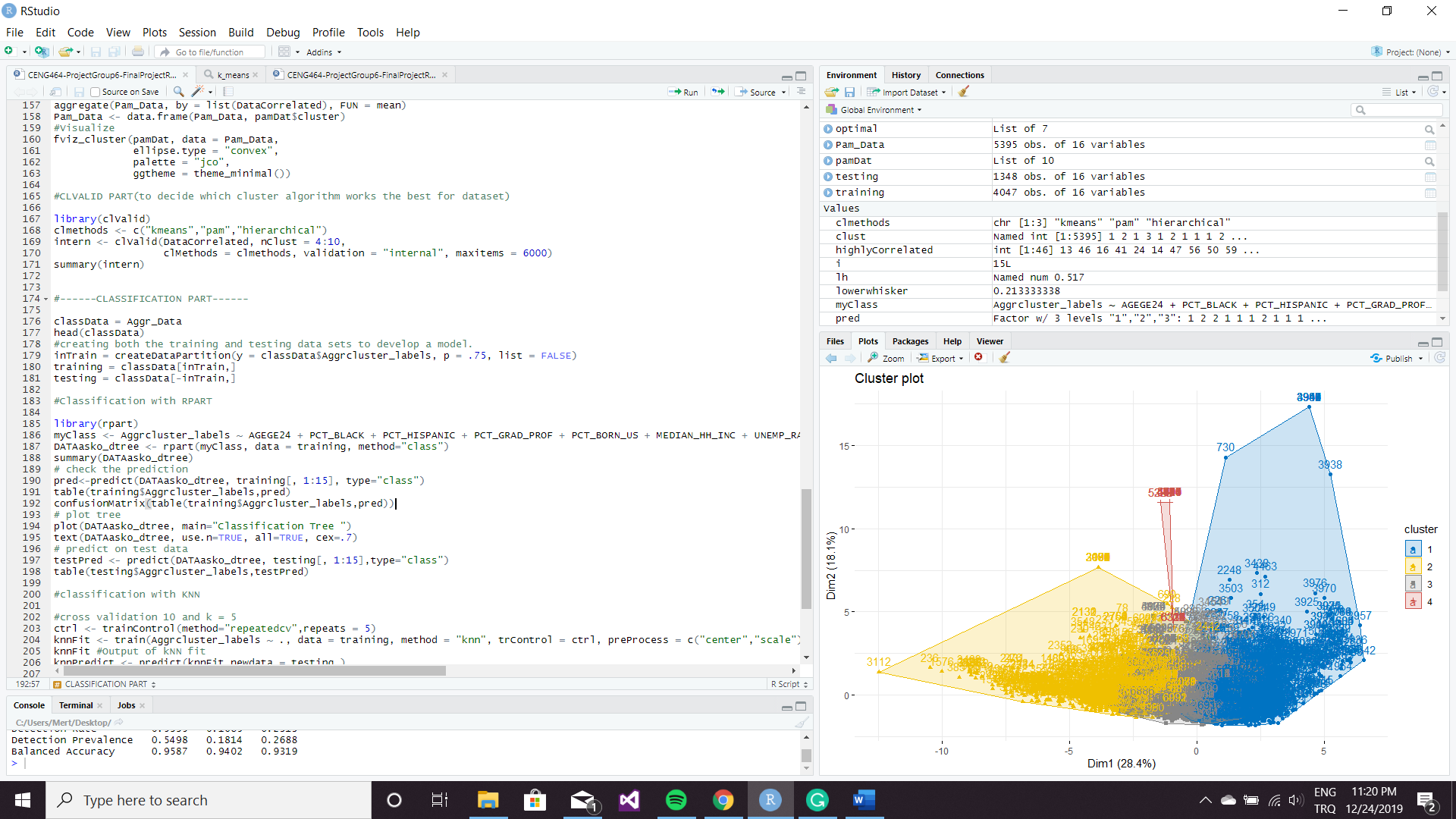
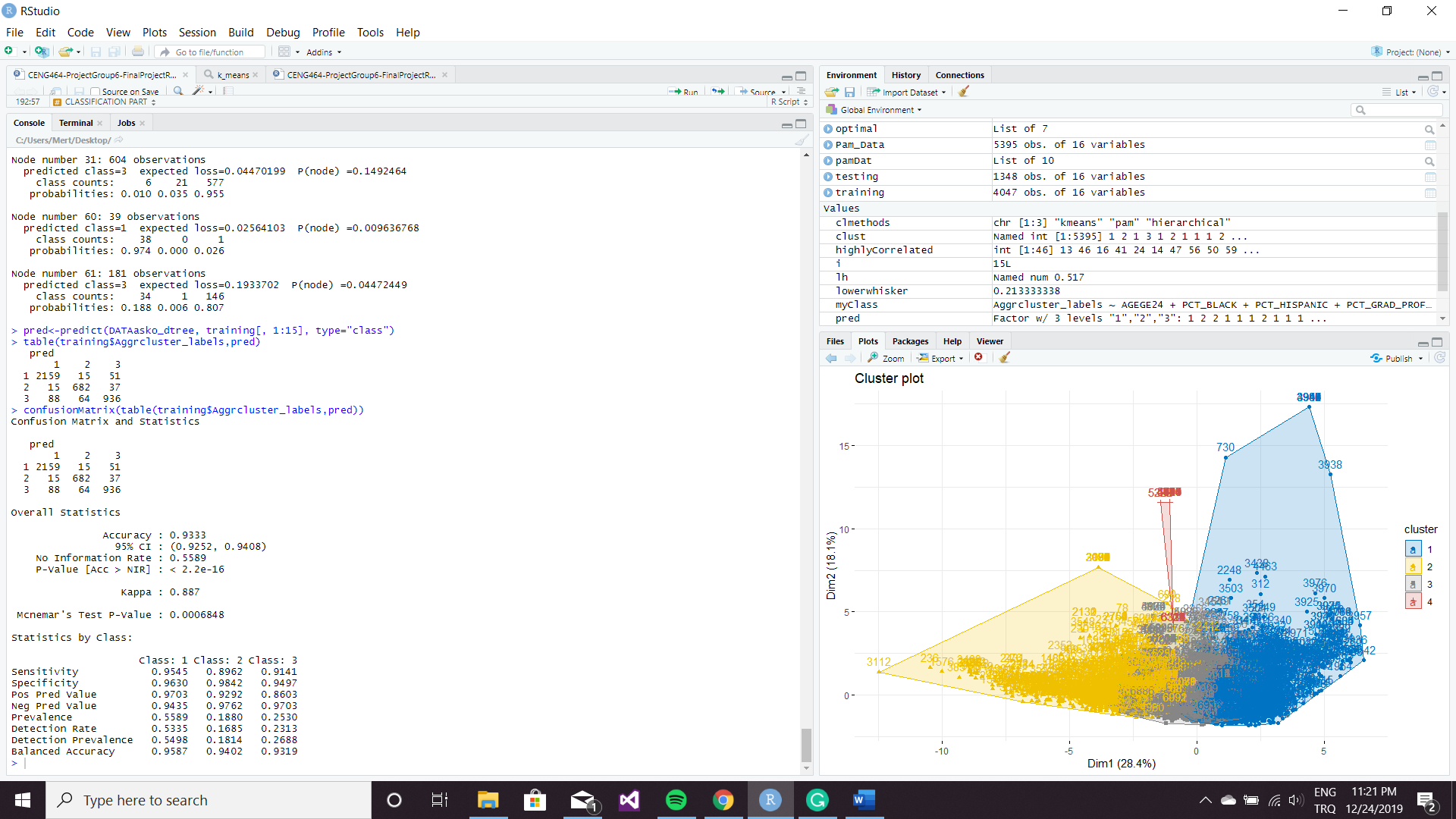
After using the function on this picture, we use summary function to see the result clearer. As a result, our code gave us this;

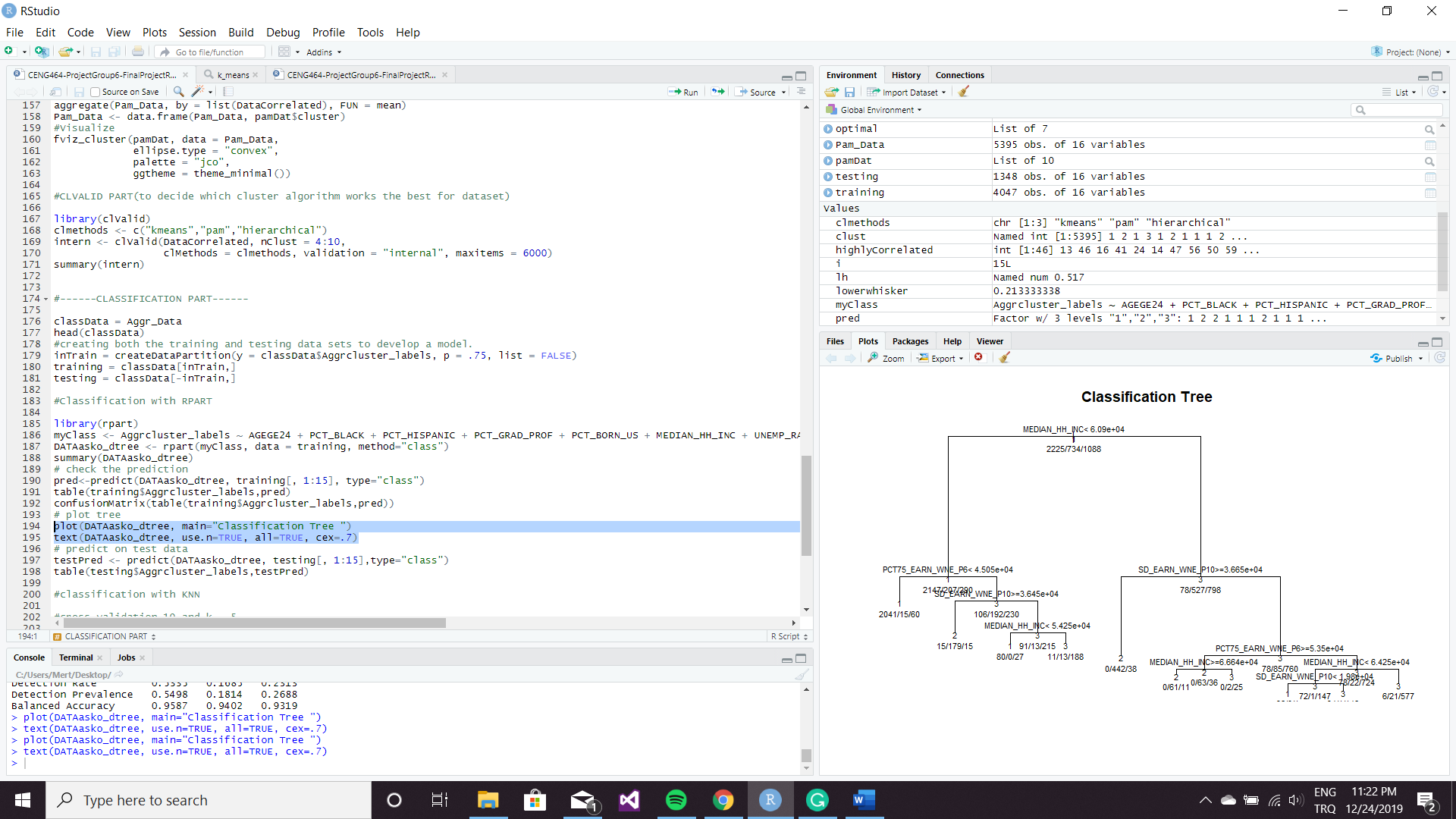


Considering this information, we decide to use Hierarchical Agglomerative clustering method for classifying. After this call, we started to search for classification methods. First, we divide our data set 75 percentage to 25 per cent for training and test. Secondly, we added Aggrcluster\_labels attribute that created by Hierarchical Agglomerative clustering method. Then we defined "myClass" that you can check in this code below;



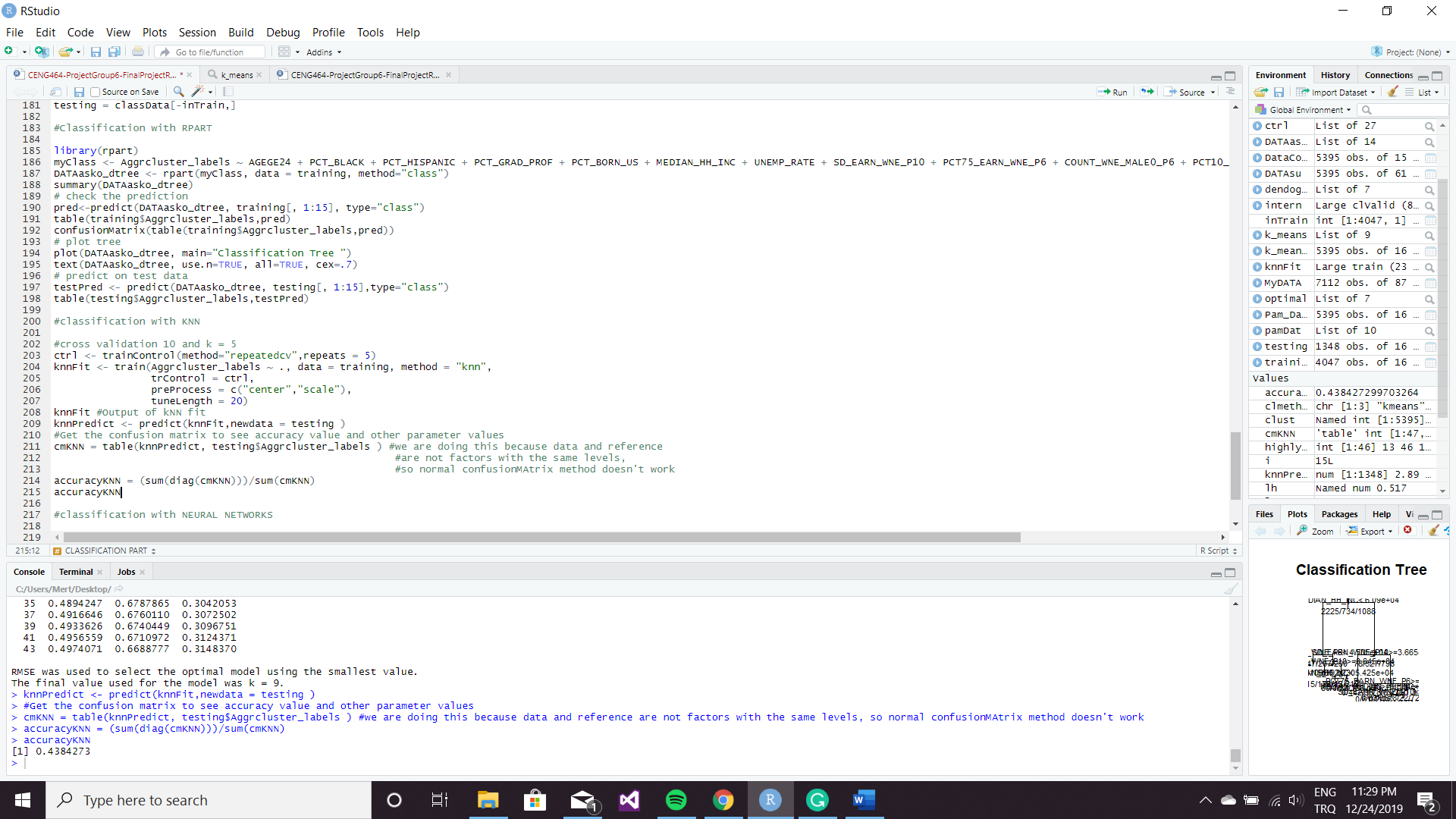
After that, we check the prediction and use confusionMatrix function. It gave us a clear understanding. You can check over the results and its classification tree on below;

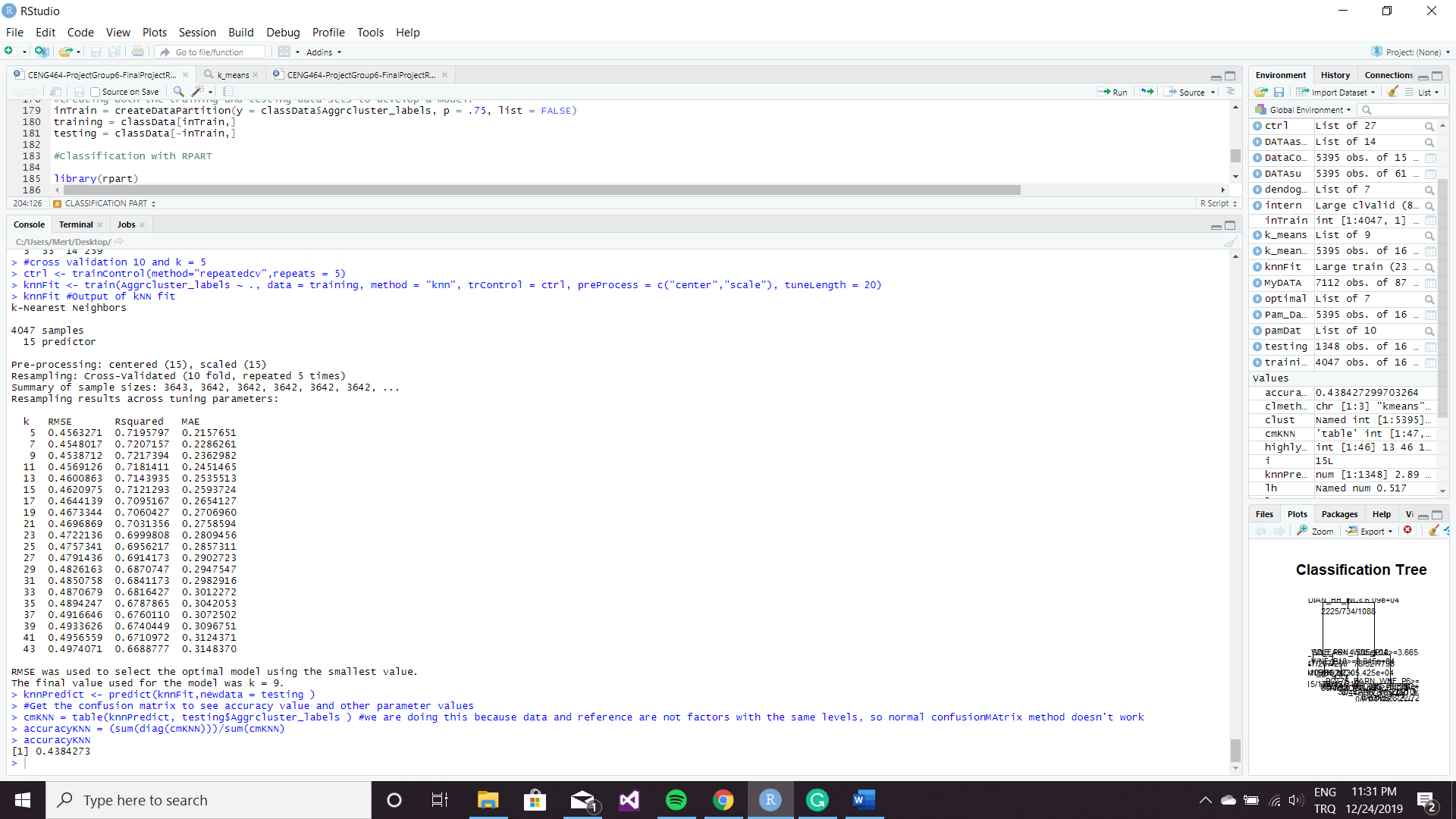




And after all the results and coding, we got 0.9333 accuracy in the decision tree.

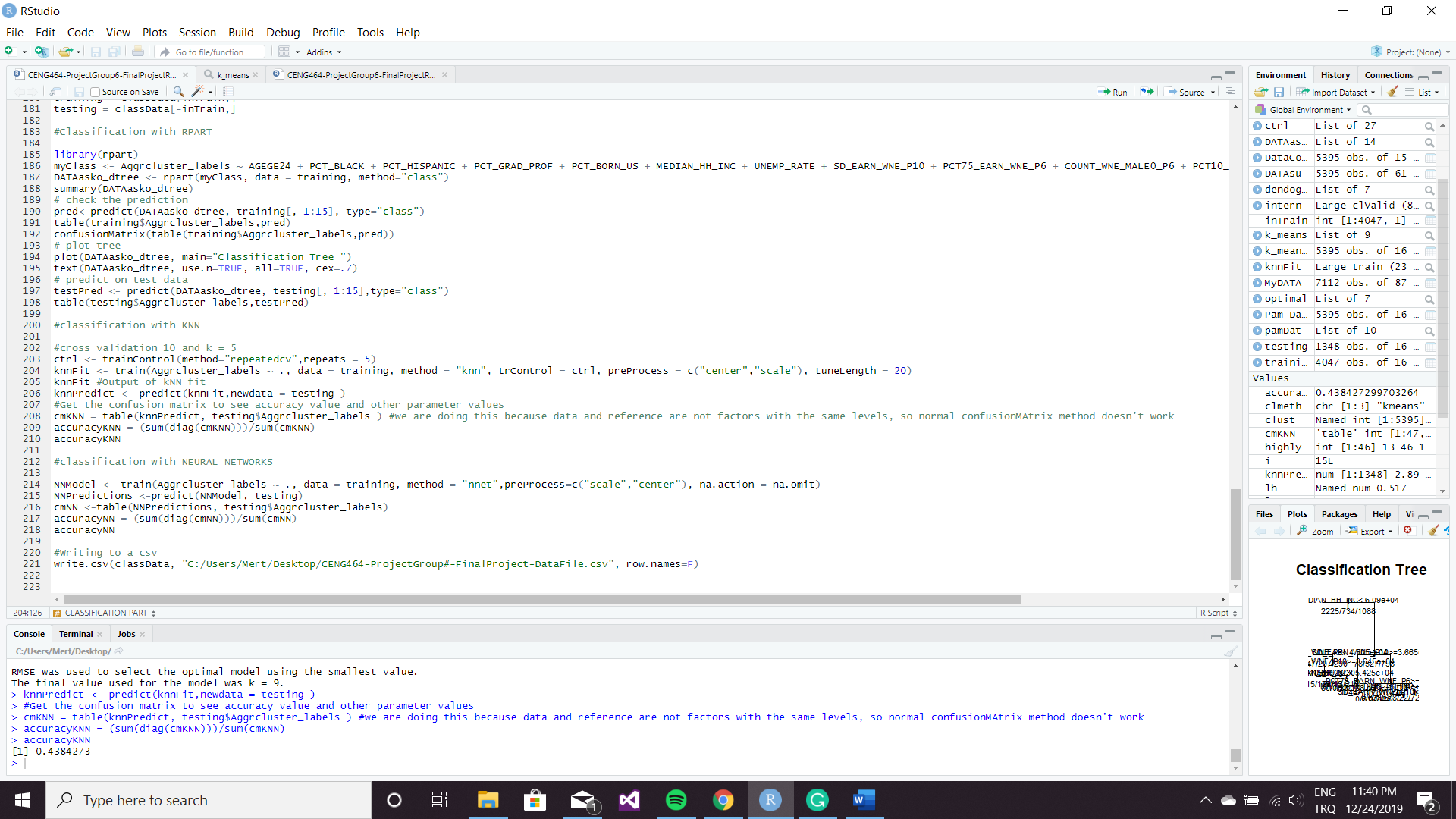
After we got the decision tree’s result, we wanted to see what can we see when we use kNN classification algorithm so we defined the cross-validation 10 and k = 5 and we predict and put it in the value that is named knnPredict by us. We want to get the confusion matrix to see accuracy value and other parameter values, but data and reference are not factors with the same levels, so normal confusionMatrix method doesn't work this is why we tried another path for it.





Again, after all the results and coding has been done, we got 0.4384271 accuracy. Definitely Decision Tree’s accuracy is much better.

Our last classification method is the neural network method. We defined a value called NNModel to train with our data set and learn it by using the neural network algorithm.

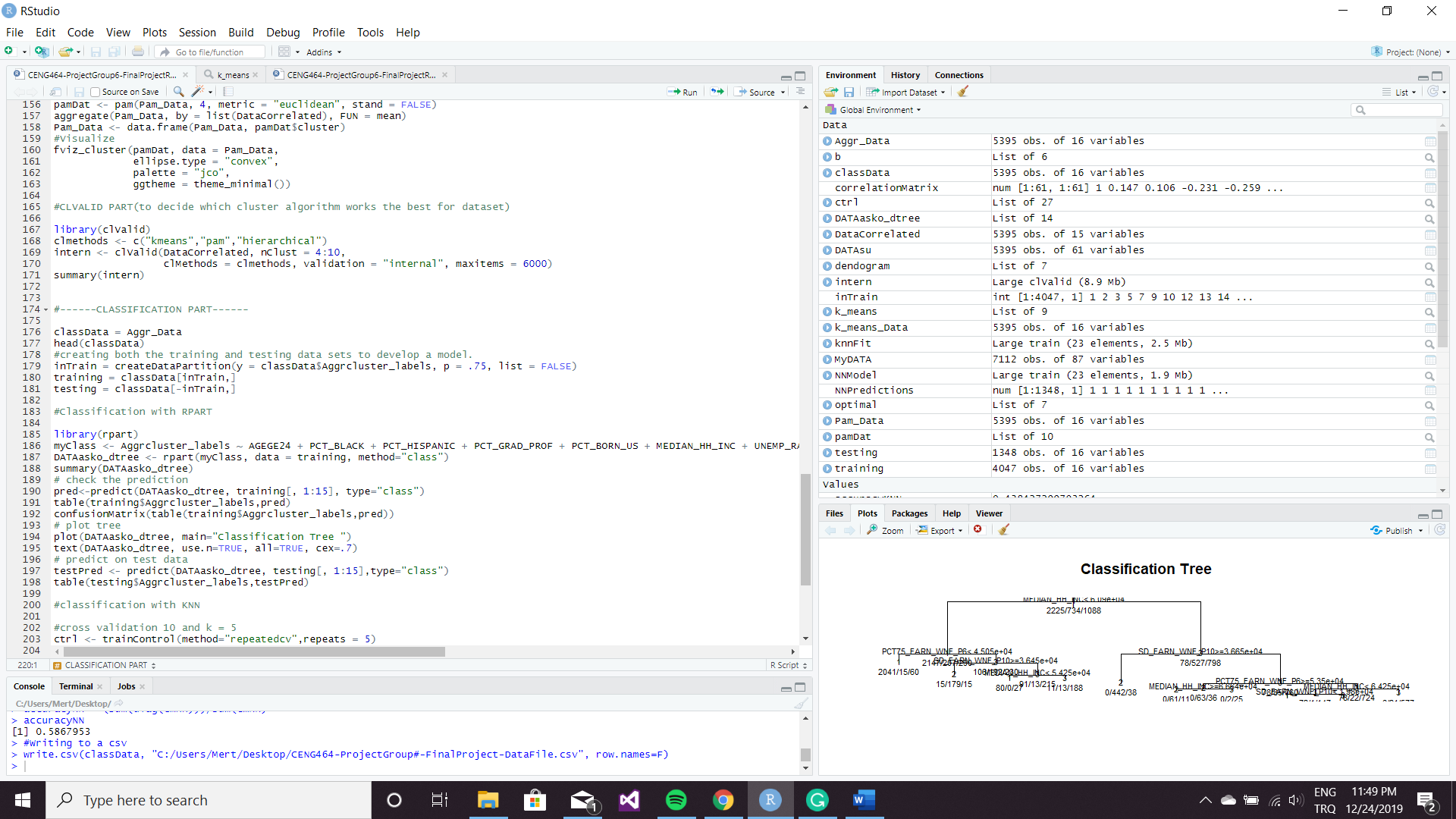


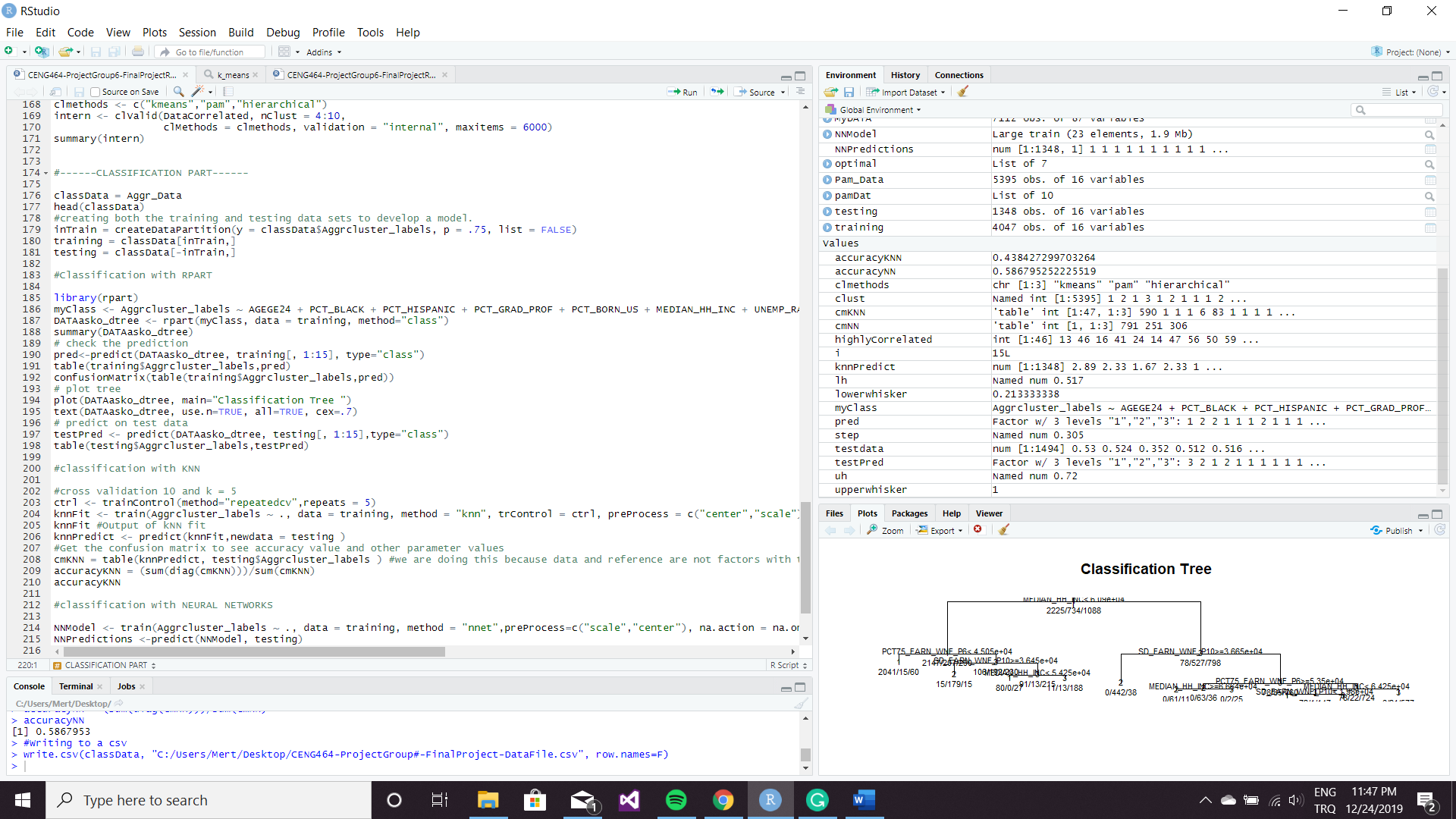
When the program finished learning, training set we tested it and got 0.5867953 accuracy.

Overall, Decision Tree’s accuracy is much higher compared with the other accuracy results.

To sum up, in this project, we preprocessed the data and made it less non-trivial data. After that, we used our clean data for getting valuable patterns. For this, we tested three different clustering method to grouping to our data and we compared which method is the best. When we have decided the best one is the Hierarchical Agglomerative clustering method, we used it to classify our data. We also used three different classification methods for it. Then we checked those methods accuracy’s and clearly see that decision tree method gave us the best result for this dataset.

Finally, this was our global environment part at the end of the program.





## REFERENCES

* <https://subscription.packtpub.com/book/big_data_and_business_intelligence/9781785286544/1/ch01lvl1sec11/data-preprocessing-techniques>
* <https://www.datanovia.com/en/lessons/choosing-the-best-clustering-algorithms/>
* <https://www.r-bloggers.com/k-means-clustering-in-r/>
* <https://www.statmethods.net/advstats/cluster.html>
* <https://www.datanovia.com/en/lessons/k-medoids-in-r-algorithm-and-practical-examples/>
* <https://en.proft.me/2017/01/22/classification-using-k-nearest-neighbors-r/>
* <https://rpubs.com/njvijay/16444>
* <https://www.youtube.com/watch?v=qLYYHWYr8xI>