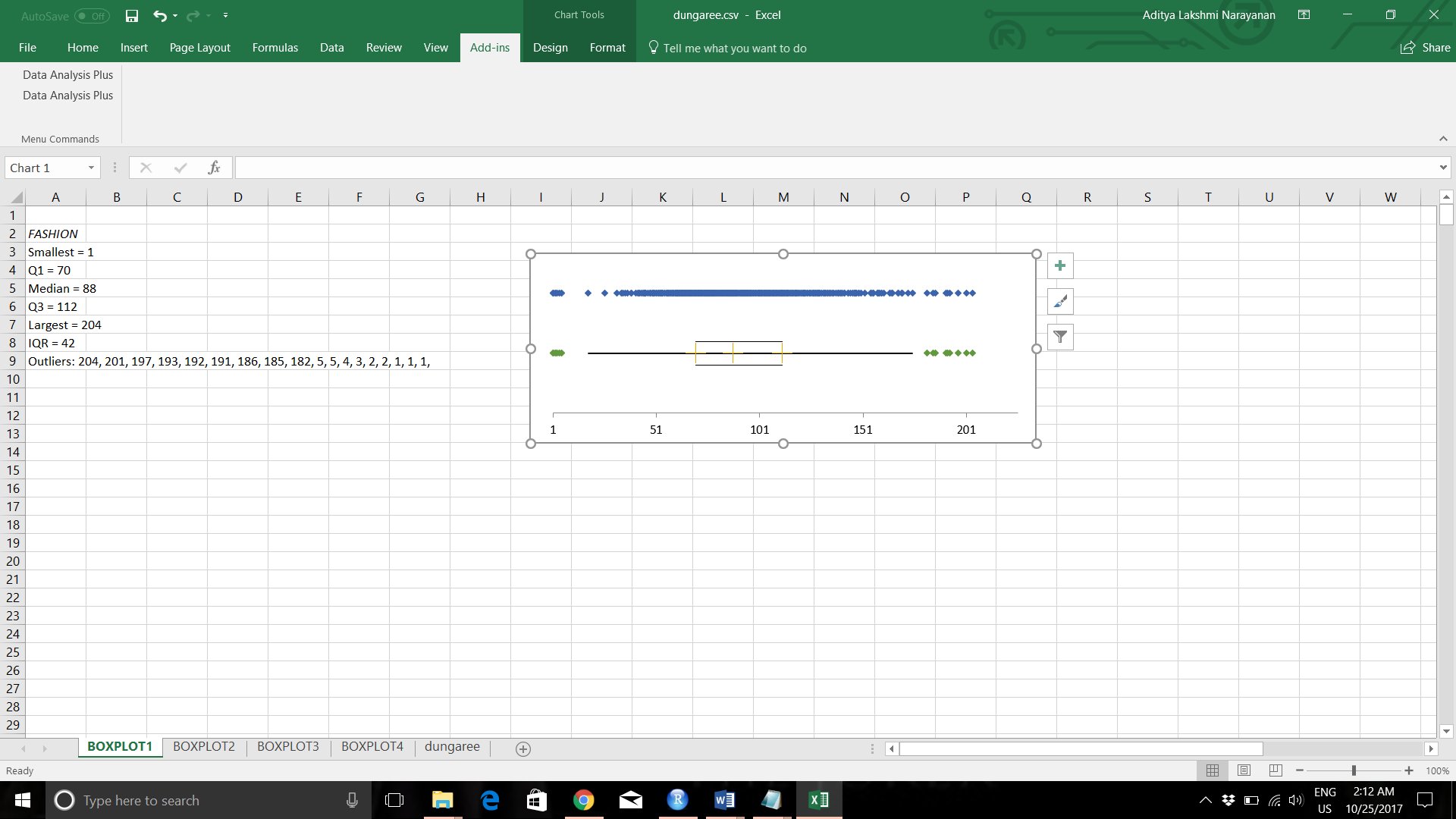
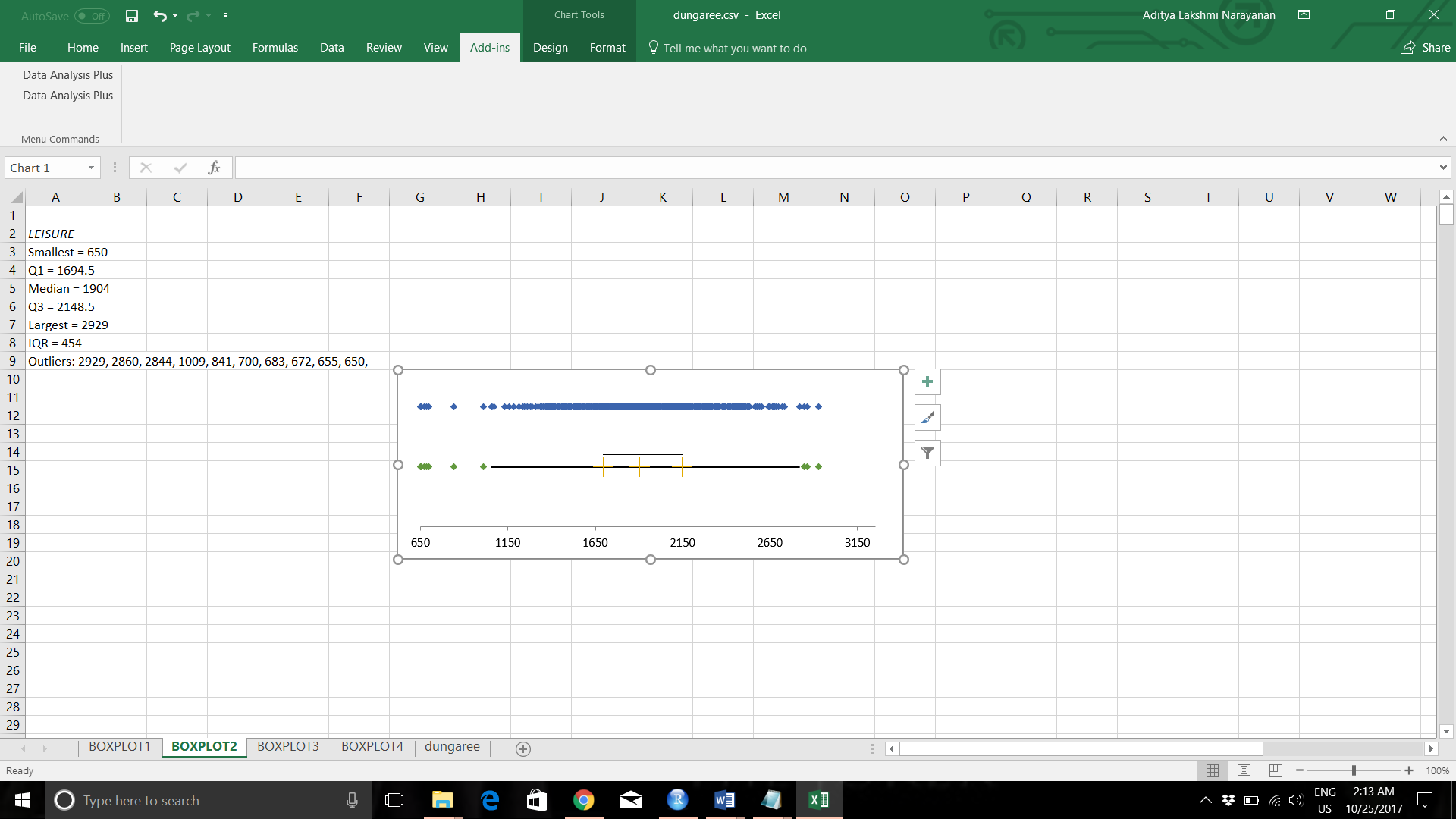
Exercise 1:

Unusual data values/outliers are identified using BoxPlot

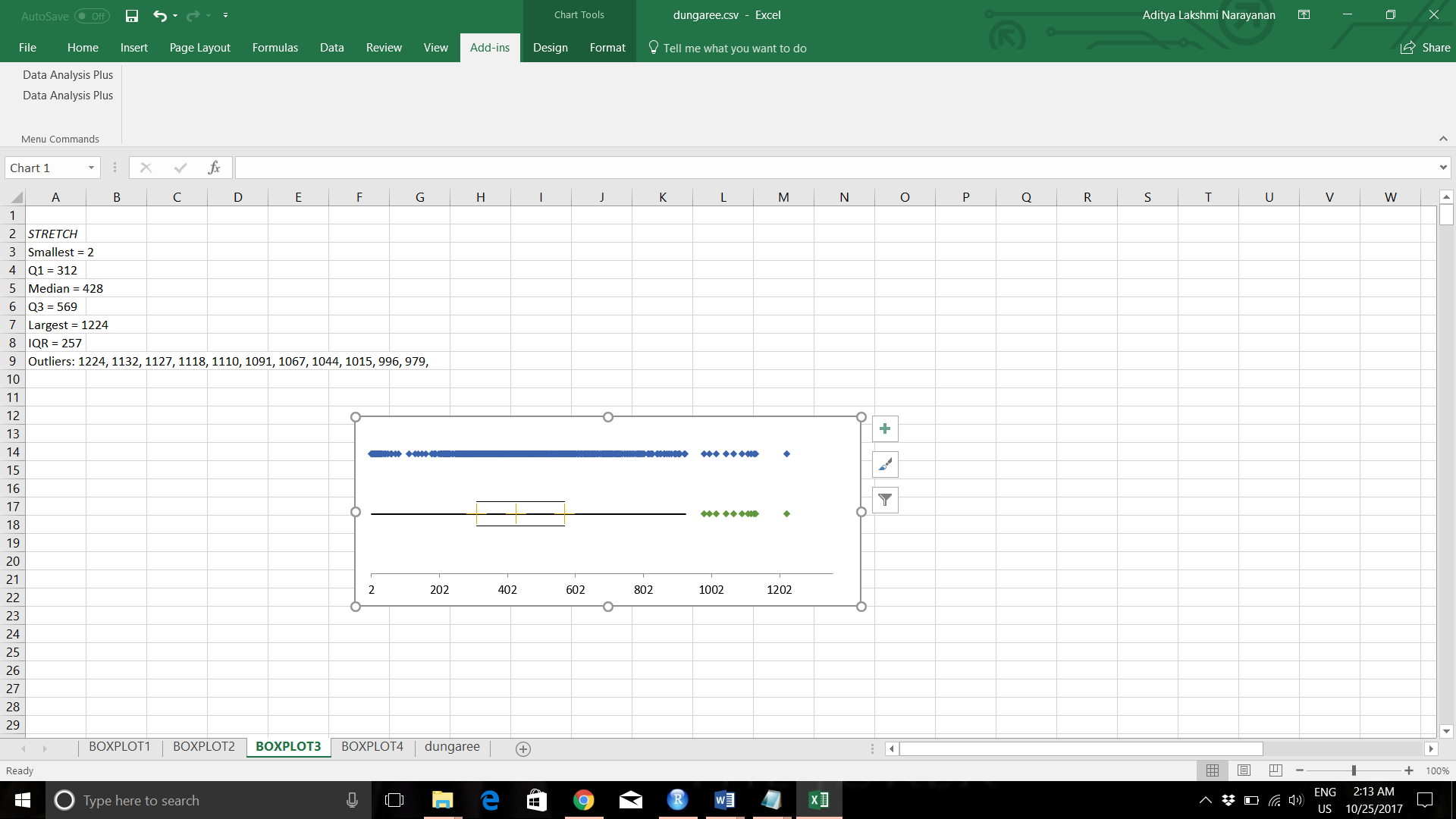
Fashion:



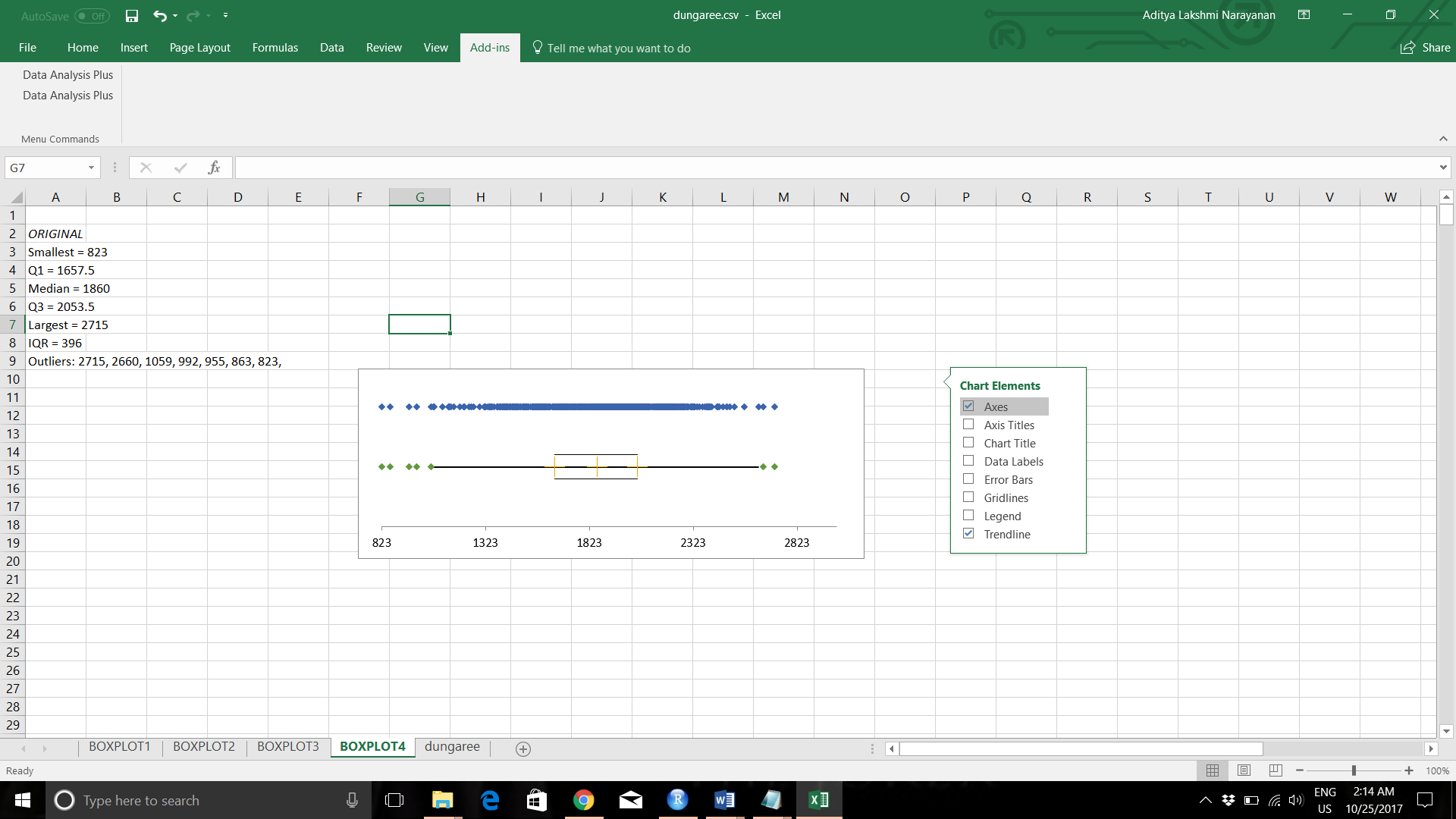
Leisure:



Stretch:



Original:



**There are no missing values in the dataset**

To eliminate the effect of outliers I winsorized the dataset wherein the top 5% and bottom 5% of the dataset are replaced by the next highest and lowest values respectively. This can be done by using **REALSTATS** add-in in excel and using the **WINSORIZE** function

Remove StoreID and SalesTot from the dataset. Normalize the data using **SAPPLY** function

Keeping max number of clusters as **10** and clustering using nbclust function, the following output is obtained:

Among all indices:

\* 5 proposed 2 as the best number of clusters

\* 6 proposed 3 as the best number of clusters

\* 2 proposed 4 as the best number of clusters

\* 4 proposed 5 as the best number of clusters

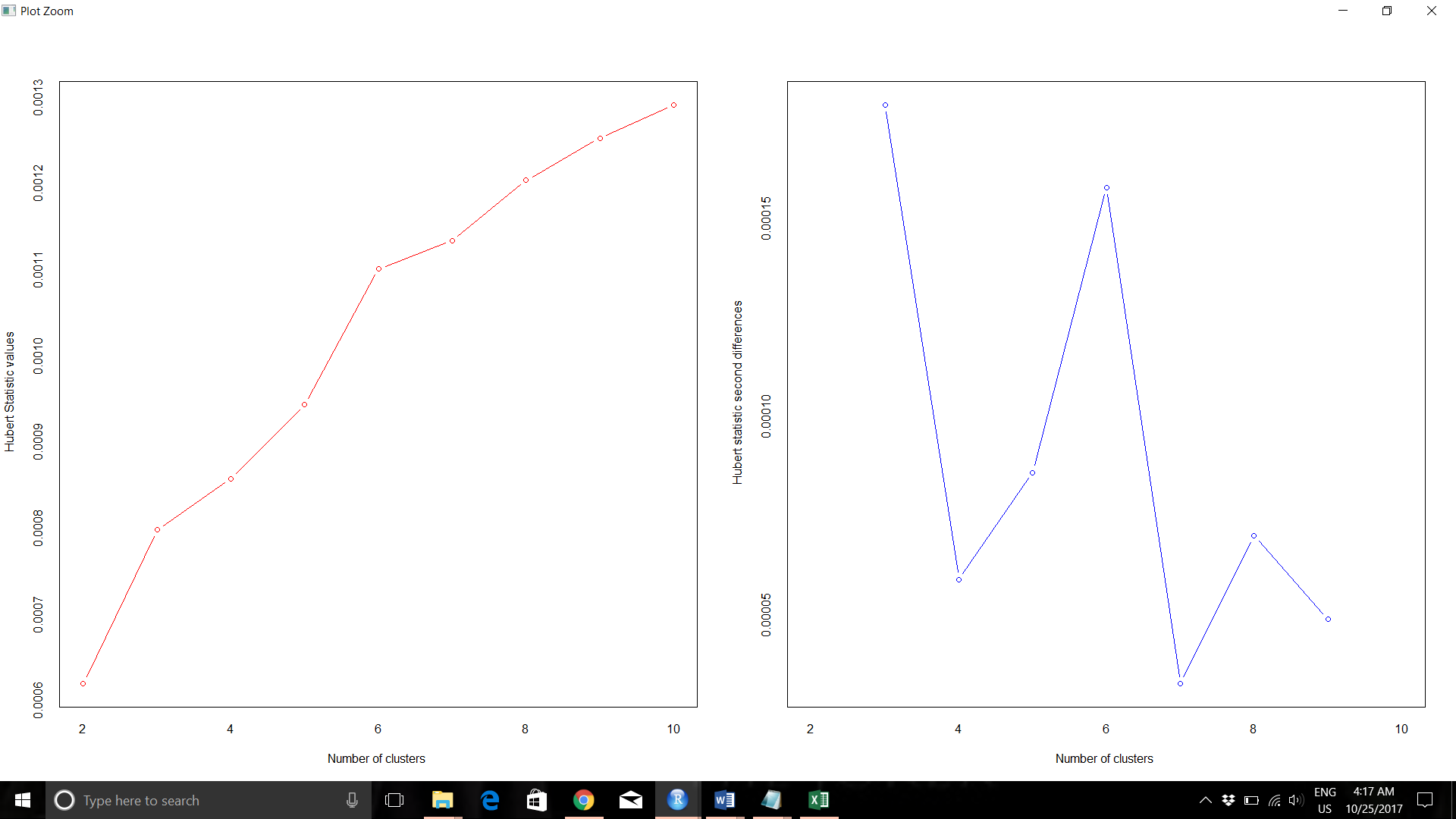
\* 2 proposed 6 as the best number of clusters

\* 2 proposed 8 as the best number of clusters

\* 2 proposed 9 as the best number of clusters

\*\*\*\*\* Conclusion \*\*\*\*\*

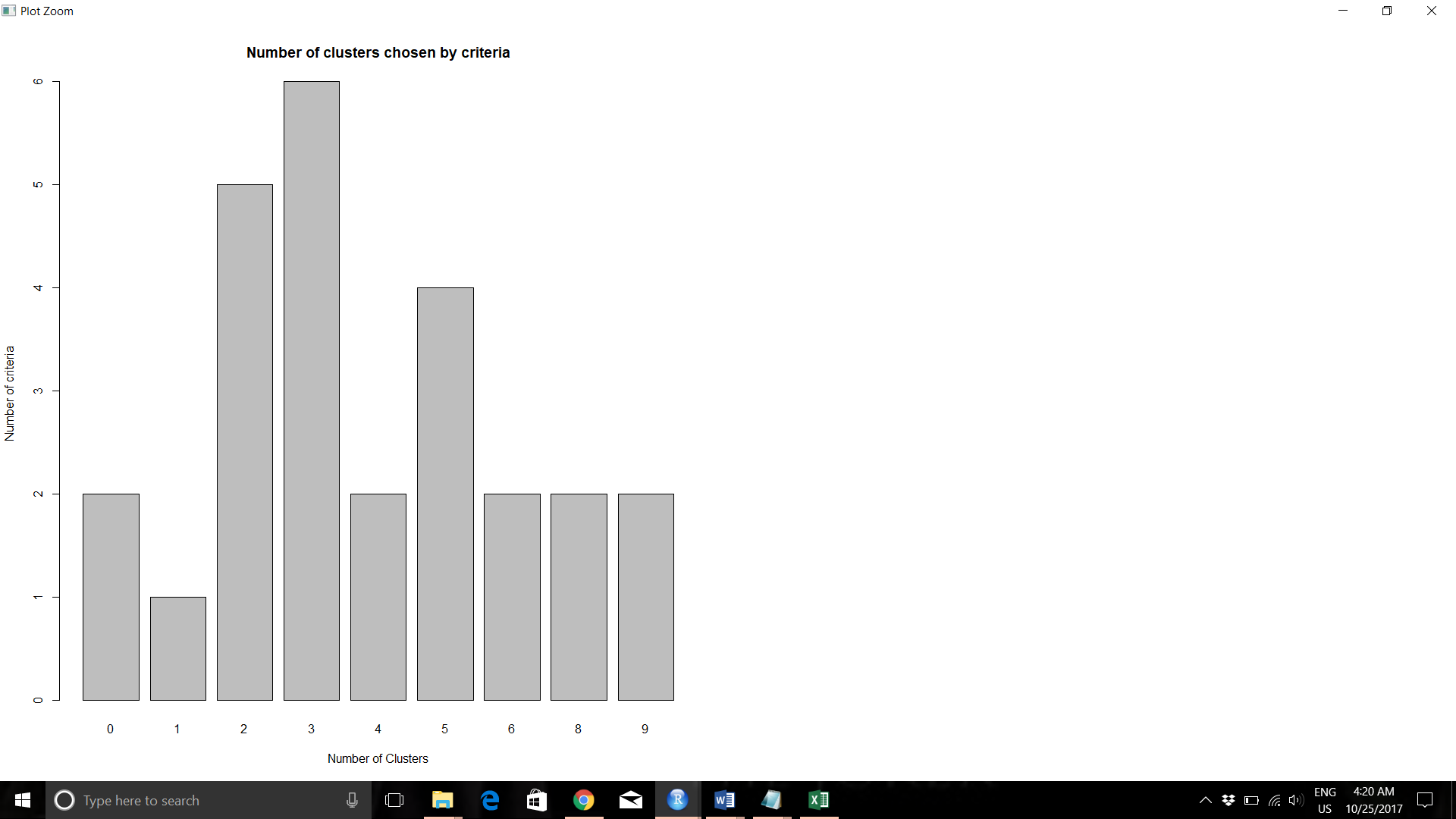
\* According to the majority rule, the best number of clusters is 3



From the left side graph we can see that the slope of graph significantly decreases at the value of 3 which indicates the best number of clusters.

From the second differences graph on the right side we can see the peak is highest when number of clusters is 3.

The bar plot gives the number of clusters chosen by number of criteria



As we can see from the graph, 3 is proposed as the best number of clusters.

|  |
| --- |
| fit.km$centers  FASHION LEISURE STRETCH ORIGINAL  1 1.1786271 -0.06071119 -0.2836218 0.2783479  2 -0.4608927 0.72599440 -0.4002778 -0.8210830  3 -0.3653970 -0.79997639 0.6839491 0.7441452 |
|  |
| |  | | --- | |  | |

On keeping the max size as 10 we are able to drill down to 3 clusters which are fairly different from each other.

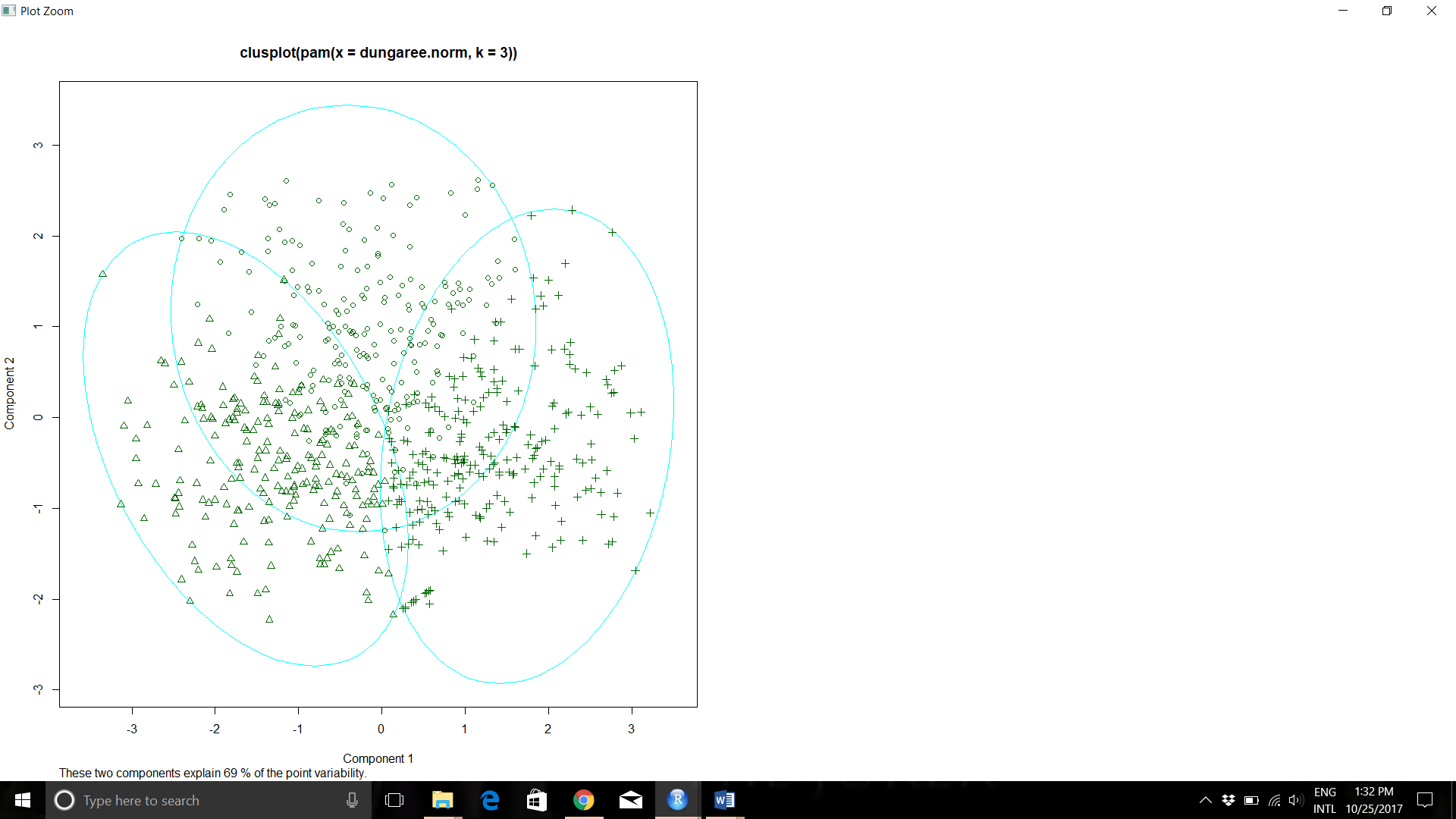
Stores belonging to cluster 1 are more focused on selling jeans with ‘Fashion’ style as indicated by the positive high value of 1.178 and are least when it comes to selling ‘Stretch’ style jeans indicated by a negative value of -0.283

Stores coming under cluster 2 sell more of ‘Leisure’ type jeans than any other type.

Stores coming under cluster 3 sell more of Óriginal’ and ‘Stretch’ type of jeans compared to other stores.

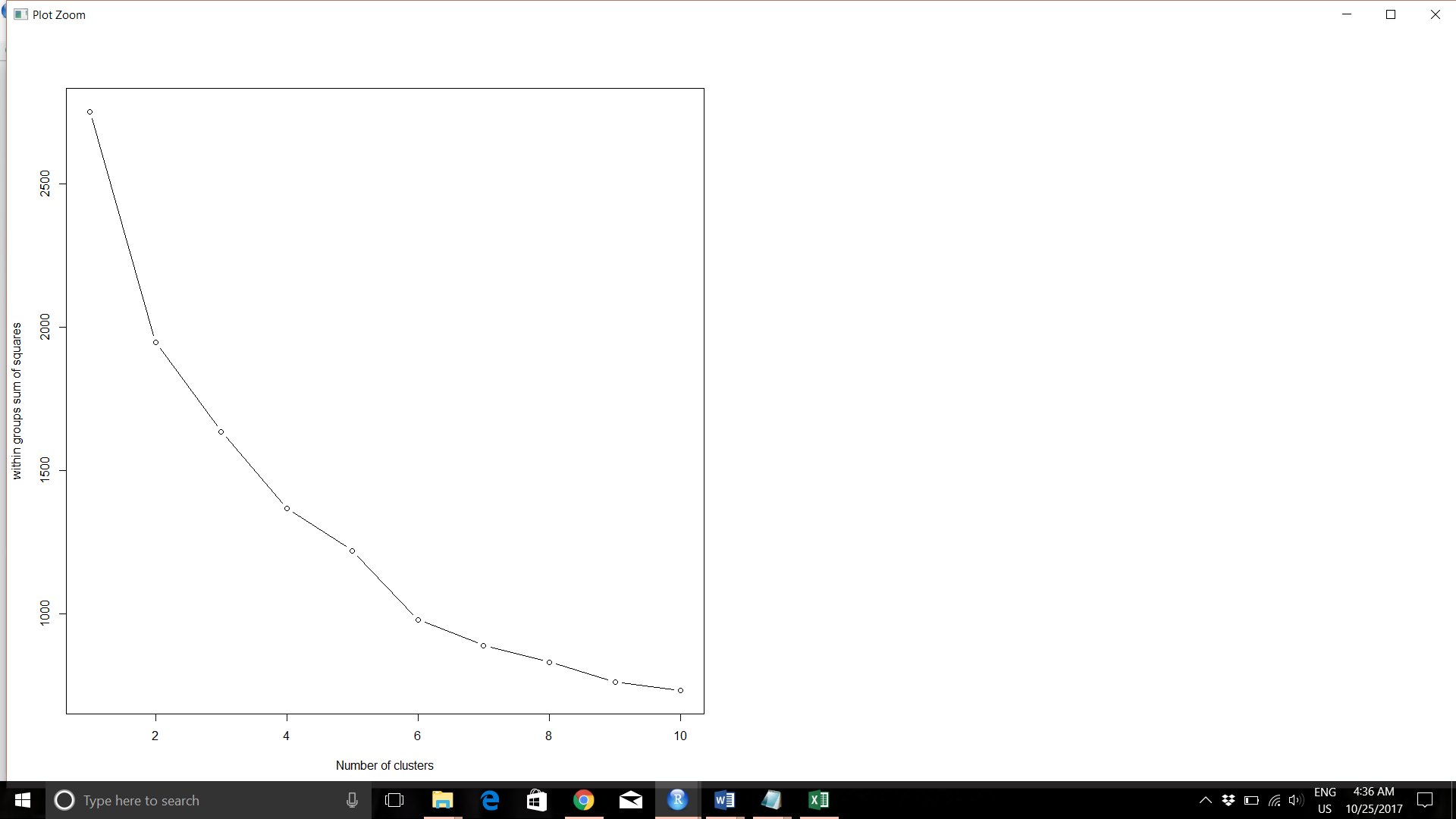
As we can see from the above 3 clusters, each cluster of stores is doing more sales on a particular type of jean which are different from each other.

Cluster Plot:



WSSPLOT with 10 clusters:

The wssplot run with 10 clusters indicates that 5 is the best number of clusters. This is determined by the “ELBOW CRITERION” where the graph tends to take a sharp drop



Keeping max number of clusters as **6** and clustering using nbclust function, the following output is obtained:

\* Among all indices:

\* 5 proposed 2 as the best number of clusters

\* 6 proposed 3 as the best number of clusters

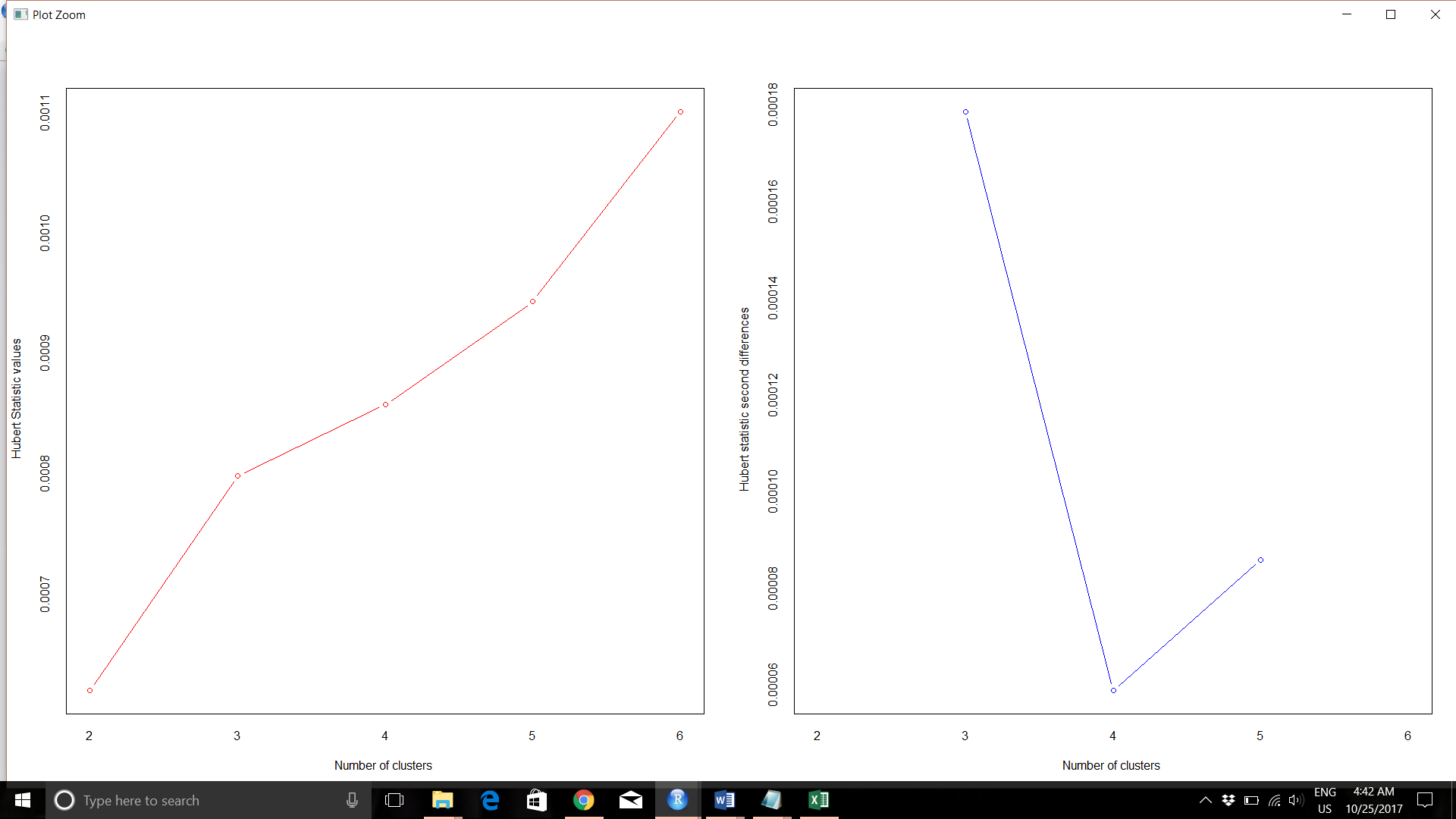
\* 2 proposed 4 as the best number of clusters

\* 7 proposed 5 as the best number of clusters

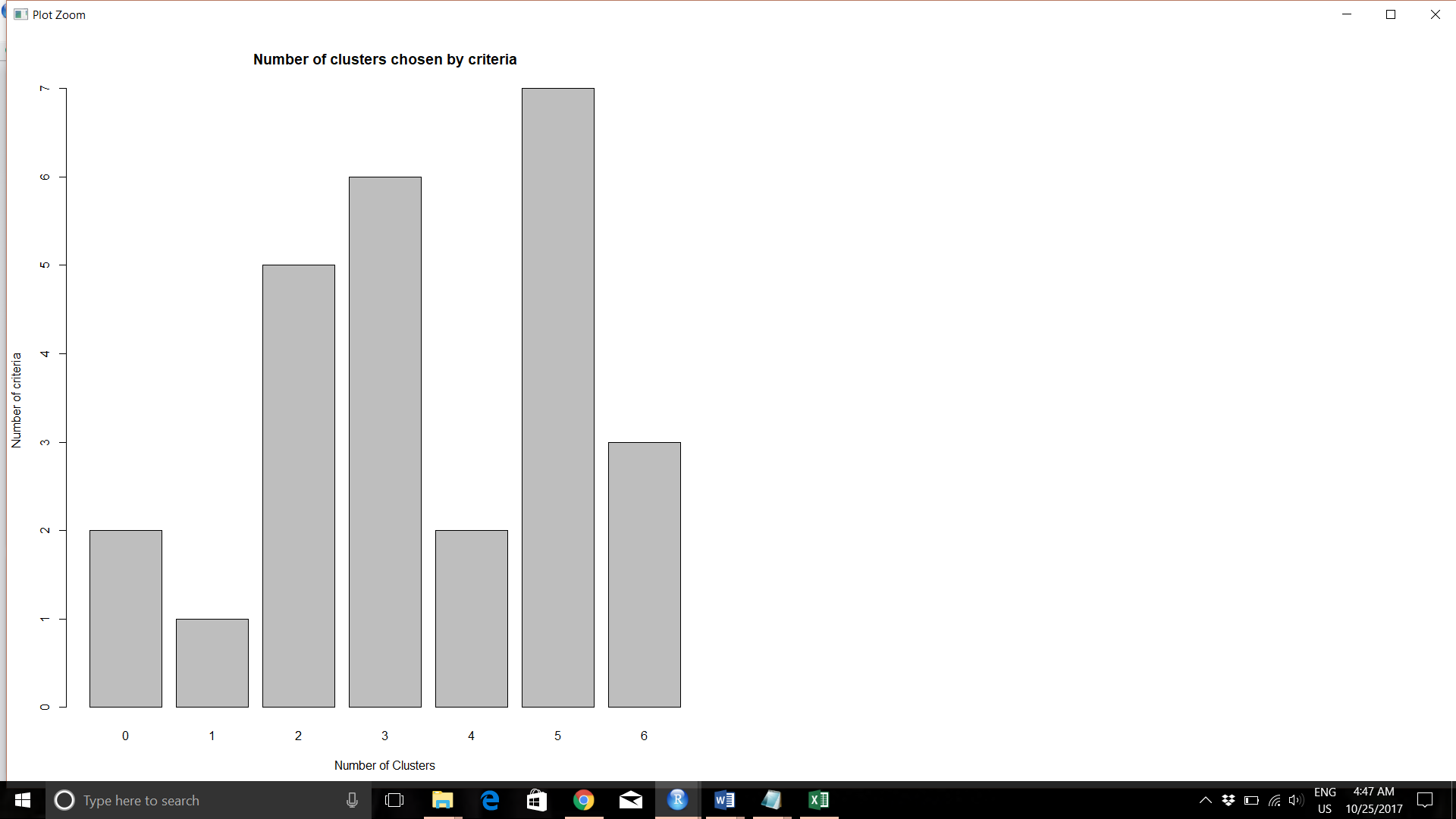
\* 3 proposed 6 as the best number of clusters

\*\*\*\*\* Conclusion \*\*\*\*\*

\* According to the majority rule, the best number of clusters is 5



However, the best number of clusters suggested is 3 as indicated by the Hubert second differences graph.



As we can see from the graph, 5 is proposed as the best number of clusters

fit.km1$centers

FASHION LEISURE STRETCH ORIGINAL

1 -0.3025592 -0.9156986 1.43300527 0.3562418

2 1.4240113 0.0478766 -0.02881116 0.1328004

3 -0.3274328 1.1392541 -0.17621988 -0.8519487

4 -0.2836576 -0.3719884 -0.25814223 0.9834344

5 -0.6201137 -0.9124261 -1.77923389 -1.1555070

Cluster 1 of stores are more inclined towards sales of Stretch and Original type of jeans.

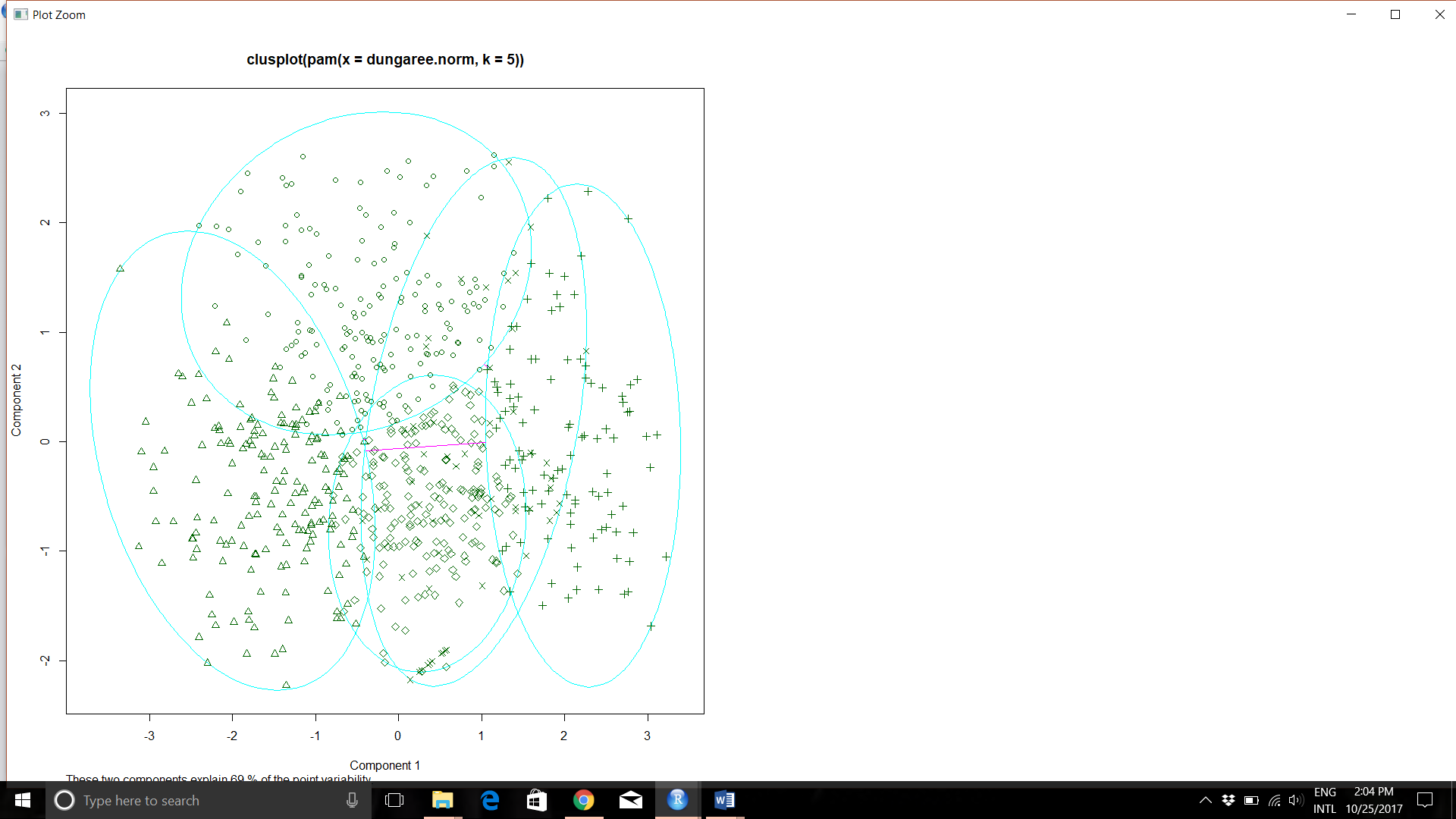
Cluster 2 of stores are more inclined towards sales of Fashion type of jeans while maintaining an almost neutral stance on Leisure and Original type of jeans.

Cluster 3 of stores are focused mainly on sales of leisure type compared to any other type of jeans.

Cluster 4 of stores are focused mainly on sales of Original type compared to any other type of jeans

Cluster 5 gives an impression that all types of jeans have less sales compared to other stores but it could also mean from the values that all types of jeans all types of jeans are sold at an equal footage at these stores. It could also mean that sales generally is low in these stores and they could be termed as non-performing.

Cluster Plot:



Based on running 2 sets with max size as 10 and max size as 6, I prefer to take the algorithm with max size of 10 as it clearly defines 3 clusters with the different types of jeans sold at a specific cluster of stores. Also, in the first cluster, the Hubert value matches with the number of clusters provided with the majority rule.

The clear demarcation in the cluster of stores and the type of jeans sold in those 3 clusters will enable us to make clear decisions related to sales. Each store can be targeted in the specific type where sales are maximum leading to better efficiency.

Another line of thought with 5 clusters can be where non-performing stores or stores where all types of jeans are sold almost equally(cluster 5) can be revamped in such a way which leads to increase in all types of jeans or can be closed down to focus more on other stores.

**Exercise 2**

Keeping the maximum number of clusters as **4** in the nbclust function, the following results are obtained:

\* Among all indices:

\* 4 proposed 2 as the best number of clusters

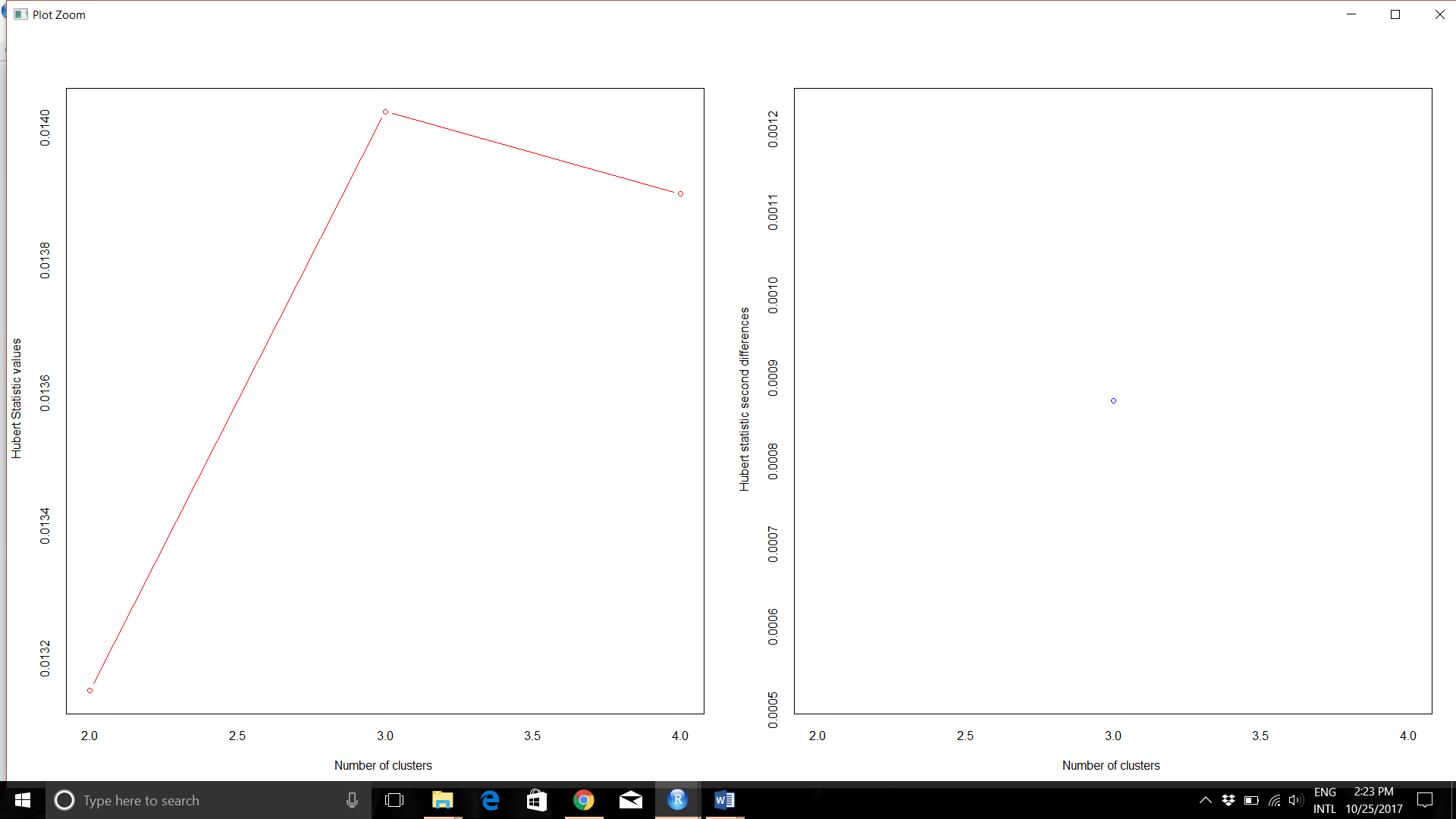
\* 14 proposed 3 as the best number of clusters

\* 5 proposed 4 as the best number of clusters

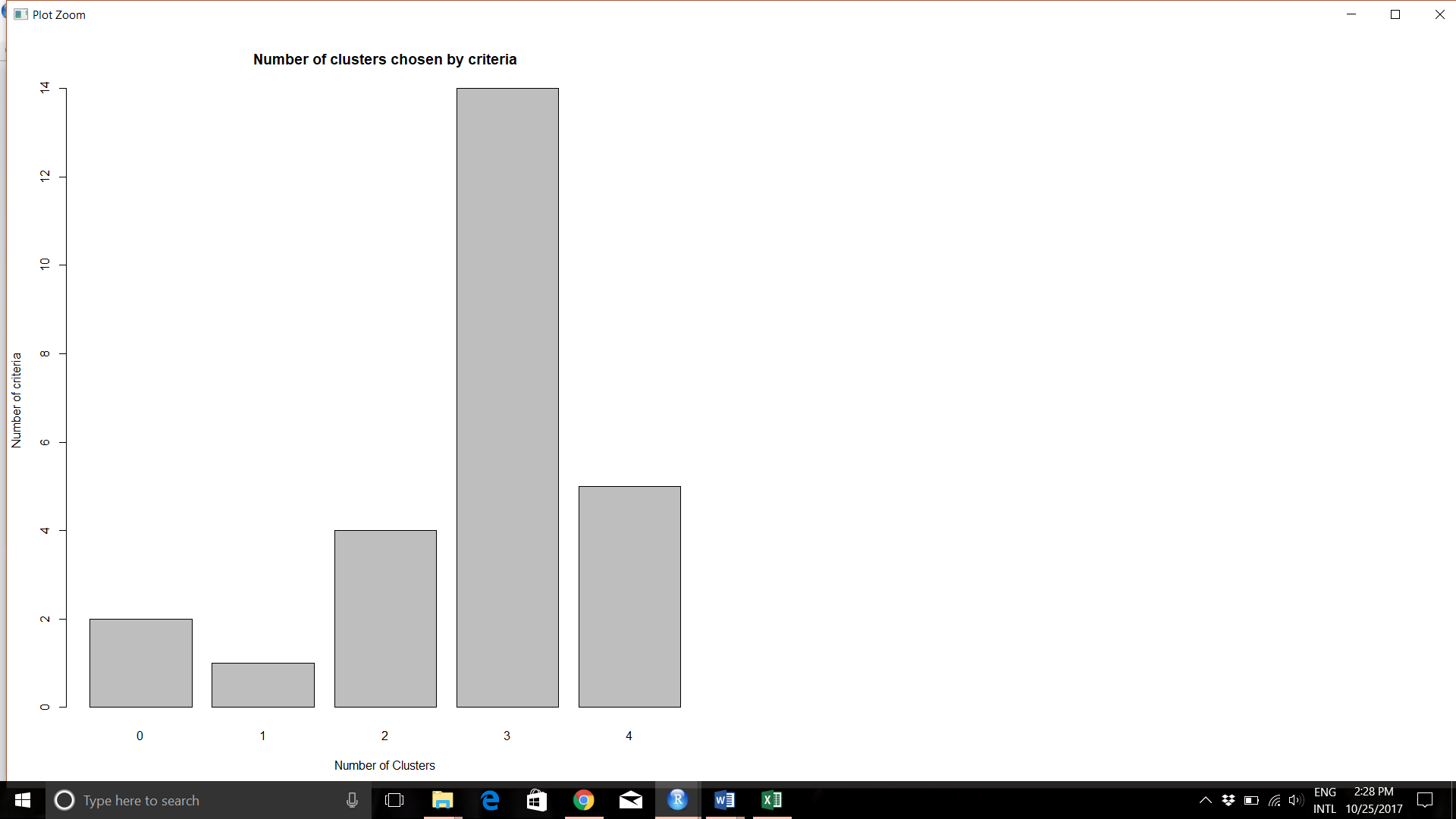
\*\*\*\*\* Conclusion \*\*\*\*\*

\* According to the majority rule, the best number of clusters is 3

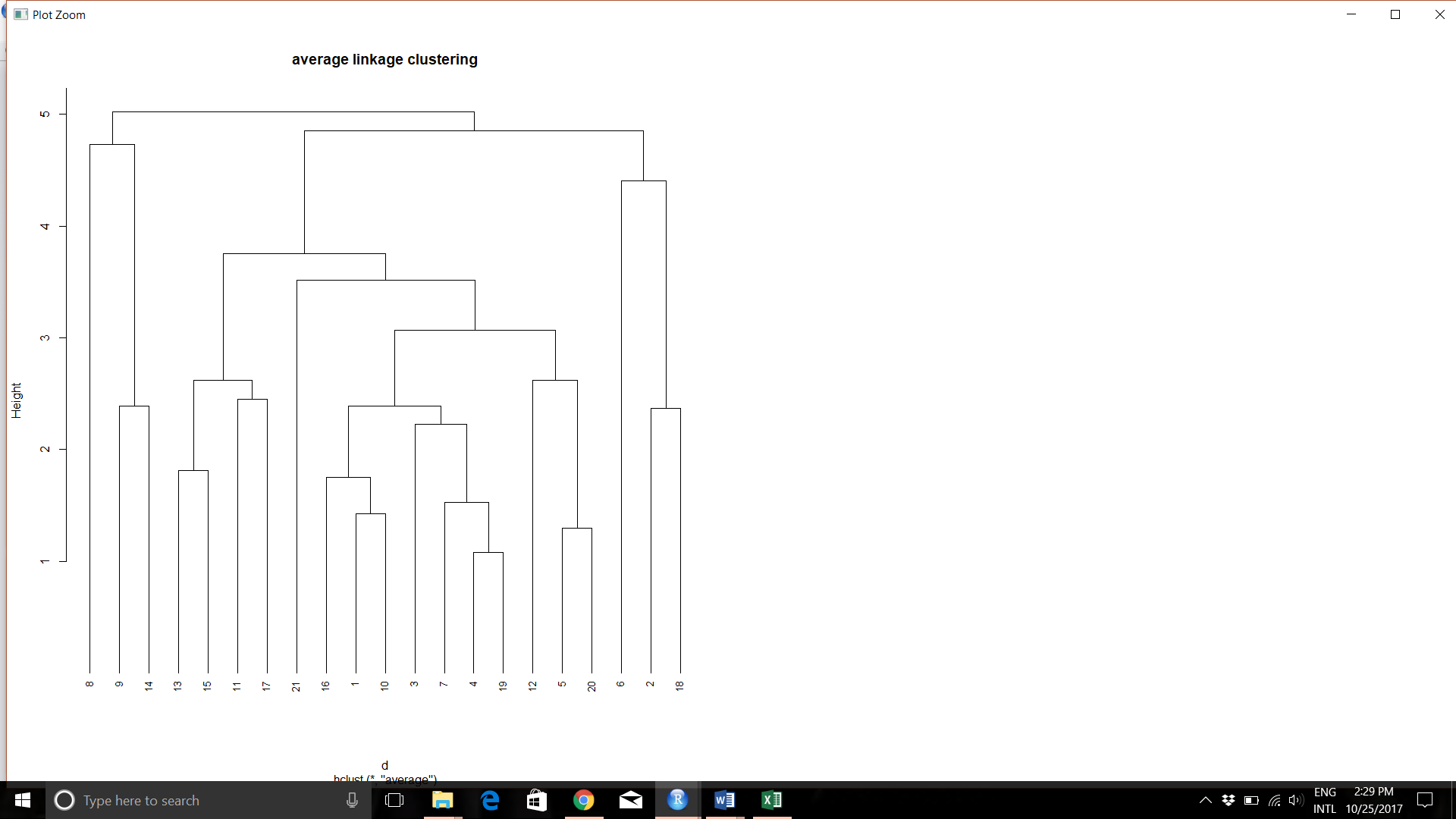
We get the same number with the Hubert graph as well



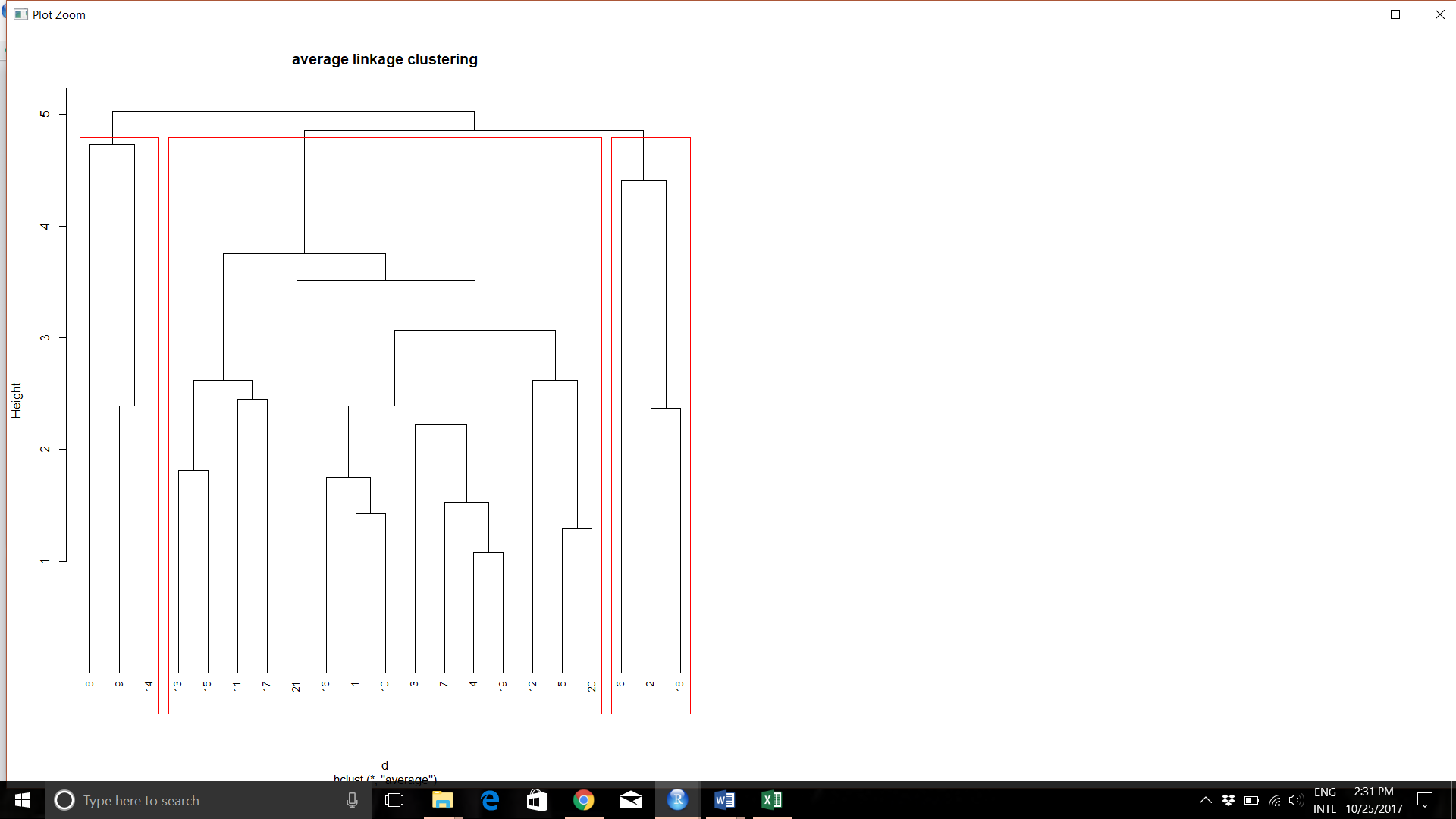
Bar Plot



Dendogram for the data



Since we have already deduced that the best number of clusters is 3 previously from the nbclust function, we can go ahead and depict it in the dendogram



The 3 clusters are indicated by the red borderline surrounding the values

Cluster 1- 8,9,14

Cluster 2- 13, 15, 11, 17, 21, 16, 1, 10, 3, 7, 4, 19, 12, 5, 20

Cluster 3- 6, 2, 18

The median is described by the Aggregate function as follows:

cluster Market\_Cap Beta PE\_Ratio ROE ROA Asset\_Turnover Leverage Rev\_Growth Net\_Profit\_Margin

1 1 0.1702742 -0.2559560 -0.32874435 0.1063815 0.5422849 0.4612656 -0.3912841 -0.3621317 0.3359869

2 2 -0.6953818 -0.4507051 1.90298017 -0.8548399 -0.9422871 -0.4612656 -0.3016910 -0.3811391 -1.5536671

3 3 -0.9704532 1.2630872 0.03299122 -0.7089994 -1.0174553 -1.8450624 0.6198379 1.5386072 -0.3650138

Cluster 1- Leader Companies

From the cluster 1 data, we can see that the companies have a high market cap value compared to other clusters. This is an indication that the company belongs to the big league. The ROA and ROE values are high but the revenue growth is not exactly represented by that. This does not mean that the company is doing bad as major companies tend to have a stable revenue growth. The good performance of these companies is also indicated by the high profit margins.

The only drawback of these companies is a lower leverage as big companies use their liabilities well to increase their profits and efficiency.

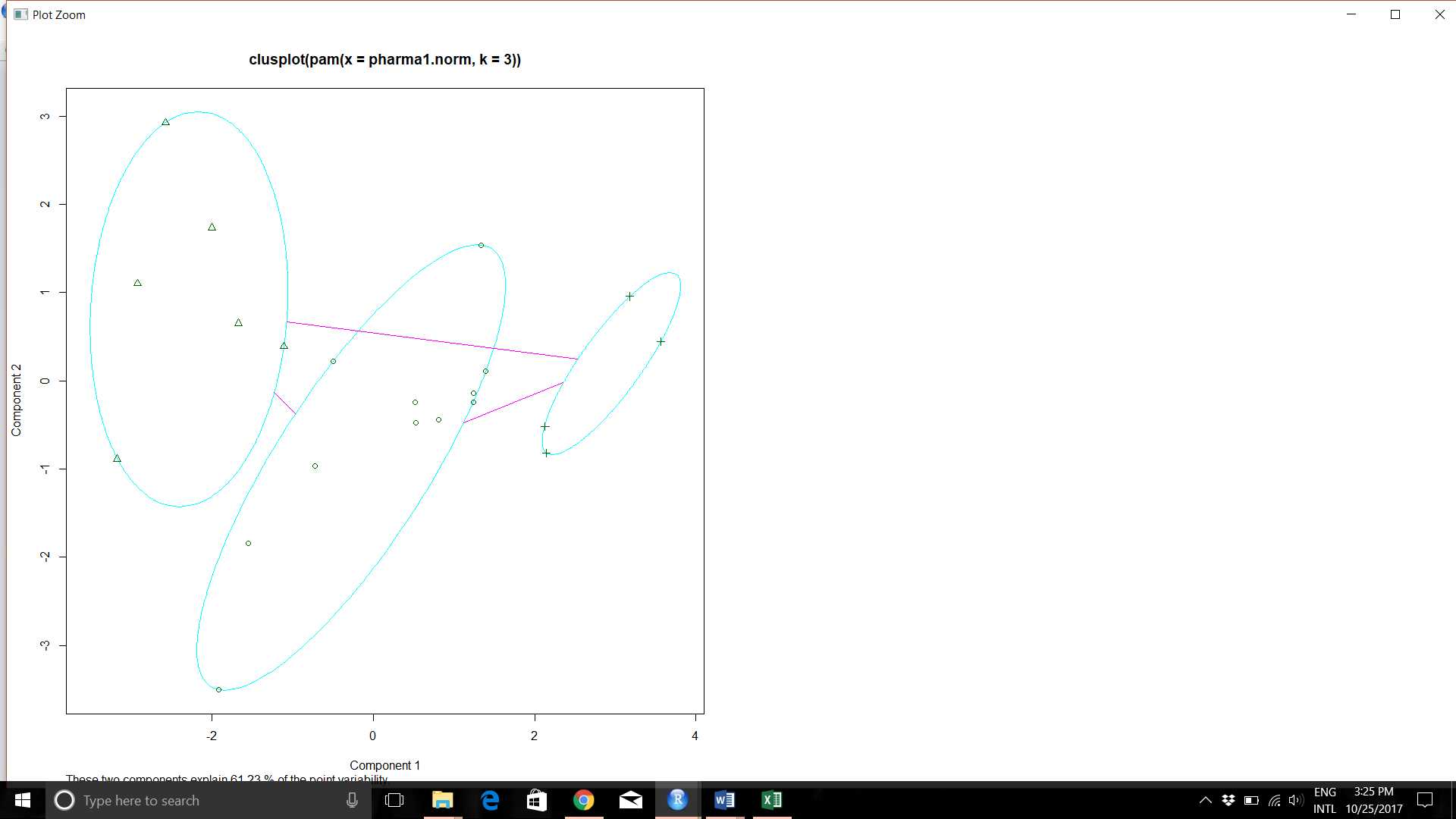
Cluster 2- Struggling Mid-tier Companies

The marketcap, ROA, ROE, Revenue growth and profit margin are found to be in the lower ranges. The PE ratio is also very high which mostly indicates that the market price relatively stays same or does not increase to effect in increase in earnings per share

Cluster 3- Start-Up Companies

The high beta value coupled with the low market cap given an idea that these companies might be in the founding stages. Beta is an indication of volatility which is very much with the case of start-up companies. The ROA and ROE for these companies are usually low as the companies have not grown big enough yet but the high leverage and revenue growth values indicate that these companies are going in the right direction.

Cluster Plot:



k-means for pharmaceuticals set:

I have run nbclust function with maximum number of clusters as 4 and obtained the following results:

\* Among all indices:

\* 12 proposed 2 as the best number of clusters

\* 8 proposed 3 as the best number of clusters

\* 3 proposed 4 as the best number of clusters

\*\*\*\*\* Conclusion \*\*\*\*\*

\* According to the majority rule, the best number of clusters is 2

fit.km$centers

Market\_Cap Beta PE\_Ratio ROE ROA Asset\_Turnover Leverage Rev\_Growth Net\_Profit\_Margin

1 0.6733825 -0.3586419 -0.2763512 0.6565978 0.8344159 0.4612656 -0.3331068 -0.2902163 0.6823310

2 -0.7407208 0.3945061 0.3039863 -0.7222576 -0.9178575 -0.5073922 0.3664175 0.3192379 -0.7505641

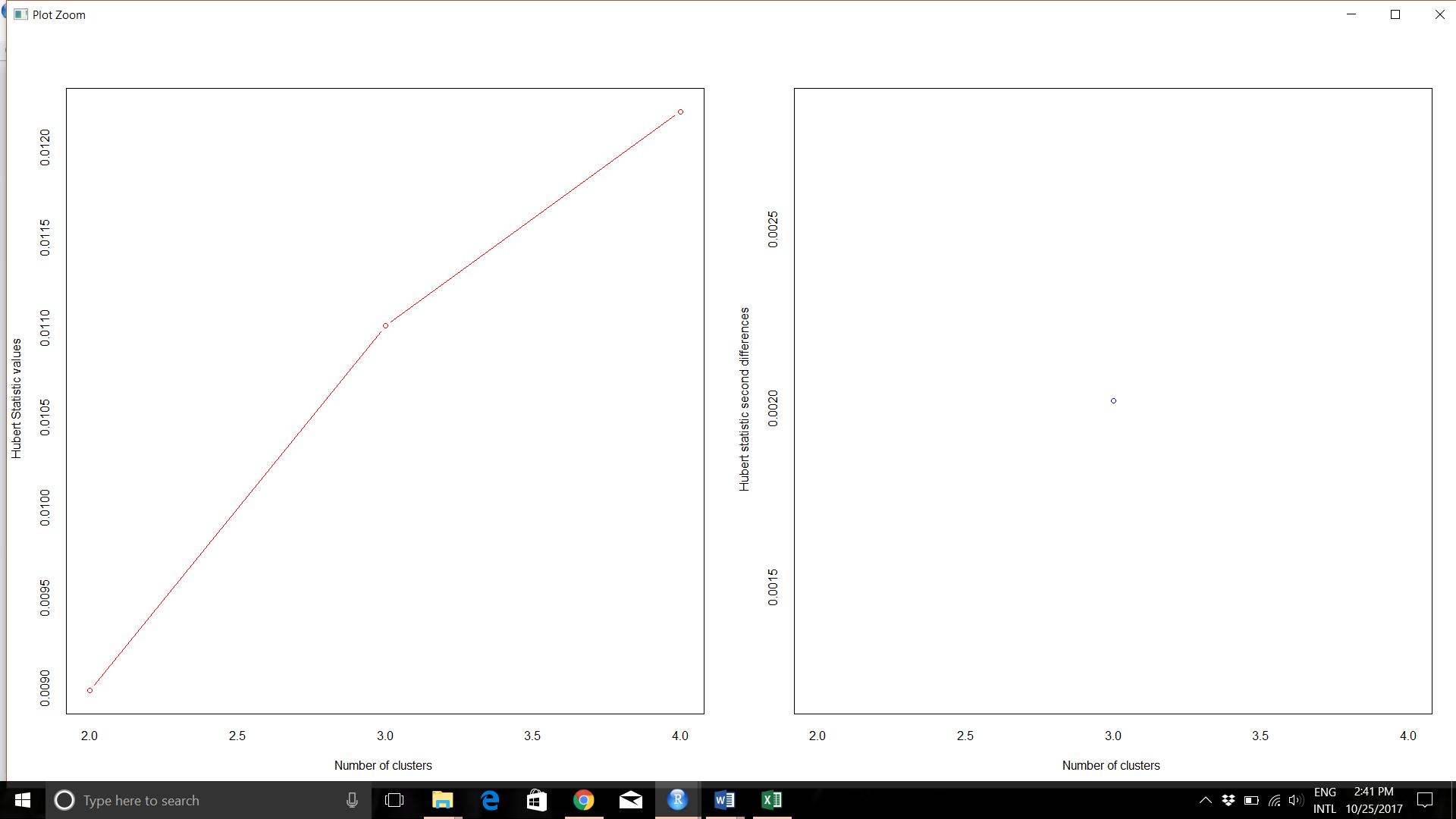
This basically splits the companies into 2 CLUSTERS:

If we notice cluster 1, the signs of good performance is shown in factors such as MarketCap, ROE, ROA, Low Beta and high profit margin. But they also have low revenue and leverage which is misleading.

The vice versa is observed with cluster 2 companies.

**A more detailed analysis can be obtained with 3 clusters as indicated above**

Also, Hubert graph shows 3 to be best number of clusters



Because of this anamoly, I have gone ahead with the heirarchial clustering result which proposes 3 clusters.

Also, reducing the number of clusters to very minimum range is not advisable as sometimes no inferences can be made from such clustering. In the 3 clusters above, we saw that it was possible to clearly differentiate the 3 groups with the relevant data at hand.

Cluster plot for 2 clusters:

