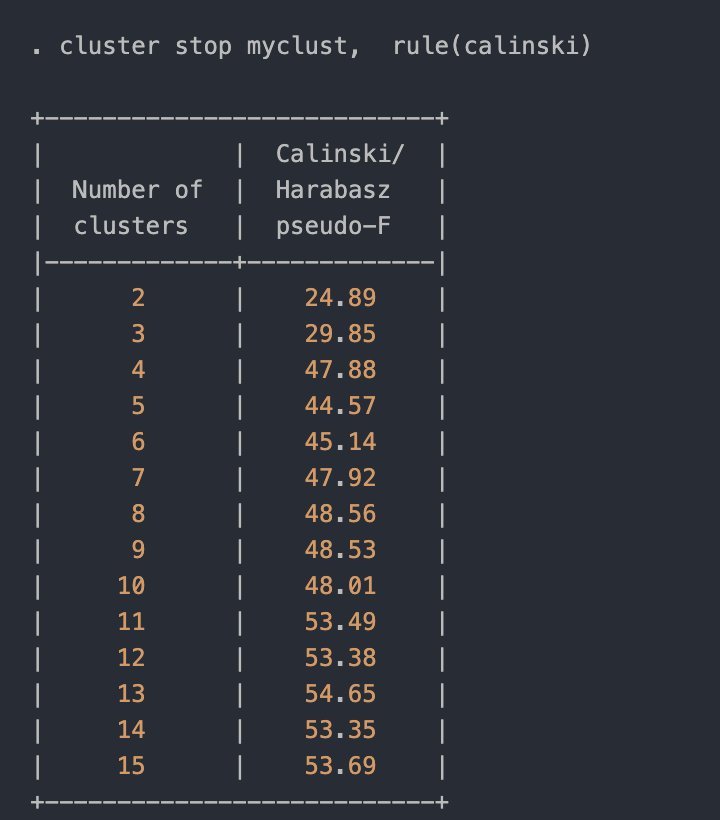
Course: Data Analysis

Student’s Name and Surname: Pavel Drankov

Please do the tasks below.

1. Open **Worldbank\_data.dta** file.

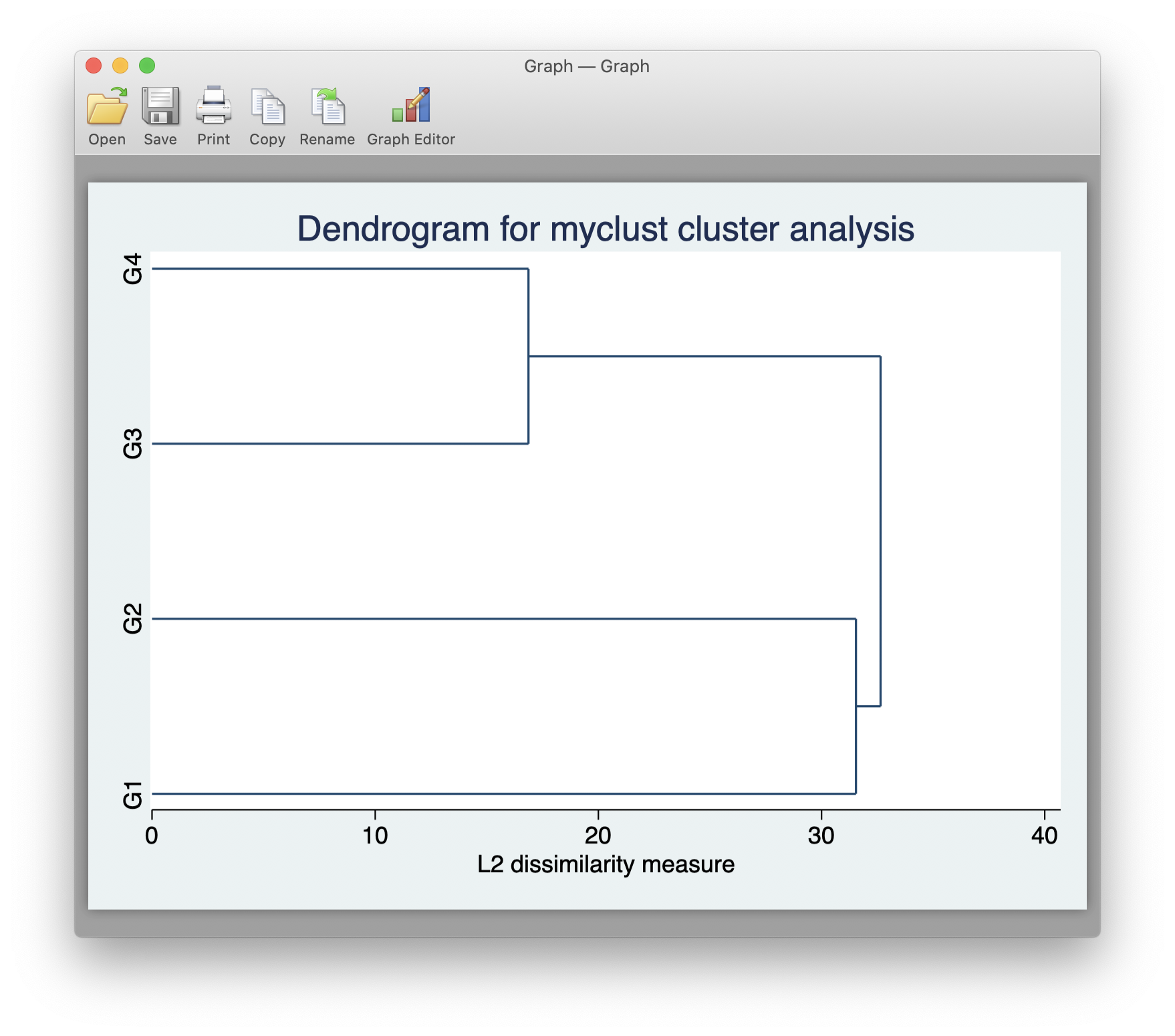
Do the cluster analysis using hierarchical clustering algorithm. Select not less than 4 interval variables for the analysis. Don’t forget to standardize the values of variables if needed.

1. Define the number of clusters. Explain your decision.

I chose 4, because Calinski/Harabasz index is decently hight in this point and at 5 cluster index is lower.

b. Create a dendrogram.

cluster dendrogram myclust, horizontal cutnumber(4)



c. Describe the clusters using descriptive statistics.

G1 - No computer communications import/export, extremely low amount of internet users.  
G2 - noticeable percentage of people use internet, medium import/export.  
G3 - Decent internet, import and export level are high.  
G4 - One of the best import and export level, high internet.

d. Try to name the obtained clusters.

G1 - Poor, non-developed countries.  
G2 - Underdeveloped countries  
G3 - Well developed country  
G4 - Huge economy & decent technology wise developed country. China is the only country in the cluster

f. Save cluster membership variable in the file.

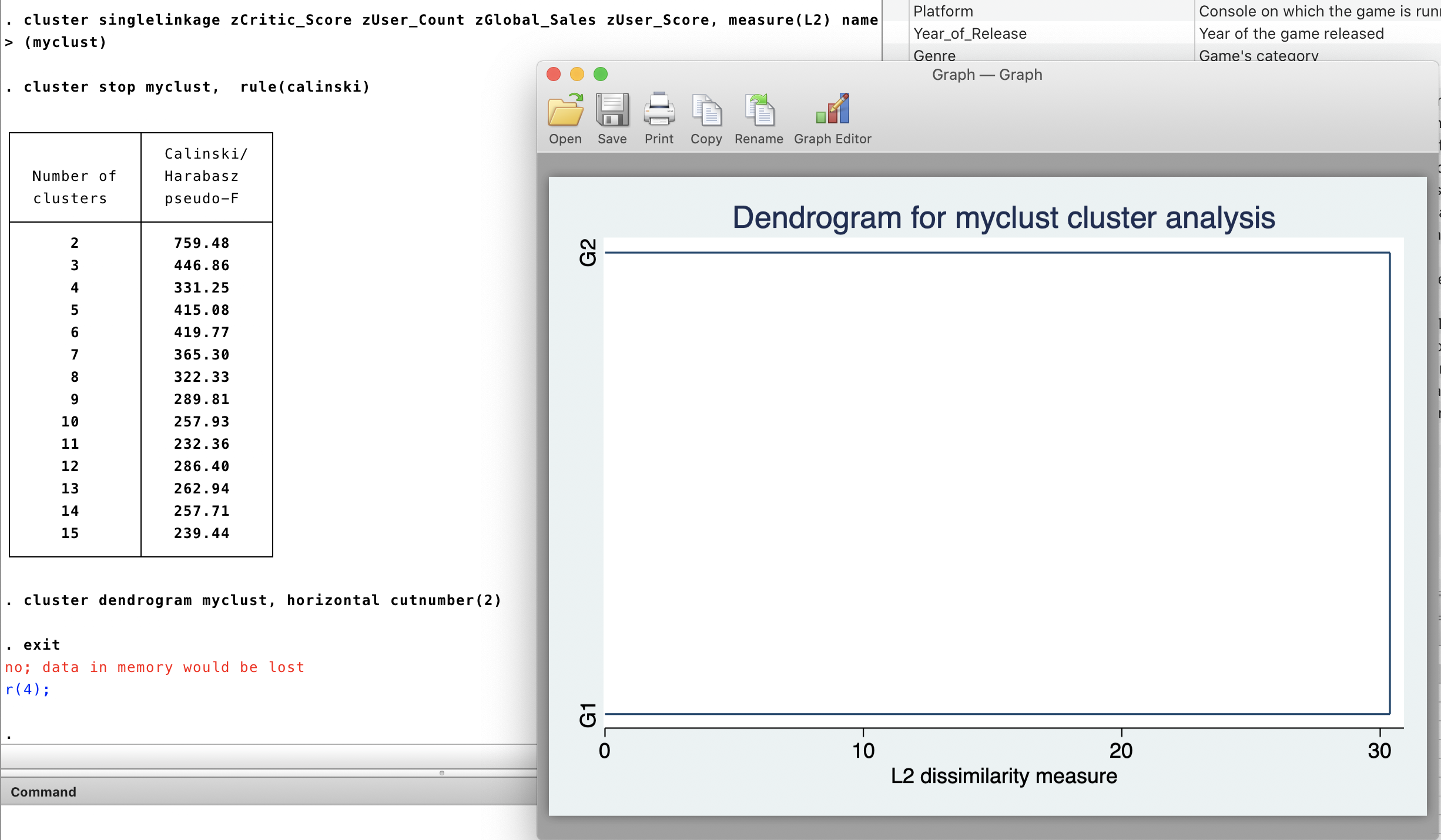
save Worldbank\_data\_with\_clusters.dta, replace

2. Open **video\_games\_sales.dta** file.

Do the cluster analysis using k-means algorithm. Select not less than 4 interval variables for the analysis. Don’t forget to standardize the values of variables if needed.

1. Define the number of clusters. Explain your decision.

With 2 clusters CH index is the highest.

2. Create a dendrogram.

3. Describe the clusters using descriptive statistics.

G1 - Below average user evaluation scores

G2 - Above average user evaluation scores

4. Try to name the obtained clusters.

G1 - Bad games

G2 - Best games

5. Save cluster membership variable in the file.

save video\_games\_sales\_with\_clusters.dta, replace

Send this file with the Stata output or do-file to amelikyan@hse.ru