**Data Analysis**

**Task 4**

**Part 1**

The file Worldbank\_tech.xlsx contains World Bank data for the year 2018.

Variables:

* country\_name: Country Name
* country\_code: Country Code
* internet: Internet users (per 100 people)
* secure: Secure Internet servers (per 1 million people)
* export: Communications, computer, etc. (% of service exports)
* import: Communications, computer, etc. (% of service imports)
* telephone: Fixed telephone subscriptions (per 100 people)
* hightech: High-technology exports (current US$)
* mobile: Mobile cellular subscriptions (per 100 people)

Please consider the following tasks

Do the cluster analysis using hierarchical clustering algorithm. Select not less than 4 interval or ratio scale variables for the analysis.

1. Define the number of clusters. Explain your decision.
2. Create a dendrogram and describe it.
3. Save cluster membership variable in the file.
4. Describe the clusters using descriptive statistics. Please note, that it’s not enough just to create a table with the descriptive statistics. The important aspect of the analysis is the interpretation of the obtained results. If you don’t provide the textual description of each cluster the whole current task will be considered not done.
5. Give names to the obtained clusters and explain your selection of names.
6. Do a post-hoc analysis to approve the validity of your results.

**Part 2**

The file Video\_Games\_Sales.csv

Source: <https://www.kaggle.com/rush4ratio/video-game-sales-with-ratings>

Variables:

* Name - The game’s name
* Platform - Platform of the game’s release (i.e. PC, PS4, etc.)
* Year\_of\_Release - Year of the game's release
* Genre - Genre of the game
* Publisher - Publisher of the game
* NA\_Sales - Sales in North America (in millions)
* EU\_Sales - Sales in Europe (in millions)
* JP\_Sales - Sales in Japan (in millions)
* Other\_Sales - Sales in the rest of the world (in millions)
* Global\_Sales - Total worldwide sales.
* Critic\_score - Aggregate score compiled by Metacritic staff
* Criticcount - The number of critics used in coming up with the Criticscore
* User\_score - Score by Metacritic's subscribers
* Usercount - Number of users who gave the userscore
* Developer - Party responsible for creating the game
* Rating - The [ESRB](https://www.esrb.org/) ratings

Please consider the following tasks

Do the cluster analysis using k-means algorithm. Select not less than 4 interval or ratio scale variables for the analysis.

1. Define the number of clusters. Explain your decision.
2. Save cluster membership variable in the file.
3. Describe the clusters using descriptive statistics. Please note, that it’s not enough just to create a table with the descriptive statistics. The important aspect of the analysis is the interpretation of the obtained results. If you don’t provide the textual description of each cluster the whole current task will be considered not done.
4. Give names to the obtained clusters and explain your selection of names.
5. Do a post-hoc analysis to approve the validity of your results.

Please send the Jupyter Notebook with comments and answers to amelikyan@hse. The task can be done **only individually**.The deadline for submitting the task is **25 March 21:00**.