Jean Monnet University

Machine Learning and Data Mining master program

Second Semester

“Knowledge Discovery and Data

Mining”

Data Mining Project with R

*Author*

Kristina Kulivnyk

*Supervisor*

Prof. Fabrice Muhlenbach

March 13, 2019

Table of Contents

[Problem understanding 3](#_Toc3755592)

[Data understanding 3](#_Toc3755593)

[Data preparation – Modeling –Evaluation for selected questions 4](#_Toc3755594)

[Question 1: “Discover attendance patterns for urban and rural locations” 4](#_Toc3755595)

[Question 2: “Make an analysis of teachers’ qualification over the years, depending on sex and educational institutions level. Analyze pupils-teachers’ ratio over the years 2012-2017” 7](#_Toc3755596)

[Question 3: “Make cluster analysis of the countries where Ukrainian student will preferably go depending on educational level (bac, license, master) over the years 2012-2018.” 9](#_Toc3755597)

[Question 4: “Discover patterns about preferred fields of studies for a male/female student over the years (2012-2018).” 11](#_Toc3755598)

[Deployment 14](#_Toc3755599)

[Conclusions: 14](#_Toc3755600)

[Sources: 15](#_Toc3755601)

# Problem understanding

For Data Mining project I decided to choose a dataset that combine information about state of education in Ukraine. It has diverse data, starting from school attendance rate, up to school infrastructure.

I’ve choose this dataset because it is compelling and indispensable personally to me to get a deeper understanding of the state of education in my native country via building various statistical and predictive models. The reason of such interest is in an increasing number of debates over the quality of education in Ukraine, its relevance comparing to European systems and causes that could help to bring it to a new, more advanced level. Since posting a real state or the education has lots of rigid drawbacks for the national government, there exist a lot of populism and manipulation both in press and web articles over such vulnerable topic. My goal is to get relevant information and to build a genuine models using independent data, collected by UNESCO.

I hope that my work could not only provide some valuable information personally for me but would also be a good start for actual changes in educational system in Ukraine. I believe that when we know reasons, we could fight for better results.

The data for my research was found on a UNESCO website (could be accessed via the link - <https://data.humdata.org/dataset/unesco-indicators-for-ukraine> ). The goal of such datasets is to provide the most recent information about state of culture, education, literacy of the country. Datasets are updated every year.

# Data understanding

Taking the first glance over my data I decided to check several questions:

1. Discover attendance patterns for urban and rural locations.
2. Make an analysis of teachers’ qualification over the years, depending on sex and educational institutions level. Analyze pupils-teachers’ ratio over the years 2012-2017.
3. Make cluster analysis of the countries where Ukrainian student will preferably go depending on educational level (bac, license, master) over the years 2012-2018.
4. Discover patterns about preferred fields of studies for a male/female student over the years (2012-2018).

After more detailed data exploration it was discovered that not all of the questions could be answered due to lack of data and features, generally bad quality of data. Another problem that I’ve faced was the absence of support documentation for a dataset. That is why the first initial step that was fulfilled - exploration of all possible variables that exist in the dataset. A full variable table could be found in Appendix 1. From 30 columns of the dataset there are 16 valuable columns with diverse variables, 7 columns with numerical data (integer and double data values) by year (2012-2018), 1 column with all countries listed and 6 columns with inconsistent data. The total number of observations (rows of the dataset) is 3318.

The size of full dataset is 1.6 MB, the file original encoded in .csv.

# Data preparation – Modeling –Evaluation for selected questions

The next important step to follow in any Data Mining project is to prepare data - it means to check the quality of data, deal with missing observations, probably select a sub-set of data and re-format it.

After initial check of the data in Ukrainian educational dataset it was discovered that several columns have inconsistent data, so it was decided to delete such columns to avoid further misunderstanding of future results of data analysis.

The quality of data set appears to be very poor. Ensuing checking the exact number of observations it was discovered that only 47 rows has no NA values. It is clear that we couldn’t simply drop all observations that has any missing value, because in such a case 98 % of information will be lost.

I came to a conclusion to split a dataset to a smaller sub-sets for each of the initially posed questions. Such preventive actions could help to be more accurate with data analysis and to deal better with incomplete information.

## Question 1: “Discover attendance patterns for urban and rural locations”

For Question 1: “Discover attendance patterns for urban and rural locations” all observations were initially filtered to select those that have “Urban” and “Rural” location markers.

The next step was to choose the data that was associated with Net attendance rate for Males and Females in percentage for each of two areas – urban and rural.

As it was already mentioned that the quality of data in the dataset appears to be very poor – the only year from which accurate numerical data could be taken appeared to be 2013. It was decided that all further explorations will be done only for this year.

After data preparation a new data frame was created. It combined information about education level, wealth level and percentage of attendance for females in rural area, males in rural area, females in urban area and males in urban area. Plots that are presented below shows some interesting patterns:

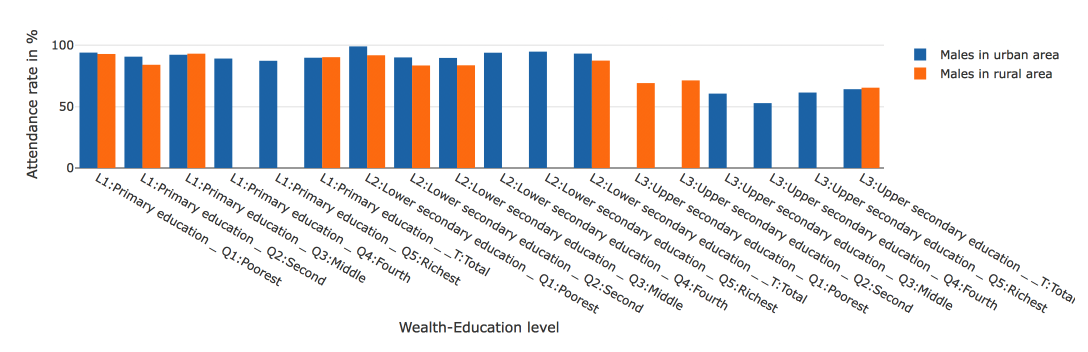


Figure 1. Male rura - urban comparison

Figure 1 shows that generally attendance rate for males in rural area is lower that for males in urban area. At the same time, another quite interesting discovery is that there is no attendance rate for Middle – Richest class in rural area for lower secondary and secondary education. That is a significant result that can be explained by 2 reasons:

1) People with a higher social status try to go to urban area and let children go to school there. But this trend holds only for families that have children in the age of 12 – 17 years old (age of lower secondary and secondary education).

2) At the same time this trend changes for middle wealth families with children of age 6-11 (Primary education). It means that people in a rural areas begin to have higher levels of incomes and that the level of primary education is good enough not to male parents move to an urban area.

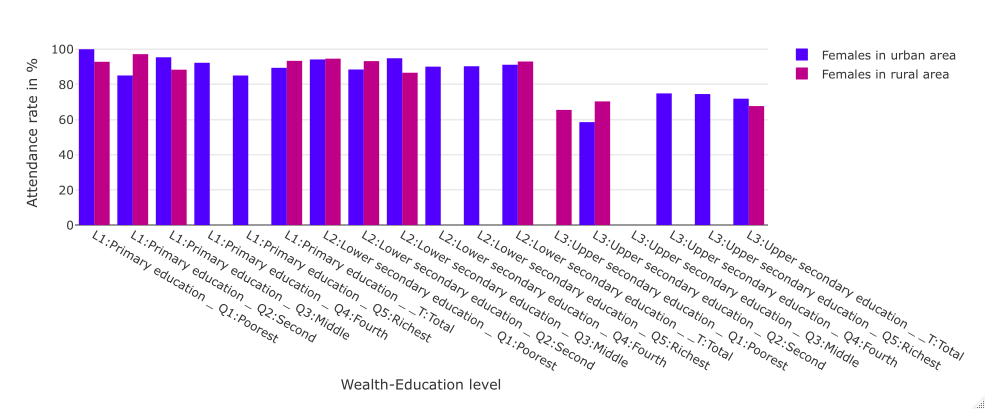


Figure 2. Female rural-urban comparison

An interesting observation from the Figure 2 shows that in 2013 in urban area there were no observations about poorest and middle class upper-secondary school girls. At the same time trend about no attendance rate for Middle – Richest class in rural area for lower secondary and secondary education still holds for females.

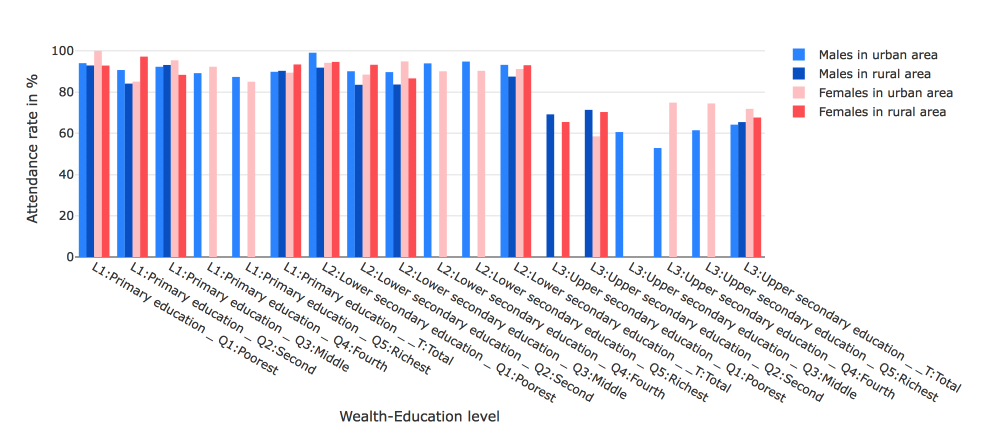


Figure 3. General comparison male-female rural-urban area

From figure 3 we can conclude that females generally have better attendance rate, except the case of lower secondary school males in urban area. Another trend is a decrease in attendance of upper secondary school for both rural and urban areas.

Table below represents mean attendance values for rural and urban females and males students.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Urban Males | Rural Males | Urban Females | Rural Females |
| Primary education | 90.70567 | 90.05229 | 91.56606 | 92.81061 |
| Lower secondary education | 93.5186 | 86.35291 | 91.59335 | 91.50249 |
| Upper-secondary education | 58.31553 | 70.28066 | 69.29369 | 67.91058 |

Table 1. Mean attendancy values

From this table we can conclude that the lowest attendance value is for urban males in upper-secondary education, but at the same time surprisingly high values for upper secondary education for rural males. Females used to attend school more often than males for both urban and rural areas. The interesting observation is that attendance rate for primary education for rural females is higher that for urban females. That could be possible due to personal parents’ reasons (usually in urban area if a small kid is ill, it stays at home, but in rural area it almost never happened, everybody goes to school).

Unfortunately, due to a small amount of available observations and nature of the data it is not possible to use any Machine Learning algorithms to make any kind of appropriate predictions.

## Question 2: “Make an analysis of teachers’ qualification over the years, depending on sex and educational institutions level. Analyze pupils-teachers’ ratio over the years 2012-2017”

The next question that was investigated concern analysis of teachers’ data. As usual the first step is to select the sub-set that could be relevant for further examination. For this dataset all statistical units that are related to teachers were selected.

The upcoming step was to perform a distinctive analysis to understand what unique statistical units were selected during the first step.

Afterwards, to make data unified, only observations with measured in percentages were selected. Eventually data was checked for the observation with the smallest amount of missing values. Those observations appear to be the ones that check pupils-teachers’ ratio. Taking into account this results our goal to proceed with teachers’ qualification analysis could not be accomplished due to the data quality. In view of that fact a new hypothesis was selected – “Analyze pupils-teachers’ ratio over the years 2012-2017”.

Following that it is interesting to check number of missing values for variables of the obtained sub-set.

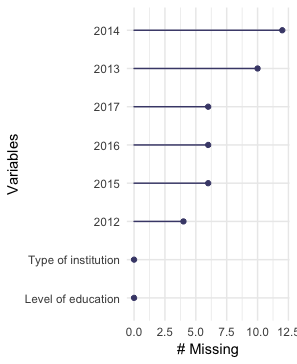
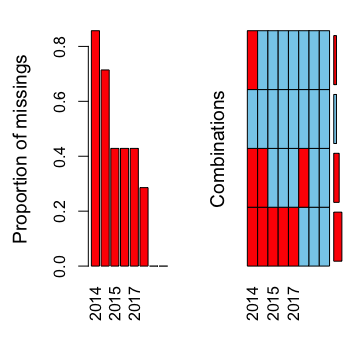
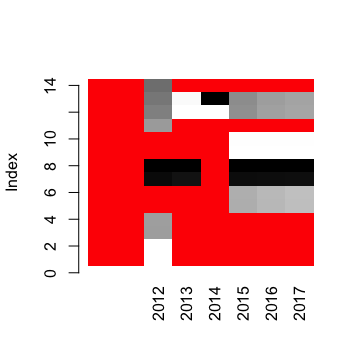


Figure 4. Analysis of missing values for teachers' sub-set

Taking into account results of the missing values analysis (Figure 4) it is clear that for 2014 there exist a big number of absent observations. This observation must be considered during all upcoming steps.

The third plot from Figure 4 shows the dependency between missing values for years and “Level of education” feature. From this plot we can deduce that for years 2012-2017 there is no values for the first four observations of the “Level of education” feature.

In the following step only the data where less that 55% of observation is missing were selected for further exploration.

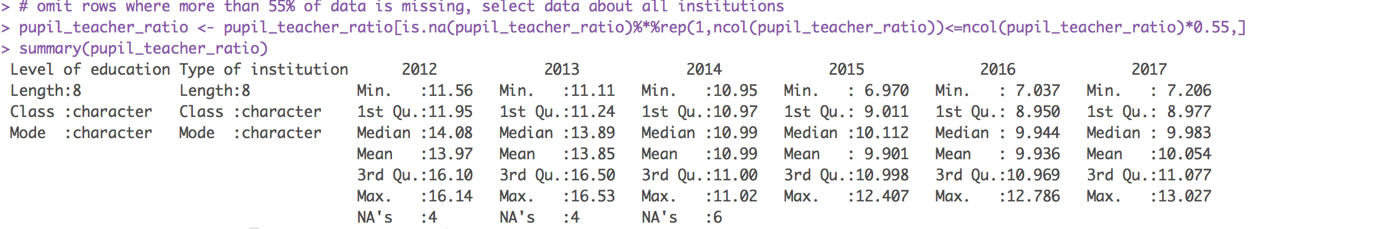


Figure 5. Pupil-teachers' ratio sub-set summary

From the summary we can conclude that the medium number of student per teacher decrease till 2016 and then start to grow up in 2017.

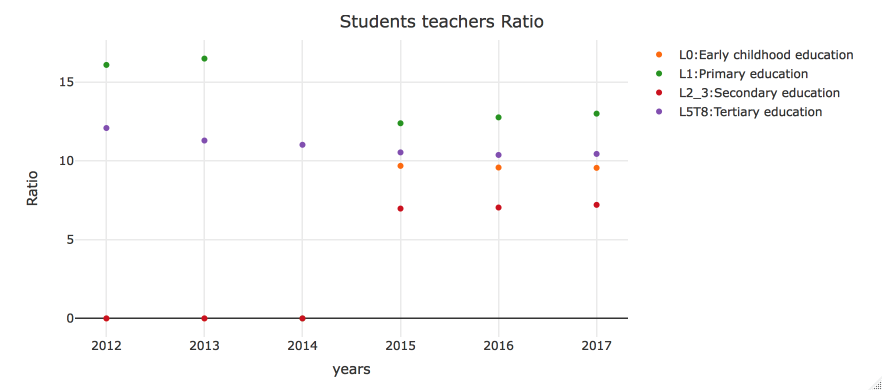


Figure 6. Pupils-teachers' ratio by education level

From the plot on figure 6 we can conclude that the number of students per teacher for early childhood education and secondary education doesn’t change much over the years of observations. For primary education a pick level was established in 2013, but than the number reduce in 2015 with a linear grows until 2017. For tertiary education a decrease of pupil-teachers’ ratio could be remarked. The most correlated reason for such trend is unstable economical situation and war that makes students leave Ukraine and not to stay for tertiary education in theirs’ native country.

Taking into account this information it is possible to proceed with the next question.

## Question 3: “Make cluster analysis of the countries where Ukrainian student will preferably go depending on educational level (bac, license, master) over the years 2012-2018.”

The new sub-set was selected. This sub-set consist of outbound internationally mobile students’ data. Searching for unusual patterns it was resolved to implement a hierarchical clustering over the sub-set.

Cluster dendrogram perfectly shows the level of similarity over number of mobile student over the six years (2012 - 2017). From the observations we can conclude that number of departures for Central and Eastern Europe and Arab states are on the same level with departures to North America and Western Europe. That confirms by the fact of popularity of this destinations among students, similarities of languages (for Central Eastern Europe) and cultures. At the same time departures to East Asia and Oceania are on low level due to several reasons, among which small number of exchange programs and low level of recognizability of the East Asia universities among Ukrainian students. Another reason for such result could be absence of observations for several years for this countries.

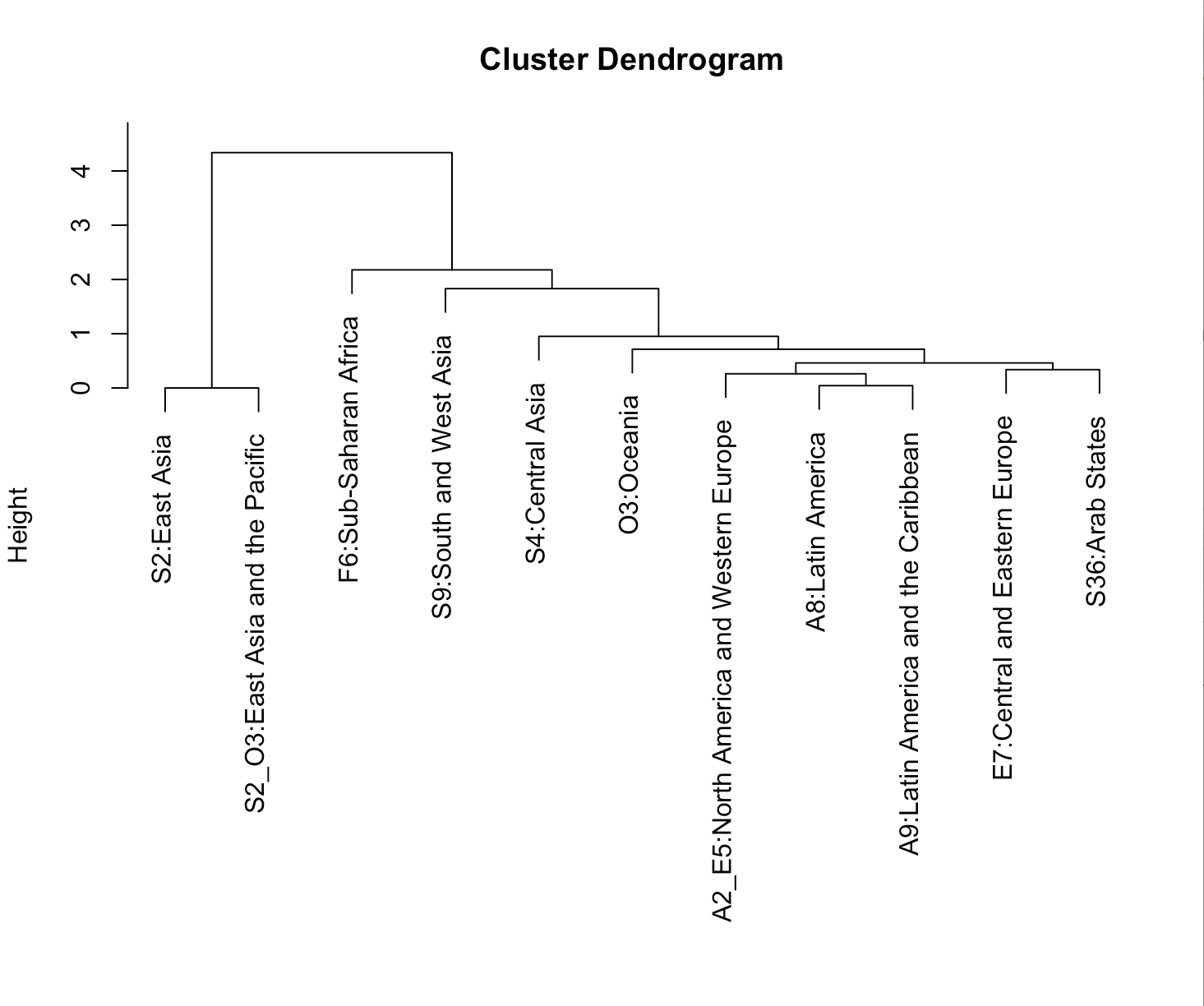


Figure 7. Mobility dendogram

To understand the global mobility trend a log scaled scatter plot was build. A log scale was needed because Western Europe and Central Europe mobility observations are almost 500 times bigger than any other observations (due to country preference level).

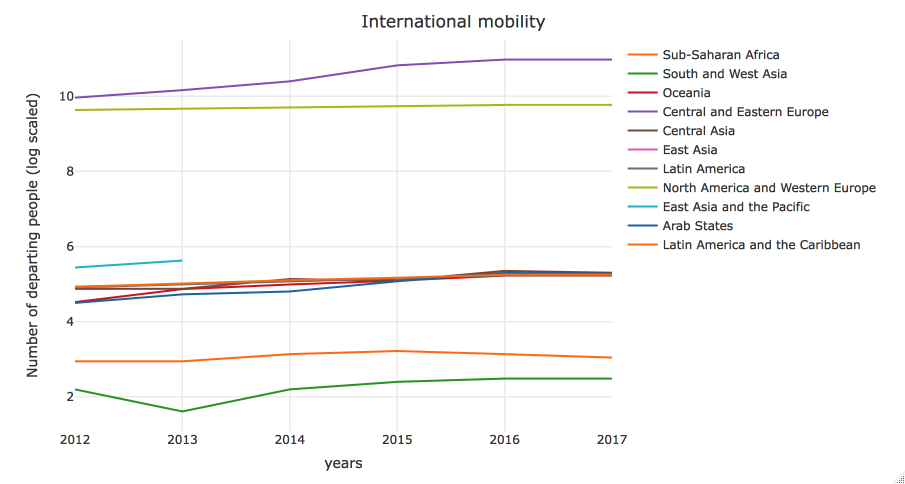


Figure 8. Mobility trends

From the Figure 8 it is possible to conclude that generally mobility grows up each year (number of departure students increase). For North America and Western Europe, the number of departures considered to be almost constant, but a significant increase could be observed for Central and Eastern European countries. The main reason – declaration of visa-free short term traveling and simplified procedure for education visas for Poland, Czech Republic and Hungary for Ukrainian students. And the same time the number of departures to Latin America decrease, probably due to a tough political situation in several countries.

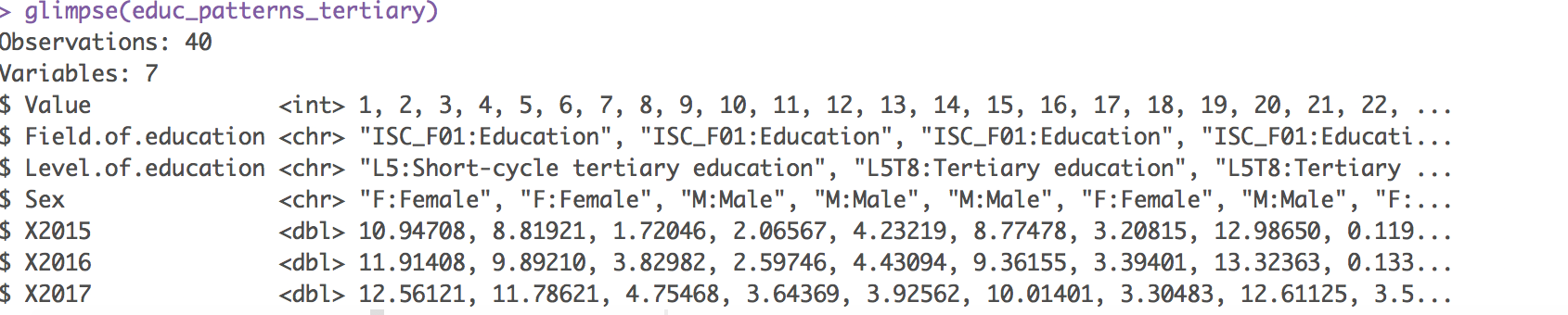
Unfortunately, the number of observation of mobility sub-set is quite small, it is meaningless to proceed with any kind of regression to make some predictions about 2019. Also the number of observed features decreased significantly after data cleaning that is why it makes more sense to proceed with additional sub-set to perform any machine learning technics.

## Question 4: “Discover patterns about preferred fields of studies for a male/female student over the years (2012-2018).”

For the last question a sub-set of “Preferred fields of education” was formed. As a main statistical unit “Distribution of students in tertiary education by field of education” was selected due to a fact the biggest number of not-corrupted observations were associated with this unit.

For the better interpretability “Field of education” were placed to be a first column in a data frame.

Our goal would be to use k-means clustering to distinguish patterns in data. This solution was inspired by the articles on R-bloggers website [10]. Using glimpse function the sub-set was observed.

Figure 9. Data observation with glimpse function

To proceed with further analysis, it is needed to make first tree features as factors. Afterwards it would be possible to build a Gower distance matrix that could be used for clustering. Gower matrix also allow us to check the most similar and de-similar data pair.

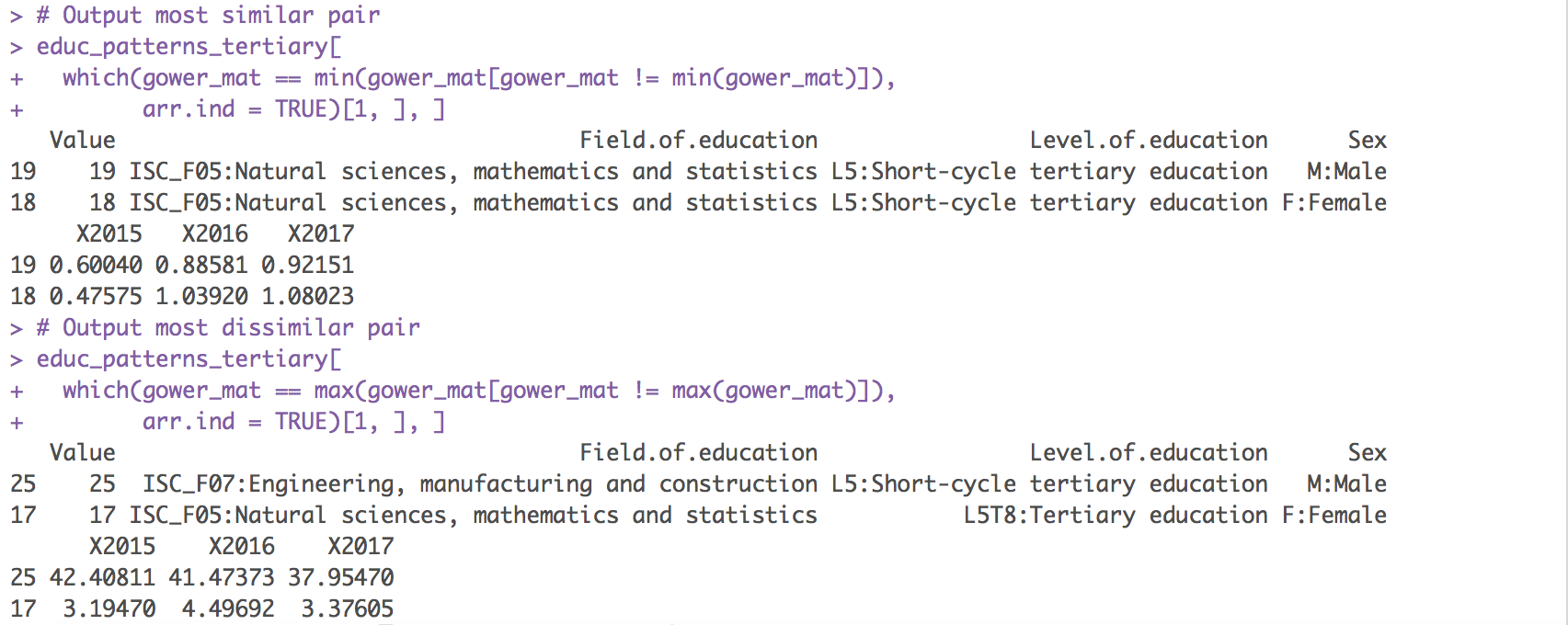


Figure 10. Most similar vs most desimilar data pair

Our next step is evaluation for clustering. Using silhouette, we would like to get the best fitted value of clusters for modeling.

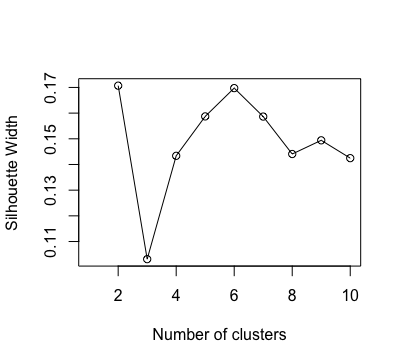
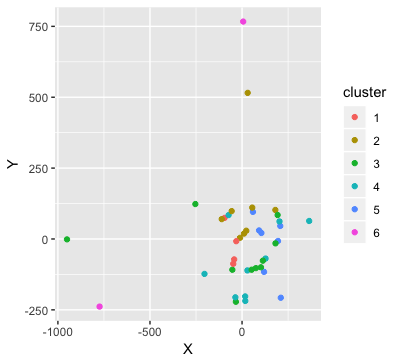
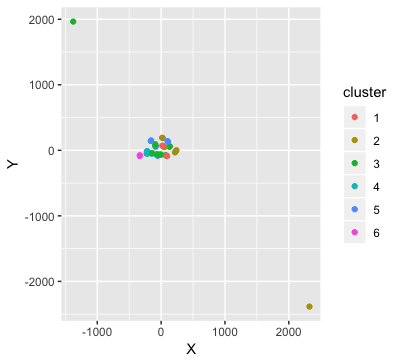
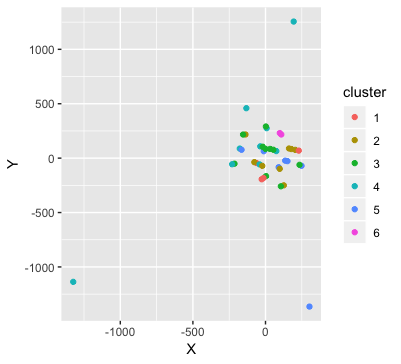


Figure 11. Evaluation of number of clusters

The evaluation results show that the best number of clusters could be either 2, either 6. It would be interesting to test both possibilities.

As a cluster algorithm a partitioning around medoids algorithm was selected. This algorithm is very familiar to k-means, but cluster centers for PAM are restricted to be the observations themselves. The pros of this algorithm is that it is easy to understand, more robust to noise and outliers when compared to k-means, and has the added benefit of having an observation serve as the exemplar for each cluster [10], which is perfect solution for preferred fields of education sub-set.

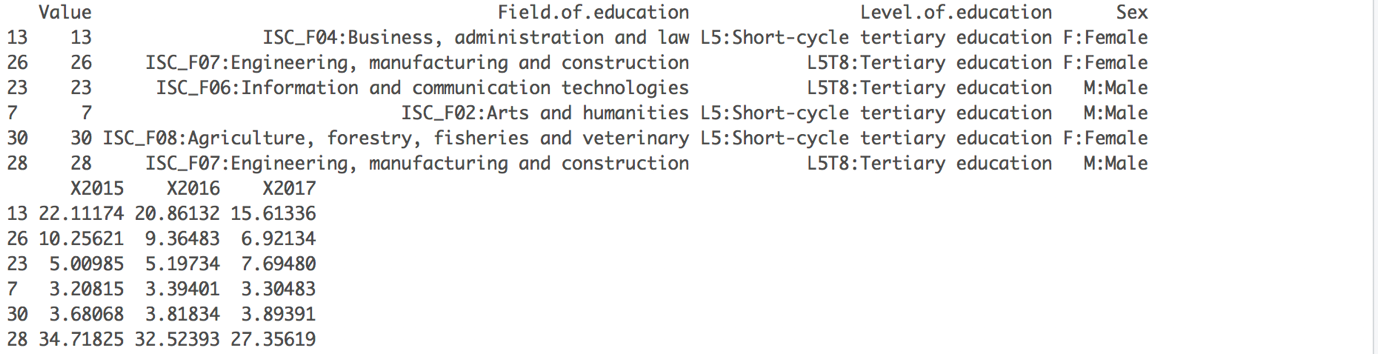
First run the clustering algorithm with 6 clusters. All computation process takes only 2.906 seconds because of a size of a “preferred fields of education” sub-set.



# 

Figure 12. Tree runs of clustering with 6 clusters

With each new run the clusters will drastically change, but methoids remains the same:



For the second run we’ll select two clusters. Computations take less time - 1.229 sec this time.

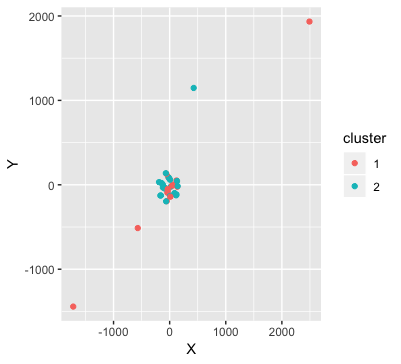
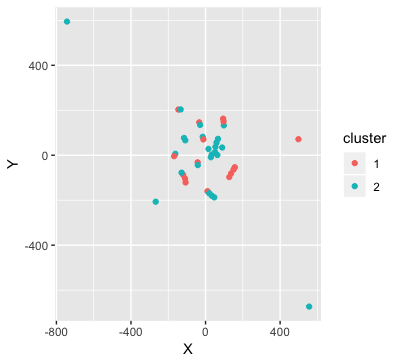
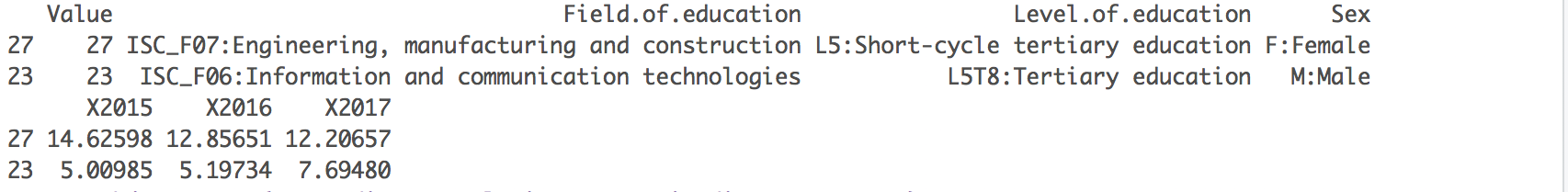


Figure 13. Tree runs of clustering with 2 clusters

Metoids for all tree runs remains the same:



Both of six and two – cluster PAM clustering provides logical results and can be used for this kind of data.

# Deployment

To make the work with the project more structured and efficient I’ve created a github repository that hosts report file, data and R code itself. It could be accessed via the link: <https://github.com/KrisKuliv/MLDM_Data_Mining>

There exist three branches:

* **Master branch** with the final version of the project.
* **Docs branch** where only report file is stored.
* **Data\_code branch** where R project files are stored.

Working with git repository helped me to keep all changes secure and not to be afraid of loosing any precious results. It is also interesting to check the statistics of commits to understand the personal productivity level.

As using git is a best-practice in the biggest part of IT companies, working with this project also made me train my git-usage skills.

# Conclusions:

Working with Data Mining projects helped me acquire a huge number of skills, such as manipulation with data, data cleaning, data understanding and general data analysis, usage of machine learning techniques for a given dataset. There are several important points that are needed to be highlighted:

1. Data selection. Ukrainian education data set was definitely not the best choice due to the quality of data. For the next projects I wish I could choose the datasets with more numerical data that could be more easily to manipulate and build machine learning models.
2. Documentation. Due to a dataset support documentation non-existence a vast number of hours were spend on additional data investigation to get the full understanding o features.
3. Results. During the project I’ve obtained various interesting results concerning educational trends in Ukraine. For the biggest part of the obtained results a direct correlation between political and socio-economical situation can be seen.

# Sources:

1. Ukraine - Sustainable development, Education, Demographic and Socioeconomic Indicators - Humanitarian Data Exchange. *Data.humdata.org*, 2019. <https://data.humdata.org/dataset/unesco-indicators-for-ukraine>.
2. Google's R Style Guide. *Google.github.io*, 2019. <https://google.github.io/styleguide/Rguide.xml#filenames>.
3. R, R. Consistent naming conventions in R. *R-bloggers*, 2019. <https://www.r-bloggers.com/consistent-naming-conventions-in-r/>.
4. Romeo, V. Preparing the data for modelling with R. *R-bloggers*, 2019. <https://www.r-bloggers.com/preparing-the-data-for-modelling-with-r/>.
5. 11 Most Useful Steps to Create Data Exploration in R | Methods | Example | Definition. *EDUCBA*, 2019. <https://www.educba.com/data-exploration-in-r/>.
6. Data Preparation — A crucial step in Data Mining. *Medium*, 2019. <https://medium.com/@chhavi.saluja1401/data-preparation-a-crucial-step-in-data-mining-dba35772f281>.
7. The 10 Mining Techniques Data Scientists Need for Their Toolbox. *Towards Data Science*, 2019. <https://towardsdatascience.com/the-10-mining-techniques-data-scientists-need-for-their-toolbox-ae15a5733b02>.
8. 8 Data Mining Techniques You Must Learn To Succeed In Business. *Medium*, 2019. <https://medium.com/@onix_systems/8-data-mining-techniques-you-must-learn-to-succeed-in-business-ae4032bf6469>.
9. 15 Easy Solutions To Your Data Frame Problems In R. *R-bloggers*, 2019. <https://www.r-bloggers.com/15-easy-solutions-to-your-data-frame-problems-in-r/>.
10. r, W. Clustering Mixed Data Types in R. *R-bloggers*, 2019. <https://www.r-bloggers.com/clustering-mixed-data-types-in-r/>.
11. blog, A. 5 ways to measure running time of R code. *R-bloggers*, 2019. <https://www.r-bloggers.com/5-ways-to-measure-running-time-of-r-code/>.
12. Factoextra R Package: Easy Multivariate Data Analyses and Elegant Visualization - Easy Guides - Wiki - STHDA. *Sthda.com*, 2019. <http://www.sthda.com/english/wiki/factoextra-r-package-easy-multivariate-data-analyses-and-elegant-visualization>.
13. K-Means Clustering in R: Algorithm and Practical Examples - Datanovia. *Datanovia*, 2019. <https://www.datanovia.com/en/lessons/k-means-clustering-in-r-algorith-and-practical-examples/>.
14. K-means Clustering in R with Example. *Guru99.com*, 2019. https://www.guru99.com/r-k-means-clustering.html.
15. Correlation Test Between Two Variables in R - Easy Guides - Wiki - STHDA. *Sthda.com*, 2019. http://www.sthda.com/english/wiki/correlation-test-between-two-variables-in-r.
16. Linear Regression With R. *R-statistics.co*, 2019. http://r-statistics.co/Linear-Regression.html.
17. Quick-R: Multiple Regression. *Statmethods.net*, 2019. https://www.statmethods.net/stats/regression.html.
18. 5 Amazing Types of Clustering Methods You Should Know - Datanovia. *Datanovia*, 2019. https://www.datanovia.com/en/blog/types-of-clustering-methods-overview-and-quick-start-r-code/.
19. Hierarchical Cluster Analysis · UC Business Analytics R Programming Guide. *Uc-r.github.io*, 2019. https://uc-r.github.io/hc\_clustering.
20. Cluster Analysis in R. *Xperimentallearning.blogspot.com*, 2019. http://xperimentallearning.blogspot.com/2018/04/cluster-analysis-in-r.html.
21. Performing a cluster analysis in R. *Instant R*, 2019. http://www.instantr.com/2013/02/12/performing-a-cluster-analysis-in-r/.
22. (first.last@ucr.edu), F. Cluster Analysis in R. *Girke.bioinformatics.ucr.edu*, 2019. http://girke.bioinformatics.ucr.edu/GEN242/pages/mydoc/Rclustering.html#2\_data\_preprocessing.
23. RPubs - Cluster Analysis in R: Examples and Case Studies. *Rpubs.com*, 2019. https://rpubs.com/gabrielmartos/ClusterAnalysis.
24. Josse, J. Handling missing values with R. *Juliejosse.com*, 2019. <http://juliejosse.com/wp-content/uploads/2018/06/DataAnalysisMissingR.html#4)_multilevel_(mixed)_data_with_missing_values>.