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

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:

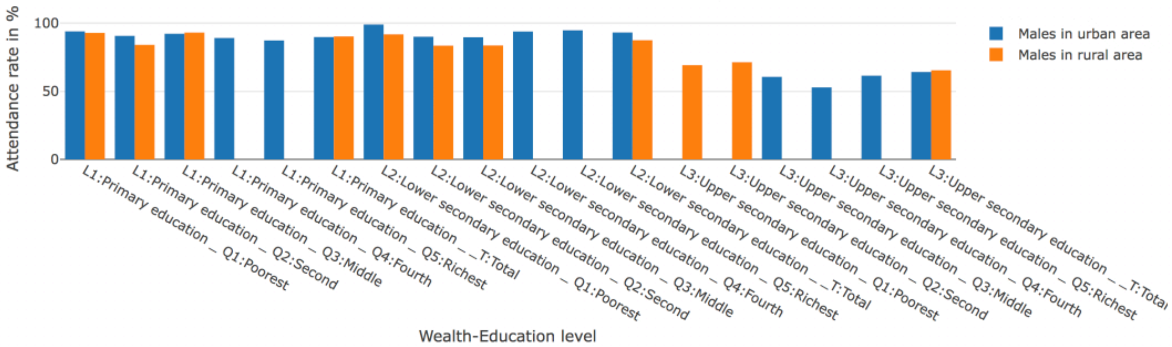


Figure 1. Male rura - urban comparison

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.

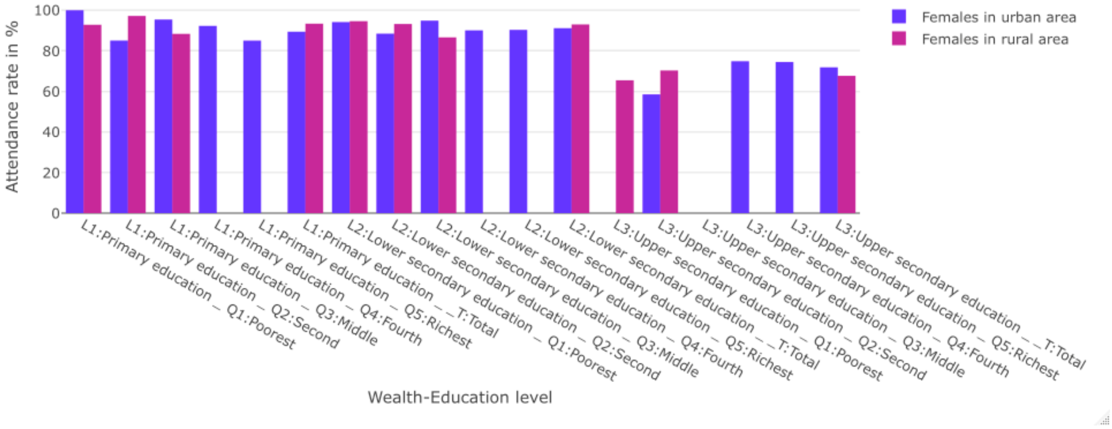


Figure 2. Female rural-urban comparison

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.

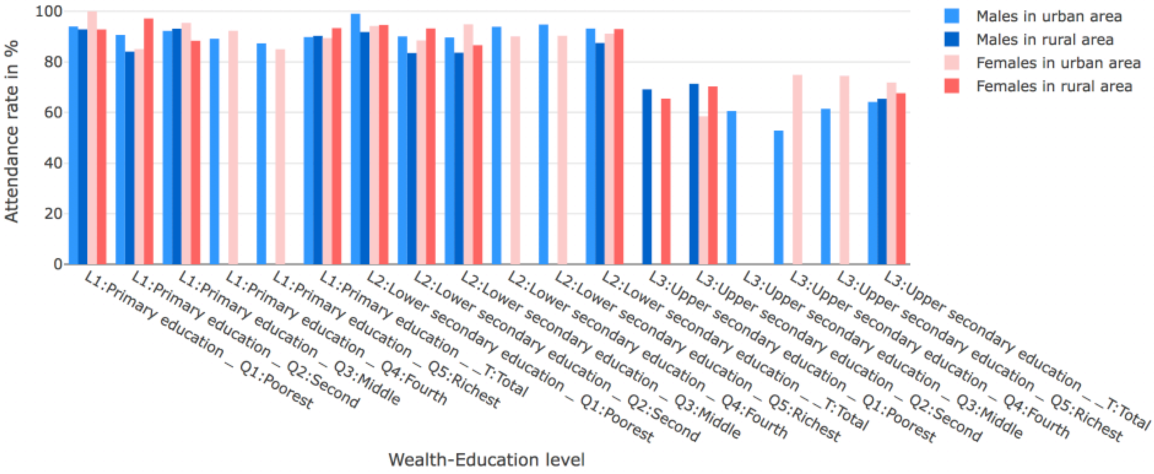


Figure 3. General comparison male-female rural-urban area

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.

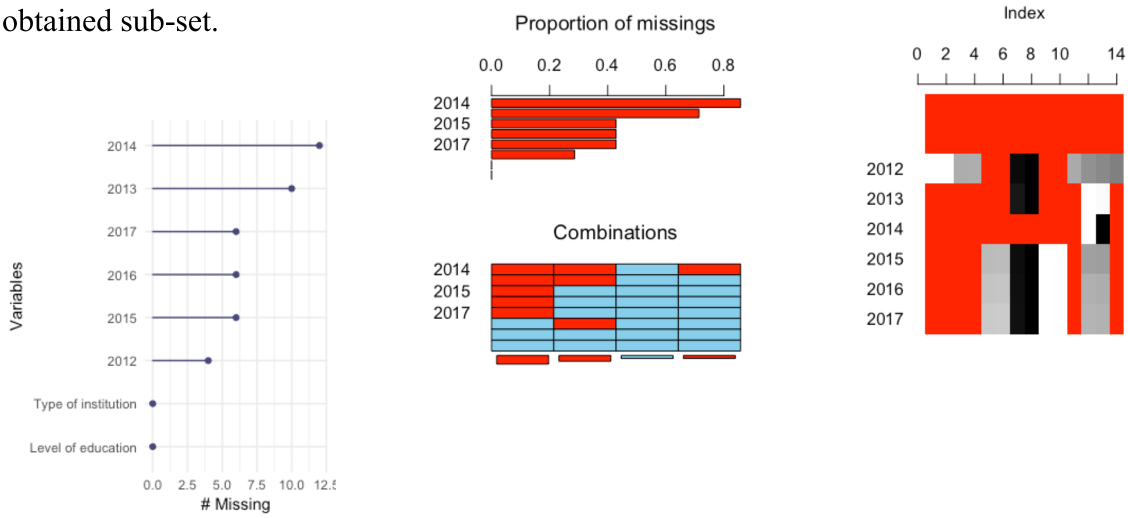
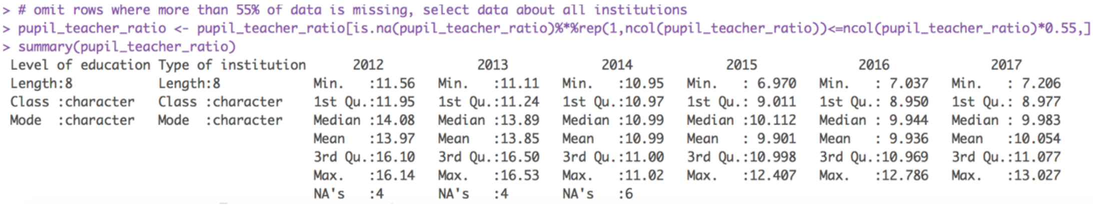


Figure 4. Analysis of missing values for teachers' sub-set

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.



```
> # omit rows where more than 55% of data is missing, select data about all institutions
> pupil_teacher_ratio <- pupil_teacher_ratio[is.na(pupil_teacher_ratio)%*%rep(1,ncol(pupil_teacher_ratio))<=ncol(pupil_teacher_ratio)*0.55,]
> summary(pupil_teacher_ratio)
 Level of education Type of institution      2012            2013            2014            2015             2016             2017       
 Length:8           Length:8           Min.   :11.56   Min.   :11.11   Min.   :10.95   Min.   : 6.970   Min.   : 7.037   Min.   : 7.206  
 Class :character   Class :character   1st Qu.:11.95   1st Qu.:11.24   1st Qu.:10.97   1st Qu.: 9.011   1st Qu.: 8.950   1st Qu.: 8.977  
 Mode  :character   Mode  :character   Median :14.08   Median :13.89   Median :10.99   Median :10.112   Median : 9.944   Median : 9.983  
                                       Mean   :13.97   Mean   :13.85   Mean   :10.99   Mean   : 9.901   Mean   : 9.936   Mean   :10.054  
                                       3rd Qu.:16.10   3rd Qu.:16.50   3rd Qu.:11.00   3rd Qu.:10.998   3rd Qu.:10.969   3rd Qu.:11.077  
                                       Max.   :16.14   Max.   :16.53   Max.   :11.02   Max.   :12.407   Max.   :12.786   Max.   :13.027  
                                       NA's   :4       NA's   :4       NA's   :6                                                          
```

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.

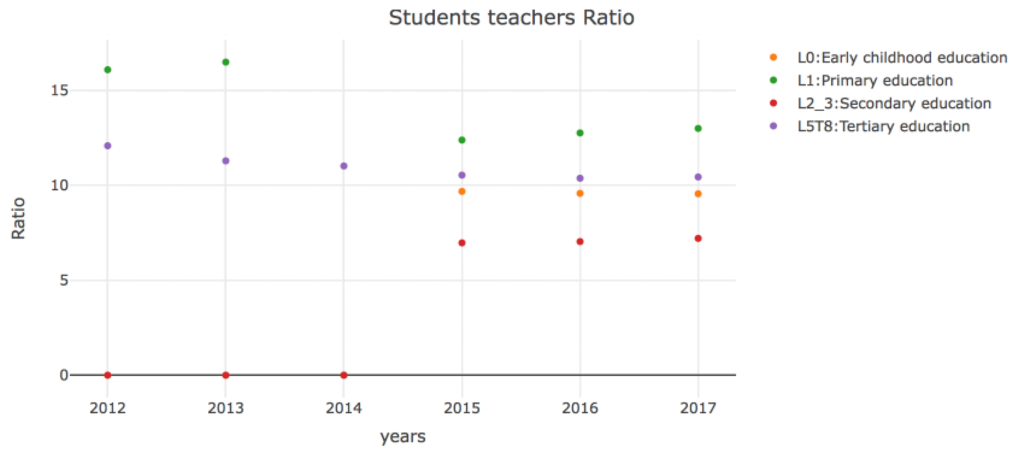


Figure 6. Pupils-teachers' ratio by education level

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.

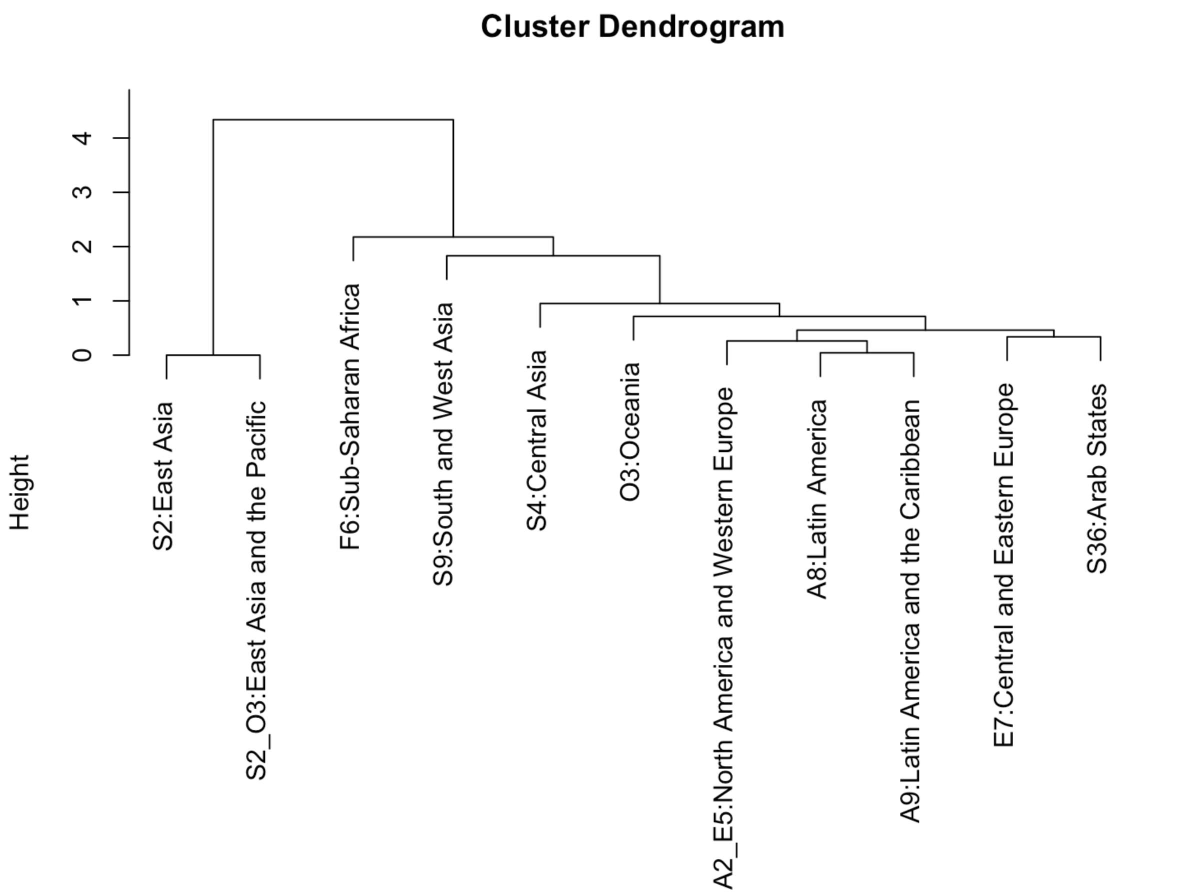


Figure 7. Mobility dendogram

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).

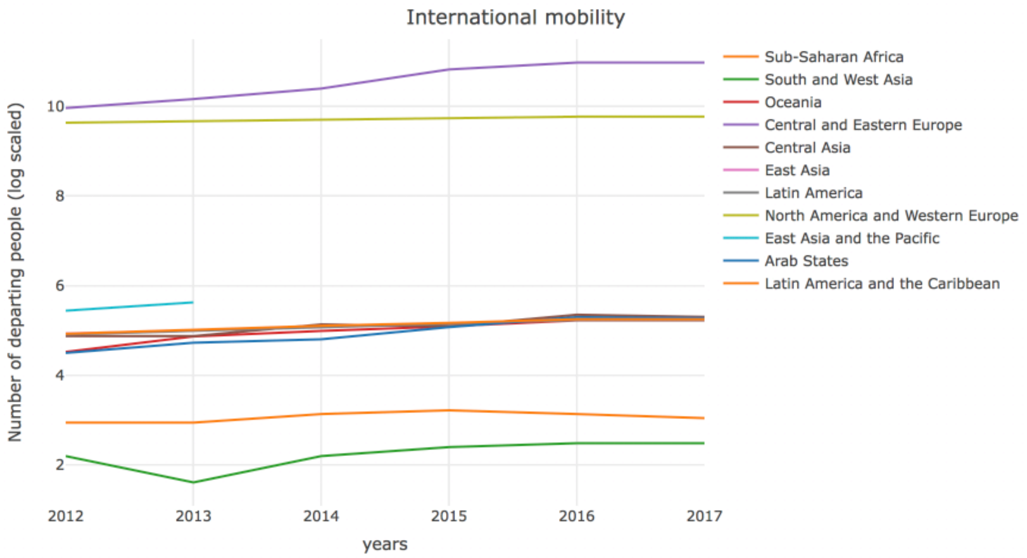


Figure 8. Mobility trends

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.

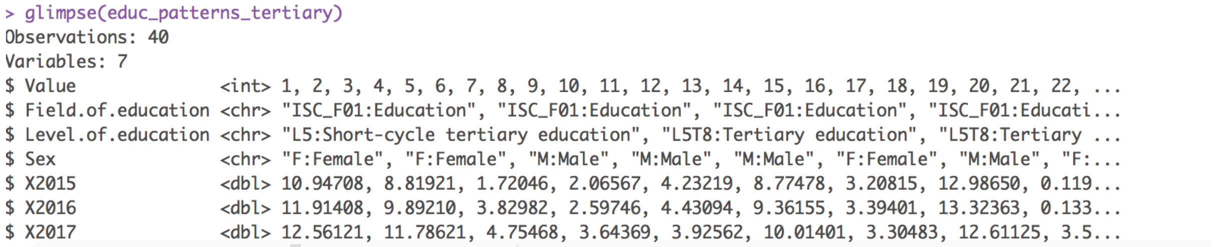
Unfortunately, the number of observation of mobility sub-set is quite small, it is meaningless to proceed with any kind of regression to make some predictions about 2019. Also the number of observed features decreased significantly after data cleaning that is why it makes more sense to proceed with additional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

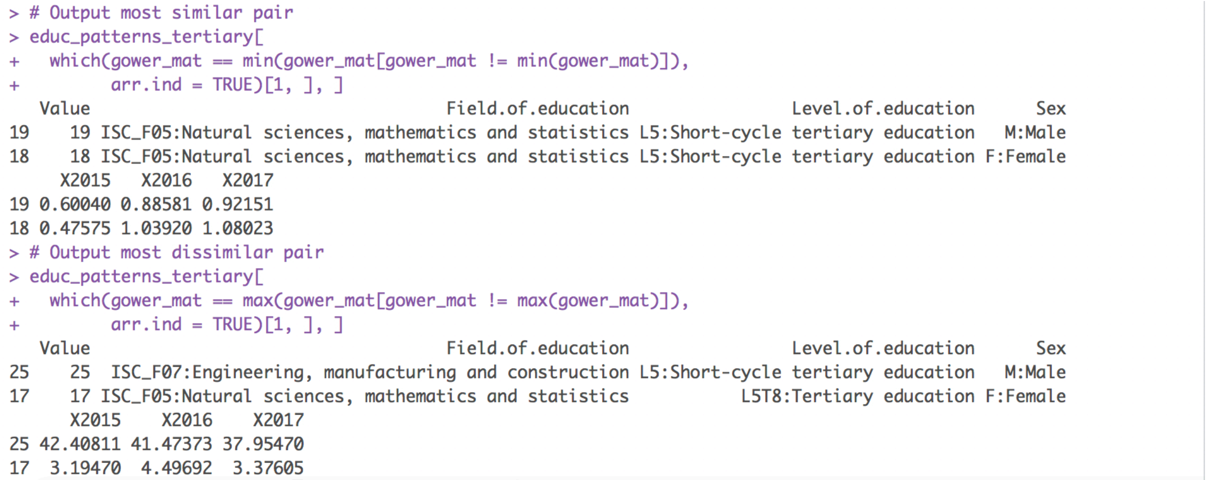
Our goal would be to use k-means clustering to distinguish patterns in data. This solution was inspired by the articles on R-bloggers website [10]. Using glimpse function the sub-set was observed.



```
> glimpse(educ_patterns_tertiary)
Observations: 40
Variables: 7
$ Value              <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, ...
$ Field.of.education <chr> "ISC_F01:Education", "ISC_F01:Education", "ISC_F01:Education", "ISC_F01:Educati...
$ Level.of.education <chr> "L5:Short-cycle tertiary education", "L5T8:Tertiary education", "L5T8:Tertiary ...
$ Sex                <chr> "F:Female", "F:Female", "M:Male", "M:Male", "M:Male", "F:Female", "M:Male", "F:...
$ X2015              <dbl> 10.94708, 8.81921, 1.72046, 2.06567, 4.23219, 8.77478, 3.20815, 12.98650, 0.119...
$ X2016              <dbl> 11.91408, 9.89210, 3.82982, 2.59746, 4.43094, 9.36155, 3.39401, 13.32363, 0.133...
$ X2017              <dbl> 12.56121, 11.78621, 4.75468, 3.64369, 3.92562, 10.01401, 3.30483, 12.61125, 3.5...
```

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.



```
> # Output most similar pair
> educ_patterns_tertiary[
+   which(gower_mat == min(gower_mat[gower_mat != min(gower_mat)]),
+         arr.ind = TRUE)[1, ], ]
   Value                                    Field.of.education                Level.of.education      Sex
19    19 ISC_F05:Natural sciences, mathematics and statistics L5:Short-cycle tertiary education   M:Male
18    18 ISC_F05:Natural sciences, mathematics and statistics L5:Short-cycle tertiary education F:Female
     X2015   X2016   X2017
19 0.60040 0.88581 0.92151
18 0.47575 1.03920 1.08023
> # Output most dissimilar pair
> educ_patterns_tertiary[
+   which(gower_mat == max(gower_mat[gower_mat != max(gower_mat)]),
+         arr.ind = TRUE)[1, ], ]
   Value                                    Field.of.education                Level.of.education      Sex
25    25  ISC_F07:Engineering, manufacturing and construction L5:Short-cycle tertiary education   M:Male
17    17 ISC_F05:Natural sciences, mathematics and statistics           L5T8:Tertiary education F:Female
      X2015    X2016    X2017
25 42.40811 41.47373 37.95470
17  3.19470  4.49692  3.37605
```

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.

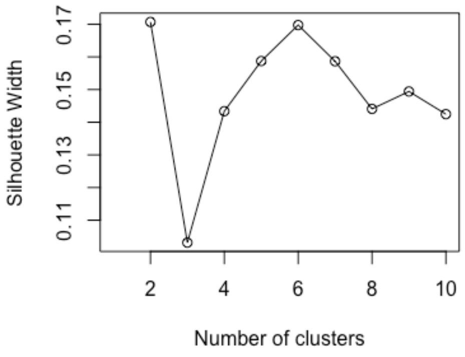


Figure 11. Evaluation of number of clusters

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.

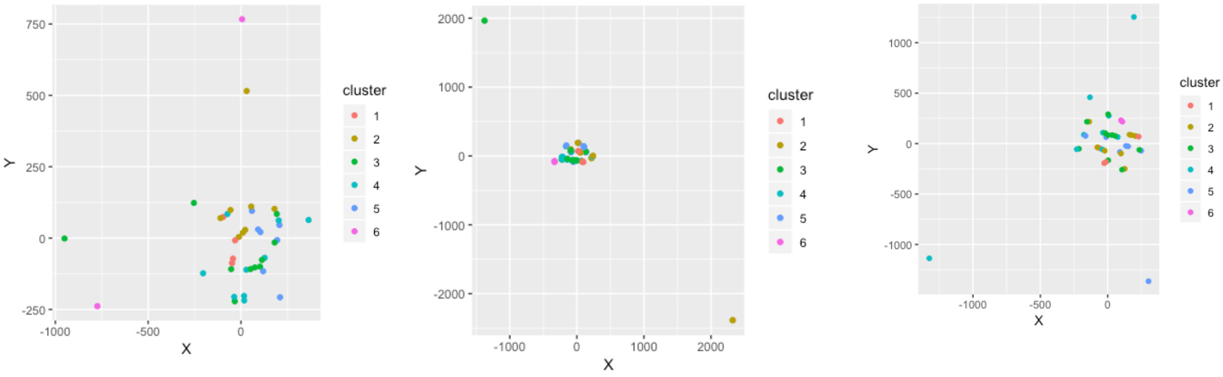
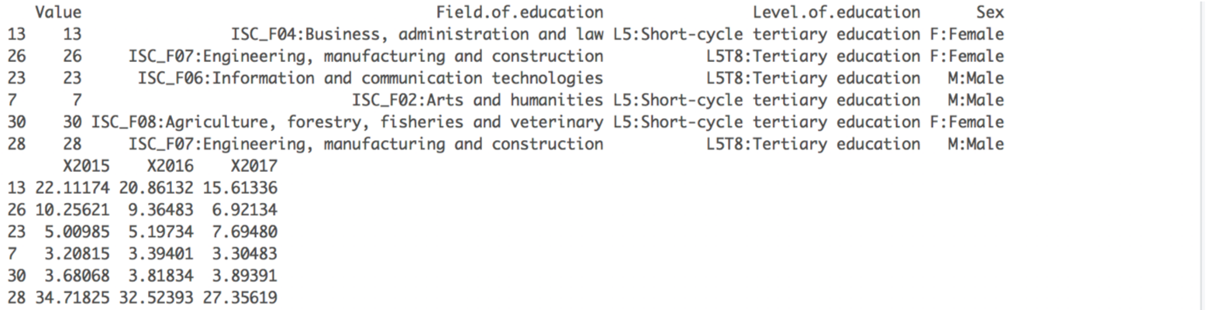


Figure 12. Tree runs of clustering with 6 clusters

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


```
   Value                                     Field.of.education                  Level.of.education      Sex
13    13                ISC_F04:Business, administration and law L5:Short-cycle tertiary education F:Female
26    26      ISC_F07:Engineering, manufacturing and construction            L5T8:Tertiary education F:Female
23    23       ISC_F06:Information and communication technologies            L5T8:Tertiary education   M:Male
7      7                         ISC_F02:Arts and humanities L5:Short-cycle tertiary education   M:Male
30    30 ISC_F08:Agriculture, forestry, fisheries and veterinary L5:Short-cycle tertiary education F:Female
28    28      ISC_F07:Engineering, manufacturing and construction            L5T8:Tertiary education   M:Male
      X2015    X2016    X2017
13 22.11174 20.86132 15.61336
26 10.25621  9.36483  6.92134
23  5.00985  5.19734  7.69480
7   3.20815  3.39401  3.30483
30  3.68068  3.81834  3.89391
28 34.71825 32.52393 27.35619
```

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

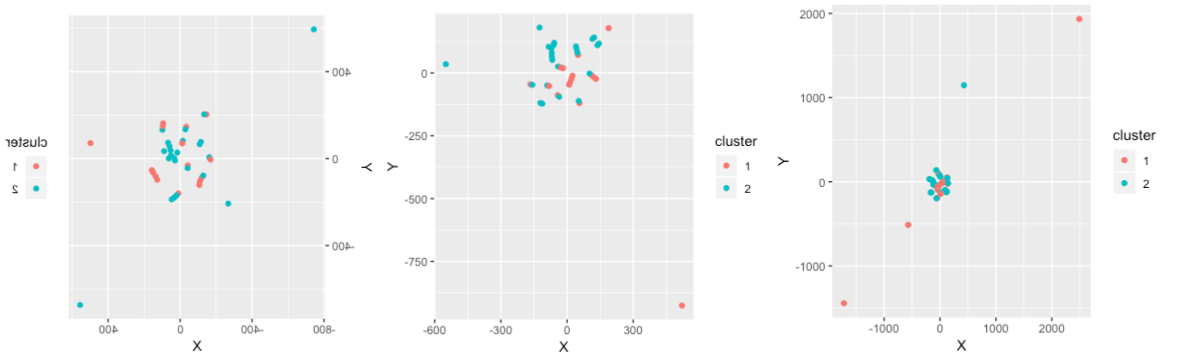
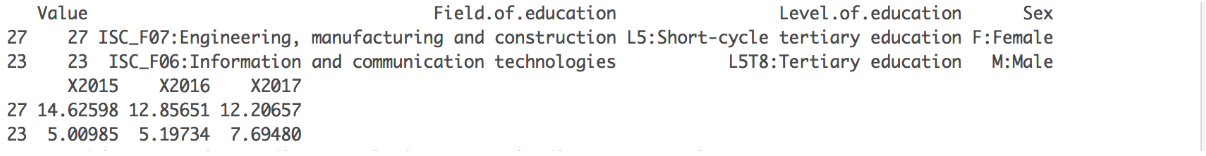


Figure 13. Tree runs of clustering with 2 clusters

Metoids for all tree runs remains the same:



```
   Value                                Field.of.education              Level.of.education      Sex
27    27 ISC_F07:Engineering, manufacturing and construction L5:Short-cycle tertiary education F:Female
23    23  ISC_F06:Information and communication technologies           L5T8:Tertiary education   M:Male
      X2015    X2016    X2017
27 14.62598 12.85651 12.20657
23  5.00985  5.19734  7.69480
```

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.

№	Statistical unit	Unit of measure	Level of education	Orientation	Sex	Age	Grade	Type of institution	Wealth quintile	Location	Type of education	Field of education	Infrastructure	Socioeconomic background	Destination region	Immigration status
1	ROFST_PHH:Rate of out-of-school children (household survey data)	PER:Number of persons	L3:Upper secondary education	_T:Total	_T:Total	SCH_AGE_GROUP:School-age population	T:Total	INST_T:All institutions	Q2:Second	_T:Total	_T:Total	_T:Total	_Z:Not applicable	_Z:Not applicable	W00:All countries	_Z:Not applicable
2	FEP:Percentage of female students	GLPIA:Adjusted location parity index	L8:Doctoral or equivalent level	_Z:Not applicable	F:Female	_T:Total	G6:Grade 6	_Z:Not applicable	_Z:Not applicable	_Z:Not applicable	NIT:Initial education	ISC_F08:Agriculture, forestry, fisheries and veterinary	ELEC:Electricity	LOW:Very poor	E7:Central and Eastern Europe	_T:Total
3	STU_MOBILE:International (or internationally mobile) students	YR:Years	L5T8:Tertiary education	C4:General programmes	M:Male	Y11T15:11-15 years	_Z:Not applicable	INST_PRIV:Private institutions	Q5:Richest	RUR:Rural	_Z:Not applicable	_Z:Not applicable	COMP_PP:Computers for pedagogical purposes	_T:Total	S2:East Asia	NIMM:Without an immigrant background
4	NARA:Adjusted net attendance rate	GPIA:Adjusted gender parity index	L1:Primary education		_Z:Not applicable	Y25T64:25-64 years	G1:Grade 1	INST_PUB:Public institutions	Q3:Middle	URB:Urban	ADULT:Adult education	ISC_F10:Services	NET_PP:Internet for pedagogical purposes	HIGH:Very affluent	F6:Sub-Saharan Africa	IMM:With an immigrant background
5	PR:Promotion rate	WPIA:Adjusted wealth parity index	_Z:Not applicable			TH_ENTRY_AGE:Official entrance age	GLAST:Last grade		Q1:Poorest			ISC_F07:Engineering, manufacturing and construction	ADAPT_INFR_MAT_DIS:Adapted infrastructure and materials for students with disabilities		S9:South and West Asia	
6	BULLIED_STU:Bullied students	GPI:Gender parity index	L2:Lower secondary education			UNDER1_AGE:One year younger than official entry age	G2:Grade 2		_T:Total			ISC_F02:Arts and humanities	HWASH:Handwashing facilities		A8:Latin America	
7	NART:Total net attendance rate	SESPIA:Adjusted SES parity index	L02:Pre-primary education			TH_ENTRY_GLAST:Official entrance age to the last grade	G3:Grade 3		Q4:Fourth			ISC_F04:Business , administration and law			S2_O3:East Asia and the Pacific	
8	STU:Students	NB:Number	L5:Short-cycle tertiary education			UNDER_AGE:Under age	G5:Grade 5					ISC_F05:Natural sciences, mathematics and statistics			O3:Oceania	
9	GAR:Gross attendance ratio	IPIA:Adjusted immigration parity index	L2_3:Secondary education			ISC1_IN_ISC02:Primary school age in pre-primary education	G4:Grade 4					_X:Unspecified			S36:Arab States	
10	FOSEP:Distribution of students in tertiary education by field of education		L4:Post-secondary non-tertiary education			Y_GE15:15 years and over	G7:Grade 7					ISC_F03:Social sciences, journalism and information			S4:Central Asia	
11	REPR:Repetition rate		L0:Early childhood education			Y3T7:3-7 years	_U:Unknown					ISC_F01:Education			A2_E5:North America and Western Europe	
12	SLEN:School life expectancy (excluding repetition)		L6:Bachelor?s or equivalent level			Y_LT5:Less than 5 years						ISC_F06:Information and communication			A9:Latin America and the Caribbean	

												technologies				
13	GRE:Effective graduation rate		L6T8:Bachelor?s to Doctoral or equivalent level			GE2_OVER_AGE:At least 2 years over age						ISC_F09:Health and welfare				
14	NAR:Net attendance rate		L1T3:Primary and secondary education			Y_GE65:65 years and over						ISC_F05_06_07:Science, technology, engineering and mathematics				
15	TEACH:Teachers		L1T8:Primary to tertiary education			Y12T14:12-14 years						OTH_ISC_F05_06_07:Other fields than science, technology				
16	CR:Completion rate		L7:Master?s or equivalent level			TH_END_AGE:Official ending age										
17	ILLPOP:Illiterate population		_T:Total			SCH_AGE_L02:School-age population of pre-primary education										
18	GECER:Gross early childhood enrolment ratio		L1_2:Primary and lower secondary education			OVER_AGE:Over age										
19	AIR:Gross intake ratio		L3_CWD:Upper secondary education, with direct access to tertiary			CE:Compulsory school age										
20	NER:Net enrolment rate		L01:Early childhood educational development programmes			Y15T64:15-64 years										
21	TRTP:Percentage of trained teachers		L6_7:Bachelor?s to Master?s or equivalent level			Y6T11:6-11 years										
22	QTEACH:Qualified teachers		L3_PC:Upper secondary education, partial completion			Y15T17:15-17 years										
23	SCH:School		L02T3:Pre-primary education to secondary education			Y15T24:15-24 years										
24	GTVP:Distribution of enrolment by orientation of education programme		L02T4:Pre-primary to post-secondary non-tertiary education			OVER1_AGE:One year older than official entry age										
25	SLE:School life expectancy		L2T5:Secondary, post-secondary non-tertiary and short-			_Z:Not applicable										

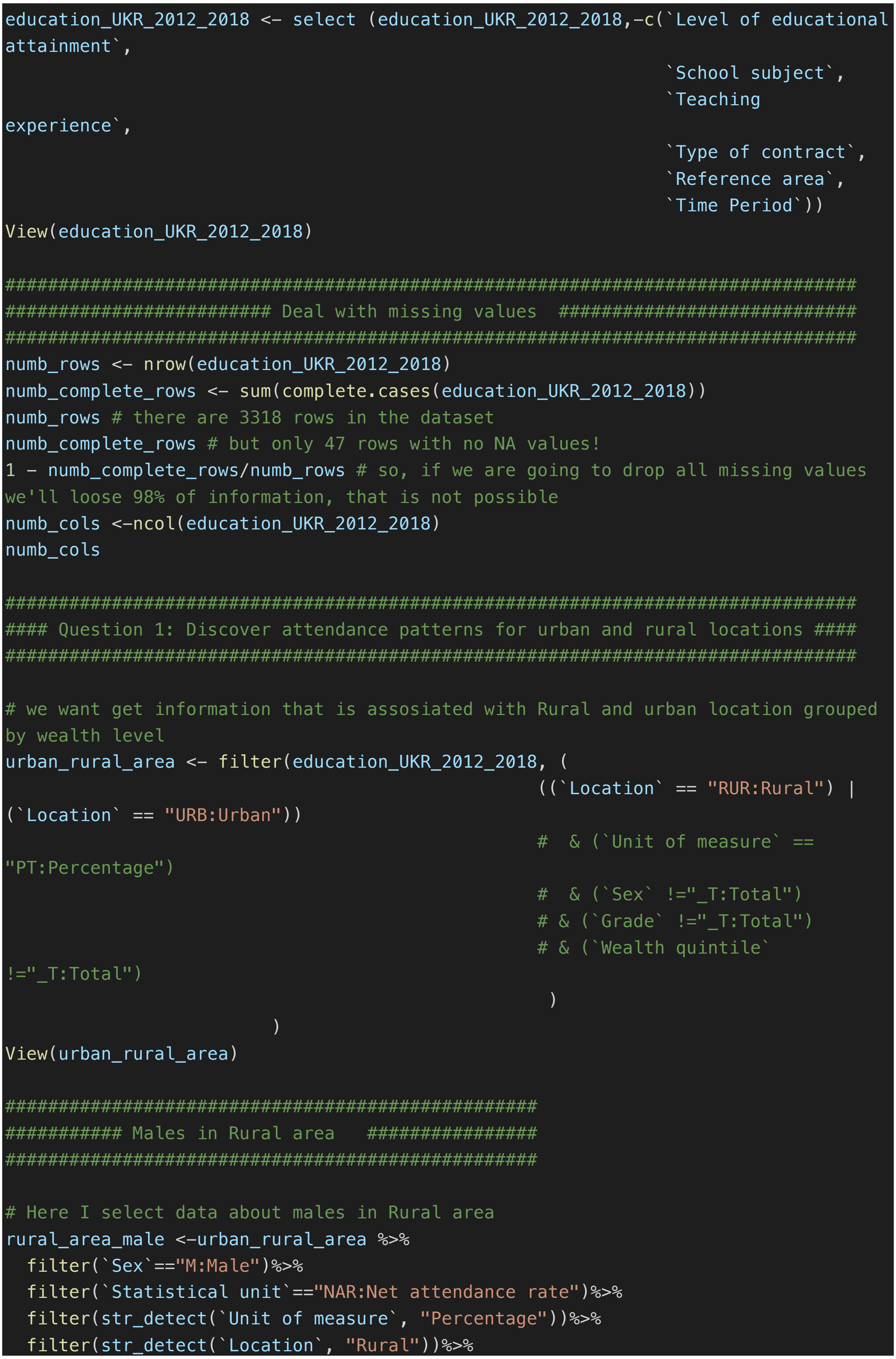
			cycle tertiary education													
26	TH_ENTRY_AGE:Official entrance age		L3_CWOUTD:Upper secondary education, without direct access to tertiary			UNDER2_AGE:Two years younger than official entry age										
27	NIRA:Adjusted net intake rate		L02T8:Pre-primary to tertiary education			Y0T7:0-7 years										
28	DR:Drop-out rate		L5T7:Short-cycle tertiary education to Master?s or equivalent level													
29	NIR:Net intake rate															
30	SR:Survival rate															
31	NENT_P:New entrants who have attended any early childhood education programme															
32	GER:Gross enrolment ratio															
33	LR:Literacy rate															
34	REPP:Percentage of repeaters															
35	PRP:Percentage of private enrolment															
36	TATTRR:Teacher attrition rate															
37	SAP:School-age population															
38	RPTR:Repeaters															
39	COMP_EDU:Compulsory education															
40	FOSGP:Distribution of graduates in tertiary education by field of education															
41	GTRANR:Gross transition ratio from secondary to tertiary education															
42	CCR:Cohort completion rate															
43	GRAD:Graduates															
44	TRTEACH:Trained teachers															
45	NENT:New entrants															
46	ECDP:Percentage of new entrants to Grade 1 of primary education with early childhood education experience															
47	ROFST:Rate of out-of-school children															
48	MOGER:Gross outbound enrolment ratio															
49	OFST:Out-of-school children															
50	NERT:Total net enrolment rate															
51	MSEP:Inbound mobility rate															

52	CHILD_TRACK_T:Children on track in health, learning and psychosocial well-being															
53	NERA:Adjusted net enrolment rate															
54	FREE_EDU:Free education															
55	PTR:Pupil-teacher ratio															
56	FGP:Percentage of female graduates															
57	TH_DUR:Theoretical duration															
58	QUTP:Percentage of qualified teachers															
59	TEP:Distribution of students in tertiary education by ISCED level															
60	CHILD_TRACK_LEARN:Children on track in learning															
61	OE:Outbound internationally mobile students															
62	TH_END_AGE_FREE_EDU:Official ending age of free education															
63	FTP:Percentage of female teachers															
64	GGR:Gross graduation ratio															
65	PTTR:Pupil/trained teacher ratio															
66	GAEP:Gross enrolment ratio in formal adult education															
67	GR:Graduation rate															
68	GAP:Graduation age population															
69	FRP:Percentage of female repeaters															
70	TH_ENTRY_AGE_FREE_EDU:Official entrance age to free education															
71	ASER:Age specific enrolment rate															
72	ESL:Early school leavers															
73	FTRTP:Percentage of female trained teachers															
74	START:Start of the academic school year															
75	MOR:Outbound mobility ratio															
76	TRANRA:Effective transition rate															
77	ATTCK:Attacks on students, personnel and institutions															
78	PQTR:Pupil/qualified teacher ratio															
79	CHILD_TRACK_HEALTH:Children on track in health															
80	END:End of the academic school year															
81	MENFR:Net flow ratio of internationally mobile															

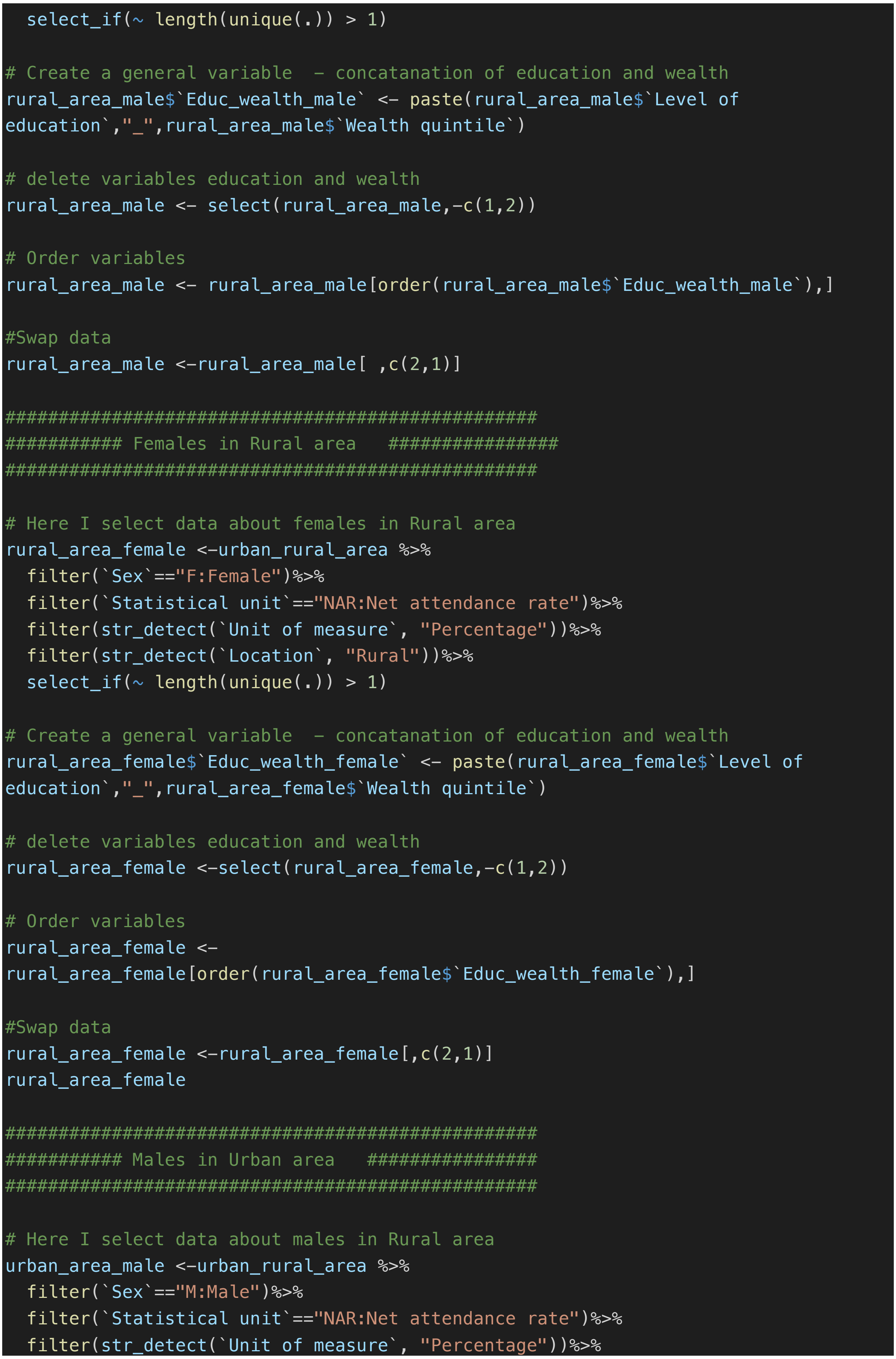
	students															
82	MENF:Net flow of internationally mobile students															
83	CHILD_TRACK_PS_WB:Children on track in psychosocial well-being															

Appendix 2. Code listing

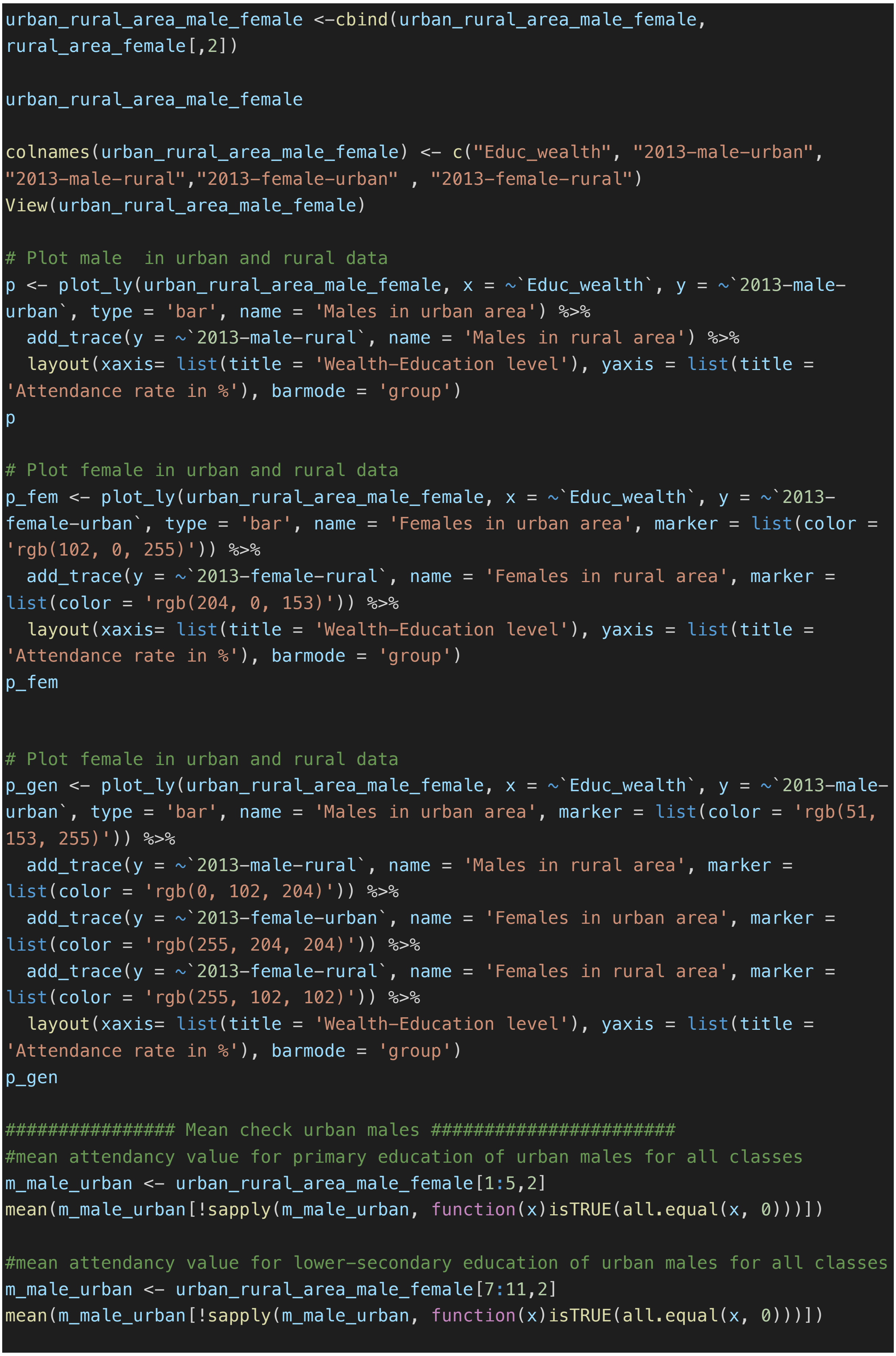
```
################################################################################
############# Header of code file. Adding all needed libraries #################
################################################################################

#install.packages("tictoc")
library(readr)
library(dplyr) # for data cleaning
library("stringr",
lib.loc="/Library/Frameworks/R.framework/Versions/3.5/Resources/library")
library("plotly",
lib.loc="/Library/Frameworks/R.framework/Versions/3.5/Resources/library")
library(naniar)
library(VIM)
library(FactoMineR)
library(missMDA)
library(cluster) # for gower similarity and pam
library(Rtsne) # for t-SNE plot
library(ggplot2) # for visualization
library(tictoc)
################################################################################
######################## Import educational dataset  ###########################
################################################################################

education_UKR_2012_2018 <- read_csv("education_UKR_2012-2018.csv")
class(education_UKR_2012_2018) #check the class of data we get
education_UKR_2012_2018 = as.data.frame(education_UKR_2012_2018) #make all data to
be a data.frame
class(education_UKR_2012_2018)

################################################################################
############################ Data understanding  ###############################
################################################################################

# first I want to understand what are the unique statistical units of each of 30
columns that I can analyse.
list_of_unique_values<- list(1:numb_cols)
list_of_unique_values
for (name in 1:numb_cols) {
  unique_value<-education_UKR_2012_2018[name]%>%
    unique()
  list_of_unique_values[[name]]<- unique_value
}
View(list_of_unique_values)

# After checking the unique values we can conclude that columns "Level of
educational attainment",
# "School subject", "Teaching experience", "Type of contract", "Reference area" and
"Time period" are redundant, so we can drop them
```



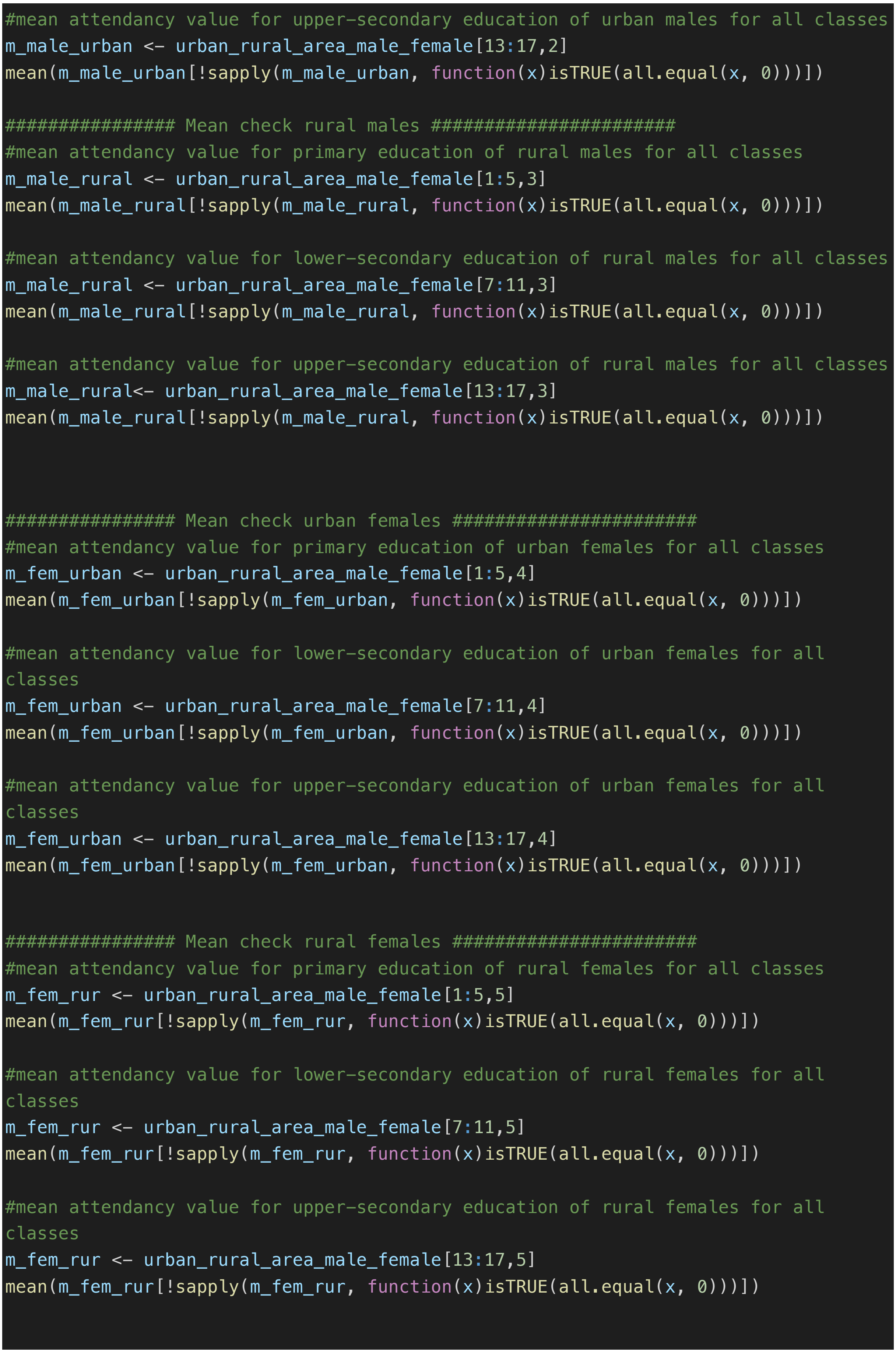
```
education_UKR_2012_2018 <- select (education_UKR_2012_2018,-c(`Level of educational
attainment`,
                                                              `School subject`,
                                                              `Teaching
experience`,
                                                              `Type of contract`,
                                                              `Reference area`,
                                                              `Time Period`))
View(education_UKR_2012_2018)

################################################################################
######################### Deal with missing values  ############################
################################################################################
numb_rows <- nrow(education_UKR_2012_2018)
numb_complete_rows <- sum(complete.cases(education_UKR_2012_2018))
numb_rows # there are 3318 rows in the dataset
numb_complete_rows # but only 47 rows with no NA values!
1 - numb_complete_rows/numb_rows # so, if we are going to drop all missing values
we'll loose 98% of information, that is not possible
numb_cols <-ncol(education_UKR_2012_2018)
numb_cols

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

# we want get information that is assosiated with Rural and urban location grouped
by wealth level
urban_rural_area <- filter(education_UKR_2012_2018, (
                                                  ((`Location` == "RUR:Rural") |
(`Location` == "URB:Urban"))
                                                  #   & (`Unit of measure` ==
"PT:Percentage")
                                                  #   & (`Sex` !="_T:Total")
                                                  # & (`Grade` !="_T:Total")
                                                  # & (`Wealth quintile`
!="_T:Total")
                                                   )
                         )
View(urban_rural_area)

##################################################
########### Males in Rural area   ################
##################################################

# Here I select data about males in Rural area
rural_area_male <-urban_rural_area %>%
  filter(`Sex`=="M:Male")%>%
  filter(`Statistical unit`=="NAR:Net attendance rate")%>%
  filter(str_detect(`Unit of measure`, "Percentage"))%>%
  filter(str_detect(`Location`, "Rural"))%>%
```



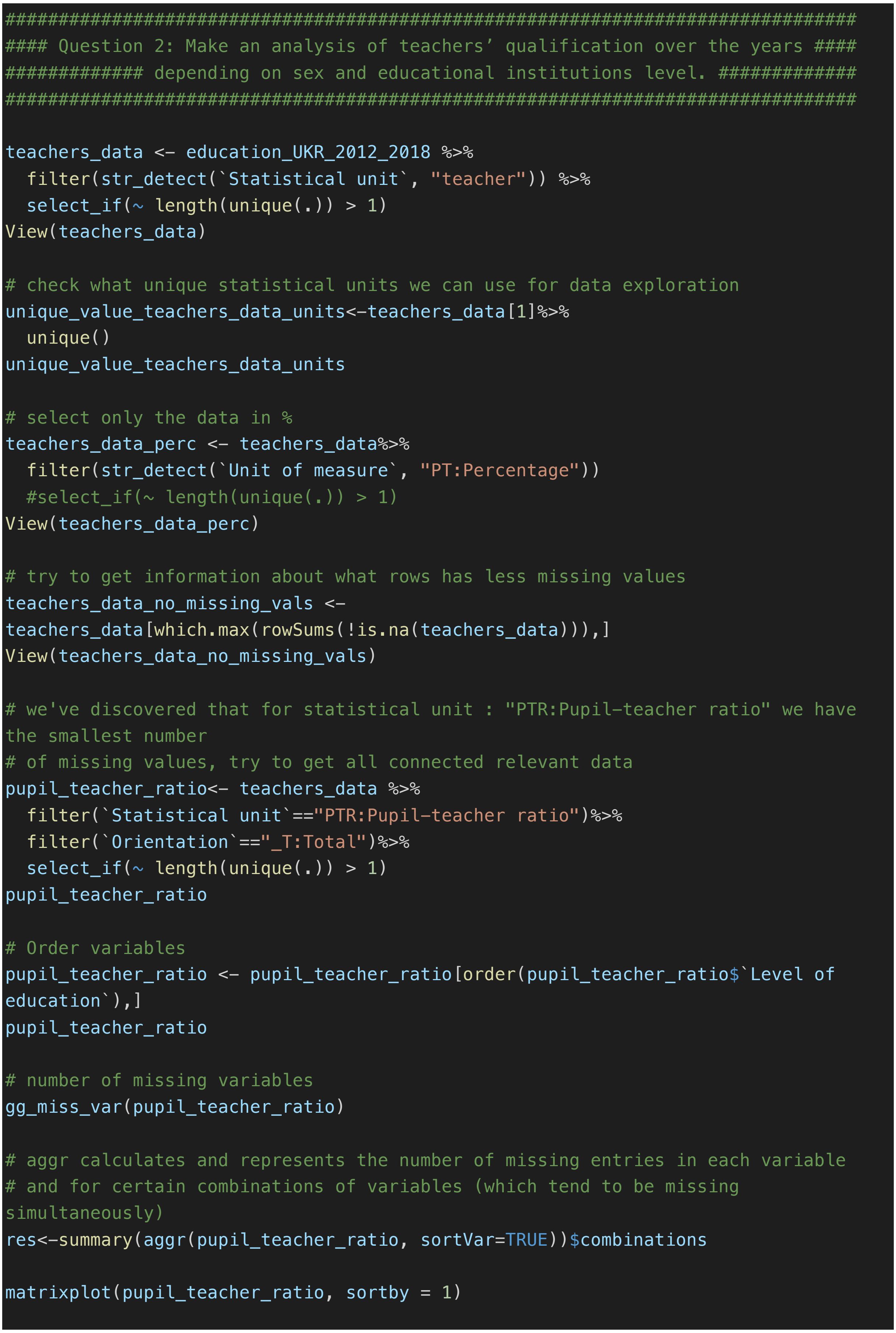
```
  select_if(~ length(unique(.)) > 1)

# Create a general variable  - concatanation of education and wealth
rural_area_male$`Educ_wealth_male` <- paste(rural_area_male$`Level of 
education`,"_",rural_area_male$`Wealth quintile`)

# delete variables education and wealth
rural_area_male <- select(rural_area_male,-c(1,2))

# Order variables
rural_area_male <- rural_area_male[order(rural_area_male$`Educ_wealth_male`),]

#Swap data
rural_area_male <-rural_area_male[ ,c(2,1)]

##################################################
########### Females in Rural area   ################
##################################################

# Here I select data about females in Rural area
rural_area_female <-urban_rural_area %>%
  filter(`Sex`=="F:Female")%>%
  filter(`Statistical unit`=="NAR:Net attendance rate")%>%
  filter(str_detect(`Unit of measure`, "Percentage"))%>%
  filter(str_detect(`Location`, "Rural"))%>%
  select_if(~ length(unique(.)) > 1)

# Create a general variable  - concatanation of education and wealth
rural_area_female$`Educ_wealth_female` <- paste(rural_area_female$`Level of 
education`,"_",rural_area_female$`Wealth quintile`)

# delete variables education and wealth
rural_area_female <-select(rural_area_female,-c(1,2))

# Order variables
rural_area_female <- 
rural_area_female[order(rural_area_female$`Educ_wealth_female`),]

#Swap data
rural_area_female <-rural_area_female[,c(2,1)]
rural_area_female

##################################################
########### Males in Urban area   ################
##################################################

# Here I select data about males in Rural area
urban_area_male <-urban_rural_area %>%
  filter(`Sex`=="M:Male")%>%
  filter(`Statistical unit`=="NAR:Net attendance rate")%>%
  filter(str_detect(`Unit of measure`, "Percentage"))%>%
  filter(str_detect(`Location`, "Urban"))%>%
  select_if(~ length(unique(.)) > 1)


# Create a general variable  – concatanation of education and wealth
urban_area_male$`Educ_wealth_male` <- paste(urban_area_male$`Level of
education`,"_",urban_area_male$`Wealth quintile`)

# delete variables education and wealth
urban_area_male <- select(urban_area_male,-c(1,2))

# Order variables
urban_area_male <- urban_area_male[order(urban_area_male$`Educ_wealth_male`),]

#Swap data
urban_area_male <-urban_area_male[,c(2,1)]
urban_area_male

##################################################
########### Females in Urban area   ################
##################################################

# Here I select data about females in Rural area
urban_area_female <-urban_rural_area %>%
  filter(`Sex`=="F:Female")%>%
  filter(`Statistical unit`=="NAR:Net attendance rate")%>%
  filter(str_detect(`Unit of measure`, "Percentage"))%>%
  filter(str_detect(`Location`, "Urban"))%>%
  select_if(~ length(unique(.)) > 1)

# Create a general variable  – concatanation of education and wealth
urban_area_female$`Educ_wealth_female` <- paste(urban_area_female$`Level of
education`,"_",urban_area_female$`Wealth quintile`)

# delete variables education and wealth
urban_area_female <-select(urban_area_female,-c(1,2))

# Order variables
urban_area_female <-
urban_area_female[order(urban_area_female$`Educ_wealth_female`),]

#Swap data
urban_area_female <-urban_area_female[,c(2,1)]
urban_area_female

# Create new data frame that combine male and female in urban and rural data
urban_rural_area_male_female <-urban_area_male
urban_rural_area_male_female <-cbind(urban_rural_area_male_female,
rural_area_male[,2])
urban_rural_area_male_female <-cbind(urban_rural_area_male_female,
urban_area_female[,2])
```



```
urban_rural_area_male_female <-cbind(urban_rural_area_male_female,
rural_area_female[,2])

urban_rural_area_male_female

colnames(urban_rural_area_male_female) <- c("Educ_wealth", "2013-male-urban",
"2013-male-rural","2013-female-urban" , "2013-female-rural")
View(urban_rural_area_male_female)

# Plot male  in urban and rural data
p <- plot_ly(urban_rural_area_male_female, x = ~`Educ_wealth`, y = ~`2013-male-
urban`, type = 'bar', name = 'Males in urban area') %>%
  add_trace(y = ~`2013-male-rural`, name = 'Males in rural area') %>%
  layout(xaxis= list(title = 'Wealth-Education level'), yaxis = list(title =
'Attendance rate in %'), barmode = 'group')
p

# Plot female in urban and rural data
p_fem <- plot_ly(urban_rural_area_male_female, x = ~`Educ_wealth`, y = ~`2013-
female-urban`, type = 'bar', name = 'Females in urban area', marker = list(color =
'rgb(102, 0, 255)')) %>%
  add_trace(y = ~`2013-female-rural`, name = 'Females in rural area', marker =
list(color = 'rgb(204, 0, 153)')) %>%
  layout(xaxis= list(title = 'Wealth-Education level'), yaxis = list(title =
'Attendance rate in %'), barmode = 'group')
p_fem


# Plot female in urban and rural data
p_gen <- plot_ly(urban_rural_area_male_female, x = ~`Educ_wealth`, y = ~`2013-male-
urban`, type = 'bar', name = 'Males in urban area', marker = list(color = 'rgb(51,
153, 255)')) %>%
  add_trace(y = ~`2013-male-rural`, name = 'Males in rural area', marker =
list(color = 'rgb(0, 102, 204)')) %>%
  add_trace(y = ~`2013-female-urban`, name = 'Females in urban area', marker =
list(color = 'rgb(255, 204, 204)')) %>%
  add_trace(y = ~`2013-female-rural`, name = 'Females in rural area', marker =
list(color = 'rgb(255, 102, 102)')) %>%
  layout(xaxis= list(title = 'Wealth-Education level'), yaxis = list(title =
'Attendance rate in %'), barmode = 'group')
p_gen

################ Mean check urban males #######################
#mean attendancy value for primary education of urban males for all classes
m_male_urban <- urban_rural_area_male_female[1:5,2]
mean(m_male_urban[!sapply(m_male_urban, function(x)isTRUE(all.equal(x, 0)))])

#mean attendancy value for lower-secondary education of urban males for all classes
m_male_urban <- urban_rural_area_male_female[7:11,2]
mean(m_male_urban[!sapply(m_male_urban, function(x)isTRUE(all.equal(x, 0)))])
```

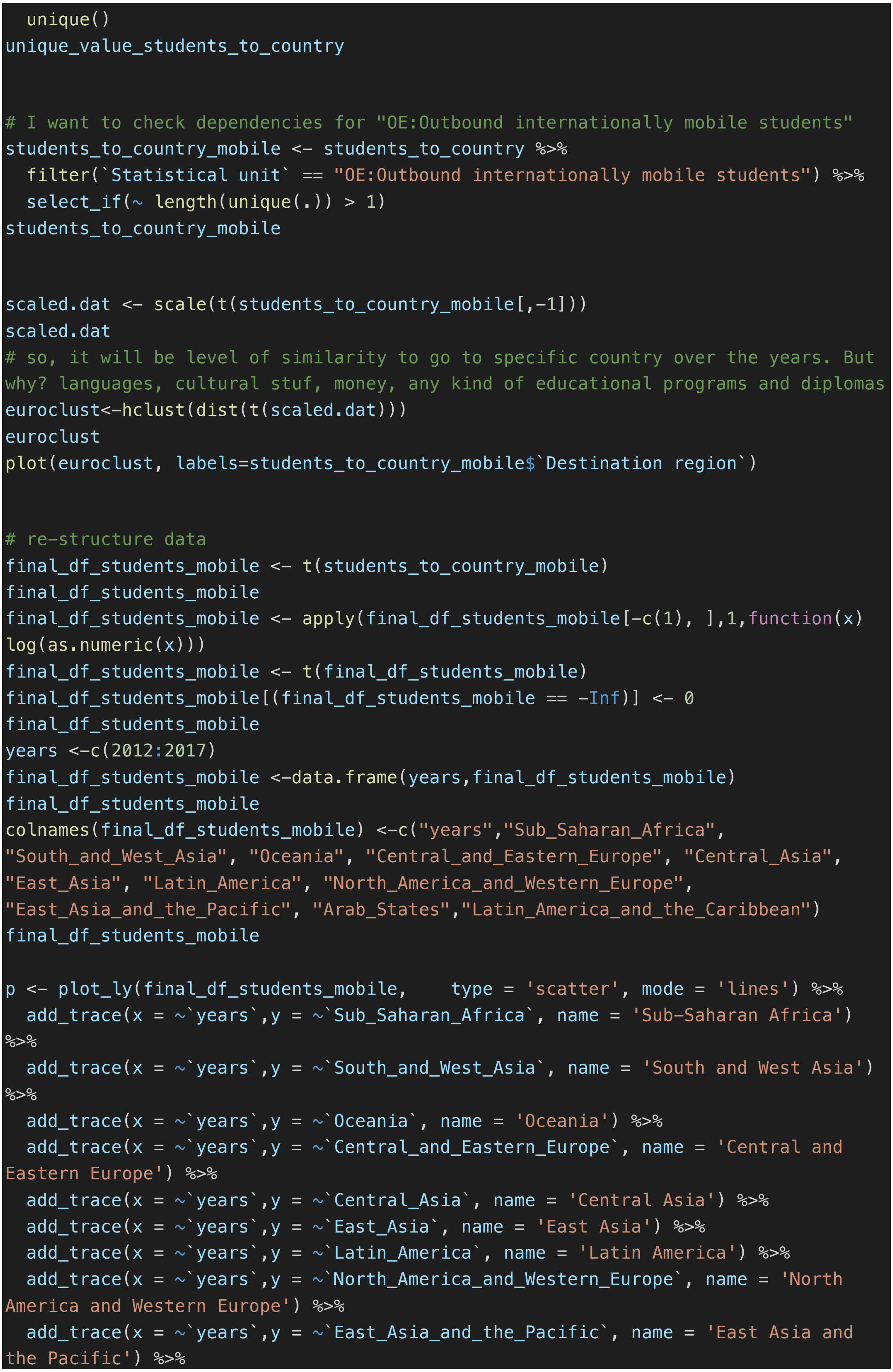


```r
#mean attendancy value for upper-secondary education of urban males for all classes
m_male_urban <- urban_rural_area_male_female[13:17,2]
mean(m_male_urban[!sapply(m_male_urban, function(x)isTRUE(all.equal(x, 0)))])

################ Mean check rural males #######################
#mean attendancy value for primary education of rural males for all classes
m_male_rural <- urban_rural_area_male_female[1:5,3]
mean(m_male_rural[!sapply(m_male_rural, function(x)isTRUE(all.equal(x, 0)))])

#mean attendancy value for lower-secondary education of rural males for all classes
m_male_rural <- urban_rural_area_male_female[7:11,3]
mean(m_male_rural[!sapply(m_male_rural, function(x)isTRUE(all.equal(x, 0)))])

#mean attendancy value for upper-secondary education of rural males for all classes
m_male_rural<- urban_rural_area_male_female[13:17,3]
mean(m_male_rural[!sapply(m_male_rural, function(x)isTRUE(all.equal(x, 0)))])



################ Mean check urban females #######################
#mean attendancy value for primary education of urban females for all classes
m_fem_urban <- urban_rural_area_male_female[1:5,4]
mean(m_fem_urban[!sapply(m_fem_urban, function(x)isTRUE(all.equal(x, 0)))])

#mean attendancy value for lower-secondary education of urban females for all
classes
m_fem_urban <- urban_rural_area_male_female[7:11,4]
mean(m_fem_urban[!sapply(m_fem_urban, function(x)isTRUE(all.equal(x, 0)))])

#mean attendancy value for upper-secondary education of urban females for all
classes
m_fem_urban <- urban_rural_area_male_female[13:17,4]
mean(m_fem_urban[!sapply(m_fem_urban, function(x)isTRUE(all.equal(x, 0)))])


################ Mean check rural females #######################
#mean attendancy value for primary education of rural females for all classes
m_fem_rur <- urban_rural_area_male_female[1:5,5]
mean(m_fem_rur[!sapply(m_fem_rur, function(x)isTRUE(all.equal(x, 0)))])

#mean attendancy value for lower-secondary education of rural females for all
classes
m_fem_rur <- urban_rural_area_male_female[7:11,5]
mean(m_fem_rur[!sapply(m_fem_rur, function(x)isTRUE(all.equal(x, 0)))])

#mean attendancy value for upper-secondary education of rural females for all
classes
m_fem_rur <- urban_rural_area_male_female[13:17,5]
mean(m_fem_rur[!sapply(m_fem_rur, function(x)isTRUE(all.equal(x, 0)))])
```



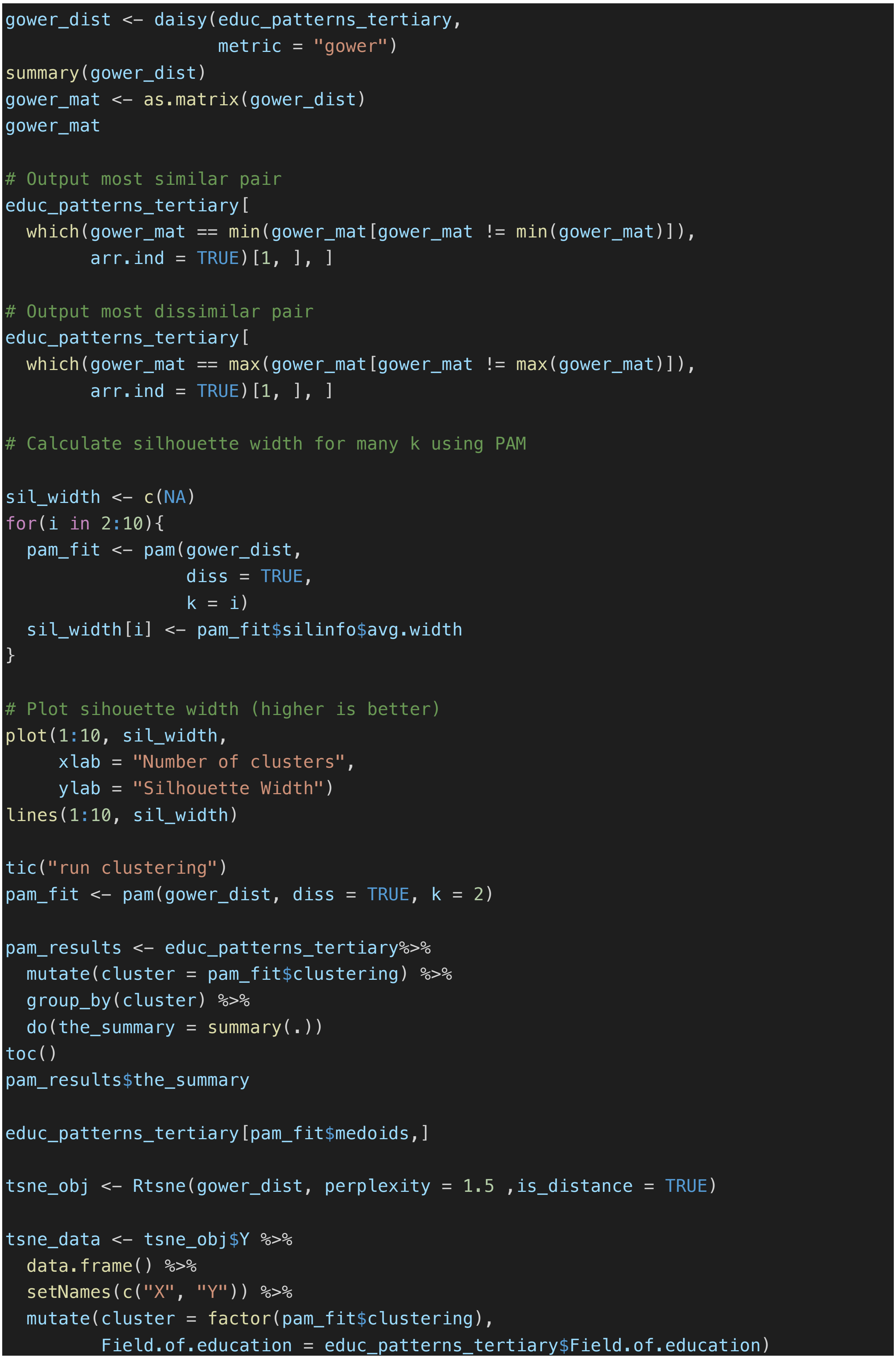
```
################################################################################
#### Question 2: Make an analysis of teachers’ qualification over the years ####
############# depending on sex and educational institutions level. #############
################################################################################

teachers_data <- education_UKR_2012_2018 %>%
  filter(str_detect(`Statistical unit`, "teacher")) %>%
  select_if(~ length(unique(.)) > 1)
View(teachers_data)

# check what unique statistical units we can use for data exploration
unique_value_teachers_data_units<-teachers_data[1]%>%
  unique()
unique_value_teachers_data_units

# select only the data in %
teachers_data_perc <- teachers_data%>%
  filter(str_detect(`Unit of measure`, "PT:Percentage"))
  #select_if(~ length(unique(.)) > 1)
View(teachers_data_perc)

# try to get information about what rows has less missing values
teachers_data_no_missing_vals <-
teachers_data[which.max(rowSums(!is.na(teachers_data))),]
View(teachers_data_no_missing_vals)

# we've discovered that for statistical unit : "PTR:Pupil-teacher ratio" we have
the smallest number
# of missing values, try to get all connected relevant data
pupil_teacher_ratio<- teachers_data %>%
  filter(`Statistical unit`=="PTR:Pupil-teacher ratio")%>%
  filter(`Orientation`=="_T:Total")%>%
  select_if(~ length(unique(.)) > 1)
pupil_teacher_ratio

# Order variables
pupil_teacher_ratio <- pupil_teacher_ratio[order(pupil_teacher_ratio$`Level of
education`),]
pupil_teacher_ratio

# number of missing variables
gg_miss_var(pupil_teacher_ratio)

# aggr calculates and represents the number of missing entries in each variable
# and for certain combinations of variables (which tend to be missing
simultaneously)
res<-summary(aggr(pupil_teacher_ratio, sortVar=TRUE))$combinations

matrixplot(pupil_teacher_ratio, sortby = 1)
```

```r
# omit rows where more than 55% of data is missing, select data about all
institutions
pupil_teacher_ratio <-
pupil_teacher_ratio[is.na(pupil_teacher_ratio)%*%rep(1,ncol(pupil_teacher_ratio))<=
ncol(pupil_teacher_ratio)*0.55,]
summary(pupil_teacher_ratio)
pupil_teacher_ratio_all_institutions<-pupil_teacher_ratio %>%
  filter(`Type of institution`=="INST_T:All institutions")%>%
  select_if(~ length(unique(.)) > 1)
pupil_teacher_ratio_all_institutions

#change all NA values to 0 values
pupil_teacher_ratio_all_institutions[is.na(pupil_teacher_ratio_all_institutions)]
<- 0

# re-structure data
final_df <- t(pupil_teacher_ratio_all_institutions)
final_df <- final_df[-c(1), ]
years <-c(2012:2017)
final_df <-data.frame(years,final_df)
colnames(final_df) <-c("years","L0:Early childhood education", "L1:Primary
education", "L2_3:Secondary education", "L5T8:Tertiary education")
final_df

p <- plot_ly(final_df,    type = 'scatter', mode = 'markers') %>%
  add_trace(x = ~`years`,y = ~`L0:Early childhood education`, name = 'L0:Early
childhood education') %>%
  add_trace(x = ~`years`,y = ~`L1:Primary education`, name = 'L1:Primary
education') %>%
  add_trace(x = ~`years`,y = ~`L2_3:Secondary education`, name = 'L2_3:Secondary
education') %>%
  add_trace(x = ~`years`,y = ~`L5T8:Tertiary education`, name = 'L5T8:Tertiary
education') %>%
  layout(
    title = " Students teachers Ratio",
      yaxis = list(title = "Ratio"))
p


################################################################################
##### 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. ##############
################################################################################

students_to_country <- education_UKR_2012_2018 %>%
  filter(`Destination region` != "W00:All countries") %>%
  select_if(~ length(unique(.)) > 1)

# check what unique statistical units we can use for data exploration
unique_value_students_to_country<-students_to_country[1]%>%
```



```
  unique()
unique_value_students_to_country


# I want to check dependencies for "OE:Outbound internationally mobile students"
students_to_country_mobile <- students_to_country %>%
  filter(`Statistical unit` == "OE:Outbound internationally mobile students") %>%
  select_if(~ length(unique(.)) > 1)
students_to_country_mobile


scaled.dat <- scale(t(students_to_country_mobile[,-1]))
scaled.dat
# so, it will be level of similarity to go to specific country over the years. But
why? languages, cultural stuf, money, any kind of educational programs and diplomas
euroclust<-hclust(dist(t(scaled.dat)))
euroclust
plot(euroclust, labels=students_to_country_mobile$`Destination region`)


# re-structure data
final_df_students_mobile <- t(students_to_country_mobile)
final_df_students_mobile
final_df_students_mobile <- apply(final_df_students_mobile[-c(1), ],1,function(x)
log(as.numeric(x)))
final_df_students_mobile <- t(final_df_students_mobile)
final_df_students_mobile[(final_df_students_mobile == -Inf)] <- 0
final_df_students_mobile
years <-c(2012:2017)
final_df_students_mobile <-data.frame(years,final_df_students_mobile)
final_df_students_mobile
colnames(final_df_students_mobile) <-c("years","Sub_Saharan_Africa",
"South_and_West_Asia", "Oceania", "Central_and_Eastern_Europe", "Central_Asia",
"East_Asia", "Latin_America", "North_America_and_Western_Europe",
"East_Asia_and_the_Pacific", "Arab_States","Latin_America_and_the_Caribbean")
final_df_students_mobile

p <- plot_ly(final_df_students_mobile,    type = 'scatter', mode = 'lines') %>%
  add_trace(x = ~`years`,y = ~`Sub_Saharan_Africa`, name = 'Sub-Saharan Africa')
%>%
  add_trace(x = ~`years`,y = ~`South_and_West_Asia`, name = 'South and West Asia')
%>%
  add_trace(x = ~`years`,y = ~`Oceania`, name = 'Oceania') %>%
  add_trace(x = ~`years`,y = ~`Central_and_Eastern_Europe`, name = 'Central and
Eastern Europe') %>%
  add_trace(x = ~`years`,y = ~`Central_Asia`, name = 'Central Asia') %>%
  add_trace(x = ~`years`,y = ~`East_Asia`, name = 'East Asia') %>%
  add_trace(x = ~`years`,y = ~`Latin_America`, name = 'Latin America') %>%
  add_trace(x = ~`years`,y = ~`North_America_and_Western_Europe`, name = 'North
America and Western Europe') %>%
  add_trace(x = ~`years`,y = ~`East_Asia_and_the_Pacific`, name = 'East Asia and
the Pacific') %>%
  add_trace(x = ~`years`,y = ~`Arab_States`, name = 'Arab States') %>%
  add_trace(x = ~`years`,y = ~`Latin_America_and_the_Caribbean`, name = 'Latin
America and the Caribbean') %>%
  layout(
    title = "International mobility",
    yaxis = list(title = "Number of departing people (log scaled)"))
p


# regression
students_to_country_mobile
students_to_country_mobile_norm <-students_to_country_mobile[-c(1)]
students_to_country_mobile_norm
students_to_country_mobile_norm <-
t(apply(students_to_country_mobile_norm,1,function(x) log(as.numeric(x))))
students_to_country_mobile_norm
scatter.smooth(x=students_to_country_mobile_norm[,1],
y=students_to_country_mobile_norm[,2], main="Latin America ~ Western Europe")

# has no meaning

students_to_country_mobile_norm <- as.data.frame(students_to_country_mobile_norm)
students_to_country_mobile_norm

final_df_students_mobile
library("ggpubr")
require(gridExtra)
plot1 <- ggscatter(final_df_students_mobile, x="years", y = "Oceania",
          add = "reg.line", conf.int = TRUE,
          cor.coef = TRUE, cor.method = "pearson",
          xlab = "years", ylab = "Oceania")
plot2 <- ggscatter(final_df_students_mobile, x="years", y =
"North_America_and_Western_Europe",
          add = "reg.line", conf.int = TRUE,
          cor.coef = TRUE, cor.method = "pearson",
          xlab = "years", ylab = "North_America_and_Western_Europe")
plot3 <- ggscatter(final_df_students_mobile, x="years", y = "Arab_States",
                   add = "reg.line", conf.int = TRUE,
                   cor.coef = TRUE, cor.method = "pearson",
                   xlab = "years", ylab = "East_Asia")
grid.arrange(plot1, plot2, plot3)



################################################################################
### Question 4: Discover patterns about preferred fields of studies for a ######
############# male/female student  over the years (2012-2018) and make a    #######
################ prognosis about 2019.##############
################################################################################

educ_patters <- education_UKR_2012_2018 %>%
```

```r
  filter(`Field of education` != "_T:Total") %>%
  filter(`Field of education` != "_X:Unspecified") %>%
  filter(`Field of education` != "_Z:Not applicable") %>%
  filter(`Sex` != "_T:Total") %>%
  select_if(~ length(unique(.)) > 1)

View(educ_patters)

educ_patterns_tertiary <- educ_patters %>%
  filter(`Statistical unit` == "FOSEP:Distribution of students in tertiary education by field of education") %>%
  select_if(~ length(unique(.)) > 1)

View(educ_patterns_tertiary)

# check what unique statistical units we can use for data exploration
unique_unit_educ_patterns_tertiary<-educ_patterns_tertiary[1]%>%
  unique()
unique_unit_educ_patterns_tertiary

unique_field_educ_patterns_tertiary<-educ_patterns_tertiary[3]%>%
  unique()
unique_field_educ_patterns_tertiary

# change char values to numbers
rows_patterns <- nrow(educ_patterns_tertiary)
rows_patterns

#Swap data
educ_patterns_tertiary <-educ_patterns_tertiary[ ,c(3,1,2,4,5,6)]
educ_patterns_tertiary

# Order variables
educ_patterns_tertiary <-
educ_patterns_tertiary[order(educ_patterns_tertiary$`Field of education`),]
educ_patterns_tertiary

#append extra variable
Value <-c(1:40)
educ_patterns_tertiary <-data.frame(Value,educ_patterns_tertiary)
educ_patterns_tertiary

glimpse(educ_patterns_tertiary)

educ_patterns_tertiary <- educ_patterns_tertiary%>%
  mutate(Field.of.education = factor(Field.of.education)) %>%
  mutate(Level.of.education = factor(Level.of.education)) %>%
  mutate(Sex = factor(Sex))
educ_patterns_tertiary
glimpse(educ_patterns_tertiary)
```



```
gower_dist <- daisy(educ_patterns_tertiary,
                    metric = "gower")
summary(gower_dist)
gower_mat <- as.matrix(gower_dist)
gower_mat

# Output most similar pair
educ_patterns_tertiary[
  which(gower_mat == min(gower_mat[gower_mat != min(gower_mat)]),
        arr.ind = TRUE)[1, ], ]

# Output most dissimilar pair
educ_patterns_tertiary[
  which(gower_mat == max(gower_mat[gower_mat != max(gower_mat)]),
        arr.ind = TRUE)[1, ], ]

# Calculate silhouette width for many k using PAM

sil_width <- c(NA)
for(i in 2:10){
  pam_fit <- pam(gower_dist,
                 diss = TRUE,
                 k = i)
  sil_width[i] <- pam_fit$silinfo$avg.width
}

# Plot sihouette width (higher is better)
plot(1:10, sil_width,
     xlab = "Number of clusters",
     ylab = "Silhouette Width")
lines(1:10, sil_width)

tic("run clustering")
pam_fit <- pam(gower_dist, diss = TRUE, k = 2)

pam_results <- educ_patterns_tertiary%>%
  mutate(cluster = pam_fit$clustering) %>%
  group_by(cluster) %>%
  do(the_summary = summary(.))
toc()
pam_results$the_summary

educ_patterns_tertiary[pam_fit$medoids,]

tsne_obj <- Rtsne(gower_dist, perplexity = 1.5 ,is_distance = TRUE)

tsne_data <- tsne_obj$Y %>%
  data.frame() %>%
  setNames(c("X", "Y")) %>%
  mutate(cluster = factor(pam_fit$clustering),
         Field.of.education = educ_patterns_tertiary$Field.of.education)
```

```
ggplot(aes(x = X, y = Y), data = tsne_data) +
  geom_point(aes(color = cluster))
```