Faculty of Mathematics and Information Science

WARSAW UNIVERSITY OF TECHNOLOGY

Group Project
Documentation: part 2

Authors: *Agata Makarewicz, Jacek Wiśniewski*

Thesis title: *Application for Analysis of the Economic Growth Indexes for European Countries*

Supervisor: *Agnieszka Jastrzębska, Ph.D. Eng.*

version 1.5

8.11.2021

Table of Contents

1 Abstract

This document contains a general design specification for the engineering group diploma thesis entitled “Application for Analysis of the Economic Growth Indexes for European Countries”. It consists of the following parts:

- Solution proposal – short description of the project flow
- Data understanding and preprocessing – detailed information about first steps in the project
- GUI design - user interface vision
- Technology selection - languages, libraries, platforms and other technologies used

As the continuation of the previous document „Group Project Documentation: part 1”, mentioned chapters provide more detailed information about the realisation of the project. The aim is to give guidance to the potential developer so that it is possible to reconstruct the process from scratch. Furthermore, there is a whole chapter dedicated to data understanding and preprocessing which are the steps in the Cross-industry standard process for data mining (CRISP-DM) method.

1.1 History of changes

Date	Author	Description	Version
30.10.2021	Agata Makarewicz	Template	1.0
1.11.2021	Agata Makarewicz	Data sources & GUI design description added	1.1
3.11.2021	Agata Makarewicz, Jacek Wiśniewski	GUI design chapter finished, collect data and data quality chapters added	1.2
5.11.2021	Agata Makarewicz, Jacek Wiśniewski	Describe, explore, construct, integrate & select data chapters added	1.3
7.11.2021	Agata Makarewicz, Jacek Wiśniewski	Solution proposal & preprocess data chapters added	1.4
8.11.2021	Agata Makarewicz, Jacek Wiśniewski	Final version for the checkpoint	1.5

2 Vocabulary

Homepage - a webpage presented after turning on the application. It will have all of the functionalities like filtering data and generating the report.

"Read about the project" page – a webpage that will present all of the information about the project, authors and contact email addresses.

Report – content from homepage consisting of charts and results of clustering algorithms with comments.

Clustering - the task of dividing a set of objects into several groups called clusters in such a way that objects within the same cluster are more similar to each other than to objects in other clusters.

Business cycle - intervals of expansion followed by a recession in economic activity. Fluctuations are usually characterized by general upswings and downturns in a span of macroeconomic variables.

Segmentation – i.e. **time-series segmentation** is a method of time-series analysis in which an input time-series is divided into sub-series (sequences) with hypothetically homogeneous statistical properties.

Model – machine learning algorithm used for clustering.

Cross-industry standard process for data mining (CRISP-DM) – process model with six phases describing standard data mining methodology

K Nearest Neighbors (KNN) – imputation algorithm. K chosen neighbours' values are used to calculate new estimates to be imputed.

Mean squared error – error measure. It is calculated by the equation $\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$

Interpolation – imputation algorithm. It has different methods specifying how the data should be imputed. For instance, the linear method fills the gap between known data linearly.

Correlation - statistical relationship between two random variables. In this project it is calculated using Pearson's coefficient, defined as follows: $\rho_{X,Y} = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y}$, where X,Y represent variables, σ - standard deviation, and *cov* stands for covariance

Correlation matrix – table containing correlation for pairs of variables

UNDP – United Nations Development Programme, a United Nations organization which is helping countries eliminate poverty, exclusion and inequalities

HDI – Human Development Index, a summary measure of health, education, and economic conditions, developed by UNDP

PWT – Penn World Table, a dataset containing many important economic indicators, developed by researchers from the University of Groningen and University of California

ISO Code – 3-letter (alpha-3) country code defined by ISO 3166-1 standard

3 Solution Proposal

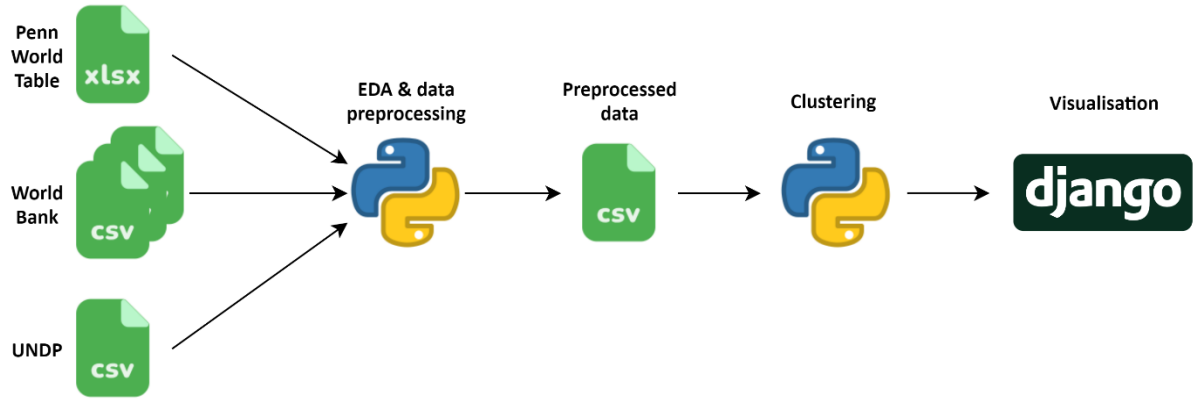

Figure 1. Solution diagram

The first step of the project is the overview of open-access data such as World Bank or Eurostat databases, in search of indicators which are useful in business cycles identification. After choosing potentially relevant data, datasets are downloaded to predefined folders, in Excel or CSV format, with a modified filename if needed. The next step is to explore collected data, verify its quality and perform necessary preprocessing. To this end, the Jupyter Notebook file is prepared. Firstly, datasets are loaded using the Pandas library and represented by data frames. Secondly, variables statistics, their distribution, missing values and correlation are analyzed in order to identify variables which need to be dropped. In parallel with this process, new variables are constructed to deal with high correlation cases. After data integration and final selection, necessary preprocessing is performed to rectify quality issues. This process includes mainly imputation of missing values, data normalization, anomaly detection and removal. All mentioned operations are performed within a single Jupyter Notebook, after running which final dataset in CSV format is returned. The next step of the project is to apply various clustering algorithms to the chosen and pre-processed time series of economic growth, in order to group European countries and verify the previously proposed divisions. Models will be implemented in a Python script, using the Scikit-learn library and fuzzy-c-means package. As an input file, the dataset returned by the previously mentioned notebook is taken. There are to be three different clustering methods implemented: k-means, hierarchical clustering and fuzzy c-means. Those algorithms will be evaluated using the existing cluster analysis assessment indexes - inertia, silhouette score, GAP statistic, PBM and Rand index – to verify the quality of performed grouping and homogeneity of obtained clusters. The analysis will cover complete time series and selected segments, predefined ones as well as identified by existing segmentation algorithms. The implemented models and their evaluation will be part of the web application with a graphical user interface written in Django. It will present the results of the work, allow user to compare the results of the methods used, select variables and parameters for the models, as well as the development indicators presented in the charts. Visualizations of the clusters obtained with different clustering methods will also be available.

4 Data understanding and preparation

4.1 Collect initial data

Some of the widest and most popular sources of open-access data are World Bank and OECD databases, and when it comes to Europe, also Eurostat. In terms of economic data, however, Penn World Table is the most established data source, and therefore suitable for this project. Checking the

data using the online viewer available on the webpage shows that only a few indicators important in business cycles identification are missing. These indicators, such as inflation and unemployment are to be found in World Bank Open Data, or other mentioned sources. However, one of the crucial indexes – the Human Development Index – is only available directly on the webpage dedicated to the report in which it is published (Human Development Report). Given the topic of this thesis, as well as the amount and quality of relevant data offered by different open sources, three following data sources have been chosen:

- **Penn World Table** – a database with indicators on relative levels of income, capital, employment, national accounts, population and productivity, covering 183 countries between 1950 and 2019 (version 10.0). It is developed and maintained by researchers from the University of California, Davis and the Groningen Growth Development Centre of the University of Groningen.
- **World Bank Open Data** – a collection of databases developed and maintained by Development Data Group of World Bank Group, containing indicators on a variety of topics, including health, climate, education, economic sectors and more. Data mainly comes from World Bank Group surveys and data collection efforts, other international organizations such as UN specialized agencies or the statistical systems of member countries.
- **Human Development Reports** – annual reports published by the Human Development Report Office of the United Nations Development Programme (UNDP). They have been released since 1990, exploring different themes through the human development approach and publishing one of the key development indicators – the Human Development Index.

OECD and Eurostat databases have also been considered but it turned out that World Bank Open Data offered the same information (indicators) for more countries and longer periods.

Loading chosen data requires a brief look at the raw Excel/CSV files to understand their structure.

- Penn World Table dataset is downloaded as an Excel file, containing data in the „Data" sheet.
- World Bank Open Data datasets are downloaded as CSV files (XML and Excel options available) within a Zipped folder alongside metadata files containing information about given indicators and countries for which its values are provided. For each indicator, files need to be extracted and data is loaded from the file with a name starting with API_SI; four rows need to be skipped.
- Human Development Index data is downloaded as a CSV file; while loading, five rows need to be skipped.

All datasets are placed in a dedicated folder, in corresponding subfolders. Penn World Table file is loaded with no changes. In the case of World Bank data, filenames are changed to the names of the indicators to make them easier to identify. Since for each indicator from this source there is a separate file, data is loaded by iterating over the dedicated folder, and filenames become variables' names in the process. Additionally, the HDI data filename is replaced by *"Human_Development_Index"* to make it more user-friendly.

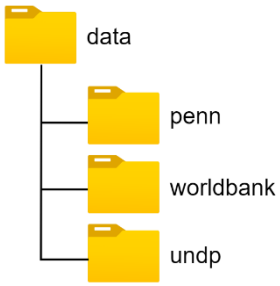

Figure 2. Data folders structure

4.2 Describe data

- **Penn World Table** is a dataset with 44 numerical columns, 8 character columns, and 12810 records. The share of missing values in the columns varies from 0% to 92% with 34 columns having 25%. Columns in this table are divided into 8 groups:
  - Identifier variables
  - Real GDP, employment and population levels
  - Current price GDP, capital and TFP
  - National accounts-based variables
  - Exchange rates and GDP price levels
  - Data information variables
  - Shares in CGDPo
  - Price levels, expenditure categories and capital
- From **World Bank Open Data** have been chosen 13 datasets with important indicators missing from the Penn World Table. Each of them has 4 identifier variables (country name, country ISO code, indicator name and code) and multiple numerical columns, each one corresponding to one year, with information about:
  - CO2 emission
  - Employment by economic sectors
  - Export and Import
  - Inflation
  - Migration
  - Population by age
  - Unemployment
  - Urban population
- Dataset downloaded from **Human Development Reports** webpage has 32 columns containing information about country name, the value of Human Development Index for every year from 1990 to 2019, and rank according to latest HDI value. It has 206 rows, every row representing one country, region, world or level of human development.

4.3 Explore data

Data described above is available for countries from all around the world, therefore to conduct insightful exploration, filtering of the European countries needs to be done at the very beginning, as they are the subject of this thesis. For that purpose provided ISO codes and country names are useful. It appears that there are 46 European countries for which any data is available. However, seven countries have any information on only five indicators, and another eight countries do not have any information on at least one indicator. These cases can lead to the problems in further steps and are described in detail in the following chapters.

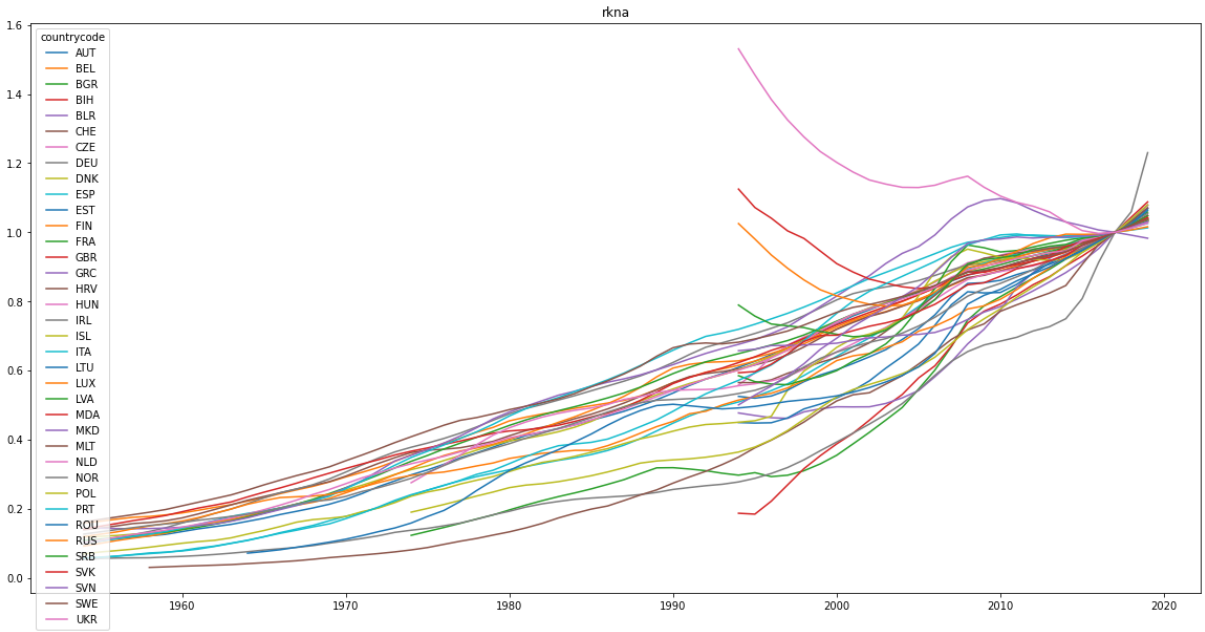

Figure 3. Plot presenting values of one of the indicators (Capital services at constant 2017 national prices). Each line represents one country, described in legend by country ISO code.

Analyzing indicators' values in time, there are several conclusions which can be drawn by just looking at the plots:

- there are 16 indicators with relative values – either value for 2017 or USA is taken as a baseline (presented above)
- most of the indicators have a lot of missing values before 1990-1995 (presented above); it is explainable because a lot of European countries were formed or gained full independence in those years

Another step in the analysis is correlation matrix calculation to investigate the dependence between indicators. There are groups of variables present which pairwise has a very high correlation coefficient value, and they are mostly the ones from PWT. Based on the data description these groups contain different versions (calculated differently) of one indicator, for instance, GDP variations.

4.4 Data quality

Data exploration showed that there are the following data quality issues, which need to be addressed:

- **Different measurement units** – datasets contain multiple indicators in different units, on different scales, for instance, the population is given in millions whereas import value in a share of GDP and HDI on scale 0-1; therefore data needs to be standardized before further processing and modelling
- **Missing values** – there are multiple missing values across analyzed datasets, mostly because some of the European countries have gained full independence (or has been formed) around 1990-1995; there is also an indicator which was not proposed until 1990 (HDI) therefore there is no previous data on it; another case is that for some countries there is no data at all on some indicators; on top of that, all those missing values are represented by different symbols, for instance, whitespace or colon; they need to be replaced by one value (for instance NaN) to obtain consistent representation

- **Different geographical entities** – World Bank data contains indicators' values not only for individual countries but also for the regions (for instance South Africa, Central Europe); for those regions, there are no officially assigned ISO codes and therefore they are not recognized by Python packages; they need to be filtered out using a dictionary of countries available in one of the packages
- **Improper location data** – one of the datasets (HDI) does not contain a column with country codes, instead of that only countries' names are provided, however, there are leading whitespaces present and unnecessary elaboration on countries' names (for instance 'The republic of') is added, which make them unrecognizable for Python packages; such data needs to be cleaned before further processing to be able to assign ISO codes
- **Different granularity** – almost all indicators' values have been collected yearly, however, for one of them (Net migration) there are only five-year estimates available; such data needs to be resampled and imputed to obtain consistent granularity

4.5 Construct data

Due to the fact, that there are many economic indicators available, there was a need to construct only two new variables:

- GDP per capita – equal to GDP value divided by population; created not to mix data sources (there is data available for this indicator, but not in Penn World Table, from which GDP and population values are taken)
- Percentage of employed – equal to the number of people engaged (employed) divided by population; created to deal with high correlation between mentioned variables

4.6 Integrate data

At this step data from each source has a column *countrycode* containing ISO 3166-1 alpha-3 codes for the analyzed countries. Data from Penn World Table and World Bank already had it, whereas, in case of HDI data from UNDP, it was added based on the cleaned column with countries' names. All three datasets are merged on the *countrycode* column into one dataset structured as shown below.

	countrycode	country	currency_unit	year	rgdpe	rgdpo	pop	emp	avh	hc	...	export	import	inflation
0	ALB	Albania	Lek	1990	12005.756836	12096.718750	3.286073	1.324078	NaN	2.516159	...	15.405064	24.031900	-0.431369
1	ALB	Albania	Lek	1991	9891.153320	10822.847656	3.280395	1.317463	NaN	2.515733	...	7.484819	28.585701	35.514247
2	ALB	Albania	Lek	1992	7674.183594	9767.368164	3.245886	1.052518	NaN	2.515308	...	12.499591	96.285881	232.984659
3	ALB	Albania	Lek	1993	9402.027344	11222.619141	3.195199	0.991653	NaN	2.514883	...	15.978830	64.539503	125.650814
4	ALB	Albania	Lek	1994	11298.698242	12672.190430	3.146519	1.068879	NaN	2.514457	...	11.983694	41.118890	35.842475

Figure 4. A table containing the first 5 rows from an integrated dataset, with multiple columned collapsed due to high dimensionality

4.7 Select data

All data combined, there are over 60 indicators available. Relevant data is chosen in a few stages.

I. Nonnumerical variables are dropped, such as *currency_unit, indicator_name* and ones from PWT's *Data information variables* group.

II. Based on the data description and literature, not all indicators from PWT are important for business cycles identification – not needed ones are dropped. Moreover, exploration and again data description shows that some of the indicators have relative values (values for 2017 are denoted as 1); such variables are also dropped.

III. The correlation matrix is calculated to investigate the dependence between indicators. Based on the values of the coefficient highly correlated pairs of variables (approximately on 0.9 and higher level) are identified and one of them is dropped. For instance, PWT provides GDP values calculated in 5 different ways, all highly dependent therefore only one of them is left for further analysis.

IV. Another important factor in the data selection process is investigating the number of missing values to verify variable's completeness and usability. There are 3 dimensions regarding which amount of missing data needs to be examined:

   a. amount of missing values for a given indicator (variables with only 60-70% of data present, or less, are dropped; imputation on such scale would lead to bias and artificial similarity of countries)

   b. amount of missing values for a given year (if there are no values on most of the indicators for a given year, it is left out of further analysis)

   c. amount of missing values for a given country (as above)

   Except for the above, there is one particular situation, that needs to be considered. Some variables, regardless of the percentage of missing values, might not have any values present for a particular country. In such a case reasonable imputation is impossible, so either variable or country must be dropped. In general, the aim is to characterize as many countries as possible with a maximum number of variables. Given the task, it is less desirable to leave a country out, because the fewer countries, the less interesting analysis. However, it is important to have a sufficient number of variables to identify cycles.

Taking all those issues into account, there are 26 indicators left, collected for 39 countries, in the years 1990-2019.

4.8 Preprocess data

Before creating the first clustering models, it is necessary to get rid of missing values. Depending on the distribution of missing values, different methods are to be used. Situations, where countries do not have any data in one variable, are described in the previous chapters. For the rest scenarios, the recommended option is to impute data. The best imputation algorithm is to be chosen by the special research. In this research, some known data are removed and then imputed. The best algorithm is the one that will achieve the lowest mean squared error. Below there are presented results of the research.

- Some countries do not have data for the last or the first measured date. In this situation, it appeared that the best imputation algorithm is to impute the last known value.

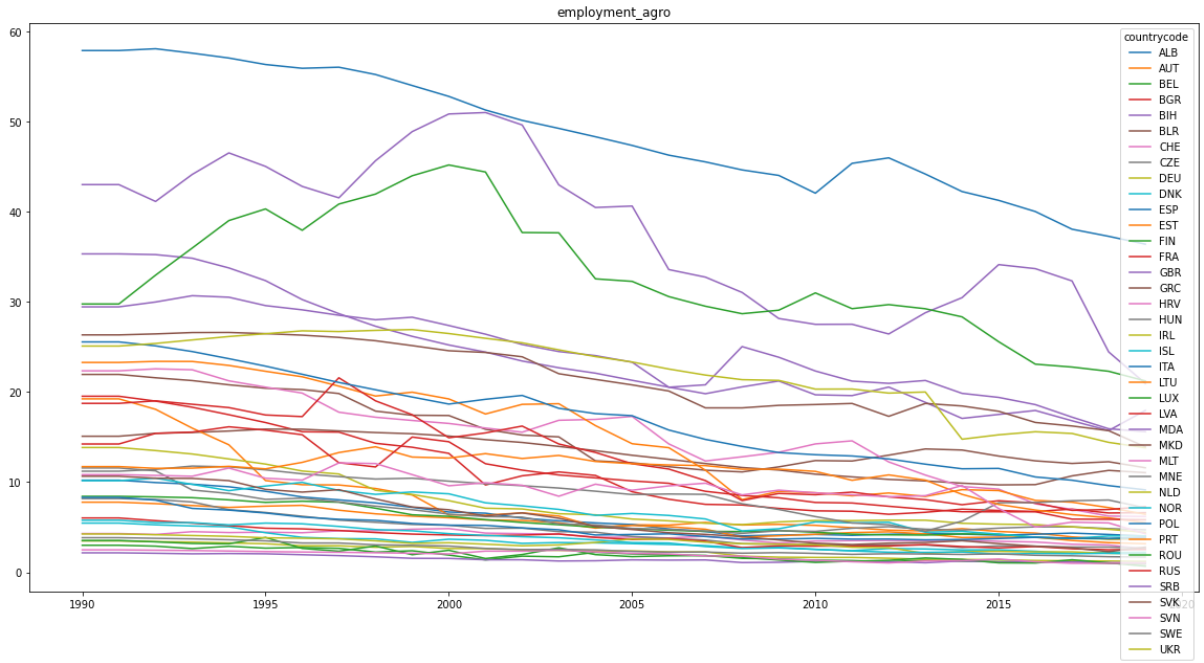

Figure 5. Data imputed in a share of employment in agriculture in 1990

- There are some indicators which started to be collected for a few countries couple of years later than for other countries. In such cases, the best option is usually to interpolate data. The only indicator that belongs to the group described above but has not been interpolated is inflation. Due to the irregular behaviour of the indicator around the 1990 year, it appeared that the best imputation option is the "K nearest neighbours" algorithm.

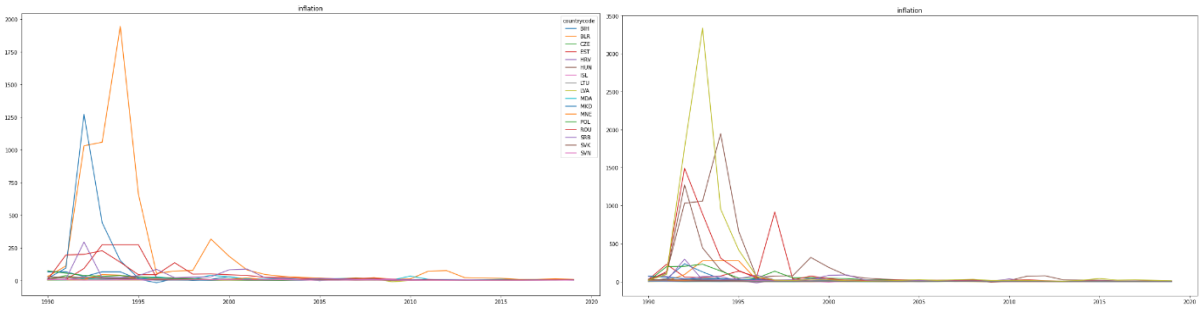

Figure 6. KNN imputation for inflation indicator. The plot on the left presents data only for countries for which imputation is performed, whereas the plot on the right – for all countries available.

- The last indicator that needed imputation is net migrations. In this scenario, data for this indicator was collected once every 5 years. To fill the gaps there is performed linear interpolation.

Another important step will be data normalization. As already mentioned in the data quality chapter, the final dataset contain multiple indicators in different units, on different scales. Therefore data needs to be normalized before passing it to clustering models, in order to unify feature range and avoid the biased contribution of indicators with greater values.

Last but not least, before the modelling phase, outlier detection needs to be performed. To find abnormal indicator values in the analyzed dataset, *IQROutlierDetector* from the *skfda* Python package is used. It is a standard method which marks as an outlier every record (in this case –

a country in terms of given indicator) which has at least one point outside of the 1.5*IQR range where IQR stands for interquartile range. Since this part of the project is not yet finalized, except for dedicated algorithms, standard and well-known models will be tested such as Isolation Forest or One-Class SVM. Given the nature of analyzed data, some outliers will probably be explainable due to countries' structural changes. If a country will be recognized as an outlier in terms of most of the indicators, there may be a need to leave it out of modelling.

5 GUI Design

The results of the work will be presented in the form of an application with a graphical user interface written in Django, which allows the user to compare the indicators and clustering results for different countries and algorithms. The application will include two web pages:

- **Homepage**
  Webpage containing:
  - section with filters enabling the user to choose variables, machine learning algorithm and its parameters to perform clustering.
  - section where the report will be displayed, with clustering results and multiple charts with which user can interact to get more insight.
- **'Read about the project' page**
  Webpage divided into two sections:
  - an introductory section containing information about the diploma thesis, such as topic, authors, supervisor and abstract.
  - an explanatory section containing a list of economic indicators available with description, as well as a list of available clustering algorithms and their parameters with description; the user can scroll through the lists or use a search box to get the needed information.

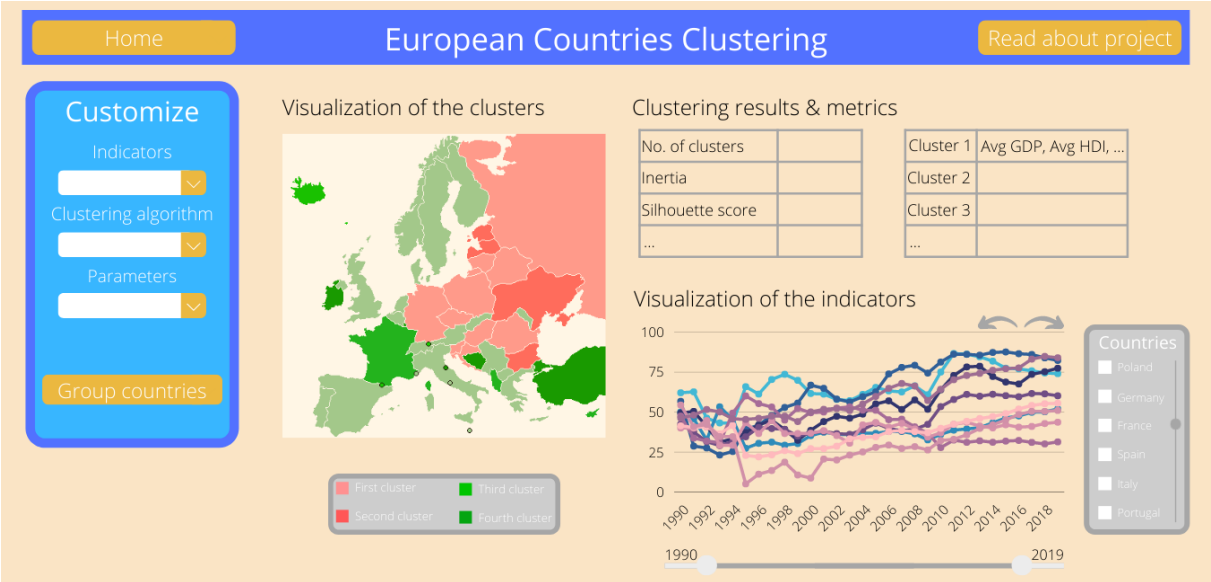

Figure 7. Homepage vision

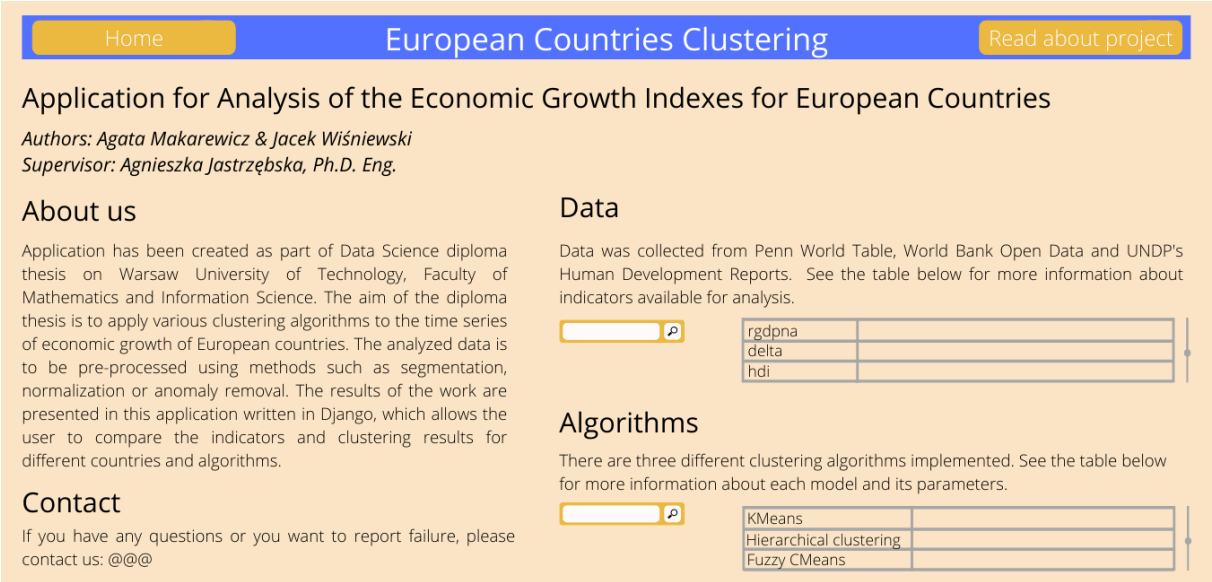

Figure 8. 'Read about the project' web page vision

6 Technology selection

- Language: Python
- Libraries: NumPy, Pandas, SciPy, Scikit-Learn, Matplotlib, Seaborn, Plotly, PyOD
- Modules & packages: pycountry, fuzzy-c-means, seglearn, statsmodels.tsa, skfda
- Framework: Django
- OS: Windows/Linux/Mac OS
- Version control: Github

7 References

[1] Kaushik, S. (2016). *Analytics Vidhya*. An Introduction to Clustering and different methods of clustering: https://www.analyticsvidhya.com/blog/2016/11/an-introduction-to-clustering-and-different-methods-of-clustering/

[2] Robert C. Feenstra, Robert Inklaar, Marcel P. Timmer. (2021). *Penn World Table*. Pobrano z lokalizacji https://www.rug.nl/ggdc/productivity/pwt/

[3] *United Nations Development Programme*. https://www.undp.org/

[4] *Wikipedia*. Business cycle: https://en.wikipedia.org/wiki/Business_cycle

8 Bibliography

[1] Aghabozorgi, Saeed, Shirkhorshidi, Ali S., and Wah, Teh Y. Time-series clustering – A decade review. Information Systems 53 16-38, 2015.

[2] Gräbner, C., Heimberger, P., Kapeller, J., and Schütz B. Structural change in times of increasing openness: assessing path dependency in European economic integration. Journal of Evolutionary Economics 30, 1467–1495, 2020.

[3] Bartlett, W. and Prica, I. Interdependence between Core and Peripheries of the European Economy: Secular Stagnation and Growth in the Western Balkans. LSE'Europe in Question' Discussion Paper Series, LEQS Paper No. 104/2016, 2016.

[4] Hamilton, James Douglas Time Series Analysis. Princeton University Press, 1994.

[5] Pal, Avishek, Prakash, PKS. Practical Time Series Analysis. Master Time Series Data Processing Visualization and Modelling Using Python. Packt, 2017

[6] Fu-Lai Chung, Tak-Chung Fu, V. Ng and R. W. P. Luk, "An evolutionary approach to pattern-based time series segmentation," in *IEEE Transactions on Evolutionary Computation*, vol. 8, no. 5, pp. 471-489, Oct. 2004.