

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 1.0

30.10.2021

Table of Contents

[1 Abstract 3](#__RefHeading___Toc2796_2030408508)

[1.1 History of changes 3](#__RefHeading___Toc800_779315115)

[2 Vocabulary 4](#__RefHeading___Toc802_779315115)

[3 Specification 5](#__RefHeading___Toc6747_630044125)

[3.1 Executive summary 5](#__RefHeading___Toc6749_630044125)

[3.2 Functional requirements 5](#__RefHeading___Toc6751_630044125)

[3.3 Non-functional requirements 7](#__RefHeading___Toc6753_630044125)

[4 Project schedule 8](#__RefHeading___Toc6755_630044125)

[5 Risk Analysis 10](#__RefHeading___Toc6757_630044125)

[6 Bibliography 11](#__RefHeading___Toc6793_630044125)

# Abstract

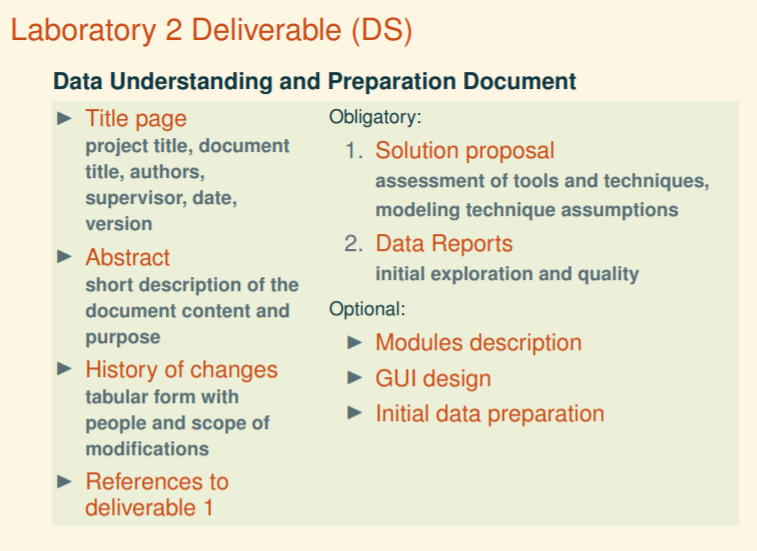
This document contains general design specification for the engineering group diploma thesis entitled “Application for Analysis of the Economic Growth Indexes for European Countries”. It consists of following parts:

* TODO
* TODO
* TODO
* GUI design - user interface vision
* TODO
* TODO

The main goal of the project is to apply several standard clustering methods such as k-means, hierarchical clustering and the fuzzy c-means method to time series of economic growth to group European countries and verify the previously proposed divisions, based on different criteria such as GDP per capita, the level of industrialization or HDI. The algorithms will be evaluated using the existing cluster analysis assessment indexes, e.g. inertia, silhouette score, GAP statistic and PBM index. The thesis will be based on publicly available data, including the Penn World Table. The selection of variables itself is one of the tasks. The analysis will cover complete time series and selected segments (e.g. before and after 2008 - the year of the last financial crisis). Another issue to examine will be the aspect of similarity of time series in the context of the assessment of synchronization or non-synchronization of business cycles of selected groups of countries before and after the crisis. The implemented models will be part of the web application in which the user will be able 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.

TODO:

Obraz zawierający stół

Opis wygenerowany automatycznie 

Obraz zawierający stół

Opis wygenerowany automatycznie

## History of changes

|  |  |  |  |
| --- | --- | --- | --- |
| **Date** | **Author** | **Description** | **Version** |
| 30.10.2021 | Agata Makarewicz | First version | 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 |
|  |  |  | 1.3 |

# 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** - is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar to each other than to those in other groups (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](https://en.wikipedia.org/wiki/Time_series#Analysis) in which an input time-series is divided into a sequence of discrete segments in order to reveal the underlying properties of its source.

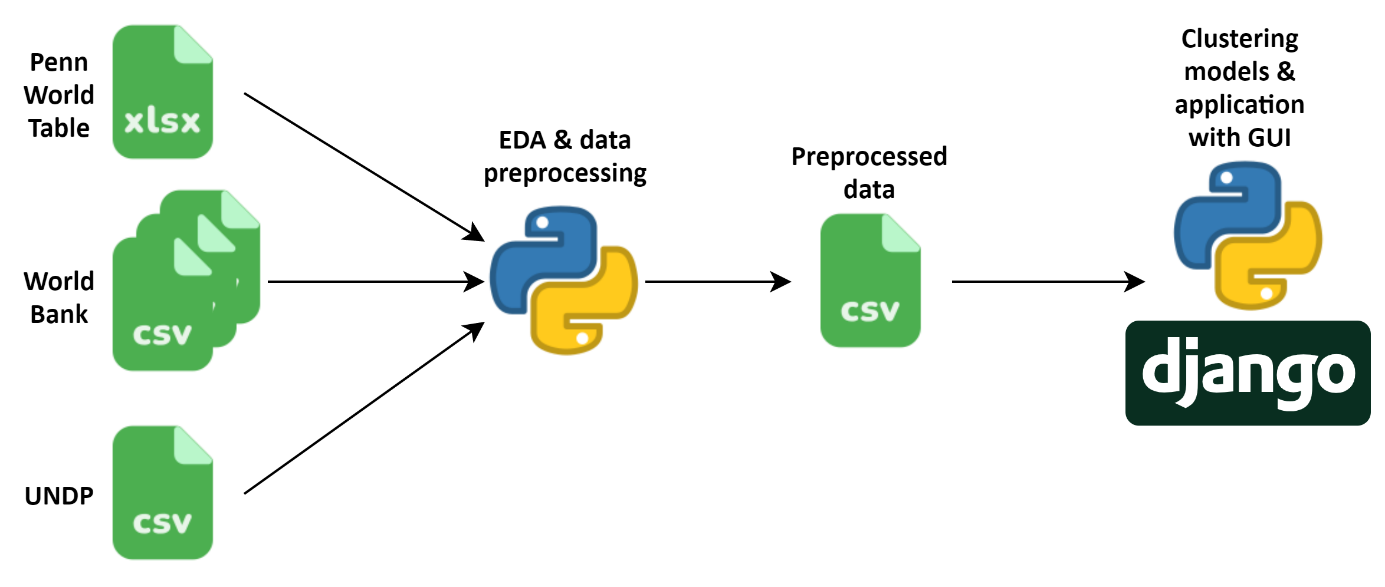
**Model** – machine learning algorithm used for clustering.

**Risk analysis** - the science of risks and their probability and evaluation. Probabilistic risk assessment is one analysis strategy usually employed in science and engineering.

**Schedule**  - a basic time-management tool, consists of a list of times at which possible tasks, events, or actions are intended to take place, or of a sequence of events in the chronological order in which such things are intended to take place.

**Vocabulary** - a set of familiar words within a person's language. A vocabulary, usually developed with age, serves as a useful and fundamental tool for communication and acquiring knowledge

# Solution Proposal



# Data reports (Data understanding & preparation)

## Collect initial data

The most popular sources of publicly available data are Penn World Table, World Bank Open Data, and OECD. After checking Penn World Table, using the online viewer available on the webpage, it appeared that only a few important indexes are missing. Missing information like inflation and unemployment was taken from World Bank Open Data. The human development index had to be taken from a separate webpage dedicated to this report. OECD did not offer more information than World Bank Open Data, therefore it was omitted in the following steps. This analysis leads to three chosen data sources:

* **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 – Human Development Index.

After data sources selection it was important to load data:

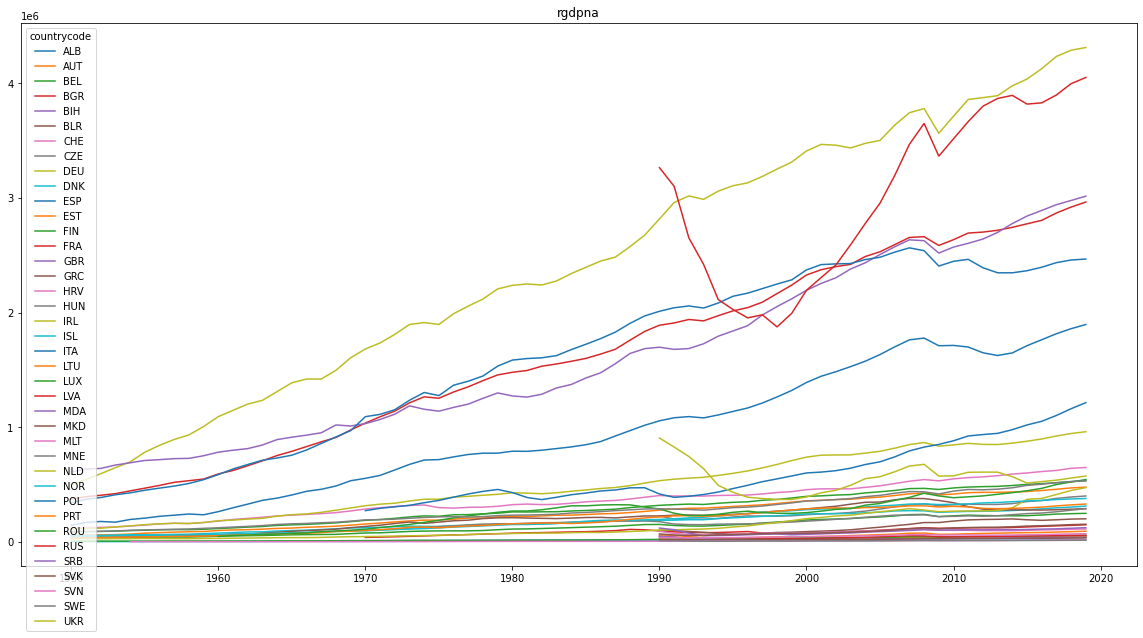
* The recommended way to use Penn World Data is to download excel file from rug.nl webpage and load content of „Data” excel page.
* To use World Bank Open Data it was necessary to download zip file for each economic index following with extracting and loading data from excel file with name starting with API\_SI. It is also important to remember about the header and footer in the files.
* TODO

## Describe data

* **Penn World Table** is a dataset with 44 numerical columns, 8 character columns, and 12810 records. Share of Null values in the columns is 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 downloaded 13 datasets with important indicators missing in the Penn World Table. Each of them has 2 identifier variables (year and country code) and numerical columns with information about:
  + CO2 emission
  + Employment by economic sectors
  + Export and Import
  + Inflation
  + Migration
  + Population by age
  + Unemployment
  + Urban population
* **Human Development Reports** is a dataset with 31 columns containing information about country name and value of Human Development Index for every year from 1990 to 2019. It has 188 rows, every row representing one country.

## Explore data

For the beginning of the data exploration it was necessary to filter European countries out of all datasets because they are the subject of this thesis. It appeared that there are 46 European countries in all mentioned before tables. 7 countries have not null information about less than 5 indicators while 8 more countries do not have any information about at least one indicator which can lead to the problems with imputation in further steps.



Most of the indicators had a lot of missing values before 1990 (like Gross Domestic Product presented above). It is explainable, because a lot of countries did not exist before this year.

TODO

## Data quality

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

* Different measurement units – datasets contain multiple indicators in different units, on different scales, for instance population is given in millions whereas import value in share of GDP and HDI on scale 0-1; therefore data needs to be standardized before further processing and modeling
* Missing values – there are multiple missing values across analyzed datasets, mostly due to the fact that 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 3166-1 codes and therefore they are not recognized by Python 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 3166-1 codes
* Different granularity – almost all indicators’ values has been collected on a yearly basis, 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

## 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)
* Percent of employed – equal to number of people engaged (employed) divided by population; created to deal with high correlation between mentioned variables

## 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 showed below.

Obraz zawierający stół

Opis wygenerowany automatycznie

## Select data

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

1. Nonnumerical variables are dropped, such as *currency\_unit, indicator\_name* and ones from PWT’s *Data information variables* group.
2. 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.
3. 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.
4. Another important factor in data selection process is investigating the amount of missing values to verify variable’s completeness and usability. There are 3 dimensions regarding which amount of missing data needs to be examined:
   1. 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)
   2. 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)
   3. 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 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 maximum number of variables. Given the task, it is less desirable to leave a country out, because the less countries, the less interesting analysis. However, it is important to have 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.

## Clean data (Preprocessing)

TODO

# GUI Design

1. 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. Application will include two webpages:

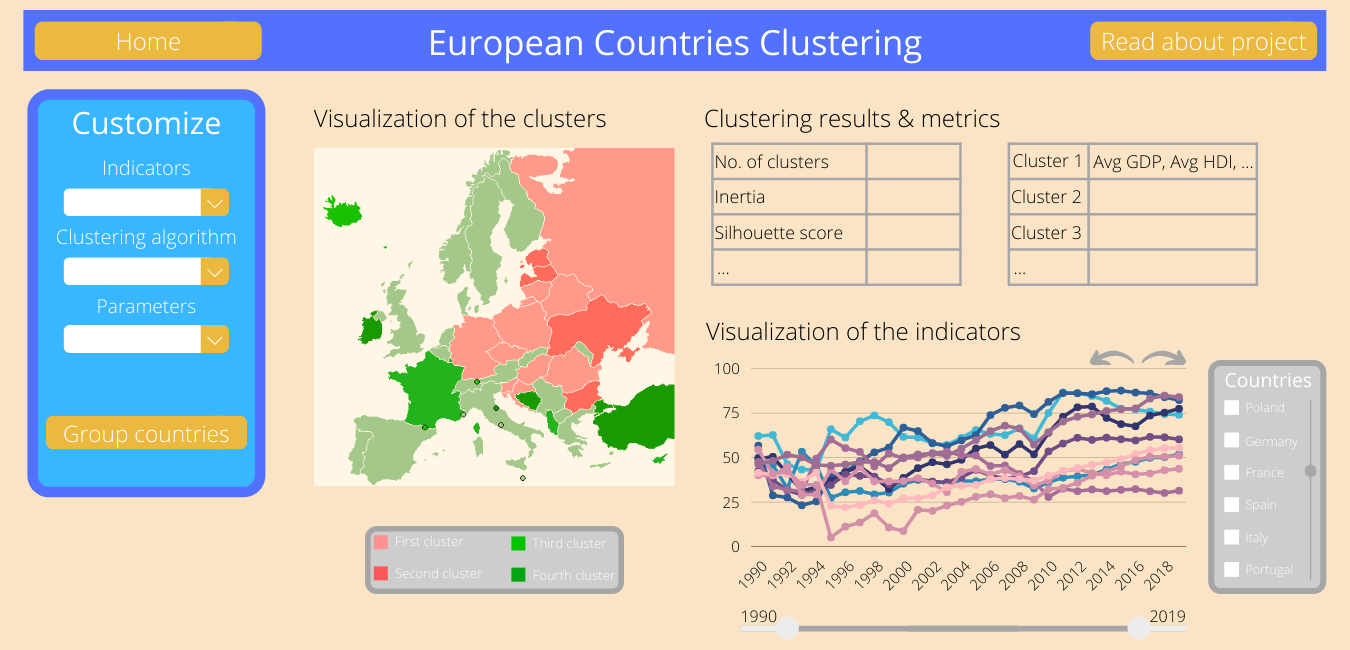
* **Homepage**

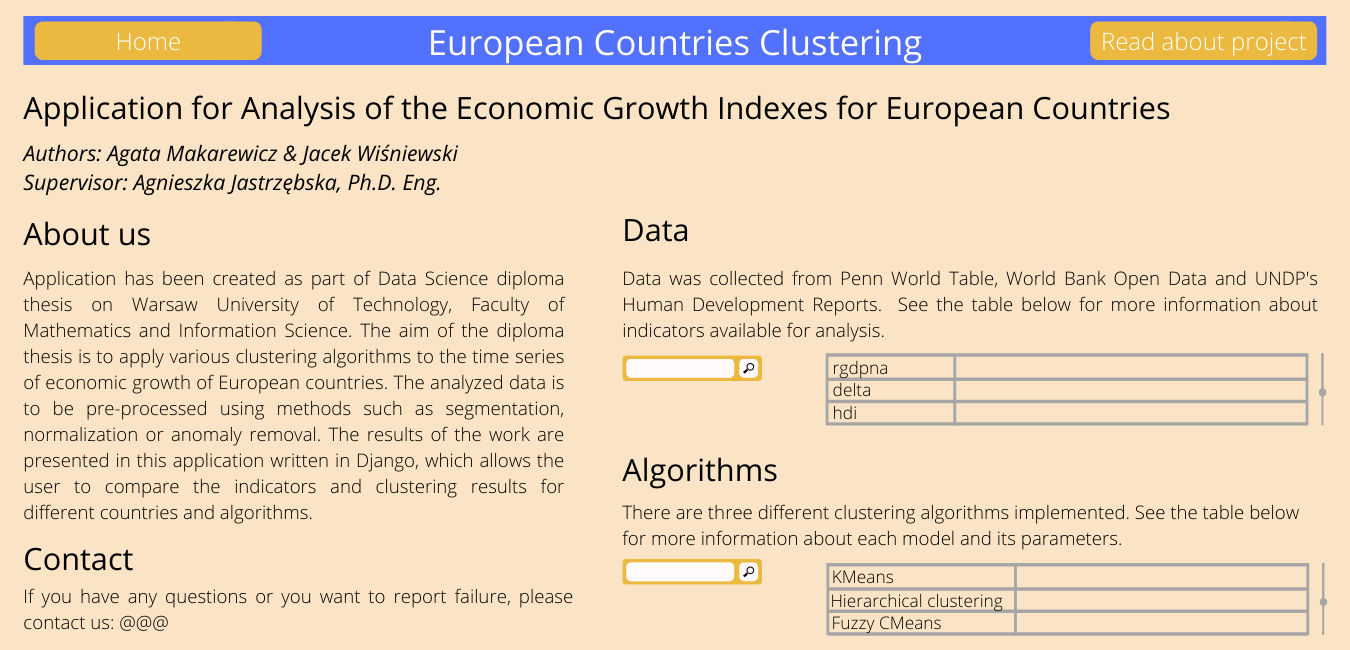
Webpage containing:

* + section with filters enabling user to choose variables, machine learning algorithm and its parameters in order to perform clustering.
  + section where 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:

* + introductory section containing information about the diploma thesis, such as topic, authors, supervisor and abstract.
  + explanatory section containing list of economic indicators available with description, as well as list of available clustering algorithms and their parameters with description; user can scroll through the lists or use a search box to get needed information.

  
*Figure x. Homepage vision*

*****Figure x. ‘Read about the project’ webpage vision*

# References

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

# References

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

# 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