DOI: https://doi.org/10.53052/9788367652018.11

Iva KOSTADINOVA[1], Vasyl MARTSENYUK[2], Georgi DIMITROV[1],
Dejan RANCIC[3], Oleksiy BYCHKOV[4], Vasil TOTEV[1],
Tomasz GANCARCZYK[2], Damian Mateusz GRYGIERCZYK[2],
Kacper PALKA[2], Mateusz DAMEK[2], Wojciech KŁÓSKO[2],
Tamara RISTOVSKA[1], Gabriela KALINOVA NAYDENOVA[1],
Genadiy PENEV GOSPODINOV[1], Slaven KRASIMIROV PANOV[1],
Andjela KOSTIC[3], Djordje ANTIC[3], Danica PEJCIC[3], Mina KRSTIC[3]

PROJEKTOWANIE I REZULTATY KRÓTKIEGO SZKOLENIA C3 Z BIG DATA: PODEJŚCIE OPARTE O KOMPETENCJE

Streszczenie: W wyniku realizacji projektu Erasmus+ nr. 2020-1-PL01-KA203-082197 pt. „Innowacje dla Big Data w realnym świecie", powstało szkolenie. Ogólnym celem tego krótkoterminowego działania szkoleniowego jest poprawa zdolności studentów do rozpoznawania koncepcji biznesowych i trudności związanych z przepływem pracy Big Data, a także studiowanie i znajdowanie kreatywnych rozwiązań problemów na dużą skalę. Trenerzy powinni kierować uczestników szkolenia do procesów biznesowych związanych z Big Data, aby studenci mogli je rozpoznać i wiedzieć, jak mogą być przetwarzane.
W tym celu został opracowany przewodnik, który poprowadzi nauczycieli, jak zorganizować kurs oparty na rzeczywistych przypadkach, aby przygotować takich specjalistów. Szkolenie jest zorientowane na kompetencje.

Keywords: iBIGworld, kurs, szkolenie, Big Data, podejście zorientowane na kompetencje, studium przypadku

DESIGN OF THE SHORT BIG DATA TRAINING COURSE C3 AND RESULTS: COMPETENCE ORIENTED APPROACH

Summary: As a result of an Erasmus+ project no. 2020-1-PL01-KA203-082197 entitled "Innovations for Big Data in a Real World", a training was created. The overall goal of this short-term training activity is to improve students' capacity to recognize business concepts and

[1] University of Library Studies and Information Technologies (ULSIT), Sofia, Bulgaria: (i.kostadinova, g.dimitrov, v.totev, 46397r, 46331r, 46011r, 46222r)@unibit.bg
[2] Department of Computer Science and Automatics, University of Bielsko-Biala, Poland: (vmartsenyuk, tgan)@ath.bielsko.pl, damiangrygiercz@gmail.com, palka99kacper@gmail.com, mateuszdamek@onet.pl, wk054421@student.ath.edu.pl
[3] University of Niš (UNi), Nis, Serbia: (dejan.rancic, endziko, djordje.antic)@elfak.rs, danicaborcina@gmail.com, mina_krstic@hotmail.com
[4] Taras Shevchenko National University of Kyiv (TSNUK), Kiev, Ukraine: oleksiibychkov@knu.ua

difficulties associated with Big Data workflow, as well as to study and find creative solutions to large-scale problems. The trainers should guide the trainees to business processes related to Big Data so that the students can recognize them and know how they can be processed.
For this purpose, a handbook has been developed to guide teachers on how to organize a course based on real cases to prepare such specialists. The training is competences oriented.

Keywords: iBIGworld, course, training, Big Data, competences oriented approach, case study

1. Introduction

Big Data processing is a complex activity involving different experts or an expert with various competencies, and the art of value extraction is the heart of Big Data Analytics. The last is a whole scientific palette of advanced methods applied, all or selected, depending on the defined problem from different fields to unlock valuable insights.
Understanding the processes related to Big Data, the skills for processing and analysis of data extracted from Big Data are key to many areas of the global economy. Through Big Data analysis, analyses of processes such as consumer preferences, analyses of the quality of a product placed on the market, analyses, and forecasts for the development of companies and many other applications can be made.
One of the most natural ways for skills acquiring is real use case study. Lately a lot of works have appeared related to application of use cases in Artificial Intelligence and Machine Learning education [1-9].
Analytics is a team sport. Data needs to be located and cleansed, models have to be created, tested, monitored, and updated. All this requires teamwork.
Working on Big Data projects is impossible without team collaboration, especially in the Analytics stage, as it interlinks all processes, depends on them, creates insights and responds for decision-making. Each activity is essential.
More or less, all of the discussed methodologies can be adapted for the purposes of different data-driven projects. Which of them to use depending on the particular analysis goals, knowledge of experts, V's characteristics of data available, and accessible technologies?
Based on the mini-course objective, what is needed to remember is the Big Data Analytics (BDA) lifecycle and data-intensive project's drivers.
As a result of an Erasmus+ project no. 2020-1-PL01-KA203-082197 entitled "Innovations for Big Data in a Real World" [10], a training was created. The aim of the project iBIGworld was to conduct a study in relation to the use and operation of Big Data and the need for the labor market for such specialists. In line with the project, the following trainings were developed:

- training for teachers,
- training for students,
- training for business representatives.

The organized training was in the form of a summer school - a seminar, where teachers and students met to exchange experience and new knowledge.

2. Students training organisation

Participants of the training activity were students, namely:
- 4 students from UBB, Poland,
- 4 students from ULSIT, Bulgaria,
- 4 students from UNi, Serbia,

and trainers:
- 4 trainers from UBB, Poland.
- 4 trainers from ULSIT, Bulgaria.
- 4 trainers from UNi, Serbia.
- 3 trainers from TSUNK, Ukraine (online).

The scope of the academis classes and workshop covers Big Data knowledge and skills development.
The scope of the workshop on analytics tools includes skills-building in Data Mining and Machine Learning for Big Data Analytics.
Format of the training consists of Lectures, Lab Sessions and Workshop.
The organization of the training includes the activities:
- Theoretical Background: Lectures, Individual Research Tasks, Quizzes,
- Practical Sessions: Live and Video Demonstrations, Guidelines, Labs, Individual and Group Tasks,
- Assessment: Final Group Project.

We apply the learning methods:
- learning-by-doing,
- case study,
- project-based learning.

The training aims to achieve two primary goals in the learning path of Big Data. Deepening the interdisciplinarity in the Big Data domain where Data Mining, Machine Learning, Data Science, and Advanced Analytics play a role as an approach palette to knowledge discovery. The training for students was organized in two stages. In the first stage, a theoretical basis in the field of Big Data was given, and in the second part, practical examples in the field were presented.
The lectures provide an overview of the Knowledge Discovery Paradigm based on Big Data, interdisciplinary links between fields, actors, and processes involved in Analytics, and potential applications, impact, and importance for business digital transformation, Industry 4.0, and Society 5.0.
Participating students had the opportunity to examine and process different types of data. In the learning process, international teams of students were formed to develop a solution together.
Accelerating skill-building in Big Data Analytics by applying supervised and unsupervised approaches for regression, classification, clustering, and feature engineering through particular software tools (Orange, Tableau) following the learning-by-doing and project-based methods. The training is competences oriented.
Below you can see the topics of the lectures:

1. Data Analytics Overview (part one)
   - Introduction,
   - Multi-Disciplinary Nature ,
   - Actors & Processes,
   - Categories,

- Methodologies,
- Applications,
- Trends.

2. Data Analytics Overview (part two)
   - Data Analytics Terminology,
   - Exploration Data Analysis (EDA) through Summary Statistics,
   - Exploration Data Analysis (EDA) through Visualization Techniques,
   - Data Quality Strategy,
   - Machine Learning for Big Data Analytics: Approaches, Techniques & Algorithms.
3. Analytics Tools Overview (part one)
   - Big Data & AI Tools Landscape,
   - Popular BDA Solutions: Cloudera, SAP HANA, SAS Viya, Alteryx, Apache Ecosystem, Azure ML,
   - Languages: R & Python.
4. Analytics Tools Overview (part two)
   - IBM Watson,
   - KNIME,
   - Orange,
   - Tableau.

3. Teacher and business representatives training – part of Big Data academic class & workshop

Theoretical aspects of Big Data were emphasized in the teacher training. The following topics were displayed:

- Approval of the Elaborated Big Data Requirements,
- Big Data Requirements and how to find the best solution to problems,
- Big Data specialists in the Data Lake ecosystem,
- iBigData framework for training in HE,
- Big Data Smart Job Hub.

4. Practical session (based on a real case study)

4.1. Orange Lab Sessions

Installation & Familiarization of Orange3 included the following stages:

- Installing the software,
- Workspace (canvas) and components (widgets) familiarization,
- Creating a workflow,
- Work with built-in datasets,
- Basic Data Exploration with Orange,
- Feature Statistics,
- Data Preparation,
- Classification,
- Regression,
- Cluster Analysis.

Loading Data in a Orange3. As soon as we open a File widget, we can load our data. It will show up in the canvas when you click on File. Double-click it to open the widget. Orange comes with many data files, and load one of them. You can, naturally, load your own data in simple steps (Fig. 1).

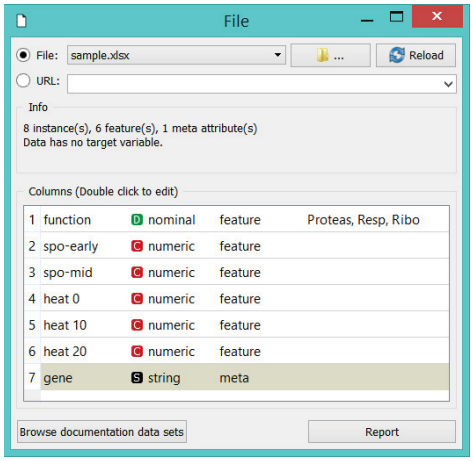

Figure 1. Loading Data in a Orange3

So, double click the File widget (Fig. 2) icon to open it, then click the file browser icon ("...") to locate the downloaded (in the current case, called sample.xlsx) file on the disk.

The File widget allows for the configuration of file types and roles. Attributes have roles (input features, meta attributes, and target/class) and can be numeric, categorical, date/time, or textual. Additionally, they can be modified via the File widget.

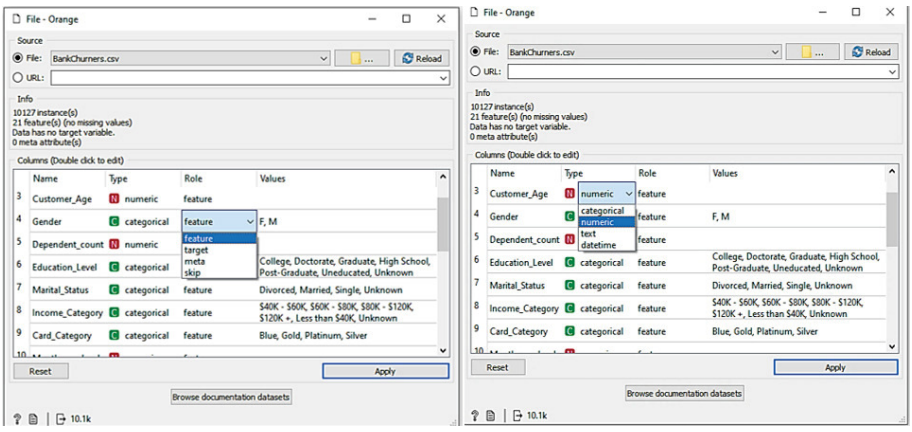

Figure 2. File widget

The data contents can be seen through the Data Table widget. For that purpose, we are connecting both widgets.

To see the contents of the Data Table (Fig. 3), double-click it:

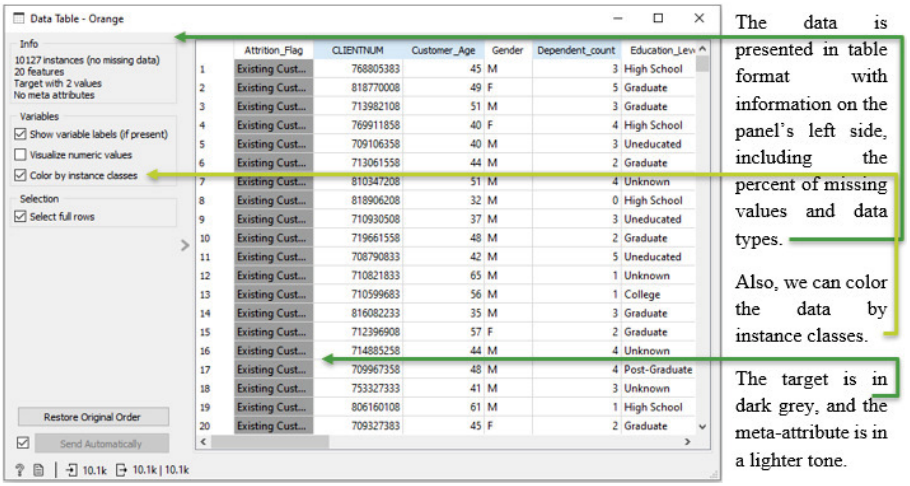

Figure 3. Data Table

Building Workflows. Analytical workflows are executed from left to right by placing and connecting widgets on the canvas. In Orange, data does not flow backward (Fig. 4).

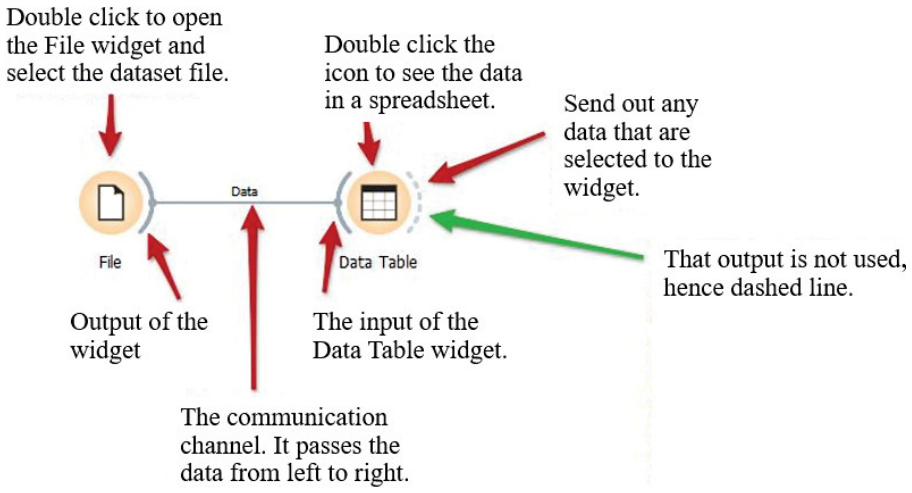

Figure 4. Building Workflows

Feature Statistics and Data Preparation. Following the instructions in the demonstration, we can see several steps through applying appropriate widgets to prepare the data for analysis and ensure data quality.

For this purpose, a prebuilt **dataset called 'heart_desease'** is used. It has a subset of 12 attributes from the Cleveland database. The 'goal' field refers to the presence of heart disease in the patient. It is integer-valued from 0 (no presence) to 1 (presence). The associated attributes are age, gender, four types of chest pain (typical Angina, atypical Angina, non-Anginal pain, Asymptomatic), values of measurement: serum cholesterol, resting systolic blood pressure, maximum heart rate. Also, is the fasting

blood sugar higher than 120 or not; are the resting electrocardiogram results normal, have left ventricular hypertrophy, or have an ST-T wave abnormality; is thalassemia described as a fixed defect (no blood flow in some parts of the heart), normal blood flow, or reversible defect (blood flow is observed but it is not normal). The slope of the peak exercise ST segment is presented in the data as upsloping, down, and downsloping. Also, the number of major vessels colored is counted from 0 to 3, and the presence or absence of exercise-induced Angina.

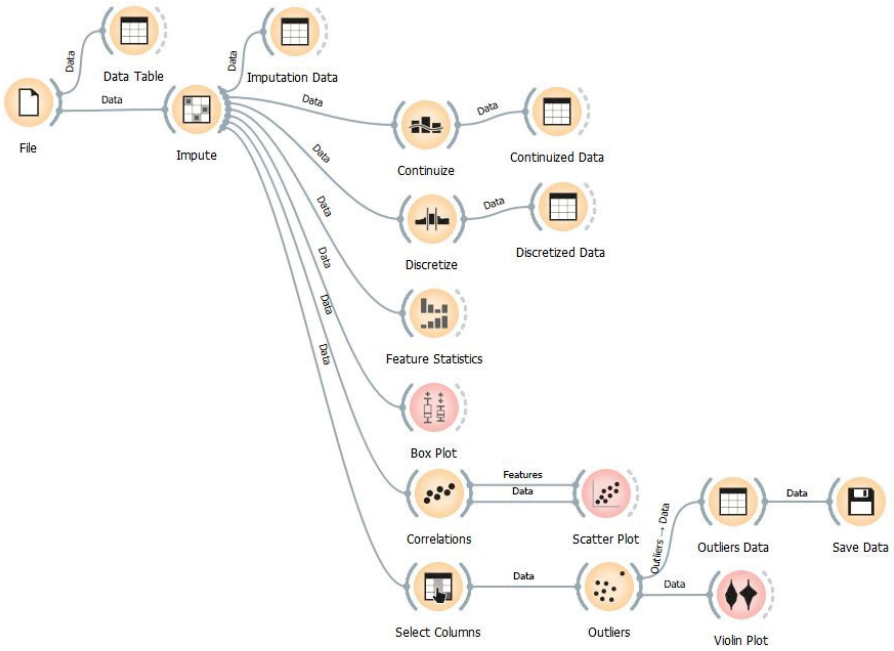

Figure 5. Workflow with different imputation data

Then students should do the following steps:

- loading the data,
- presenting the data in a tabular format (Table widget),
- observing the missing values. The question mark signals a missing value. We guess that such values are less than 0.1% of all the data. Otherwise, the information about that would be given in the File widget. Such a percent is not significant and would not provide a bias. But we have enough available values for all features to deal with the missing data,
- imputing the missing values (Impute widget) and applying the "Average or Most-frequent" method,
- specifying imputation technique. In the top-most box, "Default method", the user can specify a general imputation technique for all attributes. It is possible to specify individual treatment for each attribute, which overrides the default treatment set. The imputation methods for individual attributes are the same as the default methods. As almost all features have a few missing data of a different type (numerical and categorical), apply the Average or Most-frequent method. It

uses the average value (for continuous attributes) or the most common value (for discrete attributes),
- checking for missing values connecting the imputation output to the Table widget,
- transforming the categorical attributes into numeric (Continuize widget),
- receiving a data set in the input and outputs the same data set in which the discrete variables (including binary variables) are replaced with continuous ones:
  1. Apply "Treat as ordinal" and check the result with the Table widget. It converts the variable into a single numeric variable enumerating the original values.
  2. Normalize the numeric values (to the interval [0,1]) and check the output.
- discretizing the numeric data features (try Entropy-MDL, Equal-frequency, and Equal-width methods),
- getting helpful statistical information for features (Feature Statistics) and make conclusions,
- visualizing the imputed data with the box plot and make conclusions,
- finding the pairwise attribute correlations (Correlations widget). The widget computes Pearson or Spearman correlation scores for all pairs of features in a dataset. These methods can only detect monotonic relationships:
  1. Visualize the most correlated pair of attributes with the scatter plot.
- removing outliers from the age column (Outliers widget), save and visualize the data.

4.2. Tableau Lab Sessions

Installation & Familiarization of Orange3. Students are performing the following steps:
- Installing the software – Tableau Public,
- Data workspace and loading data,
- Using limited preprocessing functionality,
- Familiarization with visualization and analysis workspaces,
- First visual analysis,
- Exploring different visualization techniques,
- Forecasting,
- Clustering,
- Dashboards,
- Story.

Loading data and Panes. Students followed all the described below steps. The first screen is called Connect pane. Notably, the exceptional variety of data sources by structure and the ability to connect with various repositories such as IBM, Microsoft, Teradata, Spark is available. The students should do:
- downloading the Superstore data from https://public.tableau.com/. It opens a web storage of various sample data,
- loading the data to Tableau by using the Excell option in the Connect pane,
- looking at the data in the next pane called Data Source.

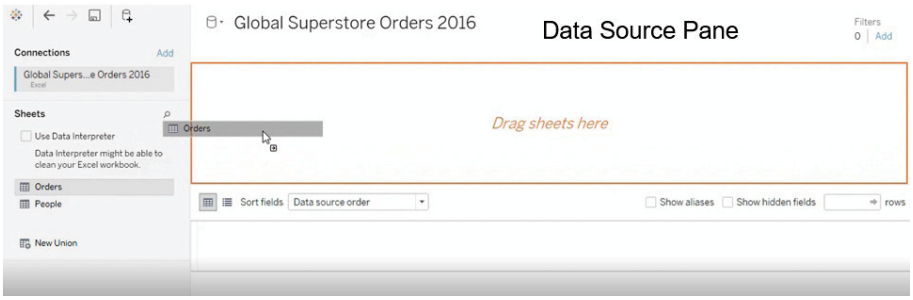

Figure 6. Loading data and Panes in Tableau

- dragging the order table to the canvas
  The primary set used for this training contains a list of worldwide company transactions described with 24 attributes: Row ID, Order Priority, Discount, Unit Price, Shipping Cost, Customer ID, Customer Name. Ship Mode, Customer Segment, Product Category, Product Subcategory, Product Container, Product Name, Product Base Margin, Country, Region, State or Province, City, Postal Code, Order Date, Ship Date, Profit, Quantity ordered new, Sales, Order ID.

Preprocessing functionality includes:

- extracting more information from the same data source. Drag the other table onto the workspace,
- integrate the data by adding a connection to the other source. A text file of returned orders is saved in a CSV format,
- editting the join in the appropriate icon.

We choose a left join, so we get all the information from the orders table and only relevant returns information. It's already based on order id as the join clause, but we could change this if desired.

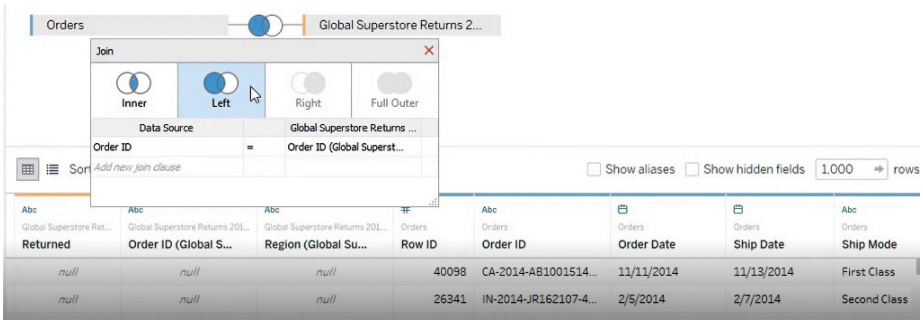

Figure 7. Preprocessing functionality in Tableau

In the integrated table, the join parts are coded in color. The order return data is in the orange line, and the information on all transactions is marked with a blue line. In this grid view, we can do some essential metadata management.

- splitting the order ID field. It has multiple parts, such as the code of distribution center, the year and two additional codes. Use custom split and a split on a hyphen. Rename the field to a distribution center.

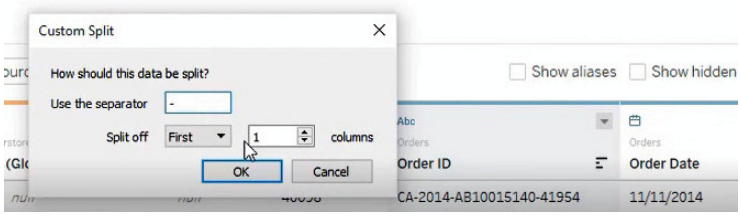

Figure 8. Split in Tableau

- connecting to Live and click on the sheet tab down at the bottom.

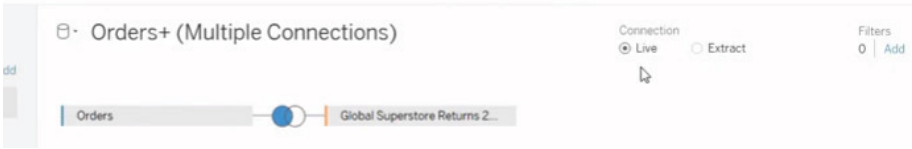

Figure 9. Connect to Live in Tableau

Another option in the Data Source Pane is the kind of connection to the data we desire: live or extract the data. Connecting live is excellent when we have constantly changed data or want to leverage the high-performance database we are connected to. Alternatively, we may choose to import data into Tableau's fast engine with an extract that takes the data offline and minimizes performance impact while still allowing regularly scheduled refreshes to keep the data up to date.

Dimensions and Measures are Tableau's way of distinguishing Categorical and Numerical features from the dataset. After going to the worksheet, we see dimensions being categorical features responsible for a graph's different dimensions or axes, and measures being the continuous values representing a datapoint plotted along an axis.

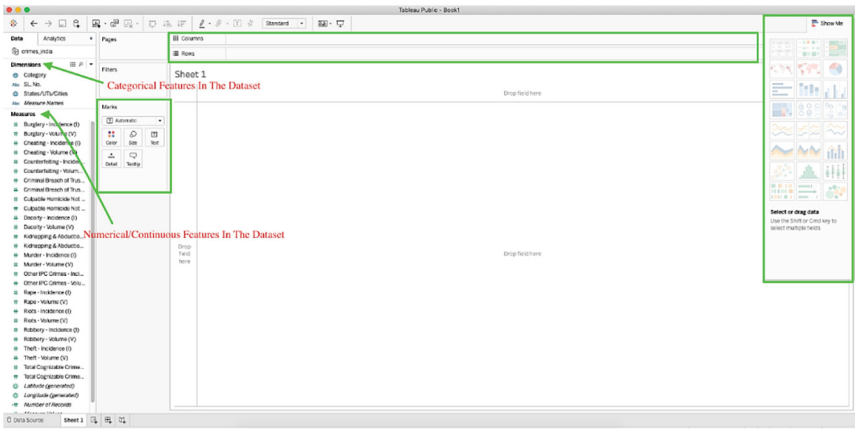

Figure 10. Tableau project

Students should bring category to rows, segment to rows, quantity to columns, market to columns and bring market to color.

Visual Analysis & Techniques. It's that easy to visualize how the sales look per category, customer segment and market. We can quickly see that Africa is an emerging market.

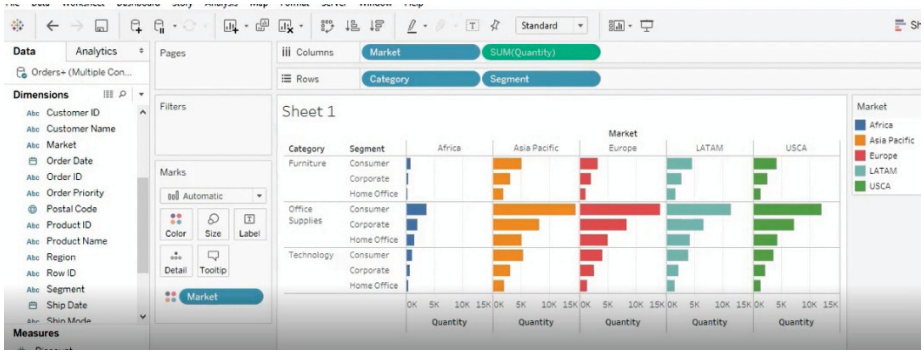

Figure 11. Visual Analysis & Techniques in Tableau

The left pane is broken up into dimensions and measures that represent the column headers in the excel sheet
In this case, the dimensions are categorical fields such as date, customer and category. They are often discrete fields and create labels in the chart also are colored in blue. The measures, on the other hand, are our metrics. They are the numbers we want to analyze. Measures are often continuous fields that create axes in the chart and are colored green. Students should do the following steps:
- comparing what quarterly growth looks like over the years,
- presenting as a cross-table the following visualization of Sales Seasonality,

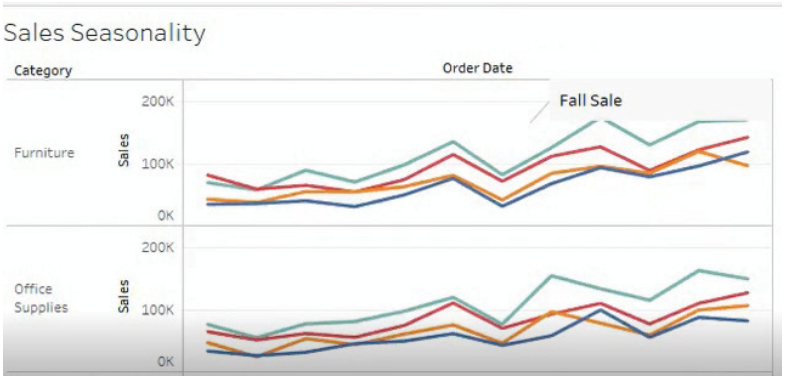

Figure 12. Present as a cross-table the following visualization of Sales Seasonality in Tableau

- exporting the table in an Excel spreadsheet,

- activating the map window from Show Me Pane and load the data on profit and location where the sale took place in the canvas. Indicate the area to which the settlement belongs. Use the size shadow and color settings to present the data more understandably. Different colors should be according to the sales profit.

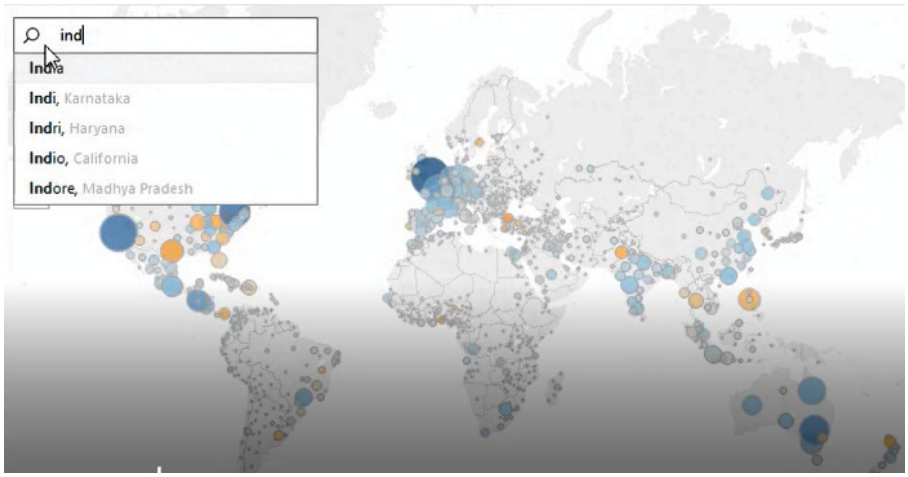

Figure 13. Activate the map window from Show Me Pane in Tableau

The map is interactive, and we can choose the desired area or location.

4.3. Resourses and steps of training

The training is based on the usage of the following resources:

- specially designed videos or live demonstrations illustrating all tasks for the Lab Sessions,
- primary steps in training (for learners),
- Lab Session Notes for trainers,
- Workflows of tasks,
- Prebuild datasets and Links to data.

In turn, we can summarize the following steps of training:

- Installation,
- Choice widget and canvas,
- Loading data,
- Building workflows,
- Basic data exploration,
- Features statistics and data preparation,
- Preprocessing in one widget,
- Prediction,
- Model testing and evaluation,
- Clustering,
- Forcasting,
- Regresion,
- Visual analysis and techniques.

5. Results

Training has included the workshop, which consists of morning sessions for team building & final projects objective and tasks and afternoon session for work in teams and preparing the project in informal environment. Final projects were presented in morning session in such a way.
Team 1 has presented the project dealing with the development of Kaggle+Python workflow for Big Data use case on Bank churners.
Team 2 has reported on the project aiming to construct Big Data workflow for the purpose of binary classification of COVID x-ray data.
Team 3 has presented the project devoted to the visualization facilities of Big Data.
Team 4 has displayed the project devoted to the problem of developing workflow for regression and cauterization of Big Data. The data from the Human Development Index were used.

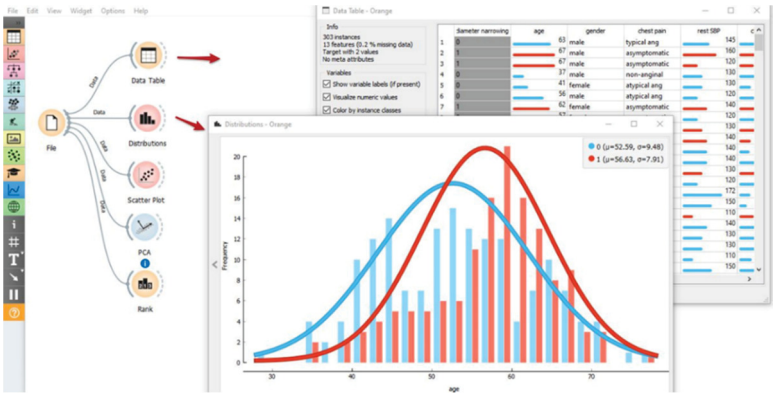

Figure 13. Basic data exploration in Orange

As a result of training the following competences were achieved:
- ability to select an efficient algorithm(s) for Big Data problem, which takes into consideration the scale,
- ability to model, analyze, and evaluate an organization's business processes,
- capability to choose the best sampling and filtering method(s) for a given Big Data analysis case,
- effectively use a variety of data analytics techniques (Machine Learning, Data Mining, Prescriptive and Predictive Analytics),
- apply quantitative techniques (statistics, time series analysis, optimization, and prediction),
- using a wide range of Big Data analytics platforms.

Students obtained the skills:
- capable of quickly adapting activities to new technologies,
- able to perform an objective analysis of a data-driven problem and take appropriate actions to solve it through analytics tools,
- compare analytics tools and specify differences between them by purpose, features, application domain, limitations and training,
- identify, compare, and apply open-source and automated machine learning data analytics tool(s),

- select and apply the most appropriate analytics tool(s) for a specific data-driven problem,
- critically assess the data source, usefulness, and potential problems associated with the data,
- upload, edit, save, and export data using analytics tools,
- assure data quality through analytics tools,
- apply and fit ML techniques to the analytical problem using the appropriate tool (s),
- apply adequate model evaluation metrics and accurately interpret the analytics output,
- use analytics tools for data visualization to present concepts/ideas/phenomena from a new perspective to decision-makers.

6. Conclusions

Big Data processing is a complex activity involving different experts or an expert with various competencies, and the art of value extraction is the heart of Big Data Analytics. The last is a whole scientific palette of advanced methods applied, all or selected, depending on the defined problem from different fields to unlock valuable insights.
Understanding the processes related to Big Data, the skills for processing and analysis of data extracted from Big Data are key to many areas of the global economy.
This training is to improve students' capacity to recognize business concepts and difficulties associated with Big Data workflow, as well as to study and find creative solutions to large-scale problems. The trainers guide the trainees to business processes related to Big Data so that the students can recognize them and know how they can be processed. The training is competences oriented.

REFERENCES

1. IBANEZ M.-B., DI-SERIO A., DELGADO-KLOOS C.: Gamification for engaging computer science students in learning activities: A case study." IEEE Transactions on learning technologies 7.3 (2014): 291-301.
2. Webpage: *https://www.nagarro.com/en/blog/ai-ml-education-real-life-use-cases*
3. Webpage: *https://www.technavio.com/report/artificial-intelligence-market-in-the-us-education-sector-analysis-share-2018?tnplus*
4. Webpage: *http://dilab.gatech.edu/a-suite-of-online-learning-tools/*
5. Webpage: *https://www.forbes.com/sites/ilkerkoksal/2018/04/17/ai-has-already-started-reshaping-the-special-education/#530adcf229d5*
6. Webpage: *https://dl.acm.org/doi/10.1145/3298689.3347030*
7. Webpage: *https://fred.stlouisfed.org/series/SLOAS#0*
8. Webpage: *https://www.worldometers.info/gdp/gdp-by-country/*
9. Webpage: *https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3510012*
10. iBigWorld: Innovations for Big Data in a Real World (Erasmus+ project 2020-1-PL01-KA203-082197) documentation *https://ibigworld.ath.edu.pl/index.php/en/home-english/*