UN
Global Working Group
on Big Data for Official Statistics

Task Team on Training, Competencies and Capacity Development

Competency Framework

for big data acquisition and processing

Table of Contents

1. Background

Dynamic socio-economic changes which originated *inter alia,* from progressive and ubiquitous digitalization of most areas of life, have tremendously transformed the data environment. We can now speak of a *data revolution* at every stage of data management and processing, and a vibrant *data industry* where private entities act as data owners at an unprecedented scale. One of the most visible manifestations of the new circumstances is a shift in data users' expectations. There is now increasing demand from commercial entities, the public and government, for real-time information, that goes beyond traditional statistical production. It is worth noting that the pace at which data is collected, processed and made available is often key to stakeholders.

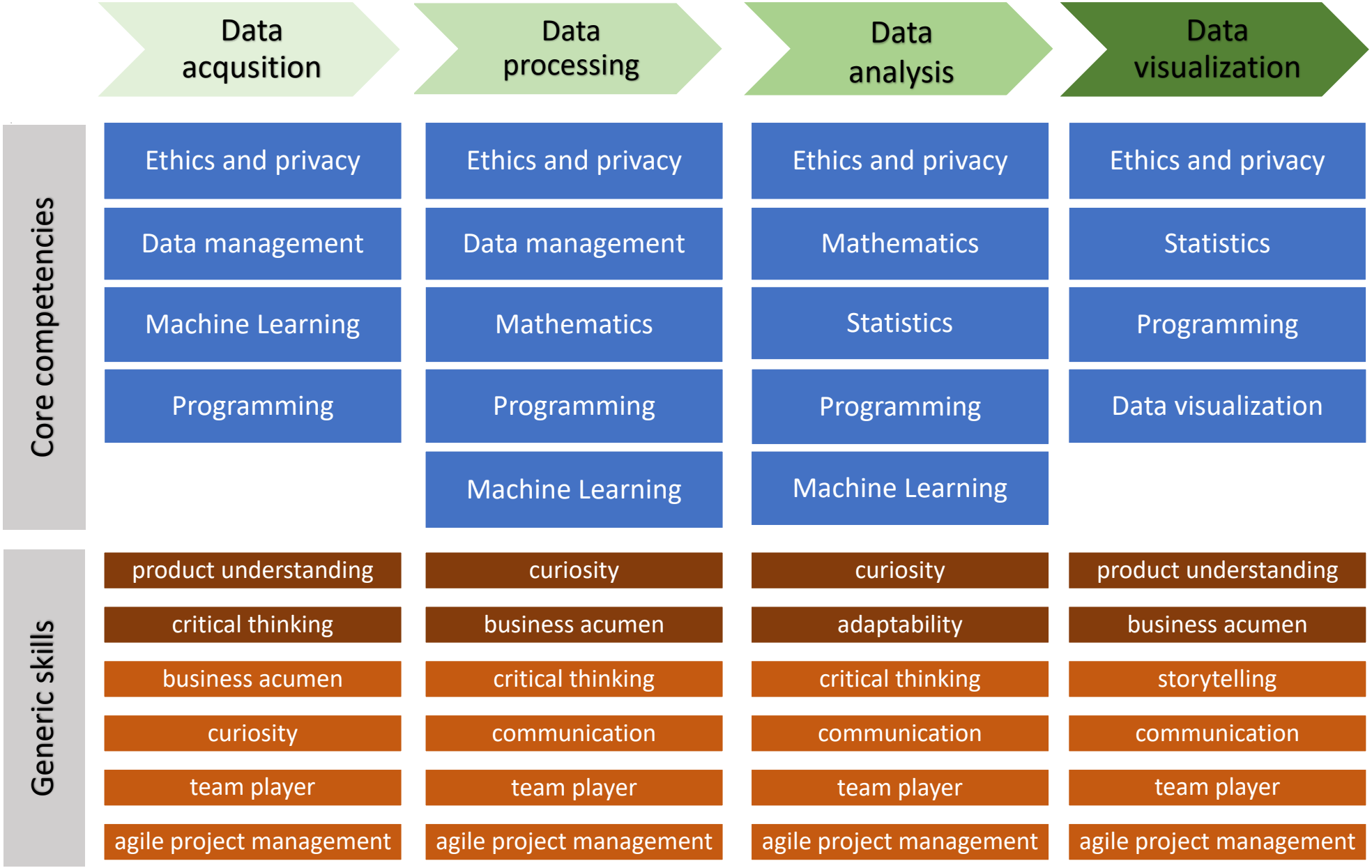
The above-stated circumstances impose immense pressure on official statistics. On the one hand, its role to ensure the highest standards and quality of statistical information becomes even more vital in the era of *fake news* and *post-truth*. On the other hand, it is expected to keep up with the growing demands of data users. To this end, attempts to modernize statistical production have been increasingly undertaken by national statistical organizations (NSOs), since they recognize the potential of novel data processing techniques and new data sources. Among the latter, big data have been of particular interest to the NSOs. Yet, they entail another challenge, not only at the level of their acquisition and implementation into the statistical production, but also in the realm of sustaining relevant skills which reach beyond the traditional set of statistical competencies.

To address this challenge, the UN Global Working Group Task Team on Training, Competencies and Capacity Development has developed this Competency Framework for use by NSOs. It covers the wide array of skills and knowledge considered relevant for those working with big data acquisition and processing. The proposed framework involves core competencies, as well as a more general set of soft skills. They are outlined with reference to a simplified statistical production process, and are followed by thematic blocks. The framework is accompanied by the appendix with the list of selected IT packages and tools which are neither obligatory to apply nor exhaustive, but might prove useful as a reference catalogue of existing applications.

2. How to use this Competency Framework

This Competency Framework sets out an extensive set of skills and knowledge that the UN GWG Task Team for Competencies, Training and Capacity Development considers useful for acquiring and processing big data. It is not a requirement that each data specialist must possess all of them. The framework is intended to provide general guidance for the NSOs, for use when hiring, assessing knowledge gaps, and training staff in specific areas. This will help the NSO to achieve their strategic business goals, now and in the future. It is fully recognised that different NSOs will be running different projects, in different thematic areas. They will also have different types of data specialist, e.g. data analyst, data engineer, data scientist, etc., and each will require different compositions of skills and knowledge.

3. Big data-related competencies according to the statistical production process



4. Core competencies – areas of knowledge and skills

Dimension 1			
Name of the area	**Ethics and privacy**		
Dimension 2			
Competence title and description	To possess a basic level of ethics and privacy knowledge in below-listed issues: 1) Basic definitions of issues related to the processing of big data (personal data and anonymous data, active and passive big data, dimensions of big data, consciously and not-consciously transferred data, etc.) 2) Philosophical aspects of collecting and processing big data (ethic control and a pragmatic view of the impact on the life of people and organizations: privacy, impact on personal capabilities and freedom, rights between data owner and data explorer) 3) Legal framework for management of big data (personal data processing steps and principles, privacy and transparency policy, data processing purposes) 4) Technical aspects of work with private customer and identity data (obtaining and sharing private information, transparent view of how our data is being used, openness of data)		
Dimension 3	A - Foundation	B - Intermediate	C - Advanced
Proficiency levels	Demonstrate knowledge and understanding of basic rules of philosophical, legal of collecting, processing and sharing of big data.	Demonstrate knowledge, understanding and putting into practice philosophical, legal and technical rules of collecting, processing and sharing of big data.	Thorough knowledge of the application of personal data protection law, proficiency in personal data management and skillfulness in performing operations on varied data sets respecting the law, ethical norms, while maintaining the highest technical standards. Advises others on the ethical and privacy considerations of data.
Dimension 4			
Knowledge examples	• Know the rules for the processing of personal data • Understands the ethical basis of managing large customer data sets • Describe the advantages and disadvantages of the use of record level data to achieve business purposes		
Skills examples	• Able to develop a method of collecting, storing and sharing data in accordance with law regulations and ethical standards in the organization • Able to assess whether the acquired data sets have personal data that allow the identification of units • Describe and uses software that protects data against uncontrolled disclosure		
Attitude examples	• Pragmatic view of the impact of personal data regulations on the life of people and organizations • Critical thinking around ethics • Understanding and acceptance for rights between data owner and data explorer • Awareness of the responsibility for the use of private data • Awareness of disclosure control methods if outputs are identifiable		

Dimension 1			
Name of the area	Mathematics		
Dimension 2			
Competence title and description	To possess a basic level of mathematics knowledge in a range of below-listed issues: 1) Basis of algebra: matrices and linear algebra, algebra of sets 2) Probability: theories (conditional probability, Bayes rule, likelihood, independence) and techniques (Naive Bayes, Gaussian Mixture Models, Hidden Markov Models)		
Dimension 3	A - Foundation	B - Intermediate	C - Advanced
Proficiency levels	Demonstrate knowledge and understanding of algebra.	Demonstrate knowledge and understanding of algebra and methods, and ability to apply some of them.	Thorough knowledge of algebra, and skillfulness in performing operations on varied data sets. Is able to advise others on the possible solutions and application of methods to particular problems.
Dimension 4			
Knowledge examples	• Know the rules for creating matrices • Know sentence logic and first order logic • Describe the theoretical basis of probability theories		
Skills examples	• Carry out operation on matrices (addition, scalar multiplication and transposition) • Able to study the basic properties of functions and relations • Able to indicate classes of equivalence relations abstraction		
Attitude examples	• Prepared for independent study of connection issues in mathematics language • Understand significant limitations in defining concepts and mathematical attitude		

Dimension 1			
Name of the area	Data management		
Dimension 2			
Competence title and description	To possess data management knowledge in a range of below-listed issues: 1) Database systems: database management systems, data models – definition and types, entity relationship model, models implementation (pre-relational, relational and object-oriented models) 2) Basics of cryptography: hash function, binary tree 3) Database: relational database, tabular data, data frames and series, shard, on-line analytical processing, data warehousing, data lakes, data vaults, logical multidimensional data model, extract, transform and load (ETL), NoSQL 4) Varied data formats: (Json, shp, XML, csv)		
Dimension 3	A - Foundation	B - Intermediate	C - Advanced
Proficiency levels	Demonstrate knowledge and understanding basic data management skills.	Demonstrate knowledge and understanding of, data base management tools and methods, and ability to apply some of them.	Thorough knowledge of proficiency in data base management and skillfulness in performing operations on varied data sets. Is able to advise others in finding data management solutions.
Dimension 4			
Knowledge examples	• Know the basic concept of SQL and NoSQL databases (such as table, column, row, field, field type, primary and foreign key, relations) • Understand the consequences of using the hash function • Know the basic elements of the SQL language • Define functional dependencies occurring among the analyzed data • Describe the existing database and indicate the appropriate transition keys for the use for official statistics • Describe the advantages and disadvantages of a dataset in various formats		
Skills examples	• Able to create database structures in selected database management systems (e.g. MySQL, MongoDB, *more in annex*) • Select the most used method of going deeper through all the binary tree nodes • Able to present the logical structure of the database using tables and graphical relationships in selected programs (e.g. MS Access, Hbase, *more in annex*). • Able to place and search specific information in the database • Use simple administrative tasks related to databases, e.g. backing up structures and the data itself • Apply query to relational and non-reactive databases • Apply ETL techniques - acquisition, processing (including pre-purification) and loading data from non-statistical sources		
Attitude examples	• Systematically supplement knowledge of new trends in the field of computer science on the subject of computer data storage • Identify data sources and assess their usefulness in complementing studies at hand • Carefully analyze the data and adjust them to the needs of database users • Use metadata to clarify data processing. • Aware of logged data import, export, edit, processes		

<table>
<tr><td>Dimension 1</td><td colspan="3"></td></tr>
<tr><td>Name of the area</td><td colspan="3">Statistics</td></tr>
<tr><td colspan="4">Dimension 2</td></tr>
<tr><td>Competence title and description</td><td colspan="3">To possess a certain level of statistical knowledge in a range of below-listed techniques, to understand and be able to apply selected techniques, to know their underlying assumptions and limitations:
1) Descriptive statistics (mean, median, range, SD, var)
2) Analysis of variance (ANOVA, MANOVA, ANCOVA, MANCOVA);
3) Multiple regression, time-series, cross-sectional
4) Other multivariate techniques: principal components analysis, factor analysis, clustering techniques; discriminant analysis
5) Stochastic Processes: e.g. Markov chains, queuing processes; Poisson processes, random walks
6) Time Series Analysis: time series models; ARIMA processes and stationarity; frequency domain analysis
7) Generalized linear model; any of: log-linear models; logistic regression, probit models, Poisson regression
8) Hypothesis testing: formulation of hypotheses; types of error; p-values; common parametric (z, t, F) or non-parametric (χ^2, Mann-Whitney U, Wilcoxon, Kolmogorov-Smirnov) tests
9) Index numbers: Laspeyres/Paasche indices, hedonic indices; chaining; arithmetic and geometric means as applied to indices.</td></tr>
<tr><td>Dimension 3</td><td>A - Foundation</td><td>B - Intermediate</td><td>C - Advanced</td></tr>
<tr><td>Proficiency levels</td><td>Demonstrate knowledge and understanding of underlying assumptions of at list two of the above-listed areas/techniques.</td><td>Demonstrate knowledge, understanding of underlying assumptions and ability to apply at least four of the above-listed techniques.</td><td>Demonstrate knowledge and understanding of underlying assumptions in own area of expertise as well as, more generally, in other statistical areas. Is able to advise others and use network of contacts to ensure that the most appropriate methodology is applied.</td></tr>
<tr><td colspan="4">Dimension 4</td></tr>
<tr><td>Knowledge examples</td><td colspan="3">• Understand the theoretical basis of analysis of variance (e.g. ANOVA)
• Describe the assumptions underlying the logistic regression
• Understand the consequences of the assumptions not holding
• Depict the expected output of factor analysis</td></tr>
<tr><td>Skills examples</td><td colspan="3">• Compare selected statistical methods and specify differences between them
• Select most relevant statistical method for a specific analytical problem
• Deploy most relevant statistical technique for a specific data set and analytical problem
• Effectively and accurately interpret statistical output</td></tr>
<tr><td>Attitude examples</td><td colspan="3">• Identify new statistical needs and develop statistical analyses to meet them
• Provide critique of statistical analyses produced or received
• Provide guidance on the selection of data sources and matching them with relevant statistical techniques to meet the goals of the analysis at hand</td></tr>
</table>

Dimension 1			
Name of the area	Machine Learning (ML)		
Dimension 2			
Competence title and description	To possess a combination of knowledge and skills in developing self-learning algorithms, including: 1) Programming: data structures (stacks, queues, multi-dimensional arrays, trees, graphs, etc.), algorithms (searching, sorting, optimization, dynamic programming, etc.), computability and complexity (P vs. NP, NP-complete problems, big-O notation, approximate algorithms, etc.) 2) Data modelling: finding useful patterns (correlations, clusters, eigenvectors, etc.) and/or predicting properties of previously unseen instances (classification, regression, anomaly detection, etc.) 3) Model evaluation: e.g. validation accuracy, precision, recall, F1-score, MCC, MAE, MAPE, RMSE, PCC2 4) Application of ML algorithms and libraries: identification of a suitable model (e.g. decision tree, nearest neighbor, neural network, SVM, etc.), selecting a learning procedure to fit the data (e.g. linear regression, gradient descent, genetic algorithms, bagging, boosting), controlling for bias and variance, overfitting and underfitting, missing data, data leakage, among others 5) Understanding the digital product the ML solution will constitute part of		
Dimension 3	A – Foundation	B - Intermediate	C – Advanced
Proficiency levels	Demonstrate knowledge and understanding underlying assumptions of basic probability theories and most common statistical methods and machine learning techniques, programming skills in one of the ML-related applications.	Demonstrate knowledge and understanding of applying probability theories and variety of the statistical methods and machine learning techniques. May have developed further programming skills in at least two of the packages and ability to apply them to resolve ML-related analytical problem.	Demonstrate knowledge, understanding of probability theories and most of the statistical methods and a variety of ML techniques. Demonstrates the ability to apply various ML techniques in various scenarios, and is able to advise and lead others. Have the understanding and skills to fit the ML solution into a system of product/service at hand.
Dimension 4			
Knowledge examples	• Understand Bayes rules • Understand the assumptions underlying model evaluation (quality) indicators, e.g. accuracy, recall, F1 score • Understand the differences between neural networks and SVM		
Skills examples	• Develop a statistical model and fit relevant ML techniques to the analytical problem at hand (e.g. classification and coding, data edition and imputation, image recognition optimization process) • Apply adequate model evaluation indicators		
Attitude examples	• Proactive in searching for optimization opportunities in statistical production with the use of ML • Monitor predictive performance of the employed model to ensure its quality control, being up to date and ability to generate valid results		

<table>
<tr><td>Dimension 1</td><td colspan="3"></td></tr>
<tr><td>Name of the area</td><td colspan="3">Programming</td></tr>
<tr><td colspan="4">Dimension 2</td></tr>
<tr><td>Competence title and description</td><td colspan="3">To possess a certain level of proficiency in programming languages and tools in terms of their functionality and employment for acquisition, processing and visualizing data, as follows:
1) Basic programs to handle the data and create databases: MS Office (e.g. Excel Analysis ToolPak, Access)
2) Relational database management language: SQL
3) Integrated development environments (IDE): R-Studio, Anaconda, more in annex
4) Programming languages, statistical computing environments and results visualization (e.g. Python, R), including:
a) Basics of programming: variables, functions, expressions, loops (break, continue and for expressions)
b) Data structures: vectors, matrices, arrays, factors, lists, data frames
c) Uploading, editing, saving and exporting data (also use of the API)
d) Functions: built-in functions, User-Defined Functions (UDFs)
e) Factor analysis</td></tr>
<tr><td>Dimension 3</td><td>A – Foundation</td><td>B – Intermediate</td><td>C - Advanced</td></tr>
<tr><td>Proficiency levels</td><td>Demonstrate knowledge and understanding of the basic functionalities of analysis tools with graphical interfaces</td><td>Apply the appropriate programs, tools and perform intermediate operations (loading, editing, saving, exporting data)
Use of built-in functions or define own function (UDFs) and perform factor analysis.</td><td>Demonstrate knowledge and understanding of the advanced functionality of selected tools.
In work with data, use advanced functionalities of libraries and packages.
Able to advise others on the best tool to use for the job in hand.</td></tr>
<tr><td colspan="4">Dimension 4</td></tr>
<tr><td>Knowledge examples</td><td colspan="3">• Know the types of queries used in relational databases
• Understand the differences between sorts of data structures: vectors, matrices, arrays, factors, lists, data frames
• Describe the functionality of selected libraries and packages in Python, R
• Depict the expected output of factor analysis</td></tr>
<tr><td>Skills examples</td><td colspan="3">• Upload, edit, save and export data using Python and R programming language
• Develop and create a relational database using dedicated programs
• Deploy selected library or package for in-depth data analysis
• Obtain data for R package, determine their quality, build and graphically present the model</td></tr>
<tr><td>Attitude examples</td><td colspan="3">• Automate processes related to the development of raw statistical data
• Discover dedicated libraries to facilitate statistical analysis with various file formats
• Systematically increase knowledge related to the technical process in the field of coding practices to build scalable digital products
• Use version control platforms to assist with collaboration
• Understand the need to expand technological knowledge in order to improve the skills of using new computer tools</td></tr>
</table>

Dimension 1			
Name of the area	**Data visualization**		
Dimension 2			
Competence title and description	To possess the skills to create graphical representation of the information derived from big data sources (e.g. trends, outliers, patterns), based on the ensemble of the following areas of knowledge and competencies: 1) Mathematics basics: trigonometric function, linear algebra, geometric algorithm, graph theory, etc. 2) Data management and analysis: data cleaning, statistics, modelling 3) Graphics: Canvas, SVG, WebGL, computational graphics, etc. 4) Programming (libraries and packages): e.g. R (ggplot2) and Python, Tableau, Power BI, ArcGIS (*more in annex)* 5) Essential design principles: aesthetic, color, interaction, cognition, etc. 6) Visual solutions: coding, analysis, graphical interaction		
Dimension 3	A - Foundation	B – Intermediate	C - Advanced
Proficiency levels	General knowledge of visual solutions related to big data Programming skills to develop simple visual representation of the data (e.g. charts, graphs, box plots, histograms, infographics). Good understanding of when to use which graph.	Demonstrate knowledge of specific visual solutions related to big data. Programming skills to apply a selection of more complex visual methods (e.g. area chart, bubble cloud, heat map, treemap, word cloud) and an understanding of when to apply which visual method.	Thorough knowledge of visual solutions related to big data. Programming skills to deploy a wide array of appropriate visual methods. General knowledge of graphic design, color regimes applicable in certain domains (e.g. map making). Able to advise others on the most appropriate data visualization tool to apply.
Dimension 4			
Knowledge examples	• Understand trigonometric functions and their relation to data visualization • Understand graph theory • Understand visualization functions of analysis software • Understand color regime in maps development		
Skills examples	• Prepare data sets for visualization purpose • Generate heat map • Apply adequate visualization technique to the data/analytical output at hand • Able to simplify complex theories/data through visualization		
Attitude examples	• Proactive in searching for the most attractive, yet clear visualization techniques • Critically assess the match between the target audience and the purpose of the information to be presented in order to utilize the most adequate visualization forms • Proactive in exploring new data visualization techniques and packages in order to enhance data presenting		

5. Generic skills

Generic skills	Description
Communication	• Able to link business orientation with the scientific, analytical, and technical facets • Skillfully communicate findings to data users and decision-makers • Describe and explain, with influence, the value of work to the stakeholders • Able to effectively convey information to both, technical and non-technical audiences
Curiosity	• Intellectually curious to look for answers to address statistical research questions • Able to go beyond the initial assumptions of research and results • Keen to seek solutions for hidden, overlooked queries
Business Acumen	• Able to deal with a massive amount of knowledge and translate it effectively for a non-technical audience • Equipped with knowledge of current and upcoming trends • Able to acquire foundations of relevant disciplines, concepts and tools • Possess knowledge and analytical skills of organization's business objectives in order to provide answers to current problems • Able to use data to accelerate the growth of the organization
Storytelling	• Convey results of work coherently and understandably • Use data visualization to present decision-makers concepts/ideas/ phenomena from a new perspective • Able to use different approaches to build narratives in order for stakeholders to attain a new sense of clarity and identify the best course of action
Adaptability	• Able to quickly adapt activities to the latest technologies • Respond to varying business trends
Critical Thinking	• Able to perform an objective analysis of a problem at hand and take appropriate actions to solve it • Understand the need to take a closer look at the data source and critically asses its quality, usefulness and potential problems associated with it • Logically identify strengths and weaknesses of ideas and technical approaches and make effective decisions based on these attributes
Product Understanding	• Work with the customer to fully understand their needs, and regularly report on progress for feedback • Able to propose actionable insights that can improve product quality • Understand the need to adapt the production process to the expected product and its functionality • Ensure that a plan is in place for implementation of the new product, with customer involvement
Team Player	• Understand importance of teamwork • Able to collaborate effectively with others • Able to manage a team effectively
Agile project management	• Work closely with the customer to deliver in small increments • Manage work and delivers to plan

6. References

1) 5 Skills You Need to Become a Machine Learning Engineer. (2020). Retrieved May 19, 2020, from https://blog.udacity.com/2016/04/5-skills-you-need-to-become-a-machine-learning-engineer.html

2) 5 Skills You Need to Become a Machine Learning Engineer. (2020). Retrieved May 19, 2020, from https://blog.udacity.com/2016/04/5-skills-you-need-to-become-a-machine-learning-engineer.html

3) Carretero S., Vuorikari R., Punie Y. (2017). DigComp 2.1, The Digital Competence Framework for Citizens. European Commission. Retrieved May 19, 2020, from http://publications.jrc.ec.europa.eu/repository/bitstream/JRC106281/web-digcomp2.1pdf_(online).pdf

4) Competency framework for the Government Statistician Group. (2016). Retrieved May 19, 2020, from: https://gss.civilservice.gov.uk/policy-store/government-statistician-group-gsg-competency-framework/

5) Competency profiles created by the Modernisation Committee for Organisational Frameworks and Evaluation. (2016). Retrieved May 19, 2020, from https://statswiki.unece.org/display/bigdata/Competency+Profiles

6) Curriculum for data science. (2016). Retrieved May 19, 2020, from https://www.cyfronet.krakow.pl/cgw16/presentations/S8_02_presentation-Edison-CGW-26-10-2016.pdf

7) Data visualization beginner's guide: a definition, examples, and learning resources. (2020). Retrieved May 19, 2020, from https://www.tableau.com/learn/articles/data-visualization

8) Ferrari A. DIGCOMP: A Framework for Developing and Understanding Digital Competence in Europe. European Commission. (2013). Retrieved May 19, 2020, from https://www.rebiun.org/sites/default/files/2017-11/JRC83167.pdf

9) OECD Competency Framework, *Talent.oecd – Learn. Perform. Succeed.* (2018). Retrieved May 19, 2020, from https://www.oecd.org/careers/competency_framework_en.pdf

10) Proposing a framework for Statistical Capacity Development 4.0. (2017). Retrieved May 19, 2020, from https://paris21.org/sites/default/files/inline-files/CD4.0-Framework_final.pdf

11) Rybnicka D., Wilczek G. Python. (2015). Podstawy programowania. Retrieved May 19, 2020, from https://python101.readthedocs.io/pl/py3/basic/basic.html

12) Rybiński M. (2020). Krótkie wprowadzenie do R dla programistów, z elementami statystyki opisowej. Retrieved May 19, 2020, from https://www.mimuw.edu.pl/~trybik/edu/0809/rps/r-skrypt.pdf

13) Roland van Loon. (2020). The Soft Skills That Are An Asset to Every Data Scientist. Retrieved May 19, 2020, from https://www.simplilearn.com/soft-skills-for-data-scientist-article

14) Statistician Competency Framework. Government Statistical Service. (2012). Retrieved May 19, 2020, from https://gss.civilservice.gov.uk/archive/wp-content/uploads/2012/12/Statistician-competency-framework.pdf

15) The European Qualifications Framework. (2020). Retrieved May 19, 2020, from https://www.cedefop.europa.eu/en/events-and-projects/projects/european-qualifications-framework-eqf

16) The 30 Best Python Libraries and Packages for Beginners. Retrieved May 19, 2020, from https://www.ubuntupit.com/best-python-libraries-and-packages-for-beginners/

17) Top 7 key skills required for Machine Learning jobs. (2018). Retrieved May 19, 2020, from https://bigdata-madesimple.com/7-key-skills-required-for-machine-learning-jobs/

18) Tutorials in DataCamp. Retrieved May 19, 2020, from https://www.datacamp.com/community/tutorials/functions-python-tutorial\

19) UN Competency Development – A Practical Guide. UN Office of Human Resources. (2010). Retrieved May 19, 2020, from: https://hr.un.org/sites/hr.un.org/files/Un_competency_development_guide.pdf

20) U.S. Census Data Science & Visualization Curriculum. Retrieved May 19, 2020, from
- https://datavizcatalogue.com/
- https://www.census.gov/dataviz/
- https://www.census.gov/data/adrm/what-is-data-census-gov.html

Appendix – List of programs and tools

Area	Programs and tools
Data management	MS Excel, Access, SQL Server, MySQL, Python (*arrow, numpy, pandas),* R (*DBI, Dplyr, stringr*), MS Azure, Apache Hadoop, Ataccama, Profisee, SAS, Cassandra, MongoDB, Oracle NoSQL DB, Hbase
Statistics	MS Excel (*Analysis ToolPak*), MS Access, Statistica, SPSS, Stata, Python (*scikit-learn, SciPy, numpy, matplotlib, statsmodels, pandas*), R (*stats*), SAS, Statistica Big Data Analytics
Machine Learning	Python (*scipy, keras, TensorFlow, NLTK),* TensorFlow, Apache Spark, Torch/Pytorch, Keras
Programming	Python, R, Linux commands, R-studio and Anaconda, Git and Github, SQL
Data visualization	Python *(pillow, matplotlib, bokeh*), R (*ggplot2, esquisse*), Tableau, Power BI, ArcGIS, D3, HighCharts, Echarts, Vega