DataScience
in Human & Social Science for Women Empowerment

datascience-project.eu

Analysis for the definition of competences in the field of data science in the human and social sciences

Project Result 2

2021-1-IT02-KA220-HED-000023199

Co-funded by the European Union

Abstract

Data Science is a multidisciplinary field that combines technical skills, domain knowledge, and analytical thinking to extract insights from data and solve complex problems. Data Science has applications in various domains, such as business, health, engineering etc.

In recent years there is a growing use of Data Science also in the areas ofeducation, social sciences, and humanities. However, despite the growing demand for data scientists, there is a lack of diversity and inclusion in this field, especially among female students.

In this report, we intend to present the overall work that led to the creation of a Framework of competences in Data Science, which aims to provide a comprehensive and coherent set of learning outcomes and assessment criteria for data science education at the undergraduate level. The Framework is based on a literature review of existing data science curricula and competencies frameworks, as well as on the input of experts and stakeholders from academia and industry. The Framework covers different dimensions such us: data management, data analysis, data communication, and data ethics.

The following Report also includes the analysis of a questionnaire that measures the degree of skills, knowledge and Interest of students in Data Science of the Universities of Salento (Italy), Sannio (Italy), Oviedo (Spain), and Academia de Studii di Economice of Bucharest (Romania). The questionnaire was designed to assess the students' self-perception of their data science competencies according to the Framework, as well as their motivation, interest, and confidence in pursuing a career in data science. The questionnaire was administered online to a sample of 440 students from different disciplines and backgrounds.

The report has been divided into Four (4) parts: - The first part is dedicated to the analysis of the literature on the state of the art of the world of education to bring female students closer to Data Science. We review the main challenges and barriers that women face in entering and advancing in data science careers, as well as the best practices and initiatives that aim to foster gender diversity and inclusion in this field. We also discuss the

potential benefits and opportunities that data science can offer to female students from different disciplines and backgrounds. - The second part explores various case studies that apply methods and tools of data analysis to the humanities and social sciences. We showcase some examples of how data science can enhance the understanding and interpretation of cultural, historical, cultural, social, and economic phenomena. We also highlight the importance of interdisciplinary collaboration and communication between data scientists and domain experts.

- The third part is dedicated to the presentation of the Questionaire and the scale used – The fourth and last part is dedicated to the measurement of students' skills, Interest, Attitude in Data science. We present the results of the questionnaire analysis, focusing on the descriptive statistics, the reliability and validity of the instrument, and the differences among groups based on gender, discipline, country, and level of education. We also discuss the implications and limitations of our findings, as well as some suggestions for future research and practice.

Keywords: Graduates, Labour Market, Data Science, STEAM, Gender Gap, Structural Equation Models, Human & Social Sciences, Women in STEM, Gender diversity in data science, Data-driven solutions for gender equality, Gender bias in machine learning, Women in data science

INDEX

INTRODUCTION AND GOALS

The Present Report represents part of the Framework of Competencies in Data Science, one of the main Outputs of the project.

Specifically, the Report has 3 main Objectives:

1. To provide an overview of the State of the Art of the Gender Gap in STEM with particular reference to the area of Data Analysis;
2. To present the results of the analysis of the questionnaire administered to students for the purpose of gathering opinions and measuring attitude and interest toward data analysis;
3. To define the basic and specific skills needed to integrate Data Analysis into Humanities and Social Sciences (HSS) curricula.

The Framework

The Framework is a shared methodology pathway for defining the basic and specific skills needed to integrate in-depth study on Data Science into Human and Social Sciences (HSS) curricula. It was constructed with the aim of proposing a flexible conceptual framework that is applicable to different socioeconomic and cultural contexts and disciplinary backgrounds.

A two-method approach was used to define the Framework:

1. A bottom-up method through a fact-finding survey to be carried out comparatively among the project partner universities.

2. A top-down method through literature study, secondary source analysis, and case studies to identify the best determinants of successful data science courses in the humanities and social sciences.

The construction of the Framework has followed the following steps

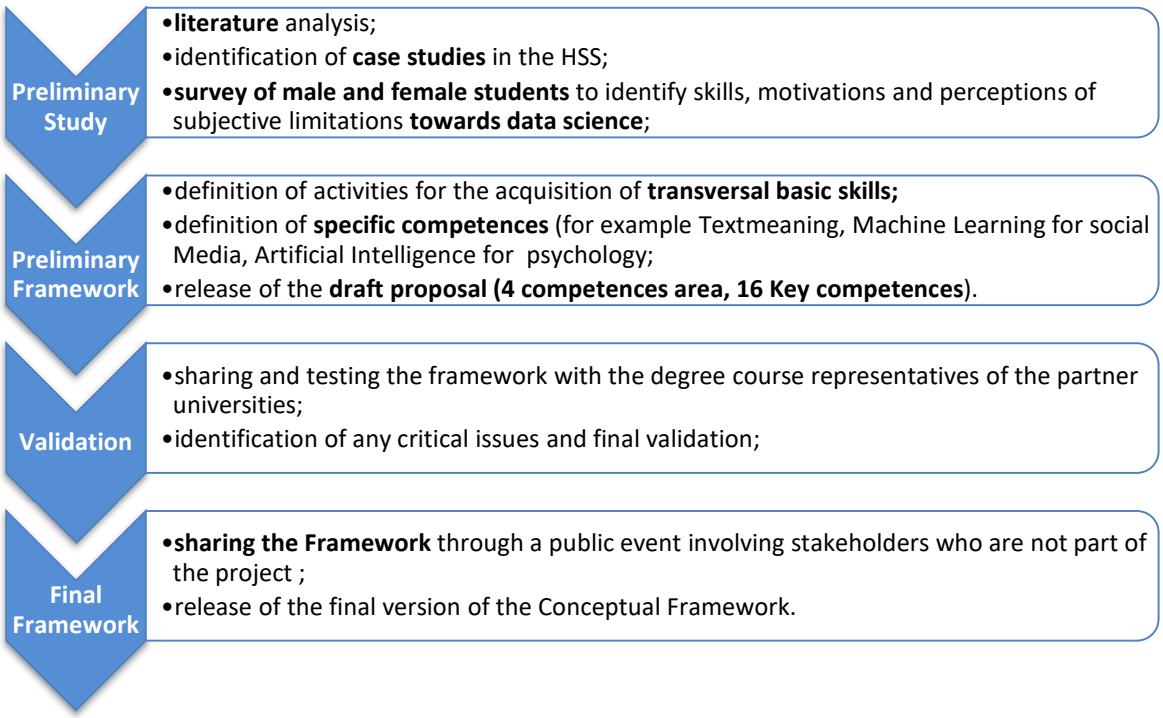


Figure 1

Methodology

As can be seen from Figure 1, the construction of the Framework was divided into four stages:

1. Preliminary study: Through the analysis of relevant academic literature and the in-depth study related to case studies concerning the integration of Data Science in the Humanities and Social Sciences, the set of basic (cross-cutting) and specific (related to the application to different disciplinary fields) competencies that serve to specifically define the educational and research objectives of the project has been defined.

A crucial section of the preliminary study is the structured survey (qualitative and quantitative) aimed at defining the Competencies, Motivations, and individual perceptions by male and female students in HSS courses of study of the subjective limitations toward theoretical and applied in-depth study in Data Science. Above all, the survey allowed us to understand the orientation of

students (with particular attention to female students) toward the study of Data Science and to see whether, through appropriate training in data analysis, they would be interested in learning more about the interaction between the social sciences and humanities and Data Science-related disciplines. The survey sample was constructed using and valuing collaboration among the project partner universities. The research results are also instrumental in validating the content of the courses and modules provided in Outcome 3 (R3).

2. Preliminar Framework: in light of the outcomes of the previous phase, the learning objectives to be achieved and the activities to be carried out in university courses required for students' acquisition of transversal (basic) and specific (applied to different disciplinary contexts) competencies were defined in detail. The overall competency framework has been divided into 4 competency areas, comprising a total of 16 Key Competencies, which develop 60 Learning Outcomes.

3. Validation of the Framework: the Framework validation process was shared among all project partners. This made it possible to identify and overcome any critical issues related to the adaptation of the Framework to different national contexts and study paths.

4. Definitive Framework: the release of the final version of the final Conceptual Framework was achieved through a public event organized by the University of Salento (hybrid modality, online and face-to-face) involving stakeholders and associates of the project. The event contributed, on the one hand, to the dissemination of the outcomes of the project phase to other Italian and European Universities potentially interested in the dissemination of Data Science in HSS curricula and, on the other hand, to the involvement of companies in the Information and Communication Technology sector most interested in the interaction between Data Science and Human and Social Sciences.

PART 1: GROUNDING - LITERATURE REVIEW

In licterature it is widely recognized that the social construct of gender includes many spheres of individuals' personalities and working lives because it is closely related to perceptions, views and perspectives of the private and public spheres. Within the general theme of gender gaps in the labor market, special attention is assigned to the so-called issue of women's work participation in STEM (Science, Technology, Engineering, Mathematics) fields, which are socially perceived and predominantly characterized by a male presence in both undergraduate and postgraduate education and the labor market. From the perspective of the implications that this state of affairs generates from the socioeconomic point of view, it is crucial to specify that the low participation of women in education and the labor market in STEM fields is partly to be attributed to cultural stereotypes referring to gender and to dynamics of discrimination and self-discrimination that are transformed into specific personal aspirations and choices (Thébaud and Charles 2018).

The effects of the gender gap in STEM fields generate diverse economic, social, and cultural effects. Reference is often made in the literature to the persistence of pay gaps between men and women caused by educational choices that tend to segregate women in 'female' fields characterized by lower career prospects and lower pay levels. Another area of implications relates to the reinforcement of stereotypes based on presumed innate abilities that differ between men and women, which produce different outcomes in terms of aspirations, work choices and, more generally, perpetuate the culture of gender segregation even in subsequent generations. Finally, the female component and other components of the labor market, would represent an untapped resource within the labor market itself that, from a human capital perspective, needs a gradually increasing sharing of workers specialized in

STEM fields to foster economic development, ecological transition and digital and meta-digital transformation of economic activities. For the purposes of the activities, the consortium proposed to carry out, through the 'Data Science in Human & Social Science for Women Empowerment' project, two elements to be developed with respect to this issue. First, in the general context of university education and employment in STEM disciplines, it is clear from the literature that, at present, the most persistent and pronounced gender gap is mainly observed in advanced societies with high income levels where there is from a formal point of view greater gender equality (Charles 2011, 2017; Stoet and Geary 2018). This means that in order to incentivize greater participation of women in STEM education and work fields, alternative educational strategies need to be planned than in the past. Furthermore, and this is the second point, in some countries such as the United States, it has been found that, especially in the area of computer science, there is a progressive decrease in women's training at the university level. According to Thébaud and Charles (2018) "Women's share of US bachelor's degrees in computer science declined from 28% to 18% between 2000 and 2015."

Given the dynamics of the labor market, in which Data Analysis-related jobs are one of the areas of work expected to develop most in the coming years (World Economic Forum 2018, 2020), this trend represents a definite prospective employment disadvantage for women.

A crucial issue, therefore, concerns the specifics of training and mentoring women in STEM (Beck et al 2022). The effectiveness of these activities is crucial for reversing the trend observed so far and for fostering not only greater participation of women in STEM fields training but also for supporting their labor participation and reduction of dropout in these fields. The strategies adopted in different contexts are diverse and in the literature emerges 'observation that there is no specific winning strategy but that there is a need to develop a coordinated set of training, mentoring and leadership training actions. At present, the main activities that training and mentoring can be

classified into peer mentoring (horizontal approach), role modeling, one-on-one vertical mentoring, or group mentoring. These activities helped women and other segregated groups to achieve greater awareness of their career prospects, opportunities arising from specific goals of personal fulfillment, increased awareness about the needs of balancing work activities and caregiving burdens, recognition of gender biases, and also access to a network of specialists who support women and minorities in their prospects for self-actualization (Beck et al. 2022, Dennehy and Dasgupta 2017, Welch et al. 2012).

With regard to the specific case of computer science, the phenomenon of an increased propensity of female graduates not to pursue employment in STEM fields should also be recorded (Sassler et al. 2017). The phenomenon is also widespread for other educational paths (such as engineering), but the persistence and even low participation of women in these fields of work seem to confirm the difficulty in shaking the stereotype of masculinity that surrounds the field (DuBow and James-Hawkins 2016). Some universities in the United States have implemented targeted interventions involving the reorganization of introductory computer science courses, the activation of mentoring and peer-to-peer support pathways. The result has been a remarkable increase in the percentage of female graduates, from 10 percent to 40 percent five years after the start of the activities (Cheryan et al. 2013; 2015).

References

- ✓ Beck Makini, Jillian Cadwell, Anne Kern, Ke Wu, Maniphone Dickerson, Melinda Howard (2022) "Critical feminist analysis of STEM mentoring programs: A meta-synthesis of the existing literature" Gender, Work & Organization, 29: 167-187, https://doi.org/10.1111/gwao.12729.
- ✓ Charles Maria (2011) "A World of Difference: International Trends in Women's Economic Status". Annual Review of Sociology, 37: 355–71.
- ✓ Charles Maria (2017) "Venus, Mars, and Math: Gender, Societal

Affluence, and Eighth Graders' Aspirations for STEM". Socius: Sociological Research for a Dynamic World, 3: 1–16.

- ✓ Cheryan, Sapna, Benjamin J. Drury, and M. Vichayapai. 2013. "Enduring Influence of Stereotypical Computer Science Role Models on Women's Academic Aspirations". Psychology of Women Quarterly, 37: 72–29.
- ✓ Cheryan, Sapna, Allison Master, and Andrew N. Meltzoff. (2015) "Cultural stereotypes as gatekeepers: Increasing girls' interest in computer science and engineering by diversifying stereotypes". Frontiers in Psychoogy, 6: 49. Available at: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4323745.
- ✓ Dennehy, T. C. and Dasgupta, N. (2017) "Female peer mentors early in college increase women's positive academic experiences and retention in engineering". Proceedings of the National Academy of Sciences of the USA, 114 (23): 5964– 5969. https://doi.org/10.1073/pnas.1613117114.
- ✓ DuBow, Wendy M., and Laurie James-Hawkins. 2016. "What Influences Female Interest and Persistence in Computing? Preliminary Findings from a Multiyear Study". Computing in Science and Engineering, 18: 58–67.
- ✓ Sassler Sharon, Katherine Michelmore and Kristin Smith (2017) "A Tale of Two Majors: Explaining the Gender Gap in STEM Employment among Computer Science and Engineering Degree Holders", Soc. Sci. 2017, 6, 69; doi:10.3390/socsci6030069.
- ✓ Stoet, Gijsbert, and David C. Geary. (2018). "The Gender-Equality Paradox in Science, Technology, Engineering, and Mathematics Education". Psychological Science, 29: 581–93.
- ✓ Thébaud Sarah and Maria Charles (2018) "Segregation, Stereotypes, and STEM", Reprinted from: Soc. Sci. 2018, 7, 111; doi:10.3390/socsci7070111.
- ✓ Welch, J. L., Jimenez, H. L., Walthall, J., & Allen, S. E. (2012). "The women in emergency medicine mentoring program: An innovative approach to mentoring". Journal of Graduate Medical Education, 4 (3):

362– 366.

- ✓ World Economic Forum (2018) The Future of Job report, available at https://www.weforum.org/reports/the-future-of-jobs-report-2018/
- ✓ World Economic Forum (2020) The Future of Job report, available at https://www.weforum.org/reports/the-future-of-jobs-report-2020/

PART 2: CASE STUDY

Nowadays Data science is making its way into our businesses, our communities, and, in many cases, our lives.

Modern Data Science has emerged in technology and computing as one of the most widely used and sought-after disciplines by companies, starting with the big giants of the digital age such as Google, Amazon or Facebook.

Today, it is transforming all sectors, from retail and telecommunications to agriculture, health, culture, transportation, and the criminal justice system.

Data Science is thus an interdisciplinary field that uses and combines data inference, statistical and scientific methods, algorithm development, programming skills, and information technology in order to solve highly complex problems and extract meaningful information and value from data.

In this section we will show some examples of case studies and initiatives (In Europe and beyond) collected by project partners that demonstrate applications of quantitative data science methods in the humanities and social sciences.

Hate speech: measures and counter-measures

The "Hate speech: measures and counter-measures project" is aimed at developing and applying advanced computational methods to conduct a systematical measurement, analysis and counter hate speech across different online domains, including social media and news platforms. Hateful content online is, indeed, a growing problem that is polluting civic discourse, inflicting harm on targeted victims, creating and exacerbating social divisions, and eroding trust in the host platforms.

The aim of the project is to understand the scale and scope of online hateful content, taking into account its different forms, from 'everyday' subtle actions to overt acts of aggression and criminality, and the different targets, such as ethnic minorities and women. To this purpose, researchers from the

Universities of Oxford, Surrey, Sheffield and the George Washington University, led by The Alan Turing Institute's "Hate Speech: Measures & Counter-measures" project, have developed a tool that uses deep learning to detect East Asian prejudice on social media.

The project also aims to understand the dynamics and drivers of hate, providing granular insight into when, where and why it manifests.

This use case is of particular use for data science techniques as it addresses the problem of racial discrimination by building a model on Python from a 20,000 tweets-large dataset.

How was it done

This research uses advanced computational methods, including supervised machine learning, stochastic modelling and natural language processing, to detect and analyse hate speech.

The paper presents a report on the creation of a classifier that detects and categorizes social media posts from Twitter into four classes (Hostility against East Asia, Criticism of East Asia, Meta-discussions of East Asian prejudice and a neutral class). The classifier achieves an F1 score of 0.83 across all four classes.

Subsequently, a final model (coded in Python) is provided, as well as a new 20,000 tweet training dataset used to make the classifier, two analyses of hashtags associated with East Asian prejudice and the annotation codebook.

Challenges

Working with qualitative indicators, as for this very case analysing abusive tweets, using a binary schema (i.e. prejudiced or not) to classify the data resulting from such tweets could be theoretically problematic because distinct types of behaviour, with different causes and impacts, are collapsed within one category.

It can also negatively impact classification performance because of substantial

within-class variation. The researchers have therefore developed the taxonomy and codebook that they have used iteratively, moving recursively between existing theoretical work and the data to ensure the detection system can be applied by social scientists to generate meaningful insights.

Visit the web site

https://www.turing.ac.uk/research/research-projects/hate-speech-measures-and-counter-measures

References:

Bertie Vidgen, Austin Botelho, David Broniatowski, et al. (2020), Detecting East Asian Prejudice on Social Media, https://arxiv.org/pdf/2005.03909v1.pdf

https://www.turing.ac.uk/research/research-projects/hate-speech-measures-and-counter-measures

Suggestions for exploitation

A prior knowledge of basics of statistics might be recommended as the paper might not be accessible to people with low understanding of highly advanced statistics-related glossary. Anyways, the non-technical parts are pretty clear and explanatory for everybody and the conclusions are very reader-friendly and accessible.

Lots of external references, which are useful to deep-dive into the background of the analysis.

As in the use case #1 that we have proposed, this use case too is pretty timely relevant as there are potential linkages with numerous other articles on the issue.

For instance, the following:

o Barometro dell'odio: intolleranza pandemica

https://d21zrvtkxtd6ae.cloudfront.net/public/uploads/2021/04/Amnesty-barometro-odio-2021.pdf

an opportunity to narrow the scale and scope of the analysis and applying the

same methodology to other relevant domains of interest (general hate speech VS hate speech targeted at minorities, i.e., LGBT community, migrants, etc.)

DEXIBIT: Data analytics for visitor attractions

Dexibit is a company that applies machine learning and artificial intelligence to exhibitions and museums. Dexibit began when founder Angie Judge noticed a museum guard with a clicker counter. With a background in network analytics, she saw the need for transformation in data for visitor attractions. Years on, Dexibit doesn't just help attractions count visitors accurately, but predicts and analyzes visitor behavior too.

Dexibit It is a significant example of the usefulness of big data for more effective management of cultural organisations and better allocation of public funds.

How was it done

Dexibit offers data analytics services to study the audiences of cultural venues such as museums and predict their flows.

Dexibit helps cultural institutions deliver world class visitor experience while delivering on loyalty expextations and meeting the realities of commercial sustainability.

The data collected together with online traffic, social media and business transactions, allow a snapshot of the visitor experience to be developed. And also to build predictive models of visits.

Instead, the technology developed by Dexibit, which combines machine learning and artificial intelligence, provides digital snapshots of visitor presence in cultural venues (mainly museums) returned in the form of customised dashboards and automated reports. This data, together with online traffic, social media and business transactions, enables the development of a picture of the visitor experience. If cultural venues have historical data, Dexibit is also able to build predictive visitor models.

Challenges

Dexibit began when founder Angie Judge noticed a museum guard with a clicker counter. With a background in network analytics, she saw the need for transformation in data for visitor attractions.

Years on, Dexibit doesn't just help attractions count visitors accurately, but predicts and analyzes visitor behavior too.

Core to Dexibit is our philosophy of data ethics. We believe in insight inspired decisions based on understanding visitors and their experiences at large instead of micro-targeting individuals at the expense of privacy.

Dexibit's experience is part of the Data science for good philosophy and a cultural data initiative. Dexibit has a data ethic approch.

Importantly, in order to meet the needs of museums at such a difficult and unprecedented time as the pandemic, Dexibit committed to offer the post-COVID scenario simulation service free of charge.

Visit the web site

https://dexibit.com/

Data for Good

Data for Good Italia is an initiative that aims to illustrate the opportunities of data science at the service of social issues through data, analysis and news.

The main objective is to tell how data and their analysis can contribute to a better society by contributing to the impact of social initiatives.

In summary, Data for good selects and collects socially relevant data analysis.

Visit the website

https://www.dataforgood.it/

Women in Data Science

Women in Data Science (WiDS) elevates women in the field by providing inspiration, education, community, and support. WiDS started as a one-day technical conference at Stanford in November 2015. Seven years later, WiDS is a global movement that includes a number of worldwide initiatives:

A conference with nearly 200 regional events worldwide in more than 50 countries, reaching 100,000 participants annually.

A datathon, encouraging participants to hone their skills using a social impact challenge.

A podcast series, featuring data science leaders from around the world talking about their work, their journeys, and lessons learned along the way.

A Next Gen program to encourage secondary school students to consider careers in data science, artificial intelligence (AI), and related fields.

A workshop series to build your data science skills, inspiring women and girls with role model instructors.

Challenge

WiDS Next Gen strives to inspire secondary school students to take relevant courses and consider future careers involving data science, Artificial Intelligence (AI) and other related areas. We particularly encourage younger women and girls to consider these STEM fields by showing examples of successful women in technology roles.

Visit the website

https://www.widsconference.org/

Gender bias in children's books: leveraging AI to help create diverse textual material

Gender bias is highly present in children's books. Lead characters are about twice as likely to be male than female and non-binary characters are basically absent. It is not only a matter of frequency, but also of the way genders are portrayed. We all know a few stories of princesses that need to be saved by male heroes. By the age of 7, children's aspirations are shaped by gender-related stereotypes. These are carried into adulthood and contribute to gender power gap, gender pay gap, discrimina-tion, harassment, etc.

We want to help publishers and authors to create more diverse and inclusive content.

This project was born during the Hackathon 'AI for Gender Equality' organized by Vinnova (Sweden's Innovation agency) and Women in AI in December 2020, where it won second place. The project is carried under the name of Paige by Annalisa Cadonna, Camilla Damian and Laura Vana.

The project goal was in the scope of the 'AI for Gender Equality' hackathon. The project leveraged the AI field of Natural Language Processing (NLP) which deal with unstructured data, in this case text.

How was it done

We have used both python and R, and developed a shinyR app. In python and R, we have used NLP open-source libraries, such as nltk and spacy. We have also used transformers (specifically BERT).

Challenges

- Only books without copyrights are quite old and do not reflect the modern children's book land-scape
- Both books and models consider gender as binary, which it is not. However, there is not enough data for including other gender expressions in the analysis
- Pre-trained language models are trained on biased data

The goal should determine which technology to use and not vice versa.

Collaboration between people with different expertise and background is fundamental for the suc-cess of such a project.

Increase the quality of fraud reporting with AI

The right information at the right time helps companies minimize risks and thus protects the organization. With a strong corporate culture in which employees stand up for the company instead of remaining silent, companies are more successful economically. Because everyone knows the value of a good hint to minimize business risks and improve the organization.

Sometimes, however, whistleblowers' descriptions are very vague and often lack the details to actually take appropriate action. This means that case handlers have to invest a lot of manual work to analyse the case appropriately. And sometimes, despite a lot of manual work, a reported case cannot be fully resolved.

The team of Compliance 2b is convinced that there is a better way to identify the valuable information!

Our internal reporting channel uses trustworthy Artificial Intelligence (AI) to increase the quality of each report, and the system asks whistleblowers to provide missing details if necessary. Each report is assessed for its relevant content, and the relevant content is color-coded and displayed to the case manager. In addition, the AI learns successively with new cases and supports the case handlers in choosing the effective analysis activities to investigate the reported case, whereby the final decision always remains with the humans. Furthermore, the anonymization of names enables a neutral processing of the information by focusing on the facts instead of the messenger.

A detailed description can be found on our homepage https://compliance2b.at/

Small organizations have several disadvantages. On the one hand, they have less specialists who can process important reports appropriately and without any bias, i.e. with no negative consequences for the whistleblower. On the

other hand, fewer reports are submitted, from which the organization could grow.

But the important thing is that the focus is on resolving wrongdoing and that the organization can learn from these hints.

If there are no specialists to do this work, then an AI can support the case handlers in this activity and the organization can minimize unethical behaviour. Our use case shows how this can be done effectively with limited data while addressing the specific needs of the organization.

Particularly in these times, it is very important and has a high impact on society that it is not the nasty individual who gain the upper hand, but those who follow the rules and do the right thing for everyone.

How was it done

We have chosen a solution via semantic textual similarity (STS) models. Such a model uses two given text documents to decide how similar the contents of the two documents are on a scale from 0 to 100%. We use STS by comparing, for a reported category, a set of pre-defined phrases with the incoming report text. The category with the best matching phrases is finally selected by the system.

The advantage for the employing organization is that the phrases can be different for each organization, depending on the business model and individual risk situation.

Semantic similarity of free texts can be determined efficiently since several years with so-called document embeddings. Such an embedding is a vector in the mathematical sense to which the text is mapped with a special model. Two free texts are similar in content exactly when the embedding vectors have a small cosine distance to each other. Most of the computational effort is in creating the embeddings. This is also the advantage of this method. Instead of passing texts in pairs to a similarity model, texts are mapped individually to embeddings from which the similarity can be determined with minimal computational effort. Thus, the computational effort does not scale quadratically with the number of free texts, but linearly.

In 2018, the Sentence-BERT (SBERT) model created a new state-of-the-art in modeling STS and other regression problems on pairs of documents. SBERT is an evolution of Google BERT and has been trained via supervised learning to produce document embeddings specifically suited for STS. SBERT provides many times more accurate results than BERT and the runtime of similarity determination in a corpus of 10,000 documents could be reduced from 65 hours to 5 seconds because, unlike BERT, SBERT processes 10,000 individual documents instead of 50 million pairs of documents. For these reasons, we decided to use an open-source variant of SBERT made available on Huggingface.

Challenges

The majority of modern text classifiers must be trained to map texts to the correct clue category using supervised learning. This requires a training data set specific to the use case with several 1,000 texts and the corresponding category (label). On the one hand, we did not have such a large labeled training dataset available in the context of this project. On the other hand, the decision of a classifier trained with supervised learning can only be seen indirectly and with additional explanatory models like LIME. Therefore, we chose a different solution path via semantic textual similarity (STS) models. Such a model is also more adaptable to a specific organization and thus has individual advantages.

References

https://compliance2b.at/

https://www.awsconnect.at/KI-Anbieter

https://www.aws.at/service/cases/gefoerderte-projekte-auswahl/digitalisierung/compliance2b/

SoGooD²ata

SoGooD²ata (is a Spanish NGO, founded by data scientist Ana Laguna Pradas in 2019. Dr.

Laguna Pradas's goal was to apply data science to social and human sciences in such a way as to use data and their analysis as a new tool for solving problems of great social importance. Most of the social projects that the SoGooD²ata team has dealt with in these 3 years and 4 months of activity are projects that are part of the ONU priorities defined in 2018. Some examples are: Health and wellbeing, quality education, gender equality, sustainability, reduction of inequalities, etc. Over time the team of this NGO has grown to now count 13 collaborators (7 of them women), whose experience ranges from data science to research and social sciences. To achieve its objectives, moreover, SoGooD²ata collaborates with realities such as the Comillas Pontifical University (https://www.comillas.edu/), COCEMFE (https://www.cocemfe.es/), which is a NGO whose goal is to achieve the full inclusion and active participation of people with physical and organic disabilities, and Nielsen (https://global.nielsen.com/), an American society of information, data and market measurement. At the moment, the projects they are working on are 9 and they concern pollution, collaborative transport, health and nutrition, history (Spain, colonization and uninhabited villages), translation of new born cries, inclusion of people with disabilities, education, bullying and citizens awareness of artificial intelligence. The latter is based on the assumption that analytical challenges of a social nature can be faced through data science and, for this reason, notions on this subject should be disseminated and within the reach of people involved in other fields of study, such as, for example, the social sciences, but also much more. Therefore, the SoGooD²ata team also takes care of organizing and holding master classes and sessions on Big data, artificial intelligence, machine learning and more, in schools, institutions, business schools, etc.

Why was this use case chosen

In recent years, various disciplines in the field of data science have shown significant progress. Data is gaining more and more value, but its use is often limited to marketing, digital banking, private companies, etc. With SoGooD²ata,

Ana Laguna Pradas and her team members were able to highlight the role that data science could and should play in other fields, such as the social sciences. Thanks to the pioneering work of the collaborators of this NGO, various projects and researches in the field of social data science have come to life and this has allowed to pave the way for the innovation that data analysis tools and techniques can bring to social research. Furthermore, in addition to contributing to social projects through data science, SoGooD²ata is a prime example of how the most innovative ideas derive from a heterogeneous working context with different approaches to the problem and different points of view. It is probably precisely this diversity and multidisciplinarity that has allowed the creation of a team of differentiated professionals with a good balance also from the point of view of the gender of researchers: 50% is made up of women.

How was it done

The techniques used for data analysis vary according to the projects. For some projects, they used correlational and factorial statistical analysis techniques, which are perfect for studying any relationships between variables and factors. This is the case of the studies on the pollution level of the city of Madrid. Another example is the project concerning education: aimed precisely at studying the relationship between the quality of education and socio-economic factors. For the bullying project, on the other hand, it was necessary to carry out textual data analysis. These analyzes can also be carried out with classic statistical analysis software that have modules for analyzing textual data (such as text mining). For another project, analyzes of sound data (voices of people with disabilities) were carried out using voice recognition tools and, in particular, voice analysis using artificial intelligence (AI) algorithms. For the project on the crying of newborns, they used a technology based on deep learning: AMSI (Acoustic MultiStage Interpreter), which, after processing the sound (specifically, cries of babies aged 0 to 6 months) through deep learning models, extracts a meaning from it. Finally, the analysis of big data using various software and algorithms is very useful in the case of research on

collaborative transport.

Challenges

There are not yet many examples of the application of data science to fields such as social sciences, so it may not be so easy to find the right idea and start a project in this area. The SoGooD²ata team, however, managed to exploit the diversity of the fields of experience of the various collaborators to create a heterogeneous context in which it is possible to work with an innovative and multidisciplinary approach. The people working on these projects are volunteers who have an interest in society at heart and want to leverage their knowledge in data science to contribute to research in multiple fields. This led them to face the problem of the lack of data on which to work in a proactive way. Not only do they help each other in the search for data sources, but they also take advantage of the contribution of people outside the organization. Through their site, in fact, they can be contacted to suggest innovative ideas and to share or indicate interesting data sources.

Results

SoGooD²ata pursues its aim of bringing together data and experts from heterogeneous domains to achieve worldwide social challenges not only within their team, but also through the organization and participation in various events, reaching a wide and diversified audience. In fact, they organized the Data Science for Social Good Summit in October 2020, together with DSSG Portugal and Nova SBE. The event was intended to inspire NGOs and governmental institutions, by using practical examples on how to apply Data Science. Dr. Laguna Pradas has also participated in events such as I4DS-Inspiration for Data Scientists (organized by the company SAS "Statistical Analysis System"), and WIDS (2019), an initiative founded in 2015 at Standford University that aims to inspire and educate data scientists worldwide, support women in the field and encourage them to connect one another. On both occasions, she gave a presentation on artificial intelligence to interpret the

cries of newborns. This project is one of the most successful and has also led to the creation of the startup Zoundream (https://zoundream.com/), thanks to which Dr. Laguna Pradas was included in the Mujeres Referentes del Emprendimiento Innovador en España in 2021.

Tips for others

The idea of applying data science to other fields, such as the human and social sciences, has certainly proved successful. The greatest strength of SoGooD²ata, however, is the composition of the team and the percentage of women within the team (50%) is a figure that should not be underestimated. In fact, very few women study or work in data science and for some time initiatives and events have been proposed in order to encourage women's participation in data science. In this sense, will SoGooD²ata have found the right formula? It was probably their purpose and their way of applying data science also to human and social sciences (where the percentage of women is higher) that led to this result. It might be appropriate, however, to deal more with the advertising so that more people get to know this NGO and its projects, which are only mentioned on their website. The projects, research and technologies they deal with could be presented in more detail on the website. It would be important both for mere knowledge and because they can be a source of inspiration for others.

Visit the website

https://sogooddata.org/

Aula Magna Business School

Aula Magna Business School (https://aulamagnabs.com/) is a specialization school founded by Clara Lapiedra in 2020. Aula Magna Business School's goal is to reduce the gender gap at the executive level, guiding the education of women in any of their fields. With its all-female team, it offers

several training programs for professionally active women and for companies who, by valuing diversity, push forward equality plans. Their commitment is not only towards professional women and companies: AMBS is committed to an evolution of society, which has a direct impact on 4 of the 17 Sustainable Development Goals of the 2030 Agenda, established by the ONU: quality education, gender equality, decent work and economic growth, reduction of inequalities. The courses, subsidized by Fundae - Fundación Estatal para la Formación en el Empleo (https://www.fundae.es/) are the following: Executive Development Program, Wompreneur Program, Sustainability and New Economic Models Program, Data Analysis for managers Program. The latter arises from the awareness of an alarming fact: women represent only 21.4% of data scientists (TechRepublic, 2019). The program is carried out completely online and includes 50 hours of total work, between theory and practice, divided into 10 weeks. The contents are based on the idea of creating value starting from data. It involves training on the functioning, roles and components of a data office, on data profiles and capabilities (from data scientist to data architect) and on business applications with data strategies. In this, as in all their programs, the AMBS professors apply the Case Method, established by Harvard University as early as 1870, which consists in providing information on a specific moment in the history of a company in which some decision is needed. It favors the immediate application of the concepts learned. The cases used as well as being current, global, digital and of world-famous companies, are cases in which the protagonists are always women.

Why was this use case chosen

2 out of 10 Data Scientists are women, the lowest figure in STEM positions (Global Diversity Report. Harnham, 2021). AMBS was born with the intent to promote and improve women's access to positions of responsibility, management and consulting and to reduce the gender gap with a powerful tool: training. The AMBS team of professors and professionals therefore felt that to achieve the set goals it was necessary to organize a training course on data

science, in particular on data analysis. AMBS is a success story both for women, as well as for companies and society. With the inclusive approach proposed by AMBS, society wins with the fight for the wage gap and quality education, reducing inequalities. In an interview with elEconomista (2021), Clara Lapiedra in fact says: "the numbers are always right. Diverse companies produce more diversified products or services and, therefore, are able to better meet the needs of their customers and this returns to their profits". The fact that diversity is a resource and source of innovation is now more than ever a reality.

How was it done

The Data Analysis Program for AMBS Managers is entirely delivered through their ecampus. The topics are treated with a company and business-oriented approach. The first part of the program, in fact, focuses on data-driven organizations and presents data as a strategic resource for creating value in organizations and the keys to a successful data-driven transformation, with a data strategy that allow to meet the business needs with technology. The program continues with the studying in detail of the main functions related to data: data governance, data engineering, data visualization and analysis. Here there's a switch from theory to practice. Practical workshops then follow, which are useful for developing basic knowledge in data visualization and science through the use of the most popular tools on the market. After that, they explore the ways in which the data strategy is made operational and how a great diversity of profiles is coordinated to create value from the data. Finally, there's a work on the ethical and legal aspects of Big Data, such as privacy, intellectual property, protection of personal data, etc.

Challenges

AMBS is aimed at working women, who are already professionally active, so it was necessary to think of flexible methods that allowed school clients to be able to combine work and lessons, without taking away too much free time (hence the solution of adapting the program to a single day weekly: always on

Friday afternoon). Consistent with its objectives, this business school has benefited from an e-learning training model via their e-campus. A great challenge arises with online training: avoiding demotivation and boredom. For this particular reason, AMBS has adapted the contents to a winning scheme. The courses are delivered with a highly interactive model both between the participants and with the teachers. The immediate application of the contents (thanks to the Case Method) and the interaction between the participants that occurs to provide a solution to the cases is a formula that seems to have the power to keep attention and motivation high. In addition, available EdTech solutions and technology are exploited and updated from edition to edition to maximize the user experience. Finally, all AMBS programs aim to be as current as possible and consistent with the needs of women and companies today. In this regard, before the launch of this business school, a survey was carried out for women and companies who usually enroll in training programs.

Results

In just one year of activity, AMBS won the Talent Award in the Equality category from the Barcelona Chamber of Commerce and appeared in a special edition of the Financial Times, where they talked about training programs. Furthermore, thanks to their data analysis program, AMBS has become ambassador of the Women in Data Science (WiDS) program of the University of Standford, an important initiative that aims to inspire and educate data scientists worldwide, support women in the field and encourage them to connect one another. Clara Lapiedra in 2021 stated that 3 out of 4 participants change jobs in the 6 months after graduating from their business school, whether it is a change of company or a promotion within the same, with the attribution of greater powers and responsibilities. Another indicator of the success of AMBS's work is the solid network that the participants have the opportunity to build. This is because a network full of discussion and support is often an excellent opportunity for empowerment. It is no coincidence that this is one of the assumptions on which WiDS itself is based on. Finally, AMBS has

quickly gained a lot of exposure and the city council of Santa Coloma de Gramenet, a city in the province of Barcelona, will be awarding scholarships to its citizens for the Data Analytics for Managers program. This initiative is part of their 'Competents' campaign, based on several pilot projects aimed at reducing the employment gap.

Tips for others

Given the excellent results, it is clear that AMBS has identified a need (of women, companies and society), and then successfully filled it. The example of this business school shows how it is necessary and socially positive to offer tools and social networks to support women in their ascent to management positions and, in the case of the data science course, to enter fields with little female presence. In a world of work where more and more transversal and multidisciplinary profiles and skills are required and valued, it could also be possible to organize something not so limited to the context of private companies. The data analysis program for managers was born, in fact, because of the low presence of women in managerial positions and in data science, therefore from the desire to bring them closer to a digital and data-driven world. But the value and need for a digital and data-driven approach exists in all sectors. Therefore, we could also try to encourage the use of data science in fields where the percentage of women is high, such as the human and social sciences, pursuing a concept of transversality rather than specialization with the idea of only a vertical ascent.

Visit the website

https://aulamagna.us/

Big Data analytics for better informed qualitative research: case studies from the European Statistical System (ESS)

Lead by the National Statistics Institute of the Netherlands and implemented along with all other EU Members States' National Statistics Institutes, the

ESSnet Big Data II project is an transnational initiative carried out within the framework of the European Statistical System aimed at exploring and valorising the use of big data source – and following applications – for better informed decision making systems.

Within this initiative, two clusters of pilot actions are foreseen:

Implementation projects – which implements advanced Big Data analytics in the following fields:

- online job vacancies
- enterprise characteristics
- measuring electricity consumption, identifying energy consumption patterns
- maritime and inland waterways statistics, environmental statistics

Experimental projects – which implements advanced Big Data analytics in the following fields

- Financial Transactions Data
- Earth Observation
- Mobile Networks Data
- Innovative Tourism Statistics

The methodological approach and the quality of deliverables is coordinated by a stand-alone Work Package, namely "Methodology and Quality", which includes specifically key coordinates to strengthen a coordinated approach to:

- Literature overview
- Quality issues and potential alternative solutions
- Methodologies applied and challenges identified

For the scale and scope of the DATA SCIENCE project, three pilot projects stand out specifically: the one on online job vacancies, (implementation pilot), the one on enterprise characteristics (implementation pilot), and the innovative

tourism statistics (experimental pilot). These three initiatives manage to tackle those dimensions and fields of interest that, more than the others listed, can trigger interesting insights for other professionals not necessarily involved in quantitative analysis, but highly impacted by the quality and nature of data for qualitative research.

Moreover, being these three initiatives carried out by National Statistics Institutes of EU Member States, readers and users of this case study can remain assured on the quality of deliverables and the relevance of results and the methodology behind it beyond what might be territory-specific circumstances.

Target groups of the DATA SCIENCE project might found themselves involved in either the employability or the tourism sector, or at the intersection of both.

How was it done

Each of these three pilot initiatives has its own methodology and work breakdown structure:

- Innovative tourism statistics
  - Task 1, Inventory of big data sources related to tourism statistics
    - Web Scraping
    - External data administration
    - Source characteristics – variable identification, variable taxonomy, variable mapping, variable ontology (relationship and hierarchies description)
  - Task 2, Examining availability, legal aspects and the quality of the new identified data sources used in the project
  - Task 3, Developing a methodology for combining and disaggregating data from various sources
    - Combining Data through statistical methods – Non Extensive Cross Entropy, Adjusted Spatio-Temporal Disaggregation Model
    - Spatial-temporal disaggregation of tourism data
  - Task 4, Flash estimates in the field of tourism compared with existing

statistics

- Task 5, Use of big data sources and developed methodology to improve the quality of data in various statistical areas
  - Verification of estimations

Source: https://ec.europa.eu/eurostat/cros/content/WPJ_Overview_en

- Enterprise characteristics
  - Task 1, Design and development of transparent webscraping policies
  - Task 2, Generalised and extended methods, procedures and implementation requirements for webscraping on enterprise characteristics
  - Task 3, Experimental statistics, including reference metadata (i.e., Enterprise URLs Inventory, E-Commerce in Enterprises, Social Media Presence on Enterprises webpages, sustainable enterprises, etc.)
  - Task 4, Starter Kit for webscraping
    - Procedures for testing and maintenance of web scraping
    - Implementing the functional production prototypes
  - Task 5, Quality template for statistical outputs
    - Quality management template for webscraped enterprise characteristics

Source: https://ec.europa.eu/eurostat/cros/content/WPC_Overview_en

- Online job vacancies
  - Task 1, Setting-up of the framework
    - Identification of statistical production processes and capabilities
    - Definition of the conceptual production processes
    - Developing and evaluating scenarios for data governance
    - Data management in the wider aspect of data sharing
    - Collaboration processes management
    - Software for obtaining and processing third party data sets

- Transforming web data from job portals into the structure for analysis
- Computing statistical outputs

o Task 2, Statistical output

- Investigation and selection of potential use cases and define indicators(s) in the field of job vacancies statistics
- Investigation of the methodology for the calculation of the indicator(s) in the field of job vacancies statistics
- Calculation of the indicator(s) in the field of job vacancies statistics
- Quality assessment of statistical outputs (e.g. accuracy, sensitivity, specificity)

o Task 3, Implementation requirements of prototypes in the relevant statistical production processes at European and national level

- Definition of the implementation requirements of prototypes in the relevant statistical production processes at European and national level
- Outline of the architecture, processes and infrastructure for future production of statistical outputs in other statistical domains

Source:
https://ec.europa.eu/eurostat/cros/content/WPB_Overview_en#Task_1_.E2.80.93_Methodological_framework

Challenges

Challenges relate to the specific scenario of implementation.

Innovative Tourism Statistics:

- Preparation of inventory of data sources related to tourism statistics
- Identification and classification of source characteristics
- Design of scalable approaches for data collection from the Web through web scraping
- Linking and matching data integration techniques
- Implement spatial-temporal disaggregation related to the capacity of

tourist accommodation establishments

- Combining data for expenditures estimation per country-specific analysis

Enterprise characteristics

- Handling, processing and managing rapidly growing volumes of data – generating in turn higher consumption of computing and storage resources
- Webscraping increases in complexity as the internet evolves, as more and more data become available and diversify in its format (some of which increasingly more difficult to extrapolate compared to others)
- The "natural" element of non-representativity of some web-scraped data, and the biases in the estimations that might come from it. A key challenge has been represented by the recognition and intervention on this biases and how to adjust them to reduce the margin of error
- The lacking of training data for machine learning of sufficient quality

Online job vacancies

- Online job advertising might not be necessarily "100%" representative of the labour market dynamics as not all job vacancies are announced online – i.e., some job positions are more likely to be advertised online compared to others. As such, the dataset risks to be biased towards certain professions and positions (i.e., typically those requiring higher qualifications)
- Some job vacancies are often published simultaneously on more than one job-searching website (i.e., recruitment agencies and corporate websites'), an element raising the issue of the so-called data duplication
- Publishers may use different formatting of the advertisement, favouring some highlights and details of the publication over others. From a statistical perspective this requires the challenge of designing classification algorithms that takes into consideration such diversity
- In a large sample of cases, the online job advertising does not include information about the employer, with no possibility of including data on the

sector / industry to which that advertisement refers to

The full work-breakdown structure of the project is available here: https://ec.europa.eu/eurostat/cros/content/essnet-big-data-1_en

Tips for others

The amount of information available is of great extent. As such, it is critical to learn how to navigate the library of resources with relative ease and effectiveness. For case studies as the one here discussed, users are recommended to look at first into the table of content, so as to get a better sense on what the case is, and what it deals with.

From there, is it of great use to start assessing the background and objectives of the given case: typically, the background provides for inputs on the motivation and overall rationale of the implemented activities, with insights on the needs addressed and opportunities to which capitalize on; the objectives' section introduce readers to what will be the expected results and means / tasks designed for their achievements.

Once readers have done approaching the overall scale and scope, they can step into the more technical snapshot of the case, namely the list of work packages. Each work package has its own title and general description, objective, list of tasks foreseen for its implementation, list of milestones and concrete deliverables. Usually the information available from the general description are enough to understand if that given work package might be of use or not, when in doubt readers are recommended to move on with the description and content of tasks – this will provide for them even further awareness on the relevance of that case.

Once the spin-off of interest has been identified, the next most relevant field of interest is most definitely represented by the full list of deliverables produced during implementation. Deliverables are reports, analysis, files, and documents from which the users can extrapolate the needed information. In general, each

WP comes with the following deliverables:

1. Structuring of the methodological approach foreseen for that WP
2. Intermediate and Final report
3. Quality assurance report
4. Others (documents, files, etc. relating specifically to the technical workload planned for that WP)

Each of this deliverable represent a precious source of information that readers can filter and scout based on the lens of their analysis. In general, the most convenient way to assess the evidences from a report as one of those here discussed is by browsing the text after setting up a specific key word of interest (i.e., challenges, opportunities, results, etc.).

PART 3: CONCEPTUAL MAP AND QUESTIONNAIRE

The Data Science Questionnaire has been developed to provide an initial assessment of students' knowledge, skills, attitudes, needs and awareness of data science.

The data collection allows for a practical feedback and integration of the findings of the desk research and the country snapshot provided by each partner.

In fact, several conceptual models have been formulated and applied in the literature, specifically developed to study the impact of information systems by users and to understand the relationships between the different factors and the measured outcome.

In order to create the questionnaire, we started from the analysis of the literature and identified the following 3 reference scales, already tested and validated:

UTAUT (Unified Theory of Acceptance and Use of Technology)

The unified theory of acceptance and use of technology (UTAUT) is a technology acceptance model formulated by Venkatesh and others in "User acceptance of information technology: Toward a unified view".[1] The UTAUT aims to explain user intentions to use an information system and subsequent usage behavior. The theory holds that there are four key constructs: 1) performance expectancy, 2) effort expectancy, 3) social influence, and 4) enabling conditions.

[1] Venkatesh, Viswanath; Morris, Michael G.; Davis, Gordon B.; Davis, Fred D. (2003). "User Acceptance of Information Technology: Toward a Unified View". MIS Quarterly. 27 (3): 425–478. doi:10.2307/30036540. JSTOR 30036540. S2CID 14435677.

In our case we have adapted the UTAUT scale to Data Science, in other words we used the UTAUT model to analyze user intention to accept Data Science tool in their research.

Going into detail, the UTAUT model consists of four constructs that determine the level of acceptance of a technology by users, both in terms of attitude to the system (*behavioral intention* - BI), and use of the same (*use behavior* - USE).

Turkish Validation of STEAM Scale and Examination of Relations Between Art Attitudes, STEM Attitudes, STEM Awareness and STEAM Attitudes among Pre-Service Teachers

The aim of this study was to adapt the STEAM Attitude Scale developed by Kim and Bolger (2017) in order to explain the STEAM attitudes of preservice teachers and test a structural equation model composed of the attitude towards art and STEM awareness and some other variables.

We have included this scale in our questionnaire to measure students' attitudes towards data science.

Knowledge a Skills

In addition to the two scales described above, a section on "Knowledge and Skills" was added, useful for measuring the level and technical knowledge of the students of the universities being measured.

The questionnaire was therefore constructed by putting together these 3 described scales, which led to the conceptualization of the following structural model:

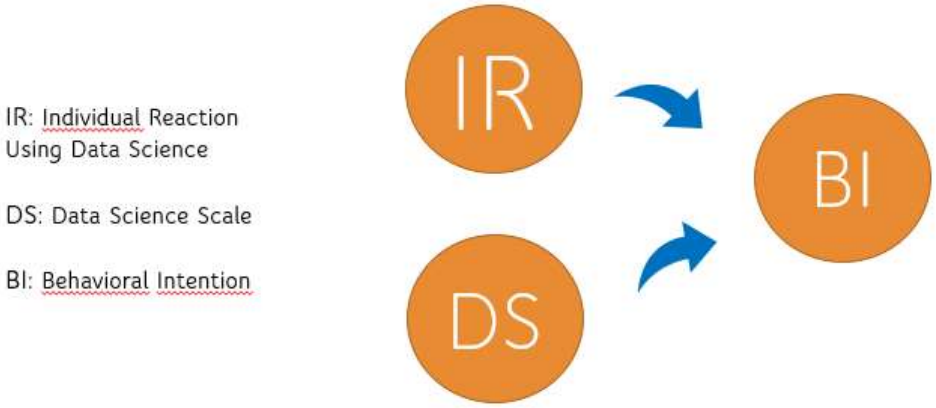


Figure 2

As for UTAUT, our model consists of 2 main constructs that determine the level of acceptance of Data Science within their courses of study by students, in terms of both aptitude for using DS models (behavioral intention - BI).

The questionnaire:

The questionnaire consists of the following 5 sections:

- The first section is dedicated to personal information (Gender, age, Country, Education level, field of degree of the studies).

Specifically, it consists of 9 questions.

- The second sub-scale is dedicated to Skill & Knowledge in order to MEASURING STUDENT
KNOWLEDGE AND SKILLS about Data science.
Specifically, students were asked to:

- Indicate which exams they have taken in subjects related to Data Science (or which they will take);
- Knowledge of specific Data Science software;
- The degree of skills in Data Wrangling, Data Visualization;
- Answer some questions of Descriptive Statistics.

This section consists of 9 closed-ended questions and one open question.

- The third scale measures Individual Reaction Using Data Science.
It consists of 6 questions; each question intends to measure a specific construct:

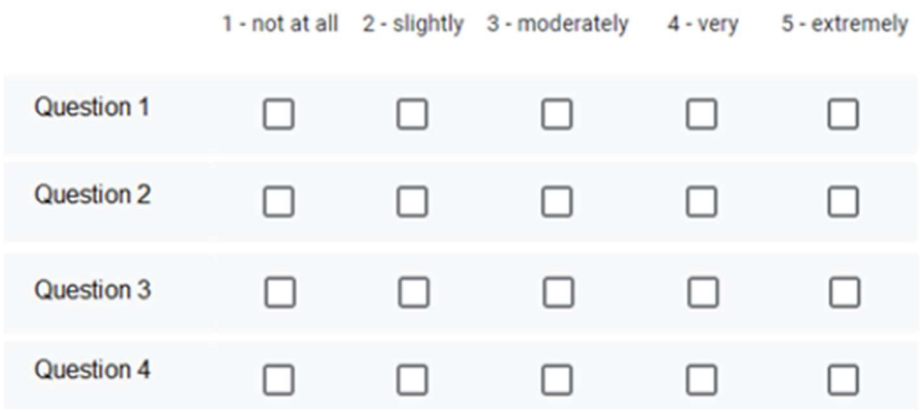
- performance expectancy
- effort expectancy
- self-efficacy
- social influence
- facilitating conditions
- anxiety

- The fourth sub-scale called" Data Science" measures:

- ✓ Attitude in DS;
- ✓ Interest in DS;
- ✓ Perceived ability
- ✓ Value.

- The fifth section, finally, measures Behavioral Intention to use Data Science, and is a single-item construct.

As shown in the figure below, items were on a 5-point Likert scale, ranging from Not at all to Extremely and passing by the neutral category Moderately.



PART 4: DATA ANALYSIS

This section is dedicated to the analysis of the data collected through the questionnaire described above.

In particular, the answers provided by the participants will be examined in order to identify any trends, common models and training gaps.

In this section the results of the analysis will be presented and discussed in order to provide a complete overview of the students' responses on the degree of knowledge and acceptance of Data Science.

All the partners were committed to data collection (in particular the Universities). Data were collected by the different partner countries, were then back translated in English, properly coded, and melt together allowing to evaluate the overall knowledge, attitude, Individual reaction, and behavioral Intention about using data Science.

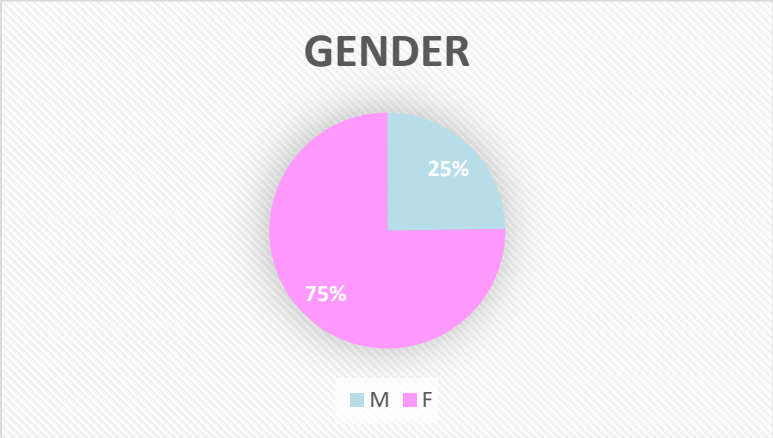
According to the project planning, the minimum threshold for the sample size would have been 300 participants.

The final sample includes instead 440 subjects where:

- 239 were provided by the ACADEMIA DE STUDII ECONOMICE DIN BUCURESTIU (ASE) - Romania;
- 154 were provided by the University of Salento -Italy;
- 47 were provided by the University of Oviedo -Spain;

Below is an overview of the data analysis broken down by Section and Question.

SECTION 1: Descriptives



From a gender point of view, data were initially unbalanced. Indeed

We have maschi 109 femmine 331

In any case, this data is consistent with the logic of the Project that intends to analyze and include women in particular in order to encourage them in training.

in STEM disciplines and· Increase their skills in data science. Therefore, it is strategic for our analysis to have a female majority.

AGE

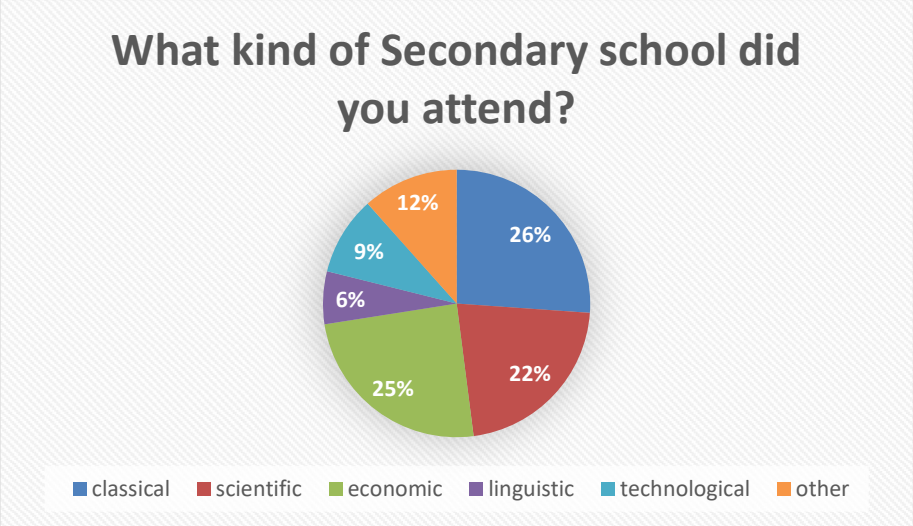
Respondents' mean age is 21 with a sd =3,20; the median, 21, reveals however that half of the overall sample is 27 years old or younger. The Table below reports the descriptive statistics of the *Age* among the different countries involved:

Descriptives statistics *of Age*

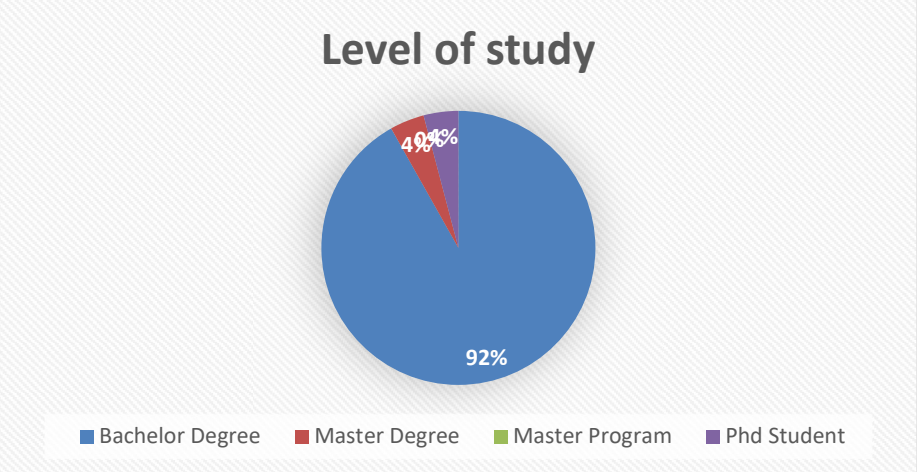
	1.2 AGE
N	440
Missing	0
Mean	21.2
Median	21.0
Standard deviation	3.20
Minimum	18
Maximum	51

With respect to the Education in Secondary School, it is possible to note a slight prevalence of scientific sectors.

However, we can say that the sample is quite heterogeneous.

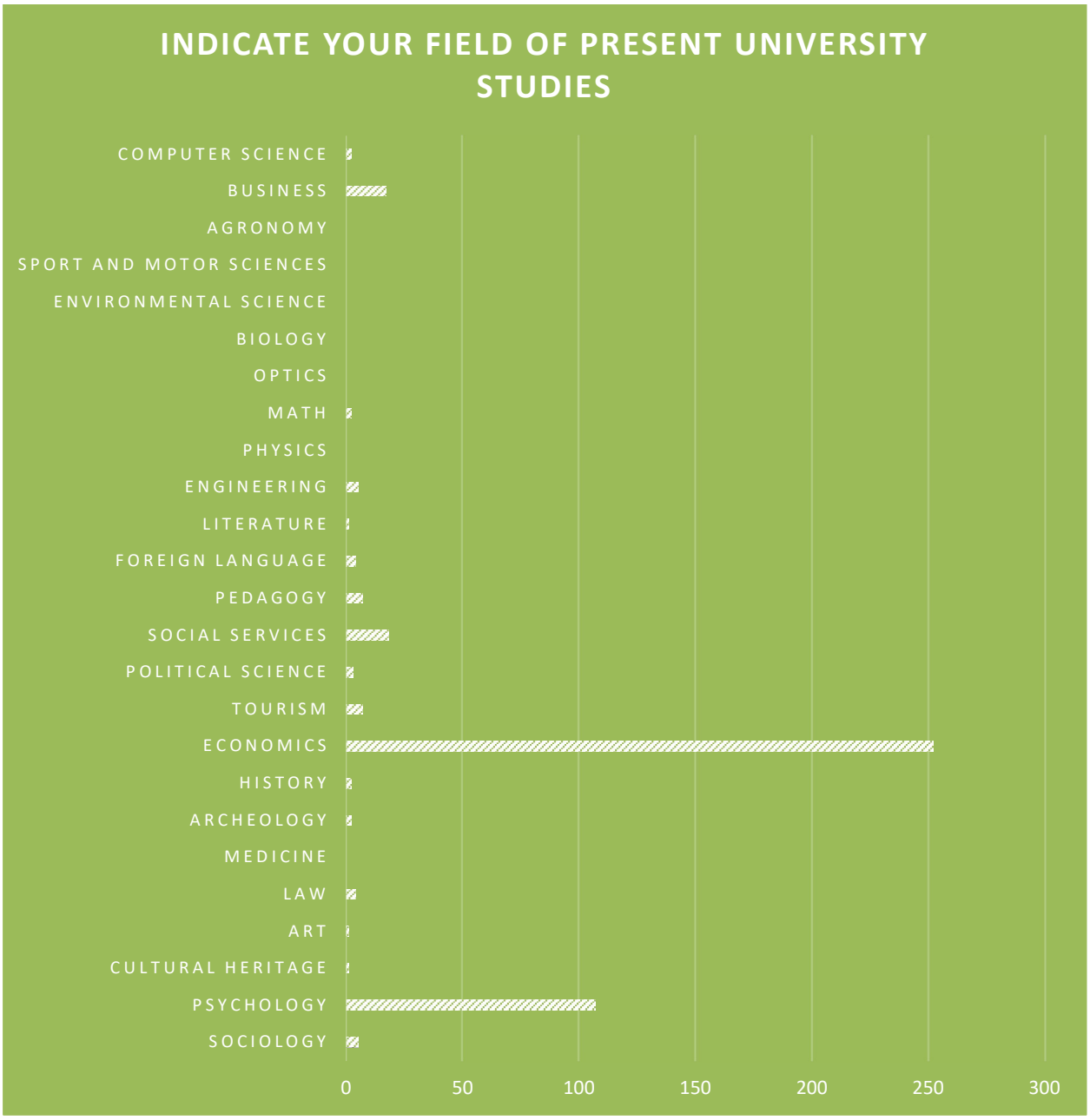


Concerning the current level of stud of the respondents, as shown as shown in the graph 92% of the respondents are enrolled in a three-year bachelor's degree program.



Relating the field of Studies, as we can see from the graph below that most of the interviewees belong to the courses of Economics, Psychology and Social Services. As for other courses in the humanities and social sciences, it is possible to detect answers from courses in Foreign Languages, Sociology, Pedagogy, History, Tourism Sciences, Political Science, Archaeology (37

answers in total).

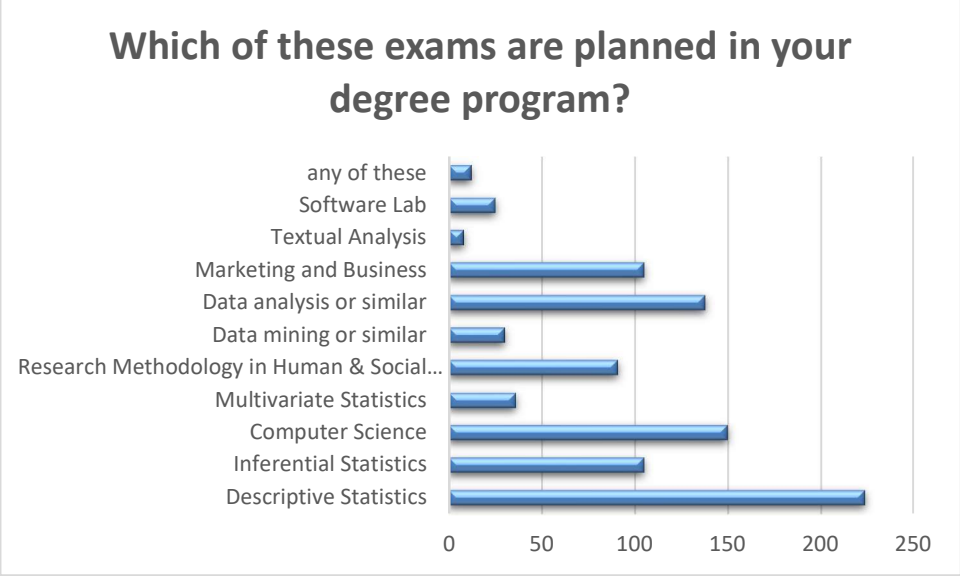


SECTION 2: Knowledge & Skills

The aim of this section is to assess students' knowledge and skills in data science, as well as their level of interest and motivation towards data science.

Specifically, students are asked to demonstrate their understanding of various concepts and techniques related to data analysis, visualization, modelling, and machine learning. This section consists of 11 questions.
Below is a question-by-question description of what each question expects from you and how it will be graded.

Question 2.1 asked:



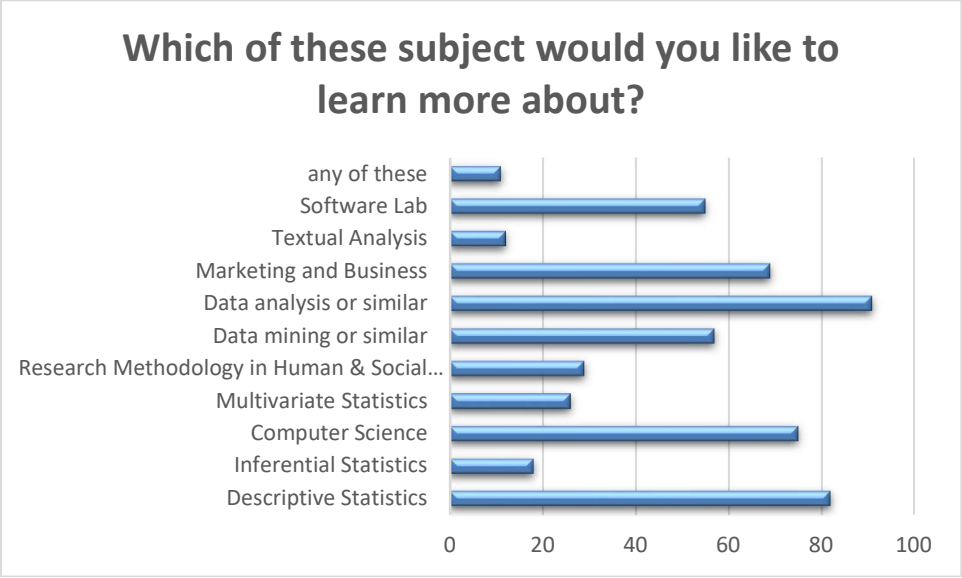
One of the objectives of this question was to analyze the distribution of planned exams in different degree programs related to data science. To this end, we asked them to indicate which exams they had to take in their degree program. The results are summarized in the graph above.

As shown in the graph, most respondents indicated that in their degree program are planned exams of Descriptive Statistics (224), Inferential Statistics (105) and Computer Science (140).

The graph shows that there is a high degree of consistency among degree programs related to data science in terms of the planned exams. However, there are also some differences that reflect the specific focus and orientation of each program. Therefore, students should carefully consider their interests and career goals when choosing a degree program in data science.

It is notable that those coming from more humanistic degree courses, such as archaeology, history, and law, have very few or none of these examinations in their syllabus. Among the most common exams in these courses is computer science.

Question 2.2 asked:



As shown in the graph above, most respondents indicated that they would like to learn more about Data analysis or similar (91), Descriptive Statistics (82),

Computer Science (75), Software lab (55).

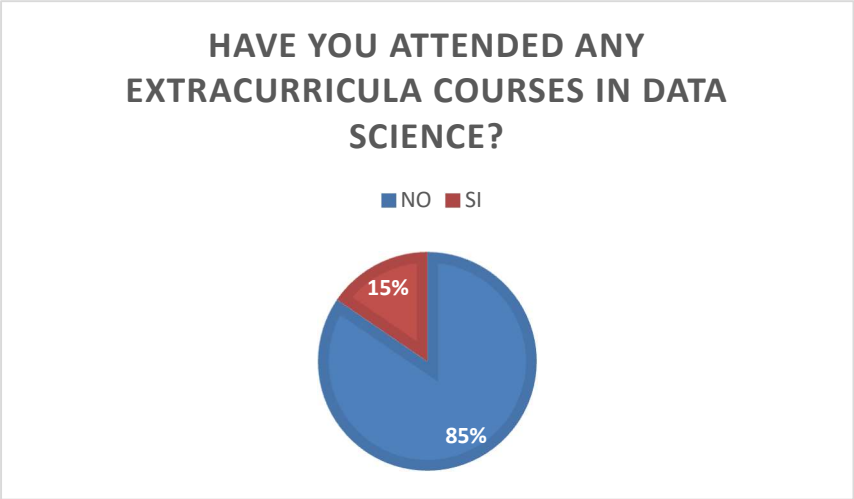
It is also interesting to note that:

- Among the purely humanistic students, descriptive statistics and data analysis are the most popular courses.
- Among economics students, Methodology of Research in the Humanities and Social Sciences is the most popular,

Another interesting aspect that emerged from the analysis is that all the respondents (except for 12 students) indicated that they wanted to deepen different subjects in addition to those included in the curriculum.

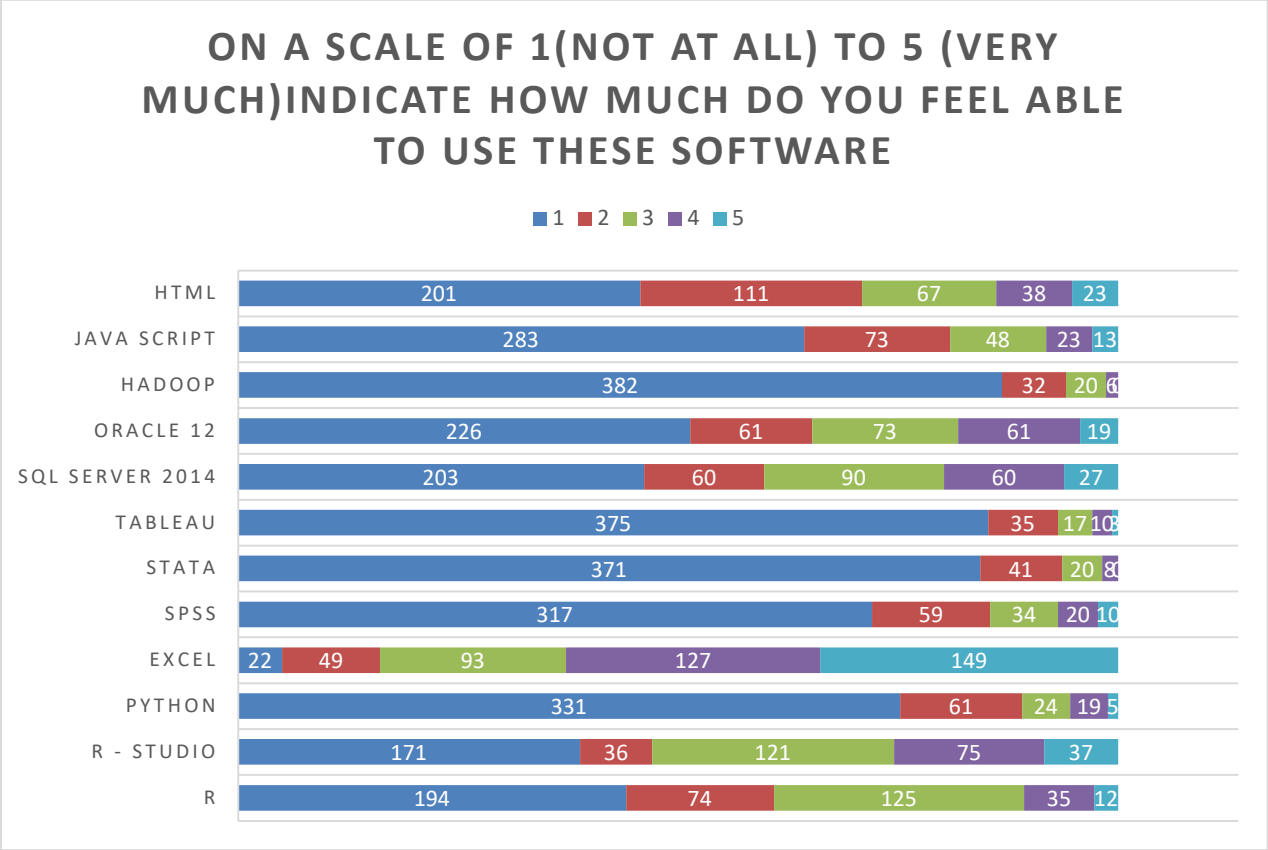
In General, the analysis of the question suggests that among students there exists a good attitude and curiosity to learn more about data science.

Question 2.3 asked:



One of the findings of this question is the low participation rate of students in data science courses outside the regular curriculum. As shown in the graph, only 68 students out of 440 stated that they attended an extra-curricular course in data science in the past year. This represents a mere 15.5% of the total sample. We believe that this is a missed opportunity for students to enhance their skills and knowledge in a rapidly growing field that offers many career opportunities and benefits.

Question 2.4 asked:



This question is dedicated on the knowledge about statistical software. Statistical software is a type of software that allows users to perform various tasks related to data analysis, such as data collection, data manipulation, data visualization, data modeling, and statistical inference. Statistical software can be used for different purposes, such as scientific research, business analytics, education, and decision making.

The question results reveal the level of familiarity and confidence that the respondents have with different software tools for data analysis. As the graph shown, the software they feel most confident with are *Excel* and *R studio*,

which are widely used and accessible programs for spreadsheet manipulation and statistical computing respectively. On the other hand, the software they know least about are *Tableau*, *Stata*, and *Python*, which are more specialized and advanced tools for data visualization, econometrics, and programming respectively. These findings suggest that there is a need for more training and education on how to use this software effectively and efficiently for data analysis.

Question 2.5 asked:

If using another software related to Data Science, please specify.

This question was answered by only 4% of respondents indicating the following software.

- Java,
- C/C++
- CSS

Questions 2.6 and 2.7 are dedicated to the self-assessment of skills in Data Wrangling and Data Visualization.

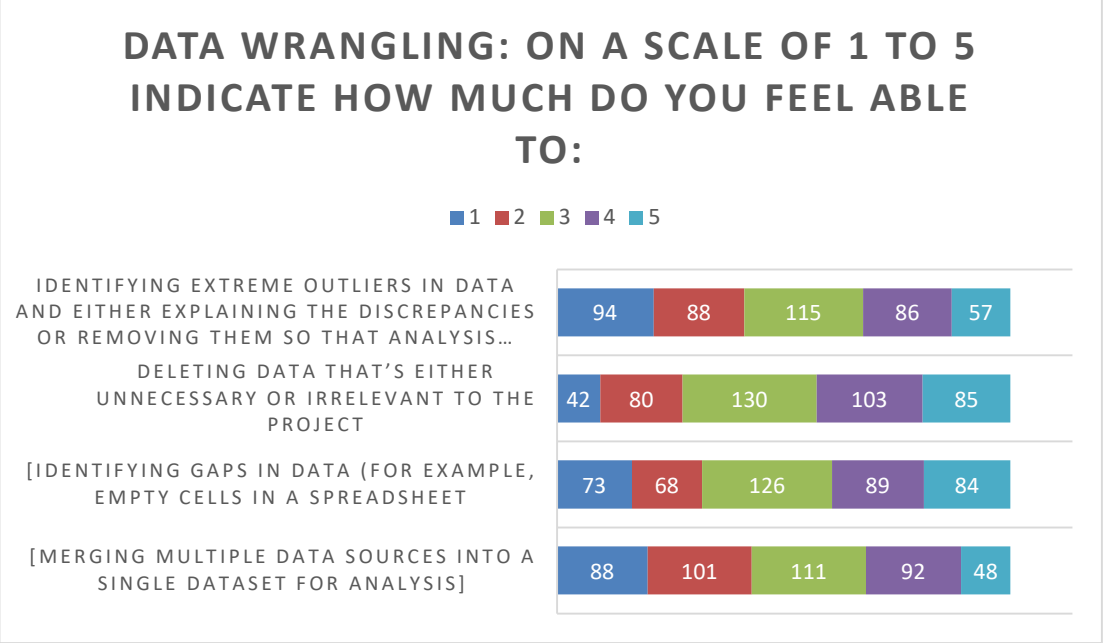
Data wrangling is the process of cleaning, organizing, and transforming raw data into a format that is easier to use for things like business analytics or machine learning. It involves transforming and mapping data from one format into another with the intent of making it more appropriate and valuable for a variety of downstream purposes such as analytics. It can be defined as the process of cleaning, organizing, and transforming raw data into the desired format for analysts to use for prompt decision-making.

Data visualization is the visual presentation of data or information. It is the graphical representation of information and data using charts, graphs, maps, and other visual tools. The goal of data visualization is to communicate data or information clearly and effectively to readers. By using visual elements like

charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.

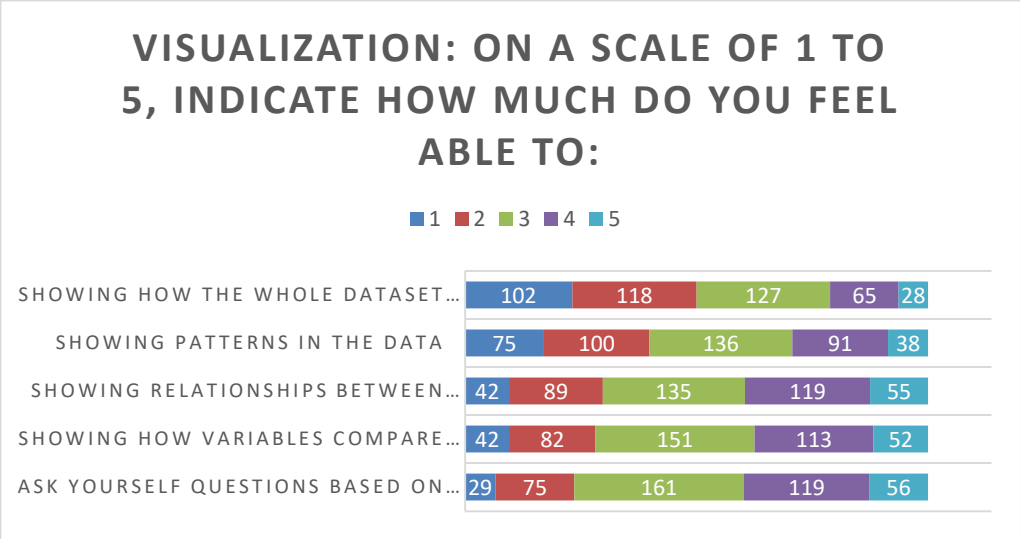
Specifically:

Question 2.6 asked:



As the graph shown, the majority of the students choose the neutral answer. In general, the answers are very heterogeneous. This indicates that there is a wide range of opinions and experiences among the students regarding their data wrangling skills. However, 42% of respondents say they have good skills in Deleting Data, and 39% are able to identify gaps in their data sources. These are two important aspects of data wrangling that can help students avoid errors and biases in their analysis. Therefore, we can conclude that some students have a high level of confidence and competence in working with data, while others may need more guidance and support.

Question 2.7 asked:

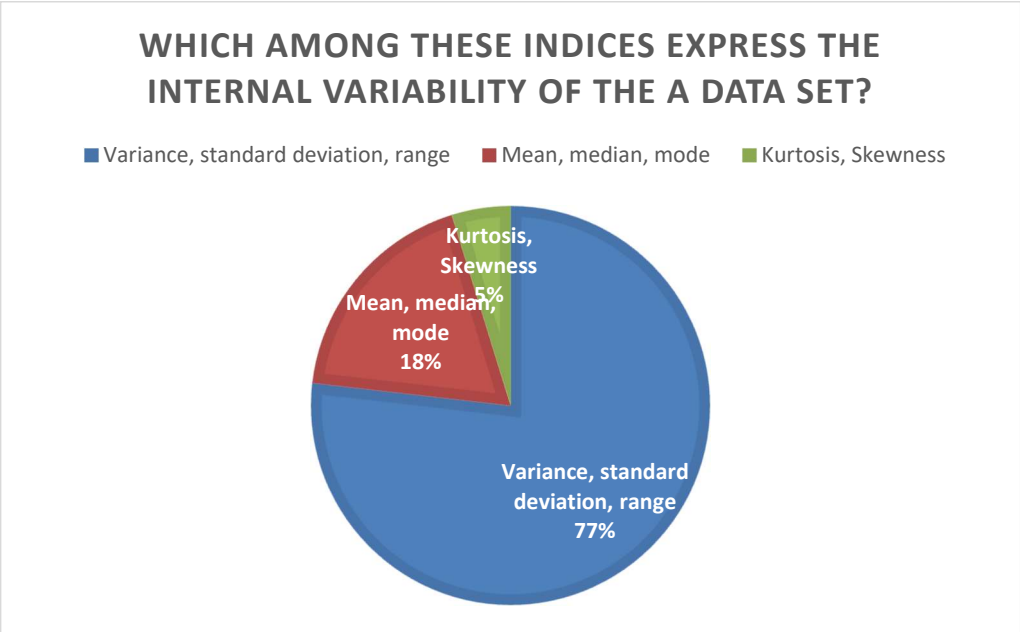


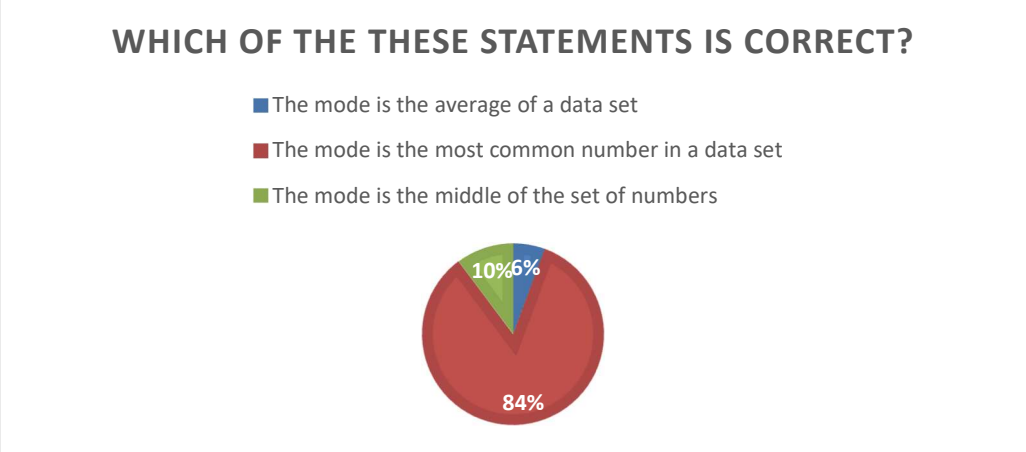
As with question 2.6, the majority of students again chose the neutral answer. In general, we can draw the same conclusions as for the previous question, namely that some students have a high level of confidence and competence in working with data, while others may need more guidance and support.

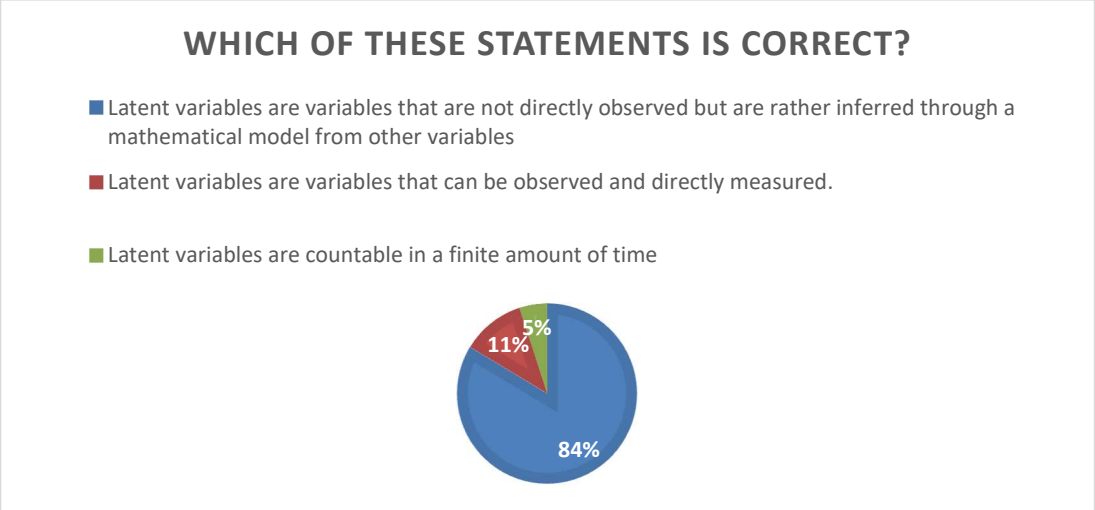
Questions 2.8, 2.9 and 2.10 of the questionnaires were designed to assess the respondents' familiarity with some basic concepts of descriptive statistics, such as mean, median, mode, standard deviation, variables etc. These concepts are useful for summarizing and comparing numerical data in various fields of study and research.

The results show that the majority of the respondents (80%) have a good knowledge of descriptive statistical indices and were able to answer all three questions correctly. However, there were some significant differences among the respondents according to their academic background. The most common errors were found among the students who belong to the more humanistic disciplines, such as tourism, languages, history, social services and so on.

These students usually do not have any statistical courses in their curriculum and may lack the basic mathematical skills required for understanding and applying descriptive statistics.







.

The question 2.11 asked:

What is your opinion about introducing or adding more data analysis/science into HSS curricula?

The majority of respondents indicated that they were somewhat or very interested in data science and would support the introduction of data science courses.
The most common reasons for learning or wanting to learn data science were to improve their skills, to pursue a career in data science, and to solve real-world problems.

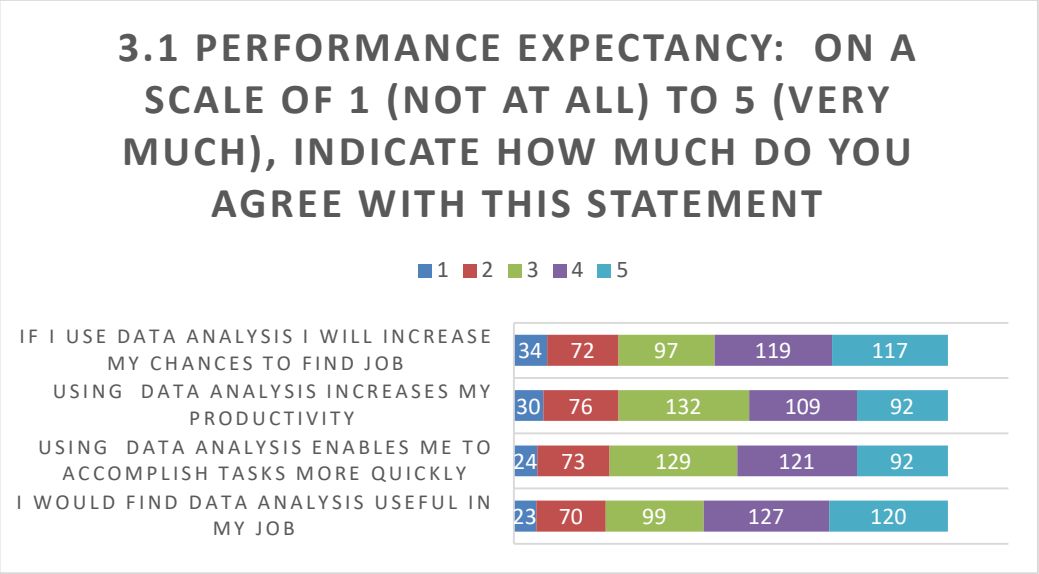
SECTION 3: Individual reaction using data science

This section is dedicated to measuring Individual Reaction Using Data Science. Individual Reaction using Data Science refers to the analysis of the response or behavior of an individual towards a particular event or stimuli, using data science techniques. It involves the collection, processing, and analysis of data to gain insights into an individual's behavior, preferences, and tendencies.

Individual Reaction using Data Science (IR) is a latent variable that captures the extent to which an individual is willing and able to use data science methods and tools in their work or study. IR is measured through 6 (six) constructs that reflect different aspects of the individual's motivation, confidence, and perceived support for data science. These constructs are: Performance Expectancy (PE), which is the degree to which an individual believes that using data science will help them achieve their goals; Effort Expectancy (EE), which is the degree to which an individual perceives data science as easy to learn and use; Self Efficacy (SE), which is the degree to which an individual feels confident in their ability to use data science; Social Influence (SI), which is the degree to which an individual is influenced by the opinions and behaviors of others regarding data science; Facilitating Conditions (FC), which is the degree to which an individual has access to the necessary resources and support for data science; and Anxiety (AN), which is the degree to which an individual feels anxious or fearful about using data science.

This section illustrates the results emerged from each of the items of the questionnaire:

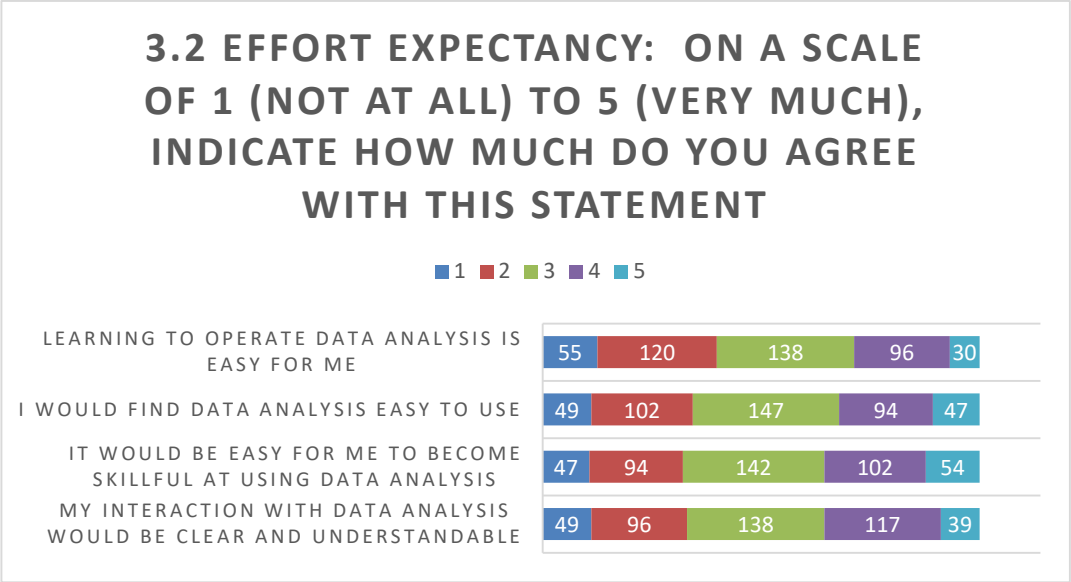
Question 3.1 asked:



The question aimed to measure the perceived benefits and challenges of Data analysis for different types of tasks and domains. The figure above illustrates one of the key dimensions of the questionnaire: Performance Expectancy, which refers to the degree to which an individual believes that using Data analysis will help him or her to attain gains in job performance. The figure shows that 49% of the answerers have a high-performance expectancy, meaning that they strongly agree or agree with the statements that Data analysis enhances their effectiveness, increases their productivity, and enables them to accomplish tasks more quickly. Moreover, 47% of the answerers with high performance expectancy also believe that Data analysis improves their employability, as it gives them a competitive edge in the job market. These results suggest that Data analysis is seen as a valuable skill and a source of competitive advantage by many answerers who use it in their work.

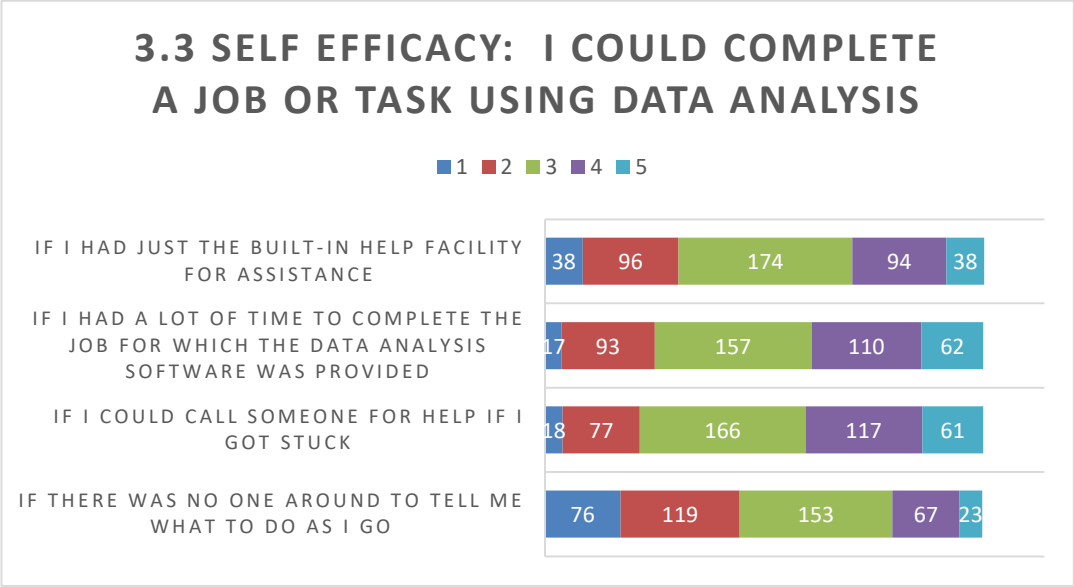
In particular, 237 people think not only that Data analysis would be useful in their job, but also that it would increase their chances to find a job.

Question 3.2 asked:



One of the factors that may influence students' motivation to learn Data Science is their expectancy about the effort required to master data analysis skills. The question results indicate that the students do not have a strong perception of data analysis as either easy or difficult to learn and to operate. Out of 301 respondents, 126 agreed that learning to operate data analysis would be easy for them, while 175 disagreed. However, the majority of the respondents (144) chose a neutral answer to this question, suggesting that they are uncertain about the effort expectancy in Data Science.

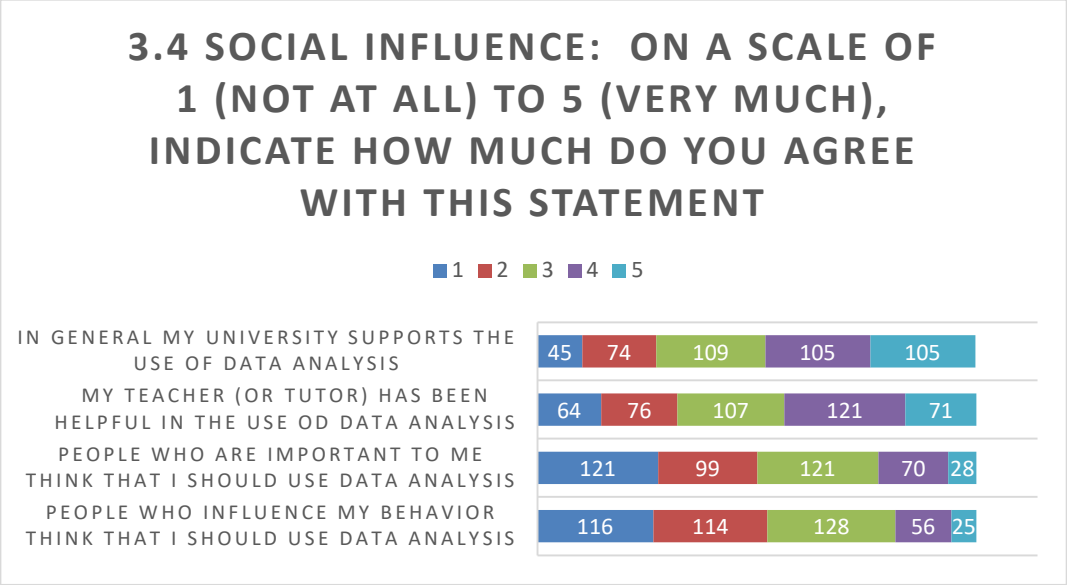
Question 3.3 asked:



Data analysis is a skill that requires confidence and practice. The figure above illustrates the level of self-efficacy of the respondents regarding this skill. Self-efficacy is the belief in one's ability to perform a task successfully. As we can see, most of the respondents have a moderate level of self-efficacy, meaning they are not sure whether they can perform data analysis tasks or not. However, a significant proportion of the respondents (more than 170) have a low level of self-efficacy, indicating that they would need a lot of time or external support to complete data analysis tasks. They would not feel

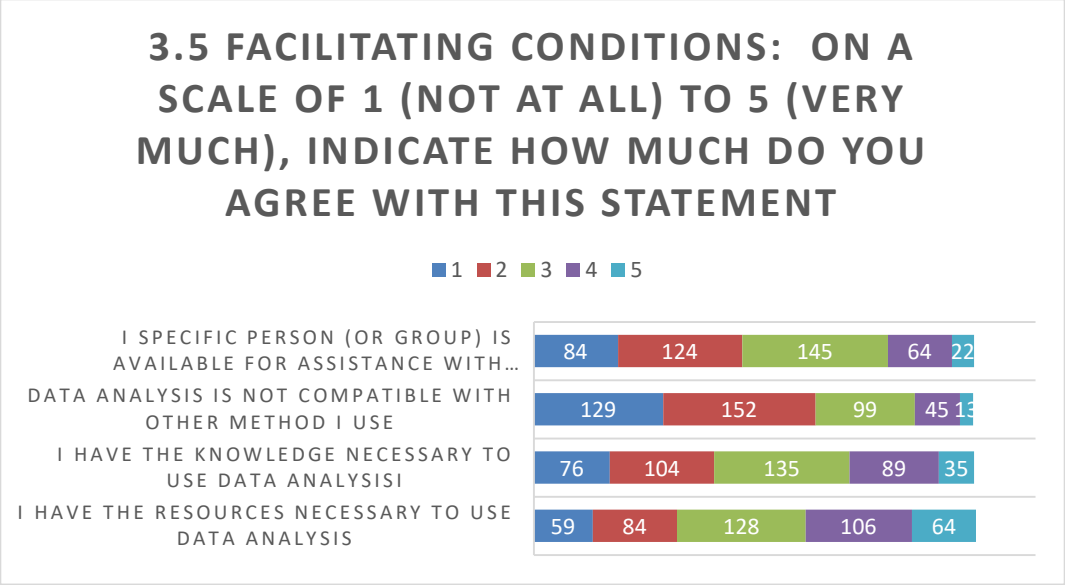
comfortable doing data analysis tasks on their own, without guidance or feedback. This suggests that there is a need for more training and mentoring in data analysis for these respondents, to help them develop their skills and confidence.

Question 3.4 asked:



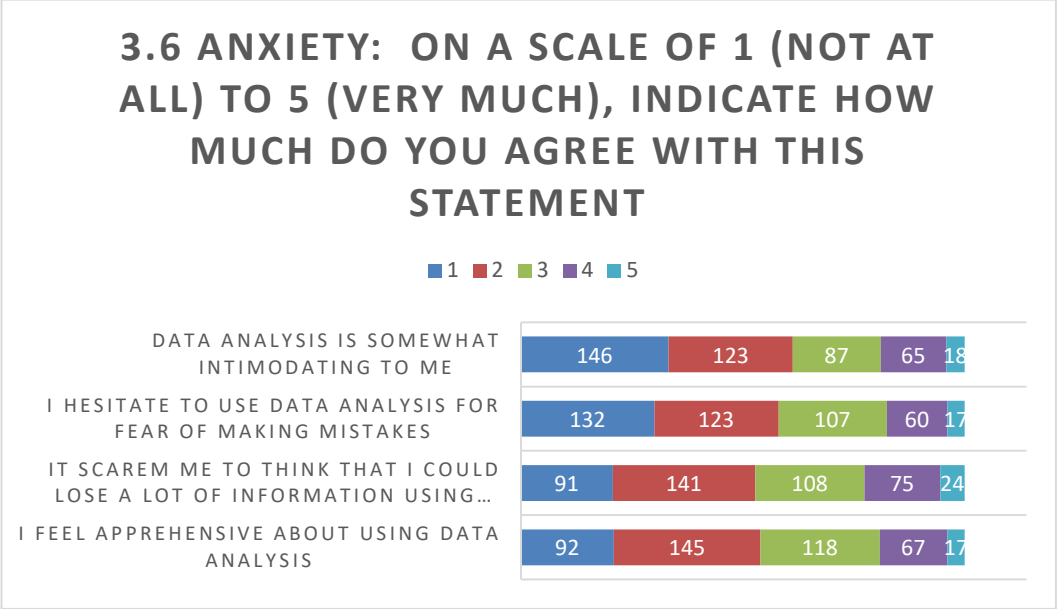
The aspect of social influence applied to Data Science refers to how the opinions and behaviors of others affect the adoption and application of data analysis methods and tools. The figure above shows the results of a question that asked respondents about the sources of influence that encouraged them to use data analysis in their work or studies. The result revealed that teachers, and the academic context in general, have a positive influence on the use of data analysis, as they provide guidance, feedback, and resources for learning and practicing data science skills. However, the figure also shows that people outside the academic context, such as peers, friends, life mentors, have a lower or negative influence on the use of data analysis. This may indicate that there are barriers or challenges for data science practitioners to communicate and collaborate with non-academic stakeholders, such as lack of trust, understanding, or support.

Question 3.5 asked:



This question asked about the facilitating conditions for using Data analysis in research. The results did not indicate a clear pattern of whether the respondents had access to the necessary resources or not. However, they did reveal that a large proportion of the respondents lacked the required knowledge to use Data analysis effectively. A noteworthy finding was that 283 respondents claimed that Data analysis was incompatible with other methods they use in their study or research.

Question 3.6 asked:



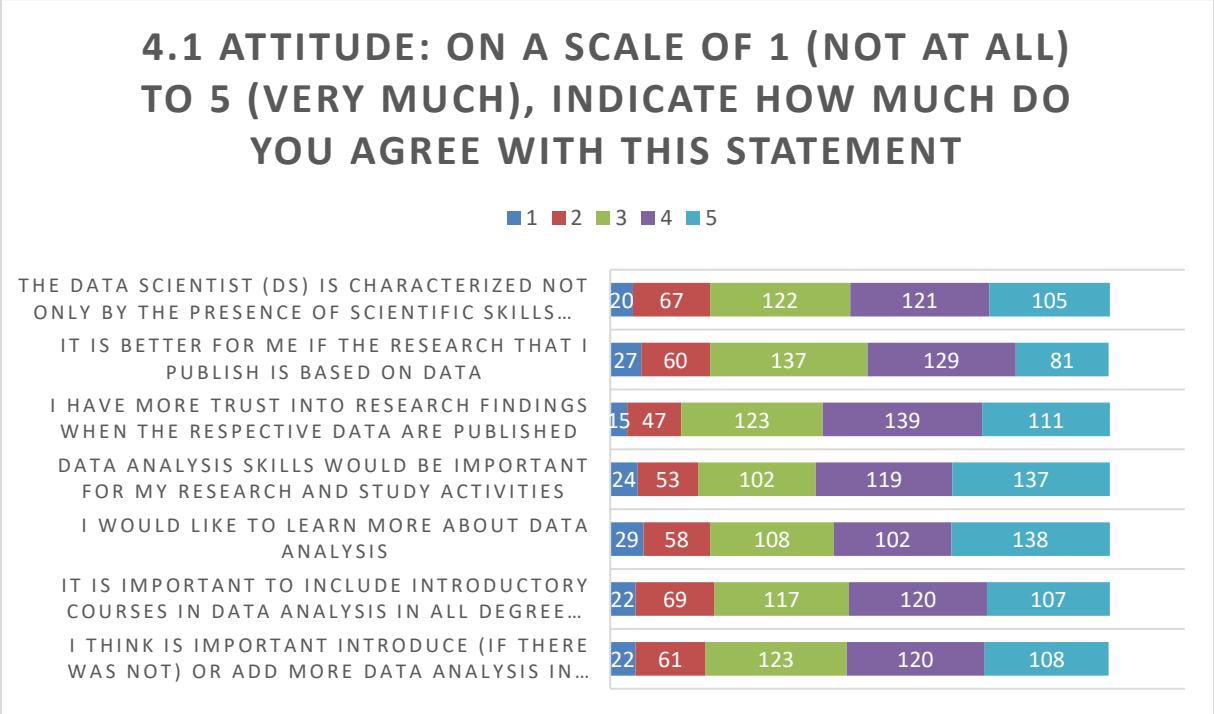
The graph above reveals some interesting insights into the attitudes and perceptions of the respondents towards data analysis. According to the graph above, 49% of the respondents believe that publishing their data would enhance the quality and credibility of their research, while 253 respondents (51%) trust more in research findings that are accompanied by published data. Furthermore, 54% of the respondents agree that data analysis requires not only technical skills, but also human skills such as creativity, communication and critical thinking. The majority of the respondents (56%) also support the idea of introducing data analysis courses in all academic disciplines, especially in the fields of humanities and social sciences. Finally, 243 respondents (49%) express their interest in learning more about data analysis, and 259 respondents (52%) consider data analysis as an important skill for their research and study activities.

SECTION 4: Data science scale

The purpose of this section is to assess the level of Data Science Scale among the participants of this study. Data Science Scale is a construct that reflects

how much a person is engaged with and motivated by data science activities and topics. It consists of four dimensions: Attitude, Interest, perceived ability and value. Attitude refers to the general feelings and opinions that a person has towards data science. Interest measures the extent to which a person enjoys learning about and doing data science. Perceived ability captures the self-confidence and self-efficacy that a person has in their data science skills and knowledge. Value denotes the importance and relevance that a person assigns to data science in their personal, academic, and professional lives. By measuring these four dimensions, we aim to understand how the participants perceive and relate to data science as a field of study and practice.

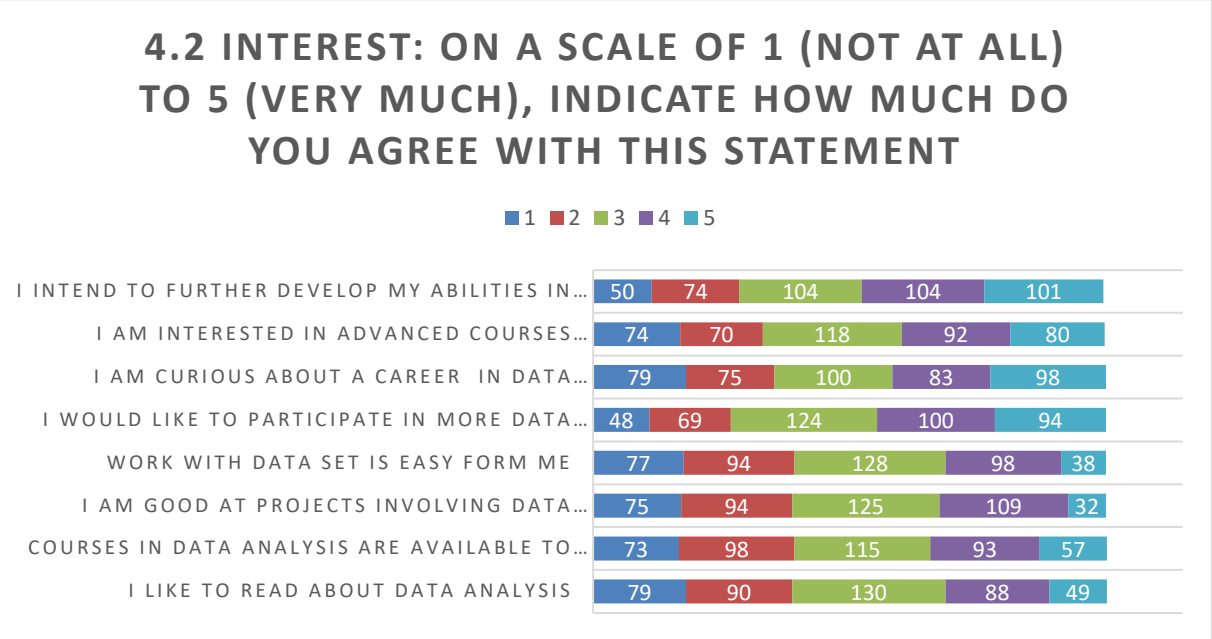
Question 4.1 asked:



The question results show a high level of interest and awareness among the respondents about the role and importance of data in research. According to the survey, 49% of the respondents believe that publishing data along with research papers would enhance their credibility and impact. Moreover, 253 respondents (52%) indicate that they trust research findings more when the

data are available and accessible. The result also reveals that the respondents recognize the importance of human skills, such as communication and collaboration, for data scientists. 56% of the respondents agree that data scientists need to have these skills in addition to technical and analytical abilities. Furthermore, the survey suggests that there is a strong demand for data analysis education across different disciplines. 57% of the respondents think that introductory courses in data analysis should be offered in all degree programs, especially in humanities and sciences. Finally, the survey indicates that the respondents are eager to learn more about data analysis and its applications. 243 respondents (50%) express their desire to acquire more knowledge and skills in this field, and 259 respondents (53%) state that data analysis would be beneficial for their research and study activities.

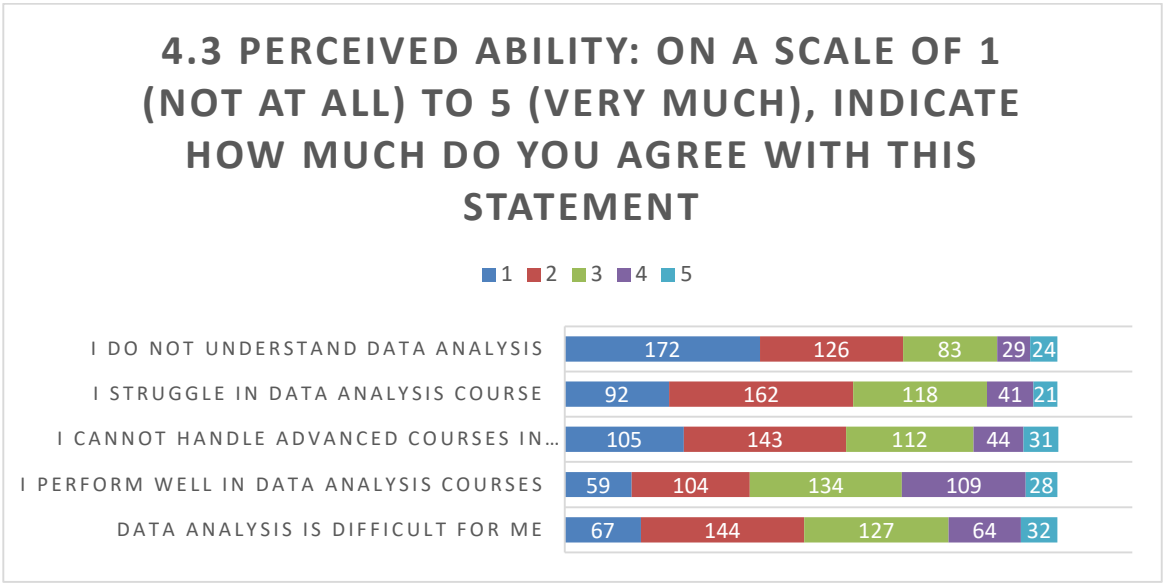
Question 4.2 asked:



The question results show that the participants have different levels of interest and experience in data analysis. Some of them have already taken several courses and applied their skills in their work or research projects, while others have only a basic knowledge and limited practice. However, regardless of their

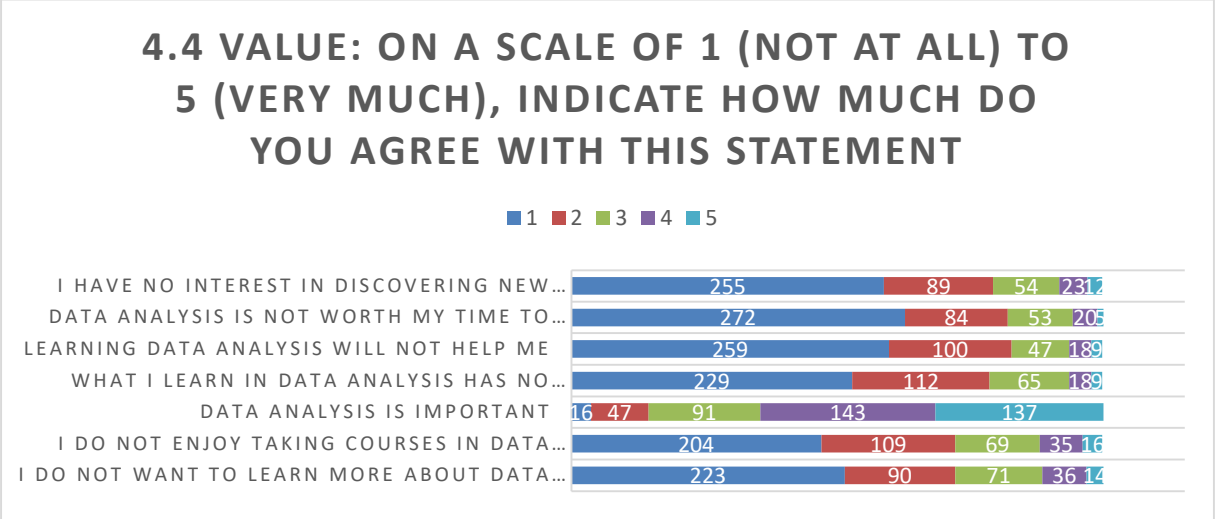
background, half of the participants expressed a desire to learn more about data analysis and improve their competencies in this field. They indicated that they would be interested in enrolling in future courses that cover topics such as data visualization, statistical inference, machine learning, and big data. This suggests that there is a demand for more advanced and diverse data analysis education among the survey respondents.

Question 4.3 asked:



As we can see from the graph this question measured the perceived ability of the students to understand and perform data analysis. The results show that most of the students do not find data analysis very hard to understand. Only 96 out of 440 respondents reported that they find it difficult, and only 53 out of 440 said that they do not understand it at all. Moreover, only about 15% of the students indicated that they struggle with data analysis and are not confident in taking advanced courses. On the other hand, 137 out of 440 students claimed that they perform well in data analysis courses, while 165 out of 440 expressed that they do not.

Question 4.4 asked:



The results of this question indicate a very clear tendency to value the importance of data science, and appreciation for data science among the participants. As shown in the figure above, only a small fraction of them (7%) expressed a lack of curiosity or motivation to explore new applications of data science, or a disregard for the value and relevance of data analysis. Moreover, only 11% of them reported a low level of enjoyment or desire to learn more about data analysis through courses. These findings suggest that data science is widely recognized as a useful and attractive field of study and practice by the majority of the respondents.

SECTION 5: Behavioral Intention

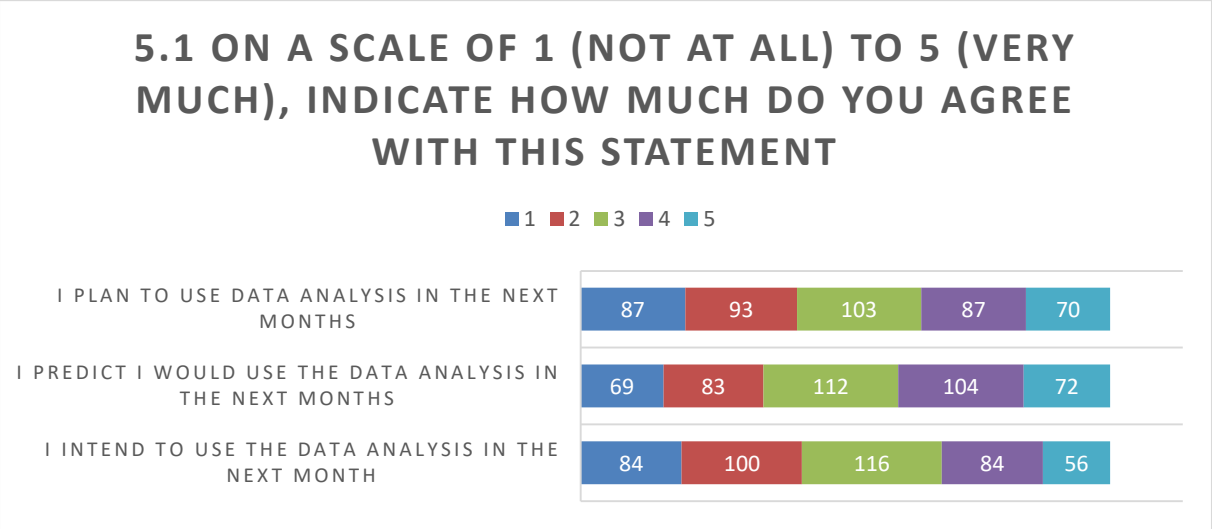
This fifth and final section has been called Behavioral intention and consists of a single item construct.

Behavioral Intention to use Data Science is a key construct used to describe an individual's intention, prediction, or plan to use data science in the future. It is a measure of an individual's willingness to engage in a particular behavior, in this case, using data science techniques. Understanding an individual's behavioral intention can help predict their future behavior and provide insights into how to encourage adoption and use of data science in different contexts.

In this study, we focus on the behavioral intention to use data science, which

refers to the extent to which an individual intends to apply data science techniques in their work or research in the future.

Question 5 asked:



The results of the survey show a mixed attitude towards Data Analysis among the participants. While a majority of them (180 out of 440) expressed no interest or intention to use Data Analysis in the near future, a significant minority (176 out of 440) indicated that they would likely use it in the coming months. This suggests that Data Analysis is not widely adopted or understood by the target population, but there is some potential for growth and improvement. The reasons for this discrepancy and the barriers to Data Analysis adoption need further investigation and analysis.

CONCLUSION

The survey data collected from 440 students across 4 universities revealed some interesting insights into their perceptions and attitudes towards data science. These results indicate that students are aware of the importance and relevance of data science in today's world, and that they have a positive disposition towards acquiring more knowledge and skills in this field. This is encouraging for educators and practitioners who aim to promote data science education and foster a culture of data literacy among students. However, there are also some challenges and barriers that need to be addressed, such as the lack of access to adequate resources, guidance, and support for learning data science, as well as the diversity and inclusion issues that affect the representation and participation of different groups of students in data science. This Report discusses these findings in detail and provides some recommendations for improving the quality and accessibility of data science education for students from various backgrounds and disciplines.

For this reason, our project Data Science in Human and Social Science for Women Empowerment aim to recommend further research and propose new actions to explore these issues and develop effective strategies to overcome them.

The main objective of this study was to develop a comprehensive and coherent framework of skills in Data science that can guide the design and implementation of educational programs and professional development initiatives.

Specifically, this study carried out:

- an in-depth analysis of the existing literature, standards, and practices in the field of Data science, as well as the current and future needs and challenges

of the labor market and society.

- a systematic and participatory process of consultation and validation with relevant stakeholders, including experts, educators, employers, and learners from different sectors and domains;
- a synthesis and integration of the findings and feedback from the previous steps into a framework of skills in Data science. The Framework includes four areas of expertise: Human and Social Competences, theoretical skills of data analysis, technical skills, non-technical skills.

In Total the framework is composed of 16 Key Competences, which are further specified by 60 Learning Outcomes that describe the knowledge, skills, and attitudes expected from a data scientist.

The Framework is available HERE

Lastly, this study will be further explored by examining the relationship between different variables related to data science. The survey data, in fact, will then be processed and analyzed using a structural equation modelling approach to validate the conceptual proposed that represents the main factors and dimensions of data science. Therefore, this study will be further explored through an analysis of the responses with R Studio in order to estimate the parameters and measure the validity and reliability of the proposed conceptual map. The study will lead to the publication of a scientific article that will be published in a scientific journal of data science.

DataScience
in Human & Social Science for Women Empowerment

2021-1-IT02-KA220-HED-000023199

Data Science

Data Science Project

datascienceeuproject@gmail.com

datascience-project.eu

Co-funded by
the European Union