Analyzing the Content of Data Science Masters' Curricula using a Dictionary-based Approach

With focus on India & the U.S.A

Dissertation – MY498

In part completion of the MSc in Applied Social Data Science

Department of Methodology

London School of Economics and Political Science

Candidate Number: 10443

Supervisor: Dr. Patrick Sturgis

Word count: 9926

August 2021

Abstract

The growing interest in data science has created a demand for well-trained professionals who can help organizations make better decisions with the help of data. As a response to this demand, data science education programs have been developed in universities across the world. These embryonic programs will greatly benefit from research that evaluates their contents to improve their effectiveness. However, there is a dearth of such research into the contents of these expensive programs and the ones that exist are mostly limited to a very small subset of the offered programs in the USA. In this project, I have made an attempt to study and evaluate these programs in a much broader sense and included programs from both India and the USA, making comparisons across the contents of data-related masters' programs as well as administrative variables such as the funding of universities and types of departments offering the program. Data on 111 programs were collected and a dictionary approach was used to quantitatively analyze and compare the programs using the Greater Data Science framework. It was discovered that current programs place great focus in computing concepts while they do not offer enough focus on topics relating to reproducibility, ethics or other concepts that make a up a "science". Differences were observed between programs across the two countries in the contents of the programs. A prototype of a web-based program recommender was developed to use the GDS dictionary created in this project to evaluate new programs. The dictionary approach was coupled with a content-based recommendation system to develop a program recommender that will connect learners to suitable programs. The tool was deployed onto the cloud. It is reasoned that this tool will be developed into a larger, more effective product in the future.

Table of Contents

Chapter 1 - Introduction

Data science as a discipline has constantly resisted definition (Donoho, 2017). A recent publication by the Harvard Data Science Review has likened data science to the popular "Blind Men and an Elephant" parable (Kolaczyk, Wright and Yajima, 2021) - Everybody sees it the way their perception allows them to. Statisticians see it as a statistical discipline using sophisticated computational tools (Cleveland, 2001; Breiman, 2001), computer scientists see it as opportunity to build data-driven products (Loukides, 2010) and generalists view it as a confluence of computing, statistics and domain expertise (Tukey, 1962; Conway, 2010; Blei and Smyth, 2017).

In 2012, Davenport and Patil published their business-oriented conceptualization of data science and the term "data scientist" was tossed into the midst of the global audience, a term readily picked up by media houses, business leaders, academic organizations and learners. Explaining the high demand-low supply characteristic of data science practitioners, the influential piece is regarded as the catalyst to the data science movement, causing much of the hype around data science as a lucrative career option (DuBois, 2010). The increased demand for data scientists and the apparent shortage of supply has motivated the establishment of several training programs that are tasked with the goal of preparing data scientists for the future. A quick search for courses about "data science" on the popular online learning platform Coursera[1] returns 2943 search. According to DataSciencePrograms[2], there are close to 1000 different programs at all study levels teaching data science and similar subjects in the USA alone. This is just the tip of the iceberg as there are a lot more other online learning platforms, and countries with universities offering these programs.

[1] https://www.coursera.org/search?query=data%20science&
[2] www.datascienceprograms.org

1.1 Formalizing Data Science Education (DSE)

A program is a "set of specific activities designed for an intended purpose with quantifiable goals and objectives" (Lodico et al., 2010). Training programs can be broadly classified into two categories - Academic and Non-academic. Academic training programs are programs that lead to a university or college degree, often requiring students to attain a specific number of credits in order to pass necessary coursework. Non-academic training programs in this context are those programs that are not taught in university or colleges, but are offered as certificate programs without the involvement of a credit system, for example MOOCs (Baturay, 2015). Non-academic programs have low entry barriers and more often than not, the only technical requirement to enrol into a program is the ability to connect to the Internet (Pappano, 2012). In general, non-academic programs tend to be less-expensive than academic ones (Pappano, 2012). Thus, choosing an academic program is a higher involvement decision than enrolling into a non-academic program. High involvement suggests a high level of investment in terms of time as well as money, and more often than not a high involvement decision cannot be reversed without undesirable consequences (Silbiger, 2016). This makes the academic data science programs an interesting and relevant area of study.

As a definition for the term has not been put forth yet, I take the liberty to define data science education here. Data science education (DSE) is *a collective set of practices involved in disseminating knowledge about data science concepts involving, but not limited to the creation of training programs (both academic as well as non-academic), the efforts of marketing these programs to a suitable market and the implementation of the programs to cultivate a responsible generation of data-driven thinkers*. The research in this project specifically deals with academic programs because of the higher stakes involved.

In section 2, a review of past work that could be brought under the umbrella of DSE is discussed. Section 3 sheds light on the theoretical framework used to evaluate current data science masters' programs and section 4 discusses the methodology of this project. Finally, section 5 and 6 discusses the results, their implications and the future of research in DSE.

Chapter 2 - Relevant Literature

Previous work in understanding the teaching of data science in academic programs can be dichotomized into research providing guidance towards creating "effective" programs and research evaluating existing programs.

2.1 Research on creating "effective" data science programs

At the undergraduate level, the focus of programs has been to introduce to students the concepts of the data science and the utility of data in the contemporary world (Anderson et al., 2014; Yan and Davis, 2019). Baumer (2015) urges program designers to create curriculums that excite students in the endeavour of working with new age data. Another recommended approach for undergraduate education is the inclusion of data management involving teaching students the process of inquiring and finding answers from "large, messy, complex, challenging data sets, for which there is no obvious goal" (Horton et al, 2015). However, given the amount of practical experience that is often expected of data scientists, researchers have argued that fully-fledged data science programs should be confined to the graduate levels of education (Irizzary, 2020). In graduate programs, the necessity of designing a curriculum focused around real-world case studies, demonstrating critical thinking ideas and popularizing reproducibility and ethical considerations has been underlined (Irizzary and Hicks, 2018; Donoho, 2017). Given the obvious interest of the industry in data science graduates, thoughts on what graduates need to know in order to get employed in the industry have also been published (Davenport et al., 2015).

2.2 Research on evaluating data science programs

Program evaluations offer two major benefits – First, program designers are able to receive feedback on their curriculums and second, this feedback can be used to improve programs and thus intellectually advance the field of DSE. Evaluations of existing program curriculums are performed using variants of curriculum analysis which in turn is an approach under content analytics in learning analytics (Lang et al.

(2017). In DSE, curriculum analysis has helped inform the field the several characteristics of data programs offered at the undergraduate as well as the graduate level (Aasheim et al., 2015; Kim 2016; Tang and Sae-Lim, 2016; Volpe and Esposito, 2020; Oliver and McNeil, 2021). While there have been attempts to understand these characteristics along with a program's compatibility to what is considered necessary for a data job in the industry (Lyon and Mattern, 2017; Börner et al., 2018; Volpe and Esposito, 2020), much research has revolved around evaluations of programs using theoretical frameworks. These evaluations are performed either at the individual course-level or at the program level. While almost similar in terms of implementation, the two approaches differ in the circumstances under which they are most useful. Evaluation of individual courses sheds light on the structure of course syllabi and its contents, thus leading to the creation of a network of concepts that are considered important in the area of instruction (Varvel et al., 2012; Kim 2016; Friedman (2018; 2019)). Such course-wise evaluation can also be used to create new syllabi by analysing existing ones (Fréchet et al., 2020). Börner et al. (2018) analysed over 2 million course syllabi records from the Open syllabus project[3] in an attempt to discover interactions between academia, industry and research in the context of data science and engineering jobs. The evaluation of programs on the other hand includes the evaluation of courses that constitute each program and these are aggregated to get an evaluation for the whole program (Tang and Sae-Lim, 2016). Program evaluations focus on how the constituent courses within a program come together to form a "balanced" whole to train a student (Zhang et al., 2017; Volpe and Esposito, 2020). It is widely accepted that a "balanced" data science program is not one that is a mere assortment of random activities, but one that incorporates several courses that together cover the full scope of data science (Cleveland, 2001; Donoho, 2017; Ortiz-Repiso et al., 2018). Program evaluations also help rank data science programs using frameworks. Popular frameworks of program evaluation include the National Academies of Sciences, Engineering & Medicine (NASEM) framework with 10 emphasis areas of data science (National Academies of Sciences, Engineering, and Medicine, 2018), the GDS framework with 6 divisions of data science (Donoho, 2017), the Data Science Action Plan (DSAP) framework suggesting 6 loci of data science

[3] https://opensyllabus.org/

activity along with necessary allocations of effort[4] and the Business Intelligence and Analytics (BI&A) framework dividing concepts, technologies and developments in business intelligence and analytics into 3 divisions (Chen at al. 2012). While the validity of the obtained rankings is yet to be proved, interpretation of these results could help unearth latent patterns and gather points on how programs are developed and how they could be improved (Oliver and McNeil, 2021).

2.3 Where does this project look to contribute in DSE?

Existing research on the content of programs offering DSE is concentrated around programs conceived at universities in the west, specifically the USA. Such a focus limits our understanding of DSE as it offers a perspective that is most likely only shared only by highly developed economies (Oliver and McNeil, 2021). This limits our understanding of DSE in developing nations such as India where the interest towards data science as high as it is in developed nations (Kaggle, n.d.). Past work has seldom considered differences arising between programs based on the departments offering the programs as well as the funding status (publicly-funded or privately-funded) of the institutions involved. Also, studies have placed heavy focus on programs that exclusively taught "data science", but often excluded programs covering very closely-related subjects such as artificial intelligence, machine learning, data analytics and business analytics. Collectively, I refer to these graduate programs as data-programs[5] here. These constraints have led much of previous work to focus on only few programs that make through these very narrow criteria of inclusion (Ortiz-Repiso et al., 2018; Volpe and Esposito, 2020).

This capstone project realizes the importance of a broader, comparative analysis between programs. Therefore, the focus of this research will be on understanding and assessing data programs in USA and India to gain a big picture of DSE and its response to the growing demand for data scientists. India and the U.S.A are known to have very strong representations in the global data science landscape when it comes to growing interest in the field of data science (Kaggle, 2020). Kaggle is an

[4] http://www2.isye.gatech.edu/~jeffwu/presentations/datascience.pdf
[5] In fact, all these subjects fall under the broad area covered by the full scope of data science (Donoho, 2017)

active virtual community of practice (VoCP) with data scientists and machine learning practitioners from all over the world (Li et al., 2021). The platform has over a million users and is the largest and most-active community dedicated to data science (Usmani, 2019). The annual Kaggle DS and ML survey is conducted to understand the state of data science in the world. The 2020 edition had over 13,000 responses from data scientists across the world. Therefore, Kaggle is a fairly reliable (and most reliable at present) source on the Internet when it comes to understanding the state of data science in the world[6]. The scope of this research is focused on masters-level taught programs as these programs tend to be more industry-oriented than undergraduate ones, and also are more common than undergraduate programs at present (Irizzary, 2020).

To summarize, the research questions this project looks to answer are

1. What are Masters' data-programs in India and the USA training students in?
    a. Are there differences in the contents of programs between India and the USA?
    b. Are there differences in the contents of programs between public-funded and private-funded institutions?
    c. Are there differences in the completeness[7] of programs based on differences in departments offering the program?
    d. Are there differences in the completeness of programs based on the offered subject?
2. What are the keywords used to describe these programs on the program webpages?

[6] https://towardsdatascience.com/why-does-everyone-use-kaggle-db1bdf1f1b1a
[7] Identified by the characteristic of a program to cover all necessary areas (areas defined in chapter 3)

Chapter 3 – Theoretical Framework

As discussed in section 2.2, there exists several different frameworks to evaluate data science programs. While ones like NASEM are specifically designed for undergraduate data science programs, the work in Donoho (2017) was explicitly conceived for a graduate-centric focus.

Figure 1. Greater Data Science Divisions (Donoho, 2017)

This research makes use of the 6-division partitioning of data science introduced in Donoho (2017) (Figure 1). According to Donoho, this framework named "Greater Data Science" (GDS) considers the complete scope of data science as an intellectually exciting field than one that is a mere assortment of tasks. The GDS is a simple, conceptual construct that envisages data science as the journey of learning from data. Oliver and McNeil (2021) have used the GDS framework in their analysis of 18 undergraduate data science programs in the USA. In the research, the authors qualitatively coded each of the 18 programs based on course titles and descriptions and conducted statistical analyses, but limited the search to only the composition of each division in the 18 programs.

Table 1 includes the different words that will help identify the existence of a particular GDS division. These words have been extracted manually from Donoho (2017). I have also added in a few words external to those explicitly mentioned in Donoho (2017), but help identify existence of a relevant GDS division.

Table 1. GDS divisions and associated words

GDS Division	Words indicating this division (as per Donoho, 2017)	Words added by the author
GDS 1	Gathering, experimental design, data resources, web scraping, image processing, social media data collection, reformatting, recoding, preprocessing, data cleaning, wrangling, preparation, exploration, EDA, insights	Data curation
GDS 2	Data representation, transformation. Data formats, modern databases, text files, spreadsheets, SQL, noSQL, data streams, distributed databases, structures, transformations, algorithms, math representations, acoustics, images, sensors, network data, fourier transform, deployment, deep learning	Data management, metrics, real-time, measurement
GDS 3	Computing, languages, R, python, specific languages for text manipulation, managing computational pipelines, efficiency, cluster computing, cloud computing, workflows, parallel jobs, documentation of steps of data analysis/research project, package development	Data engineering, deployment, real-time
GDS 4	Modeling, generative modeling, inference, traditional academic statistics, stochastic model, predictive models, machine learning	Deep learning, mining
GDS 5	EDA, plot types, developing plots, dashboard, monitor data processing pipelines visually, present conclusions, visualizations	
GDS 6	Tool evaluation, analysis of workflows, processing workflows, validity, effectiveness of workflows, new trends in data science, encoding documentation, meta-analyses, data quality, methodology, scientific understanding, reproducibility	Ethics, responsibility, privacy, problem framing, logic

A given word can indicate the presence of more than one GDS division

Chapter 4 – Methodology

As implied by the research questions, this research is posed as a comparative analysis, looking to find similarities and dissimilarities in DSE between India and the USA. Esser and Vliegenthart (2017) discuss a framework for comparative analysis which includes the explicit stating of the comparison's purpose, delineating the units of comparison clearly and comparing units with at least one common, functionally equivalent dimension. In this research, I have made attempts to incorporate these recommendations as follows.

1. **Purpose of comparison:** Stated in the research questions (section 2.3)
2. **Units of comparison:** India and the USA, both differentiated on the basis of their geographical, political, social and economic boundaries
3. **Common, functionally equivalent dimension:** Both countries have an interest in data science far higher than most other countries in the world (Kaggle, 2020)

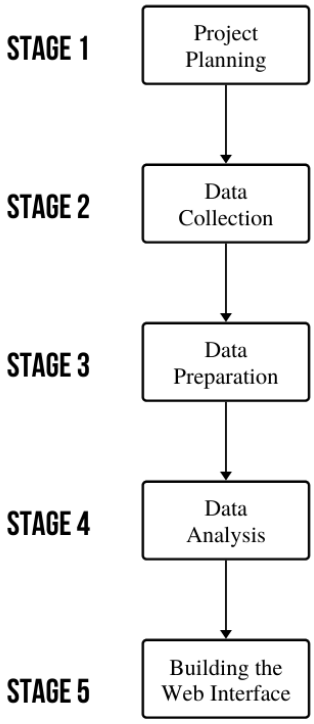


Figure 2. Methodology Diagram

4.1 Stage 1: Project Planning

The planning stage was crucial to identify sources of data, sampling strategies and required data variables. The **population** under consideration here was all Masters level academic data-programs[8] in India and the USA. These programs will be identified by the presence of at least one of the *key-terms* in its title. The selected key-terms were *data, analysis, analytics, artificial intelligence, machine learning, ai and ml*. At this stage, it was also confirmed that all data programs offered by universities in India and USA among the top 50 of their respective national QS university rankings[9] will form the **sampling frame**. There are two reasons for this decision; the QS ranking is among the most widely accepted rankings in the world (Aguillo et al., 2010), and using the QS ranking of universities in each country will help rank universities relative to only universities within the same country. The use of a ranking system ensures systematic data collection and since this ranking is independent of what I personally think of any university, conscious or unconscious bias is avoided. The choice of top 50 universities per country was motivated by a desire to have a reasonable number of programs[10].

The task of choosing a suitable academic program is informed primarily by the program's webpage (Hosein and Rao, 2021). Keeping this resourceful aspect of program webpages at the centre of focus, data collection was performed as an endeavour of extracting program details from associated webpages. While using program aggregator databases such as find-a-masters.com[11] was initially considered, the idea had to be dropped due to the limited ability of such databases to provide a detailed, valid curriculum of programs. The comparison of programs w.r.t their curriculums were made possible by the GDS framework (chapter 3). In addition to adopting the GDS framework, it was decided to only focus on the compulsory courses in a curriculum. This decision was taken based on the assumption that a

[8] Programs that focus on teaching subjects that make use of data
[9] https://www.topuniversities.com/university-rankings/usa-rankings/2020;
https://www.topuniversities.com/university-rankings/rankings-by-location/india/2020
[10] "Reasonable" here meant any number of programs greater than 50 to have a collection of data points larger than most past studies in DSE. Co-incidentally, this dataset has over 100 programs, making it the largest study in DSE until now.
[11] https://www.findamasters.com/

program's compulsory courses are more useful than its range of electives to understand what concepts the program designers consider "most important" for a graduate of the program.

Table 2. Data variables

Name of variable	Type of data
Institution Name	Categorical
Funding status	Binary (Private or Public)
Name of country	Binary (India or USA)
Program name	Textual
University ranking	Ordinal
Name of department offering program	Textual
Program description	Textual
Titles of compulsory courses	List of textual data
Descriptions of compulsory courses	List of textual data
Location (State, Country)	Categorical
Webpage HTML Structure[1]	Textual

[1]Includes HTML tags as well as text in them

Binary data: Can take on one of two possible values

Categorical: Can take on any value from a discrete set of possibilities

Ordinal: Can take on numerical values that are ordered

Numeric: Can take on numerical values that are not ordered

Textual: Can take on any kind of text data

The *program description* contains all textual content on the program webpage including the program's views on data science, its program and its goals. The *descriptions of compulsory courses* include details on what topics will be taught in the given course. It does not include data on assessment types or reading material. The data on compulsory courses were usually obtained from course directory webpages through links which were displayed in the main program webpage.

4.2 Stage 2: Data Collection

The planning stage was followed by data collection. The data for this research was collected using a 3-phase approach as depicted in figure 3.

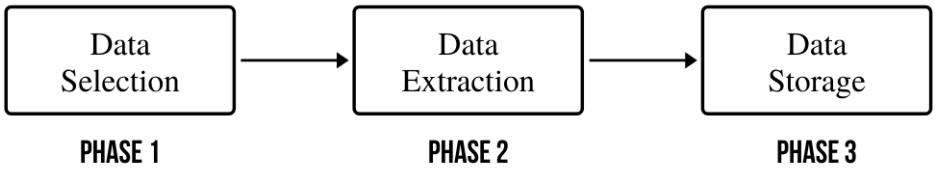


Figure 3. Data collection pipeline

4.2.1 Data Selection

Purposive sampling was used to filter the programs. Purposive sampling is a non-probability sampling technique that leverages the knowledge of the researcher when selecting units (Sharma, 2017). This kind of sampling attempts to *collect all data that is available* by using specific criterion for inclusion. It was selected over other alternatives to emphasize on the homogeneity of data points selected by choosing only those that fit into a specific profile (Palinkas et al., 2015).

Table 3. Filtering programs

Type of filtering	Filtering criteria	Implementation action
Primary	Title of program has at least one of the keywords	Search the university's graduate program directory[1]
Secondary	Offered with a full-time option	Examine program webpage
	Duration of at least 1 year	
	Predominant portion offered on campus[2]	
	Results in a single credit-based taught degree[3]	

[1]Case sensitivity is considered

[2]In normal circumstances, such as pre-pandemic times

[3]Excludes diplomas, integrated programs, dual degrees, certificate courses, executive programs

Table 3 represents the criterion for inclusion along with details on how they were implemented. Filtering was applied independently to both India and USA. The selected programs were stored in Excel sheets called Prospective Program Lists (PPLs) (Figure 4).

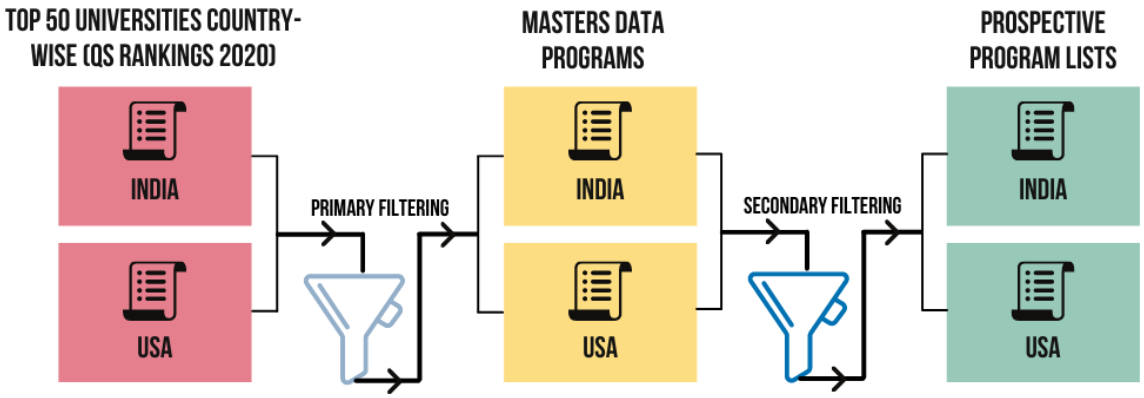


Figure 4. Creating the PPLs

PPL India had 34 qualified programs and PPL USA had 82 of them. The PPLs were raw collections of program names and corresponding webpage URLs (if they existed) and were used to inform further collection of data about each program.

4.2.2 Data Extraction

The PPLs were used as references to extract relevant data from each program using web scraping. Web scraping or web harvesting is the process of collecting data from the world wide web and storing it for later use (Zhao, 2017). There exist different web scraping techniques including manual extraction, use of regular expressions, use of APIs, HTML parsing and custom-built scraping software among others (Sirisuriya, 2015). The data here was extracted using both the manual extraction and HTML parsing variants. Manual extraction refers to the simple method of copy-pasting content from a webpage onto a file for storage. HTML parsing makes use of extracting content from a webpage based on its underlying HTML structure. A combination of these two were involved in order to make up for the individual limitations of each approach (Table 4, Figure 5).

Table 4. Web scraping approaches used

Approach	Strengths	Limitations	When was it used?
HTML Parsing	1. Low possibility of human-error 2. Fast process	1. Complex[1] 2. High initial time-investment present	1. Program webpage
Manual Extraction	1. Simple[2] 2. Low initial time investment	1. High possibility of human-error 2. Time-intensive process 3. Exhausting	1. Program webpage not present 2. Curriculum data collection

[1]Involves multiple steps such as inspecting webpage structure, writing code, evaluating code performance etc.

[2]Involves only copy-paste

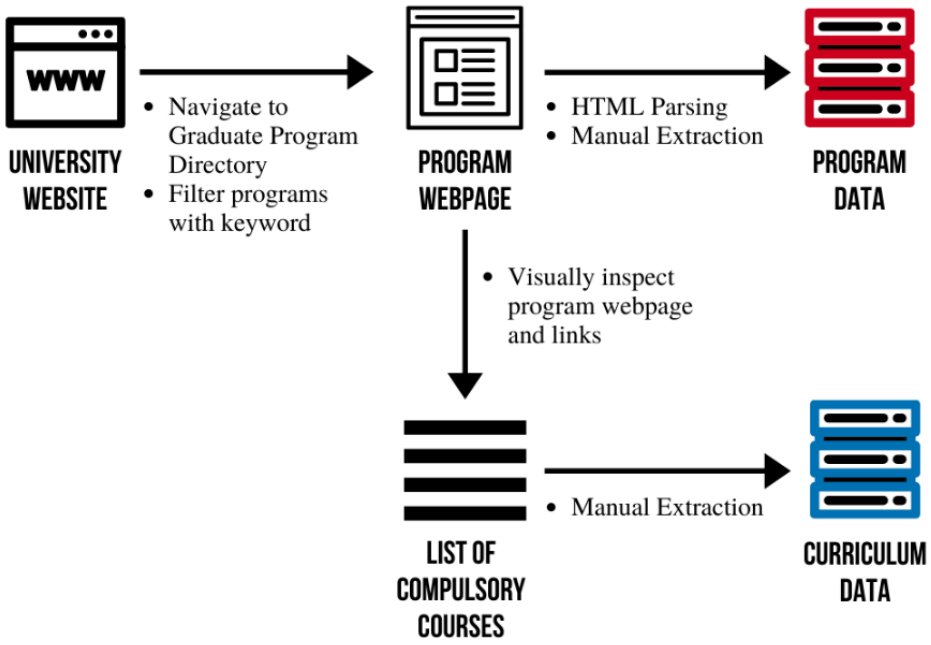


Figure 5. Web scraping to collect data

Figure 6 represents a high-level structure of most webpages. This structure was arrived at via observation and has been used to write a web scraper for data collection.

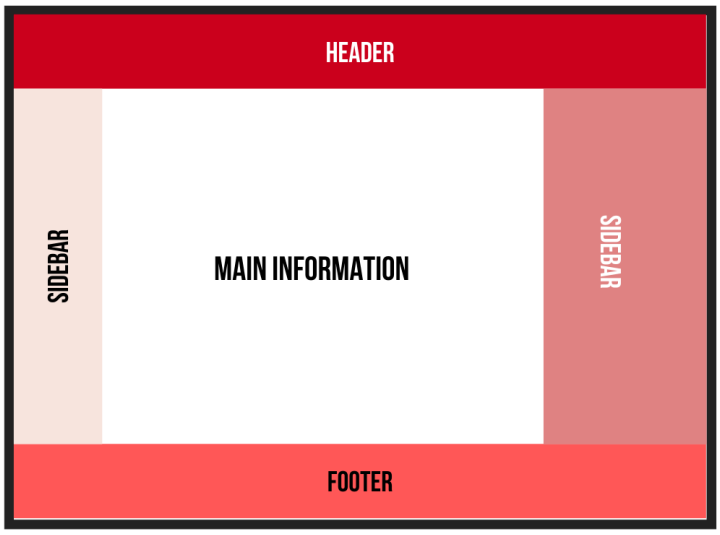


Figure 6. High-level structure of a program webpage

Web scraping with HTML parsing

HTML parsing was implemented via a scraping program written in Python using requests (Chandra and Varanasi, 2015) and beautifulsoup libraries (Richardson, 2007). Each webpage was visually inspected and the CSS selector[12] that helped identify the component that housed details of the program (Main Information as per figure 6) was retrieved. The URL of the webpage and the associated CSS selector for a given program were passed as inputs to the scraping program. The program extracted all textual content within the component identified by the CSS selector on the webpage. The textual content was later manually inspected to remove less useful information.

Web scraping with manual extraction

The UWS did have some obvious limitations owing to the high variance in the structure of the program webpages considered. Therefore, manual extraction was used to extract program descriptions. Manual extraction was also used to extract details of programs that did not have a webpage and instead used online PDF documents. All data regarding compulsory courses in the programs' curriculums were systematically collected via manual extraction. Manual extraction was also

[12] https://www.w3schools.com/css/css_selectors.asp

used to add features such as duration of programs, type of university and department involved among others.

4.2.3 Data Storage

Post data extraction, the datasets created were subject to sanity checks and minor changes were made to the data collected. The end products were 4 csv files – *Program data* and *Curriculum data* for each country. Program data contained data for each program, while Curriculum data contained data for each compulsory course. These final, processed datasets were the ones that were used for analysis in this project. Table 5 gives an overview of the final datasets. Appendix A contains names of variables, programs and universities included in the study.

Table 5. Summary of final datasets

Country	Type of data	Number of variables	Number of programs	Number of unique universities	Number of public & private universities
India	Program data	17	30	18	Public: 13
India	Curriculum data	4	389		Private: 5
USA	Program data	17	81	37	Public: 12
USA	Curriculum data	4	613		Private: 25

Both program datasets for each country were merged into Program data. Same was performed with Curriculum data.

4.3 Stage 3: Data Preparation

Data preparation is the act of transforming data into a form more suitable for a given task at hand (Zhang et al., 2003). In order to answer research question 1 regarding the content of the programs, the curriculum data was quantified using the GDS framework (chapter 3). To answer research question 2 regarding keywords in program descriptions, variables in the program data subject to textual pre-processing.

4.3.1 Quantifying Curriculum Data

Curriculum data included titles and descriptions of compulsory courses. These were quantified using the GDS framework via 2 steps. First, all the course titles and descriptions in all 111 programs collected were used to create the GDS dictionary and then this dictionary was used to score all 111 programs. The creation of the GDS dictionary was aided with the help of a web-based interface custom-made for this task (Appendix B).

Creating the GDS Dictionary

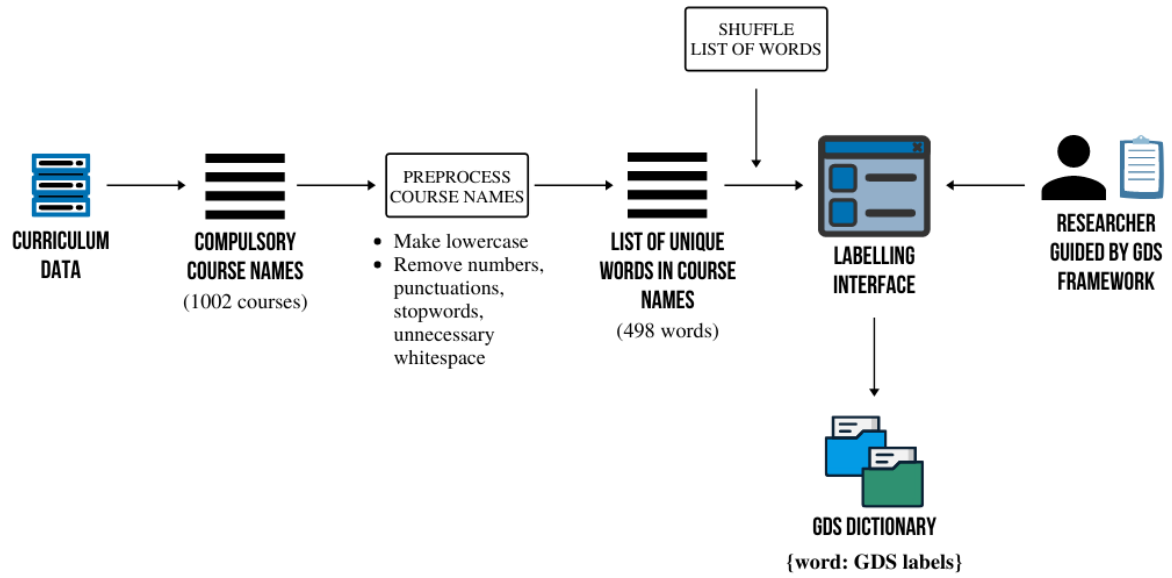


Figure 7. Creating the GDS dictionary

Figure 7 describes the GDS dictionary creation process. The curriculum data had in total 1002 course titles across all 111 programs. These course titles were processed to extract only the unique words that formed these titles, leading to 498 such words. These words were then shuffled to mix up the order of their occurrence to break any order of occurrence of these words[13]. The web-interface for labelling was used to label each of the 498 words. Each word could be labelled with one or more GDS divisions, where GDS divisions are the 6 divisions described in chapter 3. An important question the reader is expected to have is the rationale behind the decision

[13] This was performed to make the word-labelling process fair by not biasing labelling of a word based on the previous word

to label individual words rather than the courses themselves. There were two reasons for this decision:

1. To set in motion the creation of a GDS dictionary that will help objectively evaluate data science programs
2. Words in course names are far more likely to be repeated in future course names than the entire course names themselves – This makes the labelling process easier, and more robust to future updates

The interface was used to label each of the 498 words into one or more GDS divisions based on consulting the GDS framework and the details of the courses that contained this word in the title.

Scoring programs using the GDS dictionary

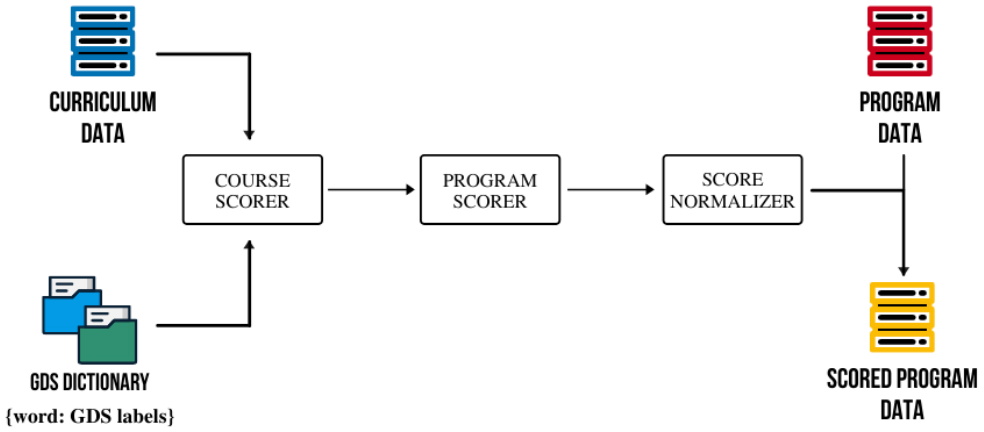


Figure 8. Scoring new programs

Once the dictionary was created, it was used to score all 111 programs. Figure 8 describes the scoring process. The `Course Scorer` block performed the task of scoring each of the 1002 compulsory courses. The `Program Scorer` aggregated scores of all 1002 courses by grouping them together based on the program that offered the courses. A course received a score of 1 for a given GDS division D where D ϵ {1,2,3,4,5,6}; if at least one of the words in the course's title has that GDS division as a label as per the GDS dictionary. Else, the course received a score of 0 for that GDS division. If a course had a 0 for all divisions, it was given a score of 1 for the "Not determinable" variable. *So, every course was scored across 7 variables (GDS divisions + Not determinable)*. Since each course was identified along with both the

course title as well as the name of the program offering it, duplicated course names did not skew the results.

A program P was scored on GDS division D with the sum of the scores its constituent compulsory courses C_1, C_2, …, C_n received for a division D. Thus, program P was also scored across 7 variables. The "Not determinable" variable was 0 for all programs scored as a program just needed the existence of one of the labelled words in at least one compulsory course's title to have a determinable score for a given division. The 6 scores per division for each program were then normalized to sum up to 1 by the `Score Normalizer` and describe each division in terms of how much of the given program's content it covered, in terms of percentages.

As per Donoho (2017), the full scope of data science is realized only when a program focuses its teaching on all 6 divisions mentioned in the framework, though the ideal proportion of each GDS division in a program is not mentioned. I assume that the most balanced data program would be one in which the proportion of all 6 GDS divisions are equal[14]. If such a program existed, the standard deviation of its 6 GDS division scores would be 0[15]. Based on this idea, it is easy to reason that the standard deviation of the 6 scores for a given program is representative of how "far-away" it is from the "ideal" program (given the "ideal" program has equal coverage of all 6 GDS division areas). Thus a GDS score is computed which is the standard deviation of the proportions of GDS divisions for a given program.

$$GDS\ Score\ for\ program\ \ P = \sum_{i=1}^{6} D_i \qquad \ldots 1$$

$$,where\ D_i\ is\ the\ proportion\ of\ GDS\ division\ i\ that\ is\ covered\ by\ program\ P$$

While the standard deviation of the GDS scores of a program has on its own, limited applicability, it does however have better utility if used as a metric of comparison between programs. A program with a lesser value for this quantity will be more "balanced" when compared to one with a higher value. More balanced, as it is closer to the ideal program that covers all divisions of data science equally.

[14] The assumption has been made to aid in the explanation of the concept here, and thus is not based on facts
[15] The standard deviation of a list of equal numbers is always 0

4.3.2 Pre-processing Program Data

Features in program data such as name of department and program title had high cardinality (large number of unique values) due to slight differences in how universities chose to name their apartments and programs. In an attempt to bring about uniformity, they were processed to extract new features, Type of department and Subject area of program[16]. Type of department was assigned to each program as follows:

1. Data-related: Presence of 'data', 'artificial intelligence' or 'machine learning' in department name
2. Engineering-related: Presence of 'computer', 'engineering' or 'info' in department name
3. Management-related: Presence of 'management', 'business' or 'commerce' in department name
4. Social-related: Presence of 'social', 'socio', 'public' or 'politics' in department name
5. Statistics-related: Presence of 'statistics or 'math' in department name
6. Other: Everything else

Subject area could assume any value or combination of values from Analytics, Artificial Intelligence, Business Analytics, Data Science, Machine Learning, Statistics and Miscellaneous and was assigned by observing program names.

Textual pre-processing of program descriptions

The program descriptions obtained from the program webpages were converted to a form similar to a document-term matrix (Anandarajan et al., 2019) with the help of the spaCy[17] "en-core-web-md"[18] pre-trained nlp pipeline[19]. The pipeline, also called processing pipeline is a feature of spaCy that facilitates the step-wise processing of text input. The pipeline segments text into tokens; assigns part-of-speech tags,

[16] Each of these features could be assigned multiple categories
[17] Natural Language Processing library in python
[18] https://spacy.io/models/en#en_core_web_md
[19] https://spacy.io/usage/processing-pipelines#pipelines

dependency labels; labels named entities and performs lemmatization. The "textrank" component was added to the spaCy pipeline to extract keywords from programs descriptions[20].

4.4 Stage 4: Data Analysis

In the pursuit of answering the research questions, 2 main types of analytical methods were used – Statistical analyses (RQ1) and Keyword extraction for research (RQ2).

4.4.1 Statistical Analyses to compare programs on the basis of GDS

To answer the research questions regarding the content of the programs and their differences across countries, departments etc., inferential statistics were. The R computing environment (R, 2013) and rstatix package (Kassambara, 2020) were used for performing statistical analyses. The 7 numerical features relating to GDS scoring of programs were compared across relevant categorical features. The aim of such a comparison was to identify if there were differences between groups within each of these categories.

The Shapiro's test for normality (Ghasemi and Zahediasl, 2012) revealed that none of the numeric features under consideration were normally distributed. Since the assumption of normality does not hold, non-parametric tests were performed to compare groups within each category (Nayak and Hazra, 2011). This comparison is most easily understood as an attempt to reject or fail to reject the statement that two (or more) groups of numerical data are obtained from the same population (Nayak and Hazra, 2011). This statement is called the **null hypothesis**. If the null is not rejected, then it means that no reasonable difference exists between the groups and hence they are not to be considered as separate entities. If the null is rejected, then the groups are said to be *statistically significantly different*. A summary of the tests performed here are depicted in table 6.

[20] https://github.com/DerwenAI/pytextrank

Table 6. Summary of statistical tests

Categorical Variable	Groups[1] and number of programs included in each[2]	Total programs covered	Statistical test used to find differences[3]
Country	India (30) USA (81)	111	Wilcoxon rank sum test
Department	Data-related (13) Engineering-related (37) Management-related (26)	76	Kruskal-Wallis H test
Type of university	Private (71) Public (40)	111	Wilcoxon rank sum test
Subject offered	Analytics (24) Artificial Intelligence (13) Business Analytics (25) Data Science (36)	98	Kruskal-Wallis H test

[1] All groups in each category were independent as the existence of one was not related to the existence of another

[2] Only groups that were considered for the final analyses are included

[3] Tests used were different based on number of groups available

The Wilcoxon's Rank Sum Test (N=2)

The Wilcoxon's rank sum test is the non-parametric equivalent of the 2-sample t-test of means. Unlike the t-test, the rank sum test's null hypothesis is *equality of medians*. This makes the rank sum test robust to outliers, a useful characteristic in the testing of GDS score differences (some programs have far higher values than others). Also, the number of observations in groups are not required to be equal. The Wilcoxon's rank sum test for two groups A and B with n_A and n_B observations of a numeric variable V respectively, is conducted as depicted in algorithm 1.

Algorithm 1: Step-by-step Wilcoxon's rank sum test

1. **Assign ranks to observations**
   Order all $n_A + n_B$ observations in ascending order, assigning a rank to each. If there is a tie, assign mean rank of tied observations to tied observations
2. **Compute w_A and w_B**
   $wA \leftarrow Sum\ of\ ranks\ for\ all\ observations\ in\ A$
   $wB \leftarrow Sum\ of\ ranks\ for\ all\ observations\ in\ B$
3. **Fix hypotheses and level of significance**
   $H0\ (null): median(A) = median(B)$ or "Both populations are the same"
   $HA\ (alternate): median(A) \neq median(B)$ or "Both populations are not the same"

α (level of significance): 0.05 in this research

4. **Compute test statistic**

$$Z = \frac{w_A - \mu_{w_A}}{\sigma_{w_A}}$$

where,

$$\mu_{w_A} = \frac{n_A\,(n_A + n_B + 1)}{2} \; and \; \sigma_{w_A} = \frac{\sqrt{n_A \,.n_B \;(n_A + n_B + 1)}}{12}$$

5. **Compute $p - value$ from standard normal distribution for the test statistic**
6. **Interpret**

If $p - value < \alpha$, "Reject the Null"

Else "Fail to reject the null"

Source: (Montgomery and Runger, 2010)

The Kruskal-Wallis H test (N>2)

In categories that had more than 2 groups, the Kruskal-Wallis H test was used. It is the non-parametric equivalent of the ANOVA and is used to compare a numeric variable across a categorical variable where number of categories > 2 (McDonald, 2009). It does not require the numeric variable to be normal. The Kruskal-Wallis makes use of similar terminology as the Wilcoxon rank sum test. If there are k groups to be compared such that n_1, n_2, …, n_k are the number of observations of numeric variable V in each group, the test statistic H for the Kruskal-Wallis H test is

$$H = \left[\frac{12}{n_T(n_T+1)} \sum_{i=1}^{k} \frac{{R_i}^2}{n_i}\right] - 3(n_T + 1) \qquad \text{… 2}$$

$k \; = \; number \; of \; groups$

$n_i \; = \; number \; of \; observations \; in \; group \; i$

$$n_T = \sum_{i=1}^{k} n_i$$

$R_i = sum \; of \; ranks \; for \; sample \; i$

The $p - value$ for a given H-statistic is computed using a chi-square distribution with (k-1) degrees of freedom. If the p-value is less than the chosen α (0.05 in this case), the null hypothesis that all groups are from the same population is rejected. If the null was rejected, then Dunn's post-hoc test with Bonferroni correction was performed to find the groups between which differences exist (Dinno, 2015).

4.4.2 Keyword Analysis of Program Descriptions

Keyword extraction is the process of selecting terms from a text that most accurately help describe the text's contents (Rose et al., 2010). This helps lay out the essence of a text by using fewer terms. The TextRank algorithm is a popular choice for keyword extraction and is built on the premise of Google's PageRank (Mihalcea and Tarau, 2004). The algorithm ranks words in a text on the basis of how "reputed" each word is. This reputation is based on how many *co-occurrence links* a word forms with its neighbouring N words (N is the window size), across iterations over the whole piece of text. Reputation of a word W is bolstered by the number of links it has with other words and also the reputation of those linked words themselves. The words are then ranked in the descending order of these reputation scores and the top L/3 of these are chosen as keywords, where L is the total length in words of the text. If there are keywords that are adjacent to each other both in the original text as well as in the list of keywords, they are merged to form multi-word key phrases. The pytextrank module (Nathan, 2016) which is an implementation of TextRank in python was used to extract keywords across all program descriptions. The keywords were then visualized using wordclouds. Wordclouds are useful text visualizations that help summarize texts and unearth patterns within text using colour-coding (Heimerl et al., 2014).

4.5 Stage 5: Building the Web Tool

A web-based tool named Data Program Selector (DPS) was developed using python and streamlit to perform the functions of scoring a new program with the GDS dictionary and also recommending suitable programs for a learner from the existing database of 111 programs based on the learner's preferences.

The user interacts with the tool which has 3 pages, namely, the home page, score a program page and find a program page (figure 9).

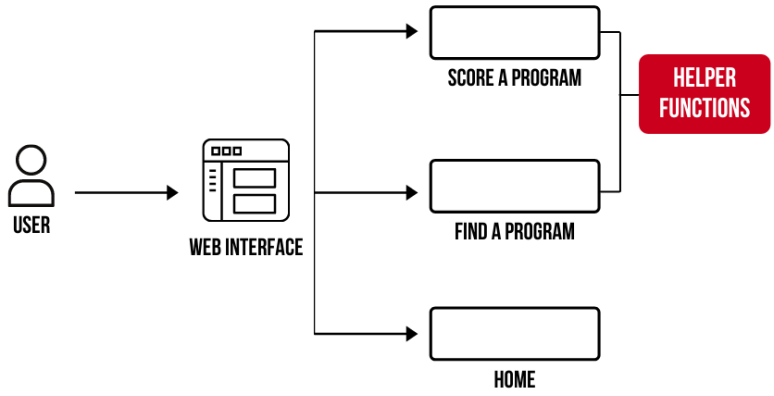


Figure 9. High-level architecture of the tool

The scoring and finding programs functionality are performed respectively by two specific components:

1. The *scorer* which scores a new program using the GDS dictionary
2. The *seeker* which performs two functions:
    1. Filter programs based on user's preferences (Basic and advanced filters are used)
    2. Recommend other programs from the database that are similar to the top filtered program

The basic filter of the seeker was designed to help a user filter the 111 programs on the basis of both the details of a program such as country, university type, subject offered and department offering the program. The advanced filter is a novel concept and is discussed below.

Novel Data Program Recommender[21]

As evidenced in Relevant Literature (chapter 2), there have been multiple views on what constitutes the ideal data program. In spite of the assumption than an ideal program will cover equally, all 6 divisions of GDS in section 4.3.1, I propose that we accept that *an ideal data program can hardly exist as long as attempts to find one are guided by the lens of the program designers*. An ideal program is specific to every individual and depends on what they wish to study. For example, a learner who is experienced in data storytelling but oblivious about data acquisition techniques might

[21] A part of the Data Program Selector

prefer that he or she learn more of GDS 1 than GDS 4 in their program of choice. This could be slightly or wildly different for another learner. Embracing these differences, the DPS allows users to filter programs based on the proportions of each GDS division covered by a program. Algorithm 2 explains this filter's working.

Algorithm 2: Finding a suitable program based on user's interests

1. **User enters priorities of GDS divisions. Example: 4,5,6,1,2,3**
   $$priorities \leftarrow from\ user\ input$$
2. **Seeker converts this string of priorities into a vector**
   This vector is called *gds_weights_user*
   The first number in the priority string is given a weight of 64, the next 32 … and the last is given 2
3. **Seeker extracts proportions of GDS divisions for all 111 programs as a matrix**
   This matrix is called *gds_proportions*
4. **Compute dot product to find a weighted GDS score for all programs**
   $$weighted\_gds\_score = gds\_proportions \,.\, gds\_weights\_user$$
5. **The program with the highest *weighted_gds_score* is presented to the user**

The user can now ask for recommendations similar in content to the program displayed by the above filter. This leads to the recommender using a content-based recommendation approach (Pazzani and Billsus, 2007) to select the top 5 programs from the database similar to the filtered program. The similarity is computed on the program descriptions as well as the 6 GDS division proportions using cosine similarity (Lahitani et al., 2016). The recommended programs belong to the country of preference as selected by a user in the basic filter, but can belong to any subject, department or university type irrespective of the user's choice. This was a deliberate design choice to help the user discover different kinds of programs that are similar to what he/she wants to learn, irrespective of administrative differences. The DPS tool was deployed onto the web using Streamlit sharing[22].

Appendix C contains the link to all code and the deployed DPS tool.

[22] https://streamlit.io/sharing

Chapter 5 – Results

5.1 Content of Masters' data-programs in India and USA

Of the 6 GDS divisions[23], the most covered across all 111 programs was GDS 3 – Computing with Data (mean: 34.4%, median: 33.3%) and the division with the least coverage was GDS 6 – Science about Data Science (mean: 1.6%, median: 0%) (figure 10).

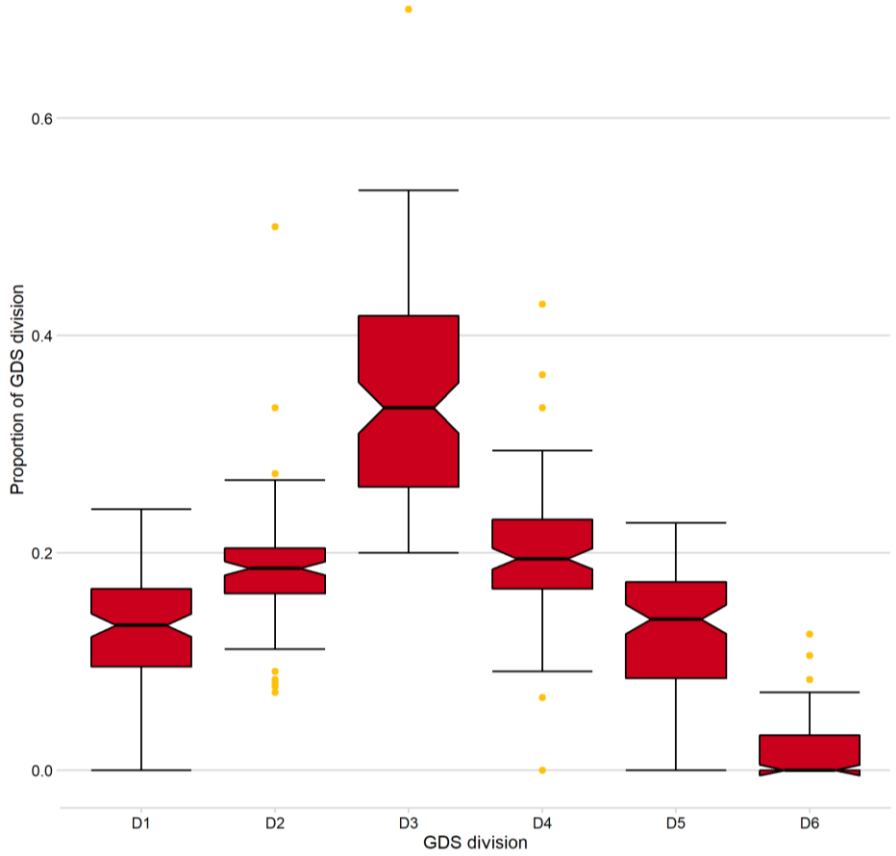
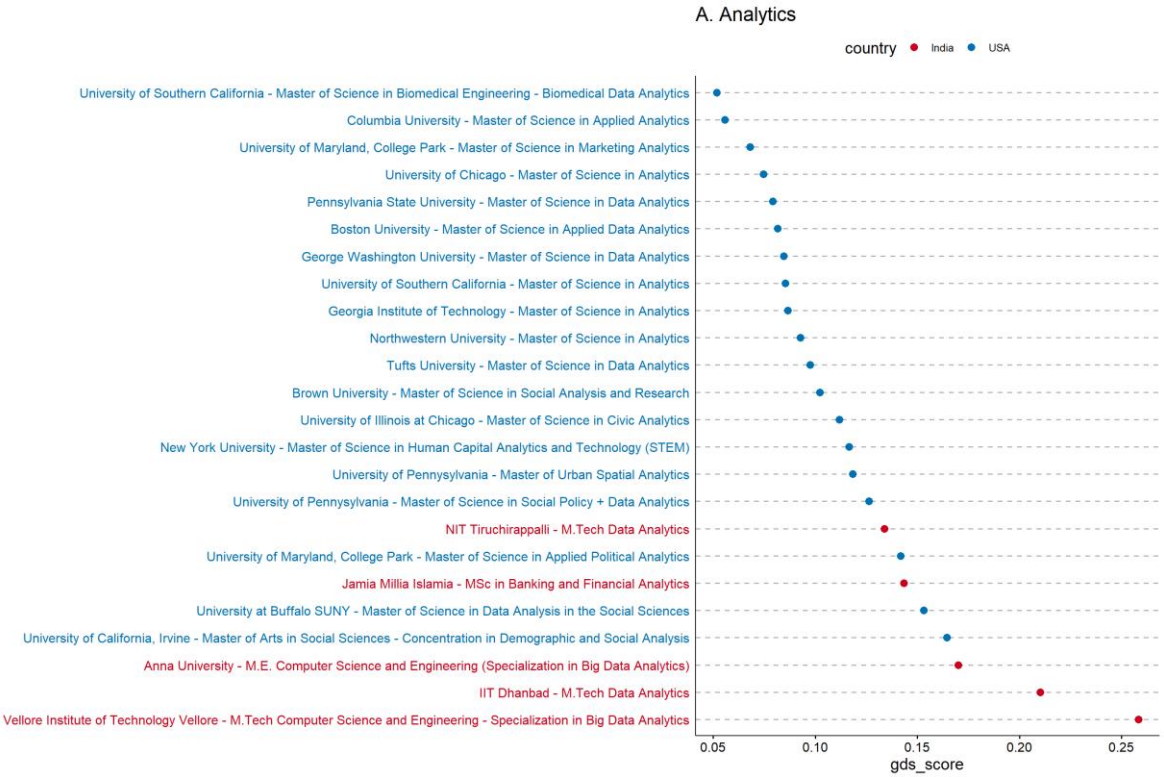


Figure 10. Distribution of proportions of GDS divisions.

The representation of GDS 6 was far lower than any of the other divisions and this indicates how the representation of topics such as reproducibility, meta-analyses, data quality, ethics, data privacy and all other components of "true science" (Donoho, 2017) are underwhelming in the collected data programs.

[23] From here onwards, every division GDS i is abbreviated as D i in the figures; where $i \in \{1,2,3,4,5,6\}$

Note: The next two pages together contain figure 11.



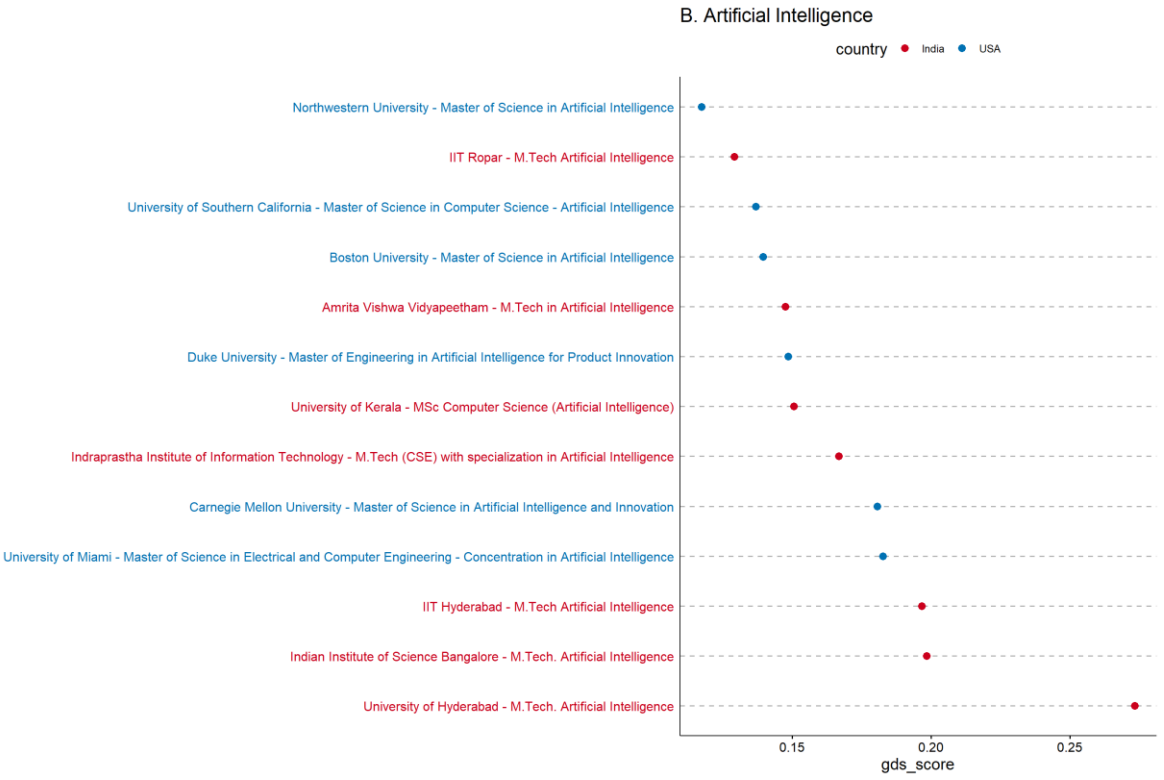


Figure 11. GDS scores of all 111 programs in the study. *Note: As per assumptions in section 4.3.1 (page 25), lower GDS score indicates more "balanced" program*

Overall, programs in USA tend to be more balanced than the ones in India across all 4 subject areas (figure 11). Of all 4 subject areas, artificial intelligence has programs that are generally more imbalanced than those in other subjects. This is in line with

what we would expect as AI programs tend to be more focused in technology than the other subjects presented here.

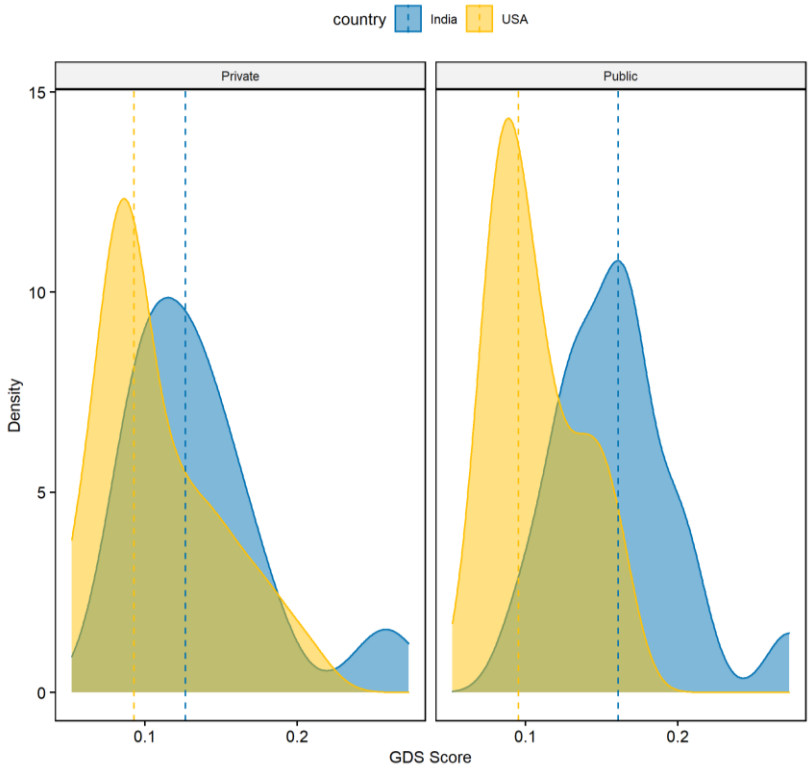


Figure 12. Distribution of GDS scores

GDS scores of programs in the USA are generally lower than those in India (figure 12). An important addition to this is that among private universities, the overlap between Indian and American programs is almost twice their overlap among public universities.

The association between a university's national ranking and the proportions of GDS divisions covered by its programs are different depending on the division considered as well as the country considered (figure 13). In India, universities with better ranks concentrate less on Data Gathering, Preparation and Exploration while in the USA, better ranked universities seem to focus less on Data Representation and Transformation[24].

[24] These are not statistically significant results, but the patterns are visually discernible

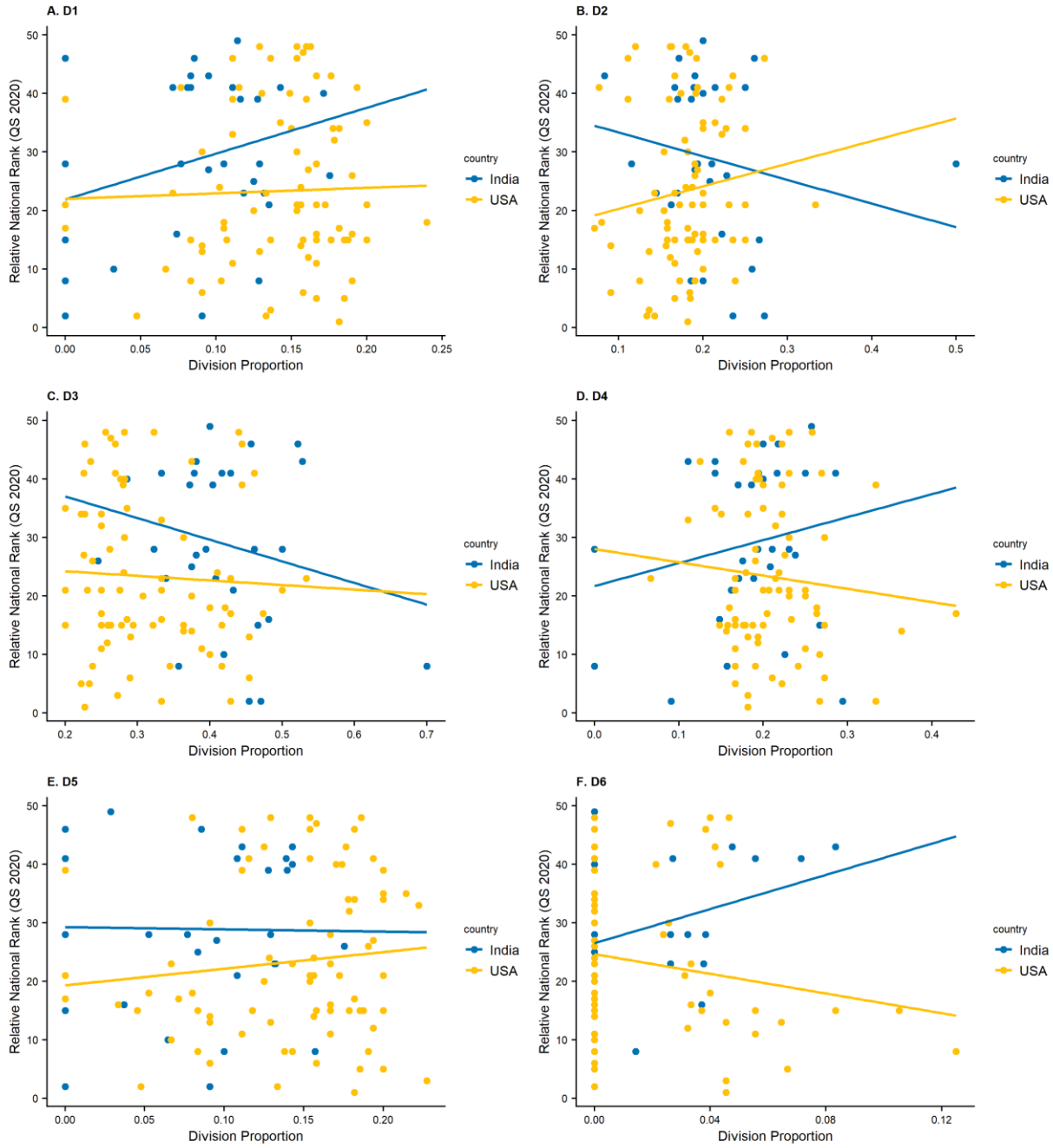


Figure 13. Rank of host university vs GDS division proportion

An interesting result however is the treatment of Science about Data Science in both countries. While in the USA, better ranked universities tend to cover more of this division than the worse ranked ones, in India, it's quite the opposite.

5.1.1 Are there differences in the contents of programs between India and the USA?

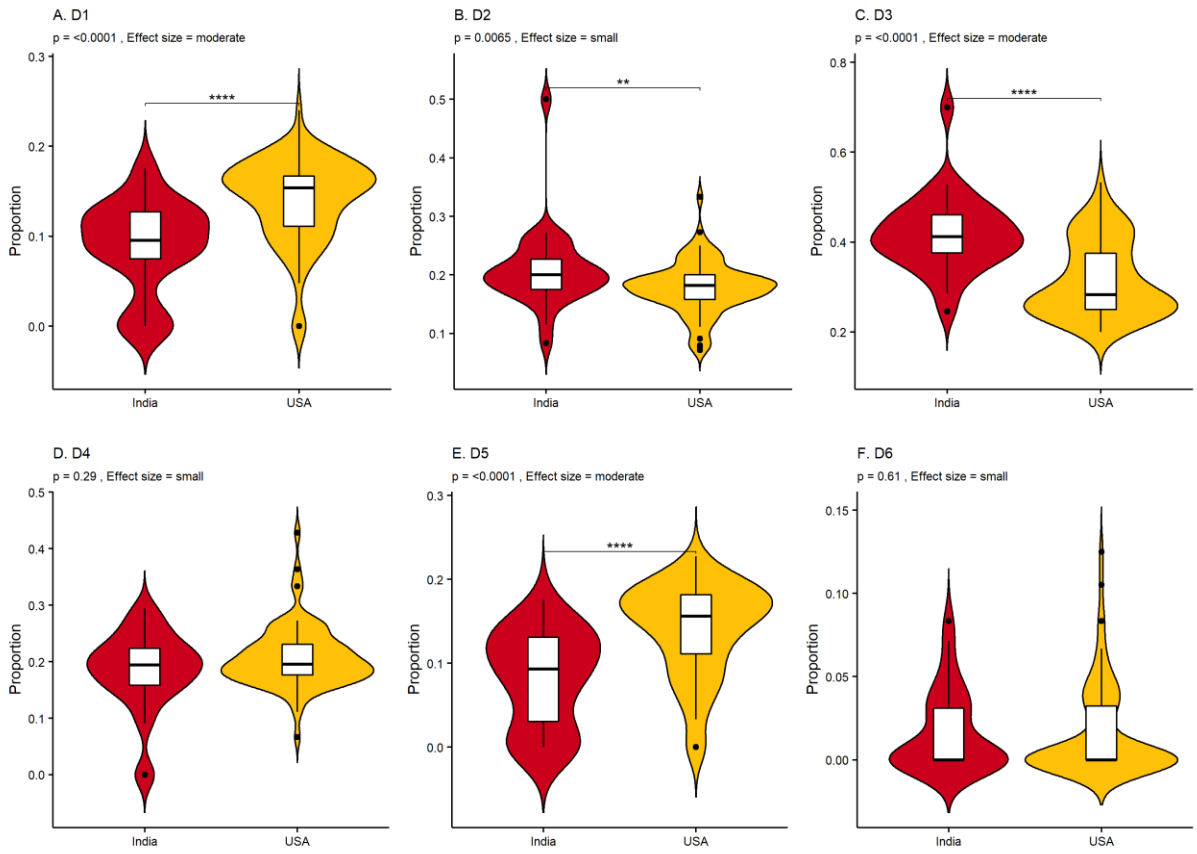


Figure 14. GDS in India and USA (Wilcoxon rank sum test). *Note: The presence of stars (*) indicates statistical significance*

The bigger picture reveals that programs in USA were more balanced than those in India. However, a division-wise comparison reveals that significant differences between the two countries existed in only 4 of the 6 divisions, with only the differences in Data Gathering, Preparation and Exploration; Computing with Data and Data Visualization and Presentation being not small in terms of effect-size[25]. Of these 3 divisions, Indian programs had grater proportions of their coursework in Computing with Data than the American ones. This pattern was reversed in the case of the other two divisions. *Thus, there are differences in the contents of programs in India and the USA.*

[25] Emphasized to make findings robust and less misleading than reporting only the p-values

5.1.2 Are there differences in the contents of programs between public-funded and private-funded institutions?

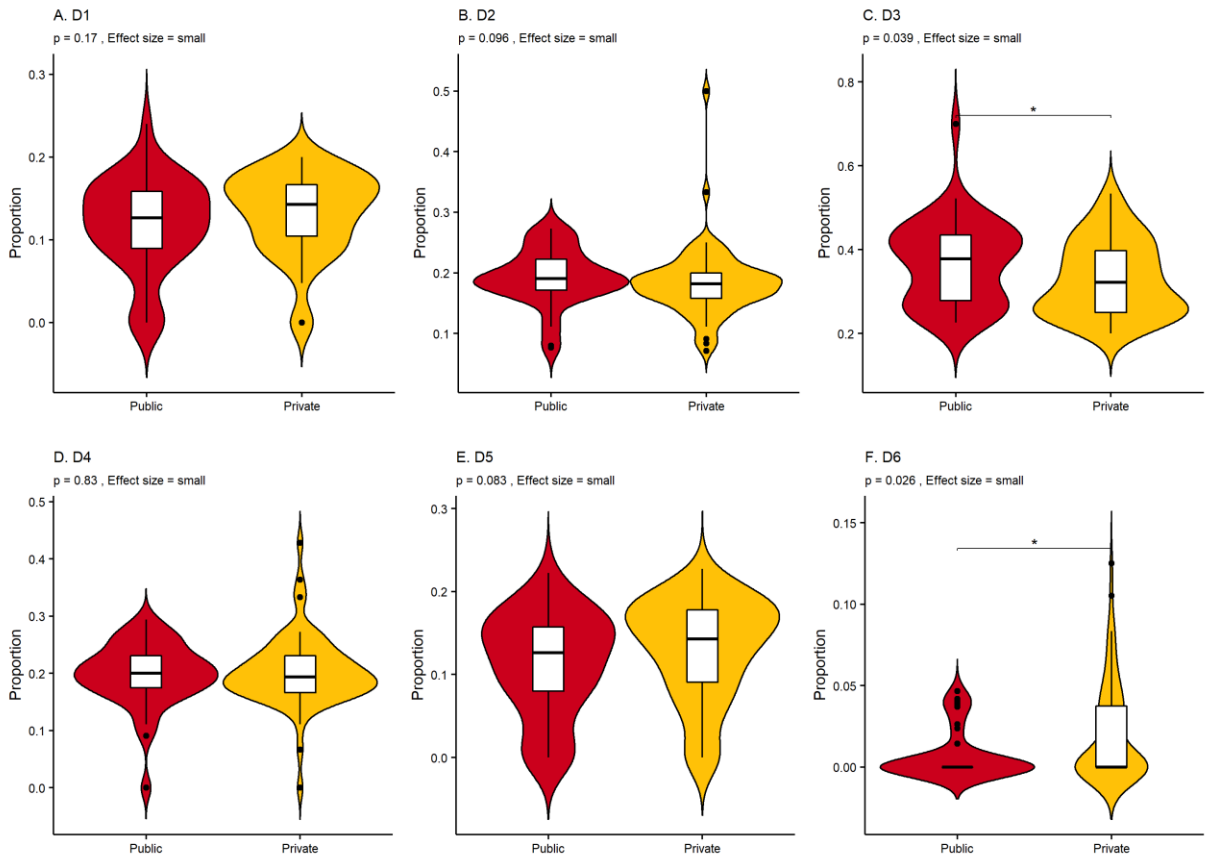


Figure 15. GDS in Public and Private universities (Wilcoxon rank sum test). *Note: The presence of stars (*) indicates statistical significance*

While there exist some divisions where there is a statistically significant difference between private and public universities, the associated effect sizes are small (figure 15). Therefore, there is very little difference in the contents of programs on the basis of whether a privately-funded institution or a publicly-funded institution is providing it.

A comparison of GDS division proportions between private and public universities in either country produced mostly uninspiring results except where there was a strong significant result differentiating public and private universities in India in proportions of Science about Data Science (p-value<0.0001, effect-size: large). As per the result, private universities in India offered programs that covered topics in Science about Data Science with greater focus than public universities.

5.1.3 Are there differences in the completeness of programs based on differences in departments offering the program?

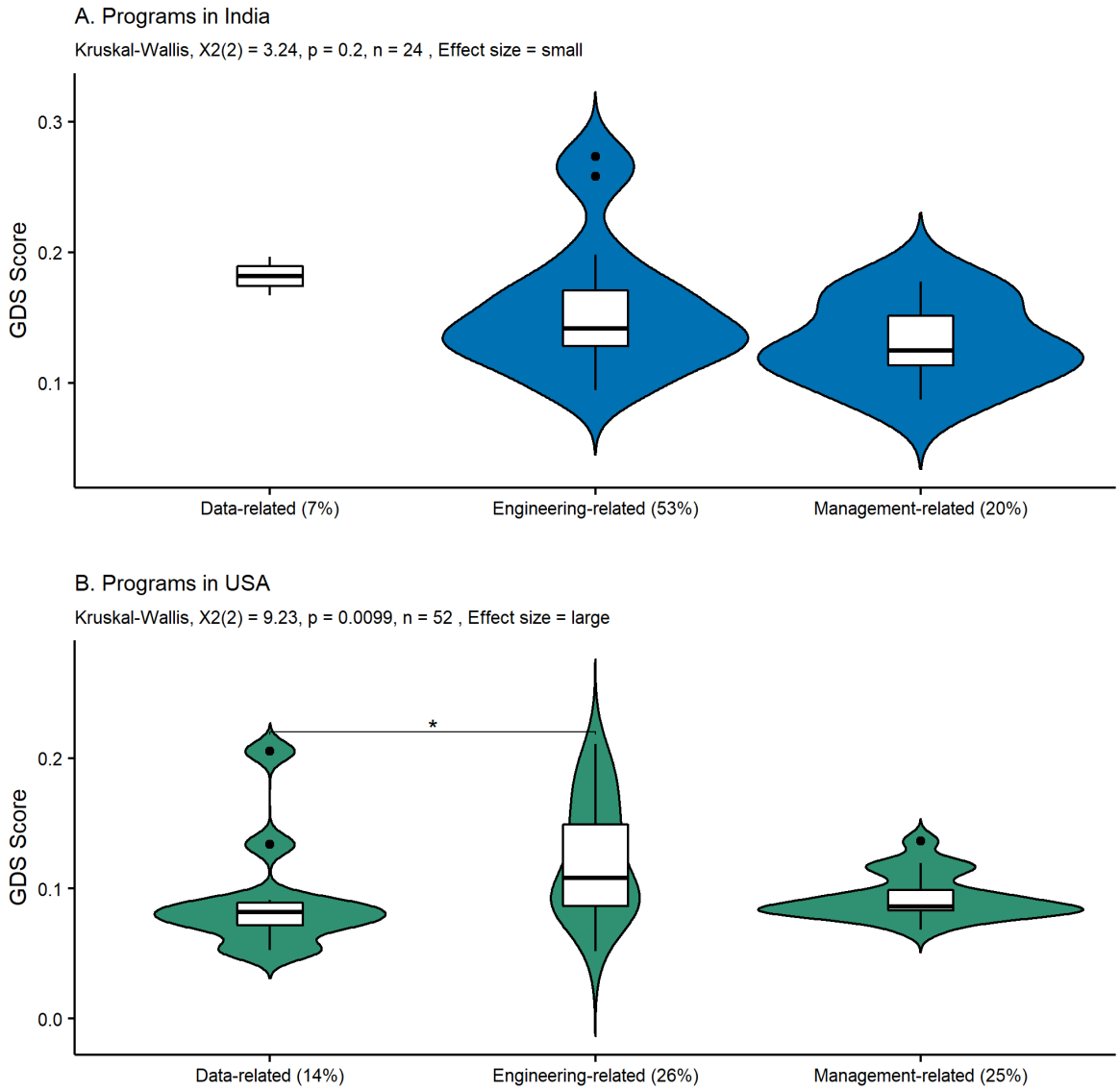


Figure 16. GDS scores across type of departments offering the program. *Note: In parentheses, the proportion of programs per department type of the concerned sample is shown*

In the sample collected, 76 programs were offered by departments related to the disciplines of engineering (37), management (26) and data (13)[26]. In India, of 24 such programs, 53% (16) were offered by engineering-related departments alone. In the USA, no type of department held such monopoly. There were no significant differences between the GDS scores of each department type in India while such a

[26] Only these were considered to avoid programs that were born of collaboration between departments from different disciplines

difference did exist between programs in the USA (p-value:0.0099, effect-size: large) (figure 16). Therefore, while there is not enough evidence of programs offered by one department being more complete than another in India, it is evident that programs offered by data-related departments are more complete than others in the USA (on account of a lower GDS score).

5.1.4 Are there differences in the completeness of programs based on the main subject offered by the program?

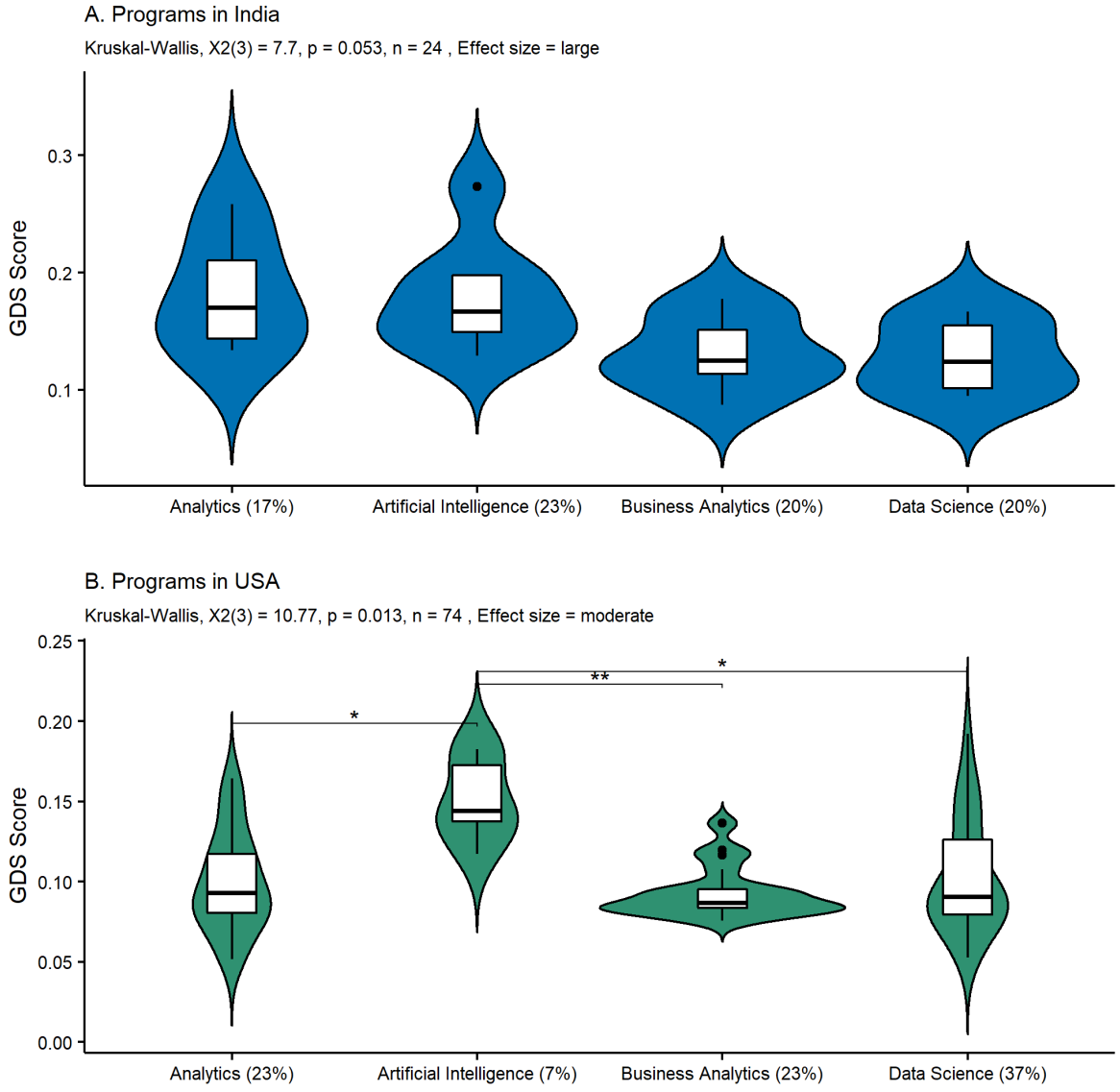


Figure 17. GDS scores across subjects offered by the programs. *Note: In parentheses, the proportion of programs offering the subject type is shown*

Data Science (36 programs), Business Analytics (25 programs), Analytics (24 programs) and Artificial Intelligence (13 programs) were the 4 main types of subjects that were discovered by analysing all program titles in the collected sample. In the USA, a large majority of programs in this sample teach Data Science while in India, such a majority subject is not evident. Figure 17 depicts the differences in GDS scores for programs across these subjects in either country. While the differences are not statistically significant in India (most likely due to the small number of programs in each subject), the differences are more pronounced in the USA. Programs offering Business Analytics look well balanced in the USA than other subjects[27]. Programs focusing on AI have higher GDS scores in general indicating less balance. However, since the number of programs offering AI in the USA is less than 10, such a finding is to be taken with a pinch of salt.

5.2 Content of program webpages of these Masters' data-programs

Of the 30 programs from India in the sample, 11 did not have dedicated web pages while all 81 programs from the USA all had dedicated web pages. The programs in India that did not have web pages had their details presented via PDF documents hosted on the university website. 3 of these did not provide any kind of program description or overview, although they did have information on the curriculum.

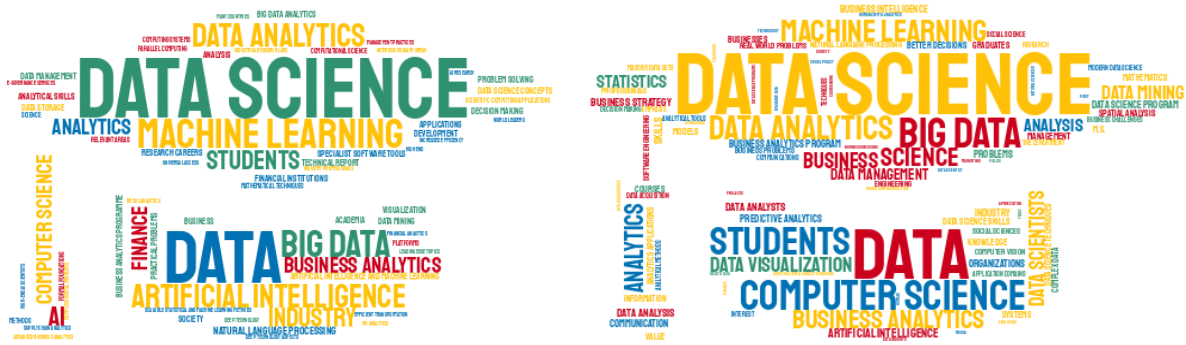


Figure 18. Keyword wordclouds of program descriptions in a) India b) USA (left-to-right)

[27] Shorter violin plot indicating less variance in the whole category

Overall, the main keywords used to describe programs include Data, Data Science, Machine Learning, Data Analytics, Big Data, Computer Science, Analytics, Students etc. (figure 18).

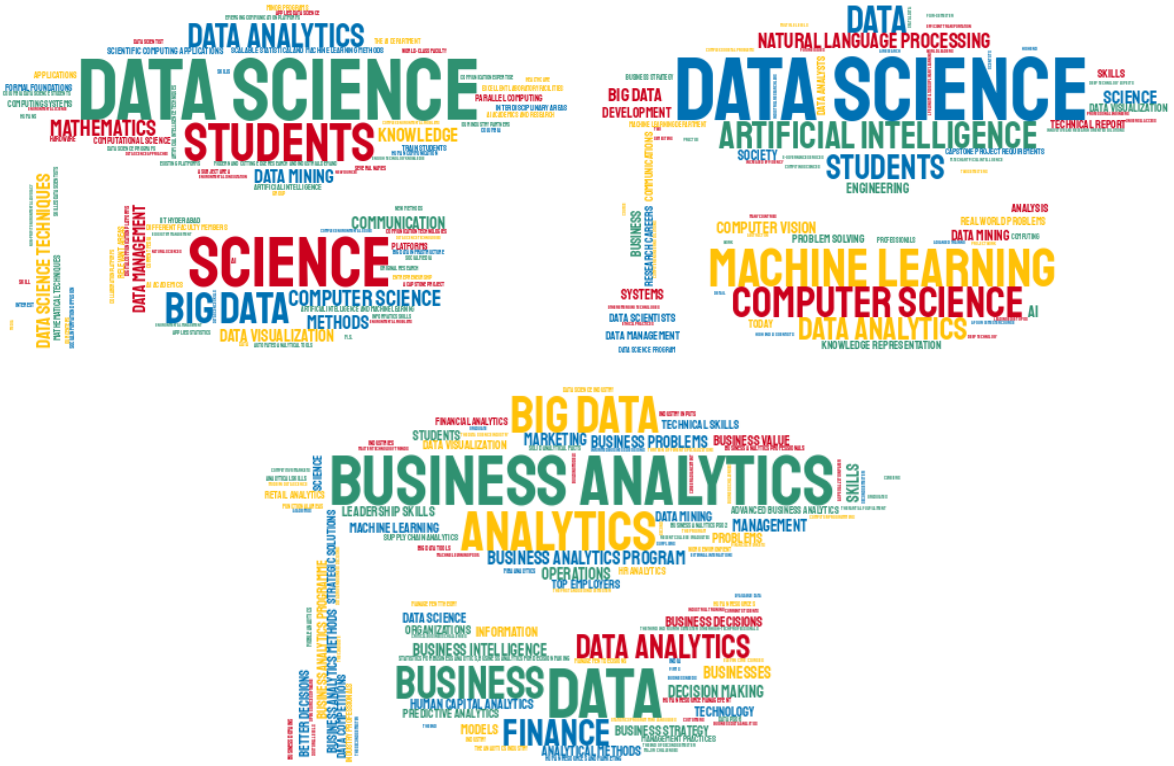


Figure 19. Keyword wordclouds of program descriptions in programs offered by a) Data-related b) Engineering-related and c) Management-related departments (clockwise)

As the size of a word is commensurate to its frequency in a given category, it is evident that based on the department offering the program, there are some subtle differences in keywords used. In engineering-related departments, main keywords used include Artificial Intelligences, Machine Learning, Computer Science, Natural Language Processing etc. (apart from those like Data that are common to other departments also). Most Business analytics programs are offered by management-related departments and this is evidenced from figure 19.c.

5.3 Using the Data Program Selector to choose a suitable Masters' data-program

This section is an overview of the DPS tool that was built. Each screenshot is presented as an attempt to take the reader through a visual demo of the prototype.

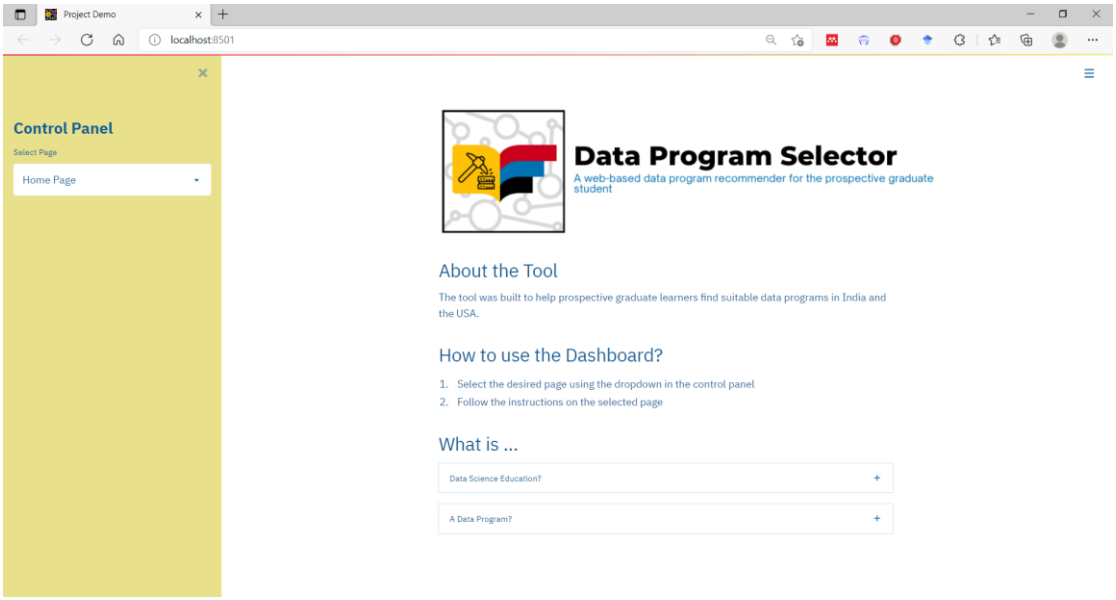


Figure 20. The home page

The Control Panel on the left-hand side is used to select the desired page.

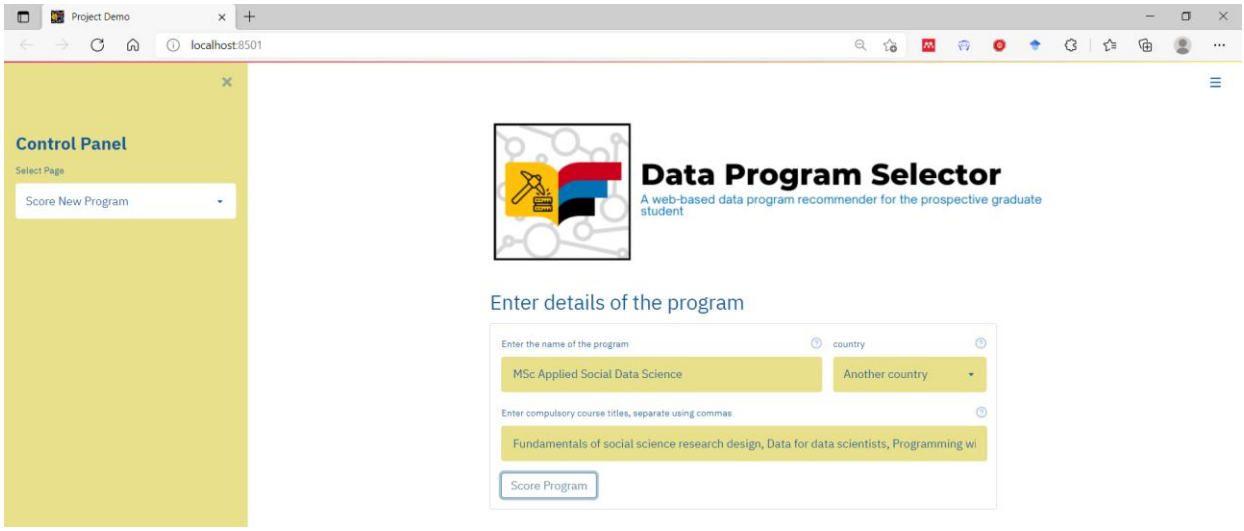


Figure 21. Scoring a new program – Entering the input

As input, the compulsory course titles have to be entered along with a name of the program and the country.

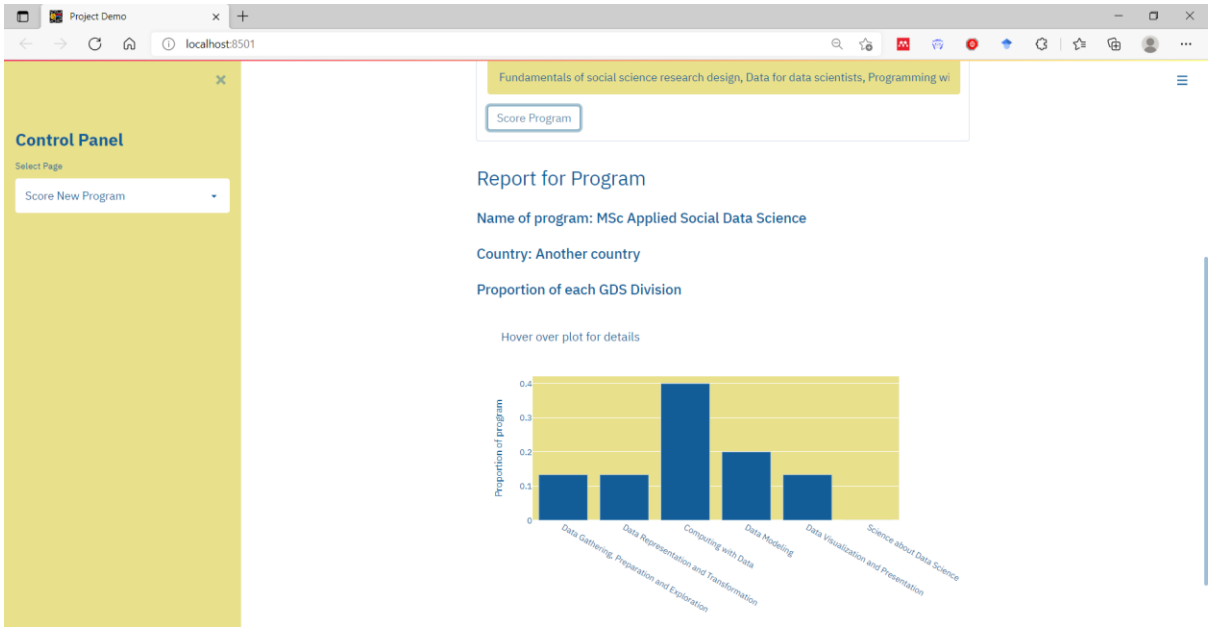


Figure 22. Scoring a new program – The output

A basic report of the GDS division proportions in the course is displayed. This internally makes use of the GDS dictionary created (section 4.3.1).

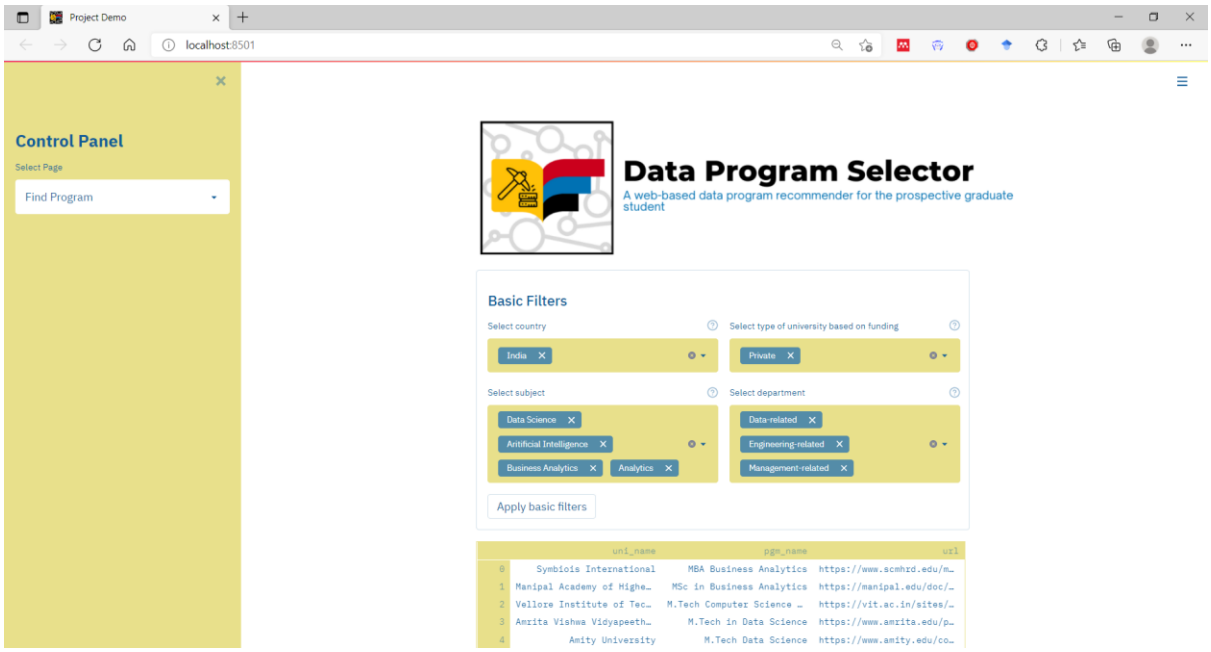


Figure 23. Find programs – Applying basic filters

Applying the basic filters displays all programs out of the 111 collected for this study that qualify after filtering. These are displayed as a table with university name, program name and URL.

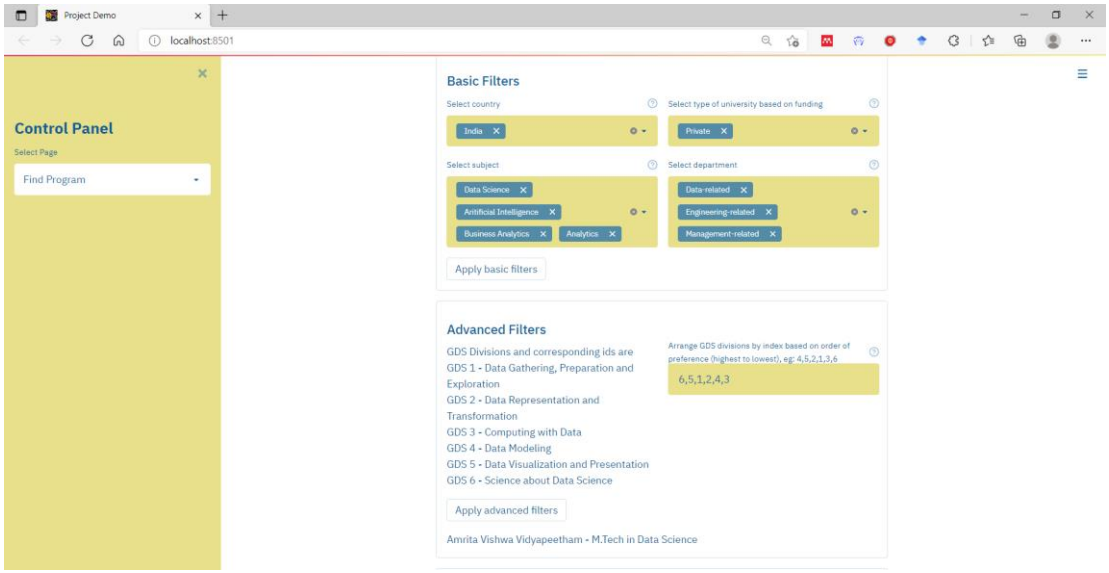


Figure 24. Find programs – Applying advanced filters

The advanced filter as described in algorithm 2 (page 29), is in fact a computation to extract programs that teach the 6 GDS divisions in a way similar to the user's priorities. In figure 25, the user's highest priority is to study GDS 6. This means that a program that does not cover GDS 6 will most likely not be a fit for this user irrespective of how well it covers GDS 3 or 2. Based on this input, the program that scores the highest compatibility with this user is presented. The user can then get content-based recommendations of other programs within his/her preferred country that are similar to the above generated program on the basis of program descriptions as well as the GDS proportions (figure 25).

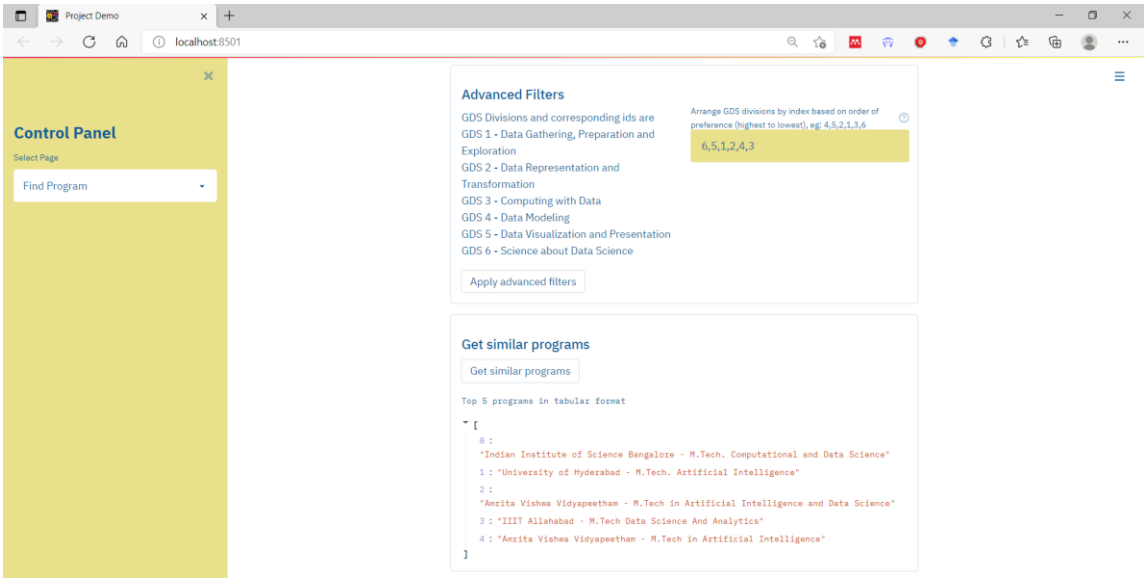


Figure 25. Find programs – Generating content-based recommendations

Chapter 6 – Discussion

Understanding data science education (DSE) is of key interest to the field of data science, mainly due to its ability to develop, evaluate and enhance curriculums to meet the demand of the growing community of data science enthusiasts (Donoho, 2017). However, DSE as a research area is still in its nascent stages and is yet to be formalized as an active attempt to compare and contrast graduate level programs. Those evaluations that have been performed have been overly concentrated on programs in the USA (Oliver and McNeil, 2021; Friedman, 2018) and seldom did little to understand differences arising in programs due to factors such as a university's funding status and the type of department offering the program. In this study, I attempted to evaluate 111 Masters' level data programs offered at the top 50 universities (as per QS rankings) from both India and the USA with the help of the Greater Data Science (GDS) framework developed in Donoho (2017). Studies in the past have attempted to evaluate and compare programs using purely qualitative methods and these are hard to reproduce, even if the coding rubric is made explicit. Hence, the attempt here was to use the dictionary approach to evaluate programs. The dictionary approach was used to set in motion the attempt to formalize a general approach to evaluate data science programs. Since a dictionary for such purposes is yet to be defined, I created a dictionary by labelling words in the titles of compulsory courses across the programs based on which GDS division(s) it implied. This dictionary was used to score all 111 programs based on the titles of their compulsory courses and for each program, the proportions of the program covering each of the 6 GDS divisions were obtained, along with a GDS score which was the standard deviation of the division proportions.

Across all 111 programs in the sample, it was gathered that most emphasis was placed on topics concerning Computing with Data while the least emphasis was on topics concerning Science about Data Science (such as meta-analysis, tool evaluation, ethics, reproducibility etc.). This is in line with what was observed by Donoho (2017) where the author made a call to expand data curriculums to include topics that would help make it a "science". An argument here could be that this lack

of representation of this division could also be attributed to programs not having course titles that explicitly depict this division, while some programs might teach them anyways as smaller components. However, this is exactly the point I wish to make. Concepts in this division are not to be relegated as second-class citizens in data science curriculums, if the goal is to create responsible data practitioners. Important differences emerged between programs in India and the USA on the basis of what was concentrated on. Programs in the USA were more balanced in their coverage of different data science areas than those in India. While programs in the USA concentrated more in Data Modeling and Data Gathering, Preparation and Transformation than those in India, Indian programs focused more on Computing with Data. One of the primary reasons to this could be the fact that over 50% of Indian programs in the sample were offered by departments of engineering (and technology) compared to the 26% in the USA. While there were no differences in programs on the basis of whether the host university was privately-funded or public-funded, a surprising result was that private universities in India covered Science about Data Science with greater effect than the public universities in the country. This might have to do with the "communication" courses placed into curriculums in private universities in India. In the USA, departments that specifically dealt with data-related studies offered the most balanced programs i.e an almost equal coverage of all 6 GDS divisions when compared to other departments. In India, such data-related departments were a rarity (only 2). From these findings, there emerges an idea that DSE in India does lag from that in the USA on the basis of how well Masters' programs train students in the different divisions of data science.

The emphasis however needs to be *less on the existence of differences* and *more on how to patch up these gaps*, especially from the perspective of a country like India that has immense potential in the data spectrum (Kaggle, 2020). A useful recommendation here could be that more data-related departments be setup in India based on the attempts in the USA in order to provide more balanced programs. Preparing the data workforce for the future by covering Science about Data Science will help both countries develop generations of responsible data practitioners (Donoho, 2017).

The development of the Data Program Selector (DPS) web-based tool was a novel way to venture into DSE via application built on theory. The tool makes use of the created GDS dictionary to score a new program. The DPS helps users understand the overall coverage of a new program they have come across on the internet. This feature could be extended to any program (including non-academic ones) in any country. The DPS also offers more than traditional program database aggregators by filtering programs to a user based on what data science areas they would like their program to focus on more and offer recommendations similar to these. This is an important step towards helping prospective learners choose the program that they feel will be the best fit.

6.1 Limitations

While this study will be helpful for those interested in DSE, it does have limitations like most pilot attempts. While the dictionary is a more generalizable approach than qualitative coding, its validity has not been formally studied. The assigning of GDS divisions to a word has been based mainly on the GDS framework in Donoho (2017) and to a limited extent, on my subjective interpretation of words. A more robust way of labelling would be useful to improve the validity of the results. Generalizability concerns exist as only universities in the top 50 in India and the USA have been considered in the study. While this was a deliberate choice for the purposes of data collection, a larger sample will most likely be better suited for generalizations. While attempts have been made to include all kinds of data programs by selecting programs with titles having one or more words from a selected keyword list, there is still a chance that some programs that covered data concepts were overlooked. Also, the study evaluates programs only on the basis of compulsory course titles. A broader outlook including contents of the courses, the final projects, assessments, experience of faculty etc. would be recommended to understand the complete scope and coverage of data programs.

6.2 Future scope

This study is to be extended in the future to include variables such as contents of the courses, the final projects, assessments, experience of faculty etc. in order to understand the complete scope and coverage of data programs. The study can be applied to countries other than India and the USA, in an attempt to understand DSE across boundaries. The DPS tool, a key deliverable in this project is a prototype of what could in the future be a complete product built to objectively evaluate data programs and connect inquisitive learners to the programs that fit their profiles and interests best. The area of DSE research is a continuously evolving endeavour and in accordance with that, it will be important to update the GDS dictionary at regular checkpoints to keep its contents relevant to the rapidly changing world of data science.

Acknowledgements

I would like to express my gratitude to Dr. Patrick Sturgis for the timely guidance that has helped me shape my research. I also thank the Department of Methodology for providing useful resources that have helped write this report.

7. References

1. 2020 Kaggle ML & DS Survey | Kaggle (no date). Available at: https://www.kaggle.com/c/kaggle-survey-2020 (Accessed: 10 December 2020).
2. Aasheim, C.L., Williams, S., Rutner, P. and Gardiner, A., 2015. Data analytics vs. data science: A study of similarities and differences in undergraduate programs based on course descriptions. Journal of Information Systems Education, 26(2), p.103.
3. Anandarajan, M., Hill, C. and Nolan, T., 2019. Term-document representation. In Practical Text Analytics (pp. 61-73). Springer, Cham.
4. Anderson, P., Bowring, J., McCauley, R., Pothering, G. and Starr, C., 2014, March. An undergraduate degree in data science: curriculum and a decade of implementation experience. In Proceedings of the 45th ACM technical symposium on Computer science education (pp. 145-150).
5. Baturay, M.H., 2015. An overview of the world of MOOCs. Procedia-Social and Behavioral Sciences, 174, pp.427-433.
6. Baumer, B., 2015. A data science course for undergraduates: Thinking with data. The American Statistician, 69(4), pp.334-342.
7. Blei, D. M. and Smyth, P. (2017) 'Science and data science', Proceedings of the National Academy of Sciences of the United States of America, 114(33), pp. 8689–8692. doi: 10.1073/pnas.1702076114.
8. Börner, K., Scrivner, O., Gallant, M., Ma, S., Liu, X., Chewning, K., Wu, L. and Evans, J.A., 2018. Skill discrepancies between research, education, and jobs reveal the critical need to supply soft skills for the data economy. Proceedings of the National Academy of Sciences, 115(50), pp.12630-12637.
9. Breiman, L., 2001. Statistical modeling: The two cultures (with comments and a rejoinder by the author). Statistical science, 16(3), pp.199-231.
10. Chandra, R.V. and Varanasi, B.S., 2015. Python requests essentials. Packt Publishing Ltd.

11. Chen, H., Chiang, R.H. and Storey, V.C., 2012. Business intelligence and analytics: From big data to big impact. MIS quarterly, pp.1165-1188.
12. Cleveland, W.S., 2001. Data science: an action plan for expanding the technical areas of the field of statistics. International statistical review, 69(1), pp.21-26.
13. Conway, D., 2010. The Data Science Venn Diagram. Available at: http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram (Accessed: 8 December 2020).
14. Davenport, M. G. et al. (2015) 'Data Driven: What students need to succeed in a rapidly changing business world', PricewaterhouseCoopers LLP, (February 2015).
15. Davenport, T.H. and Patil, D.J., 2012. Data scientist. Harvard business review, 90(5), pp.70-76.
16. Della Volpe, M. and Esposito, F., 2020. How universities fill the talent gap: The data scientist in the Italian case. African Journal of Business Management, 14(2), pp.53-64.
17. Dinno, A., 2015. Nonparametric pairwise multiple comparisons in independent groups using Dunn's test. The Stata Journal, 15(1), pp.292-300.
18. Donoho, D. (2017) '50 Years of Data Science', Journal of Computational and Graphical Statistics, 26(4), pp. 745–766. doi: 10.1080/10618600.2017.1384734.
19. Esser, F. and Vliegenthart, R., 2017. Comparative research methods. The international encyclopedia of communication research methods, pp.1-22.
20. Fréchet, N., Savoie, J. and Dufresne, Y., 2020. Analysis of text-analysis syllabi: Building a text-analysis syllabus using scaling. PS: Political Science & Politics, 53(2), pp.338-343.
21. Friedman, A., 2018. Measuring the promise of Big Data syllabi. Technology, Pedagogy and Education, 27(2), pp.135-148.
22. Friedman, A., 2019. Data science syllabi measuring its content. Education and Information Technologies, 24(6), pp.3467-3481.
23. Ghasemi, A. and Zahediasl, S., 2012. Normality tests for statistical analysis: a guide for non-statisticians. International journal of endocrinology and metabolism, 10(2), p.486.

24. Heimerl, F., Lohmann, S., Lange, S. and Ertl, T., 2014, January. Word cloud explorer: Text analytics based on word clouds. In 2014 47th Hawaii International Conference on System Sciences (pp. 1833-1842). IEEE.
25. Hicks, S.C. and Irizarry, R.A., 2018. A guide to teaching data science. The American Statistician, 72(4), pp.382-391.
26. Horton, N. J., Baumer, B. S. and Wickham, H. (2015) 'Taking a Chance in the Classroom: Setting the Stage for Data Science: Integration of Data Management Skills in Introductory and Second Courses in Statistics', CHANCE, 28(2), pp. 40–50. doi: 10.1080/09332480.2015.1042739.
27. Hosein, A. and Rao, N., 2021. Selling lemons? The relationship between learning and teaching information on university programme web pages and future students' course satisfaction. Journal of Marketing for Higher Education, pp.1-18.
28. Irizarry, R.A., 2020. The role of academia in data science education.
29. Kaggle: Your Home for Data Science (no date). Available at: https://www.kaggle.com/ (Accessed: 10 December 2020).
30. Kassambara, A., 2020. rstatix: Pipe-friendly framework for basic statistical tests. R package version 0.6. 0.
31. Kim, J., 2016. Who is teaching data: meeting the demand for data professionals. Journal of Education for Library and Information Science, 57(2), pp.161-173.
32. Lang, C., Siemens, G., Wise, A. and Gasevic, D. eds., 2017. Handbook of learning analytics. New York, NY, USA: SOLAR, Society for Learning Analytics and Research.
33. Li, X., Bai, Y. and Kang, Y., 2021. Exploring the social influence of Kaggle virtual community on the M5 competition. arXiv preprint arXiv:2103.00501.
34. Lyon, L. and Mattern, E., 2017. Education for real-world data science roles (Part 2): A translational approach to curriculum development. International Journal of Digital Curation, 11(2), pp.13-26.
35. McDonald, J.H., 2009. Handbook of biological statistics (Vol. 2, pp. 6-59). Baltimore, MD: sparky house publishing.

36. Mihalcea, R. and Tarau, P., 2004, July. Textrank: Bringing order into text. In Proceedings of the 2004 conference on empirical methods in natural language processing (pp. 404-411).
37. Montgomery, D.C. and Runger, G.C., 2010. Applied statistics and probability for engineers. John Wiley & Sons.
38. Nathan, P., 2016. Pytextrank, a python implementation of textrank for text document nlp parsing and summarization. https://github. com/ceteri/pytextrank.
39. National Academies of Sciences, Engineering, and Medicine, 2018. Data science for undergraduates: Opportunities and options. National Academies Press.
40. Nayak, B.K. and Hazra, A., 2011. How to choose the right statistical test?. Indian journal of ophthalmology, 59(2), p.85.
41. Oliver, J.C. and McNeil, T., 2021. Undergraduate data science degrees emphasize computer science and statistics but fall short in ethics training and domain-specific context. PeerJ Computer Science, 7, p.e441.
42. Ortiz-Repiso, V., Greenberg, J. and Calzada-Prado, J., 2018. A cross-institutional analysis of data-related curricula in information science programmes: A focused look at the iSchools. Journal of Information Science, 44(6), pp.768-784.
43. Palinkas, L.A., Horwitz, S.M., Green, C.A., Wisdom, J.P., Duan, N. and Hoagwood, K., 2015. Purposeful sampling for qualitative data collection and analysis in mixed method implementation research. Administration and policy in mental health and mental health services research, 42(5), pp.533-544.
44. Pappano, L., 2012. The Year of the MOOC. The New York Times, 2(12), p.2012.
45. Pazzani, M.J. and Billsus, D., 2007. Content-based recommendation systems. In The adaptive web (pp. 325-341). Springer, Berlin, Heidelberg.
46. Richardson, L., 2007. Beautiful soup documentation. Dosegljivo: https://www. crummy. com/software/BeautifulSoup/bs4/doc/.[Dostopano: 7. 7. 2018].

47. Rose, S., Engel, D., Cramer, N. and Cowley, W., 2010. Automatic keyword extraction from individual documents. Text mining: applications and theory, 1, pp.1-20.
48. Sharma, G., 2017. Pros and cons of different sampling techniques. International journal of applied research, 3(7), pp.749-752.
49. Silbiger, S., 2016. The 10-day MBA: a step-by-step guide to mastering the skills taught in top business schools. Hachette UK.
50. Sirisuriya, D.S., 2015. A comparative study on web scraping.
51. Tang, R. and Sae-Lim, W., 2016. Data science programs in US higher education: An exploratory content analysis of program description, curriculum structure, and course focus. Education for Information, 32(3), pp.269-290.
52. Team, R.C., 2013. R: A language and environment for statistical computing.
53. Tukey, J.W., 1962. The future of data analysis. The annals of mathematical statistics, 33(1), pp.1-67.
54. Varvel Jr, V.E., Bammerlin, E.J. and Palmer, C.L., 2012. Education for data professionals: A study of current courses and programs. In Proceedings of the 2012 iConference (pp. 527-529).
55. Yan, D. and Davis, G. E. (2019) 'A First Course in Data Science', Journal of Statistics Education, 27(2), pp. 99–109. doi: 10.1080/10691898.2019.1623136.
56. Zhang, J., Fu, A., Wang, H. and Yin, S., 2017. The development of data science education in China from the LIS perspective. International Journal of Librarianship, 2(2), pp.3-17.
57. Zhang, S., Zhang, C. and Yang, Q., 2003. Data preparation for data mining. Applied artificial intelligence, 17(5-6), pp.375-381.
58. Zhao, B., 2017. Web scraping. Encyclopedia of big data, pp.1-3.
59. Kolaczyk, E. D., Wright, H., & Yajima, M. (2021). Statistics Practicum: Placing "Practice" at the Center of Data Science Education . *Harvard Data Science Review*. https://doi.org/10.1162/99608f92.2d65fc70
60. Loukides, M., 2010. *What is data science?*. [online] O'Reilly Media. Available at: https://www.oreilly.com/radar/what-is-data-science/ (Accessed: 18 April 2021)

61. DuBois, J., 2020. *The Data Scientist Shortage in 2020*. [online] QuantHub. Available at: https://quanthub.com/data-scientist-shortage-2020/ (Accessed: 17 April 2021).
62. Lodico, M.G., Spaulding, D.T. and Voegtle, K.H., 2010. *Methods in educational research: From theory to practice* (Vol. 28). John Wiley & Sons.
63. Usmani, Z., 2019. *What is Kaggle, Why I Participate, What is the Impact? | Data Science and Machine Learning*. [online] Kaggle.com. Available at: https://www.kaggle.com/getting-started/44916 (Accessed: 1 May 2021).
64. Aguillo, I., Bar-Ilan, J., Levene, M. and Ortega, J., 2010. Comparing university rankings. *Scientometrics*, *85*(1), pp.243-256.
65. Lahitani, A.R., Permanasari, A.E. and Setiawan, N.A., 2016, April. Cosine similarity to determine similarity measure: Study case in online essay assessment. In *2016 4th International Conference on Cyber and IT Service Management* (pp. 1-6). IEEE.

Appendix A – Summary of Collected Data

A.1 Variables in the data collected

The Program data for both India and the USA had the following variables.

Name of variable	What it contained	Type of data
uni_name	Name of university	Categorical
pgm_name	Name of program offered	Textual
dur_yrs	Duration of program in years	Numeric
type	Funding status of university	Categorical
national_rank_qs	Relative national rank as per QS 2020	Numeric
dept_involved	Name of department(s) involved	Textual
location	Location of university	Categorical
url	Program webpage URL	Textual
descr	Description of program on the program webpage	Textual
header_tag	Types of HTML header tags	List of categorical data
header_names	Text in HTML header tags	List of textual data
links	Links on the program webpage	List of textual
date_pub	Date when the program webpage was first published	Datetime
date_last_mod	Date when the program webpage was last modified	Datetime
html_tags	Types of HTML tags	List of categorical data
html_tag_freq	Frequencies of occurrences of different HTML tags	Dictionary
html_emphasized_elements	Text in emphasized HTML tags*	List of textual data

*Tags such as <b>, , , <mark>, <i>

The Curriculum data for both India and the USA had the following variables.

Name of variable	What it contained	Type of data
url	Program webpage URL	Textual
compulsory_course	Name of compulsory course	Textual
overview	Overview of the compulsory course	Textual
topics_covered	List of topics covered in the compulsory course	Textual
Both overview and topics_covered contained more or less the same type of data and hence both were merged together at a later stage in analysis		

A.2 Programs included in study

A.2.1 From India

1. 'Indian Institute of Science Bangalore - M.Tech. Computational and Data Science',
2. 'Indian Institute of Science Bangalore - M.Tech. Artificial Intelligence',
3. 'University of Hyderabad - MBA Business Analytics',
4. 'University of Hyderabad - M.Tech. Artificial Intelligence',
5. 'IIT Guwahati - M.Tech Data Science',
6. 'IIT Hyderabad - M.Tech Artificial Intelligence',
7. 'Anna University - M.E. Computer Science and Engineering (Specialization in Big Data Analytics)',
8. 'Jamia Millia Islamia - MSc in Banking and Financial Analytics',
9. 'Symbiois International - MBA Business Analytics',
10. 'Symbiois International - MSc Data Science and Spatial Analytics',
11. 'IIT Ropar - M.Tech Artificial Intelligence',
12. 'Manipal Academy of Higher Education - MSc in Business Analytics',
13. 'NIT Tiruchirappalli - M.Tech Data Analytics',
14. 'Vellore Institute of Technology Vellore - MSc Data Science',
15. 'Vellore Institute of Technology Vellore - MSc Business Statistics',
16. 'Vellore Institute of Technology Vellore - M.Tech Computer Science and Engineering - Specialization in Artificial Intelligence & Machine Learning',
17. 'Vellore Institute of Technology Vellore - M.Tech Computer Science and Engineering - Specialization in Big Data Analytics',

18. 'Annamalai University - MSc Data Science',
19. 'Annamalai University - MBA Business Analytics',
20. 'IIIT Allahabad - M.Tech Data Science And Analytics',
21. 'Amrita Vishwa Vidyapeetham - M.Tech in Artificial Intelligence',
22. 'Amrita Vishwa Vidyapeetham - M.Tech in Artificial Intelligence and Data Science',
23. 'Amrita Vishwa Vidyapeetham - M.Tech in Data Science',
24. 'Indraprastha Institute of Information Technology - M.Tech (CSE) with specialization in Data Engineering',
25. 'Indraprastha Institute of Information Technology - M.Tech (CSE) with specialization in Artificial Intelligence',
26. 'Amity University - M.Tech Data Science',
27. 'Amity University - MBA Business Analytics',
28. 'IIT Dhanbad - MBA Business Analytics',
29. 'IIT Dhanbad - M.Tech Data Analytics',
30. 'University of Kerala - MSc Computer Science (Artificial Intelligence) ',

A.2.2 From USA

1. 'Harvard University - Master of Science Data Science',
2. 'Stanford University - Master of Science Education Data Science ',
3. 'Stanford University - Master of Science in Statistics: Data Science',
4. 'Massachusetts Institute of Technology (MIT) - Master of Business Analytics',
5. 'Columbia University - Master of Science in Applied Analytics',
6. 'Columbia University - Master of Science in Data Science',
7. 'University of California, Los Angeles - Master of Science in Business Analytics',
8. 'Yale University - Master of Science in Statistics and Data Science',
9. 'University of Pennysylvania - Master of Science in Engineering in Data Science',
10. 'University of Pennysylvania - Master of Science in Social Policy + Data Analytics',
11. 'University of Pennysylvania - MBA Business Analytics',
12. 'University of Pennysylvania - Master of Urban Spatial Analytics',
13. 'Cornell University - Master of Professional Studies in Applied Statistics - Data Science',
14. 'New York University - Master of Science in Data Science',
15. 'New York University - Master of Science in Human Capital Analytics and Technology (STEM)',
16. 'University of Chicago - Master of Science in Analytics',
17. 'Duke University - Master of Science in Interdisciplinary Data Science',
18. 'Duke University - Master of Engineering in Artificial Intelligence for Product Innovation',
19. 'John Hopkins University - Master of Science in Data Science',

20. 'John Hopkins University - Master of Science in Business Analytics and Risk Management',
21. 'University of Southern California - Master of Science in Business Analytics',
22. 'University of Southern California - Master of Science in Biomedical Engineering - Biomedical Data Analytics',
23. 'University of Southern California - Master of Science in Computer Science - Data Science',
24. 'University of Southern California - Master of Science in Applied Data Science',
25. 'University of Southern California - Master of Science in Communication Data Science',
26. 'University of Southern California - Master of Science in Environmental Data Science',
27. 'University of Southern California - Master of Science in Healthcare Data Science',
28. 'University of Southern California - Master of Science in Spatial Data Science',
29. 'University of Southern California - Master of Science in Electrical and Computer Engineering - Machine Learning and Data Science',
30. 'University of Southern California - Master of Science in Analytics',
31. 'University of Southern California - Master of Science in Computer Science - Artificial Intelligence',
32. 'Northwestern University - Master of Science in Analytics',
33. 'Northwestern University - Master of Science in Artificial Intelligence',
34. 'Carnegie Mellon University - Master of Computational Data Science',
35. 'Carnegie Mellon University - Master of Science in Artificial Intelligence and Innovation',
36. 'Carnegie Mellon University - Master of Science in Machine Learning',
37. 'University of Michigan - Ann Arbor - Master of Science in Data Science',
38. 'University of Michigan - Ann Arbor - Master of Science in Survey and Data Science - Data Science Concentration',
39. 'Brown University - Master of Science in Data Science',
40. 'Brown University - Master of Science in Social Analysis and Research',
41. 'Boston University - Master of Science in Applied Data Analytics',
42. 'Boston University - Master of Science in Computer Science with Specialization in Data-Centric Computing',
43. 'Boston University - Master of Science in Applied Business Analytics',
44. 'Boston University - Master of Science in Artificial Intelligence',
45. 'Boston University - Master of Science in Business Analytics',
46. 'Rice University - Master of Data Science',
47. 'Georgetown University - Master of Science in Health Informatics and Data Science',
48. 'Georgetown University - Master of Science in Data Science and Analytics',
49. 'Georgetown University - Master of Science in Data Science for Public Policy',
50. 'University of Washington - Master of Science in Business Analytics',
51. 'University of Washington - Master of Science in Data Science',
52. 'University of California, San Deigo - Master of Science in Business Analytics',

53. 'Emory University - Master of Science in Business Analytics',
54. 'University of California, Davis - Master of Science in Business Analytics',
55. 'University of Rochester - Master of Science in Data Science',
56. 'Vanderbilt University - Master of Science in Data Science',
57. 'Georgia Institute of Technology - Master of Science in Analytics',
58. 'University of Illinois at Urbana-Champaign - Master of Science in Business Analytics',
59. 'George Washington University - Master of Science in Data Analytics',
60. 'George Washington University - Master of Science in Data Science',
61. 'George Washington University - Master of Science in Business Analytics',
62. 'Tufts University - Master of Science in Data Science',
63. 'Tufts University - Master of Science in Data Analytics',
64. 'University of Miami - Master of Science in Business Analytics',
65. 'University of Miami - Master of Science in Electrical and Computer Engineering - Concentration in Artificial Intelligence',
66. 'University of Miami - Master of Science in Data Science',
67. 'University of Notre Dame - Master of Science in Data Science',
68. 'University of Notre Dame - Master of Science in Business Analytics',
69. 'University of California, Irvine - Master of Data Science',
70. 'University of California, Irvine - Master of Science in Business Analytics',
71. 'University of California, Irvine - Master of Arts in Social Sciences - Concentration in Demographic and Social Analysis',
72. 'University of Illinois at Chicago - Master of Science in Business Analytics',
73. 'University of Illinois at Chicago - Master of Science in Civic Analytics',
74. 'University at Buffalo SUNY - Master of Science in Business Analytics',
75. 'University at Buffalo SUNY - Master of Science in Data Analysis in the Social Sciences',
76. 'University at Buffalo SUNY - Master of Professional Studies in Data Sciences and Applications',
77. 'Pennsylvania State University - Master of Science in Data Analytics',
78. 'University of Maryland, College Park - Master of Professional Studies in Data Science and Analytics',
79. 'University of Maryland, College Park - Master of Science in Applied Political Analytics',
80. 'University of Maryland, College Park - Master of Science in Business Analytics',
81. 'University of Maryland, College Park - Master of Science in Marketing Analytics'

Appendix B – Word Labelling Interface

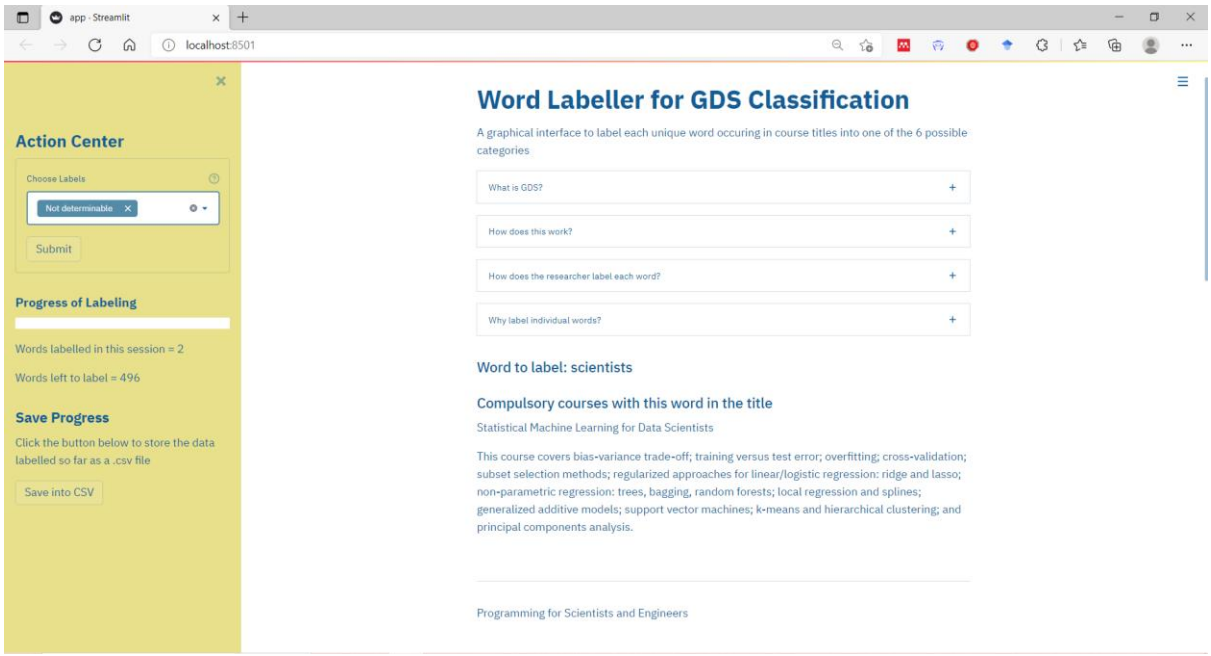


Figure a) Word labeller interface

The above interface was used to label each of the 498 unique words that made up the compulsory course titles. The interface displayed each word along with the names of the course titles it appeared in and their descriptions (if available). This helped me assign a word into a given GDS division based on which kinds of courses it appeared in. The GDS framework was consulted predominantly in the course of labelling. The interface was written in python using the streamlit framework for the front-end.

If a word was ambiguous w.r.t GDS division, it was labelled as “Not determinable” (as in figure a). A word was labelled into a GDS division only if

1. It or one of its synonyms belonged to that division as per the GDS framework in chapter 3
2. Or the courses in whose titles it appeared were teaching concepts predominantly relating to that GDS divisions

Some words were also assigned multiple GDS divisions. For example, “analysis” was put into everything except GDS 6 – Science about Data Science.

Appendix C – Code & Deployed Application

C.1 Code

All code for this project is hosted on Github.

Link to code: https://github.com/ry05/data-science-education-research

C.2 Deployed Application

The Data Program Selector app has been deployed using Streamlit Sharing.

Link to app: https://share.streamlit.io/ry05/data_program_selector_app/app.py