Defining Data Science

A Case Study in Australia

Xinrui WANG and Tsai-Chun TSOU

‘28 October 2022

Table of contents

Preface

What is Data Science? What do you learn in Data Science? What kind of jobs do Data Scientists do? These are the type of questions we've encountered during social occasions and despite spending almost two years studying Business Analytics, we still struggle to give a definite answer. Data Science, Business Analysis, and Data Analysis are new degrees that are created only in recent years thanks to the advance in technology and increasing demand for talent to work with data. The sudden rise of these degrees gave authorities little time to regulate them. Even the Australian and New Zealand Standard Classification of Occupations (ANZSCO), the Australian Bureau of Statistics (ABS) and the Department of Home Affair are slow to catch up with their description of this occupation.

After four months of gathering data, conducting analysis, and running models, we have attempted to define Data Science in Australia. Special thanks to Professor Di Cook and Professor Rob Hyndman for their support. A huge thank you to Dr. Emi Tanaka for the opportunity to work on such a special project and for her guidance throughout the semester[1].

[1]This report is written jointly by Tsai-Chun Tsou and Xinrui Wang as part of the ETC5543 research project supervised by Dr. Emi Tanaka.

1 Abstract

With the increase in demand for data scientist and the rising popularity of the degrees, how does one define Data Science in Australia. The Australian Mathematical Sciences Institute (AMSI) and the Statistical Society of Australia (SSA) are conducting a review of Australian Data Science Degrees with surveys and focus groups. The survey method is prone to biases due to the subjective nature. Our research attempts to tackle similar topic but with a more objective approach by collecting data directly from public resources. We collected Data Science related unit information from Group of Eight Universities and Data Scientist Job Listings from kaggle. We used the data to decompose the core disciplines involved in the degree as well the type of skill sets that may be required. To expand on the initial exploratory data analysis, we also build Latent Dirichlet Allocation models to construct our own text corpus.

From the exploratory data analysis, we observe a lack of homogeneity within the Universities' analysis. The inconsistent data metrics made it difficult to draw direct comparison between the employer data and university data. Nonetheless we were able to conclude that computational disciplines are more prominent on both sides.

2 Introduction

Data Science has ranked as one of the most in-demand jobs in Australia in recent consecutive years. As demands steadily grows, students are also increasingly interested in Data Science degrees, yet recruiters still seem to struggle to fill up data science positions. This leads to our main question: what is Data Science? Is there a shared structure or skill set of Data Science courses offered at Australian universities? Are students and employers' perception of data science similar? To answer these questions, we looked a data from both University and Employer perspectives.

There is no readily available data from Australian universities, so we had to build our own data set through web scraping. The initial target was to collect data from all universities in Australia including both undergraduate and postgraduate courses, however, due to time constrain, the data collected for this project only contains Master of Data Science courses from the Group of Eight (Go8) universities. The employer data was retrieved from Data Scientist Job Listings on kaggle.

By exploring the current situation and potentially a definition of Data Science in Australia from both university and employer perspectives, the findings would help students and recruiters have a clearer picture of what to expect, as well as raising attentions and awareness to potential gaps between employer demands and university offerings.

Part I

Data Collection

3 University Data

3.1 Web Scraping

In order to explore the Data Science degrees around Australia Universities, we compiled a list of universities in Australia and the Data Science or related degrees they offered, then web scraped required information from each university's website using R. In total, we collected 298 units from eight postgraduate courses in Data Science across all Group of Eight (Go8) universities.

To start off the project, Professor Tanaka provided sample code for data scraping using Monash Handbook as an example. Libraries `rvest` and `rSelenium` are two of the main tools. Initially, we studied her code and then tried to replicate her code to be applied to other university's websites.

The flow of the data scraping is as follow (example code from `uom-master-datasci.qmd`):

1. Identify the main page (url) where the degree information is contained, which usually is the most updated version of the handbook.

```
remDr$navigate("https://handbook.unimelb.edu.au/2022/courses/mc-datasc/course-structure")
sub_list <- read_html(remDr$getPageSource()[[1]])
```

2. Use functions from `rvest` to retrieve all the course unit code (or course unit url). Retrieve the degree code and formal degree name and save it for later.

```
curriculum <- sub_list %>%
      html_element("#top") %>%
      html_element(".mobile-wrap") %>%
      html_elements("table") %>%
      html_elements("a") %>%
      html_attr("href")
```

3. Use `rSelenium` functions and course unit information, to direct R to the unit information page.
4. Retrieve the following information from the page using rvest functions:

- Unit Name
- Unit Code
- Unit Overview
- Unit Learning Outcome
- Unit Prohibition/ Pre-requisite/ Co-requisite

5. Repeat step 3 & 4 with loop function.

```
for(unit in curriculum) {
      remDr$navigate(glue("{baseurl}{unit}"))
      wait_time()
      unit_html <- read_html(remDr$getPageSource()[[1]])

      # unit name
      subject_text <- unit_html %>%
        html_element("h1") %>%
        html_text()
        ...
```

6. Compile all the retrieved data from the University into a dataframe and export it as csv file.

```
data <- data %>%
        bind_rows(tibble(!!!c(list(Course = title,
                                   #Course_code = "MC-DATASC",
                                   Course_overview = paste0(coverview, collapse = " "),
                                   #Unit_code = cunit,
                                   Unit = subject_text,
                                   Overview = overview,
                                   Prerequisite = paste0(pre, collapse = ", "),
                                   Corequisite = co,
                                   Prohibition = paste0(pro, collapse = ", "),
                                   Outcomes =lo
                                   ))))
```

Despite the process being similar for each University, we soon realize the process was going to be more challenging than expected.

3.2 Inconsistent Information

Monash University's student handbook on Degrees and Courses is a spectacular website for data scraping. Its html code is clearly labeled and anything you need to know about the degree

or course can be found on the website. The same cannot be said about other universities.

The course descriptions on the handbook and universities' website page are usually structured in a different manner, since the majority of the data is collected from universities' handbooks, course descriptions are also extracted from handbooks for consistency purposes.

In addition, the required unit information listed above is not all available at the targeted universities. The handbook from University of New South Wales contains extremely limited information: unit overview is brief, unit requisites are only available for a few units, and unit outcome is not provided at all.

3.3 Difficult to Manage Websites

Each university website is unique. Sometimes the information is not straightforward. An example of this is University of Adelaide. The main website for the degree does contain the list of units that go into the degree.

However, instead of having just one page with all the unit information, the link takes you to a page with different unit information depending on when the unit is offered and on what campus.

Tina tried bypassing the pages by directly looking at the url of the final unit information page I want to be on. Unfortunately, the url is not designed or structured in a way which she was able to predict the url based on the current unit code. With that said, her only option was to code the function to jump from pages to pages before landing on the right unit information page.

It is also often found that the unit overview and learning outcomes for each unit within the same university could vary slightly in format. For example, unit overview may appears before or after campus location at University of Western Australia, empty spaces could be found after section title at the University of Melbourne, which would break the chain of extracting corresponding information.

3.4 Collected Data

The collected data contains **298 units** from 8 universities, and **8 variables** including School, Course, Course_code, Unit, Unit_code, Outcomes, Overview and Description. The full data set is made available below for exploration.

Show 10 entries

Search:

	School	Course	Course_code	Unit	Unit_code	Outcomes	Overview
1	monash	Master of Data Science	C6004	FIT9132 - Introduction to databases	FIT9132	Apply the theories of the relational database model. Develop a sound relational database design.I Implement a relational database using a sound database design.II Manage data that meets user requirements, including queries and transactions.I Contrast the key characteristics of the relational database models and the relational database model.	This unit will introduce the concept of data management in an organisation through relational database technology. Theoretical foundation of relational database analysis and design, implementation of relational database using Oracle is covered.
2	monash	Master of Data Science	C6004	FIT9136 - Algorithms and programming foundations in Python	FIT9136	Apply best practice Python programming constructs for writing correct, robust, and readable.I Restructure a computational program into smaller units using functions and classes using the object-oriented methodology.I Implement basic algorithmic strategies in a Python application and apply debugging, testing and exception handling techniques.I Investigate useful Python libraries for scientific computing and data analysis.I Experiment with data visualisation and analysis and visualisation technique to formulate business insight.	This unit introduces the Python programming and the basics of data structure and algorithms including their analysis and implementation as a preparatory experience working with Python implementation of data structures and algorithms widely used in modern programming language to solve simple problems. Topics covered in this unit are listed in the content click the Read more button below.
3	monash	Master of Data Science	C6004	FIT9137 - Introduction to computer architecture and networks	FIT9137	Describe basic concepts of computer hardware and software architecture.I Explain the role of the OS and an operating system (OS), namely, process management, memory management, I/O management.I Analyse and formulate the functions and protocols of computer networks (local area networks, wide area networks and the Internet.I Examine networks using the underlying fundamental principles and concepts for data transmission.I Use the fundamental concepts of computer and network security including identifying common threats and applying countermeasures.	This unit introduces students to fundamentals of computer systems, with emphasis on the modules: Module 1 Computer architecture and modern computer systems (4hr) Module 2 Computing technology: Module 3 CPU, memory storage and operating system basics. Module 4 Data classic TCP/IP layered architecture, MAN and LAN networking For more content click the Read more button below.
4	monash	Master of Data Science	C6004	MAT9004 - Mathematical foundations for data science and AI	MAT9004	Use trees and graphs to solve problems in computer science.I Apply counting principles in computer science.I Apply principles of elementary probability theory, evaluate probabilities and expected and use Bayes' Theorem.I Demonstrate basic knowledge of linear algebra including the manipulation of matrices, solution of linear systems and the ability to find and apply determinants.I Explain fundamental concepts in calculus including differentiation and integration, and composite, inverse and implicit functions.I Demonstrate key skills in the calculus of functions of several variables including the calculation of partial derivatives, find tangent planes and identify stationary points, real findings and convexity for optimisation.	Mathematical topics fundamental to computer science and statistics including trees and graphs, counting in, principles of elementary probability theory, linear algebra, and differential calculus concepts of calculus in one and several variables.
5	monash	Master of Data Science	C6004	FIT5126 - IT research methods	FIT5126	Review and critique research literature (from a wide range of disciplines) to develop and research findings in a sub-discipline of IT.I Communicate research findings in a form of writing, online and orally, video, in research and industry contexts, using appropriate and professional issues that may arise in IT research and practice.I Critically analyse and identify and explain major research philosophies and methodologies.I Conduct ethical and justify research methods and techniques of data collection and analysis appropriate to particular research aims and IT projects in sub-disciplines of IT.I Analyse and describe the organisational, political and socio-economic contexts that motivate research, and the implications of research outcomes.I Design and carry out a rigorous and ethical research project.I Produce a report explaining the project, its design and interpretation of the findings.I Demonstrate ability to operate effectively as a member of a research project team.	The aim of the unit is to introduce Masters students to research across the range of sub-disciplines. It provides a broad coverage of the issues, philosophies, methods, and techniques associated with Computer Science, Software Engineering, Information Systems and (information) Technology research. It introduces students to the major research paradigms, the nature of... For more content click the Read more button below.
6	monash	Master of Data Science	C6004	FIT5145 - Introduction to data science	FIT5145	Analyse the role of data in business (including ethical, cultural and managerial issues).I Apply basic tools for managing data (including data analysis and visualisation).I Apply basic tools for managing data (including data analysis and basic predictive modelling and data analysis methods).I Analyse and describe data processing requirements for a data science project.I Identify data resources and standards.	This unit looks at introduction to data studies, and simple tools to understand the many facets of working with data, and the role of data in Data Science over all. It will cover core task of Data Analysis. Working with the data as part of a business. Overview of the lifecycle in an organisation For more content click the Read more button below.
7	monash	Master of Data Science	C6004	FIT5147 - Data exploration and visualisation	FIT5147	Perform exploratory data analysis using a range of visualisation tools.I Describe the role of data visualisation in the data analysis (limitations).I Critically evaluate and interpret a data visualisation using well-known standard visualisations for qualitative, quantitative, textporal and spatial data.I Choose an appropriate data visualisation.I Implement static and interactive data visualisations using R and other tools.	This unit introduces students to statistical and visualisation techniques for the exploratory analysis of data. It will cover the role of data visualisation in data science (and its limitations. Visualisation of qualitative, quantitative, textporal and spatial data will be presented. What makes an effective data visualisation, principles of data visualisation, and creating them. For more content click the Read more button below.
8	monash	Master of Data Science	C6004	FIT5196 - Data wrangling	FIT5196	Parse data in the required format.I Assess the quality of data from various perspectives.I Resolve data quality issues ready for the data analysis.I Identify appropriate data sources for data enrichment.I Document the wrangling process for others to reproduce.I Write program scripts for data wrangling processes.	This unit introduces students to techniques for data wrangling. It will cover the problems that prevent raw data from being effectively used in analysis and first data exploration and pre-processing tasks that are necessary for analytics. These include, but not limited to, the handling of bad and missing data, data preparation and For more content click the Read more button below.
9	monash	Master of Data Science	C6004	FIT5197 - Statistical data modelling	FIT5197	Perform exploratory data analysis with descriptive statistics and visualisation.I Construct models for inferential statistical analysis.I Produce models for predictive statistical analysis.I Write computer programs for random sampling, simulation and hypothesis testing for required scenarios.I Implement the above techniques through programming and scripting.I Interpret results for a variety of models.	This unit explores the statistical foundations and methods that underlie the analytical aspects of Data Science (including Machine Learning). By working through practical examples, this unit gives a rigorous mathematical and statistical foundation to enable a thorough understanding of data science and machine learning. Students taught in later MDSIMH units will... For more content click the Read more button below.
10	monash	Master of Data Science	C6004	FIT5149 - Applied data analysis	FIT5149	Analyse data sets with a range of statistical, graphical and computational tools.I Evaluate the limitations, appropriateness and benefits of different analytic tools for given tasks.I Design solutions to real world problems with data, using analytic tools.I Assess the results of an analysis.I Communicate the results and findings for both specific and broad audiences.	This unit aims to provide students with practical statistical analytical and data modeling skills for the roles of a data analyst, business analyst, statistician. It will be introduced to established and contemporary Machine learning techniques for data analysis and presentation using widely used analysis software. They will be... For more content click the Read more button below.

Showing 1 to 10 of 298 entries

Previous 1 2 3 4 5 ... 30 Next

4 Employer Data

For employer's perception of Data Science, we decided to look at the job postings for Data Science relevant positions. We would have scraped career websites given more time. However, due to the circumstances, we found readily available data from Exploring 2 years' of Data Scientist Job Listings.

4.1 Data Science Job Postings

The data was scraped from Seek.com by Steve Condylios. The collected data contains **2,857** job posts and **52 variables**.

Exploratory data analysis was conducted exploring the salary and breakdown of Go8 emloyers. However they did not yield interesting results and thus put aside. For the purpose of this report, only 29 variables were looked at, including jobId, jobTitle, jobClassification, mobileAdTemplate, 25 programming languages. Of all the job posts, 535 are for senior or managerial positions, 25 for graduate positions, and the rest not specified in the job title.

Condylios also included Data Analyst job posting in the data set. 92 jobs are for Data Analyst and 82 jobs are labeled Data Analyst/Data Scientist. This is another interesting topic for comparison between Data Analyst and Data Scientist but for the scope of the project, we do not delve deeper into the data collection decisions.

The a fraction of the data set is made available below for exploration.

Show 10 entries   Search:

	jobId	jobTitle	jobClassification	mobileAdTemplate	R	Python	Matlab
1	38098375	Senior Method Development Immunoassay Scientist - Adelaide Location	Science & Technology	My client is a leading Australian owned and rapidly growing full service CRO based in South Australia, they pride themselves on being the only Australian CRO to offer a full suite of in house services for conducting early-phase clinical trials with efficacy. On offer for the right candidate is this newly created role and a very competitive salary and pension (In Australia known as: Supperanuation which is 9.5% of the base salary), Covered visa expenses + sponsorship, relocation assistance and compensation included too. Role / Responsibilities: Development, transfers and pre-validation of robust bioanalytical methods (mainly for New Biological Entities, Biosimilars, Anti Drug Antibodies, Neutralising Antibodies, and Biomarkers); that will meet the Bioanalytical Method Validation guidelines of the regulatory authorities (FDA, EMA) in a timely and cost-effective manner. Lead troubleshooting and technical investigations for bioanalytical laboratory methods with support from the Research and Development Manager and/or senior staff members. Assists Study Director and the bioanalytical laboratory team in method transfers as required. Write and Review Analytical Laboratory Methods. Timely and effective communication with clients and internal Bioanalytical Regulatory Group on project status in an environment with rapidly changing priorities. Support Operational Management to implement training programs for new staff. Contribute to internal strategic innovation meetings and initiatives. Escalate troubleshooting and other issues with assays to the Research and Development Manager that lead to additional activity outside of scope of work or significant increase in timelines. Perform scientific review of SOPs. Support the Research and Development Manager by following processes for managing R&D projects and escalation of any resourcing concerns as required. Support the bioanalytical laboratory team and Research and Development Manger by providing scientific input into assay development, troubleshooting and recent regulatory guidance. Support review of new projects as required. Adhere to OECD GLP regulations, guidelines and SOPs, as applicable. Departmental and Organisational Responsibilities Embrace the core values of the company and endeavor to display those attributes at all times with clients and staff alike. Commit to our training programs to ensure we are at the forefront of technology in the Asia Pacific region Commit to adhere to our quality systems. Remain current with required reading of SOPs relevant to the role and applicable duties. Other duties as directed by Line Management or delegate. Take reasonable care for your own health and safety and that of other personnel who may be affected by their conduct. These include general staff responsibilities and those additional responsibilities that apply for Managers, Team Leaders and other Personnel. Skills / Experience: Ordinary degree in the appropriate discipline. Minimum 4 years experience of working in an analytical laboratory (CRO experience highly desirable). Minimum 2 years experience in analytical method development, specifically ligand binding assays. High level of written and verbal communication skills. Proven ability to deliver tasks on time and with a high degree of accuracy. Demonstrate efficient problem solving and scientific data interpretation abilities. Demonstrated experience in communicating with clients on high level science issues. Proven ability to maintain strict adherence to client confidentiality issues. Ability to multitasking High attention to details. Demonstrated initiative, trust and reliability and the ability to work largely unsupervised. Experience gained in a regulated Bio-analytical laboratory desired.If you feel you have the right skills and experience for this position then please apply now, for further information please email: daniel.evans@hobsonprior.com.au	1	0	0
2	38259053	Pricing Data Scientist - Lower North Shore	Information & Communication Technology	Leading financial services organisation is hiring a Data Scientist to join their risk management team in St Leonards.The Data Scientist will be responsible for monitoring and developing risk based pricing models.Responsibilities include:Developing solutions that provide insights to make critical decisionsHands on coding and developing high quality modelsMonitoring Economic Capital models and reportingProvide analytical expertise and supportSkills required:Pricing optimisation, GLM and GBM or other decision tree modelling experienceExperience applying statistical modelling, machine learning, deep learning and data mining to commercial problemsExtensive experience in “R” and SAS for data extraction and manipulationExcellent communication and presentation skillsPlease hit apply below or for more information contact Alexa Cauvin @ Randstad Technologies on 02 8235 3326.At Randstad, we are passionate about providing equal employment opportunities and embracing diversity to the benefit of all. We actively encourage applications from any background.	1	0	0
3	38267899	Insights Analyst – Online Video Streaming	Information & Communication Technology	The Company:A big brand with a reputation for leading the way with Big Data is looking for an Insights Analyst who is passionate about using the latest Data Analysis tools in the market to deliver business critical insights.This company is investing significantly into their digital media steaming capabilities, with a huge project environment and a push to be a market leader in this very competitive and fast-moving space to work in. Your analysis will play a key part in the strategic direction of this company’s flagship products and services! Hit the Apply for this job button now to get more information! The Role: Working with data sets generated from digital streaming behavior, your job will be to create unique insights that can be used by a range of teams to improve the product performance and UX.We are looking for a passionate Data Analyst with experience working in media / broadcasting or an enterprise environment on large data sets. This role will give you the freedom to utilise a range of technologies in order to deliver these insights.We are ideally seeking people with experience using advanced analytics tools and strong coding ability using the R programming language.You will not only run deep dive analysis with a fresh approach but also produce visualisations to communicate your insights clearly to senior leaders. Essential Skills:+3 years working as an Insights Analyst, Data Analyst or Data Scientist.SQLCommercial coding experience - R programming language skills essentialData Visualisation – Tableau, Power BI etcStrong communication and presentation skills.Ideally a background in TV, Media, Online Streaming, Digital.Relevant Bachelor’s degree – Maths, Stats etc What’s in it for you:This is one of the busiest and most talked about organisations in the country right now. The project has a huge scope for the coming years! You'll be joining a large Agile team with a manager that believes in continual development. This team has excellent staff retention rates and the manager looks for people that want to grow with the company for years to come. They have a reputation for being a great place to work where you can benefit from exceptional career growth opportunities and internal promotion. With an awesome site, they have a gym, onsite restaurants, bars and cafes. This is a company that genuinely promotes flexibility and work/life balance, really valuing their employees. Hit the ‘Apply for this job’ button now to get more information or call Rebecca on +61 2 9993 1071.	1	0	0
4	38267995	Credit Risk Analyst	Banking & Financial Services	Responsibilities: Design and development of scorecards and models across the credit life cycle (e.g. application, behaviour, fraud and collection scorecards) for the retail credit portfolioReview and update portfolio risk monitoring models and assist in the reporting of portfolio riskContribute to the Financial Planning processes e.g. loss & IFRS9 compliant reserve forecasting, recovery forecasting, capacity planning etc.Development and monitoring of credit scoring, risk segmentation, predictive models and other modelling techniques to enhance credit decision making as well as optimize acquisition, collections and recovery activitiesPartner with Credit Risk and Collection Teams to support credit strategies through account level performance based analytics and requirementsDesign of appropriate champion/challenger tests as well as preparing policy options/changes/recommendations based on analytics and/or regulatory Requirements:Educated to a degree level in data science/quantitative disciplines such as statistics, finance, engineering, mathematics etcStrong experience in statistical model building is essentialStrong working knowledge of SAS, R, SQL, Excel Sthree Australia is acting as an Employment Agency in relation to this vacancy.	1	0	0
5	38269206	Data Scientist	Information & Communication Technology	• Large, high profile organisation• Unique Project• Senior roleThe RoleOur client, a large Sydney based telecommunications organisation, is looking for an experienced Data Scientist with coding skills.Who you areYou will have the following skills/experience• Proven experience within Data Science and Analytics • Coding skills in SQL- Tableau- Python • Industry experience in a Telco or large enterprise environment • Scoping and defining business problems with stakeholders• Preparing data for analysis including loading, cleaning, manipulating and exporting• Analysing data including data exploration, modelling and testing• Preparing and presenting results including developing visualisations , dashboards and recommendations for actionWhat’s on offerThis is a 12 months + engagement with a high profile organisation with highly competitive remuneration on offer.Who we areFinXL IT Professional Services is an established, innovative Australian company, providing technology enabled business solutions and consulting services across a number of industries. These include; Government, Telecommunications, Commercial and Finance. We currently have more than 1200 consultants assisting our clients across Australia and are continuing to grow at a rapid pace. Through the delivery of services and implementation of new systems, processes and technology, we assist our clients to gain competitive advantage and reach new heights.What’s nextIf you would like to be considered, please send applications to julial@finxl.com.au or contact Julia Leung on 02 8243 6808 to discuss further.	0	1	0

Showing 1 to 5 of 5 entries   Previous 1 Next

Part II

Text Analysis

To find out what is data science from universities' and employers' perspectives, whether there is a shared structure or common skill set of Master of Data Science degrees offered at Go8, what employers expect from a data scientist in the workplace, an exploratory data analysis, in particular text analysis has been conducted.

For universities, the main purpose is to identify shared skills or concepts offered by Master of Data Science degrees through exploring faculty of the units, detailed teaching contents etc. Whereas for employer data, the focus is to extract information regarding skills and programming languages in demand.

5 Unit Text Analysis

5.1 Unit Code Analysis

To explore the teaching contents of Master of Data Science at Go8, an analysis based on faculty of units offered is conducted to see what components are included in this degree.

Unfortunately faculty information is not directly available on the unit handbooks, in this case, unit code is taken as a surrogate identification. As shown in the sample data below, unit code is a combination of letters and numbers, the first few characters such as FIT, MAT, usually represents the faculty this unit belongs to, we could then make relatively educated assumptions on the content of the unit.

The grouping was performed manually using the code listed below. It is a subjective choice made under careful considerations, we are aware that the grouping is not 100% accurate. For example, the code 'DATA' from University of Sydney is classified under IT, however, some of the units start with DATA are actually more related with statistics, which means 'DATA' belongs to multiple departments. Although there would be misclassifieds units, the results could still provide a meaningful guidance regarding the teaching components of Master of Data Science at Go8 universities.

```
math <- c("STAT", "MATH", "MATHS", "STATS", "MAT", "MAST", "ACTL", "QBUS")
it <- c("COMP", "FIT", "CITS", "INFS", "COSC", "CSSE", "CSYS", "EDPC", "INMT", "PHIL", "PH
commerce <- c("ECON", "FINS", "MARK", "ACCT", "FINM", "MGMT", "MKTG")
spatial <- c("GEOM", "ITLS")
science <- c("EDUC", "SCIE", "SOCR")
health <- c("BINF", "BMS", "HTIN", "PUBH")
```

It is clear from Figure **??** that IT and Stat/Math are the two dominating components in the Master of Data Science degrees at Go8. Most units (165 out of 298) fall under the IT faculty, followed by Math and Stats, which has 79 units.

Similar findings could be observed from some but not all Go8 universities. Figure **??** shows the faculty breakdown by university. Since the total number of units offered by each university is different, instead of showing the actual number, proportions are plotted to make better comparisons across universities.

School	Course	Unit
monash	Master of Data Science	FIT9132 - Introduction to databases
monash	Master of Data Science	FIT9136 - Algorithms and programming
monash	Master of Data Science	FIT9137 - Introduction to computer arch
monash	Master of Data Science	MAT9004 - Mathematical foundations fo
monash	Master of Data Science	FIT5125 - IT research methods
monash	Master of Data Science	FIT5145 - Introduction to data science
monash	Master of Data Science	FIT5147 - Data exploration and visualis
monash	Master of Data Science	FIT5196 - Data wrangling
monash	Master of Data Science	FIT5197 - Statistical data modelling
monash	Master of Data Science	FIT5149 - Applied data analysis
monash	Master of Data Science	FIT5201 - Machine learning
monash	Master of Data Science	FIT5202 - Data processing for big data
monash	Master of Data Science	FIT5205 - Data in society
monash	Master of Data Science	FIT5212 - Data analysis for semi-structu
monash	Master of Data Science	FIT5230 - Malicious AI
monash	Master of Data Science	BMS5021 - Introduction to Bioinformati
monash	Master of Data Science	BMS5022 - Advanced bioinformatics: eff
monash	Master of Data Science	FIT5126 - Masters thesis part 1
monash	Master of Data Science	FIT5127 - Masters thesis part 2
monash	Master of Data Science	FIT5228 - Masters thesis part 3
monash	Master of Data Science	FIT5229 - Masters thesis final
monash	Master of Data Science	FIT5120 - Industry experience studio pr
monash	Master of Data Science	FIT5122 - Professional practice
unimelb	Master of Data Science	Methods of Mathematical Statistics
unimelb	Master of Data Science	A First Course In Statistical Learning
unimelb	Master of Data Science	Programming and Software Development
unimelb	Master of Data Science	Algorithms and Complexity
unimelb	Master of Data Science	Elements of Data Processing
unimelb	Master of Data Science	Database Systems & Information Model
unimelb	Master of Data Science	Statistical Modelling for Data Science
unimelb	Master of Data Science	Multivariate Statistics for Data Science
unimelb	Master of Data Science	Computational Statistics & Data Science
unimelb	Master of Data Science	Cluster and Cloud Computing
unimelb	Master of Data Science	Advanced Database Systems
unimelb	Master of Data Science	Statistical Machine Learning
unimelb	Master of Data Science	Data Science Project Pt1
unimelb	Master of Data Science	Data Science Project Pt2
unimelb	Master of Data Science	Data Science Research Project Pt1
unimelb	Master of Data Science	Data Science Research Project Pt2
unimelb	Master of Data Science	Foundations of Spatial Information
unimelb	Master of Data Science	Spatial Databases
unimelb	Master of Data Science	Spatial Analysis
unimelb	Master of Data Science	Information Visualisation
unimelb	Master of Data Science	Analysis of High-Dimensional Data
unimelb	Master of Data Science	Advanced Statistical Modelling
unimelb	Master of Data Science	Mathematics of Risk
unimelb	Master of Data Science	Optimisation for Industry
unimelb	Master of Data Science	Practice of Statistics & Data Science
unimelb	Master of Data Science	Stochastic Calculus with Applications
unimelb	Master of Data Science	Advanced Probability
unimelb	Master of Data Science	Random Processes

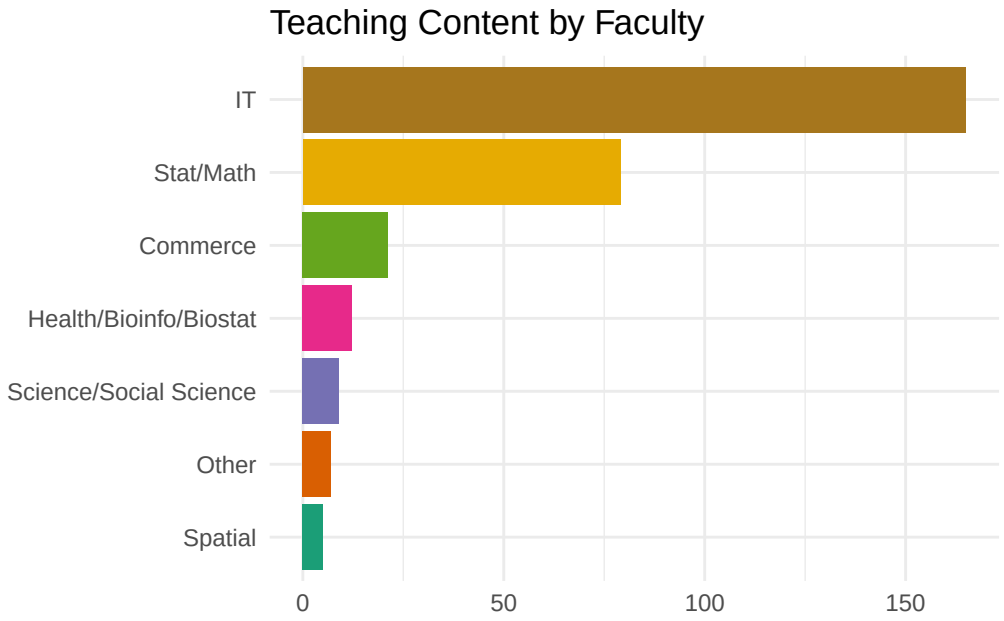

Figure 5.1: Teaching Content by Faculty

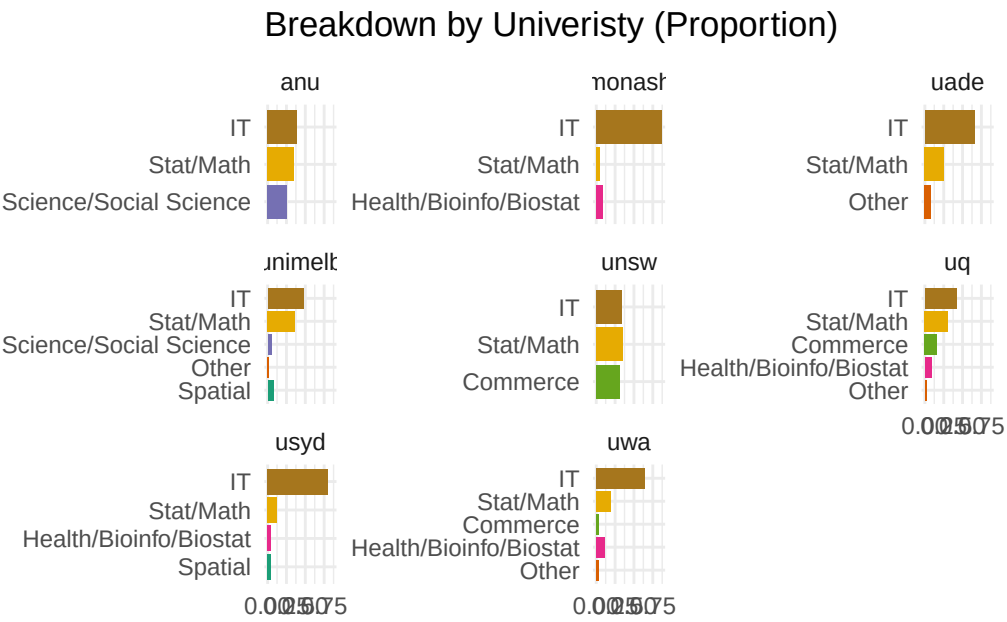

Figure 5.2: Faculty Breakdown by University (Proportion)

At Monash University (monash), University of Adelaide (Uuade), University of Sydney (usyd) and University of Western Australia (uwa), IT is apparently dominating the Master of Data Science degree, especially at Monash University. However, University of Melbourne (unimelb) and University of Queensland (uq) offers relatively higher proportion of statistical and mathematical (Stat/Math) units, whereas units offered at the Australian National University (anu) and UNSW Sydney (unsw) are more evenly distributed across IT, Stat/Math, Science/Social Science and Commerce respectively.

Based on the findings above, it seems that there is a shared structure across Go8 that Master of Data Science is a IT based, computational degree, but the proportion it occupies varies by universities. Monash University tends to be heavily focused on IT and computational aspects, whereas the Master of Data Science degree at UNSW Sydney and ANU are more balanced across IT, statistics and math, as well as science and commerce.

5.2 Unit Overview and Learning Outcome - Bigram

After having a rough idea of the bigger picture, we then moved to explore what exactly are the teaching contents. Single word analysis, bigram and trigram are all produced in order to identify the frequently mentioned words and/or terms. `distinct` function is applied to unit and term, so that same terms are only counted once in a unit to avoid duplicated counting. Words and terms such as 'student', 'successful completion' add more noises than values to the results, are removed in the pre-processing step.

The bigram, which is Figure **??** below provides the most informative results among all. Machine learning appears quite often, as well as software development, linear models, statistical analysis, spatial data. It seems that these frequently mentioned terms are associated with both computational and statistical concepts and skills, which aligns with the findings from the unit code analysis in previous section.

Unfortunately, due to the limited number of observations in the collected data set, the count for each term is too low to make meaningful interpretations or justifications. In addition, similar terms such as research findings, research designs and research literature are supposed to be grouped and counted together, but are not in the bigram. This issue is later solved by introducing the text2vec technique for natural language processing, which will be discussed in the next chapter.

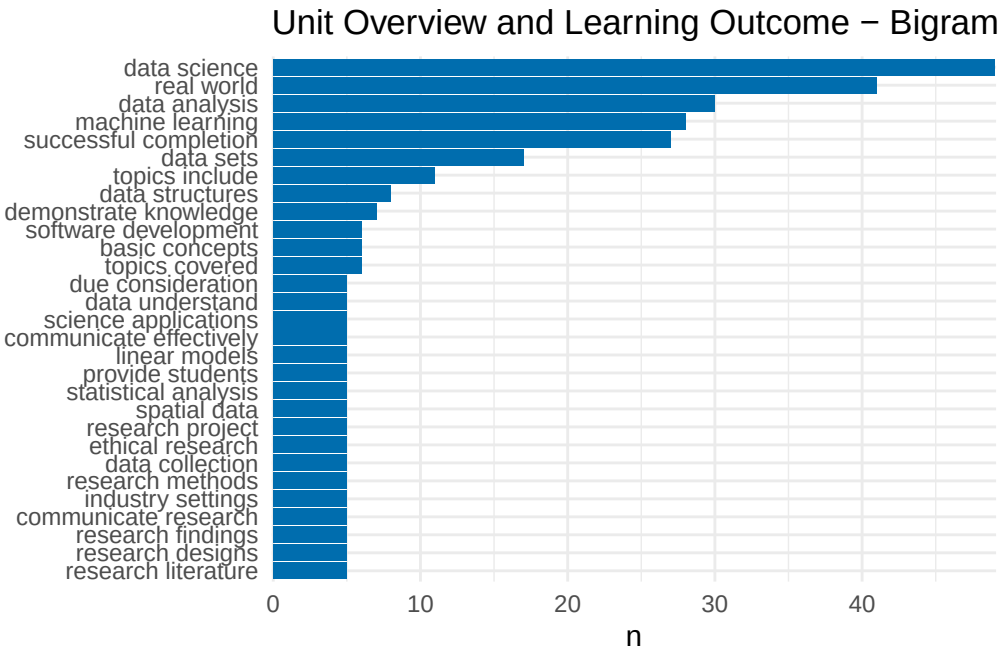

Figure 5.3: Unit Overview and Learning Outcome - Bigram (n > 4)

6 Job Text Analysis

To explore the skills and programming languages in demand from employers, we focused on the mobileAdTemplate and the 25 columns of programming languages. Similar to before, single word analysis, bigram, and trigram are all produced.

6.1 Word Frequency

Figure 6.1: Job Word Frequency (freq >200)

Word frequency was calculated using the same method as Section **??**. Programming languages Python and SQL seems prominent. Potentially due to the amount of senior positions in the data, *experience* is mentioned a lot. In terms of other knowledge or skills, *statistics*, *modelling*, *analysis* are some terms that seems to be standing out. To ensure frequency is meaningful, we looked at bigram and trigram.

6.2 Bigram

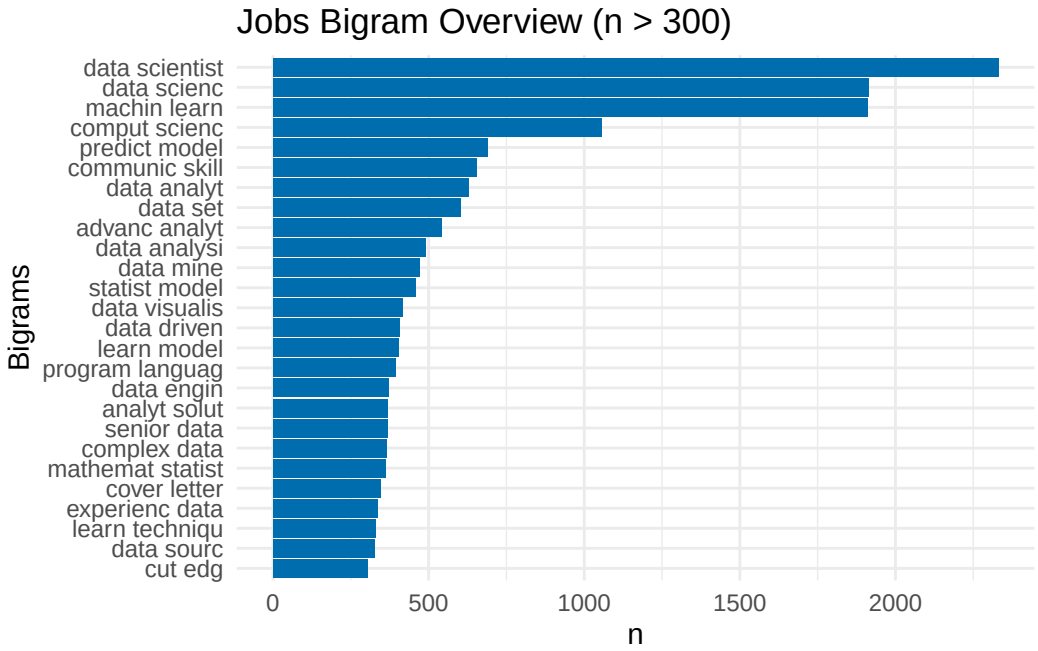

Figure 6.2: Job Description - Bigram (n > 300

For Figure **??**, we stemmed the words before joining them into bigram in attempt to avoid under-counting. However, from the figure we can still see that some terms like *data analyt* and *data analysi* are still counted separately. From this bigram, the popular skills or knowledge mentioned are *machin learn*, *predict model*, *communic skill*, *data analyt* and *data mine*. Mathematical skill, *mathemat statist* is also mentioned quite often. Some of these terms on the list are vague and can mean be grouped together. Trigram yeilded similar results as the bigram. With the same problem of under-counting the n-grams due to insufficient groupings.

6.3 Programming Languages

Figure **??** shows the count of each language mentioned by employers. 89% of the jobs mentioned Python, 69% mentioned R and 68% mentioned SQL. Tableau's popularity is not a surprise given the high frequency for *data visualis* in the previous section.

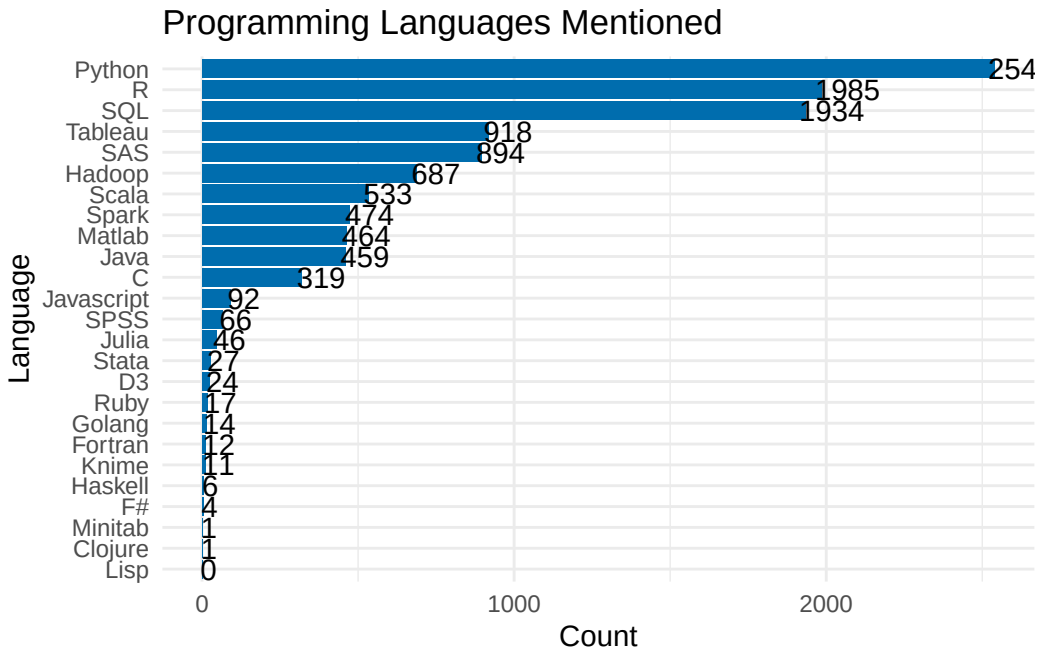

Figure 6.3: Programming Skills Mentioned by Employers

Part III

text2vec

As mentioned in Section Section **??**, since the collected university data is relatively small, to make more educated and meaningful interpretations, similar words shall be grouped together and counted by groups. This is usually computed using text corpus, which is a language resource consisting of a large and structured set of texts, since data science is a new term waiting to be defined, there is no available text corpus on this topic. Therefore, we adopted the concept of word2vec and tried to build our own text corpus.

There are multiple publicly available models and packages to conduct similar computations, however, each model takes hours to fit. Due to time constrains, we have only fitted the Dirichlet Allocation (LDA) model with a few parameter adjustments using the `text2vec` package with the concepts illustrated by Das (2016).

Algorithm and Model Fitting

According to Das (2016), the algorithm behind the LDA model is to convert words to document-term matrix (DTM), where the rows, columns and entries correspond to documents, terms and counts respectively. LDA then fits a probabilistic model that assumes a mixture of latent topics, where each topic has a multinomial distribution for all words. The number of topics is a parameter that could be adjusted by needs.

The initial code to build the LDA model was provided by Professor Tanaka, the major part of the code to build the first version of LDA model is also provided below.

```
list(
  tar_target(wiki_stats, get_wiki_articles("https://en.wikipedia.org/wiki/List_of_statisti
  tar_target(wiki_sociology, get_wiki_articles("https://en.wikipedia.org/wiki/Index_of_soc
  tar_target(wiki_computing, get_wiki_articles("https://en.wikipedia.org/wiki/Index_of_com
  tar_target(clean_wiki_stats, map(wiki_stats, clean_wiki_article), format = "rds", reposi
  tar_target(clean_wiki_sociology, map(wiki_sociology, clean_wiki_article), format = "rds"
  tar_target(clean_wiki_computing, map(wiki_computing, clean_wiki_article), format = "rds"
  tar_target(clean_stats, preprocess_text(clean_wiki_stats)),

  tar_target(clean_ssc, preprocess_text(c(clean_wiki_stats, clean_wiki_sociology, clean_wi

  tar_target(itoken_ssc, itoken(clean_ssc, tokenizer = stem_tokenizer),
             cue = tar_cue(mode = "thorough")),

  tar_target(vocab_ssc, create_vocabulary(itoken_ssc, ngram = c(1, 3), stopwords = stopwor
             cue = tar_cue(mode = "thorough")),
  tar_target(vocab_ssc_prune, prune_vocab(vocab_ssc, n_min = 40),
             cue = tar_cue(mode = "thorough")),
```

```
tar_target(dtm_ssc, create_dtm(itoken_ssc, vocab_vectorizer(vocab_ssc_prune)),
           cue = tar_cue(mode = "thorough")),
tar_target(tcm_ssc, create_tcm(itoken_ssc, vocab_vectorizer(vocab_ssc_prune),
                               skip_grams_window = 5L),
           cue = tar_cue(mode = "thorough")),
tar_target(word2vec_model_ssc, model_glove(vocab_ssc_prune, tcm_ssc),
           cue = tar_cue(mode = "thorough")),
tar_target(word2vec_dist_ssc, dist2(t(word2vec_model_ssc$components), method = "cosine")
           cue = tar_cue(mode = "thorough"), format = "rds", repository = "local"),
tar_target(word2vec_res, find_close_words("statistics", word2vec_dist_ssc, 10),
           cue = tar_cue(mode = "thorough")),
tar_target(lda_model03_ssc, model_lda(dtm_ssc, ntopics = 3),
           format = "rds", repository = "local"),
tar_target(lda_model20_ssc, model_lda(dtm_ssc, ntopics = 20),
           format = "rds", repository = "local")
           )
```

The model must be trained before it could be used, we web scraped over 4448 Wikipedia articles as training data, including 2816 articles in statistics, 1005 articles in sociology and 627 in computing. The functions used in the codes above such as `get_wiki_articles`,`clean_wiki_article`, `get_clean_combined_wikis`, `model_lda`, `preprocess_text`, `stem_tokenizer`, `prune_vocab`, `model_glove` and `find_close_words` are constructed by Professor Tanaka for pre-processing purposes, the original scripts could be found from the project repository.

Model Adjustments

We have tested using different values for parameter `ntopics` and tried out training the LDA models with different combinations of data.

The results provided differs from models, Figure **??** compares the results produced by the full model and model without sociology data on ten topics.

From the results computed by the full model, Topics 9, 10, 2 and 8 occupies relatively higher proportion compare with the others, but the order varies across universities, and their proportions are not significantly larger than the rest of other topics, makes it hard to draw meaningful interpretations. On the right hand side, results from the model without sociology data demonstrates a better picture: Topics 6, 8 and 9 in this case are the top 3 topics across all Go8 universities, however, proportions of Topic 10, 5, 7, 3 and 4 are also obvious higher in some of the universities, brings in difficulties to make justifications.

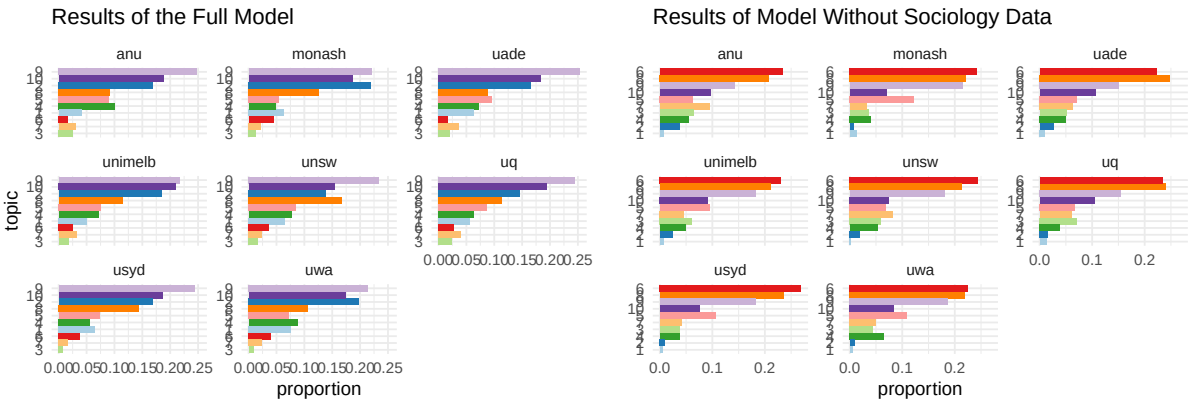

Figure 6.4: Compare Results Between Full Model and Without Sociology Data

As sociology data tends to brings in noises to the model, and is not closely relevant to the data science topic compare with statistics and computing, another two models are fitted using only statistics data and computing data respectively, the results of both models are shown in Figure **??**.

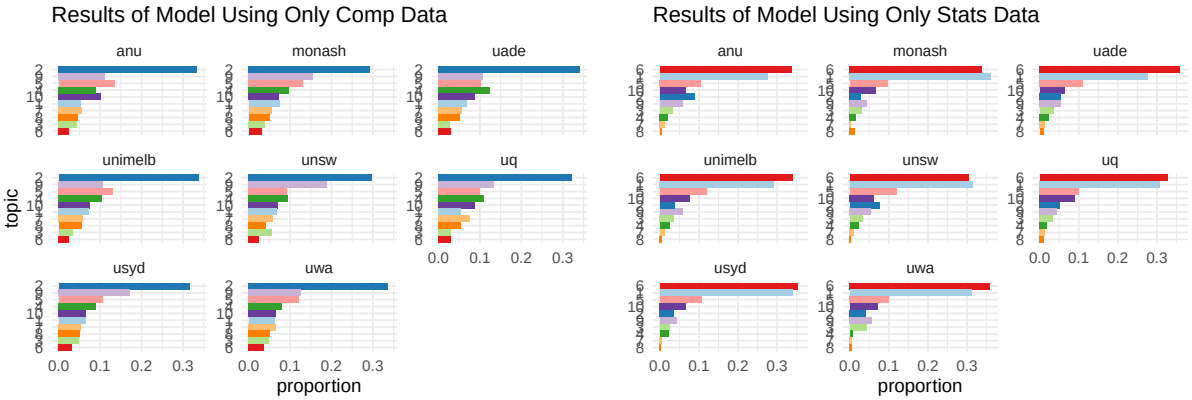

Figure 6.5: Compare Results Between Full Model and Without Sociology Data

Topic 2 is the only dominating topic based on the results provided by the model using only computing data, which provides a clearer picture than the previous models. The table below listed the top ten words for each topics, it turns out Topic 2 contains words like comput (computation etc.), system, program, softwar (software), which are associated with computational aspects, especially software. This model provides a more meaningful results than the prior ones, however, there is not much interpretations could be made for the other topics, the information it offers is still not very satisfying.

In terms of the model trained by only data in statistics, there are also dominating topics across all eight universities: Topics 6 and 1, besides, topics 5, 10, 2, 9 together took a relatively higher proportion compare with the rest of other topics. Both models using only computing

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9
window	comput	ibm	algorithm	network	bit	format	intel	softwar
system	system	comput	can	use	instruct	use	chip	compani
version	program	system	number	can	memori	imag	design	appl
releas	use	disk	function	web	use	digit	processor	free
support	machin	drive	set	data	address	can	bit	use
user	design	use	state	internet	regist	video	core	also
oper_system	process	control	languag	secur	processor	standard	mhz	develop
oper	develop	machin	use	access	oper	disc	use	open
os	inform	card	symbol	protocol	can	data	introduc	sourc
file	softwar	unit	problem	link	data	file	microprocessor	user

or statistical data delivers better results, model trained by only statistical data provides more information than the other, hence is selected to use for further analysis on our university data.

Note that it requires highly skilled linguists and huge efforts to establish a proper text corpus, the model we built is still fairly basic and could be further optimised by adjustments.

7 Apply the Selected Model to Collected University Data

Before applying the fitted LDA model to our university data set, words from unit overview and learning outcomes are stemmed using the `SnowballC` package, so that noises like plurals and part of speech are removed. The stemmed words are then assigned to the corresponding topic with the highest probability, instead of counting the appearance of words, the new counts generated are based on topics.

Similar with the university breakdown in Section **??**, to make more objective comparisons, counts are converted to proportions due to different number of units scraped for the eight universities. Figure **??** suggests that Topics 1 and 6 are obviously the dominating topics in Master of data science at all Go8 universities, whereas Topics 2, 5, 9, 10 together also occupies a relatively large proportion.

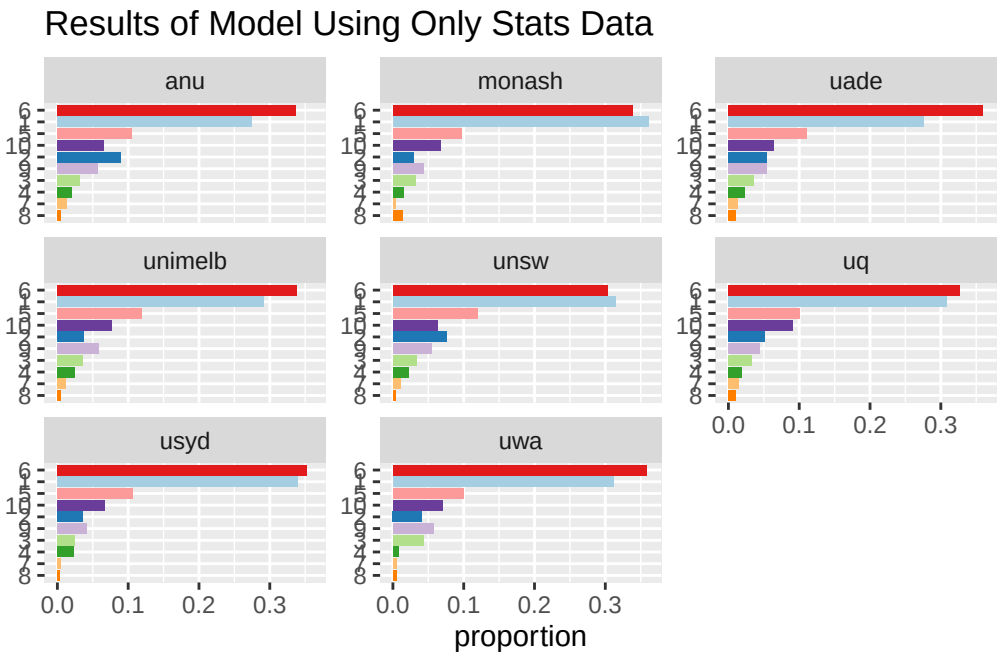

Figure 7.1: Topics Proportion by University

The top ten words based on probabilities for each of the ten topics are provided below, colours