UNIVERSITY OF TORONTO

MIE 1624 PROJECT REPORT

Wenqing Wang 1004551362
Xihao Liu 1003463682
Yifan Shao 1004534435
Yichun Cai 1003911485
Zhihao Zhang 1004131628
Xiafei Luo 1004675336

Table of Contents

Table of Figures

Table of Tables

1. Summary of Data

In this project, we used three types of data: job information recruitment sites (www.indeed.ca and www.glassdoor.com), curriculum information from the 14 data science graduate programs in North America, and survey data from Kaggle ML and Data Science Survey, 2017.

1.1 Job information

We collect job information from indeed.ca and glassdoor.com by web scraping. We have collected job titles, job descriptions and recruitment companies. The keywords used in the query are: data analyst / data scientist / data engineer (for Technical Program in part 2), data manager / business analyst / ai system designer (for Business Program in part 3). This part of the data includes more than 8,000 recruitment information and the data needs data cleansing to get the job description keywords used for analysis.

1.2 Course Information

We refer to the top 30 North American data science graduate programs and have taken 14 representative programs. The collected course information includes the title of the course and the name of the project. The title of the course is the most general description of the content of the course and there are few noise words. This part of the data includes more than 300 courses and also requires data cleaning.

1.3 Kaggle ML and Data Science Survey

Survey data is multiple choice responses about question on data science. There are 16717 responses from users in kaggle.com. Since this part of the data is a multiple choice, this part of the data does not require data cleaning

The survey data is analyzed from following perspective: the top jobs in data science, the most relevant topics, the most required skills in data science, the top software's and platforms used in data sciences. The most frequently mentioned keywords from the data science's survey are shown in **Appendix A**.

2. MIE1624 Course Curriculum Design

2.1 Background

MIE1624 course will be redesigned as an introduction type course that aims to any student with math background and want to get general and basic knowledge about data science. The course will be the entry-level requirement for both our Technical Data Science program and Managerial and Business Data Science program.

This course teaches about data science basic concepts and related math concepts. It contains training about using data science software, and the hot and trending ML topics are also included. This course composed by two parts which are lectures and optional tutorials. By finishing this course, students understand the fundamentals of data science, and they are able to apply the basic data science knowledge and skills learned in this course works.

2.2 Course Topic Selection

By reading articles from various data science professionals, we have found out that if a person wants to be a successful data analyst, technical skills are important while soft skills along with some other skills, such as excel, are equally important. Thus, we divided the key words extracted from job websites and Kaggle dataset into three groups (**Appendix B**). By using k-means method, keywords extracted from Kaggle and Job websites are divided into three parts according to their importance. The three categories can be seen intuitively from Figure 1. Skills belonging to Category 2 in red refers to the most fundamental skills, which must be taught in MIE 1624. Skills classified in Category 1 are important varies from programs. The courses corresponding to these skills would be arranged in part 3 and part 4. Courses related to the skills in category 0 in grey would be elective courses according to students' interest.

As the most basic course, MIE 1624 aims at imparting the most important skills for students who want to work on data analysis in the future. The redesigned Syllabus of the course is in Appendix.

3. Master of Data Science and Analytics (M.D.S.A) Program Curriculum Design

The M.D.S.A program and M.B.A.I program curriculums are designed by using the same algorithm:

1. **Occupation analysis**. According to our analysis, the six most popular jobs in data analysis are AI system designer, Business Analyst, Data Analytic, Data Engineer, Data Manager and Data Scientist.
2. **Skills mining**. Find out the key skills needed for each job **(Appendix D)** and rank the importance of skills for each program (**Appendix E**). According to our analysis of the program, we found the most important seven skills: programming, statistics, machine learning, algorithms, data visualization, business management and soft skills. Due to the fact that each program has different emphasis on skills, this will be reflected in the curriculum arrangement in the following step.
3. **Keywords for each skill**. By analyzing other school requirements for each program and the high frequency keyword extracted from Kaggle and job dataset (**Appendix B**) and combining our group members' knowledge of each skill, we added keywords to each skill in Python.
4. **Courses classification**. Courses Analysis courses from 14 different Master programs in Data science or management and business analytics offered by universities in North America. By analyzing the keywords appearing in course title, each of them corresponds to one or more skills set in step 2. This allows them to be categorized into seven groups of courses satisfying corresponding skill sets.

5. **Courses grading**. Since each skill have several courses corresponding to it, courses are divided into three sections based on their difficulties. Titles with keyword ‘introduction’ are automatically classified as simple courses while those with keyword ‘advanced’ are advanced courses. Also, courses contain more than one topic is considered to be more advanced. The classification of other courses is solved by k-means method. This would allow courses to be arranged in proper sequence.
6. **Curriculum Customization**. Then we use keywords for each skill obtained from step 3 to connect to course category, therefore we obtain the importance level of each course category toward each job title. Since the job titles are related to M.D.S.A program and M.B.A.I program, we can find the level of importance of course category toward each program. The result is shown in **Appendix E**. We can determine what and how many courses are required in the new program. The course category with largest blue dot, are required course. Course category with “somewhat important” will require one course. Course categories with “not that important” dots are not required by the program.

The proposed M.D.S.A program has following curriculum: the program can be completed by full time student in 4 semester time or 16 months. A student has to take 10 half course credit worth of courses plus a practicum capstone project in the last semester to graduate. The students are required to take 5 required core courses, and 4 of the elective courses have to be chosen from business management, machine learning, algorithm and data visualization. The detailed composition of the courses is shown in the table below. Detailed course curriculum can be found in (**Appendix H**). The sequence of courses example can be found in **Appendix F**.

5 required	1 required	1 required	1 required	1 required	1 free elective	1 required
Intro to Data Science Cloud Computing Data Wrangling Statistics Business Communication & Analytics Consulting	Data Visualization Data Visualization II	Machine Learning for Data Science Probabilistic Graphical Models Regression Deep Learning Predictive Modeling Statistical Learning Data Mining Foundations of A.I. Natural Language processing Statistical Learning	Foundations of Data Management Analytics of Finance Project Management Operation Management Marketing Analytics	Discrete Math and Algorithm Analysis of Algorithms Algorithms and Data Structures	Available courses in Appendix F can be selected	Capstone Project

Table 1: Course Selection of M.D.S.A. program

4. Master of Management and Business in Analytics and A.I. program Curriculum Design

Use the steps mentioned in section 3, we developed the curriculum for M.B.A.I program. The program can be completed by full time student in 4 semester time or 16 months. A student has to take 10 courses plus a practicum capstone project at the last semester to graduate. For Master of Management and Business in Analytics and A.I., the students are required to finish 4 required core courses and 6 elective courses. 4 out of the 6 elective course have to cover topics in statistics, machine learning, business and management and algorithms. The composition of the courses is shown in the table below.

Detailed course curriculum can be found in (**Appendix H**). The sequence of courses example can be found in **Appendix G**.

5 required	1 required	1 required	1 required	1 required	2 free electives	1 required
Intro to Data Science Foundations of A.I. Business Communication & Analytics Consulting Knowledge, Visualization and Communication	Probability Statistics Advanced Optimization Modern Applied Statistics	Machine Learning for Data Science Probabilistic Graphical Models Regression Deep Learning Predictive Modeling Statistical Learning Data Mining Foundations of A.I. Natural Language processing Statistical Learning	Foundations of Data Management Analytics of Finance Project Management Operation Management Marketing Analytics	Discrete Math and Algorithm Analysis of Algorithms Algorithms and Data Structures	Available courses in Appendix H can be selected	Capstone Project

Table 2: Course Selection of M.B.A.I program

5. Data Science Education EdTech

Assume we established EdTech startup and aim to develop Data Science education. The main service of EdTech is online education, the website provides basic courses and a platform for connect potential employers and employees to connect. The types of users our education website is trying to attract are:

- People who have no idea about data science
- People who want to improve their skills in data science area
- Potential employers and employees in data science area

We propose following questions to ourselves:

1. How to help people to know data science and improve their technical skills?
2. How to connect employers with users?

We answer questions proposed above with following answers:

1. According to data from job information which was analyzed in Part 1, the top 10 entry level skills people needs to know in data science area are showed in table below.

Keyword	Kaggle Word Frequency	Job Word Frequency	Class
Python	1800	2571	2
R	1424	2197	2
SQL	1130	2413	2
Jupyter notebooks	945	41	1
TensorFlow	758	432	1
C/C++	610	1022	0
MATLAB/Octave	561	314	1
Unix shell / awk	545	277	1
Amazon Web services	505	576	1
Java	499	1202	0
NoSQL	457	434	1

Table 3: Word Frequency and Class for Top 10 Keywords

Therefore, we provide those online courses for users. We will provide practice sets and discussion board. Users can also upload their problem sets and may provide rewards for it if they want.
According to K-Means Algorithm, the higher-level class skills are more popular and important than lower level class skills. People will more interested in those level 2 class courses. Thus, we will provide more about those skills to attract more users.

2. Students who enrolled in data science programs can find their project topic in our websites, because potential employers are encouraged to provide their own question and real datasets for users. This can not only help students to solve realistic problems and finish their programs, but also get the references when they apply those companies. In addition, users can upload their resumes on websites directly for target companies.

Appendix A: Keywords Extracted from Kaggle and Job Websites

	keyword	kaggle_word_frequency	job_word_frequency	k_means
0	Python	1800	2571	2
1	R	1424	2197	2
2	SQL	1130	2413	2
3	Jupyter notebooks	945	41	1
4	TensorFlow	758	432	1
5	C/C++	610	1022	0
6	MATLAB/Octave	561	314	1
7	Unix shell / awk	545	277	1
8	Amazon Web services	505	576	1
9	Java	499	1202	0
10	NoSQL	457	434	1
11	Tableau	412	999	0
12	Spark / MLlib	360	1183	0
13	Hadoop/Hive/Pig	354	2067	0
14	Microsoft Excel Data Mining	332	1895	0
15	SAS Base	198	1020	0
16	IBM SPSS Statistics	177	290	1
17	Microsoft Azure Machine Learning	176	325	1
18	Perl	133	170	1
19	IBM Watson / Waton Analytics	81	43	1

Table 4: Keyword Frequencies in Kaggle and Job Dataset

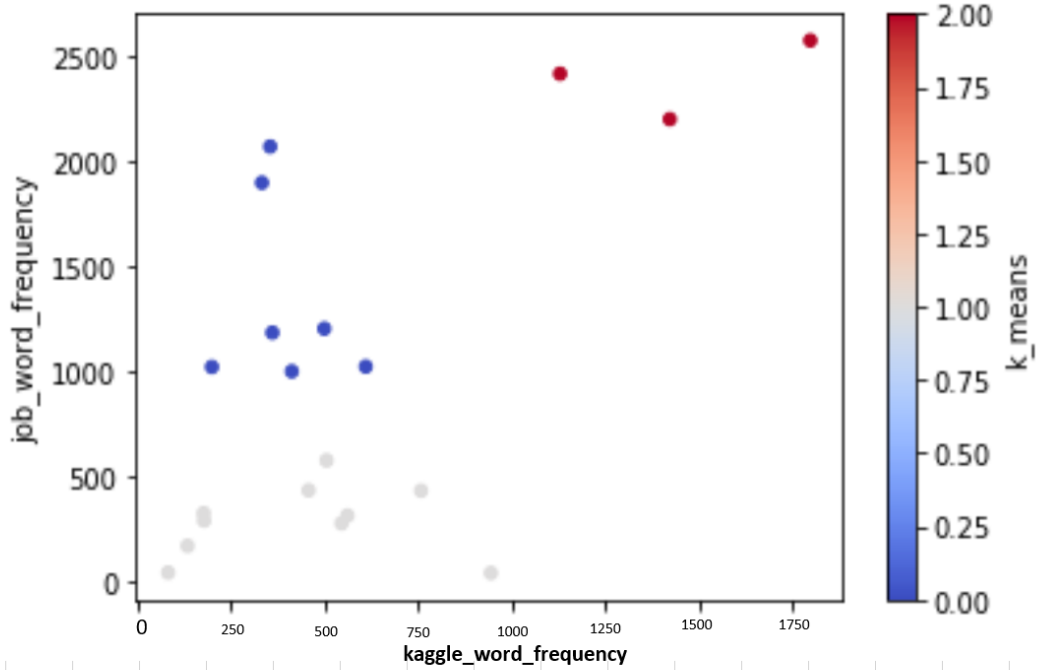
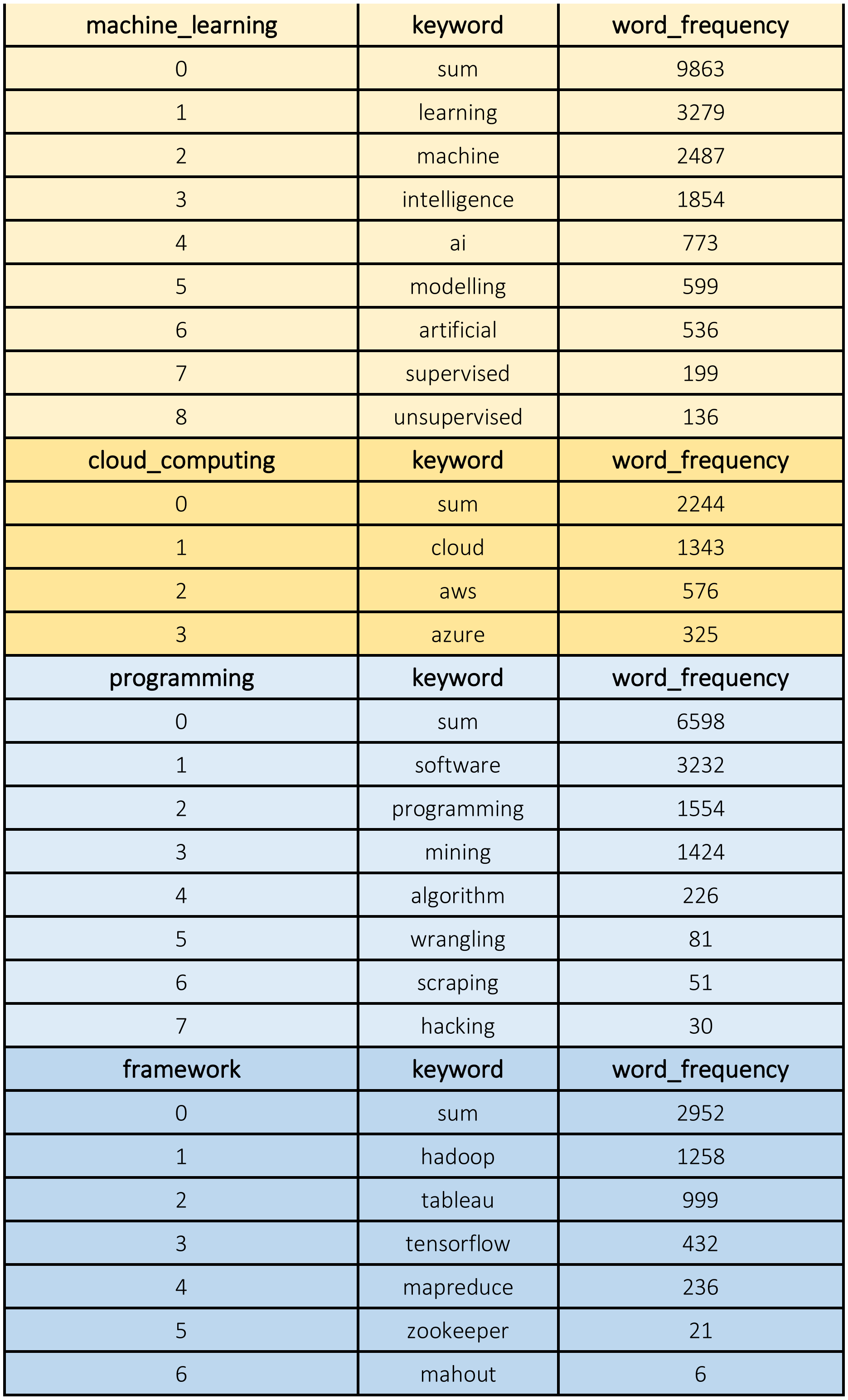


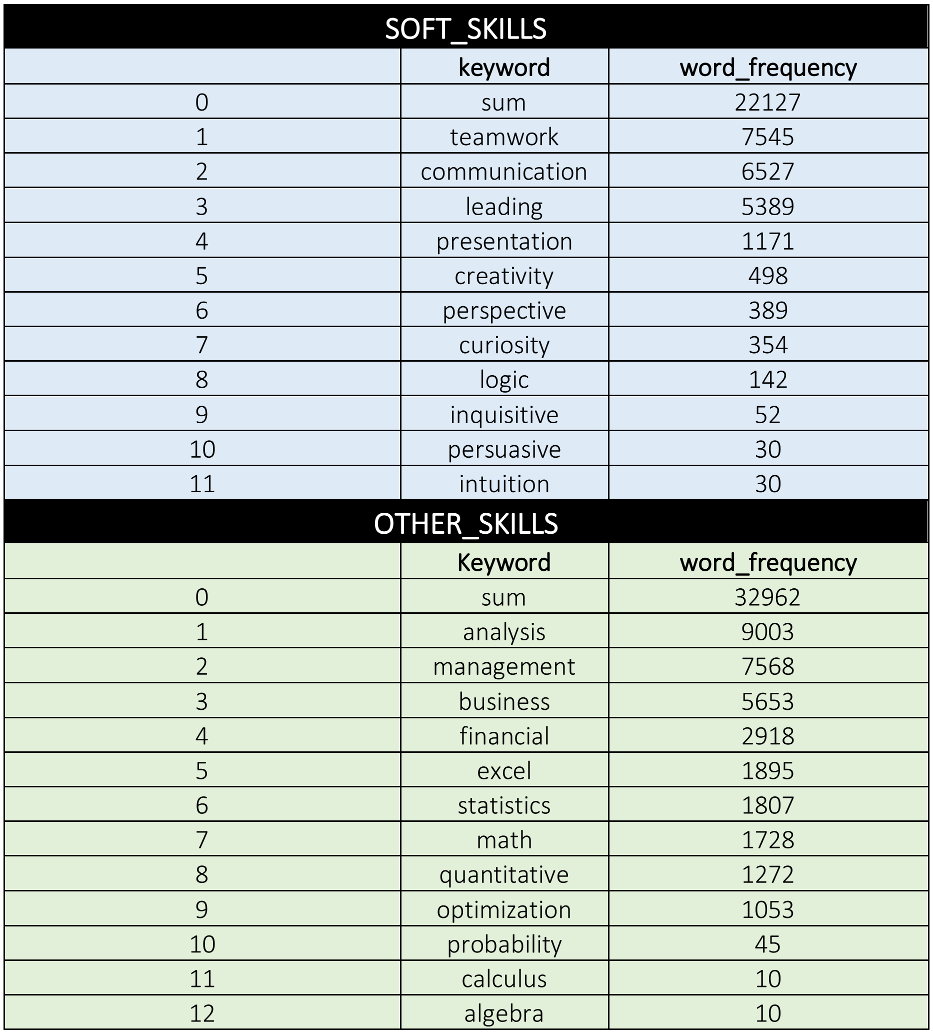
Figure 1: Keywords Divided into Three Groups by Using k -means Metho

Appendix B: Most Important Skills for Data Analysts

Technical Skills		
Scripting language	Key word	Word frequency
0	sum	13785
1	python	2571
2	sql	2413
3	r	2197
4	java	1202
5	spark	1183
6	c	1022
7	sas	1020
8	javascript	704
9	scala	552
10	matlab	313
11	pig	241
12	ruby	196
13	perl	170
14	octave	1
data_visualization	keyword	word_frequency
0	sum	1434
1	visualization	1240
2	d3	194
database	keyword	word_frequency
0	sum	3112
1	database	1516
2	hive	568
3	nosql	434
4	hbase	265
5	mysql	223
6	mongodb	106
tools	**keyword**	**word_frequency**
0	sum	318
1	shell	277
2	jupyter	41



machine_learning	keyword	word_frequency
0	sum	9863
1	learning	3279
2	machine	2487
3	intelligence	1854
4	ai	773
5	modelling	599
6	artificial	536
7	supervised	199
8	unsupervised	136
cloud_computing	keyword	word_frequency
0	sum	2244
1	cloud	1343
2	aws	576
3	azure	325
programming	keyword	word_frequency
0	sum	6598
1	software	3232
2	programming	1554
3	mining	1424
4	algorithm	226
5	wrangling	81
6	scraping	51
7	hacking	30
framework	keyword	word_frequency
0	sum	2952
1	hadoop	1258
2	tableau	999
3	tensorflow	432
4	mapreduce	236
5	zookeeper	21
6	mahout	6



SOFT_SKILLS

	keyword	word_frequency
0	sum	22127
1	teamwork	7545
2	communication	6527
3	leading	5389
4	presentation	1171
5	creativity	498
6	perspective	389
7	curiosity	354
8	logic	142
9	inquisitive	52
10	persuasive	30
11	intuition	30

OTHER_SKILLS

	Keyword	word_frequency
0	sum	32962
1	analysis	9003
2	management	7568
3	business	5653
4	financial	2918
5	excel	1895
6	statistics	1807
7	math	1728
8	quantitative	1272
9	optimization	1053
10	probability	45
11	calculus	10
12	algebra	10

Table 5: Important Skills and Word Frequency

Appendix C: MIE1624 New Syllabus

MIE 1624: Introduction to Data Science and Analytics

Instructors: Olesandr Romanko, oleksandr.romanko@utoronto.ca

Office hours: After the lecture or before tutorials

Prerequisites: None

Lectures: Monday 18:00 – 21:00 in GB244

TA: Sanjif Rajaratnam sanjif.rajaratnam@mail.utoronto.ca, Siyang Xu siyang.xu@mail.utoronto.ca, Zheng Huang <kennyv.huang@mail.utoronto.ca>

Course description

This course illustrates the application of data science and analytics in various fields, such as Managerial and Business Data Science. This course composed by two parts which are lectures and optional tutorials. Supervised and unsupervised learning, time series and neural networks will be approached in lectures. **Python, SQL** and **R** languages will be approached in tutorials. Unique aspects of managerial and business compared to other industries will be discussed. Real-world datasets will be provided to illustrate the complexity of applying these methods to different fields.

Course goals

- Build deep understanding of future careers
- Learn data science basic concepts and related math concepts
- Train about using data science software
- Learn some hot and trending topics
- Learn what is the data science and how to develop in different fields
- Learn to apply the basic data science knowledge and skills to science related works, like consulting

Grading

Assessment	Weight	Date
Homework (2 assignments, 20% each)	40%	See schedule of topics; Individual
Project	30%	See schedule of topic; Teams of 6
Exam	30%	See schedule of topics; Individual

All assessments are due by the start of class

Schedule of topics

The schedule of topics below is subject to change without notice

Week	Lecture	Tutorial	Due
1	Introduction	/	/
2	Tools	Basic Python	
3	Fundamental Maths	Pandas & NumPy	
4	Logistic Regression	Matplot & Seaborn	
5	Decision Tree – Random Forest	Scikit-learn	Assignment 1- Solving an analytics problem in Python
6	Decision Tree – Gradient Boosted Machine	Notebook&Tensorflow	/
7	Others	SQL	
8	Time Series	R	Assignment 2 – Solving an analytics problem in any other languages
9	Unsupervised – K mean	/	/
10	Neutral Networks		
11	Review		Project – Consulting in the field you are interested in
12	Test		/

Course outline

Special lecture from guest lecturer Geoffrey Hinton:

What skills do you need to be a successful data analyst?

Data mining and Machine learning

1. Supervised learning
   - Use Python to perform supervised learning
   - Build predictive models
     - Tune their parameters
     - Assess their performance on unseen data
     - Using real-world datasets
2. Unsupervised learning:
   - Clustering

        - K means
        - mixture models
        - hierarchical clustering
    - Dimension reduction

3. Data visualization
    - Customizing of plots: axes, annotations, legends
    - Overlaying multiple plots and subplots
    - Visualizing 2D arrays, 2D data sets
    - Working with color maps
    - Producing statistical graphics
    - Plotting time series
    - Working with images
4. Deep learning with Jupyter Notebooks in the Cloud

   Programming language

- Introduction to Python
- Introduction to QSL
- Introduction to R Programming
- Comparison of Python, QSL and R usage in data science

5. Soft skills

- Group communication skills
- Leading
- Creativity and Curiosity

Other skills

- Data analysis ability
- Math and statistics knowledge

Appendix D: Key Skills and Importance of skills for Each Occupation

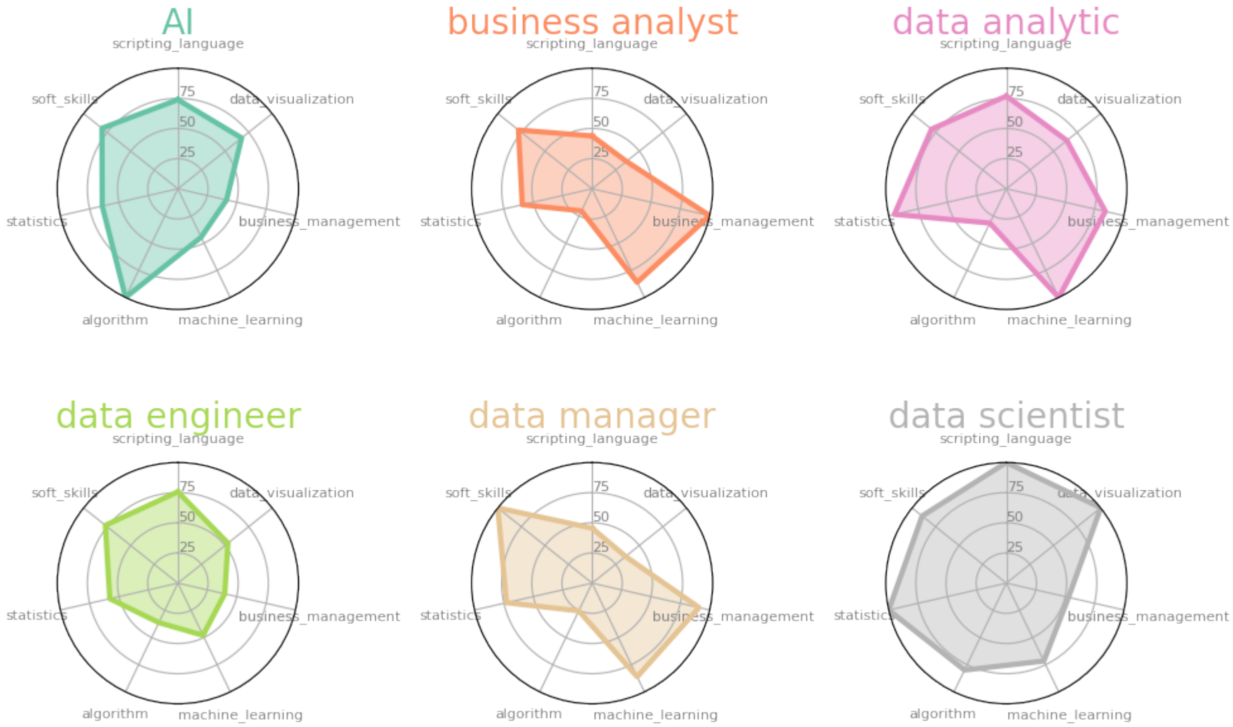
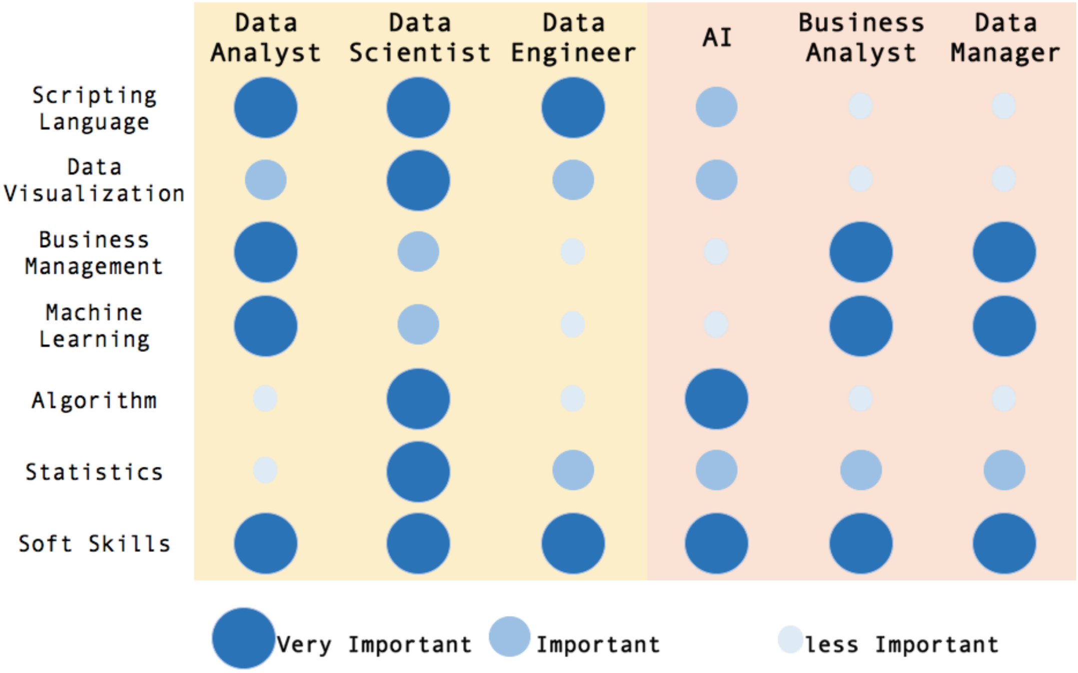


Figure 2: Key Skills for Each Job Title



	Data Analyst	Data Scientist	Data Engineer	AI	Business Analyst	Data Manager
Scripting Language	●	●	●	◍	○	○
Data Visualization	◍	●	◍	◍	○	○
Business Management	●	◍	○	○	●	●
Machine Learning	●	◍	○	○	●	●
Algorithm	○	●	○	●	○	○
Statistics	○	●	◍	◍	◍	◍
Soft Skills	●	●	●	●	●	●

● Very Important ◍ Important ○ less Important

Figure 3: The Importance Level of Skills for Each Job Title

Appendix E: The Importance of Skills for Each Program

Program	Technical Data Science				Managerial and Business Data Science			
Job	Data Analyst	Data Engineer	Data Scientist	Avg	AI	Business Analyst	Data Manager	Avg
Scripting Language	3	3	3	3	2	1	1	1
Data Visualization	2	2	3	2	2	1	1	1
Business Management	3	1	2	2	1	3	3	2
Machine Learning	3	1	2	2	1	3	3	2
Algorithm	1	1	3	2	3	1	1	2
Statistics	3	2	3	3	2	2	2	2
Soft Skills	3	3	3	3	3	3	3	3

Table 6: Importance Level for Each Job Title and Program

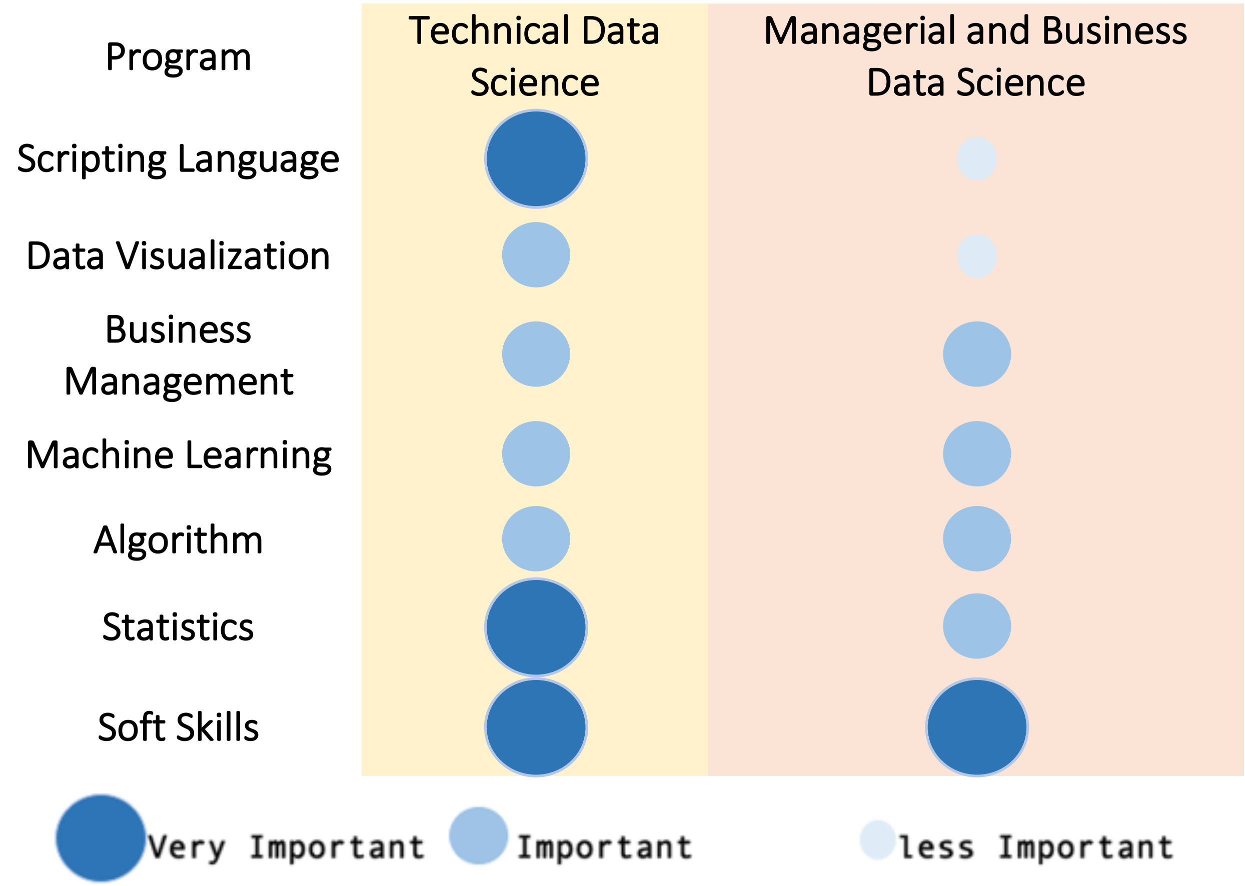
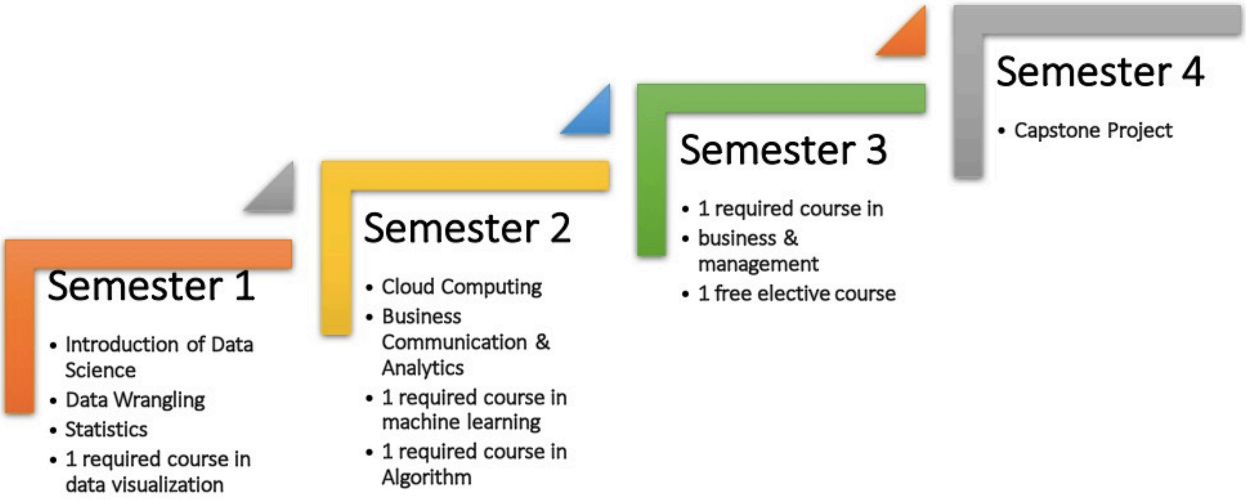


Figure 4: The Importance Level for Each Program

Appendix F: Courses Sequence in M.D.S.A.



Semester 1

- Introduction of Data Science
- Data Wrangling
- Statistics
- 1 required course in data visualization

Semester 2

- Cloud Computing
- Business Communication & Analytics
- 1 required course in machine learning
- 1 required course in Algorithm

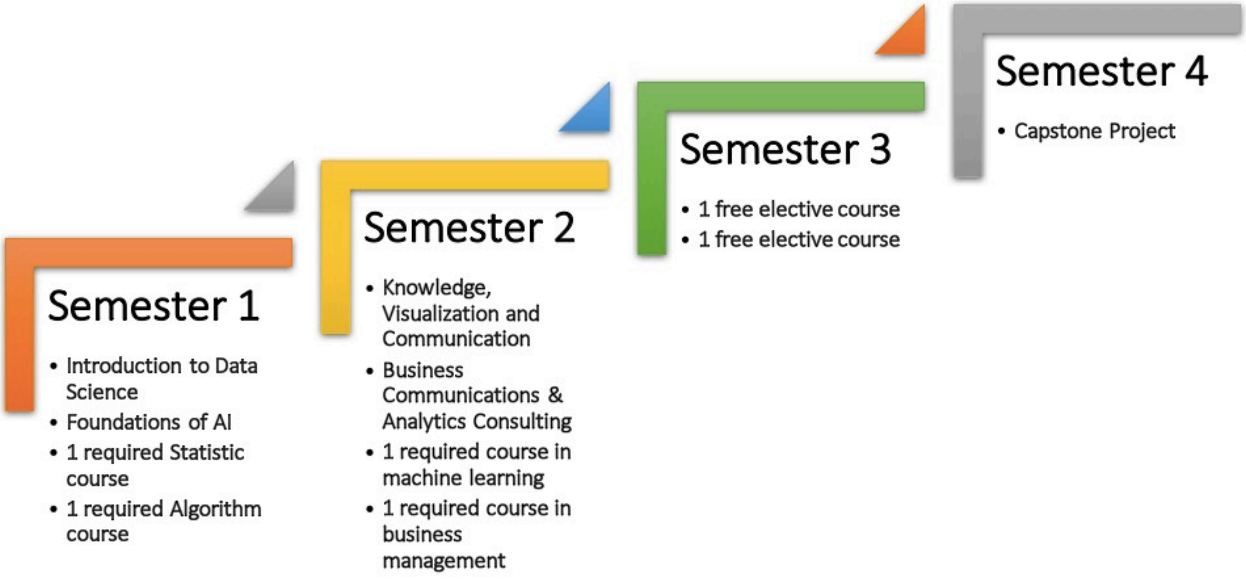
Semester 3

- 1 required course in
- business & management
- 1 free elective course

Semester 4

- Capstone Project

Appendix G: Courses Sequence in M.B.A.I.



Semester 1

- Introduction to Data Science
- Foundations of AI
- 1 required Statistic course
- 1 required Algorithm course

Semester 2

- Knowledge, Visualization and Communication
- Business Communications & Analytics Consulting
- 1 required course in machine learning
- 1 required course in business management

Semester 3

- 1 free elective course
- 1 free elective course

Semester 4

- Capstone Project

Appendix H: Courses Offered in Program

The following is the course curriculum for the courses offered in M.D.S.A and M.B.A.I.

Statistics Category:

Advanced Optimization
Level: Advanced

Optimization problems are widely used in many decision problems including in manufacturing systems, such as decisions on the acquisition, utilization and allocation of production resources. This course covers advanced optimization techniques in many different fields, such as logistics, manufacturing and transportation. This course included popular optimization methodologies like linear programming, integer programming, and constraint programming. The objectives for this course are: a) Improving the capacity of modelling complex optimization problems in such a way that they can be solved using standard software packages; b) Providing an understanding of principal optimization problem solving procedures and c) Develop specialized solution procedures for non-standard problems. Furthermore, students learn the best way of solving Sudoku.

Statistics
Level: Introduction

Statistics is a branch of mathematics dealing with the collection, analysis, interpretation, presentation and organization of data. Statistics knowledge is important for data analyze no matter for which area. Main objectives for this course is to cultivate statistical thing, develop skills in handling complex problems in data analysis and research design and prepare students for future courses having quantitative components.

Modern Applied Statistics
Level: Intermediate

This course will focus on the theoretical foundations as well as the practical implementation of numerical methods in statistics. Topics to be covered include: pseudo-random number generation, simulation methods for inference, bootstrapping, numerical optimization, estimation of functions/density estimation and Markov chain Monte Carlo methods.

Probability
Level: Introduction

Learn the language and core concepts about probability. Basic principles about Bayesian and frequentist contained. This course will use the software R to do simulation.

Business and Management Category:

Foundations of Data Management
Level: Introduction

This course focuses on providing facets of data management with solid and robust mathematical foundations. it is prepared for those who want to get fundamental knowledge on the data management. Storage system, file system, network file system and data modeling methods are included in lectures.

Analytics of Finance
Level: Intermediate

This course covers the main quantitative methods of finance. The course covers three broad sets of topics: derivative pricing using stochastic calculus, dynamic optimization, and financial econometrics. The emphasis is on rigorous and in-depth development of the key techniques and their application to practical problems.

Project Management
Level: Introduction

This course guides students through fundamental project management concepts and behavioral skills needed to success-fully launch, lead, and realize benefits from projects in profit and nonprofit organizations.

Operation Management
Level: Intermediate

This course is an introduction to the concepts, principles, problems, and practices of operations management. Emphasis is on managerial processes for effective operations in both goods-producing and service-rendering organization. Topics include operations strategy, process design, capacity planning, facilities location and design, forecasting, production scheduling, inventory control, quality assurance, and project management. The topics are integrated using a systems model of the operations of an organization.

Marketing Analytics
Level: Intermediate

This course aims to cover topics in marketing analytics, an area that remains the decision enabler of utmost importance for many of the offline and online companies' marketing and merchandising divisions. The objective of the course is to give students a general understanding of this vital area in marketing while demonstrating critical application areas in online and offline marketing channels.

Machine Learning Category:

Machine Learning for Data Science
Level: Intermediate

Since machine learning is a growing field leaking into almost all fields such as web searching, advertisement placing, credit scoring and so on. This course will develop a deeper understanding of practical solutions using predictive analytics. Perquisite is MIE 1624. The objective for this course is to get further understanding of machine learning and data analysis. Frequently using algorithmic techniques including sorting, searching, greedy algorithms and dynamic programming in assignments and projects are required.

Probabilistic Graphic Models
Level: Intermediate

Probabilistic graphical models are a powerful framework for representing complex domains. The aim of this course is to develop the knowledge and skills necessary to design, implement and apply these models to solve real problems. Basic probability theory, statistics, programming, algorithm design and analysis knowledge are required before attending this course.

Regression
Level: Intermediate

This graduate level course offers an introduction into regression analysis. Regression knowledge is important for researchers to predict a future value for some dependent variable. The process of finding this mathematical model that best fits the data involves regression analysis. This is an applied linear regression course that emphasizes data analysis and interpretation.

Deep Learning
Level: Advanced

Deep learning is a branch of machine learning concerned with the development and application of modern neural networks. Deep learning is behind many recent advances in AI, including Siri's speech recognition, Facebook's tag suggestions and self-driving cars. We will cover a wide range of topics from basic neural networks, convolutional and recurrent network structures, deep unsupervised and reinforcement learning, and applications to problem domains like speech recognition and computer vision. Prerequisites for this course is that students have a strong mathematical background in calculus, linear algebra, and probability & statistics (students will be required to pass a math prerequisites test), as well as programming in Python and C/C++.

Data Mining
Level: Introduction

Data Mining studies algorithms and computational paradigms that allow computers to find patterns and regularities in databases, perform prediction and forecasting, and generally improve their performance through interaction with data. It is currently regarded as the key element of a more general process called Knowledge Discovery that deals with extracting useful knowledge from raw data. The knowledge discovery process includes data selection, cleaning, coding, using different statistical and machine learning techniques, and visualization of the generated structures. The course will cover all these issues and will illustrate the whole process by examples.

Predictive Modeling
Level: Intermediate

This class is an introduction to supervised learning, ie, predictive models ideas. Our main goal is to introduce the main concepts and the students familiarize with the most popular tools in this area. KNN, Regression, Logistic Regression, etc. will included.

Statistical Learning
Level: Advanced

provide an introduction to the theory of statistical learning and practical machine learning algorithms with applications in signal processing and data analysis.

Foundations of Artificial Intelligence
Level: Introduction

This course will give the overview about the field of Artificial Intelligence. It contains the foundations of symbolic intelligent systems, search, logic, knowledge representation and so on.

Natural Language Processing
Level: Intermediate

Students will understand syntax and semantics in NLP. Current methods about statistical to machine translation will also be talked about. Machine learning techniques used in NLP is contained.

Soft Skills Category:
Business Communication & Analytics Consulting
Level: Introduction

A detailed study of business communication and consulting for practical analytics problems are included in this course. The course objective is to create a professional, public-facing business message, develop a persuasive application packet, improve students' fundamental data analysis ability. Soft skills like communication skill will be developed through team project and presentation.

Communication Theories in Technology and Society

Level: Introduction

Explores impact of communication on social action, corporate environments and interpersonal relationships. Formation and management of online identities discussed. Introduction to online media construction and analysis with particular emphasis on the world wide web.

Knowledge, Visualization and Communication

Level: Introduction

This course is about the introduction of visual analytics and related basic techniques. It includes computer graphics, information analysis and so on.

Algorithms Category:

Discreet Math and Algorithm

Level: Introduction

Introduce students to ideas and techniques from discrete mathematics that are widely used in science and engineering. This course teaches the students techniques in how to think logically and mathematically and apply these techniques in solving problems. To achieve this goal, students will learn logic and proof, sets, functions, as well as algorithms and mathematical reasoning. Key topics involving relations, graphs, trees, and formal languages and computability are covered in this course.

Analysis of Algorithms

Level: Intermediate

This course will talk about the analysis about the related algorithms, writing proofs for algorithms and applying algorithms design paradigms.

Algorithms and Data Structure

Level: Intermediate

This course will talk about three main topics, which are data structures, algorithms and generic programming. It contains algorithms design, complexity analysis and correctness proof in details. This course teaches coding for both data structures and algorithms in C ++.

Data Visualization Category:

Data Visualization

Level: Introduction

You will understand how to request data sets from different sources. You will learn how to clean up, format and analyze data. How to use related tool to build static and online visualizations will also talked about.

Data Visualization II

Level: Intermediate

Interactive visualization, design choices, dynamic change over time, multiple views, data reduction, dealing with complexity.

Programming Category:

Introduction to Data Science

Level: Introduction

This course illustrates the application of data science and analytics in various fields, such as Managerial and Business Data Science. This course composed by two parts which are lectures and optional tutorials. Supervised and unsupervised learning, time series and neural networks will be approached in lectures. **Python, SQL** and **R** languages will be approached in tutorials. Unique aspects of managerial and business compared to other industries will be discussed. Real-world datasets will be provided to illustrate the complexity of applying these methods to different fields.

Cloud Computing

Level: Intermediate

In this course, we use open source implementations of highly available clustering computational environments, as well as RESTFul Web services, to build very powerful and efficient applications. We also learn how to deal with not trivial issues in the Cloud, such as load balancing, caching, distributed transactions, and identity and authorization management. In the process we will also become very familiar with Linux operating system. Students need to familiar with python, java or C.

Distributed System

Level; Intermediate

The course will contain the following topics: Remote Objects and Remote Invocation; Clocks and Clock Synchronization; Logical time and Logical Clocks; Global States; Replication; Transactions and Concurrency Control; Coordination and Agreement and Multi-cast.

Data Wrangling

Level: Introduction

This course will enable you to design database schemas, implement database schemas, navigate data management issues and so on. It contains new technologies and programming in Python.

Computer Security

Level: Intermediate

This course provides an introduction to security and privacy issues in various aspects of computing, including programs, operating systems, networks, databases, and Internet applications. It examines causes of security and privacy breaches and gives methods to help prevent them. Students completing this course should be better able to produce programs that can defend against active attacks, and not just against random bugs.

Graph Theory and Application

Level: Intermediate

This course will talk about basic graphic theoretical concepts: paths and cycles, trees, bipartite graphs, spanning subgraphs and etc. Algorithms related also will contained in discussion. Matching theory, Coloring and so on. Applications talked in this course are in biology and social sciences area.

Introduction to Database

Level: Introduction

This course talked about how to model data using different model. Also, this course will teach implementation of schema using SQL. Using SQL do query to your data. The topic about RDBMS will also talked in this course.

References:

Coyle, P. (2015, January 29). *THE IMPORTANCE OF SOFT SKILLS IN DATA SCIENCE*. Retrieved from DATACONOMY: http://dataconomy.com/2015/01/the-importance-of-soft-skills-in-data-science/

Holtz, D. (2014, November 7). *8 Skills You Need to Be a Data Scientist*. Retrieved from UDACITY: https://blog.udacity.com/2014/11/data-science-job-skills.html

Iqbal, R. (n.d.). *WHAT ARE THE KEY SKILLS OF A DATA SCIENTIST*. Retrieved from dataschiencedojo: https://datasciencedojo.com/key-skills-of-a-data-scientist/

Mittal, S. (2016, April 9). *Soft Skills that the Most Successful Data Analysts Have*. Retrieved from ANALYTIXLABS: https://www.analytixlabs.co.in/blog/2016/04/08/soft-skills-that-the-most-successful-data-analysts-have/

Shron, M. (2014). Thinking with Data.

Terdoslavich, W. (2017, March 15). *Key Skills That Data Scientists Need*. Retrieved from Dice: https://insights.dice.com/2017/03/15/key-skills-data-scientists-need/

Wickham, H. (2014, September 4). *Data science: how is it different to statistics*. Retrieved from IMS Bulletin online: http://bulletin.imstat.org/2014/09/data-science-how-is-it-different-to-statistics%E2%80%89/