MIE 1624 Introduction to Data Science

The Design of Data Science Courses and Programs

A data-driven approach

Project Report

Produced by:

Sundeep Pothula
Nathan (Jun Lin) Guan
Kelly (Kexin) Zhang
Emmett Borg

Date:
March 23, 2018

Professor:
O. Romanko

1. Problem Definition

As the use of data science and analytics spread across industries, demand for qualified individuals increases. The steps to becoming qualified as a data science professional, however, are not so clearly defined. Compared to a field such as law, data science as a study is in its infancy. Without the time (centuries in the case of law) to hone course curriculums or program structures, educators must move quickly to meet the growing demand.

As such, this consultation paper for the University of Toronto focuses on designing the following:

(1) a graduate level data science and analytics course curriculum,
(2) a technically focused data science graduate program curriculum,
(3) a managerial focused data science graduate program curriculum.

The objective of all three curriculums is to educate individuals to a strong level of job readiness in the field. Employability and marketability of graduating students is therefore a key consideration in all design choices.

2. Data Sources

As these curriculums are tailored towards industry employability, public job boards are a natural data source. For this we used the following source for job descriptions:

Indeed: an international job board with high user traffic, simple design, and is free to use.
Stackoverflow: a Q&A site with large user-base in the technology field, including a job board.

We are also interested in currently offered academic programs and courses related to data science. We used the following sources for course curriculums:

University Sites: A total of 10 graduate program curriculums were scraped from various educational institution websites. Included were 7 focused on technical data science, and 3 focused on management analytics.

And finally, the 2017 Kaggle ML and Data Science survey dataset includes a variety of useful data regarding the current state of data scientists. The Kaggle dataset included freeform text and multiple choice responses, and is free to the public. A full list of sources can be found in the appendix.

3. Data Analysis

For identifying different skills from job and course descriptions, advanced Natural Language Processing technique - IBM Watson Natural Language Understanding (NLU) is used. Since Watson NLU is not trained to determine different Skill Set entities, A custom Model is created in Watson Knowledge studio and is trained to identify skill sets in each document as Technical skills, Coding skills or Soft skills. After testing in Watson and improving prediction accuracy to 95%, Model is deployed in Watson NLU. Deployed model is called using Watson API and can be used to process text, html and url links for identifying skill sets.

To design course curriculum in a data driven approach for Technical and Business programs on the basis of skills required for jobs we followed a two fold strategy. First, in an Optimization way, based on the frequency of each technical, coding and soft skills in job descriptions, each course obtained from university sites is assigned a weighted score. The scores of the 121 courses in the 7 technical graduate programs are calculated based on skills extracted from technical data science job postings (2969 jobs used). On the other hand, the scores of the 63 courses in the 3 business-oriented graduate programs are calculated based on skills extracted from business-oriented data science job postings (1101 jobs used). Second, we used KNN algorithm to predict the courses that cover maximum skill sets required by the jobs. We combined and matched the two algorithm course predictions to finalize the course curriculum. Our analysis showed that different skill sets are preferred for technical and business-oriented job positions. Therefore, the two different types of curriculums are designed accordingly. For

example, it was found that Python is the most referenced programming languages in technical job descriptions and we recommend that Python be predominantly used in teaching practical applications in all course curriculums listed below.

4. Recommended Introductory Data Science Course Curriculum

Our analysis produced a list of topics which were most relevant based on data scientist skills sought after in the workforce. These topics should serve as a guideline to the lecture topics for the redesign of MIE 1624: Intro to Data Science and Analytics.

The topics include the following:

- Databases
- Programming
- Big Data
- Cloud Computing
- Supervised Machine Learning
- Unsupervised Machine Learning
- Statistics
- Data Visualization
- Simulation Modeling

As lecture time is limited, we also recommend the following list of topics be used for short presentations to be completed by small groups. These topics may be encompassed by one of the eight above, but would benefit from further dedicated lecture time. Currently the MIE 1624 course implements group presentations of this nature, and has proved itself useful for covering as much content as possible. We are not recommending a redesign of the presentation structure, only a different set of topics. These topics are chosen from Kaggle datasets which provides us the information of top skills used in current data science related jobs. A detailed graph is provided in Fig 2.1.

- Regression/Logistic Regression
- Decision Trees
- Random Forests
- Time Series Analysis
- Natural Language Processing
- Ensemble Methods
- Outlier detection (e.g. Fraud detection)
- Neural Networks
- Gradient Boosted Machines
- Bayesian Techniques
- Support Vector Machines
- Computer Vision

5. Recommended Technical Data Science Program Curriculum

The University of Toronto is also considering the launch of a technically-oriented Master of Data Science and Analytics (M.D.S.A) program. Note that the faculty of Applied Science and Engineering currently offers a variety course-based Masters degrees, which will be consistent with the general structure of our recommendation.

We recommend that M.D.S.A students complete either:

(A) 10 courses, including at least 5 core courses

(B) 7 courses, including at least 4 core courses and a extra project-based course

Course topics were found through the analysis of technically-focused positions posted on the job board datasources. Specifically, results found using search results such as “data scientist”, “data analyst”, and “data engineer” would generally be technical in nature. Core courses topics were the most common required skills in these results. Remaining course requirements may be filled with any of the elective courses listed below. We do not consider these courses core as they focus on specific applications or industries that may not be relevant for all students.

All courses from university technical programs are ranked based on weighted scores assigned, as explained in Section 3. Top 30 courses are listed and compared with the top 30 courses obtained using nearest neighbour approach. By cross-referencing the two lists, a final list of courses that reflect the most common required skills are

produced, from which a list of core courses and elective courses are chosen. For elective courses, we incorporated courses from other disciplines, for example Data Driven Medicine and Sustainability Technology, in order to cover a broader career scope. The technical program courses are listed below:

Course Names	**Main Topics to cover**
Core Courses	
Intro to Data Science	Programming,, Statistics, Machine Learning
Scalable Data Systems & Algorithms	Databases,, Cloud technologies
Topics in Modern Statistics: Applied Machine Learning	Statistics, Machine Learning
Mining Massive Data Sets	Algorithms, Big Data, Spark, Hadoop
Advanced Machine Learning	Machine Learning, Algorithms
Management of Big Data and Big Data Tools	Big data, Hadoop, NoSQL
Exploratory Data Analysis & Visualization	Visualization Techniques, Tableau
Elective Courses	
Topics in Quantitative Finance: Big Data in Finance	Big Data, Data science for Finance services
Topics in Information Processing: Deep Learning for Computer Vision, Speech, and Language	NLP, Text Analytics, Computer vision
Analytics for Big Data	Big Data, Hadoop
Business Intelligence from Big Data	Probability, Statistics, SQL
Data Science Visualization Lab	Visualizations, Data Exploration Projects
Business Communication & Analytics (Communication)	Communication, Teamwork, soft skills
Sustainability Technology - Urban Analytics	Projects, Design of Intelligent cities
Data Driven Medicine	Big Data, Healthcare ,Biostatistics, Social research
Project-Based Courses	
Research Project	Students have the option to conduct research in data science with a faculty professor.
Capstone Project	Students have the option to propose a capstone project either under the supervision of a faculty member, or sponsored by an industry partner

Table 1: Proposed Technical-oriented Master's Program Curriculum

6. Recommended Managerial and Business Data Science Program Curriculum

To appeal to students with a business and soft-skills focus, the University of Toronto is also considering the launch of a Master of Business and Management in Analytics and AI (M.B.A.I) program.

Course topics were found through the analysis of managerial or business focused positions posted on the job board datasources. Specifically, results found using search results such as "business analyst" and "data analytics manager" would generally require a combination of soft and technical skills.

Note that the University of Toronto's Rotman School of Management currently offers a variety of elective graduate level courses for the Master of Business Administration program. Should there exist collaboration between faculties, the M.B.A.I program may utilize some existing offerings for elective courses.

We recommend that M.B.A.I students complete either:

(A) 10 courses, including at least 6 core courses

(B) 7 courses, including at least 5 core courses and a project-based course

Similarly, by cross-referencing the two lists, a final list of courses that reflect the most common required skills are produced, from which a list of core courses and elective courses are chosen. Based current job trends from Kaggle Survey, data science is largely applied in fields of marketing, finance, health care etc. Therefore courses that cover the corresponding fields are also selected as electives. The business program courses are listed below:

Course Names	**Main Topics to cover**
Core Courses	
Business Analytics Strategy	Management, Consulting in Analytics
Acquisition and Analysis of Data	Databases, Visualizations, SQL
Predictive Modelling	Machine Learning, programming, Business cases
Big Data Analytics	Big Data, Data Analysis, NoSqL
Coding Foundations for Analytics	Programming, Algorithms
Project Management	Leadership, Management
Data Mining of Visualization	Business Intelligence, Visualizations, Tableau
Elective Courses	
Managing Data Analytics Teams	Leadership, Teamwork, Communication, Management
Marketing Analytics	Customer Theory, Marketing
Pricing Analytics	Marketing ,Mathematical Optimization, Case study
Data-Driven Quality Management	Strategic Management, Six Sigma, Statistics

Operations & Supply chain Analytics	Supply chain Management, Projects
Healthcare Analytics	Projects in Healthcare
Decision Analytics	Business Decision Mapping, Projects
Project-Based Courses	
Business Case Competition Project	Students can participate in a business case competition in collaboration with a business partner
Capstone Project	Students have the option to propose a capstone project either under the supervision of a faculty member, or sponsored by an industry partner

Table 2: Proposed Business-oriented Master's Program Curriculum

7. Connecting Students with Jobs and Internships: Startup Pitch:

In the previous sections we have mentioned the importance of employability, and the recommendation of internships for M.D.S.A. and M.B.A.I. students. However, connecting students with employers is no easy task. Many schools offer private job boards available only to alumni or current students, but do not take an active or data-driven approach to match students and employers or jobs.Our Startup aims to solve this pain point using Machine Learning approach.

We are looking to create a personalized start to end experience in the field of Data Science, from learning phase to Job search by building a fully functional student job board. This tool has the potential to provide multiple benefits:

(1) Higher employment rates for graduates and internship seekers
(2) Better understanding of required skills for a job a students wants, and how to achieve them
(3) More efficient use of time and resources during a job search, as you are applying to jobs which are seeking something near your skillset.

Through our analysis for the previous sections, we found it possible to determine necessary skills from a job description. Further, we can use the same process to determine skills listed in a students résumé. From here, we are using text analytics and KNN machine learning algorithm to match a student résumé with the most closely matched job posting on a skills basis. For any skills that the student currently does not have listed, we can also recommend relevant blogs on the topic. This would help educate the student to match the position requirements more closely if they are interested in doing so.

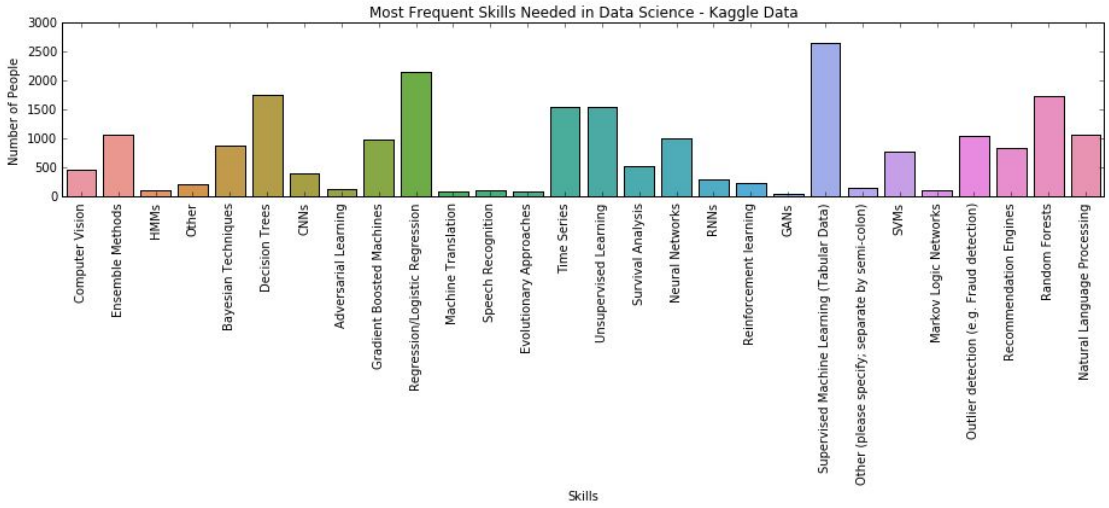
Appendix

1. Data Sources

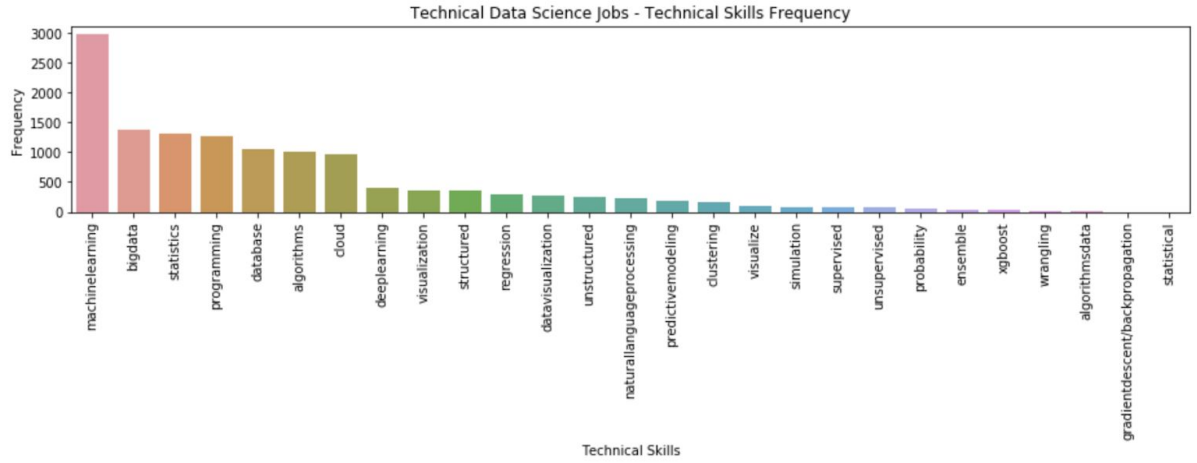
Name	**Link(s)**
Indeed	For {city name} in ['Toronto', 'Vancouver', 'Waterloo', 'New-York', 'Chicago', 'San-Fransisco', 'Boston', 'Seattle', 'San-Diego', 'Atlanta'] And for {job title} in [‘data+scientist’, ’data+analyst’, ‘data+engineer’, ‘business+analyst’, ‘data+analytics+manager’] http://www.indeed.ca/jobs?q={job title}&l={city name} for Canadian cities http://www.indeed.com/jobs?q={job title}&l={city name}
Graduate Curriculums with Data Science focus	http://www.mccormick.northwestern.edu/analytics/curriculum/courses.html https://statistics.stanford.edu/academics/ms-statistics-data-science https://masterdatascience.science.ubc.ca/program/courses https://www.ryerson.ca/graduate/datascience/courses/ http://datascience.columbia.edu/course-inventory#Stat and CS https://www.datasciencemasters.uw.edu/program-details/courses-curriculum/course-descriptions/ https://analytics.stat.tamu.edu/for-students-2/
Graduate Curriculums with Management Analytics focus	https://smith.queensu.ca/grad_studies/mma/program_structure_and_content/curriculum.php https://www.mcgill.ca/desautels/programs/mma/program-structure https://wpcarey.asu.edu/masters-programs/business-analytics/curriculum
Kaggle ML and Data Science Survey 2017	https://www.kaggle.com/kaggle/kaggle-survey-2017/data

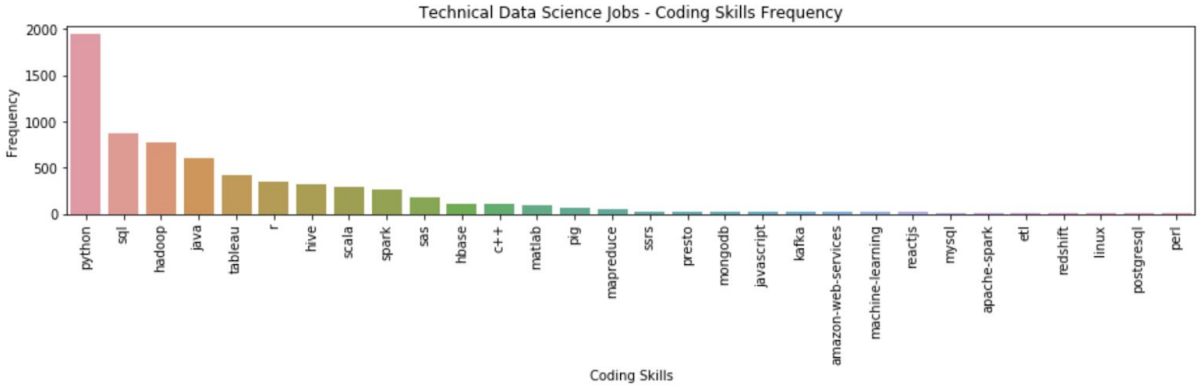
2. Figures

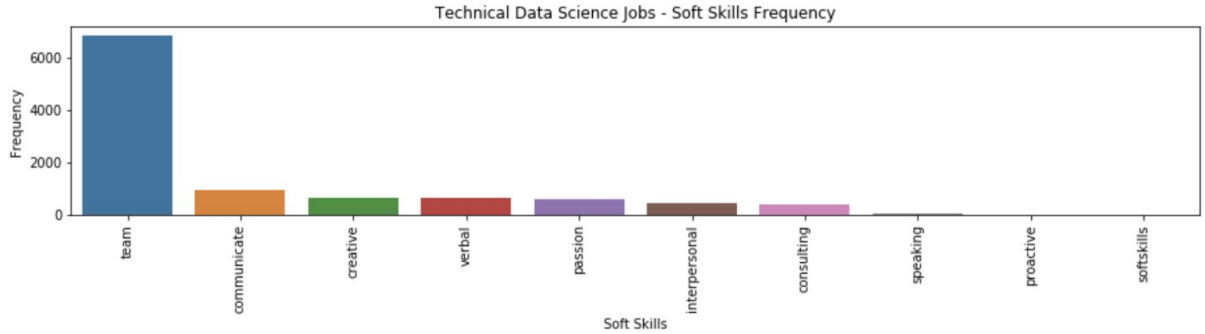
2.1 Kaggle Data Set for Frequent Job Skills

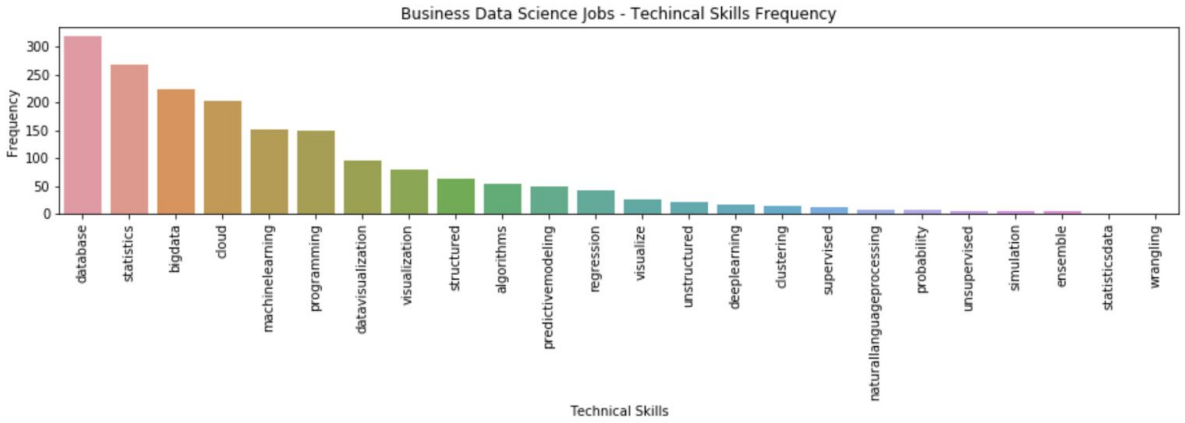


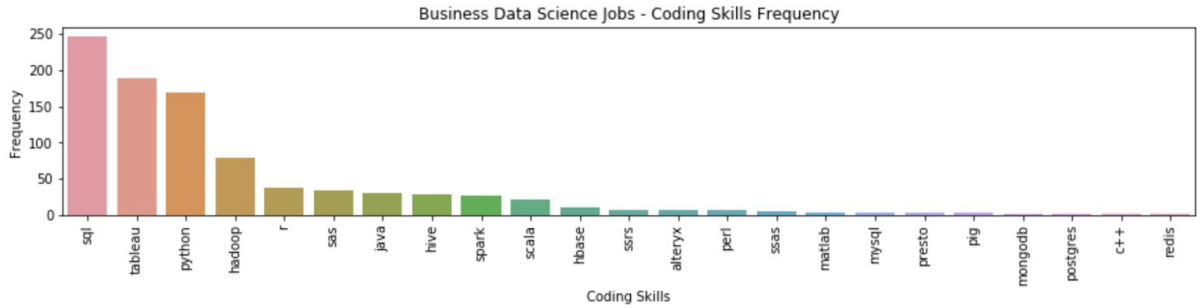
2.2 Skills Frequency Summary from Indeed Job Postings

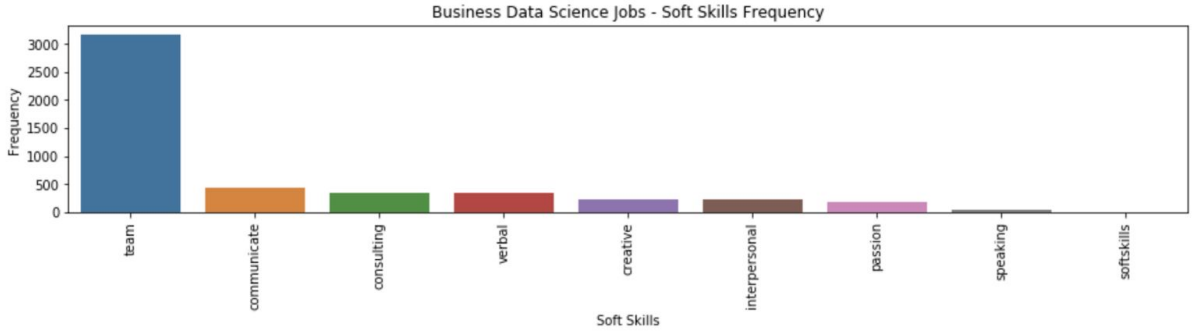







Business Data Science Jobs - Techincal Skills Frequency
Frequency
0
50
100
150
200
250
300
database
statistics
bigdata
cloud
machinelearning
programming
datavisualization
visualization
structured
algorithms
predictivemodeling
regression
visualize
unstructured
deeplearning
clustering
supervised
naturallanguageprocessing
probability
unsupervised
simulation
ensemble
statisticsdata
wrangling
Technical Skills


Business Data Science Jobs - Coding Skills Frequency
Frequency
0
50
100
150
200
250
sql
tableau
python
hadoop
r
sas
java
hive
spark
scala
hbase
ssrs
alteryx
perl
ssas
matlab
mysql
presto
pig
mongodb
postgres
c++
redis
Coding Skills


Business Data Science Jobs - Soft Skills Frequency
Frequency
0
500
1000
1500
2000
2500
3000
team
communicate
consulting
verbal
creative
interpersonal
passion
speaking
softskills
Soft Skills

2.3 Treeplot of Different Skill Set for Technical and Managerial Positions

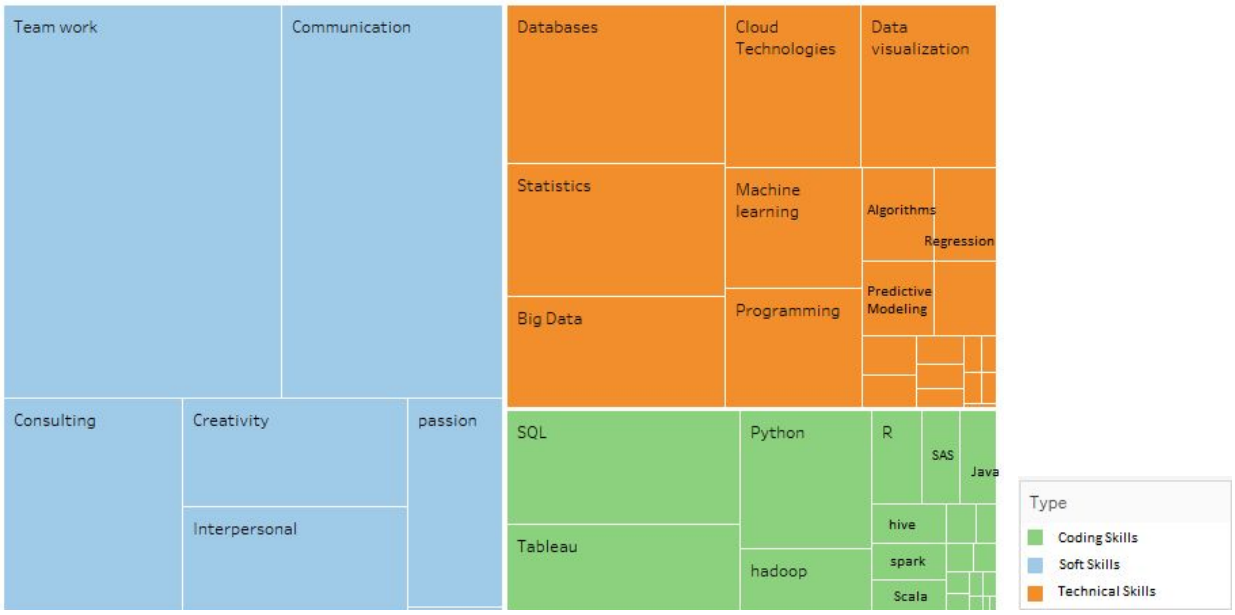


Fig 2.3.1 Top Skills (Technical Position) - Treeplot

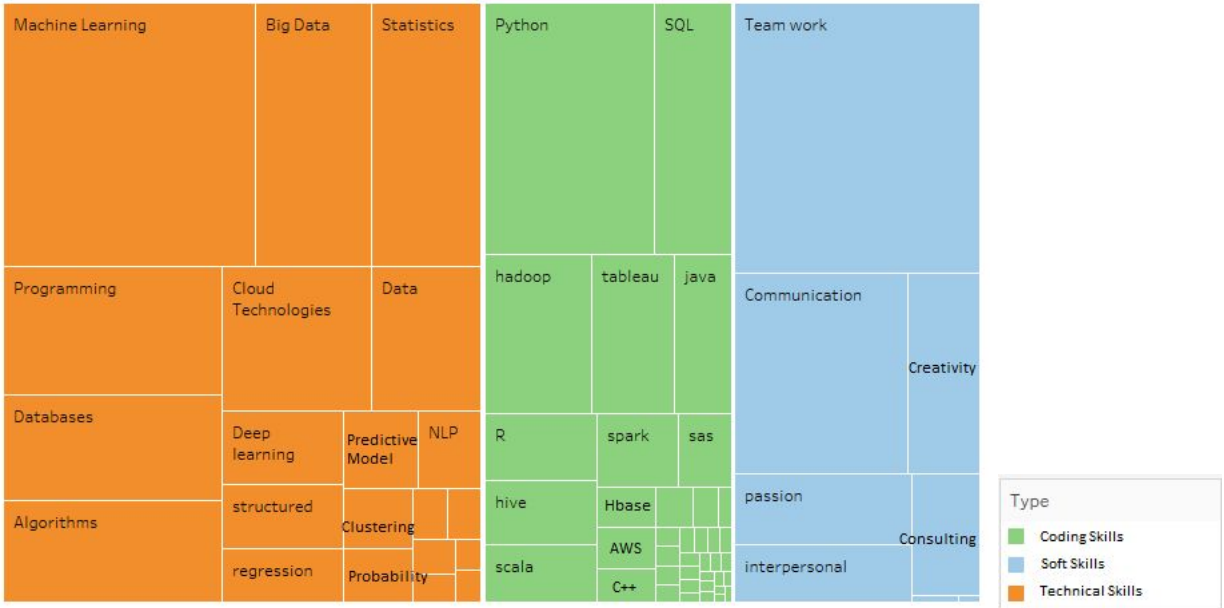


Fig 2.3.2 Top Skills (Managerial Position) - Treeplot