**Below are the Details to be moved to Jupyter Notebook**

**1- Scrape job postings from Indeed.ca**

We scrapped a total of 1136 job postings listed on indeed.ca by searching the following Job titles setting the search location to be “Canada”:

Job titles that indicate the skills required in industry for the graduates of a technical Master’s program or an introductory technical course in Data Science:

1. Data Analyst
2. Data Scientist

XXX job listings that resulted from searching each of the above technical job titles were scraped and processed using our indeed\_web\_scraper routine.

The indeed\_web\_scraper routine does the following:

1. Collects the first 3-4 result pages from searching each of the job titles and returns a list of result page URLs
2. Each of the result pages contains 20 job postings
3. The web scraper parses the text in each of the 20 job postings that appear in each one of the result pages collected
4. The web scraper then looks for certain keywords that are defined in a Python dictionary called skills\_dict
5. The skills\_dict was constructed to include a comprehensive list of skills that relate to programming languages, technical skills, and business, management, and soft skills that we would like to capture the frequency of their occurence in the listed jobs
6. The web scraper forms a Pandas DataFrame with its rows representing the job postings, and columns representing each of the skills in the skills\_dict (the skills we are interesting in capturing their presence, or lack of, in each job posting)
7. For each job posting scraped, “1” is assigned to a skill if it has been mentioned in that job posting, and “0” is assigned if the skill was not mentioned in the job posting.

Two plots are then generated that help us interpret the results of the scraped job postings:

1. A bar plot showing the frequency at which each skill was mentioned in the scraped job postings.

This plot indicates the skills that are most in demand in industry and which should be included in our program and course curriculum as core topics.

1. A dendrogram visualizing the hierarchical clustering of the skills.

This plot indicates how the skills demanded by employers in the job postings relate to one another. Jobs that co-occur in multiple job postings end up in the same cluster. As the frequency of co-occurrences of a pair of skills increases, the clusters to which the skills are assigned become closer.

We use this information to design the curriculum of each course by teaching the skills in similar/closer clusters together in the same course.

The above process is repeated by searching indeed.ca for job titles that indicate the skills required in industry for the graduates of a business/management Master’s program. The job titles searched for this time are:

1. Management Analytics
2. Business Analyst
3. Data Manager
4. Analytics Manager
5. AI System Designer

A bar plot and a dendrogram are produced similar to the ones above for these job titles that are used to design the business/management Master’s degree curriculum.

**2- Collect and analyze data from Kaggle\_ML 2018 survey responses:**

The second source of data we use in designing our course and program curriculums is the Kaggle\_ML 2018 survey responses. Certain questions were selected based on the relevance of the answer they provide to skills and/or topics that a graduate of our programs should know to equip them with the knowledge required to work in industry.

Below is a list of the questions that were selected:

* fdjff
* dfjkd
* sdfljd

A was formed with its rows representing the survey respondents and columns representing the tool, skill, or other categorical answer to the selected questions.

Bar plots were then produced representing the frequency of answers to each questions. These bar plots help us deduce the most important skills, topics, or other type of categories currently utilized in industry. Our programs and courses would then teach these high frequency skills as core topics due to their importance in the industry that the programs/courses are meant to equip our graduates for.

A co-occurrence matrix with highlighting the skills that co-occur in the respondents’ answers is then formed to obtain a dendrogram that visualizes how each of the skills/topics relate to one another. This is used to assist us in combining the skills in the same, or close, clusters together in our curriculum.

**Technical program design**

**Section 1: program “input” -- profiles of incoming candidates**

This project is intended as a design proposal for a new professional Master of Data Science and Analytics (MSDA) program with technical emphasis. Data science is an interdisciplinary field that combines expertise from different technical and expert domains. Data science is an interdisciplinary field that combines expertise from different technical and expert domains. As can be seen in figure 1, while the majority of Data Scientists come from classical technical backgrounds (Math, engineering and CompuSci), many field practitioners are also from such fields as Business and Life and Social sciences.

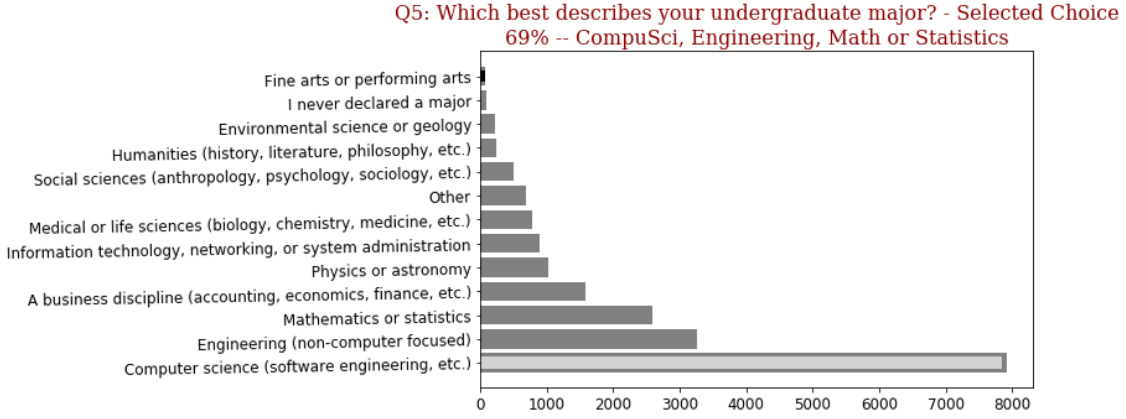
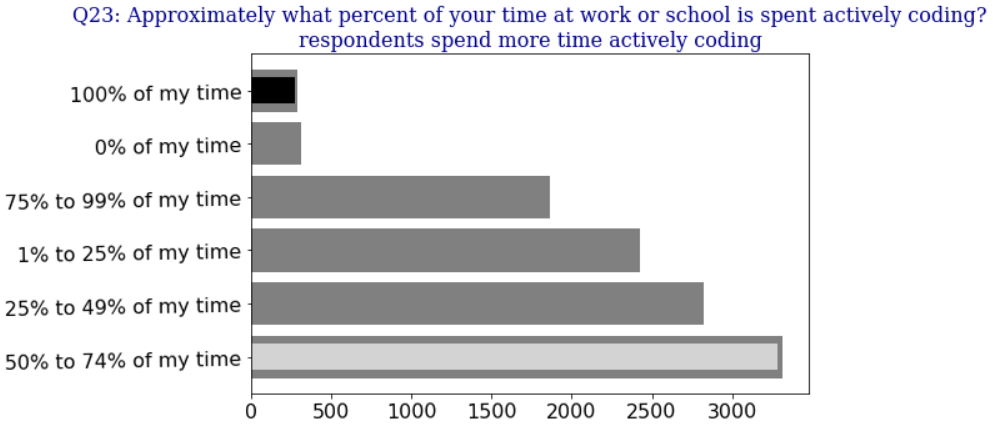
****

Fig. X. MSDA program “input”: undergraduate majors of Data Scientists.

Business, life and social science backgrounds can be in some cases less intense on the quantitative and programming preparation. However, as can be seen on Figure 2 from the results of Kaggle’s ML and DS 2018 industry survey, programming plays an important part of day-to-day activities of industry professionals, with the majority of responders saying that they spend 25% to 75% of their work time is spent actively coding. On the other hand, many candidates from classical technical background possess significant training in these disciplines, and might not want to spend as much time doing the revision on math and basic programming.



To address this gap, this project proposes an alternative approach which gives candidates introduction to math and programming for ML in a form of an optional summer bootcamp that ensures that all the students coming into the program, regardless of their background and experience, are equiped with the right tools to succeed in it. For details about the bootcamp, see Section X of this report.

**Section 2: program body -- MSDA program structure**

The Master of Data Science and Analytics (MSDA) consists of 3 main elements: 4 manatory core courses, 4 electieve technical courses (at least 2 must be picked), and 4 hands-on projects. The students are required to complete at least 6 courses and all 4 projects, final of which is the capstone project. Capstone project acts as the main evaluation of a prospective candidate at the end of their program, and goes on to become their main portfolio work to date, as Kaggle’s survey shows that industry professionals value personal projects as much, or even more as academic accomplishments, when assessing Data Scientist’s proficiency.

Kaggle survey results show that while Data Science reamins to be a truly diverse and evolving filed, there are clearly some industry “staples” when it comes to technology, such as use of Python, matplotlib, scikit-learn, TensorFlow and Keras. Survey also shows that soft skills, such as communication and management, come up on Indeed.ca job postings clustered together with technical skills, such as programming and ML. A good graduate program would take all of these factors into account, recognize the core stable elements that provide foundations, while updating the “shell” of skills and tools around them to stay relevant in the field for years to come.

The four mandatory courses are focused on building background for candidate’s core technical competencies: they cover Algorithms and Programming, Machine Learning (ML) and Data Mining (DM), Applied Statistics, Probability, and Linear Algebra, and RDBMS and massive databases. This mix is intended to bring math and computer science together to build strong foundation for advenced skillset required for an aspiring Data Scientists. It also takes hints from the job market by including RDBMS and massive datasets in the mandatory core program, as SQL/databases is the most frequently mentioned skill, according to scraped Indeed.ca’a posts for Data Scientist and Data Analyst. This is not surprising, given the pervasiveness of relational databases in many industries, and the trend towards Big Data applications.

The four elective courses offer candidates an opportunity to expand their interest to a particular sub-field of Data Science, such as Natural Language Processing (NLP) or Time Series Analysis. Each of the elective courses also comes with a practical component, such as modeling or analyzing social media data. Elective courses are intended as a bridge from the theoretical core to hands-on projects, which in turn act as a bridge from academia to the industry. Projects are based on industry parterships, and involve working in teams to solve real world issues, while engaging with all main data types, as Kaggle’s survey shows that a typical Data Scientist handles a variety of data sources, and spends more time gathering and cleaning data, then building actual ML models. The program also includes two mandatory seminars on ethical, legal, and social aspects of working with data, and on algorithm bias, as many of Kaggle’s survey responders indicated that they recognize it as an important issue.

**Section 3: program body -- hands-on projects**

Ryerson University in Toronto is offering a MSc (Master of Science) program in Data Science and Analytics. This program is similar to what this section aims to design, so an analysis of Ryerson's program is required to improve the understanding of market and competition in DS and ML education. It could also offer insights into potential ways of facilitating DS and ML education.

Ryerson’s program combines lecture-based classroom and hands-on lab environments to give students exposure to real life data, projects, and organizations. Viewed from the point of view of the job market, this is the correct approach, since according to Kaggle’s ML and DS 2018 survey, the majority of industry professional in DS and ML view personal projects to be equally or even more important in assessing candidate’s profficiency when compared to academic accomplishments alone (as can be seen on Figure 3).

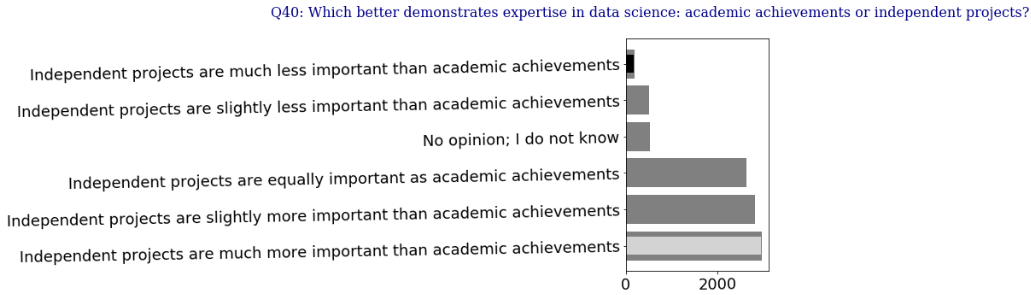
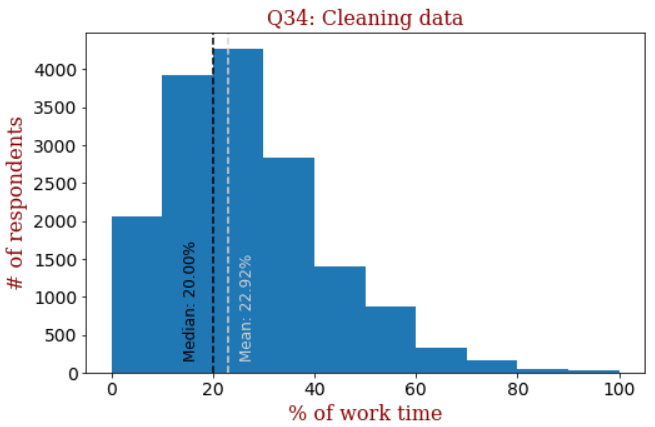
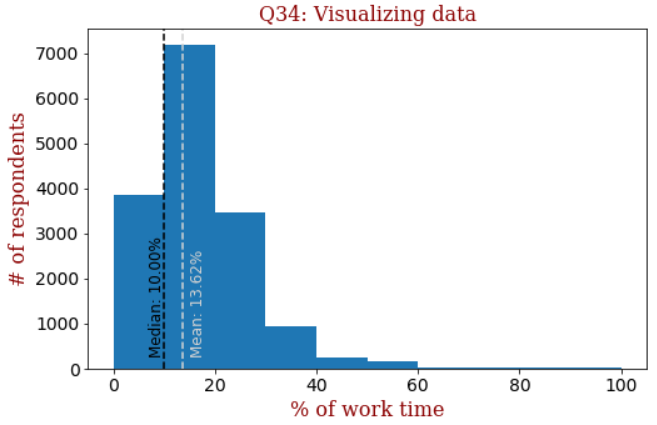
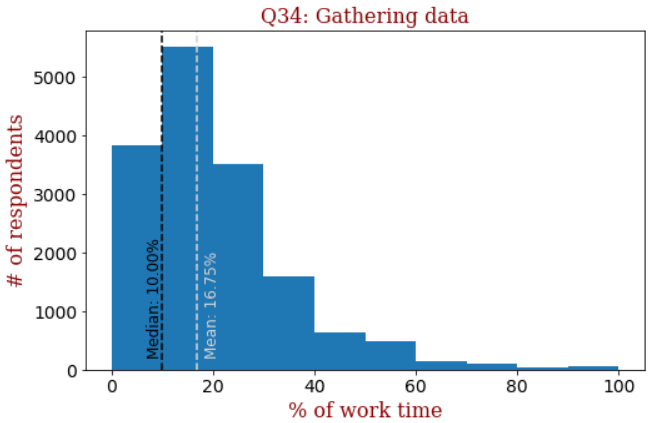
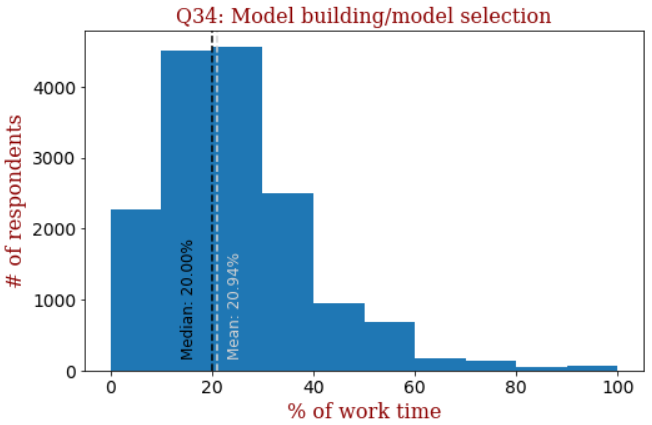
****

Figure 3. Most respondents on Kaggle’s ML and DS survey view individual project as more important when compared to academic accomplishments

In terms of tasks that students should be focused on during the projects, Kaggle’s ML and DS 2018 survey offers insights into what amount of their work time do industry professionals spent performing particular tasks. The top 4 tasks for professional Data Scientists are:



In terms of the tools and skillset that the hands-on part of the MSDA program will focus on can be established by looking at the Hierarchical Clustering data obtained from the scraped DS and ML job postings from Indeed.ca. According to the data,

In terms of data types, many Kaggle survey responders encounter all of the main data types together, rather than specializing on one particular data type, so all of them should be covered during the program practicums.



Figure 5. Par of the Dendrogram from Kaggle’s survey results

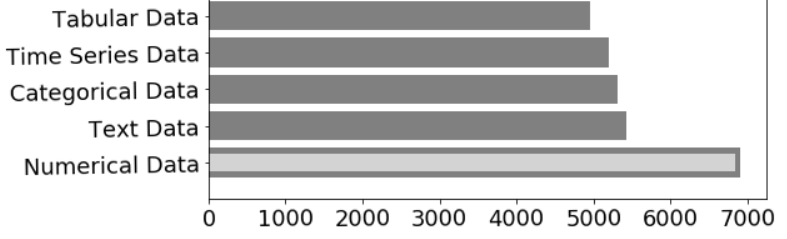


Figure 6. Par of the Dendrogram from Kaggle’s survey results

In addition to that, as can be seen from scraped job data from Indeed.ca on Figure 8, the most frequent skill that appears on the job postings is SQL and databases, indicating that this area of expertise posseses significant relevance to Data Scientists working in a large number of organizations, and should therefore be covered heavily within the program curriculum.

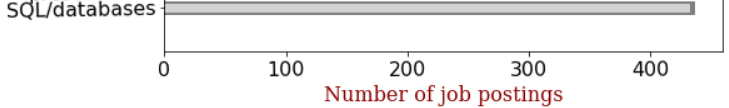
****

Figure 8. Most frequent skill appearing on Indeed.ca’s job postings is SQL and databases

From the point of view of particular Python libraries and ML frameworks to be used during assigments and projects under MSDA program, the list of the most relevant frameworks can be determined from the dendrograms and Kaggle survey resutls, as was done in Section X of this report.

**Section 4: program body -- soft skills**

In addition to that, as Hierarchical clustering of Data Scientist and Data Analyst job postings from Indeed.ca shows, even for technical specialists in DS, combinations of soft skills, such as presentation, management, and consulting, appear with technical skills, such as Python, R, and Data Mining. Soft skills seem to play an important part in Data Scientist’s proficiencies, according to job postings from Indeed, so these components should also be introduced in all the projects undertaken by the students as part of MSDA program, along with the core technical competencies.

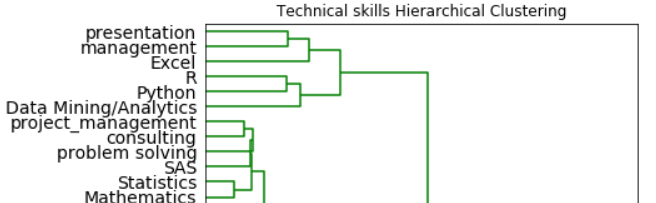


Figure 7. Part of dendrogram from Indeed.ca’s technical DS job postings -- soft skills in the same cluster with technical skills

**Management Program:**

Total of 7 courses

4 Core courses

3 Electives

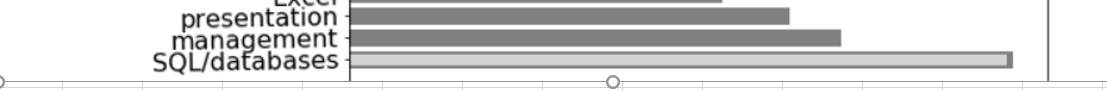
1 Internship or capstone project

4 Core Courses:

**1- Introduction to Management and Leadership**

* The main functional areas of business including strategy, operations and finance.
* Introduction to critical thinking and problem solving techniques
  + Root cause analysis
  + Bottleneck analysis
  + Problem statement formulation
* Understand the concept of emotional intelligence and its role in leadership effectiveness
* Introduction to Excel
* Mindful coaching
* Managing technical teams
* Introduction to software development life-cycle SDLC

Notes: This course is designed based on the findings from scrapping the job titles related to management analytics. It was observed that management and presentation/communication skills showed up in majority of the postings and were clustered together by the hierarchical clustering algorithm.

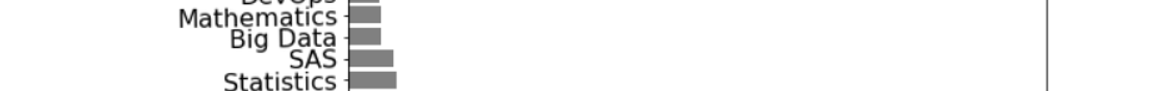
****

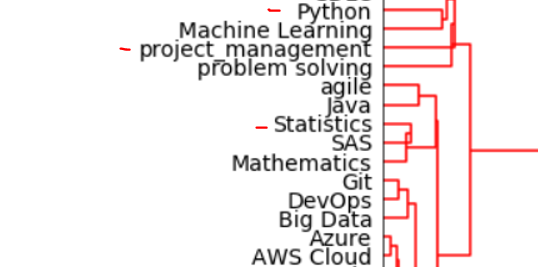
****

**2- Introduction to Statistical Data analysis**

* Introduction to descriptive statistics (mean, median, range, standard deviation)
* Exploratory data analysis
* Histograms, Percentiles, and Outliers
* Probability theory
* Random Variables
* Probability Distributions (Normal, Poissons, Uniform)
* Central Limit theorem
* Hypothesis Testing
* Introduction to regression
* Introduction to SAS and Python

Notes: This course was selected to be a core course based on the frequency of occurrence of the skills related to statistical analysis, mathematics, and statistical software resulting from scrapping the job titles related to management analytics.

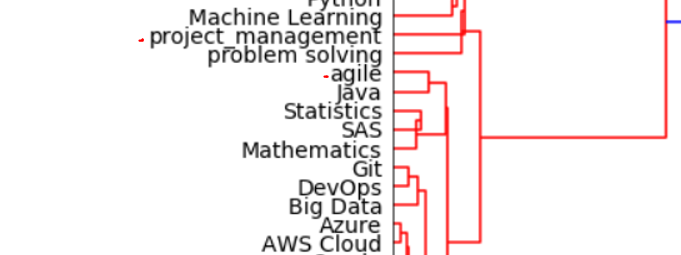
****

****

**3- Agile Project Management**

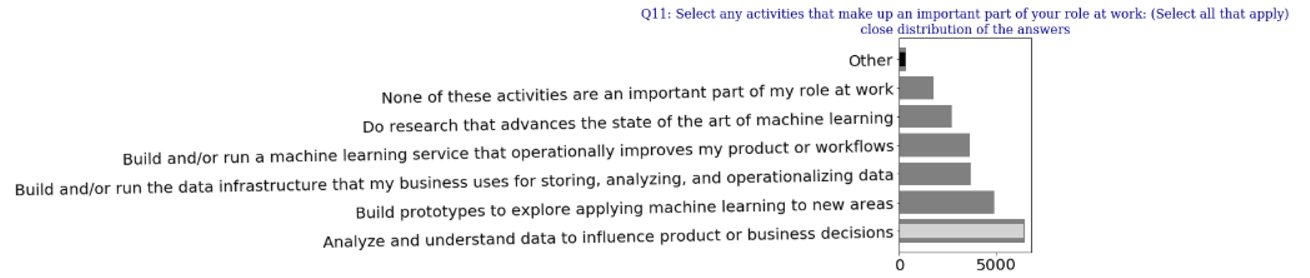
* Translating ideas into business applications through contemporary project management and software development techniques.
* The “Agile” approach to software development.
* Enabling teams to break down complex applications into simple components that are iteratively developed through Agile in complex and frequently changing business environments.
* Agile processes and techniques, design approaches, data requirements, test and deployment.
* Managing uncertainty and risks associated with AI implementations.

****

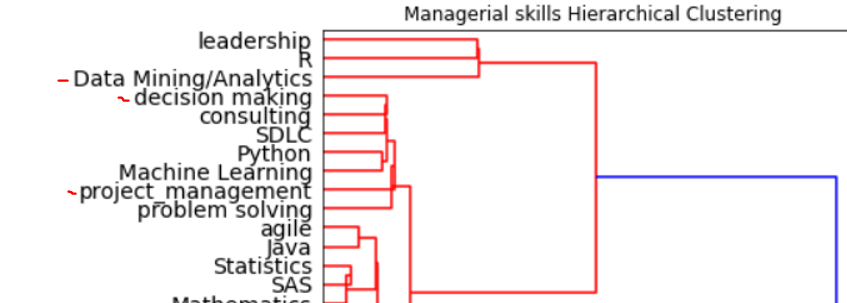
****

**4- Analytical Decision Making**

* Fundamental concepts analytical thinking.
* The use of analytical methods in management problem-solving.
* How to construct an analytical model of a problem that can be used to identify a decision.
* Decision-making frameworks including optimization, simulation, and decision analysis.
* Understanding variation, perceiving relative risk of alternative decisions.
* The challenges of communicating and implementing results in an organizational context.

****

****

****

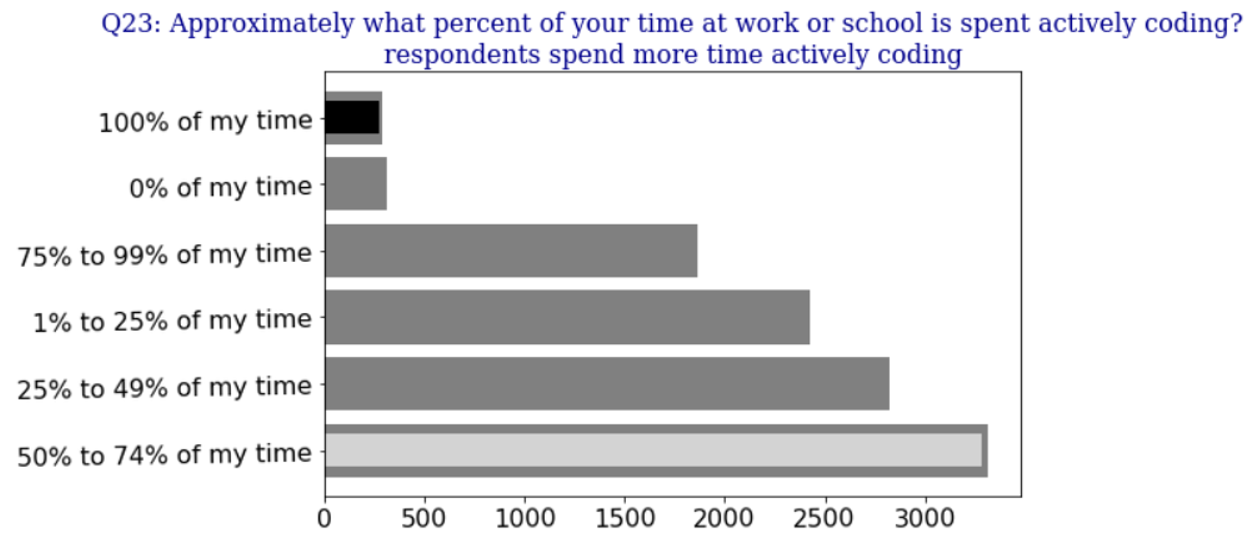
**List of elective courses:**

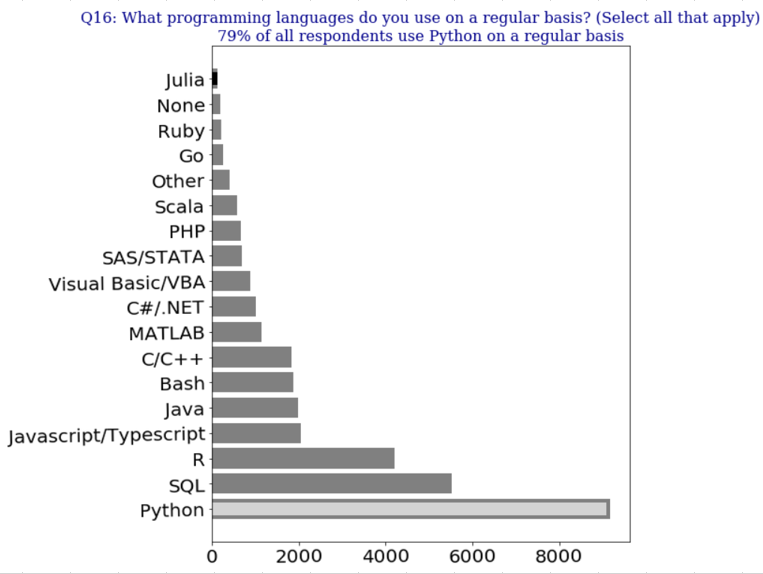
1. **Practical Data Analytics:**

This course will expose the learner to a broad range of technical skills that are required to prepare data for advanced analysis. Using a combination of theory and practical exercises and case studies, the learner will develop the data acquisition and preparation skills that are a necessary pre-requisite to applying advanced statistical modelling, data mining techniques, or machine learning algorithms to their data. Demonstrate the ability to prepare, explore and validate sample data for advanced analysis.

This course provides practical coding exercises in Python to familiarize the students with the most popular coding language in the industry.

Note: This course is designed to provide the students with practical experience working on data analysis project using Python

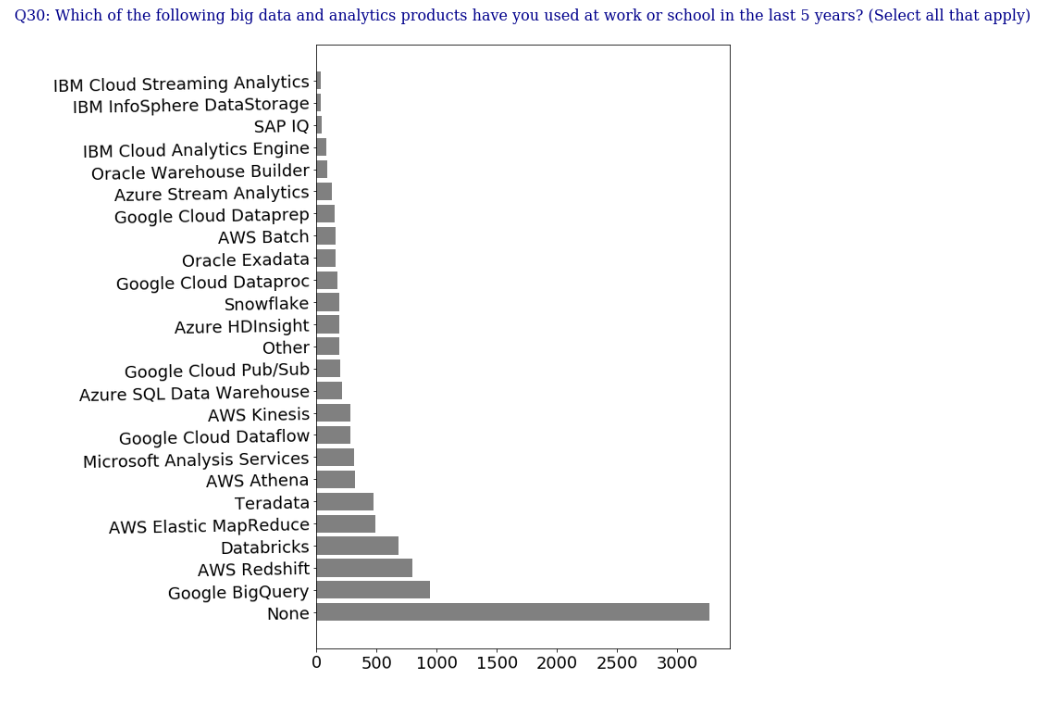
****

****

**2. Big Data Analytics**

This course will introduce the students to a diverse uses of big data techniques. These techniques are often aimed at identifying and quantifying various structures in the data, What are the key similarities between certain business units with respect to customer satisfaction, What are the characteristics of important customer segments. Model validation and effective communication of model-based results will be stressed. This course will introduce the students to a diverse uses of big data techniques. These techniques are often aimed at identifying and quantifying various structures in the data, What are the key similarities between certain business units with respect to customer satisfaction? What are the characteristics of important customer segments. Model validation and effective communication of model-based results will be stressed.

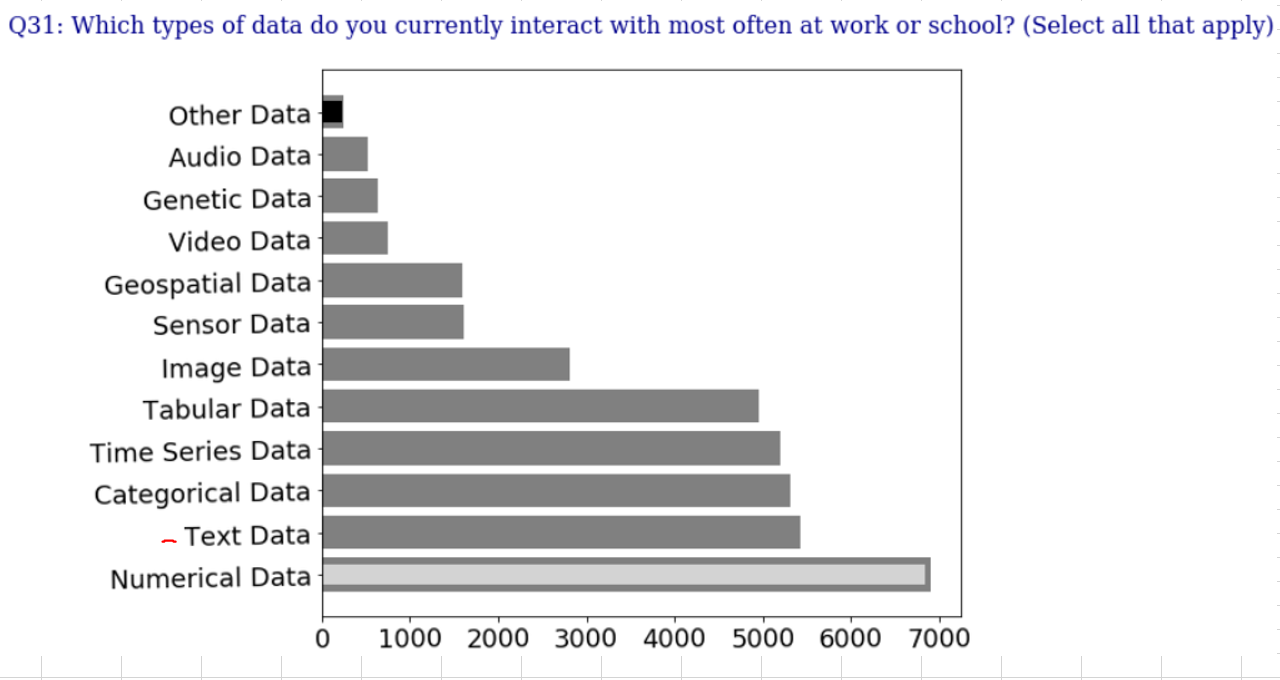
Note: This course was chosen to be an elective based on the fact that most respondents to the Kaggle survey responded to have used “None” of the big data products in the survey question below.

****

**3- AI and Analytics:**

* APIs and Google Analytics.
* Hadoop for Analytics.
* Text mining and Sentiment Analysis.
* Graph Analysis in Machine Learning.
* Neural Networks I: Tuning and Topology.
* Current topics and themes in Machine Learning.

****

****

**4. Improving Customer Value with Analytics**

* How to use data from multiple sources to make more effective business decisions
  + Sales data, historic consumption data, transactions data, and marketing data.
* Understand the basic principles of data-driven marketing in financial services and banking.
* Applications in targeting decisions, segmentation decisions, and customer relationship management.
* How to allocate marketing budget across acquisition and retention.

**Internship or Capstone Project:**

**Option 1: Capstone Project**

The capstone project will allow students to demonstrate and apply their AI management learning by focusing on an area of interest including potential projects to pursue after graduation. Students will be supported by a capstone mentor as a guide and domain expert, and research assistants.

**Option 2: 4 months Internship**

A student may choose to gain real life experience by working in industry for 4 month working in a role relevant to the program (i.e. business analyst, data scientist, data analyst, etc.)

**Course Curriculum Design:**

**10 chapters:**

**Course Curriculum Design:**

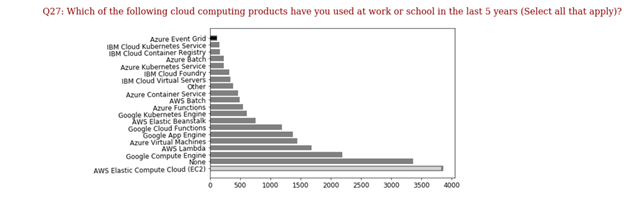
**10** **Chapters:**

## **1.** **Introduction to Data Science and Analytics:**

**a. Applications of data science**

**b. Data science concepts**

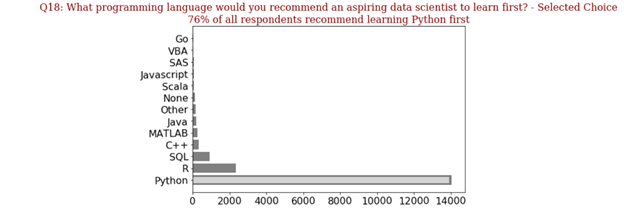
**c. Overview of Cloud computing Services**

****

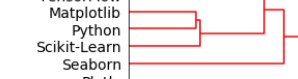
**Based on the Kaggle survey data, Amazon Web Services (AWS) will be utilized for the introduction to Cloud Computing Services.**

## 

## **2.** **Python for Data Science:**



**Based on the Kaggle data survey, Python is the most recommended first programming language. Therefore, all programming in this course will be performed in Python.**

****

**Furthermore, based on the clustering results from the Kaggle survey, these are the most mentioned libraries. These libraries will be focused on, in our course.**

## 

## **3.** **Basic statistics:**

**a. Introduction to regression**

**b. Probability Distributions (Normal, Poissons, Uniform)**

**c. Central Limit theorem**

**d. Hypothesis Testing**

**e. Statistics case studies in IPython**

## **4.Overview of linear algebra:**

**a. Linear algebra and matrix computations**

**b. Vector Spacec and linear transformations**

**c. Functions, derivatives, convexity**

## **5. Optimization:**

**a. Unconstrained non-linear optimization algorithms**

**b. Overview of constrained optimization algorithms**

**c. Optimization case studies in IPython**

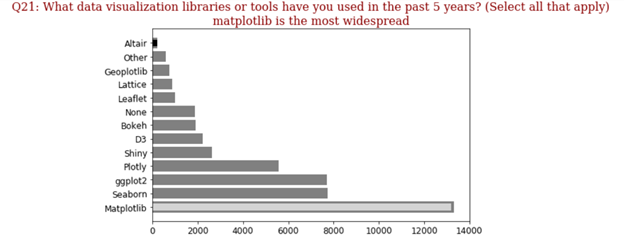
## **6. Data Visualization:**

**a. Matplotlib**

**b. Seaborn**

**c. Plotly**

**d. ggplot2**

****

## **7. Machine Learning:**

**a. Linear regression**

**b. Linear classification**

**c. Trees**

**d. Support Vector machine**

**e. Kernals**

## **8. Machine Learning (Continued):**

**a. Unsupervised learning**

**b. Validating and resampling**

**c. Optimization**

**d. Hyperparameter optimization**

**9. Simulation Modeling:**

**a. Monte Carlo simulations**

**b. Real world simulation modeling case studies**

## **10. Cognitive computing and artificial intelligence:**

**a. Neural Networks**

**b. Learning and neural neural networks**

**c. Perceptrons / Multilayer perceptrons**

**d. Convolutional Neural Networks**

**e. Recurrent Neural Networks**

**f. Case studies with Neural Networks**

****

**Based on the Kaggle data Survey results, Keras and TensorFlow (Neural Networks libraries) are required tools for data scientists.**

**Data Science education EdTech effort**

Based on our assessment of the curriculums designed in the above sections (see section # “Profile of incoming candidates”), we’ve identified the following gaps that an EdTech startup may consider resolving to help aspiring and professional data scientists:

* Lack of experience and familiarity with relevant software languages, libraries and application programming interfaces (APIs)
* Lack of experience working with real world data and the full cycle of data consulting projects

In order to help aspiring data scientists learn relevant sofware the EdTech startup would focus on:

* Bootcamps to ramp individuals on relevant software languages, libraries and APIs
* Partnering with the following online learning platforms based on the following figure below:
  + Coursera
  + Udemy
  + DataCamp
  + Udacity

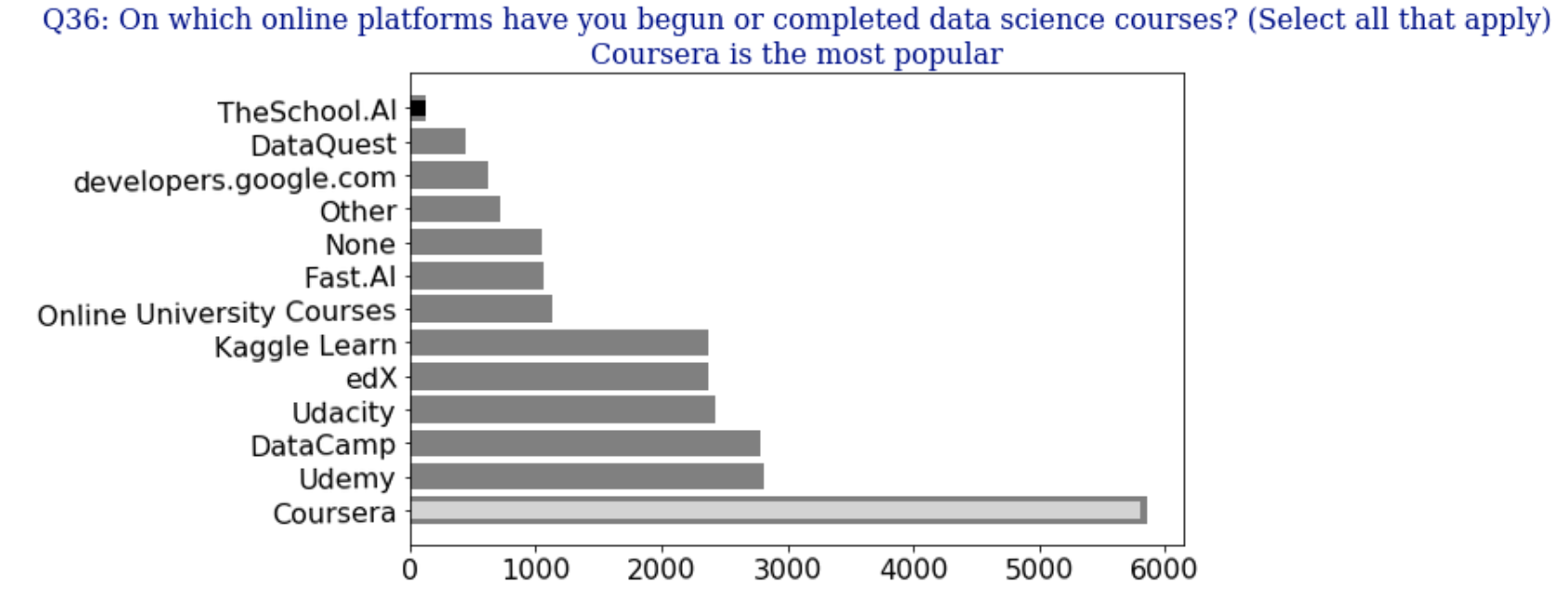


Figure #. Online learning platform data from Kaggle Data Survey

In order to help aspiring and professional data scientists gain invaluable hands-on experience in working with real world data and experiencing the full life cycle of a data consulting project, the EdTech startup would focus on:

* Hosting competitions with industry partners to solve problems from the following industries based on the figure below:
  + Computers/Technology
  + Academics/Education
  + Accounting/Finance
  + Online/Internet based serivces
  + Medical/Pharmaceutical

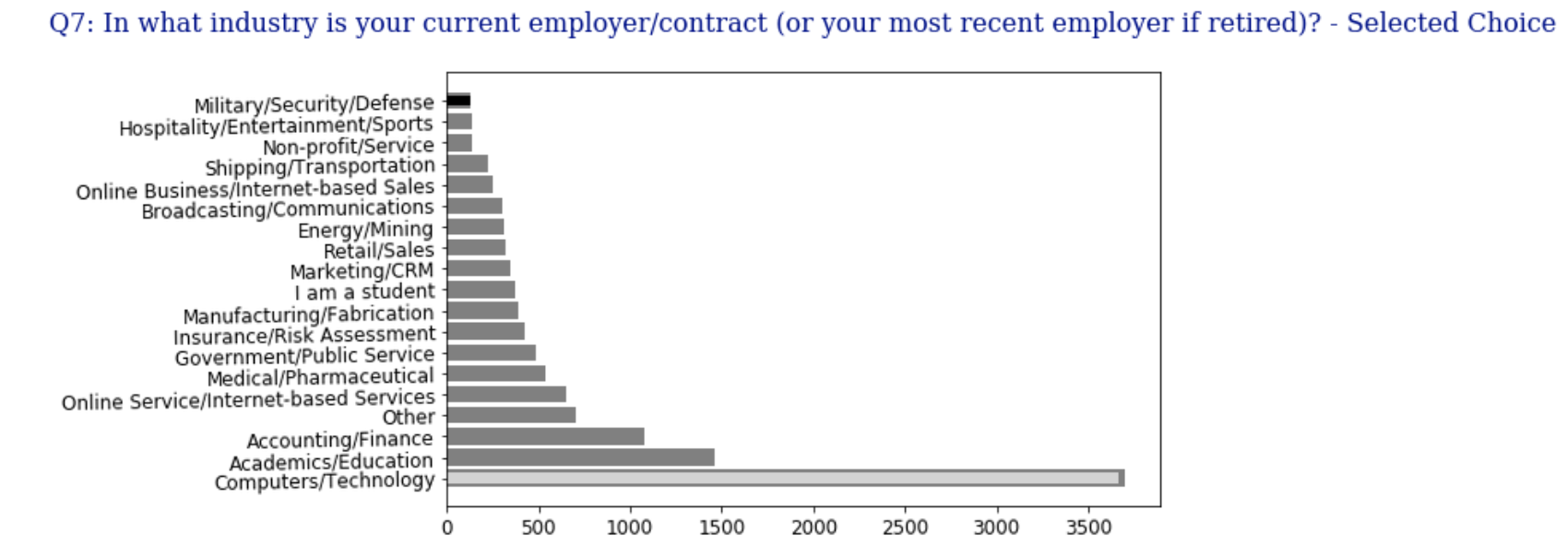
****

Figure #. Relevant datat science industries from Kaggle Data Survey