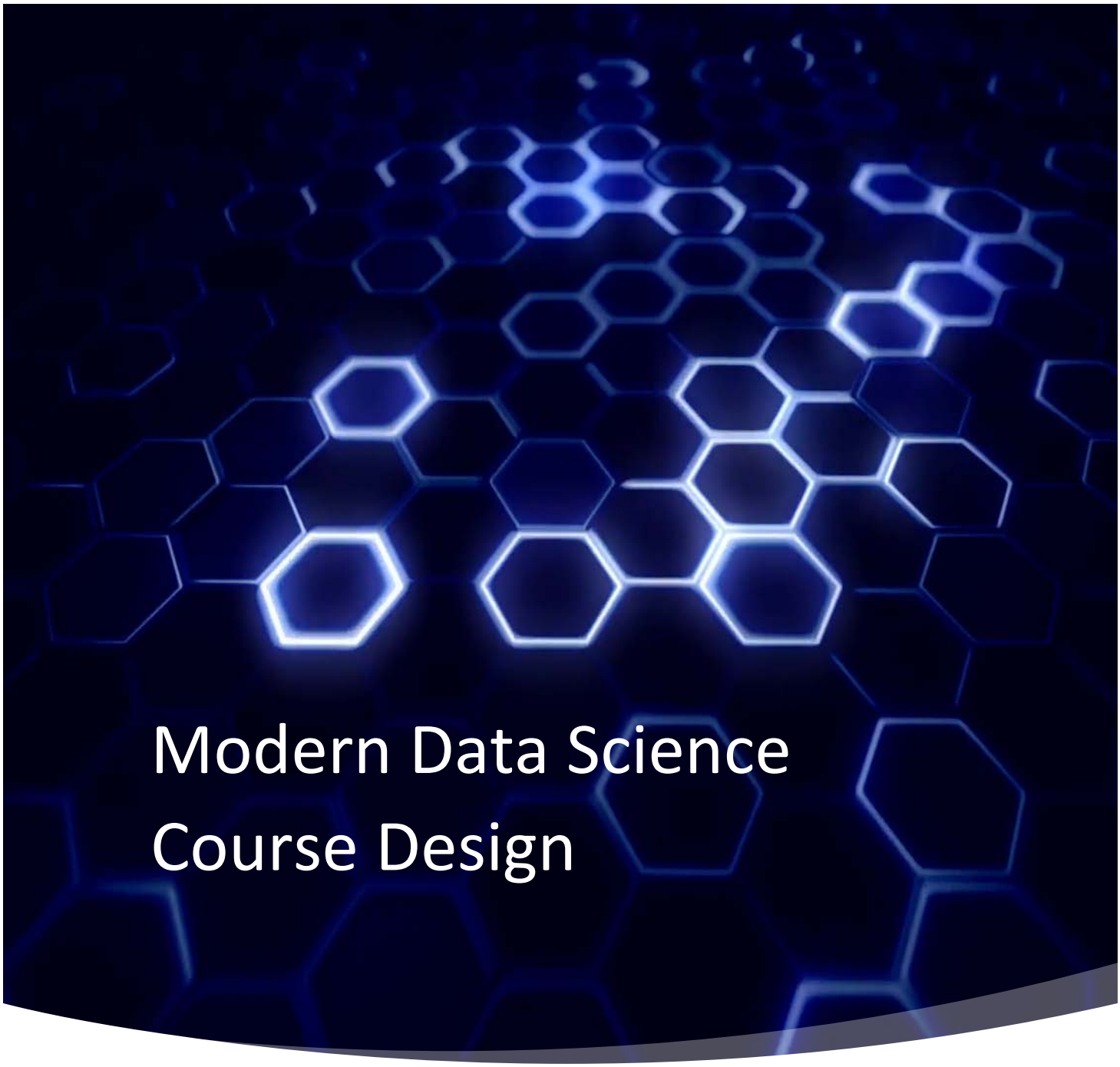


Modern Data Science Course Design

VAUGHN GAMBETA
BRENT HUCHUK
KEON YOUNG PARK
TARUN SINGHAL
PRASHANTH SRIDHAR
MOHAMMAD TAHLE

MIE 1624 INTRODUCTION TO DATA SCIENCE & ANALYTICS | Winter 2018 Course Project

Contents

Introduction and Motivation

There has been concerns over the lack of modern data science and analytics courses being offered at the University of Toronto. A recent analysis of the data science industry indicates a lack of skills from graduating students and a 500% growth in data science positions[1].

In order to remain competitive there is a need to re-design the current introduction to data science course and to introduce comprehensive Master's Degree programs to prepare students for the burgeoning field of data science and AI.

The Masters of Data Science & Analytics (M.D.S.A) will focus on the technical skills sought after in the industry while the Masters of Business and Management in Analytics and AI (M.B.A.I) will enhance the soft skills and technical skills of future managers in the field.

The objective of the report, summarized below, is to outline the current industry trends in data science and determine the skills and tools to incorporate into a program to ensure that what is being taught in the courses meets the needs of the industry.

#	Objectives	Outcomes
1	Current Industry Trends	Re-Design Intro. Course
2	Relevant Industry Skills	Technical Masters
3	Most Used Tools	Management Masters
4	Competitive Courses	Education Start-Up

To determine the skills deemed most important, 14,000 online job postings for the previous sixty days from across thirty major North American cities are analyzed. Also, the results from a 2017 third-party industry wide survey[2] is examined that was conducted on participants with careers or are students in the data science field. The survey produces real-world relevant tools and opinions within the data science field to assist in the design of the course and programs.

Further analysis provides deeper insights. For example, a classification of expected salary based on work assignment, found users spending a majority of their time working with production and visualizations were more likely to be in a higher income bracket (Appendix 8.0).

[1] Source: https://blog.linkedin.com/2017/december/7/the-fastest-growing-jobs-in-the-u-s-based-on-linkedin-data
[2] Source: http://www.kaggle.com/kaggle/kaggle-survey-2017/

What skills are currently being sought after in industry?

The job posting reveals that a number of key hard and soft skills overlap most positions. The industry survey reveals what the most commonly used skills, tools and methods are in the industry.

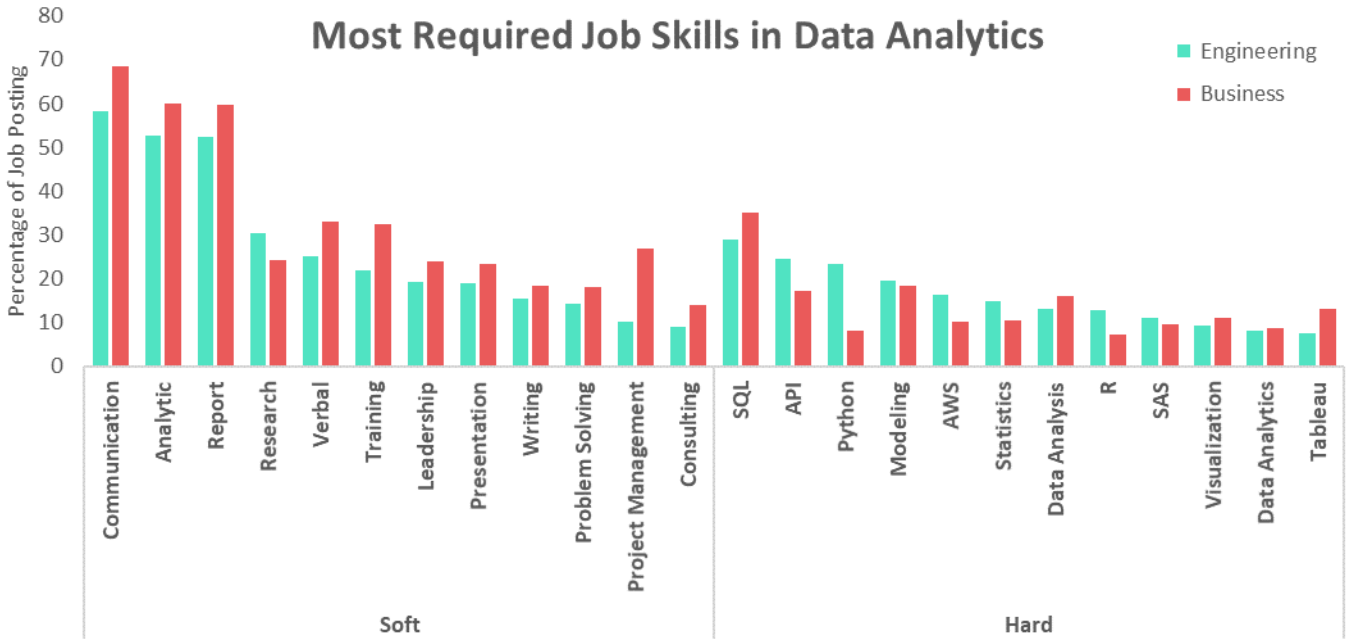


Figure 1: Job Skill Analysis - 14,000 Job Postings

Having strong communication, analytics and reporting skills ranks high for both business and engineering positions. The opinion of industry professionals is weighed heavily on the selection of topics and courses that will make up the new programs. These opinions reveal that machine learning, high-level programming and a broad range of methods are heavily used in practice.

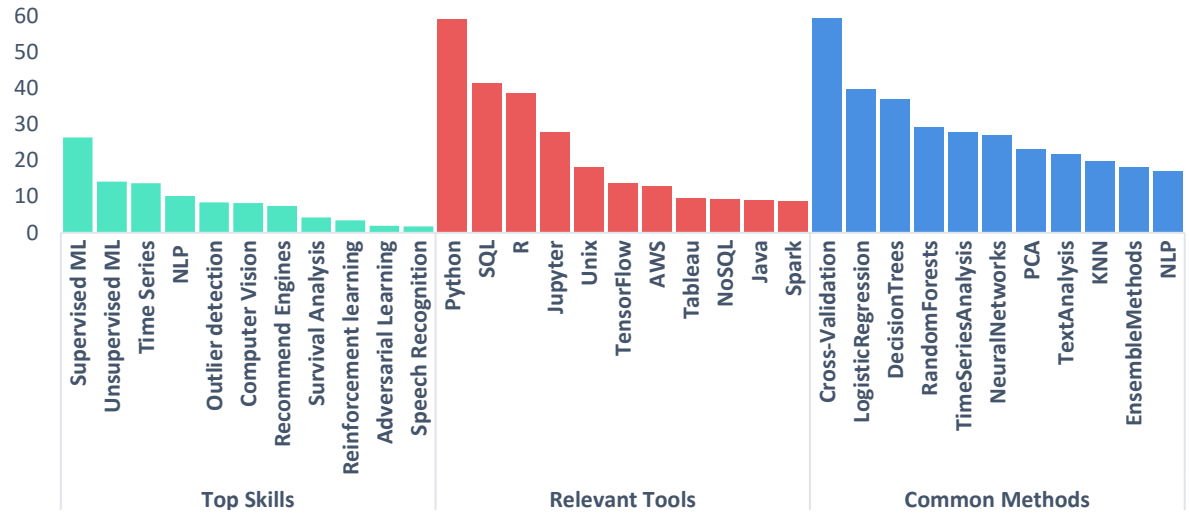


Figure 2: Industry Professional Survey Results (Source: Kaggle)

Is the introductory course according to market's needs?

The research conducted indicates that some changes to the course curriculum for MIE1624 – Introduction to Data Science & Analytics is necessary to align with the market needs. The desired changes are highlighted in green that the data indicated as highly utilized, while the red items indicates key topics required by industry but are already in the existing course design.

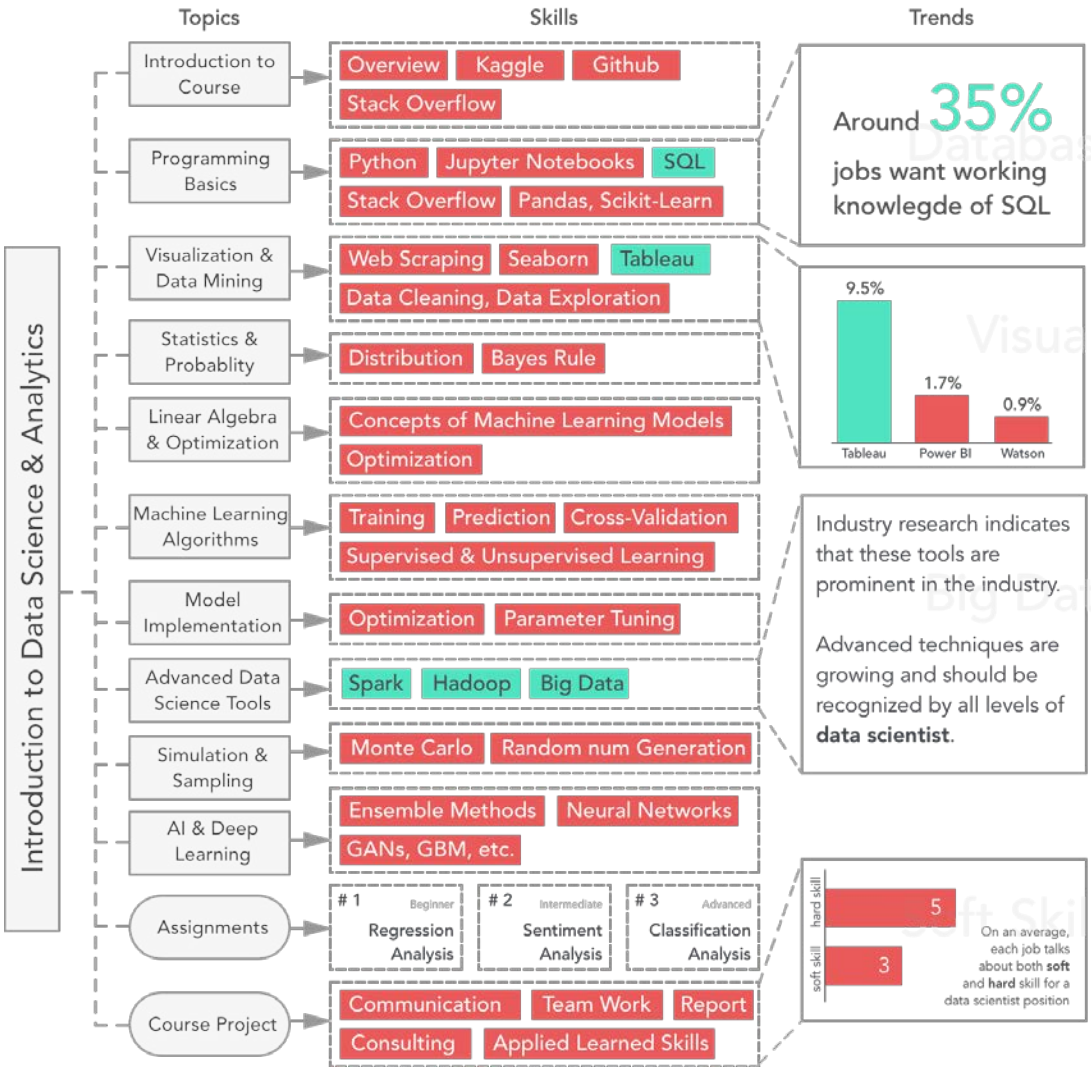


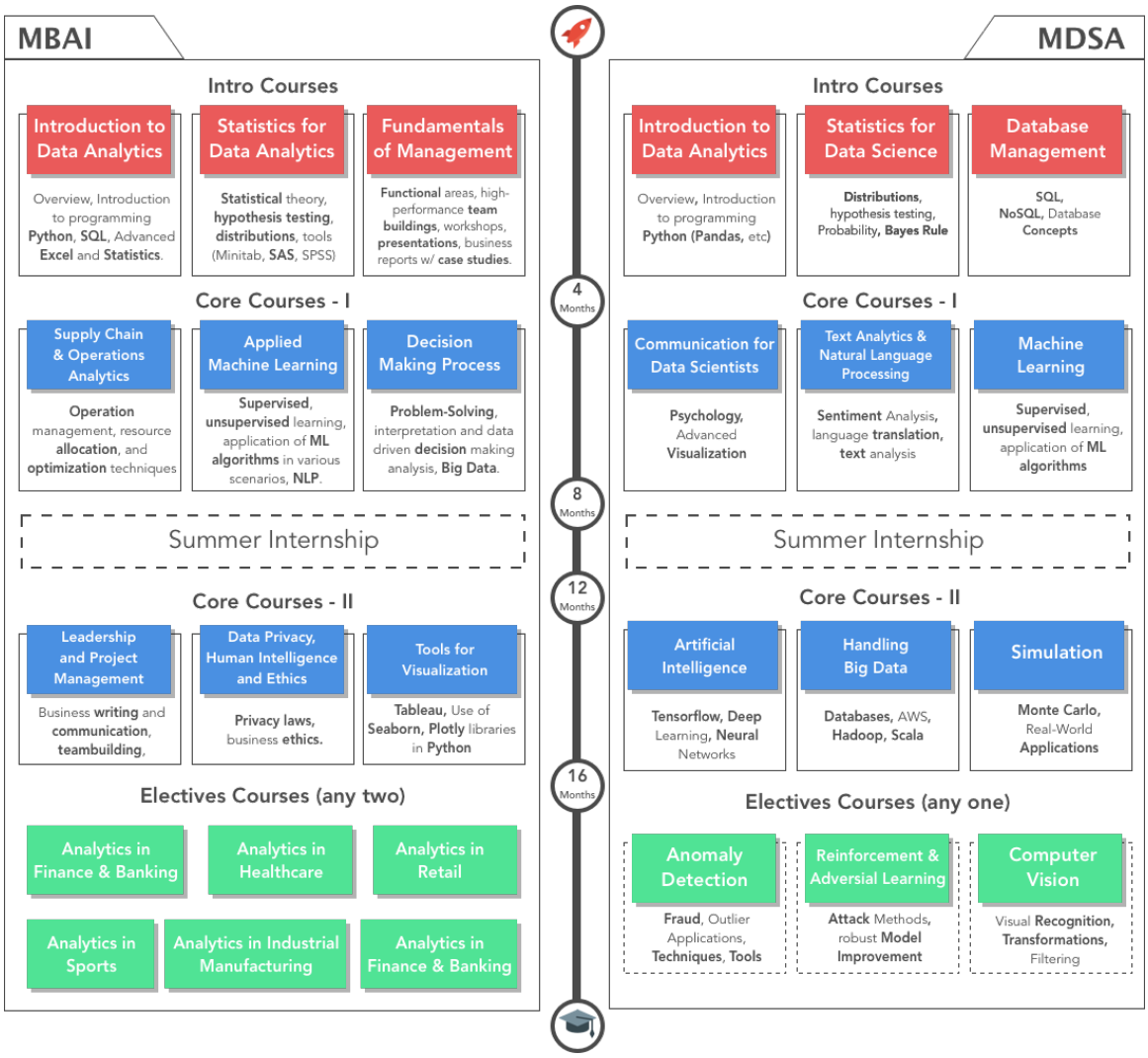
Figure 3 Course Outline MIE1624 - Introduction to Data Science & Analytics - Market Alignment

Natural language processing, classification and prediction with time series analysis are each some of the highest ranked skills according to the data (Appendix 2A) and are applied in the assignments. The survey highlights that supervised machine learning is the most competent area of professionals (Appendix 2A) and all data scientists should be introduced to this topic.

Big Data is a top-rated skill in getting employed (Appendix 1A) and utilizes Hadoop and Spark (Appendix 1B). These concepts should be recognizable by all modern data scientist and will be included here as a formal topic. The above course design will prepare students for an entry level position across a broad range of industries and with the recognition of advanced tools.

How do we prepare the next generation of data scientists and managers?

Introducing Master of Business and Management in Analytics and AI (MBAI) and Master of Data Science and Analytics (MDSA) programs will extend the competitiveness of the university and the graduating students. The following program designs extend deeply into relevant topics according to the industry data analysis and competitive benchmarking with other programs.



Master of Business & Management in Analytics & AI (MBAI)

The master's program focuses on training the essential software tools used in the industry for data science and analytics. The focus will be applying leadership and effective team building (Figure 1) to reach desired goals. It provides technical expertise to blend business and analytics.

Master of Data Science and Analytics (MDSA)

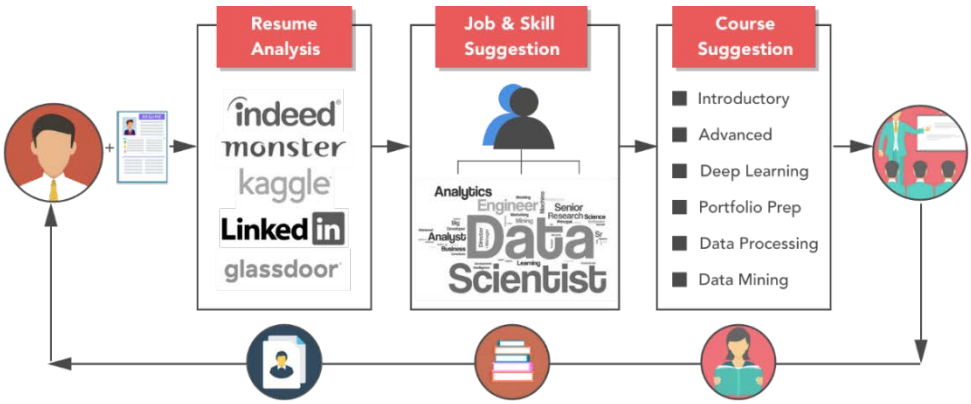
This course extends the level of detail from the introduction topics, specifically to overcome technical industry challenges (Appendix 6A) such as advanced data cleaning. The program

focuses on the top items in work skills (Figure 1), relevant tools (Figure 2) and most common methods (Figure 3) used by industry professionals.

Both masters programs have a mandatory 4-month internship that will provide real-world experience sought after in industry (Appendix 2B). Each master's program includes elective courses to provide students the opportunity to focus their interests.

Data Scientist in 30 days – D'eXpert

D'eXpert wants to help prepare and further develop careers of data scientists. Achieved by analyzing candidate resumes based on the most recent trends in industry and recommending skill gap improvements while providing courses and online resources to fill those gaps.



The majority of industry professionals and students indicate that they use popular platforms to enhance their data science skills (Appendix 3B).

D'eXpert will adopt various Kaggle competitions as its course curriculum (Appendix 4) by providing detailed breakdown of each data analytics topic and tool used to solve. Utilizing online courses (i.e. Coursera) and videos (i.e. Youtube) to train out candidate gaps.

D'eXpert bots will scrape data from major job postings and social media to have up to date industry requirements and trends in order to suggest *personalized* courses for each candidate. It will also offer individual courses for specific skills that are essential in data science such as data preprocessing, advanced visualization and web scraping for those who only want to focus on those particular topics.

D'eXpert prediction engine leverages temporal jobs and skills data and forecasts future demands. This not only helps candidates to learn skills based on the current markets needs but also prepares them for the skills that may be in demand in future.

Conclusions

As the data science field expands, more effort will need to be exerted to prepare future data scientists for success. These proposed programs will touch upon all current areas of data science from the introduction course through to the detailed master's programs. Armed with a thorough set of skills and view of the field, graduating students will better serve the companies whom hire them.

APPENDIX

Appendix 0 - Program & Data Science Career

The industry wide survey of data science students and employees provides an insight into the satisfaction and job salaries to be expected. These two metrics are important to the measure of success in this field. The satisfaction within the field appears high based on the self-reported survey data (Appendix 5). It should be cautioned however, that the data collected is only for those in the field and we are unable to directly compare to other professional categories. The salary distribution across job titles is broad and most job titles do reach in excess of $100,000 salary range (Appendix 7B). Most notably are the following careers,

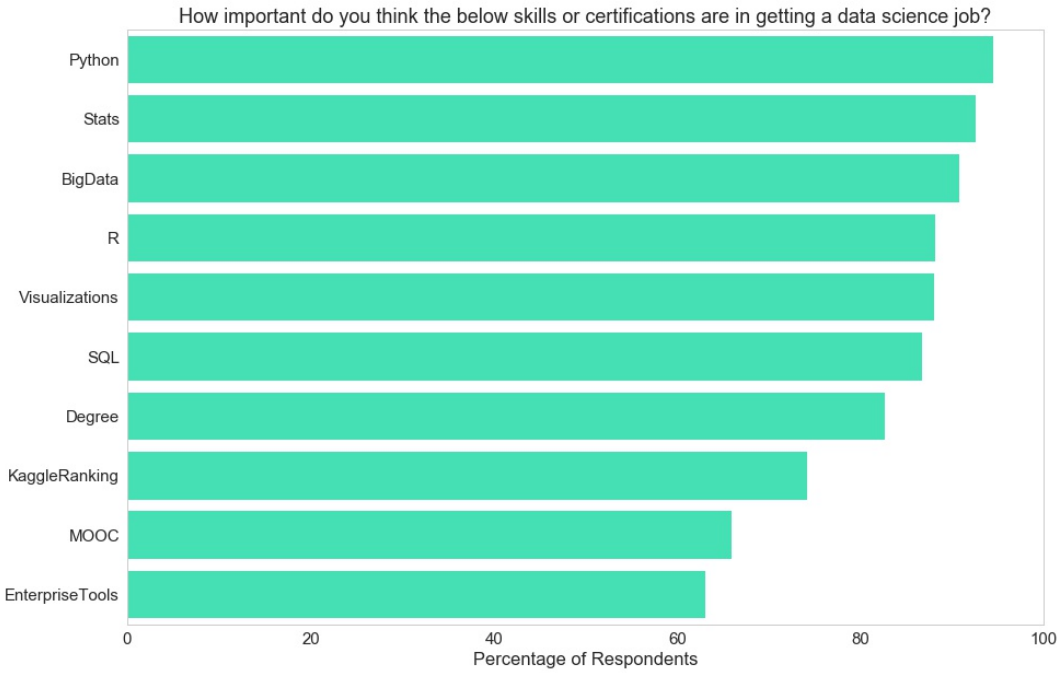
Data Miner, Data Scientist, Machine Learning Engineer, Predictive Modeler, Software Developer

These positions make up a large proportion of the data science industry and will be the focus of what skills will be taught in the introduction course and technical masters.

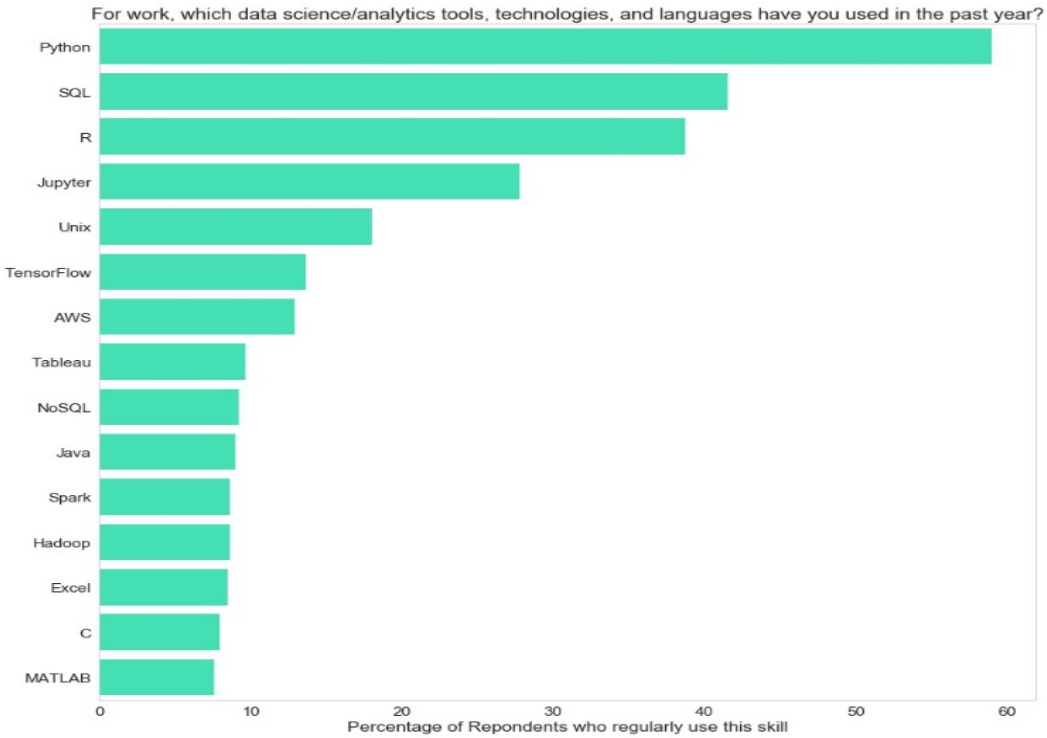
The recommended length of time spent studying to learn the skills required in practice is 1 -2 years from the opinions of professionals already in the field (Appendix 7C) and it is recommended that a masters degree be acquired to be successful. This is justification that a 2-year masters program will suffice to learn the required skills needed to be successfully in the field. The design of the courses will take into account the recommendations of industry professionals to select appropriate course content and masters program courses.

The new programs will be fit into the current educational system structures examined at major universities offering data science degrees. The introduction course being designed will have a duration of 10 weeks and include all high-level concepts deemed most relevant through industry research. The Master's programs will consist of 10 – 12 courses with up to 10 weeks of study with a master's project per course and could include a co-op or internship.

Appendix 1.0

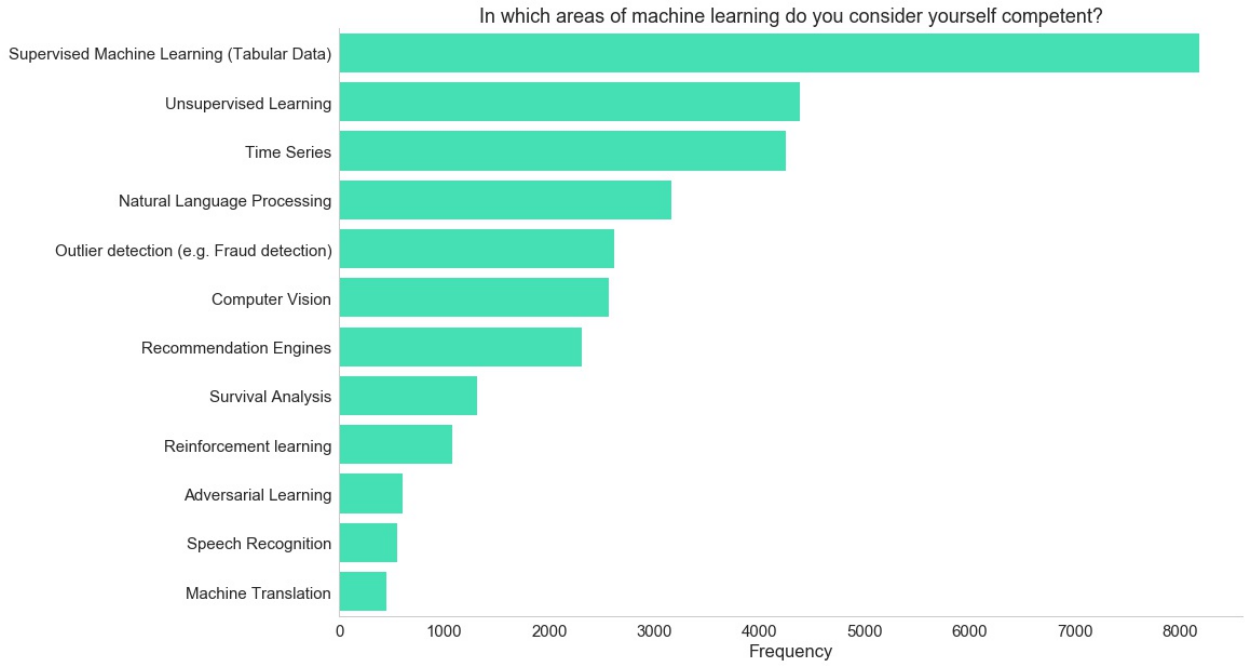


Appendix 1A: Important Skills & Certifications

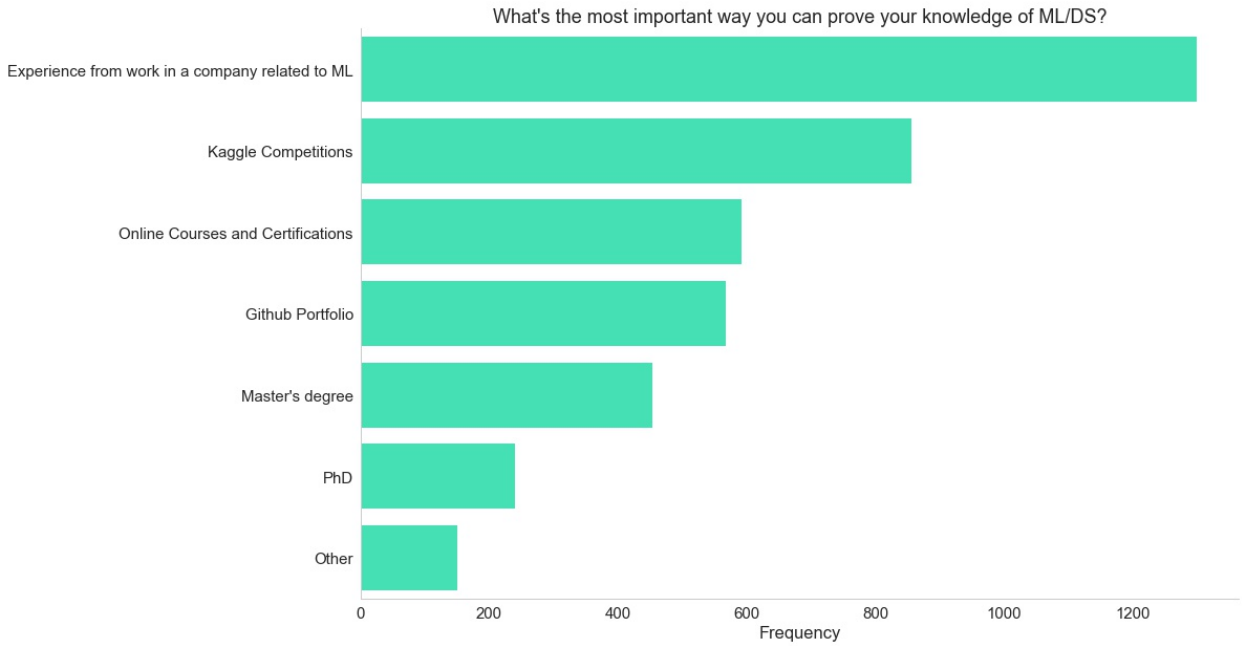


Appendix 1B: Most Common Used Tools, Technologies & Languages

Appendix 2.0

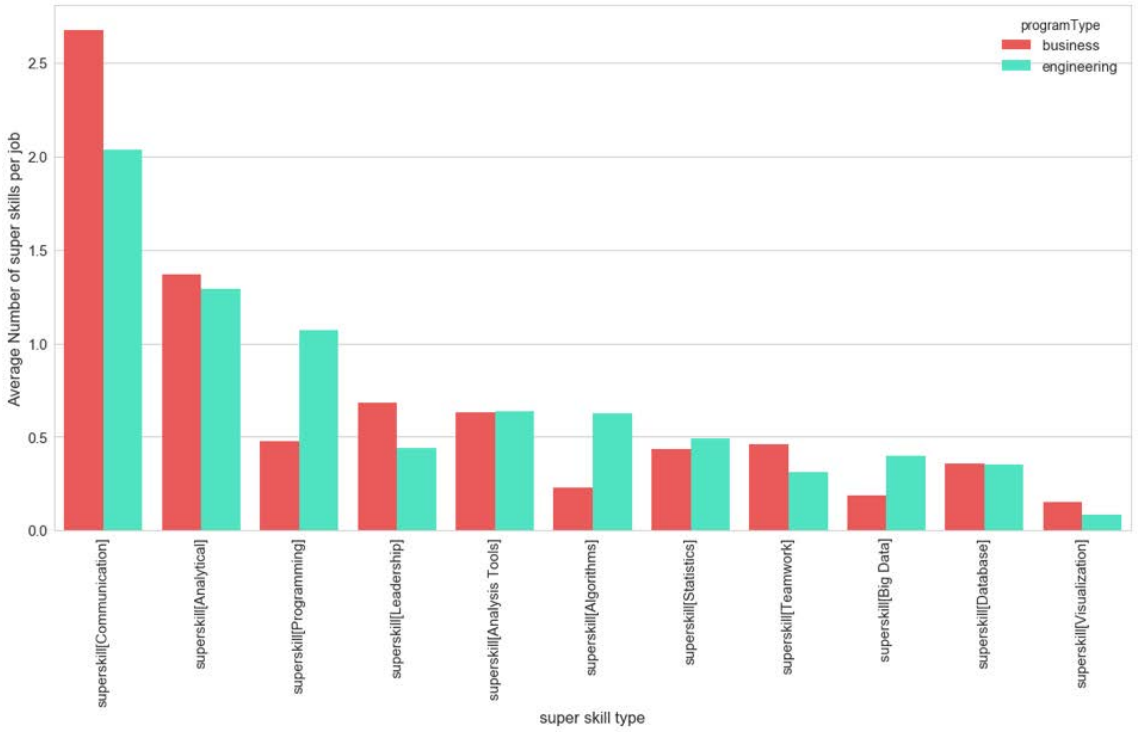


Appendix 2A: Competent Machine Learning Skills (Kaggle)

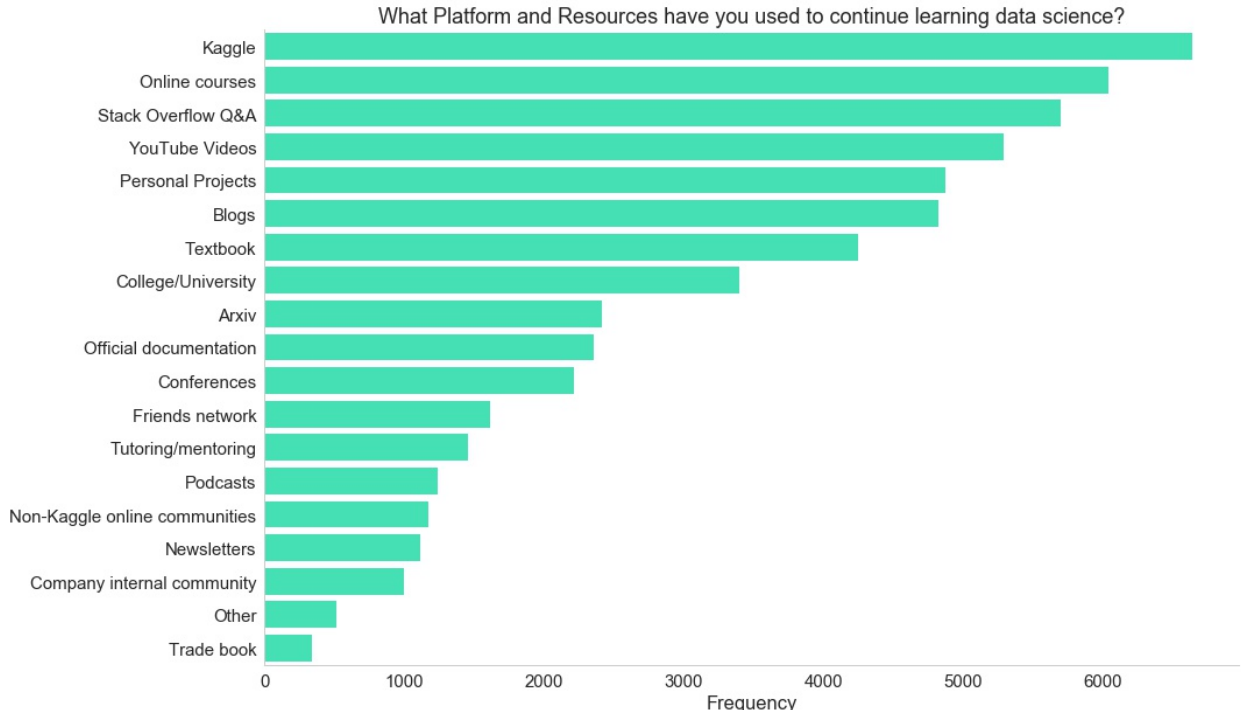


Appendix 2B: Most Important Way to Prove ML Knowledge

Appendix 3.0



Appendix 4A: Super Skills Classification & Frequency per Job

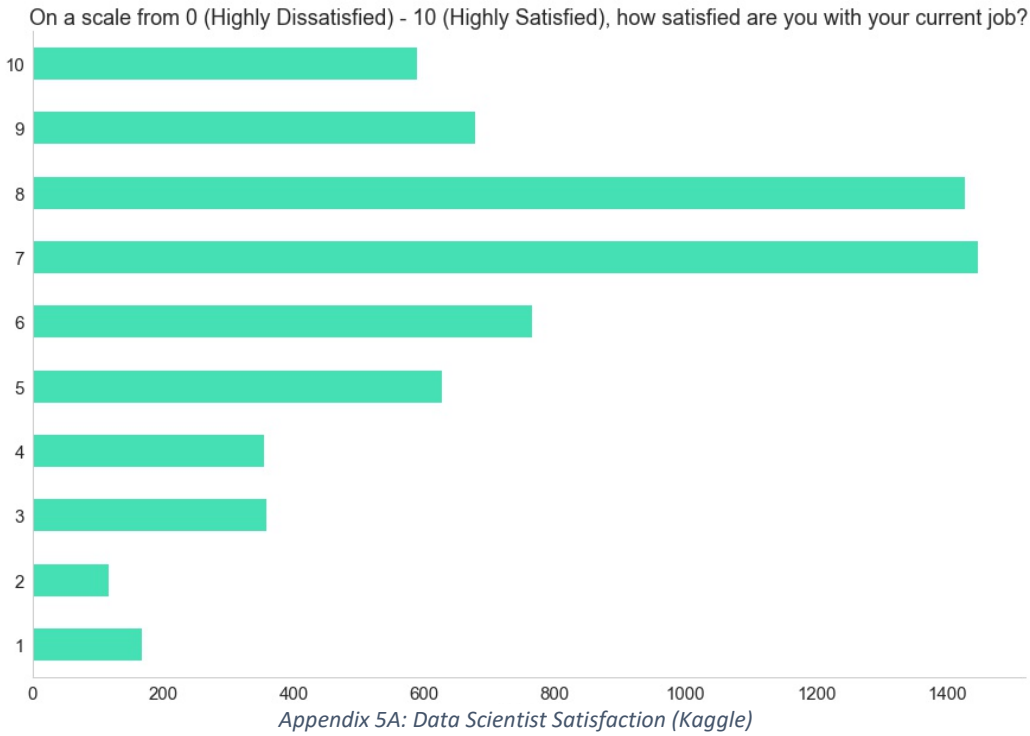


Appendix 3B: Popular Platforms & Resources to Learn Data Science

Appendix 4.0 – D'eXpert Course Curriculums

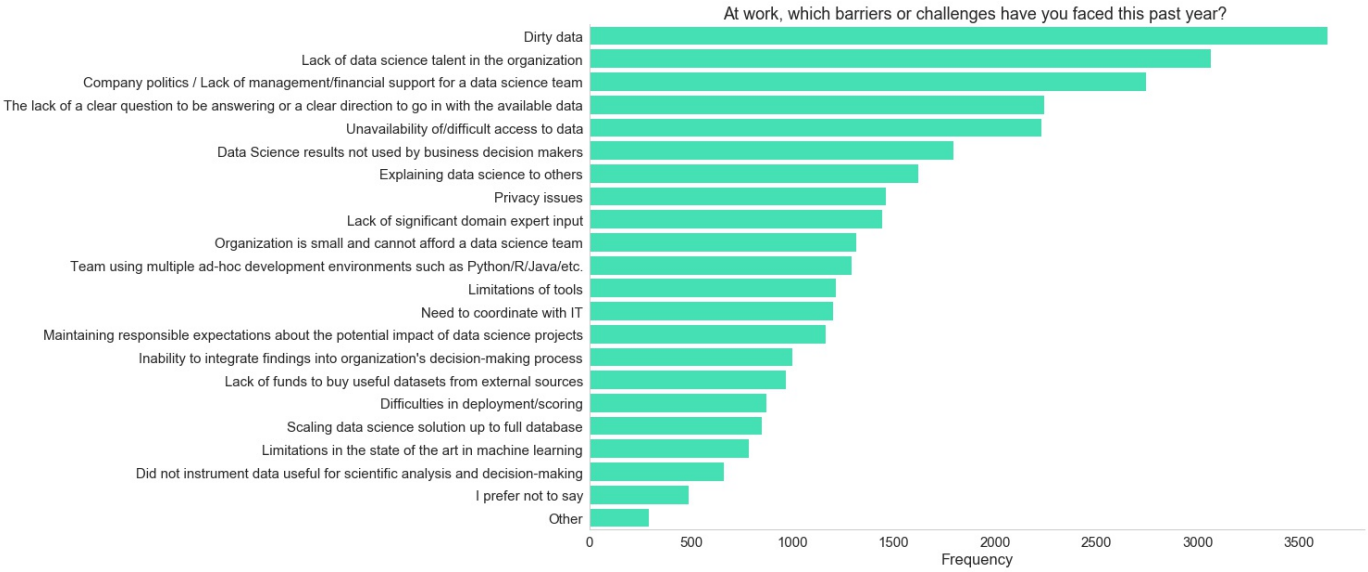
D'eXpert Online Courses	Course Description
Beginner Course (4 weeks)	• Basic necessary concepts in linear algebra, statistics, programming • Detailed Tutorials on 2 Kaggle competitions (basic level)
Intermediate/Advanced Course (4 weeks)	• 3~4 detailed tutorials on Kaggle competitions • detailed strategies on model improving skills and getting high scores on Kaggle • Portfolio preparation guideline
Deep Learning Course (6-8 weeks)	• Neural networks, deep learning tools, image processing techniques • Unsupervised learning • Real world data analytics problems using Deep Learning
Data Scientist Prep. Course (12-16 weeks)	• Targets people who have no background in data science but wants to develop a career in data science industry • Includes beginner, intermediate/advanced, deep learning course • Full preparation of data analytics portfolio for job interviews
Data Management/Preprocessing Course (2-3 weeks)	• Tutorials on specific libraries such as Pandas which is useful in handling big data • Advanced Data Visualization
Web Scraping Course (2-3 weeks)	• Tutorials on web scraping specific libraries

Appendix 5.0



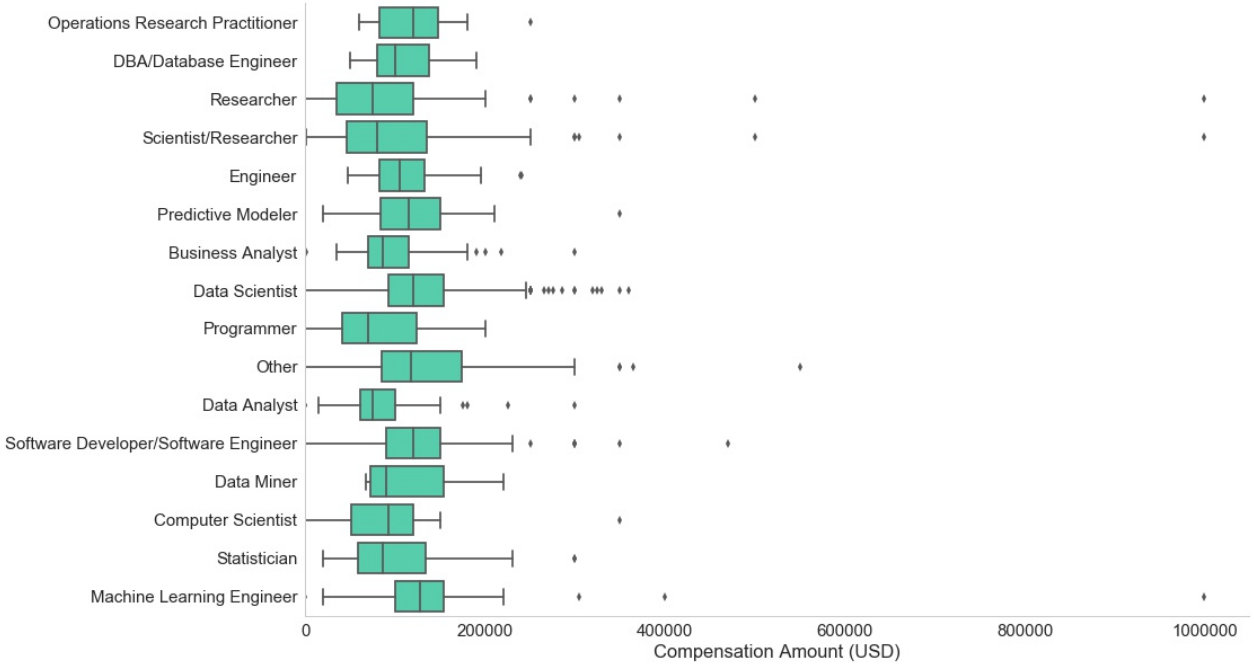
Appendix 5A: Data Scientist Satisfaction (Kaggle)

Appendix 6.0

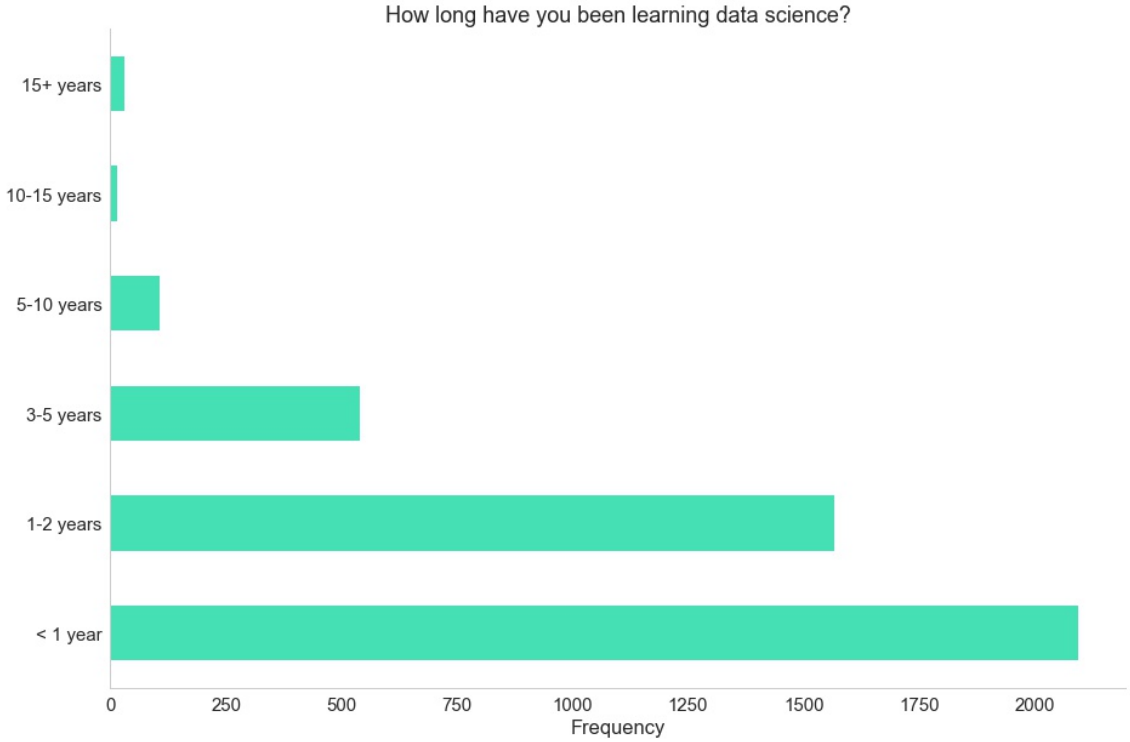


Appendix 6A: What is the most common challenges in Data Science Career? (Kaggle)

Appendix 7.0



Appendix 7B: Data Science Industry Salary Ranges (Kaggle)



Appendix 7C: Length of Time to Learn Data Science

Appendix 8.0 – Salary Bracket Feature Analysis (Specific Task Times)

The following Decision Tree Classifier attempts to identify which tasks best translate into a higher salary bracket. Here three salary brackets were defined, a low, medium, and high. Based on the training time spent productionizing and time spent data visualizing were the highest significant values, with individuals spending more time on both generally staying in a higher bracket for salary.