Table of Contents

Introduction

Since Harvard Business Review referred to it as "the sexiest job in the 21st century" back in 2012, Data Science has evolved from a promiscuously undefined branch of analytics to an independent and well–established discipline with its methods, processes and systems inherited from mathematics, statistics and computer science among others. Inspired by the way tech giants in Silicon Valley harnessed Big Data to explore unprecedented opportunities, organizations in various sectors are well on their way to building up similar capabilities within their business operation in which Data Scientists play a key role. There is nevertheless a lack of consensus on the definition of Data Science, and so–called Data Scientists frequently find themselves at the intersection of multiple areas where the skill set required increases in variety and sophistication, especially in the age of Artificial Intelligence when the scope of work has extended beyond traditional analysis and reporting. This remains overwhelming to any entry–level practitioners and particularly daunting to those without computing or quantitative background that wish to break into the field. Career pathway is also quite ambiguous because companies have their own ways to define the role of Data Scientists, and the term "Data Scientist" is often used interchangeably with Data Analyst, Data Architect or Machine Learning Engineer.

Thus, how to launch a Data Science career is one of the top concerns among aspiring novices. Though one might say that career advice is only one Google search away, it is most likely to be company–specific and subject to personal experience. A more effective job search strategy and comprehensive career plan can be made if one has a real insight into the job prospects and current demand for Data Scientist in the labour market. Intuitively, where would be the better place to find it than in data itself and by utilizing what Data Science has to offer? Analyzing data on Data Scientist job postings on LinkedIn – the most popular site for job postings and professional networking, this project aims to shed light on the following questions

1. Where is the demand for Data Scientist the highest?
2. What is a typical Data Scientist expected to do?
3. What skills and qualifications are required for Data Scientist?
4. How is job description for this role different among experience levels?
5. How is job description for Data Scientists different from Data Analyst, Data Engineer and Machine Learning Engineer?

Data Wrangling

By using Selenium webdriver and Python library Beautiful Soup, I was able to collect 11,697 job postings for Data Scientist available (until March 26th, 2020) on LinkedIn across 16 countries: United States, United Kingdom, Canada, Spain, Australia, Singapore, India, China, Japan, France, Netherlands, Germany, Sweden, Switzerland, Italy and South Africa. Data

scraping was based on the search results of query "Data Scientist", which means that the dataset also includes postings for other data–related positions displayed in the same LinkedIn result page. Each job posting is characterized by 11 attributes among which 1 has URL format, 1 is spatial, 1 is temporal, 1 has HTML format, 5 are single–valued text and the other 2 are multi–valued. Though the values for *description* attribute could have been extracted as text, HTML tags are stored instead in order to retain the structure of the content for further pre–processing. The dataset can be found in tabular format at https://bit.ly/2xH8arO. The dataset contains no numerical attributes, so the bulk of the project involves processing textual data. Python is used for data scraping, wrangling and cleansing while Tableau and R are mainly used for visualization.

Language Translating

Not all job postings were in English, and some companies in non–native English speaking countries preferred to use their own languages. To assure consistency, it was crucial to have the contents translated to English across all textual attributes. Two currently popular Python translation APIs are Google's and Microsoft's. Google algorithm is more accurate, but there is limitation on API calls per day. As speed and accuracy are equally important in this project, I decided to make use of Google Translate website instead by first writing non–English data into multiple Excel files categorized by languages and manually uploading each document on the website for translation. The translated results were gathered and read back into Python, with which the original contents are replaced entry by entry.

Categorizing and Reducing Cardinality

The categorical data were nevertheless messier after translated as synonyms were now introduced, which affected 4 attributes *title*, *exp level*, *job type* and *industry*. For example, translation produced various versions of the category *Associate* in *exp level* attribute such as *medium level*, *collaborator* or *assistant manager*. As for *title* attribute, there were no fixed categories and companies could be very creative in writing their job titles. Applying regular expression matching, I managed to group the values into these categories *Data Scientist / Data Analyst / Data Engineer / Machine Learning Engineer / Consultant / Big Data Developer & Administrators* / *Software Engineer / Researcher.* On the other hand, the categories of the other attributes are pre–determined by LinkedIn, so I only needed to eyeball the unique values of each attribute and manually map them to the corresponding categories.

Deriving additional attributes

The last step in the wrangling process was to derive from current attributes key information needed for analysis and store in new ones. Two attributes *country* and *area* (states or cities) were generated from *location* as the analysis is mainly conducted on the country level.

Regarding *description* attribute, extra work needed to be done. Apart from job responsibilities and requirements, a job posting content may contain other kinds of information such as company introduction, business description or job benefits. Analyzing the entire bodies of text would have prevented us from answering question 2 and 3 distinctly and thoroughly. It would also have introduced more noises to the visualizations and caused misleading intepretation. Thus, I found a strong need to be able to extract information on job responsibilites and job requirement seperately. It was observed in most job postings that these pieces of contents are presented in bullet points and each section is assigned a bold sub–heading. These sub–headings contain specific keywords that indicate whether the block of text is about job responsibilities or job requirement. For instance, that of a paragraph describing job responsbilities is likely to include such keywords as *responsibilities*, *duties* or *roles* while those as *requirements*, *qualifications* or *must have* are more frequently found in contents on job requirements. Based on this structure and corresponding HTML tags, for each job description, I was capable of extracting and categorizing two blocks of text and storing them in two attributes – one as *responsibilities* and the other as *requirements*. Next, the texts were cleaned by removing HTML tags, stop words (e.g., a, an, the) and special characters, including punctuations. They were then "lemmatized" – which refers to removing inflectional endings and returning the dictionary form of a word (e.g., responsibilities → responsibilities).

I also derived 2 additional textual attributes containing 2 indispensable pieces of information in job requirement – *minimum years of experience* and *minimum degree level* . This was performed by searching and matching regular expressions. While *minimum years of experience* is numerical attribute, *minimum degree level* is categorcal which has one of the 4 values *Bachelor / Master / PhD / Not Mention*.

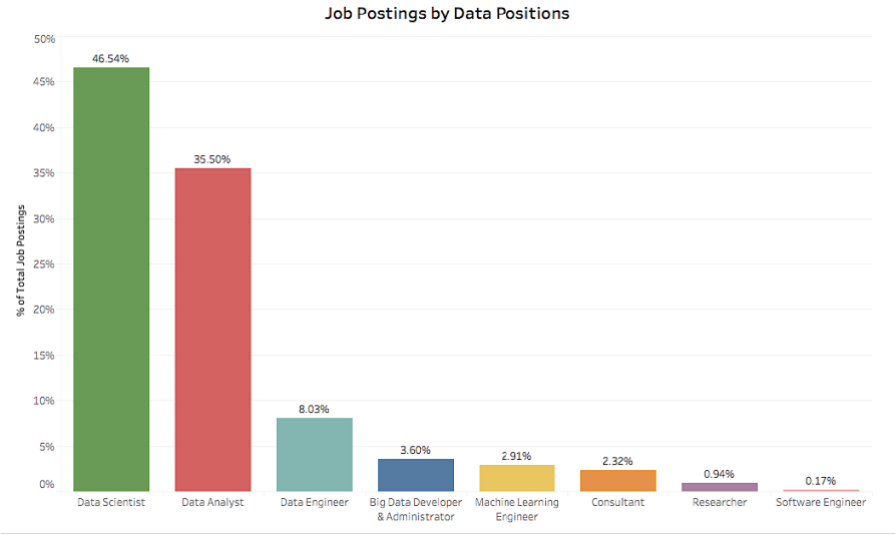
Data Checking

Two errors found in the wrangled data were duplicated entries and null values. Duplication may have occurred during the scraping process or because a company posting the same job with the same description mutlitple times to increase visibility. All duplications were deleted regardless. Null values came from two sources – one from the original data and the other from the derived attributes (except *country* and *area*). Null values in the original data existed in *exp level, job type, job func* and *industry* attributes, which was because the company was unable to classify their job posting under any specific category. In the second case, null values were generated due to the absence of the key information I wished to extract. Not all job descriptions were written in bullet point format. Some only allowed for extracting either job responsibilities or requirements whereas in others it was impossible to break text into chunks. This means that entries with such unclassifed text would have null values for one or both of *responsibilities* and *requirements* attributes. Similar situations applied for *minimum years of experience* and *minimum degree level*

though the number of missing cases were only a few. It is important to assume that the missing of these values is random and independent of the values of complete–case attributes. For example, it is nearly impossible that companies in Asian countries are more likely to structure their job descriptions in bullet format than those in Europe, or job postings for Data Scientist are less likely to contain information on years of experience than for other positions! Thus, entries with missing values are retained in the dataset, but those cases will be excluded when the respective attribute is being analyzed. However, *responsibilities* and *requirements* will be treated slightly differently, which will be discussed in Data Exploration section.

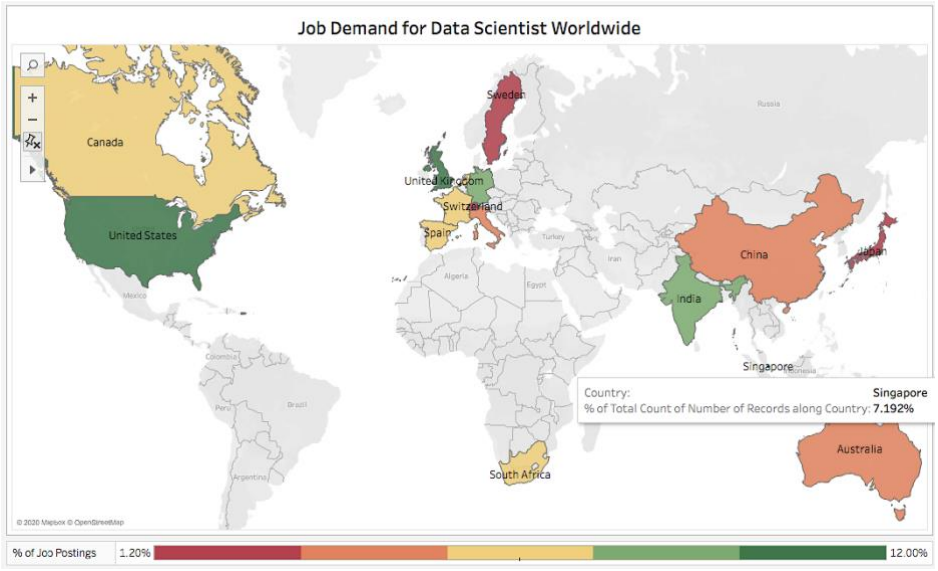
Data Exploration

In search for Data Scientist, Data Analyst deemed to be the most related position, and together they made up of 82.4% of jobs postings. To answer questions 1 to 4, the analysis will include data on Data Scientist and Data Analyst only, and I will use "Data Scientist" / "Data Science" as a general term. For the final question, I will examine these positions separately to compare the differences.



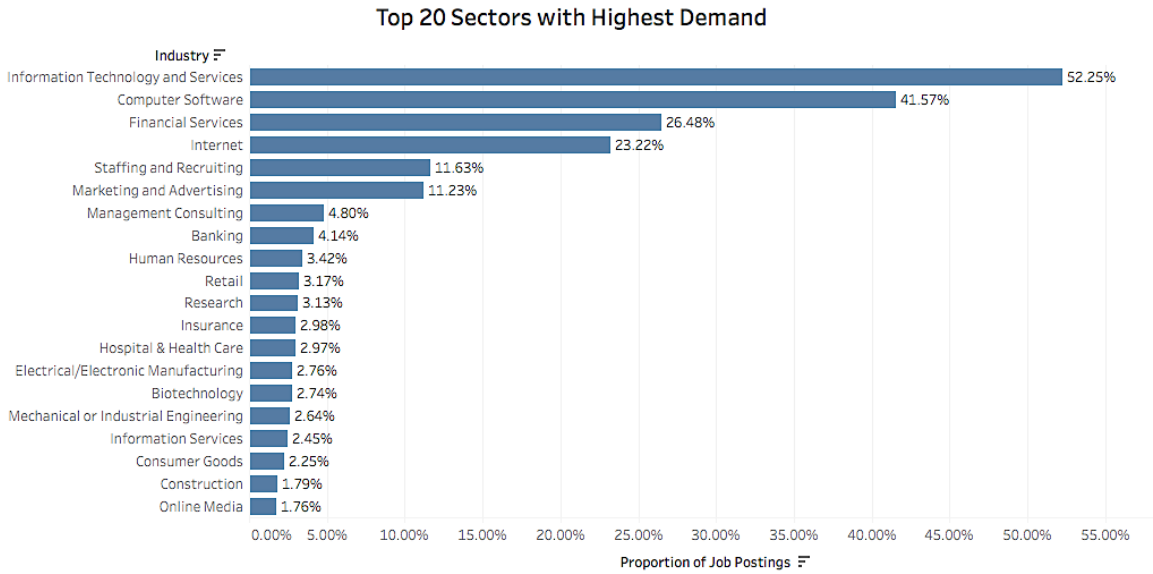
Job Demand

As Data Science is an emerging field in IT, opportunities for Data Scientist are more available in developed tech job markets. It is no surprising that United States top the list again with 11.2% of total job postings. Data Scientist was the best job in the U.S in 2019, according to both rankings of LinkedIn and Glassdoor. The outlook for Data Scientist is also extremely promising in the U.K (10.4%), India (9.6%) and Germany (8.6%), India, though a developing country, has become one of the fastest growing tech hubs in the world with an increasing pool of young IT professionals. It is also noticeable that despite being of smaller size than China and Australia, Singapore (7.2%) has cemented its position as APAC's central tech hub with a thriving innovation culture and strong demand for Data Science.



It is important to notice that the number of job postings highly depend the popularity of LinkedIn in each country. The proportion of job postings may not reflect the true demand in such countries as China, in which LinkedIn is not as widely used as local job sites like Zhaopin.com or 51job.com. The analysis can also be subject to selection bias in the sense that the majority of LinkedIn users are from the U.S, which explains why the country dominates in terms of job postings.

Technology and internet services companies are in greatest need of Data Scientist. However, as technological advances have transformed all kinds of businesses, opportunities are prevalent in non-tech sectors as well.



Big Data benefits Marketing and Advertising companies with respect to online tracking and understanding of customer behavior, which allows for more effective targeting and advertising strategies. Healthcare industry takes advantage of Big Data to improve the quality of treatment through faster access to patients' medical history as well as early detection and more accurate diagnosis of serious illness. Data Science also fits itself well into human resources practices. Job search sites like LinkedIn need skilled data professionals to curate employee profiles and produce

better matching with potential employers and vice versa. Implementing analytics within internal HR department helps improve quality of new hires as well provides insights into current employee engagement and productivity. It is undeniable that Data Science is a transferable skill nowadays.

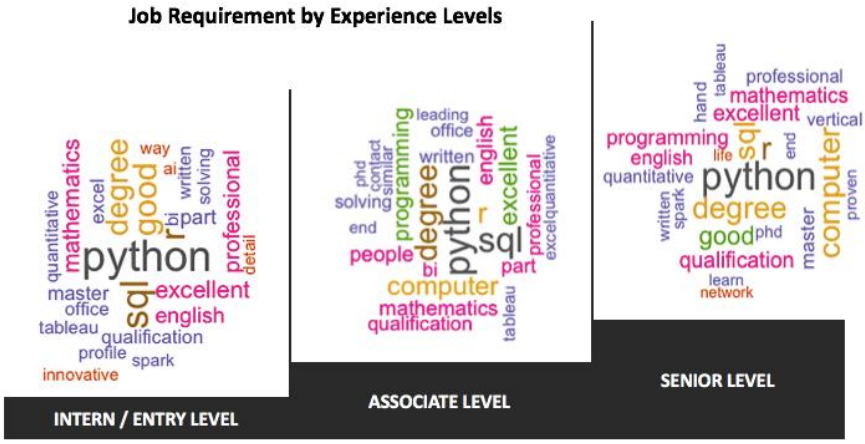
Job Responsibilities and Requirements

I used word cloud packages in R to visualize the data of these heavily textual attributes. As mentioned above, null values or unclassified descriptions were not omitted. For each attribute, I selected top 300 most frequently mentioned from the set of classified text, and count the frequencies of these words in the unclassified set. The combined values of frequencies from both sets were used for word cloud visualizations. Besides, I also removed generic or non–self–explanatory words with high frequencies, leaving only meaningful keywords.

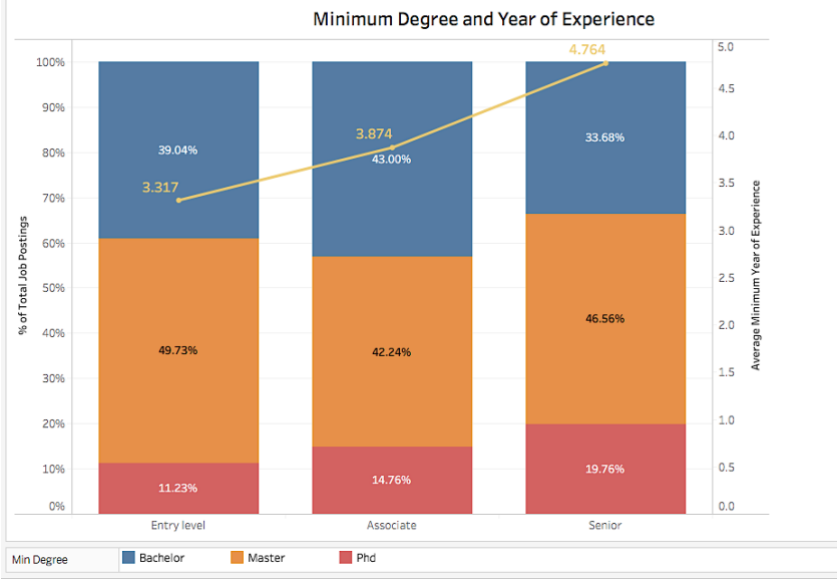
Data Scientist is a teamwork position, which requires effective collaboration with such roles as Software Developer, Product Manager, Data Engineers as well as other Data Scientists to deliver technological solutions. Analytics remains at the core of a Data Scientist's job. A Data Scientist is first and foremost expected to be able to convert massive volumes of data into actionable insights using statistical techniques and a wide range of technical tools. A large number of Data Scientist roles are closely related to product development in which experiments and strategies are executed based on the findings generated by Data Scientists which should show an in–depth understanding of the business and customers / end users. Building machine learning models is another key responsibility of a Data Scientist, which puts the role midway between Data Analyst and Machine Learning Engineer.

Python is the number 1 programming language for Data Scientist, followed by R, SQL and Java. Most commonly used data analysis libraries and machine learning packages are built on top of these languages, so a Data Scientist is required to be fluent in at least one of them. Though itself not a programming language, SQL is a must–have in a Data Scientist's toolkit due to the popularity of relational database management systems, and a large part of a Data Scientist's job involves retrieving and manipulating data from data warehouse. In terms of qualifications, degrees in Computer Science or Mathematics are preferred though other quantitative fields are sometimes acceptable. This guarantees that the candidate has solid background to comprehend and perform machine learning or advanced data modelling tasks. Data Scientists are also expected to be familiar with big data analytics and processing tools such as Excel, Tableau or recently Apache Spark, Hadoop and AWS. Data Scientist is not a fully technical role, so an ideal candidate should also demonstrate desirable soft skills include effective spoken and written communication, problem solving along with ability to work both in teams and independently.

52.6% of job postings are for Internship and Entry level positions, 21.5% for Associate and 15.7% for senior levels or higher. Although job responsibilities across experience levels are closely similar, the requirements are slightly different for each level. Python is a must–have for all levels, however it is significantly required for Senior levels over SQL and R. On the other hand, SQL is most important for Associate levels, equally as important as Python. Excel and Power BI are more commonly used by employees at Entry or Associate levels while Tableau is popular for all different job levels. Even a Data Scientist who has just started at the Entry level is expected to have good Math background whereas Degree in Computer Science is highly preferred for higher levels. Although university degree is an essential part of qualification, 58.43% employers never mentioned which levels are required. Among those concerned about education level, 46.37% expected at least a Master degree and 13.5% expected at least a PhD. Nearly half of the Entry

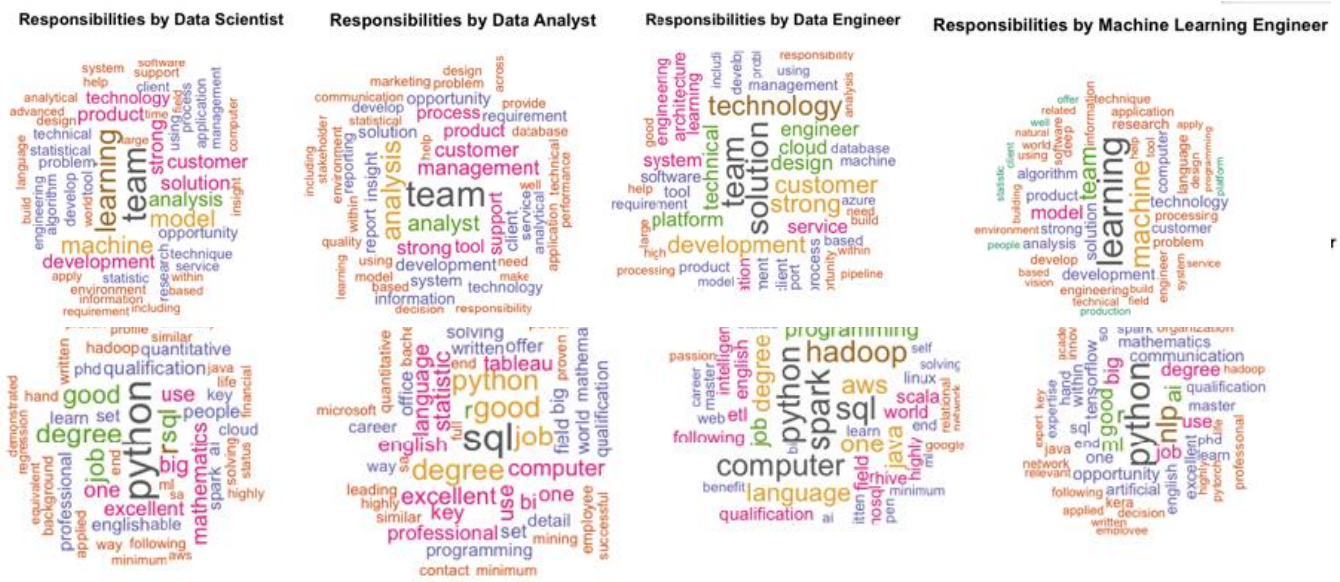


level and Associate level jobs required Master level at the minimum. PhD is more likely to be requested for Senior positions, but it can also be seen in job descriptions of some research–focused positions at lower levels. Scanning those job descriptions for Data Scientist, I notice that almost all companies expect candidates to have prior experience, including Entry level ones. Despite the fact that only 48% specified years of experience required, the bar is very high. On average, a qualified applicant for an Entry level position should have a minimum of around 3 years of experience. The number increases to 4 and 5 respectively for the other two levels. When analyzing the data on job requirement, I did not distinguish between must–have's and nice–to–have's, but whether it is mandatory or optional, one must attempt to fit into these requirements to be at an advantage.



Data Scientist vs. Data Analyst vs. Data Engineer vs. Machine Learning Engineer

One of the beginners' most frequently asked questions is how Data Scientist is different from other Data positions. Answering this question will help us make a though–out plan for our career, explore the suitable areas to specialize in and the right set of skills and knowledge to develop.



- Data Analyst position is focused entirely on analysis which entails interacting with database systems, exploring insights and reporting results. Therefore, fluency in SQL or BigQuery is a must along with strong knowledge of statistics. A crucial component of reporting is visualization, so it is clear that the toolkit also includes Tableau and Power BI. Python is preferable but very little programming or mathematics is required.
- The role of Data Engineers is to design, develop and maintain data architectures. They work closely with cloud platforms, data warehouse software and large–scale processing systems such as AWS, Hive, Hadoop or Spark especially, which is more often seen in the skillset of a Data Engineer than other positions. In terms of programming languages, SQL and Python are equally important and heavily required. Java and Scala are sometimes mentioned in the job description, which is rare for Data Scientist and Data Analyst. This is a highly technical position, so an ideal candidate is expected to have a computer–related degree.
- Machine Learning Engineer, as in the title, is in charge of building and implementing machine learning systems. It is strongly associated with modern–day buzzwords including Artificial Intelligence and Deep Learning. Machine Learning Engineer is a technical position, but the scope of work may also involve research aspects. They can be found reading research papers to comprehend algorithms and to be updated with new AI trends or cutting–edge technologies. In the data appear some new keywords that refer to the tools or frameworks frequently used by Machine Learning Engineer – TensorFlow, Pytorch and Keras. Python is the must–known programming language along with excellent knowledge of popular AI fields like Computer Vision and Natural Language Processing (NLP). R or SQL is not very much concentrated, and PhD is more likely to be a part of the requirements for this position.

In comparison with other data positions, Data Scientist can be considered to be "a little bit of everything". The responsibilities of Data Scientist tend to overlap with those of Data Analyst and Machine Learning Engineer, which involve both analysing data and building machine learning models. The modelling aspect seems to be of higher importance as Python remains a must–have along with proficiency at mathematics. However, the analytics part is what makes an aspiring Data Scientist cannot ignore specialized languages like R and SQL. Machine Learning Engineer is somewhat an independent technical role while Data Scientist, Data Analyst and Data Engineer are more business–oriented and teamwork–based. In fact, in a "full–stack" IT company in which roles are specialized, these 3 positions normally collaborate with each other in order to deliver product solutions.

Conclusion

Data Scientist is in high demand in most developed economics and more often found in big cities where tech ecosystem is thriving the most. Every continent has a dominant hunting spot for Data Scientist jobs. The U.S dominates North America and the world not only with respect to Data

Scientist but tech roles in general. India and Singapore stand out across Asia whereas the U.K, specifically England is the most desirable location to start a Data Science career in Europe. There are tremendous job opportunities for Data Scientist in non–IT sectors, notably Financial Services, Human Resources and Marketing / Advertising. Data Scientist and Data Analyst are more often called interchangeably as both of them have analytics at the core of their jobs. However, the key difference is that Data Scientist is also expected to work with machine learning models, have experience with programming languages, mainly Python as well as be familiar with a wide range of Big Data tools. There are no clear boundaries among data positions, and they are all sometimes referred to as a Data Science job. I would prefer to consider Data Scientist as a generalist in the sense that the role is of more advanced analytics than Data Analyst, but less technical and AI–focused than Machine Learning Engineer while expected to be familiar with data processing tools in a toolkit of Data Engineer, with which Data Scientist shares the least in common. Across all experience levels, no remarkable differences are found in the job responsibilities, and surprisingly not much as well in requirements. Getting an Entry level job is not easy. It not uncommon to come across a job posting that requires a PhD or a long list of technical skills for an Entry level position. This means aspiring novices must step up their game and build up a variety of skills and experience so that they can adjust to different requirements from employers. Another useful advice is to choose an area which you can specialize in, preferably towards analytics or towards machine learning. The field is high in demand but short in supply. Getting into it is challenging but once you do and find it suitable, I believe the journey ahead will be rewarding.

Reflection

Throughout this project, I got to improve the programming skills of processing and visualizing textual data using Python and R. It was highly useful to start an analysis with questions and hypotheses. This approach guided me on which attributes to look at, what forms of data wrangling to perform and which visualizations to choose that could best present the information I wanted. Nevertheless, the analysis has some limitations. LinkedIn is not representative of the entire job demand and may be subject to selection bias due to imbalanced user base across countries. Categories of other attributes are not comparable in size as well. For example, the data was collected based on query for Data Scientist, so there was insufficient amount of data for other positions. Thus, conclusions on roles and requirements for these positions may not reflect the reality. There are 2 solutions to this: one is to collect more data from other job sites – both global such as Glassdoor or Seek etc. and popular local ones; the other is to conduct statistical tests that could validate and quantify reliability of my results. Furthermore, there may be unobserved errors from the process of deriving *responsibilities* and *requirement* attributes, which may have affected the quality of word clouds. A better text classification text that can identify whether the content of a block of text is about job responsibility or requirement would help improve data quality.

Bibliography