IPTEC – Revista Inovação, Projetos e Tecnologias
e-ISSN: 2318-9851
Doi: https://doi.org/10.5585/iptec.v8i1.17263
Recebido: 21/5/2020 – Aprovado: 3/6/2020

What is a Data Scientist? Analysis of core soft and technical competencies in job postings

Carolina Coelho da Silveira[1]

Carla Bonato Marcolin[2]

Matheus da Silva[3]

Jean Carlos Domingos[4]

Abstract

The advancement of technologies has enabled companies to transform a large amount of data generated into important information for making strategic decisions. With this, the Data Scientist has been demanded as a piece of fundamental value for the organization. However, the skills necessary for this professional to work in the market are not yet consolidated in the literature. This research aims to map and analyze the soft skills and technical competencies of Data Scientists through a descriptive approach, using both a qualitative and quantitative typology. By collecting job postings, it was possible to verify that most companies are not concerned with the candidate's degree and educational level, but with the necessary soft skills and technical competencies. In this sense, the trend is to value a multidisciplinary profile. Among the most important skills for these professionals are Good Communications, Team Player, Problem Solver, Python, English, and SQL. We compilated the main skills aiming to contribute to the profile of the Data Scientist, that is still something new to be understood both by companies and by the literature.

Keywords: Data Scientist; Professional Competences; Data Analysis; Big Data.

[1] Graduanda em Administração, Faculdade de Gestão e Negócios, Universidade Federal de Uberlândia – FAGEN/UFU. Uberlândia, MG – Brasil. ORCID: https://orcid.org/0000-0001-6365-9871 | carolinasilveira004@gmail.com

[2] Doutora em Administração, Faculdade de Gestão e Negócios, Universidade Federal de Uberlândia – FAGEN/UFU. Uberlândia, MG – Brasil. ORCID: https://orcid.org/0000-0003-0260-5073 | carla@ufu.br

[3] Especialização em Marketing, Universidade de São Paulo – USP. Barueri, SP – Brasil. ORCID: https://orcid.org/0000-0001-6036-2706 | matheusbianchirs@gmail.com

[4] Doutor em Engenharia de Produção, Faculdade de Gestão e Negócios, Universidade Federal de Uberlândia – FAGEN/UFU. Uberlândia, MG – Brasil. ORCID: https://orcid.org/0000-0001-5013-7329 | jdomingos@ufu.br

Como citar:
Silveira, C. C. da, Marcolin, C. B., Silva, M. da, & Jean Carlos Domingos, J. C. (2020). What is a Data Scientist? Analysis of core soft and technical competencies in job postings. *Revista Inovação, Projetos e Tecnologias*, *8*(1), 25-39. https://doi.org/10.5585/iptec.v8i1.17263

1 Introduction

The rise of Big Data has pervasively impacted a myriad of aspects of human life, ranging across science, economics, culture and society, in both positive and negative ways (Mauro, Greco, Grimaldi, & Ritala, 2018). Bernard Marr, from Marr (2016, p. 01), says that "as the value of Data Analytics becomes apparent in all fields of activity, a growing number of people will want to be able to extract insights from their data".

First, the goal of Big Data Analytics is to enhance organizational decision-making and decision-execution processes: informed decision-making is one of the building blocks of organizational success and the importance of comprehensive analysis of information before making operational and strategic decisions has been highlighted in the works of many organizational researchers and practitioners (Tabesh, Mousavidin, & Hasani, 2019). In its turn, Data Science concerns with the collection, preparation, analysis, visualization, management and preservation of large collections of information, enabling the creation of data products; moreover, Data Science incorporates from multi-disciplines aspects such as statistics, computer science, applied mathematics and visualization, and these disciplinary areas are utilized by Data Scientists, who, gathering data, are able to transform it into a tractable form, making it tell its story and presenting that story to others (Ecleo & Galido, 2017).

Business ecosystems can be complex and, faced with this data torrent revolution, organizations must quickly adapt to the new system dynamics and environment to survive (Vidgen, Shaw, & Grant, 2017). The advent of new sources of data coupled with the renewal of methods and technologies used for business-impacting analytics require the development of new interdisciplinary competencies spanning from IT skills to business domain knowledge and communication skills (Chen *et al*., 2012 as cited in Mauro et al., 2018), posing as a talent challenge for companies, seeking to upgrade their human capital.

Consequently, these companies need to quickly secure the appropriate competencies in the area of Big Data: such a race for acquiring the right talent does not seem to slow down while the labor market is unable to cope with an exponentially increasing demand (Mauro et al., 2018). On the other hand, individuals who seek to fill in into this gap might have a hard time figuring out what skills to acquire, or, at least, which of them most matter to companies. However, with no common definition to "Data Scientist" due to the newness of the field, organizations and individuals have been forced to come up with their own definition and list of skills leading to ambiguity and incorrect resource fit (Ho, Nguyen, Paffod, & Slater, 2019).

Given this context, the purpose of this work is to map and organize the core soft and technical competences required for Data Scientists in order to create value from data inside and outside the organizations. The transformations brought by new technologies also affect business structures, creating opportunities and challenges for organizations. Increased data and the digitalization of different processes help organizations to improve results analysis and to process efficiency actions, but at the same time, organizations are under pressure to show a more transparent approach to internal controls and quality management of information provided to different stakeholders; within that, organizations must acquire the right people, with the right skills to support their analytics transformation, so there will be intense competition for Data Scientists that are technically competent and able to create innovative and practical solutions to business problems through data analytics (Vidgen et al., 2017). Helping organizations and individuals given this imminent competition and the lack of a solid Data Scientist's profiling are issues that this article aims to contribute.

2 Literature review

2.1 Big Data and Business Analytics as trends

As stated by Andrew McAfee and Erik Brynjolfsson (2012), you can't manage what you don't measure. The ability to not only collect, but most importantly, process massive amounts of data have rapidly created a new gold rush, and many organizations have devoted considerable amount of time, money and effort to obtain profitable insights into the behavioral and structural characteristics of their environments (Tabesh et al., 2019). This is what we call Big Data: the massive, complex and real-time streaming data that requires sophisticated management, analytical and processing techniques to extract insights (Gupta & George, 2016).

In this context, the concept and practice of Business Analytics had a significant growth in the last decade, attracting the attention of researchers and managers from different areas (Mortenson, Doherty, & Robinson, 2015). Business Analytics can be seen as an emerging phenomenon reflecting the exponential growth of data in terms of volume, variety and velocity and, inside this concept, is the extensive use of data driven quantitative methods, specially statistics and mathematics (Marcolin, Becker, Wild, Schiavi, & Behr, 2019). Hereupon, Business Analytics have the power to allow data analysis from different data sources with different data formats, making it possible to improve companies' performance and identify business opportunities (Bayrak, 2015).

The Big Data revolution is far more powerful than the analytics that were used in the past. As pointed by McAfee and Brynjolfsson (2012), we can measure and therefore manage more precisely than ever before, we can make better predictions and smarter decisions, we can target more effective interventions and can do so in areas that have been dominated so far by gut and intuition rather than by data and rigor. As the tools and philosophies of Big Data spread, they will change long-standing ideas about the value of experience, the nature of expertise and the practice of management. Smart leaders across industries will see using Big Data for what it is: a management revolution. As data becomes cheaper, the complements to data becomes more valuable. Some of the most crucial of these are Data Scientists and other professionals skilled at working with large quantities of information; although, as important as statistics are, many of the key techniques for using Big Data are rarely taught in traditional statistics courses.

In line with these trends, organizations are racing to keep up with the changes and leverage the benefits of useful information embedded within large volumes of data, as the confluence of data proliferation, algorithmic advancement and more powerful computing and storage facilities have opened new possibilities for the transformation of data into business insights, decisions and actions (Tabesh et al., 2019; Chui, Kamalnath, & Mccarthy, 2018). Advanced analytics is likely to become a decisive competitive asset in many industries and a core element in companies' efforts to improve performance and it's a mistake to assume that acquiring the right kind of Big Data is all that matters (Barton & Court, 2012). Data-driven strategies can already be seen as an increasingly important point of competitive differentiation, thus, in order to create value from data and maintain competitive advantage, the focus in analytics skills is an important capability to obtain (Barton & Court, 2012; Provost & Fawcett, 2016), both for organizations and individuals.

McAfee and Brynjolfsson (2012) stated that companies succeed in the Big Data era not simply because they have more or better data, but because they have leadership teams that set clear goals, define what success looks like and ask the right questions. Big Data's power does not erase the need for vision or human insight. On the contrary, as the authors continue, we still must have business leaders who can spot a great opportunity, understand how a market is developing, think creatively and propose truly novel offerings, articulate a compelling vision, persuade people to embrace it and work hard to realize it, and deal effectively with customers, employees, stockholders and other

stakeholders. The successful companies of the next decade will be the ones whose leaders can do all that while changing the way their organizations make many decisions.

As for Barton and Court (2012), the era of Big Data is evolving rapidly and they suggest that most companies should act now. But rather than undertaking massive overhauls of their companies, executives should concentrate on targeted efforts to source data, build models and transform the organizational culture. Such efforts will play a part in maintaining flexibility. That nimbleness is essential, given that the information itself - along with the technology for managing and analyzing it - will continue to grow and change, yielding a constant stream of opportunities. As more companies learn the core skills of using Big Data, building superior capabilities may soon become a decisive competitive asset.

2.2 The Data Scientist

Over the past few years, businesses have begun to realize that Data Science is the key to solving some of their most pressing problems. Data formats are multiplying and connectors are becoming crucial. Given the nature of the work, Data Science is an inherently collaborative and creative field and professionals have to work within interdisciplinary team environments to find innovative ways to complete projects. Having said that, Data Scientists can have a variety of roles and responsibilities in the business and for different roles, different skill sets often are needed (Chuprina, Postanogov, & Kostareva, 2017; Meyer, 2019).

We can define a Data Scientist as someone who is able to extract patterns and trends from data through certain data-related tasks, regardless of its characteristics, formats and consequently challenges. The Data Scientist communicates and disseminates the findings, creates data artifacts or optimizes existing ones, improving business management and performance through the enrichment of the decision-making process. Data Scientists must also be able to deal with Big Data in all the stages of data flow and topics like ethics, privacy and security should also be constantly on their minds. Aspects related to the computing field such as artificial intelligence, machine learning, programming, databases and other data driven aspects have also a strong presence in this profile. In order to communicate findings, Data Scientists must have strong social and personal capabilities, like communication, business acumen and curiosity (Costa & Santos, 2017).

A case study promoted by Vidgen et al. (2017) showed that organizations want Data Scientists who are curious, who should also have a problem-solving orientation, be capable of working independently and be able to work with the business to co-create plausible, convincing stories through data that lead ultimately to actionable insights.

The Data Science Association - DSA (2016 as cited in Costa & Santos, 2017) presents the Data Scientist as a person who knows about: machine learning; algorithms; modeling; statistics; analytics; visualization; math; business acumen; scientific method; large datasets. DSA acknowledges that a Data Scientist is a professional that is able to: create, validate and transform data to create meaning; liberate and create meaning from raw data; play with data, spot trends and learn truths few others know; explore, ask questions, do "what if" analysis, question existing assumptions and processes; communicate findings to both business and IT leaders in a way that can influence how an organization approaches a business challenge; extract actionable and valuable intelligence from large data sets.

It's quite clear that being a Data Scientist requires a multidisciplinary profile, it's not expected to be found concentrated in only one competence area (Costa & Santos, 2017). Chatfield, Shlemoon, Redublado and Rahman (2014) complement this idea by encapsulating the skills of a Data Scientist as a set of six common attributes: entrepreneurship and business domain knowledge; computer

scientist; effective communication skills; create valuable and actionable insights; inquisitive and curious; and statistics and modeling.

Mauro et al. (2018) suggest that Data Scientists and their deep expertise on analytical methods are far from being sufficient in granting companies a real competitive advantage. The authors complement that Data Scientist is not a homogeneous profession, but includes both hard and soft skills, as well as different connotations towards organizational processes, technologies and value creation. Highlighting the multidisciplinary profile, Mendelevitch, Stella and Eadline (2016) propose that, instead of hiring Data Scientists with the combined skill sets of Data Engineers and applied Scientists, build a team comprised of Data Engineers and applied Scientists, and focus on providing a working environment and process that will drive productivity for the overall team.

For Vidgen et al. (2017), while the Data Scientist undoubtedly needs strong statistical and mathematical skills, they also need IT skills, notably an ability to program and an ability to manipulate data. However, being a Data Scientist is not merely about being good with numbers, they also need to be a "*bricoleur"*, be curious, problem-focused, able to work independently and capable of co-creating and communicating stories to the business that form the basis for actionable insight into data. Future Data Scientists must have the ability to work cross-functionally across business silos and focus on the end goal, like creating solutions and delivering business value. Their role and field extend beyond the boundaries of the IT department. This has significant implications in the future of the Data Scientist role, such as recruitment, training and managing the analytics talent pipeline, and thus, the HR strategy.

In the words of Chuprina et al. (2017), even though organizations are trying to bring more and more Data Scientists into the workforce, the gap between Data Science's supply and demand remains quite substantial. From the authors' point of view, it's actually impossible to gather in only one person, during a four-year University program, skills related to programming, statistics and business communications. This is reinforced by Meyer (2019, p. 384) who says:

> Employers are struggling to meet demand for Data Scientists. This shortage is compounded by the hybrid nature of Data Scientist positions; that is, needing a mix of analytic skills and domain-specific expertise, which is difficult to develop in one individual. This difficulty in finding qualified Data Scientist candidates is leading organizations to seek creative ways to not only find, but also develop and grow workforce talent in-house.

Facing the intellectual trends needs many of the same skills as facing the commercial ones and seems just as likely to match future student training demand and future research funding trends. The would-be notion takes Data Science as the science of learning from data, with all that this entails. This larger vision, called by Donoho (2017) "Greater Data Science" cares about each and every step that a professional must take, from getting acquainted with the data all the way to delivering results based upon it, and extending even to that professional's continual review of the evidence about best practices of the whole field itself. There are six dimensions which encompass the activities of the Greater Data Science.

1. Data Exploration and Preparation: the greater effort devoted to Data Science is expended by diving into the messy data to learn the basic of what's in them, so that data can be made ready for further exploitation. This requires exploration, to sanity-check its most basic properties and to expose unexpected features, and preparation, as many datasets contain anomalies and artifacts so it's necessary to identify and address such issues.

2. Data Representation and Transformation: data sources can assume a very wide range of formats, so Data Scientists have to often implement an appropriate transformation restructuring the originally given data into new and more revealing forms. Doing so, the Data Scientists end up developing skills in areas such as modern databases (where they need to know the structures, transformations and algorithms involved in all kinds of data representations) and mathematical representations (there are some useful mathematical structures for representing data of special types).

3. Computing with Data: every Data Scientist should know and use several languages for data analysis and data processing. But beyond knowing the basics, they need to keep current on new idioms for efficiently using those languages and need to understand the deeper issues associated with computational efficiency. Also, cluster and cloud computing and the ability to run massive numbers of jobs on such clusters has become an overwhelming, powerful ingredient of the modern computational landscape.

4. Data Visualization and Presentation: modern practice has taken the techniques of data visualization and presentation to much more elaborate extremes. Data Scientists not only develop plots, but also create dashboards for monitoring data processing pipelines that access streaming or widely distributed data and develop visualizations to present conclusions from a modeling exercise.

5. Data Modeling: Data Scientists should work with both generative modeling, in which one proposes a stochastic model that could have generated the data and derives methods to infer properties of the underlying generative mechanism and predictive modeling, in which one constructs methods which predict well over some given data universe.

6. Science about Data Science: it's the study of what data analysis "in the wild" is actually doing, when it's possible to identify commonly occurring analysis/processing workflows.

Donoho (2017), to conclude, brings up that these categories of activity, when fully scoped, cover a field of endeavor much larger than what is currently done by the Data Scientists, since a single category, the Data Modeling, dominates the representation of today's Data Science teachings, researching and techniques. Combining those six dimensions with job postings analysis allows to understand the core competences to form data scientists considering both academical and commercial views, helping managers and individuals to fill the talent gap foreseeing for the next years (Mauro et al., 2018; Meyer, 2019; Vidgen et al., 2017).

3 Method

This research had a descriptive approach, attending both quantitative and qualitative analysis. As Nassaji (2015) states, the goal of descriptive research is to describe a phenomenon and its characteristics, concerning more with "what" rather than "how" or "why" something has happened, using tools such as survey. More frequent than not, in such research the data is often analyzed quantitatively, using frequencies, percentages, averages or other statistical analyses to determine relationships. Qualitative research, however, is more holistic often involves an inductive exploration of the data to identify recurring themes, patterns or concepts and then describing and interpreting

those categories.

The data was collected through LinkedIn, a social network that focuses in professional networking and career development. This data source has been increasingly used and Bradbury (2011) reinforces that LinkedIn is a valuable source of business network information, being a perfect example of Big Data: a very large dataset that cannot be mined using traditional relational database management tools.

Table 1: Contents extracted from the job opportunities.

Web page's link	Used a consulting company?	Job description
Job opportunity title	Company's line of business	Degree required
Company	Company's overview	Previous experience
Company's headquarters location	Hard skills	Desirables
Language	Soft skills	Criteria provided by job poster

Source: Research Data.

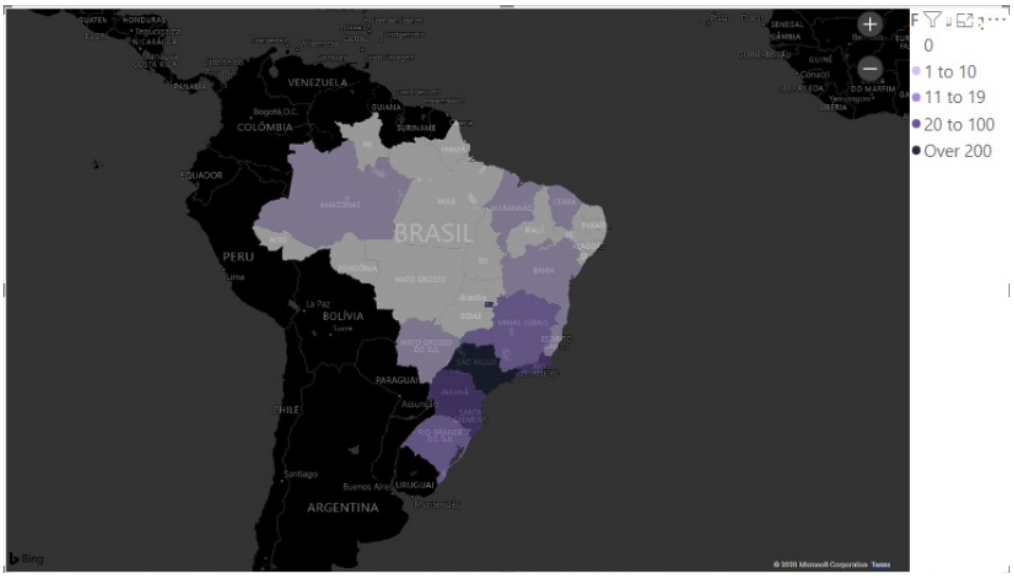
To conduct this study, it was necessary to survey Data Scientists' jobs from LinkedIn. To acquire these jobs, we limited the search location to "Brazil", filtered by "Jobs" and searched for the terms "Data Scientist" and "Cientista de Dados" (the Portuguese translation). The URL for each result was collected and saved in a spreadsheet, as the jobs offered in LinkedIn change daily and the content extraction would happen throughout the month. All the duplicated jobs were excluded, resulting in 248 results for the first term and 96 for the second. All through November of 2019, for each job, the contents were extracted, including the topics as presented in Table 1.

4 Results

Interestingly enough, we observed that not all job opportunities were named Data Scientist or its Portuguese translation, "Cientista de Dados". A few variations such as big data analyst, big data architect, big data consultant, big data engineer, big data specialist, computer scientist, data analyst, data engineer, data strategist, data science analyst, data science consultant, decision strategist, machine learning engineer and others showed up in almost 40% of the results. This actually reassures Ho et al. (2019) affirmative that the definition or list of skills for a Data Scientist is so sparse and inconsistent that it becomes evident when we do a simple job about it and it returns job title postings different from the search term.

About the jobs' locations, when informed, was the same as the company's headquarters. The Brazilian states with more job opportunities were, respectively: São Paulo, Rio de Janeiro, Federal District, Paraná and Santa Catarina. Graphic 1 shows the frequency for each state's appearance.

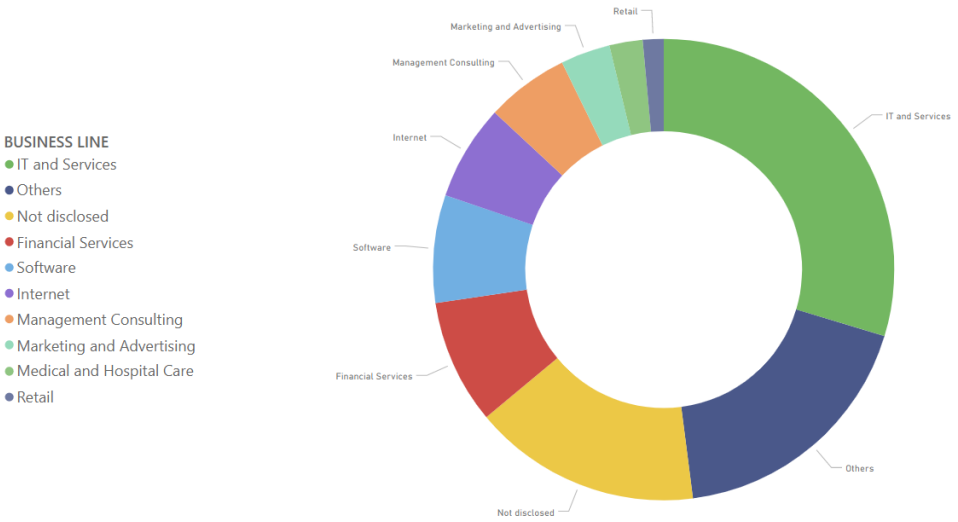
Of all the opportunities, 79% were in Portuguese and only 20% were being publicized by consulting companies. A few companies offered more jobs than others. More specific, 14 companies offered 84 jobs, that is, those companies were responsible for nearly 25% of the opportunities, demonstrating a relative concentration, and what could indicate that in Brazil there could have a future expansion for this professional, given its growth in other business areas.



Graphic 1 – Map of Brazil, showing the frequency of jobs offered in each state

Source: Research Data.

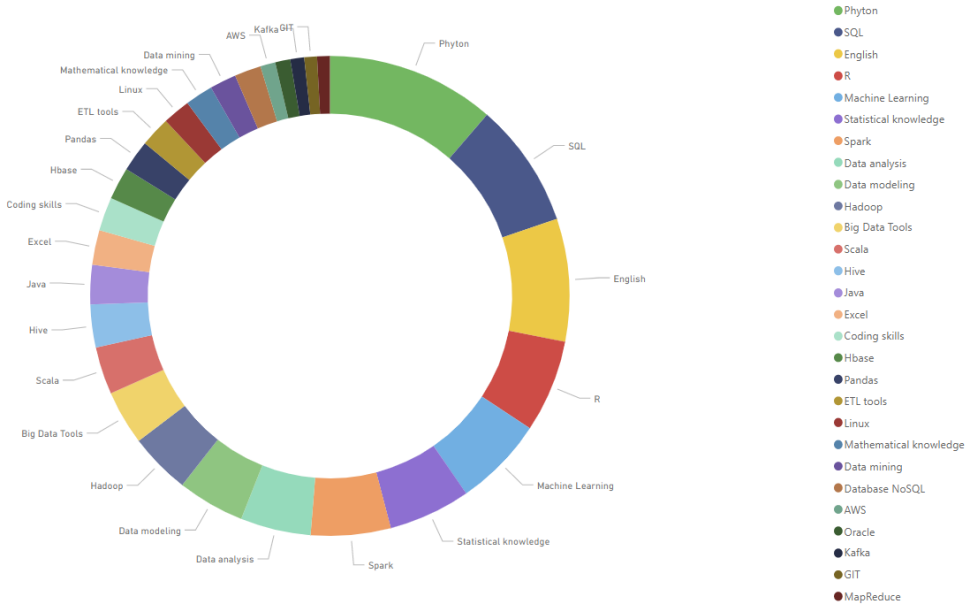
As for the companies' line of business, the results are detailed in the Graphic 2. Besides those, other industries appeared in the results, such as accounting, oil and energy, insurance, cosmetics, airlines, legal practice and others, but not in relevant percentages.



Graphic 2 – Companies' line of business

Source: Research Data.

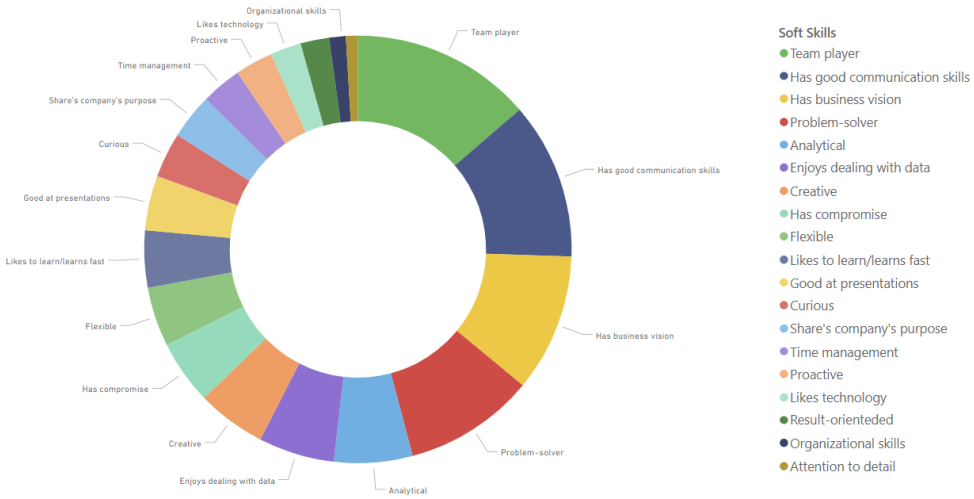
When analyzing the hard and soft skills asked by the job posters, we found a myriad of content and see that the jobs' requirements are vast and diverse. First, about hard skills: in general, knowledge in a few fields is required, like: database, Big Data tools, coding languages, programming, ETL tools, statistical and mathematical knowledge, machine learning, operational systems, business intelligence methodologies and others. A particular trait for the hard skills is that job posters are very flexible about them. In a few cases, they do not list precise tools, saying only that it is a requirement to know tools in a specific field or, if they do list a number of tools, they are only suggestions. The main tools and knowledges asked for hard skills are shown in Graphic 3.



Graphic 3 – Main requirements for hard skills

Source: Research Data.

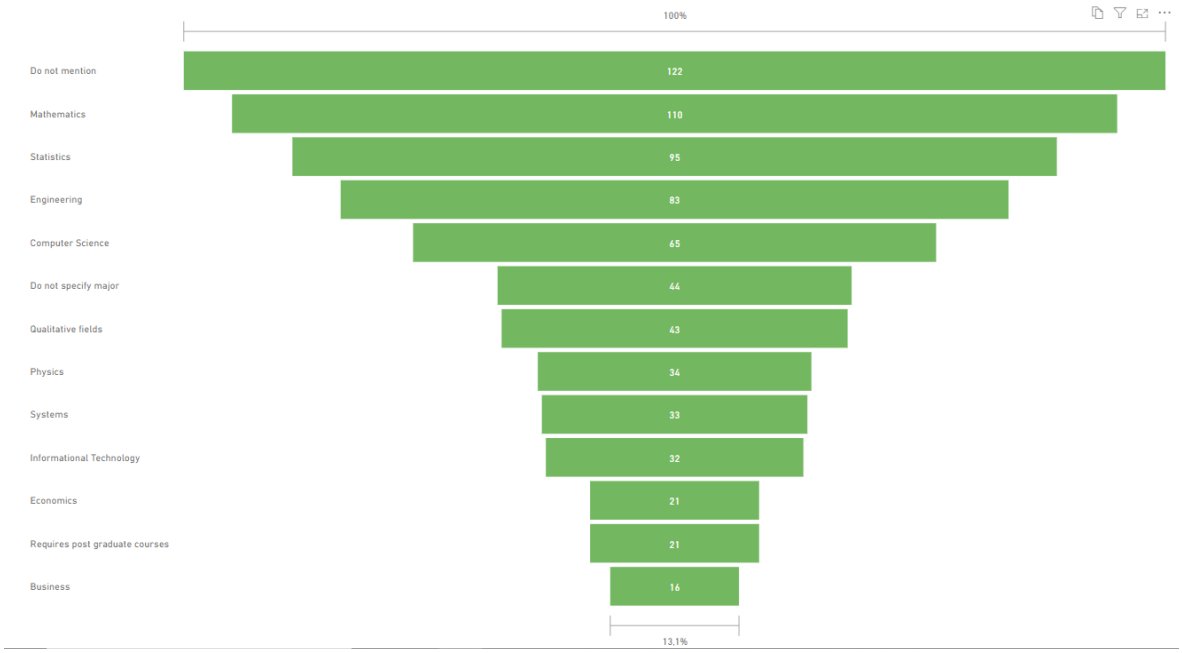
As for the soft skills, 53% of the job opportunities did not mention any at all. The ones that did, list characteristics such as the ones shown in the Graphic 4. Here we see that almost half of the companies understand and agree with Mauro et al. (2018) definition, that Data Scientist is not a homogeneous profession, including both hard and soft skills (see section 2.2).



Graphic 4 – Main soft skills seeked by the companies

Source: Research Data.

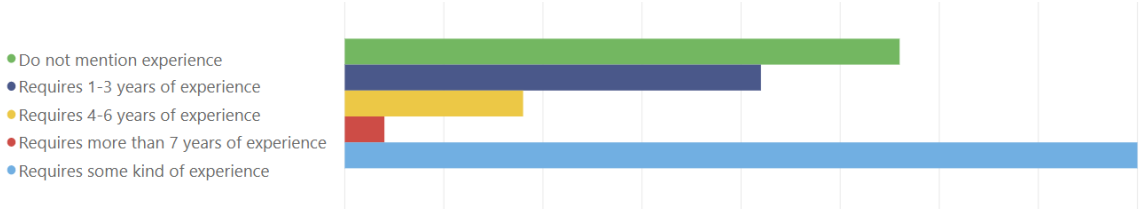
Despite all above findings, when analyzing the education level required that we found one interesting outcome of this study. Considering the mention about any specific specialization area, 35% of the jobs do not mention any. In addition, 13% asked for a University Degree but did not specified a major, and 12,5% asked only for a Degree in a quantitative field. For graduate courses, 6,1% requires some kind (specialization, masters, doctorate or PhD). Only the remaining jobs did mention some kind of major, and Graphic 5 shows the main results.



Graphic 5 – Educational levels required

Source: Research Data.

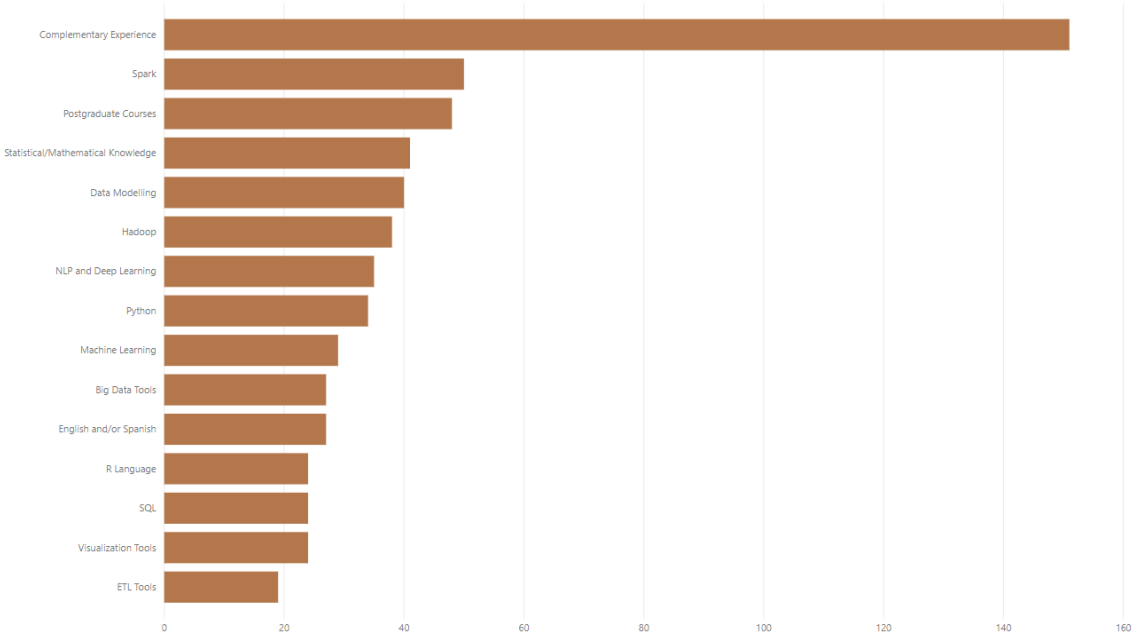
Complementary to the academic education, there's the level of previous experience required from the applicant. Again, we found a myriad of content and see that the jobs' previous experiences requirements are diverse, fairly related to the job's description. In a more quantitative way, we analyzed how much time of experience was required. Graphic 6 shows the results.



Graphic 6 – Levels of previous experience required

Source: Research Data.

Also, job posters often add some plus skills they would like applicants to have, although they are not mandatory: 55% of the jobs analyzed have chosen desirable characteristics and they vary accordingly, mostly, to the job description. It's relevant to highlight that even though 28% of the opportunities do not mention previous experience and 40% only mention it generally, almost half of them mention it as a differential they would like the applicant to have. This means that even if not mandatory, having some previous experience is important to the hirers. Graphic 7 shows the details.



Graphic 7 – Main desirable characteristics

Source: Research Data

With the analysis of job postings, it is possible to identify a set of skills required of Data Scientists to occupy a job opportunity in Brazil. Among the skills categorized as hard, it is identified that many discriminate specific techniques or tools in the computational area, and others point to a field of knowledge. These skills can be classified and grouped into classes when considering their purpose or main functionalities.

Just as Donoho (2017) identified when proposing the six dimensions for the activity of Data Scientist, the requirements identified in this research also direct a much greater field of effort for the competencies of a data scientist, which are beyond the area of modeling of data, an area that currently dominates the essential skills for a Data Science professional.

Seeking to elucidate the fundamental competencies to train this professional, when considering the six dimensions proposed by Donoho (2017) and the skills required in the business environment, we can identify in Table 2 a relationship between the six dimensions and the skill classes identified in this research, as well how to relate them to soft skills.

Table 2: Relationship between the six dimensions and the skill classes

Dimensions	Hard Skills		Soft Skills
	Groups	**Skills**	
Basic skills in all dimensions	Common skills	Coding skills Excel English	Flexible Time management Proactive Good at presentations Has compromise Team player
	Operational system	Linux	
	Distributed version control system	GIT	
1. Data Exploration and Preparation:	Data manipulation	Database NoSQL Big Data Tools ETL tools SQL Hive	Attention to detail Organizational skills Curious Share's company's purpose

	Data modeling and analysis	Data modeling Data analysis Mathematical knowledge Statisticalknowledge	Enjoy dealing with data Analytical Has business vision Has good communication skills
2. Data Representation and Transformation	Programming Language	Pandas (library) Java Scala R Python	Result-orientated Creative Enjoys dealing with data Has business vision
	Artificial Intelligence Techniques	Data mining Machine Learning	
	Distributed database systems	Oracle Hbase	
	Data manipulation	Database NoSQL Big Data Tools ETL tools SQL Hive	
3. Computing with Data	Frameworks for clustering and large-scale data analysis	MapReduce Spark	Likes technology Likes to learn/learns fast Enjoys dealing with data Problem-solver
	Platforms for distributed systems	AWS Kafka Hadoop	
	Distributed database systems	Oracle Hbase	
	Programming language	Pandas (library) Java Scala R Python	
4. Data Visualization and Presentation	Programming Language	Pandas (library) Java Scala R Python	Proactive Share's company's purpose Creative Analytical Has business vision Has good communication skills
	Data manipulation	Database NoSQL Big Data Tools ETL tools SQL Hive	
	Data modeling and analysis	Data modeling Data analysis Mathematical knowledge Statistical knowledge	
5. Data Modeling	Artificial Intelligence Techniques	Data mining Machine Learning	Share's company's purpose Likes to learn/learns fast Creative Analytical Problem-solver
	Data modeling and analysis	Data modeling Data analysis Mathematical knowledge Statistical knowledge	

	Frameworks for clustering and large-scale data analysis	MapReduce Spark	Likes technology Proactive
6. Science about Data Science	Artificial Intelligence Techniques	Data mining Machine Learning	Curious Share's company's purpose Likes to learn/learns fast Analytical Problem-solver Has business vision
	Data modeling and analysis	Data modeling Data analysis Mathematical knowledge Statistical knowledge	

Source: Research Data.

The Table 2 also highlights a set of skills considered as basic for a data scientist, that is, a set of skills that are necessary for all the dimensions of this professional's performance. The analysis of the framework can allow managers and individuals to understand how the skills of a data scientist have been delineating, in order to conduct actions for training and recruiting professionals who meet the demand for talent expected in the coming years as well as to select the main skills to independently develop in the individual's case.

5 Conclusion

Even with this myriad of results, it's already possible to draw a few conclusions and identify similarities between the results obtained and the literature gathered. First, we can see that almost all job opportunities are set in Brazil's economic centers (that are also states' capitals). This shows that Data Science is still something new, yet to be absorbed by medium and smaller cities and companies. Also, it's possible to see that, even though half of the opportunities is being offered by IT and services, financial services, software and internet companies, the other half is being offered by a great variety of companies, showing that they recognize the Data Scientist's role and that they want this professional in their team, what could indicate a second talent run still to come.

Showing resemblance to the literature presented, we have the hard skills. Costa and Santos (2017) have said that Data Scientists should have a profile tending to programming skills, data bases and machine learning knowledges. Google and Facebook (2016 as cited in Costa & Santos, 2017), have also highlighted the analytical, mathematical and statistical knowledge and the strong interest in data, using it to make decisions and solve problems. All of those characteristics (and others) are mentioned, recurrently, in the jobs analyzed. We can also see that the hard skills are what matters the most to the job posters, seeing that they are mentioned in all the opportunities, even if briefly.

The soft skills showed up in a relevant number of opportunities, evincing behavioral skills which make the Data Scientist a professional who is even more skilled and integrated to the team. These skills are what differentiate the Data Scientist from every other professional from a quantitative field, such as engineers, statistics or computer scientists. Costa and Santos (2017) also brought up how necessary the soft skills are for a Data Scientist, being a few of them strong social and personal capabilities, good communication and business acumen.

But the most interesting occurrence, until now, happens when we begin to analyze the educational level required: a third of the jobs do not bother to even mention the applicant's major. This could relate to what was said by Chuprina et al. (2017) (see section 2.2), that a simple University program could not condense all that is needed in a Data Scientist, or it could simply imply that the hirers care more about the applicant's skills and knowledge than about his or hers formal

academic education.

It's certain that we will continue this study with an in-depth analysis, as there is so much more to understand from the data collected. But because being a Data Scientist is still something new and to be formally understood, and as Big Data and Business Analytics become more of essential techniques and not only trends, it's imperative that new studies take place in this field so we can, finally, properly address it.

This study presents some limitations. The data itself has a very dynamic nature, since new jobs are being posted and being fulfilled, which causes them to be deleted from the platform. We tried to capture the data in the shortest data frame, in order to keep a genuine photography of the scenario. For future studies, we plan to compare this profile with other countries, with more or less mature markets, in order to understand similarities and differences among data scientists worldwide.

References

Barton, D., & Court, D. (2012). Making advanced analytics work for you. *Harvard Business Review*, 90(10), 78–83.

Bayrak, T. (2015). A review of business analytics: A business enabler or another passing fad. *Procedia-Social and Behavioral Sciences*, 195, 230-239.

Bradbury, D. (2011). Data mining with LinkedIn. *Computer Fraud & Security*, 2011(10), 5-8.

Chatfield, A. T., Shlemoon, V. N., Redublado, W., & Rahman, F. (2014). Data scientists as game changers in big data environments. *Proceedings of the Australasian Conference on Information Systems*, Auckland, New Zeland, 25.

Chui, M., Kamalnath, V., & Mccarthy, B. (2018). *An executive's guide to AI*. McKinsey & Company.

Chuprina, S., Postanogov, I., & Kostareva, T. (2017). A way how to impart data science skills to computer science students exemplified by obda-systems development. *Procedia Computer Science*, 108, 2161-2170.

Costa, C., & Santos, M. Y. (2017). The data scientist profile and its representativeness in the European e-Competence framework and the skills framework for the information age. *International Journal of Information Management*, 37(6), 726-734.

Donoho, D. (2017). 50 years of Data Science. *Journal of Computational and Graphical Statistics*, 26(4), 745-766.

Ecleo, J. J., & Galido, A. (2017). Surveying LinkedIn Profiles of Data Scientists: The Case of the Philippines. *Procedia Computer Science*, 124.

Gupta, M., & George, J. F. (2016). Toward the development of a big data analytics capability. *Information & Management*, 53(8), 1049-1064.

Ho, A., Nguyen, A., Paffod, J., & Slater, R. (2019). A Data Science Approach to Defining a Data Scientist. *SMU Data Science Review*, 2(3).

Marcolin, C., Becker, J. L., Wild, F., Schiavi, G., & Behr, A. (2019). Business analytics in tourism: Uncovering knowledge from crowds. *BAR-Brazilian Administration Review*, 16(2).

Marr, B. (2016). *Big Data: Will We Soon No Longer Need Data Scientists?*. Retrieved 30 November, 2019, from https://www.forbes.com/sites/bernardmarr/2016/04/27/will-we-soon-no-longer-need-data-scientists/#2242e2e26897.

Mauro, A. de, Greco, M., Grimaldi, M., & Ritala, P. (2018). Human resources for Big Data professions: A systematic classification of job roles and required skill sets. *Information Processing & Management*, 54(5), 807-817.

Mcafee, A., & Brynjolfsson, E. (2012). Big data: The management revolution. *Harvard Business Review*, 90(10), 61–67.

Mendelevitch, O., Stella, C., & Eadline, D. (2016). *Practical Data Science with Hadoop and Spark: Designing and Building Effective Analytics at Scale.* Addison-Wesley Professional.

Meyer, M. A. (2019). Healthcare data scientist qualifications, skills, and job focus: a content analysis of job postings. *Journal of the American Medical Informatics Association*, 25(5), 383-391.

Mortenson, M. J., Doherty, N. F., & Robinson, S. (2015). Operational research from Taylorism to Terabytes: A research agenda for the analytics age. *European Journal of Operational Research*, 241(3), 583-595.

Nassaji, H. (2015). Qualitative and descriptive research: data type versus data analysis. *Language Teaching Research*, 19(12), 129-132.

Provost, F., & Fawcett, T. (2016) *Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking*. Alta Books.

Tabesh, P., Mousavidin, E., & Hasani, S. (2019). Implementing big data strategies: A managerial perspective. *Business Horizons*, 62(3), 347-358.

Vidgen, R., Shaw, S., & Grant, D. B. (2017). Management challenges in creating value from business analytics. *European Journal of Operational Research*, 261(2), 626-639.