Term Paper

University of Applied Science - Online

Study-branch: Data Science

Differences and Similarities in Jobs in the Data Domain

Koen Bothmer

Student Number: 92014567

Ankersweg 39

realized in Holtum (NL)

Advisor: Tim Schlippe

Code and environment available at https://github.com/KoenBothmer/job_analysis

Delivery date: 02-05-2021

Contents

I List of Figures

1 Introduction

Data Scientists have been in high demand over the past decade, so much that the job has even been called 'sexy' [Davenport and Patil, 2012]. For a job that is so highly sought after there is still surprisingly little consensus on what the position of a 'Data Scientist' entails. In example [Taylor, 2016] illustrates the difficulties in defining this field by describing the most popular attempts that have been made to visualize it in Venn diagrams.

This lack of a definition makes the hiring process difficult as the interview process is highly dependent on the point of view on the definition of a Data Science position of the hiring manager. This gives ground to online communities like r/datascience [Reddit, 2021] with, at this moment, 468.807 members, whose focus is on discussing Data Science career questions.

The confusion is enhanced by the rise in demand of job positions in MLOps and Data Engineering [Mäkinen et al., 2021]. While there is some consensus that these are specialists that are able to engineer pipelines and deploy models created by Data Scientists it remains unclear what the differences and similarities in these roles are.

Given such a lack of definition, for those that call themselves Data Scientists, it makes sense to try and define themselves using a data driven approach. Such attempts have been made [Ho et al., 2019], [De Mauro et al., 2016]. One example of this approach is shown in the work by [De Mauro et al., 2016] who use scraping and text mining techniques to classify vacancies with the term 'Big Data' in their job title into different job families. This source provides the most common bigrams in each of it's job families.

In essence, the approach by [De Mauro et al., 2016] made use of web scraping and expert knowledge to classify vacancy texts into job families. To expand on this technique, in this work, these job families were used as input search terms for a similar experiment. Instead of using expert knowledge, the three bigrams, which are essentially job positions, were used as provided by [De Mauro et al., 2016]. Web scraping and text mining techniques were developed to examine what skills are required by employers in their vacancy texts for these roles.

This approach allows for research on the differences and similarities among these job positions allowing to formulate an answer to the question *What are the differences and similarities in skills employers require in Data Analysts, Data Scientists and Data Engineers?*

The methodology described in this paper is simple to reproduce and fully available at the Github repository accompanying this work [Bothmer, 2021].

The remainder of this paper is structured as follows: Section 2 describes the background of the question this paper seeks to answer, Section 3 describes the method used to mine vacancy texts in order to perform the research, Section 4 describes the results and in Section 5 the work is concluded.

2 Background

2.1 Defining the Data Scientist

In 2012 [Davenport and Patil, 2012] concluded that it is difficult to define Data Science given there is no broadly accepted academic context. Because of this, the focus in the job market is very much on tools required. The idea of [Davenport and Patil, 2012] is that, in order to flourish, those involved in the Data Science field should focus less on tools and more on the core principles underlying these tools instead. It is described that, in essence, the task of the Data Scientist is to enable better decision making by extracting knowledge from data.

To this avail Data Scientists have not always been in demand; Extracting knowledge and especially value from data used to be an endeavor for Business Intelligence Analysts. However, the technical skills needed have grown beyond the scope of a typical Business Intelligence specialist as data sources and datasets have grown too large to store and analyze using traditional technologies [Debortoli et al., 2014]. This has given rise to a whole field of Big Data professionals among whom the Data Scientist plays a central role. This shows as the demand for Data Scientists has grown with the supply of data [Roberts, 2000] and so has the whole Big Data Field. The skills required for Big Data Professionals have grown into a vast field of specializations as shown by [Debortoli et al., 2014]

2.2 Splitting the Data Scientist

An often cited complaint from past years has been that companies tried to hire so called "Unicorn Data Scientists" [Baškarada and Koronios, 2017], among job seekers the sentiment was that employers expected to hire someone with skills covering the whole Big Data Analytics work field. As the name suggests, these Unicorns are extremely rare, if not impossible to find [Baškarada and Koronios, 2017]. Adjacent to this topic a recent discussion has developed around the Data Science team and the specialists these teams contain. It seems to be the consensus that the most frequent roles in these teams are Data Analyst, Data Scientist and Data Engineer [Saltz and Grady, 2017]. As one would expect, the required skills for these positions have often been shown to overlap [Roberts, 2000][De Mauro et al., 2016][Ho et al., 2019]. How these roles overlap is an interesting topic for discussion as most of the cited work analyzes the different job positions in isolation, this work proposes a methodology to compare positions in relation to each other.

2.3 Analyzing Job Postings

This work expands on the ideas of [De Mauro et al., 2016] which shows how web scraping can be combined with text mining techniques to extract knowledge from job postings in the Big Data domain. This source provides the most common bigrams in the job titles of each of it's job families. In the Data Science job family these are 'Data Scientist', 'Data Analyst' and 'Data Engineer'. To expand on this work I zoomed in on these bigrams by web scraping specifically for jobs with these categories in their title. The idea for this paper is similar to that of [Ho et al., 2019] which uses similar pre-defined job search terms. The natural language processing and modelling techniques used by [Ho et al., 2019] are quite sophisticated and beyond the scope of this paper.

The works that are similar to this paper [Roberts, 2000][De Mauro et al., 2016][Ho et al., 2019] all employ methodology similar to the techniques used in this paper but they offer two opportunities to expand on:

- Expert Knowledge Required and complexity: These cited works all start by web scraping job postings after which text mining techniques are employed. After the text mining techniques, in all these works, a complex classification or clustering modelling step is applied. Eventually, these works all use expert knowledge to make the final classifications of job requirements.

- Reproducibility: The sections in the cited works that describe the methodology are clear but complex and not very in-depth, the scraped data and mined information and the code to produce these results are not shared. This makes these works difficult to reproduce.

This work offers a methodology that requires no expert knowledge and no complex modelling techniques making it applicable to many different contexts. As the technique is fully automated this can be done by just changing one line of the source code. For example, one could use the exact same technique to compare any three given job titles by just changing the search term. It is also convenient that this technique is easily repeatable so that changes in job requirements are easily tracked over time by repeating the experiment by just rerunning the provided code. This is important because Data Science is an evolving field [Saltz and Grady, 2017] [Davenport and Patil, 2012] which makes the skills required likely to be subject to change.

An opinion of not only me but many others is that researchers should share code and acquired data wherever possible [Peng, 2015]. One of the reasons not to publish code is that the code has a complex set of dependencies and is therefore difficult to share [LeVeque, 2013]; As experimental code is often very dependent on the environment it was developed in it can be difficult to get a piece of code to work on another environment. A common and stable way to solve this issue is to virtualize the developers environment in a Docker container that can be shipped with the experiment's code [Boettiger, 2015]. A more extensive discussion about sharing code and containerization is beyond the scope of this paper but I did include a GitHub link to this paper's notebook and Docker Container in order to maximize reproducibility [Bothmer, 2021].

3 Methodology

In this chapter the methodology and techniques are described that were used to acquire and analyze job postings from the internet. This chapter aims to explain the underlying assumptions and reasoning behind the choices that were made. For an in-depth explanation of the code itself I would like to refer to the code posted on the Github repository supporting this paper [Bothmer, 2021]. The included documentation and comment sections are extensive and I could not better explain them in plain text.

3.1 Web Scraping Job Postings

All job postings analyzed in this work were retrieved from Indeed [Indeed, 2021]. This choice was made because Indeed is the largest job site in the world [Indeed, 2021] and operates in a large number of countries. As it was strived for to web scrape only one job site, choosing the largest platform ensured that the maximum number of applicable job postings were found.

The scraping tool was developed using the Python package Beautiful Soup [Richardson, 2007] which enables the retrieval of data from common structures in the websites html code.

The developed tool takes in a search term and then scrapes the returned results pages for unordered list elements (list items "li" in unordered lists "ul") in the job description texts. List elements prove to be very likely to be requirements as all of the most common results show to be requirements.

The search terms that were used to retrieve the analysis data were 'Data Scientist', 'Data Analyst' and 'Data Engineer' as justified in chapter 2.

The scraping tool was used to analyze 50 vacancy texts for each category from Indeed's main page. The number of analyzed job postings is relatively low as opposed to similar works like [Debortoli et al., 2014][Ho et al., 2019]. This choice was made because the purpose of this work is explorative, the 150 scraped job postings showed to be sufficient to demonstrate the technique and generate enough support to formulate an answer to the research question. A more extensive analysis of a deeper dataset is beyond the scope of this work, some interesting recommendations to expand on this work can be found in chapter 4.

3.2 Pre-processing Listed Requirement Elements

A simple pre-processing step was performed on the scraped list element results in order to make them more comparable over different job categories. This step contains the removal of common non-alphanumeric characters and turning all the remaining elements to lower case, to enable better comparison of the results. From the alphanumerical, lower case elements common stop words were removed using a stop word dictionary offered by NLTK [Loper and Bird, 2002].

It must be stressed that no substantive changes that require human attention were made to the data, no returned elements were filtered out or altered making the analysis very easy to be repeated without needing substantive expertise.

3.3 Text Mining of Job Postings

In this paper some explanatory data analysis was performed by visualizing most common single words and most common bi-grams. In order to extract information from text elements bi-grams have been a popular choice [Tan et al., 2002]. The explanatory data analysis of the job requirements results clearly shows bigrams to be much more informative then single word counts. For example, the top 3 most common single words in 'Data Scientist' job postings as on 29-04-2021 were 'Data', 'Experience' and 'Science'. The top 3 most common bigrams in the same job postings show to be more informative as they are: 'Machine Learning', 'Data Science' and 'Computer Science'. The results of the explanatory data analysis are further discussed in chapter 4.

3.4 Visualizing Differences and Similarities

All visualizations were made using Python's counter collection, which offers the convenient 'Counter.most_ common()' method [Van Rossum and Drake, 2009]. Common terms in texts and counter values were used to plot Matplotlib bar charts [Hunter, 2007].

A popular and natural choice for visualizing intersections between skill sets in the data science domain are Venn diagrams [Taylor, 2016]. Works similar to this paper also show Venn diagrams [Debortoli et al., 2014][Ho et al., 2019]. Without the source code, it cannot be said with certainty, but these visuals look to be manually produced. Matplotlib offers the convenient matplotlib-venn extension [Hunter, 2007] which was used to fully automate the creation of a Venn diagram from nothing but the input search terms.

4 Results

This chapter provides an overview of the most common bigrams in requirements in job postings for the search terms provided in chapter 2: "Data Scientist", "Data Engineer", "Data Analyst". The choice for bigrams is supported by the results of a short exploratory data analysis that was performed. After this the most common bigrams for each category are visualized in isolation, finally a Venn diagram is shown that visualizes the differences and similarities among job categories in one overview.

4.1 Exploratory Analysis: Bigrams Over Single Words

The choice to represent requirements by most common bigrams was based on results from a short exploratory data analysis in which some text mining techniques [Tan et al., 2002] were compared in the context of this problem. Namely the difference between counting single word versus bi-grams shows clearly how the bigrams are a more appropriate technique for this paper's context. Comparing figure 4.1 and figure 4.2, which show most common words and most common bigrams respectively, it is clear that bigrams provide a lot more information. The bigram plot is much less distorted by disproportionately large categories and the actual categories are more informative.

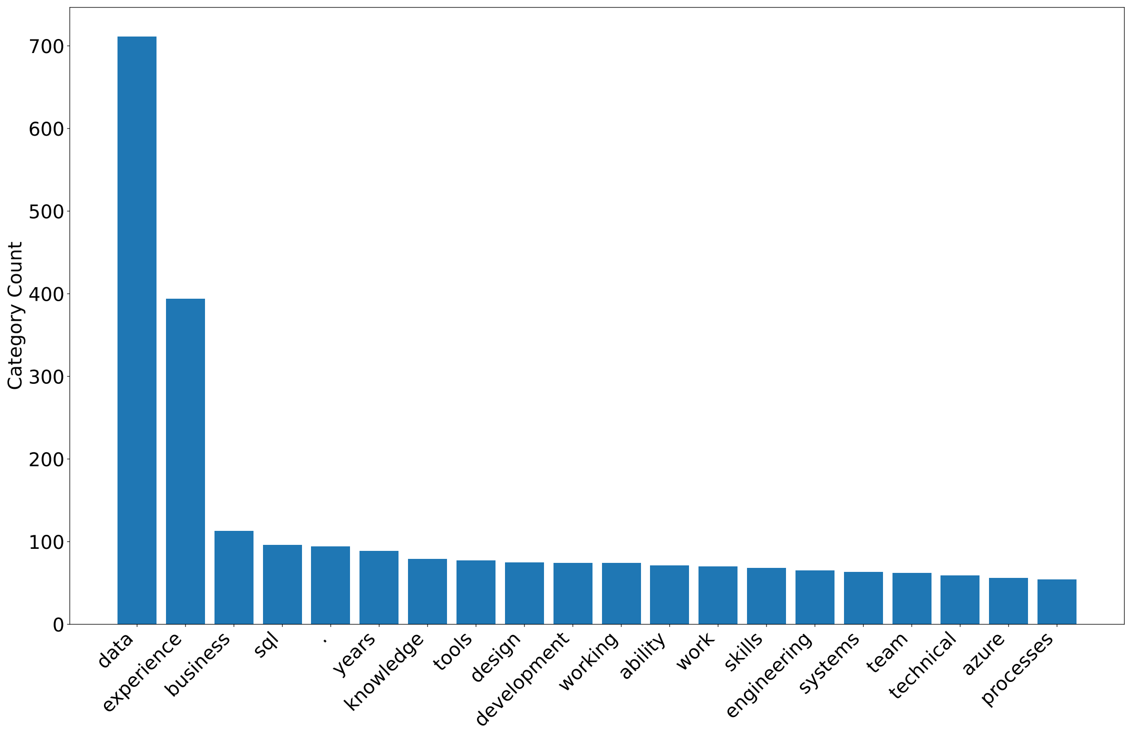

Figure 4.1: The 20 most common terms in Data Engineer job postings

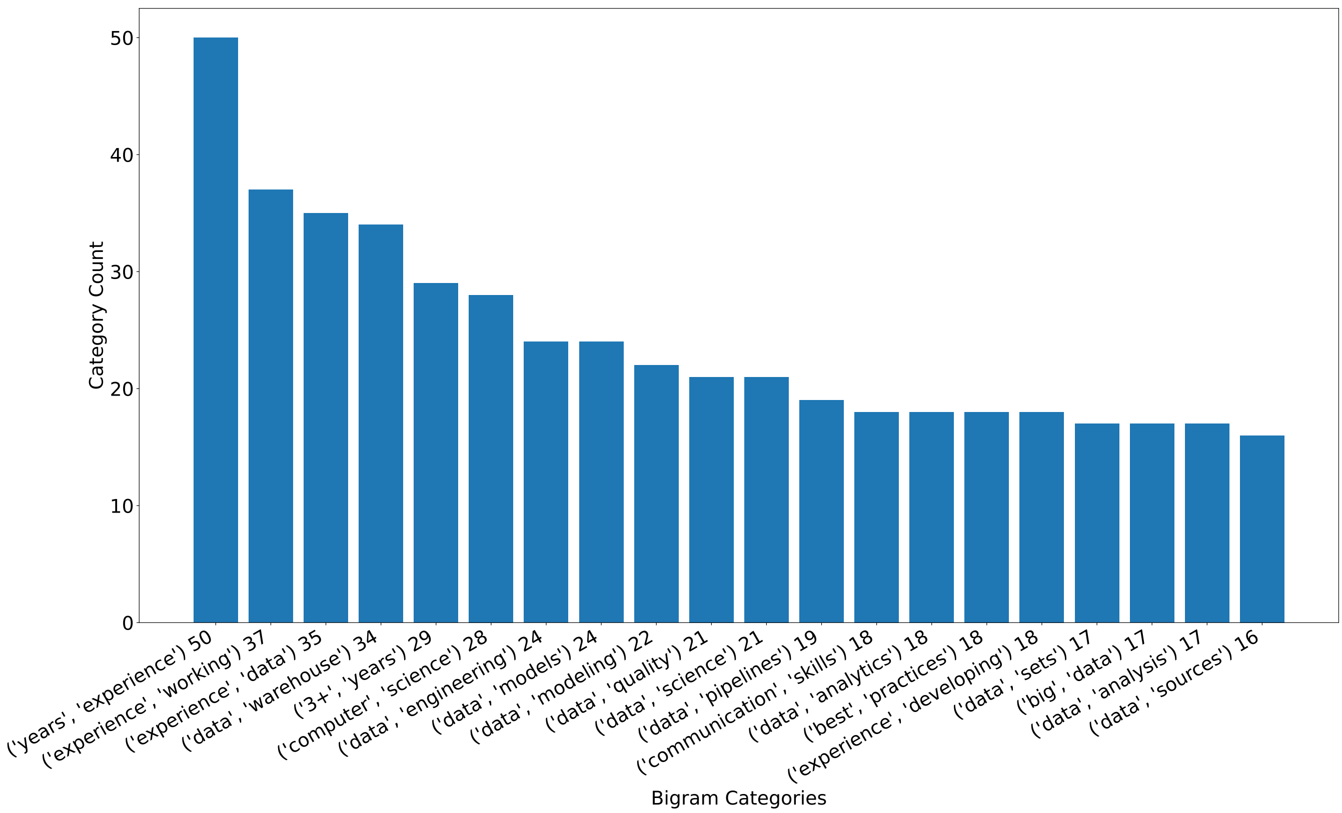

Figure 4.2: The 20 most common bigrams in Data Engineer job postings

4.2 Job Categories in Isolation

As described in chapter 3, for each of the job categories "Data Engineer", "Data Scientist" and "Data Analyst" the 20 most common bigrams are visualized in figure 4.2, figure 4.3 and figure 4.4 respectively. From these visuals, some differences and similarities start to emerge but the main purpose of these visuals is to get an overview of the most common bigrams in each of the job categories.

4.2.1 Data Engineer

Figure 4.2 shows a picture of the Data Engineer as an experienced worker, who requires the most technical skill set of the three categories. Experience in data warehousing, computer science and data modelling are very frequently required as is the ability to set up data pipe lines. Soft skills, while still present in "Communication Skills" are not as frequent as the technical skills described.

4.2.2 Data Scientist

In figure 4.3 it can be observed that the "Data Scientist" is the relatively most common bigram in the whole dataset. While an obvious requirement it is still a remarkable observation that the term is much more frequent then "Data Analysis" and "Data Engineering" for the "Data Analyst" and "Data Engineer" categories.

Figure 4.3 also shows "Machine Learning" is in particular high demand for Data Science roles. An image emerges of an employee who applies advanced analytics techniques and the backbone of his tools appear to be Python, R and SQL.

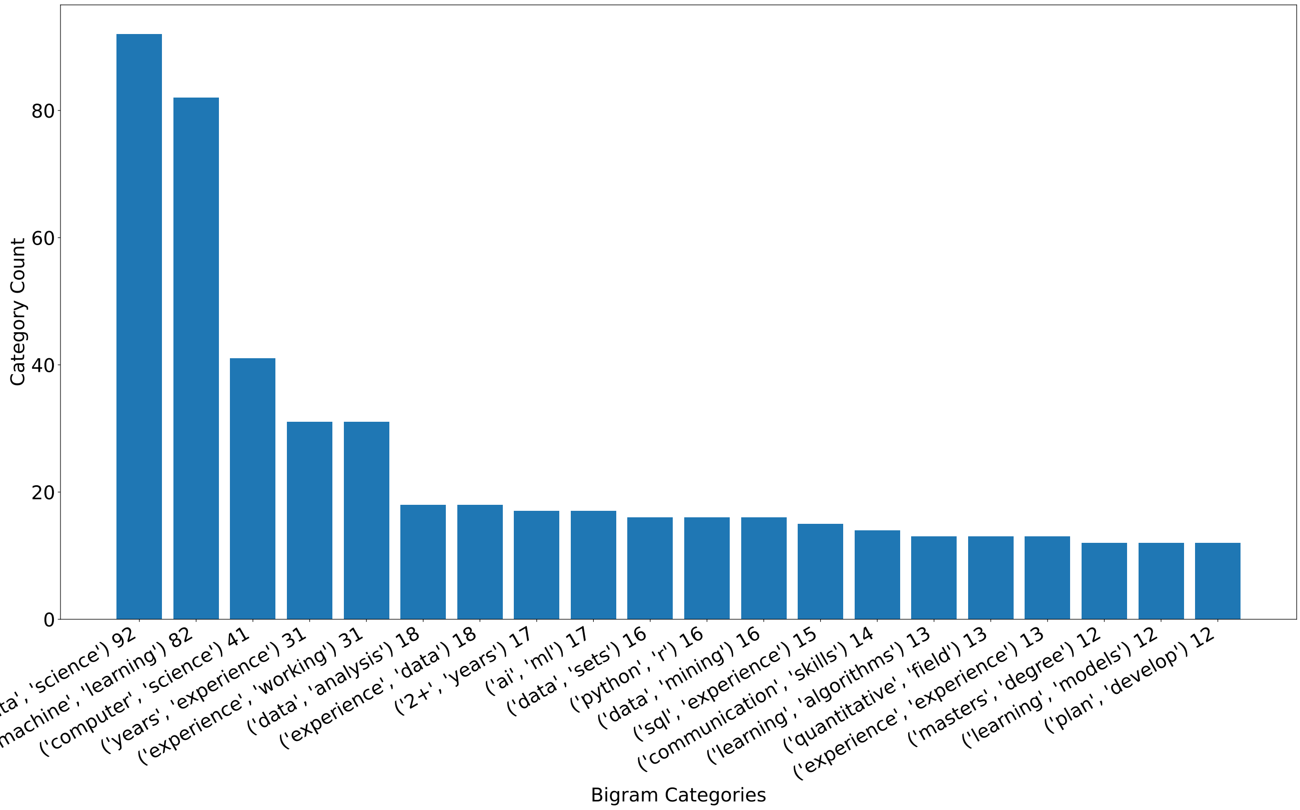

Figure 4.3: The 20 most common bigrams in Data Scientist job postings

4.2.3 Data Analyst

The most frequent skills required for Data Analysts are shown in figure 4.4. Still a technical role for experienced workers which requires abilities in computer science. Focus in this role seems to be more on creating value from existing datasets, as bigrams like "Data Analysis", "Data Visualization" and "Vizualization Tools" are frequently present in the requirements.

It is noteworthy that "Data Science" is a very frequent bigram in Data Analyst job postings.

4.3 Comparing Job Categories: Job Similarities

Figure 4.5 shows the differences and similarities of the 30 most common bigrams for each of the analyzed job categories. The Venn diagram clearly shows the similarities and differences among the three researched job categories. In this section these results are discussed starting with the similarities. The similarities are shown in the intersecting planes of figure 4.5. In the subsections that follow each intersecting plane is discussed.

4.3.1 Data Scientist ∪ Data Engineer ∪ Data Analyst

At the center of the diagram the bigrams that were common in all of the analyzed job postings are displayed. As discussed in section 4.2 all roles are data and data analysis centered and have a strong computer science component. It is noteworthy that the "Data Science" bigram was frequent in all of the analyzed job categories. Alltough there is a high number of Data Scientist job postings available online, it seems that some skills in the Data Science domain are considered common ground for all of the analyzed positions. "Communication Skills" is also a bigram that is present in the top 30 of all of the job categories.

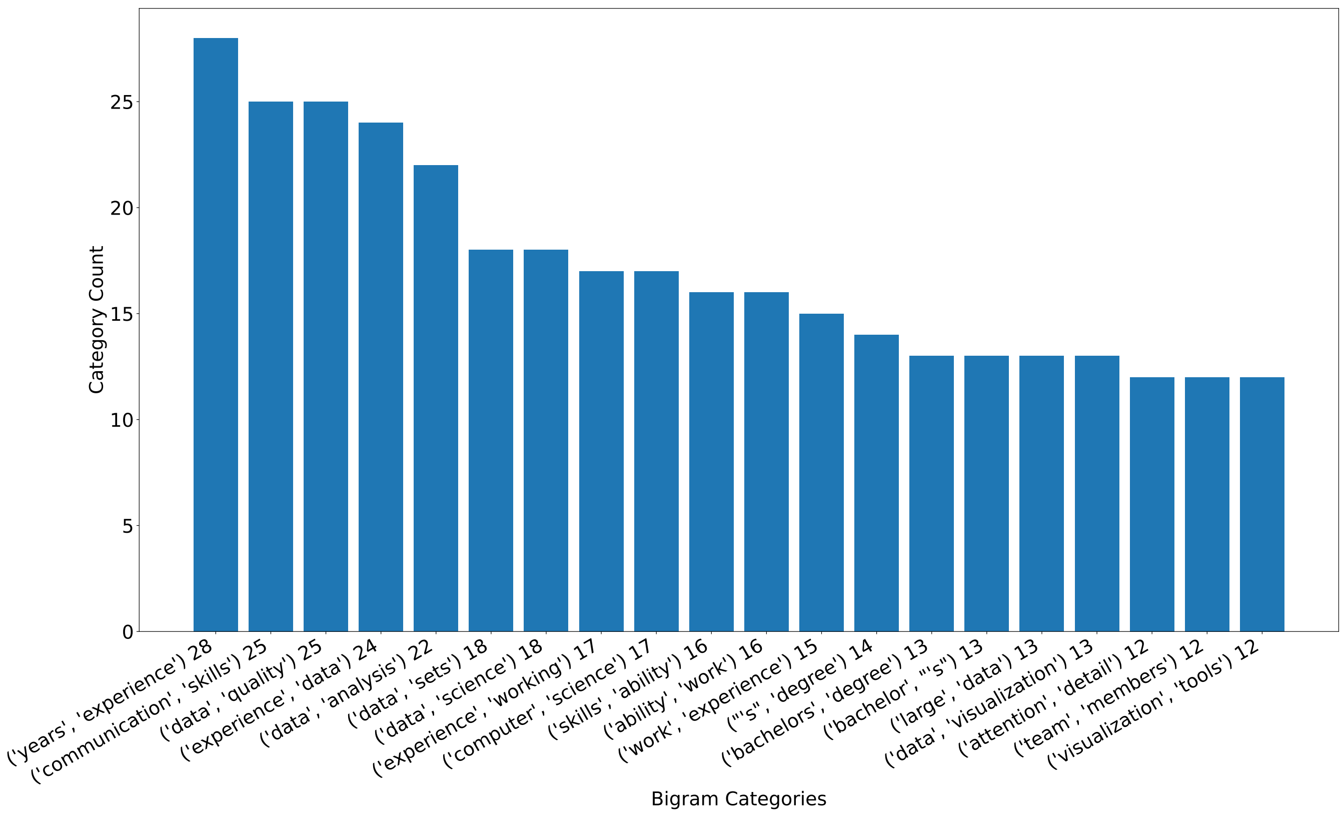

Figure 4.4: The 20 most common bigrams in Data Analyst job postings

4.3.2 Data Scientist ∪ Data Analyst ∩ Data Engineer

Bigrams that are common in both Data Scientist and Data Analyst are "Statistical Analysis", "large data" and "Python, R". Especially the scripting requirements for Data Analysts were surprising to the author.

It must be noted that none of these bigrams are present in the Data Analyst top 20 bigrams as shown in figure 4.4, so alltough present in the top 30 of Data Analyst, these bigrams all came from the lowest ranking 10 positions.

4.3.3 Data Scientist ∪ Data Engineer ∩ Data Analyst

To the surprise of the author, the intersection of Data Scientist and Data Engineer only shows the not very informative bigram "data sources". Based on this analysis of job postings, it seems that job requirements for Data Scientists and Data Engineers are well defined and consistent throughout the analyzed vacancy texts.

4.3.4 Data Analyst ∪ Data Engineer ∩ Data Scientist

The bigrams at this intersection show a demand for data quality, the collaborative nature of both jobs is shown by the "team members" bigram. Data analytics is a common bigram for both job categories. It is remarkable that the most frequent education level is a bachelor's degree where Data Scientist shows the requirement of a master's degree in it's most frequent bigrams.

4.4 Comparing Job Categories: Job Differences

Each of the three job categories have a long list of bigrams that are distinctive in relation to the other two job categories. Where the intersections show a common 'base knowledge' required for any of these job categories,

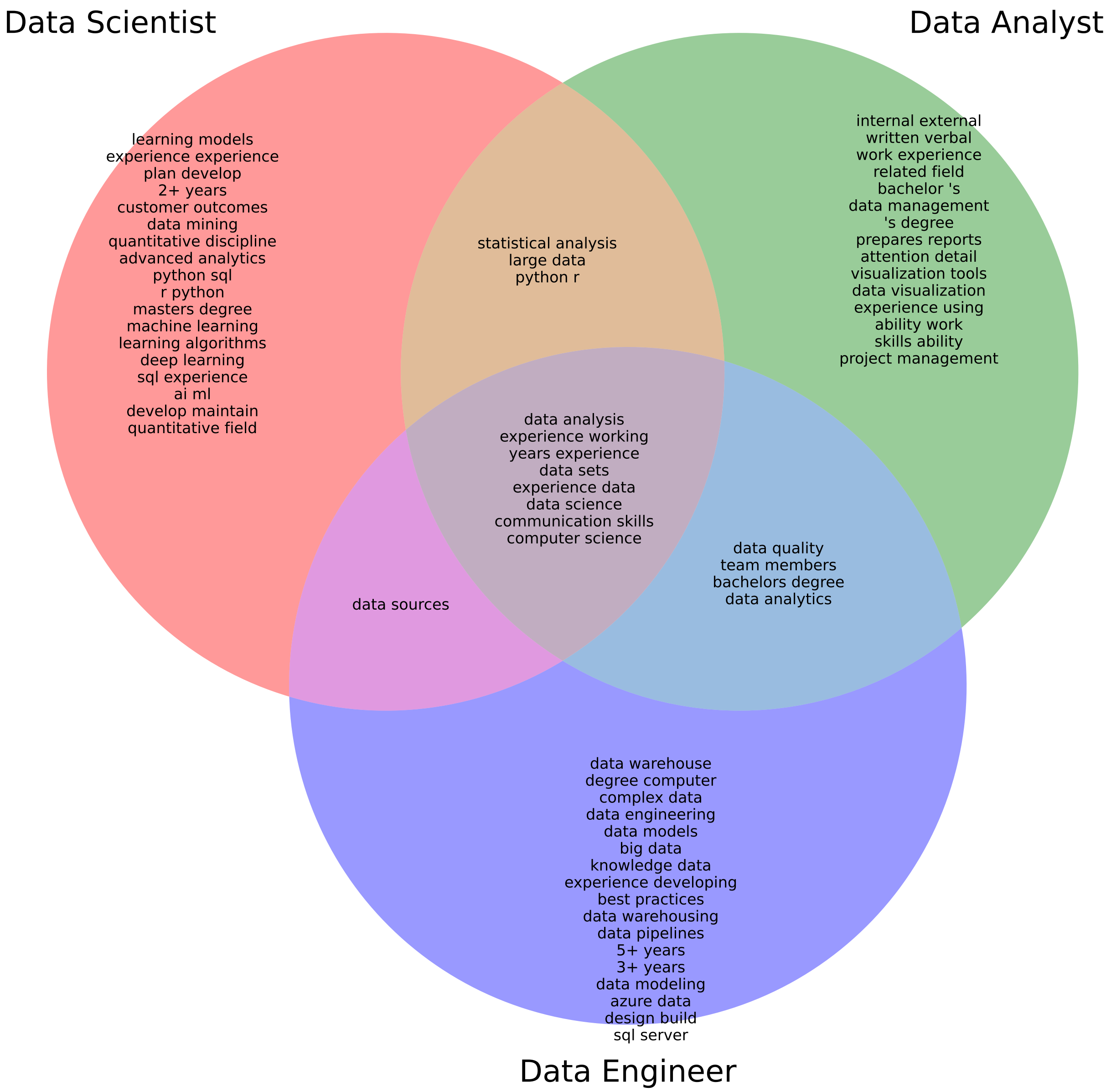

Figure 4.5: A Venn Diagram of the 30 most frequent bigrams in 3 job categories

the disjoint sets of each job category are interesting as they give a better sense of what exactly is the distinct specialty of each job category based on the requirements of employers.

4.4.1 Data Scientist ∩ Data Analyst ∩ Data Engineer

The disjoint set of the data Scientist reveals a specialization of advanced analytics, Machine Learning and Deep learning models. Tools to perform these advanced methods are Python, R and SQL. An image emerges that confirms the views of [Davenport and Patil, 2012], the Data Scientist is a specialist that uses statistical techniques to turn structured and unstructured data into actionable information. It is noteworthy that this is the only job category in which a Master's degree is a common requirement.

4.4.2 Data Analyst ∩ Data Scientist ∩ Data Engineer

The bigrams that are only frequent in Data Analyst job postings show the most business and soft skills related requirements of all of the skillsets. Bigrams like "written verbal", "attention detail", "project management" and "prepares reports" all show this.

All tough there are some technical requirements at the intersecting planes, the specialization of the Data Analyst seems to be their ability to visualize and communicate results.

4.4.3 Data Engineer ∩ Data Analyst ∩ Data Scientist

The Data Engineer specialization shows a very large amount of highly technical requirements in it's most frequent, disjoint, bigrams. "data warehouse", "data pipelines" and "data modelling" are some examples of this but almost all of the most common bigrams show the requirement of a very strong technical skillset for the Data Engineer.

The image that arises confirms the views of [Saltz and Grady, 2017] of the data engineer as a specialist that enables the advanced analytical work of data scientists to be used to create value by setting up the required data warehouses and data pipelines.

Of all roles, the highest number of years of experience are demanded for Data Engineer positions.

5 Conclusions

Returning to the main question of this paper: *What are the differences and similarities in skills employers require in Data Analysts, Data Scientists and Data Engineers?* While figure 4.4, figure 4.2 and figure 4.3 give a general idea of what the requirements for each job entail, the most extensive answer that can be given based on the performed research is to refer to figure 4.5. The created Venn diagram is the cornerstone of this work and also instrumental to the answer as formulated:

5.1 The Differences Among Data Scientists, Data Analysts and Data Engineers

As argumented in chapter 4, the differences in job classes are represented by the bigrams in the disjoint sets of each category in figure 4.5. These disjoint classes show the specialization of each job category relative to the other categories.

Data Scientists specialize in advanced and predictive analytics. Requirements for Data Analysts are the most business and communications oriented. Data Engineers require the deepest technical skillset.

5.2 The Similarities Among Data Scientists, Data Analysts and Data Engineers

The similarities among the job classes can be found at the intersections of figure 4.5. The center of the diagram, Data Scientist $\cup$ Data Analyst $\cup$ Data Engineer shows a common experience base that is required for all three researched job categories.

It is surprising how the bigrams "Data Science" and "Computer Science" are present in all three job classes, this might indicate that some extend of experience in data science is required even in Data Analyst job postings. Adjacent to this central requirement the bigram "python r" was found in the intersection of Data Scientist and Data Analyst. This also indicates that the technical demands for Data Analyst are more advanced then the author would have expected.

5.3 Reccomendations

In conclusion, it is impossible to give an exhaustive answer to the research question based on just one technique. It was shown that analyzing job postings can contribute to an *up to date* understanding of the differences and similarities among different specializations in a common field like the data domain.

This work offers an interesting opportunity to expand on this repeatable approach by iteratively conducting this experiment over time to build a dataset of how job requirements change over time. This is interesting as Data Science is a young and evolving field [Saltz and Grady, 2017], the required skillsets are expected to shift and change.

Because the setup of the experiment is relatively easy to repeat in different contexts it would be interesting to make use of this by making some alterations to the change terms and analyzing the results. Differentiating in geographic location could be interesting, to research whether job requirements differ in different locations.

The proposed methodology also offers interesting research opportunities to analyze intersecting job categories in fields outside of the data domain.

The results are in line with the findings of similar works [Roberts, 2000][De Mauro et al., 2016][Ho et al., 2019]. However, these works rely on more complicated methodology and expert knowledge to make classifications. The code used in the experiments is not shared. The employed methodology for this paper makes this work reproducable without expert knowledge. To facilitate this all code is published in a reproducable way on this paper's Github repository [Bothmer, 2021]. This is unique in this line of work.

Bibliography

[Baškarada and Koronios, 2017] Baškarada, S. and Koronios, A. (2017). Unicorn data scientist: the rarest of breeds. *Program*.

[Boettiger, 2015] Boettiger, C. (2015). An introduction to docker for reproducible research. *ACM SIGOPS Operating Systems Review*, 49(1):71–79.

[Bothmer, 2021] Bothmer, K. (2021). Job analyis. `https://github.com/KoenBothmer/job_analysis`.

[Davenport and Patil, 2012] Davenport, T. H. and Patil, D. (2012). Data scientist. *Harvard business review*, 90(5):70–76.

[De Mauro et al., 2016] De Mauro, A., Greco, M., Grimaldi, M., and Nobili, G. (2016). Beyond data scientists: a review of big data skills and job families. *Proceedings of IFKAD*, pages 1844–1857.

[Debortoli et al., 2014] Debortoli, S., Müller, O., and vom Brocke, J. (2014). Comparing business intelligence and big data skills. *Business & Information Systems Engineering*, 6(5):289–300.

[Ho et al., 2019] Ho, A., Nguyen, A., Pafford, J. L., and Slater, R. (2019). A data science approach to defining a data scientist. *SMU Data Science Review*, 2(3):4.

[Hunter, 2007] Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95.

[Indeed, 2021] Indeed (2021). About indeed.

[LeVeque, 2013] LeVeque, R. J. (2013). Top ten reasons to not share your code (and why you should anyway). *Siam News*, 46(3).

[Loper and Bird, 2002] Loper, E. and Bird, S. (2002). Nltk: The natural language toolkit. *CoRR*, cs.CL/0205028.

[Mäkinen et al., 2021] Mäkinen, S., Skogström, H., Laaksonen, E., and Mikkonen, T. (2021). Who needs mlops: What data scientists seek to accomplish and how can mlops help? *arXiv preprint arXiv:2103.08942*.

[Peng, 2015] Peng, R. (2015). The reproducibility crisis in science: A statistical counterattack. *Significance*, 12(3):30–32.

[Reddit, 2021] Reddit (2021). A place for data science practitioners and professionals to discuss and debate data science career questions. `https://www.reddit.com/r/datascience/`. Accessed: 2021-04-25.

[Richardson, 2007] Richardson, L. (2007). Beautiful soup documentation. *April*.

[Roberts, 2000] Roberts, L. G. (2000). Beyond moore's law: Internet growth trends. *Computer*, 33(1):117–119.

[Saltz and Grady, 2017] Saltz, J. S. and Grady, N. W. (2017). The ambiguity of data science team roles and the need for a data science workforce framework. In *2017 IEEE international conference on big data (Big Data)*, pages 2355–2361. IEEE.

[Tan et al., 2002] Tan, C.-M., Wang, Y.-F., and Lee, C.-D. (2002). The use of bigrams to enhance text categorization. *Information processing & management*, 38(4):529–546.

[Taylor, 2016] Taylor, D. (2016). Battle of the data science venn diagrams. *KDNuggets News*.

[Van Rossum and Drake, 2009] Van Rossum, G. and Drake, F. L. (2009). *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA.