Visualizing Data Science in the Modern Day: SIADS 593 Milestone I

By: Patrick Thornton, Pu Zeng, Braedon Shick

I. Motivation

Our project is pieced together with personal passion and curiosity within our career field of choice. Each team member is bringing a unique perspective into the field of data science and is interested in learning about the past and present of data science. As a group, we have read articles such as Forbes' *The Top 5 Data Science and Analytic Trends in 2023* and Deloitte's *Marketing Data Science Trends* independently concluded there is a lack of merging academia and industry trends. This void of information led us to scour the web to find data sources that pertained to either the industrial, or academic world. Finding promising sources for each division, we were driven to create this report showing how they might be related. After our project, one will be able to understand these trends, and capitalize on them.

II. Data Sources

A. Data Sets Overview

Data Set	File Type	Location	Time Period	Records
CSRanking	CSV	https://github.com/emeryberger/CSrankings	1969 - 2022	81,904
QS Ranking	CSV	https://www.kaggle.com/data sets/jkanthony/world-university-rankings-202223	2023	1,422
Stanford AI Report	XLSX	https://docs.google.com/spreadsheets/d/1AAIebjNsnJj_uKALHbXNfn3_YsT6sHXtCU0q7OIPuc4/edit#gid=0	2023	537
Google Scholar data set	CSV	https://www.kaggle.com/data sets/victoryerz/ms-data-science-universities	1741 - 2022	22,764
Lookup Tables for Visualization (Country and university information)	CSV	1)https://github.com/lukes/ISO-3166-Countries-with-Regional-Codes/blob/master/all/all.csv 2)http://goodcsv.com/education/north-american-colleges-universities/ 3) Data Science Universities across US	2019 2023 2016	1) 250 2) 137 3) 954
Email domain of Different Universities	JSON	https://github.com/Hipo/university-domains-list	2017-2023	9,993
Kaggle Machine Learning & Data Science Survey	CSV	https://www.kaggle.com/competitions/kaggle-survey-2022	2017 - 2022	130,298

B. Academic Data Sets

The **CSRanking** data set provides the annual publications of each scholar in top universities together with their domains and institutions.

The **QS Ranking** data set provides key information about the top 1,500 institutions around the world. This data set is used as a lookup table so we can find the location of universities in the CSRanking data set.

The **Stanford AI Report** Model data set provides valuable information about Parameter, Computing, and Data Trends in Machine Learning.

The **Google Scholar** data set is extracted from Google Scholar profiles using python modules to obtain this information. We used the data to compare the strength of data science research between different universities. Given the data set doesn't provide us with a good variable about the university name, we need to use the author name or email domain as keywords to search the university name in other lookup tables.

The **Email domain of Different Universities** data set has a list of all universities and their domain names. The data set has over 180 contributors helping maintain and consistently updated.

C. Important Academic Data Sets Features

Data set	Variable Name	Data Type	Attribute Type	Description
CSRanking	Author	string	Nominal	Author Name of the Publication
	Year	datetime	Quantitative	Publication Year
	Title	string	Nominal	Publication Title
	University	string	Nominal	The University Name of the Scholar
Google Scholar Data Set	Name	string	Nominal	Author Name of the Publication
	Cited By	int	Quantitative	Total Citations of the Publication
	Publication_cites _per_year	dict	Quantitative	The Number of Paper's Citations each year
	Email_domain	string	Nominal	The Email Domain of Each Scholar
Model Data Set	System	string	Nominal	The Name of the Model
	Organization	string	Nominal	The Organization Name of the Model
	Organization Categorization	string	Nominal	The Categorization of the Organization (Academia, Industry, or Both)
Lookup Tables	Institution	string	Nominal	The Institution Name of Scholars
	Country	string	Nominal	The Country of the University
	Domain	string	Nominal	The Email_Domains of Universities
	Longitude	float	Quantitative	The Longitude of the University
	Latitude	float	Quantitative	The Latitude of the University

D. Industry Data Set

The **industry data set** we used is a combination of the Kaggle Machine Learning & Data Science Survey from 2017 to 2022. Each year the survey was conducted between September 16th to October 16th. These data sets were downloaded from Kaggle's website and the links to each specific data set is included in the works cited section of this document. The survey has over 300 questions where each sub question is in its own column. The questions themselves have also evolved over the years and there are only 17 identical questions from 2018 and 76 from 2019.

In the raw data file, the columns represent each question, or sub-question, and are represented by their question number (Ex. Q7, Q7a, etc.), and change numbers based on the year. For clarity's sake we will be referring to them as their names post data manipulation. The variables of interest were Country, YearlyComp, EduLevel, Title, and Age.

E. Important Industrial Data Sets Features

Variable Name	Data Type	Attribute Type	Description
Country	string	Nominal	Respondents Country of Residence
YearlyComp	float	Quantitative	Salary of Respondent in USD
EduLevel	string	Nominal	Respondents highest level of education
Title	string	Nominal	Current job title of respondent
Age	float	Ordinal	Respondents Age
Various Topics	string	Nominal	Various skills or technologies respondents through were relevant to the question's topic

III. Data Manipulation Methods

A. Academic data set

[Notebook-Academic]

1. Publications

The CSRanking data set provides CSV files for publications and scholars. We combined these two CSV files on the scholar's name to reconstruct the institution information of each paper.

The CSRanking data set has an incomplete CSV file for institutions and their corresponsive country. To get the country variable, a good solution is to scrape the Nominatim website by the name of the institutions. It can provide nearly all the data needed, but an error of 403 or 502 would soon occur as there are 21,477 unique institutions in the data set.

Our solution is to combine the CSRanking data set with the QS Ranking of Universities data set. The university names have to be manipulated as CSRanking uses "Univ." instead of "University." There are other differences, such as "Univ. of Texas at Austin" can be found instead of "The University of Texas at Austin" and "University of California - Berkeley" instead of "University of California, Berkeley." Regex is used to replace all the non-word characters and convert the names to lower cases to build a "key" variable to join the universities in different data sets. The QS Ranking data set only includes 1,477 universities compared with 21,477 institutions in the CSRanking data set.

2. Models

The model data set provides the names, dates, organizations, and organization categorizations of each important model, so we still need to find the country of each model. As nearly half of the models are built by companies, the data set in publications is not very useful. Though there are only 356 unique organizations in this data set, the Nominatim cannot work correctly since large companies have branches worldwide, and it's challenging to find the headquarters automatically.

We scraped Bing search to get the country variable. If a question like “Which country is Google belongs to?” is asked, Bing will highlight the answer to the problem. Therefore, it’s easier to use libraries such as BeautifulSoup to get the answer. However, Bing cannot provide the country for each organization. It can answer where Google is, but it failed with Google Brain. We will replace the nan value, if the organization name is contained by another organization name with a valid country variable. Finally, the country variables for all the organizations were successfully reconstructed.

Bing also returns different names for the same country, such as “U.S.,” “The U.S.,” “American,” or even “U.S. state of South Carolina” for “United States.” A dict variable is constructed to replace the different names with one.

3. Scholar

The scholar data set is obtained from Google Scholar and does not provide information about scholars' affiliated universities directly. Instead, it offers descriptions like "Professor of EECS, University of Michigan." We combined two methods to reconstruct the institution's information.

First, join the scholar's name with the CSRanking data set, which has a CSV file for all the scholars and their institutions in the CSRanking.

Then use the Email domain on the scholar's profile to get the university information. A complete match may not be appropriate, as "cs.stanford.edu" and "standford.edu" both belong to Stanford, but the email domain data set only has the latter. So, the university name is constructed if it contains a particular email domain, like "stanford.edu." The shorter domain must be matched first, and parallel computation should not be used, so even if "nyu.edu" is recognized as Yeshiva University because it contains "yu.edu," it will be corrected with the for-loop of "nyu.edu."

With these two steps, we reconstructed 92.84% of university information for the 9,794 scholars.

B. Kaggle data set

[Notebook-Kaggle]

The primary hurdles with the Kaggle data set were formatting of the columns, and the variety of questions over a five-year span. In the data sets, the questions with multiple answers had their answers encoded to their own columns. As an example, assume the main question asked, “What is your preferred computer language to work with?” If there were a list of 5 options, then there would be 5 sub-questions with the different options as answers. The difficulty caused by having multiple years was that the questions order and phrasing would change year to year, with some questions being added or removed entirely.

To utilize the information in these data sets, there were three approaches. The first, to analyze each year on their own, and compare the resulting visuals. The second, was to find the columns that were similar enough between the data sets, and combine the data sets using only those columns. Finally, the last solution was to recombine the encoded columns, manually identify questions that were worded differently but asking the same question between the years, rename those columns, and then combine the different years.

While the first option was the simplest, the third provided us with the most useful information, and is what we decided to go with. Decoding the sub question columns was the brunt of this approach, but prior to getting to this step, we cleaned up the data set significantly by removing any free text answers. This process was done by identifying any text response column using regex with the pattern “- Text”. Before we could combine the encoded columns, we needed to adjust the encoded column names back to their base name. An example of this would be taking “What is your preferred computer language to work with? – Selected Option – Python” to “What is your preferred computer language to work with?”. Luckily the structure of these questions was always similar, with the question, followed by the “– Selected Option – Python”. Utilizing this pattern, we split the columns on hyphens, and renamed the columns to the first element in the list resulting from the split. From here we applied a merge function between columns with the same name, inserting a comma between the columns data points.

Once the issue with the sub questions was fixed, we moved on to selecting only columns that were consistent between the years. Some columns, like compensation and country were included in all 6 years of the survey, but others only became consistent post 2019. To get the most out of the data, we decided to do a few visuals with all 6 years to show a longer timeline, but for most of the work we decided to only look at 2019-2022. With these 4 years as our focus, we manually read the remaining questions names and responses, identified which ones were the same across all the years, and filtered down to those 29 columns. Finally, we created a mapping for each year to rename the columns with consistent naming and combined them. A byproduct of only using the last 4 years was that Kaggle had also became consistent with response options. Previously, there were conflicts of country names across survey options in different years, such as "United States of America" and "United States" can be found, but in the last 4 years, they are all "United States of America".

For a few of our visuals, we needed numeric data points, but most of the columns gave us a range (Ex. Yearly Compensation = $50,000-60,000). In order to make these features usable, we wrote a function to take in a column of a data set, strip the values of the unwanted text, split the column on the hyphen, and find the mean between the resulting split. There were a few cases where we needed to replace the values manually, such as when we replaced “I have never written code” with 0 to indicate the number of years the respondent had used data science techniques when writing code.

To count the frequency of the Top3 answers, some inaccurate results from previous steps need to be manipulated. For example, the option "Jupyter (JupyterLab, Jupyter Notebooks, etc)" in the original columns are split into a list like ['Jupyter (JupyterLab', ' Jupyter Notebooks', ' etc) ']. Unstacking the list will cause repetitive counting. We need to replace 'Jupyter (JupyterLab' as 'Jupyter' and set others to '' value to avoid this problem.

1. Outliers

As the respondents volunteered to finish the survey, the result cannot represent the whole population. For example, Zimbabwe has a high average salary in 2022 and no data for the other years. It’s more reasonable to drop countries with more than three missing data in 5 years to compare salaries worldwide.

IV. ANALYSIS

ACADEMIA

[Notebook-Academic]

A. The Number of Publications and Scholars

In the year 2022, publications are primarily written by scholars in North America, East Asia, and Europe. While these regions have all made significant contributions to data science research, the United States and China are the most dominant players in the space. In Figure 1 below, we can see this distribution across organizations in the United States and China. One prominent trend in the past ten years in the quantitative perspective of data science is the rapid growth of China and China’s universities.

If we calculate the sum of publications in history, the top universities are similar to 2022. Carnegie Mellon University (CMU) ranks top 1 with 907 publications, Tsinghua University, Peking University, University of Illinois Urbana-Champaign (UIUC), and Stanford rank 2-5 with 682, 679, 611, and 595 publications in data science.

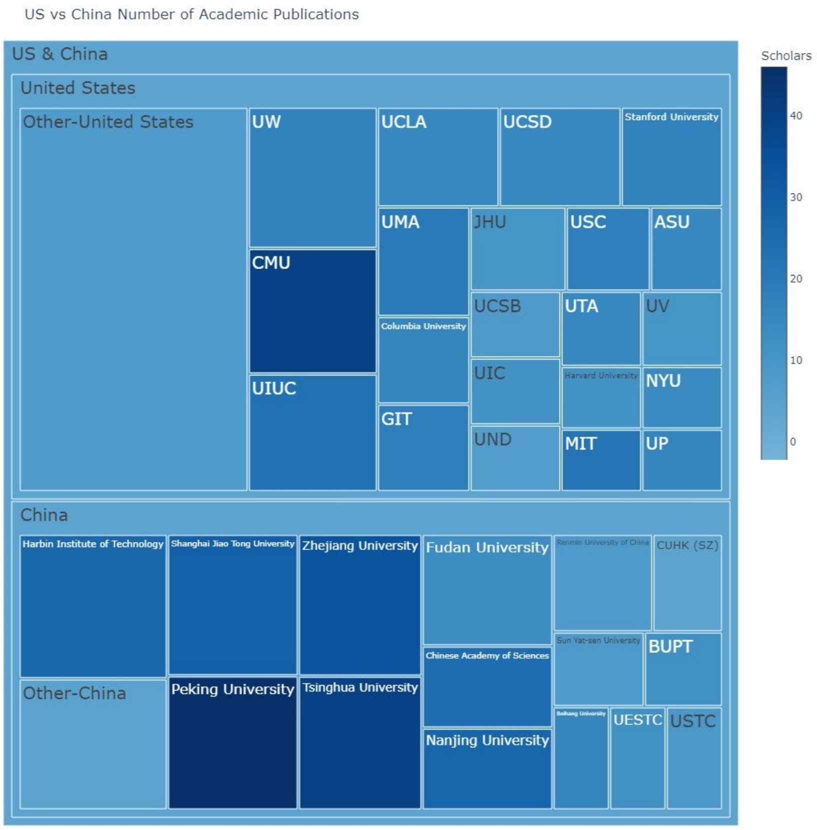

Figure 1 :US vs China Number of Academic Publications

B. Model Title Analysis, US vs China

The strength of data science research is more evenly distributed in the United States than in China. The number of publications is more or less the same across top universities in the US, while the top 10 universities contributed almost 80% of publications in China.

Chinese and American publications also differ in the keywords used in their titles. In the US titles are more likely to contain words like “model”, “using”, “algorithm”, “analysis”, and “data” while Chinese scholars are more likely to use “based”, “via”, “network”, “aware”, and “recommendation” in their titles. There may be a difference in original work and application work here, and we will discuss it in the next section.

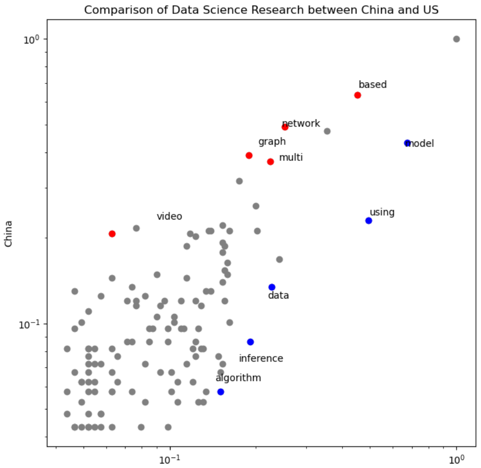

Figure 2: Comparison of Data Science Research between China and US

C. Important model origins

Stanford University’s AI Index Report 2023 identifies the most important AI models in history, which helps trace the origin of significant progress in data science. The number of important models is a good indicator for original work. This data set includes models from Theseus in 1950, Perceptron Mark I in 1957, to GPT 4 in different machine learning domains (Figure 3). Using this indicator, we can distinguish breakthroughs from cases where a breakthrough discovery was used.

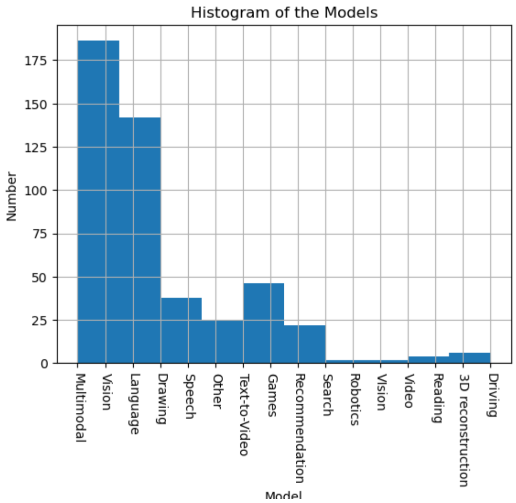

Figure 3: Histogram of the Models

Compared with the number of publications, which are more evenly distributed across countries, it’s interesting to find that the US dominated the area of important models with 496 out of 738 authors of the models. The numbers of original models worldwide are shown in Figure 4 with a log scale, as the differences between the United States and other countries are significant.

Though China shows rapid growth in the number of publications and scholars, the trend is less significant in the area of original models (Figure 5). This change in trend relative to the previous section indicates that while the field of data science is expanding in China, the cutting-edge research still takes place in the US.

Another interesting trend is that though academia and industry created roughly the same number of models in total (294 vs. 241), industry outperformed academia since 2017. In 2022, 34 models were contributed by the industry, while only two were from academia.

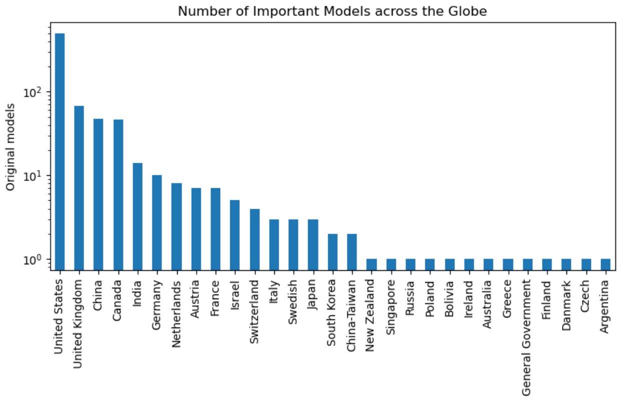

Figure 4: Number of Important Models across the Globe

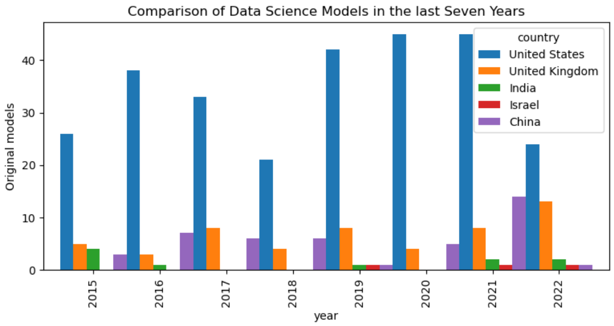

Figure 5: Comparison of Data Science Models in the last Seven Years

D. Density of citations by publication origin

The number of citations is another way to evaluate the quality of data science work, as essential publications tend to be cited more. Using a data set obtained by Google Scholar, we can estimate the number of citations each publication received. Insights from this data set can only be directional however, as not every country has equal access to Google, so papers that those countries might cite, wouldn't be captured in this data set.

With this bias in mind, we narrowed the scope of this analysis to the United States only, to see how the citation activity corresponded to organization location. Looking at Figure 6 below, we see that citation data is primarily focused around the east and west coast, along with the great lakes area. Referencing back to Figure 1, we saw a considerable number of publications can be attributed to universities like UC Berkeley (UCB), Stanford, or Carnegie Mellon (CMU). This trend aligns with the map shown in Figure 6 as there is significant activity in Northern California where a few of the higher publication count universities are located. What this overlap tells us is that not only are these universities producing a significant number of publications, but that they are also high quality.

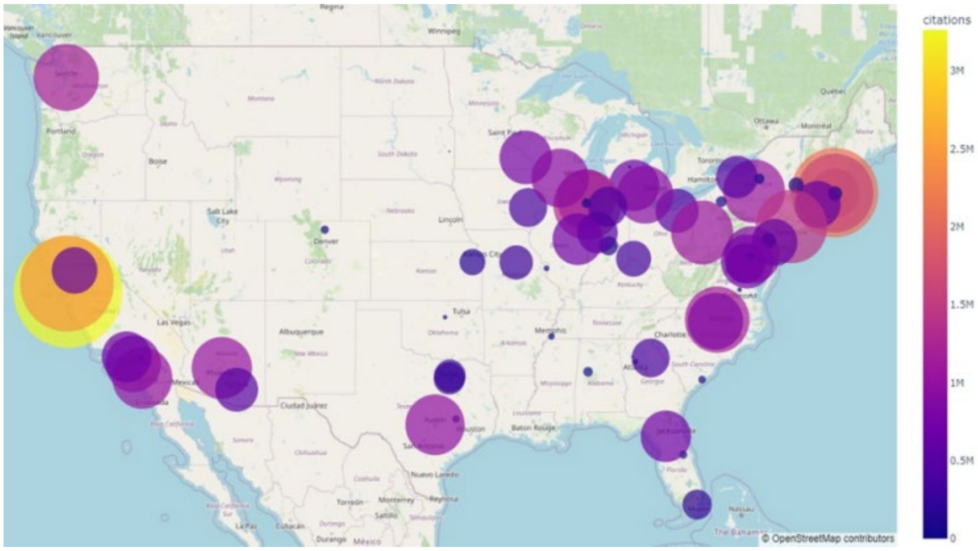

Figure 6: Location and Quantity of Citations in the US

Another insight we can glean from Figure 6 is that universities around the great lakes produce higher quality reports given the high number of citations relative to the number of publications seen from these universities in Figure 1.

INDUSTRY

[Notebook-Kaggle]

E. The distribution of salaries across countries

Moving from the academic world to the industrial one, we can better measure the strength of data science across the world using job data. To obtain this data, we are using data sets from Kaggle's annual Machine Learning and Data Science Survey. Kaggle has been running this survey since 2018, but over the years, and especially the earlier ones, the survey went through several changes in formatting and content. For this reason, we will be focusing primarily on 2022's result, only including the previous years where a timeline is indicated.

To begin our analysis, we wanted to look at the nationality of the respondents of the survey, to see how much representation the data set had. In 2022, India had the most representation with 8,792 respondents, and 2,920 were from the United States. Brazil, Nigeria, Pakistan, Japan, China, Egypt, Mexico, Indonesia, Turkey, Russia, and South Korea are other countries with more than 300 respondents. Part of the gap in respondents can be explained by the difference in population. For example, the difference in respondents between India and the US is right around a 3:1 ratio in India's favor, but the population difference is closer to 4:1.

With the Kaggle Survey data set, the average salaries in different countries can be calculated and shown in a geographical heat map. It's unsurprising that data scientists earn more in the developed world. The United States ranks first with an average salary of $144,999 per year; Australia and Israel are the other countries with an average salary of more than $100,000 (Figure 7).

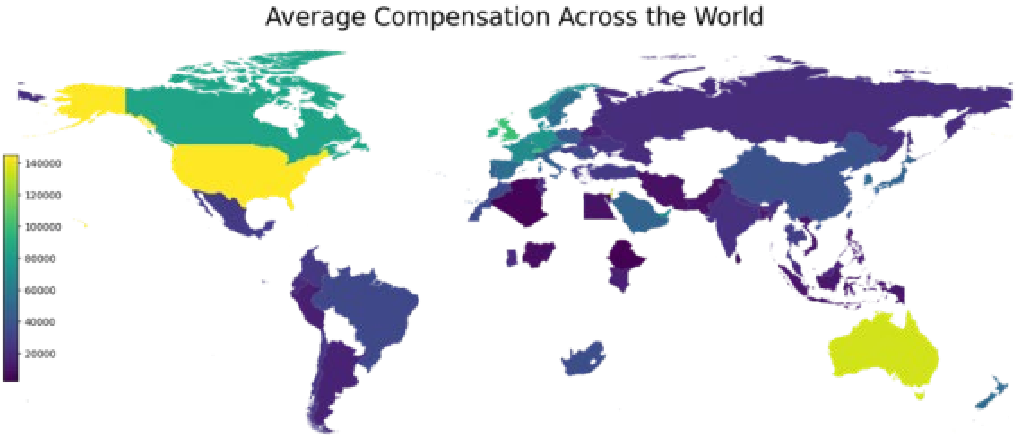

Figure 7: Geographical Heatmap of Data Science Salaries

If the ratio of the average salary of a data scientist to GNI per capita is used, the pattern seems to be reversed. Nepal, India, Morocco, and Ghana rank in the top 4 countries, with 24, 13,11, and 10 ratios. Interestingly, this means that while the US, Australia, and Israel are the highest paid, the quality of living in those three countries doesn't even break the top 4.

Though it may be possible that the respondents to the Kaggle survey can not represent the whole population, as India has over 8,792 respondents, we may have some confidence to claim the pattern may reflect the more urgent need for data scientists in underdeveloped countries.

F. Effects of education and occupation

Location isn't the only influencer on salary, nor is it the only data point captured in the Kaggle survey. Experience is a common criteria employers look for when bringing new talent to a company and is often quoted as a big influence in how much you can make. We can measure experience in several ways, but some primary indicators include education and job title. In Figure 8 below, we represent the salary distribution between different levels of experience using a tree diagram. This diagram showcases the educational background within different job titles, and how much each level of education is slated to earn you within the role.

The first thing to notice in Figure 8 is the different sizes of the containers. These varying sizes represent the number of respondents within each role and education level. Data Science and Data Analytics have the most representation in the survey, along with a very even spread of education levels. What this visual also shows us is that if an individual only has their bachelor's degree, they are likely stuck at being a data analyst or software engineer, whereas getting your masters opens you up to almost any role you would like.

For most occupations, the salary difference between levels of education is small, except in the cases of Managerial, Software Engineering, and Machine Learning roles. In these scenarios, a PHD degree is worth more. Outside of education, we can see that more specialized roles, such as MLE, Data Scientist, Software engineers, and Management do get paid more, but the difference isn't huge.

If the effects of education and occupation are compared between countries, it's interesting to find that occupation explains more difference in salary in underdeveloped countries, and little difference can be found between bachelor, master, and doctoral degree in these countries. This is perhaps partly due to the difference between the role of leader and follower in data science

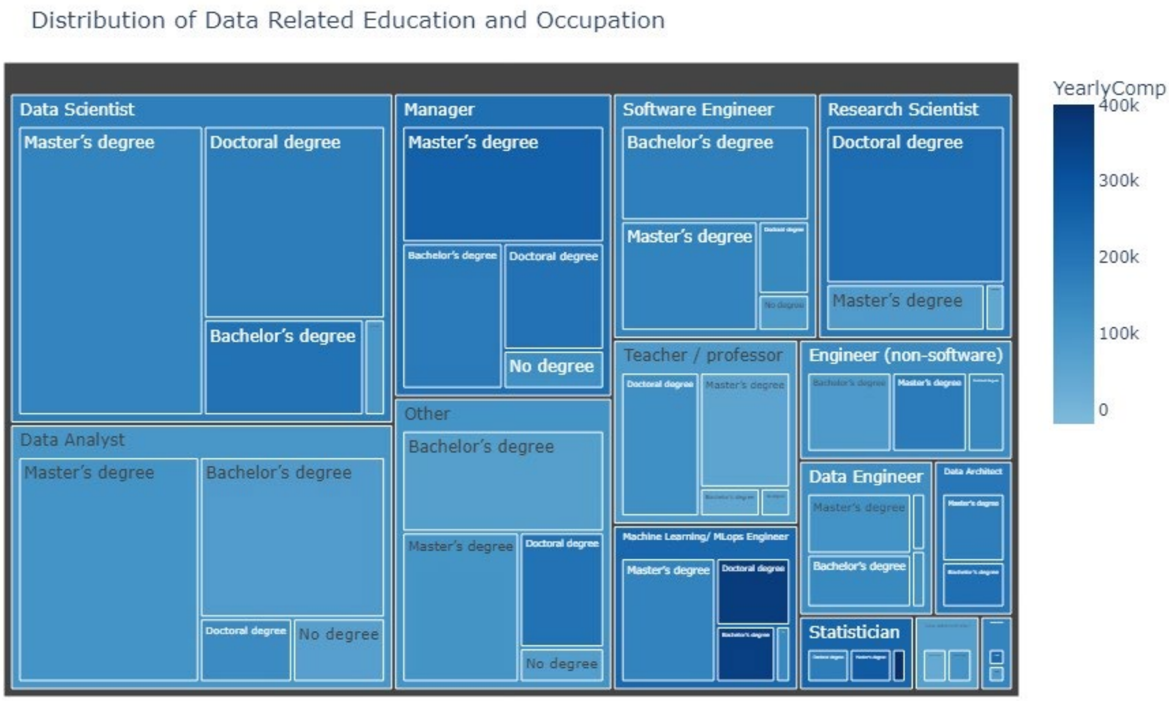

Figure 8: Distribution of Data Related Education and Occupation

G. Most used tools by professionals

Lastly, we wanted to see how respondents view which skills were necessary to succeed in the industry. To do this we calculated the frequency of respondents' choices in various Kaggle survey topics. From here, we ranked the choices by frequency and compared them in Figure 9. A key takeaway from this visual is that while there is a large variety of tools used in the industry, several categories are dominated by the top three options.

Another interesting thing is that though the survey was conducted during 2018-2022, classical algorithms still dominated the field. Linear regression or Logistic regression is the most popular machine learning algorithm, followed by Decision Trees or Random Forests, which was developed in 1995, according to the academic data set. For other specific fields, such as computer vision and Natural Language Processing, the top algorithms are mostly developed before 2016.

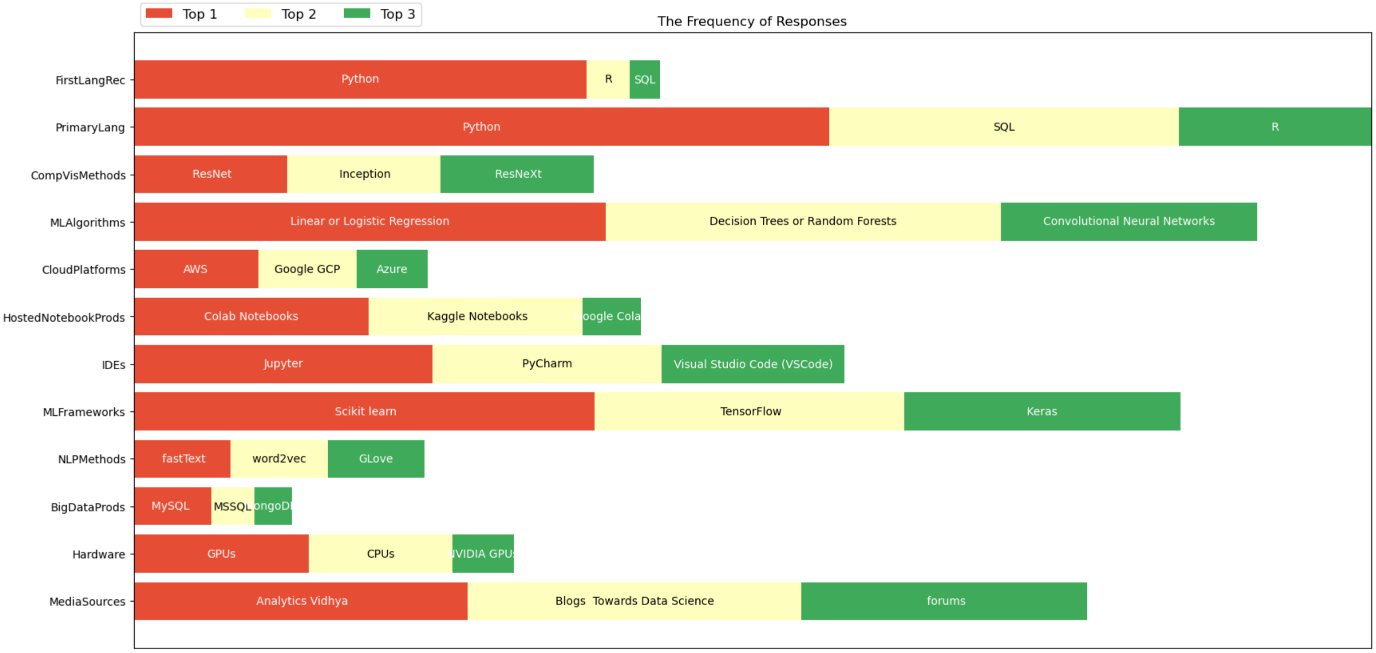

Figure 9: The Frequency of Responses

CONCLUSION

Throughout this report we have represented several facets of data science, in both academic and industrial settings. A common theme that we saw is that this field has become global over the last several years and is continuing to expand in areas where it already flourished. As the universities publishing research, and the topics of the publications themselves, become more diverse, the professional organizations in their respective countries begin utilizing data science techniques more and more. Being educated and experienced in these techniques opens up a world of rich research, and well-paid opportunities. While the level of education, role, and country of residence play a large role in yearly compensation, what most professionals agree is that proficiency in core techniques such as Python, or Linear Regression, and technologies such as Amazon Web Service or Scikit Learn is crucial to finding success in the field.

V. Statement of Work

Our team worked together to Analyze the Trends of Data Science through Academic and Industry data sets. Each member of the team has a unique work and academic background and we leveraged that to help ensure the success of this milestone project.

Pu Zeng was responsible for the data collection, clearing, manipulation, analysis, and the notebook of the academic data sets. He ensured that the data was properly visualized in the report to extract important trends and patterns and combined different libraries like BeautifulSoup, and Requests to collect extra data and achieve complete analysis. He also did some auxiliary work on the Kaggle data, including data exploration and visualization generation. Pu also helped set up, write, and edit the project report and maintain the GitHub and Colab notebooks.

Patrick Thornton was responsible for the processing and manipulation of the Kaggle data sets. During this process he applied data exploration, cleaning, and manipulation techniques both in Python and manually using Excel. The results of this work were the formatted and combined Kaggle data sets. Using these data sets, visuals were generated that were used in the analysis portion of this report. Cleaning and commenting on the industrial notebook were also part of this process. Outside of coding, Patrick also helped write and edit format the report.

Braedon Shick was responsible for the creation of the rough draft, exploration of initial data sets, and effectively communicating our data related findings from both data sets. He was responsible for the analysis in Google Scholar data set and visualizing Kaggle response statistics. His valuable experience in English Literature allowed for organized documents to display our project through all phases.

Overall, we communicated and collaborated throughout the last seven weeks by Slack or Google Meets to effectively meet the milestones required for each portion of the project. In future, we could improve our communication of the Milestone 1 Oral exam, as the misaligned Oral Exam preparation caused a significant decrease in progress on the report.

V. Reference

Emeryberger. (n.d.). Emeryberger/CSrankings: A web app for ranking computer science departments according to their research output in selective venues, and for finding active faculty across a wide range of areas. GitHub. https://github.com/emeryberger/CSrankings

Jatin. (2022, June 11). World University Rankings 2023. Kaggle. https://www.kaggle.com/datasets/jkanthony/world-university-rankings-202223

NOMINATIM documentation. (n.d.). https://nominatim.org/release-docs/develop/

Beautiful Soup documentation¶. Beautiful Soup Documentation. (n.d.). https://tedboy.github.io/bs4_doc/

Ai index report 2023 – artificial intelligence index. Stanford University | AI Index. (n.d.). https://aiindex.stanford.edu/report/

Google. (n.d.). Parameter, compute and data trends in machine learning. Google Sheets. https://docs.google.com/spreadsheets/d/1AAIebjNsnJj_uKALHbXNfn3_YsT6sHXtCU0q7OIPuc4/edit#gid=0

Yerz, V. (2022, November 21). Data Science in universities. Kaggle. https://www.kaggle.com/datasets/victoryerz/ms-data-science-universities

Hipo. (n.d.). HIPO/university-domains-list: University domains and Names Data List & API. GitHub. https://github.com/Hipo/university-domains-list

Marketing data science trends: Deloitte Us. Deloitte United States. (2020, August 7). https://www2.deloitte.com/us/en/pages/deloitte-analytics/articles/marketing-data-science-trends.html

Marr, B. (2022, November 1). The top 5 data science and analytics trends in 2023. Forbes. https://www.forbes.com/sites/bernardmarr/2022/10/31/the-top-5-data-science-and-analytics-trends-in-2023/

Kaggle. (2017, October 27). 2017 Kaggle Machine Learning & Data Science Survey. Kaggle. https://www.kaggle.com/datasets/kaggle/kaggle-survey-2017

Kaggle. (2018, November 3). 2018 Kaggle Machine Learning & Data Science Survey. Kaggle. https://www.kaggle.com/datasets/kaggle/kaggle-survey-2018

2019 Kaggle Machine Learning & Data Science Survey. (n.d.). https://www.kaggle.com/competitions/kaggle-survey-2019

2020 Kaggle Machine Learning & Data Science Survey. (n.d.). https://www.kaggle.com/competitions/kaggle-survey-2020

2021 Kaggle Machine Learning & Data Science Survey. (n.d.). https://www.kaggle.com/competitions/kaggle-survey-2021

2022 Kaggle Machine Learning & Data Science Survey. (n.d.). https://www.kaggle.com/competitions/kaggle-survey-2022