Sarah Wright

Data 231-00

Professor Frazier

Project Proposal : The Emerging Data Scientist

Data Science is listed as one of the sexiest careers of the 21st century. It takes a simple Google search to see the over saturated list of course recommendations and tutorials. In addition to that, companies and individuals take advantage of this emerging field by selling tools and platforms that are “hot” and will “optimize” the data science experience. I would like to see which platforms, languages, and other programs are good predictors of a persons longevity in the field. While the length of time a person works for will not say much about the satisfaction, this project assumes that the longer a person works in Data Science, the more experience they will have coding. Thus, this project will determine which platforms, languages, and other resources are good predictors for the longevity/retention of a Data Professional. Overall, I hope to answer the question: What should an aspiring Data Scientist student (a person who is learning DS) learn for a long career? In other words, What are the odds of “success” in data science for people who use different platforms, languages, and software ? Before diving into the data set, I would like to cite two articles.

This first article was written by Ahalt, Stan, and Kip Kelly. The article gives a depiction of the emerging Data Science industry and the gap between talented data scientist and the number of career openings for Data Science. This article begins by explaining how several no-tech and tech industries have recently expanded the field. Kelly and Stan give the example of Google, stating that “‌because it [Google] receives more than three billion search queries on a daily basis … Google was able to identify 45 search terms that, when used in a mathematical model, showed a strong correlation between their predictions and the CDC’s flu outbreak statistics,” (Ahalt and Kelly 1) it also mentions several other examples of Big Data in other fields such as in the companies: Target, Yahoo, Ebay, and others as way increase sales and customer loyality. Its seems that there is a very strong demand for Data Scientist in several fields. However, the article states that, “the United States could face a shortage of up to 190,000 workers with deep analytical skills and 1.5 million managers and analysts with the ability to use the big data analytics to make effective decisions.” (Ahalt and Kelly 10). Additionally, access to hard and software often limit the opportunities of many individuals. Fields like law, mathematics, and other disciplines are limited to the amount of tools people have access to for instance access to books, computers, and computing software, however this doesn’t seem to be the case with Data science. “It is cheaper than ever to purchase memory and storage, and good quality, open-source software is competing with commercial software, putting pressure on commercial software developers to keep their prices down “(Ahalt and Kelly 1). On top of that, “The rapid growth of big data has outpaced colleges’ and universities’ ability to develop and implement new curriculums. “(Ahalt and Kelly 8). Overall, Ahalt and Kelly seem to claim that there is a need for Data Scientist in the industry, however the people who can fill that roll are limited. That is why courses and company support could play a big role in the hiring of Data Scientist. Which begs the question, How do we increase the number of individuals to fill these roles? Are the people who fill these roles compensated well, or does a lot of money go into the training only for the field? How accessible is Data Science? These questions are further elaborated in the Article, however this is also the question that I would like to answer with the Data Set. However, many of these questions are answered in this next article.

For many, education can be an inhibitor in pursuing a career. According to Rawlings-Goss, institutions like Colleges and Boot camps feel this surge as well, stating that because of this gap,“ there is no standard definition of a Data Science curriculum” (Rawlings-Goss) and as a result, colleges, courses, and Boot Camps scramble to fill this void- for a fee. Additionally, while there are not many tools that a Data Scientist needs to purchase, there is a large time cost that comes with the numerous languages, tools, and skills that a Data Scientist needs to use to complete the work expected of them. Some of these skills that data scientist need to be familiar with are: “maximum likelihood estimators, statistical tests, distributions, calculus, and algebra” (Rawlings-Goss) along with computer language skills such as,“ Python, Machine Learning, Deep Learning, TensorFlow, Neural Networks, Computer Science, and Computational Linguistics” (Rawlings-Goss). While some of these skills are dependent on the industry, these online courses and classes, attempt to sell and teach as many as possible so that people entering the industry are well-equipped. Despite this, Kelly states, “On-the-job training and self-teaching may not be adequate in developing existing staff, particularly if they ‘don’t know what they don’t know.’” (Ahalt and Kelly 10). For many individuals, the process of getting into a new field can be daunting, and with Data Science it can be even more so, since it is an emerging field. In this second article by Gross-Ross, he explains the specifications of becoming a Data Science. It lists education paths, job positions and languages for what each type of Data Professional uses most frequently. While there is not much to cite from this article besides list and “How-to-guides”, I included it because it goes into details about several of the variables mentions in the Data Set, such as Cloud Computing and the differences between Data Scientist and Machine Learning Specialist. This leads us to the central question: What tools are actually used?

To answer these questions, I plan to run multiple logistic regression to compute the odds of having a long data career using the Years of Coding (the *Q6 variable) as the response variable and the following as the explanatory values: Gender, Countries, age, Formal Education, Current Role, Program languages used, Computing Platforms, Hardware, Industry, Size of company, Data Data Support, Year Compensation, Money spent computing at home, Course, and Public Sharing Platforms.*

The data set selected for this work comes from the Kaggle Machine Learning & Data science 2021 survey. In 2021, the Kaggle organization conducted a survey to gain insight on individuals in Machine Learning & Data science. The data set consists of 25,973 individual responses, 369 variables, and 38 questions (listed in kaggle\_survey\_2021\_answer\_choices.pdf ). To conduct this survey, Kaggle sent out: an anonymous survey to people opting into the Kaggle email list, they created promotional ads website banners asking people to fill out the survey on other Data & Computer Science websites, and posted to their twitter page. The individuals in this survey are limited to the people who saw the ads/promotions or who use coding websites and resources where the ad would have been displayed. Overall, this data is only representative of people who are within or connected to the Kaggle community. The people who filled out the survey are from 171 different territories and countries have varying experience levels and various demographics. Overall, I would say that data is diverse and representative of the Kaggle community. However, some responses are not collected in the dataset specifically from individuals who answered less than 15 questions, spent less than 5minutes completing the survey, or individuals identified as “Spam” or “Duplicates”.

Before any data analysis, the data needs to be clean and reorganized. Since the question of interests is about Data Science, the first step is to filter the variables of interest so that the ones left are only about Data Science and not machine learning nor software engineering. These questions are listed in the reference document included with the data (kaggle\_survey\_2021\_methodology.pdf) along with questions that are not related to the topic of interest. The remaining variables of interest thus far are as follows:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ***Variable name*** | ***Original definition*** | ***Units*** | ***Range or Levels*** | ***Possible re-coding*** | ***Rationale*** |
| Q6 | For how many years have you been writing code and/or programming? | Years | Never - 20+ | As is | Response variable of interest |
| Q2 | What is your gender? | N/A | Male / female/ Other | As is | Potential confounding variable (Some genders may not be represented) |
| Q3 | What country do you currently reside? | N/A | Over 50 different countries | As is | Potential confounding variable. (Location may influence access to jobs or training) |
| Q1 | What is your age in years? | Years | 18 - 70+ | As is | Potential Confounding variable (We expect people over a certain age to have job security) |
| Q4 | What is your highest level of formal education? | N/A | No formal education - professional doctorate, also “I prefer not to answer” | There are multiple responses for some individuals, I would like to group them:  No Formal Education = yes  Education = yes | Main explanatory variable of interest |
| Q5 | Select the title most similar to your current or (most recently held) role? | N/A | Business Analyst, Data Analyst, Data Engineer, Data Scientist | Filter by Data related careers and working individuals  <- | Helps to filter the data out, so we are only looking at Data Science Careers |
| Q7\_Part\_*1 - 12 - OTHER* | What programming languages do you use on a regular basis? | N/A | Python, R, SQL, C, C++, Java, JavaScript, Julia, Swift, Bash, MATLAB, None, Other | There are multiple responses for some individuals, I would like to list them as individual variables Python = Yes/ No  R = Yes / No  Etc. | Main explanatory variable of interest |
| Q9\_Part\_1 *- 12 - OTHER* | Which of the following integrated development environments (IDE's) do you use on a regular basis? | N/A | JupyterLab, RStudio, Visual Studio, VSCode, PyCharm, Spyder, Notepad++, Sublime Text, Vim/Emacs/Similar, MATLAB, Jupiter Notebook, None, Other | There are multiple responses for some individuals, I would like to list them as individual variables MATLAB = Yes /No  Spyder = Yes/No | Main explanatory variable of interest |
| Q10\_Part\_1 *- 16 - OTHER* | Which of the following hosted notebook products do you use on a regular basis? | N/A | Kaggle, Colab, Azure, Paperspace/Gradient, Binder/JupyterHub, Code Ocean, IBM Watson Studio, Amazon Sagemaker Studio, Amazon EMR, Google Cloud, Google Cloud Datalab, Databricks, Zeppelin/Zepl, Deepnote, Observable, None, Other | There are multiple responses for some individuals, I would like to list them as individual variables 1 Kaggle = Yes /No  Binder = Yes/No  Etc. | Main explanatory variable of interest |
| Q11 | What type of computing platform do you use the most often for your data science projects? | N/A | A laptop, A personal computer/desktop, A deep learning workstation, A cloud computing platform, None, Other | There are multiple responses for some individuals, I would like to list them as individual variables Laptop = Yes/ No  None = Yes / No  Etc. | Main explanatory variable of interest |
| Q12\_Part\_1 *- 5 - OTHER* | What type of specialized hardware do you use on a regular basis? | N/A | NVIDIA GPUs, Google Cloud TPUs, AWS Trainium Chips, AWS Inferentia Chips, None, Other | There are multiple responses for some individuals, I would like to list them as individual variables GPUs = Yes/ No  None = Yes / No  Etc. | Main explanatory variable of interest |
| Q14\_Part\_1 *- 11 - OTHER* | What data visualization libraries or tools do you use on a regular basis? | N/A | Matplotlib, seaboard, Plotly/Plotly Express, Ggplot / Ggplot2, Shiny, D3.js, Altair, Bokeh, Geoplotlib, Leaflet/Folium, None, Other | There are multiple responses for some individuals, I would like to list them as individual variables D3.js= Yes/ No  None = Yes / No  Etc. | Main explanatory variable of interest |
| Q25 | What is your current yearly compensation? | Dollars / USD | $0 - $1,000,000 | Need to segment into 5 groups | Potential alternative Response Variable (odds of increasing earning potential) |
| Q27\_A\_Part\_1 *- 11 - OTHER* | Which of the following cloud computing platforms do you use on a regular basis? | N/A | AWS, Microsoft Azure, GCP, IBM Cloud/RedHat , Oracle Cloud, SAP Cloud, Salesforce Cloud, VMware Cloud, Alibaba Cloud, Tencent Cloud, None, other | There are multiple responses for some individuals, I would like to list them as individual variables SAP Cloud = Yes/ No  Tencent = Yes / No  Etc. | Main explanatory variable of interest  (Maybe for ML more than Data Science) |
| Q32\_A\_Part\_1 *- 20 - OTHER* | Which of the following big data products (relational databases, data warehouses, data lakes, or similar) do you use on a regular basis? | N/A | MySQL, PostgreSQL, SQLite, Oracle DataBase, MongoDB, Snowflake, IBM Db2, Microsoft SQL Server, Microsoft Azure SQL Database, Microsoft Azure Cosmos DB, Amazon Redshift, Amazon Aurora, Amason RDS, Amazon DynamoDB, Google Cloud BigQuery, Google Cloud SQL, Google Cloud Firestore, Google Cloud Big Table, Google Cloud Spanner, None, Other | There are multiple responses for some individuals, I would like to list them as individual variables MySQL = Yes/ No  MongoDB = Yes / No  Etc. | Main explanatory variable of interest |
| Q34\_A\_Part\_1 *- 16 - OTHER* | Which of the following business intelligence tools do you use most often? | N/A | Amazon QuickSight, Microsoft Power BI, Google Data Studio, Looker, Tableau, Salesforce, Einstein Analytics, Qlik, Domo, TINCO Spitfire, Alteryx, Sisense, SAP Analytics Cloud, Microsoft Azure Synapse, Thoughtspot, None, Other | There are multiple responses for some individuals, I would like to list them as individual variables Looker = Yes/ No  Qlik= Yes / No  Etc. | Main explanatory variable of interest |
| **Q40\_Part\_1 *- 11 - OTHER*** | On which platforms have you begun or completed data science courses? | N/A | Coursera, edX, Kaggle, DataCamp, Fast.ai, Udacity, Udemy, LinkedIn Learning, Cloud-certification programs, University Courses, None, Other | There are multiple responses for some individuals, I would like to list them as individual variables Udemy = Yes/ No  Fast.ai = Yes / No  Etc. | Main explanatory variable of interest |
| Q41 | What is the the primary tool that you use at work or school to analyze data? | N/A | Basic statistical software, Advances statical software, Business Intellignence software, Local developmen environments, Cloud-based data software & API, other | There are multiple responses for some individuals, I would like to list them as individual variables API = Yes/ No  Other = Yes / No  Etc. | Main explanatory variable of interest |
| Q39\_Part\_1 *- 9 - OTHER* | Who/what are your favorite media sources that report on data science topics? | N/A | Twitter, Email, Reddit, Kaggle, Course Forums, Youtube, Podcasts, Blogs, Journal Publications, Slack communities, none, others | There are multiple responses for some individuals, I would like to list them as individual variables Twitter = Yes/ No  Slack = Yes / No  Etc. | Main explanatory variable of interest |

Work Cited

<https://link.springer.com/chapter/10.1007/978-3-030-22407-3_2>

Rawlings-Goss, Renata. “Building Data Careers.” *Data Science Careers, Training, and Hiring*, 2019, pp. 5–30, link.springer.com/chapter/10.1007/978-3-030-22407-3\_2, 10.1007/978-3-030-22407-3\_2. Accessed 27 Oct. 2021.‌

<https://renci.org/wp-content/uploads/2013/08/The-Big-Data-Talent-Gap-White-Paper1.pdf>

Ahalt, Stan, and Kip Kelly. *The Big Data Talent Gap*.