Data Science Career Paths & Skillsets in 2021

A glimpse of Kaggle's *State of Data Science and Machine Learning Survey*

Jingsong Gao, jg2109@georgetown.edu
Ercong Luo, el890@georgetown.edu
Rui Qiu, rq47@georgetown.edu

Introduction

Each year, Kaggle conducts an annual worldwide survey among active users to provide a comprehensive view of the data science (DS) and machine learning (ML) community. After preliminary cleaning, a total of 25,973 responses were recorded for analysis. Following the survey, Kaggle hosts an annual Data Science Survey Challenge that awards top notebook authors who tell a rich story about a subset of the DS and ML community.

For our project, we would like to tell a data science story about data science professionals working in the US. The questions we raise pertain to comparing and contrasting job titles that fall within the general umbrella of data science, such as data scientist, data analyst, machine learning engineer, statistician. The key features with which to tell the data science story include but are not limited to salary, preferences on programming languages and environments, hardware/cloud utilization, diversity, and inclusion. Not only are these interesting questions to ask, but they also provide insights for fellow students in our DSAN program to narrow down job search.

Dataset

This research draws data from the 2021 Kaggle Machine Learning & Data Science Survey conducted anonymously among the Kaggle community from 09/01/2021 to 10/04/2021. The dataset consists of 25,973 valid responses to the survey questionnaire encompassing 42 questions about personal background and preference on programming language, IDE, cloud service, computing hardware, database, machine learning package, workflow. Due to the nature of our data science story, a subset of data will be sliced as the data of interest.

Statistical Questions and Methods

Data science is a broad and interdisciplinary field. As a result, there is a lot of variation in the day-to-day responsibilities, workflows, skill sets, and salaries for jobs related to data science. We would like to do some key comparisons among the following job titles: data analyst, data engineer, data scientist, machine learning engineer, software engineer, statistician. Preliminarily we are interested in the following questions regarding career paths, with proposed analytical methods in parentheses:

(1) What percentage of the survey respondents are working under these job titles? (EDA)

(2) What are the statistics on salaries for these job titles? (EDA, ANOVA, hypothesis testing)
(3) What levels of education are required for these job titles? (EDA)
(4) Is there a significant income gap between genders for these jobs? (EDA, hypothesis testing)
(5) What is the typical skill set for these jobs? How does it affect the pay rate? (regression, ANOVA)
(6) Is there a certain correlation between industry and the need for these jobs? (EDA, bootstrap, hypothesis testing)

Additionally, we would like to investigate the requisite skillsets for each of the data science careers as guidance to our program cohort on which skills and technologies to focus the career preparation on. Our inquiry includes but is not limited to the following questions:

(1) What programming languages and IDEs do they use? (hypothesis test on the two-way table)
(2) Where do they get and share the knowledge? (EDA)
(3) What data science packages do they use? (EDA)
(4) Are they using cloud computing tools? What are their preferences? Will a user's preference for cloud computing platforms affect his or her preference for big data products/cloud storage products? For example, we want to know if an AWS EC2 dedicated user will actually prefer AWS S3 over other products. (Hypothesis testing)
(5) What is the overall AWS usage percentage among DS practitioners? Is it the same for Google Cloud? (Law of Large Numbers, two proportions Z-test.)

Expected Conclusions

Admittedly, Kaggle is a popular data science learning community with a focus on machine learning competition, its user group is representative enough to reflect a partial status quo of data science practitioners in the industry. In addition, one should keep in mind that not all data science professionals are keen on participating in machine learning competitions while working "nine to five".
Nevertheless, the survey still manages to give us a snapshot of what a data science practitioner's career looks like, especially in a post-pandemic era. Hopefully, it could provide some guidance to graduate students in the States who are pursuing data science-related careers.