How much is a Data Science job candidate worth?

By Laura Gascue

Introduction

The data science and machine learning (DS/ML) field has been growing exponentially during the past 10 years. New positions have been open in basically any industry, and Data Scientists are in large demand. But not all candidates and positions are the same. The field includes experienced analysts, programmers and researchers switching into this growing field, data scientists who started in the field from the beginning and have now significant DS/ML experience, and a large number of young people just coming out of school. At the same time, the DS/ML field is composed of a large variety of tools developed in different languages, with language specific packages used in different environments, allowing to process and analyze any data from any field. As a result, skills needed to succeed in this field can range significantly. As a reflection of this large variety of skills demanded we find a wide range of salaries being offered. Job titles, though highly correlated with earnings show a large overlap. So, if you are going to hire a new recruit, how much should you offer them? We can also ask, if you want to enter the Data Science job market, what skills or characteristics will help you land those high paying jobs?

The BLS reports an average salary for Data Scientists and Mathematical Science Occupations (SOC code 152098) in the US of $103, 930, and a median salary of $98,230 with an inter-quantile range of $32,140. Other related positions add to this variation with $86,200 median salary for Operations Research Analysts (SOC code 152031) at the bottom, and on the other end, $110,140 median average for Software Developers and Software Quality Assurance Analysts and Testers (SOC code 151256).

We can find market description analysis based on job posting analysis, using actual or estimated salary offers, and applying NLP tools to identify skills required for the positions. There are few data sources that describe current positions and salaries using small surveys, with obscure methodologies (see The Burtch Works Study, Salaries of Data Scientist & Predictive Analytics Professionals, August 2020). We use a relatively large survey specifically targeting the data science and machine learning community conducted by Kaggle, and aimed to get a comprehensive view of the state of the data science and machine learning field all around the world. The survey includes a question on yearly compensation, together with respondent's demographic characteristics, experience and skills.

Before we look at the data – Where is the data coming from?

Kaggle launches a Data Science and Machine Learning (DS & ML) Survey every year as a way to learn about the DS field, with questions on the user demographics (age, gender, education level, working status), their data science and machine learning knowledge and experience, and methods and tools they use or would like to get familiar with. The survey was distributed to the entire Kaggle community through the Kaggle (opted-in) email list, and promoted on the Kaggle website and Kaggle Twitter channel.

There are 20,036 answers to the survey in 2020 from 171 different countries, including active workers, unemployed individuals and students. The number of active workers in the survey is

13,213, and after dropping observations with missing income, the sample size is reduced to 10,729 individuals.

The source data is fairly clean and well organized, but there is a significant amount of missing values in some of the questions. In some cases, the missing values just reflect the questions' patterns. Not everyone is asked the same questions: students and unemployed individuals are not asked about their employer and people that don't write code are not asked about writing code.

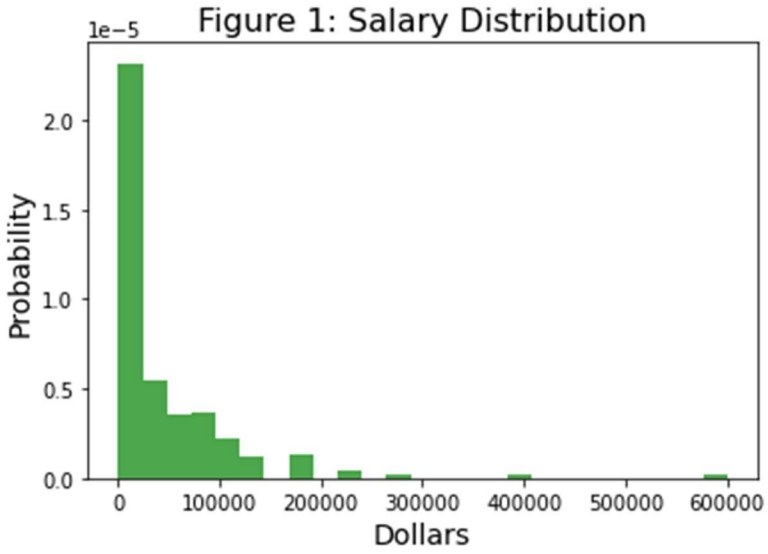
However, in some other cases, they may be actual missing values. Without knowing the nature of the missing values, any decision we make on how to handle them can't be tested. If we drop individuals with missing values, we will be probably generating a bias in the sample. A common approach would be to either impute the missing values or drop those individuals from the analysis implying that individuals that didn't answer some of the questions are similar to the rest of the individuals answering those questions. In this case, this doesn't seem to be the best approach. Many of the questions, and in particular the ones with larger numbers of missing values, ask about the use of specific methods and tools. Even though each question has an option for "Other" or "None", probably the people not answering are the ones that are not using those methods or tools. If this is the case, by dropping them, or using information from other individuals to guess, will generate a bias. In this analysis, all questions where the respondent is given a list to choose from and the respondent didn't answer, we assume the individual didn't use or know about any of the options and we impute the "None" option.

The compensation question is reported in ranges, so we use the range middle point to create a continuous measure. The question asks about "compensation", and we use compensation and salaries indistinctly, since both terms show cost of opportunity. In order to model salaries we want to consider five types of variables: demographic characteristics, geographic region, job position, experience, and DS/ML tools. Demographic variables considered are age level, education, and gender. To account for salary disparities around the globe, we created indicators for top 10 respondent countries (India, US, Brazil, Japan, Russia, UK, Germany, Nigeria, Spain, and Canada), and indicators for low income, medium-low income, medium-high income, and high income countries (data from the World Population Review). Variables considered to reflect experience are years of experience programming, years of experience using machine learning methods (ranges converted to continuous variables using range midpoint). We also consider the number of languages the person use to program, indicators for specific languages (Python, R, SQL, C, C++, Java, Javascript, Julia, Swift, Bash, Matlab), the use of visualization methods, cloud computing platform, deep learning workstation or a personal computer/laptop, the use of advance statistical software (SPSS, SAS, etc.), basic statistical software (Excel, Google Sheets, etc), business intelligence software, or local development environments. But before diving into modeling salaries we will explore their distribution in the sample.

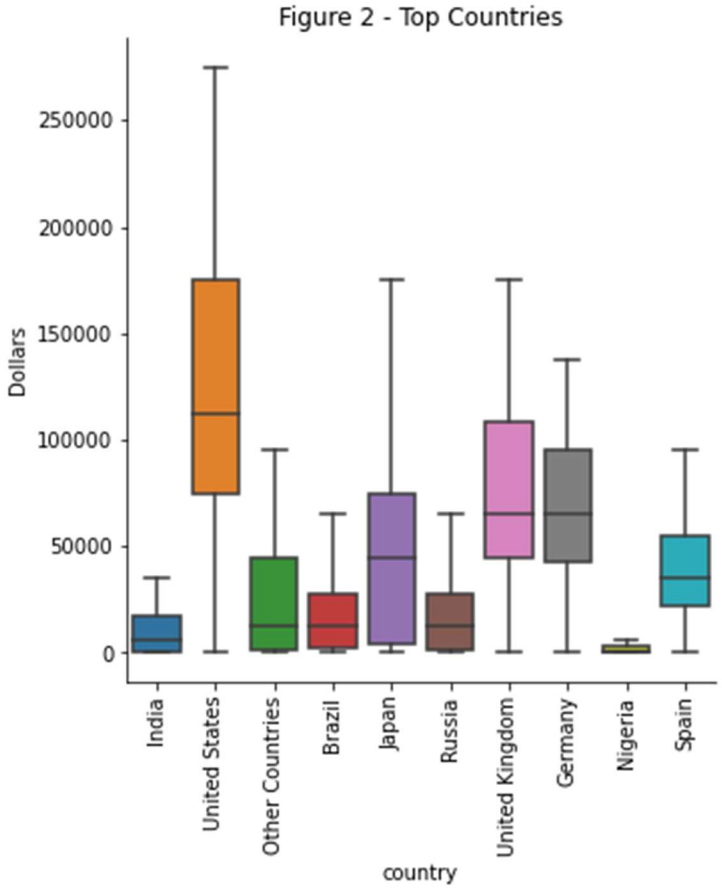
In this first report we limit the analysis to data exploration. We build all the necessary features and look at frequencies and distributions for different subpopulations.

Salary Distribution

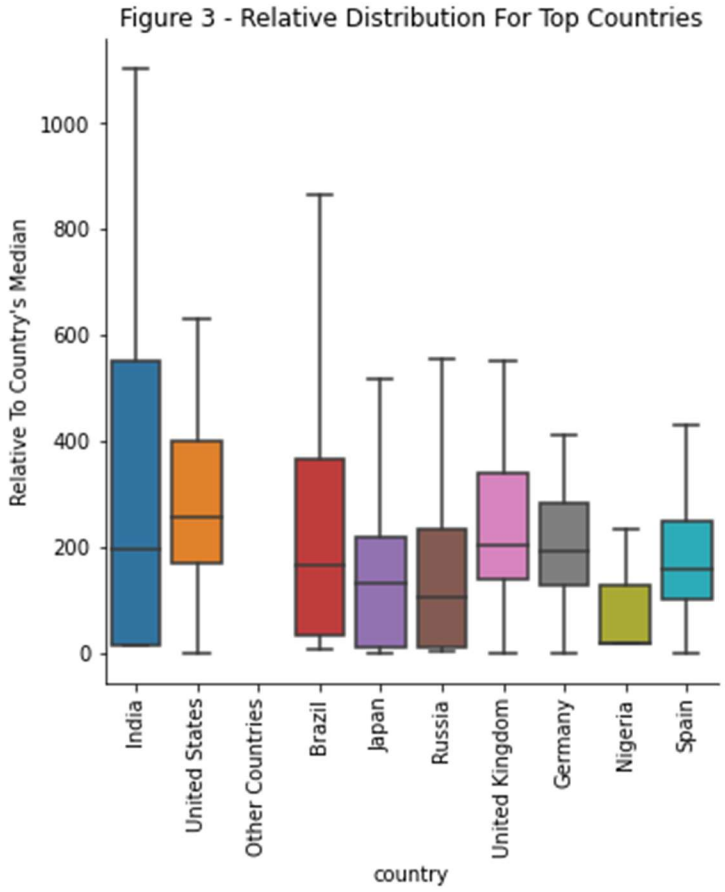
We can use the Kaggle survey to assess salaries paid to individuals in the DS/ML community. A first look at salary distribution shows large variations with a high concentration on the low end of the range, with very low compensation values. About 41% of the sample earns less than $10,000, while about 9% of the sample earns more than $100,000.



However, this plot hides large income inequalities across countries. It is now very easy to work from any part of the world for anybody else in the world. However, there are still regulations that make the substitution of workers with foreign residents harder. Figure 2 shows boxplots for salary distribution for the top ten respondent countries. Salaries in the US are much higher than in the rest of the countries, including other rich countries like the United Kingdom and Germany. US salaries also present larger variability.



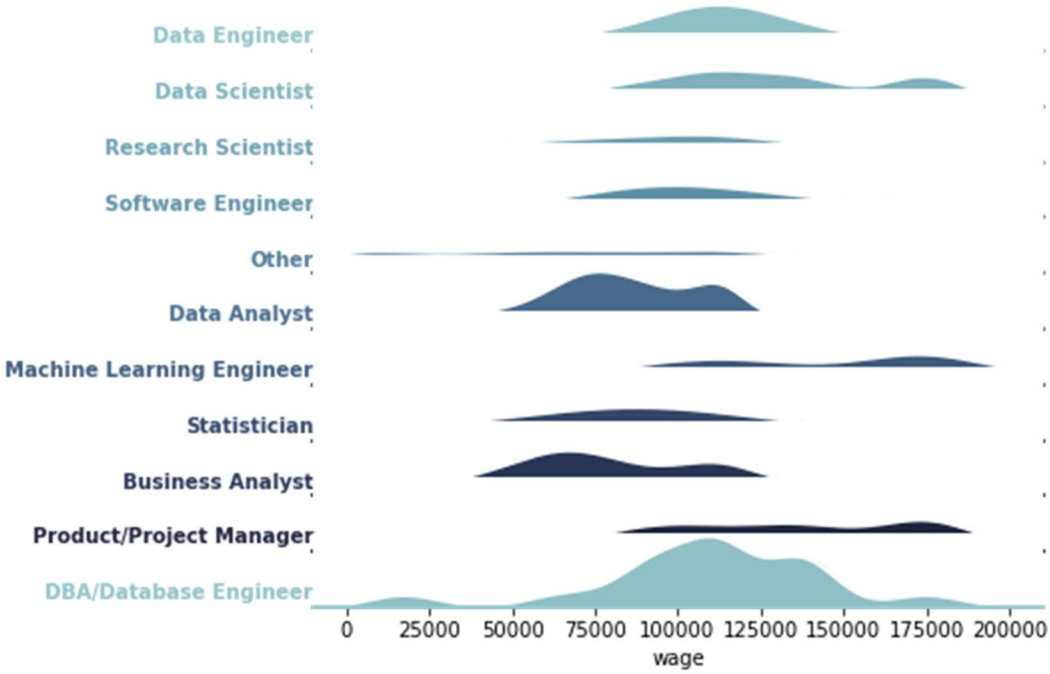
A more insightful approach when comparing salaries in the data science community across countries, is to look at wages relative to countries median income. Figure 3 shows how salaries for that with the exception of Nigeria, median salaries for the DS/ML community are above countries' median income levels, with United States, United Kingdom and Germany being more than double their median income.



The two countries with the most responses by far are India with 2,353 responses and the US with 1,484 responses. These are also the two countries on the extreme of the compensation levels, with the US having an average salary of $121,130, six times larger than India's $18,503 average salary, and a median salary of $112,500, eighteen times larger than India's $6,250 median salary.

Salary distribution for a given job title, even within a given country, has a large variation. Different positions can be associated with higher or lower salaries, but there is still large overlap between job titles. Look at figure 4, with salary distribution in the US. Database administrators and engineer compensation expands over a wide range, going from a low mode below the bulk of all the other job positions, to a high mode on top of all other positions. The three other titles that present a mode on the higher end of the distribution (around $175K) are Data Scientist, Machine Learning Engineer, and Product or Project Manager. Job titles with most of their compensation distribution being on the lower end of the range are Data Analysts, Business Analysts and Statisticians.

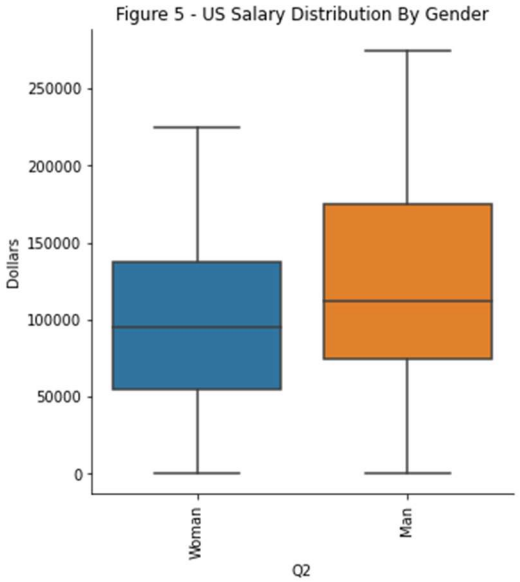
Figure 4 US Salary Distribution By Job Title



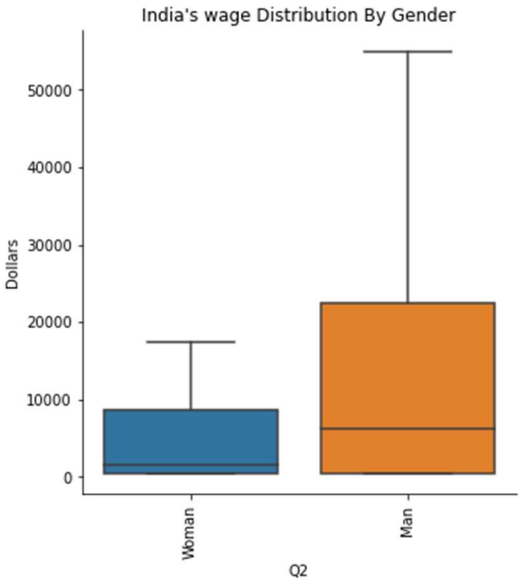
From the distribution chart we can also observe that the most frequent position observed is the DBA/Database Engineer, followed by Data Analysts, and the multiple modes observed are probably related to the country of residence.

Another common source of salary differences is gender. Gender disparity is clearly present in the DS/ML community. The salary gap between men and women is strong in this field, with most survey participants being males with median male compensation of $22,500, more than three times larger than women, and average compensation of $46,771 against $33,340 average compensation for women.

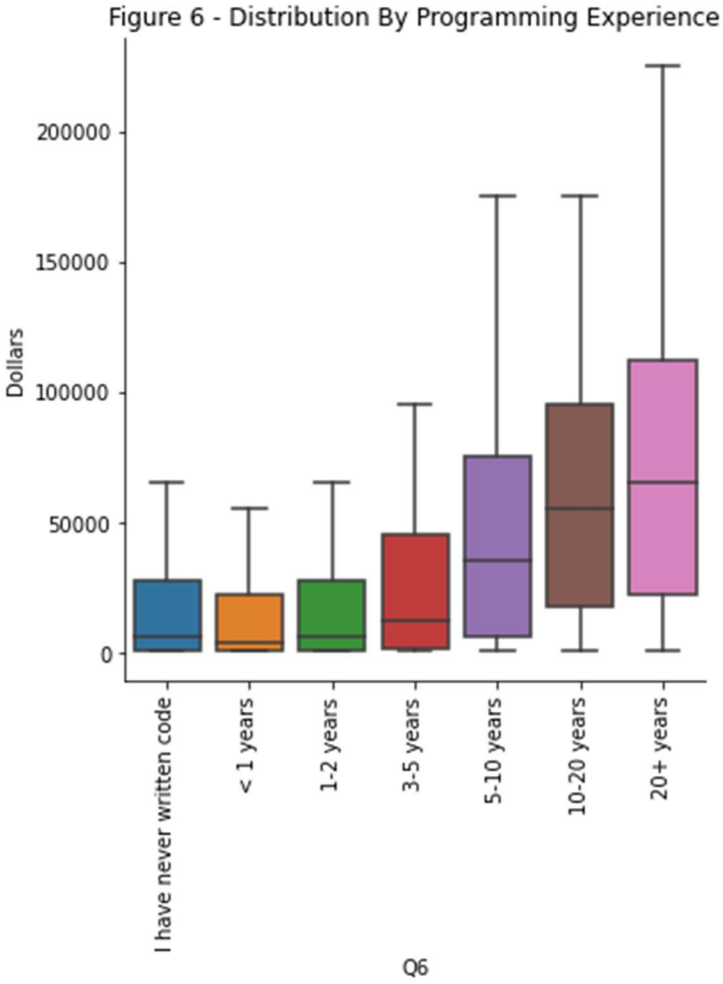
Figure 5 - US Salary Distribution By Gender



India's wage Distribution By Gender



The gap is also present when we look at the two examples, India and the US, being larger in India than in the US. Still, US men salaries in the DS/ML community are on average 32% higher than those for women, with the top quartile of men reaching salaries around $175K or above, while women's top quartile earnings are around $137.5K or above.



Finally, we will expect that more experienced programmers will earn higher salaries. This is true for our sample and can be observed in figure 6, but the boxplots also show an increasing variability as experience grows. These large differences between years of experience are also present when we look at just the US. An individual with just one or two years of experience earns in the US a median salary of $75,000, while somebody with 10 or more years of experience earns a median salary of $137,500, or on average $73,430 more.

Conclusion

When trying to predict salaries in the DS/ML field we should consider the obvious characteristics: country of residence, gender, experience, and position title. However, these aspects are not enough

to explain the high variability in salaries observed. Other data present in the survey, looking at individual’s skills and tools used could be relevant to explain this variation.