A Look at Data Scientists

Ben Merrill

Abstract

The Data Science community is growing and changing constantly. In 2019, Kaggle took a survey of Data Scientists from around the world in more than 170 countries in hopes to measure this emerging industry. This survey data begs the question: 'What are the biggest differences in Data Scientists from country to country?'

In this project, we us Python packages to explore Kaggle's survey data and make some general conclusions about the field of Data Science. We find that there are many young coders, salary greatly differs by country, and many Data Scientists are in school.

Motivation

Data Science has only been a popular way to make business decisions for a few years. It is a new and high-paying field of work that is needed in tech, finance, health, and many other industries. As it emerges, employers hoping to hire data scientists, students are beginning their studies of data, and many others may be interested in measuring this new field of work.

Dataset(s)

The dataset for this project is Kaggle's third annual Machine Learning and Data Science Survey. It is a survey of nearly 20,000 Data Scientists around the world in 2019. The data can be found and downloaded here: https://www.kaggle.com/c/kaggle-survey-2019/data. It contains categorical and written question responses on topics like Machine Learning, age, sex, coding experience, salary, education, and much more.

Data Preparation and Cleaning

This data was generally clean, as it had been edited by the data team at Kaggle.com. Even so, there are still many limitations and editing that needed to be done. Many of the responders only answered a handful of the questions that were posed, because there were differing backgrounds of responders, some with much industry experience and some with none. This led to a differing number of 'NaN' values in each column of the dataset.

I also cleaned the dataset to format and select only the questions that had a large number of responders, was relevant to general conclusions, and didn't have 'write in' and 'select all that apply' response options.

Research Question

What are the biggest differences in Data Scientists from the 7 most surveyed countries?

Research Questions Note

In the notebook, you will find two research questions. The second is, ‘Can we use factors and Machine Learning to accurately predict employee salary?’ The short answer (for now) is no.

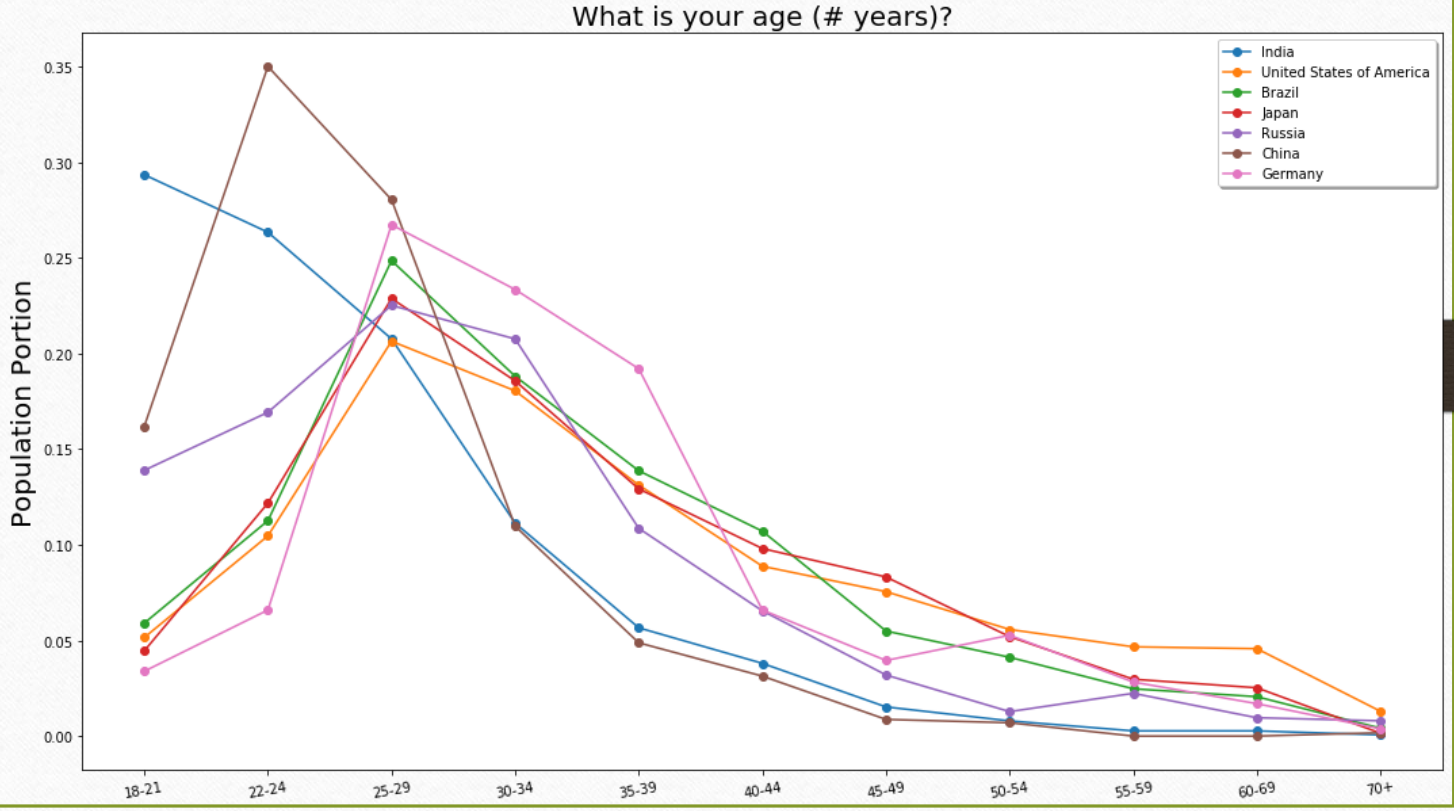
You will find in my Jupyter Notebook that I attempted to respond to this question, but we have yet to learn a Machine Learning methods to make decision tree, cluster, or linear regression based on categorical data. We can use numerical data to predict categorical data, but we still must learn how to train a model using categorical data. It will be fun to return to this question after having taken the 'Machine Learning’ Edx course through UCSD.

Methods

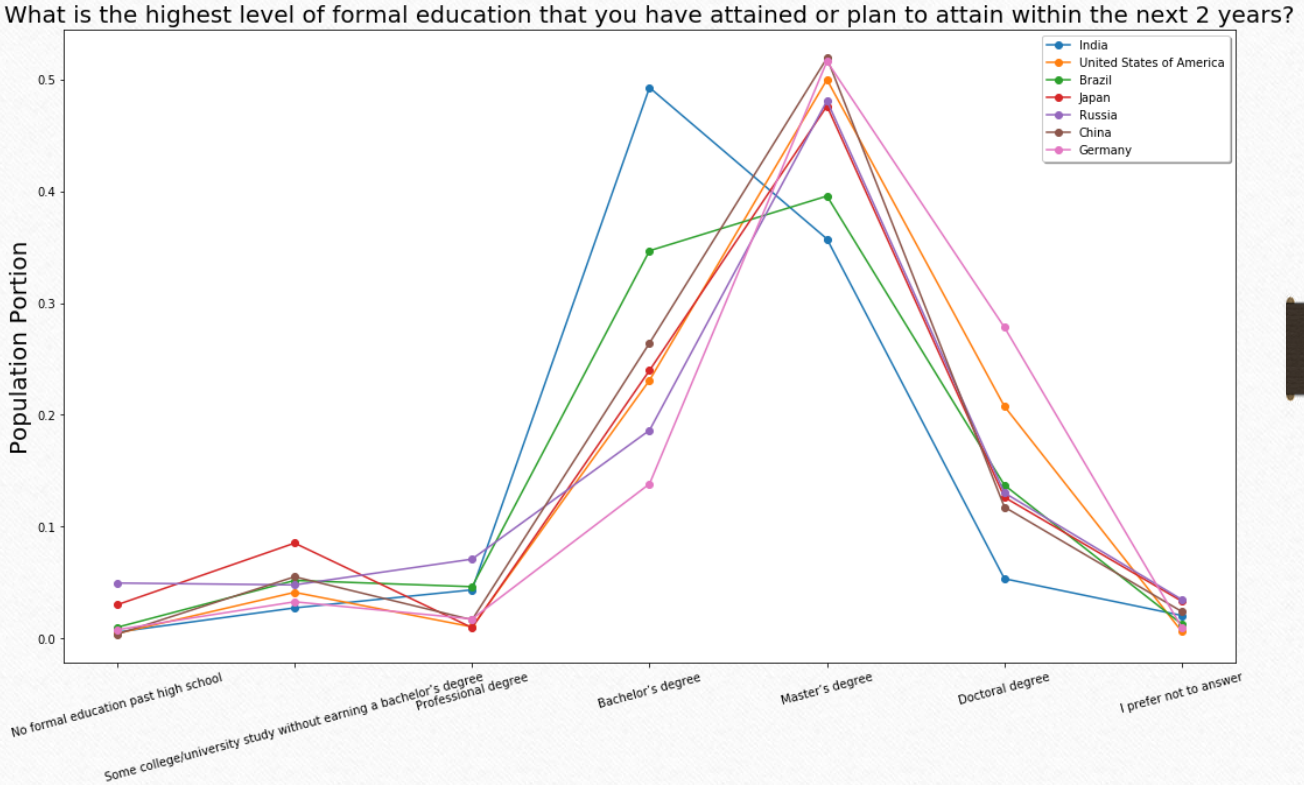
Using the filtered and cleaned data, we selected the 7 most surveyed countries. We then created some functions to return a columns data sorted by county and category. Due to differing sample sizes, the best way to measure a country's presence was by measuring the proportion of the population of that country in each category, i.e. 29% of India's surveyed were in the ages of 18-21.

From there, the goal was to display the proportional data simply, elegantly, and accurately. Because there were too many categories, there are 5 selected for this presentation: Age, Education, Salary, Current Role, and Coding Experience.

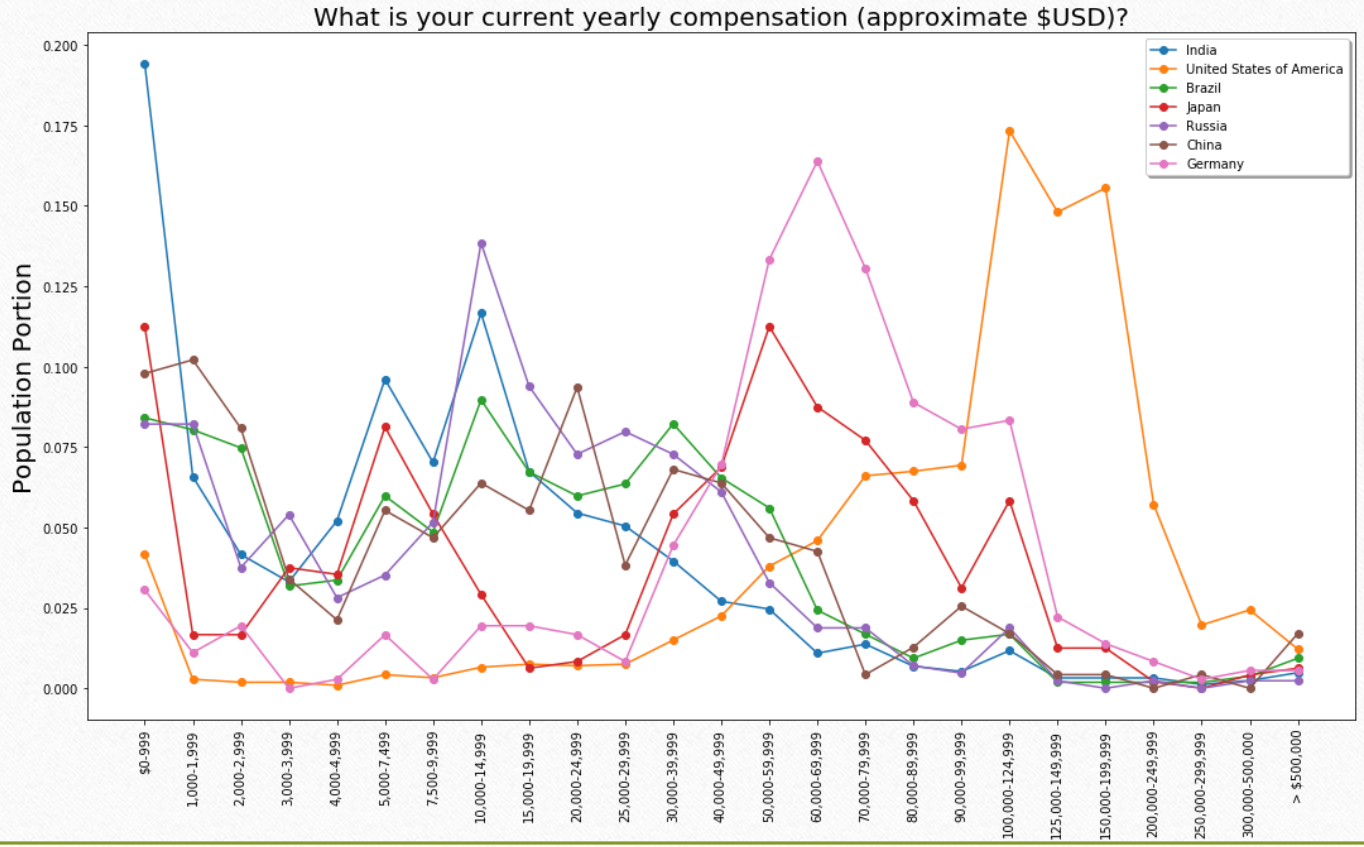
Age


What is your age (# years)?
Population Portion
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
18-21
22-24
25-29
30-34
35-39
40-44
45-49
50-54
55-59
60-69
70+
India
United States of America
Brazil
Japan
Russia
China
Germany

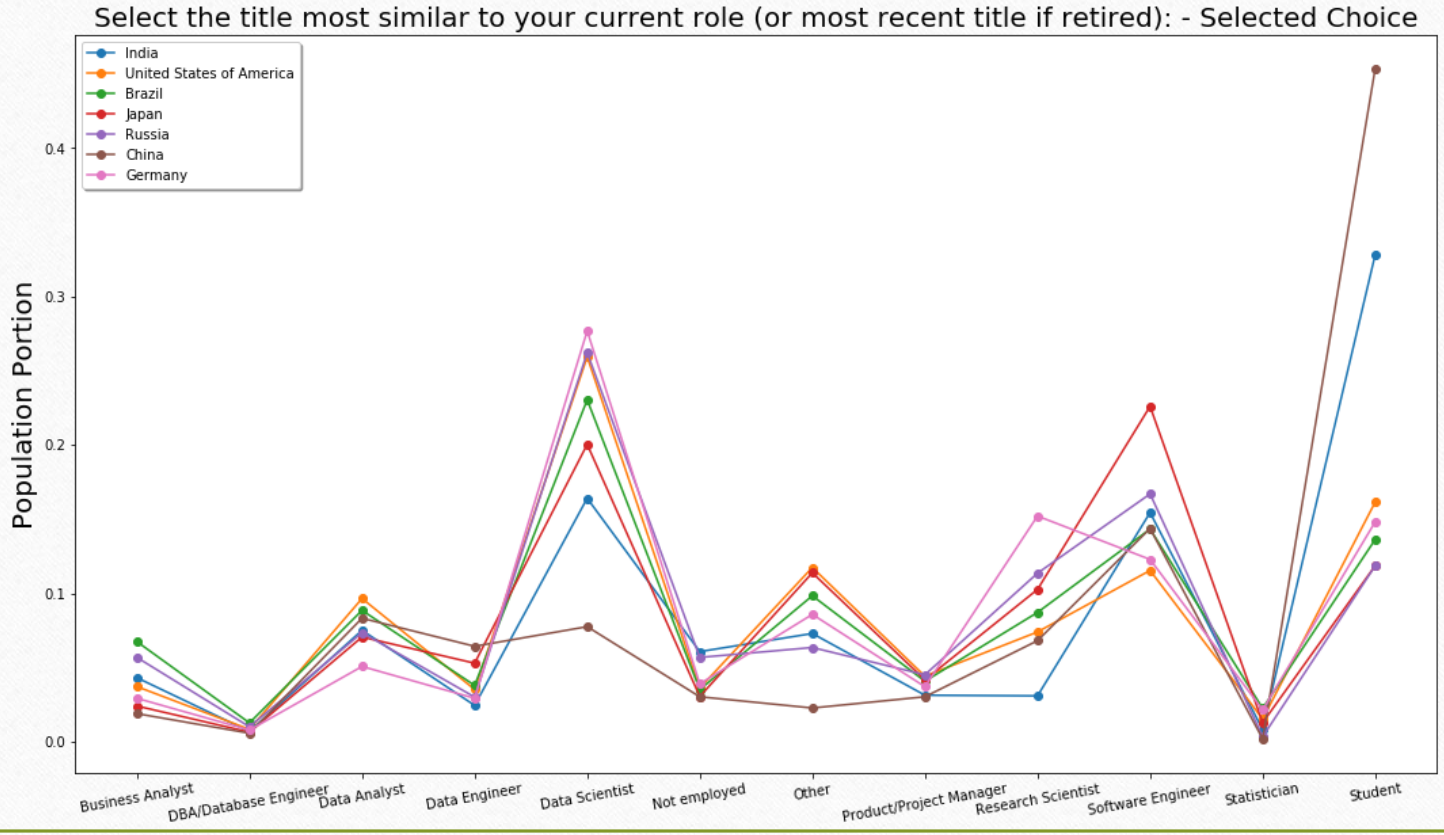
Education


What is the highest level of formal education that you have attained or plan to attain within the next 2 years?
Population Portion
0.0
0.1
0.2
0.3
0.4
0.5
India
United States of America
Brazil
Japan
Russia
China
Germany
No formal education past high school
Some college/university study without earning a bachelor's degree
Professional degree
Bachelor's degree
Master's degree
Doctoral degree
I prefer not to answer

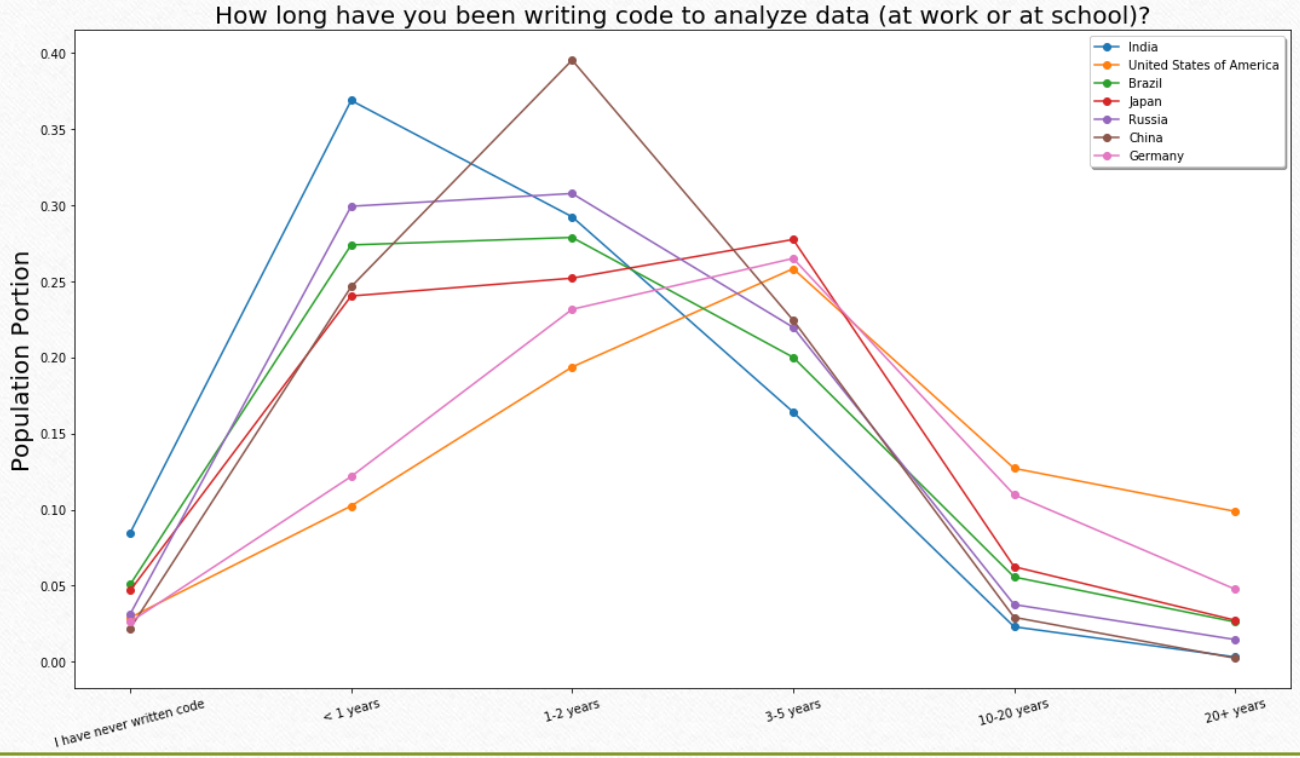
Salary


What is your current yearly compensation (approximate $USD)?
Population Portion
0.200
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
India
United States of America
Brazil
Japan
Russia
China
Germany
$0-999
1,000-1,999
2,000-2,999
3,000-3,999
4,000-4,999
5,000-7,499
7,500-9,999
10,000-14,999
15,000-19,999
20,000-24,999
25,000-29,999
30,000-39,999
40,000-49,999
50,000-59,999
60,000-69,999
70,000-79,999
80,000-89,999
90,000-99,999
100,000-124,999
125,000-149,999
150,000-199,999
200,000-249,999
250,000-299,999
300,000-500,000
> $500,000

Title


Select the title most similar to your current role (or most recent title if retired): - Selected Choice
India
United States of America
Brazil
Japan
Russia
China
Germany
Population Portion
0.0
0.1
0.2
0.3
0.4
Business Analyst
DBA/Database Engineer
Data Analyst
Data Engineer
Data Scientist
Not employed
Other
Product/Project Manager
Research Scientist
Software Engineer
Statistician
Student

Coding Experience



Findings

To address our question of biggest differences from country to country, we can generalize this from the data:

- India and China have younger populations, with a large proportion as students
- Japan, Germany, and US have the older populations, with high age, salary, and experience
- The majority of individuals have been coding less than 3 years
- Many data scientists are software engineers, research scientists, and students
- Most people working in the field have a masters degree

Limitations

The largest limitation has to do with the sampling and the 'NaN' values. The only respondents were people who use Kaggle. It's possible that some data scientists prefer to use other sites to get their data.

Not every question was shown to every respondent since some respondents didn't have the experience to answer a question. This means that for each category there was a different number of respondents. These differing respondents make it more difficult to compare our graphs to one another.

For example, a student may not answer the question about salary because they may not be working. You can see this clearly by a larger percentage of people identifying as 'student' in the Title slide than people with salaries between '$0-999'.

Conclusions

Although our research question is generally broad, we can conclude that populations in various countries differ greatly from one another. It seems that as the fields of Machine Learning and Data Science are beginning to grow, new, young coders are rising up to fill the need. Also, countries with older populations and more experience in the fields tend to have higher salaries.

It will be exciting to see how the field of Data Science will continue to grow and change over the coming years.

Acknowledgements

I did this project on my own. The data was found thanks to Kaggle at the site

https://www.kaggle.com/c/kaggle-survey-2019/overview

References

‘2019 Kaggle ML & DS Survey’. Kaggle Inc., October 2019. 17 November 2019, https://www.kaggle.com/c/kaggle-survey-2019/overview.