The Current State of Data Science: Tools of the Trade, Compensation and Representation

2024-12-08

by Rosey Gutierrez and Jill Nomura

I. Introduction

As Data Science & Machine Learning evolves as one of the most dynamic and impactful domains in technology, we want to analyze trends and patterns using the 2020 Kaggle DS & ML Survey, which gathered responses from over 20,000 participants across 171 countries and territories. The survey provides information capturing the experiences, aspirations, and tools of individuals engaged in or aspiring to work within the data science ecosystem. The survey methodology and flow logic, as well as a full breakdown of the questions (which can be used to asses missingness mechanisms as some questions were only asked to more experienced participants) can be found in the `supplementary_data` folder.

Problem Description

Our analysis focuses on answering key questions about the demographics, career choices, tools, and compensation of respondents, providing a comprehensive view of the current state of the field.

Specifically, we aim to:

1. Explore the **distribution of age and gender across different roles** to understand representation in roles such as Data Scientist, Machine Learning Engineer, and Research Scientist.
2. Identify the **top careers among respondents** and examine how these vary based on geographic region.
3. Analyze the **most commonly used tools, programming languages, and products among data scientists** and compare these with the skills and tools that other respondents in the data most commonly use.
4. Investigate the **relationship between educational background and job compensation**, and assess whether salary can be predicted solely based on one's level of education.

We wish to assess the **demographic distribution of respondents' gender, and country of residence**, and critically evaluate whether the data science field is representative of the broader population.

II. Analysis

Questions

1. What is the distribution of age and gender across different roles (e.g., Data Scientist, ML Engineer, Research Scientist, etc.)

Methods To analyze the distribution of age and gender across different roles, we will begin by filtering the data to include only relevant columns: age, gender, and role. We will remove rows with missing or invalid entries in these columns to ensure the integrity of the analysis. Specifically, we will identify and exclude invalid entries in the role column, such as blank or whitespace-only values.

Next, we will map the age data into pre-defined age groups (e.g., "18-21," "22-24") to enable categorical analysis. Any rows with unmatched or ambiguous age values will also be excluded.

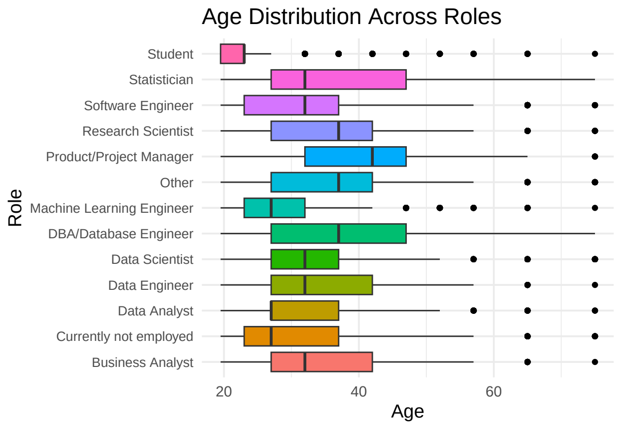
After cleaning the data, we will group the responses by role and calculate the total count for each gender and age group within each role. For gender, we will compute the proportion of each gender within each role and identify the gender with the highest proportion. Similarly, for age groups, we will calculate the proportion of responses for each age group within each role and determine the age group with the highest proportion.

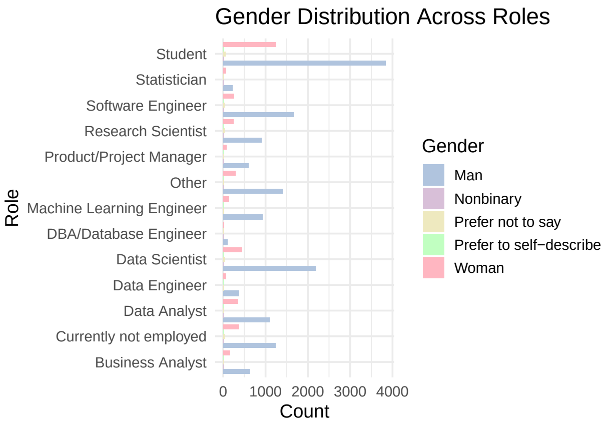
We will visualize the distributions for age and gender across roles for a better understanding of the differences.

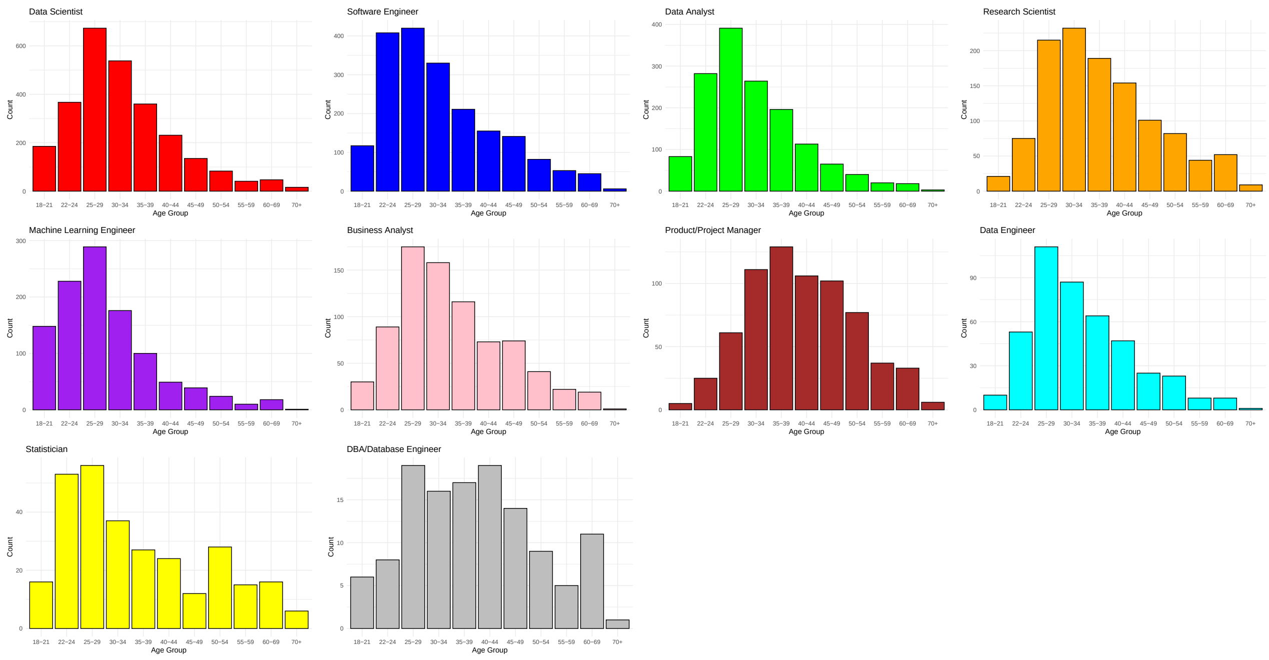
Analysis

Table 1: Distribution of Age and Gender among Industry Professionals

Role	Gender	g_prop	Age	a_prop
Business Analyst	Man	0.7857143	25-29	0.2192982
Currently not employed	Man	0.7475787	25-29	0.2615012
DBA/Database Engineer	Man	0.8240000	25-29	0.1520000
DBA/Database Engineer	Man	0.8240000	40-44	0.1520000
Data Analyst	Man	0.7525424	25-29	0.2650847
Data Engineer	Man	0.8443936	25-29	0.2540046
Data Scientist	Man	0.8198804	25-29	0.2514948
Machine Learning Engineer	Man	0.8576710	25-29	0.2670980
Other	Man	0.8163500	25-29	0.1899827
Product/Project Manager	Man	0.8713873	35-39	0.1864162
Research Scientist	Man	0.7657581	30-34	0.1976150
Software Engineer	Man	0.8485772	25-29	0.2134146
Statistician	Man	0.7551724	25-29	0.1931034
Student	Man	0.7431831	18-21	0.4898472







Conclusion Across every single role, men represent an overwhelming majority often comprising of 75% or more of the respondents in most occupations. Roles such as ML Engineer (85.8%) and Product Manager (87.1%) exhibit a particularly high male ratio. This trend is consistent across most roles, with the highest proportion of women observed in roles such as Statistician and Research Scientist, but even those proportions are below 25%.

In terms of age distribution, the 25-29 age group consistently represents the largest proportion among most roles. Exceptions to this trend include Product/Project Managers, and Research Scientists, where the largest proportions shift to slightly older age groups such as 35–39 and 30–34, respectively. Database Engineers are tied for the age groups of 25-29 and 40-44 making it a more spread-out profession among age groups. Students are also very obviously overwhelmingly in the lowest 18-21 age-group. Additionally, looking at the distribution of age across each role with its frequency yield interesting findings. These histograms are ordered by role, starting with the one with the highest count of respondents and ending with the one with the lowest. According to the data, “Data Scientist” is the most common role, with 2,676 professionals represented in the survey, making it the red histogram. Conversely, “DBA/Database Engineer” is the least common role, with only 125 respondents, and is displayed as the gray histogram. Examining the histograms reveals a notable trend: a progression from right-skewed distributions to approximately normal distributions. A right-skewed distribution indicates that a role is in high demand, attracting a younger workforce eager to enter the field.

In contrast, a more normal distribution suggests that the role is less in demand, with fewer new entrants required to meet the labor needs. This shift in distribution type serves as an indicator of current demand trends across roles. Aspiring students interested in data-related careers might benefit from focusing on roles with highly skewed distributions, as these are indicative of growing demand. Higher skewness magnitudes denote heavier skew. According to the analysis, the role with the highest skewness in age distribution is "Machine Learning Engineer," with a skewness value of 1.56.

The analysis shows a very clear gender imbalance in the data science and machine learning fields, with men significantly outnumbering women and other genders in every role in the data. Moreover, the concentration of respondents in the 25–29 age group suggests that the field might be primarily populated by early-career professionals.

2. What are the top careers among respondents, and how do they vary across region?

Methods To analyze the distribution of career roles across regions and identify the most popular roles, we will take the Q5 and Q3 columns and rename them `Role` and `Region` respectively and filter out any missing values. We will then group the data by `Region` and count the number of responses from each country, so that we can choose the top 20 in order to create a reasonable visualization. The 'Other' category will need to be excluded to focus on named regions.

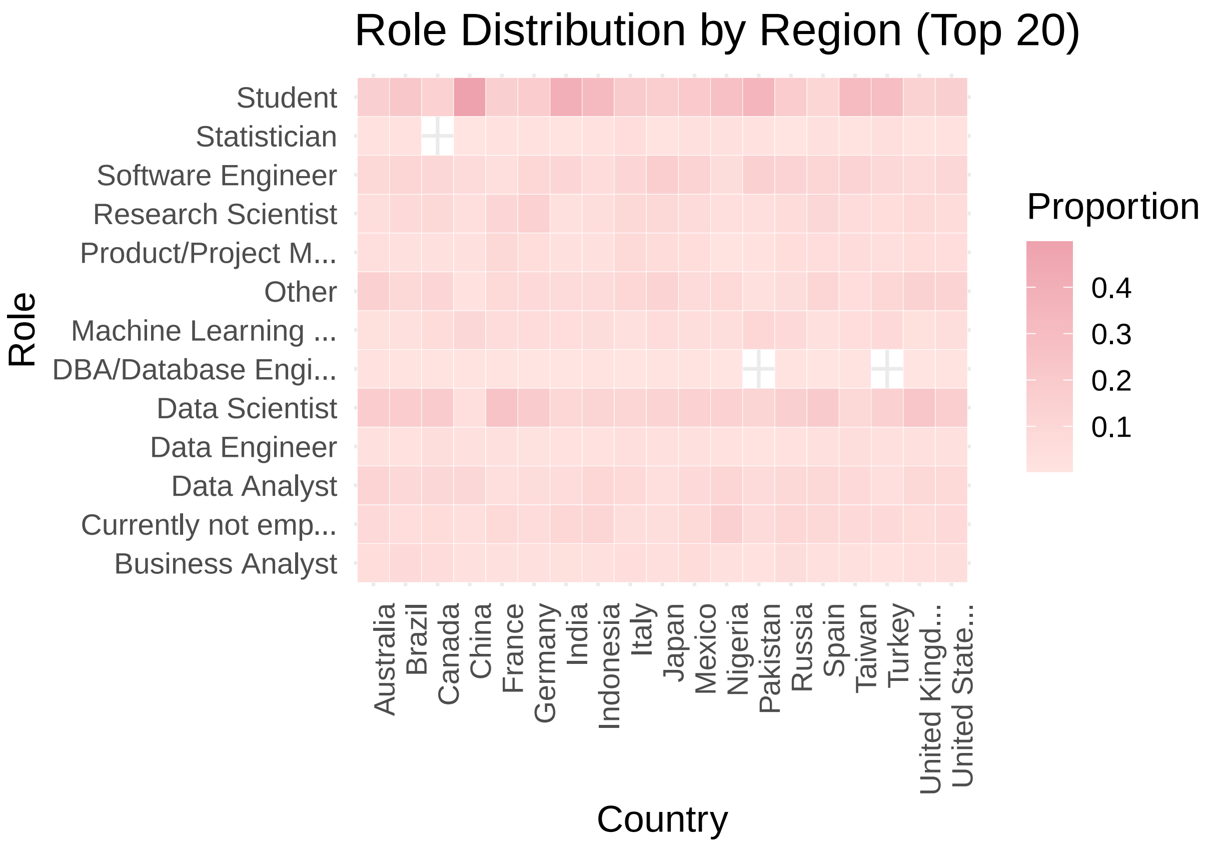
For each region among the top 20 by respondents, we will group the data by `Role` and calculate the proportion of responses for each role relative to the total number of responses in that region.

We will use a heatmap to compare the distribution of roles among regions, and identify the most popular role by region using a table.

Analysis .

Table 2: Most Common Role among Respondents by Region

Region	Role	Proportion
Australia	Data Scientist	0.1785714
Brazil	Student	0.2165414
Canada	Data Scientist	0.1896552
China	Student	0.4976190
France	Data Scientist	0.2492114
Germany	Data Scientist	0.1917098
India	Student	0.3974473
Indonesia	Student	0.3212996
Italy	Student	0.1886792
Japan	Student	0.1755594
Mexico	Student	0.2062780
Nigeria	Student	0.2727273
Pakistan	Student	0.3492647
Russia	Student	0.1826401
Spain	Data Scientist	0.1963190
Taiwan	Student	0.3095238
Turkey	Student	0.2953846
United Kingdom of Great Britain and Northern Ireland	Data Scientist	0.2333333
United States of America	Data Scientist	0.1768986



Conclusion We can observe from the analysis that “Student” and “Data Scientist” are the two most prominent roles globally. Students are particularly dominant in regions such as China, India, Pakistan, and Indonesia, where it accounts for a substantial proportion of responses, with the highest proportion being in China (49.76%). This likely reflects the growing interest in data science and machine learning among students in these regions, quite possibly fueled by the increasing availability of educational resources and career opportunities.

The “Data Scientist” role is more prominent in countries like Australia, Canada, France, Germany, UK, and the United States, where it consistently emerges as the most popular role. For example, in Canada, France, and Germany, the “Data Scientist” role accounts for approximately 19–25% of the responses. It might be reasonable to infer that that these regions have a more established data science industry, with professionals actively pursuing careers as data scientists in these countries. This is further evident from the heatmap also illustrating the diversity of roles across countries, with smaller proportions distributed across other roles, such as “Software Engineer”, “Machine Learning Engineer”, and “Research Scientist”, particularly in regions with established technology industries like Germany, United States, and United Kingdom.

The prevalence of students in emerging economies suggests a growing pipeline of future professionals, while the dominance of data scientist roles in developed economies highlights the maturity and demand of the field in these regions. The large student populations in some countries might indicate future change in who leads the industry should these regions invest in the industry and nurture their emerging professionals; or ultimately risk losing them to regions that will provide them with the opportunities to grow and contribute to the field.

3. What are the most common tools, programming languages currently used by data scientist respondents? Is there a difference among tools used between Data Scientists compared to other roles?

Methods In order to investigate what the most common tools, programming languages, and products used by data scientists we will filter the survey data to include relevant columns such as roles taken from `Q5`, programming languages taken from `Q7`, tools from `Q9_`, and cloud products from `Q26_` excluding all rows with missing data.

We'll then segment the data into two groups, 'Data Scientists' (those whose occupation was explicitly marked as Data Scientist) and 'All Other Respondents'. For each group we can then reshape the data to create a frequency table for count and proportion for each response.

For visualization purposes, we will identify the top 20 most commonly used tools, languages and products for data scientists vs all other respondents. Then, in order to test whether data scientists use the same tools as all other respondents, we'll construct a contingency table of tool usage by group. We'll perform a chi-square test with Monte Carlo simulation (in order to avoid errors brought on by the small count values in some of the tools that might make the chi-squared test inaccurate) to determine if there's a statistically significant difference between the two groups.

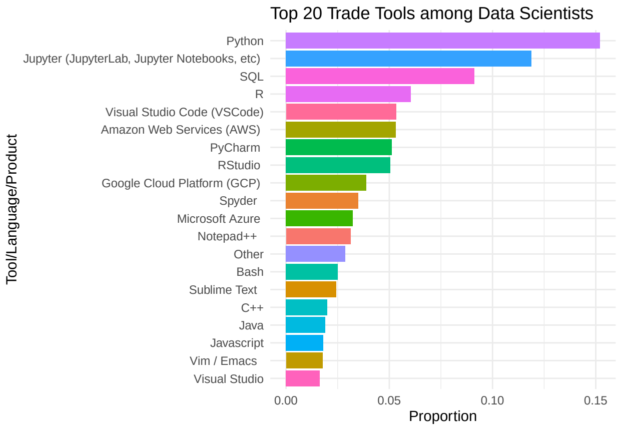
A p-value below 0.01 will lead us to reject the null hypothesis, indicating significant differences in tool usage patterns between data scientists and all other respondents. Otherwise, we will fail to reject the null hypothesis, suggesting similar tool usage.

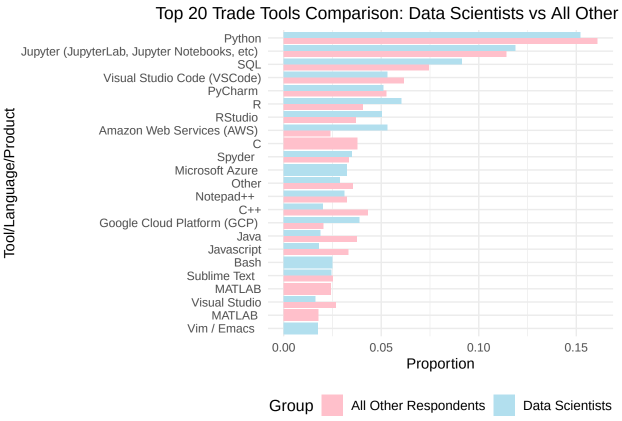
Analysis

```
## [1] "Chi-Square Test Results with Simulated P-Values:"
```

```
##
##  Pearson's Chi-squared test with simulated p-value (based on 10000
##  replicates)
##
## data:  contingency_table
## X-squared = 2017.8, df = NA, p-value = 9.999e-05
```

```
## [1] "Reject the null: Data scientists use different tools compared to all other respondents."
```





Conclusion Based on our observations, the most commonly used tools among Data Scientists include Python, Jupyter Notebooks, SQL, R, and Visual Studio Code (VSCode), with Python emerging as the clear leader. These tools are heavily utilized for data analysis, coding, and machine learning tasks, highlighting their relevance to the daily responsibilities of data scientists. The bar plot illustrating the top 20 trade tools among data scientists highlights their strong preference for tools tailored to data manipulation, statistical analysis, and cloud computing, such as AWS and Google Cloud Platform.

When comparing data scientists to respondents in other roles, we can find some significant differences in tool usage. The comparison plot indicates that while Python and Jupyter Notebooks dominate across groups, data scientists show a greater reliance on specialized statistical and machine learning tools like R and PyCharm. Conversely, other respondents display a more diverse usage pattern, including tools like C and MATLAB, which are less frequently used by data scientists.

The chi-square test further confirms these distinctions, rejecting the null hypothesis that data scientists and other respondents use tools in similar proportions (p-value < 0.0001). This statistical evidence supports the observation that data scientists adopt tools that align with their analytical and modeling-focused workflows, differentiating them from other roles in the field.

Our findings emphasize the unique technological preferences of data scientists compared to their peers in other roles, reflecting their specialized focus on data-driven tasks.

4. What is the relationship between educational background and current job compensation? Is it possible to predict salary based on someone's level of education alone?

Methods

In order to analyze the relationship between education level and compensation, we'll start by selecting the relevant columns for Education (`Q4`) and Compensation (`Q24`).

Compensation in particular will need to be cleaned and standardized since respondents entered their income in different ways and encoding for symbols like $ dashes and commas might vary across regions. We will handle invalid UTF-8 encoding in the compensation column by replacing problematic characters with placeholders, non-numeric characters such as dollar signs, commas, and extra spaces will be removed, and non-standard dashes will be replaced with standard ones. Compensation ranges (e.g., "50,000–60,000") will be converted into numeric values by calculating their midpoint. Rows with invalid or non-parseable compensation values will be filtered out.

We will also simplify Education levels into short categories to facilitate analysis.

We'll start with some descriptive statistics so we can get some insights into the central tendencies and variability of compensation by education level. Then, we'll create some visualizations like a bar and box plots to show the distribution of salaries across education levels.

By conducting an ANOVA test we'll determine whether the mean compensation differs across education levels, and a linear regression model will be used to asses whether education level can predict compensation. We'll then evaluate the model using Mean Absolute Error to measure its predictive performance.

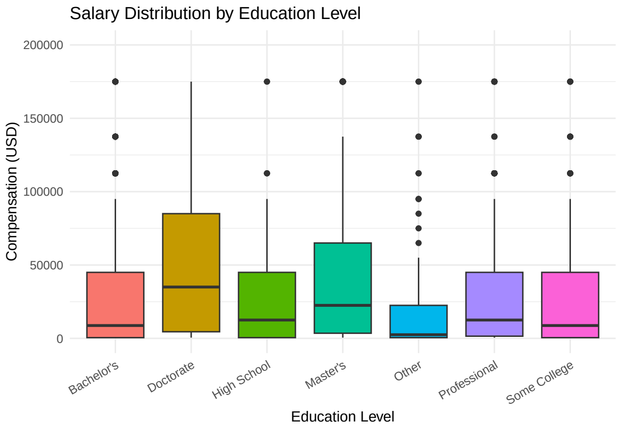
Analysis

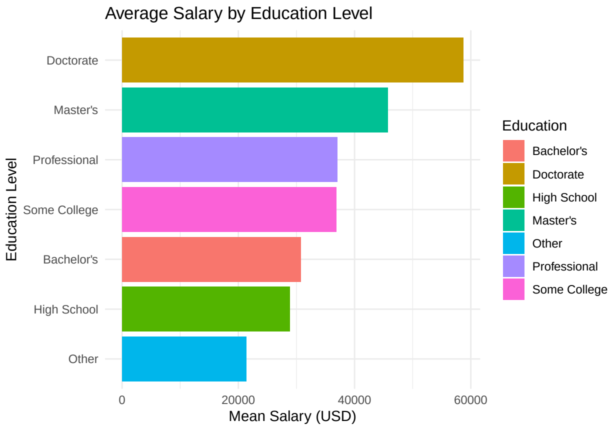
```
##                Df    Sum Sq   Mean Sq F value Pr(>F)
## Education       6 1.021e+12 1.701e+11   52.76 <2e-16 ***
## Residuals   10672 3.441e+13 3.225e+09
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## [1] "Mean Absolute Error (MAE): 40861.5676871915"
```

Table 3: Model Results: Relationship Between Education and Compensation

Term	Estimate	Std. Error	t-value	p-value
(Intercept)	30798.053	1036.429	29.7155375	0.0000000
EducationDoctorate	27877.673	1723.356	16.1763866	0.0000000
EducationHigh School	-1930.628	5612.124	-0.3440103	0.7308454
EducationMaster's	14891.007	1318.220	11.2962947	0.0000000
EducationOther	-9388.293	4663.193	-2.0132755	0.0441107
EducationProfessional	6226.504	2819.555	2.2083287	0.0272425
EducationSome College	6031.971	3074.092	1.9621957	0.0497656





Conclusion

As we observe the relationship between Education and Compensation, the box plot reveals that compensation distributions do vary significantly by education level, with a higher range of salaries belonging to individuals with higher degrees of education like Doctoral and Master's degrees when compared to other education levels. The bar plot further emphasizes these differences by showcasing that individuals with Doctoral degrees far outperform everyone by earning the highest average salary (approximately $58,000), followed by those with a master's degree (around $45,000). In contrast, individuals with only a High School education or "Other" qualifications tend to have the lowest average salaries. The average compensation seems low due to how the income values were aggregated (midpoint between income ranges) but the overall patterns are preserved.

The ANOVA results confirm that educational background has a statistically significant impact on compensation (p-value $<$ 2e-16), indicating differences in mean compensation across education levels. This is further supported by the linear regression results, which show that holding a doctoral degree is associated with a significant increase in compensation compared to the baseline (bachelor's degree), while other education levels, such as "Professional" and "Some College," have smaller but still statistically significant effects. Notably, individuals with only a high school education do not exhibit a statistically significant difference in compensation compared to those with a bachelor's degree.

Despite the significance of education in predicting compensation, the regression model explains only a small portion of the variance in compensation (adjusted $R^2 = 2.83\%$). This low R^2 value, combined with a high mean absolute error (MAE = $40,861), suggests that while education level does contribute to compensation differences, other factors such as industry, job role, location, or experience likely play a more substantial role in determining salaries.

In summary, education is an important but limited predictor of compensation. Advanced degrees, particularly doctoral and master's degrees, are associated with higher earnings, but further factors need to be taken into account to better explain the variability in compensation.

III. Advanced Analysis

What is the distribution of respondents' gender, and country of residence? Is the data science field representative of the overall population? Would this have any consequence?

Methods To examine the distribution of respondent's age and gender and compare them to the population at large we obtained a `world_population.csv` file containing the population for each country for several years. Since the survey was conducted in 2020 we focused on the Year 2020 column in order to compare the representation by region for survey takers and the population at large.

As far as gender goes, since most countries have a female share of the population between 49 and 51 percent (within one percentage point of an equal share of men and women), we decided to split global male and female ratios evenly, allocating 4 percent of the overall global population to other genders like nonbinary, trans etc in order to offer some degree of representation. The exact figure is hard to measure given that this metric has not been formally investigated, so we wanted to at least give it some consideration while not affecting the female /male ratio, and the overall results given current gender parity.

We will extract and merge the relevant columns from the survey and the world population data, filter out rows with missing or invalid data, and standardize country names so that we can merge appropriately.

We'll create visualizations and numerical summaries to compare the distributions if they are significant.

Analysis .

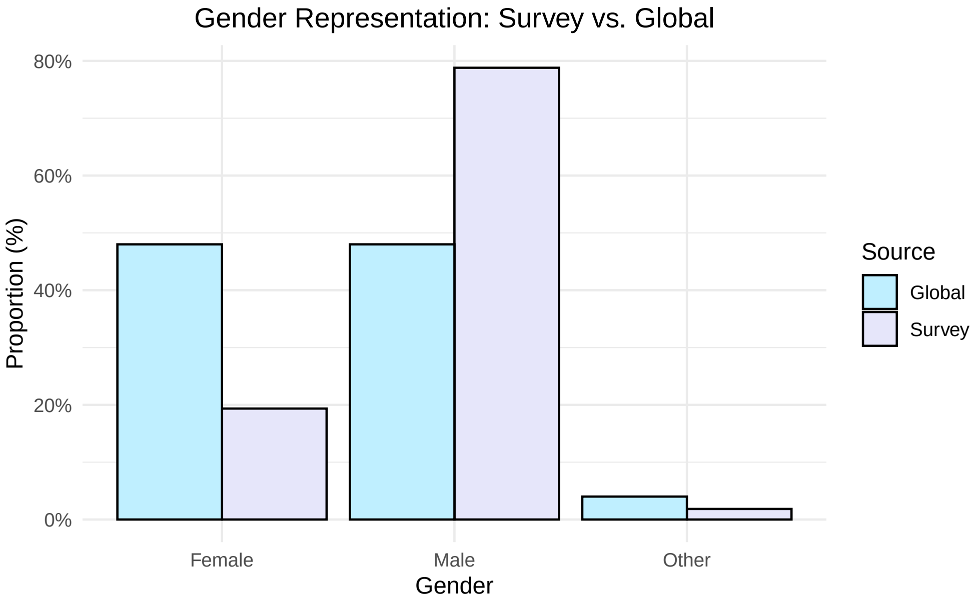
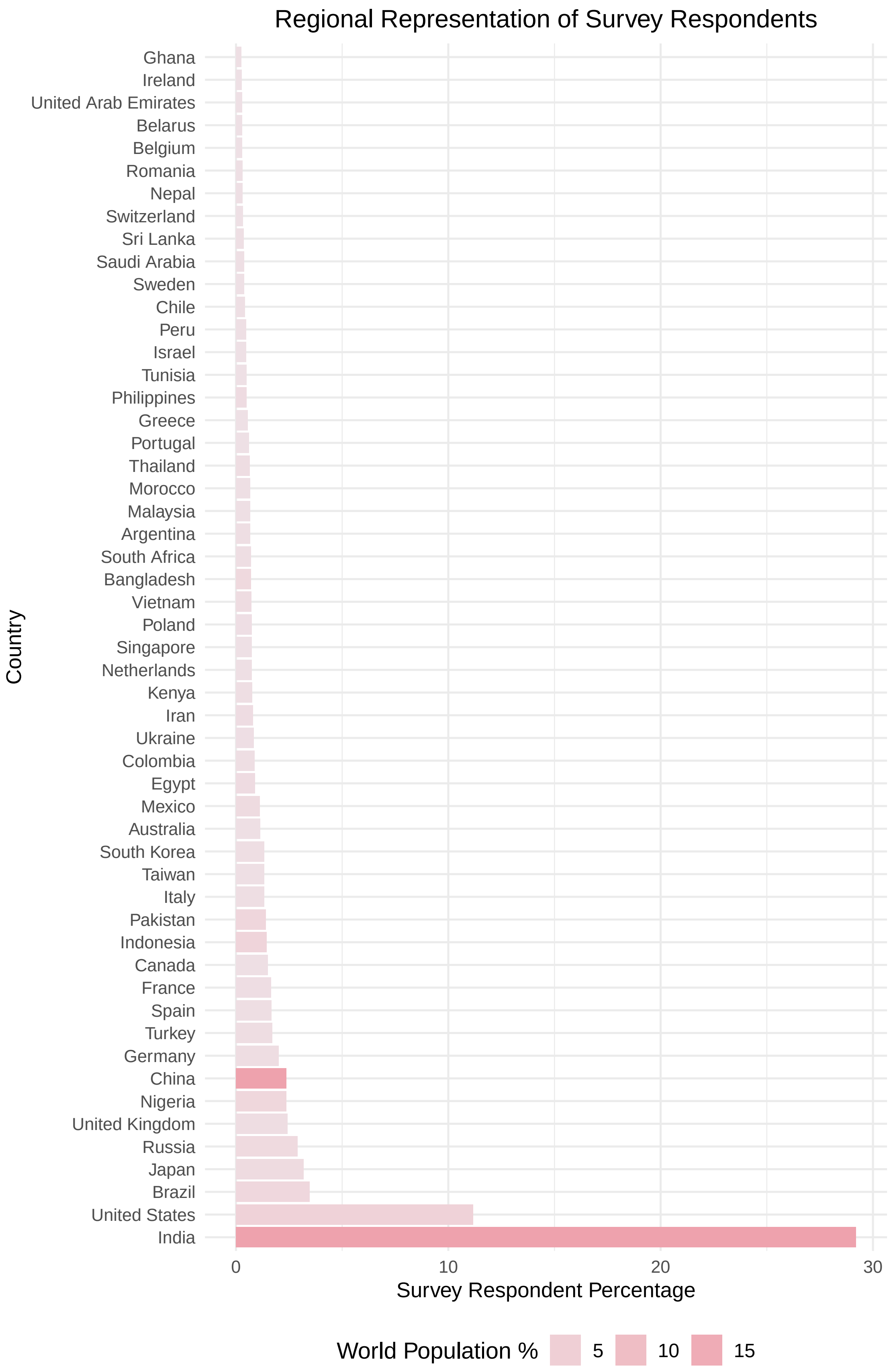


Table 4: Numerical Analysis of Gender Representation Disparities

Gender	Survey Proportion (%)	Global Proportion (%)	Absolute Difference (%)	Percentage Difference (%)
Female	19.36	48	28.64	-59.68
Male	78.80	48	30.80	64.17
Other	1.84	4	2.16	-53.96


Regional Representation of Survey Respondents
Ghana
Ireland
United Arab Emirates
Belarus
Belgium
Romania
Nepal
Switzerland
Sri Lanka
Saudi Arabia
Sweden
Chile
Peru
Israel
Tunisia
Philippines
Greece
Portugal
Thailand
Morocco
Malaysia
Argentina
South Africa
Bangladesh
Vietnam
Poland
Singapore
Netherlands
Kenya
Iran
Ukraine
Colombia
Egypt
Mexico
Australia
South Korea
Taiwan
Italy
Pakistan
Indonesia
Canada
France
Spain
Turkey
Germany
China
Nigeria
United Kingdom
Russia
Japan
Brazil
United States
India
Country
0
10
20
30
Survey Respondent Percentage
World Population %
5
10
15

Conclusion The analysis of survey respondents' demographics highlights a significant disparity between the representation of genders and global population distributions. Based on the survey data, males constitute 78.8% of respondents, a figure that far exceeds the assumed global gender balance of 48%. Females, who represent an estimated 48% of the global population, account for only 19.36% of survey respondents—highlighting major underrepresentation. Gender non-conforming individuals identifying as "Other" are underrepresented as well, comprising only 1.84% of survey respondents compared to the assumed global figure of only 4%.

This gender imbalance is further reinforced by the numerical analysis, which reveals a 64.17% overrepresentation of males and a -59.68% underrepresentation of females in the survey. Such disparities align with broader trends discussed in studies, such as the underrepresentation of women and nonbinary individuals in STEM fields, particularly data science and related technical disciplines. Articles from MIT Professional Programs emphasize systemic barriers, including cultural stereotypes, unequal access to resources, and biases in hiring practices, which may explain these trends.[1]

Additionally, the regional representation plot underscores that the majority of respondents hail from countries like India, the United States, and Brazil, while many regions, particularly in Africa and smaller nations globally, remain grossly underrepresented. This concentration of respondents from select regions could limit the survey's ability to provide insights into the global data science workforce and its inclusivity.

In conclusion, the underrepresentation of females and gender non-conforming in the survey reflects persistent gender disparities in data science, which could have significant implications for equity and innovation in the field as well as persisting biases in models that will fail to account for real world data given the limited scope of the developer workforce.[2] Addressing these gaps requires targeted efforts to foster greater inclusivity through outreach, mentorship, and systemic changes to reduce barriers to entry for underrepresented groups. Improving regional diversity in survey participation could help capture a more comprehensive picture of the global data science workforce.

IV. Conclusion / Discussion

Our report has explored various facets of the data science profession, aiming to understand its current state, representation, and disparities, as well as its broader implications. By analyzing survey data and comparing it to global population benchmarks, we addressed multiple questions and uncovered critical insights into the demographics, tools, compensation, and regional representation of data science professionals.

Data scientists predominantly rely on Python, SQL, and Jupyter-based tools, reflecting the field's focus on programming and statistical analysis. However, when compared to the broader survey population, data scientists exhibit a distinct preference for domain-specific tools like R and Jupyter, with statistically significant differences confirmed via a chi-squared test. This underscores the specialized nature of data science and its reliance on technical expertise.

Educational background has a measurable impact on compensation, with individuals holding doctorates earning significantly higher salaries compared to those with other degrees. Despite a statistically significant relationship between education level and compensation (as shown by ANOVA results), the linear regression model demonstrated only modest predictive power, suggesting that other factors, such as experience, industry, and geographic location, likely play critical roles.

Our analysis revealed significant disparities in the demographics of survey respondents compared to global population data. Males overwhelmingly dominate the data science field (78.8%), while females (19.36%) and gender non-conforming individuals (1.84%) are underrepresented. This is a stark contrast to the global gender distribution of 48% male, 48% female, and our hypothetical 4% of gender non-conforming individuals. Moreover, regional representation is concentrated in a few countries like India and the United States, while many regions, particularly in Africa and smaller nations, are significantly underrepresented. These disparities highlight systemic barriers that limit inclusivity and diversity within data science and pose developmental risks and bias permanence and normalization.

The gender and regional imbalances observed have far-reaching consequences. The lack of representation of women and nonbinary individuals in data science may hinder innovation by excluding diverse perspectives.

Similarly, regional disparities suggest unequal access to data science opportunities, which could perpetuate global inequities in technological advancement. Addressing these gaps will require targeted interventions, including educational outreach, mentorship programs, and systemic changes in hiring and training practices.

In conclusion, while the field of data science continues to grow and innovate, it is evident that it has not yet achieved equitable representation across genders, regions, and socio-economic backgrounds. The tools and compensation analysis reflects the technical depth and financial potential of the profession, but the demographic disparities raise important questions about inclusivity, and bring into question the reliability of developed technology where such obvious biases endure. To ensure that data science evolves as a globally representative and inclusive field, stakeholders must commit to reducing barriers, fostering diversity, and investing in underrepresented communities. These actions will not only promote fairness but also enhance the field's ability to solve complex, global challenges, and better capture the world we live in.

Appendix

[1] The Gender Gap in STEM: Still Gaping in 2023 - MIT.edu

[2] Bias and fairness in machine learning and artificial intelligence