Group Two

Assignment One

Presented to Dr. Andy Asare; Khizer Kamran

Camila Delgado, Sayad Subhan Shah Sadat, Joshua Smith, Amaka Onuchukwu

Introduction to Actionable Data Visualizations

DATA 605

DataThon 1

TABLE OF CONTENTS

INTRODUCTION

The 21st century has witnessed an unprecedented surge in technological advancements and innovations. Among the various transformative forces shaping industries across the globe, Data Science has emerged as an important catalyst creating a platform for organizations seeking to gain a competitive edge. With the expansion of data acquisition efforts and the increasing demand for informed decision making, the role of data scientists and machine learning experts has become irreplaceable. The 2022 Kaggle Machine Learning & Data Science survey dataset, consisting of responses from 23,997 participants offers a panoramic view of the dynamics driving this field.

Over the years there has been a paradigm shift from the traditional data analysis methods to more advanced methods involving machine learning and artificial intelligence which can be attributed to improvement in cloud computing and technology. Today, we see a convergence of disciplines such as Computer science, statistics and mathematics creating solutions for business problems.

In this study, we aim to deeply explore the intricacies of the data science landscape by analyzing the Kaggle dataset. By addressing the following guiding questions, we would uncover valuable insights into the relationship between several factors such as company size, tool used and programming languages preferences. Understanding these relationships can uncover actionable opportunities for individuals and organizations to thrive in our data- driven world.

In subsequent sections, we would discuss the methodology, data analysis techniques, our findings and implications of our review including actionable recommendations for improvement.

OBJECTIVES AND QUESTIONS

In order to better understand the contemporary data science landscape better, it is vital to have well defined objectives and goals that should be met by the end of the paper. In this paper, it was concluded that key information could be derived from investigating both the general trends of the world of data science supplemented by an exploration of the trends of the top 7 countries that are frontrunners in technology. According to one source, the leading countries in terms of nominal GDP in technology are the following (Wallach, 2021):

- 1: United States of America
- 2: China
- 3: Japan
- 4: Germany
- 5: India
- 6: United Kingdom
- 7: France

Consequently, the objectives of this paper are as follows:

Objectives

- Investigate the relationship between level of education and yearly compensation for Canadians
- Examine the difference in job titles with respect to machine learning experience and level of education.
- Explore the correlation between respondents' educational attainment, experience in Machine Learning and their respective Job titles.
- Analyze the usage of different programming languages (Python, Javascript, R) across the top 7 Tech giants around the globe.
- Inspect the level of education (Bachelor's, Master's, so on) across the top 7 Tech giants.

Questions

Naturally, these objectives are followed with their own respective questions which will assist in addressing them as best as possible. The questions that this paper will look to answer are as follows:

1. How does current yearly compensation for Canadian Respondents change based on their education level
2. How do respondents' job titles differ based on their education level and machine learning experience?
3. How does respondents' job title change the set of data visualization tools (Tableau, PowerBI, Google Data Studio) they use?
4. What programming languages are popular amongst the top 7 global tech giants across the globe?
5. How is level of education distributed amongst the top 7 global tech giants across the globe?

DATA COLLECTION AND CLEANING

Data Collection

The data collection process for "DataThon 1" was a relatively rudimentary process. The original dataset titled "2022 Kaggle Machine Learning & Data Science Survey" was extracted from Kaggle and served as a data source for a Kaggle competition. The original data source was in structured CSV file format; consisting of 23999 rows across 239 columns.

However, students were provided with a reduced version of this dataset via the Desire to Learn (D2L) platform for the function of increased simplicity. This reduced version of the original data, also in structured CSV file format, consisted of 23998 rows across 19 columns; drastically reducing the file size of the data (3.47MB from the original 24.81MB). The final dataset prior to cleaning consisted of two numerical columns/attributes and 17 categorical columns/attributes.

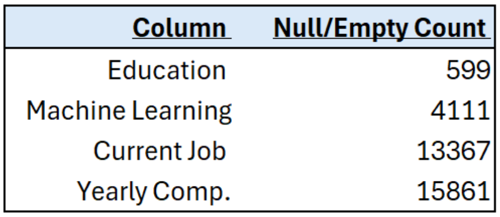
Data Cleaning

To be consistent with the respective themes of the course, data cleaning was conducted in Tableau Prep as opposed to Python or R before being loaded into Tableau for analysis. As the reduced (final) dataset had already received an initial round of preliminary cleaning, there was little work to be done; however, three primary considerations were taken into account.

Firstly, only particular columns or attributes were being considered for the analysis section of the report; subject to the questions we either elected to choose or formulate. Thus, not all columns were utilized in the analysis section and could be dropped from the data via Tableau prep. Ultimately duration, age, “[are you currently a student]”, programming experience, current industry, and current company size were dropped.

Looking to the second cleaning step, we are presented with the notion of column titles. Many of the column titles present, although descriptive, do contain unnecessary characters. Six of the 13 remaining columns after the first cleaning step above were subject to change. Particularly, the three columns pertaining to Programming languages (Python, R and JavaScript) and the three columns regarding data-oriented software (Power BI, Google Data Studio, and Tableau) underwent alteration. All programming languages underwent the removal of the “Used Regularly” phrase (i.e. Python Used regularly) to simply reflect the language (i.e. Python). A similar process was undertaken for the data-oriented software columns by removing the word “used” to merely reflect the software.

When considering the final and most crucial potential step toward the cleaning process, one must consider the notion of null or missing values in the dataset that follows the two cleaning steps above. Of the remaining columns, we observe that there is a total null or empty count of 33,938 missing values across the dataset. By the remaining columns that contain null values following the above two cleaning steps, we observe the following distribution of Nulls:



Column	Null/Empty Count
Education	599
Machine Learning	4111
Current Job	13367
Yearly Comp.	15861

Figure 1: Null/Empty Count via Column

It is crucial to note that if one were to drop all rows that contain null or empty values, one would reduce the dataset from 24,000 rows to 7,000 rows. Dropping 70% of the dataset would be unacceptable as we would be losing critical values in other columns by dropping an entire row. One may argue that we can go column by column and perform a mean or linear interpolation to fill these missing values, however, interpolating such a high degree of data points within the dataset could lead to inaccurate, skewed, or misleading results. Thus, all empty values shall remain present in the cleaned, final version of the dataset and dealt with for individual cases of each visualization if required.

DATA VISUALIZATION PROCESSES AND TECHNIQUES

Objective 1:

When observing our first research objective: “Investigate the relationship between the level of education and yearly compensation for Canadians”, one can look to the provided question of “How does current yearly compensation for Canadian respondents change based on their education level?”.

To perform such a task, one needs to consider two basic dimensions: a respondent’s level of education (categorical), and the range that a respondent's annual/ yearly salary falls into (technically a category as the attribute exists as a string-based range; allowing the attribute to be categorical in nature opposed to numerical). Once these two variables have been identified, subject to the requirement that the individual’s country of residence is Canada, one can split the salary range column into two columns; the first column representing the first value in the salary range (the minimum – converted to a whole integer) and the other representing the second value (the maximum – converted to a whole integer) of the salary range. An average of both columns can then be computed to determine the average minimum and maximum salary for each educational level or degree obtainment.

Objective 2:

In leveraging Tableau to visualize how respondents' job titles differ based on their education level and machine learning experience, several key steps were undertaken to ensure a comprehensive and insightful representation of the data. Columns representing respondents' education level, machine learning experience, and job titles which were required to meet the objective were identified as key variables for analysis. The visualization involved the x-axis representing the various job titles, the y-axis displaying the different education categories (e.g., Bachelor's, Master's, Doctoral) and the distinct count measure applied to Machine learning experience for clarity. This allowed for a clear comparison of job title distributions across different education levels and machine learning experience.

Objective 3:

The third objective of this data required the effect of job titles of the respondents over their usage of visualization technologies to be inspected and investigated. To accomplish this investigation a tally approach of the different visualization technologies with respect to each job title within this dataset was visualized in horizontal and grouped bar charts. This approach accomplished two goals at the same time such that not only did it display the relative ratio of the usage of each visualization technology per job title compared to each other, but it also displayed the frequency of responses for each job titles as well.

Objective 4:

The fourth objective of this paper required the investigation of the different programming languages that the residence of the top 7 global technology leaders mentioned above most frequently use or are proficient in. To visualize the popularity of these languages, initially a simple tally was taken for each programming language for each of the 7 countries. It was quickly found that this approach would not yield easily interpretable and concise results since the number of respondents for each country were different compared to the rest. For instance, India had more than 7,000 responses while the other countries would have less than 1,000.

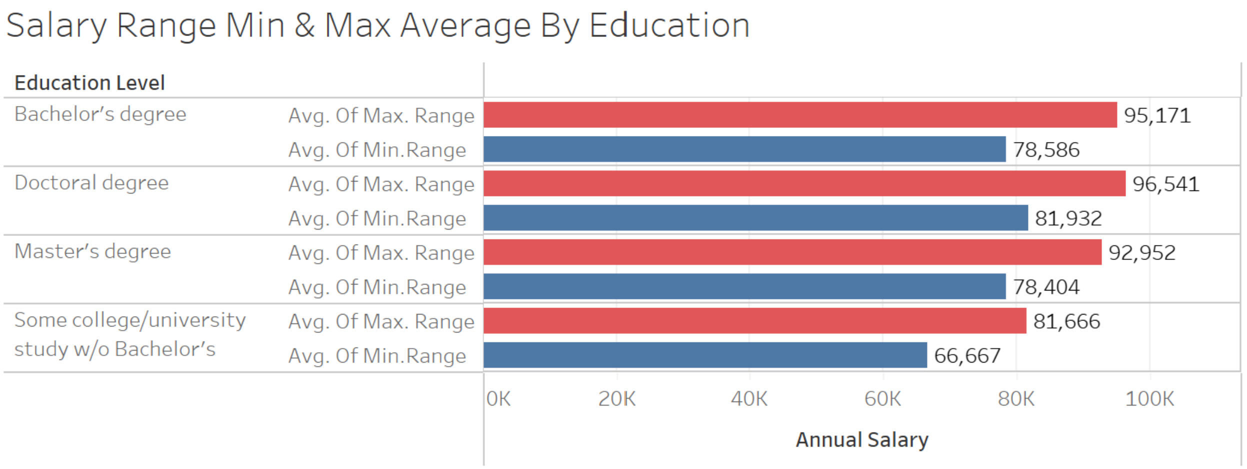
In order to combat this issue, the ratios of the respondents' positive answers compared to their total answers were considered instead of their total count. The reasoning behind this approach is to offset the discrepancies in the number of responses since a ratio-based approach would only look at the percentages of each response for each country compared to their total count. This approach seems to have resolved the issues found with the first approach. Hence, the percentages for each programming language for each respective country were visualized with grouped bar charts.

Objective 5:

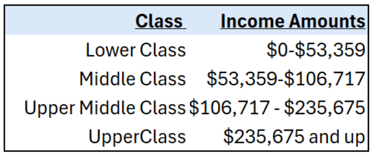
For the fifth objective of the paper, the frequency of each level of education for the respondents of each of the 7 countries mentioned above were required to be explored. A similar issue was found when the tally of each level of education was visualized with grouped bar charts for each country. Similar to the issue faced above, there appeared to be a large discrepancy between the number of responses recorded for India compared to the rest of the countries listed on this paper. Again, the same approach of visualizing the ratios of each level of education for each country was adopted and appeared to yield the appropriate results. The resulting grouped bar charted depicted the relative frequency of each level of education compared to the rest of the countries for better visualization and interpretability.

DISCUSSION OF VISUALIZATION FINDINGS

Objective 1:



When reviewing the visualization above, the results/findings provide clear and distinct insight into the research question and objective at hand. To begin this analysis, one can refer to the Canadian income brackets below (Government of Canada, 2023):

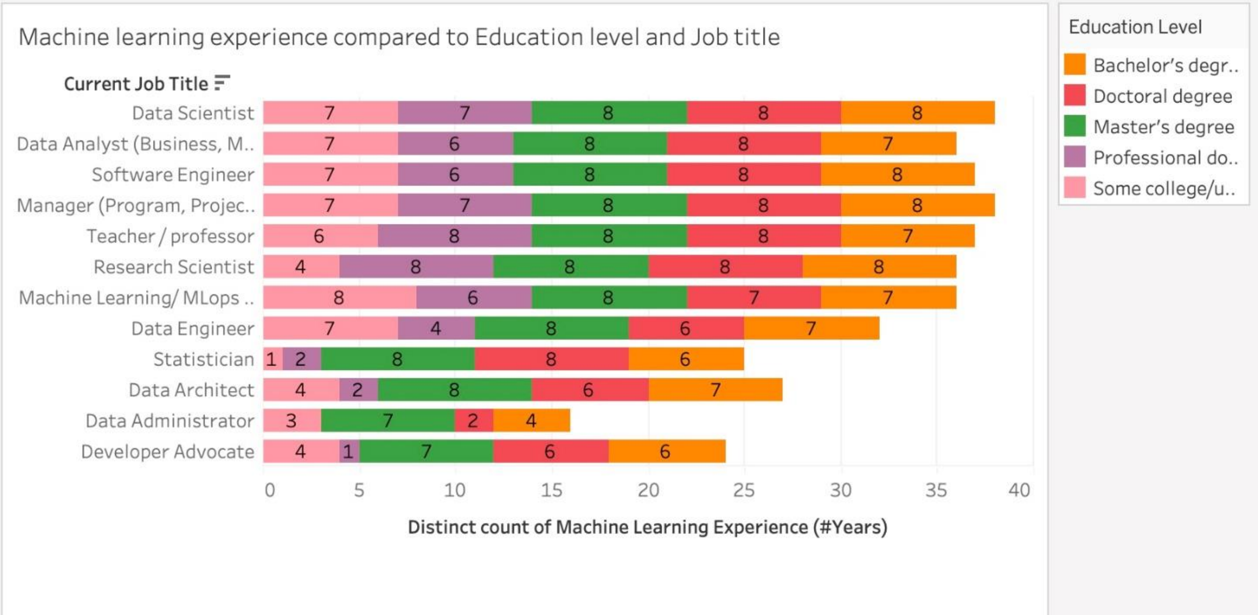


Class	Income Amounts
Lower Class	$0-$53,359
Middle Class	$53,359-$106,717
Upper Middle Class	$106,717 - $235,675
UpperClass	$235,675 and up

Figure 2: Canada Income Brackets
Source: Government of Canada

Looking to the visualization in comparison to the census data above, insights can be derived. Firstly, one observes that the average minimum and maximum salary across educational levels falls within the "middle class income bracket" as defined by the government of Canada (Government of Canada, 2023). We can see that the average minimum and maximum salaries across bachelor's degrees, master's degrees, doctoral degrees, and those who have some degree of college or university without earning a bachelor's is relatively consistent across the educational categories. It is noteworthy to mention that doctoral degrees do possess a slightly higher average minimum and maximum salary compared to the other three educational categories. Moreover, those who have some degree of college or university without earning a bachelor's do have a consistently lower average minimum and maximum salary compared to those who have completed their degrees (bachelors, masters, and doctoral degrees). Thus, based off the data (despite limitations), it is evident that those who have doctoral degrees earn greater wages than those with lower educational obtainment and those who have "some college/ university" (diploma, incomplete, etc.) earn less than those with formal bachelors, masters, and doctorate degrees. As a result, there appears to be a clear positive relationship between higher educational obtainment and higher annual salary/ wages. It is noteworthy to mention that "professional doctorates" and "no formal education past high school" have been removed from the visualization and analysis as they contained drastically fewer data points than other educational levels and could result in skewed data or inaccurate results of averages; for instance, prior to removal, the average maximum and minimum salaries of those with a high school education was 250,999 and 150,500.

Objective 2:

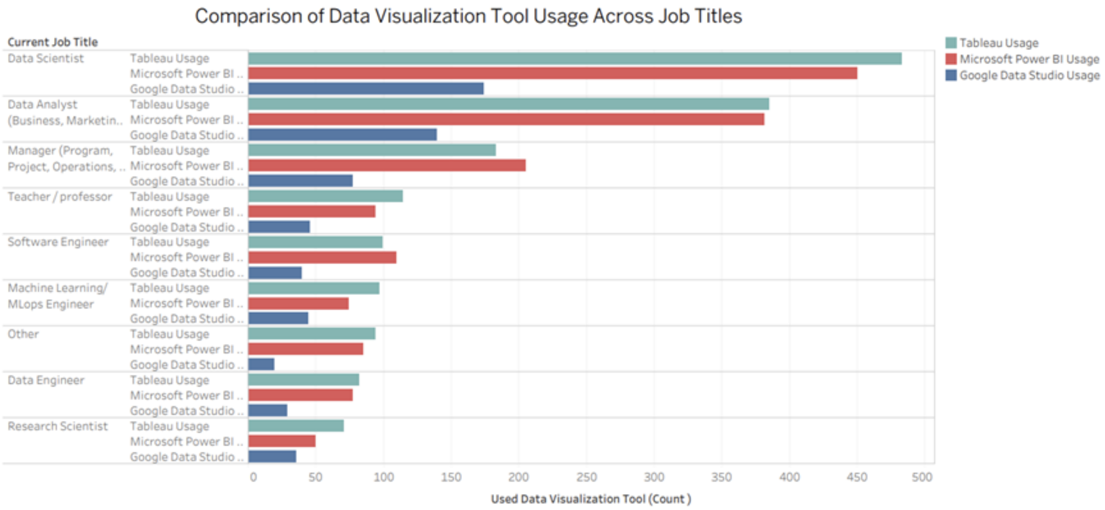


The graph above shows a stacked bar chart of the years of Machine learning experience for various job titles and educational qualification ranging from bachelor's degree to Professional degrees. To answer the question of how respondents' job titles differ based on their education level and machine learning experience, we explored the correlation between respondents' educational attainment, Machine learning experience and job titles. By cross-referencing respondents' education backgrounds with their machine learning expertise, we can uncover patterns that reflect the interplay between academic qualifications and hands-on skill. To create meaningful visualization of this correlation, we considered the distinct count of the Machine learning experience of the respondents. From the results, the different levels of education and machine learning experience spread across the survey group and influence career paths. Clearly, all job categories had bachelor's and master's degrees with Seven and eight years of Machine learning experience being the most prevalent.

Interestingly, we observe that individuals with less machine learning experience (1-3 years) took up roles such as Data Administrator, Developer advocate, Statistician, Data Architect and Engineer(non-software) which required fewer coding efforts. On the other hand, respondents with both higher education levels e.g. master's degrees, Doctoral degree and Professional doctorate and significant machine learning experience tend to gravitate towards more technical and senior roles as Data Scientists, Software Engineers, Professors, Machine learning Engineer etc.

The output also provides some insights to jobs role that may not require extensive machine learning experience and those for which machine learning experience is critical.

Objective 3:

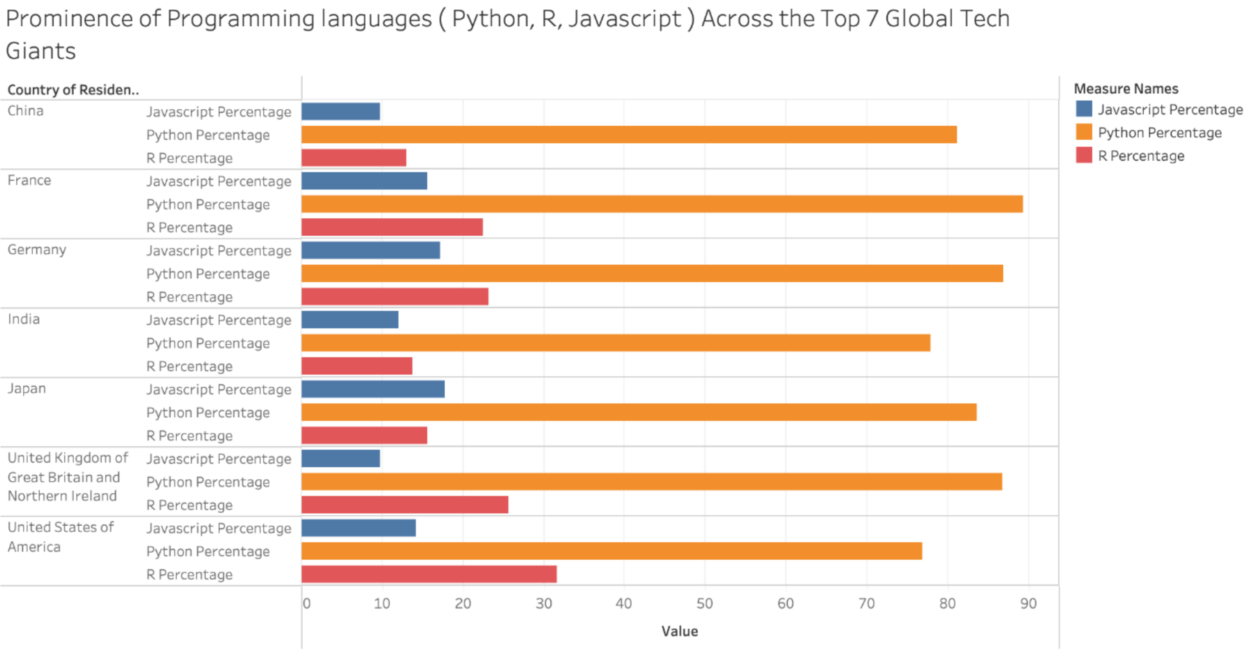


Objective three aimed to Compare Data Visualization Tool Usage Across Job Titles. We utilized a bar chart to depict the count of "True" statements indicating the usage of three visualization tools across various roles. The visualization underscores variations in tool usage across different job titles. For instance, Data Scientists exhibit the highest usage of all three tools (Google Data Studio, Microsoft Power BI, Tableau), followed closely by Data Analysts and Managers, suggesting role-specific preferences based on job responsibilities and data analysis needs. This highlights the need for adoption in data visualization tool selection within data-intensive roles.

Key takeaways from this visualization reveal that Tableau emerges as the most widely used tool across all job titles, with consistently higher usage counts compared to Google Data Studio and Microsoft Power BI. This underscores Tableau's dominance and widespread adoption in the data visualization industry. Additionally, while Tableau remains the most popular tool overall, Data Scientists and Data Analysts demonstrate relatively high usage counts for all three tools, indicating a tendency to leverage multiple tools for diverse data analysis tasks. Moreover, Managerial positions, particularly those in Program, Project, Operations, and Executive-level roles, exhibit a pronounced preference for Microsoft Power BI, suggesting its perceived suitability for generating business intelligence reports. These insights provide valuable understanding into the relationship between job titles and data visualization tool usage, enabling informed decision-making processes.

Objective 4:

As discussed above, for the fourth objective of this paper, the popularity of each programming language was required to be investigated to provide an answer for the paper's fourth question. The visualization below contains some key findings which can be derived from the data.



Prominence of Programming languages (Python, R, Javascript) Across the Top 7 Global Tech Giants

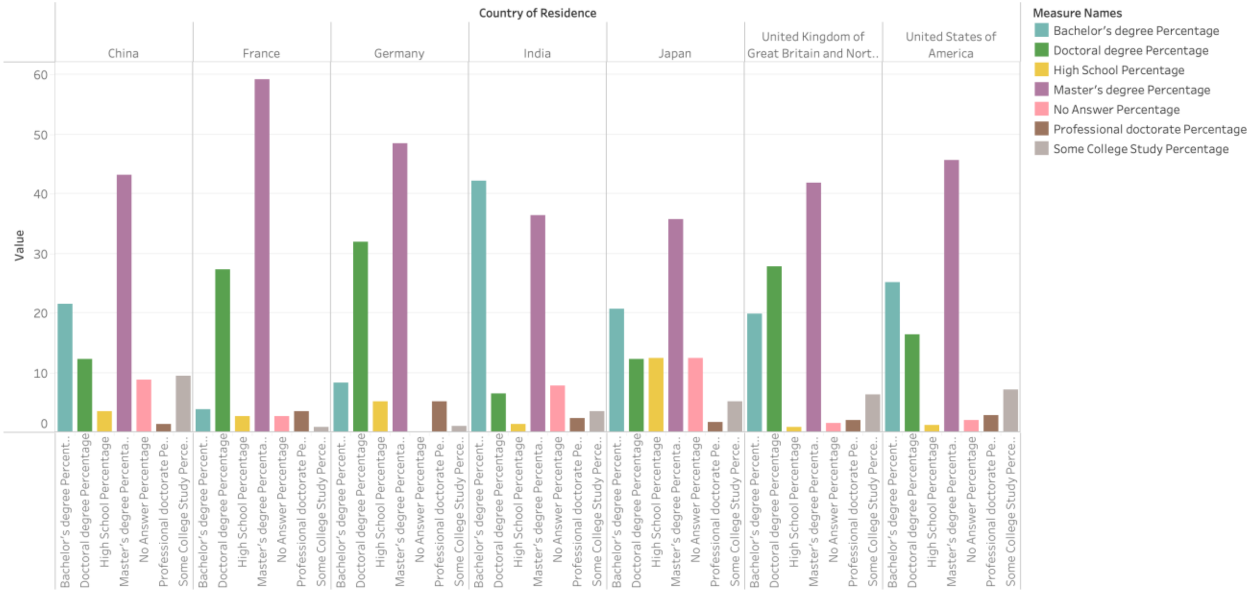
Javascript Percentage, Python Percentage and R Percentage for each Country of Residence. Color shows details about Javascript Percentage, Python Percentage and R Percentage. The data is filtered on Top 7 Tech Giants, which keeps 7 members.

As it can be observed, there is a general trend for Python to be the most popular programming languages across the top 7 countries in technology. In general, the percentage of positive responses for each programming language can be seen to be uniform across the different nations, there are two notable exceptions that can be observed. While R is the favored programming language compared to JavaScript for most of the countries above, the opposite is true for Japan. Furthermore, India also is close to not following this general preference. However, in general, the key takeaways from this visualization is that Python is the most popular and transferable skill across the tech giants and there further appears to be a general preference for R as well. These takeaways help provide a baseline for up-and-coming data scientists to focus more on improving their proficiency in these two programming languages primarily.

Objective 5:

For the fifth and final objective of this paper, the distribution of levels of education across the top 7 technology giants around the globe were to be investigated. The following visualization provides some key similarities and differences between the countries.

Frequency of Level of Education Across the Top 7 Global Tech Giants



Bachelor's degree Percentage, Doctoral degree Percentage, High School Percentage, Master's degree Percentage, No Answer Percentage, Professional doctorate Percentage and Some College Study Percentage for each Country of Residence. Color shows details about Bachelor's degree Percentage, Doctoral degree Percentage, High School Percentage, Master's degree Percentage, No Answer Percentage, Professional doctorate Percentage and Some College Study Percentage. The data is filtered on Top 7 Tech Giants, which keeps 7 members.

As it can be observed, like the programming languages visualization with respect to its uniformity. However, there are still some irregularities that can be observed alongside some interesting differences between Western and Asian nations. As it can be observed, in general, there are a larger number of master's degree holders compared to bachelor's degree holders across the different nations. However, one country that stands out in this respect is India where the ratio of bachelor's degree holders is higher compared to the number of master's degree holders. Another key observation from this visualization is the slight but noticeable difference between Western nations such as France, Germany and United States and Asian countries such as China, India, and Japan in that there appears to be a larger number of Doctoral degree holders in Western countries compared to their Asian counterparts.

PRACTICAL RECOMMENDATIONS

Objective 1:

A handful of practical recommendations exist despite the limitations of the data. As we see a relatively clear link between higher education attainment and a higher income, a potential recommendation would be a policy alteration to increase the accessibility of higher education or post-secondary education (increased tuition subsidies increased scholarship availability and financial aid to target low-income students. Other practical recommendations could include education institution counseling; perhaps aiding students in looking at the financial benefits of certain educational levels or workplace initiatives for organizations to support continuing education (however this could increase a firm's salary expenses).

Objective 2:

Based on the analysis, the following recommendations can be made to enable professional smoothly transition into specialized roles in data science. These recommendations may include designing tailored training programs, adopting targeted recruitment strategies, or skill development initiatives designed to address skill gaps and promote career advancement within the data science field. Specifically, Companies can invest in skill development programs tailored to employees' career aspirations and educational backgrounds. Data professionals should prioritize continuous learning and upskilling to remain competitive in the evolving job market.

Objective 3:

The analysis of tool usage across diverse job titles highlights the importance of optimizing training programs to accommodate varying tool preferences. To address this, organizations should develop comprehensive training initiatives covering multiple visualization tools, empowering employees to select the most appropriate tool for their specific data analysis needs. Additionally, there is an opportunity to enhance tool adoption among Data Engineers and Machine Learning/MLops Engineers through targeted training efforts tailored to their unique roles and requirements. Furthermore, continuous evaluation of the organization's tool portfolio is recommended to ensure alignment with evolving job roles, data analysis demands, and industry trends. By adopting these strategies, organizations can empower employees to navigate the evolving data visualization landscape proficiently.

Objective 4:

The key findings for the fourth question provide something to consider since a few of the countries appear to not follow the general trend of preferring R over Javascript. For this reason, specifically, the recommendations of this paper with respect to its fourth objective are the following:

1) Individuals that are looking into data science should focus primarily on python since it is the most frequent programming language across the globe. Developing high skill in python may prove to be more fruitful compared to the rest of the programming languages due to its transferability.
2) Region specific career strategies can be a deciding factor for individuals getting into the data science field to be successful in industry. For instance, a data scientist in Japan may tailor their strategy to be more JavaScript oriented compared to a data scientist in the United States.

3) A general level of importance given to R also appears to be noteworthy since most of the countries prefer it compared to JavaScript. The respondents in biggest technology giant that is the United States in this list appear to favor highly towards R compared to the rest of the countries.

Objective 5:

The key findings for the fifth question of the paper were far more interesting since it exposed the irregular behavior of India compared to the rest of the countries within this list which appear to be mostly similar to one another. Consequently, the recommendations of this paper regarding the fifth objective are centered around India and they are the following:

1) One plausible reason behind the lack of Master's and Doctoral degree holders in India may be because of a lack of institutions. Hence, the Indian government should focus on investing into the country's graduate/postgraduate institutions to alleviate any lack of higher education institutions within the country.

2) Another probable reason behind this phenomenon can be the lack of roles and job positions that require a higher level of education within the country. To offset this issue, the Indian government can also focus on providing incentives for the private sector to expand into services and goods that require higher levels of education to provide or produce.

3) Since there can be many distinct factors contributing to a lower level of higher education degree holders in the country such as immigration, government policies, lack of investors

and so on, the Indian government should investigate to shed light on the root of this phenomenon.

LIMITATIONS OF STUDY

Although the findings of the paper help in providing the answers to some important questions for individuals that are just starting their data science journey, it does come with some limitations that are as follows:

1) Incomplete Data: The dataset had an imbalanced number of responses for the different columns and variables that it had gathered information for. For example, it was found that there was a significantly considerably large number of responses for individuals that reside in India compared to the other nations that this paper focused on. Although these imbalances within the dataset were offset with the adaptation of ratio-based approaches, it could still significantly affect the key findings of the paper.

2) Response Bias: As it appeared with the investigation of the dataset's variables, there is a difference in the number of responses for different job titles as well. One contributing factor for this difference can be a response bias where individuals with certain job titles may have not been as inclined to take the survey as the others. This could also have adverse effects on the findings of the paper.

3) Data credibility: The dataset that this paper conducted its analysis with was sourced from Kaggle where the dataset was shared with the public. Although Kaggle is a well-known

source for a variety of different datasets, the data collecting practices performed by the individuals responsible for gathering the dataset may have not been ideal. As a result, the credibility of the data goes on to reduce the credibility of the findings through it.

4) Quality of Data: The dataset itself, although preliminarily cleaned, did possess an immense degree or amount of missing data, as mentioned in the data-cleaning process; in our unique case, after having dropped all irrelevant rows: specifically, 33,938 missing (Null/ empty) are still present (see Figure 1).

REFERENCES

Wallach, O. (2021). The World's Tech Giants Compared to the Size of Economies. *Visual Capitalist.* Retrieved from: https://www.visualcapitalist.com/the-tech-giants-worth-compared-economies-countries/

Government of Canada. (2023). *Income Tax Brackets 2023.* Retrieved from: https://www.canada.ca/en/revenue-agency/services/tax/individuals/frequently-asked-questions-individuals/canadian-income-tax-rates-individuals-current-previous-years.html

Kaggle. (2022). *2022 Kaggle Machine Learning & Data Science Survey*. Retrieved from: https://www.kaggle.com/competitions/kaggle-survey-2022/data