CFG PROJECT REPORT: DATA CAREER TRENDS FROM 2020 - 2022

Maria Contreras, Dabrowka Goral, Fussilat Ibrahim, Sarah Lavelle, Liman Li, Clare Xie

Introduction

The aim of our project is to explore data career trends from 2020-2022, using cross-sectional survey data from Kaggle. We will generate insights into the evolution of data careers, understand the demographics of data professionals, identify skills and tools required to transition into the data field and gain knowledge of compensation trends across the industry.

As such, the objectives of this project are to:

- Understand the raw data, what it is describing and how we can derive insights from this.
- Clean and normalise data across all years.
- Call an API to adjust compensation values for inflation, ensuring equal comparison across all years.
- Combine the datasets to form a master dataset with data from all years and allow for joint analysis and visualisation.
- Analyse our master dataset to understand the according to our question set.
- Visualise data in a manner that is meaningful, representative and understood by our target audience.

A brief overview of our project roadmap can be seen below. Please see our project log for further details.

- Pre-planning (28-30/4/23): Project scoping and initial research to identify source data and publicly available APIs.
- Week 1 (1-7/5/23): Finalise project topic and data sets to be used. Devise an initial question that poses specific and measurable questions, allocate team tasks and discuss and execute data cleaning methods.
- Week 2 (8-14/5/23): Clean individual data sets, ready to collate into a master data set.
- Week 3 (15-21/5/23): Create a master dataset, perform exploratory analysis, and refine our question set.
- Week 4 (22-28/5/23): Finalise visualisations and analysis and complete report.

Background

This project topic was chosen as it is particularly relevant to each of us, as prospective data professionals. To help us enter into the data industry, we are interested in understanding the tools that are currently most used within the industry and whether we can identify any trends to anticipate which skill sets may be in demand. Further, we chose to explore demographics of data professionals in order to understand the state of the data industry and the workforce we can expect to join.

Our target audience is prospective data professionals. We aim to provide an overview into the state of the Data field during 2020-2022, including demographics, tools used by existing professionals and potential compensation for different Data careers.

Through this project, we aim to answer the question: “What are the trends in data professionals across 2020-2022?”, specifically:

1. Is data science a male dominated field, and is that changing?
2. What skills/tools are needed for each of the different roles?
3. What is the relationship between yearly compensation and other relevant demographic information (age, gender, highest level of education, team size, company size, region of employment)?
4. What factors lead to higher compensation?

Our analysis will be performed to answer the above question set. This includes analysing data based on different factors (e.g. sex, tools used, education levels, employment information, compensation) and how these may influence one another. All raw data is qualitative data, requiring manipulation to convert to representative quantitative data prior to further analysis. From this, raw numerical data, calculated percentages/ratios and qualitative “bins” will be used for analysis.

Steps Specifications

Data gathering

Due to the limitation of working with publicly available data, our first step was to identify a high quality dataset. We chose Kaggle survey responses from 2017-2022 as it provides insights into data career trends across multiple years, allowing for comparisons over time. The Kaggle datasets collate demographic information, tools/technologies studied and used within the workplace, employment details, compensation and others from Kaggle users. To supplement our core datasets, we have utilised an API to pull inflation rates from the World Bank database, allowing us to compare compensation data across the years.

Framing questions

Initially, we assigned each team member a dataset from each of the years spanning 2017-2022. From here, we each took the time to understand our individual data sets in terms of: questions asked; types of responses given; and quality of the dataset.
We then reconvened to discuss key questions could be meaningfully answered, with final questions based on three criteria:

1. Availability of data across multiple years: data must be available across the years in order to accurately compare. This required discussions about the differences in questions and whether these could be fairly compared. For example, in

2017 the question, “At work, how often did you use <tool/technologies> this past year?” was shifted to be “What <tools/technologies> do you use on a regular basis?” from 2020 onwards.

2. Requirement for data manipulation: if heavy manipulation was required to answer one of our key questions, this was disregarded. If only slight manipulation was required, this was deemed acceptable for our acceptance criteria. For example, comparison of compensation across years would only require a simple inflation adjustment.
3. Relevance to our target audience. As the aim of this project is to generate insights into the state of the data industry, we ensured that each key question chosen will be of use to those pursuing careers in this field.

Pre-processing

These datasets required heavy cleaning before analysis could be completed. Each team member was given a dataset to clean, according to specific criteria. Details can be found in the code, however a brief overview includes:

- Standardisation of column names (e.g. tool frequency was named as “tools_used_<tool_name>”).
- Addition of a “year” column to track data the correct year, once combined into the master dataset.
- Categorisation of columns across years into the same “bins” (e.g. Age: 18-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50+).
- Dropping rows if there is a NaN value for any survey question - with exceptions for questions with multiple parts.
- Dropping rows if there is a NaN value for all of the multiple choice questions (i.e. the person doesn’t frequently use any of the listed tools).
- Encoding the dataset with numerical values rather than strings after dropping the rows with null values so that our machine learning model could work with the data.
- Nominal categorical variables were re-coded as dummy variables, for inclusion in the ordinal logistic regression model.

During the cleaning process, we decided to reduce our dataset to only include data from 2020-2022 as earlier data was formatted quite differently to these final three years, requiring a large amount of manipulation before meaningful comparisons could be drawn. Further, the 2020-2022 datasets alone contained enough data to generate valuable insights. A more in-depth cleaning was then completed to then merge the 2020-2022 datasets into a master dataset.

Analysis and visualisation

With our four research questions as our guide, we designed our approach to delve deep and explore each one meticulously.

To answer Question 1, we calculated the non-male representation in the field annually and charted this over time. The resulting line graph allowed us to visually depict the trends, highlighting any prominent shifts in gender representation. This visualisation was complemented by an in-depth analysis of potential contributing factors to these observed trends.

Question 2 was addressed through the utilisation of count plots. This gave us a visual representation of which tools were most frequently used among the most popular job titles. We also conducted an analysis to discern which job titles most frequently used the most popular tools. These visualisations proved instrumental in providing insights into the skills currently in high demand, as they clearly illustrated the relationship between job roles and the corresponding toolsets.

For Question 3, we used both stacked distribution interactive graphs to explore the relationship between yearly compensation and relevant demographic data. The former, constructed with the help of Plotly, offered an engaging and informative way to visualise the compensation distribution within each demographic group, and across multiple categories such as age, gender, job title, and coding experience. This interactive graph not only provided an overview of the compensation distribution but also allowed for a more detailed exploration on demand.

Finally, for Question 4, we implemented an ordinal logistic regression model. This model was designed to predict the chances of an individual falling into a specific income level category, based on independent variables like gender, region, job title, company size and team size. We additionally added in year dummy variables into the model to control for year related factors that may bias the model (i.e. year fixed effects). The model formed a critical part of our analysis, giving us a predictive lens through which to view our data.

Implementation/Execution

Development approach and team member roles

The team has managed all code via Google Colab. Each team member was assigned a year of the survey data to understand and complete an initial round of cleaning. Further, we used a range of shared documents (Jamboard, Google Sheets, Google Docs) to scope the project and to share information that is not written code. Following this, we had regularly scheduled calls to ensure that our individual code is cohesive with the wider project, and to discuss challenges encountered and how to overcome these.

We undertook the majority of the processes of combining the 2020-2022 data sets, performing analysis and producing visualisations, over group Zoom calls. This allowed us to write and execute code, provide a forum to air concerns and problem solve as an entire group.

We further allocated work based on our own individual strengths. Following initial data cleaning, merging and more detailed cleaning and manipulation (e.g. standardising columns, calling an API, normalising data) was conducted by Liman, Maria and Sarah. Creating and fitting the regression model was also created by this group. Analysis and visualisation was allocated to different team members by question: Question 1 was completed by Liman and Fussilat, Question 2 by Clare and Liman, Question 3 by Maria and Question 4 by Sarah. All group members contributed to aspects of report writing.

Tools and libraries

We used Pandas, Numpy, Matplotlib, Scikit-Learn, Seaborn, Re, Statsmodels, Plotly libraries, called an API to pull data from the World Bank database and also implemented an ordinal logistic regression model. Further, we utilised Jupyter Notebook, Google Colab and Git to share code.

Implementation process

Following the data cleaning process, we decided to remove 2017-2019 datasets in favour of analysing the 2020-2022 data. This was largely due to the 2020-2022 survey questions and variable coding being similar, allowing for accurate comparison across years.. We then combined the 2020-2022 datasets and conducted all analysis and visualisation as a group.

The analysis of our dataset and visualisation of the results was conducted according to our key questions:

For Question 1, respondents were classified as “Male”, “Female” or “Other”, and the proportion of the non-male population was calculated for each year. A line graph was then plotted.

Question 2 first required a count of the tools which were deemed frequently used, to identify the top three most used tools. Dictionaries were created to include the count of respondents within job titles that reported using these tools and then a percentage was calculated to determine the proportion of the top three job titles as users of the respective tools. Results were visualised as a normalised stacked bar plot for these three job titles only. Second, an additional visualisation was created to show the top three tools used within specific data-related jobs. We identified the three tools that were used for each job title and calculated the percentage that this job title represented, as a portion of total users of each tool. This was then visualised as bar graphs across six subplots.

For Question 3, we performed an initial calculation of the yearly inflation-adjusted compensation with an API. We then used bar charts and box plots to track the salary change across age groups and coding experience. Pie charts were used to visualise gender and continent to better understand the composition of our population sample. Then, for the nominal demographic categories, we plotted stacked bar charts of salary ranges for gender and continent.

Finally, for Question 4, we implemented an ordinal logistic regression model with yearly compensation as the dependent variable and gender, region and job title, team size, company size and year fixed effects as independent variables. One dummy variable from every group of related dummy variables (e.g region dummy variables) was excluded from the model so that it would act as the reference category to avoid the multicollinearity dummy variable trap[1]. A test for multicollinearity was conducted (VIF test) to understand multicollinearity between variables. Following this and a heatmap visualisation, we did not remove any variables from the model as the correlations were not strong enough to warrant this. The proportional odds assumption was also considered as regards model testing, however, no appropriate library is currently available in python to conduct the Brant Test.

Agile development

This project was completed in a highly iterative manner, with our process and methods being reviewed at each step.

1. Concept: The scope of our project shifted significantly within the first five days of receiving the task. We initially aimed to analyse changes in consumer behaviour during COVID-19, however were unable to source publicly available datasets to provide meaningful insights. We landed on our current project after meeting twice to discuss our separate findings.
2. Inception: Each team member took ownership over one dataset (according to survey year) and cleaned their respective data. We implemented a ticketing system on Jamboard to keep track of our tasks, with categories including: “To-Dos”, “In-Progress”, “On-Hold”, “Done”, “Submitted”.
3. Iteration: During the data cleaning phase, we met many times through group calls to discuss challenges faced and often resulted in us pivoting the direction of our project. The most notable example of this was our decision to drop the 2017-2019 data sets.
4. Release: When joining the 2020-2022 data, we encountered many issues in our data cleaning code (see “implementation challenges”). We made many attempts to analyse and visualise our data to best answer the key questions. While we conducted numerous calculations and employed many analytical methods, our final process (see

[1] LearnDataSci. Accessed at: https://www.learndatasci.com/glossary/dummy-variable-trap/#:~:text=machine%20learning%20courses.-,What%20is%20the%20Dummy%20Variable%20Trap%3F,coefficient%20variables%20in%20regression%20models.

"implementation process") was governed by the method which most accurately visualised our results. For example, the graph describing the composition of users of the three most popular tools began as a stacked bar graph containing all job titles, however it now only visualises the top three job titles of which respondents reported using the tools.

5. Maintenance and Retirement: the reporting of our results was conducted at the completion of our project and describes the iterative nature of the overall process.

Implementation challenges

The majority of the challenges faced while conducting this project related to inconsistent data collection and labelling across the survey question sets and responses. For example, in 2020 a question was posed about whether the user frequently uses H20 Driverless AI. However, in 2021 and 2022, "H2O" was stated in the question rather than "H20". This, combined with other wording differences, resulted in difficulties in standardising column names and values when merging the dataset, as our columns were named based on the tools/methods stated in the question.

All data collected is categorical, with a lack of continuous data to analyse. We convert the compensation data from buckets (e.g. 50,000-59,000) into an average value, calculated an inflation adjusted value, and utilised this as a continuous variable.

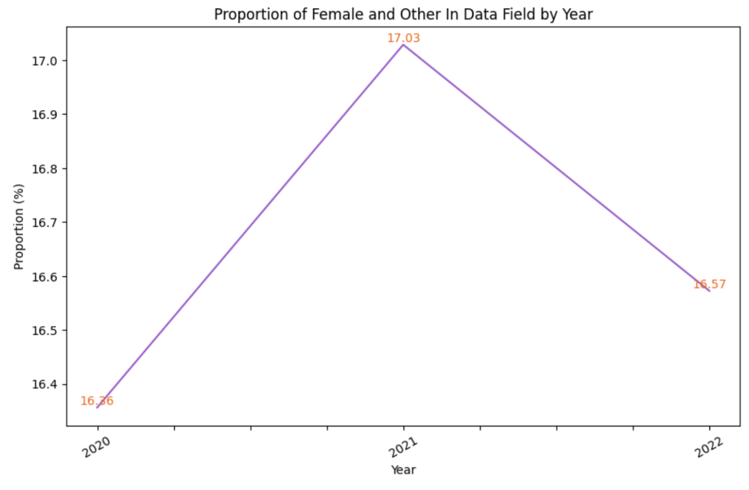
There is also an element of subjectivity in the responses to these survey questions. For frequency of tool usage, the term "regularly" differs in how each individual chooses to define this. As this is merely a consequence of the survey design, this is a consideration we must take into account when drawing insights from our data.

Results

Question 1: Is data science a male dominated field, and is that changing?

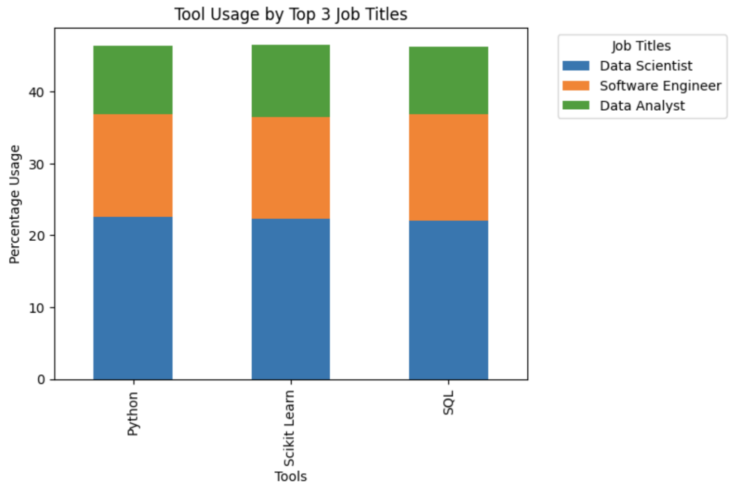
Between 2020 and 2022, we observed a unique trend: an initial increase in non-male participants followed by a decrease.

In 2020, the non-male representation was moderate and we observed a significant growth in 2021. This increase signalled a potentially positive trend to better gender diversity and inclusion in the data industry, particularly within non-technical roles. Multiple factors could have contributed to this rise, including more inclusive hiring practices, broader diversity initiatives, or changes in the societal perception of gender roles in this field.



Graph 1. Line Graph for Gender Ratio Change Over Time

However, the growth trajectory shifted in 2022 when the proportion of non-male participants declined compared to the previous year. This fluctuation is intriguing and prompts further inquiry. The decrease might be attributed to a multitude of factors, potentially involving changes in the job market dynamics, or effects of global events such as the ongoing COVID-19 pandemic, which has been reported to disproportionately impact non-male workers across various industries.



Graph 2. Composition of Three Most Commonly Used Tools, by Top Three Job Titles

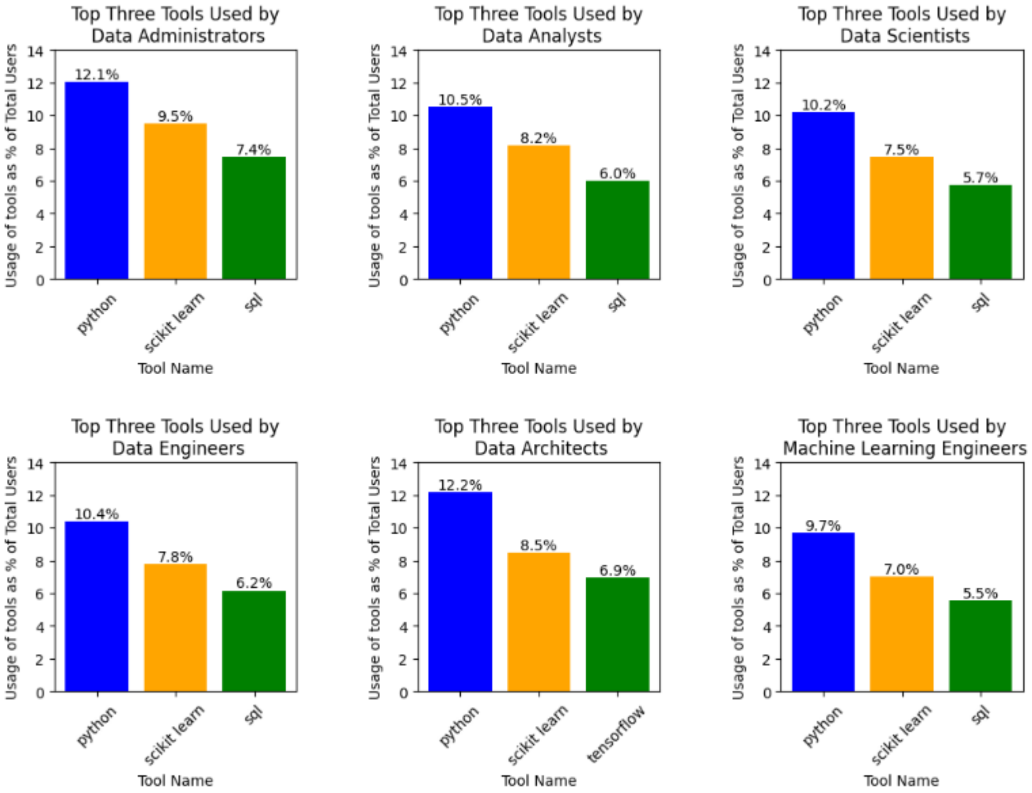
In general, our findings underscore the need for continued efforts towards promoting diversity and inclusion in the data industry, as the proportion of females and others in the data field is less than 18%. A diverse workforce is critical to fostering innovative solutions and it's crucial that we continue monitoring these trends closely and take steps towards achieving a balanced representation in this dynamic field.

Question 2: What skills/tools are needed for each of the different roles?

Of all the tools included in the survey, Python, Scikit Learn and SQL were the three most frequently used tools by Kaggle survey participants. Python was the most used tool, with 27,708 users reporting frequent use, SciKit Learn had

20,422 frequent users and SQL had 15,937. While these tools were used by people from a wide range of job titles, the three most common were “Data Scientists”, “Software Engineers” and “Other” (see Graph 2). Data Scientists represent 22.53% of total Python users, 22.32% of SciKit Learn users and 22.01% of SQL users. Similarly, Software Engineers represent 14.38% of total Python users, 14.21% of SciKit Learn users and 14.90% of SQL users. The similarity of the user composition is likely due to the fact that if a survey participant is a frequent user of one, they are likely to be a frequent user of many tools, including the top 3 listed here.

As our target audience includes prospective data professionals, we narrowed the focus of our analysis to the most commonly used tools among particularly data-focused job titles (Data Administrator, Data Analysts, Data Engineers, Data Architects and Machine Learning Engineers) (see Graph 3).



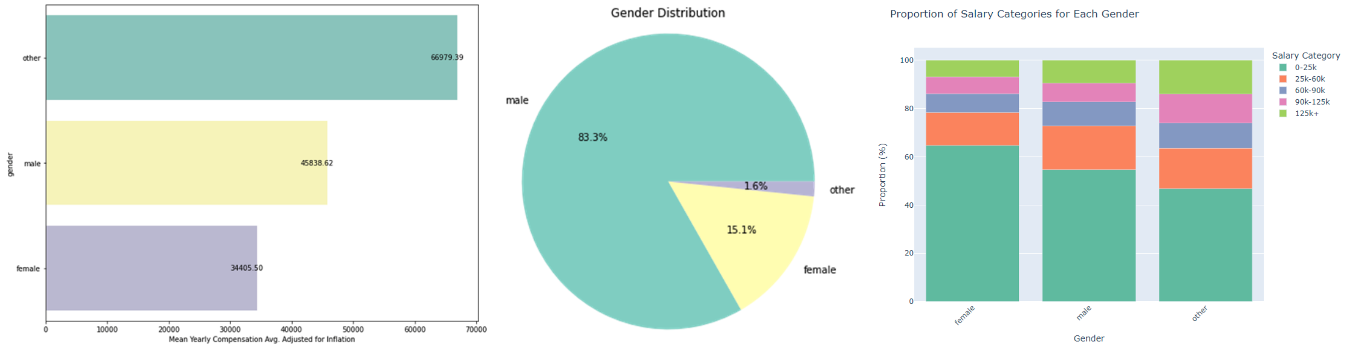
Graph 3: Three Most Frequently Used Tools for Data Professions

Across all six professions, we found that Python and SciKit Learn were the top two most used tools. SQL was the third most frequently used tool across all job titles, except Data Architects where the third most used tool was TensorFlow.

Question 3 - What is the relationship between yearly compensation and other relevant demographic information?

To explore this question, we decided to investigate how the following factors affected the yearly compensation of professionals in the data industry.

Gender:



Graph 4: (1) Yearly Avg. Compensation, (2) Data Population Split and (3) Salary Proportion for each Gender Category

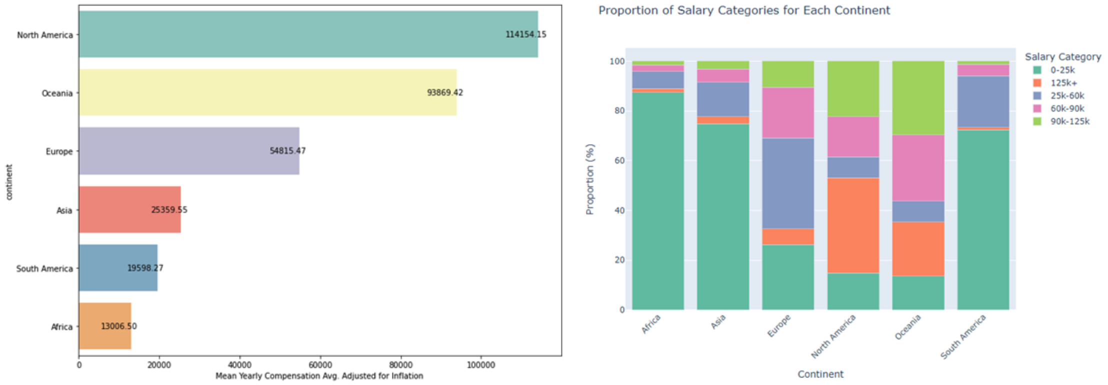
Across genders, we found that the highest average yearly compensation (~67k) corresponded to the gender category of ‘Other’. This category encompasses respondents who identified themselves as non-binary, preferred not to say, preferred to self describe. On the other hand, male respondents reported an average yearly compensation of about 46k, while female respondents had an average yearly compensation of approximately 34.5k.

The pie chart revealed the majority of respondents, constituting approximately 83% identified as male, while female respondents accounted for 15.1% and individuals in the 'other' category represented 1.6% of the respondents. These proportions emphasise the gender imbalance in our dataset, with a significantly higher representation of male respondents compared to females and individuals who identified differently. Furthermore, the disparities in compensation highlight the existing gender pay gap within the data careers surveyed.

The small sample size of the 'other' category in the dataset could lead to skewed results and make it difficult to generalise the factors contributing to their higher compensation. To gain a comprehensive understanding of this, it would be necessary to have a more representative sample that includes a larger proportion of individuals from diverse gender identities.

The final plot shows the proportion of each salary bucket for different genders, revealing that approximately 65% of females earn between 0-25k, indicating a higher concentration in the lower salary range compared to males (55%) and 'Other' (47%). A decreasing proportion of the higher salary buckets are also shown to be higher for males than that for females, further corroborating the hypothesis of gender disparities in compensation.

Region (Continents):

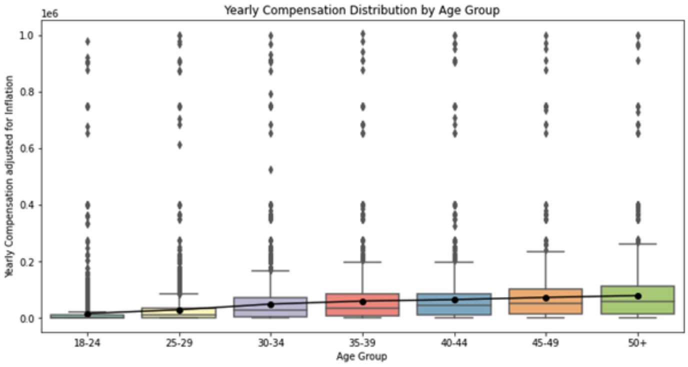


Graph 5: (1) Yearly Avg. Compensation, (2) Data Population Split and (3) Salary Proportion for each Continent

Across continents, North America is shown to be the continent with highest average yearly compensation (~115k) with Oceania following closely behind, with the lowest being South America and Africa, with averages of approximately 20k and 13k respectively. These results are corroborated by the salary proportions for each continent, where the continents with the highest proportion of the lowest compensation range (0-25k) are South America and Africa, each at 72.3% and 87.4% respectively, with these continents also having the lowest proportion of the highest compensation range (125k+), each with 0.97% and 1.5% respectively. While on the other hand, the higher earning continents seen in the first graph (North America and Oceania) follow the opposite trend for the proportion of compensation buckets.

However, when taking into account the population split shown in the pie chart, it might be worth noting that a very low proportion of respondents originate from 'Oceania' raising similar concerns as the ones related to the 'other' category in the gender graphs. These results are indicative of the regional disparities in compensation, highlighting the influence of regional economic factors and job market dynamics on salary levels within the data industry.

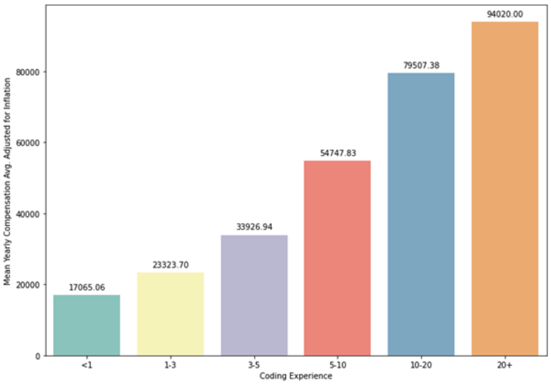
Age:
The distribution of wages has a fairly linear increase with age. This is expected as this is the case for most industries, where wages increase with years of experience, which are also expected to be linearly correlated with age. Furthermore, each compensation range for each age is fairly narrow, with outliers all at the higher end of the compensation range. This is to be expected, as like we have seen previously, significantly higher wages always make the lowest proportion of compensation ranges.



Graph 6: Box Plot of Average Yearly Compensation for Each Age Range

Coding Experience:

The graph below shows an almost exponential increase in average compensation with an increase in years of coding experience. Salaries for people with less than 1 year of coding experience are approximately 17k, with an approximately 6k increase for 1-3 years, a further 10k for 3-5 years, 20k for 5-10 years, 25k for 10-20 years and a smaller increase of 15k for higher than 20+ years of experience. While the 20+ salary increase seems to be somewhat smaller than for 10-20 years, it is expected that this includes people who are retired and overall less respondents in general.



Graph 7: Average Yearly Compensation for Different Coding Experiences

Question 4: What factors lead to higher compensation?

From the VIF test, highest education appeared to have a moderate multicollinearity level. However, the moderate (~0.5) correlations between highest education and team size, and also with company size were too weak to warrant removal from the model.

The model shows that relative to being male, being female decreases the log-odds of being in a higher category of yearly compensation by 0.445, while holding all other variables constant. This finding is significant at the p<0.05 level.

As regards region, compared to being based in Europe, being in Africa, Asia and South America is associated with a decrease of 1.609, 1.151 and 1.444 in the log-odds of being in a higher category of yearly compensation respectively, holding other variables constant. Contrastingly, compared to being based in Europe, being in North America and Oceania is associated with a 2.109 and 1.691 increase in the log-odds of being in a higher category of yearly compensation respectively, holding other variables constant. These findings are significant at the p<.05 level.

As regards job title, compared to being a Data Administrator, being a Data Analyst or a Software Engineer is associated with a 0.244 and 0.256 decrease in log-odds of being in a higher category of yearly compensation, while being a Product Manager, Product/Project Manager , Program/Project Manager or a Data Scientist is associated with an increase of 0.916, 0.648 and 0.368 increase in log-odds of being in a higher category of yearly compensation respectively, holding other variables constant. These findings are significant at the p<0.05 level.

As regards education level, having an additional level of education (‘high school', 'some college/university study without earning a bachelor's degree', 'Professional degree', 'Bachelor's degree', 'Master's degree', 'Professional doctorate', 'Doctoral degree') increases the log-odds of being in a higher category of yearly compensation by 0.093, holding other variables constant. For every additional year spent writing code, the log-odds of being in a higher category of yearly compensation increase by 0.513, holding other variables constant.This finding is significant at the p<0.05 level.

As regards company and team size for, every one unit increase in the company size (moving up the following categories: '0-49', '50-249', '250-999', '1000-9,999', '10,000+') and every one unit increase in the team_size (moving up the following categories: '0', '1-2', '3-4', '5-9', ‘10-14’, ‘15-19’, ‘20+') the log-odds of being in a higher category of yearly compensation increase by 0.157 and 0.134 respectively, holding other variables constant. These findings are significant at the p<0.05 level.

The Nagelkerke's Pseudo R-squared = 0.0, suggesting that our independent variables did not well explain the variance in the target variable.

Limitations

It is important to note that despite the insights generated above, there are some limitations in the results produced.

Our insights should be taken within the context that participants in the Kaggle survey represent a unique sample of the broader Data population. For example, there are a myriad of factors that could influence the gender trends we identified. Future work should look into detailed factors influencing the increase, then decrease, in non-male population, using both qualitative and quantitative methods to get a deeper understanding of the underpinning causes. Additionally, investigating other aspects of diversity beyond gender would provide a more holistic view of inclusivity in the non-technical data field.

Additionally, our data is quite unbalanced and should be considered when applying the results of our analysis to the broader Data population. For example, the largest proportion of survey participants were Data Scientists and Software Engineers which is likely to skew any analysis and predictions that we conduct from this dataset.

While we achieved the objective with our ordinal logistic regression model, we would have liked to develop several models which would additionally include variables (software tools used, industry) to additionally explore how these factors would predict the likelihood of a specific income bracket, holding all other variables constant. However, due to time constraints, we were not able to fully realise this. We also identified that it was important to run a brant test on the regression model to check that the proportional odds assumption is met, however no python package provided this functionality.

Conclusion

From our findings, we can conclude that during 2020-2022 there was an initial increase in non-male participation, followed by subsequent decrease. As such, there is a need for continued efforts towards promoting diversity and inclusion in the data industry, as the proportion of females and others in the data field is less than 17% in all years.

The top three tools used across the industry were Python, SciKit Learn and SQL. This is further reflected across the various Data focused job titles, with the exception of TensorFlow as the third most popular tool for Data Architects.

In relation to average yearly compensation, the “other” category of gender was found to have the highest average yearly compensation at ~$67,000, though it is not possible to conclude that this is representative of the overall data industry as the category ‘Other’ made up less than 2% of our sample. On the other hand, females showed both having the lowest average yearly compensation as well as being the largest proportion of earners in the lowest salary range of 0-25k. Contrastingly, men had both higher averages and higher proportions of higher earnings. This reveals disparities in compensation across genders within the data industry, indicating the presence of inequities and potential challenge in achieving gender pay parity. Survey participants based in North America and Oceania were found to have the highest average yearly compensation, with South America and Africa the lowest. The regional disparities highlight the influence of regional economic factors and differing earning potentials in each region. Age and average yearly compensation are relatively linearly related, though yearly compensation shows a fairly exponential increase with years of coding experience. These conclusions emphasise the need for continued efforts to address gender disparities, promote diversity and inclusion in the data industry, and work towards fair compensation practices. Additionally, the findings highlight the importance of considering regional economic contexts when analysing salary distributions and developing strategies to improve economic opportunities for data professionals worldwide.

Of the variables tested for a relationship with yearly compensation, the log-odds of being in a higher category of compensation bracket was increased for: users based in Europe, North America and Oceania; Product Manager, Product/Project Manager, Program/Project Manager and Data Scientist job titles; increasing levels of education and years writing code; and larger team and company sizes. Conversely, log-odds of being in a higher category of compensation relatively decreased for females and Data Analyst and Software Engineer job titles.

While we were able to answer all the questions posed, the largest challenge we faced was completing this project within the given timeframe. Despite already reducing our datasets to the three most recent years (2020-2023), cleaning and normalising that data still required a large amount of time. Further, it was often difficult to organise meetings with all six team members due to conflicting timetables and resulted in a large amount of written communication to ensure we were all aligned. This meant that our approach to completing our assigned tasks often differed and complicated the data cleaning process.

If we were to repeat this project, it would be beneficial to reduce the scope of the project and be more targeted in our approach to data cleaning and visualisation. More specifically, we spent a large amount of time individually cleaning a separate dataset. In practice, a lot of our code could have been reused across all datasets, rather than having team members write the same code for a different years’ dataset. Further, we would take more time to better understand the data structure to reduce time taken to ensure consistency across all datasets. Lastly, we would also decide on the exact visualisation to plot earlier, allowing for more even distribution of work.

Regardless, we agree that the results generated will be of use to other prospective data professionals and provide insight into the state of the industry currently, as well as what it may look like in the future.