Women in Data Science:
Gender Pay Gap by Kaggle Survey (2017-2021)

Zoe Zhang & Jin Ye

University of Pennsylvania

MSSP 607-002 | Practical Programming for Data Science

Table of Content

Introduction

The field of STEM lacks gender diversity.

According to the Global Gender Gap Report published by the World Economic Forum (2021), Finland, which ranks 2nd in the overall Global Gender Gap Index and has closed 86.1% of its overall gender gap, is still on the bottleneck in improving women's participation in STEM disciplines. The data shows that only 12.4% of women graduates in Finland preferred STEM, compared to 49.8% of men graduates.

We are interested in examining the gender factor in one subcategory of STEM: data science and machine learning. We are especially curious about the gender pay gap.

We plan to use Kaggle's annual Machine Learning and Data Science Survey, conducted from 2017 to 2021, to help us answer the following questions:

- Who is working in data science and machine learning? In other words, what are the demographics of the field?
- Is there a gender gap in participation?
- Is there a gender pay gap? Can we measure it?
- If so, in which country is the pay gap broader?

We hope that by investigating the Kaggle survey, we will be able to have a general idea about people in the industry and to map out the gender pay gap.

Key Findings

We are using static figures in our report. Please kindly visit this Colab notebook for our AWESOME interactive pictures. Notes: this Colab notebook is only for visualization; we have our data cleaning processes uploaded in a .ipynb notebook.

- **Women are generally underrepresented in the data science community, with less than 20% of participation in all data-related jobs.**
- **From 2017 to 2021, the average compensation of women dropped from 91% to 71% of men's, indicating a worsening gender pay gap in the data science field.**
- **The median compensation of women is only 9% of men in 2020, and is now 54% in 2021. In 2017 & 2018, the median wage used to be very close to men and women.**
- **Across all educational levels, women earn less than men, with a median wage ranging from 40% to 55% that of men.**
- **However, the higher the degree is, the higher percentage women take, and the narrower the gap becomes.**
- **Across different professions, women working as developer advocates and researchers earn more than men. Women working as software development engineers only earn a 36.7% median wage compared to their male counterparts.**
- **Among all countries, the higher-earning the job, the fewer women.**
- **Compared by median wage in 2021, women in India earn 27% of that of men, women in the USA earn 85% that of men, women in China earn 71% that of men.**
- **The severity of the gender pay gap does not necessarily follow a developed/ developing country pattern. Sometimes, countries that achieved higher gender equality according to other international indexes have a larger wage gap in the data science community.**
- **Gender may or may not be the most important factor determining yearly income, but the gender pay gap is REAL and is getting worse over the years.**

Data

We are using Kaggle's Machine Learning and Data Science Survey, which was first conducted in 2017, as the data for our study. The industry-wide survey aims to “establish a comprehensive view of the state of data science and machine learning” (Kaggle, 2017). It asks detailed questions about those working in the field, from their age, gender, country of residence, educational background, job title, compensation, to their employer’s industry and size, to the coding languages and machine learning methods they prefer.

From 2017 to 2021, the five annual datasets together depict a profile of those working in the data industry. The survey received over 16,000 responses each year. Combined, there are over 106,000 responses.

Tasks

Sub-Task 1: Examine and Restructure

There are two kinds of structures in the five datasets from Kaggle’s annual survey: 2017-2019 survey follows the same structure while 2020 and 2021 are of another structure.

After looking at the questions in these datasets and cross-examining their differences, we narrowed down the columns we need for our study on the gender pay gap, subsetted them out of the original datasets, and renamed them consistently. The following table shows the original column names.

Renamed	Original Column Names in Annual Datasets				
	2017	**2018**	**2019**	**2020**	**2021**
Gender	GenderSelect	Q1	Q2	Q2	Q2
Country	Country	Q3	Q3	Q3	Q3
Age	Age	Q2	Q1	Q1	Q1
Education	FormalEducation	Q4	Q4	Q4	Q4
JobTitle	CurrentJobTitleSelect	Q6	Q5	Q5	Q5
LearningTime	LearningDataScienceTime	Q8	Q15	Q6	Q6
EmployerIndustry	EmployerIndustry	Q7	** not asked*	** not asked*	Q20
EmployerSize	EmployerSize	** not asked*	Q6	Q20	Q21
CompensationAmount	CompensationAmount	Q9	Q10	Q25	Q25
CompensationCurrency	CompensationCurrency ** only in 2017*	** CompensationAmount clarified USD only*			

The missing questions are filled with NaN values instead. We'll discuss in detail how we dealt with 2017's compensation data in Sub-Task 2. We also added a column named "SurveyYear" to each dataset.

Sub-Task 2: Clean and Merge

Sub-Task 2 is divided into three parts. First, we fix the unique problems found in only some of the datasets. Next, we perform universal adjusting to all five datasets, which is mainly to drop null data and to rename the string values so that they are consistent. Finally, we merge the datasets into one and prepare it for further analysis.

Part 1: Unique Problems

1. **Gender**

   The survey question on gender includes non-binary options, yet choices provided are inconsistent throughout the five years: "Nonbinary", "Prefer to self-describe", "A different identity", "Non-binary, genderqueer, or gender non-conforming", and "Prefer not to say". Apart from inconsistency, they are also of very limited information which bounds us in further exploring. So we decided to drop these non-binary groups.

2. **Age (2017)**

   2017 is the only dataset that age is filled with a number instead of being chosen from a list of ranges. So we had to examine this column, especially the minimum and maximum.

   On the minimum side, we have some geniuses who are 0 years old but already have 3-5 years of coding experience. We also found some 16-year-old students whose entire data looks authentic but should not be working a full-time job. That's for child labor law, we are trying to focus on the gender gap. So we decided to drop all those under 18. Kaggle also noticed and fixed the problem: starting from 2018, age is set as a selection from a list of ranges starting from 18-21 to 70+.

   On the maximum side, we are seeing 10 answers of 100-year-olds. It's harder to tell the credibility of these answers, although one in particular also entered the highest compensation amount: 1.00E+11 (in ILS, 1 ILS is roughly 0.3 USD). We finally decided to keep them except the oldest with the highest income one, which frankly has a lot of zeroes. Because those above 70 only take up a very small percentage in all answers --about 0.4% in 2017 and about 0.5% in all the datasets, to be exact --and some of them may be authentic and valuable data.

   Finally, we arranged the age numbers into ranges in accordance with other datasets.

3. CompensationAmount

The CompensationAmount data varies each year: some are numbers, others are ranges; some ranges are of 5k USD differences, some are of 200k USD differences; some are in various currencies, others are in USD.

Eventually, we adjusted the column CompensationAmount into three new columns: CompensationAmount_mean (the mean of the original survey response range, which is inconsistent across the five datasets, in USD; or, in 2017's case, the given compensation exchanged into USD), CompensationAmount_thin (in thinner slices of ranges unified across the datasets, 2020 data are marked as NaN, in USD) and CompensationAmount_thick (in thicker slices of ranges unified across the datasets, in USD).

a. 2018, 2019, and 2021

2018, 2019, and 2021 have quality data that are put in similar thinner ranges. So we simply adjusted them to the same standards.

The adjusted ranges for CompensationAmount_thin are (in USD): <10000, 10000-20000, 20000-30000, 30000-40000, 40000-50000, 50000-60000, 60000-70000, 70000-80000, 80000-90000, 90000-100000, 100000-125000, 125000-150000, 150000-200000, 200000-250000, 250000-300000, 300000-500000, 500000+.

The adjusted ranges for CompensationAmount_thick are (in USD): <10000, 10000-100000, 100000+.

b. 2020

The 2020 survey on CompensationAmount is initially to be chosen from very broad ranges. We adjusted the data into the same ranges as 2018, 2019, and 2021 for CompensationAmount_thick. We fill it with NaN for CompensationAmount_thin.

c. 2017, a very messed-up year

2017 data on CompensationAmount were entered as numbers and also in various currencies (the CompensationCurrency column) while all the other years have data in USD ranges.

Step one, we get rid of non-numbers entered by listing out all unique characters in the answers. Then we inspected all the rows with these weird characters (.E+-) and replaced those that do not make sense with null values: negative income, a simple “-”, and the 100-year-old with the highest income of "1.00E+11" which we decided it’s junk survey results.

Step two, we looked at the CompensationCurrency column and found 87 kinds of currencies (all listed as a 3-character abbreviation like USD). We built up a new column named CompensationAmount_exchangerate to store their average exchange rate to USD in 2017. With website scraping, we were able to fetch the average exchange rate of 74 currencies from this site (taking AUD as an example): https://www.exchangerates.org.uk/AUD-USD-spot-exchange-rates-history-2017.html. For the remaining 13 currencies including USD (the exchange rate is 1) that cannot be acquired from the website, we manually looked up the average exchange rate and filled in the column. We finally calculated the 2017 USD equivalent, stored them in column Compensation_USD, and renamed the columns.

Step three, we arranged the USD numbers into ranges in accordance with CompensationAmount_thick and CompensationAmount_thin.

Part 2: Universal Adjusting

With unique problems fixed, we now move on to performing universal adjusting to all five datasets by creating a function called universal_cleaning. The function achieves the following goals:

1. Replacing answers like “I prefer not to answer” and "I don't know" with null values.

2. Renaming the string values so that they are consistent across the datasets. For example, in the EmployerIndustry column, some of the datasets have a category named "CRM/Marketing" while others have a "Marketing/CRM" category, etc. We selected our standard and renamed all values accordingly.

3. Renaming the string values so that they are fit for further processing. We plan to use GeoPandas to map the gender pay gap, which is why the Country names need to be adjusted to GeoPandas standards. For example, “United States of America” instead of “US”; no individual “Hong Kong” but merged into “China”, etc.

4. Adjusting overlapping in the ranges. For example in the LearningTime column, some datasets have “3-4” and “4-5” while others only have “3-5”. We adjusted them so that they don’t overlap with each other.

5. Creating a new column named JobTitle_Broad which organizes the somewhat messy JobTitle into fewer categories.

6. Finally, we dropped the rows with specific or null values in the following columns:
    a. Gender: null (non-binary already dropped in Part 1 of Sub-Task 2)
    b. Country: I do not wish to disclose my location, null
    c. JobTitle and EmployerIndustry: Student, unemployed

Part 3: Merge & Final Adjust

Finally, after all five datasets went through the universal_cleaning function, we merged them into one. We dropped the original forms of some adjusted columns.

Columns and Definition

After the restructuring and cleaning in the first two sub-tasks, we now have the following 16 columns:

1. **Gender**: Male or Female. Non-binary and null dropped.

2. **Country**: 68 countries and territories, including one named “Other”, null dropped. If there are less than 50 respondents that came from the same country or territory in the same year, Kaggle will group them as “Other” for anonymity. Although we can’t locate it on the world map, we still think it would be interesting to look at how these “tech minorities” are doing. All the “Other” in the five years combined together makes the third-largest population cluster, right after India and the United States.

3. **Age**: grouped into 11 ranges, null dropped: 18-21, 22-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59, 60-69, 70+.

4. **Age_mean**: for 2017 which age is entered as a number, this is the original age; for the rest of the years which age is selected from a list of ranges (inconsistent across the four years), this is the mean of the selected range (for example, range 22-24 will be logged as 23). 69 unique values.

5. **Education**: 7 unique values and null: No formal education past high school, Some college/university study without earning a bachelor's degree, Bachelor's degree, Master's degree, Doctoral degree, Professional degree, Professional doctorate, null.

6. **LearningTime_adjusted**: 8 unique values/ranges (in years) and null: 0, 0-1, 1-3, 3-5, 5-10, 10-20, 20-30, 30+, null.

7. **LearningTime_mean**: 17 unique values (in years) and null. The mean of the selected learning time range. Because the original ranges vary among the datasets (some might have 3-4 and 4-5 while others only have 3-5) and are merged into LearningTime_adjusted, so LearningTime_mean has more unique values than LearningTime_adjusted.

8. **SurveyYear**: 2017, 2018, 2019, 2020, 2021.

9. **CompensationAmount_mean**: for 2017 which compensation is entered as a number and a currency code, this is the calculated compensation in USD; for the rest of the years which compensation is selected from a list of ranges (inconsistent across the four years), this is the mean of the selected range (for example, range 10-20k will be logged as 15000). 1537 unique values and null.

10. **CompensationAmount_thin**: 17 unique values (in USD) and null. Adjusted into thinner slices of ranges unified across the datasets: <10000, 10000-20000, 20000-30000, 30000-40000, 40000-50000, 50000-60000, 60000-70000, 70000-80000, 80000-90000, 90000-100000, 100000-125000, 125000-150000, 150000-200000, 200000-250000, 250000-300000, 300000-500000, 500000+. All 2020 data are null.

11. **CompensationAmount_thick**: 3 unique values (in USD) and null. Adjusted into thicker slices of ranges unified across the datasets: <10000, 10000-100000, 100000+.

12. **JobTitle**: 28 unique values and null: Operations Research Practitioner, Computer Scientist, Data Scientist, Software Developer/Software Engineer, Business Analyst, Engineer, DBA/Database Engineer, Scientist/Researcher, Other, Data Analyst, Machine Learning Engineer, Statistician, Predictive Modeler, Programmer, Data Miner, Consultant, Software Engineer, Research Assistant, Chief Officer, Manager, Data Engineer, Developer Advocate, Marketing Analyst, Program/Product/Project Manager, Principal Investigator, Salesperson, Data Journalist, Developer Relations/Advocacy, null.

13. **JobTitle_Broad**: organizing JobTitle into fewer categories. 14 unique values and null: Computer Scientist, Data Scientist, Software Development Engineer, Business Analyst, Data Engineer, Other, Data Analyst, Machine Learning Engineer, Statistician, Consultant, Researcher, Chief Officer, Program/Product/Project Manager, Developer Advocate, null.

14. **EmployerSize_adjusted**: employer's employee number grouped into 3 ranges and null: <1000, 1000-9999, 10000+, null.

15. **EmployerSize_mean**: 12 unique values and null. The mean of the selected employer size range. Because the original ranges vary among the datasets and are merged into EmployerSize_adjusted, so EmployerSize_mean has more unique values than EmployerSize_adjusted.

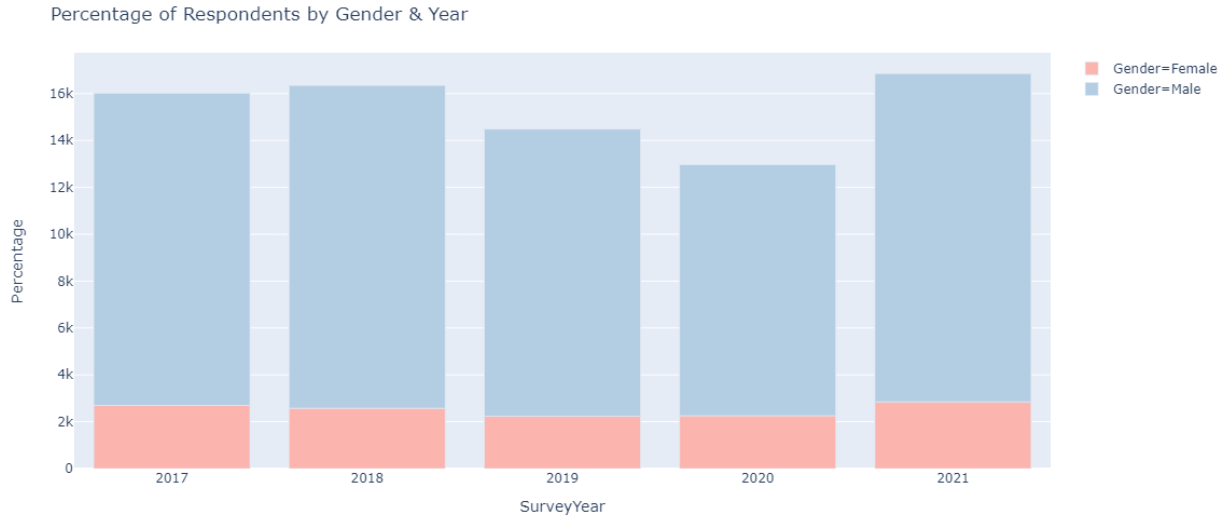
16. **EmployerIndustry**: 20 unique values and null representing the employer's industry: Mix of fields, Computers/Technology, Academics/Education, Government/Public Service, Non-profit/Service, Internet-based Business/Services, Financial, Retail/Sales, Telecommunications, Medical/Pharmaceutical, Military/Security/Defense, Marketing/CRM, Manufacturing/Fabrication, Hospitality/Entertainment/Sports, Insurance/Risk Assessment, Other, Energy/Mining, Broadcasting/Communications, Accounting/Finance, Shipping/Transportation', null.

Sub-Task 3: Who is working with data?

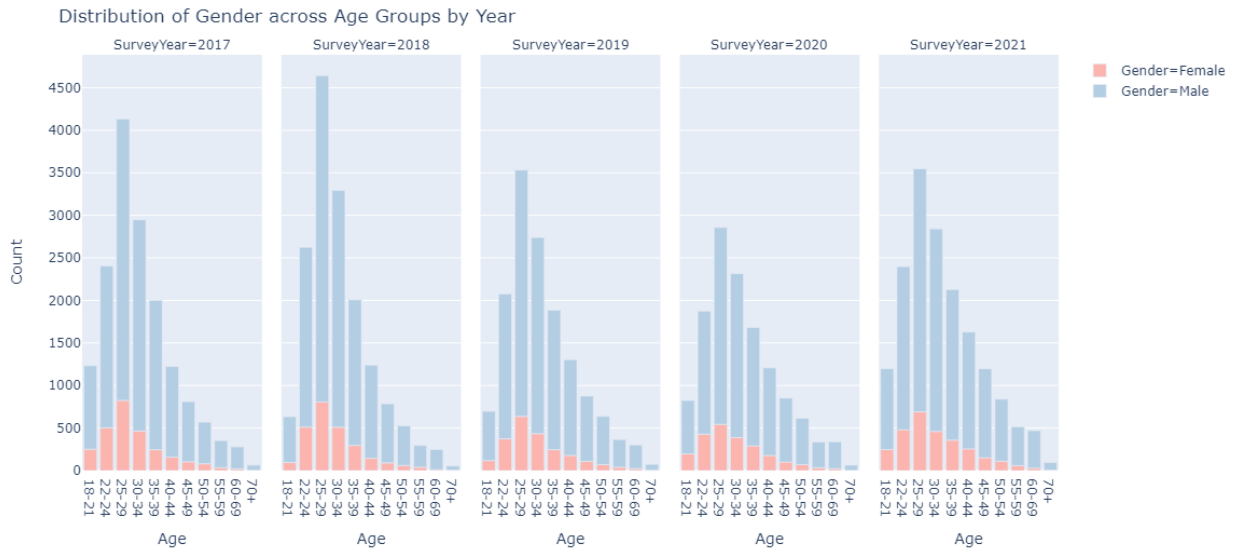
We now proceed to our analyses. Sub-Task 3 visualizes the descriptive statistics of those who responded to the survey and tries in doing so, to present a general picture of the people working in the machine learning and data science industry.

Gender, Age, and Country

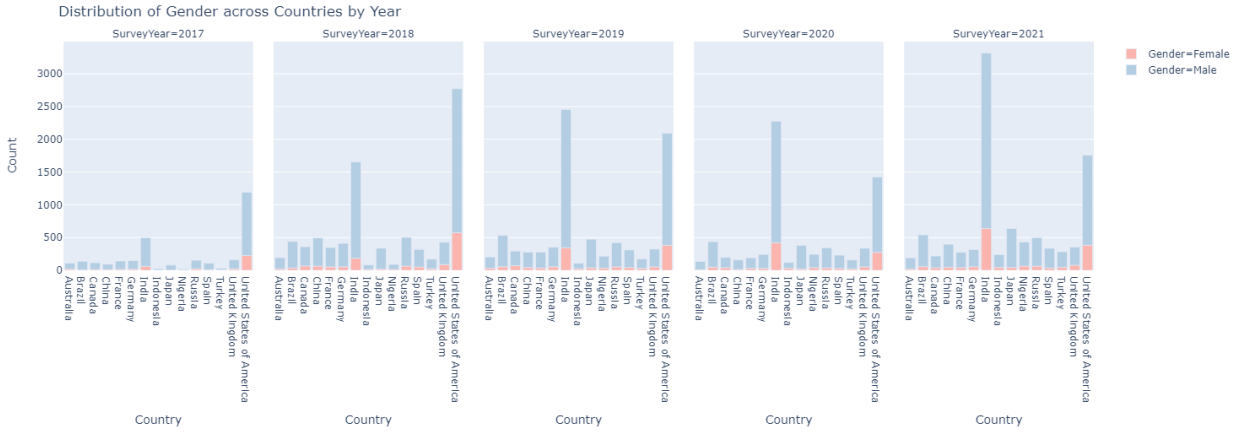
The Kaggle survey datasets gathered 106,301 responses over the five years. For each year, the survey received over 16,000 responses. After our data cleaning, roughly 75% remained for further analyses. The following figure shows the number of respondents and composition by gender after our cleaning. The percentage of female respondents falls between 15%-18% over the five years.



There's a clear peak in the distribution of age groups. For both genders and the overall number, the age group 25-29 has the largest number of respondents, followed by 30-34 and 22-24. By a closer examination, we can see women taking up a larger percentage in the age group 22-24 than in 30-34 –and there is an overall trend of more women in younger generations.



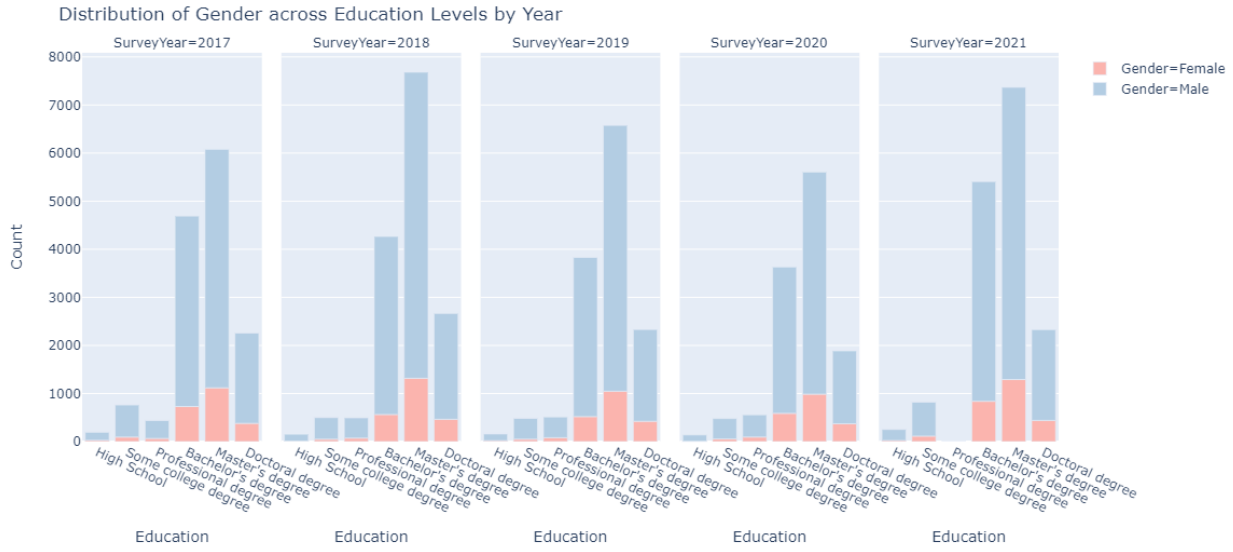
The following figure shows the number of female and male respondents from the top 15 countries with the highest number of female respondents. Of the two largest countries, 20.5% of all US respondents over the five years are women; 17.0% of all Indian respondents over the five years are women.



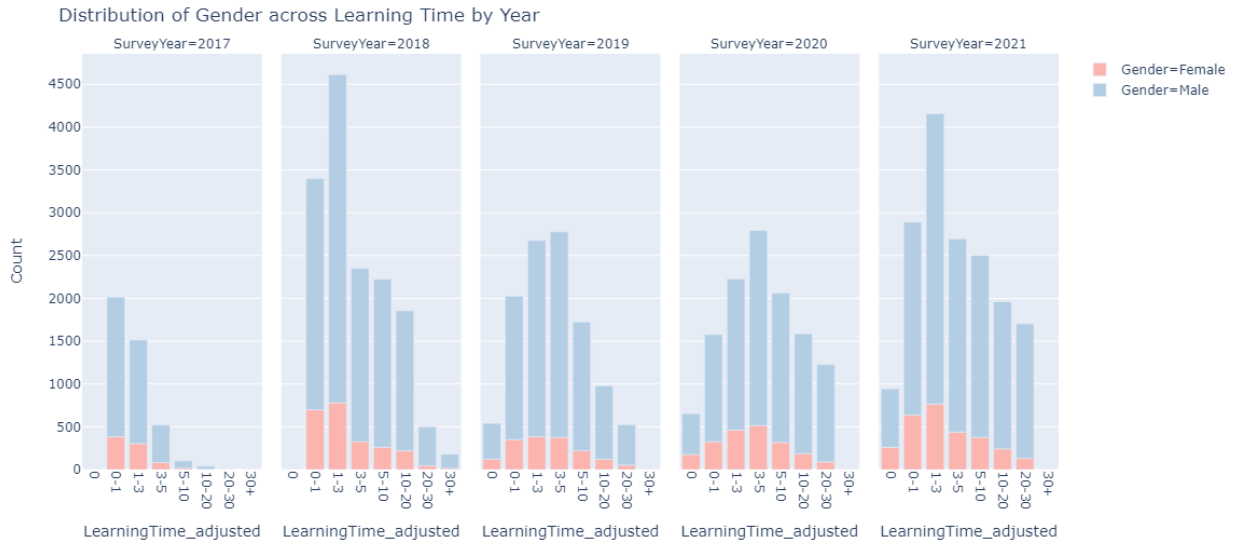
Educational Backgrounds, Job Title, and Employer

Most of the survey respondents have earned at least a Bachelor's degree, with the majority always being those with a Master's degree. Some have gained a professional degree or doctorate. Very few only graduated high school.

Interestingly, in the three most common degrees: Bachelor's, Master's, and Doctoral, the trend over the five years is that: the higher the degree, the more women.

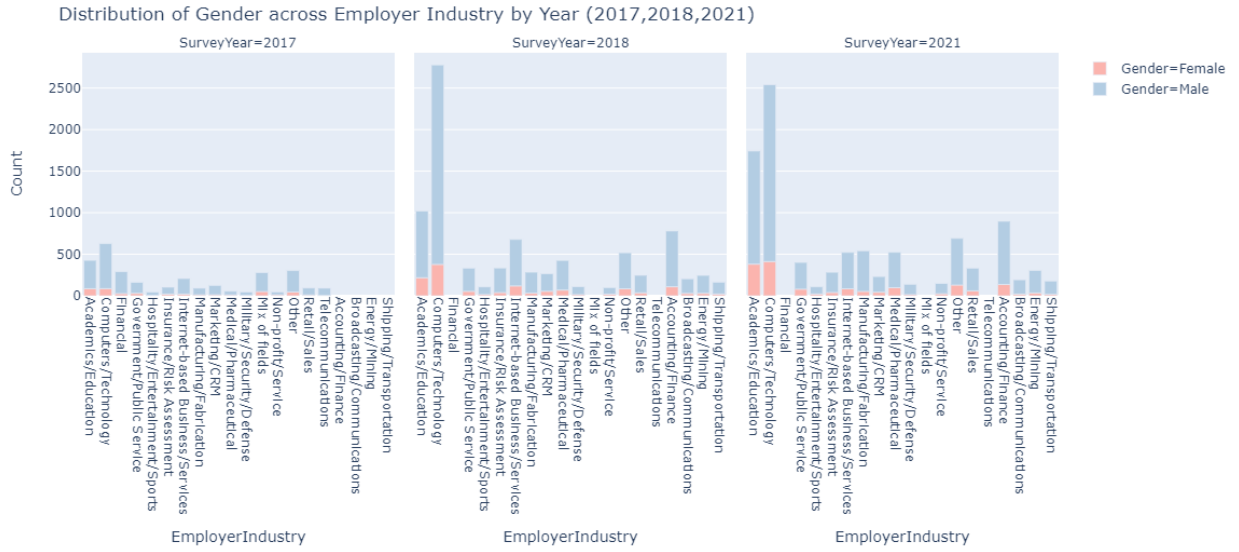


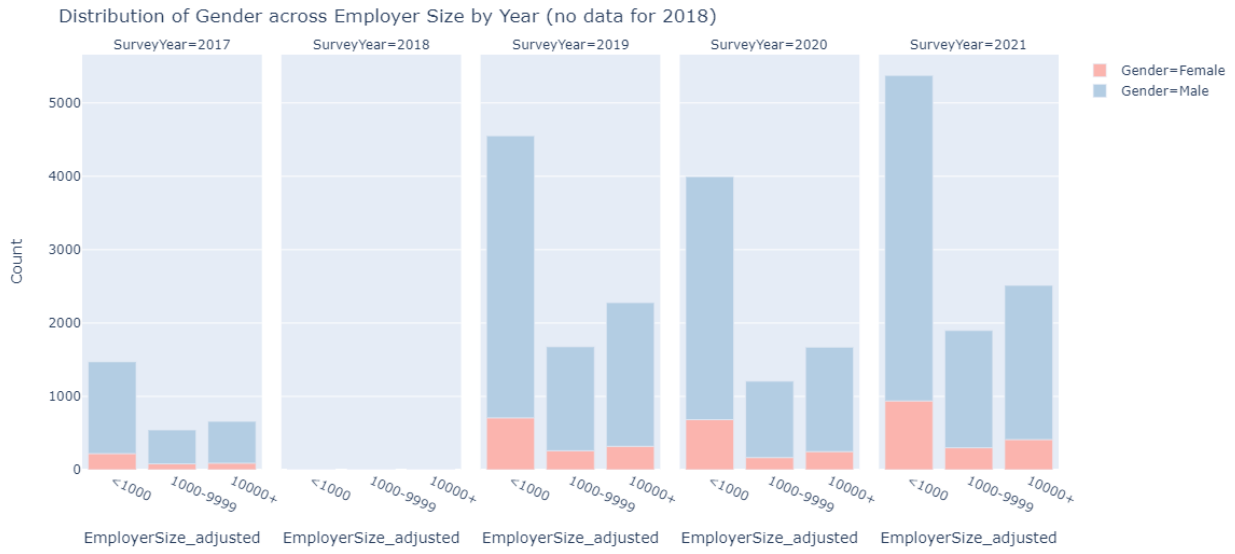
As for the distribution of learning time, the peak for both genders is in either 1-3 or 3-5 years.



In the 14 kinds of job titles, most of the respondents are data scientists, followed by software development engineers. The percentage of women in data analysts is noticeably high.

The next two figures are about the employers. Most survey respondents work in computation/technology, followed by academic/education. When it comes to employer size, it's a somewhat polarized distribution, medium-sized employers (1,000-9,999) are the least. In the four years that have employer size data, the percentages of women in smaller employers (less than 1,000) are always higher than larger employers (more than 10,000).





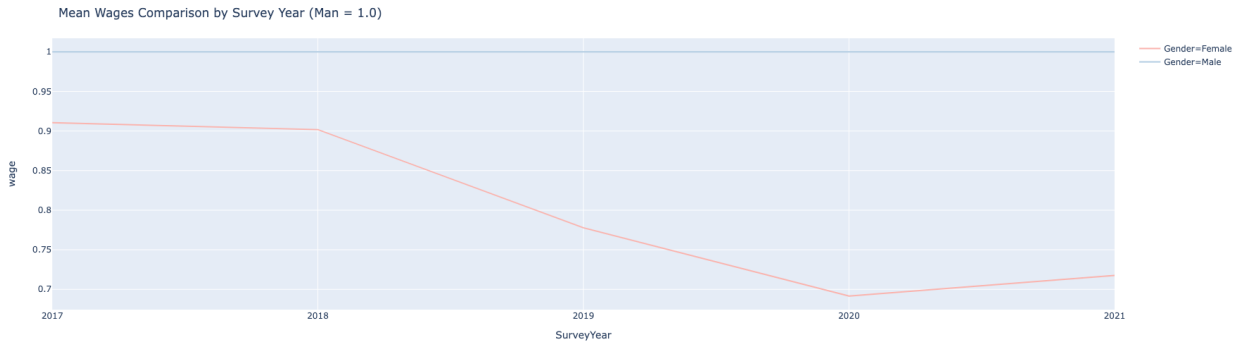
Sub-Task 4: Investigating the Gender Pay Gap

Sub-Task 4 brings us to the variable we want to focus on: CompensationAmount. We will examine the gender pay gay in intersection with other factors. We'll discuss the country factor in Sub-Task 5 and 6.

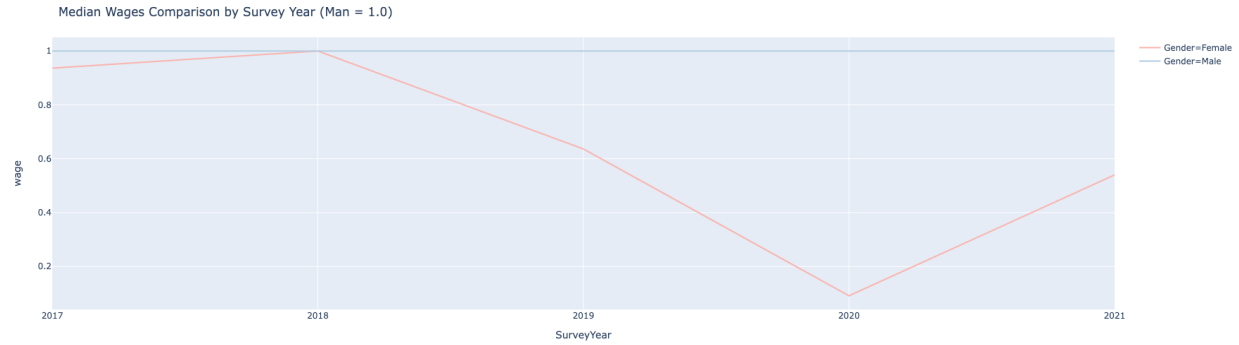
It should be noted that not all qualified respondents answered the compensation question. For 2017, only 26.7% of all respondents answered the income question, among which women are also less represented as the answer rate is 21.8% compared to men's 27.8%. But the overall answer rate has risen since 2017, with a 74.2% answer rate in 2018, 84.9% in 2019, 80.2% in 2020, and 89.5% in 2021. Therefore, compensation information should be more accurate from 2018 to 2021. However, in all these four years, the answer rates of women are all lower than that of men (with a lag from 3% to 9.2%), which might impact data accuracy.

Survey Year

From 2017 to 2021, the average compensation of women dropped from 91% to 71% of men's, indicating a worsening gender pay gap in the data science field.



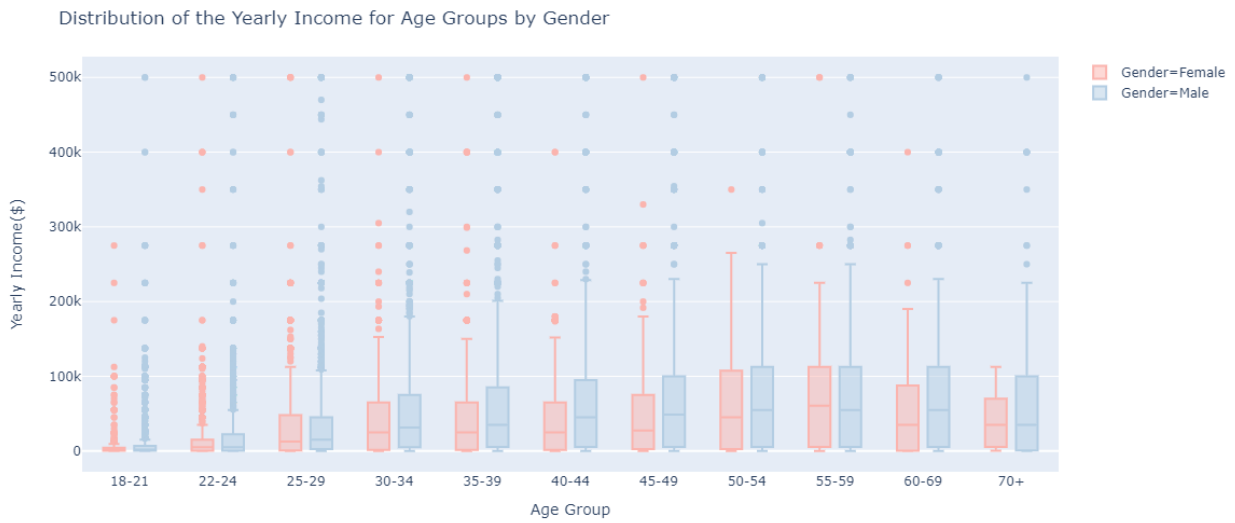
The median compensation of women is only 9% of men in 2020, and is now 54% in 2021. In 2017 & 2018, the median wage used to be very close to men and women. The extreme low median wage of women may be explained by covid-19, which caused a lot of unemployment all over the world and troubles of child care which is primarily done by women.



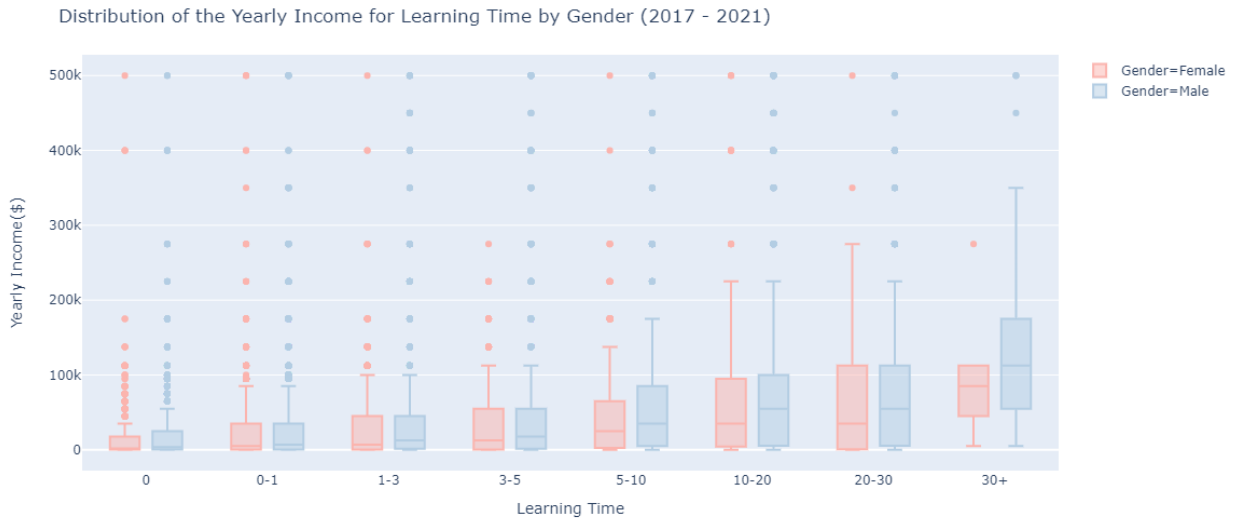
Age and Experience

In all age groups except 55-59 (about 2.4% are in this age group), the median female compensation is lower than that of male.

The median male compensation rises from the youth, keeps growing, and hits the plateau in 50-69 at 55k USD, and decreases after 70 (although the population in that age group is very few). The median female compensation hits the plateau much earlier and much lower: in 30-44 at 25k USD. From age 45 and above, the median female compensation rises again but keep in mind that the number of older female respondents is very few.

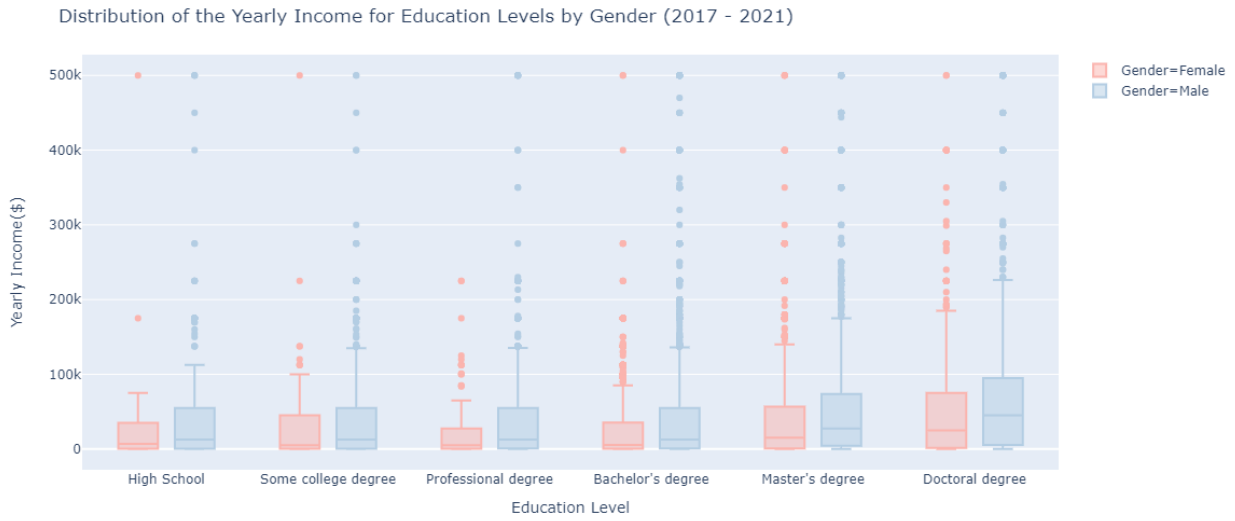


Experience perhaps shows the growth in compensation better than age.



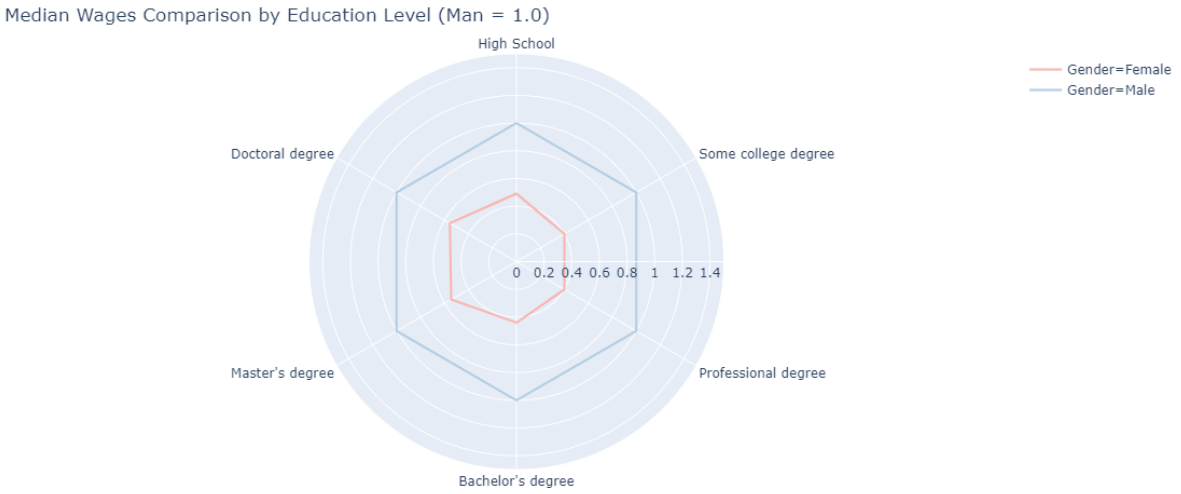
Education

Higher educational backgrounds come with high compensation –that is for the group median. Yet if we look at the gender pay gap, the median income of a male high school graduate is more than twice the median income of a female with a Bachelor's degree.



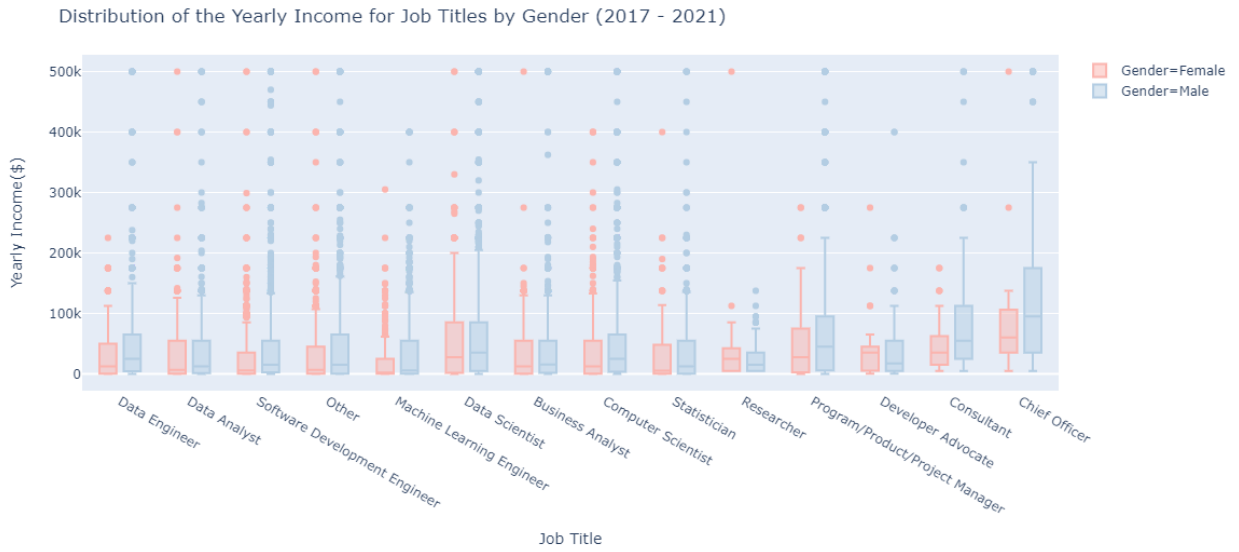
The following figure perhaps shows the gender pay gap by education level better. Women, regardless of educational background, earn less than their male peers.

We've discussed in Sub-Task-3 that the trend in the three most common degrees (Bachelor's, Master's, and Doctoral) is: the higher the degree, the more women. We can now add to the trend: the higher the degree, the narrower the gap.

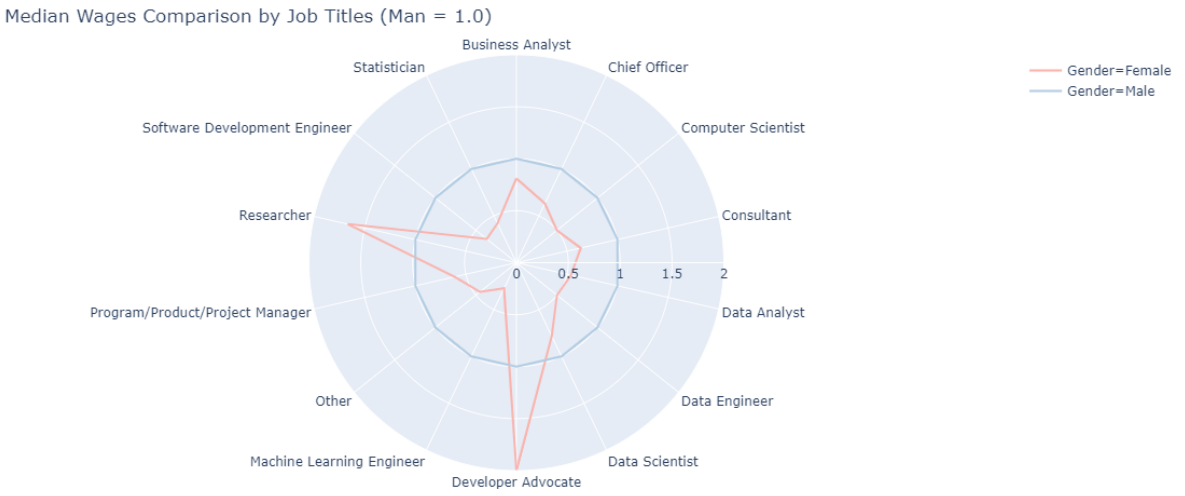


Job Titles

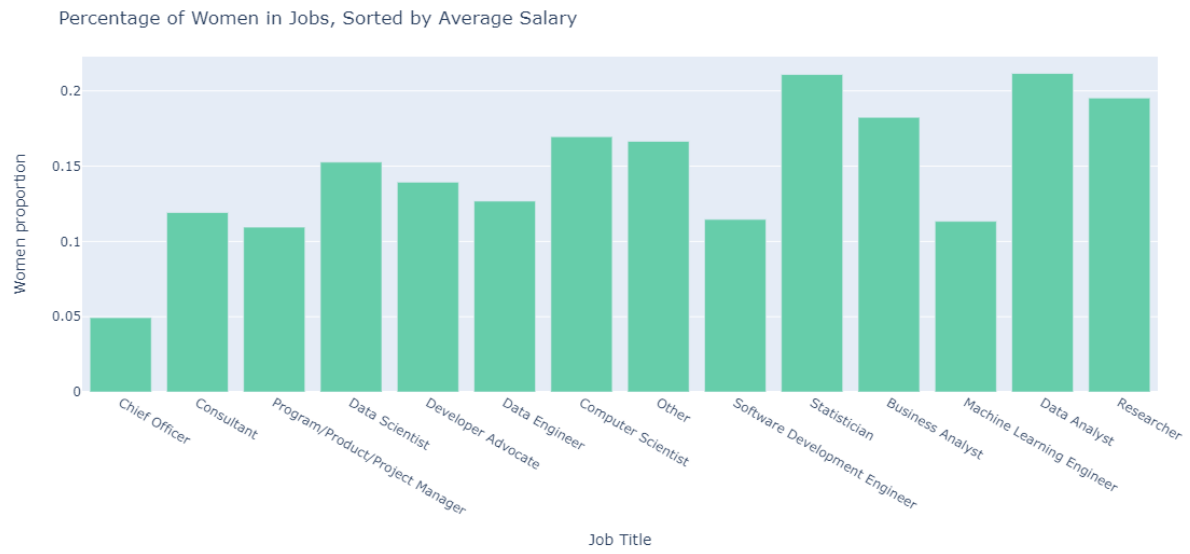
Next, we have pay gaps under different job titles. Chief Officer and Consultant are exceptionally well-paid jobs.



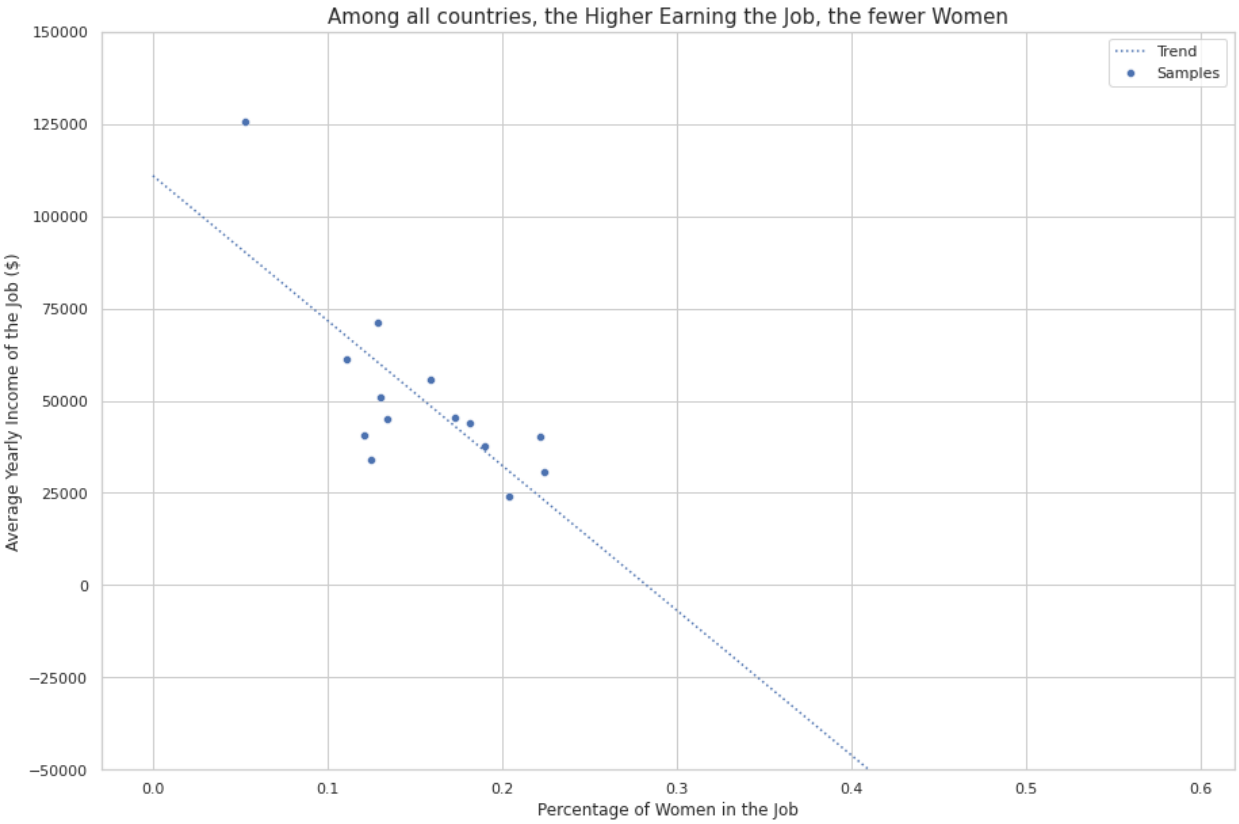
We again bring out the median wage comparison figure for a better look. Only in Researcher and Developer Advocate is the female median wage higher. Only about 0.66% of all respondents are researchers; about 0.27% of all respondents are developer advocates.



When we rank all the jobs by their average compensation (higher compensation on the left side), there's a vague trend that in general, better-paid jobs have fewer women.



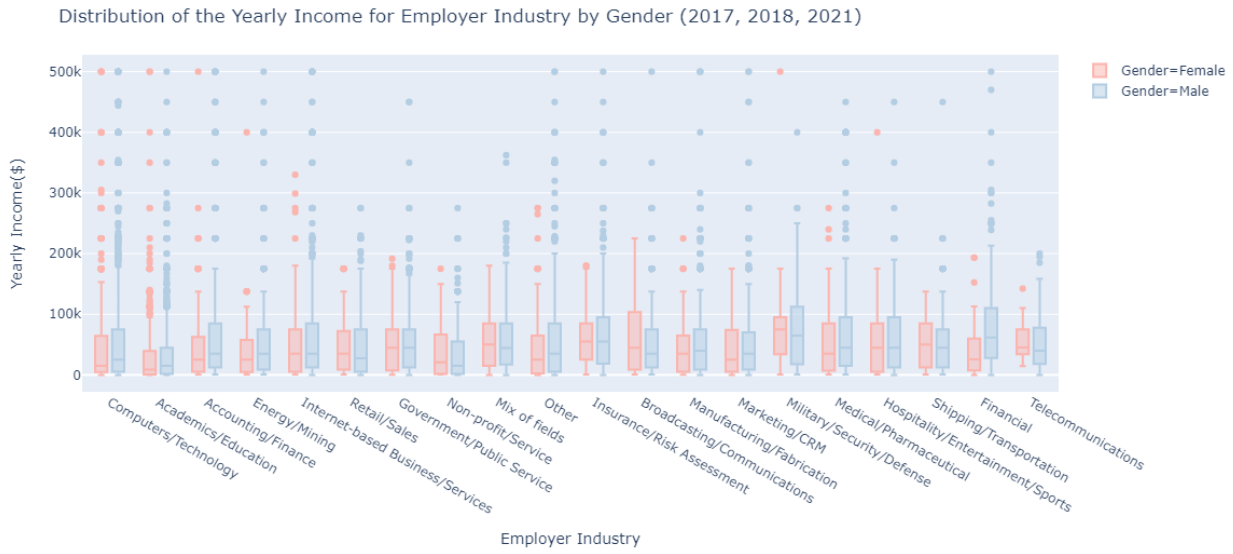
When we run regression between the average compensation of the jobs and the percentage of women in the jobs, this vague trend becomes obvious: the higher-earning the job, the fewer women.



Employer

Lastly in this Sub-Task, on the industry and size of employers.

We are seeing several industries where the median female compensation is higher than male. Some of these reverse gaps are pretty surprising, like in Broadcasting/Communications, it's 75k (female median) to 45k (male median). But when it comes to Financial, the gap is again back to a stunning 76k (male median) to 22k (female median).



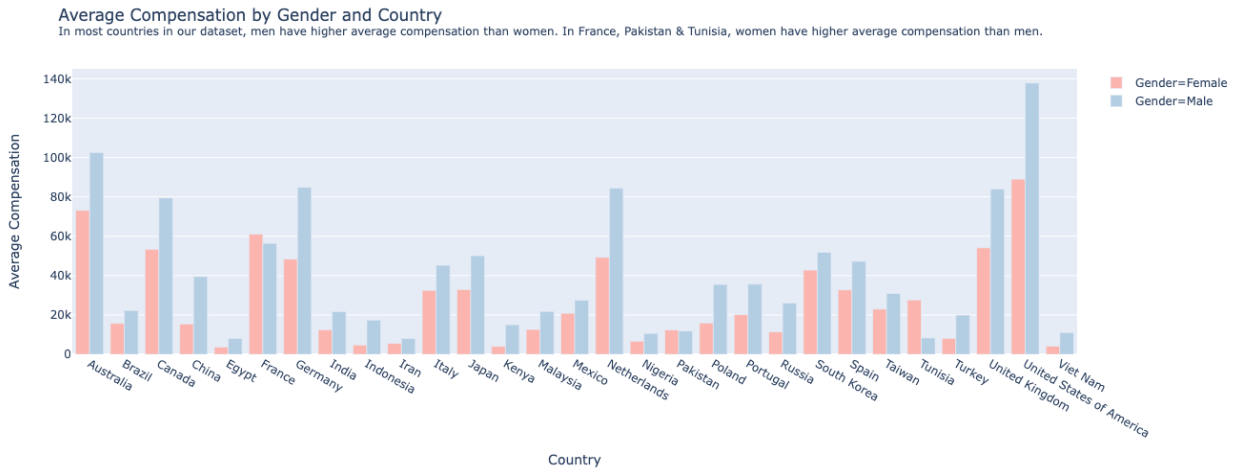
The median male compensation rises as the employer size increases. But the highest median female compensation is in medium-sized employers (1,000-9,999), which has the lowest number of respondents of the three kinds.

Sub-Task 5: Examining the Gender Pay Gap by Country

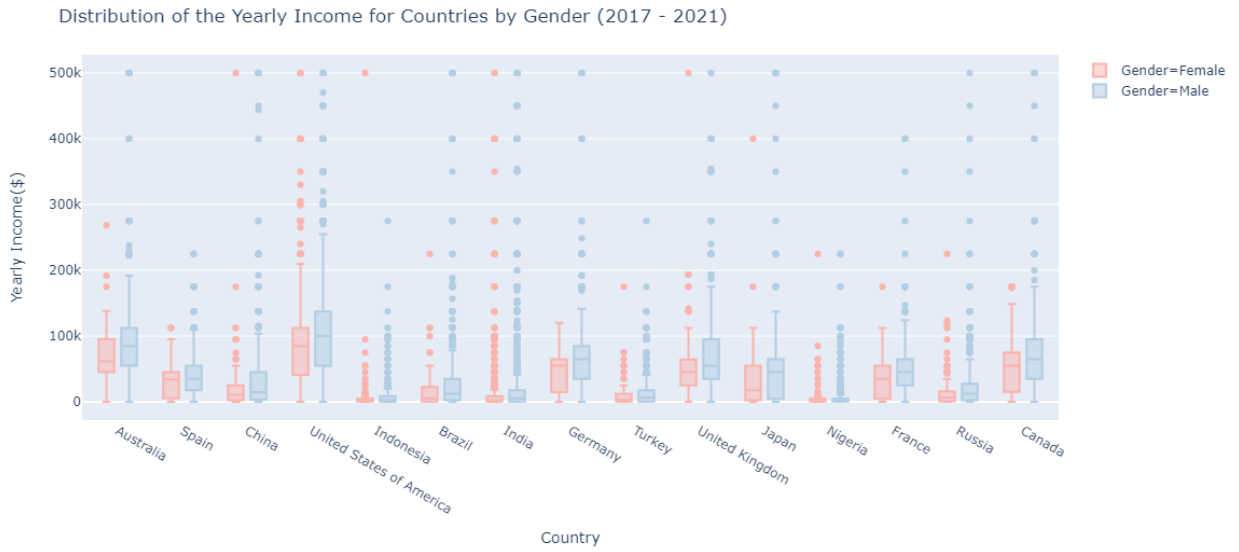
In this task, we use the 2021 dataset as an example to explore the pay gap between men and women's yearly income by country.

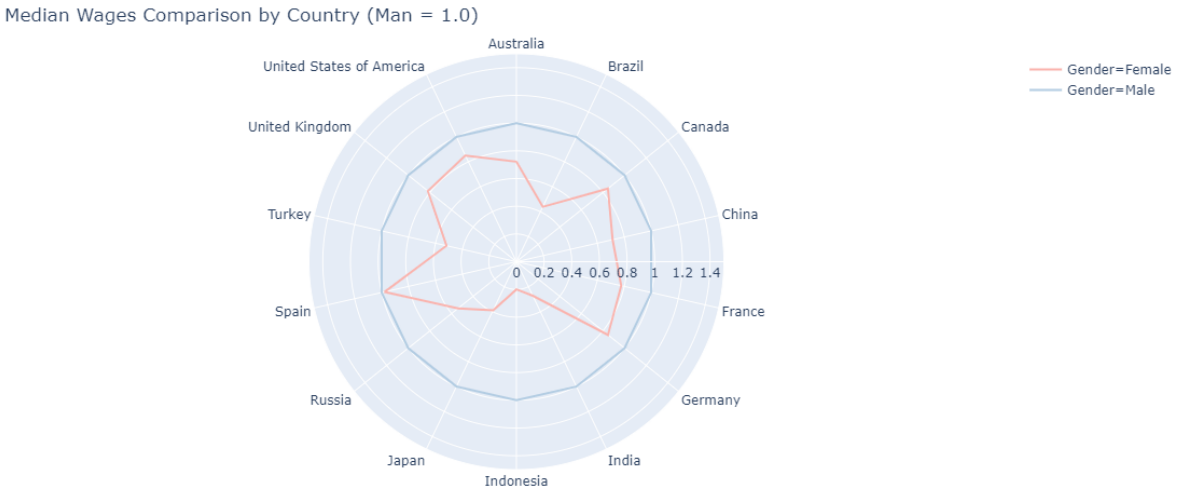
Firstly, we choose to drop countries which have fewer than 20 female respondents as the female respondents are too few to make the data representative (which also reflects how male-dominated the industry is). We also choose to drop countries marked as "other" even though there are a total of more than 600 respondents being categorized in this column. This categorization is made by the survey maker to cluster countries that have fewer than 60 respondents, and is not of much help to picture them here in this task. The cleaned dataset has 29 countries in total, mostly in Asia, America and Europe.

We compare the average compensation between the two genders. In most countries in our dataset, men have higher average compensation than women. In France, Pakistan & Tunisia, women have higher average compensation than men by 9%, 4.5% and 231.5%. The data in Tunisia looks really abnormal in this case, so we double checked that the total respondents from the country are only 62, with 26 female and 36 male. Therefore, the data may be inaccurate due to the insufficient sample size. Still, in most countries, women only earn 40%~80% that of men on average. Even in countries that are renowned for gender equality, women in tech only make 67% of men's salary in Canada, 57% in Germany, 58% in the Netherlands in the survey.



The two following figures show the gender pay gaps in the top 15 countries with the highest number of female respondents, both in the distribution of compensation and in the median comparison.



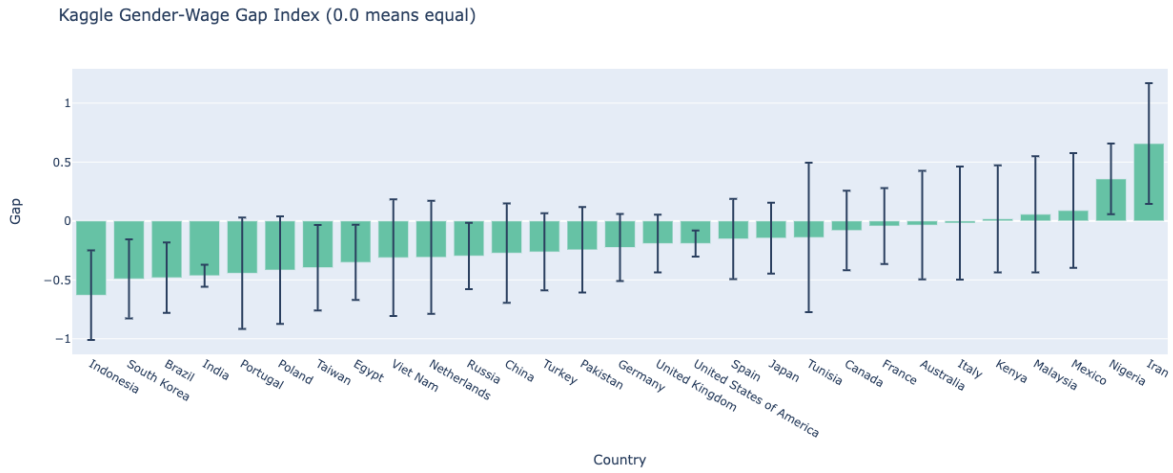


As we can see here, even though the regression results may or may not be statistically significant in some countries, the median wage gap is real for different genders. Apart from Spain, women in the other 14 countries are all receiving lower wages than men, with the worst of only 20% of men in Indonesia. In the USA, women earn 80% of what men do. In India, women earn 30% of what men do. In China, women earn 70% of what men do. No country has witnessed a reverse median salary comparison.

To better compare the data, we ran a regression on the gender pay gap among the countries that have more than 20 female respondents. The dataset has a total of 13,577 qualified respondents from 29 countries. As we want to see how the gender pay gap is deviant from the median wage (average wage can sometimes be very deceiving for too high or too low values), we centered the compensation to the median value to avoid multicollinearity. The controlled

variables are age, education level, country, job title(profession), learning time(experience), employer industry and employer size.

The bar and error bar chart above shows the estimated “gender-wage gap”, and we call it “Kaggle Gender-Wage Gap Index”.The bar chart indicates the gender pay gap. A negative value means that women earn lower wages than men by the coefficient times 100 percent. Error bars show the estimation error (1.645 times standard deviations, 90% confidence level). If error bars cross zero, that means it is not statistically significant.



Overall is estimated 0.3-ish and statistically significant. This indicates the existence of a global level gender wage gap. Based on the result, women's wages are lower by around 30% of median wage than men after controlling for other factors such as education, experience, job title, and industry.

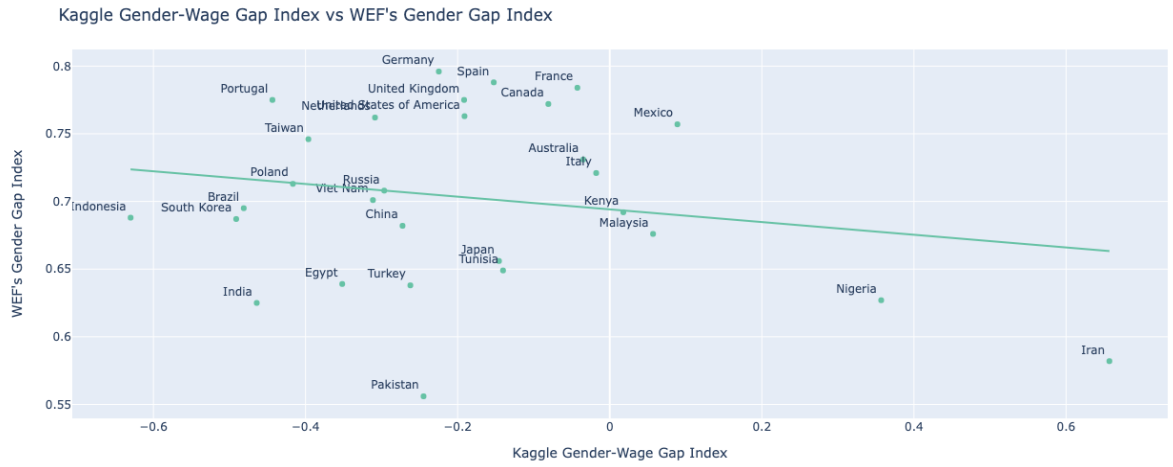
Indonesia, South Korea, Brazil and India have the largest wage gaps and these gaps are statistically significant. Women's wages are lower than men's by 50% of the median wage. Wage gaps in Taiwan, Egypt, Russia and the United States are also statistically significant, ranging from 20% to 40%.

Portugal, Poland, Vietnam, Netherlands, China, Turkey, Pakistan, Germany, United Kingdom, Spain, Japan, Tunisia, Canada, France, Australia & Italy have wage gaps as well but the gaps are not statistically significant. So, it cannot be said that the gender gaps are systematic in these countries.

Kenya, Malaysia, Mexico, Nigeria & Iran do not have a wage gap, but the results are also far from being statistically significant.

In general, the gender pay gap may or may not be statistically significant in some countries, but the gap exists and holds true in several major countries that have booming tech industries.

Finally, we compare the Kaggle Gender-Wage Gap Index with the World Economic Forum's Global Gender Gap Index[1]. WEF's Index “ranks countries according to a calculated gender gap between women and men in four key areas: health, education, economy, and politics to gauge the state of gender equality in a country.” Therefore, this index is a comprehensive gender gap index while the Kaggle Index is specialized in the data science field with a sole focus on the pay gap. The comparison is somewhat rough as the WEF index counts women’s educational attainment, political empowerment, and health and survival as weighted values while the Kaggle index only considers the economic part. But still, it may give us some insights.



What is very intriguing is that those two indexes are slightly negatively correlated. This means the country which achieved higher gender equality overall has a larger wage gap in the data science community. Again, some indexes have no statistical significance in terms of the gender pay gap and do not take other gender equality aspects into consideration. But it may shed some light on reflecting how the most “promising” industry may grow into another unbalanced situation.

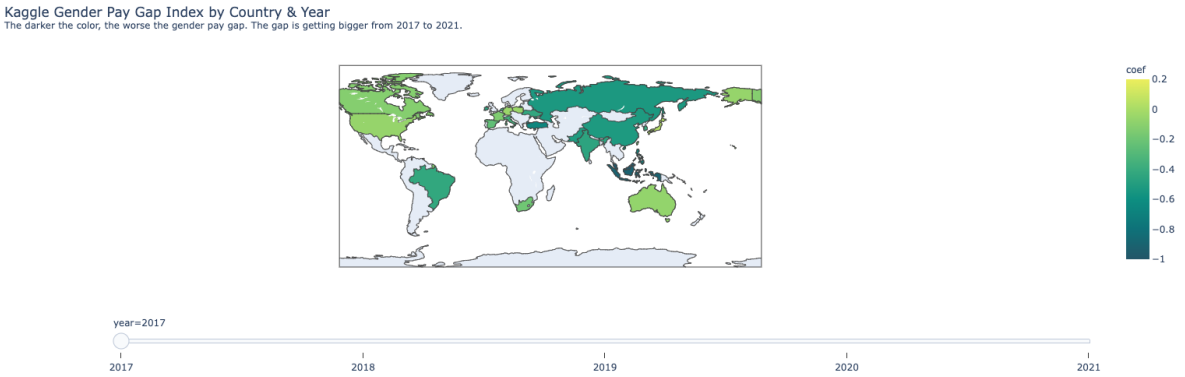
Sub-Task 6: Mapping the Gender Pay Gap

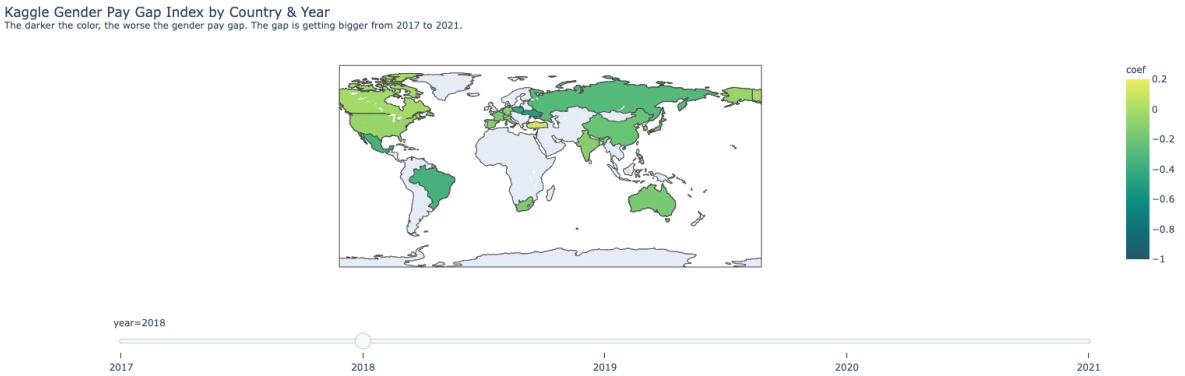
Finally, it comes to our last Sub-Task which shows the gender pay gap in different countries using an interactive map.

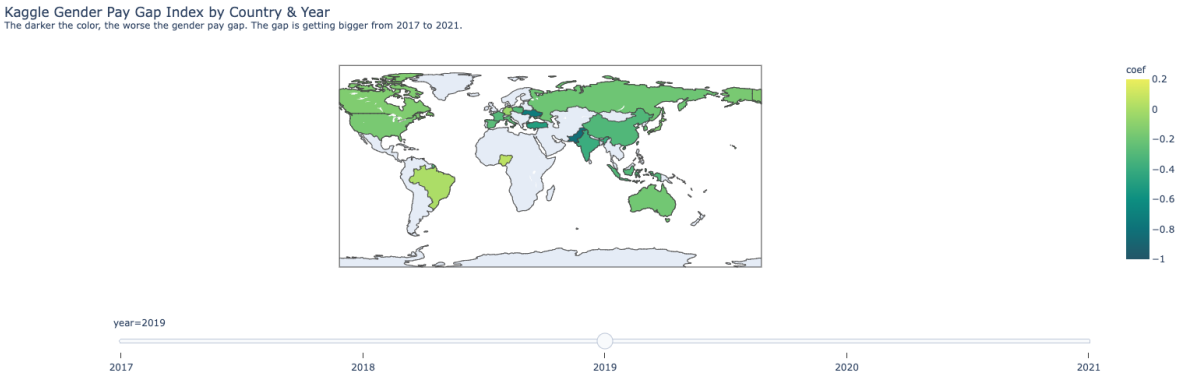
We then run a similar regression as we did for the 2021 data for the 2017 - 2020 data. The controlled variables are slightly different as some questions are not asked in certain years. But in general, the number of controlled variables is similar. Again, we chose countries that have more than 20 female respondents of that year, and the countries selected may be slightly different, but the major countries stay the same.

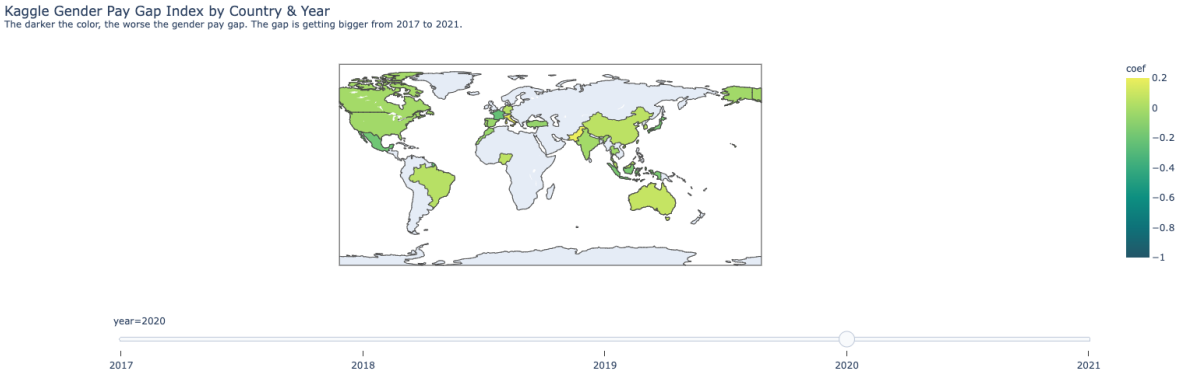
[1] https://en.wikipedia.org/wiki/Global_Gender_Gap_Report

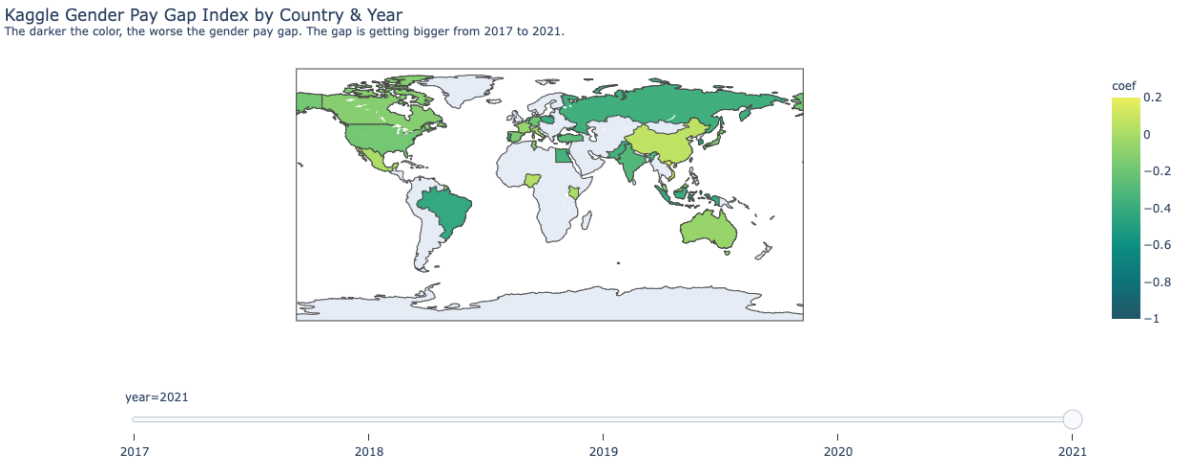
We then merge the regression results to one dataset to plot the Kaggle Gender-Wage Pay Gap indexes. We find the iso code for each country contained to match the map plotting requirements using the plotly package.











This is an interactive map in which you can use “Year” as a scroll bar to check the changes (Please see our google colab). The darker the color is for the country, the worse the gender pay gap is in the data science community. The year 2020 seems pretty chill as the economy for that year is not ideal due to the hit of covid-19, which may explain the narrowed gender pay gap index.

Conclusion

Through these analyses, we were able to have a general idea about the people working in the machine learning and data science industry; and to examine and present the gender pay gap in intersection with other factors.

We are seeing the gap across different age groups, educational backgrounds, job titles, and countries. The few exceptions, like age group 55-59 or under job titles researchers, are of some of the lowest, and likely insufficient sample sizes. We are seeing in our graphs female compensation hitting the plateau much earlier in one's lifetime and much lower on the money scale. We are seeing in our regression results that the higher-earning the job, the fewer women.

These findings, though anticipated, are still mildly disturbing for us as young women pursuing a STEM degree to see. We are witnessing in our study the proof of gender inequality.

But we were also deeply inspired.

We are seeing in these data that the higher the degree is, the higher percentage women take, and the narrower the gap becomes. We are seeing women striving in new industries. We are seeing women making career progress in their late forties. And perhaps most importantly, we are seeing a higher percentage of women in younger generations.

We want to emphasize once again, before diving into the reflection part, that many of our findings and discussions on the gender gap present only the trends. Some of these correlations may not be statistically significant, let alone that we are bound by the limitation of our data and methods.

Reflection

On the Kaggle Survey

The Kaggle survey datasets gathered 106,301 responses altogether. After cleaning, only 76,680 were left for our analysis, meaning roughly one in four responses did not pass our standards.

Partly, it was for a solid reason. Due to the limited information on non-binary gender and the limitation of our study, we had to drop the responses with non-binary or null gender status. Also, as we aim to map out the gender pay gap, we had to let go of all the data with null country status and those who are not working (students and unemployed).

We also found flaws in the survey design, some of which were fixed by Kaggle in the following years. 2017's data were especially messy since it was the first year of the survey. Both age and compensation were to be entered by the surveyed, there wasn't even a filter for non-numeric characters. It resulted in both junk data (such as the 0-year-old geniuses who already have 3-5 years of coding experience) and in fewer data gathered: in 2017, only about 27% of the respondents disclosed their compensation, but once Kaggle changed the question into a selection from a list of ranges, the percentage answered immediately rose to well above 70%.

Other flaws lived on. One that dramatically increased the amount of dirty work for us was the inconsistency throughout the five surveys. In numeric ranges, whether it's age, years of experience with data, or compensation amount, the ranges were not kept the same. Trying to fix their overlapping over the years means we had to sacrifice part of the value. Similar problems were also found in non-numeric questions such as employer industry and job title, the selections were also wordy and too often crossing with each other. The inconsistency in the surveys has largely limited the potential for longitudinal studies.

On Our Study

Due to the limited information gathered by the survey and the limitation in our study questions, we were not able to present the gender gap in a spectrum. Altogether, 1,968 of the original 106,301 responses are of non-binary gender or undisclosed, which is roughly 1.85%.

Our entire study is built on Kaggle's survey result datasets, which, although we hope for the best, will never be truly representative of the field of data science and machine learning. We only gained a rough understanding of those Kaggle users who filled the survey, which may have represented their fellows only to some extent. Also, as a result, the biases and the flaws within the survey methods lived on in our report.

One major tough decision we had to make was to replace what was a range (of compensation amount, of age, of learning time) with its mean; and in those cases like over 70 years old, we replaced it with the lowest bar of 70. This method has made many of our specific analyses possible but also brought in additional errors.

Reference

Kaggle. (2017). 2017 Kaggle Machine Learning & Data Science Survey. https://www.kaggle.com/kaggle/kaggle-survey-2017
Kaggle. (2018). 2018 Kaggle Machine Learning & Data Science Survey. https://www.kaggle.com/kaggle/kaggle-survey-2018
Kaggle. (2019). 2019 Kaggle Machine Learning & Data Science Survey. https://www.kaggle.com/c/kaggle-survey-2019
Kaggle. (2020). 2020 Kaggle Machine Learning & Data Science Survey. https://www.kaggle.com/c/kaggle-survey-2020
Kaggle. (2021). 2021 Kaggle Machine Learning & Data Science Survey. https://www.kaggle.com/c/kaggle-survey-2021
World Economic Forum. (2021, March). Global Gender Gap Report 2021. https://www.weforum.org/reports/global-gender-gap-report-2021

Appendix: Table of Contents (Canvas Files)

Women in Data Science.pdf

Notebooks

1. MSSP 607 - Final Project - Part 1 Data Processing - Shiyu & Jin.ipynb
2. MSSP 607 - Final Project - Part 2 Visualization - Shiyu & Jin.ipynb

    ** Please open the second notebook directly through **this Colab link** for better interactive visualization. It's the same notebook online, but opening it with Jupyter Notebook would require you to install plotly.express, read-in the cleaned datasets, and run it again to present our graphs.*

Original Datasets

1. multipleChoiceResponses_2017.csv
2. multipleChoiceResponses_2018.csv
3. multipleChoiceResponses_2019.csv
4. kaggle_survey_2020_responses.csv
5. kaggle_survey_2021_responses.csv

Cleaned and Merged Dataset

1. df_tech.csv