Desirée Maestri

23 May 2025

Data Science / week 3

GENDER GAP IN DATA SCIENCE & MACHINE LEARNING

EDA analysis

SQL, Pandas, Matplotlib, Seaborn,

DATASET

2022 KAGGLE MACHINE LEARNING & DATA SCIENCE SURVEY.

This survey received **23,997 responses,** collected between 16/Aug and 16/Sep 2022.

Not all questions were asked of all respondents. The file explains the survey logic, and it will be referenced as needed for this project.

If a country or territory received fewer than 50 respondents, we grouped them into a group named “Other” for privacy reasons.

DATASET LIMITATIONS

The data have been collected using a Volunteer Sample and/or a **Convenience Sample**, which could potentially skew the results.

The open-ended responses were separated to anonymize the respondents. And it seems this file was not shared. Thus, I won't consider the open-ended responses.

QUESTIONS

MAIN QUESTION: IS THERE A GENDER GAP AMONG PROFESSIONALS IN MACHINE LEARNING AND DATA SCIENCE?

- What is the distribution?
  - gender
  - country
  - students/professionals
  - age
  - salary range
- Is there a pay gap between genders?
  - If so, how big is the gap?
- Is there a difference in education level between genders?
- Are men more likely to hold leadership positions than women and non-binary people?
- Does gender have any impact on job title?

DATA QUALITY & CLEANING

	Duration (in seconds)	Q2	Q3	Q4	Q5	Q6_1	Q6_2	Q6_3	
0	Duration (in seconds)	What is your age (# years)?	What is your gender? - Selected Choice	In which country do you currently reside?	Are you currently a student? (high school, uni...	On which platforms have you begun or completed...	On which platforms have you begun or completed...	On which platforms have you begun or completed...	com
1	121	30-34	Man	India	No	NaN	NaN	NaN	
2	462	30-34	Man	Algeria	No	NaN	NaN	NaN	

- Removed and exported the first row: questions text
- Delete duplicate rows
- Check missing values (a lot columns missing: multiple choice) → cleaning as needed.
- Later, I created other subsets with cleaner data.

EXPLORATORY DATA ANALYSIS (EDA)

UNIVARIATE ANALYSIS

- Country distribution
  - Total, percentage, top 15
- Gender distribution
  - Total, percentage
- Salary range

BIVARIATE ANALYSIS

- Education level
  - grouped by gender and by level
- Age distribution by gender
- Students and professionals by gender
- Salary distribution by gender
- Job title by gender
- Correlation between gender and salary

SQL PANDAS NUMPY

MATPLOTLIB SEABORN SCIPY.STATS

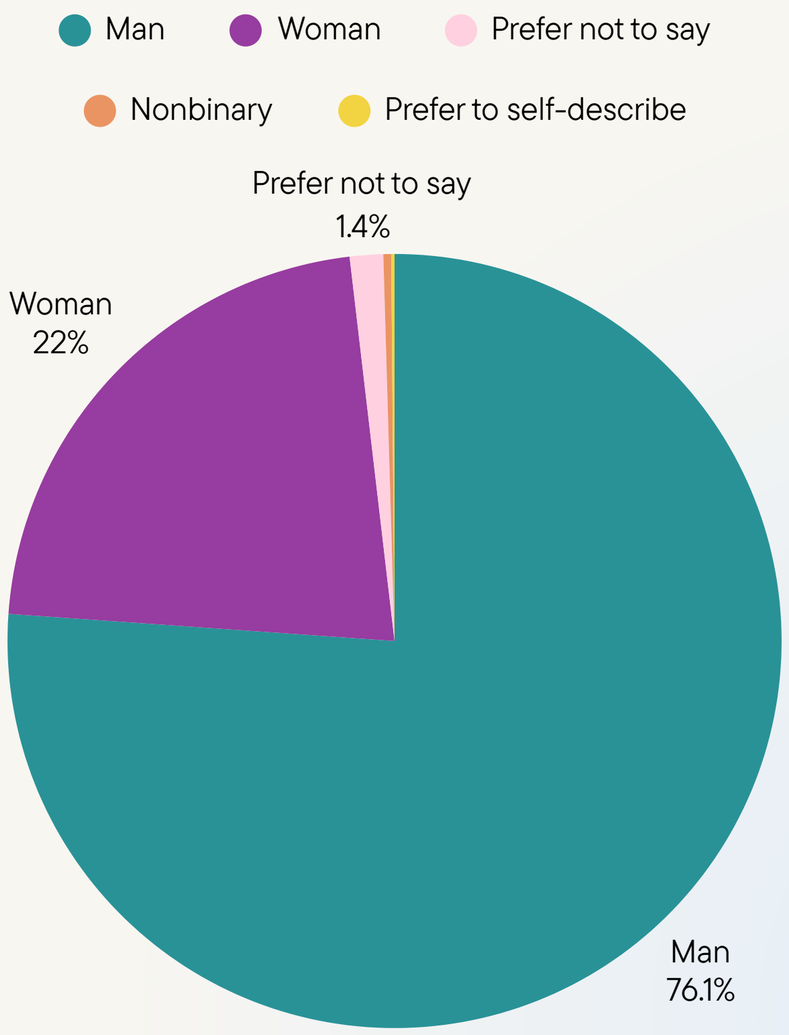
GENDER DISTRIBUTION

Man 76%,
Woman 22%,
Nonbinary 0.3%
Self-describe 0.1%
Prefer not to say 1.4%

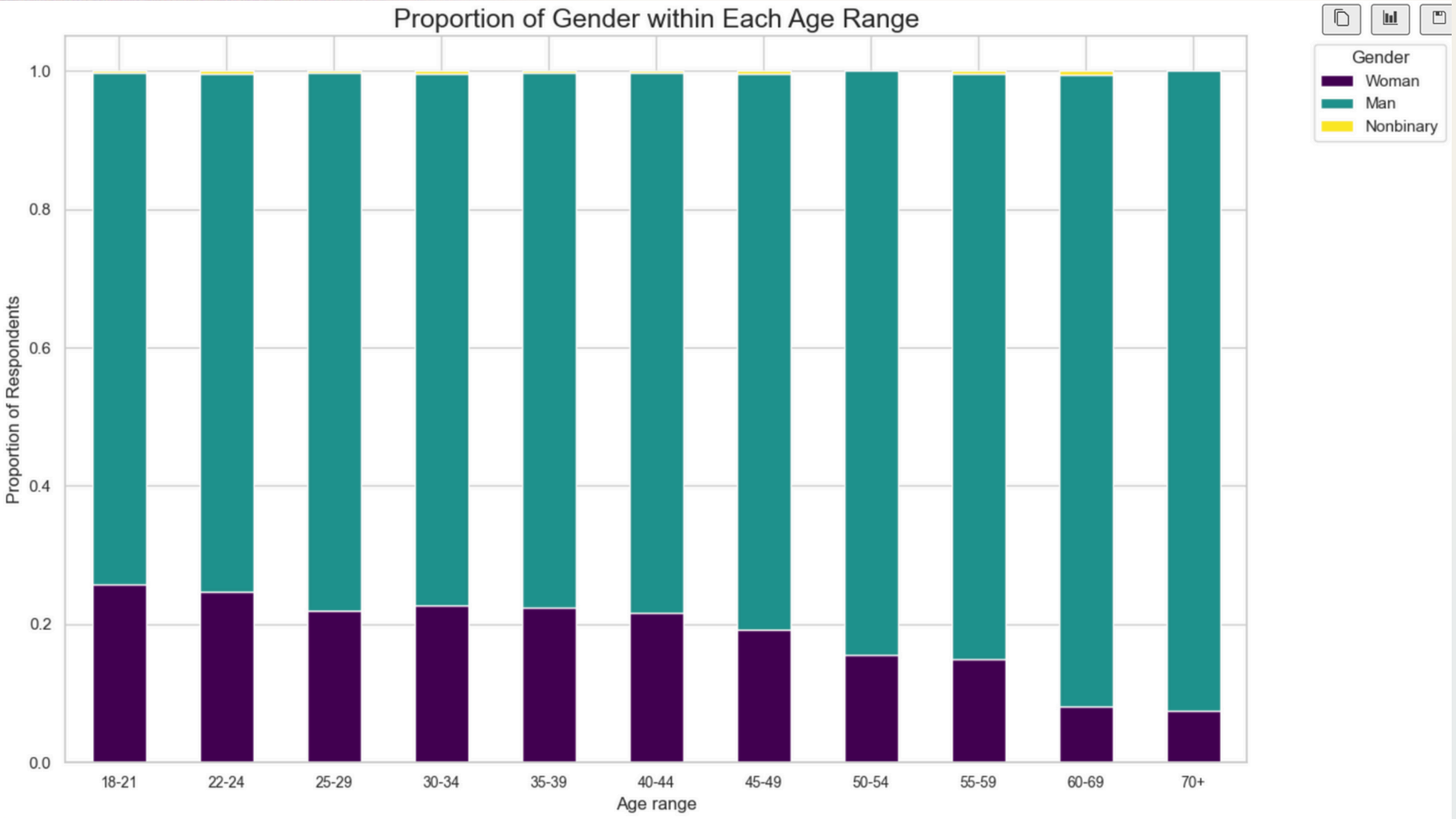
GLOBALLY PROPORTIONAL
22%-26% for women in AI/ML/DS

Self-describe data not available.
Prefer not to say can't be considered

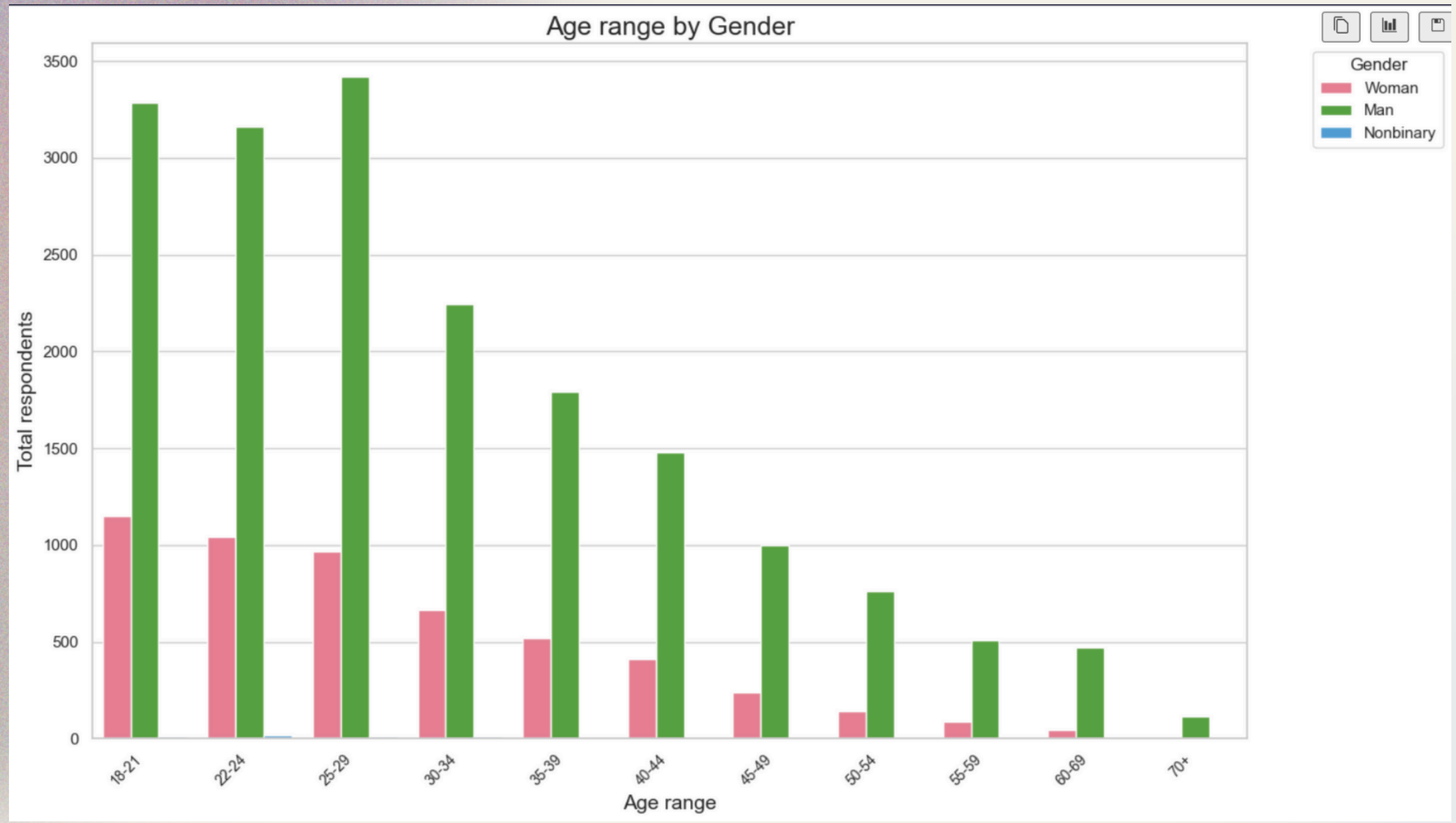
New categories:
Woman
Man
Nonbinary



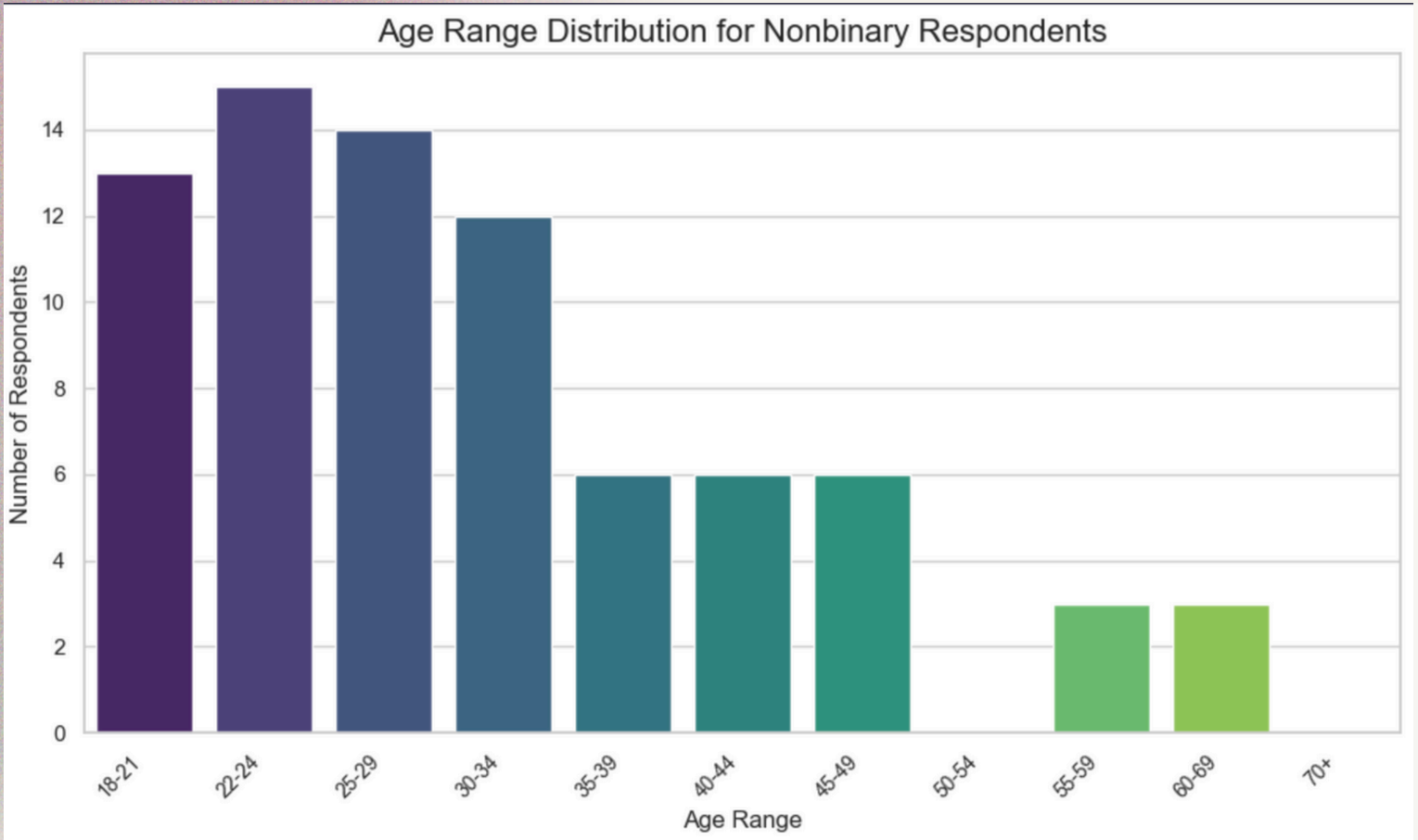
AGE BY GENDER



AGE BY GENDER



AGE BY GENDER



Higher concentration 18-34 years

COUNTRY DISTRIBUTION

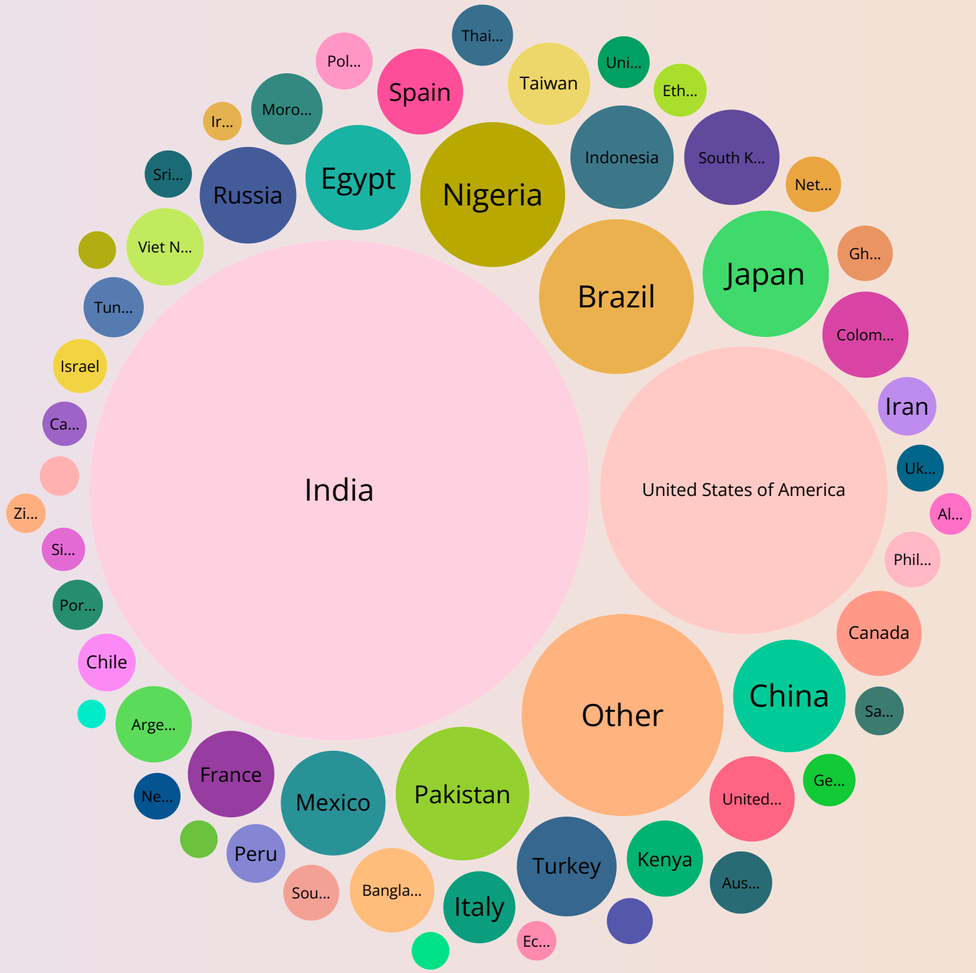
- **56 distinct countries/territories**

- **India (36.64%) and the United States (12.17%)** together, nearly half of the sample

- Countries with less than 50 entries = **"Other"**

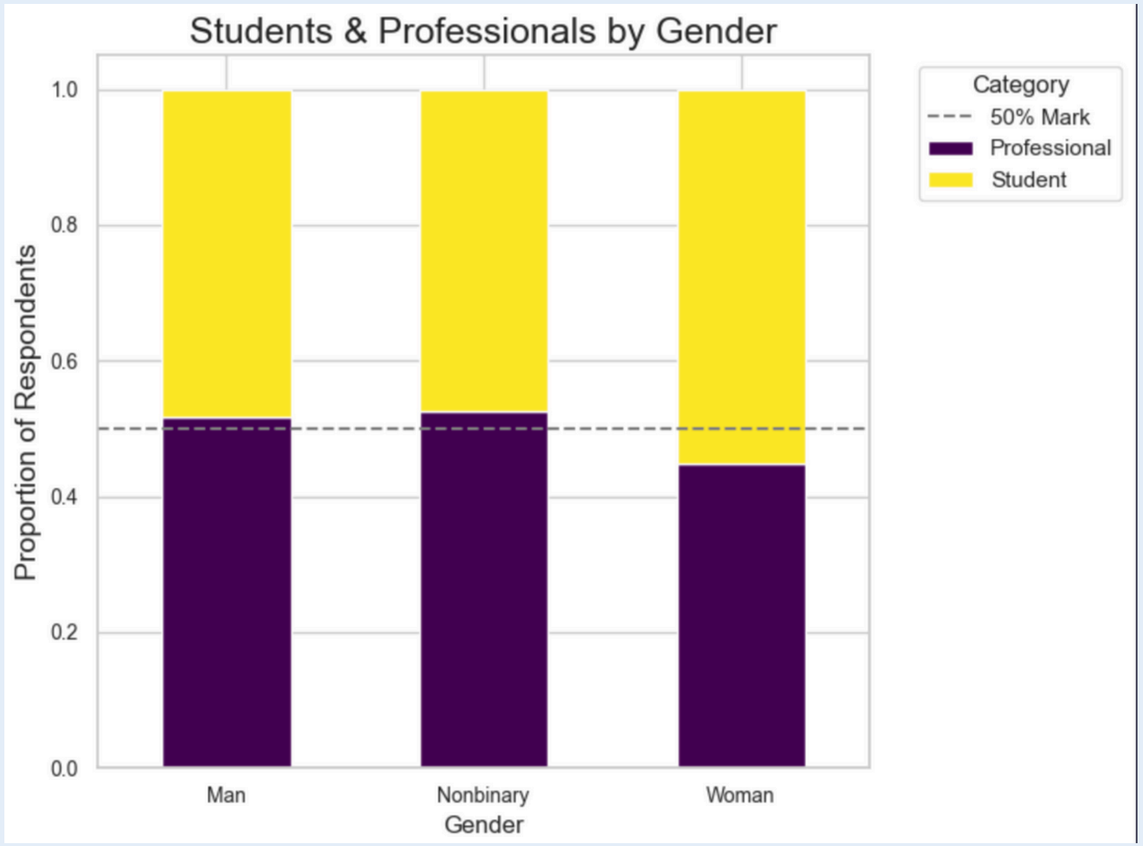
- The **European Union** is underrepresented
15th France (1.09%)

- **Methodological Bias:** Convenience sample



STUDENTS & PROFESSIONALS

- The proportion is almost 50%.
- There are slightly more female students.



EDUCATION LEVEL

Proportional among genders:
Master's degree ~ 35-39%
Bachelor's degree ~ 31-33%

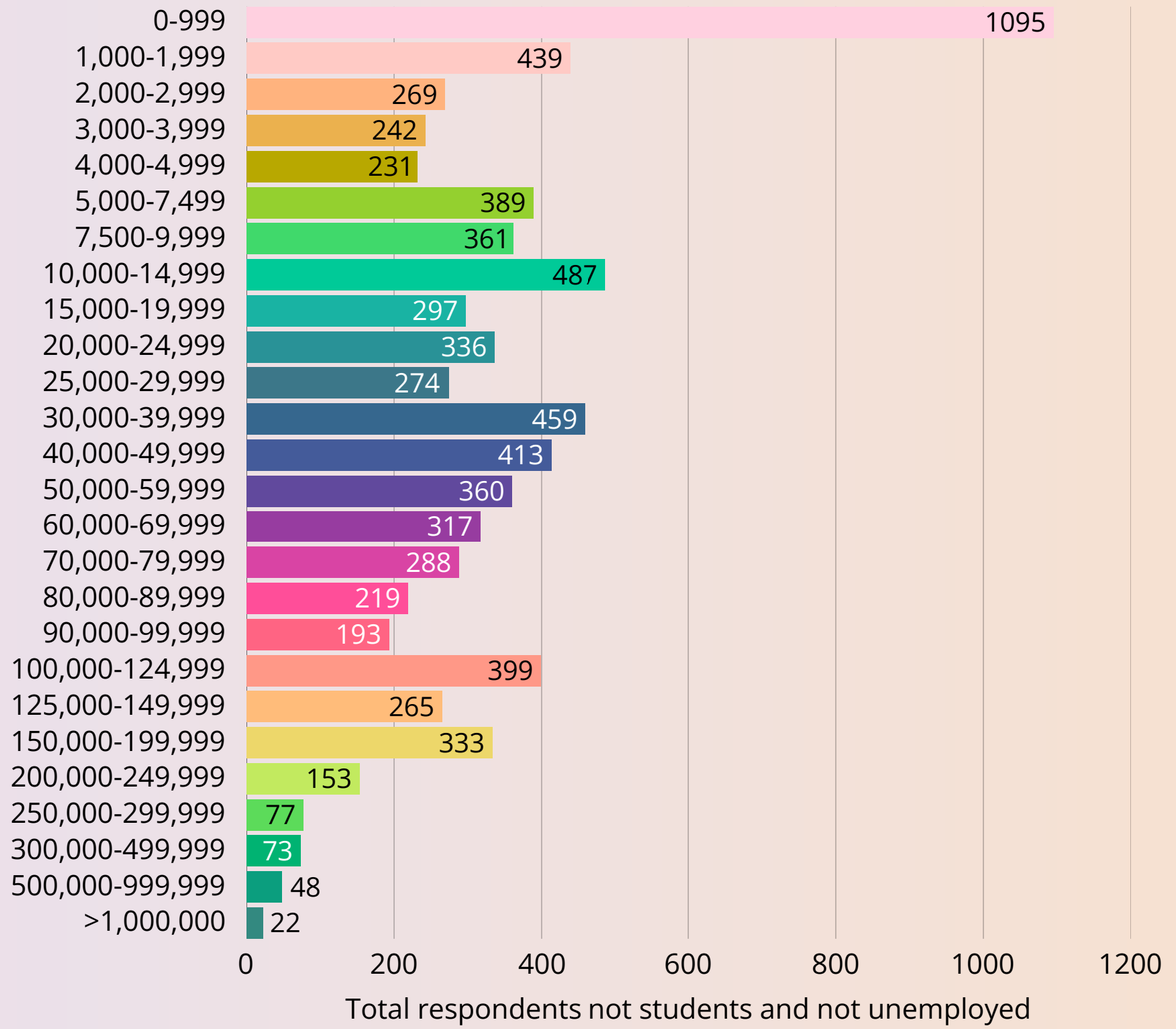
Gender	Education Level	Total...	% by Gender
Man	Master's degree	6968	39.12
Man	Bachelor's degree	5889	33.06
Man	Doctoral degree	1993	11.19
Man	Some college/university study without earning...	1109	6.23
Man	I prefer not to answer	964	5.41
Man	No formal education past high school	472	2.65
Man	Professional doctorate	419	2.35
Nonbi...	Master's degree	28	36.36
Nonbi...	Bachelor's degree	25	32.47
Nonbi...	Some college/university study without earning...	12	15.58
Nonbi...	Doctoral degree	5	6.49
Nonbi...	Professional doctorate	5	6.49
Nonbi...	No formal education past high school	2	2.6
Woman	Master's degree	2016	39.23
Woman	Bachelor's degree	1605	31.23
Woman	Doctoral degree	623	12.12
Woman	I prefer not to answer	378	7.36
Woman	Some college/university study without earning...	289	5.62

SALARY DISTRIBUTION

- Salary question asked only of respondents who are NOT students and are NOT unemployed
- Annual salary ranges from 0-999 to >1milion
- Converted to midpoint to work with continuous

count	8039
mean	56193
std	98818
min	500
25%	3500
50%	22500
75%	75000
max	1000000

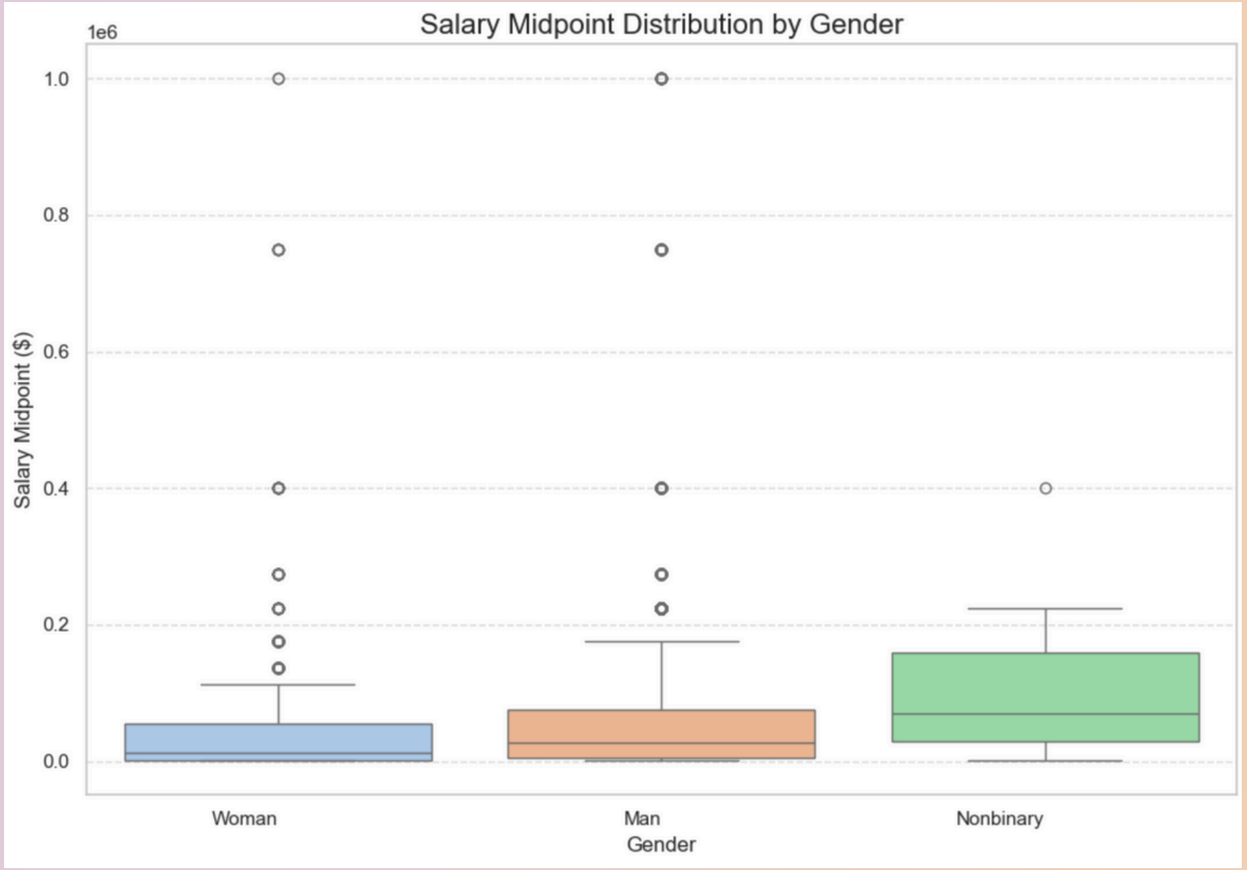
And, yes, there are 22 people who make 7 figures annually as a DS/ML/AI and are active in the Kaggle community.



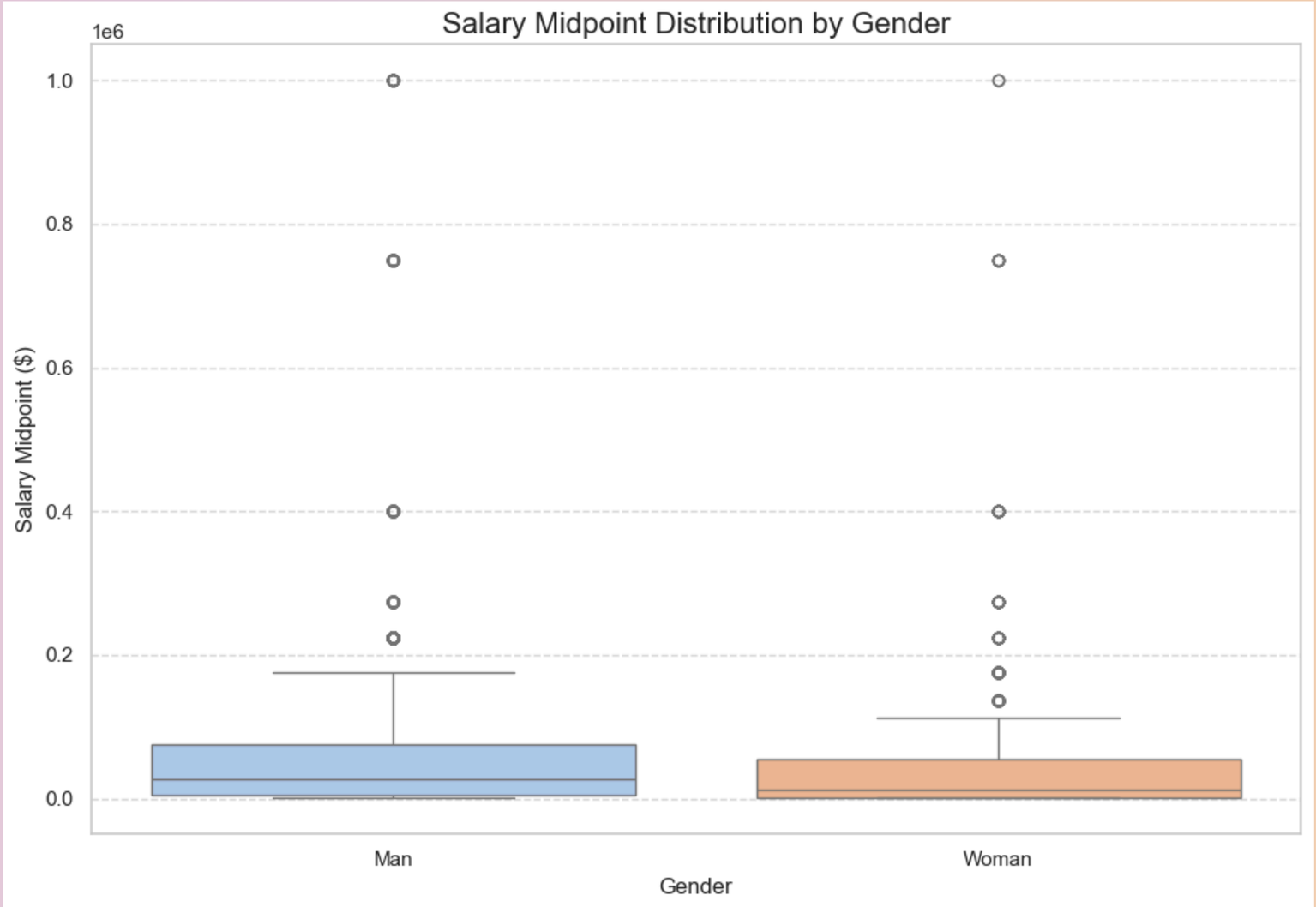
SALARY DISTRIBUTION BY GENDER

- Women earn **less than half** of Men's median
- The difference between median and mean also suggests skewness in the distribution, which is common for income data
- Nonbinary, employed and who answered this survey are potentially an exception, in comparison to the global market

Median Salary Midpoint		**Mean** Salary Midpoint	
Nonbinary	69,999.5	Nonbinary	101,863
Man	27,499.5	Man	58,756
Woman	12,499.5	Woman	43,160



SALARY DISTRIBUTION WOMEN AND MAN



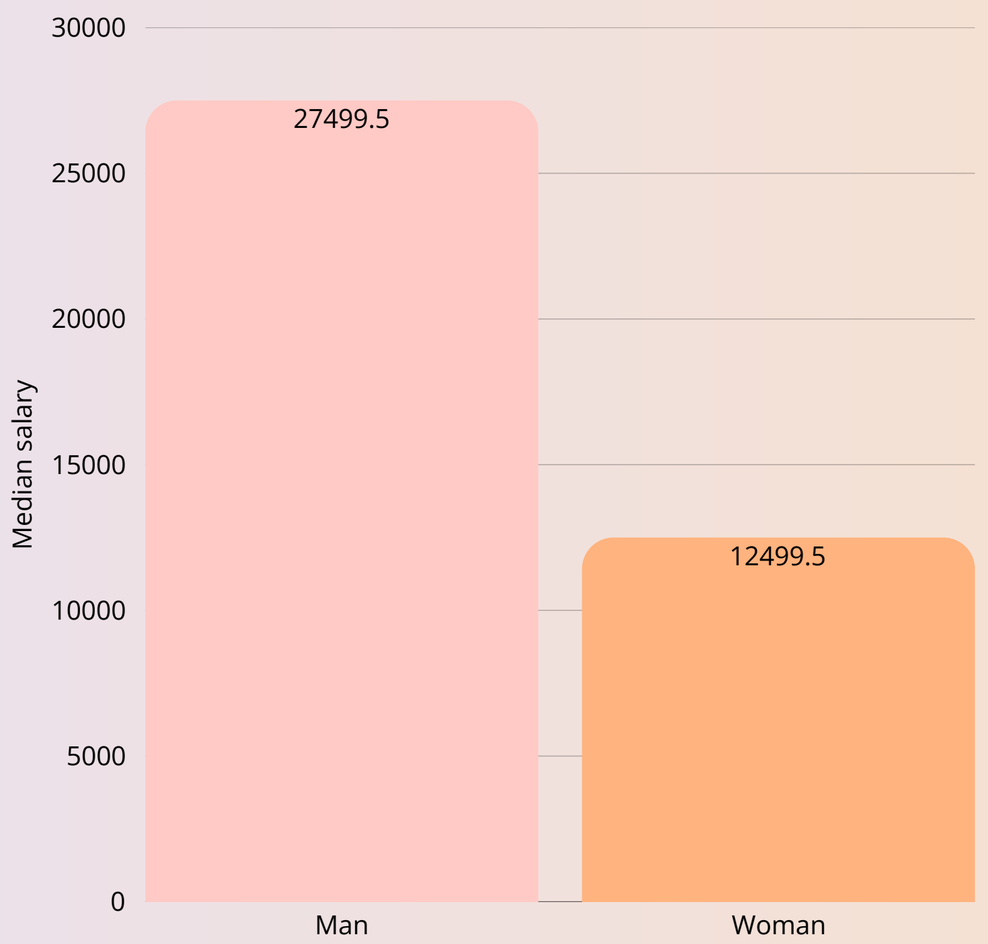
THE GENDER PAY GAP

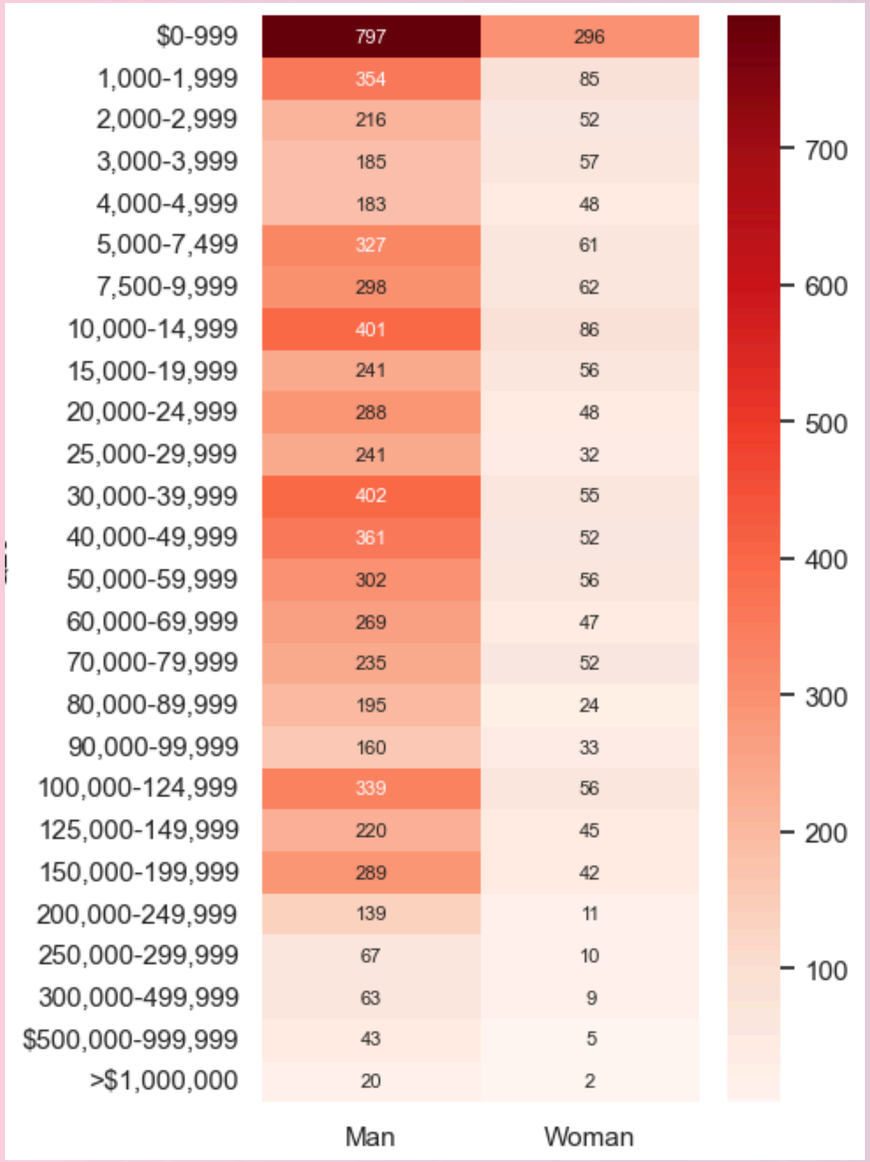
50% of the women earn **USD 12.5k** or less, per year

while

50% of the men earn **USD 27.5k** or less, per year

which is more than double





CHI-SQUARE TEST: MEN, WOMEN, NONBINARY

```
2.0422246648812616e-16
p-value < 0.05 == True
```

There is a significant correlation between gender and salary

CHI-SQUARE TEST: MEN, WOMEN

```
6.146576306657101e-18
p-value < 0.05 == True
```

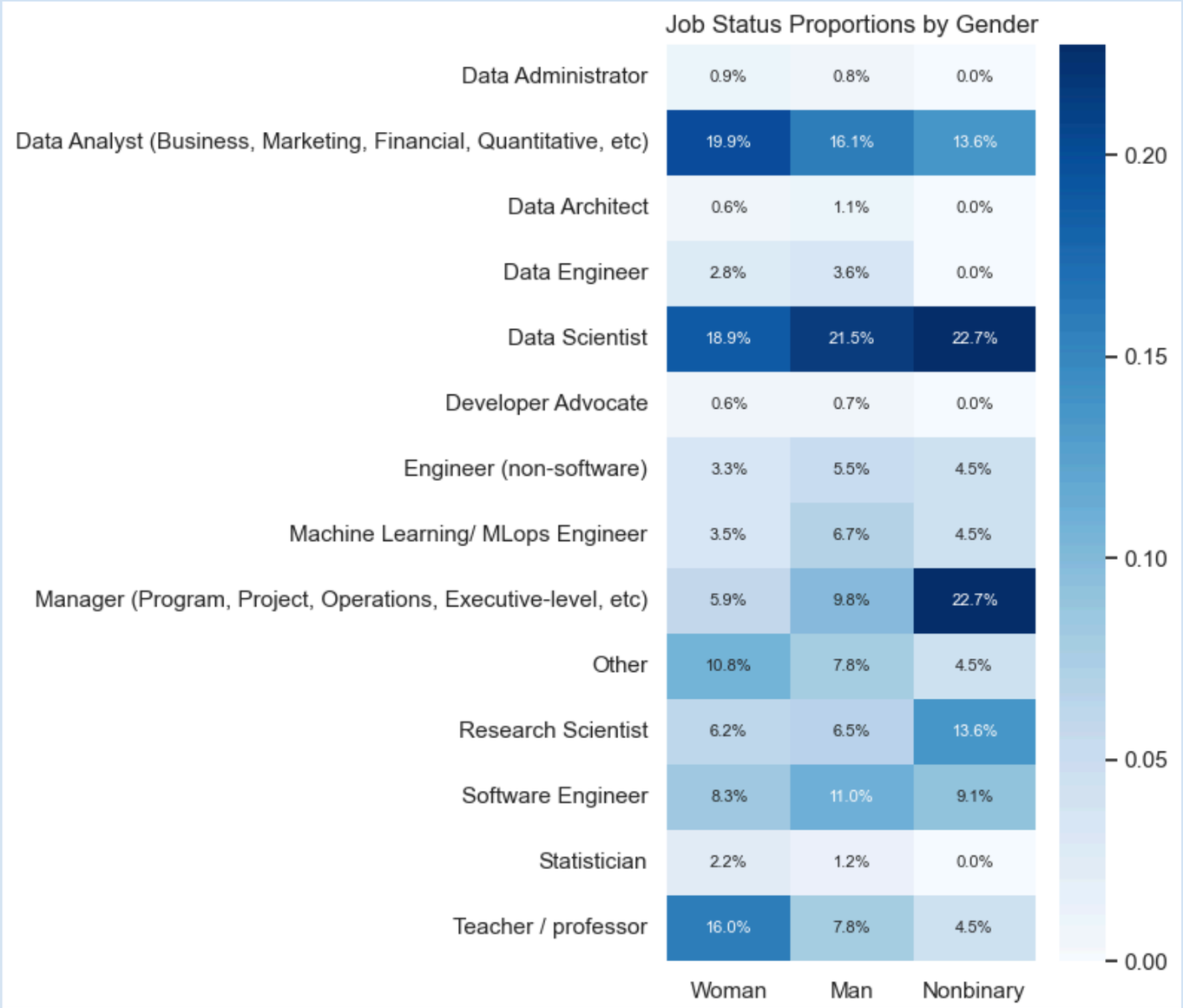
The correlation is even stronger when we analyze only Men and Women

JOB TITLE

TOP 3 WOMEN
Data Analyst = 275 - 19.9%
Data Scientist = 261 - 18.9%
Teacher/Professor = 221 - 16%

TOP 3 MEN
Data Scientist = 1424 - 21.5%
Data Analyst = 1065 - 16.1%
Software Engineer = 732 - 11%

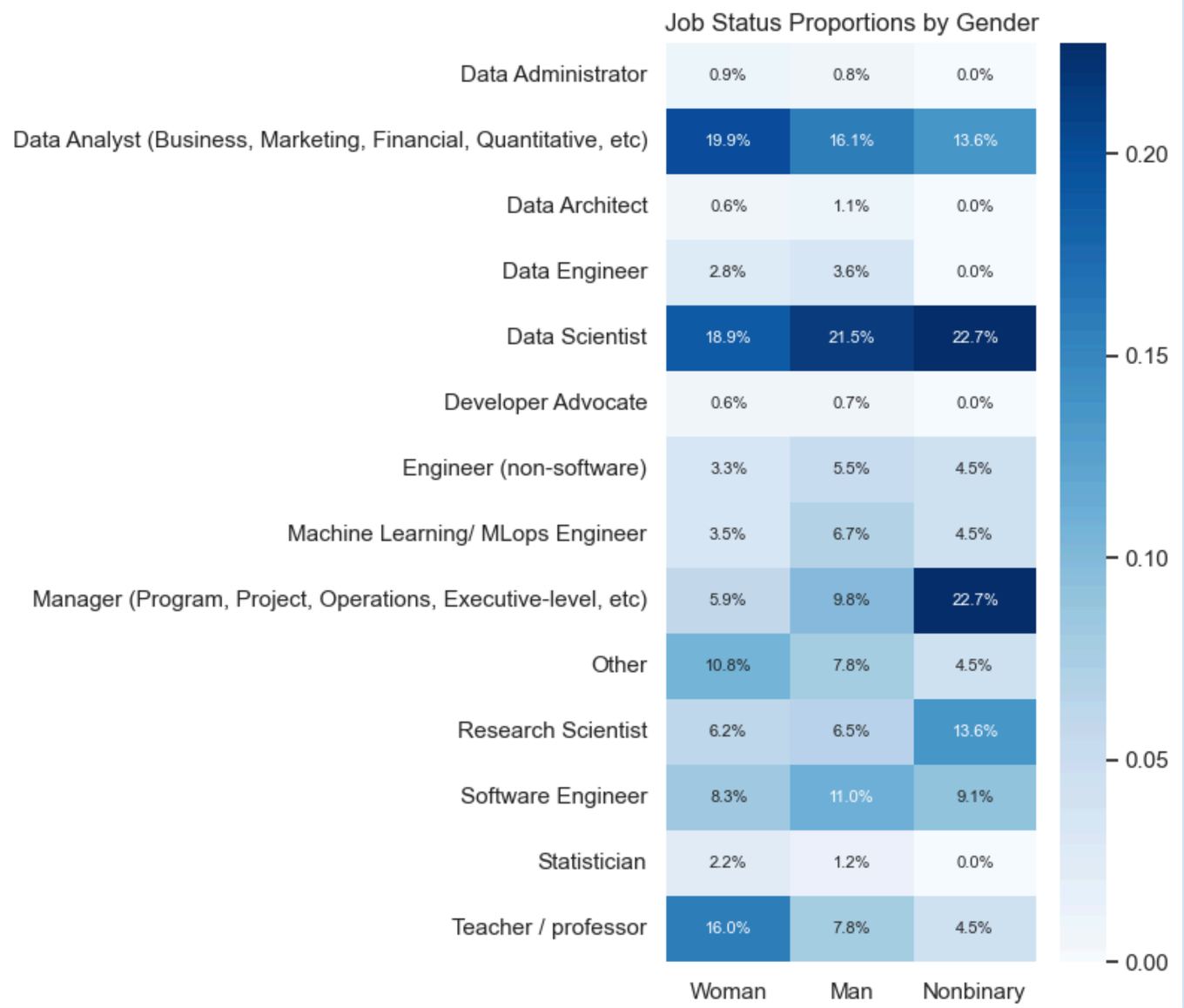
TOP 3 NONBINARY
Manager = 5 - 22.7%
Data Scientist = 5- 22.7%
Data Analyst = 3 - 13.6%
Research Scientist = 3 - 13.6%



JOB TITLE

LEADERSHIP

Data Analyst = 275 - 19.9%
Data Scientist = 261 - 18.9%
Teacher/Professor = 221 - 16%



LEARNINGS

- A lot
- Choose the dataset wisely - good for exercise, invalid for results (biased)
- How to make better plots (finally)
- Formulating questions before starting the analysis is the way to keep on track
- SQL is not that bad and is actually useful to create plots/csv files

THANK YOU

Questions?