Gender Profiling in WikiData

Mar 14, 2024

Queer Data Days

Isabella Lu, Lane Rasberry, Peter Alonzi

DATA SCIENCE

OMIDYAR NETWORK

WILLIAM + FLORA Hewlett Foundation

Goal

Objective

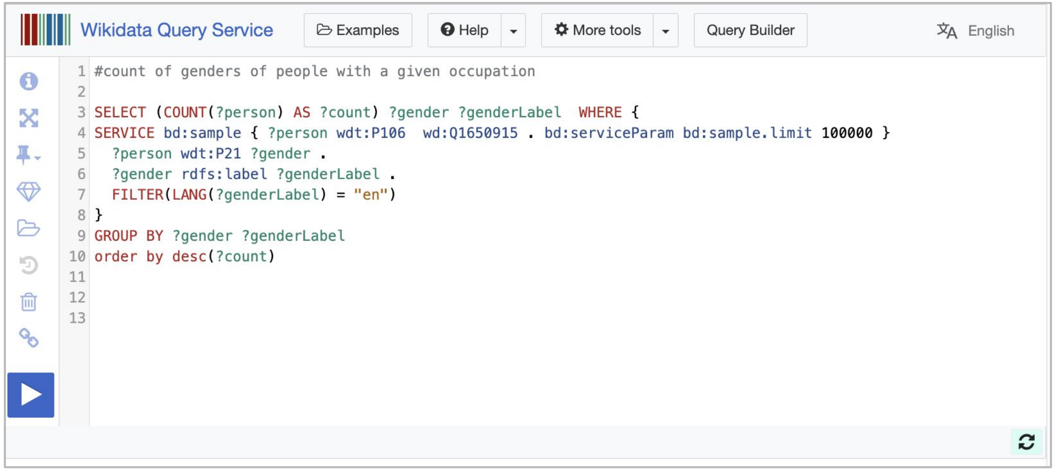
- Showcase Wiki’s capacity to encompass representative data of society
- Profiling occupations by gender

Not:

- Expose demographic trends to advance an agenda

Wikidata Query

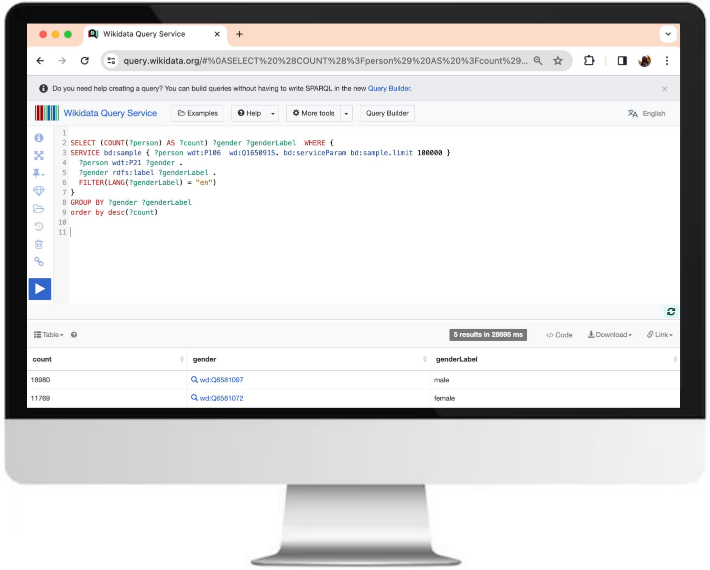
- SPARQL
    - RDF (Resource Description Framework) query programming language
    - Subject-predicate-object triples
- Purpose:
    - Used to retrieve and manipulate RDF datasets
    - Query through Wiki Database
- [Query Service]()
    - Limited > Use Samples



```
#count of genders of people with a given occupation

SELECT (COUNT(?person) AS ?count) ?gender ?genderLabel  WHERE {
SERVICE bd:sample { ?person wdt:P106  wd:Q1650915 . bd:serviceParam bd:sample.limit 100000 }
  ?person wdt:P21 ?gender .
  ?gender rdfs:label ?genderLabel .
  FILTER(LANG(?genderLabel) = "en")
}
GROUP BY ?gender ?genderLabel
order by desc(?count)
```

Accessing WikiData


Wikidata Query Service
query.wikidata.org/#%0ASELECT%20%28COUNT%28%3Fperson%29%20AS%20%3Fcount%29...
Do you need help creating a query? You can build queries without having to write SPARQL in the new Query Builder.
Wikidata Query Service
Examples
Help
More tools
Query Builder
English

```
SELECT (COUNT(?person) AS ?count) ?gender ?genderLabel  WHERE {
SERVICE bd:sample { ?person wdt:P106  wd:Q1650915. bd:serviceParam bd:sample.limit 100000 }
  ?person wdt:P21 ?gender .
  ?gender rdfs:label ?genderLabel .
  FILTER(LANG(?genderLabel) = "en")
}
GROUP BY ?gender ?genderLabel
order by desc(?count)
```

Table
5 results in 28695 ms
Code
Download
Link

count	gender	genderLabel
18980	wd:Q6581097	male
11769	wd:Q6581072	female

EXAMPLE SPARQL QUERY:

Count of genders of people with a given occupation

```
SELECT (COUNT(?person) AS ?count) ?gender ?genderLabel  WHERE {

SERVICE bd:sample { ?person wdt:P106  wd:Q1650915. bd:serviceParam
bd:sample.limit 100000 }

  ?person wdt:P21 ?gender .

  ?gender rdfs:label ?genderLabel .

  FILTER(LANG(?genderLabel) = "en")

}

GROUP BY ?gender ?genderLabel

order by desc(?count)
```

TRIPLE PATTERN: outlines what you are querying for

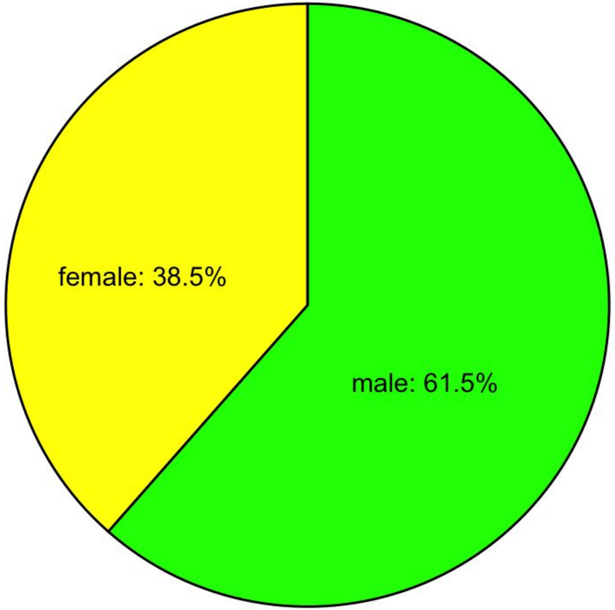
RANDOM SAMPLE: causes limitations

MATCH: values with their respective values

ORDER: determines how you want to order your query

Gender Distribution of People by Occupation

(Researcher)



For this example graph, I only included data related to genders Male and Female. I took out all other genders that made up less than 0.1% of the sample. Results not shown are addressed elsewhere.

Github

Application

Prevalence in Education Industry

Why Is This Important?

Pew Research Center

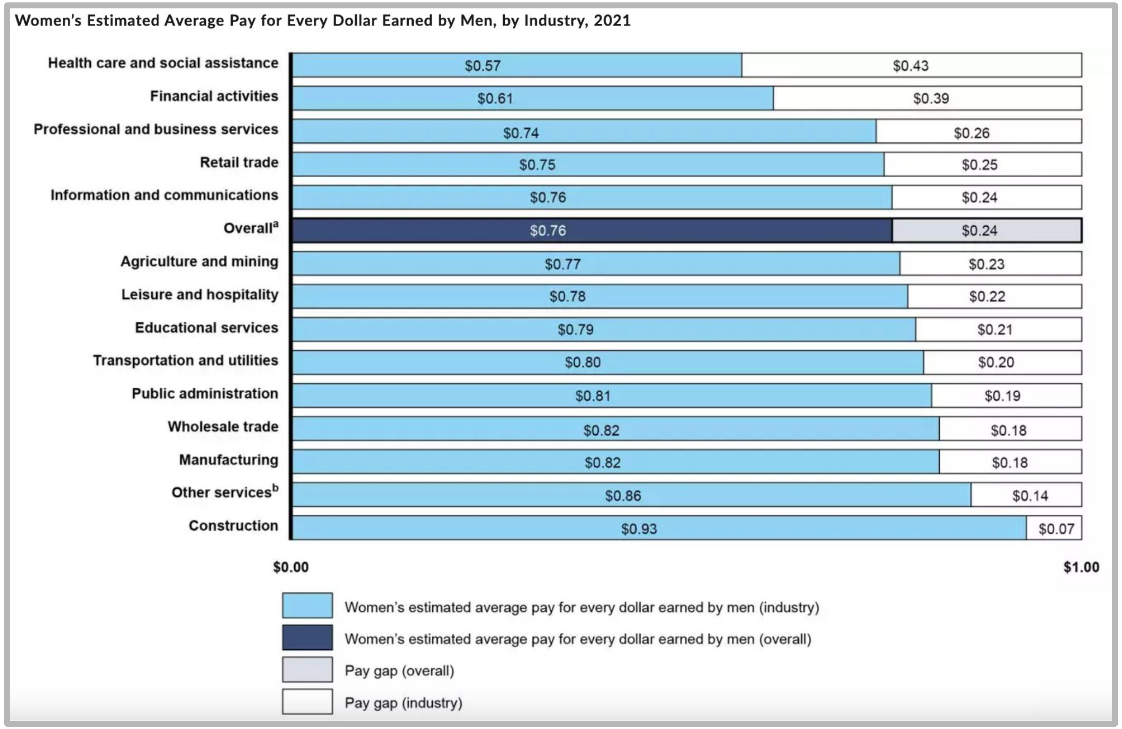
“Incomes are rising in the U.S., but the increase is not being felt equally by all Americans. “

How is income inequality being pictured in Wikidata?

Demographic

Gender Gap

Forbes : “In 2022, women earned 17% less than men on average.”



GAO

Highlight:

In Educational Services, women make $0.79 for every dollar earned by men on average

Data Outside of Wiki

NCES: National Center for Education Statistics

- Women are the majority

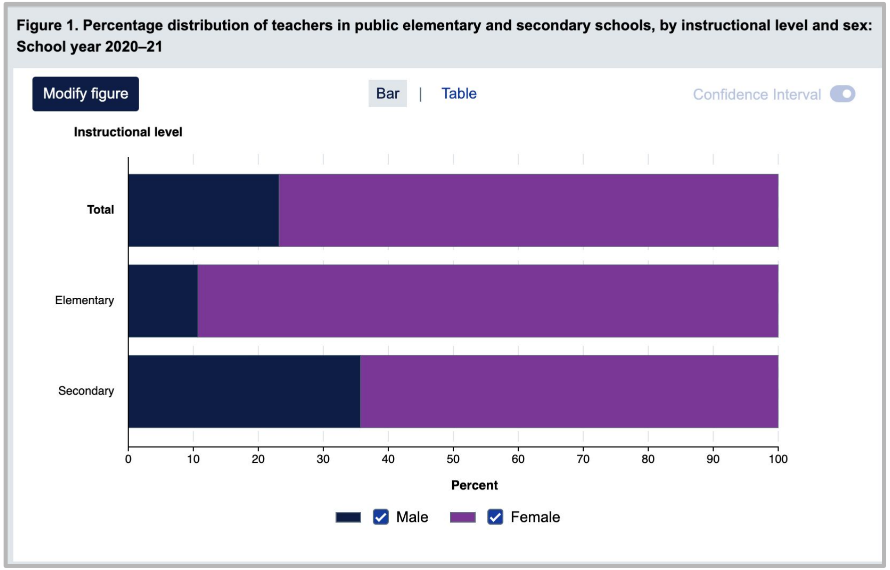


Figure 1. Percentage distribution of teachers in public elementary and secondary schools, by instructional level and sex: School year 2020–21

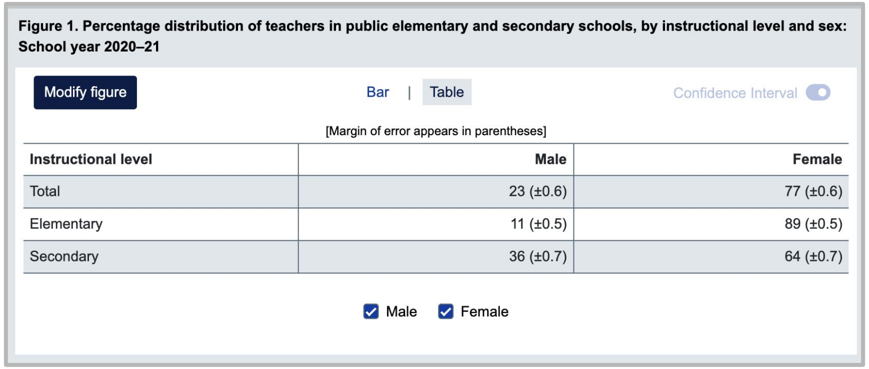


Figure 1. Percentage distribution of teachers in public elementary and secondary schools, by instructional level and sex: School year 2020–21

[Margin of error appears in parentheses]

Instructional level	Male	Female
Total	23 (±0.6)	77 (±0.6)
Elementary	11 (±0.5)	89 (±0.5)
Secondary	36 (±0.7)	64 (±0.7)

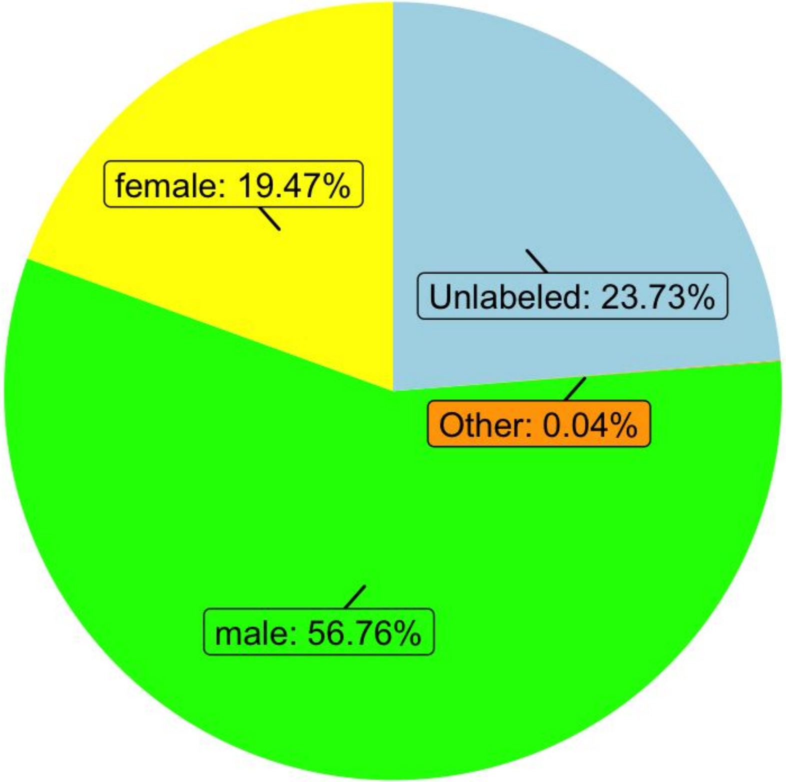
Gender Distribution of Teachers

Demographic Statistics:

- Women are NOT the majority
  - Men: 56.76% ± .1%
  - Women: 19.47% ± .1%

Data Query Details:

- Sample size 100,000
- Random sample of WikiData
- Extra: Those with no gender label



Github

Wikidata Possible Instability

Takeaways:

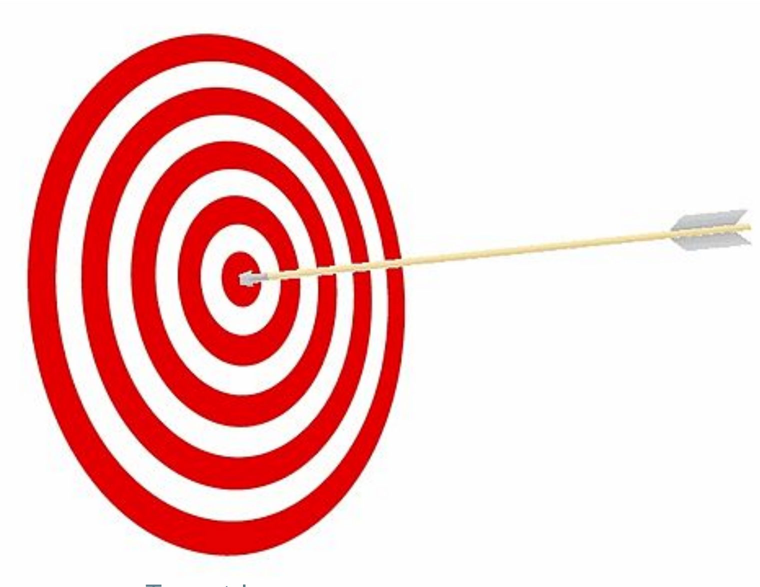
- Challenges arise with data gaps within WikiData
- Caution is needed when interpreting statistical analysis from data
  - Misinterpretation
- Data may not be fully representative of the entire population

Challenge Image

Moving Forward

Next Steps:

- Run with larger sample sizes for more accurate analysis
- Create more analyses on different demographic areas of interest
- Filter more succinctly: obtain accurate timely data



Target Image

Thanks

The William and Flora Hewlett Foundation + Omidyar Network
Funding Project at the School of Data Science at the University of Virginia