



Exploring the Notability Gender Gap

Freebase, BigQuery, Maps (Berlin Buzzwords)

Google Developer Relations:

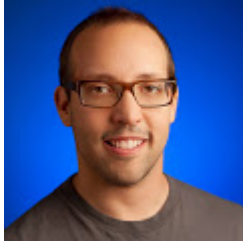
[Felipe Hoffa](#)

[Ewa Gasperowicz](#)

[@felipehoffa](#)

[@devnook](#)

Who



Felipe Hoffa



Ewa Gasperowicz

Google Developer Relations

Who are the most visited
female politicians on Wikipedia?

Most visited female politicians



Elizabeth_I_of_England	251539
Indira_Gandhi	163030
Margaret_Thatcher	141632
Sonia_Gandhi	138239
Hillary_Rodham_Clinton	124649
Shirley_Temple	101610
Sarah_Palin	79619
Angela_Merkel	76521
Hema_Malini	66406
Aung_San_Suu_Kyi	64234
Julia_Gillard	61955
Eleanor_Roosevelt	60447

Which were the most visited books
written by a woman before 2010,
on Wikipedia, on February 11th 2014?

Most visited books written by a woman



Pride_and_Prejudice	4830
The_Hunger_Games	634
Frankenstein	392
Vampire_Academy	368
To_Kill_a_Mockingbird	317
Mockingjay	237
Jane_Eyre	228
Catching_Fire	172
The_Lovely_Bones	124
Wuthering_Heights	115
Emma	110
Gone_with_the_Wind	92

Most visited books written by a woman



```
SELECT title, SUM(requests) c
FROM [fh-bigquery:wikipedia.wikipedia_views_20140211_21]
WHERE title IN (
  SELECT REGEXP_REPLACE(obj, '/wikipedia/id/', '')
  FROM [fh-bigquery:freebase20140119.triples_nolang]
  WHERE sub IN (
    SELECT a.sub FROM (
      SELECT sub, obj
      FROM [fh-bigquery:freebase20140119.triples_nolang]
      WHERE pred = '/book/written_work/author') a
    JOIN EACH (
      SELECT sub FROM [fh-bigquery:freebase20140119.people_gender]
      WHERE gender='/m/02zsn') c ON a.obj=c.sub
    JOIN EACH (
      SELECT sub, INTEGER(REGEXP_EXTRACT(obj, '([0-9]{4})')) pubyear
      FROM [fh-bigquery:freebase20140119.triples_nolang]
      WHERE pred = '/book/written_work/date_of_first_publication'
      HAVING pubyear < 2010) d ON a.sub=d.sub)
  AND obj CONTAINS '/wikipedia/id/' AND pred = '/type/object/key'
GROUP BY 1) GROUP BY 1 ORDER BY 2 DESC;
```

Exploring the Notability Gender Gap

1

Who, what, why

2

What is Freebase

3

Querying Freebase with BigQuery

4

Visualizing with Maps

What



Data source



Data processing



Data
visualisation

The process

What



Freebase



Google BigQuery



Google Maps



google.com/diversity

Why

- Exploring a dataset is fun
- Don't accept aggregated data
- Meet the tools and dataset
- Ask
- Act

Data source



About 2,530,000 results (0.25 seconds)

[Grace Hopper - Wikipedia, the free encyclopedia](#)

en.wikipedia.org/wiki/Grace_Hopper ▾ Wikipedia ▾

Grace Murray Hopper (December 9, 1906 – January 1, 1992) was an American computer scientist and United States Navy rear admiral. A pioneer in the field, ...

[USS Hopper \(DDG-70\)](#) - [Cobol](#) - [Harvard Mark I](#) - [Rear admiral](#)

[Grace Hopper Celebration](#)

gracehopper.org/ ▾ Grace Hopper Celebratio... ▾

The **Grace Hopper** Celebration of Women in Computing is the World's Largest Gathering of Women Technologists. It is produced by the Anita Borg Institute and ...

[Grace Hopper - Wikiquote](#)

en.wikiquote.org/wiki/Grace_Hopper ▾ Wikiquote ▾

Rear Admiral **Grace Hopper** (9 December 1906 – 1 January 1992) was a U.S. Naval officer, and an early computer programmer. She was the developer of the ...

[Grace Murray Hopper - Computer Science - Yale University](#)

www.cs.yale.edu/~tap/Files/hopper-story.html ▾ Yale University ▾

Rear Admiral Dr. **Grace Murray Hopper** was a remarkable woman who grandly rose ... **Grace Brewster Murray** was born on December 9, 1906 in New York City.

[Grace Hopper on Letterman - YouTube](#)



www.youtube.com/watch?v=1... ▾ YouTube ▾

Dec 12, 2012 - Uploaded by TheLazlo101

Grace Hopper on Letterman ... Play all **Grace Hopper** by JoeLumbley; 7:37 ... Rear Admiral **Grace Hopper** by ...

[US People--Hopper, Grace Murray.](#)

www.history.navy.mil/...us/.../g-hoppr.ht... ▾ Naval History & Heritage... ▾

formal and informal photographic portraits of **Grace Murray Hopper** and a picture related to her early computer work, first computer bug Women, U.S. Navy, US ...

[Biography - Rear Admiral Grace Murray Hopper, USN](#)

www.history.navy.mil/.../hopper_grace.ht... ▾ Naval History & Heritage... ▾

Transcript of Naval Service For Commodore **Grace Murray Hopper**, US Naval Reserve. 9 December 1906 - Born in New York, New York. 30 May 1944 ...

[Admiral Dr. Grace Murray Hopper - US Navy](#)



Grace Hopper

Computer Scientist

Grace Murray Hopper was an American computer scientist and United States Navy rear admiral. A pioneer in the field, she was one of the first programmers of the Harvard Mark I computer, and developed ... [Wikipedia](#)

Born: December 9, 1906, [New York City, NY](#)

Died: January 1, 1992, [Arlington County, VA](#)

Spouse: [Vincent Foster Hopper](#) (m. 1930–1945)

Awards: [National Medal of Technology and Innovation](#), [IEEE Emanuel R. Piore Award](#)

Education: [Yale University](#) (1934), [Yale University](#) (1930), [Vassar College](#) (1924–1928), [Wardlaw-Hartridge School](#)

Parents: [Mary Campbell Van Home Murray](#), [Walter Fletcher Murray](#)

People also search for



[Ada Lovelace](#)



[Charles Babbage](#)



[Alan Turing](#)



[Howard H. Aiken](#)



[Anita Borg](#)



Free and open. Licensed as **CC-BY**

Open for anyone to contribute.

A source for Google's Knowledge Graph

Download the entire graph as RDF

42.9M people places and things

2.4B triples about those things

Feebase is a graph database of facts



- Every fact can be represented as an RDF triple
- Every triple consists of SUBJECT - Predicate - OBJECT

Feebase is a graph database of facts



appears in



Daft Punk - appears in - Tron

Feebase is a graph database of facts



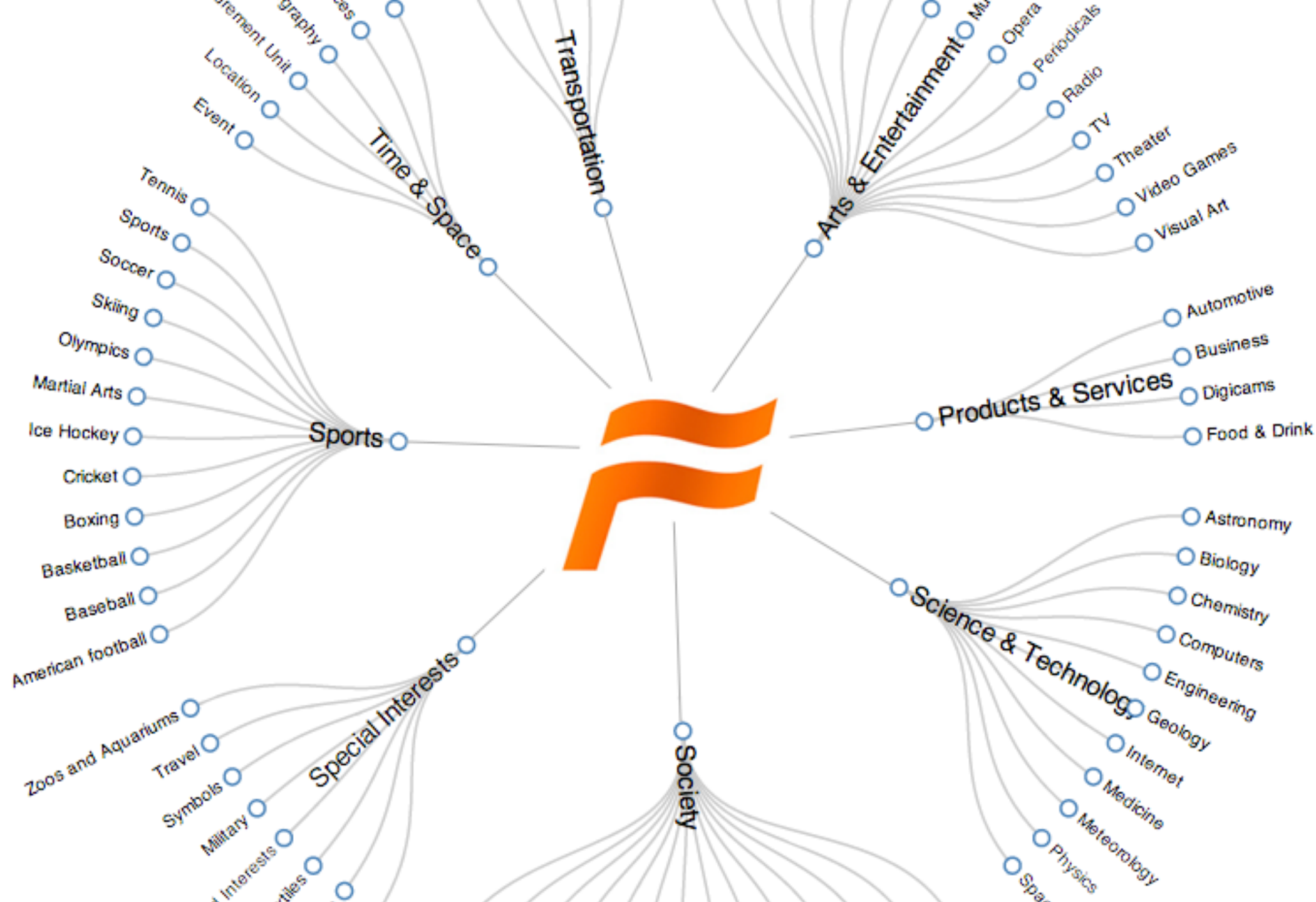
appears in



Daft Punk - appears in - Tron

[/m/016j7m](#) - [/film/music_contributor](#) - [/m/0gxrns](#)





What is / isn't in Freebase

No notability requirements.

Many topics are automatically imported from Wikipedia.

The most detailed areas of Freebase are:

- Music
- Film
- TV
- Books
- Celebrities

Data Quality

All data contributed to Freebase must be at least 99% reconciled with existing data.

Must be less than 1% duplicated or conflated topics.

Factual errors are easier to fix by the community or by bots.

Data processing



Let's write some queries



Google BigQuery is

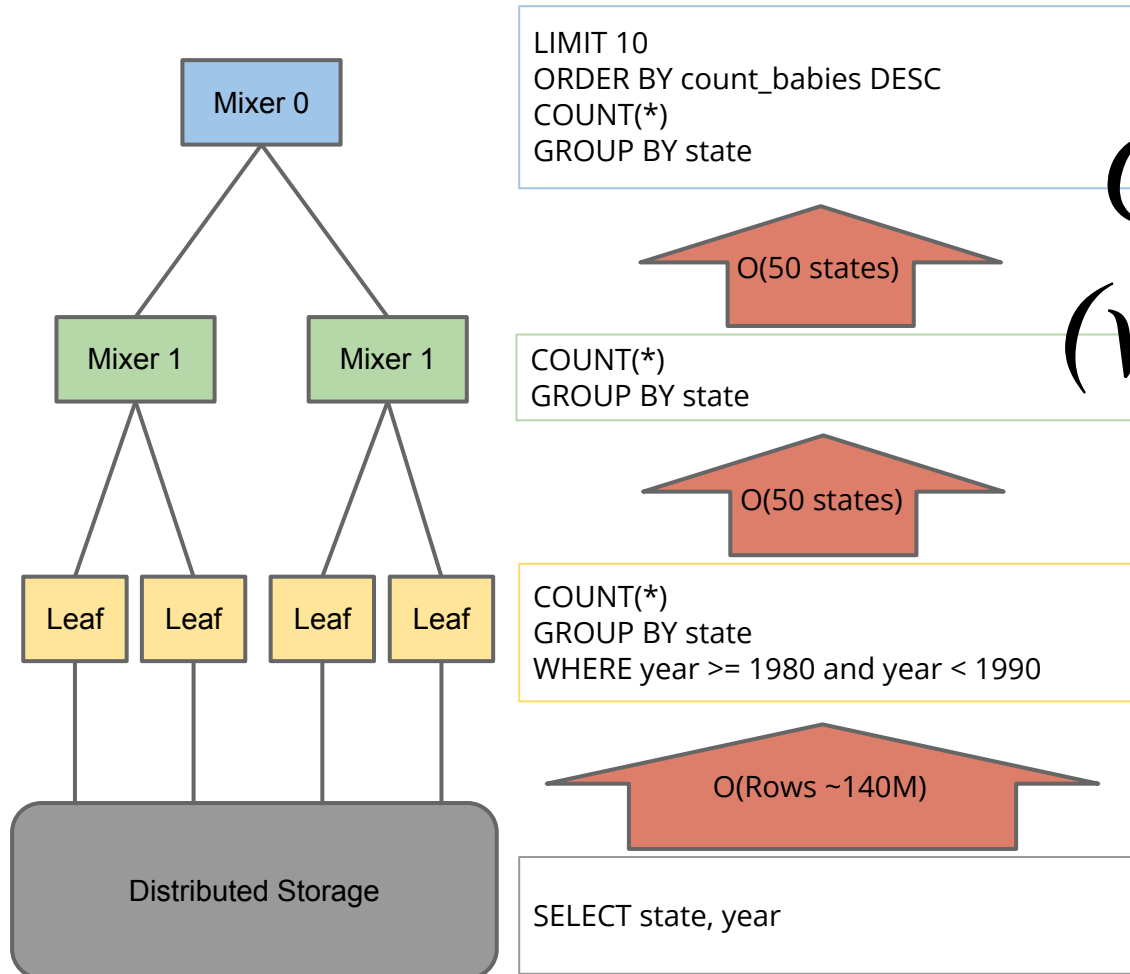
- Analytical database as a service
- Understands SQL
- Analyzes terabytes of data in seconds
- Imports JSON, CSV, data streams
- ~~\$0.08~~ \$0.026/GB month storage
- ~~\$0.035~~ \$0.005/GB queried data
- REST API: Pandas, R, ActiveRecord, Fluentd...

bigquery.cloud.google.com

developers.google.com/bigquery/

How BigQuery works

Tree Structured Query Dispatch and Aggregation

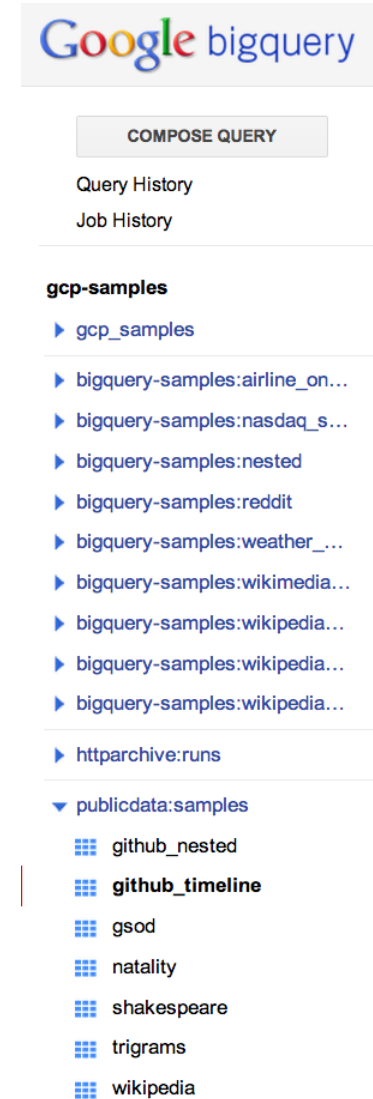


*Coming up next!
(with Dirk Primbs)*

BigQuery for Open Data

Massive open datasets available on BigQuery

- Wikimedia pageviews (68B rows)
- HTTP Archive (1.2B rows)
- NASDAQ stock quotes (903M rows)
- GitHub push logs (202M rows)
- Natality in US (128M rows)
- GSOD Weather (121M rows)
- GDELT
- etc



The screenshot displays the Google BigQuery interface. At the top, the 'Google bigquery' logo is visible. Below the logo, there is a 'COMPOSE QUERY' button. Underneath, there are links for 'Query History' and 'Job History'. The main content area shows a list of datasets under the 'gcp-samples' project. The 'publicdata:samples' project is expanded, showing a list of datasets including 'github_nested', 'github_timeline', 'gsod', 'natality', 'shakespeare', 'trigrams', and 'wikipedia'. A red vertical line is present on the left side of the 'github_timeline' dataset entry.

3 steps to data processing

- Load
- Query
- Output

How many triples do we have?



```
SELECT COUNT (*) triples  
FROM [fh-bigquery:freebase20140119.triples]
```

2,123,637,994 facts

How many triples are people?



```
SELECT COUNT(sub) people
FROM [fh-bigquery:freebase20140119.triples]
WHERE obj='/people/person'
AND pred='/type/object/type'
```

3,036,682 people

What do we know about them?



```
SELECT TOP (a.pred), COUNT(*)
FROM [triples] a
JOIN EACH (
  SELECT sub
  FROM [triples]
  WHERE obj='/people/person'
  AND pred='/type/object/type'
) b
ON a.sub=b.sub
```

/type/object/key	18492679
/type/object/type	10624553
/common/topic/topic_equivalent_webpage	10161398
/type/object/name	6039443
/music/artist/track	5202847
/common/topic/description	4398659
/common/topic/notable_types	3033217
/common/topic/notable_for	3033061
/award/award_nominee/award_nomi..	2794197
/people/person/gender	2031455
/book/author/works_written	1824391
/freebase/valuenotation/has_value	1559924
/people/person/profession	1438080
/common/topic/article	1406994
/award/award_winner/awards_won//...	1332039
/people/person/date_of_birth	1281298

Counting people by gender



```
SELECT obj gender, count(*) c
FROM [triples]
WHERE pred='/people/person/gender'
GROUP BY 1
```

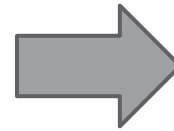
gender	c
/m/05zppz	1521700
/m/02zsn	511361
/m/04j3vhk	230

4.5s elapsed, 106 GB processed

Same with their dates of birth



```
SELECT a.sub sub, a.obj, date_of_birth
FROM [triples] a
JOIN EACH (
  SELECT sub
  FROM [triples]
  WHERE obj='/people/person'
  AND pred='/type/object/type'
) b
ON a.sub = b.sub
WHERE a.pred =
  '/people/person/date_of_birth'
```

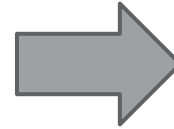


```
[fh-bigquery:
freebase20140119.
people_date_of_birth
]
```


Transforming dates into ages

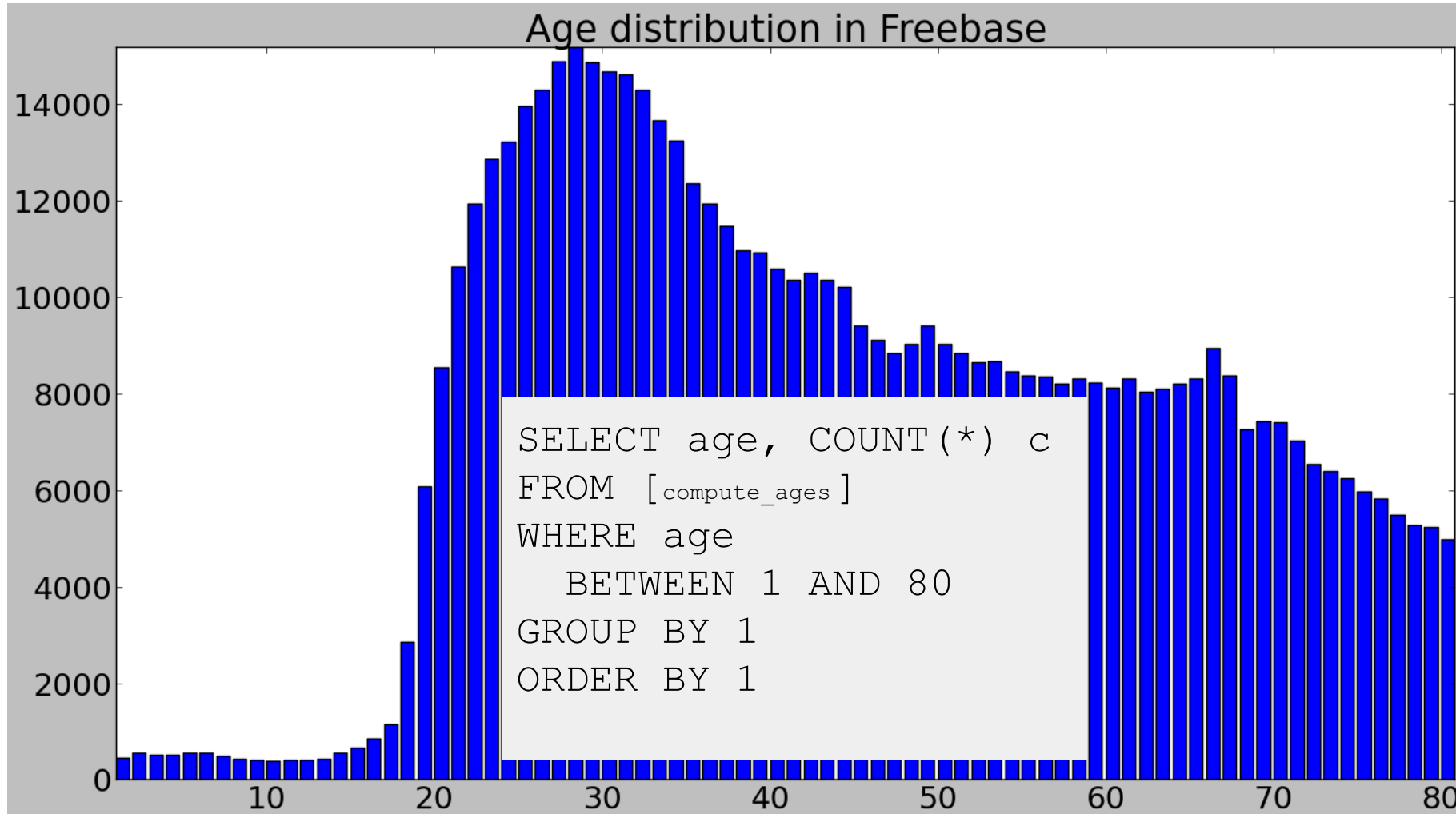


```
SELECT
  sub,
  TIMESTAMP(date_of_birth +
    ' 00:00:00') date_of_birth,
  INTEGER(DATEDIFF(
    USEC_TO_TIMESTAMP(NOW()),
    TIMESTAMP(date_of_birth + ' 00:00:00'))
    / 365.5) age
FROM [people_date_of_birth]
HAVING date_of_birth IS NOT NULL
```

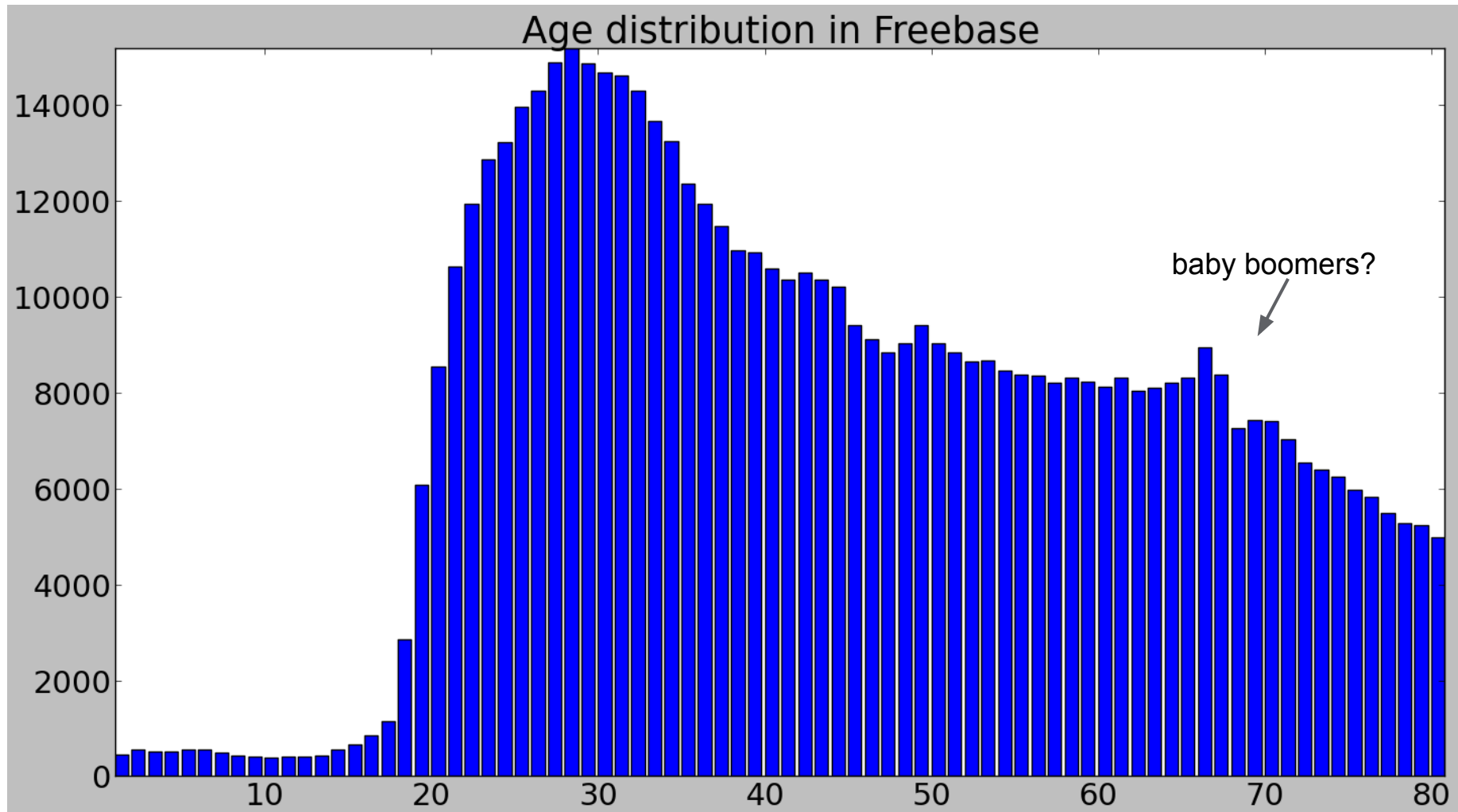


```
[fh-bigquery:
freebase20140119.
compute_ages]
```

Age distribution



Age distribution

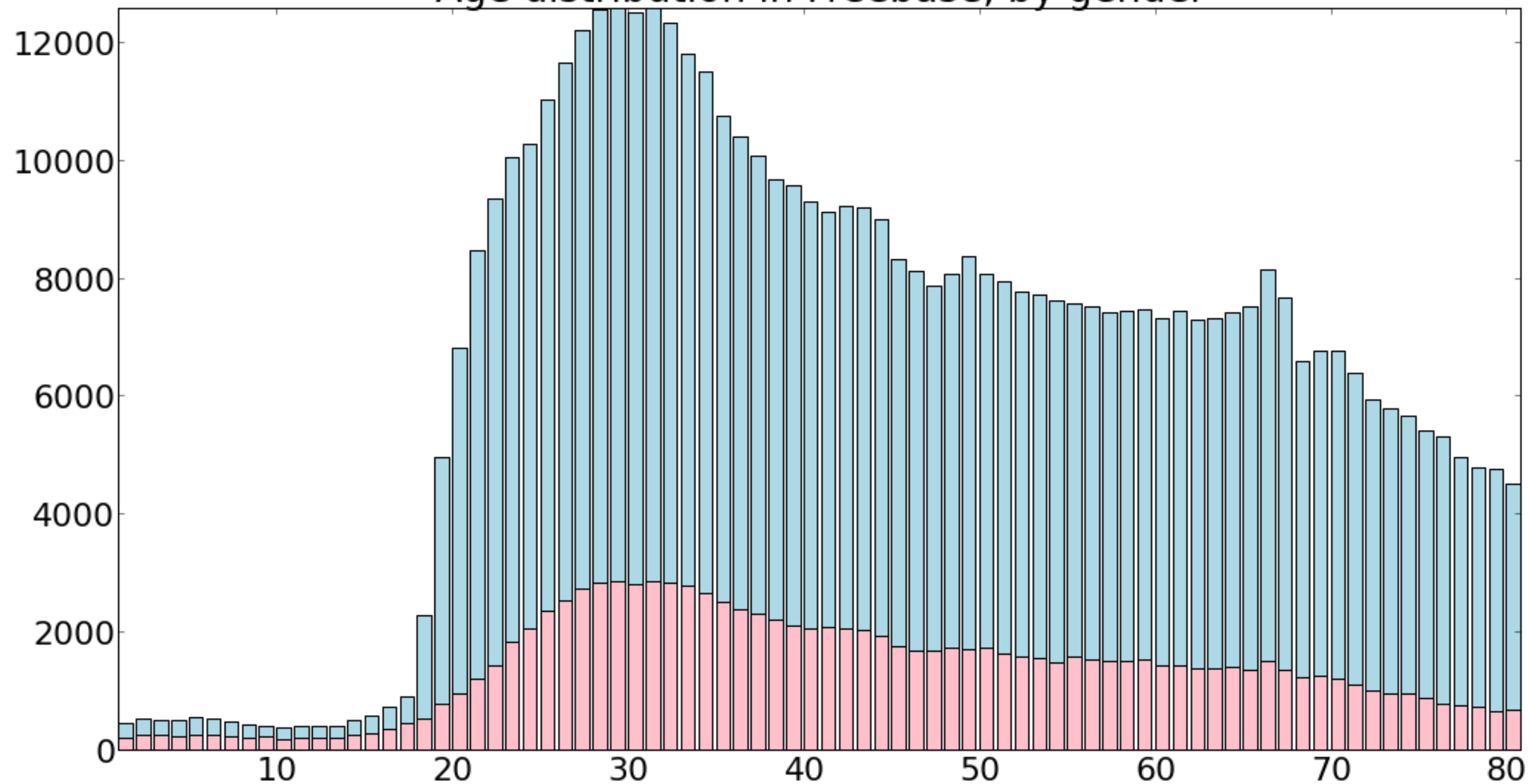


Same, divided by gender

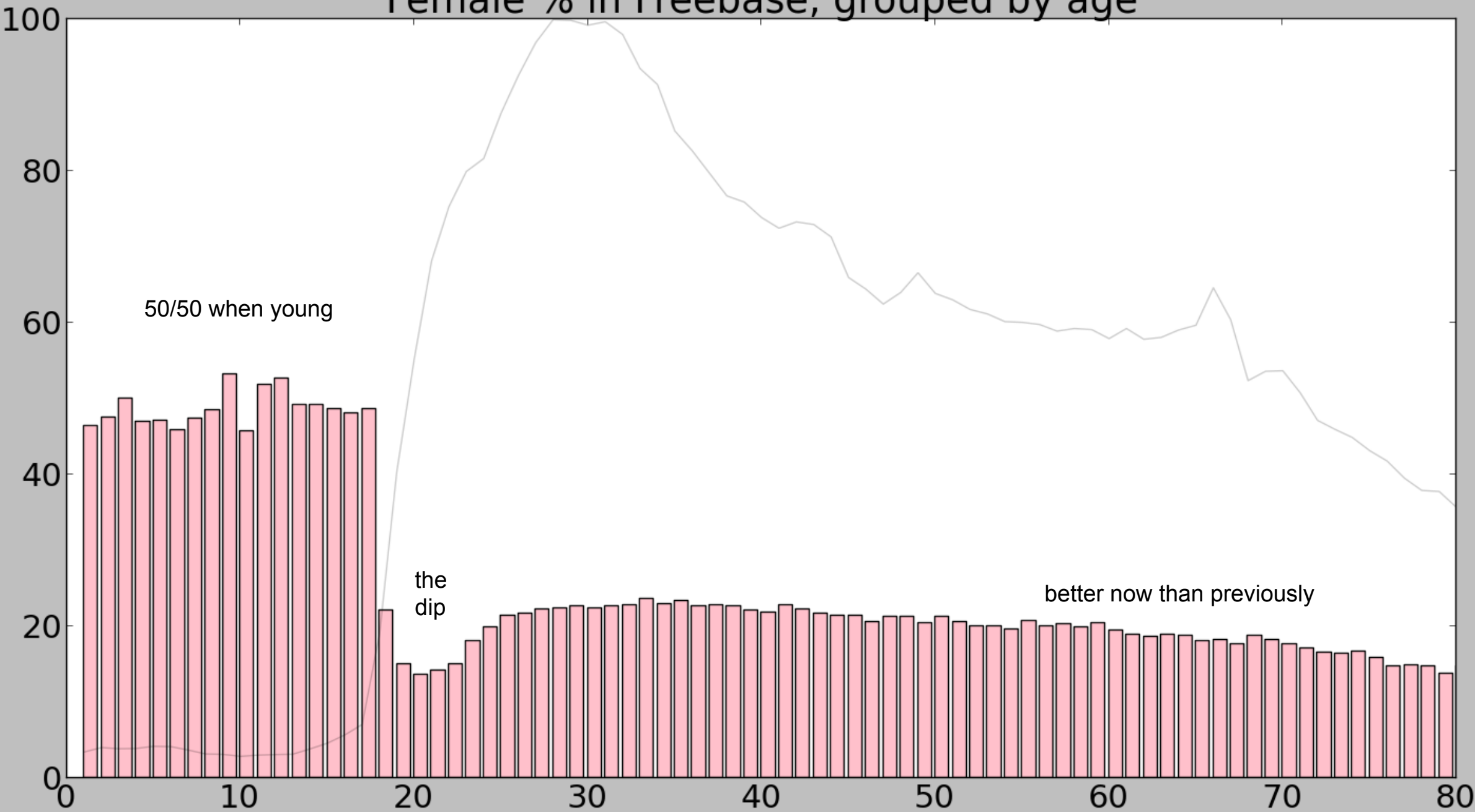


```
SELECT age,  
       COUNT (IF (gender=' /m/02zsn' , 1, null)) female,  
       COUNT (IF (gender=' /m/05zppz' , 1, null)) male,  
       COUNT (*) c  
FROM [fh-bigquery:freebase20140119.compute_ages] a  
JOIN EACH [fh-bigquery:freebase20140119.people_gender] b  
ON a.sub = b.sub  
WHERE age BETWEEN 1 AND 80  
GROUP BY 1  
ORDER BY 1
```

Age distribution in Freebase, by gender



Female % in Freebase, grouped by age

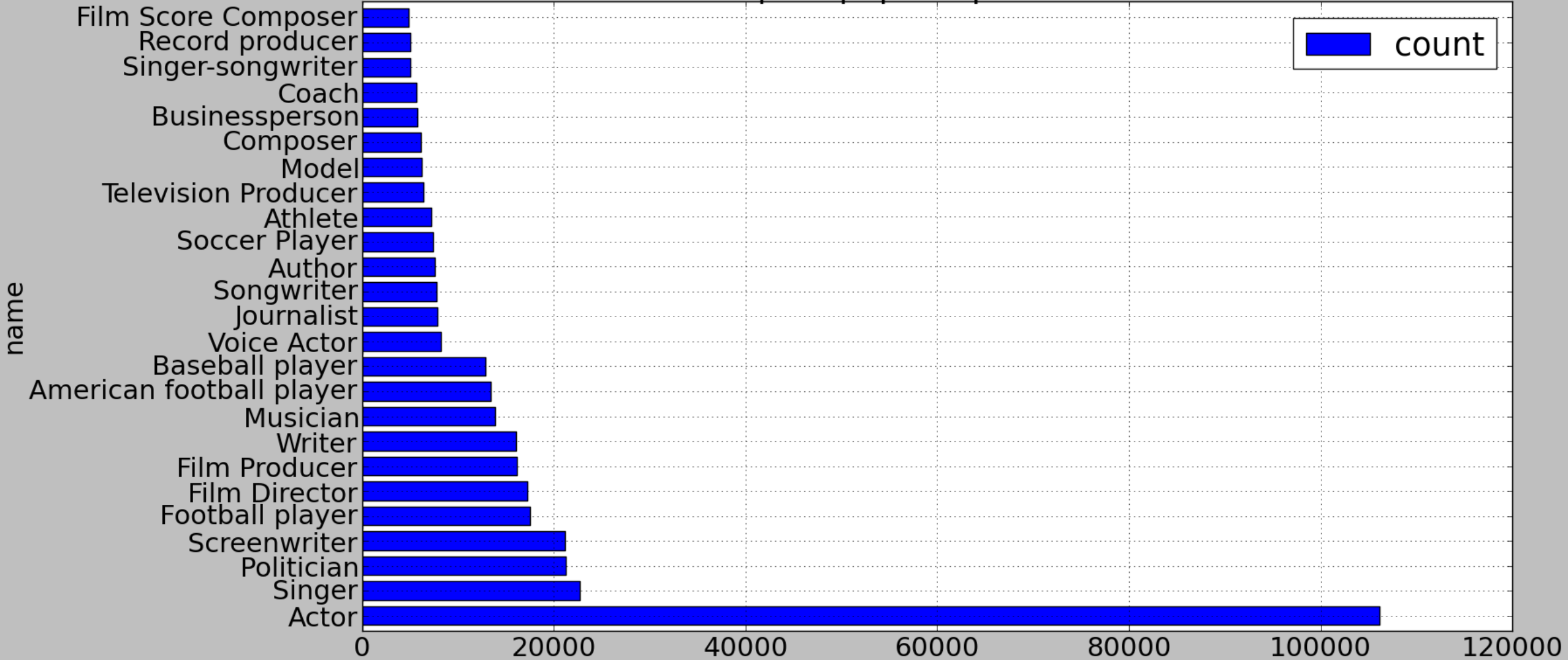


Top 25 professions

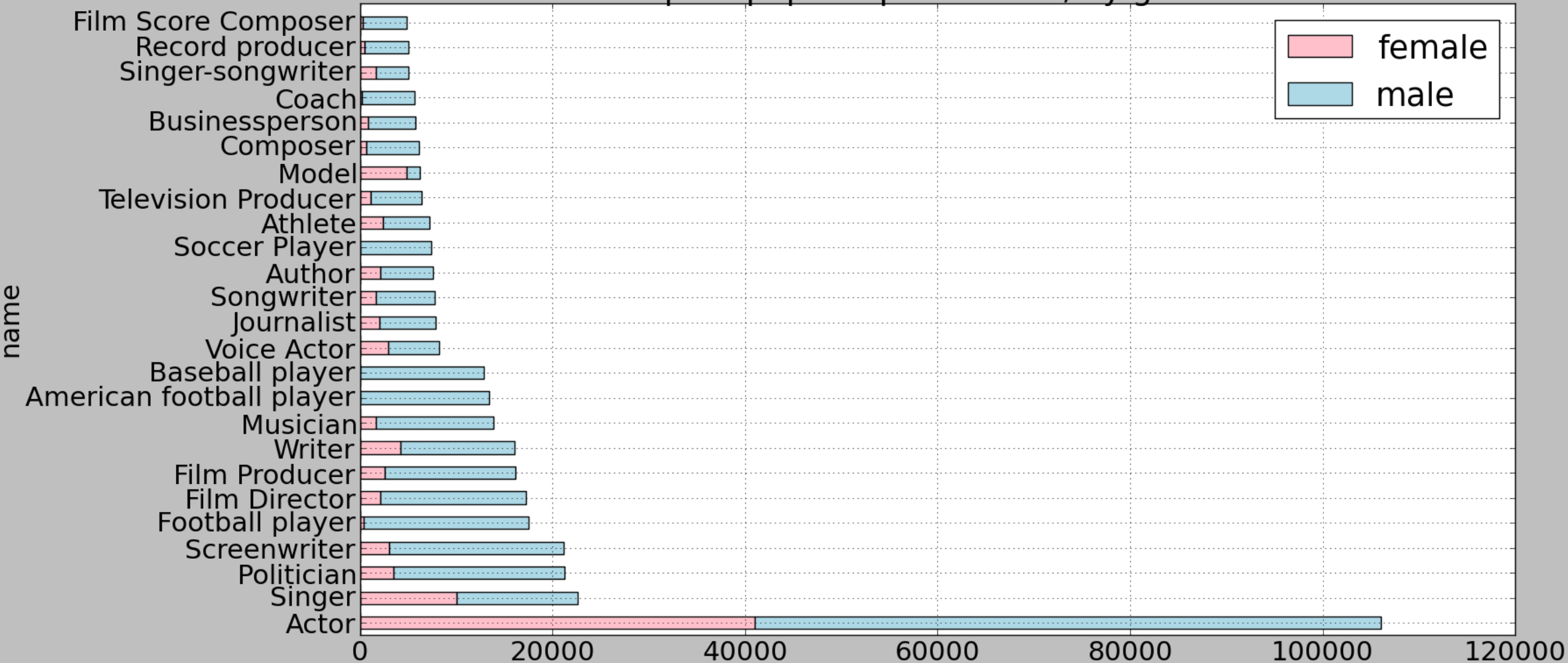


```
SELECT profession, name,  
       COUNT(IF(c.gender='/m/02zsn', 1, null)) female,  
       COUNT(IF(c.gender='/m/05zppz', 1, null)) male,  
       COUNT(*) count  
FROM [people_profession] a  
JOIN [profession_names] b  
ON a.profession=b.sub  
JOIN EACH [people_gender] c  
ON a.sub=c.sub  
JOIN EACH [compute_ages] d  
ON a.sub=d.sub  
WHERE d.age BETWEEN 0 AND 100  
GROUP BY 1, 2 ORDER BY 5 DESC LIMIT 25
```

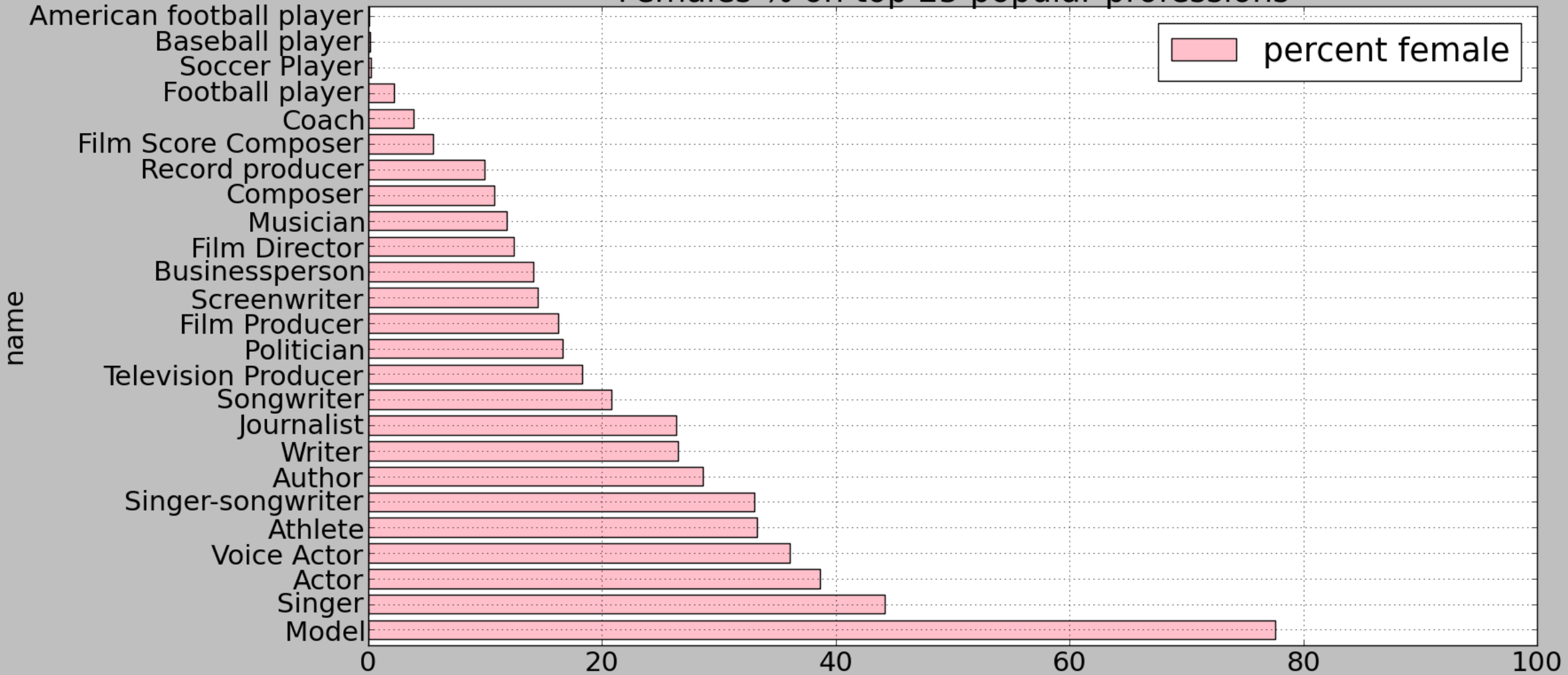
Top 25 popular professions



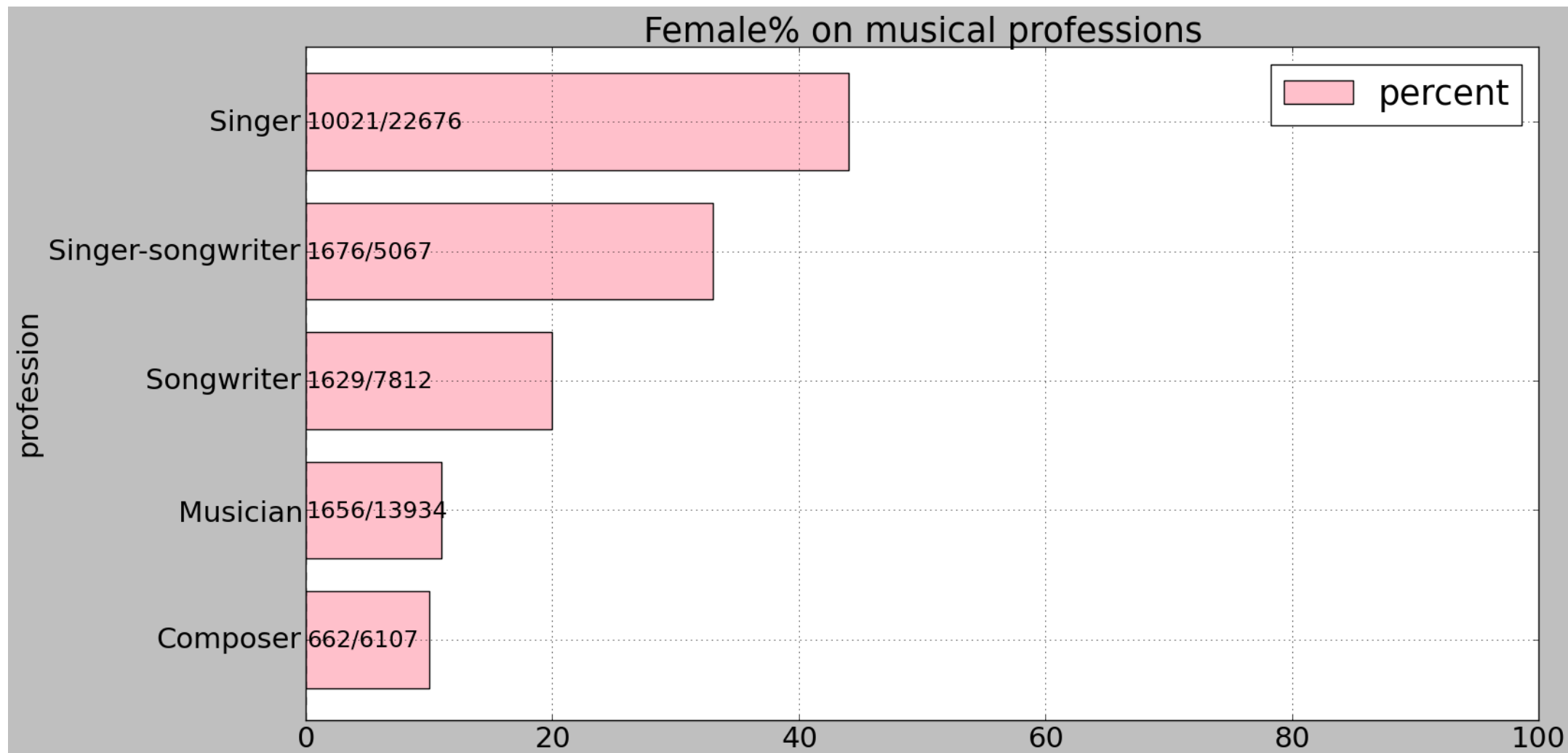
Top 25 popular professions, by gender



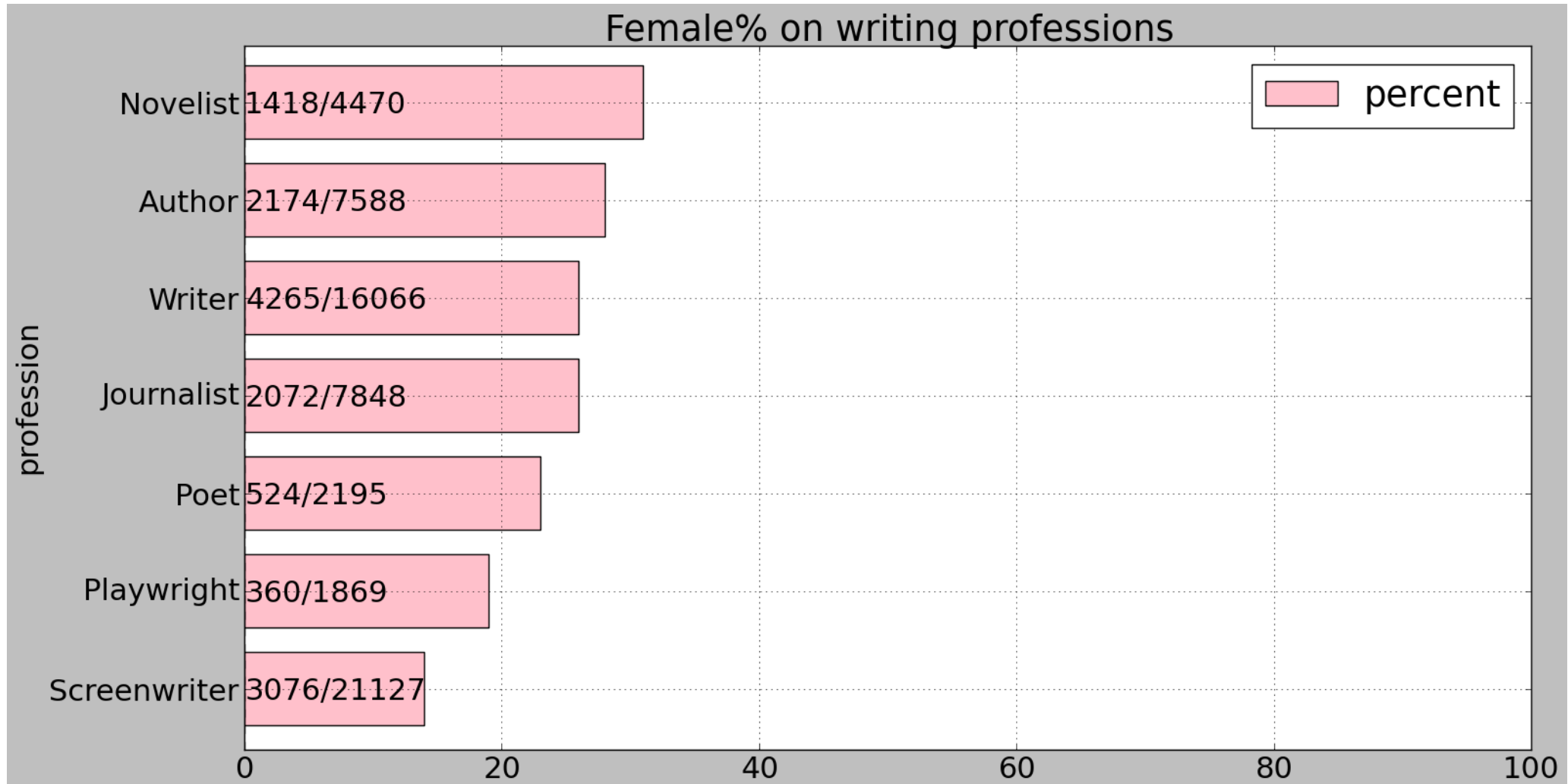
Females % on top 25 popular professions



Focus on music



Focus on writing



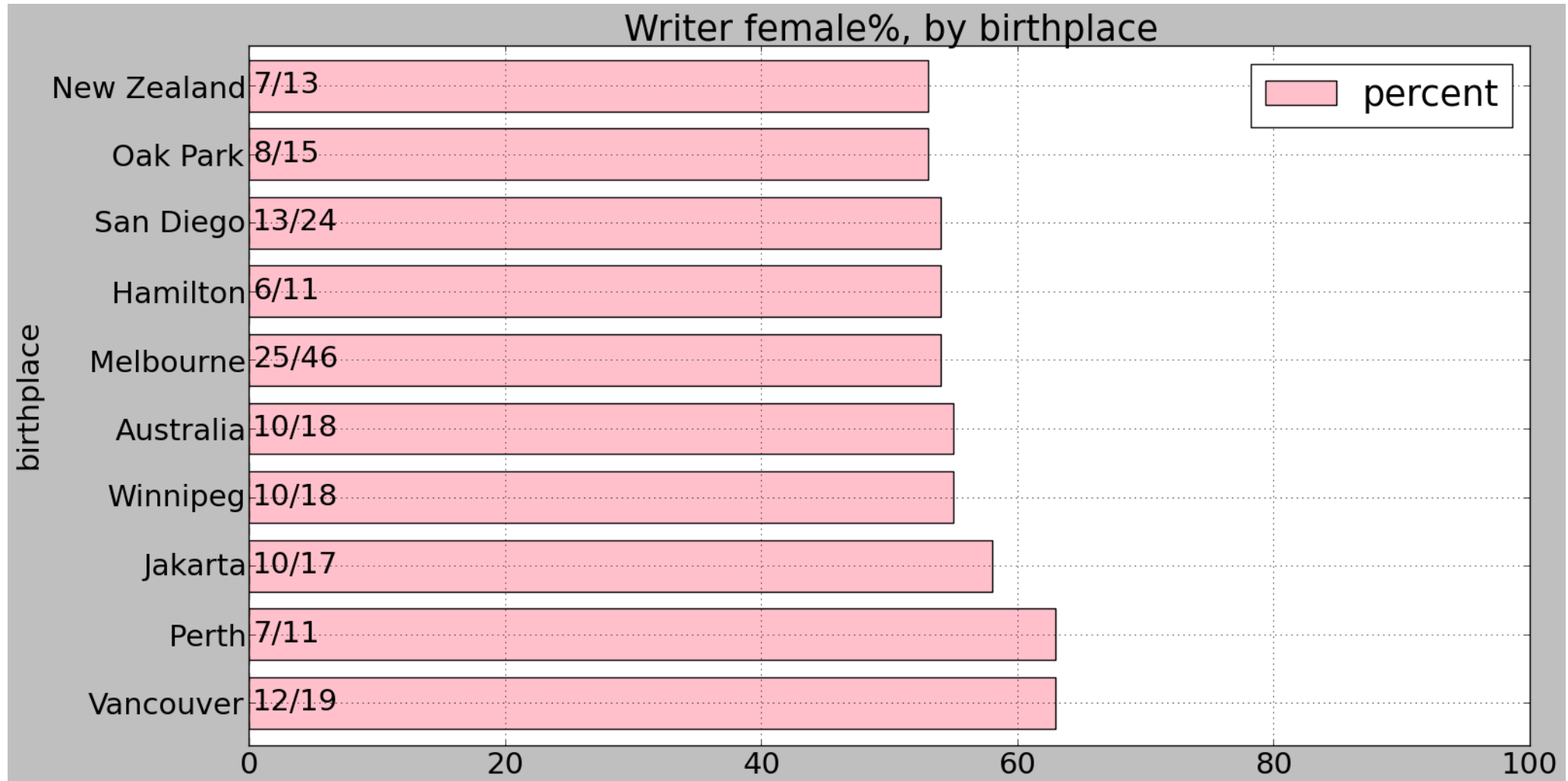
But it's different across the world... let's see by
place of birth

Writers by place of birth

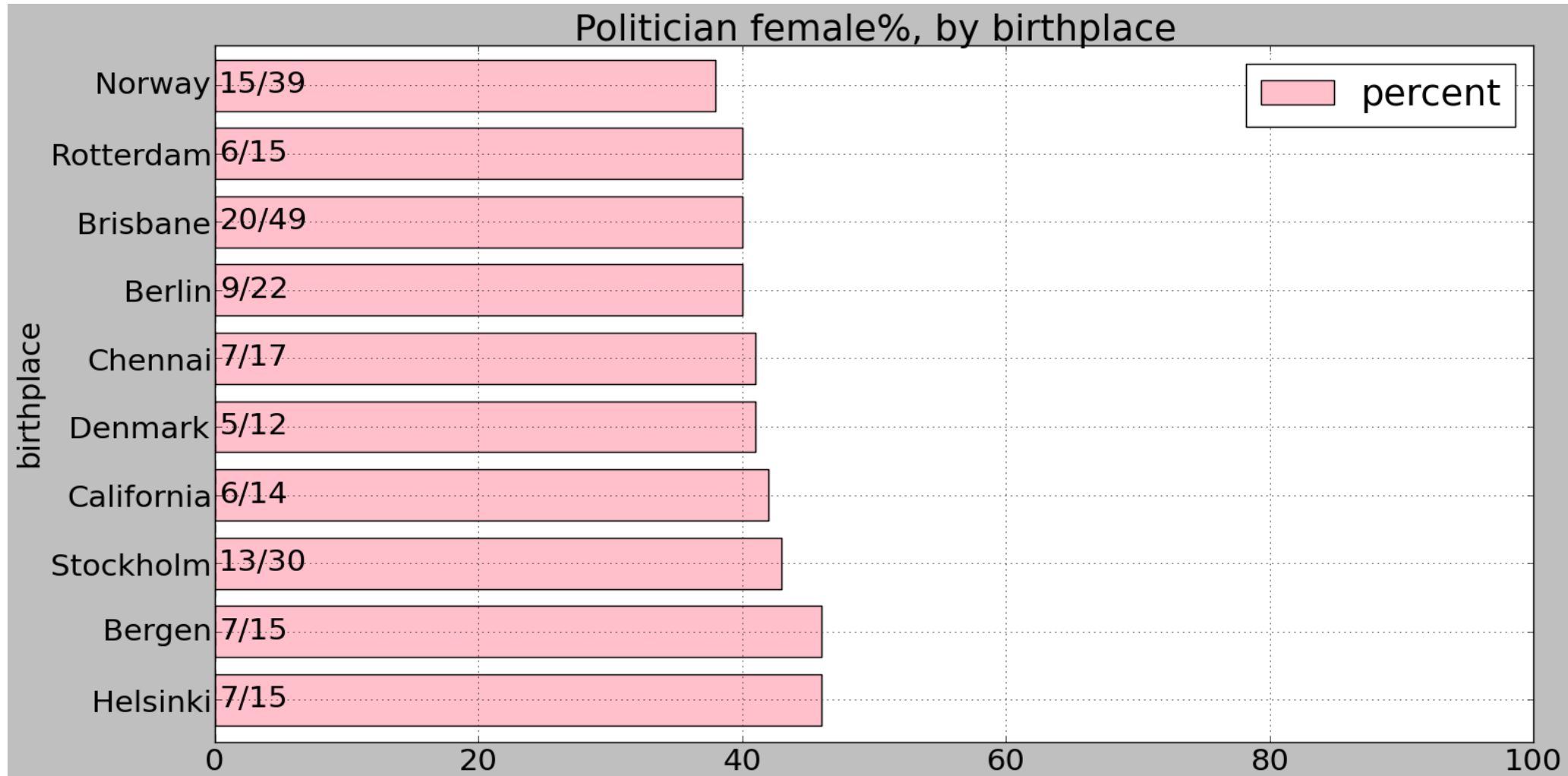


```
SELECT REGEXP_REPLACE(f.name, '[^a-zA-Z ]*', '') birthplace,
       COUNT(IF(c.gender='/m/02zsn', 1, null)) female,
       INTEGER(100*COUNT(IF(c.gender='/m/02zsn', 1, null))/COUNT(*)) percent,
       COUNT(*) c
FROM [fh-bigquery:freebase20140123.people_profession] a
JOIN [fh-bigquery:freebase20140123.profession_names] b
ON a.profession=b.sub
JOIN EACH [fh-bigquery:freebase20140123.people_gender] c
ON a.sub=c.sub
JOIN EACH [fh-bigquery:freebase20140123.compute_ages] d
ON a.sub=d.sub
JOIN EACH [fh-bigquery:freebase20140123.people_place_of_birth] e
ON a.sub=e.sub
JOIN [fh-bigquery:freebase20140123.place_of_birth_names] f
ON e.place_of_birth=f.sub
WHERE d.age BETWEEN 0 AND 100
AND b.name IN ('Writer')
GROUP BY 1 HAVING c > 10 ORDER BY 3 DESC LIMIT 10
```

Writers by place of birth



Politicians by place of birth

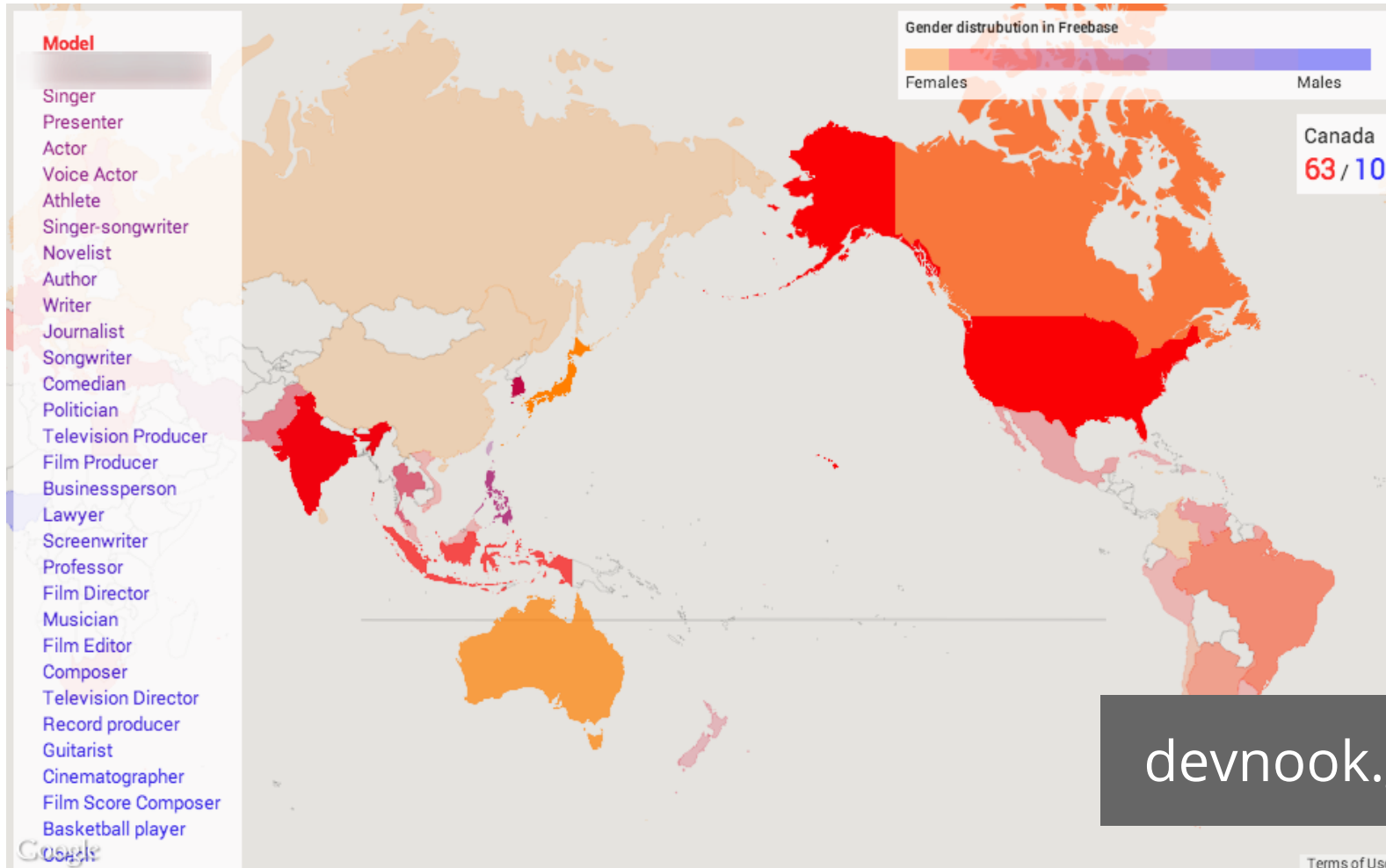


As we are talking about geo... let's put it on a map

Data visualization



Map Visualization



- Sanity check
- Explore your data
- Encounter surprising results

devnook.github.io/GenderMaps

Explore the data on the map



Sanity check

- *Is some data missing?*
- *Are some areas over-represented and therefore skewing the results?*
- *Are there regional or cultural inconsistencies in the dataset?*

Explore gender gap geographically

- *Songwriter*
- *Politician*

How does it work



Countries dataset

+

Gender gap data



GeoJSON

`google.maps.Map.data.addGeoJson`



www.naturalearthdata.com



BigQuery



Geojson file



Google Maps API v3

Encountered problems



- Not all data points in the dataset have the same granularity
 - *country, state, city levels*
 - *need to perform aggregation*
- Absolute numbers can be misleading
 - *can present numbers relative to population*

The future

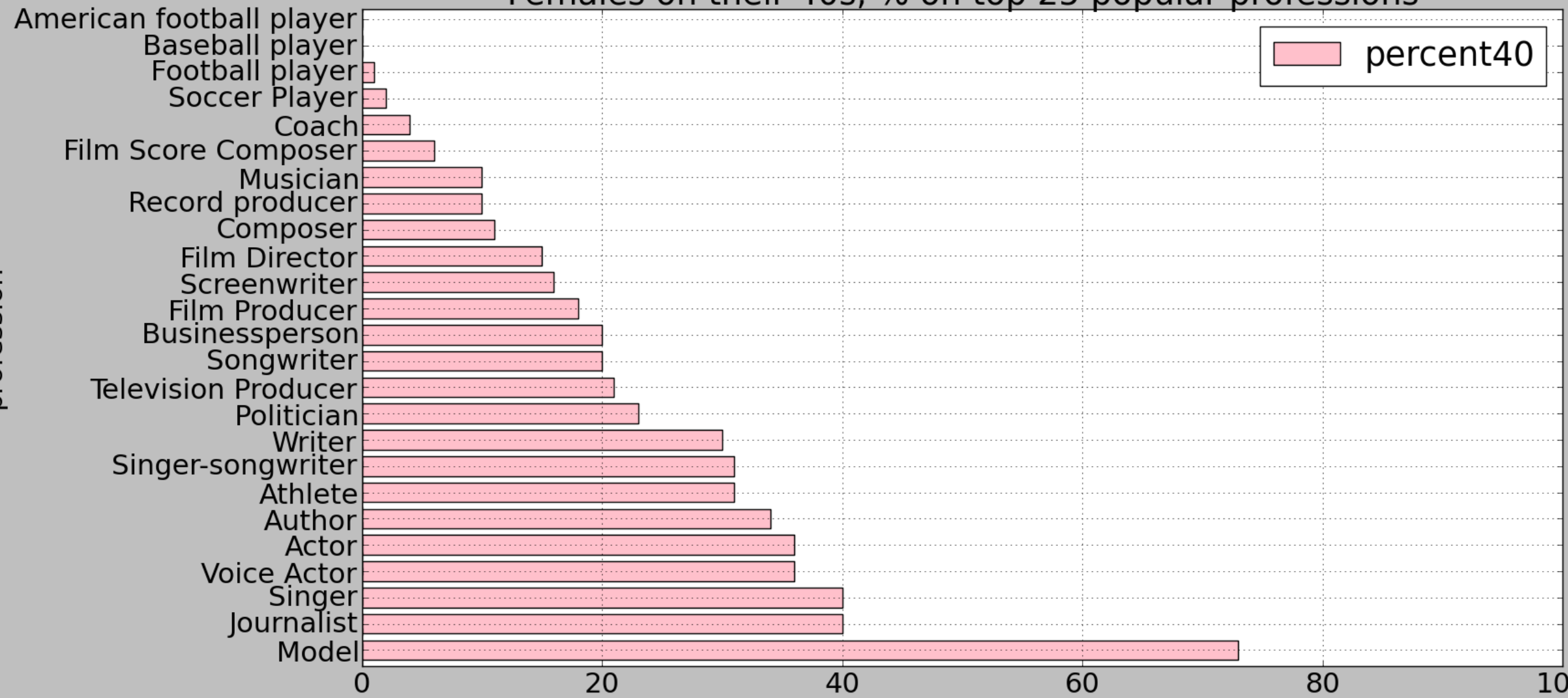


Can we look into the future?

What are the trends, how things are changing

What's the picture if we focus only within people between
40 and 50 years old:

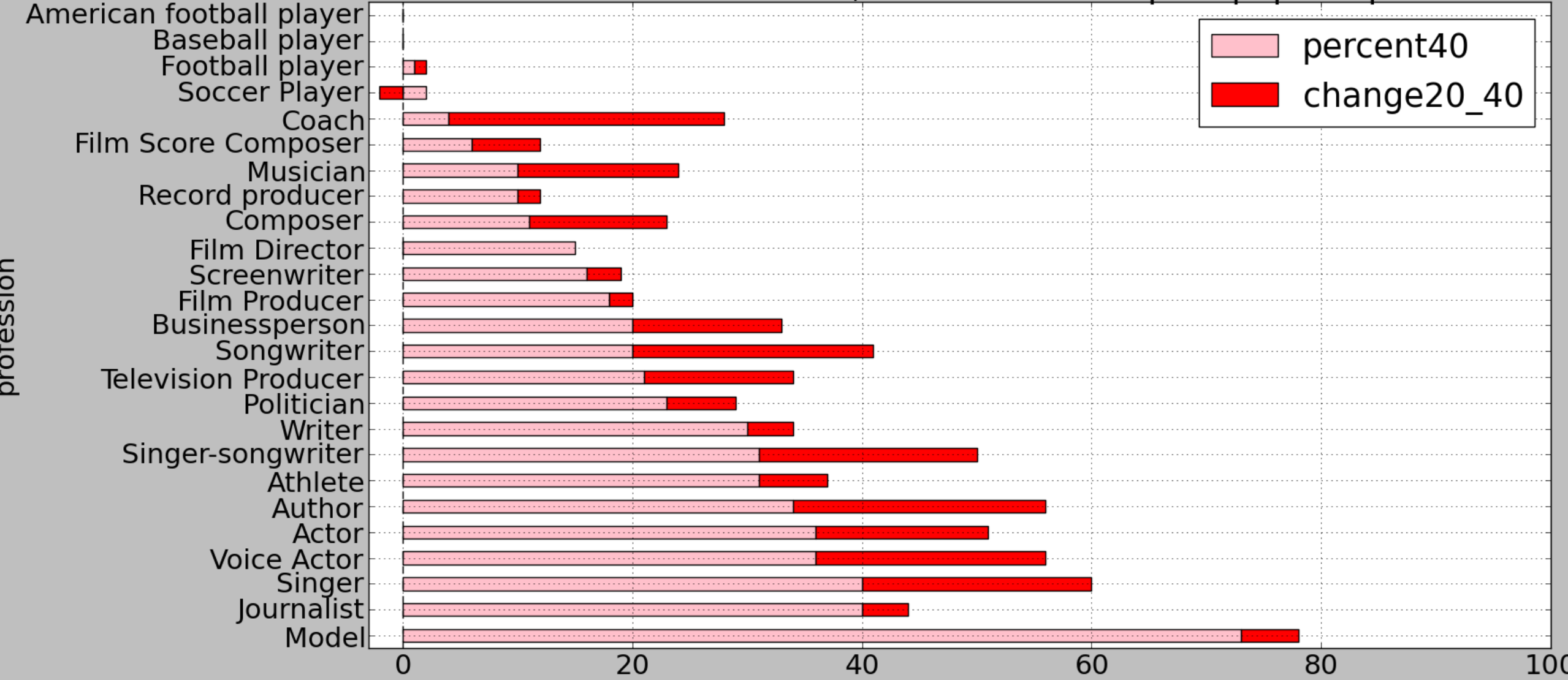
Females on their 40s, % on top 25 popular professions



That's the picture if we focus only into people between 40 and 50 years old.

How the balance changes, if we compare it with 20-30 year olds.

Females on their 40s and 20s, % increase on top 25 popular professions



In summary



You learned:

- What's Freebase.
- How to use BigQuery to explore Freebase.
- Visualize on maps

Action items:

- Explore Freebase, BigQuery.
- Visualize, use maps.
- Change the world: <http://www.google.com/diversity/>



Exploring the Notability Gender Gap

Freebase, BigQuery, Maps (Berlin Buzzwords)

Google Developer Relations:

[Felipe Hoffa](#)

[Ewa Gasperowicz](#)

[@felipehoffa](#)

[@devnook](#)

Most visited female politicians



```
SELECT title, SUM(requests) c
FROM [wikipedia_views_201308_en_top_titles_views]
WHERE title IN
( SELECT REGEXP_REPLACE(obj, '/wikipedia/id/', '')
  FROM [triples_nolang]
  WHERE sub IN
    ( SELECT a.sub sub
      FROM [people_profession] a
      JOIN EACH [people_gender] b
      ON a.sub=b.sub
      WHERE profession = '/m/0fj9f'
        AND b.gender = '/m/02zsn')
  AND obj CONTAINS '/wikipedia/id/'
  AND pred = '/type/object/key'
GROUP BY 1) GROUP BY title ORDER BY c DESC
```

Most visited books written by a woman



```
SELECT title, SUM(requests) c
FROM [wikipedia_views_201308_en_top_titles_views]
WHERE title IN (
    SELECT REGEXP_REPLACE(obj, '/wikipedia/id/', '')
    FROM [triples_nolang] WHERE sub IN (
        SELECT sub FROM [triples_nolang]
        WHERE pred = '/book/written_work/author'
        AND obj IN (
            SELECT sub FROM [people_gender]
            WHERE gender = '/m/02zsn'))
    AND obj CONTAINS '/wikipedia/id/'
    AND pred = '/type/object/key'
    GROUP BY 1)
GROUP BY title
ORDER BY c DESC;
```