OkCupid Scraper

Web Scraping Project by Fangzhou Cheng

We found love in a hopeless place

- 44% of adult Americans are single, which means 100 million people out there!
  - in New York state, it's 50%
  - in DC, it's 70%

- 40 million Americans use online dating services. That's about 40% of our entire U.S. single-people pool.

Source: http://www.match.com/magazine/article/4671/

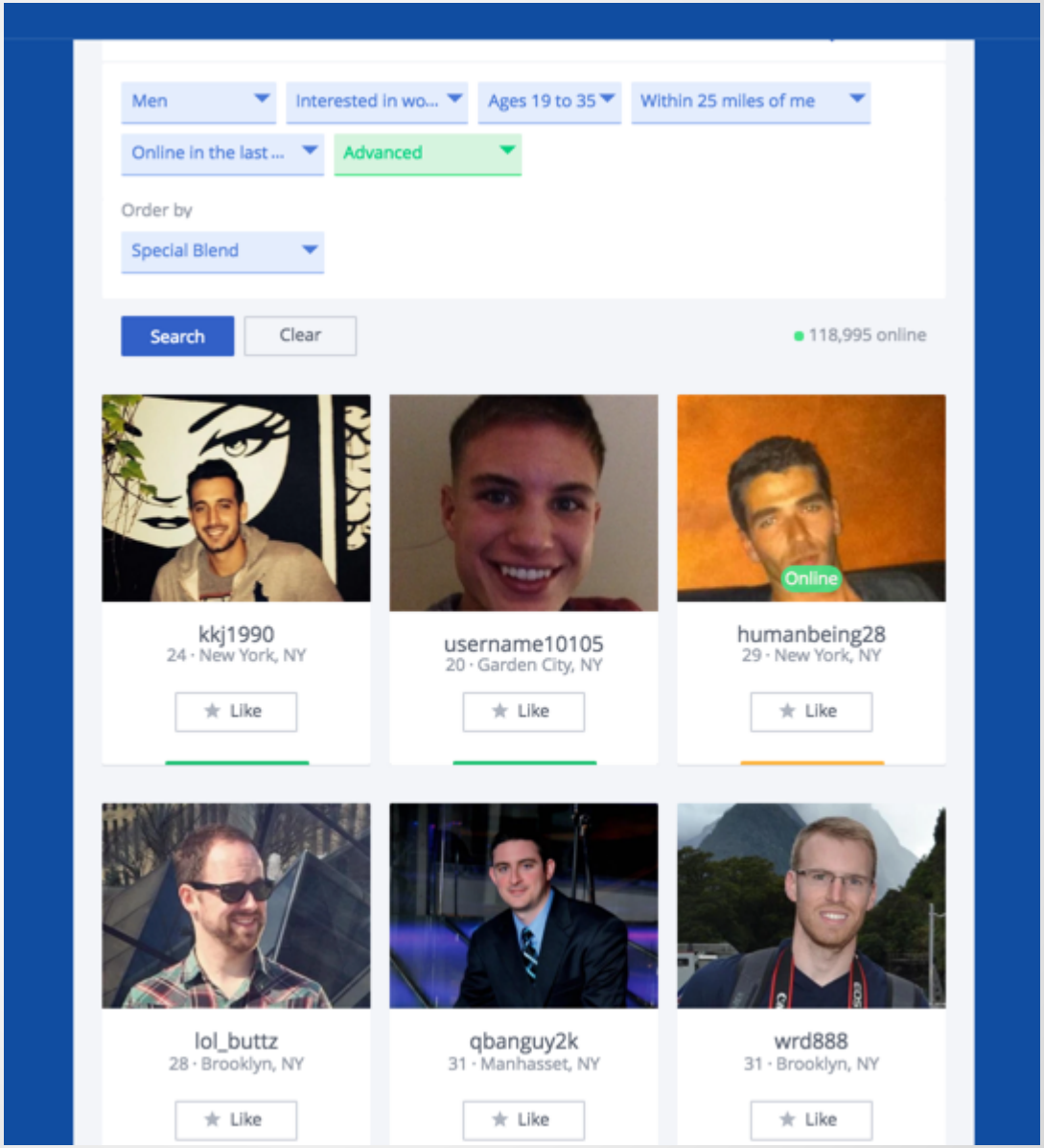
Why choose OkCupid?

- Word of mouth from my friends

- Large user base: OkCupid has around 30M total users and gets over 1M unique users logging in per day.
  - its demographics reflect the general Internet-using public

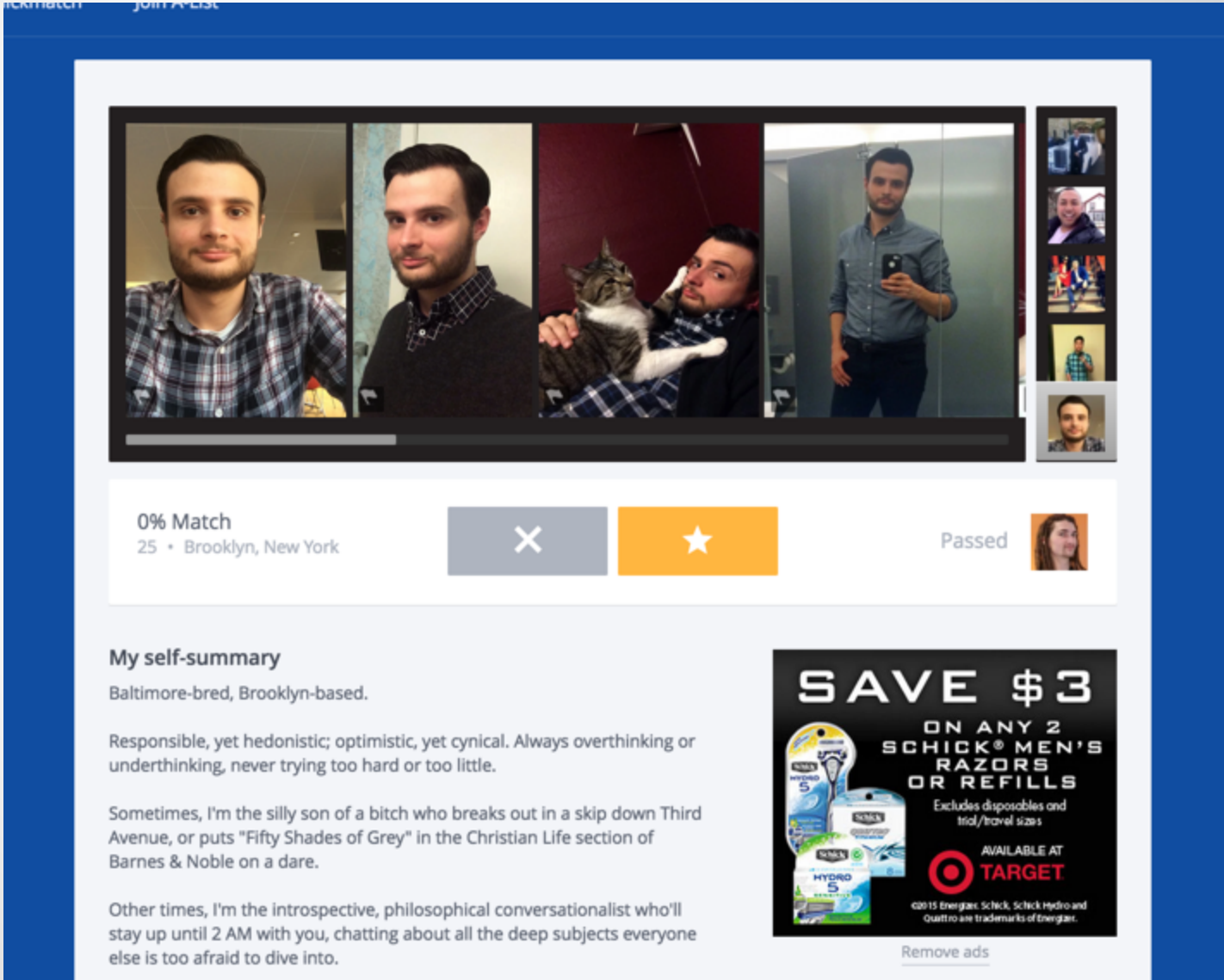
Page

Browse Matches



Page

Quick match



Page

Profile

0% Match | 0% Enemy

Send a Message

Like

Tudo_Bom_

27 • New York, NY (3 miles) • Man

About | Photos | The Two of Us

My self-summary

Style guy next door seeks feisty trouble bringer full of sass and moxy for some high jinx and highballs.

Born and raised on Long Island, but despite my Italian roots I've always been more of a pretty boy badass than other stereotypes you might imagine. I played polo at Cornell, used to shuck oysters professionally and now I take trainees through 'Suit School' at a Dutch fashion company.

Things are busy but I have a couple days off to fill, so get in touch!

What I'm doing with my life

I've been through a few story worthy jobs; used to organize high-end events with oysters; now I'm selling suits and training other sales professionals how to be successful.

I'm really good at

The Jeopardy! Teen Tournament

The first things people usually notice about me

I'm like a young Clark Kent

Favorite books, movies, shows, music, and food

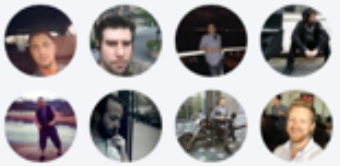
Lonesome Dove

The most private thing I'm willing to admit

Browse invisibly

Remove ads

Similar users



I'm looking for

- Women
- Ages 21–32

I'm looking for

- Women
- Ages 20–38
- Near me
- For new friends, long-term dating, short-term dating

Steps

1. Web Scraping
2. Data Cleaning
3. Data Manipulation
4. Geographic Visualization

Step 1. Web Scraping

1. Get usernames from matches browsing.
    - Create a profile with only the basic and generic information.
    - Get cookies from login network response.
    - Set search criteria in browser and copy the URL.
2. Scrape profiles from unique user URL using cookies

    www.okcupid.com/profile/username

Problems Encountered

Speed!

varies from 1 to 2 sec per profile

------after visiting about 2,000 profiles------->

varies from 5 min to forever per profile, which equals killing my program

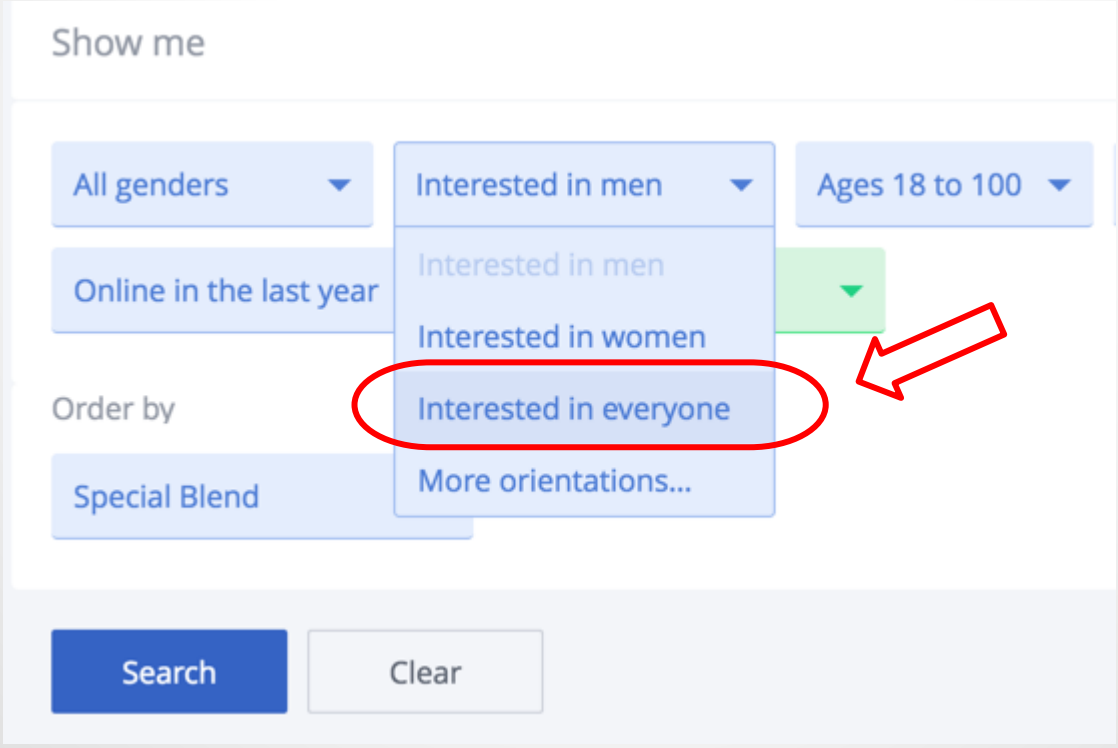
What I did:

- Create multiple profiles and generate a set of cookies
- Keep the speed around 2,000 profiles every a few hours
- Monitor the progress frequently
- Change cookies when the program is killed
- It eventually took my about 2 days to make the html5 parser and cookies work, and finish all scraping

Problems Encountered

Searching Strategy!

E.g. how do you understand this?



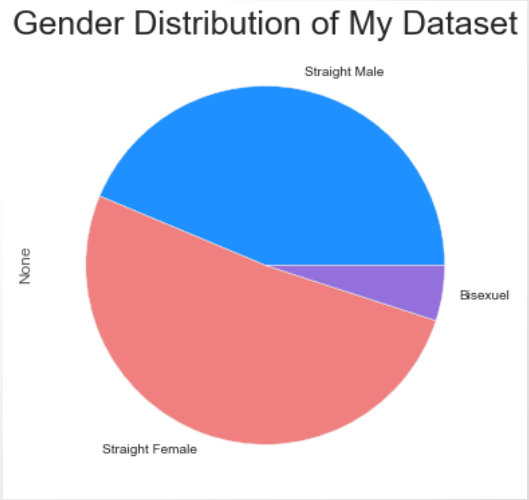
Samples

Usernames:

User Group	Total Usernames	Unique Usernames	Unique Percentage
Bisexual All Gender	300,000	23,565	7.855%
Straight Male	30,000	20,565	68.55%
Straight Female	30,000	24,195	80.65%

Profiles:

User Group	Total
Bisexual All Gender	782
Straight Male	2,054
Straight Female	2,412



What does the data look like?

- User basic information
  - gender, age, location, orientation, ethnicities, height, body type, diet, smoking, drinking, drugs, religion, sign, education, job, income, status, monogamous, children, pets, languages
- User matching information
  - gender orientation, age range, location, single, purpose
- User self-description
  - summary, what they are currently doing, what they are good at, noticeable facts, favourite books/movies, things they can’t live without, how to spend time, friday activities, private thing, message preference

Step 2. Data Cleaning

- Data cleaning in web scraping
  - Missing values
  - Consistent data types

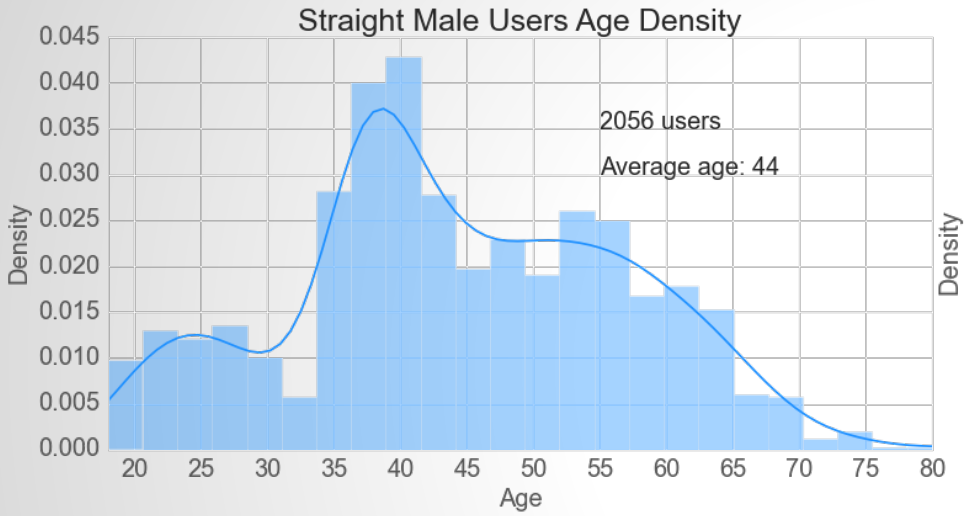
- Data cleaning in data manipulation
  - Get latitude and longitude of user location from python library geopy
  - Use regular expression to get height, age range and state/country information from long string

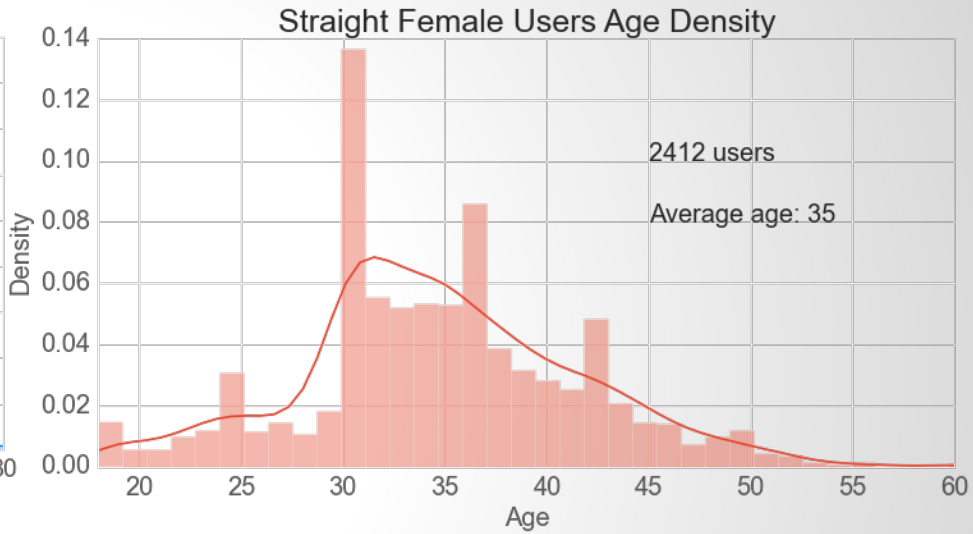
Step 3. Data Manipulation

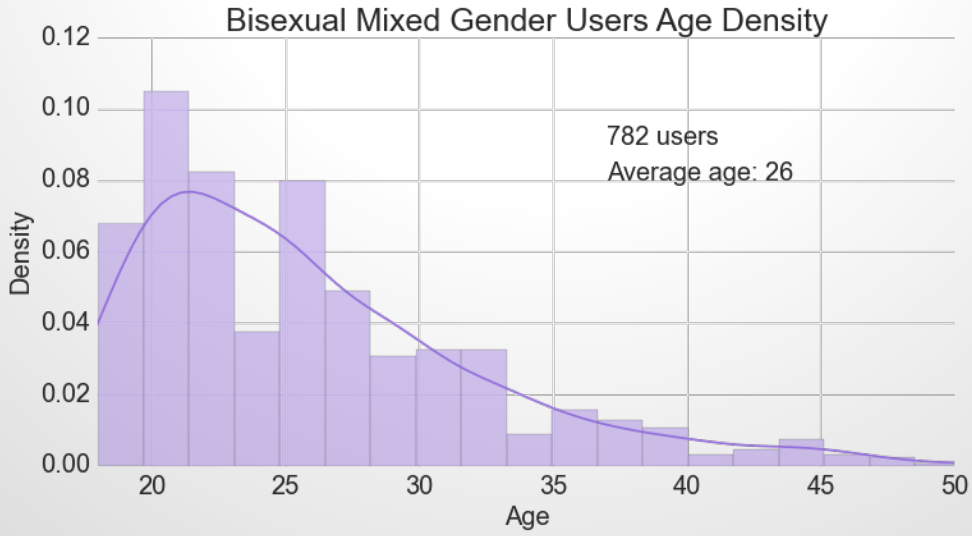
1. Demographics Analysis
   a. How old are they?
   b. Where are they located?

2. Psychological Analysis
   a. Who are pickier?
   b. Who are lying?

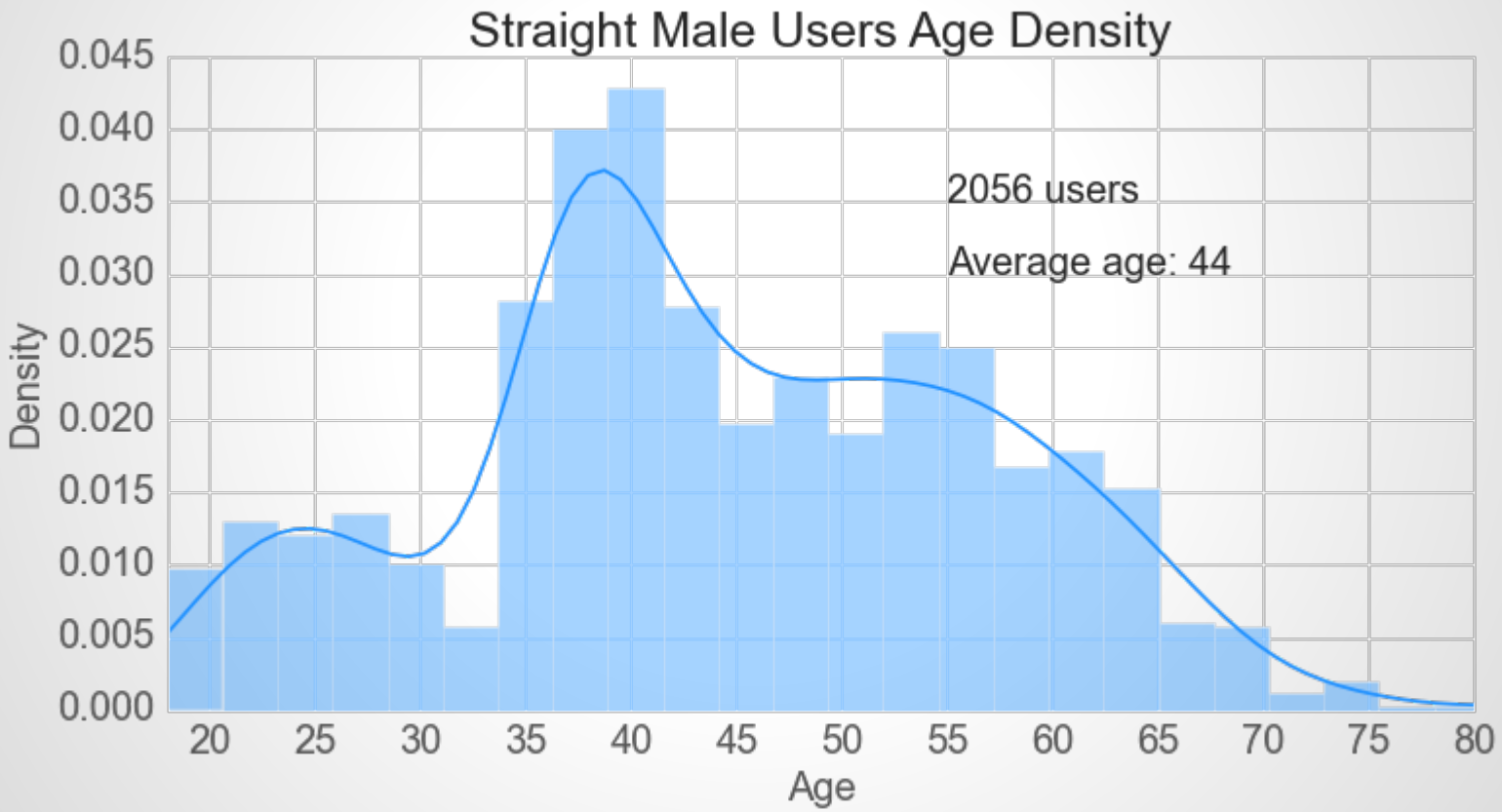
Age


Straight Male Users Age Density
2056 users
Average age: 44
Density
Age
0.045
0.040
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
20 25 30 35 40 45 50 55 60 65 70 75 80

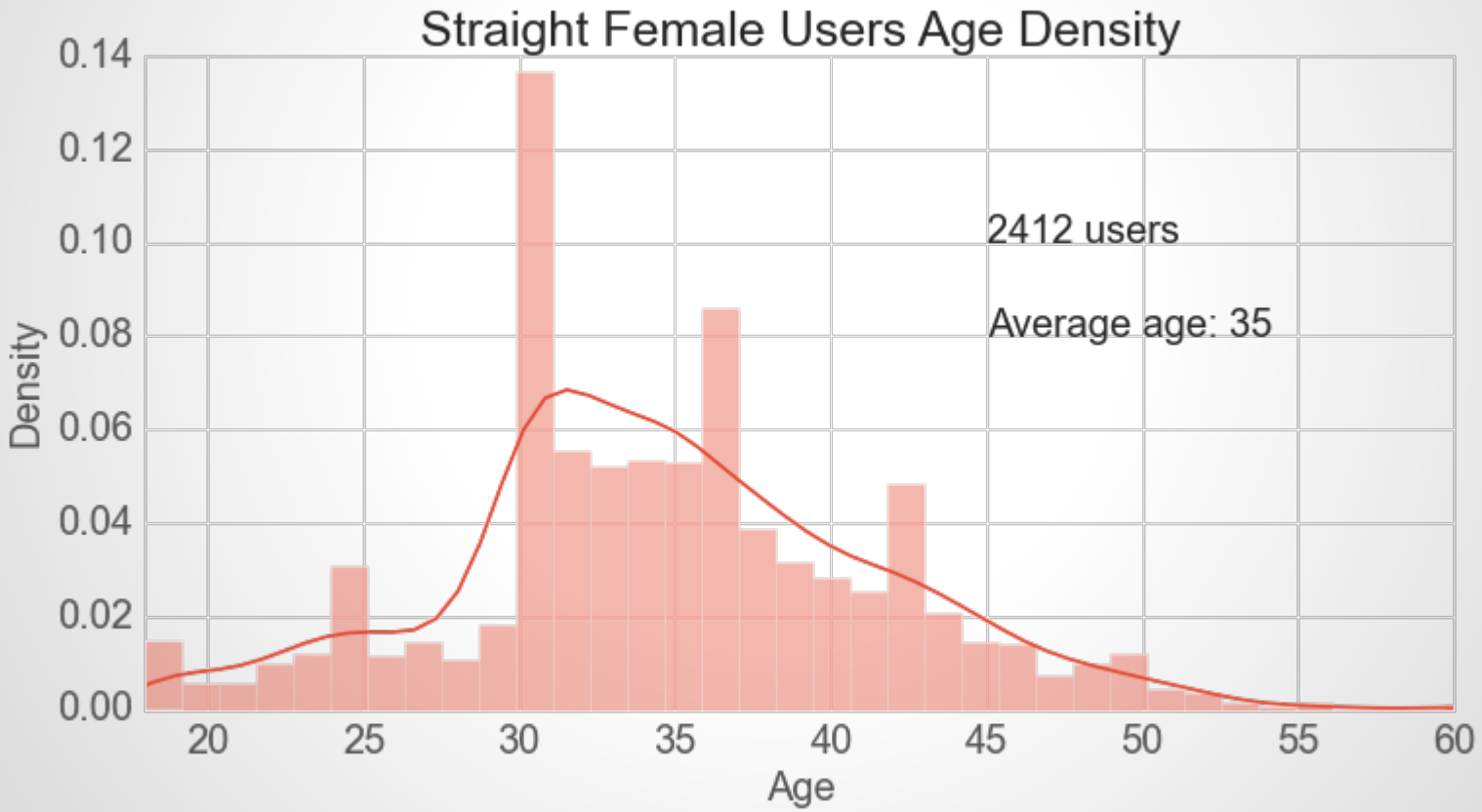

Straight Female Users Age Density
2412 users
Average age: 35
Density
Age
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
20 25 30 35 40 45 50 55 60


Bisexual Mixed Gender Users Age Density
782 users
Average age: 26
Density
Age
0.12
0.10
0.08
0.06
0.04
0.02
0.00
20 25 30 35 40 45 50

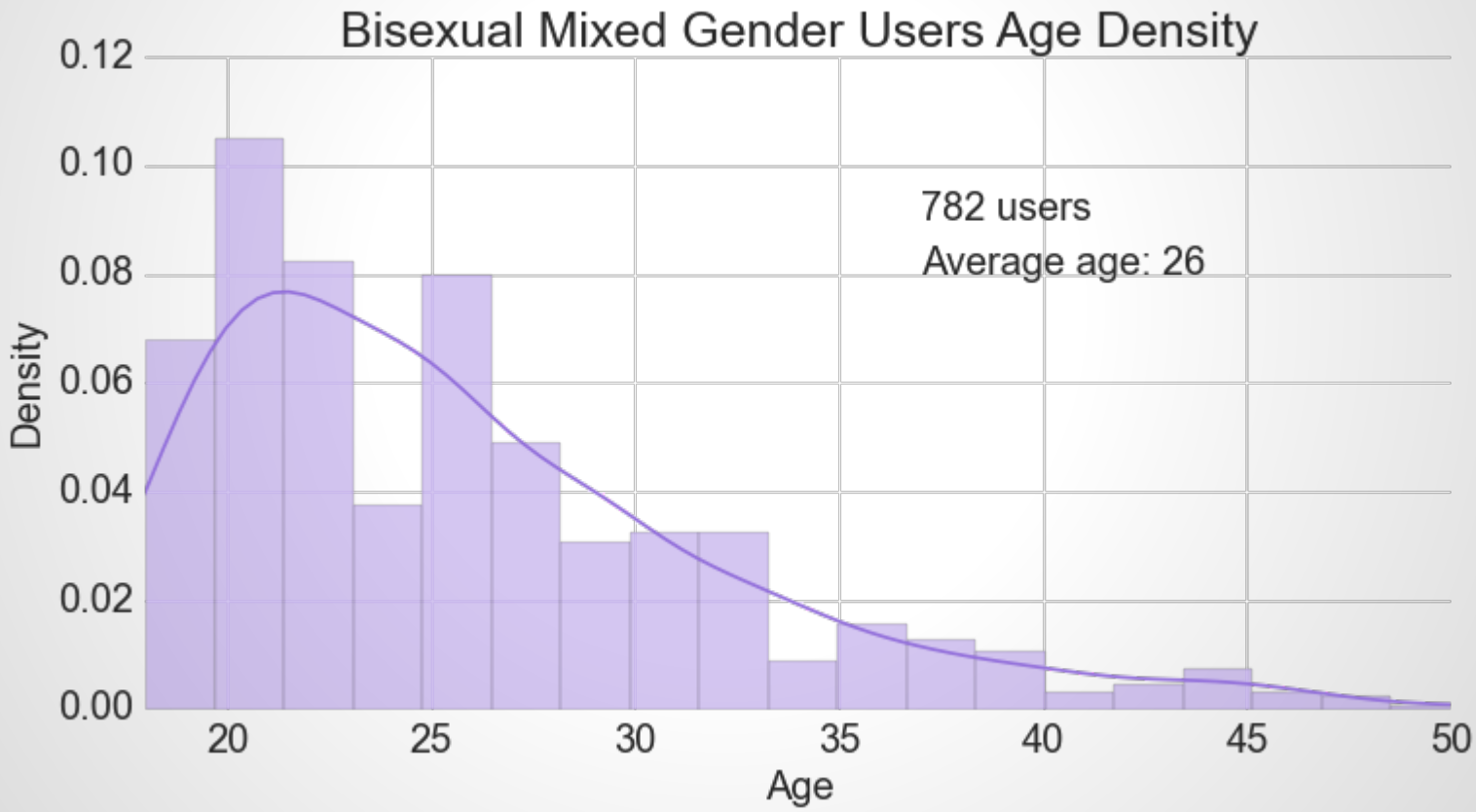
Straight Male



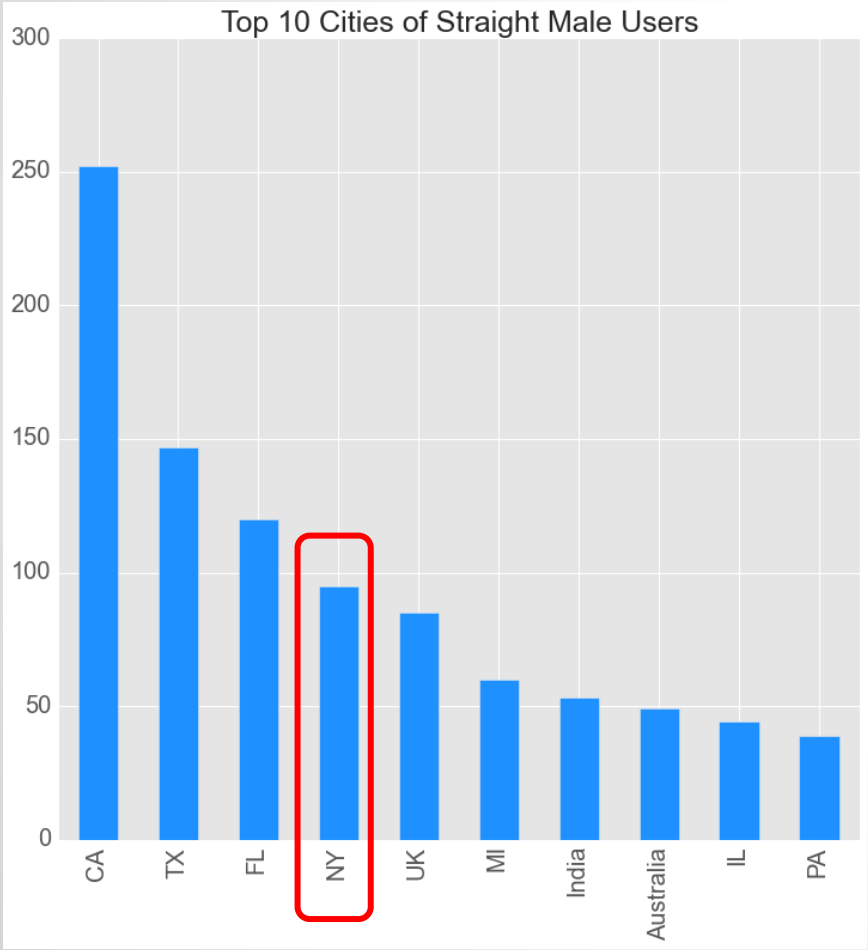
Straight Female

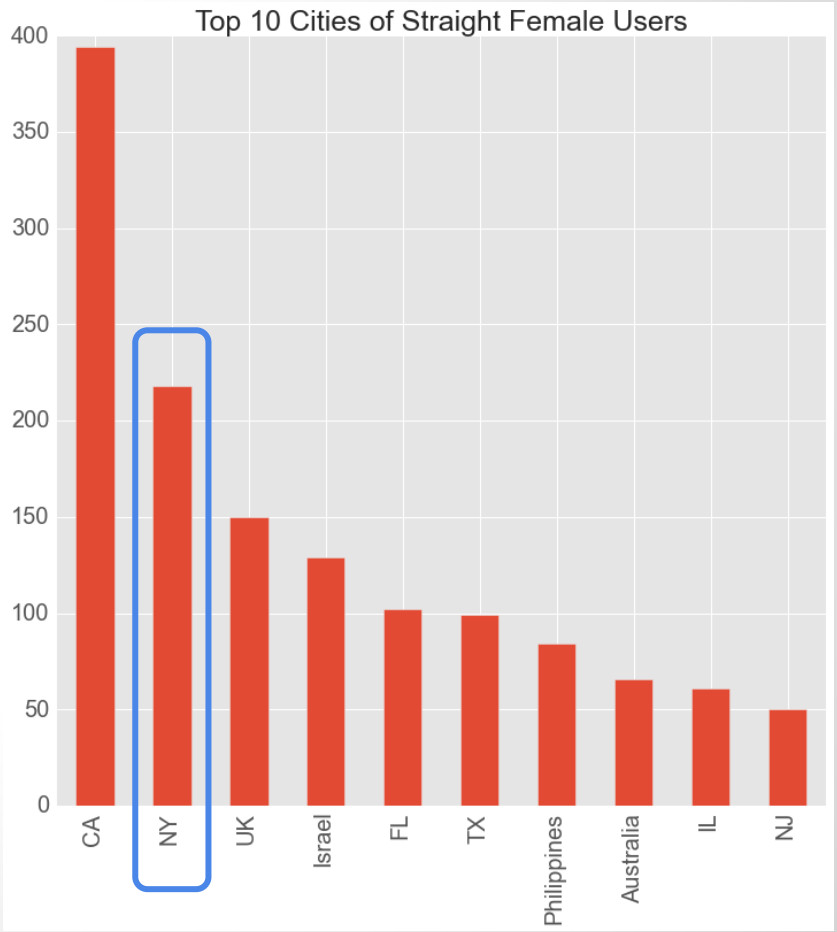


Bisexual Mixed Gender



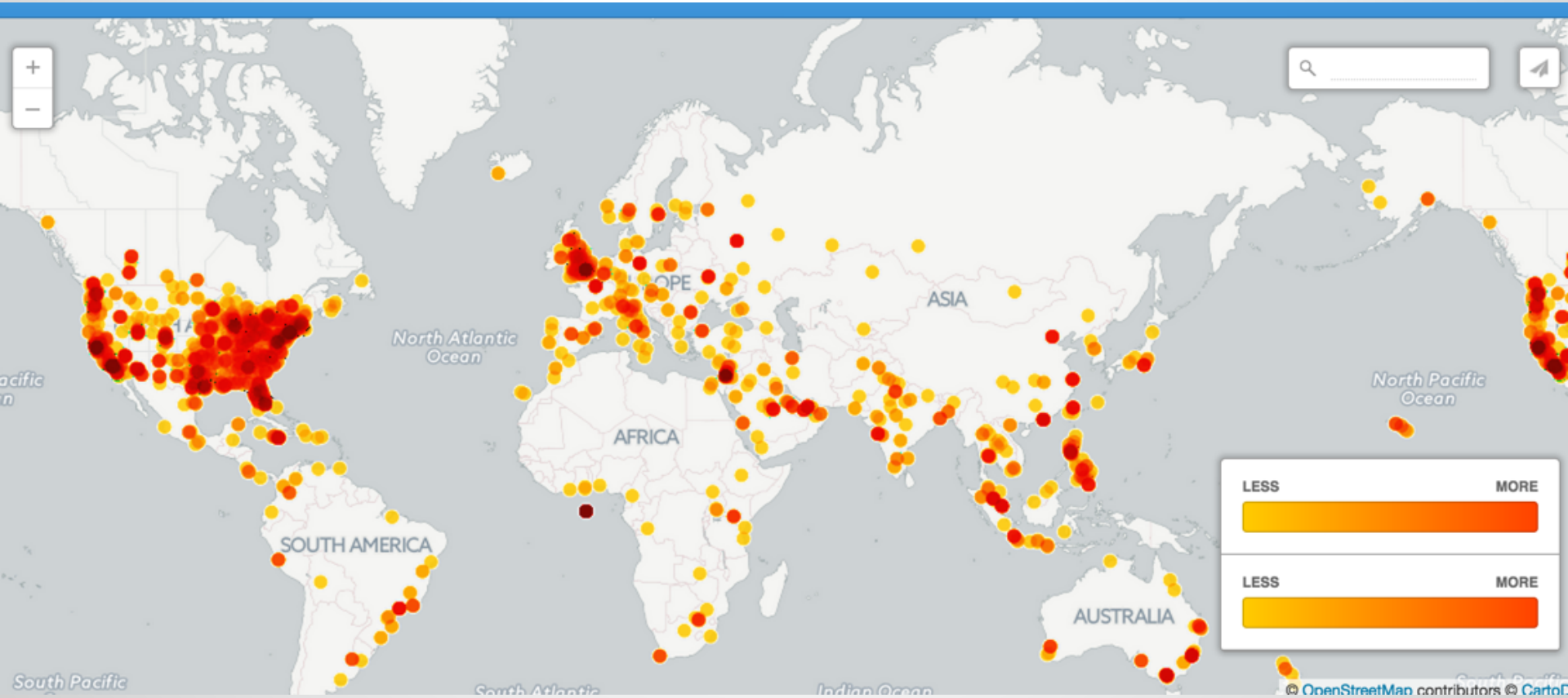
Location


Top 10 Cities of Straight Male Users
300
250
200
150
100
50
0
CA
TX
FL
NY
UK
MI
India
Australia
IL
PA


Top 10 Cities of Straight Female Users
400
350
300
250
200
150
100
50
0
CA
NY
UK
Israel
FL
TX
Philippines
Australia
IL
NJ

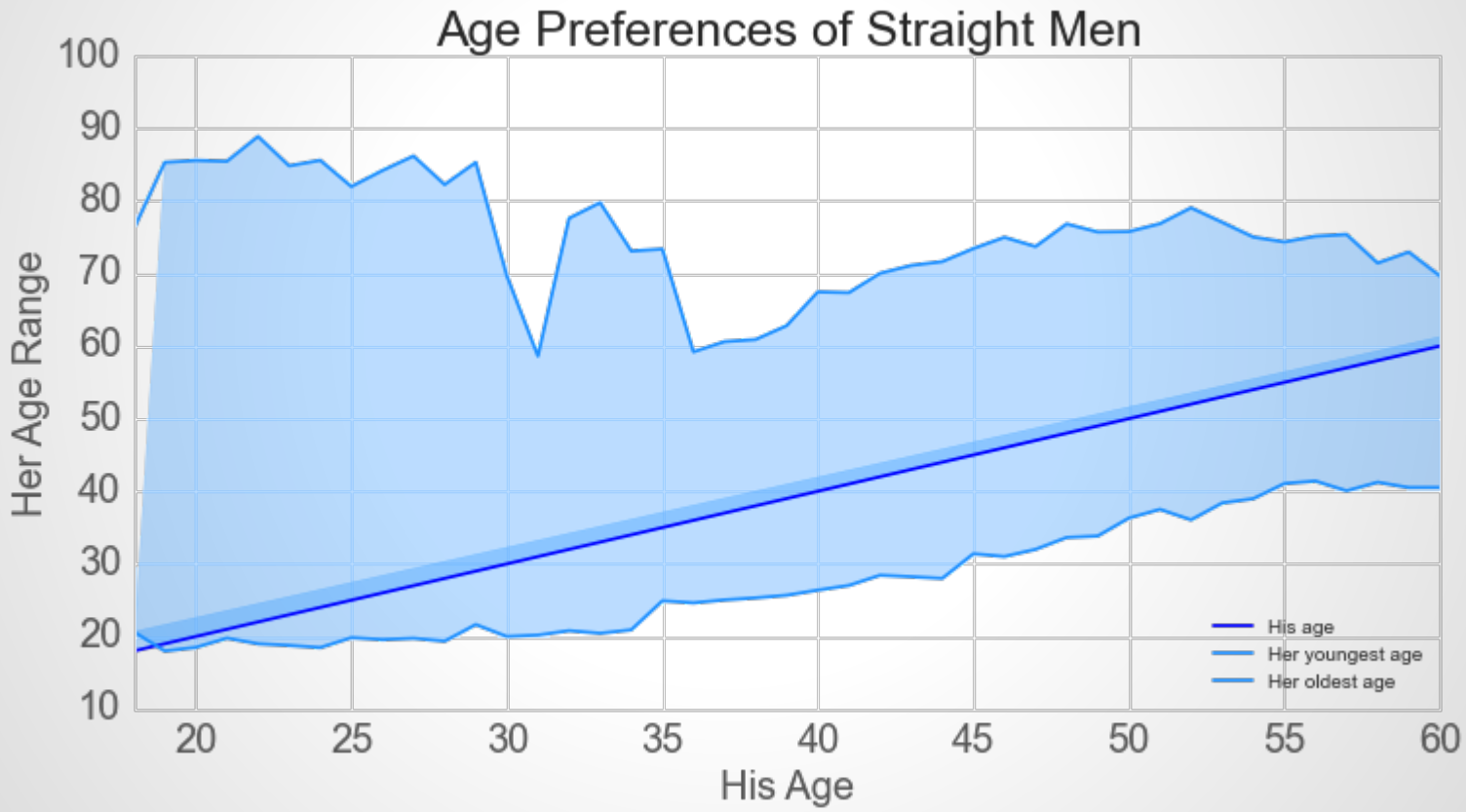
Location - Heat Map

https://chengfzchn.cartodb.com/viz/9b404876-1acd-11e5-9ef1-0e5e07bb5d8a/public_map

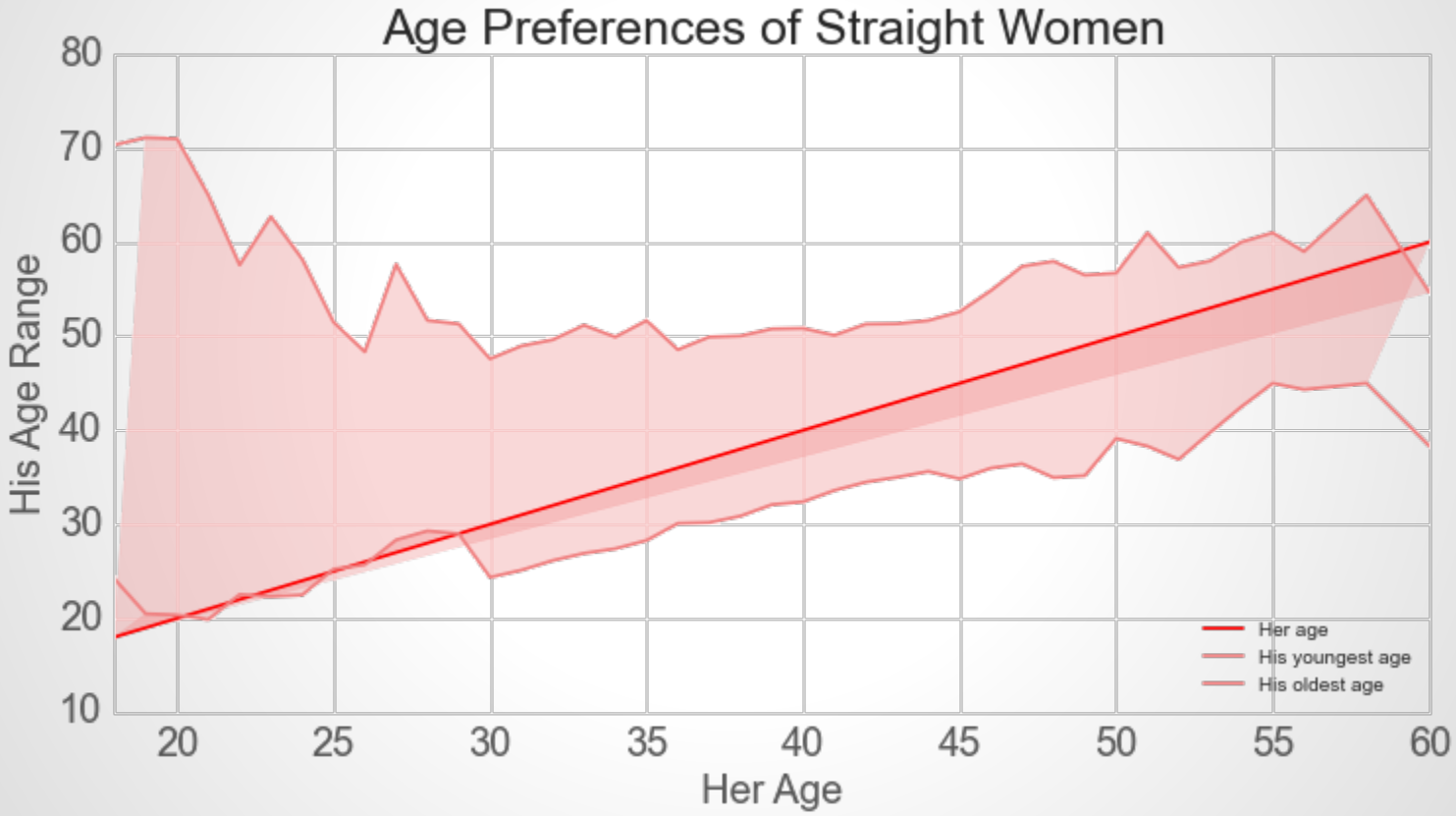


Who do you think is pickier?
Man or Woman?

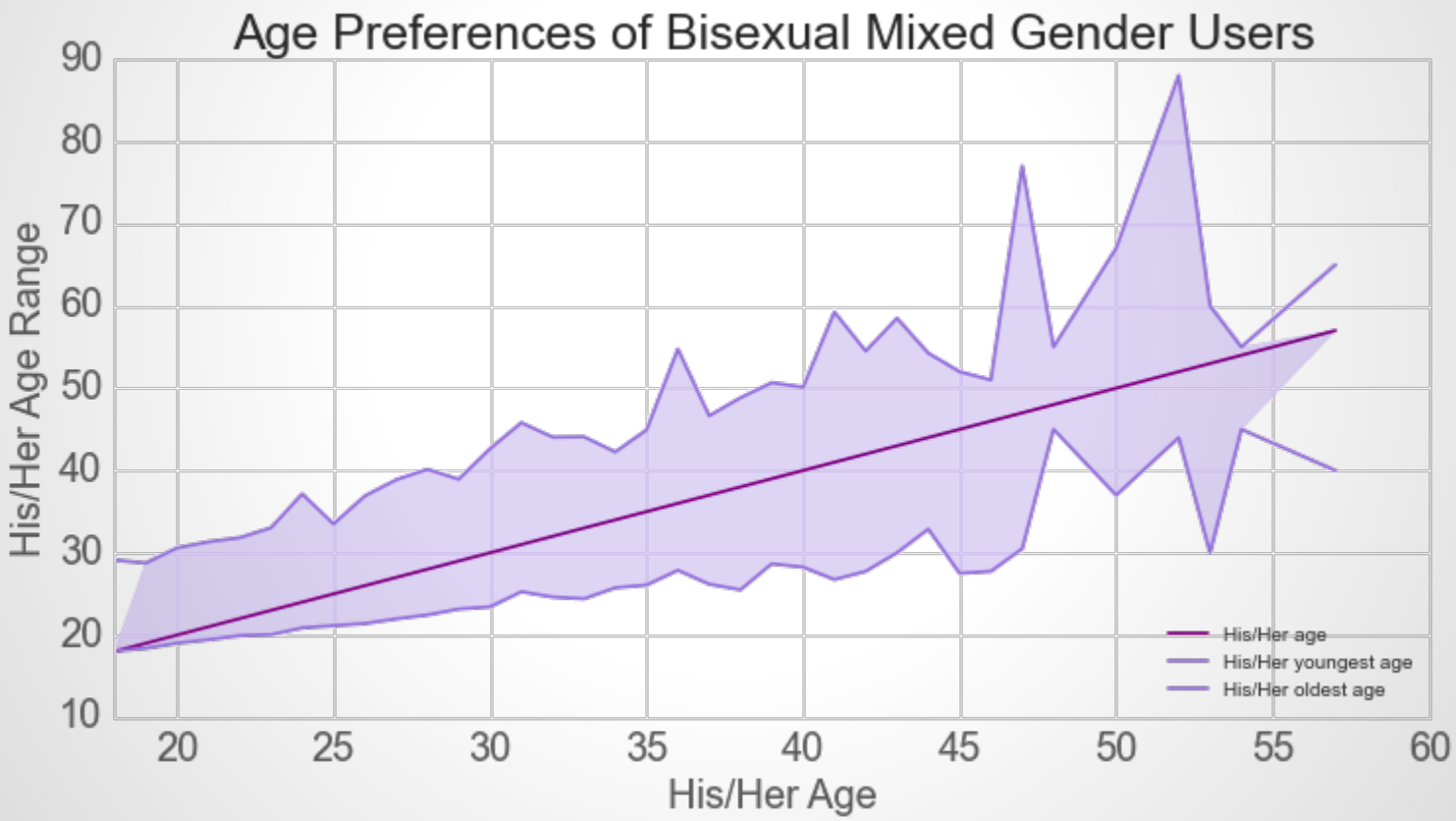
Man?


Age Preferences of Straight Men
Her Age Range
His Age
100
90
80
70
60
50
40
30
20
10
20
25
30
35
40
45
50
55
60
His age
Her youngest age
Her oldest age

Woman?

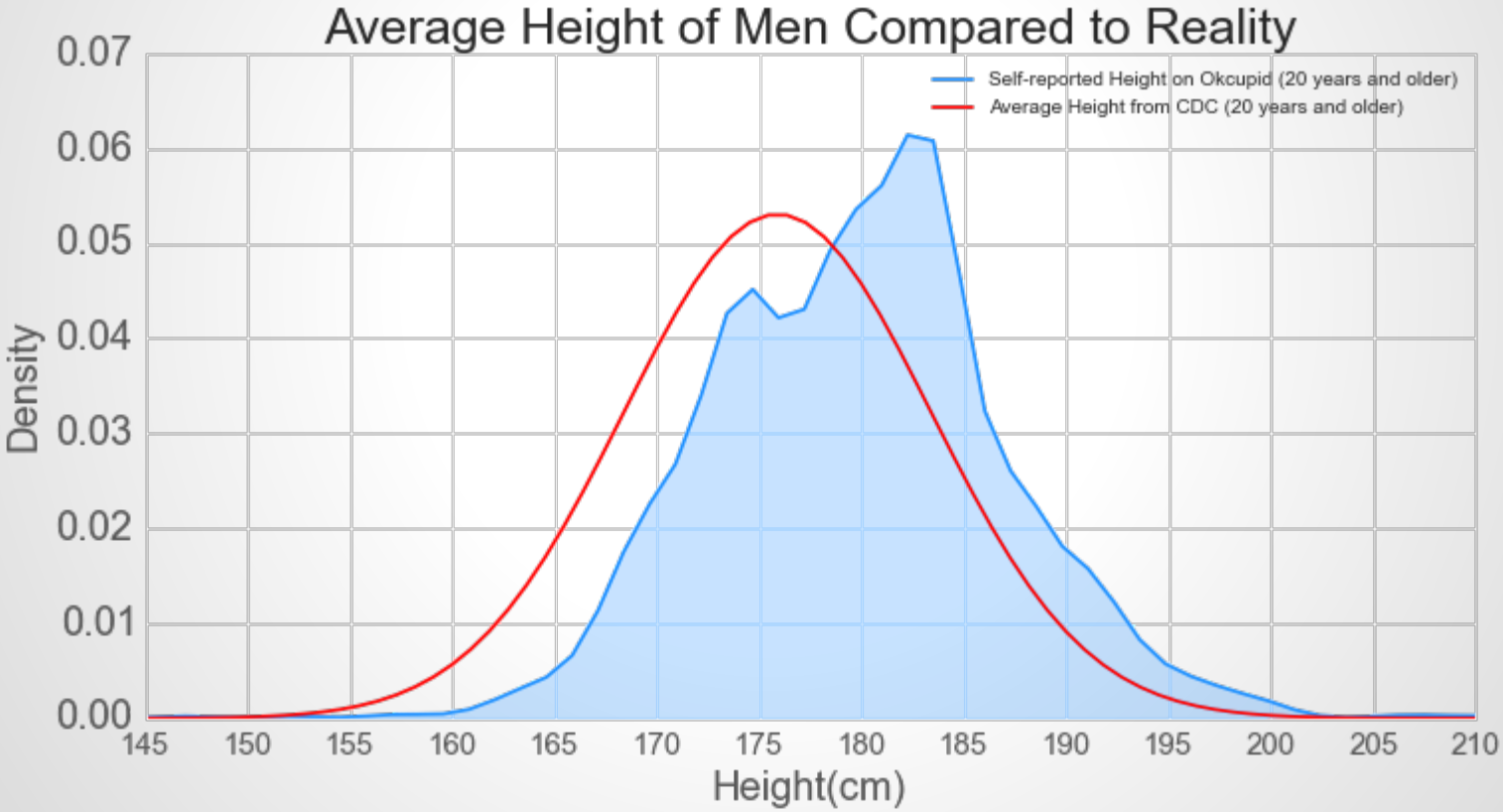

Age Preferences of Straight Women
His Age Range
80
70
60
50
40
30
20
10
20
25
30
35
40
45
50
55
60
Her Age
Her age
His youngest age
His oldest age

The Pickiest?

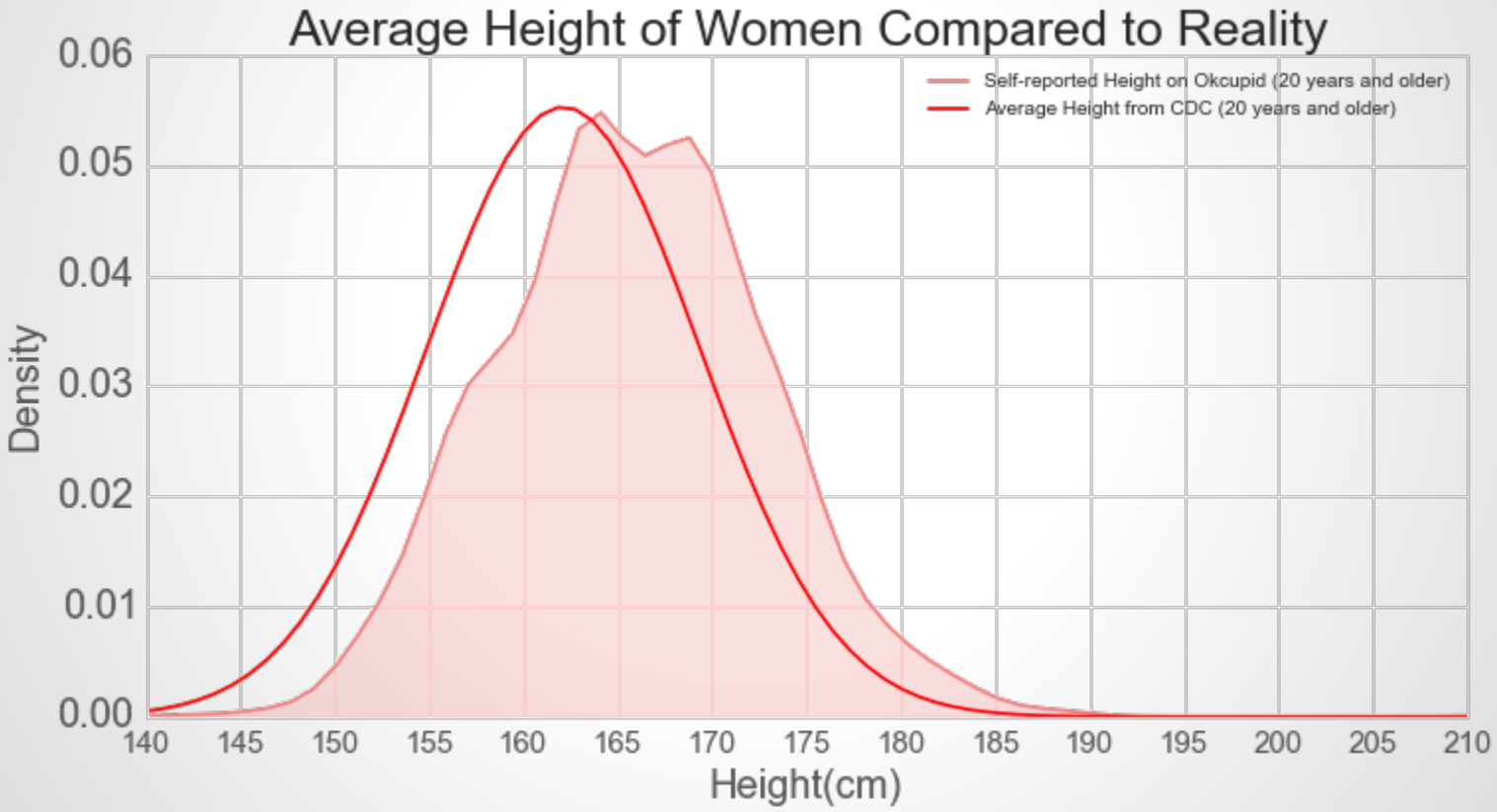


Who do you think is taller online than reality?
Man or Woman?

Men are 5 cm/ 2 inches taller!



Women are taller too!



Possible reasons:

- People lied about their heights on Okcupid.
  - Based on Okcupid cofounder, Christian's blog, higher people claimed to have more sexual partners, and they really do receiver more messages on Okcupid.
- Biased data collection.
- People who use Okcupid really are taller than the average!

Source: http://blog.okcupid.com/index.php/the-biggest-lies-in-online-dating/