

# **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

Men

Interested in wo...

Ages 19 to 35

Within 25 miles of me

Online in the last ...

Advanced

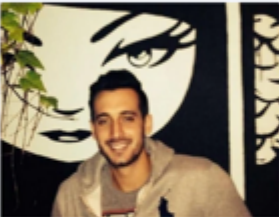
Order by

Special Blend

Search

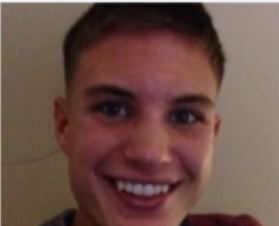
Clear

118,995 online



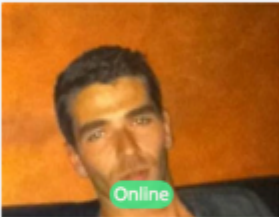
kkj1990  
24 · New York, NY

★ Like




username10105  
20 · Garden City, NY

★ Like




humanbeing28  
29 · New York, NY

★ Like




lol\_buttz  
28 · Brooklyn, NY

★ Like



qbanguy2k  
31 · Manhasset, NY

★ Like



wrd888  
31 · Brooklyn, NY

★ Like

# Page

# Quick match

0% Match  
25 • Brooklyn, New York

×

★

Passed

### My self-summary

Baltimore-bred, Brooklyn-based.

Responsible, yet hedonistic; optimistic, yet cynical. Always overthinking or underthinking, never trying too hard or too little.

Sometimes, I'm the silly son of a bitch who breaks out in a skip down Third Avenue, or puts "Fifty Shades of Grey" in the Christian Life section of Barnes & Noble on a dare.

Other times, I'm the introspective, philosophical conversationalist who'll stay up until 2 AM with you, chatting about all the deep subjects everyone else is too afraid to dive into.

## SAVE \$3

ON ANY 2  
**SCHICK® MEN'S  
RAZORS  
OR REFILLS**


Excludes disposables and trial/travel sizes

AVAILABLE AT  
**TARGET**

©2015 Energizer, Schick, Schick Hydro and Quattro are trademarks of Energizer.

# 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

BE AN ALLSTATE AGENCY OWNER









Click to watch Alyson's story

Allstate

BE AN AGENT

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](http://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

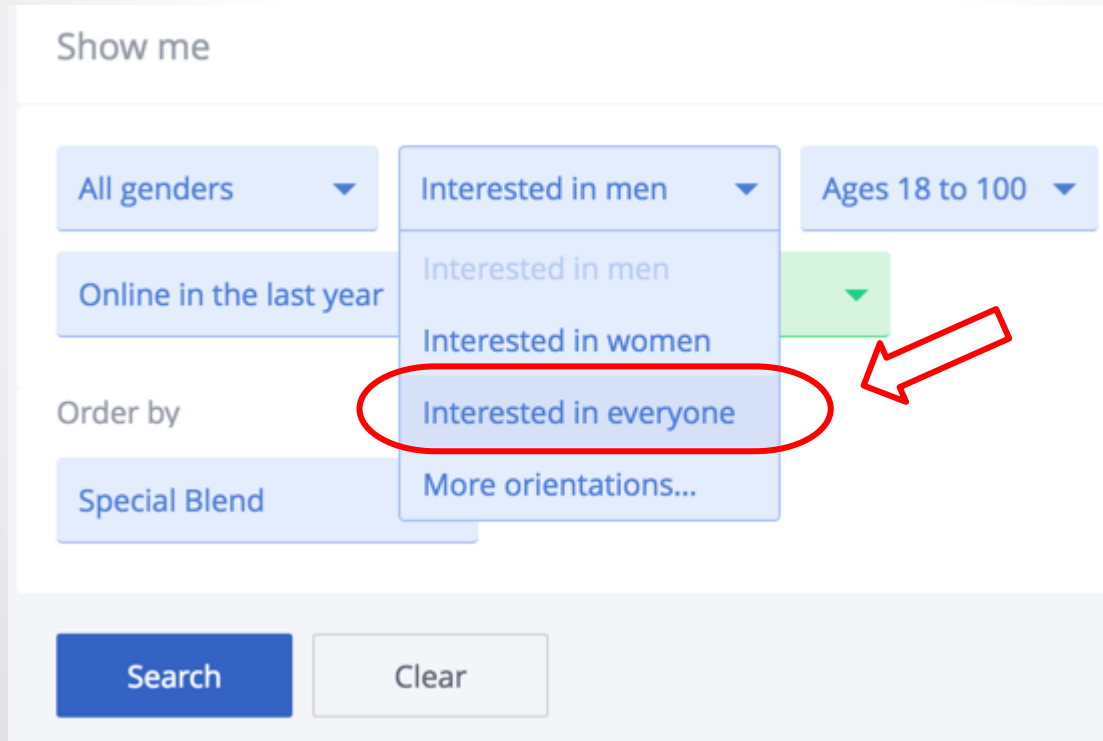
## What I did:

- 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 me 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?



The image shows a search filter interface with the following elements:

- Show me** (header)
- Filters:**
  - Gender:** All genders (dropdown)
  - Orientation:** Interested in men (dropdown menu is open, showing options: Interested in men, Interested in women, Interested in everyone (circled in red), More orientations...). A red arrow points to the 'Interested in everyone' option.
  - Age:** Ages 18 to 100 (dropdown)
  - Online:** Online in the last year
  - Order by:** Special Blend
- Buttons:** Search, Clear

# Samples

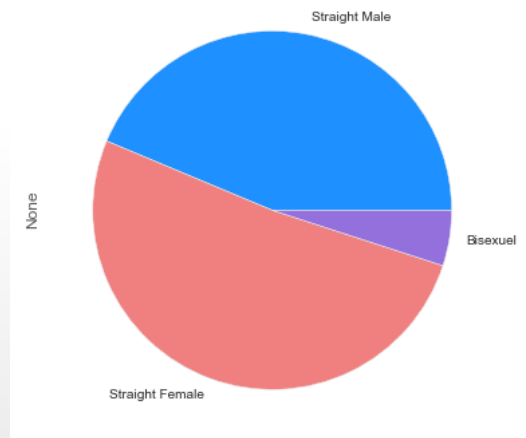
## Username:

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

Gender Distribution of My Dataset



# 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
  - User regular expression to get height, age range and state/country information from long string

# Step 3. Data Manipulation

## 1. Demographics Analysis

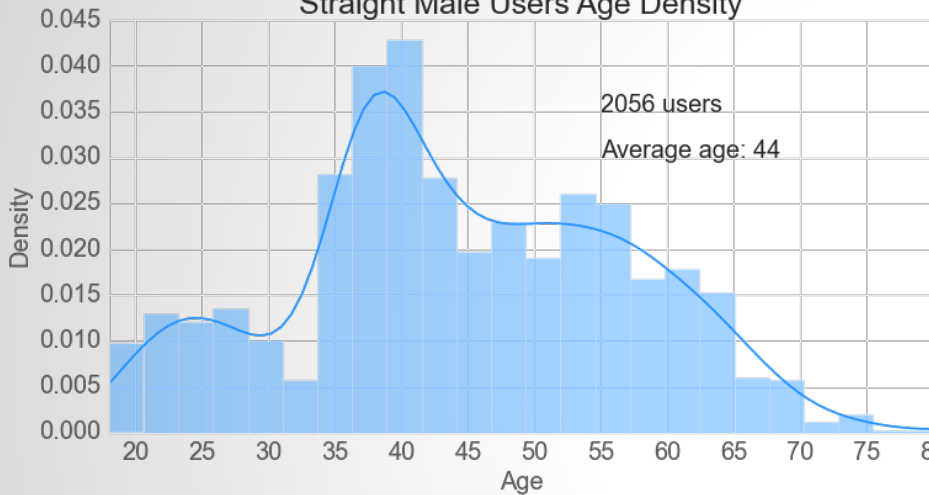
- a. How old are they?
- b. Where are they located?

## 2. Psychological Analysis

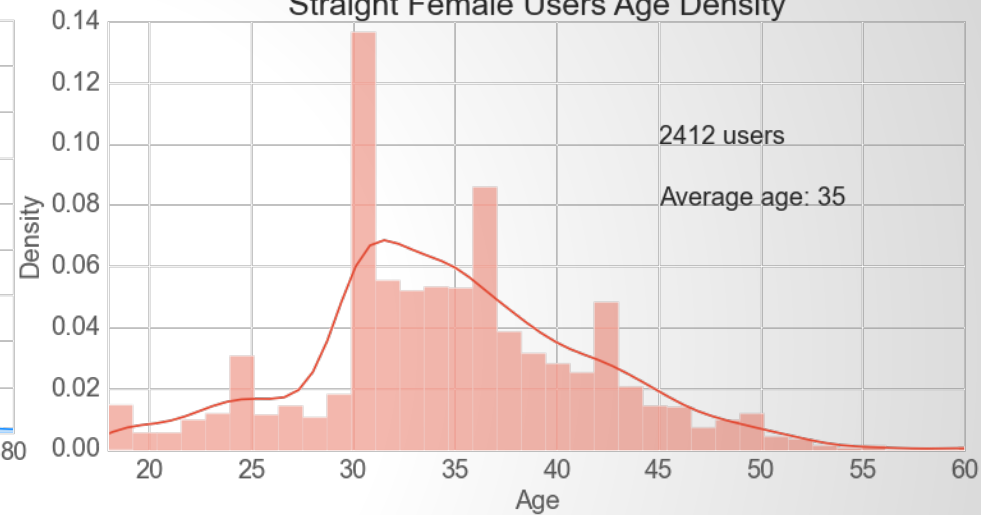
- a. Who are pickier?
- b. Who are lying?

# Age

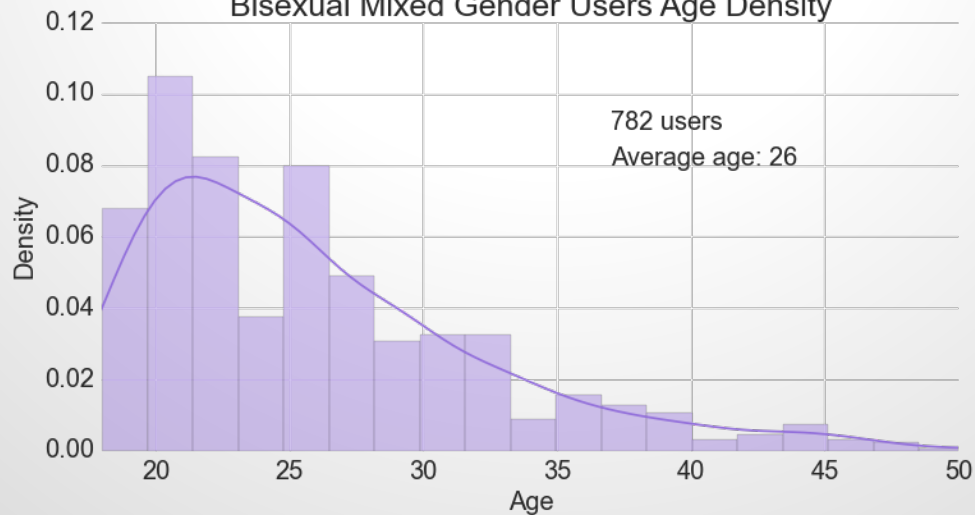
Straight Male Users Age Density



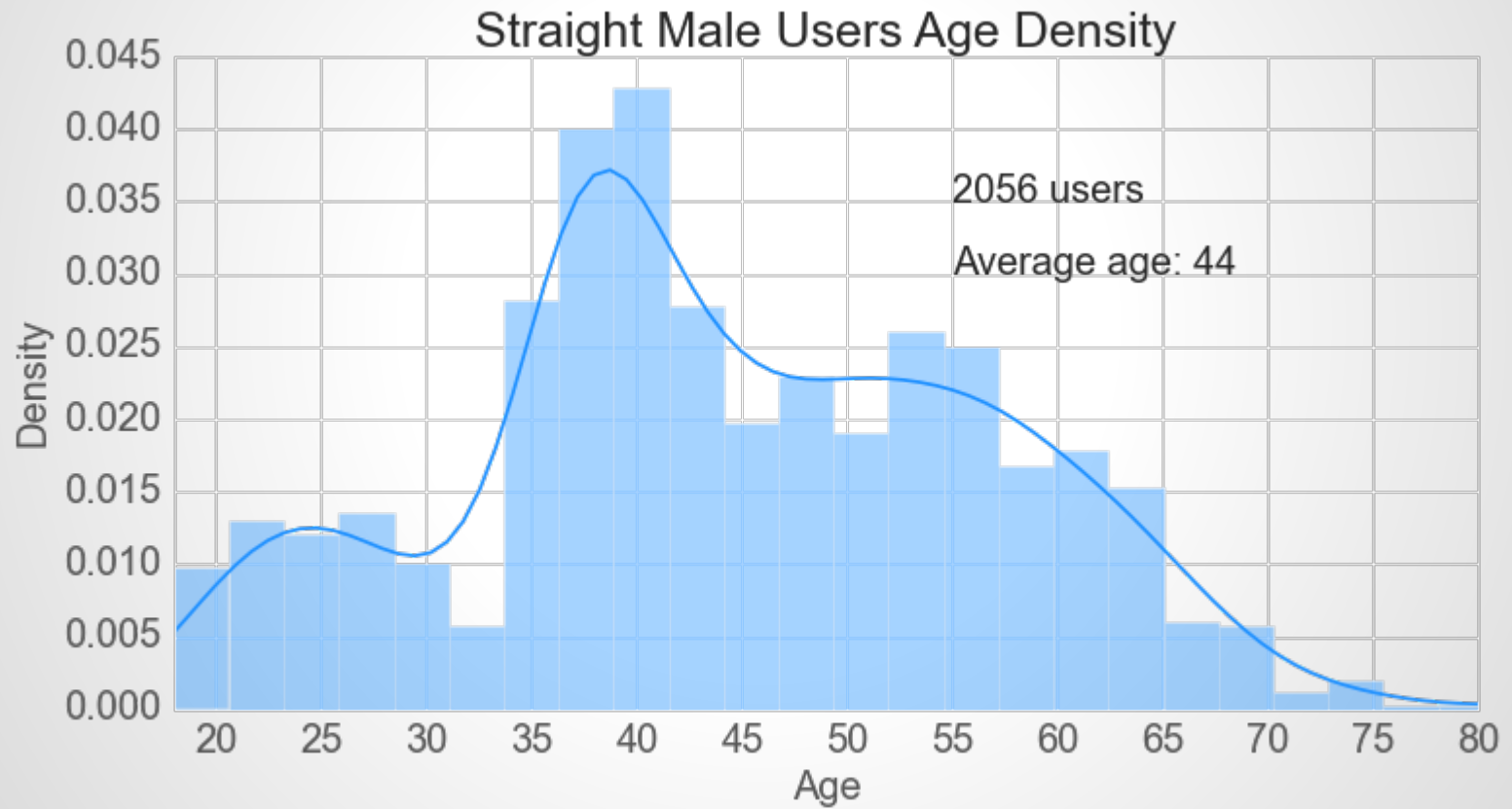
Straight Female Users Age Density



Bisexual Mixed Gender Users Age Density

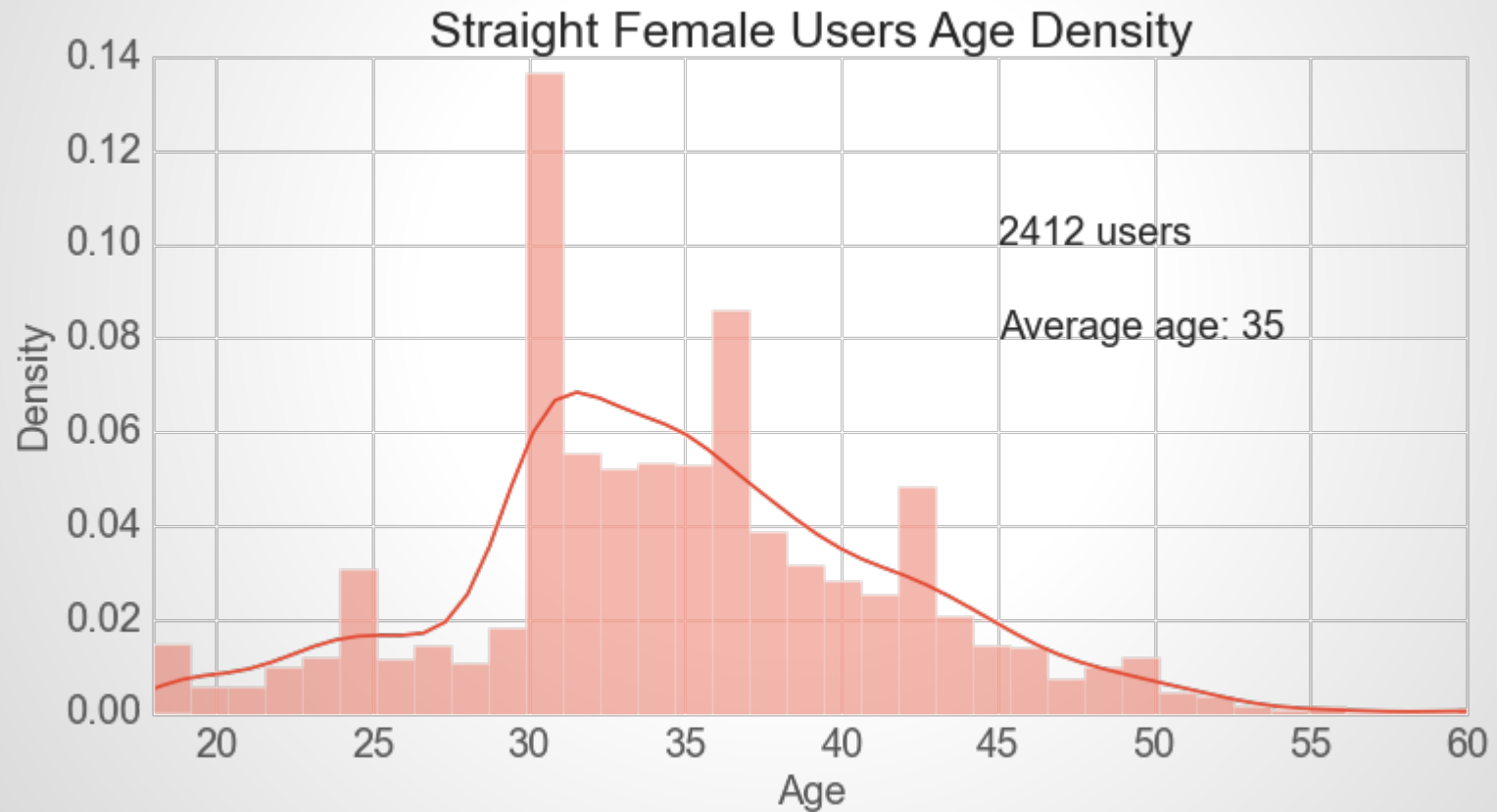


# Straight Male

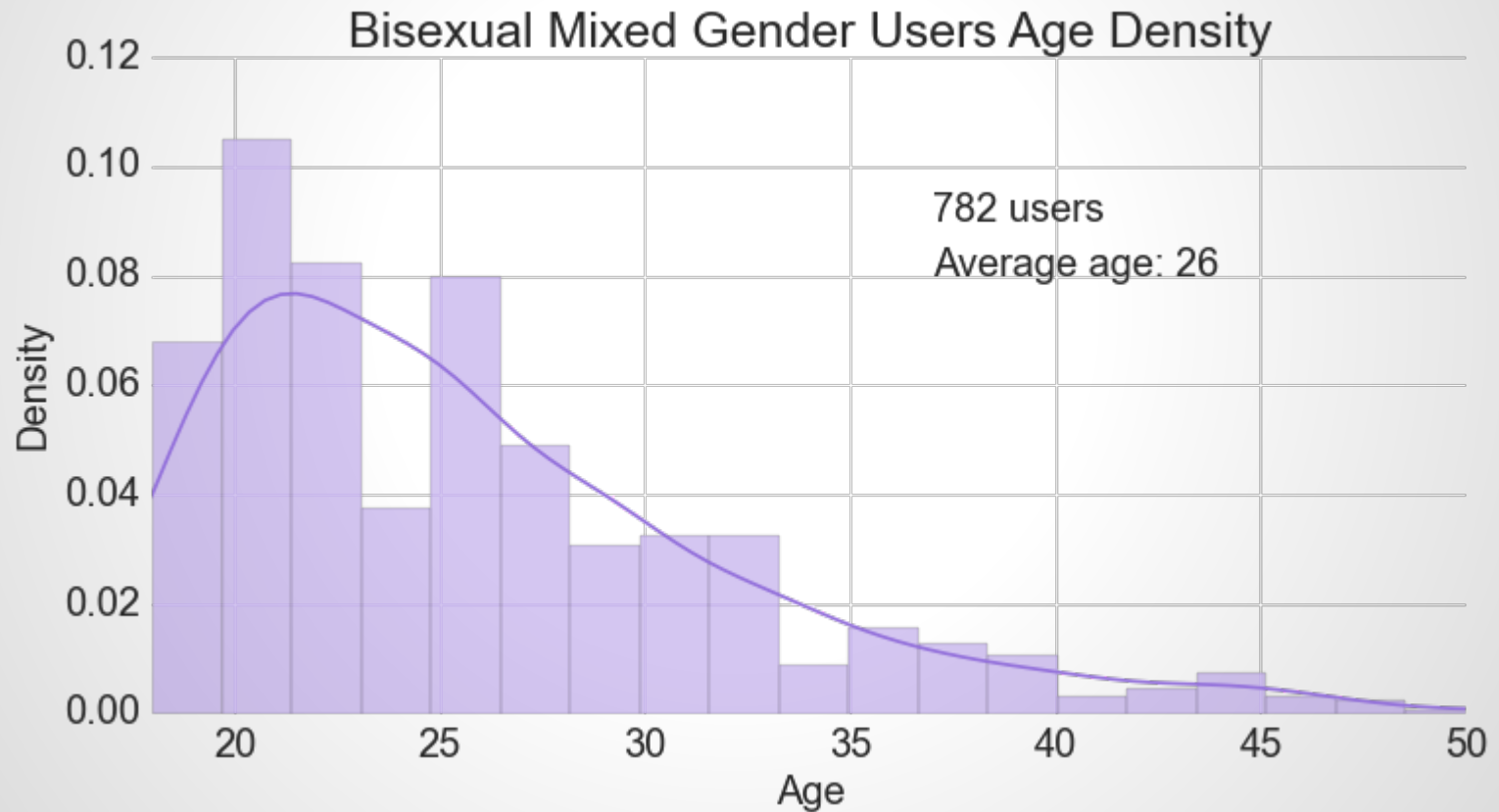




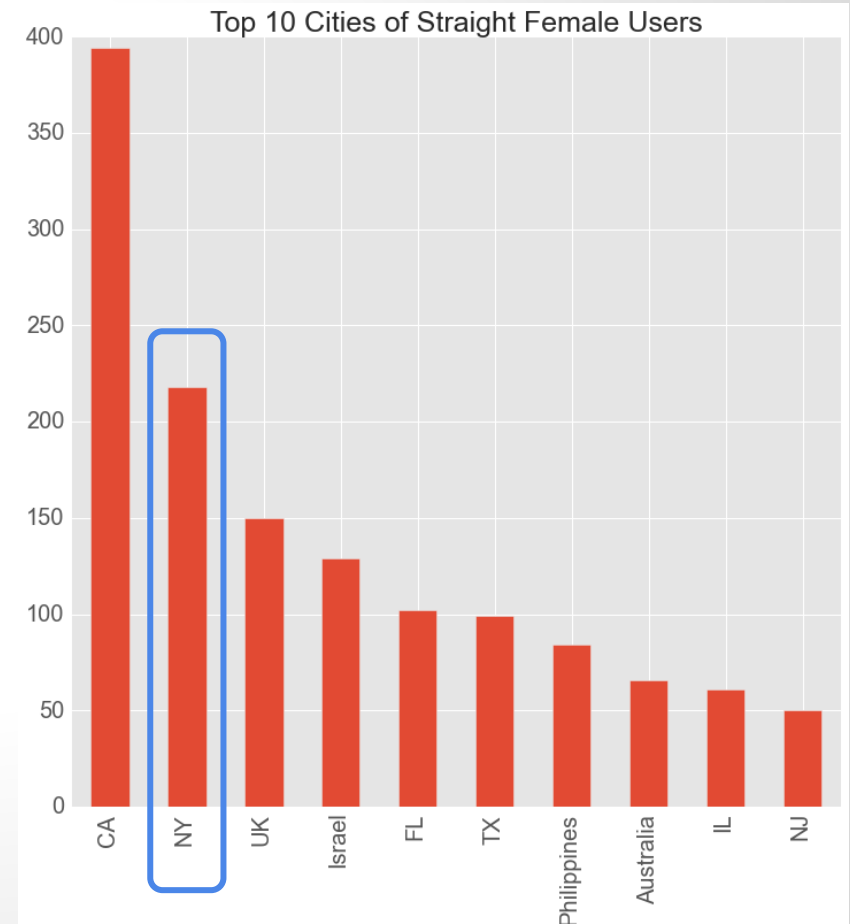
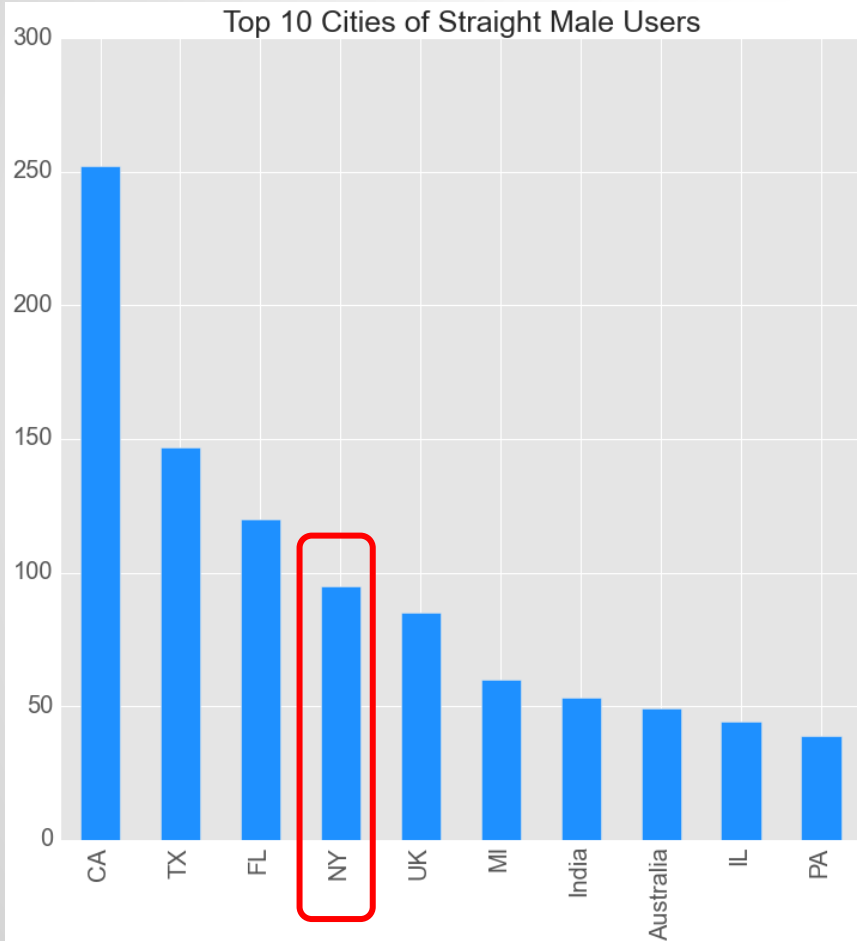
# Straight Female



# Bisexual Mixed Gender

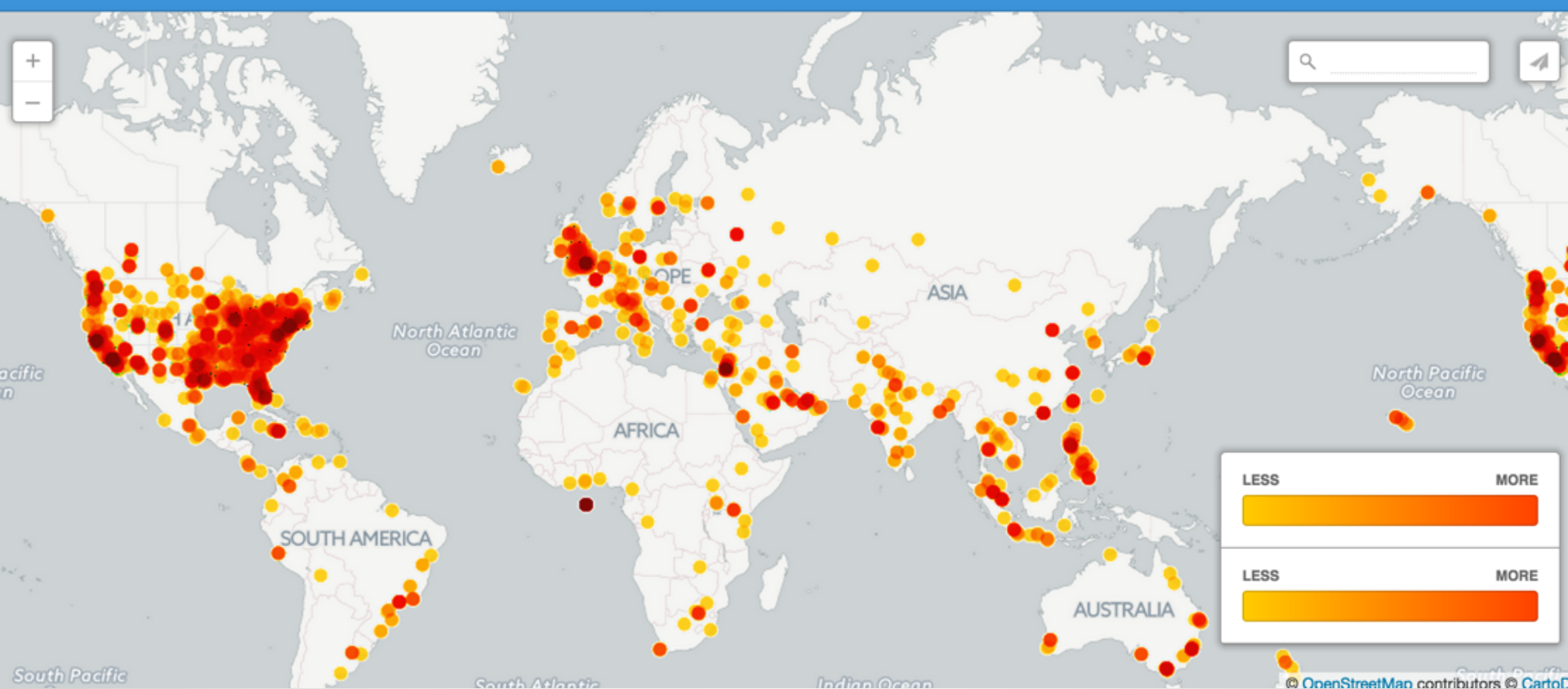


# Location



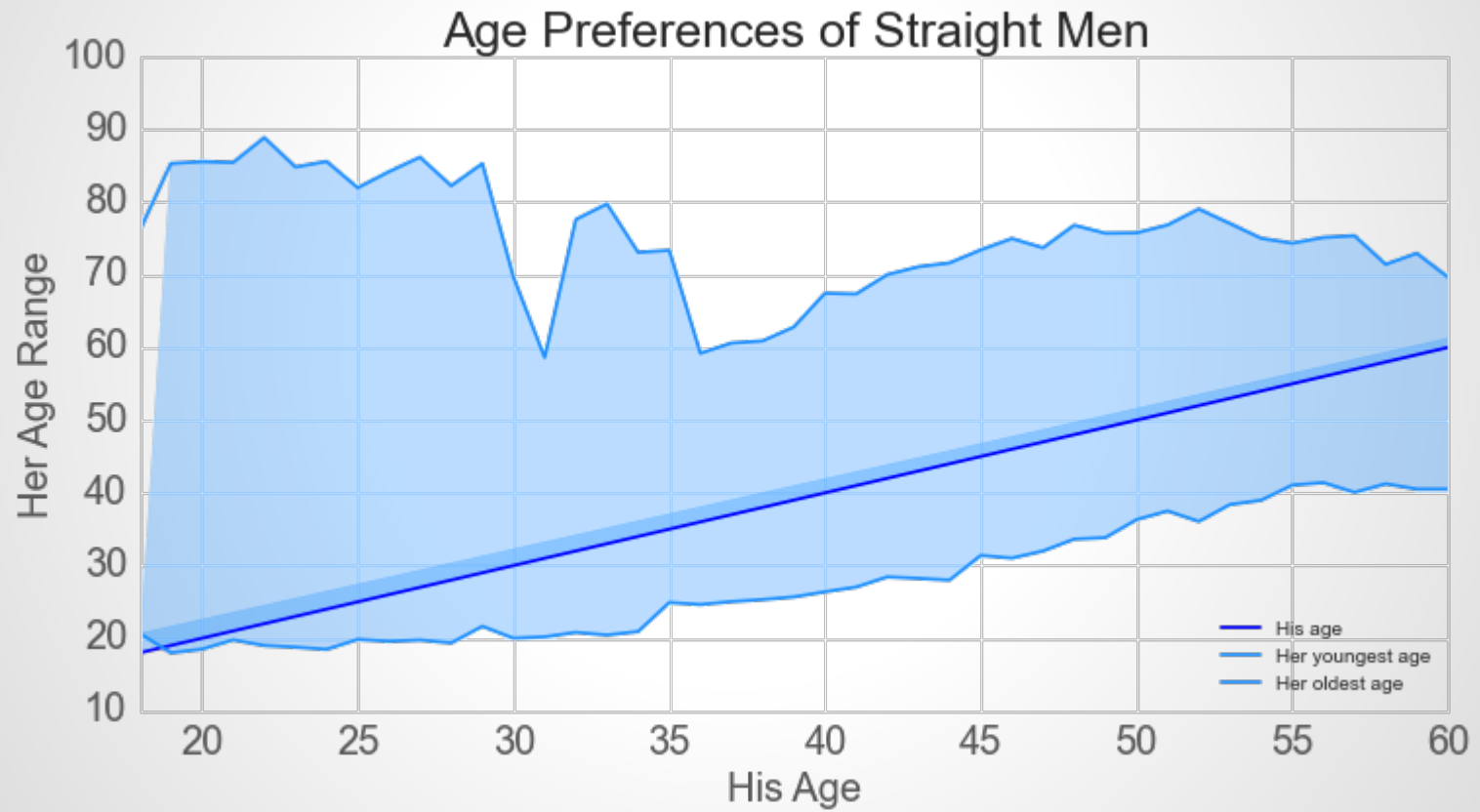
# Location - Heat Map

[https://chengfzchn.cartodb.com/viz/9b404876-1acd-11e5-9ef1-0e5e07bb5d8a/public\\_map](https://chengfzchn.cartodb.com/viz/9b404876-1acd-11e5-9ef1-0e5e07bb5d8a/public_map)

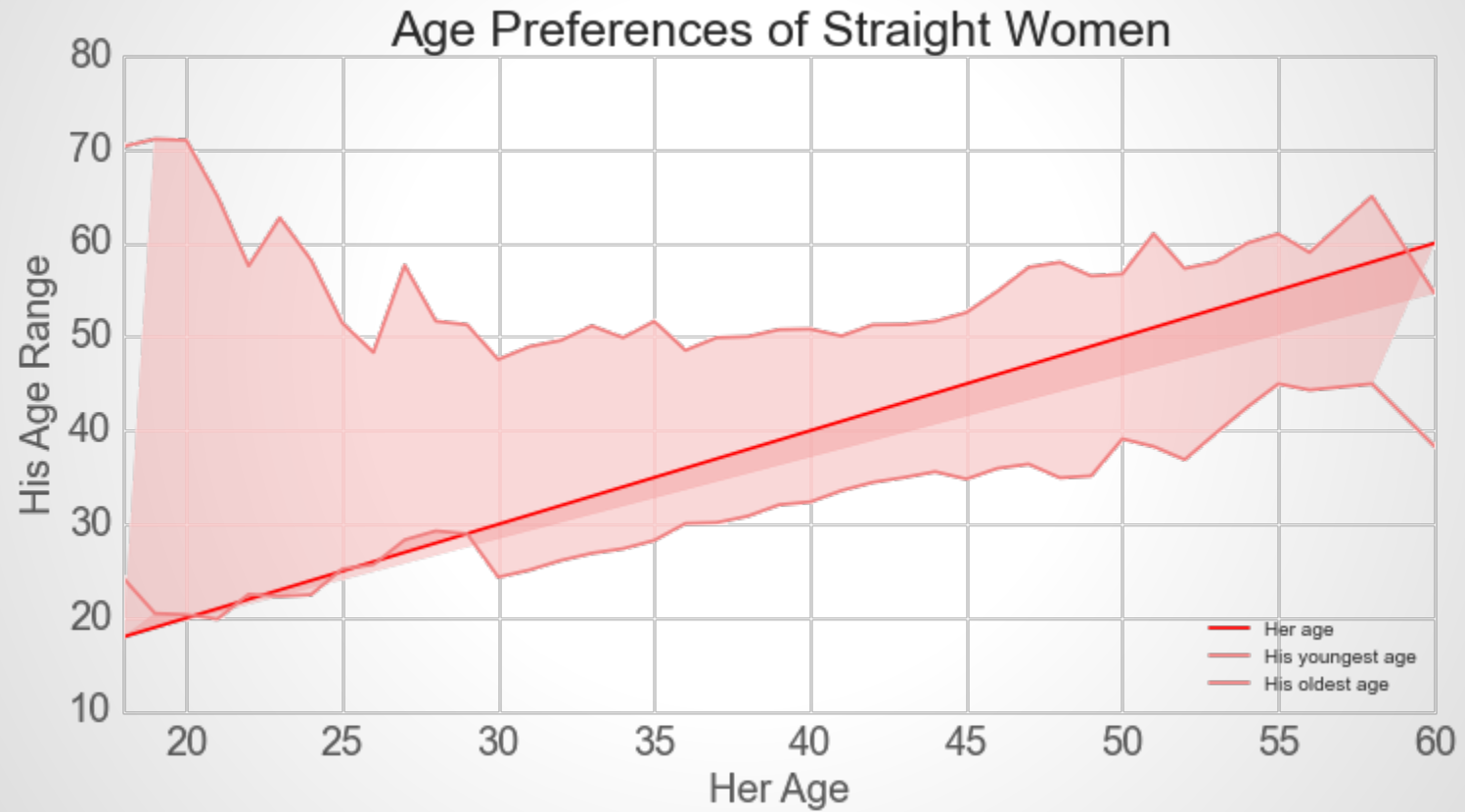


Who do you think is pickier?  
Man or Woman?

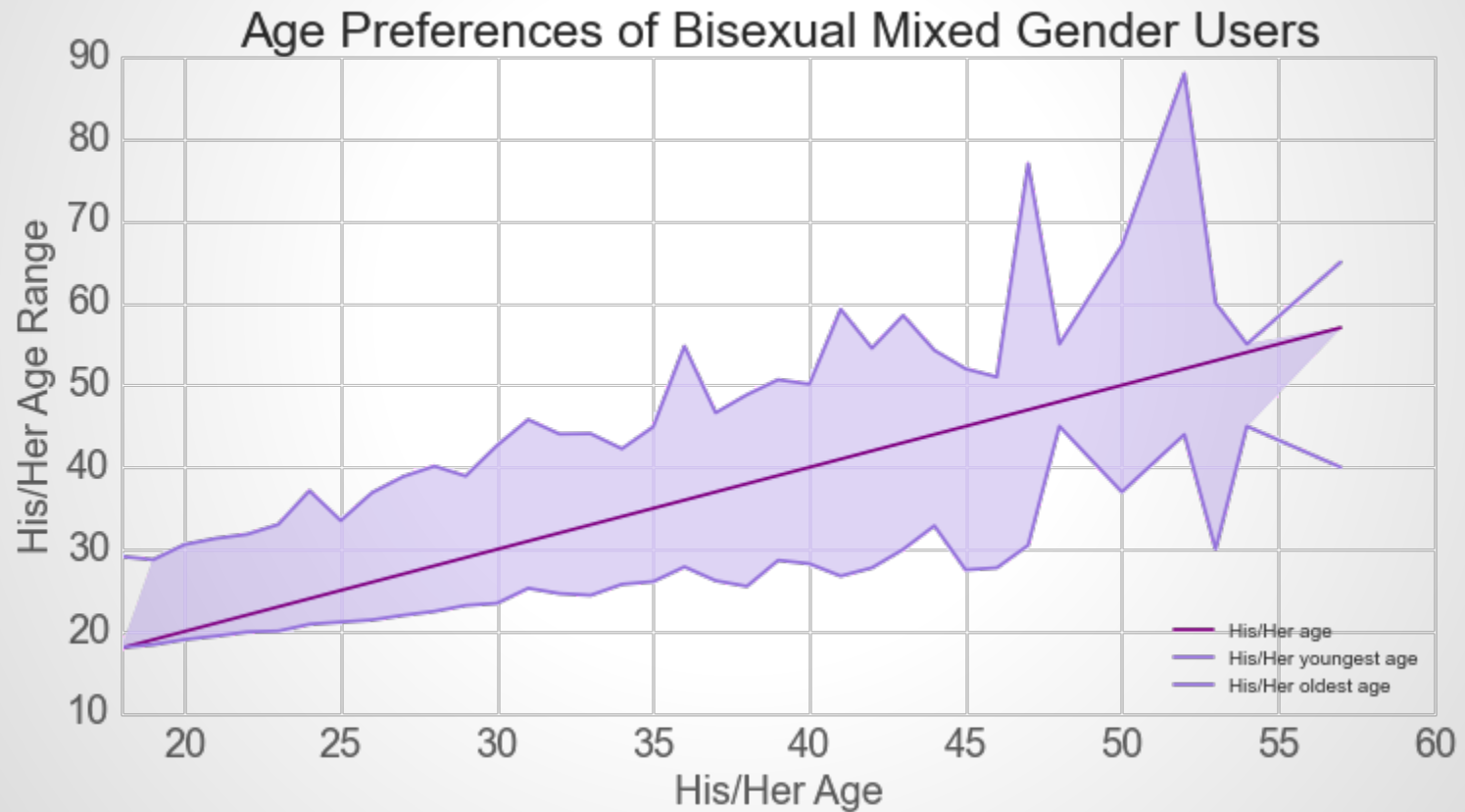
# Man?



# Woman?



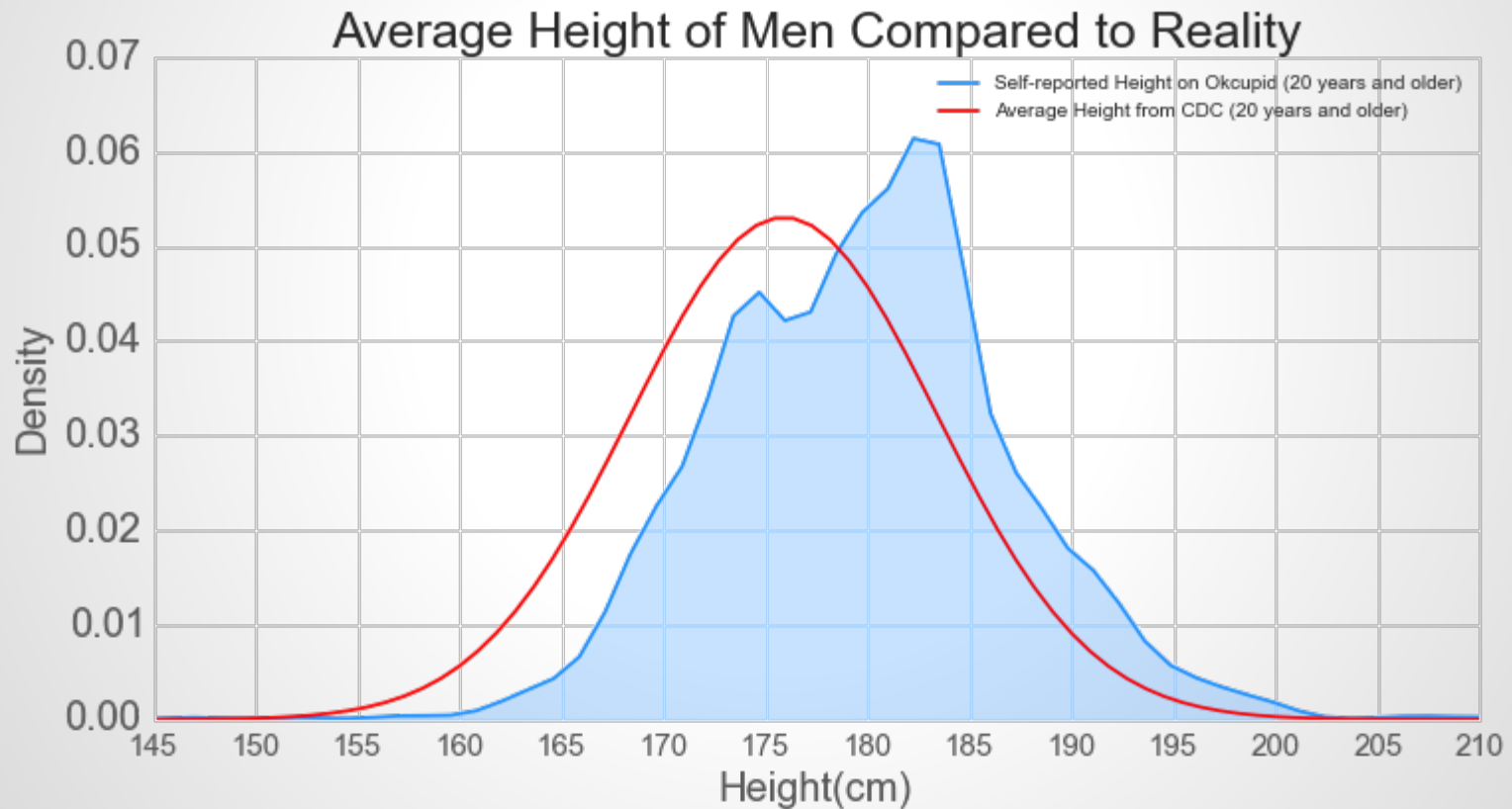
# The Pickiest?



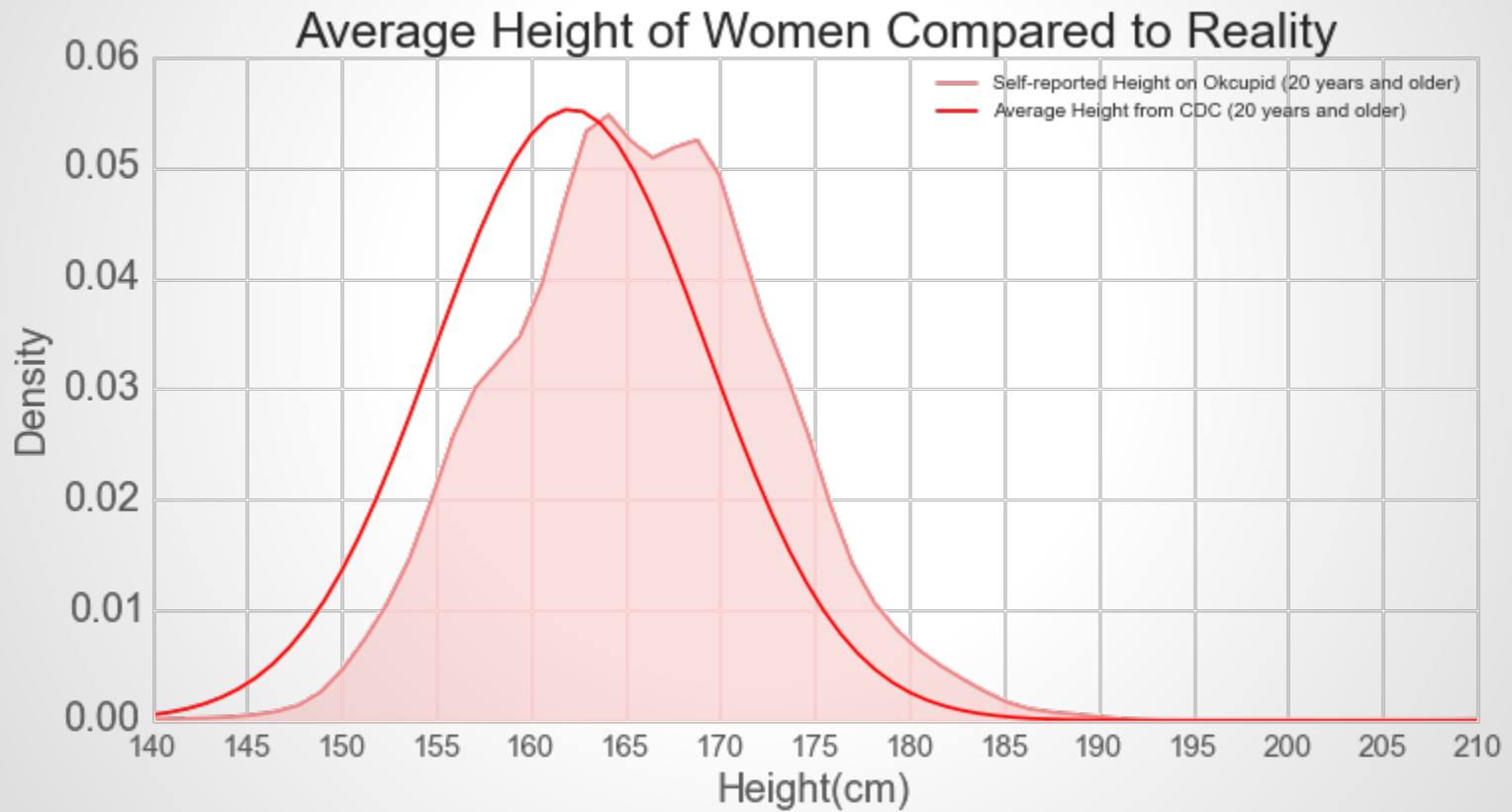


Who do you think is taller online than reality?  
Man or Woman?

# 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/>