Date-A-Scientist

By Kerry Driscoll

Table of Contents

Overview: The First Date

- OKCupid has shared self-reported data on 59,946 users looking for love and connection
  - Predominantly based in the San Francisco Bay Area
  - Used the dating service between July 2011 - June 2012

- In order to make a good match, users report some detailed information about themselves, including:
  - demographics (age, sex, ethnicity)
  - behaviors (drinking, smoking, diet)
  - attitudes (feels strongly about religion, wants kids)

- We're going to use this dating data to examine behaviors across demographic groups

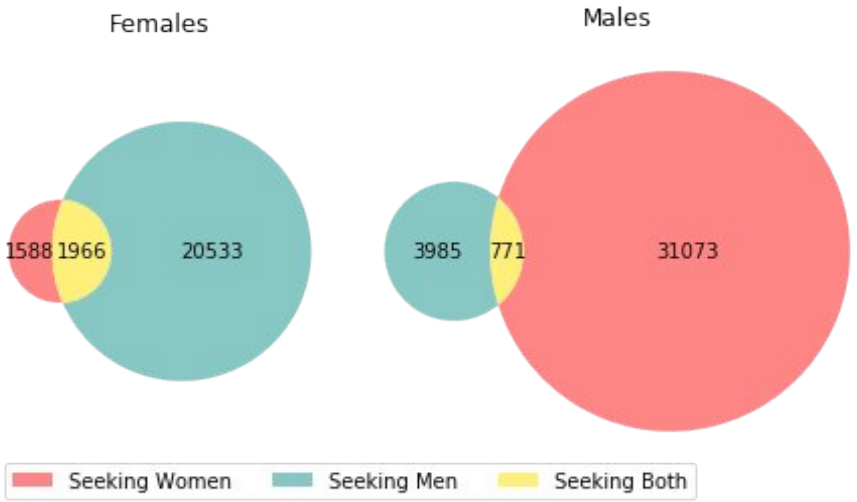
Foolin’ Around

Exploring the Dataset

Foolin’ Around

Question:

Who’s looking for who? I was curious about the size of dating pools for men vs. women and for straight people vs. queer people.



Answer:

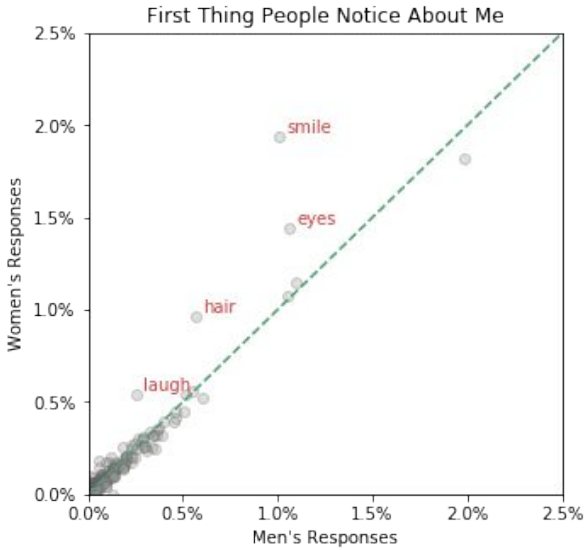
There are more fish in the sea for straight women than straight men. On the other hand, queer men have more available partners than queer women on SF’s OKCupid.

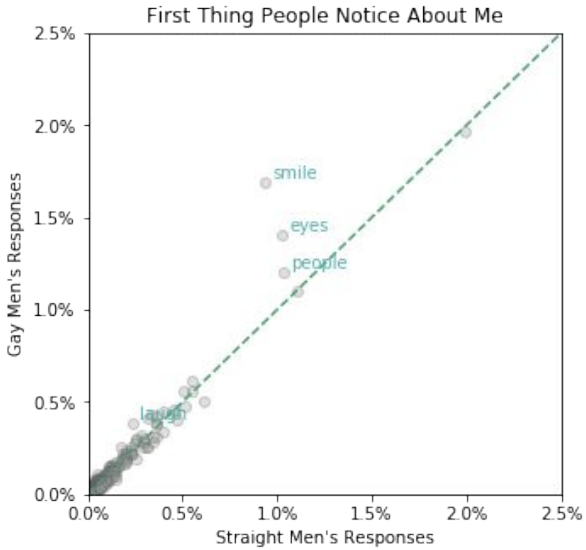
Sex	Count	Percent
Males	35,829	59.8%
Female	24,117	40.2%

Foolin’ Around

Question:

What qualities are attractive in men and women?





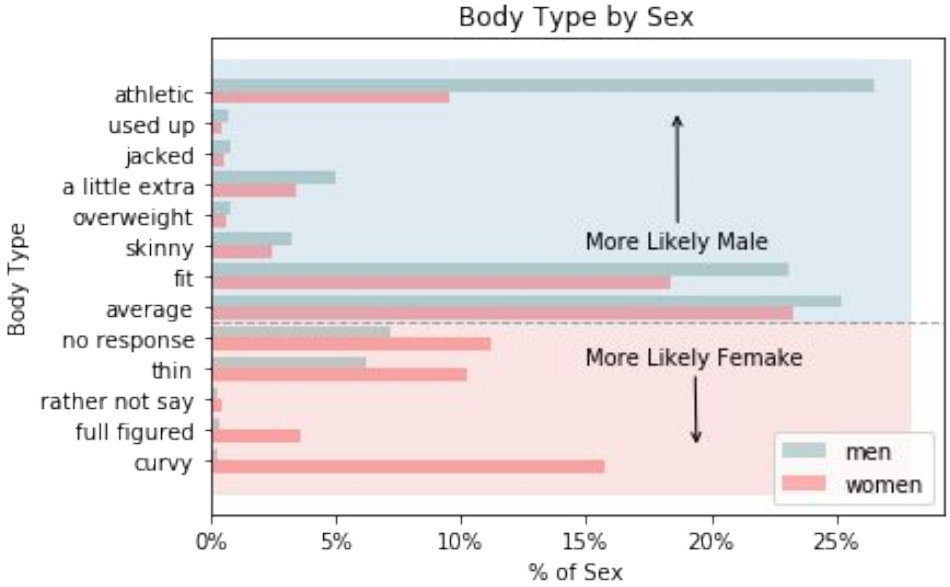
Answer:

Individuals seeking male partners (straight women and gay men) are much more likely to mention their “smile”, “eyes” and “laugh” when asked what is the first thing people notice about them. Straight men do not point out any feature at a significantly higher rate than the general population. This suggests the male gaze is more predictable than the female gaze and/or heterosexual men are uncertain how to celebrate their attractive qualities or may even be unaware of what those qualities may be.

Foolin’ Around

Question:

How do men and women describe their body types?



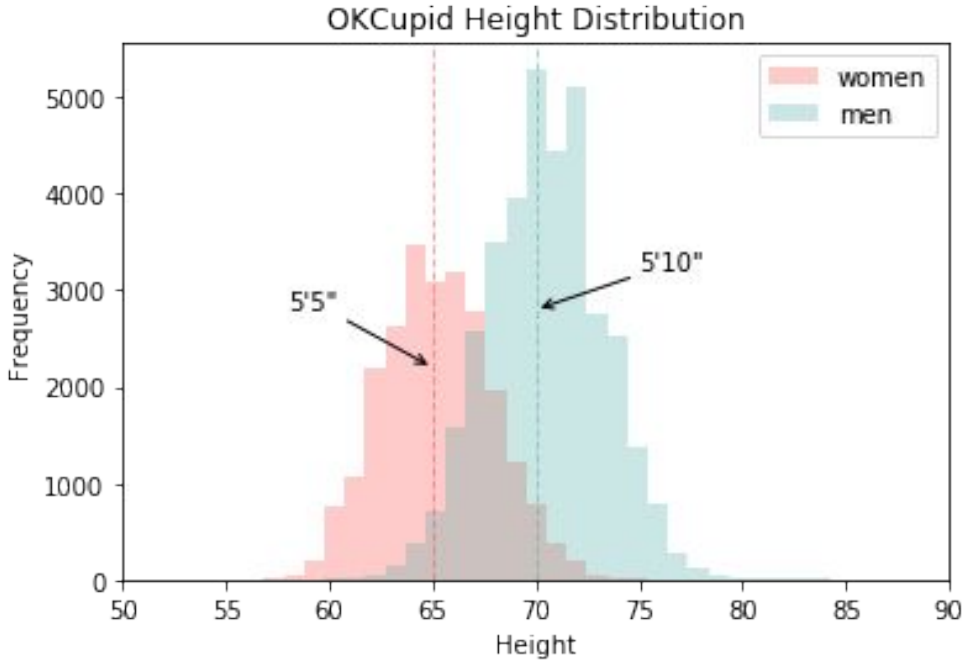
Answer:

Body types tend to be polarized by sex: for instance, men are 2.8x more likely to describe themselves as “athletic” and women are 500x more likely to describe themselves as “curvy”.

It is also interesting that women are more likely to not disclose their body type, reinforcing the idea that women are hyper-aware of their bodies and sensitive to how others perceive them.

Foolin' Around

Question: Which heights are considered "short" or "tall"?



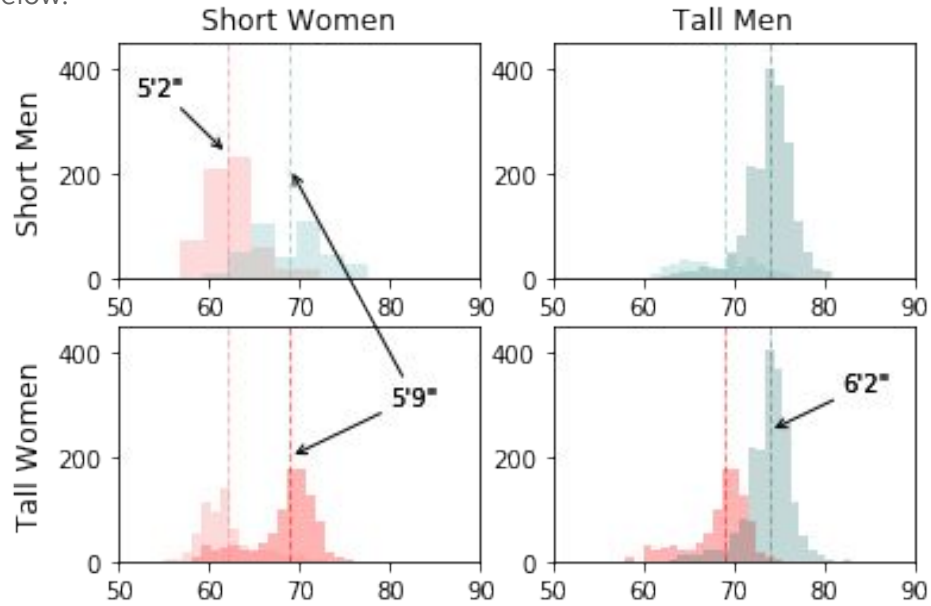
Foolin’ Around

Answer:

Men and women have different standards for “normal” height. Normal heights for women fall within a 7 inch range (62” - 69”) that encompasses 78% of their sex. However, men have a narrower standard for normal height defined by a 5 inch interval (69” - 74”) that encompasses 57% of their sex. If the same standard of “short” that women use (the 13th percentile) was applied to men’s heights, then the definition of a short man would drop 2 inches to 5’7” and below.

Interestingly, the median benchmark for a “tall” woman and a “short” man occur at the same height: 5’9”. This suggests that the conception of short vs. tall is largely based on the opposite sex.

Refers to Self as …	Women		Men	
	Median Height	Percentile	Median Height	Percentile
“Short”	5’2”	*13%*	5’9”	*31%*
“Tall”	5’9”	*91%*	6’2”	*88%*



Foolin' Around

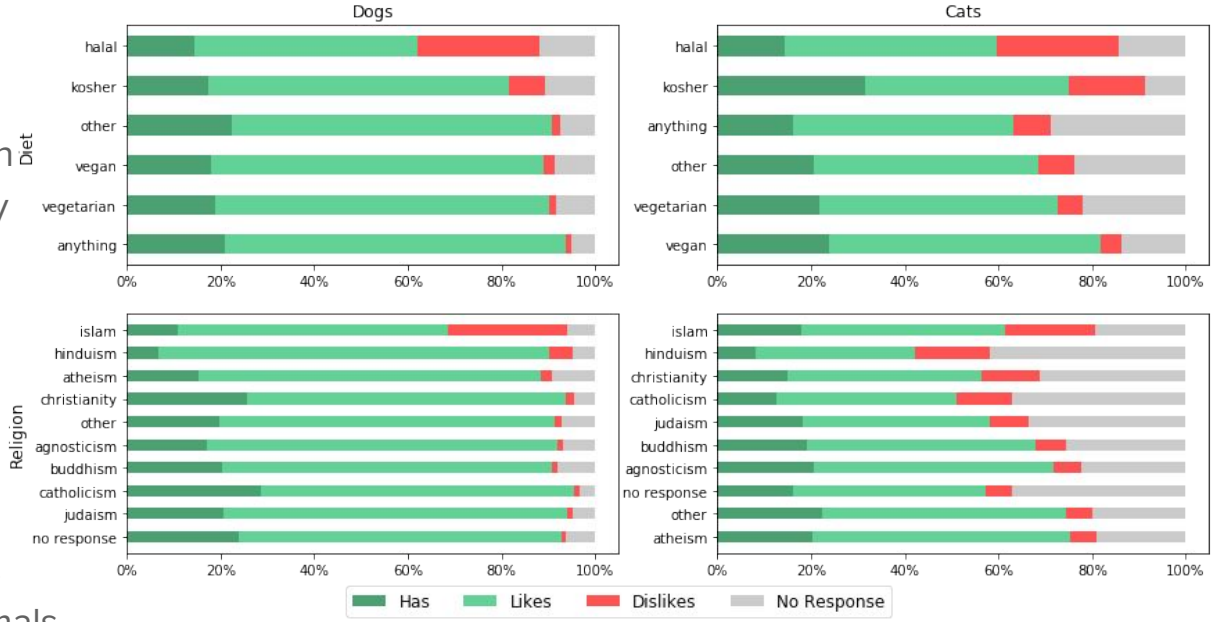
Question:

Are vegetarians and vegans more likely to like or own pets?

Answer:

Vegetarian and vegans are much more likely to like and own cats than the general population. Their affinity for dogs is similar to other diet practitioners.

One interesting and unexpected finding is that individuals who follow halal and/or practice Islam are generally less likely to own pets and more likely to dislike some animals.



Foolin' Around

How It Was Made: Columns Created Along the Way

- Definition of Short/Tall
  - Determined if the user mentions "short" or "tall" in their essay on what people notice about them

```
df['tall']=df[df['essay3'].notnull()]['essay3'].apply(lambda x: 1 if 'tall' in x else 0)
df['short']=df[df['essay3'].notnull()]['essay3'].apply(lambda x: 1 if 'short' in x else 0)
```

- Dietary Options
  - Created a dictionary of simpler diet options that mapped to the original report of diet

```
df['diet'].fillna('no response', inplace=True)
diet_options = list(df.diet.unique())
final_diet = ["anything", "vegetarian", "vegan", "kosher", "halal", "other"]
diet_dict = {}
for diet in final_diet:
    for option in diet_options:
        if diet in option:
            diet_dict[option] = diet
df['diet_final'] = df['diet'].map(diet_dict)
```

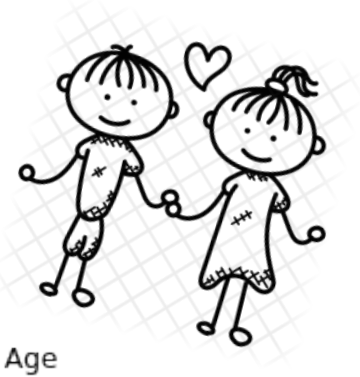
- Attitude Toward Cats/Dogs
  - Determined the word that preceded "cats" or "dogs" in the column "pets"

```
df['cats'] = df['pets'].apply(lambda x: x.split()[x.split().index('cats')-1] if (isinstance(x, str)) & ('cats' in str(x)) else 0)
df['dogs'] = df['pets'].apply(lambda x: x.split()[x.split().index('dogs')-1] if (isinstance(x, str)) & ('dogs' in str(x)) else 0)
```

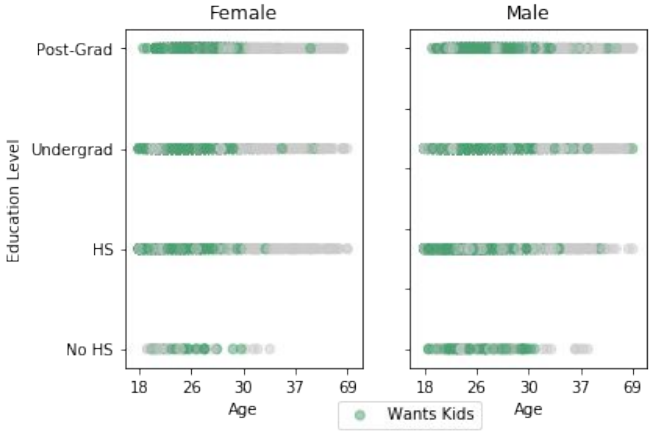
Honeymoon Phase

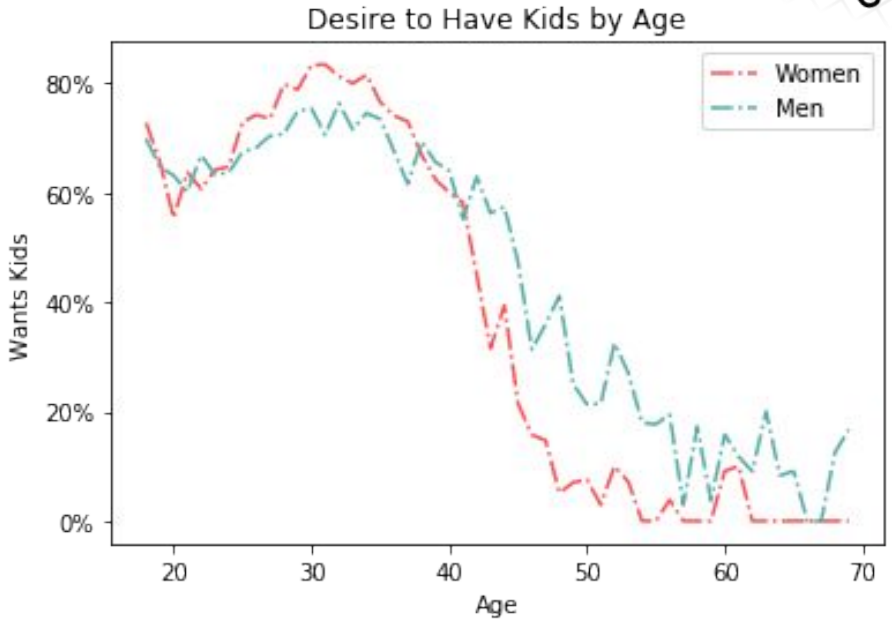
Classification

Classification



- Question: Can we determine who wants to have kids based on their age, sex and education level?
- Two Approaches
  - Support Vector Machine
  - K-Nearest Neighbor

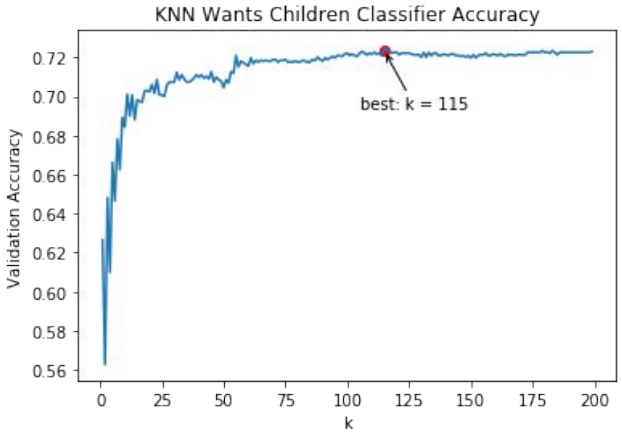




Classification

- Support Vector Machine (SVC)
  - Time to Run: 200 models in 738.5 seconds (~12 min)
  - Performance:
    - Accuracy: 72.3%
    - Precision: 80.5%
    - Recall: 31.7%
    - F_1: 0.455
- K-Nearest Neighbors (KNN)
  - Time to Run: 200 models in 23.8 seconds
  - Performance:
    - Accuracy: 72.5%
    - Precision: 79.2%
    - Recall: 33.1%
    - F_1: 0.467

Both models perform very similarly to one another. Overall, 63.4% of OKCupid users want kids, so both models outperform random. SVC's slightly better precision score means that of the individuals the model classifies as wanting to have children, SVC is more likely to be correct. On the other hand, KNN's higher recall score means out of the people who actually want to have children, KNN is more likely to correctly classify them as wanting kids. Since KNN has a significantly shorter runtime, better accuracy and F_1 scores, it is the preferred model in this case.



Gettin’ Serious

Regression

Regression

- Question: Can we predict how far someone lives from SF based on various demographic characteristics such as: whether they work in the tech industry, whether they have kids, their age, their drinking frequency and drug use frequency?

- Multiple Linear Regression
  - R^2: 0.00025
  - People in tech, with kids means live further away from SF
- K-Nearest Neighbors Regressor
  - R^2: -0.00090

It's not really appropriate to assess regressions in terms of accuracy, precision or recall because there are no "true positives", "false negatives", etc when working with linear data (not binary like classification). However, the R^2 score can give us a better idea of the superior model. Here, neither model is very good with R^2 scores hovering near 0 - meaning neither model is much better than drawing a horizontal line at the users' average distance from SF. Nonetheless, because MLR's R^2 is positive, it outperforms KNNR's negative R^2 and is slightly better than drawing a horizontal line.

Committed

Summary

Summary

- Some Findings We've Covered
  - Age, sex and education level are good predictors of the desire to have kids
  - Most everyone loves dogs, vegans/vegetarians like cats more than the average person, while halal practitioners are more likely to dislike some domestic animals
  - "Short" men and "tall" women share a height cut-off of 5'9" and men are more aggressive in their policing of normal height definitions than women
  - Straight women and gay men are more likely to highlight their "eyes", "smile" and "laugh" as noticeable features than straight men

- Next Steps
  - Expand data collection to other cities
    - Maybe there are differences in age/sex/education/industry across geographies
  - Data on the number of incoming and outgoing messages for each user profile
    - Could be used as a proxy for attractiveness
  - Data on which users made successful connections with each other
    - Could be used to determine markers of compatibility across profiles