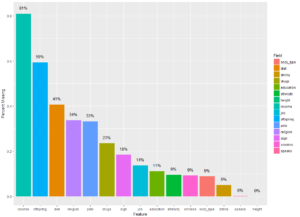
Online dating sites are an increasingly popular way to find a mate. With 17% of marriages sprouting from an online dating service, the formerly desperate activity has lost much of its stigma. More and more people are uploading profiles to attract a special someone. What insights can these curated self-portraits provide?

**The Data**

The OkCupid profiles dataset contains information from 59,946 San Francisco users. This dataset was scraped from active profiles containing at least one photo in June 2012 and made available to the public. Each profile contains information on gender, sexual orientation, ethnicity, religion, education, physical features, and more for a total of 31 possible features. Income, height, and age are the only available numeric features, making much of the data categorical. Ten possible essays were available as well, answering prompts such as "Describe your typical Friday night" and "What is the first thing people notice about you?"

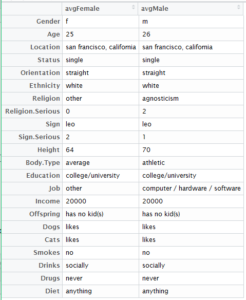
OkCupid allows the user to have fun with their profile and present a unique version of themselves, making the data messy. For some features, such as ethnicity and offspring, users are allowed to pick multiple responses, making the data messy. My first step was to clean this up so I could analyze the population by groups. I focused on ethnicity, religion, astrological sign, offspring, and pets. Observations with multiple ethnicities listed were categorized as "multiethnic" unless the only additional ethnicity was "other", in which case the ethnicity was stemmed to the primary category. Religion and astrological signs were separated into two columns each: one affiliating the user to their selected religion and sign, and the other noting the user's indicated level of devoutness to religion or astrology. Pets was split into a cats and dogs category, while offspring was split into "has kids" and "wants kids".

For certain categories, many people opted to keep information private.  In particular, OkCupid users seemed reluctant to list their income and whether or not they have or even want children.

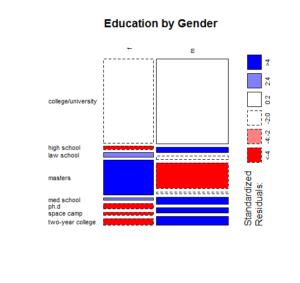
[](http://blog.nycdatascience.com/wp-content/uploads/2017/05/Capture2.png)

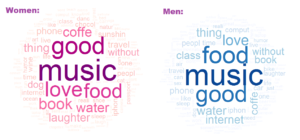
**Exploration and Visualization**

From the available data, many interesting relationships can be seen. First, we can look at the typical male and female by aggregating the most common responses for each category. Outnumbering women 3:2 in this dataset, the typical male is one year older than  his female counterpart. Much remains the same across gender lines, except that men seem to be taller.

[](http://blog.nycdatascience.com/wp-content/uploads/2017/05/average.png)

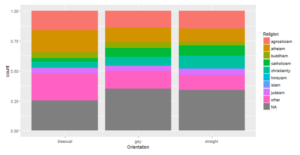
When looking at averages, however, the salary gap can be seen with mean making $24,351 more than women. The graph below shows the distribution of salary among men and women.  This disparity existed despite the fact that women on the site were more likely to have a masters degree or be in law or med school.

[](http://blog.nycdatascience.com/wp-content/uploads/2017/05/Capture3.png)[](http://blog.nycdatascience.com/wp-content/uploads/2017/05/educationbygender.png)Differences can also be seen in the way men women wrote about themselves. In a word cloud analysis of the most commonly used items that men and women "just cannot live without," it seems that women care more about love, books and coffee while men enjoy the internet and beer. Both groups like good food and good music.

[](http://blog.nycdatascience.com/wp-content/uploads/2017/05/menwomen.png)

Some men and women listed their highest level of education as space camp. Interestingly, this group showed significant differences from the general population. Those who listed space camp were 23% likely to sometimes use drugs compared to the average of 13%. Females in particular were twice as likely to not want kids if they claimed to have gone to space camp.

Another interesting group was bisexuals. Although only making up 5% of the users, this unique group showed significant differences in ethnic and gender makeup, with far more women describing themselves as bisexual than men. As shown below, they were also more likely than gay and straight people to be agnostic or atheist or ascribe to some "other" belief system. According to the data, this group was also more likely use drugs and own cats.

[](http://blog.nycdatascience.com/wp-content/uploads/2017/05/Capture4.png)

**Conclusions and Further Questions**

Although this data proved fun and interesting, it offers limited business insights. The goal of OkCupid is to subscribe and match users to spark a potential romance. The profiles do not have any information on communication between users whatsoever, so we can't conclude if a particular response or non-response correlates with higher levels of successful interaction with other users. We can say, however, that based on the ethnic and gender makeup of the profiles, OkCupid is missing out on potential members in the San Francisco area. Asians, Hispanics, women, and gay people are particularly underrepresented on the website. However, younger age groups are more diverse than older age groups, indicating this inadequate representation may change. Further research should still be conducted. Other interesting topics to explore would be message and reply rates by certain groups and why some users fail to stick with the site or even complete a profile.