OKCupid Final Project

Julian Benitez Mages, Manon LeDonne, Rose Silver, Shreya Simhadri

Northeastern University, Boston, MA, USA

Abstract

In this paper, our group analyzes an OkCupid data set from 2012 \cite{okdata}. Upon analyzing this data, our initial goals were to (1) create mechanisms to have a broad understanding of users from the data set; (2) create our own matchmaking algorithm; (3) use machine learning to predict traits amongst clusters of users; (4) look for non-obvious insights. We show success in these investigations through analysis in word clouds, correlation charts, and demographic maps. We also create tools for sentiment analysis, match-making algorithms, k-means clustering, and random essay generators to glean insights into the data set. Source code for this project can be found at https://github.com/rosesilver/OkCupid.

Introduction

OkCupid is a free online dating website which has helped over 90 million users each year find a significant other \cite{OKCUPID}. It is regarded as one of the original online dating tools that has gained substantial popularity. An important part of what makes an online dating platform so attractive and successful to users is its ability to quickly find its users suitable matches.

Matching users is a heavily data driven process. Matching algorithms take full advantage of the many records of each user's basic information, lifestyle, and personality. Over the years, OkCupid, as well as the other popular online dating websites, have hired teams of data scientists to figure out exactly how to optimize the potential partners presented to each user. With better algorithms comes more favorability among consumers, making OkCupid the obvious choice for online dating. With this status, they would be able to increase profits through a larger market presence.

In this paper, our group analyzes an OkCupid data set from Kim et. al \cite{okdata}. With over 30 different statistics for each user, the data set is powerful, allowing for many different questions to be asked and answered about the user population. We approached this data set with four goals in mind: (1) Create mechanisms to have a broad understanding of users from the data set. (2) Create our own matchmaking algorithm. (3) Use machine learning to predict traits amongst clusters of users. (4) Look for non-obvious insights.

We have ways of understanding the data at large. In Section \ref{sec:clouds1}, we introduce word clouds to get an understanding into the types of users. In Section \ref{sec:pie1}, we introduce correlations and pie charts for another way of understanding the users. In Section \ref{sec:maps1}, we also use maps to get another visual understanding of the population to look for patterns. Using this understanding, we were able to create match-making algorithms in Section \ref{sec:perf_match1}.
In Section \ref{sec:personalized_essays1}, we created personal essays to create a random user of our own. Finally, in Section \ref{sec:kcluster1}, we used k-means clustering to look for patterns among groups of people.

This project holds lots of potential significance for OkCupid and the future of online dating, as well as to maximize opportunity for advertising firms. In particular, new matchmaking algorithms can help to increase traffic on the site, which will in turn result in better matchmaking on its own, as there would be a much larger pool for the algorithms to pick from.

With greater web traffic comes greater consumers, and with this knowledge, OkCupid will want to sell more advertising space on their site. Not only will this increase profits, but those profits can be used for OkCupid to take out more ads of their own on other websites and locations. Our demographic data will make it easier for OkCupid to pinpoint their target audience, and advertising firms will have an easier time finding clients to advertise on their site. With an increased public presence, OkCupid will only be able to grow its following, in turn making our algorithms more effective, and the company, as well as its users, better off.

Data Sources

The first data set we used is from Kim \cite{okdata} published in the Journal of Statistics Education. We acquired the data from a GitHub repository from Kim's profile, rudeboybert. To download the file, we simply clicked on the file name in the repository and saved the csv file to our respective repositories. Unfortunately, within a week of finalizing this report, the Github repository used for this data source has either gone private or has been deleted. There have been subsequently new repositories also created by Kim which feature updated 2021 okcupid data, but unfortunately this is not the same as our original data. While Kim's publication in the Journal of Statistics Education \cite{okdata} originally includes links to the data, these links no longer work.

The data set includes over 4000 OkCupid users from the year 2012. Every user represents a row in the csv file. Every column of the csv file is an attribute. There are 31 attributes total. The attributes are: 'age', 'body_type', 'diet', 'drinks', 'drugs', 'education', 'essay0', 'essay1', 'essay2', 'essay3', 'essay4', 'essay5', 'essay6', 'essay7', 'essay8', 'essay9', 'ethnicity', 'height', 'income', 'job', 'last_online', 'location', 'offspring', 'orientation', 'pets', 'religion', 'sex', 'sign', 'smokes', 'speaks', and 'status'. All of these attributes were used in the project. However, the attributes had HTML

tagging and needed to be cleaned. We used a regular expression to sanitize the data for future use.

The second data set we used is from SimpleMaps, courtesy of the United States Census Bureau \cite{cities}. We found this data set by searching the internet for data sets with cities and coordinate locations. Underneath Databases, there are a list of files which can be downloaded or bought. For this project, we used the free “Basic” csv file listed in the first column of the website. To obtain the csv file, we simply clicked the “Download” button and saved it in our repository. The data set includes 28,372 entries. Each entry represents one city. Every column of the csv file is more information about the city.

The columns are: ‘city', 'city_ascii', 'state_id', 'state_name', 'county_fips', 'county_name', 'lat', 'lng', 'population', 'density', 'source', 'military', 'incorporated', 'timezone', 'ranking', 'zips', and 'id'. For this project, the only information necessary was the city, state name, state id, lat, and lng columns. We created a dataframe which only used these columns from the data set.

Methods

In this section, we introduce the computational techniques used to gather and analyze the data in our project.

Word Clouds

We chose to study word clouds as a way of understanding the users in our OkCupid data set. In particular, we wanted to create a word cloud which encapsulated all of the responses from all ten essays in the data set. In order to accomplish this, we created a frequency dictionary so that when looping through all of the essays, words would continuously be added to the dictionary. While looping through each column of essays, we would also split the words up to run it through an if statement to determine two things. (1) If the word fit the minimum letters wanted. We made sure to filter the words by the number of letters because we did not want filler words such as – I, a, the, you, me, etc– to over crowd the more important words. So, if the word fits the minimum letter requirement, then (2) determine if the word is already in the dictionary, and if so, add a plus one to the word’s frequency. But, if not, add the word to the dictionary and add a frequency of one to it. In the end, this frequency dictionary would be used in making the word cloud through one of the Word Clouds functions.

Correlation Plots, Average User traits, and Pie Charts

The OKCupid dataset used for this project had lots of information and we needed to gain a big picture understanding of our users before going any further in data analysis. This corresponds to our first goal of getting a feel for what the users are generally like. We initially created pie charts by summing up the count of each occurrence of the variety of different values and then using those values to create the pie charts for various traits.

Before implementing this, however, we had to go through the responses for various traits. We found that many of the traits were short phrases or words that could not be plotted or serve a function without a numerical value. So, we assigned numerical values to smoking habits, drinking habits, sexual orientation, and body type to be able to visualize this data. These traits were chosen because the reponses lie on a general spectrum, making it fairly easy to assign a numerical value in relation to other responses.

But, there are some limitations to the assignment of numerical values for body type, because it was somewhat difficult to decide where to place ‘athletic’ or ‘jacked’ on the spectrum from ‘thin’ to ‘overweight.’ It was decided that BMI would be the general marker for where to put values on the scale from 1-9.

Other phrases were easier to quantify like sexual orientation because the Kinsey scale was actually used for these values. The Kinsey scale is used in research to describe a person's sexual orientation based on one’s experience or response at a given time. The scale typically ranges from 0, meaning exclusively heterosexual, to a 6, meaning exclusively homosexual.

First we convert the csv file into a dataframe for easy manipulation. By iterating through each index in the dataframe, we checked the user’s answer and appended the corresponding numerical value to a list. This was done for multiple traits. If an answer was unclear regarding which value to assign, a NaN value would be appended. NaN values were also appended for all NaNs in the dataframe. Then, these lists of values were used to create new columns in the dataframe. Then, the original columns that had strings as answers were dropped and rows with NaN values were also dropped.

With the quantified data, we created an average overall user and an average male user and average female user so that we could make distinctions in trends and for gender dependent traits, like height. Values were summed up and then divided by the last index number in the dataframe. We also used this quantified data to create correlation plots with linear regression and using an individual column as the x axis or y axis value in coordinates.

Demographic Maps

One of our goals of this project was to be able to visualize our users based on where they lived and what qualities they shared. To do this, we used open source geospatial data from geopandas. We also used the United States Census Bureau data which included every city and

state in the United States as well as its longitude and latitude. We merged this dataframe with a dataframe of all of our users. We requested a public Mapbox Access Token and used Mapbox maps with plotly to visualize our data. Because many users from the same town/city were associated with the same GPS coordinates, we added random noise to each person's longitude and latitude, so that the users could be a little more spread out on the map. We created random noise proportional to how many users were from a certain city. For example, since approximately half of the users were from san francisco, we made the "noise" around san francisco larger to represent a large population of users living there. From there, we were able to color code the different users based on the qualities they listed such as age, gender, drinking habits, drug use, etc. For qualities that were non-numerical, we used sentiment analysis techniques to convert the qualities to numerical values to use for color coding.

Sentiment Analysis

In continuing to complete a goal of looking into non-obvious insights, we decided to look into sentiment analysis of the essays. More specifically, to see if there was any difference in attitude between how males and females wrote their essays. We assigned an essay a number closer to 1 if it was written with a more positive tone, while more negative responses would be assigned a value closer to -1. Since our data set is so large, we decided to go through each essay column individually instead of combining all the essays together like we did in the word clouds. We filtered the entire data set by sex, so that the result would be the sentimental values of male, "m", and female, "f". In addition, we created a sentiment function in order to make sure only strings were being used into the sentiment analysis because we found that in many of the essays, users wouldn't always write with words.

After making sure that the sentiment analysis was calculated for each essay, we then sorted the values by sex and then took the mean of all of them, to calculate the mean sentimental value of the male and female essays.

Perfect Match

Given that OkCupid is a match-making service, we thought it would be interesting if we could play matchmaker ourselves. In particular, one of our goals in this project was to find the perfect match for any chosen user in the data set. For this project, we define a "perfect match" to be a pairing of two people who share the most in common. The data set we use (ref) records 31 unique attributes for each user. Some of these attributes include age, gender, personal essays written by the users, and religion. To determine the perfect match, we wanted to find two users who have the most pairwise similar attributes across all attributes.

One of the challenges in finding a perfect match is quantifying what it means for two users to be similar. This task is easy if the pairwise attributes between two users are numerical values, such as their age; one can use any distance metric, like euclidean distance, to determine

how close two numerical values are to each other and search for the users with the smallest distance. However, the task becomes more difficult when trying to quantify similarities between written responses, such as the personal essays. There's even another layer of challenges if we consider that some attributes are more important than others in the match-making process. For example, some users might prioritize finding a partner who has a similar personality or educational background over a similar age.

Our project manages to maneuver these challenges in the following way. For attributes which hold numerical data about each user, we used a euclidean distance metric to find the distance between two user's attributes. The smaller the distance, the more similar these two users are. For attributes whose values are non-numerical (written-responses), we used a cosine similarity measurement in the following way: Let's suppose there is an attribute X that User A and User B each have a written-response for, and suppose we want to compare the similarity between these two written-responses. We can create a Counter object which keeps track of all of the unique words used in User A's response to X as well as how many times each unique word is used. We repeat with User B. We add a unique twist to traditional cosine similarity by normalizing the frequency of each word in User B's counter object based on how much it is used by all users in the data set. The goal here is to give more importance to words which are highly unique in the data set, and to give less importance to words that are frequently used by many users. For example, if two users both include the word "unicycle" in their written response to X, then this will be weighted more than if two users both include the word "like", which is more frequently used amongst all users. We also completely remove "stop words" entirely from both counter objects. Stop words include definite/indefinite articles, prepositions, conjunctions, and other words which do not add any information about a user. We then use a variant of standard cosine similarity to find the overlap between these two counter objects once they have been properly normalized.

Putting the pieces together, to find the perfect match for User A, we look at all users and one-by-one compute all pairwise distances for each attribute to User A. We tuned by hand the relative weights of different attributes in our matching program. We sum together all of a user's distances to User A (negating the cosine similarity distances so that similar users have smaller distances) to get one total distance measurement between User A and this user. Out of all of the total distances, the perfect match for User A is the user with the smallest total distance. We also included functionality in our code to prioritize only one attribute (instead of all attributes) when looking for the perfect match for a chosen user.

Personal Essays

One of our goals was to be able to create a superficial user of our own which is representative of the user data at large. Since all of the users have personalized essays, our superficial user would need to have one as well. We used the method of n-grams to write

personal essays for our superficial user. We analyzed all of the personal essays written by the users. We created a dictionary of words, where the keys in the dictionary consisted of words which were found in a personal essay. The value associated with each key is a list of words which were found to appear after the key in a sentence. In addition to this dictionary, we created a list of all words which were the start of every sentence, and we created a second list of words which were all of the words at the end of every sentence. To create a random sentence, you choose a random word from the starting list to look for in the dictionary. This will be the first word in the sentence. We then find a random value for this key in the dictionary, add this word to the sentence, and find the values associated with this word in the dictionary. We repeat this process until we find a word which appears in the end list, at which point we include that word and end the sentence. To create a random personal essay, we concatenated an arbitrary number of sentences together.

K-Means Clustering and Heatmap

We used the k-means clustering method we learned about in class, using k centroids and stabilization with repetitive clustering, to model our sample users.

One of the biggest hurdles we faced was figuring out how to graph attributes that weren't quantitative (essentially everything except for age and income). We ended up using a similar method to what we used to make pie charts, where we counted up all the results for a given attribute and placed them into numerical buckets which represented a spectrum. To minimize computing time, we quantified all our qualitative attributes in one iteration of the users. To illustrate, I'll focus on body type as an example. Like with our other quantifiable attributes, we created an empty list, which would be added to the dataframe later. For each user, we appended an integer to the list which represented their written response. In order for it to look any good on a graph, we created a spectrum to linearize the different body types, from 1 to 9 with 1 being "skinny" and 9 being "overweight." The end result was a list of integers of length n, with n being the number of users.

After we quantified the attributes we chose to graph, added them to our dataframe as columns. Because one of the parameters of this function was a list of 2 attributes that we chose to graph, we iterated through all the columns and dropped all but the two chosen ones, represented by a list called "keepers." We then returned this skinnier dataframe, and sent it through the series of functions used to create the k-means visualization. We set our k value to 5, and decided that 10 instances of the graph was enough for it to stabilize, and got some interesting results. **We later decided to tweak the K value to get some more specific clustering.**

Once we realized our issue with the k-means clustering plots, we turned our attention to making a heatmap. This was quite the task, as we essentially had to tally up each time a given

combination of two attributes occurred. To do this, we took a list of the possible results for each attribute (e.g. all the possible ages), and cycled through every possible combination. With each combination, we tallied up the number of times it appeared in our dataframe of users, creating a list of length xy, with x being the possibilities for attribute a and y being the possibilities for attribute b.

We then created a new dataframe, with each line having 3 values: the x value, the y value, and the number of times that combination appears. With this, we were able to pivot it and turn it into a heatmap. This code takes a very long time to run because of the amount of checks it has to do to tally up the occurrence for each possible attribute combination, which became one of our limitations and issues. In the end however, the heatmaps came out quite well. The last thing we had to do was change the color scale because of how so many data points were toward the lower end.

Analysis and Results

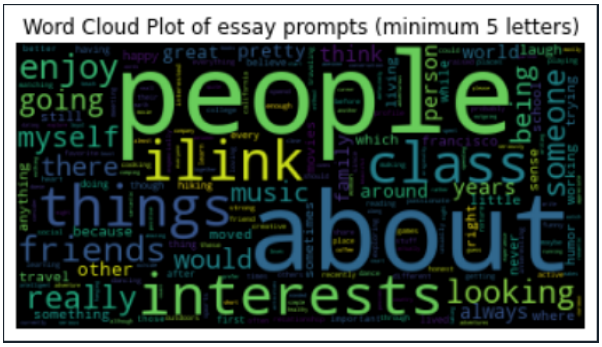
The results of processing, tabulating, and visualizing your data go here.

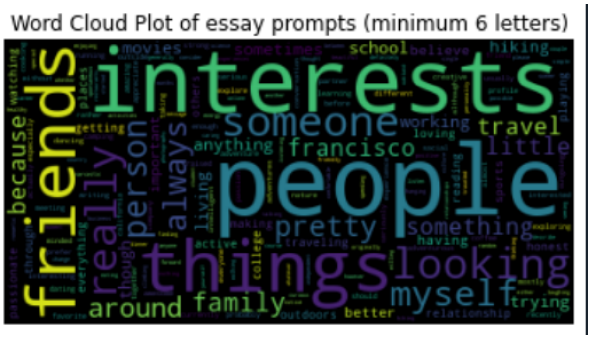
1. Word clouds MANON
2. Correlation plots and pie charts JULIAN AND SHREYA
3. Demographic maps ROSE
4. Sentiment Analysis of essays MANON
5. Perfect match ROSE
6. Personalized Essays ROSE
7. K-means clustering (targeting ads) SHREYA AND JULIAN

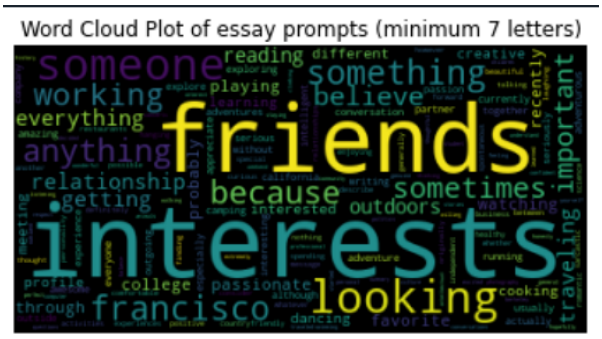
Word Clouds

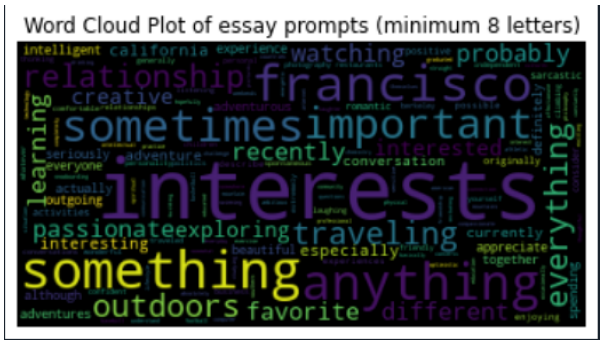
One of our first goals was to gain an understanding of our user and consumers. One way we did that was through analyzing how they wrote their essays. In the data set there were 10 different essays, each essay contained a different nuggets of information about the lives of our

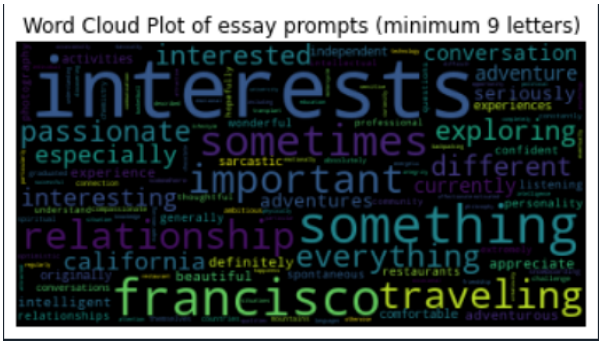
users. One way we looked at these essays was through the use of word clouds. In order to create a greater understanding and generalization of the essays, we created one large word cloud of all the essays combined. Looping through each of the essays we were able to come up with a word cloud that represented the most common words the users used together. In the end, we made about six different word clouds ranging from a minimum of five letters all the way to ten lettered words. From the visuals you can notice that the higher the minimum number of letters go the more interesting the words are. For example, the words become adjectives of the user's personality or examples of activities they like to do. Found below are the six word clouds:

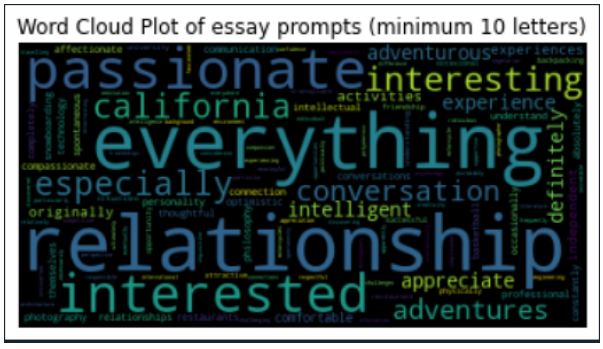






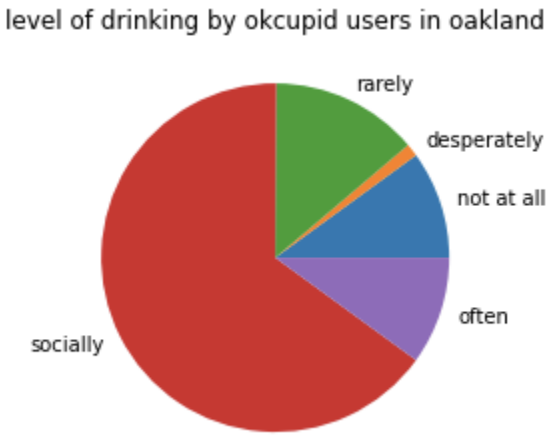


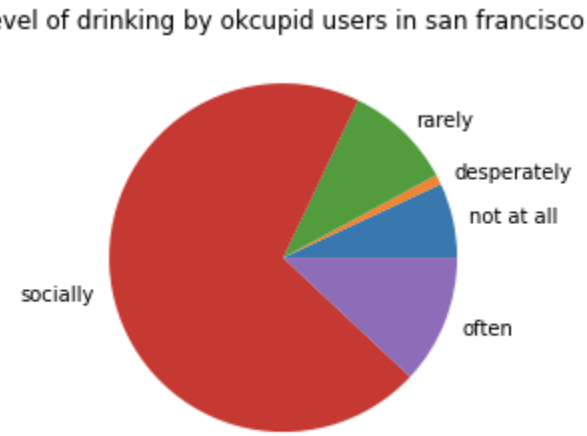




Pie Charts

A key way we wanted to analyze our data was through correlation plots and pie charts. It is easy to gather a large amount of insight on our users through these analyses, as they show interesting connections, or lack thereof, between key attributes amongst OkCupid Users. One of our first pie charts was the comparison of the rate at which people drink in the two most populated OkCupid cities, San Francisco and Oakland. To create this, we iterated through each user and tallied up the reports for drinking level, and set a different reported location as a function parameter.

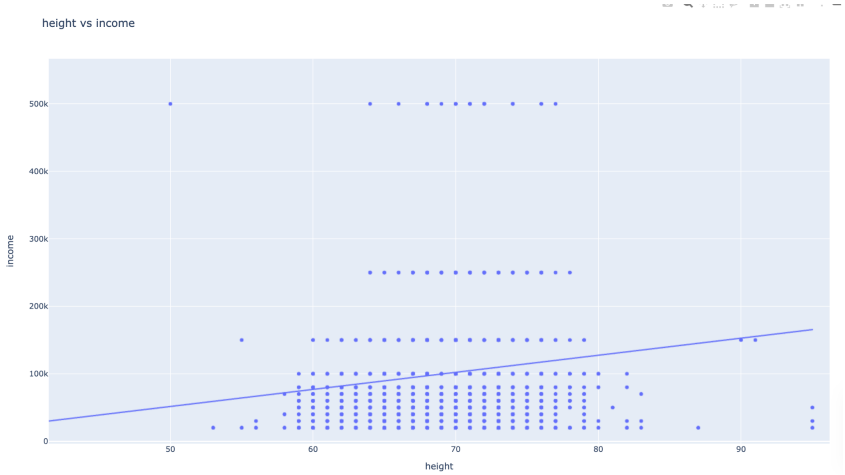




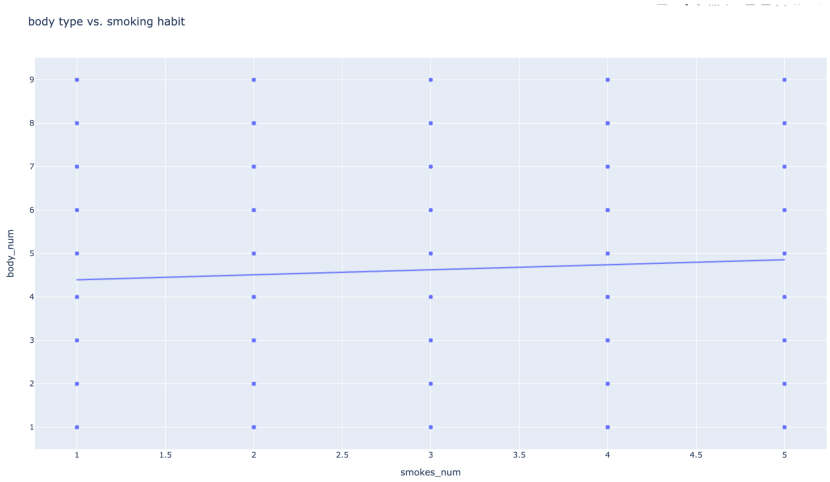
Correlation Plots

This section of the results is more focused on achieving the fourth goal, which is to find not-so-obvious insights about the users. A variety of different traits were plotted against each other, but the strongest correlations were interestingly between:

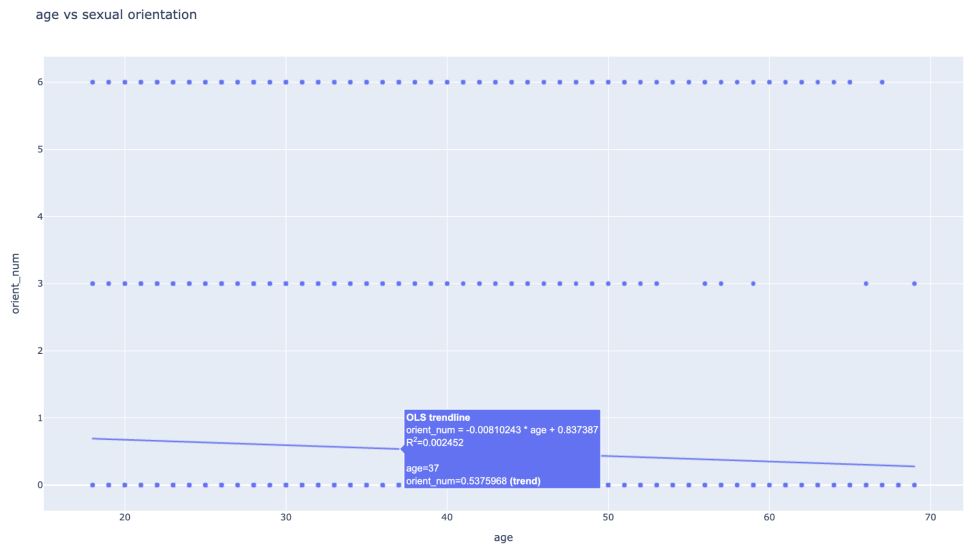
height and income, with every inch taller is a gain in $2530 in annual income (2530.9*height + -75012, r^2 = .0036),



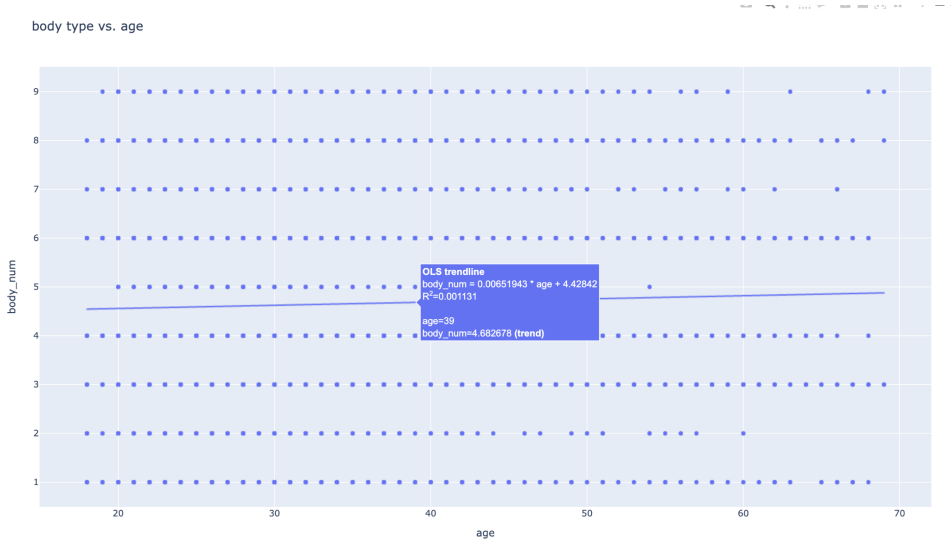
body type and smoking habits, with increasing habitual smoking correlation with a larger body type (body type = .115567*smokes_num +4.2787, r^2 = .00357)



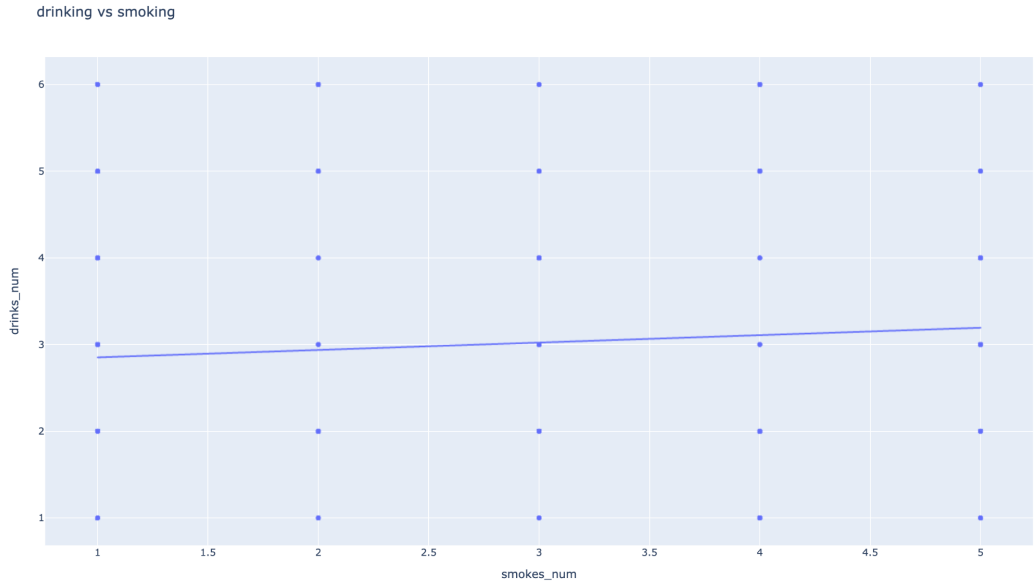
age and sexual orientation, with older people being associated with a more heterosexual orientation with a decrease in .008 on the Kinsey Scale for each year of age increased, (-.008*age +.837, r^2 = .0025)



Body type and age, with a .0065 increase in the assigned value for each year older (.0065*age +4.43, r^2 = .001131,



and smoking and drinking habits (drinks_num = .855*smokes_num +2.76, r^2 = .018).



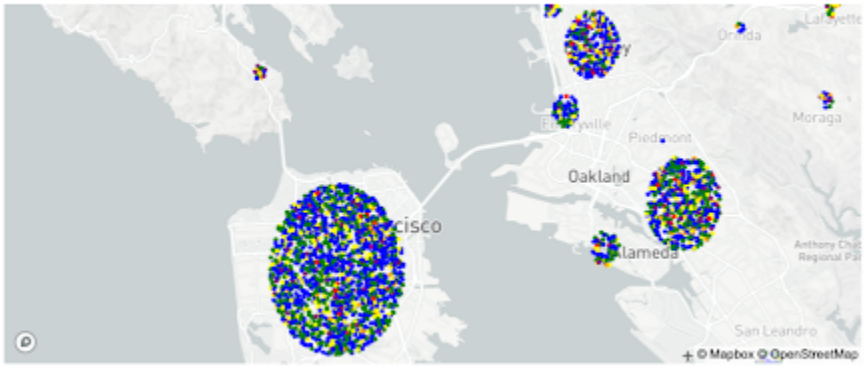
Surprisingly, there was no correlation between age and income, with an r^2 value of .000005.

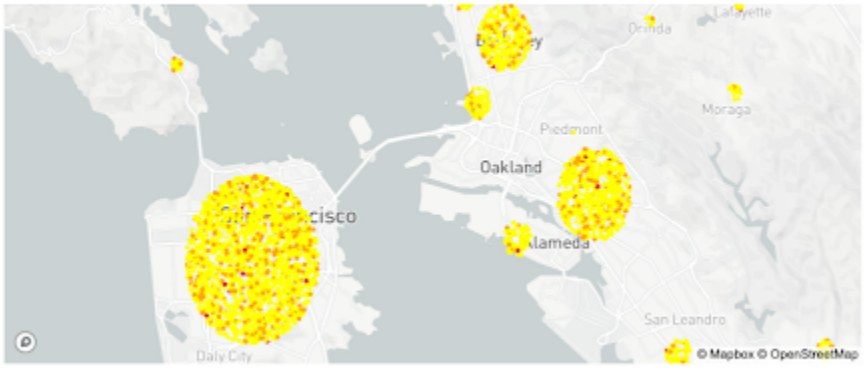
The limitations to this type of analyses was the low variability between traits because they often only fell into a single category out of a handful of possibilities after quantifying the values. Nevertheless, it was interesting to see these correlations and infer what it could mean amongst the general population as well. The R^2 values were also very small, due to this low variability. But, certain traits like drinking and smoking were correlated with each other. Height and income were correlated with each other with relative strength. Sexual orientation plotted against age was also interesting because of how it showcased the increase in heterogeneity amongst older people.

Demographic Maps

Our results gave us some insights about the data that we had not realized. For example, we did not expect for most users to be located within a 50 mile radius of each other in California. We also did not expect most people in the data set to even be in the same city!

Once we were able to understand where most of our users on OkCupid were from, our most interesting results sampled the age, drinking habits, and religions of these users. We found that most users were in their 20s and 30s, but the range of users was between 18 years old and as late as early 60s. We also found that most users drink casually and socially, but a small fraction of users prefer not to drink or prefer to drink heavily and often. Also, we see in these maps that the religious backgrounds of users are very diverse. Most users however are Agnostic, Atheist, or Christian.





Sentiment Analysis of Essays

Another way we analyzed the OkCupid data is through sentiment analysis. We wanted to understand if there was a significant difference between how women or men wrote their essays. In order to make the program more manageable, as well as, making sure that the data was mostly qualitative rather than quantitative, we used only one essay column. From looking at the answers of the essays, we can assume that essay zero is an introduction of the user themselves, and because of this the answers to this essay prompt will be very subjective making it a good candidate for the sentiment analysis. However, we found that there wasn't an overwhelming difference between female and male essays. Women had a score of about 0.218 while men had a score of about 0.196. While we believe that using essay zero will have the most substantial results, we did try the sentiment analysis on other essays as well. We found that women's score was consistently .01-.03 a little higher than that of the men's score. The only exception being essay 8, where women were lower than men. After researching, we found that essay 8's question is "The most private thing I'm willing to admit". In this case, we assume that women admit to more negative things than men do!

Perfect Match

We had a lot of success with our results. The most successful matchings were the ones where we focused on choosing partners based on 'essay0', the attribute which allows users to describe themselves. We found that many users give unique answers to this attribute, thus filtering only for this attribute makes finding unique matches possible. Below in Figure (??), we see an example of two users who had very similar 'essay0's who would probably make a great match. Finding a perfect match using all attributes also yielded some very similar pairs of users. For an example of a complete perfect matching report, see the Appendix. In the future, we would improve our match-making experience by adding filters so that people get matched with the appropriate sex. In the meantime, our algorithm has the potential to generate many compatible friendships, if not romantic relationships.

```
------------------------  essay0  -------------------------
***RANDOMLY SAMPLED HUMAN***
i am an eccentric artist type. it is actually unusual to find people who i truly connect with although i get along wi
th almost everyone i meet. i really like history and the study of it. yes, i'm a bit of a nerd who loves watching doc
umentaries and reading history books. i think i may have lived quite a few past lives.
***BEST MATCH***
i am a poli-sci and history nerd. i am not from this neck of the woods and i have a british accent.
```

```
------------------------  essay0  -------------------------
***RANDOMLY SAMPLED HUMAN***
well, for starters, my name is craig. i was born and raised in san francisco... okay... i lied... i was born in red
wood city and raised in south san francisco... but who's keeping track? i "graduated" from san diego state in 2010
with a degree in communications/journalism with an emphasis in media studies and a minor in television/film/new med
ia. sin embargo, i'm waiting to get my diploma (so i technically walked in 2010 but i won't be fully graduated unti
l 2012... stupid huh?) because i have to finish one more foreign language (spanish) course first. i'm honestly, a r
eally big kid at heart. i love me some disney, i love me some theme parks, and i love me some non-alchohalic drink
s.
***BEST MATCH***
i love photography. i am nerdy. i am a handful. i like working on media, snowboarding, playing music. i like being
romantic. i hike and camp in the desert. i am working on a degree in media studies and film. i hate fox news.  i'm
looking for a ltr or lotr. either one is fun.
```

```
------------------------  essay0  -------------------------
***RANDOMLY SAMPLED HUMAN***
hola! i'm a positive man, love is my religion. i love my life, family, job, and people. after working in restaurant's
for 7 years, i have found my balance as a chef working normal monday-friday am hours. i'm pretty busy, rock climbing
in the evenings, drinking a few pints from time to time, keeping up with the garden on the weekends, traveling around
the bay, i'm spontaneous and don't make tons of plans. i am looking forward to meeting intriguing, intelligent, and b
eautiful woman that have a stable career. someday i will be a family man, and hope you want to take the journey with
me. i believe the most rewarding experience in life is being a parent to a beautiful child. i love quotes.. recent fa
vorites: "if you can't put your heart into it, take yourself out of it" "for every minute you spend angry, you lose s
ixty seconds of peace of mind"  deal breakers: 1) you smoke cigarettes 2) you don't value family 3) zero confidence
4) low energy  deal makers: 1) you respect your body 2) you sing like nobody is listening, dance like no one is watch
ing 3) you smile 4) your motivated and can attain anything you set your heart on  i hope the above doesn't come off a
s rude, i just don't want to waste anyone's time. if we didn't work out as a couple, i love to meet new friends nonet
heless to enjoy this awesome area and world we live in. i hope to hear from all of you, don't be shy message me 8-)
***BEST MATCH***
i'm currently an acting student in san francisco(now taking time off for my job as a nanny), and i hope to start scho
ol to be a nurse as a back up plan soon.. growing up i had many ideas on what to be when i grow up. two things i coul
dn't get rid of was being a neonatal nurse, and being an actress.  i am not the type of girl to settle. i want to be
happy in life. i am not living the most healthy life, but i want to learn more about it and be able to live it. my na
nny position has taught me a lot about the benefits of living with a positive attitude and has opened up a new desire
to learn to live a healthier happier life.  my interests in special someone:  they can be whom ever they are. the onl
y thing i ask is that they are okay with: 1. be open 2. i want kids. its the most important thing to me!!! 3. i like
to cuddle and just spend time with my special someone. 4. you don't need money to have fun or be happy. 5. sorry guy
s, no sex before marriage.  interests: acting, ballroom, fishing, hiking, going on walks, rock climbing, having fun a
t the beach, listening to music, spending time with friends, anything to do with computers etc. i do a lot as i am st
ill trying to find my place in this life.  if there is anything you want to know that i didn't cover, feel free to as
k.  hope to hear from you soon! &lt;(^.^)&gt;
```

Personal Essays

As expected, it seems like most of the sentences in each personal essay were run-on sentences. However, we were surprised to see that most sentences included both a subject and a predicate. Generating an essay that made logical sense was always a gamble; perhaps one in every six iterations of the personal essay yielded an essay with some logic. It was funny to see that many randomly generated essays had a somewhat-sarcastic tone. This could be either reflecting the tone of the personal essays given by our data set, or it could arise naturally from the n-gram algorithm. Highlighted below are some of our favorite short personal essays.

```
 nothing. im an artist who can to me. i get. nothing. i just completed my passion and always focused on my political
science and exploring finding a sustainable designer adjunct faculty member helloooo san francisco.
```

```
 i make your doctor both look forward to finish up your phones ring. i still looking into aviation as i go back to ha
ving as they seem to travel plans to do. i chat with people. everything possible interested in the risk of the whole
world check it. currently working in a couple of skiing and fresno.
```

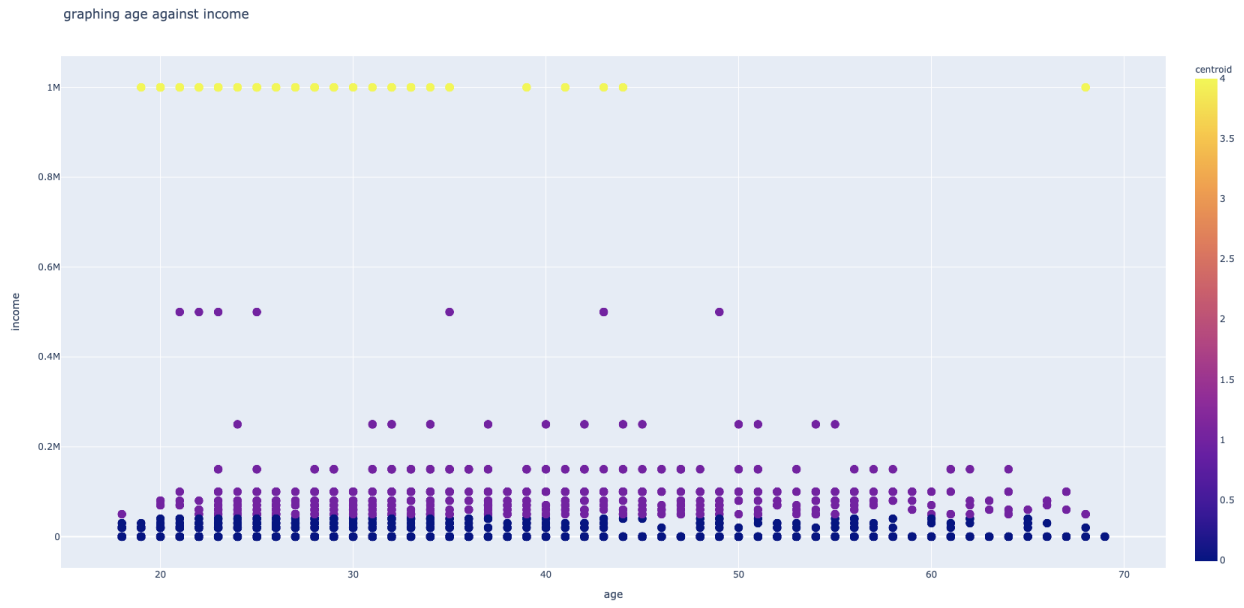
```
i write runon sentences for school. i do. im being a regular basis coping with me. learning new. living it.
```

K-Means Clustering

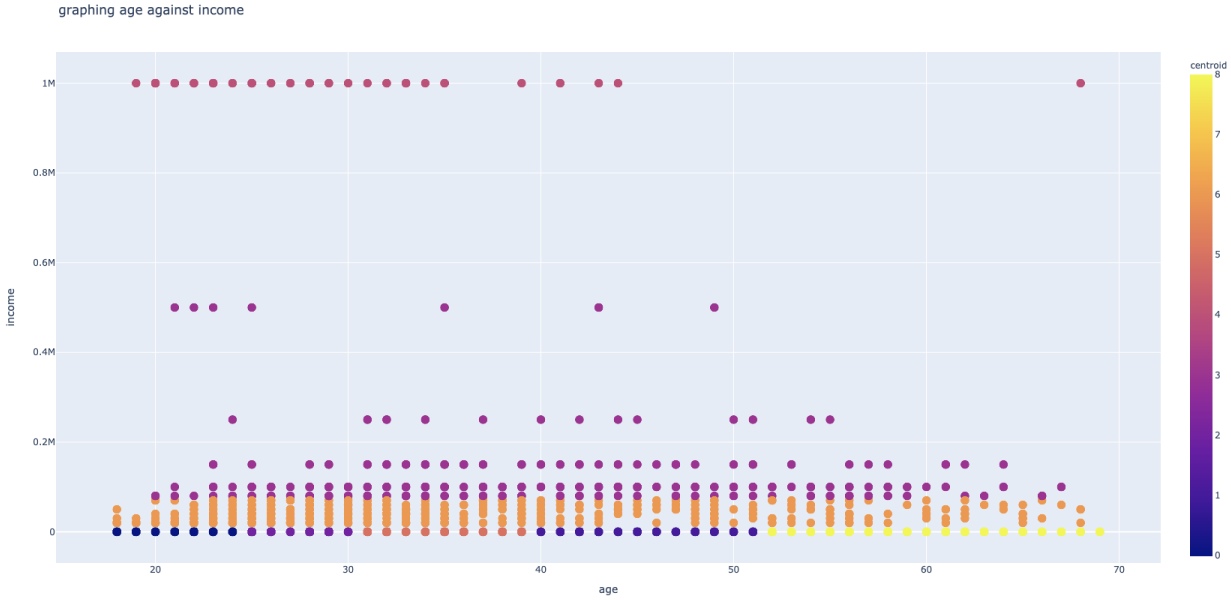
One of the biggest analyses we chose to do is k-means clustering. This is highly important for our overall motivation, as we can cluster our users by key attributes like activities, location, and income. If they are able to be clustered into large groups, it would help us determine where OkCupid should place advertisements in an effort to attract new users. Below are the results of our tests at k-means clustering, with k set to 5, and the attributes age and smoking quantity being graphed against one another.

We ran this on a few different attributes, with an example being age vs income.

We first did it with k = 4:

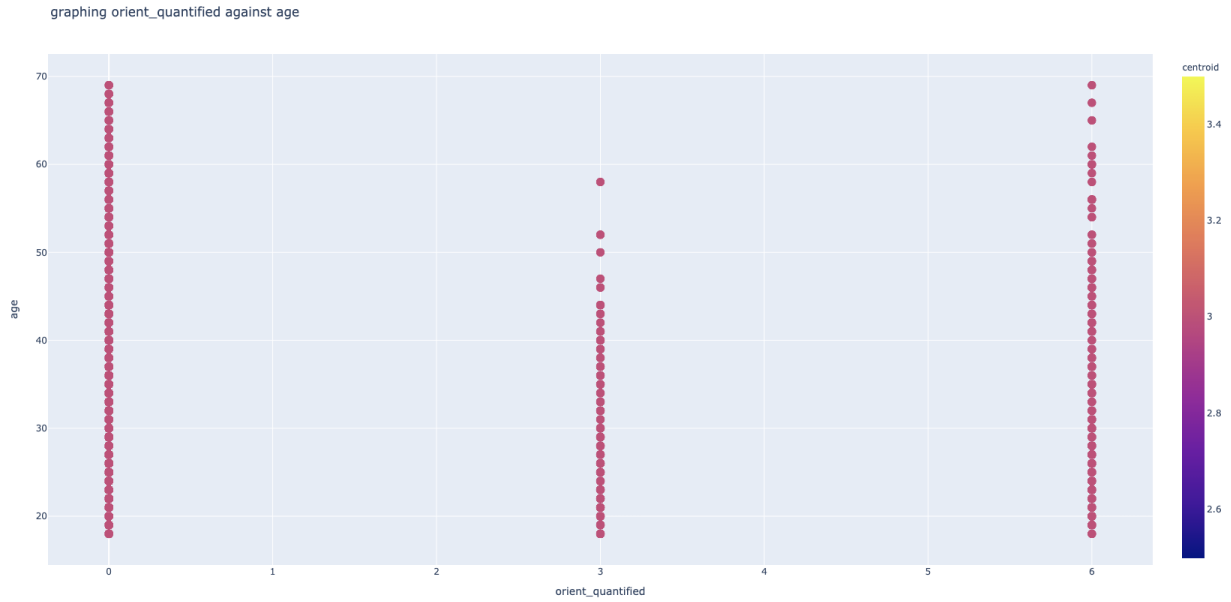


We decided it'd be smart to increase the k value, to get more clusters and see if we can group our users more specifically. Here is a plot with k = 9 of age vs income



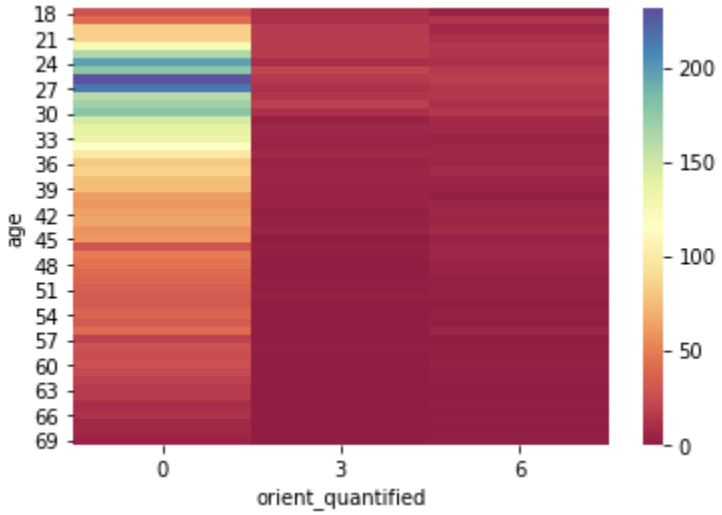
We also chose to graph height against income, with a higher k value again.

Lastly for k-means, we graphed age against sexual orientation.



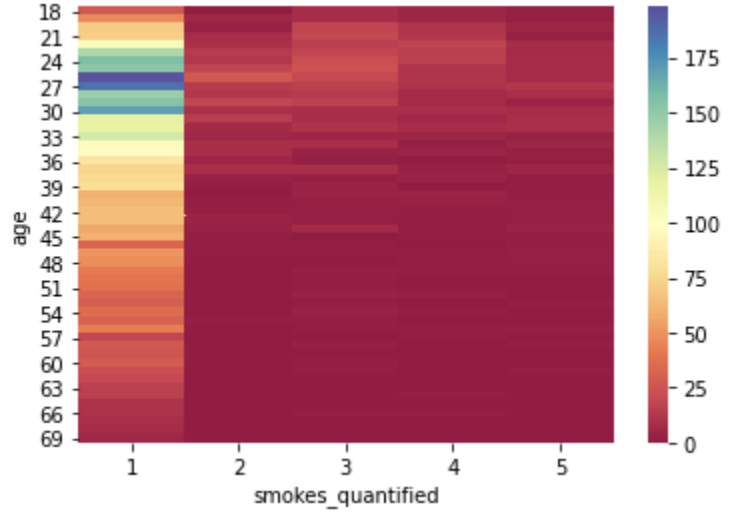
It didn't take us long to notice a major issue with our result. Because the smoking quantity can only contain 5 values, almost every value is filled, and there are no clear clusters. **With no clear clusters, this couldn't be used to gather information as to where to place advertisements, so we had to turn to heatmaps.** If we wanted to compare attributes that were quantified, a k means clustering plot wouldn't work, so we plotted heat maps comparing attributes, as well as how many people share each data coordinate.

The first heatmap we did was sexual orientation versus age. (0 = straight, 3 = bisexual, 6 = gay)

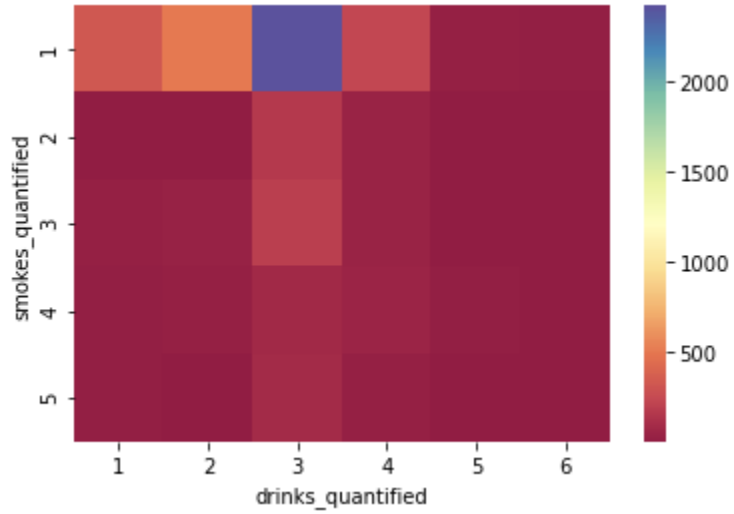


From this it's clear to see that the vast majority of users are straight, and they are also in their mid-twenties. It is also worth noting that people who report being gay or bisexual also tend to be younger. This could be due to different beliefs about sexuality in different generations, and is surely interesting to note.

Another heatmap we did was age plotted against smoking. Smoking was one of the attributes that we had to quantify, with 1 being no and 5 being the most amount of smoking reported.



Similarly to the age versus sexual orientation, this showed that the vast majority of users are in their mid-twenties. We expected older people to be the ones who smoke more, but that doesn't seem to be the case.



Lastly, we plotted drinking against smoking. It is clear that drinking in the mid-range is the most common, and that includes amongst both smokers and non-smokers. This is quite interesting, and

shows that the vast majority of OkCupid's users drink at the level called "socially." **From this, it would be a safe recommendation for OkCupid to place advertising in liquor stores.**

These were a far better analysis, and through these heatmaps we were able to make some observations, which will help our overall goal.

Conclusion

Summarize the results of your project. Was your hypothesis confirmed? Did you achieve your original aims? Be concrete about your accomplishments as well as the limitations of your work.

In conclusion, through manipulating the data we were able to create many generalizations for the data that allowed us to create and understand the average user. In doing so, some of our goals was to analyze the users data and create a user for ourselves and find them a perfect match from the data set. To achieve this, we began analyzing our data through creating word clouds, correlation plots and pie charts in order to understand how a user might advertise themselves to the public through, for example, their use of language and words in their essays or how much of a social drink they might be. In addition, looking at the correlation of some of the values gave us a better understanding of some of the trends in characteristics of users. Some of our original motivation when it came to use this OKcupid data was that potentially, Okcupid would be able to use how we analyzed their consumers and target advertising to places where their users most populate. We believe that we had been able to achieve our first goal in understanding the consumers and because of this we were able to move towards creating our own average user. We successfully created random essays that followed the words we learned through the word clouds and were able to find a "best match" with other essays. In addition, we were able to use K-means clustering, as well as heatmaps, to find out interesting correlations and information about the types of people on OkCupid. Unfortunately though, because of the small amount of variety in reponses, the R values were very small which made it difficult to find significantly strong correlations and trends. For examples, correlations plots found above in the Analysis section were found to have a positive correlation, but not such a significant one that the results were substantial. The same can be said for the sentiment analysis. While we concluded that, most of the time, women had a higher sentiment score than men. The difference was not substantial enough to have real meaning. Another limitation we faced was while finding the perfect match. While we had over 4,000 users and their data, there wasn't enough to get accurate and credible matches between users while also taking into consideration their sexual orientation and other preferences. On the other hand, over 4,000 users was sometimes too much and sometimes, while computing through the entire deck, the code would run for a long time, which was always

difficult when needing to run the program over and over again. Lastly, when it came to the heatmaps, we found that not all of the attributes worked with the same code. Each time we wanted to plot two attributes against each other we had to change the heatmap codes.

Limitations:

1. Perfect Matching: Not enough data to get really incredible matches between users while also taking into consideration both partner's sexual orientations
2. R values small: difficult to find significantly strong correlations for traits due to small amount of variety in responses (correlation plots)
3. Not all of the attributes worked with the same heatmap code, and it had to be changed around each time we wanted to plot two against each other
4.

Author Contributions

Rose implemented three parts of the project.

The first part that she implemented was the geographical mapping of users by different attributes. In order to do this, she had to use US Census Data to match people's locations with latitude and longitude coordinates; this also required her to overcome the challenge that different users inputted their location in inconsistent (sometimes ad hoc) ways. Then, in order to solve the problem that many different people were associated to the same precise coordinates of each other, she wrote a program to add random noise to each user's coordinates so that the users in a given town/city would be represented by a circle of dots whose area is proportional to the number of users in that town/city.

The second part that Rose implemented was the auto-generation of new personal essays using n-grams. Before she could feed the essays into the n-grams algorithm, she had to use regular expressions to sanitize the essays and eliminate all HTML markups.

The third part of the project that Rose implemented was the match-making algorithm. This required several creative ideas: (1) Rather than comparing essays based on standard cosine similarity, she re-weighted each vocabulary word by the frequency by which users used it, that way, words that are used more frequently would have less effect on how similar two users are perceived. (2) In order to compare users simultaneously based on essay similarity and on similarity in non-essay attributes, she assigned hand-tuned weights to each attribute in order so that the euclidean distance between the vectors of attributes would be as meaningful as possible. (3) She created a program interface in order to allow for users of the program to select which attributes they would like to use in the match-making process.

Rose also made the following additional contributions. She came up with the project idea and found the main data set to use. She created the github repository and wrote starter code for everyone to use in order to examine the data. She organized group meetings over zoom. She created an outline for the powerpoint presentation. Finally, she wrote up various sections of the proposal, the powerpoint, and the final report.

Shreya and Julian worked together primarily for coding. They found average traits of the users in the dataset, separated by sex, which corresponded with the first goal of understanding users in the set. They worked on assigning numerical values to traits that were described with words, especially for traits on a spectrum. These values could then be used in data processing that requires numbers as an input. They also helped implement the K-means clustering code in order to find meaningful and distinct groups amongst the dataset, which contributes to the third goal. They created correlation plots for various traits in the dataset to find any interesting and strong correlations, which fulfills the first and fourth goal.

Shreya also helped coordinate meeting times on zoom for group members to work together and came up with ideas and goals for the final project in the early stages. She came up with the idea of creating heatmaps for a more meaningful visualization. She came up with the idea of creating an average user with all average traits in preliminary stages. She was in charge of writing and presenting the introduction and goals sections for the proposal, presentation, and the final report. She also contributed to the methods sections on the proposal and wrote about how exactly the group was planning on accomplishing the goals.

Julian wrote the significance section of the project proposal, drawing inspiration from the business side of his major. He created the theme of our team being OkCupid's clientele, as they seek. Julian was in charge of the average users section of the presentation, constructing the breakdown of the average person, man and woman in the dataset. He also created the pie charts and future plans sections, highlighting additional results from our analysis and emphasizing where we project our project to go. Julian spearheaded the k-means clustering section of our analysis, implementing and adjusting earlier clustering work to fit our analytical goals.

After Julian found key issues with the K-means clustering analysis, he created methods to convert our data to be used for a heatmap. This made it easier to find connections between attributes, and make better potential recommendations for advertising locations for OkCupid.

Lastly, Manon focused on data analysis where she created word clouds to understand the common language used between users as well as sentiment analysis on the essays to determine if there was a difference in the way women and men wrote their essays. During presentations, Manon talked about the motivation behind the project for potential future use in the real world as well as introduced the data set to the audience and introduced her word clouds before passing the baton to Rose to continue talking about visualisations and results. Lastly, with the help of Rose

and Shreya, Manon wrote the methods section of the proposal. While everyone was assigned a section to write in the proposal, we all proof edited each other's work before handing in the finished product.

References