Preliminary Results

Li Liu, Abhishek Pandit, Adam Shelton

11/5/2019

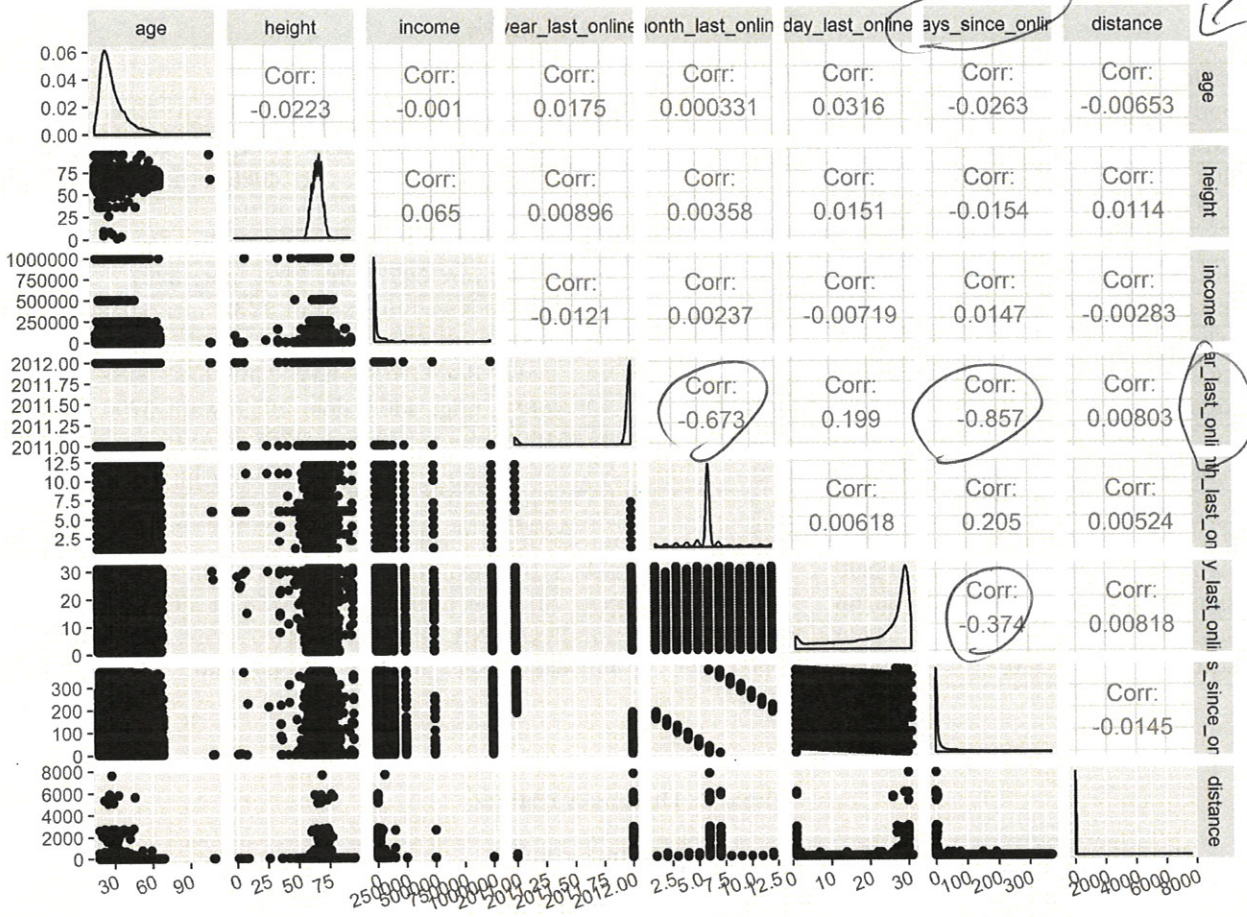
Exploratory Data Analysis

Descriptive Statistics

Overall, this dataset encompasses 54 variables for 59946 observations. Each observation is a dating profile for a person, gathered from subset of OkCupid users active on 2012/06/30, who had been active within the lst year and had at least one profile photo. Users were located within a 25 mile radius of San Francisco when the data was scraped. 59946 were "visiting" and listed a location outside of a 25 mile radius from San Francisco on their profile. Approximately 60% of the profiles identfy as male, 55% as white, 86% as straight, and 51% as multilingual. The median age specified on profiles was 30.

Visualizations

Continuous Variables



what are these telling us? what are the inputs?

Table 1: Text Variables

variable	missing	complete	min	max	empty	n_unique
education	6628	53318	10	33	0	32
essay0	5485	54461	1	48854	0	54351
essay1	7571	52375	1	7955	0	51517
essay2	9638	50308	1	6129	0	48635
essay3	11476	48470	1	4374	0	43533
essay4	10537	49409	1	44469	0	49260
essay5	10847	49099	1	30446	0	48964
essay6	13771	46175	1	11385	0	43603
essay7	12450	47496	1	3722	0	45555
essay8	19214	40732	1	13304	0	39325
essay9	12602	47344	1	11444	0	45444
job	8198	51748	5	33	0	21
location	0	59946	12	35	0	199
speaks	50	59896	7	107	0	7647
time_since_online	0	59946	2	15	0	30123

Table 2: Categorical Variables

variable	missing	complete	n_unique	top_counts	ordered
body_type	5296	54650	12	ave: 14652, fit: 12711, ath: 11819, NA: 5296	FALSE
cats	31323	28623	3	NA: 31323, lik: 18450, has: 7274, dis: 2899	FALSE
diet	24395	35551	6	any: 27881, NA: 24395, veg: 4986, oth: 1790	FALSE
diet_import	31734	28212	2	NA: 31734, mos: 21508, str: 6704	FALSE
dogs	22512	37434	3	lik: 28380, NA: 22512, has: 8493, dis: 561	FALSE
drinks	2985	56961	6	soc: 41780, rar: 5957, oft: 5164, not: 3267	FALSE
drugs	14080	45866	3	nev: 37724, NA: 14080, som: 7732, oft: 410	FALSE
ethnicity	5680	54266	11	whi: 32831, asi: 6134, NA: 5680, mul: 5051	FALSE
kids	38895	21051	3	NA: 38895, doe: 16132, has: 2461, has: 2458	FALSE
kids_import	46885	13061	5	NA: 46885, mig: 4403, doe: 4059, wan: 3790	FALSE
orientation	0	59946	3	str: 51606, gay: 5573, bis: 2767, NA: 0	FALSE
religion	20226	39720	9	NA: 20226, agn: 8812, oth: 7743, ath: 6985	FALSE
religion_import	32007	27939	4	NA: 32007, not: 12212, lau: 8995, som: 4516	FALSE
sex	0	59946	2	m: 35829, f: 24117, NA: 0	FALSE
sign	11056	48890	12	NA: 11056, leo: 4374, gem: 4310, lib: 4207	FALSE
sign_import	23180	36766	3	NA: 23180, it': 19333, it : 16758, it : 675	FALSE
smokes	5512	54434	5	no: 43896, NA: 5512, som: 3787, whe: 3040	FALSE
status	0	59946	5	sin: 55697, see: 2064, ava: 1865, mar: 310	FALSE

Table 3: Dummy Variables

variable	missing	complete	mean	count
multi_ling	50	59896	0.51	TRU: 30824, FAL: 29072, NA: 50
speaks_en	50	59896	1	TRU: 59896, NA: 50
visiting	0	59946	0.043	FAL: 57390, TRU: 2556, NA: 0

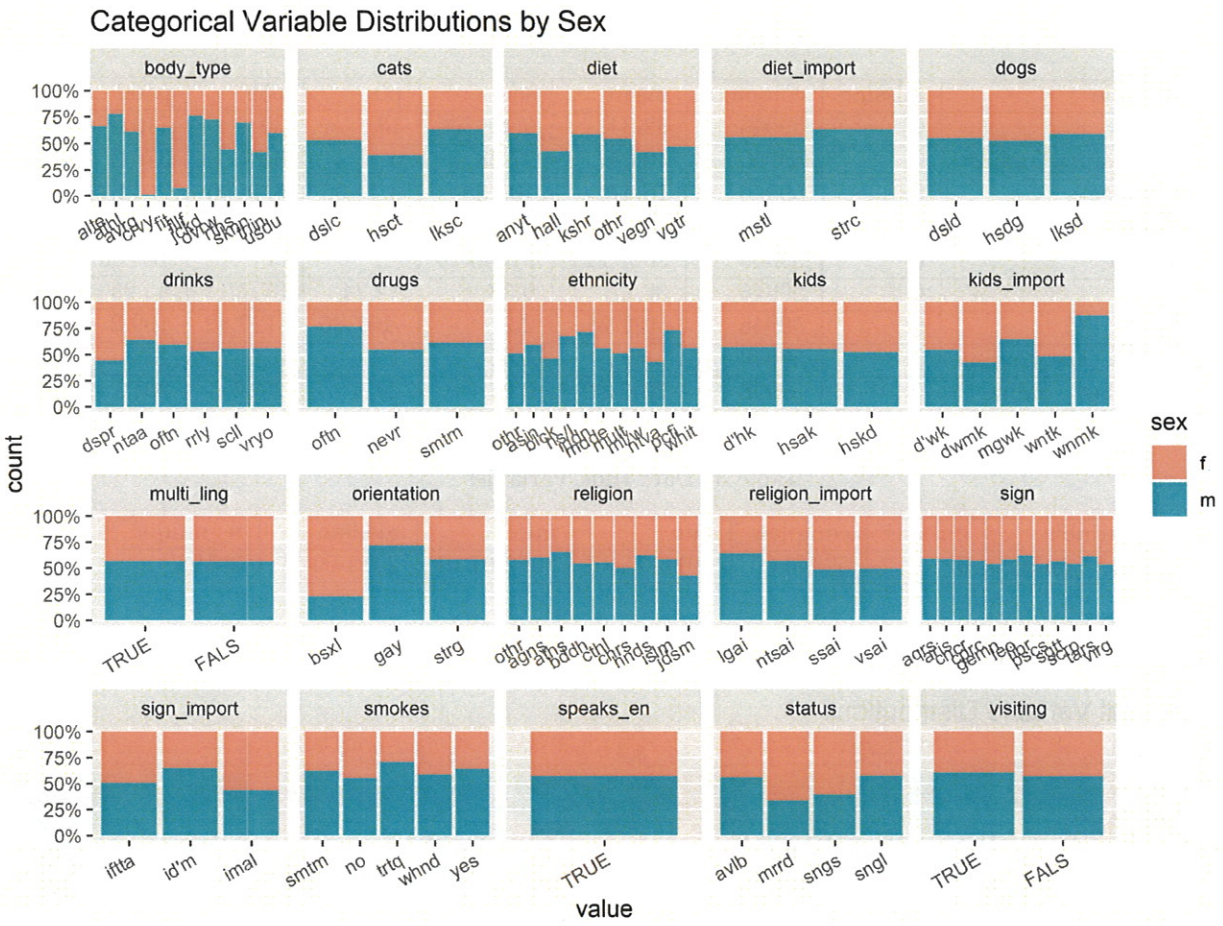
Table 4: Continuous Variables

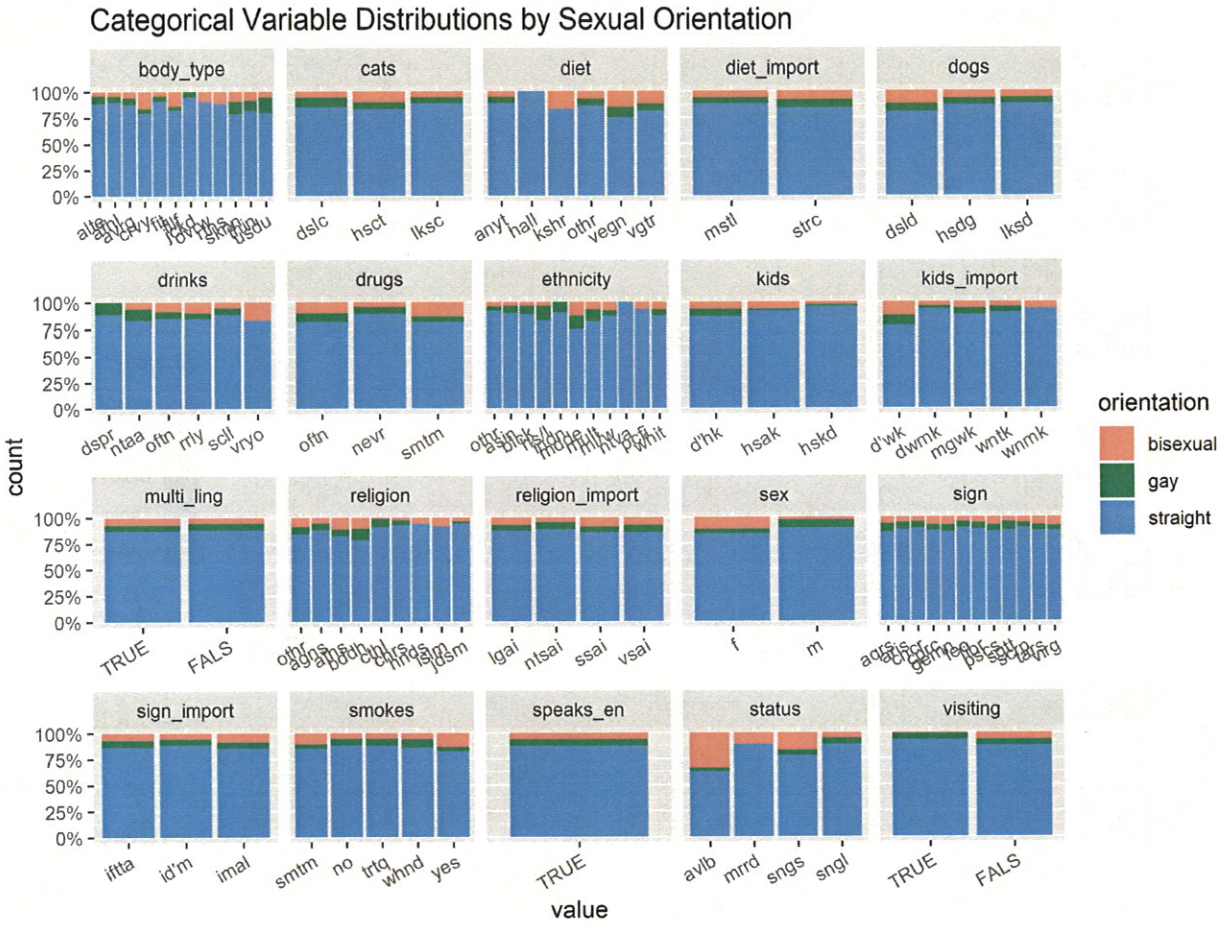
variable	missing	complete	mean	sd	p0	p50	p100
age	0	59946	32.34	9.45	18	30	110
days_since_online	0	59946	40.09	77.28	0	3.77	370.3
distance	0	59946	10.47	104.38	3.4e-05	3.4e-05	7642.77
height	3	59943	68.3	3.99	1	68	95
income	0	59946	20033.22	97346.19	-1	-1	1e+06
lat	0	59946	37.77	0.33	12.24	37.77	55.95
lon	0	59946	-122.28	2.2	-157.86	-122.42	109.2
month_last_online	0	59946	5.89	1.65	1	6	12
year_last_online	0	59946	2011.92	0.27	2011	2012	2012

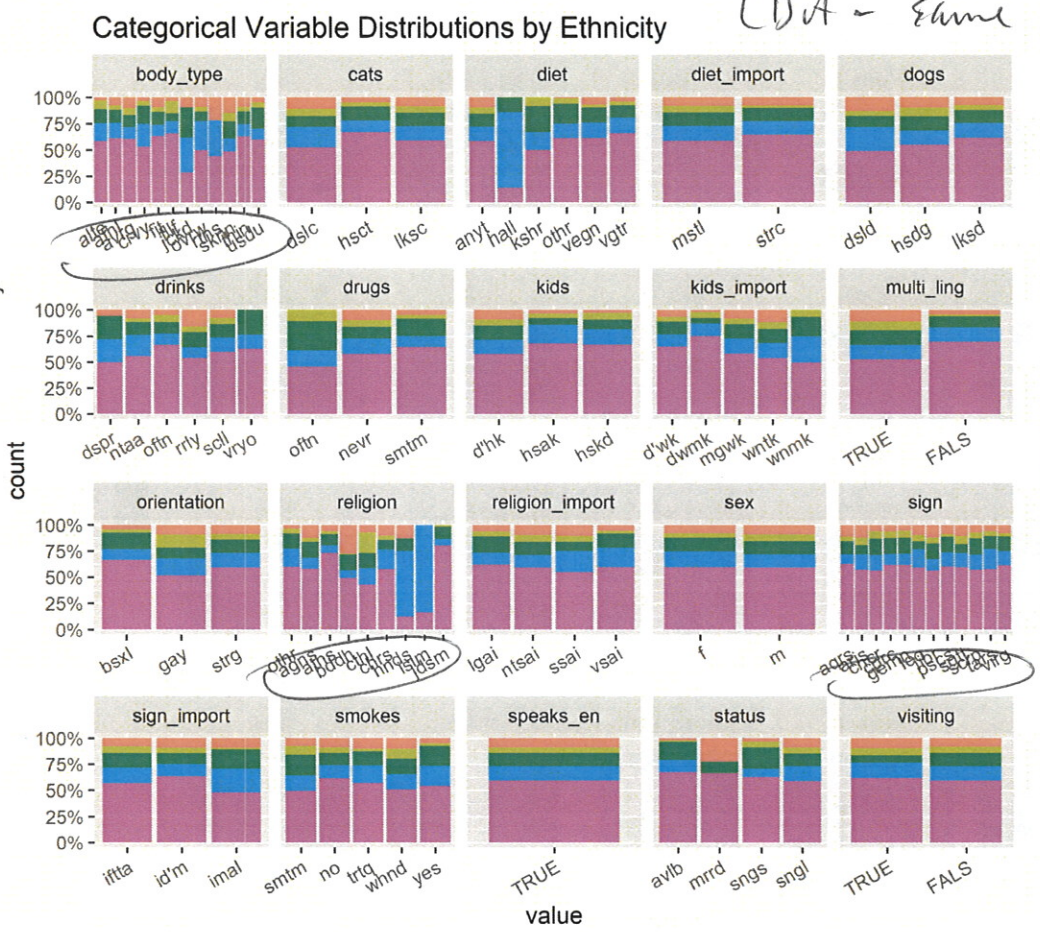
Table 5: Date-time Variables

variable	missing	complete	min	max	median	n_unique
last_online	0	59946	2011-06-27	2012-07-01	2012-06-27	30123

Categorical Variables


Categorical Variable Distributions by Sex
body_type
cats
diet
diet_import
dogs
drinks
drugs
ethnicity
kids
kids_import
multi_ling
orientation
religion
religion_import
sign
sign_import
smokes
speaks_en
status
visiting
100%
75%
50%
25%
0%
count
value
sex
f
m
dslc
hsct
lksc
anyt
hall
kshr
othr
vegn
vgtr
mstl
strc
dsld
hsdg
lksd
dspr
ntaa
oftn
rrly
scll
vryo
oftn
nevr
smtm
d'hk
hsak
hskd
d'wk
dwmk
mgwk
wntk
wmmk
TRUE
FALS
bsxl
gay
strg
lgai
ntsai
ssai
vsai
ifta
id'm
imal
smtm
no
trtg
whnd
yes
avlb
mrrd
sngs
sngl


Categorical Variable Distributions by Sexual Orientation
body_type
cats
diet
diet_import
dogs
drinks
drugs
ethnicity
kids
kids_import
multi_ling
religion
religion_import
sex
sign
sign_import
smokes
speaks_en
status
visiting
100%
75%
50%
25%
0%
count
value
orientation
bisexual
gay
straight
dslc
hsct
lksc
anyt
hall
kshr
othr
vegn
vgtr
mstl
strc
dsld
hsdg
lksd
dspr
ntaa
oftn
rrly
scll
vryo
oftn
nevr
smtm
d'hk
hsak
hskd
d'wk
dwmk
mgwk
wmtk
wmmk
TRUE
FALS
lgai
ntsai
ssai
vsai
f
m
ifta
id'm
imal
smtm
no
trtq
whnd
yes
TRUE
avlb
mrrd
sngs
sngl
TRUE
FALS



Body types of "curvy" and "full figured" are used almost exclusively by women. Women were also overwhelmingly more likely to identify as bisexual or married than men. The majority of people who identified as gay or wanting more kids were men. Those who identify their body type as "jacked", those that dislike dogs or cats, or those who want more kids are less likely to be white. Those who identified as bisexual are more likely to be white than other sexual orientations.

Text Clustering

Topic Modeling

We used the Non-Matrix Factorization method in Topic Analysis to build a topic model. We checked for differences in topics used by different models for different demographics- namely. The differences were not particularly notable.

For race, we again see similar patterns. There are a few interesting patterns- such as the high use of words to do with 'cool' by African American men, and a relative lack thereof in Asian and White men. Black men also trail behind in their references to relocation and the region, and yet use words from the topics on being new. This may imply that they wish to emphasize being new and on the lookout for friends without mentioning where they are from, which might carry the burden of judgment. For height, there are almost no visible differences between the average representation of each topic

For education, we observe some differences in a few topics, such as 'cool' and 'travel' holding higher proportions for those with less and more than high school diplomas respectively. In other areas such as 'Reach Out' and 'New Here', we observe only a few percentage points' difference

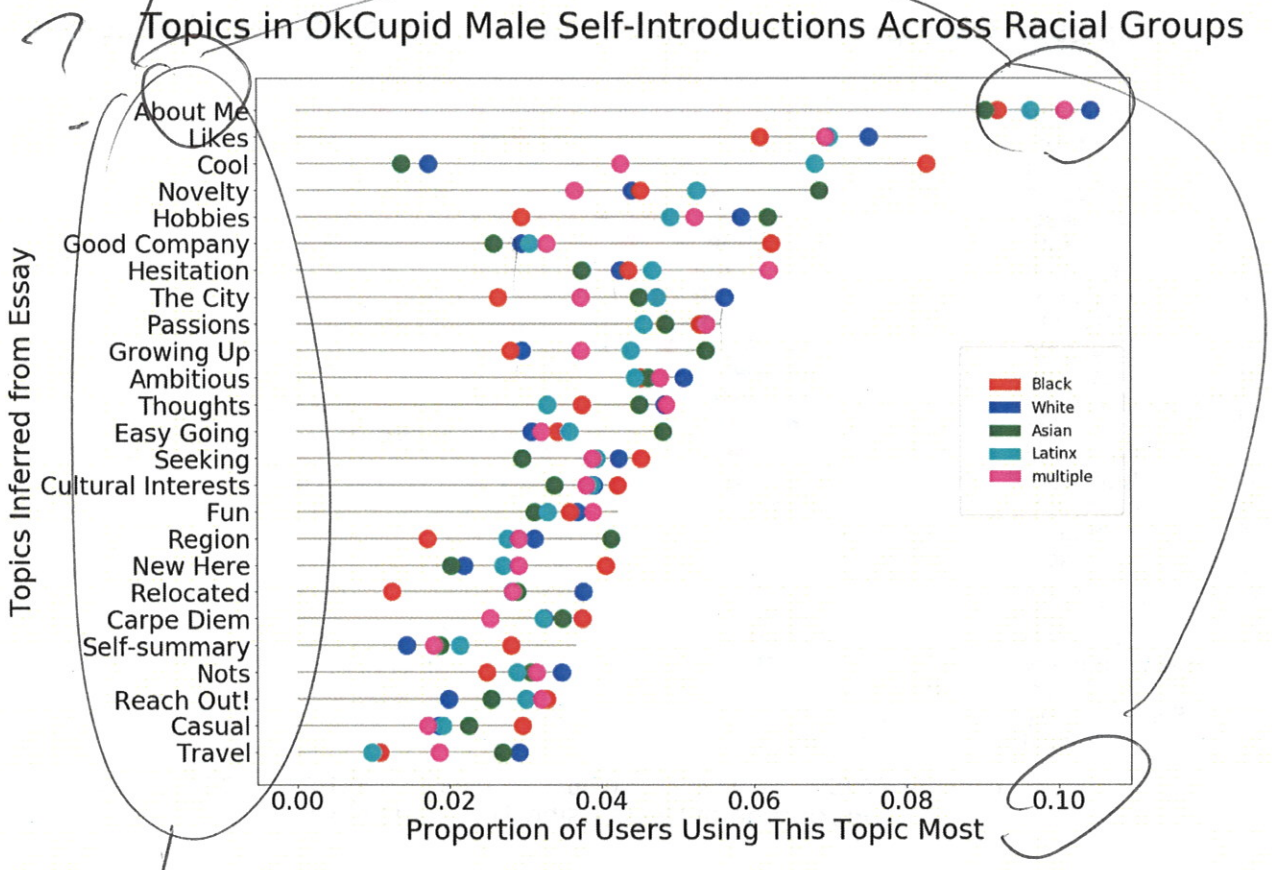


Figure 1: Topics by Race

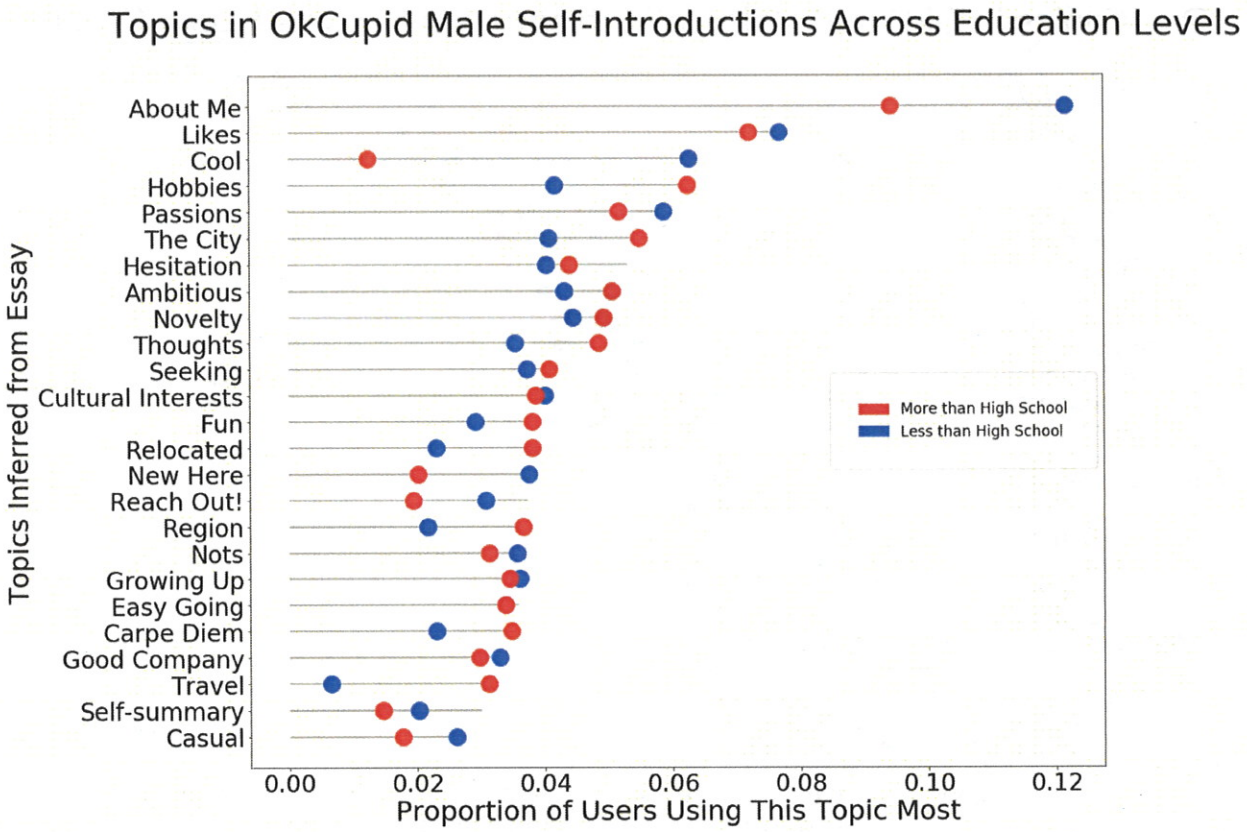


Figure 2: Topics by Education

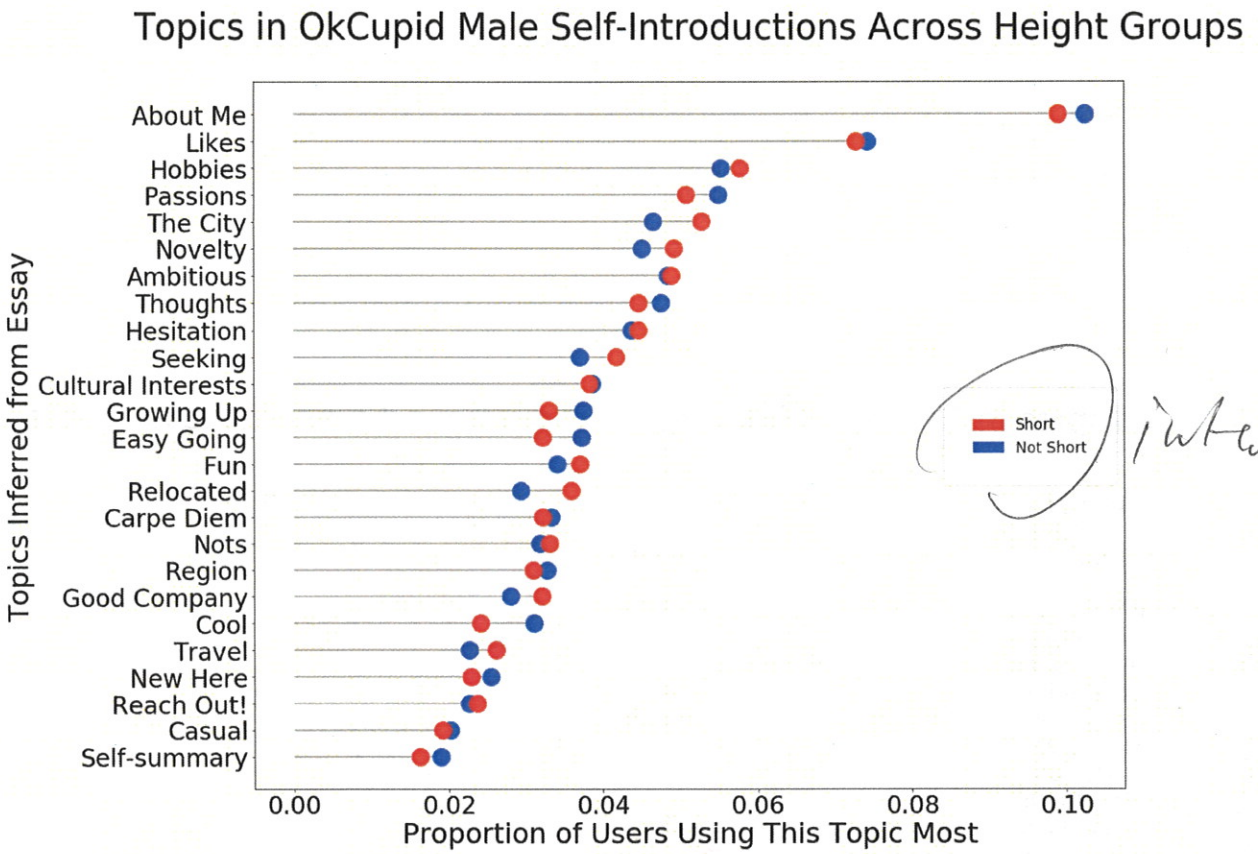


Figure 3: Topics by Height

For height, we so no noteworthy differences beyond a few percentage points for 'likes' and 'about me'.

For fitness, similarly, no clear differences emerge.

Overall, there is close to little differentiation on any of these four variables of interest.

Future Extensions- We can check for similar difference on gender, religion, etc

Vector Space Model Analysis of Profiles

We used SpaCy and other libraries to clean up the Self-Introduction essay in the profiles. We used the Doc2Vec technique- where each profile is assigned a single score based on the vector space model- averaging out across all the words in the model. We then collapse the vector space down to 2 dimensions using TSNE. We use elbow plots to check for clustering. Based on the plot, we see largely one elliptical 'blob'. There does not seem to be a clear discernible pattern

#![TSNE Blob](TSNE_blob.png)

In the future, we could consider more agglomerative clustering methods for the profile scores.

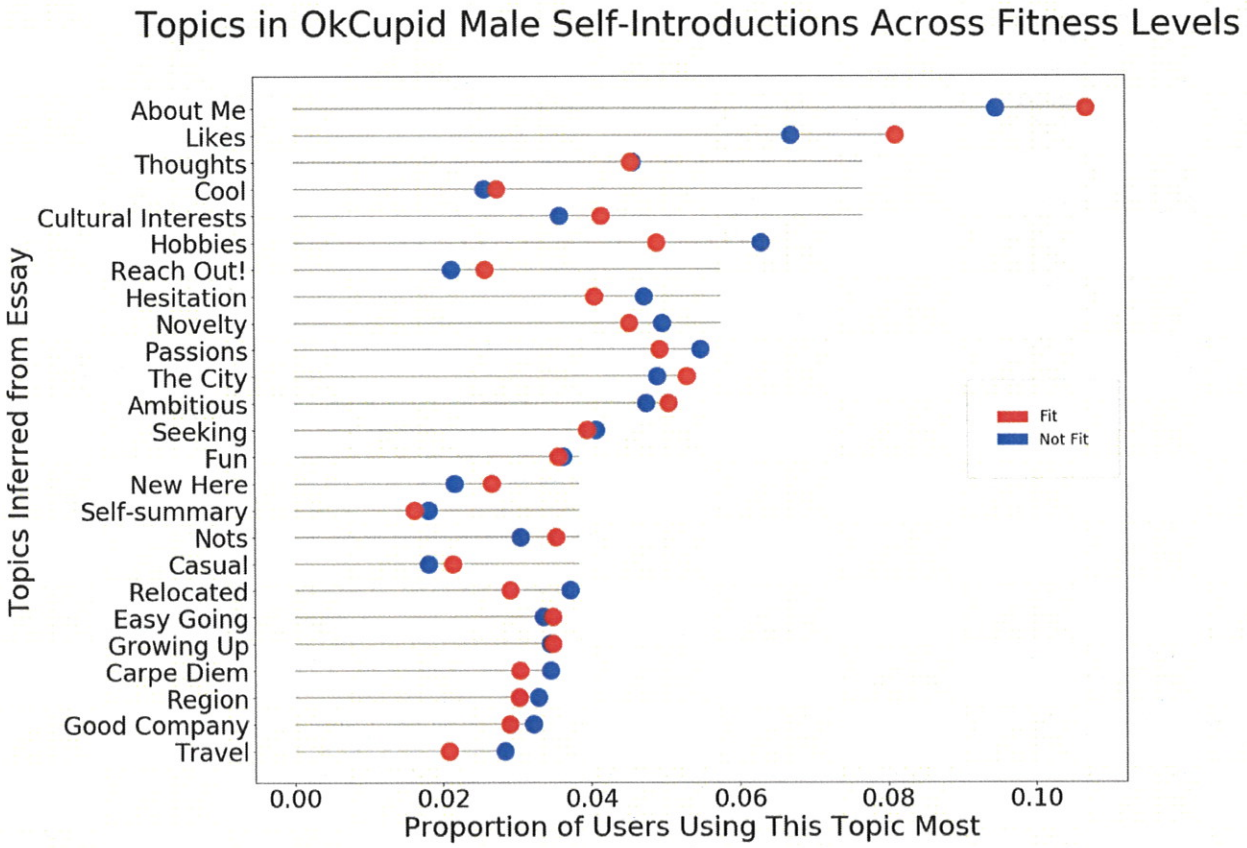


Figure 4: Topics by Physical Fitness

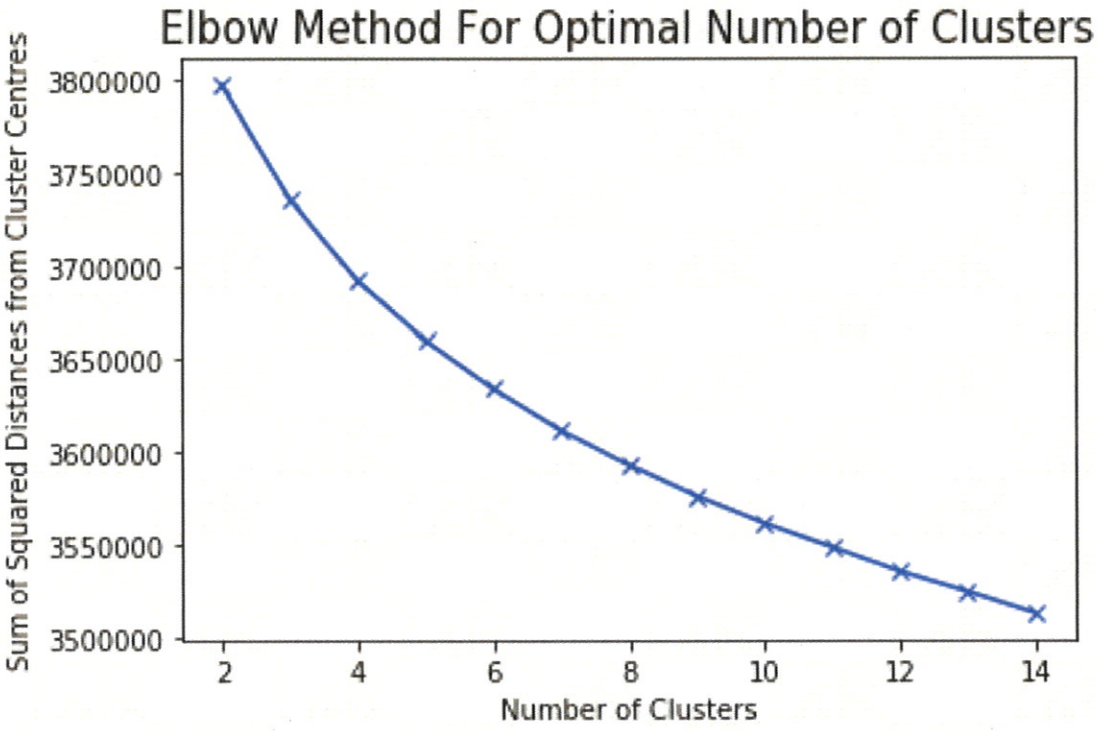


Figure 5: Elbow Plot

Demographic Clustering

The output quality from clustering depends heavily on the data we use. In the raw data, many variables are categorical. So we have to do some feature engineering with intuition from our experience using OkCupid.

1. Body type

The users could choose their body types as `average`, `fit` , `athletic`, `thin`, `curvy`, `skinny`, `jacked`, etc. Since `average` and `fit` users are the most common (46% of the sample), we labeled them as 1 and the rest as 0.

2. diet

We assume that people take diet preferences into consideration when choosing a future mate. Especially for people who are vegetarian, they might get along easier with partners who are vegetarians too. SO we labeled vegetarian users as 1 and the rest as 0. Surprisingly, this subset is almost half of the sample.

3. status

Generally, we would assume all people using online dating are single. However, there are 310 users who are married and 10 unknown. We classified all the `single` users as 1 and the rest as 0. 93% of them are single.

4. drinks

The drinks altitude is also important for choosing a partner as it might predict people's preferences for social events. We labeled `socially` as 1 the rest as 0. Although it is ambiguous of what `socially` means, we assume this is what most people would do (or pretend them to be) as 70% of them fall into this category.

5. drugs

It's highly impossible that people who don't do drugs would date someone likes drugs. So we labeled 1 for `never` drug users and 0 for the rest. 63% of them don't do drugs.

6. education

We define two groups of education level: the ones with/working on advanced degrees, and the ones who don't. Advanced degrees are Ph.D., masters, law school of medical school. 18% of the users are educated beyond the undergraduate level.

7. ethnicity

This is an important dimension when choosing the date. Many previous pieces of research also show that certain male or female races group are most/least popular on the online dating website. For simplicity, we just classified the `white` as 1 and the rest as 0. White users have the highest ratio of 55%, followed by Asian, Hispanic/Latin.

8. height

Height might not be a crucial factor. Here we labeled people as `normalheight` if their heights are in 25~75% percentile of the distribution, which means they are between 66 to 71 inches. Roughly half of the users have normal heights.

9. income

Not surprisingly, 81% of the people didn't specify their income level. So we labeled the 19% users who do as 1.

10. job

There are many job categories in the data. We are interested in two groups: the first one is students, as they certainly have a different dating preference than the other working professionals; the second one is people working in the tech and engineering industry. The two groups account for 8% and 16% of the total users, respectively.

11. Offspring

We labeled people who don't have kids as 1, which is 86% of the sample.

12. Orientation

We labeled people who are straight as 1, which is 86% of the sample.

13. Pets

We labeled people who have pets as 1, which is 57% of the sample.

14. Religion

We labeled people who not religious at all as 1, which is 34% of the sample.

15. Sex

We labeled male users as 1, which is 60% of the sample.

16. Smokes

We labeled people who don't smoke as 1, which is 73% of the sample.

17. Speaks

We calculate the number of languages users indicate they know as a new feature.

18. Distance

We use Google Map API to calculate the physical distance between users' city to San Francisco.

19. Days since Online

This feature is calculated from the `last_online` variable in the raw data.

Examine new features

The cleaned data has 59955 raws and 22 columns (features).

The heatmap shows there are no features pairs that are strongly correlated.

not surprised b/c this strategy washed out tons of variance. why not correlate across raw values instead?

Kmeans

We have started clustering with Kmeans. However, we haven't seen an optimal K by using the Elbow method and Silhouette Score. We need to further work on reducing the dimensions.

The following scatter plot illustrates the result by using 3 clusters. The y-axis is days since online. The x-axis is age. The color is their labeled clusters. We could see clusters 1 and 2 are on the left side, which represents mainly people younger than 45. Cluster 3 consists mainly of people who are older than 45.

Principal Components Analysis

We start to use principal components to reduce the dimensions. We still need further work to decide the optimal number of components.

The following plot shows the first two principal components of models with 3 PC. The color represents being male or not.

The following heatmap shows the weights loadings of the 3 PCs.

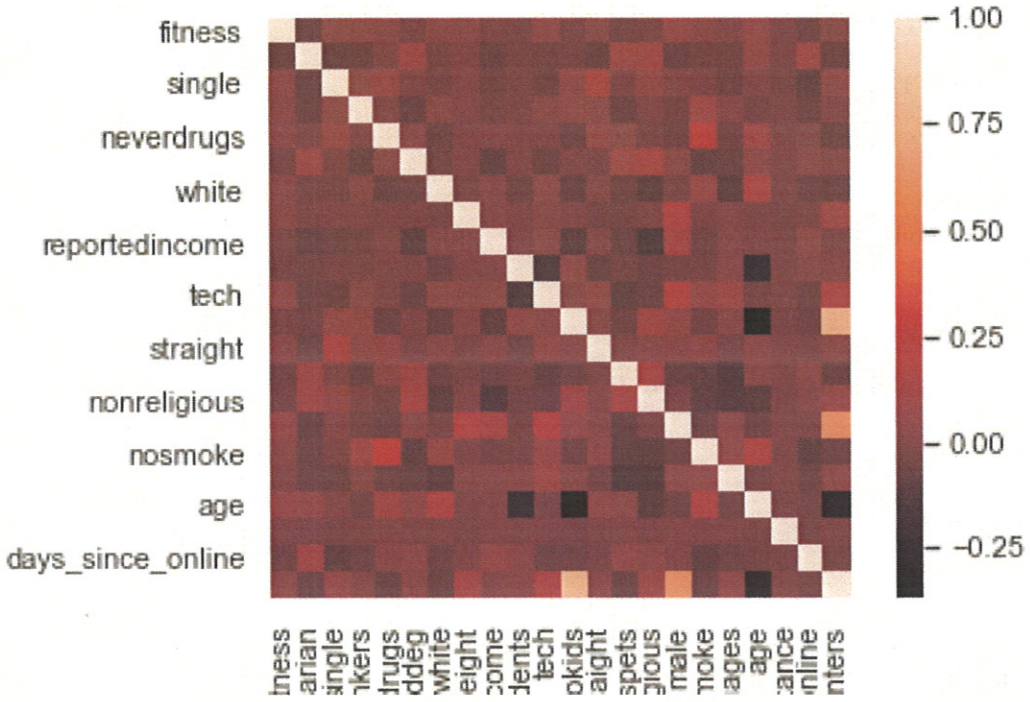
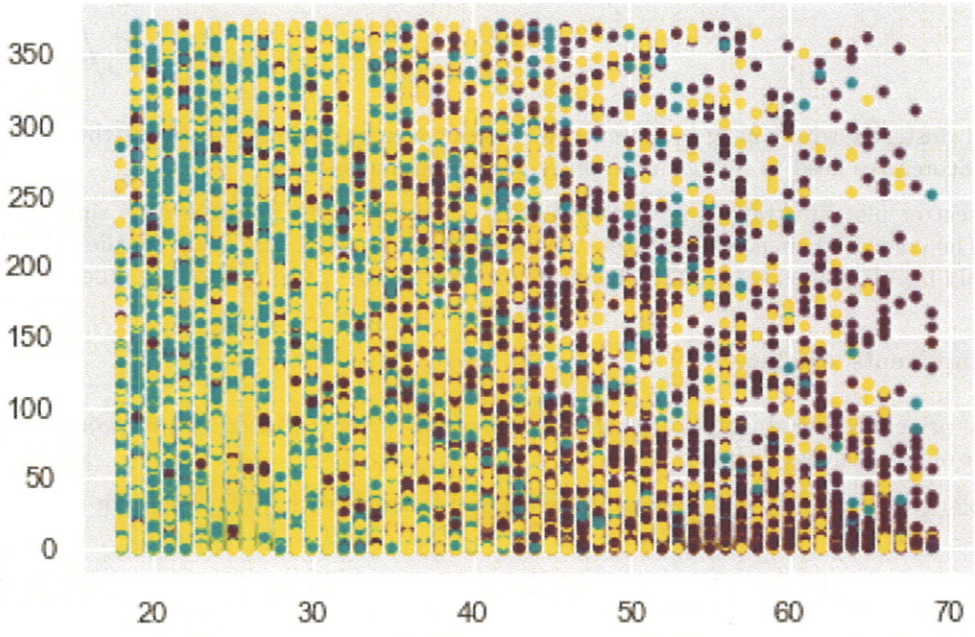


Figure 6: Correlation Matrice Heatmap



Clear as mud...

Figure 7: KMeans Visualization Example

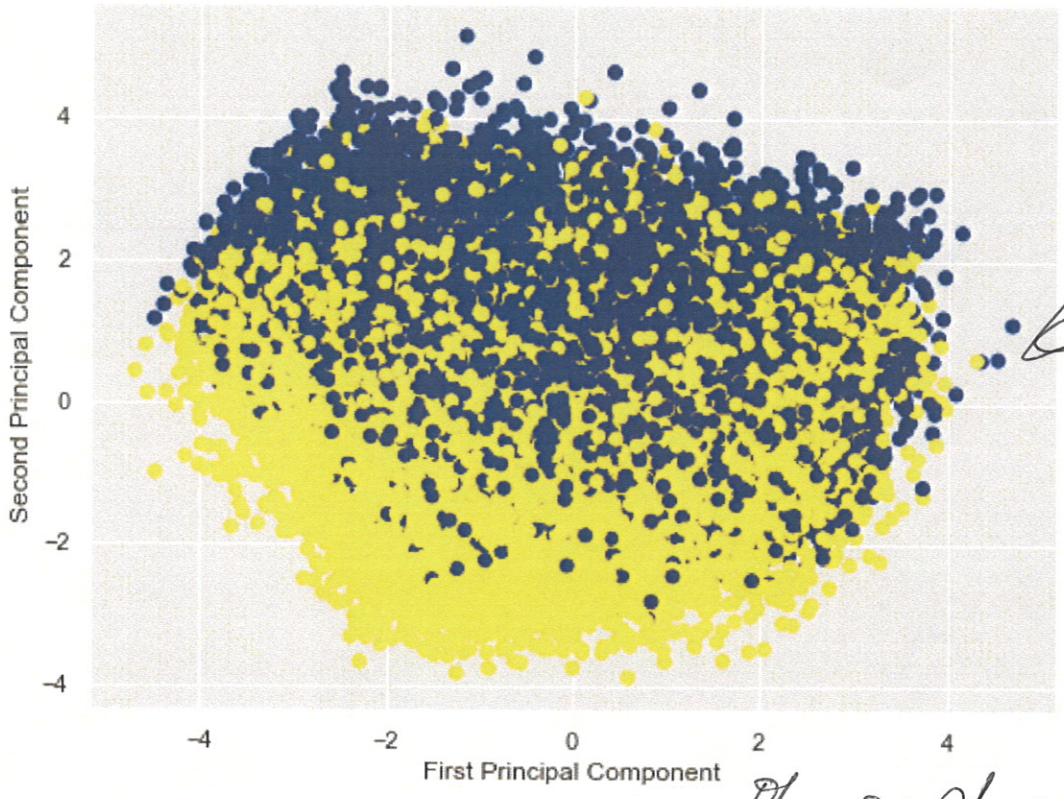


Figure 8: PCA Visualization

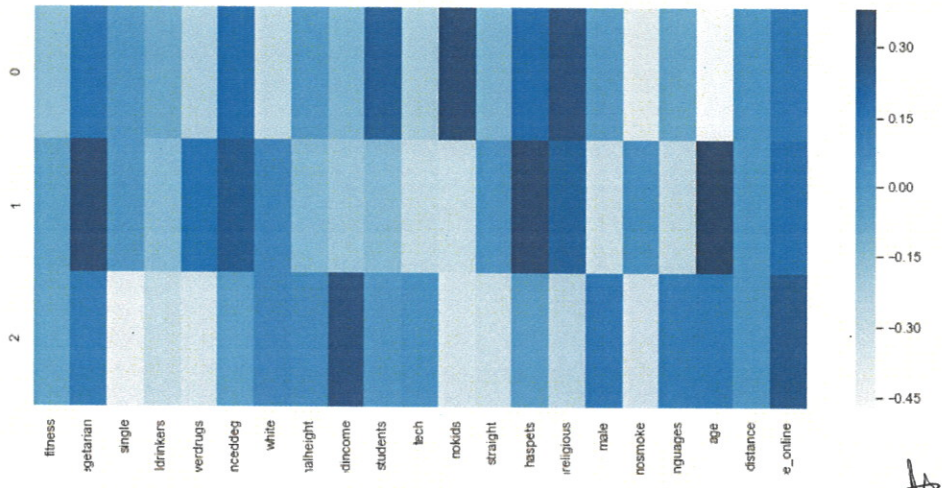


Figure 9: PCA Loadings