

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 1st 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 identify as male, NA% as white, 86% as straight, and NA% as multilingual. The median age specified on profiles was 30.

Visualizations

Continuous Variables

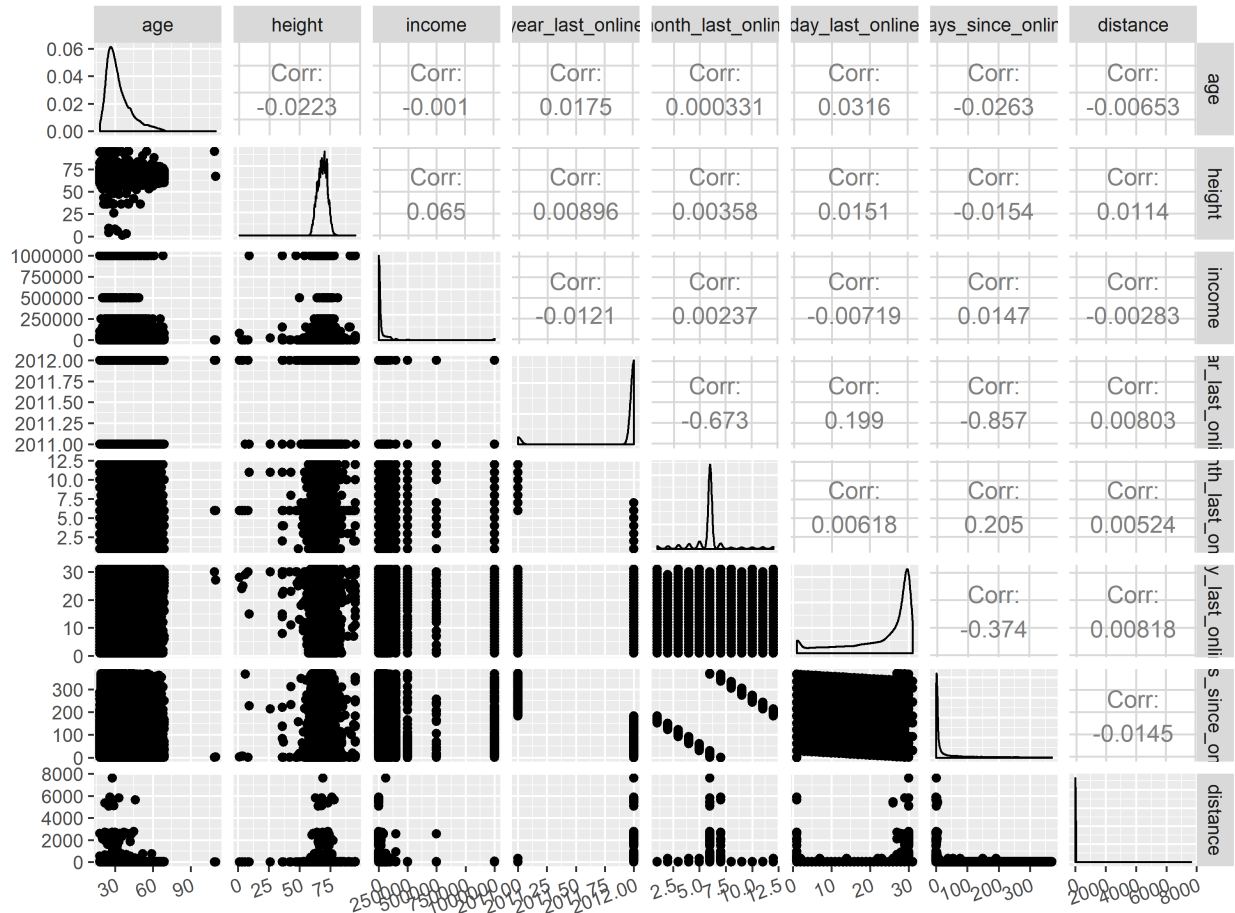


Table 1: Text Variables

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

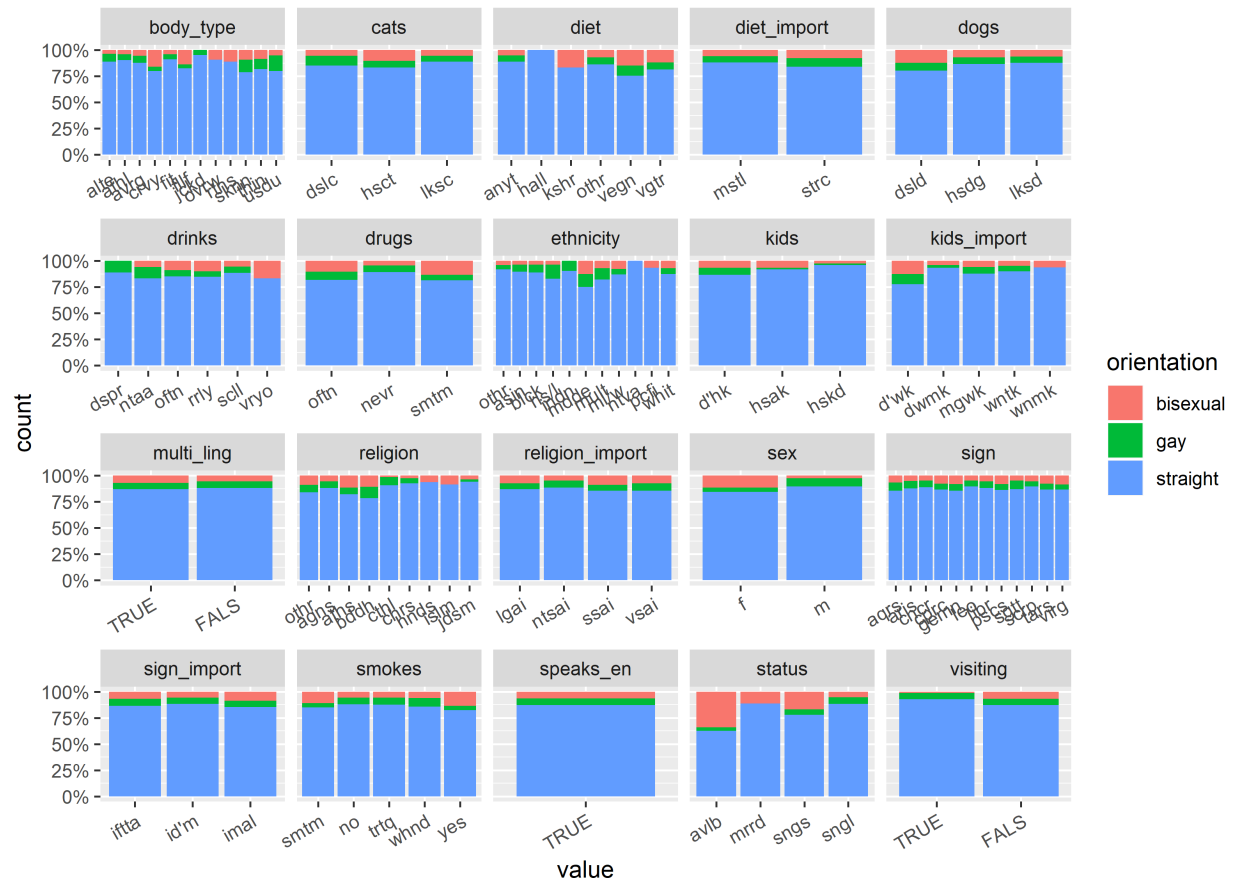
Categorical Variable Distributions



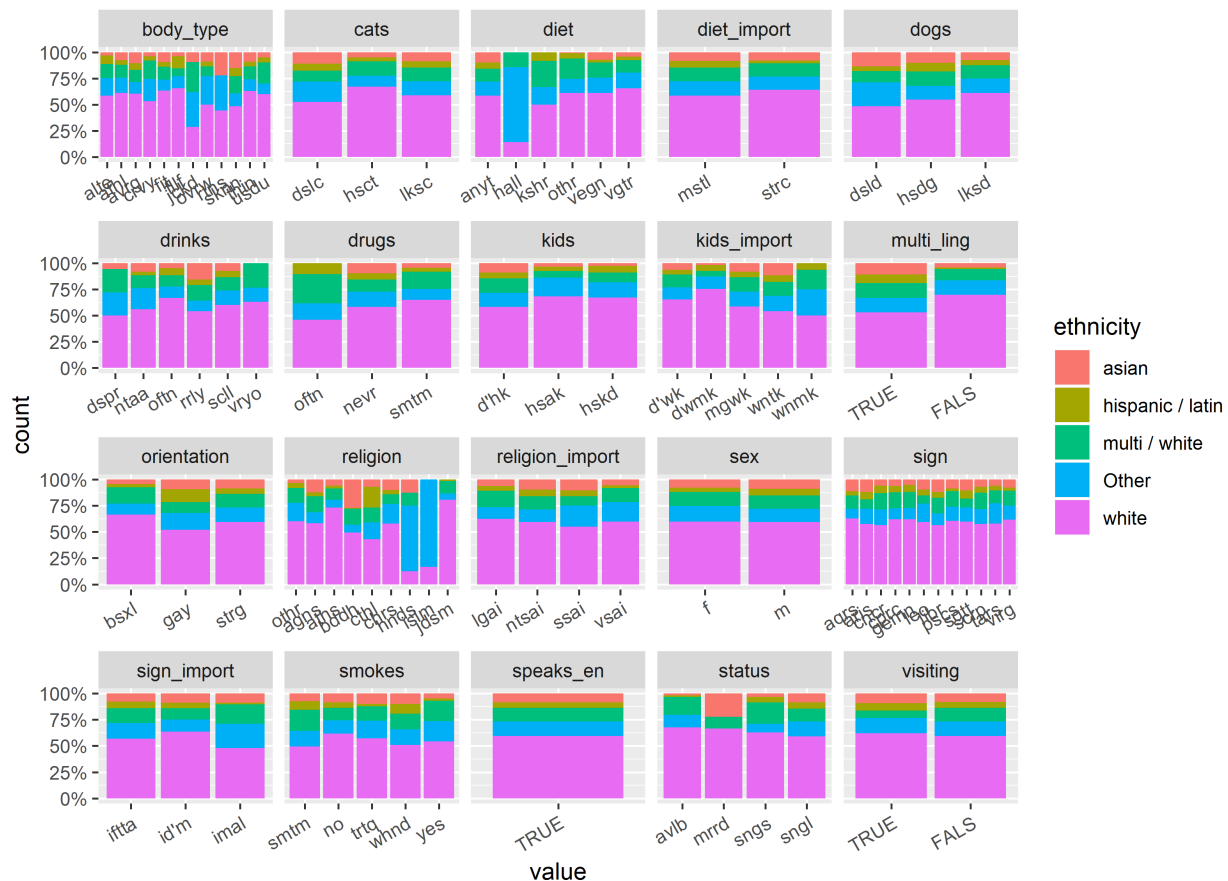
Categorical Variable Distributions by Sex



Categorical Variable Distributions by Sexual Orientation



Categorical Variable Distributions by Ethnicity



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.

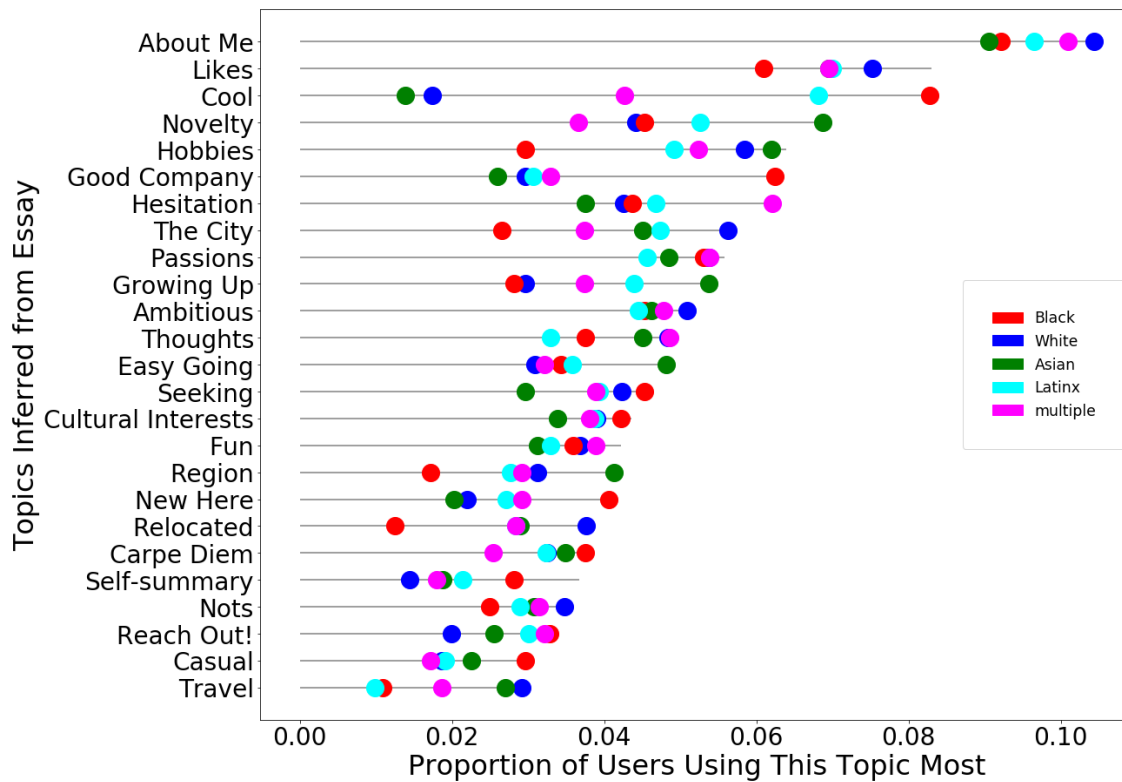
Text Clustering

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.

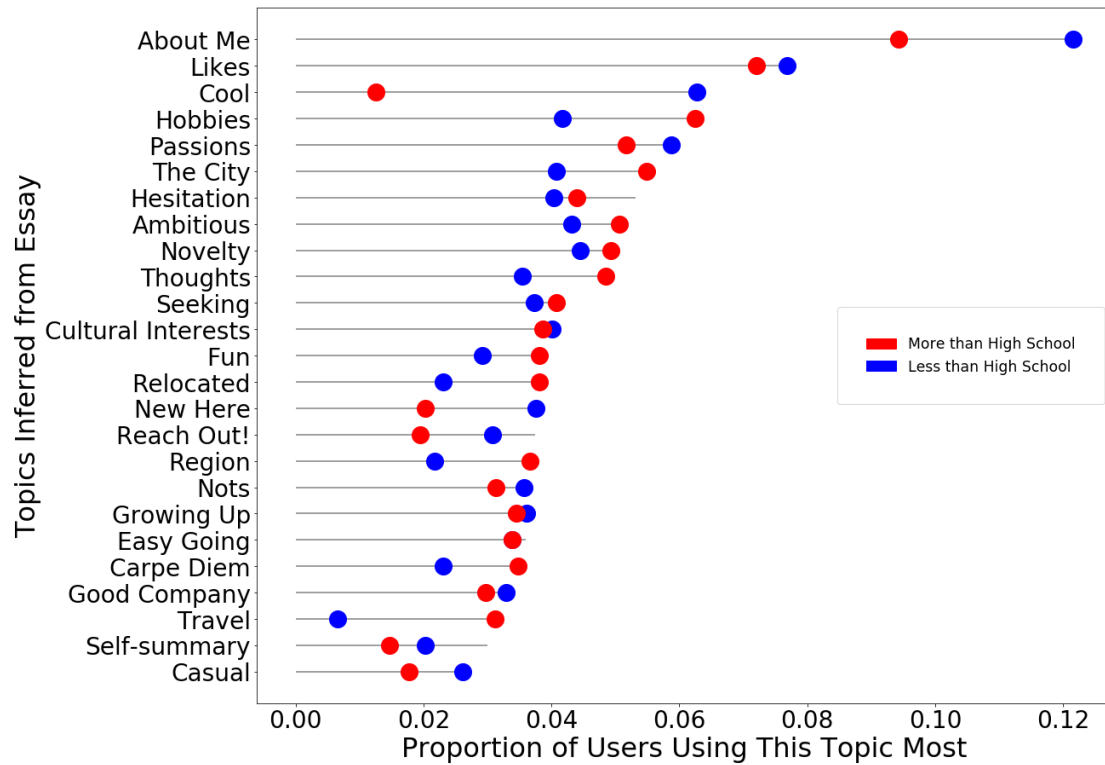
For race, we again see similar patterns. There are a few interesting patterns- such as the high use of words to do with ‘cool’ by African American men, and a relative lack thereof in Asian and White men. Black men also trail behind in their references to relocation and the region, and yet use words from 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

Topics in OkCupid Male Self-Introductions Across Racial Groups



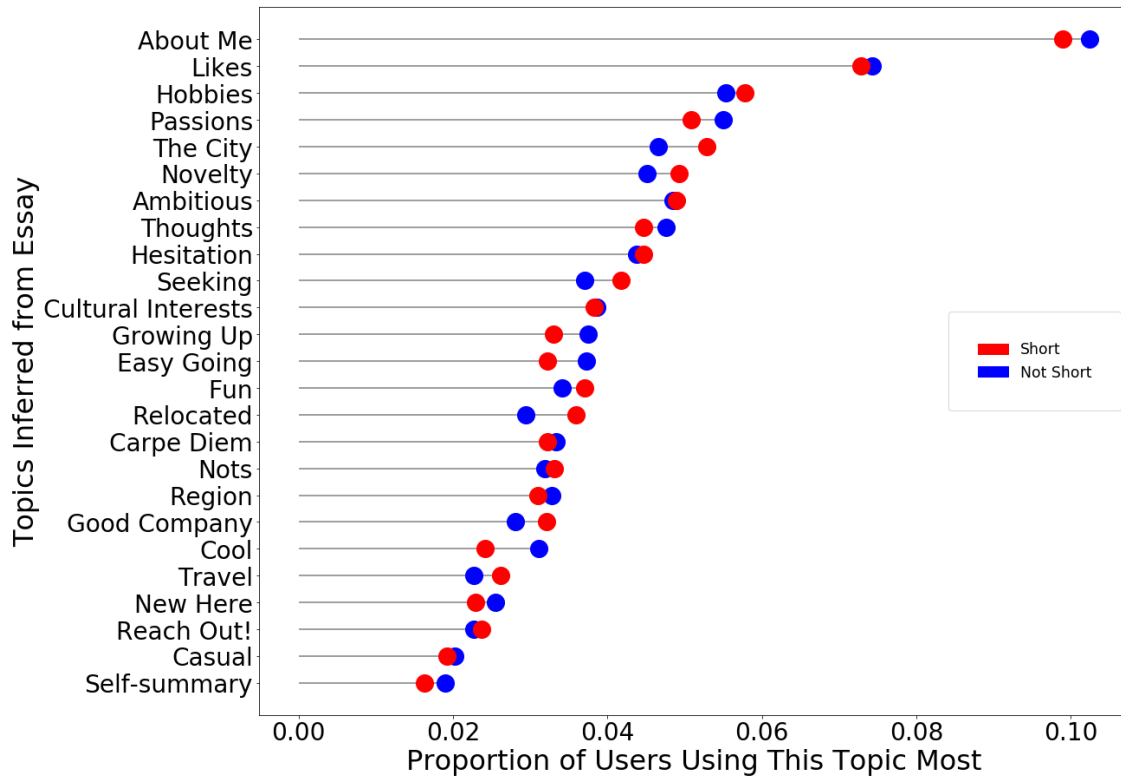
For education, we observe some differences in a few topics, such as ‘cool’ and ‘travel’ holding higher proportions for those with less and more than high school diplomas respectively. In other areas such as ‘Reach Out’ and ‘New Here’, we observe only a few percentage points’ difference

Topics in OkCupid Male Self-Introductions Across Education Levels

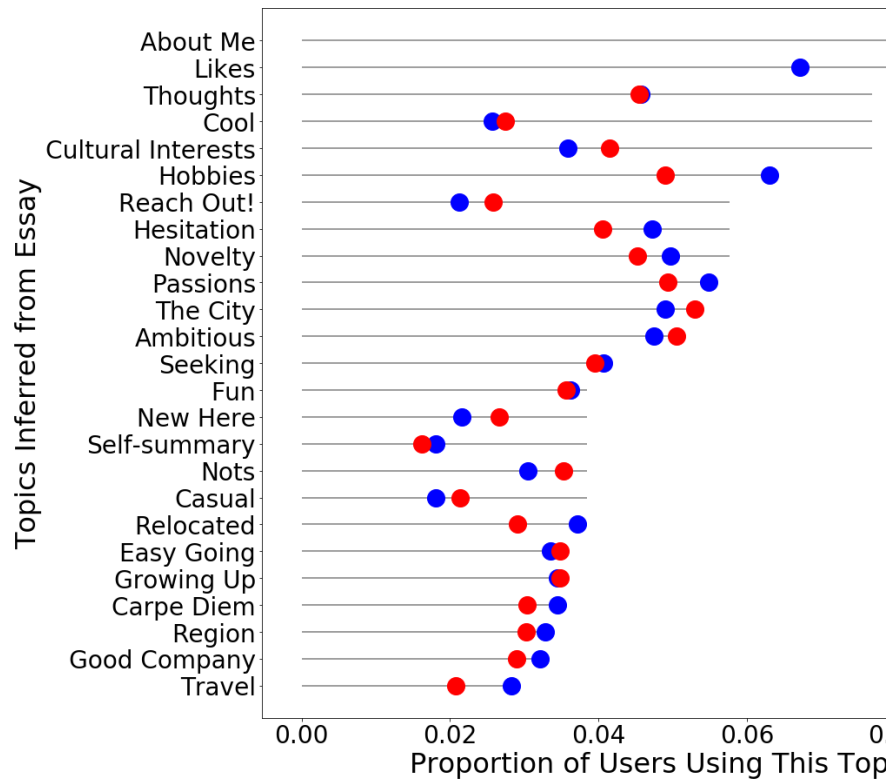


For height, we see no noteworthy differences beyond a few percentage points for 'likes' and 'about me'.

Topics in OkCupid Male Self-Introductions Across Height Groups



Topics in OkCupid Male Self-Introductions Across Fitness Groups



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

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

Unsupervised Clustering

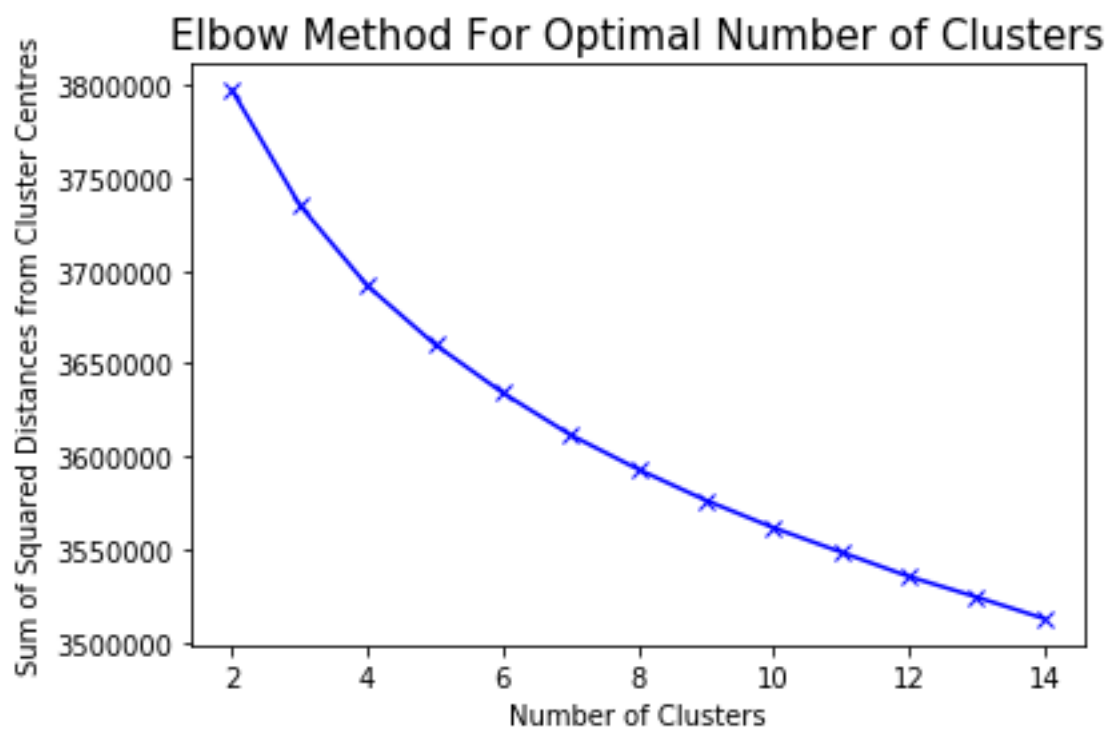


Figure 1: Elbow Plot