Module 5

Use Case 2

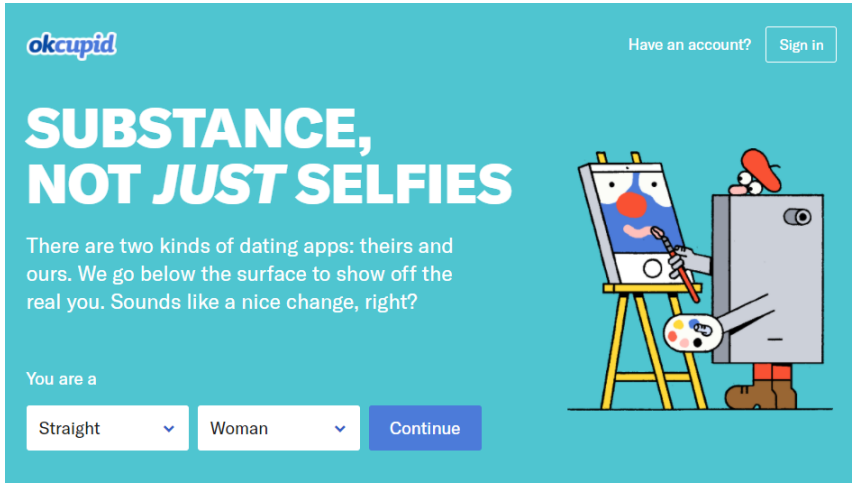
LESSON 2

José Manuel García Nieto – University of Málaga

Table of contents

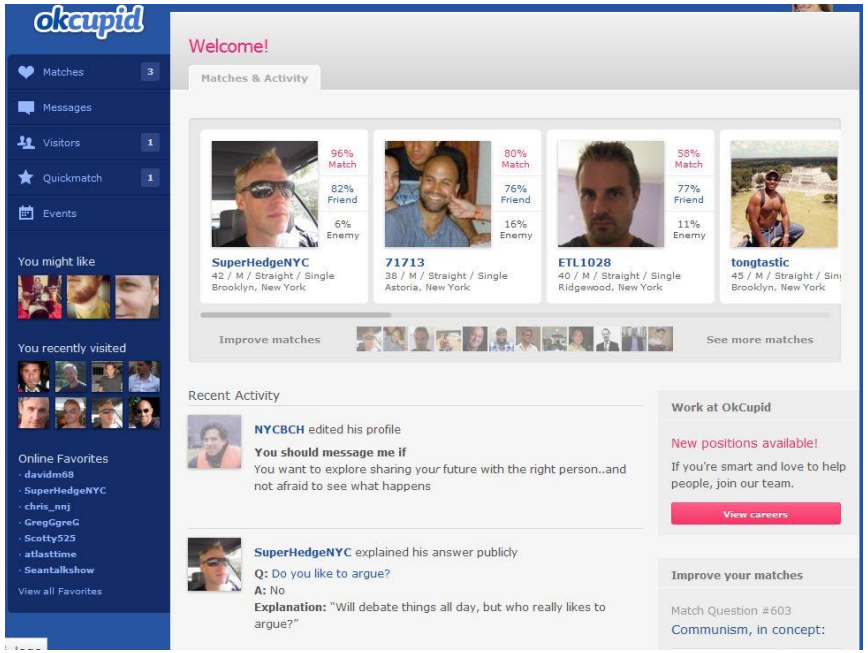
Introduction

- Use Case 1.2: OKCupid profile dataset
  - Analysis of a public dataset of almost 60,000 online dating profiles
  - Dataset published in the *Journal of Statistics Education, Volume 23, Number 2 (2015) by Albert Y. Kim et al.,*
  - Collection and distribution explicitly allowed by OkCupid president and co-founder Christian Rudder
  - Using these data is therefore ethically and legally acceptable. This is in contrast to another recent release of a different OkCupid profile dataset, which was collected without permission and without anonymizing the data (more on the ethical issues in this Wired article)

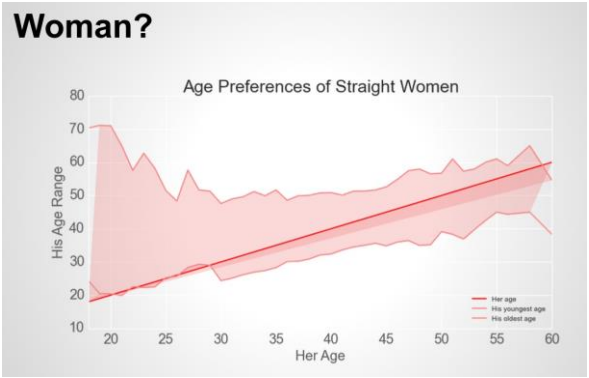


Introduction

- Use Case 1.2: OKCupid profile dataset
  - Official Site: https://www.okcupid.com/



- Developers and scientific community



Introduction

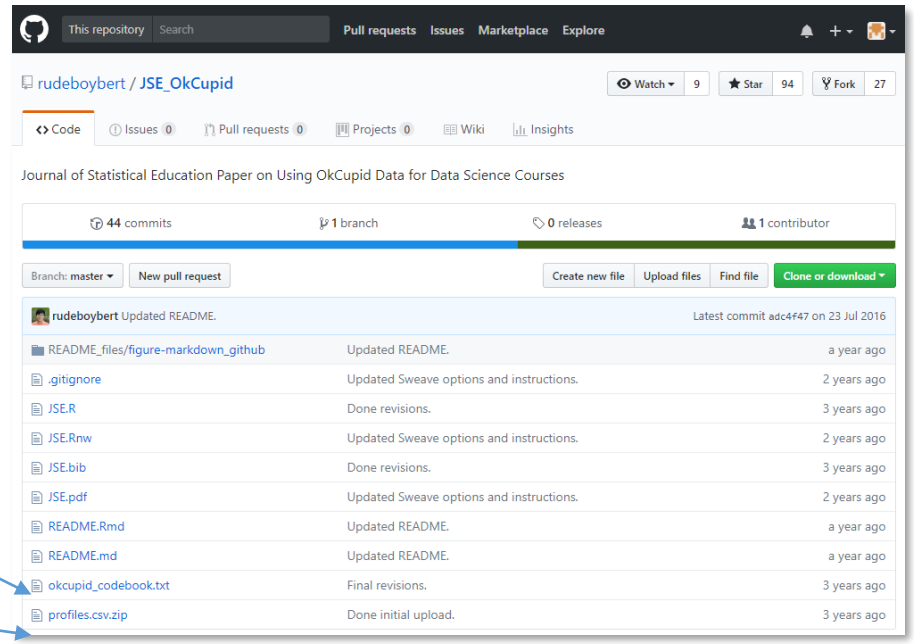
- Use Case 1.2: OKCupid profile dataset
  - Dataset details https://github.com/rudeboybert/JSE_OkCupid

    - The dataset was collected by web scraping the OKCupid.com website on 2012/06/30, and includes almost 60k profiles of people within a 25 mile radius of San Francisco, who were online in the previous year (after 06/30/2011), with at least one profile picture.

    - Metadata: codebook.txt

    - Data: profiles.csv

(link available at Virtual Campus Module5, Lesson2)



Introduction

- Use Case 1.2: OKCupid profile dataset
  - Dataset: The CSV contains a row (observation) for each profile

d - DataFrame

Index	ethnicity	height	income	job	last_online	location	offspring	orientation	pets	religion	sex	sign	smokes	speaks	status
0	asian, white	75	-1	transportati…	2012-06-28-2…	south san francisco, c…	doesn’t have kids, b…	straight	likes dogs and likes ca…	agnosticism and very ser…	m	gemini	sometimes	english	single
1	white	70	80000	hospitality / travel	2012-06-29-2…	oakland, california	doesn’t have kids, b…	straight	likes dogs and likes ca…	agnosticism but not too …	m	cancer	no	english (fluently), …	single
2	nan	68	-1	nan	2012-06-27-0…	san francisco, c…	nan	straight	has cats	nan	m	pisces but it doesn’…	no	english, french, c++	available
3	white	71	20000	student	2012-06-28-1…	berkeley, california	doesn’t want kids	straight	likes cats	nan	m	pisces	no	english, german (poor…	single
4	asian, black, other	66	-1	artistic / musical / wr…	2012-06-27-2…	san francisco, c…	nan	straight	likes dogs and likes ca…	nan	m	aquarius	no	english	single
5	white	67	-1	computer / hardware / s…	2012-06-29-1…	san francisco, c…	doesn’t have kids, b…	straight	likes cats	atheism	m	taurus	no	english (fluently), …	single
6	white, other	65	-1	nan	2012-06-25-2…	san francisco, c…	nan	straight	likes dogs and likes ca…	nan	f	virgo	nan	english	single
7	white	65	-1	artistic / musical / wr…	2012-06-29-1…	san francisco, c…	doesn’t have kids, b…	straight	likes dogs and likes ca…	christianity	f	sagittarius	no	english, spanish (oka…	single
8	white	67	-1	nan	2012-06-29-2…	belvedere tiburon, cal…	doesn’t have kids	straight	likes dogs and likes ca…	christianity but not too …	f	gemini but it doesn’…	when drinking	english	single
9	white	65	-1	student	2012-06-28-2…	san mateo, california	nan	straight	likes dogs and likes ca…	atheism and laughing abo…	m	cancer but it doesn’…	no	english (fluently)	single
10	white	70	-1	nan	2012-06-04-1…	san francisco, c…	nan	straight	nan	nan	m	taurus	yes	english	available
11	white	72	40000	banking / financial / …	2012-05-22-1…	daly city, california	nan	straight	likes cats	christianity and very ser…	m	leo but it doesn’…	no	english (fluently), …	seeing someone
12	white	72	-1	entertainment / media	2012-05-28-2…	san francisco, c…	doesn’t have kids	straight	nan	other	m	taurus	nan	english	single
13	white	66	30000	sales /	2012-06-13-1…	san	nan	straight	has dogs and	christianity	f	nan	no	english	single

Introduction

- Use Case 1.2: OKCupid profile dataset
  - Dataset: The CSV contains a row (observation) for each profile

d - DataFrame

Index	age	body_type	diet	drinks	drugs	education
0	22	a little extra	strictly anything	socially	never	working on college/univ…
1	35	average	mostly other	often	sometimes	working on space camp
2	38	thin	anything	socially	nan	graduated from masters…
3	23	thin	vegetarian	socially	nan	working on college/univ…
4	29	athletic	nan	socially	never	graduated from college…
5	29	average	mostly anything	socially	nan	graduated from college…
6	32	fit	strictly anything	socially	never	graduated from college…
7	31	average	mostly anything	socially	never	graduated from college…
8	24	nan	strictly anything	socially	nan	graduated from college…
9	37	athletic	mostly anything	not at all	never	working on two-year col…
10	35	average	mostly anything	socially	nan	nan
11	28	average	mostly anything	socially	never	graduated from college…
12	24	nan	nan	often	nan	nan
13	30	skinny	mostly	socially	never	graduated

Introduction

- Use Case 1.2: OKCupid profile dataset
  - Dataset: The CSV contains a row (observation) for each profile
  - Includes essays, which comprises input texts (written by users) with regards to questions

- *essay0*: My self summary
- *essay1*: What I'm doing with my life
- *essay2*: I'm really good at
- *essay3*: The first thing people usually notice about me
- *essay4*: Favorite books, movies, show, music, and food
- *essay5*: The six things I could never do without
- *essay6*: I spend a lot of time thinking about
- *essay7*: On a typical Friday night I am
- *essay8*: The most private thing I am willing to admit
- *essay9*: You should message me if...

d - DataFrame

Index	essay0	essay1	essay2	essay3	essay4	essay5	essay6	essay7	essay8	essay9	
0	about me: 	currently working as a…	making people laugh. …	the way i look. i am a…	books: absurdistan,…	food. water. …	duality and humorous thi…	trying to find someone…	i am new to california a…	you want to be swept off…	a
1	i am a chef: this is what…	dedicating everyday to …	being silly. having ridic…	nan	i am die hard christopher …	delicious porkness in …	nan	nan	i am very open and wil…	nan	wl
2	i'm not ashamed of m…	i make nerdy software for…	improvising in different…	my large jaw and large gl…	okay this is where the cu…	movement >conversatio…	nan	viewing. listening. d…	when i was five years o…	you are bright, open…	n
3	i work in a library and …	reading things writt…	playing synthesizers…	socially awkward but …	bataille, celine, beck…	nan	cats and german philo…	nan	nan	you feel so inclined.	wl
4	hey how's it going? curre…	work work work work + …	creating imagery to l…	i smile a lot and my inqui…	music: bands, rappers, mus…	nan	nan	nan	nan	nan	a o
5	i'm an australian l…	building awesome stuf…	imagining random shit.…	i have a big smile. i als…	books: to kill a mocki…	like everyone else, i love…	what my contribution…	out with my friends!	i cried on my first day at…	you're awesome.	wl
6	life is about the little t…	digging up buried treas…	frolicking witty	i am the last unicorn	i like books. ones with pi…	laughter >amazing	synchronicit…	plotting to take over th…	my typical friday night	nan	wl
7	nan	writing. meeting new …	remembering people's bir…	i'm rather approachable…	i like: alphabetized…	friends, family, note…	things that amuse and in…	out and about or relaxing …	nan	nan	wl
8	nan	oh goodness. at the momen…	nan	i'm freakishly b…	i am always willing to t…	sports/my softball glo…	nan	in or out... drinking wit…	potential friends/love…	http:// www.youtube.…	wl
9	my names jake. i…	i have an apartment. i…	i'm good at finding crea…	i'm short	i like some tv. i love s…	music, my guitar …	…	…	…	you can rock the bells	wl
10	update: i'm seeing someo…	i have three jobs. i've b…	hugging, kissing, lau…	my huge goofy smile	i'm constantly r…	family friends<br /…	snowboarding, food, women,…	having dinner and drinks w…	i used to wish for a j…	you are a complex woma…	wl
11	i was born in wisconsin, g…	i'm currently the youngest…	i'm really good at a li…	the way i dress. some …	books = yes. avid reader.…	guitar - even if i don't p…	a little bit of everythin…	hanging out with a small…	i'm picky when it come…	if you know who you are,…	wl
12	bang my shit bang	nan	nan	nan	nan	nan	nan	nan	nan	nan	wl
13	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	wl

Use Case Experimentation

- Use Case 1.2: OKCupid profile dataset
  - Step 1. Import required packages
    - `import pandas as pd` # high-performance, easy-to-use data structures and data analysis tools
    - `from prettypandas import PrettyPandas` # pandas to show pretty format tables **(required to pip install)**
    - `import numpy as np` # fundamental package for scientific computing
    - `import matplotlib.pyplot as plt` # 2D plotting library which produces publication quality figures in a variety of hardcopy formats
    - `import seaborn as sns` # high-level interface for drawing attractive statistical graphics
    - `from IPython.core.display import display, HTML` # Top-level display functions for displaying object in different formats

Use Case Experimentation

- Use Case 1.2: OKCupid profile dataset
  - Step 1. Import required packages
    - `import re # provides regular expression matching operations`
    - `import json # easy API to manage JSON files and encoding basic Python object hierarchies`
    - `import math # mathematic functions`
    - `import string # functions to easily manage strings`
    - `import tqdm # tool to manage intelligent iteration and progress bars (required to pip install)`
    - `from scipy.stats import kendalltau # perform statistics. Kendal tau is a ranking algorithm`
    - `import pymongo # Python distribution containing API tools for working with MongoDB`
    - `from pymongo import MongoClient, GEO2D # to obtain a MongoClient instance`

Use Case Experimentation

- Use Case 1.2: OKCupid profile dataset
  - Step 2. Perform a first data exploration
    1. Import data. Read_csv
    2. Connect to MongoBD
    3. Create mongo collection
    4. Transform dataframe to json records
    5. insert records to mongo collection
    6. Show one or several mongo documents to check that data is stored and accessible

{'_id': ObjectId('5a3b798a74ba000e6058abfa'), 'age': 32, 'body_type': 'fit', 'diet': 'strictly anything', 'drinks': 'socially', 'drugs': 'never', 'education': 'graduated from college/university', 'essay0': "life is about the little things. i love to laugh. it's easy to do\nwhen one can find beauty and humor in the ugly. this perspective\nmakes for a more gratifying life. it's a gift. we are here to play.", 'essay1': 'digging up buried treasure', 'essay2': 'frolicking
\nwitty banter
\nusing my camera to extract sums of a whole and share my perspective\nwith the world in hopes of opening up theirs
\nbeing amused by things most people would miss', 'essay3': 'i am the last unicorn', 'essay4': "i like books. ones with pictures. reading them is great too. where\ndo people find the time? i spend more time with other people not\nreading. i collect books. they sit neatly on my bookshelves.
\n
\nmovies are great. especially on movie night. with brownies.
\n
\nmusic. i love (love) it all. unless it's country.
\n
\ni love food.", 'essay5': 'laughter
\namazing people in my life
\ncolor
\ncuriosity
\nmusic and rhythm
\na good pair of sunglasses', 'essay6': "synchronicity
\n
\nthere is this whole other realm where the fabrics of our life\nstories intersect as they dance and play in a magical burst of\nenergy. this realm doesn't need you to believe in it in order to\nmaintain. it is a cluster of synchronicities and happenings. it is\na gift to those who notice them. something to be treasured\nappreciated. there is something special in each and every moment\nthat you experience in your daily waking life. this something\nbrings us back to the age old question: if a tree falls in the\nforest and no one is there to hear it, does it make a sound? this\nworks in the same way. if you are not consciously there to hear it,\nsee it, taste it, smell it, feel it none of this matters, it's\nstill there. pay attention to the little things, those that are\noften overlooked. see if you can find the magic in this gift we\ncall life.", 'essay7': 'plotting to take over the world with my army of segway riding\npandas and fire breathing kittens', 'essay8': 'my typical friday night', 'essay9': None, 'ethnicity': 'white, other', 'height': 65.0, 'income': -1, 'job': None, 'last_online': '2012-06-25-20-45', 'location': 'san francisco, california', 'offspring': None, 'orientation': 'straight', 'pets': 'likes dogs and likes cats', 'religion': None, 'sex': 'f', 'sign': 'virgo', 'smokes': None, 'speaks': 'english', 'status': 'single'}

Use Case Experimentation

- Use Case 1.2: OKCupid profile dataset
  - Step 3. Perform a first data exploration
    1. Query mongo to filter “males” and store in dataframe
    2. Query mongo to filter “females” and store in dataframe
    3. Calculate proportion of males versus females

       35829 males (59.8%), 24117 females (40.2%)

    4. Show the “Age” distribution
    5. Check, by means of mongo queries, whether there exists outliers, for example with **age > 80**

```
Age statistics:
count    59946.000000
mean        32.340290
std          9.452779
min         18.000000
25%         26.000000
50%         30.000000
75%         37.000000
max        110.000000
Name: age, dtype: float64

There are 2 users older than 80
```

Use Case Experimentation

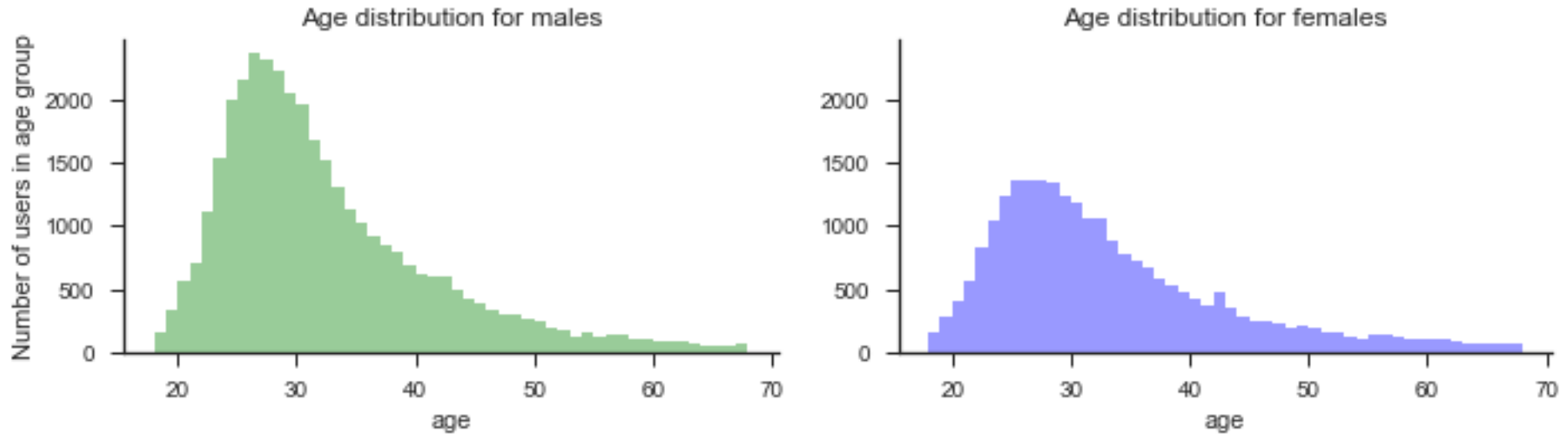
- Use Case 1.2: OKCupid profile dataset
  - Step 4. Eliminate outliers
    1. Query mongo to delete those users with age > 80

	age	body_type	diet	drinks	drugs	education	essay0	essay1	essay2	essay3	essay4	essay5	essay6	essay7	essay8	essay9	ethnicity	height
2512	110	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	67
25324	109	athletic	mostly other	nan	never	working on masters program	nan	nan	nan	nothing	nan	nan	nan	nan	nan	nan	nan	95

    2. Check that these two outliers are out of mongo collection
    3. Isolate again males and females and store in different dataframes from mongo queries
    4. Show age statistics (distributions) to check that no older than 80 remain

Use Case Experimentation

- Use Case 1.2: OKCupid profile dataset
  - Step 5. Draw age histograms for male and female users



Note that both distributions are right-skewed. Then, as is often (but not always!) the case, the mean is larger than the median.

Use Case Experimentation

- Use Case 1.2: OKCupid profile dataset
  - Step 5. Checking means and medias

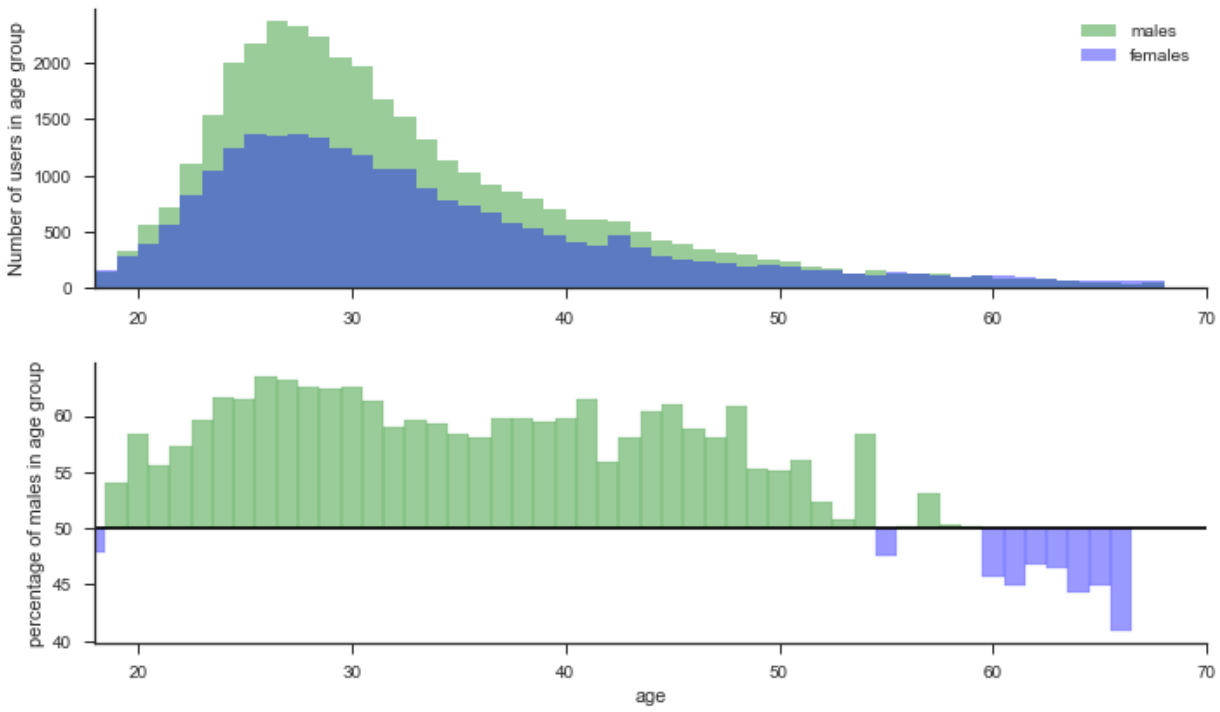
```
print("Mean and median age for males:   {:.2f}, {:.2f}".format(male["age"].mean(),male["age"].median()))
print("Mean and median age for females: {:.2f}, {:.2f}".format(female["age"].mean(),female["age"].median()))
```

```
Mean and median age for males:   32.02, 30.00
Mean and median age for females: 32.82, 30.00
```

Note that both distributions are right-skewed. Then, as is often (but not always!) the case, the mean is larger than the median.

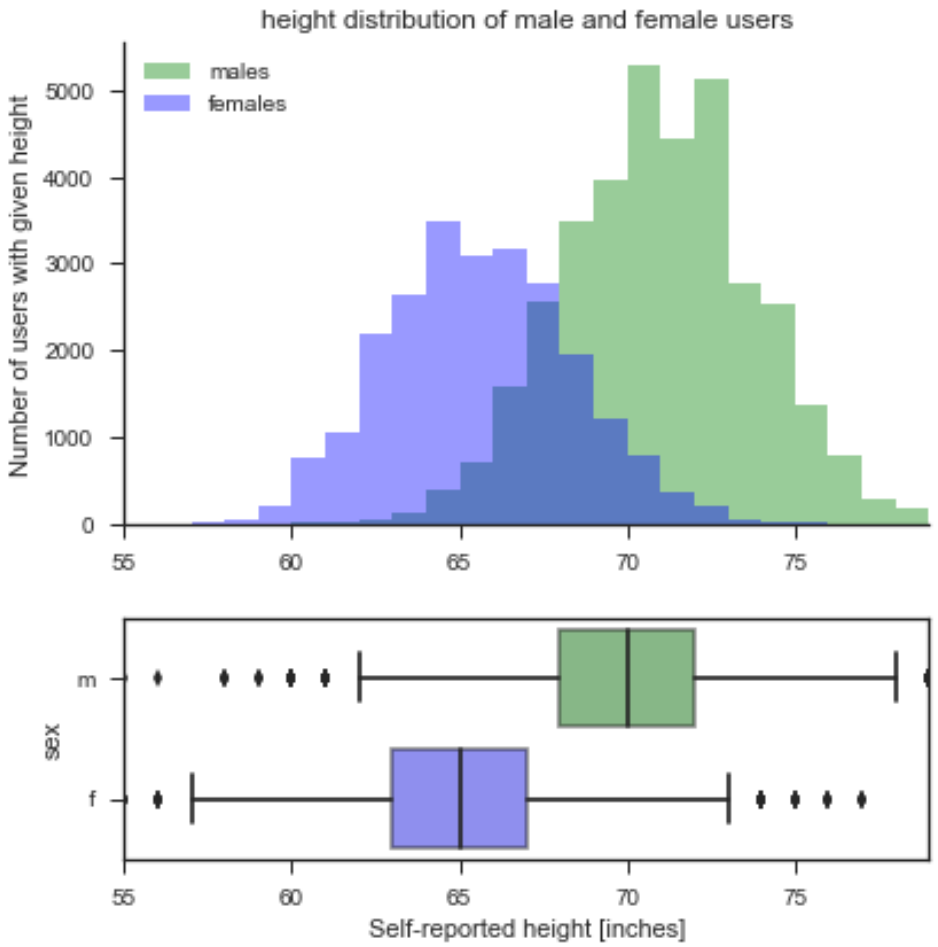
Use Case Experimentation

- Use Case 1.2: OKCupid profile dataset
  - Step 6. Females seem to be on average slightly older than males. Let's compare the age distributions in a single plot
    1. Plot the age distributions of males and females on the same axis
    2. Over-60 users are not many, but in this group there are significantly more females than males. This may be explained by the fact that, in this age group, there are more females than males in the general population



Use Case Experimentation

- Use Case 1.2: OKCupid profile dataset
  - Step 7. Analysis of height
    1. Plot the height distribution for males and females in the whole dataset



Use Case Experimentation

- Use Case 1.2: OKCupid profile dataset
  - Step 8. Cross-checking the data with external datasets (CDC)
  - https://www.cdc.gov/

Use Case Experimentation

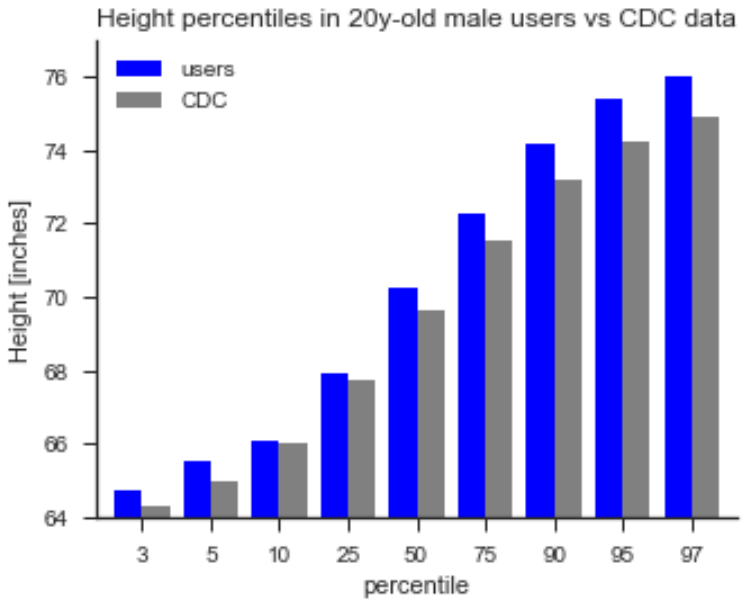
- Use Case 1.2: OKCupid profile dataset
  - Step 8. Cross-checking the data with external datasets (CDC)
    1. The CDC publishes growth charts, which contain height data for the general US population.
    2. The dataset reports statistics (3rd, 5th, 10th, 25th, 50th, 75th, 90th, 95th, 97th percentiles) for stature for different
    3. Ages from 2 to 20 years. This (and more) data is plotted by the CDC in these beautiful charts.

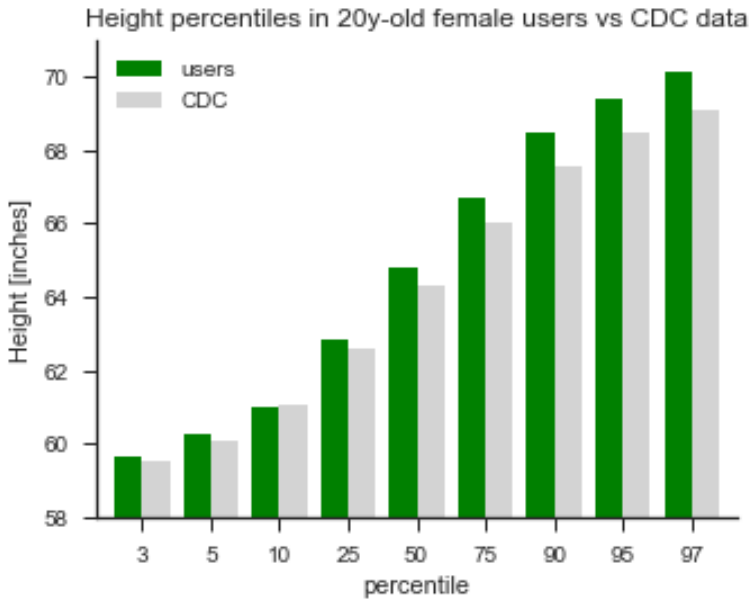
Use Case Experimentation

- Use Case 1.2: OKCupid profile dataset
  - Step 8. Cross-checking the data with external datasets (CDC)
    1. Create a new mongo collection to store and manage data https://www.cdc.gov/growthcharts/data/zscore/statage.csv
    2. Accommodate “Sex” field to match with OKCupid
    3. Compare the stats for reported heights of our 20-year-olds to the CDC stats for 20-year-olds.
    4. OKCupid height data are integers, which also causes all percentiles to be integer values. To fix this, we jitter the data by ±0.5±0.5 inches by adding random uniformly distributed noise in the range [−0.5,+0.5][−0.5,+0.5]

Use Case Experimentation

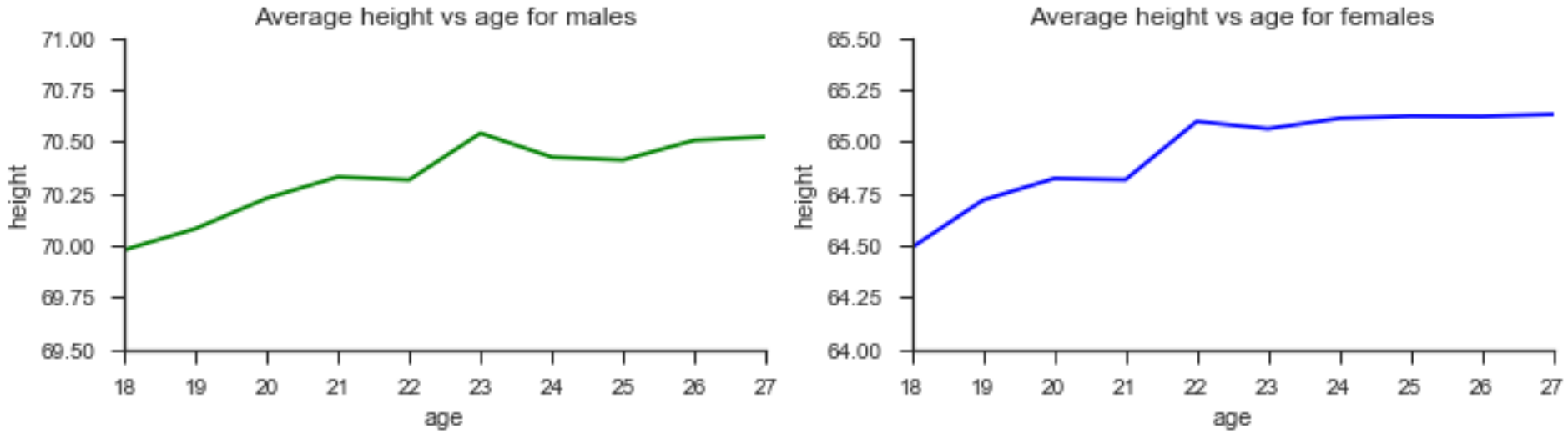
- Use Case 1.2: OKCupid profile dataset
  - Step 8. Cross-checking the data with external datasets (CDC)
    - Plot differences in percentiles





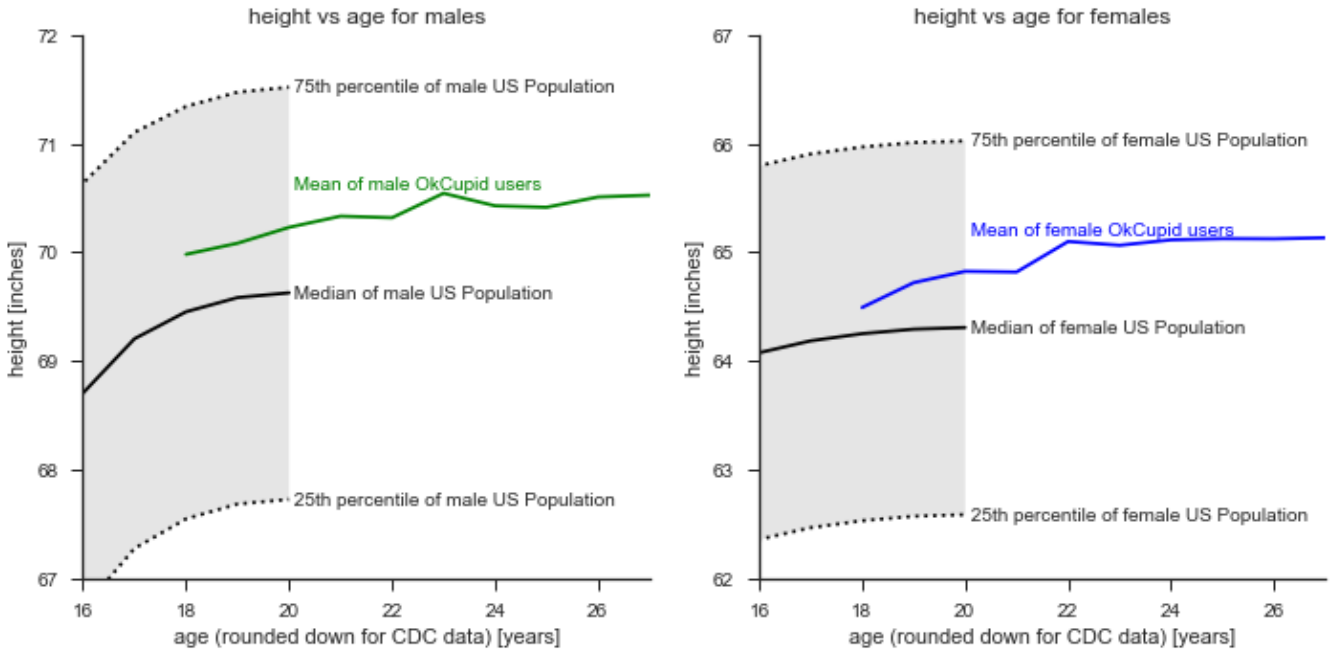
Use Case Experimentation

- Use Case 1.2: OKCupid profile dataset
  - Step 9. Study how height changes with age



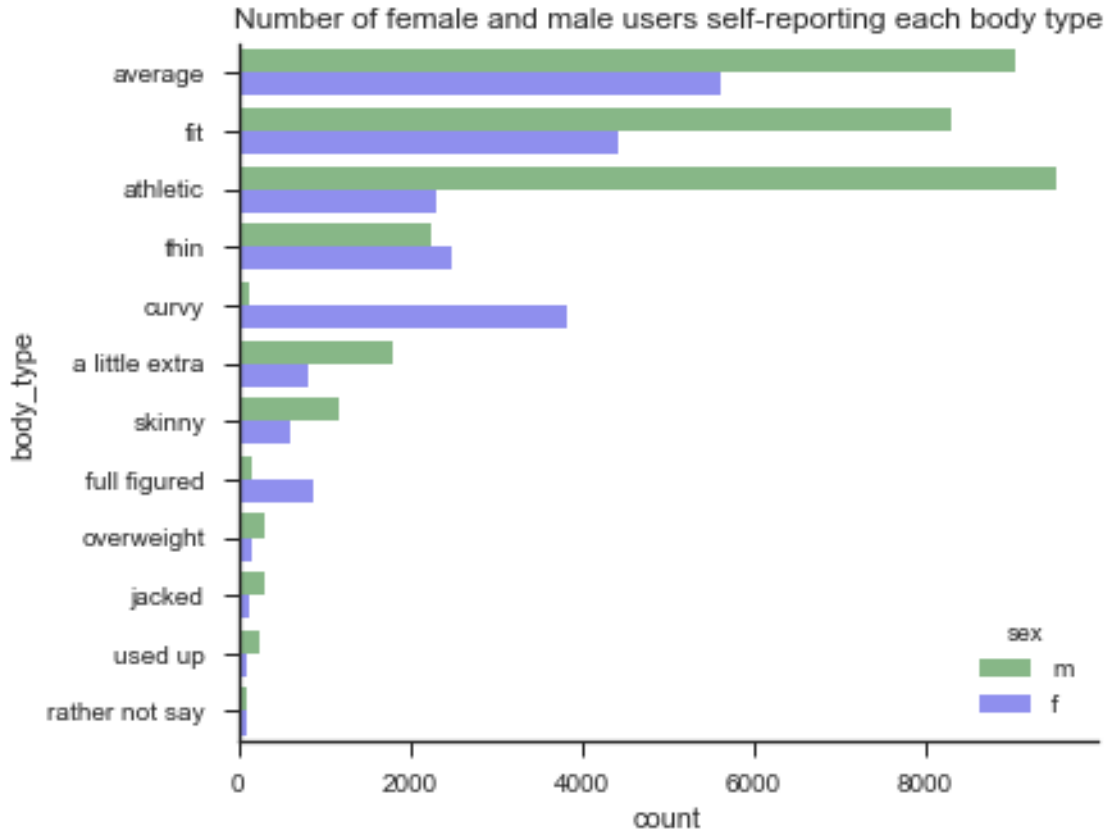
Use Case Experimentation

- Use Case 1.2: OKCupid profile dataset
  - Step 10. Compare CDC and OKCupid Percentiles



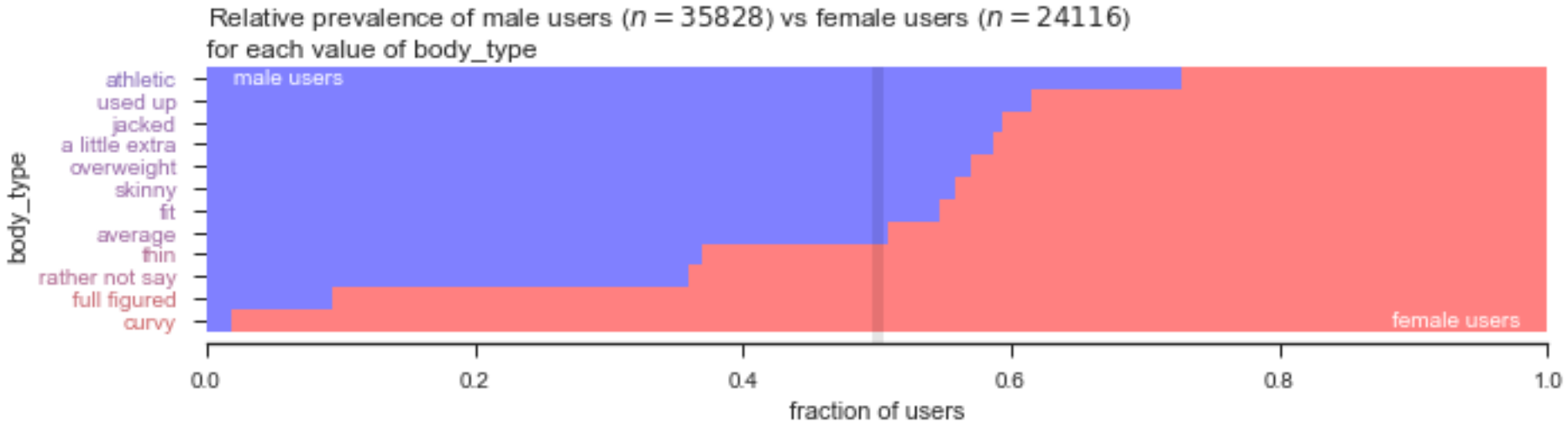
Use Case Experimentation

- Use Case 1.2: OKCupid profile dataset
  - Step 11. How do users self-report their body type?



Use Case Experimentation

- Use Case 1.2: OKCupid profile dataset
  - Step 12. Males and females are two sub-groups of the population, whereas body_type is a categorical attribute
    - It is interesting to compare how users in each of the two sub-groups (i.e. males and females) are likely to use each of the available categorical values; this is normally done through contingency tables.



Use Case Experimentation

- Use Case 1.2: The data contains essays written by the users on the following topics:

- *essay0*: My self summary
- *essay1*: What I'm doing with my life
- *essay2*: I'm really good at
- *essay3*: The first thing people usually notice about me
- *essay4*: Favorite books, movies, show, music, and food
- *essay5*: The six things I could never do without
- *essay6*: I spend a lot of time thinking about
- *essay7*: On a typical Friday night I am
- *essay8*: The most private thing I am willing to admit
- *essay9*: You should message me if...

d - DataFrame

Index	essay0	essay1	essay2	essay3	essay4	essay5	essay6	essay7	essay8	essay9	
0	about me: 	currently working as a…	making people laugh. …	the way i look. i am a…	books: absurdistan,…	food. water. …	duality and humorous thi…	trying to find someone…	i am new to california a…	you want to be swept off…	a
1	i am a chef: this is what…	dedicating everyday to …	being silly. having ridic…	nan	i am die hard christopher …	delicious porkness in …	nan	nan	i am very open and wil…	nan	wl
2	i'm not ashamed of m…	i make nerdy software for…	improvising in different…	my large jaw and large gl…	okay this is where the cu…	movement >conversatio…	nan	viewing. listening. d…	when i was five years o…	you are bright, open…	n
3	i work in a library and …	reading things writt…	playing synthesizers…	socially awkward but …	bataille, celine, beck…	nan	cats and german philo…	nan	nan	you feel so inclined.	wl
4	hey how's it going? curre…	work work work work + …	creating imagery to l…	i smile a lot and my inqui…	music: bands, rappers, mus…	nan	nan	nan	nan	nan	a: o
5	i'm an australian l…	building awesome stuf…	imagining random shit.…	i have a big smile. i als…	books: to kill a mocki…	like everyone else, i love…	what my contribution…	out with my friends!	i cried on my first day at…	you're awesome.	wl
6	life is about the little t…	digging up buried treas…	frolicking witty	i am the last unicorn	i like books. ones with pi…	laughter >amazing	synchronicit…	plotting to take over th…	my typical friday night	nan	wl
7	nan	writing. meeting new …	remembering people's bir…	i'm rather approachable…	i like: alphabetized…	friends, family, note…	things that amuse and in…	out and about or relaxing …	nan	nan	wl
8	nan	oh goodness. at the momen…	nan	i'm freakishly b…	i am always willing to t…	sports/my softball glo…	nan	in or out... drinking wit…	potential friends/love…	http:// www.youtube.…	wl
9	my names jake. i…	i have an apartment. i…	i'm good at finding crea…	i'm short	i like some tv. i love s…	music, my guitar …	…	…	…	you can rock the bells	wl
10	update: i'm seeing someo…	i have three jobs. i've b…	hugging, kissing, lau…	my huge goofy smile	i'm constantly r…	family friends<br /…	snowboarding, food, women,…	having dinner and drinks w…	i used to wish for a j…	you are a complex woma…	wl
11	i was born in wisconsin, g…	i'm currently the youngest…	i'm really good at a li…	the way i dress. some …	books = yes. avid reader.…	guitar - even if i don't p…	a little bit of everythin…	hanging out with a small…	i'm picky when it come…	if you know who you are,…	wl
12	bang my shit bang	nan	nan	nan	nan	nan	nan	nan	nan	nan	wl
13	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	wl

Use Case Experimentation

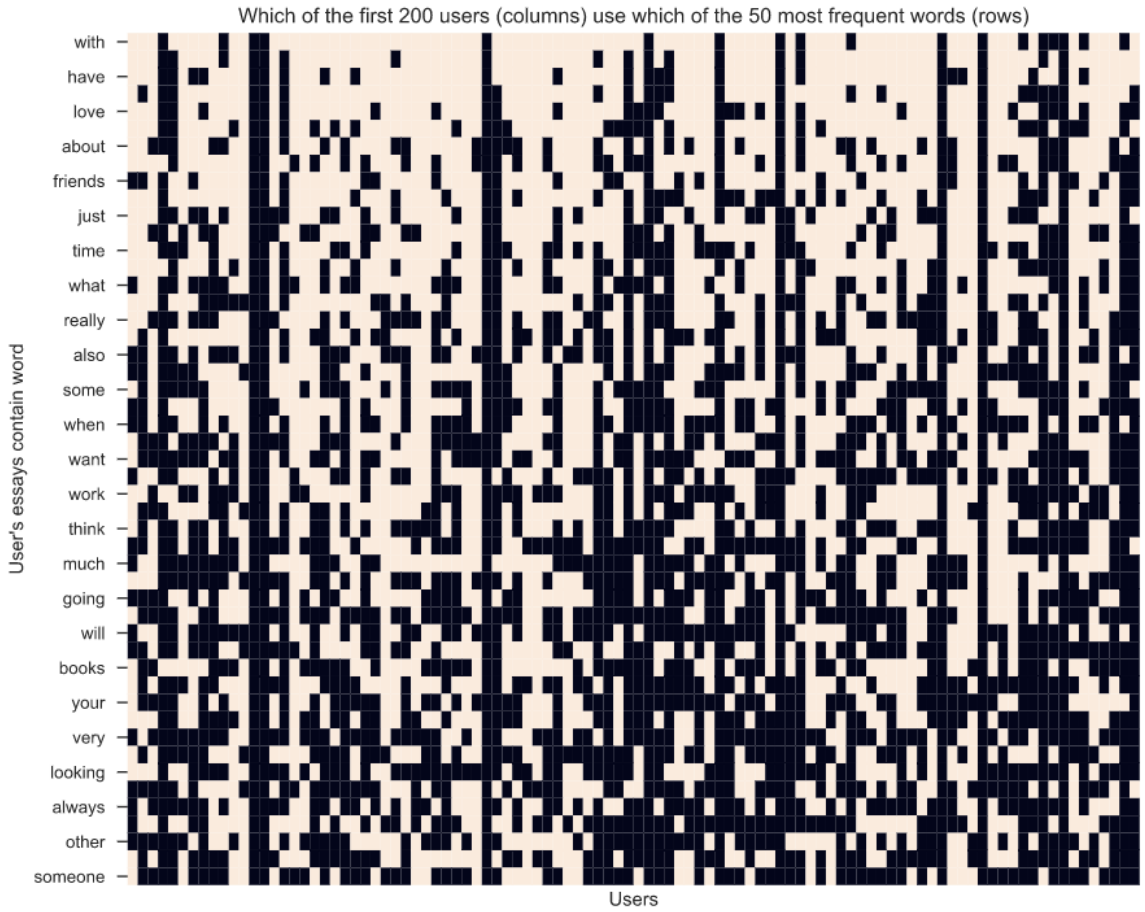
- Use Case 1.2: The data contains essays written by the users
  - We concatenate all essays to a single string and ignore the different themes

```
In [54]: for w,c in wordcounts.most_common(100):
             print(c,w)

1050302
823608 i
714584 and
679436 a
626851 the
595608 to
360865 my
356508 of
296365 in
268116 br
196373 i'm
186633 you
175822 with
171168 for
166527 that
149093 is
141352 have
134013 like
132263 it
129212 but
129154 on
127508 or
123236 me
118698 am
116091 love
115337 at
99079 be
98902 not
90958 are
90318 out
87647 good
86923 as
84950 class="ilink
82377 if
```

Use Case Experimentation

- Use Case 1.2: How frequent words are distributed among the users. i.e. how users speak?



Use Case Experimentation

- Use Case 1.2: Essays Analysis
  - Look for pattern affinities:
    - what does "Isaac" refer to? Asimov, not the game), apparently

```
isaac asimov      158
isaac hayes        16
isaac babel         5
isaac bashevis      5
isaac newton        2
isaac albeniz       2
isaac brock         2
isaac shepard       1
isaac i             1
isaac delgado       1
isaac assimov       1
isaac asimovi       1
Name: essays, dtype: int64
```

Use Case Experimentation

- Use Case 1.2: Essays Analysis
  - Mongo Queries to manage essays patters
    - Generate a collection to store essays
    - Look for affinities

```
### Stablishing patters to search and find common areas of interest between males and fmales
#pattern = "family"
pattern = "dogs"

pat = re.compile(pattern, re.I)

pipeline1 = [{"$match": {"education":"graduated from college/university" , 'essays': {'$regex': pat}}}]

pipeline2 = [{"$match": {"speaks":"english (fluently)" , 'essays': {'$regex': pat}}}]

pipeline3 = [{"$match": {"speaks":"spanish" , 'essays': {'$regex': pat}}}]
```

Discussions and Conclusions

- ✓ Motivation covered
- ✓ A first approach to real-world
- ✓ Exploring fine-grain data
- ✓ Involving open data extraction, cleaning, consolidation, transformation, enrichment, analysis and visualization
- ✓ Consolidate acquired knowledge and introduce new one
  - ✓ Bringing past and next modules

Module 5

Use Case 2

LESSON 2

José Manuel García Nieto – University of Málaga