Visualizing and Learning the Structure in Data

Rebecca Nugent
Department of Statistics, Carnegie Mellon University

PCMI Undergraduate Summer School 2016

July 1, 2016

Who am I?

- Too many Math/Stat degrees via Rice, Stanford, University of Washington (some Spanish just for fun)
- Former Swimmer, Runner, Ultimate Frisbee player
- Lots of travel and big ideas that seemed good at the time
- Currently Professor at Carnegie Mellon Statistics
- Director of the Undergraduate Statistics program
- Academic Advisor concentrating on preparation and transition
- Coach for CMU Athletics Dept
- Research Interests: clustering, record linkage, extracting high-dimensional structure, variable selection
- Applications include: characterizing civilian casualties in Syrian Civil War, modeling semantic organization in young children, modeling student interaction with online tutoring systems, early interventions for patients with unknown depression trajectories
- Learn more at `http://www.stat.cmu.edu/~rnugent`

Who are you?

On a piece of paper, please jot down quick responses:

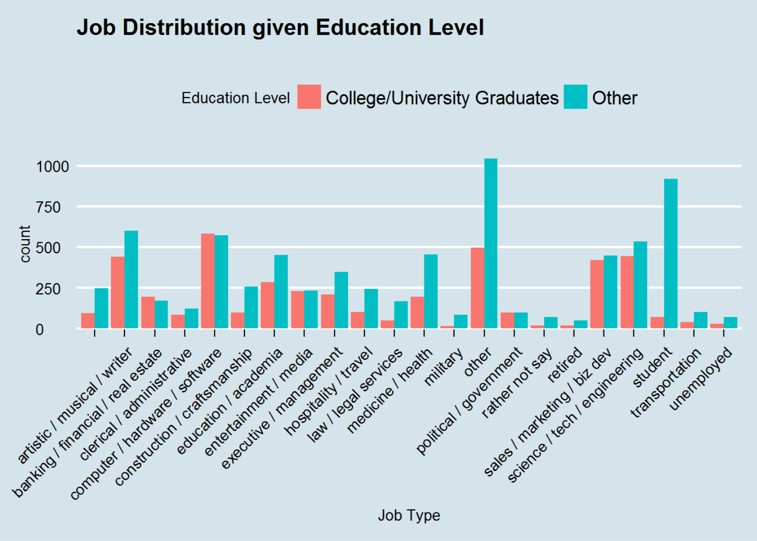
- What year are you?
- Major (current or former)
- What was your favorite/topic course in your major (so far)?
- What statistics classes have you had?
  (e.g. intro, regression, probability)
- What is/was your favorite topic from your statistics classes (if you took any)?
- How comfortable are you with the statistical programming language R?
  (1 - have no idea what you're talking about; 5 - bring it on)
- If you've graduated, what are you doing next year?
- If you haven't graduated, what do you think you want to do after school?
- Any particular topic you would like to learn more about?

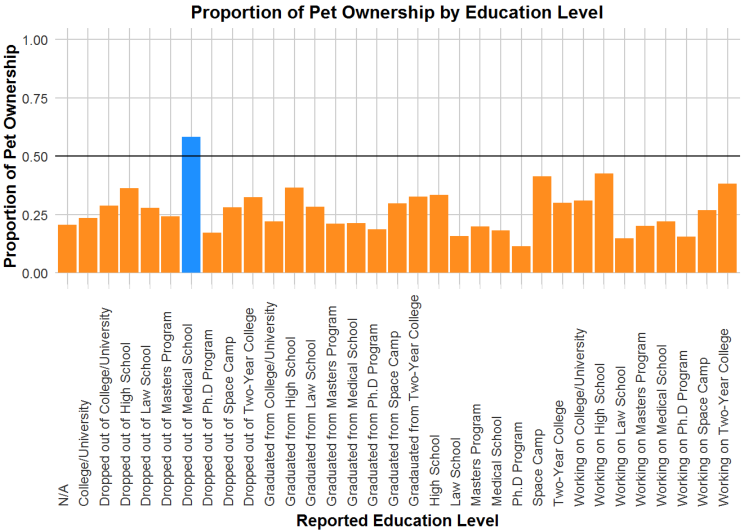
OkCupid Profiles

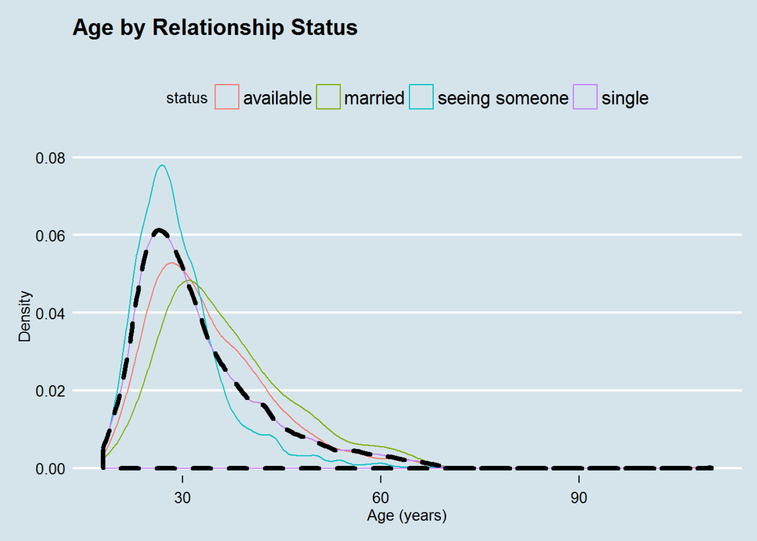
(created with OkCupid's permission by Albert Kim)

- 59,946 OkCupid users
- within 25 miles of San Francisco
- active profiles on June 26, 2012 and had been active in the previous year
- at least one picture in their profile
- Collected variables include body type, diet, education, ethnicity, offspring, sexual orientation, pets, religion, and a set of essays
  - My self summary
  - I'm really good at
  - The six things I could never do without
  - The most private thing I am willing to admit

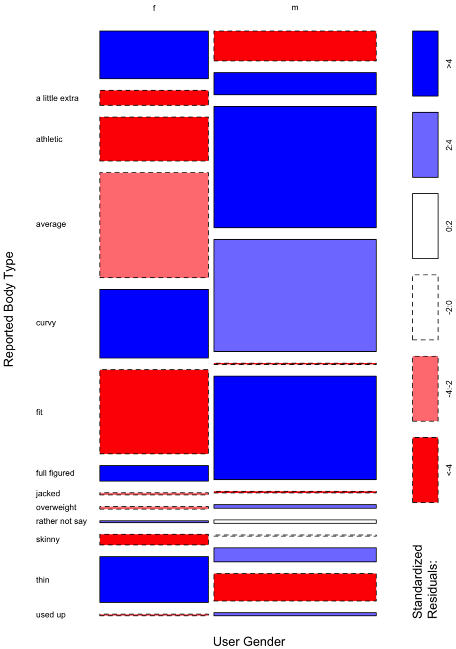
What kinds of questions might we ask about this population?

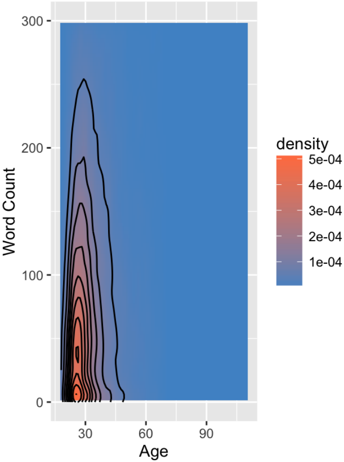

Job Distribution given Education Level
Education Level
College/University Graduates
Other
count
1000
750
500
250
0
artistic / musical / writer
banking / financial / real estate
clerical / administrative
computer / hardware / software
construction / craftsmanship
education / academia
entertainment / media
executive / management
hospitality / travel
law / legal services
medicine / health
military
other
political / government
rather not say
retired
sales / marketing / biz dev
science / tech / engineering
student
transportation
unemployed
Job Type


Proportion of Pet Ownership by Education Level
Proportion of Pet Ownership
1.00
0.75
0.50
0.25
0.00
N/A
College/University
Dropped out of College/University
Dropped out of High School
Dropped out of Law School
Dropped out of Masters Program
Dropped out of Medical School
Dropped out of Ph.D Program
Dropped out of Space Camp
Dropped out of Two-Year College
Graduated from College/University
Graduated from High School
Graduated from Law School
Graduated from Masters Program
Graduated from Medical School
Graduated from Ph.D Program
Graduated from Space Camp
Gradauated from Two-Year College
High School
Law School
Masters Program
Medical School
Ph.D Program
Space Camp
Two-Year College
Working on College/University
Working on High School
Working on Law School
Working on Masters Program
Working on Medical School
Working on Ph.D Program
Working on Space Camp
Working on Two-Year College
Reported Education Level

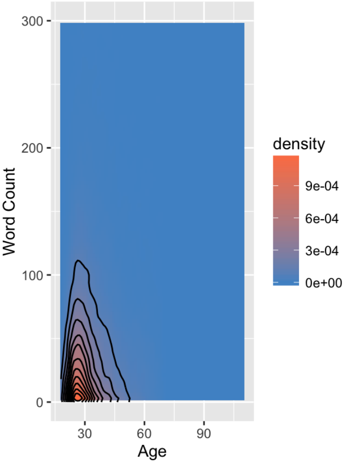

Age by Relationship Status
status
available
married
seeing someone
single
0.08
0.06
0.04
0.02
0.00
Density
30
60
90
Age (years)

Distribution of Body Type Based on Gender

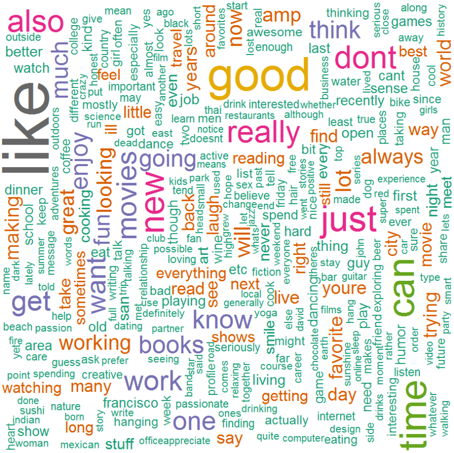




Heat Map of Age and Word Count Conditioned By Essay0: 'Self-Summary'



Heat Map of Age and Word Count Conditioned By Essay1: 'What I'm Doing With My Life'


like
good
really
just
time
can
new
movies
going
work
books
know
one
get
want
think
dont
also
much
enjoy
looking
fun
always
lot
also
outside
better
watch
dinner
making
school
keep
adventures
outdoors
coffee
great
cooking
take
sometimes
working
many
watching
name
dark
lately
summer
message
words
told
help
beach
passion
fire
yet
area
care
guess
ask
prefer
point
spending
creative
done
sushi
indian
heart
show
woman
nature
born
long
story
mexican
stuff
francisco
write
hanging
week
office
appreciate
passionate
ones
finding
drinking
say
quite
computer
actually
getting
living
internet
design
eating
mean
give
college
kind
different
crazy
honest
put
country
girl
often
feel
important
mostly
may
science
run
ill
little
especially
almost
another
film
yes
look
ago
black
easy
travel
years
lots
short
around
favorites
start
now
real
lost
amp
awesome
enough
even
job
learn
men
thai
drink
interested
restaurants
whether
although
two
notice
doesnt
east
got
dead
dance
active
means
reading
kids
tend
park
small
though
back
heads
laugh
used
hope
sex
believe
change
end
list
past
tell
friday
free
hair
stories
went
positive
nice
still
every
bit
top
find
open
least
true
places
taking
way
man
thinking
serious
close
games
along
history
away
last
best
world
business
lived
water
cant
sense
house
cool
recently
bike
since
girls
series
dog
made
experience
first
year
red
super
spent
ever
night
meet
lets
share
sure
movie
city
car
beer
john
guy
stay
thing
hard
spend
grew
let
jazz
whats
weekend
never
will
high
wine
art
fan
club
possible
loving
talk
eat
trip
writing
talking
relationship
san
full
met
use
definitely
dating
old
bad
read
playing
everything
see
etc
fiction
next
local
generally
cook
everyone
right
theres
live
bar
guitar
youre
exploring
type
smart
party
trying
partner
yoga
else
david
dancing
films
hang
play
friend
humor
order
video
future
shows
seriously
smile
might
far
course
career
chocolate
earth
favorite
sunshine
online
sleep
makes
moment
rather
listen
seeing
star
band
said
road
profile
comes
relaxing
together
game
day
need
drinks
interesting
side
whatever
walking