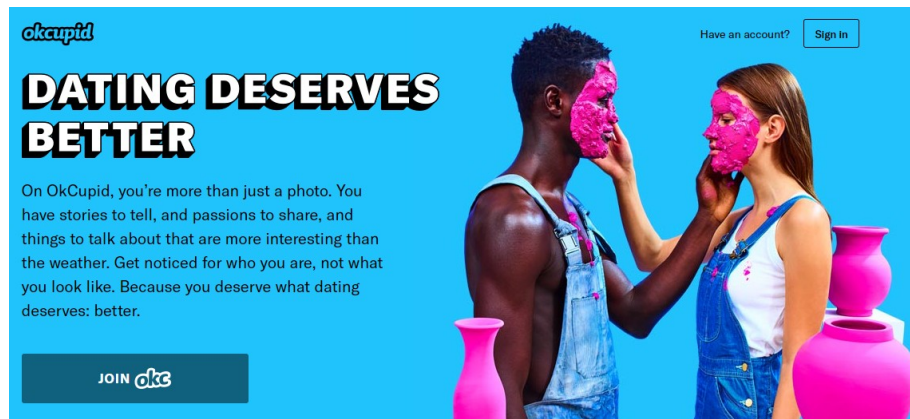


## Scenario



You are part of a business intelligence team at okcupid.com. The team has been asked to make an in-depth exploratory analysis of site users. The goal of the marketing team is to create micro segments and personas for future campaigns. Keep in mind, interesting data correlations may not be beneficial in a marketing context. For example, identifying 5 users with very specific attributes may be interesting but hardly a segment worth attracting.

You are asked to examine the data, clean it, use supplemental data to enrich the data then identify 4 or more interesting insights from the user data. All relevant cleaning, enriching and EDA steps along with the 4 insightful data nuances should be organized into a presentation. Your team will present to the head of marketing who is looking for an “ah -ha” persona or previously unknown data relationship. As the head of marketing, relevant information is consumed visually instead of in table form. Thus, your presentation should include visualizations when appropriate. Your team will need to turn in code and PowerPoint slides.

**\*\*On the day of the presentation, in addition to emailing electronic slides, print 3 copies for the professor and TAs to take notes\*\***

## Data

Source: <https://www.researchgate.net/project/The-OKCupid-dataset-A-very-large-public-dataset-of-dating-site-users>

This data set was scraped from user profiles. At the time, OKCupid did not authorize the data to be collected. After the data was released as part of academic literature, the data was authorized to be used by OKCupid.com .

***As a result, there is some moral ambiguity related to the use of the dataset.***

The data set your team is using has been authorized, cleaned and anonymized.

From the package author, "R package of cleaned profile data from "OkCupid Profile Data for Introductory Statistics and Data Science Courses": 59,946 OkCupid users who were living within 25 miles of San Francisco, had active profiles on June 26, 2012, were online in the previous year, and had at least one picture in their profile. The original data, publication, code, and codebook can be found at [https://github.com/rudeboybert/JSE\\_OkCupid](https://github.com/rudeboybert/JSE_OkCupid)"

To get the data run the following in your console:

```
install.packages('okcupiddata')
library(okcupiddata)
data('profiles')
```

### Example Abridged Data

age	body_type	diet	drinks	education	height	income	...	status
22	a little extra	strictly anything	socially	working on college/university	75	NA	...	single
35	average	mostly other	often	working on space camp	70	80000	...	single
38	thin	anything	socially	graduated from masters program	68	NA	...	available
23	thin	vegetarian	socially	working on college/university	71	20000	...	single
29	athletic	NA	socially	graduated from college/university	66	NA	...	single

Course Scripting Supplemental

You will receive an initial script with code examples to get you started since this is the first case of the course.

Criteria for Success

The presentation will be evaluated on a 5 pt scale with the following criteria.

- **Organization** – Was the presentation well organized?
- **Delivery** – Was the content delivered clearly and persuasively with the audience in mind?
- **Documentation** – Was the data mined to support the conclusion?
- **Data Mining Process** – Did the team approach the problem similar (as applicable) to steps outlined in page 19 of the book?

Another resource may be a public R-Studio examination of the data

Keep in mind this may not be helpful but code can be examined for additional ideas.

[https://rstudio-pubs-static.s3.amazonaws.com/209370\\_b62220c849b946088b463fdbc935848.html](https://rstudio-pubs-static.s3.amazonaws.com/209370_b62220c849b946088b463fdbc935848.html)