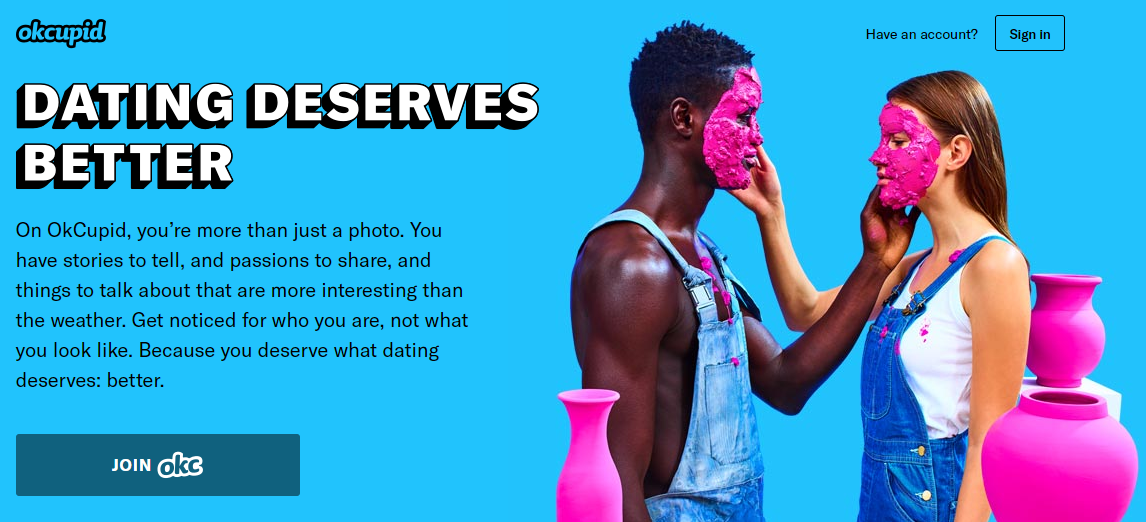
## Scenario



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 submission will include code and PowerPoint slides.**

## 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 provided has been authorized, cleaned and anonymized. The profiles are located in profiles.csv. The original data, publication, code, and codebook can be found at <https://github.com/rudeboybert/JSE_OkCupid>”

## 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.

## The Submission

* The submission will include business analyst slides covering the problem, data and 4 insights. Without a presentation, the “organization” section of the rubric will be 0. Exceptional submissions are well ordered and provide a coherent narrative covering all 4 insights.
* The submission will include a written supplemental representing the 4 insights identified and described in the business presentation. The written portion can be 3-5 sentences for each insight in a bulleted list format. Exceptional submissions include statistics from external credible sources that support the identified personas or insights. For example, “…focus on pet owners over age XX because [some org/research] says this segment will grow YY over the next 5 years…”. Without a written supplemental that coincide with the narration and supported by code the “written supplemental” section will be 0.
* The submission will include either a recorded screen narration of the business presentation, a text file with a URL to a recording (like youtube video) or audio that is embedded into the slide deck. Tone, volume, cadence, use of filler words and pronunciation will be accounted for in this section. No points will be deducted based on English proficiency (ie ESL) but technical descriptions that are incorrect will be detrimental. Failure to submit a narration, the “delivery” section of the rubric will be 0.
* An R script covering all data munging, modeling (if applicable), evaluation (if applicable) and visualization construction used to create the presentation artifacts (you do not need to use R to construct the slides but it is possible) and come to the case outcomes. Your code must use the following R functions at least once throughout your code, group\_by, aggregate & subset. Make sure to that your code contains ample comments. Failure to turn in an R script will result in a “Documentation” score of 0.

## Criteria for Success

The presentation will be evaluated on an equal weighted scale with the following criteria. For example 20 points per each category [depends on the individual course weighting found in Canvas]

* **Organization** – Was the presentation well organized?
* **Delivery** – Was the content delivered clearly and persuasively with the audience in mind?
* **Code Documentation** – Was the data mined to support the conclusion?
* **Written Supplemental** – Are the bullets clear and supported in narration and code?
* **Data Mining Proces**s – Overall, as a complete portfolio of work, is the topic interesting, organized, researched, supported and delivered effectively?

## 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.* ***Submitting these visuals and code alone will not result in a good learning outcome or rubric socre.***

[*https://rstudio-pubs-static.s3.amazonaws.com/209370\_b62220c849b946088b463fdbec935848.html*](https://rstudio-pubs-static.s3.amazonaws.com/209370_b62220c849b946088b463fdbec935848.html)