STAT 306 Group Project Proposal
March 18.2022

1. **Motivation**
   To explore whether or not any combination of our chosen "base parameters" and/or one additional categorical variable in this database serves as a more accurate and/or precise predictive model for predicting Income level than the most accurate and precise base parameters, while either controlling for predicted influential variables (such as job or education) or assuming excluded variables are independent (to be decided).

2. **The source of the data**
   Data for OkCupid Profile Data for Introductory Statistics and Data Science Courses, Journal of Statistics Education July 2015, Volume 23, Number 2. The original manuscript was subsequently revised in 2021.
   https://github.com/rudeboybert/JSE_OkCupid
   https://github.com/rudeboybert/JSE_OkCupid/blob/master/profiles_revised.csv.zip

3. **Variable description**
   All variables measured were collected by OkCupid's user profiles at an anonymized date 'X' in the 2010s, spanning the 2010s up to date X. Dataset was revised in 2021.

> *The data consists of the public profiles of 59,946 OkCupid users who were living within 25 miles of San Francisco, had active profiles during a period in the 2010s, were online in the previous year, and had at least one picture in their profile. Using a Python script, data was scraped from users' public profiles four days later; any non-publicly facing information such as messaging was not accessible.*
> *Variables include typical user information (such as sex, sexual orientation, age, and ethnicity) and lifestyle variables (such as diet, drinking habits, smoking habits). Note that random noise was added to the age variable for de-identification purposes.*
>
> Kim, A. Y., & Escobedo-Land, A. (2015). OkCupid Data for Introductory Statistics and Data Science Courses. In Journal of Statistics Education (Vol. 23, Issue 2). Informa UK Limited. https://doi.org/10.1080/10691898.2015.11889737

Response variable:

- **Income**: (US $, -1 means rather not say) -1, 20000, 30000, 40000, 50000, 60000 70000, 80000, 100000, 150000, 250000, 500000, 1000000.

Explanatory variables:

*We consider our "**base parameters**" to be four simple combinations: male*height, female*height, male*age, and female*age.*

- **age**: age of user with random noise added for anonymization (from 17 years old to 70 years old)
- **height**: inches

note: this dataset should not be confused for the infamously scraped set of ~70,000 OkCupid users collected around 2015, denounced for its breach of privacy and taken without permission of OkCupid

- **sex**: male and female
- body_type: rather not say, thin, overweight, skinny, average, fit, athletic, jacked, a little extra, curvy, full-figured, used up
- diet: mostly/strictly; anything, vegetarian, vegan, kosher, halal, other
- drinks: very often, often, socially, rarely, desperately, not at all
- drugs: never, sometimes, often
- education: graduated from, working on, dropped out of; high school, two-year college, university, masters program, law school, med school, Ph.D. program, space camp
- ethnicity: Asian, middle eastern, black, native American, Indian, pacific islander, Hispanic/Latin, white, other
- job: student, art/music/writing, banking/finance, administration, technology, construction, education, entertainment/media, management, hospitality, law, medicine, military, politics/government, sales/marketing, science/engineering, transportation, unemployed, other, rather not say, retire
- offspring: has a kid, has kids, does not have a kid, doesn't want kids; and/but might want them, wants them, doesn't want any, doesn't want more
- orientation: straight, gay, bisexual
- pets: has dogs, likes dogs, dislikes dogs; and has cats, likes cats, dislikes cats
- religion: agnosticism, atheism, Christianity, Judaism, Catholicism, Islam, Hinduism, Buddhism, Other; and very serious about it, and somewhat serious about it, but not too serious about it, and laughing about it
- sign: Aquarius, pieces, Aries, Taurus, Gemini, cancer, Leo, Virgo, Libra, Scorpio, Sagittarius, Capricorn; but it doesn’t matter, and it matters a lot, and it’s fun to think about
- smokes: yes, sometimes, when drinking, trying to quit, no
- status: single, seeing someone, married, in an open relationship

4. **Responsibilities**

Yinmengqing Ni: Data Analysis, Partial Writing, R Coding
Wayne Chen: Writing introduction and conclusion
Daniel Kennedy: Data Analysis including writeup and code, and assisting with introduction and conclusion
Hanxi Chen: Data Analysis, Partial Writing, R Coding