Gmail **Andrew Ho <kironide@gmail.com>**

Week 4 Day 1

1 message

Jonah Sinick <jsinick@gmail.com> Mon, Mar 7, 2016 at 3:35 AM
To: Ali Bagherpour <ali.bagherp@gmail.com>, Andrew Ho <Kironide@gmail.com>, Chad Groft <clgroft@gmail.com>, David Bolin <david@bolin.at>, Jacob Pekarek <jpekarek@trinity.edu>, Jaiwithani <jaiwithani@gmail.com>, James Cook <cookjw@gmail.com>, Linchuan Zhang <email.linch@gmail.com>, Matthew Gentzel <magw6270@terpmail.umd.edu>, Olivia Schaefer <taygetea@gmail.com>, Sam Eisenstat <sam.eisenst@gmail.com>, Tom Guo <tomguo4@gmail.com>, Trevor Murphy <trevor.m.murphy@gmail.com>

Continue working on the project from last week, in the same pairs (if applicable).

In the afternoon, start thinking in terms of preparing a tentative final .Rmd file in the afternoon, even if it feels like there's a lot more work to do. The goal here is to gain practice triaging. We'll have some presentations in the early evening.

==========

This project will probably be **challenging** because there are many subjective choices to make. It's our first pass at an open ended project. It will likely take you more than one day – we'll aim for student presentations on **Monday**.

I've attached the data from the National Longitudinal Study of Adolescent Health (as an .rda file, which you can just double click on to load it into R) as well as a codebook, which describes the variables and their values.

Focus in on the questions starting with "H2DA", "H2GH" and "H2NU," which concern exercise, health and nutrition. You'll probably find it useful to extract the age data implicitly in H2GI, and the sex information implicit e.g. in H2HS13.

Questions H2GH30, H2GH31, H2GH32 ,H2GH33, H2GH52 and H2GH53 touch on weight / height information and self-perceptions thereof. You may find the Wikipedia page on Body Mass Index to be useful.

Your task is to **create a model of who was of *healthy weight* in terms of age, sex, exercise and nutrition data, and interpret the model.**

Some things to think about:

1. Whether to use the features as is, or modify them in some way.
2. How to operationalize "overweight" and "underweight."
3. How to handle the missing values.
4. Whether you want to separate the analysis by gender / age or handle things simultaneously.
5. Whether to use all of the features, or just some subset. Whether to apply PCA or factor analysis for dimensionality reduction.

Write up your findings in a .Rmd file.