431 Project Instructions

Thomas E. Love

Version: 2018-11-02 05:52:43

Contents

Overview

This website contains the Fall 2018 project information for PQHS / CRSP / MPHP 431: Statistical Methods in Biological & Medical Sciences, Section 1.

- All materials related to the project (including these instructions) are maintained and linked at https://github.com/THOMASELOVE/431-2018-project.
- The direct link to this document is https://thomaselove.github.io/431-2018-project.

Your Project includes Two Studies

Your final project for this course will result in a portfolio of work related to two studies.

Study 1 - Class Survey. In the first study, you (sometimes working individually, sometimes in a group) will design, administer, analyze and present the results of a survey designed to compare two or three groups of subjects on some *categorical* and *quantitative* outcomes we will develop from your initial ideas.

Study 2 - Your Data. In the second study, you (working individually) will propose a research question and relevant data of interest to you, and then complete all elements of a data science project designed to create a statistical model for a *quantitative* outcome, then use it for prediction and assess the quality of those predictions.

You have Eight Tasks to Complete this Semester

The project involves two analyses (one for the class survey and one for your personal study), and a total of 8 tasks (deliverables.) Each task is to be completed by **12 NOON** on the specified date.

- Task B (Presentation Sign-Up) is due at noon on 2018-10-12, at http://bit.ly/431-2018-project-signup-taskB
- Task A (The Proposal) is now due at noon on 2018-10-15, via Canvas.
- Task C (Survey Editing) involves group work and is due at noon on 2018-10-23, via Canvas.
- Task D (Survey Comparison Plan) is now due at noon on 2018-11-02, at http://bit.ly/431-2018-survey-comparison-plan-
- Task E (Taking the Survey) is due at noon on 2018-11-02, as well. You'll find the survey at http://bit.ly/431-2018-survey.
- Task F (Sharing Study 2 Data) is due at noon on 2018-11-28, via Canvas.
- Task G (The Portfolio) is due at noon on 2018-12-13, via Canvas.
- Task H (Your Presentation) will be held on 2018-12-10, 2018-12-11 or 2018-12-13 in Dr. Love's office (Wood WG 82-L). The Schedule of Project Presentations is now available.

The bulk of this document contains specific instructions for each of these tasks.

Working with This Document

1. This document is broken down into multiple sections. Use the table of contents at left to navigate.

2. At the top of the document, you'll see icons which you can click to
   - search the document,
   - change the size, font or color scheme of the page, and
   - download a PDF or EPUB (Kindle-readable) version of the entire document.
3. The document is a work in progress, and will be updated occasionally through the semester. Check the Version information above to verify the last update time[1].

Need Help?

Questions about the project or the course can be directed to **431-help at case dot edu** or to Dr. Love directly at `thomas dot love at case dot edu`.

- The course home page is at https://github.com/THOMASELOVE/431-2018

[1] Note that the ePub and PDF versions will show slightly different times (but on the same day) as the HTML version.

Chapter 1

Project Objectives

It is hard to learn statistics (or anything else) passively; concurrent theory and application are essential[1].

1.1 Study 1 is about making comparisons and visualizing groups of data.

Study 1 involves data from a **class survey**, to be conducted in October. We will design, administer, analyze and present survey results designed to compare two or three groups of subjects from the class on some *categorical* and *quantitative* outcomes. In the analysis stage, everyone will be working with different parts of the same data set.

> Think of a graph as a comparison. All graphs are comparisons (indeed, all statistical analyses are comparisons). If you already have the graph in mind, think of what comparisons it's enabling. Or if you haven't settled on the graph yet, think of what comparisons you'd like to make. Andrew Gelman

In your eventual analysis of Study 1, you will be comparing both quantitative and categorical outcomes across 2-3 groups. All tools necessary for Study 1 are in Parts A and B of the course, and include the following…

- Descriptive and exploratory summaries of the data across the groups for each of your chosen outcomes, including, of course, attractive and well-constructed visualizations, graphs and tables.
- Comparisons of the population mean difference for at least one quantitative outcome across a set of two (or three) groups, including appropriate demonstrations of the reasons behind the choices you made between parametric, non-parametric and bootstrap procedures.
- Comparisons of the population proportions for at least one categorical outcome across your set of two (or three) groups, including appropriately interpreted point estimates and confidence intervals.

Note well that Study 1 is **not** about building sophisticated statistical models, and using them to make predictions. That's Study 2.

1.2 Study 2 is about building a model, and making predictions.

Study 2 involves data about a **research question that you will propose**, involving data of interest to you. Thus, everyone will be working with a different data set. You will complete all elements of a data

[1] Though by no means an original idea, this particular phrasing is stolen from Harry Roberts.

science project designed to create a statistical model for a *quantitative* outcome, then use it for prediction, and assess the quality of those predictions.

> All models are wrong but some are useful. George E. P. Box

In Study 2, you will be building a multiple linear regression model, and using it to predict a quantitative outcome of interest. The tools necessary for Study 2 appear in each Part of the course, especially Part C, and include the following…

- Describing the experimental or observational study design used to gather the data, as well as the complete data collection process.
- Sharing the complete raw data in an appropriate way with a statistician (Dr. Love). This means that, in general, data including protected health information are *not* appropriate for this project.
- Developing appropriate research questions that lead to the identification of smart measures for predictors and outcomes, and then the development of a prediction model using multiple linear regression.
- Using a training sample to develop a model, and present the process that leads to a final set of 2-3 candidate models in the training sample.
- Using a test sample to evaluate the quality of predictions from each of the candidate models, and making a final selection.
- Evaluating the adherence of the data you've collected to the assumptions of multiple linear regression, and iterating through the model-building process as necessary until the final model shows no strong violations of those assumptions.

1.3 Why Two Studies?

The main reason is that I can't figure out a way to get you to think about all of the things I hope you'll learn from this project in a single study.

1. I set different tasks for Study 1 and for Study 2, allowing us to touch on a wider fraction of the things I hope you'll learn in 431.
2. I want some of the work to be done as a class, some in groups, some as individuals.
3. Some of you have easy access to great data you want to study in this class, and in fact, that's a primary motivation for taking the class. But not all of you.
4. I have to evaluate each of your projects, and there are many students in the class. Knowing at least one of the data sets you'll be working with helps me manage this.
5. Having a broad range of activities to evaluate helps reduce the cost of a mistake on any one of them, so that we can build on what you do well.
6. All of Study 1 can be done by the middle of November, leaving the last few weeks of the semester for you to focus on Study 2.

1.4 Educational Objectives

> "Statistics has no reason for existence except as the catalyst for investigation and discovery." George E. P. Box

I am primarily interested in your learning something interesting, useful and even valuable from your project. An effective project will demonstrate:

1. The ability to create and formulate research questions that are statistically and scientifically appropriate.
2. The ability to turn research questions into measures of interest.
3. The ability to pull and merge and clean and tidy data, then present the data set following Jeff Leek's guide to sharing data with a statistician.
4. The ability to identify appropriate estimation / testing procedures for the class survey using both continuous and categorical outcomes.

5. The ability to build a reasonable model, including interactions and transformations to deal with non-linearity, assess the quality of the model and residual plots, then use the model to make predictions.
6. The ability to build a Table 1 to showcase potential differences between variables.
7. The ability to identify and (with help) solve problems that crop up
8. The ability to comment on your work within code, and in written and oral presentation.
9. The ability to build a Markdown-based report and a Markdown-based set of slides for presentation.

Chapter 2

Task A (The Proposal) Instructions

Task A requires you to accomplish the following:

On 2018-09-25, you will become part of a **group** of about five people, and your group will:

1. develop and propose 2-3 "research questions" for Study 1 (The Class Survey). Here, your research questions must clearly identify meaningful statistical comparisons.
2. propose 6-10 "homemade" survey questions for Study 1 that relate to your research questions, and
3. propose a "scale" for Study 1 that also relates to your research questions.

As an **individual**, you will also:

4. develop and propose a meaningful summary of your ideas and research question for Study 2 (Your Data). Your research question needs to clearly relate to modeling and prediction of a quantitative outcome on the basis of a set of predictor variables.
5. identify and present a detailed description of a data set that is likely to lead to an answer to the research question proposed for Study 2, and that is appropriate for use in this project.

The rest of this section contains guidance as to what sort of questions your group will need to propose for the class survey in Study 1, and as to what sorts of data sets and research questions are appropriate for your Study 2 proposal.

2.1 Deadline and Submission information

The project groups will only apply to Project Tasks A, C and D. You will become part of a project group on 2018-09-25 and the groups will disband after the survey is finalized in late October.

Task A is due at noon on 2018-10-15. (**Note the change of date.**) You will submit your Task A work through Canvas.

- The group work (Parts 1-3) need to be submitted by one of the members of your group.
  - If you are not the person submitting this information for your group, then your Task A submission should begin with the statement: "Parts 1-3 of Task A were submitted for my group by PERSON'S NAME."
- All students need to submit Parts 4 and 5 of Task A, individually. You can do this in advance of your group's submission of Parts 1-3 if you like.
- Please note that **Task A and Task B** are no longer due at the same time. Task B is due 2018-10-12, but Task A is now due 2018-10-15.

2.2 Study 1 work for Task A

In Study 1, you will survey your fellow students through an online instrument (containing somewhere around 100 items) that we will develop in Tasks A, C and D in September and October and then administer in the final week of October (Task E).

Students in the class will develop the items for this instrument in 10 groups of about 5 people per group. The final survey will include questions generated by each of the 10 groups in the class.

The course survey will be done online, and the respondents (de-identified, of course) will include all students in the current 431 class, plus the teaching assistants, and perhaps some volunteers from previous iterations of the course, in an effort to get a reasonable (but by no means random or representative) sample of graduate students at CWRU.

Remember that Study 1 is about making comparisons between groups.

2.2.1 Research Questions

The first part of Task A requires your group to develop and propose 2-3 "research questions" for Study 1 (The Class Survey).

> A research question is the fundamental core of a research project, study, or review of literature. It focuses the study, determines the methodology, and guides all stages of inquiry, analysis, and reporting. Source

- The research questions your project group will prepare for Study 1 should state the study objective in terms that allow us to apply statistical methods to test data to obtain an answer.
- Each research question should be written in the form of a comparison of 2-3 exposures or groups in terms of an outcome.
- At least one of your research questions needs to compare groups on a quantitative outcome, and at least one needs to compare groups on a categorical outcome (containing no more than 5 categories.)

Quoting Roger Peng, from Exploratory Data Analysis with R:

> Formulating a question can be a useful way to guide the exploratory data analysis process and to limit the exponential number of paths that can be taken with any sizeable dataset. In particular, a sharp question or hypothesis can serve as a dimension reduction tool that can eliminate variables that are not immediately relevant to the question. For example, (suppose that we are interested in) looking at an air pollution dataset from the U.S. Environmental Protection Agency (EPA).
>
> A general question one could ask is "Are air pollution levels higher on the east coast than on the west coast?" But a more specific question might be "Are hourly ozone levels on average higher in New York City than they are in Los Angeles?"
>
> Note that both questions may be of interest, and neither is right or wrong. But the first question requires looking at all pollutants across the entire east and west coasts, while the second question only requires looking at single pollutant in two cities. It's usually a good idea to spend a few minutes to figure out what is the question you're really interested in, and narrow it down to be as specific as possible (without becoming uninteresting).

2.2.2 Research Questions that have worked in the past

For Study 1, your research questions will need to compare two or more exposure groups on a quantitative outcome (in one case) and on a categorical outcome (in another case). Here are several examples that worked in the past:

- "Do messy people tend to have higher levels of self-described creativity than organized people?"

- "Is whether you voted in the last presidential election strongly associated with your level of interest in the current election?"
- "Does being in a committed romantic relationship result in higher self-esteem compared to not being in a committed romantic relationship?"
- "Do conscientious people spend more time every week engaged in activities, such as exercise, thought to promote health and well-being compared to others?"
- "Do graduate students who routinely pack their lunch have a lower BMI than graduate students who routinely purchase their lunch?"
- "Do individuals who spend a lot of time on social media each day have more or less social anxiety than those who do not?"

Don't boil the ocean here. You're looking for a research question that can be reasonably addressed in a survey of approximately 50 people, so it has to be pretty straightforward.

2.2.2.1 Checklist for Research Questions

1. Is our research question (RQ) something that we are curious about and that others might care about?
2. Does our RQ present an issue on which we can justify a stand prior to data collection about what we think will happen?
3. Is our RQ too broad, too narrow, or OK, given the time frame and restrictions of this survey?
4. Is our RQ measurable? What type of information do we need? Can I find a way to ask a limited number of survey questions in such a way to allow me to (after the data are collected) either support or contradict a position on my RQ?

- Adapted from this online tutorial from Empire State College

2.2.2.2 Tips on Writing Good Research Questions

- Duke has a nice overview online of key issues.
- Vanderbilt has some nice materials, built from the tutorial at Empire State College quoted earlier
- Jeff Leek provides several relevant tips in *The Elements of Data Analytic Style*
- https://researchrundowns.com/intro/writing-research-questions/ has some excellent tips on wording

2.2.3 Specifying Survey Questions

The second part of Task A requires your group to develop and propose 6-10 "homemade" survey questions for Study 1 that relate to your research questions.

Your group will need to specify the exact wording for your survey questions (and potential answers for any categorical responses.) This will likely require some editing and rework, once we have the complete set of proposed questions from all students. Be prepared to revise and resubmit, quickly, so that all items can be resolved in time for publication of the draft survey.

Of your 6-10 survey questions …

- at least two should ask the respondents to provide you with a **number** that expresses a quantitative outcome of interest to you, and these outcomes should relate closely to at least one of your research questions.
  - If you are asking people to respond to a prompt using a rating, that rating should be expressed on a wide scale. Our preference is a 0-100 scale for quantitative items, where 0 represents the most negative reaction and 100 the most positive reaction to the item.
  - One common choice is to make a statement and ask for agreement with that statement on a scale from 0 = Strongly Disagree to 100 = Strongly Agree.
  - The reason we prefer a 0-100 scale is to increase variation in our responses, as compared to, say, a 1-10 scale.

  - When responding to items using a scale like this on the actual survey, please use the whole scale.
- **at least two** should ask the respondents to provide you with a response that expresses a categorical outcome of interest to you, and these, too, should relate to at least one of your research questions.
  - You will need to specify each of the available responses that you wish to use in the survey.
  - No more than five options for your categorical outcome, please.
  - The response options you specify should be mutually exclusive and collectively exhaustive.
- **at least two** should ask the respondents to categorize themselves into one of two (or three) groups.
  - Be aware that you will need to have at least 10 students in each group in order to build a semi-reasonable analysis.
  - You should expect that 50-55 people will actually respond to the survey, in total.
  - Again, these groupings should be linked to your research questions.

You are welcome to submit exactly 6, or as many as 10 total survey questions in this part of the Task. It is likely that some of your survey questions will not correspond to some of your research questions, and that's OK, but each survey question should be linked to at least one of your research questions.

2.2.3.1 Dr. Love will include 15 Survey Questions Automatically

The following items will be included in the survey to identify "groups" of students in a reasonable way. As a result, you should not ask these questions in your proposed list, although you can and should consider whether these groupings would be good candidates for application to your research questions.

The following 7 items will have yes/no responses, and thus produce binary groups for analysis.

1. Were you born in the United States?
2. Is English the language you speak better than any other?
3. Do you identify as female?
4. Do you wear prescription glasses or contact lenses?
5. Before taking 431, had you ever used R before?
6. Are you currently married or in a stable domestic relationship?
7. Have you smoked 100 cigarettes or more in your entire life?

The next eight items generate non-binary responses. Together, after the survey is complete, we will identify "cutpoints" for these eight items to identify groups of meaningful size.

8. In what year were you born?
9. How would you rate your current health overall (Excellent, Very Good, Good, Fair, Poor)
10. For how long, in months, have you lived in Northeast Ohio?
11. What is your height in inches? (If you are five feet, eight inches tall, please write 68 inches. To convert from centimeters to inches, multiply your height in centimeters by 0.3937, and then round the result to the nearest inch.)
12. What is your weight in pounds? (To convert from kilograms to pounds, multiply your weight in kilograms by 2.2046, and then round the result to the nearest pound.)
13. What is your pulse rate, in beats per minute? (Please either use a tracking device, or count your pulse for 15 seconds then multiply by 4)
14. Last week, on how many days did you exercise? (0 - 7)
15. Last night, how many hours of sleep did you get?

2.2.3.2 Permitted Types of Items

The survey will be conducted using a Google Form, rather than Survey Monkey or some other tool. Thus, we have a somewhat restricted set of item types.

For **quantitative measures**, Google Forms permit the use of

1. a *short answer* item without any restrictions on the response (except a character limit)

2. a *short answer* item where respondents are forced to insert a number within a given range through a validation process that only accepts the response if it falls within the specified limits.
3. *linear scale* items for ordered categorical ratings (but only on a scale of up to ten points - i.e. 1 to 10)

For **categorical measures**, Google Forms permit the use of

1. *multiple choice* items for endorsing a single choice from a group of 2-10 alternatives.
2. *checkbox* items for the endorsement of one or more choices from a group of 2-10 alternatives.
3. *linear scale* items for ordered categorical ratings (often on a 1-X scale, where X is between 2 and 10)
4. *dropdown* items for selections of one option from a group of 2-3 choices.

2.2.3.3 Old Class Surveys

The surveys (each containing at least 100 items) from 2014, 2015, 2016 and 2017 are available as PDF documents on our web site.

Remember that the rules used this year have been modified from what has been used for the project previously.

2.2.4 Specifying a "Scale"

The third part of Task A requires your group to identify and propose a "scale" of items for Study 1 (The Class Survey).

Your group needs to specify a published scale (available for free public use) to generate an outcome or grouping(s) of interest from those completing the survey. You will have to provide a complete reference to the scale (online, ideally) and specify each of the items in the scale, and how the scale is then evaluated, in all necessary detail to allow us to review and replicate the scale in practice.

The word "scale" is used in many different ways. In this case, by a scale I mean a published list of items, usually accompanied by a scoring rubric that provides some sort of composite score or scores. Examples of scales we have used in the past include:

- Two Health Consciousness Scales, one due to Gould another to Dutta-Bregman Gould Health Consciousness Scale
- The Ten-Item Personality Inventory
- The Perceived Stress Scale
- The Epworth Sleepiness Scale

Your group will need to verify explicitly in your Task A materials that the scale your group proposes is freely available for use by anyone, without any fees or registration requirements.

2.3 Study 2 Work for Task A

You, individually, will present a proposal **summary** ($<$ 300 words) and a brief **data description** for Study 2 in Task A.

You will be building a multiple linear regression model, and using it to predict your outcome of interest.

2.3.1 The Proposal Summary

The fourth part of Task A is to develop and propose a meaningful summary of your ideas and research question for Study 2 (Your Data).

Your summary should begin with a title for your Study 2. Take the time to come up with a good, interesting title. You are going to work hard on this thing; please resist the temptation to kill my interest at the start by calling it "431 Statistics Project" or anything else that shows a similar lack of effort.

Provide me a very brief summary of what you're trying to accomplish - specifically, what your research question is, and what you hypothesize will happen.

The five most important things to do in the summary are:

1. Write clearly. My best advice is to finish the summary as soon as you can, and then give it to someone else to read, who can criticize it for lack of clarity in the writing.
2. Specify the topic of interest, and motivate your study of it.
3. Explicitly specify your key research question, which should be stated as a question, and which should clearly and naturally lead to a prediction model for a quantitative outcome.
4. Explain what you hypothesize will happen, and
5. Explicitly link your key research question to the data set you describe in your data description.

The summary is the heart of the proposal, and requires some care. You will need to convince me that your topic is interesting, your data are relevant, and building a model and making predictions of a quantitative outcome using the predictors available to you will be worthwhile.

- The summary ends with a statement of the research question or questions (you may have one, or possibly two.) An excellent question conveys the main objective of the study in terms that allow us to apply statistical models to describe an association between one or more predictors and a quantitative outcome. Some advice on writing a good research question is provided below, after the data set description information.
- It should be possible for me to explain your study accurately just by reading this summary. If it's not possible, it will come back to you for a REDO. Statistics is a details business. Get the details right.
- This summary should be less than 300 words.
- Use complete English sentences. Write in plain language. Use words we all know. Avoid jargon. And look at the general suggestions about writing in the Course Syllabus.

2.3.2 The Data Description

Your data description can be as long as it needs to be, although two pages is usually more than enough. It should include:

1. Your data source, which can be an online source (in which case include a working link), a published paper or journal article (in which case I need a link and a PDF copy of the paper), or unpublished data (in which case I need the details of how the data were gathered).

2. A thorough description of the data collection process, with complete details as to the nature of the variables, the setting for data collection, and complete details of any apparatus you used which may affect results.

3. Specification of the people and methods involved.

    - Who are the subjects under study?
    - When were the data gathered? By whom?
    - How many subjects are included?
    - What caused subjects to be included or excluded from the study?

4. Your planned **quantitative** outcome, which must relate directly to the research question you specified above. Provide a complete definition, including specifying the exact wording of the question or details of the measurement procedure used to obtain the outcome. If available, you can also include descriptions of secondary **quantitative** outcomes. Your outcomes must be quantitative in Study 2.

5. Your predictors of interest, which should also relate to the research question in an obvious way. Again, define the variables carefully, as you did with the outcome.

6. If you already have the data, tell me that. If you don't, specify any steps you must still take in order to get the data, and specify the date by which you will have your data (must be no later than November 1.)

2.3.3 Data Restrictions

Study 2 data sets MUST

- contain between 250 and 250,000 distinct observations,
- contain at least one quantitative outcome variable,
- contain at least four predictor variables, one of which may be identified as the "key" predictor of interest,
- include at least one quantitative predictor variable, and at least one categorical predictor variable,
- include a complete description of how the data were gathered, so that information must be publicly available,
- be in your hands no later than November 1,
- be shared with a statistician (Dr. Love) following Jeff Leek's guide to sharing data with a statistician as part of Task F. This means you need to have access to the data in the raw, and it means that I have to be able to have access to it in the raw, as well. - be capable of being fully cited for any and all data elements, including a complete codebook, as this must be provided as part of your proposal.

While there are some great resources available to some people in this class by virtue of their affiliation with one of the health systems in town, I can do nothing to get you access to health system specific data as part of your project for this class or for 432, and in general, data from those sources are not especially appropriate because of issues with protected health information.

No more than two students in the class can work on the same question for the same data. If two of you have data you would each like to work on, that may be OK, but you'll need to generate separate research questions and perform your analyses and the Project Tasks separately.

I am not interested in you using pre-cleaned data from an educational repository, such as:

- this one at the Cleveland Clinic, or this one at Vanderbilt University, or this one at UCLA, or this one at the University of Florida, or this one at Florida State University, or
- StatLib at Carnegie-Mellon University, or the Journal of Statistics Education Data Archive, or
- the data sets gathered in the fivethirtyeight package, the mosaic package, the cars package, the datasets package, or any other R package designed primary for teaching, or
- StatSci.org's repository of textbook examples and ready for teaching data, or
- any of the many textbook-linked repositories of data sets, like this one for Statistics: Unlocking the Power of Data, or
- any similar repository Professor Love deems to be inappropriate

2.3.4 Some Potentially Useful Data Sources

The ideal choice of data source for this project is a public-use version of a meaningful data set without access restrictions. With so many students in the class, I cannot be responsible for supervising your work with restricted data personally. Some appealing sources to explore include:

- the new Google Datasets Search
- https://www.data.gov/ The home of the U.S. Government's open data
- http://www.census.gov/data.html The U.S. Census Bureau has many interesting data sets, including the Current Population Survey
- http://www.healthdata.gov/ 125 years of U.S. Health Care Data
- http://www.cdc.gov/nchs/nhanes/index.htm National Health and Nutrition Examination Survey.

  - Lots of people choose to use NHANES data, and it is a great resource, but if you do use it, I will require you to look at data collected in at least two different survey forms, so that you'll have to do some merging. You may want to look at the nhanesA package in R
- http://dashboard.healthit.gov/datadashboard/data.php Office of the National Coordinator for Health IT's dashboard
- http://www.icpsr.umich.edu/icpsrweb/ ICSPR (Inter-university Consortium for Political and Social Research) is a source for many public-use data sets
  - This includes the Health and Medical Care data archive of the Robert Wood Johnson Foundation
- http://gss.norc.org/ The General Social Survey
- http://www.bls.gov/data/ Bureau of Labor Statistics
- http://nces.ed.gov/surveys/ National Center for Education Statistics
- http://www.odh.ohio.gov/healthstats/dataandstats.aspx Ohio Department of Health
- http://open.canada.ca/en Canada Open Data
- http://digital.nhs.uk/home Health data sets from the UK National Health Service.
- http://www.who.int/en/ World Health Organization
- http://www.unicef.org/statistics/ UNICEF has some available data on women and children
- http://www.pewinternet.org/datasets/ Pew Research Center's Internet Project
- http://portals.broadinstitute.org/cgi-bin/cancer/datasets.cgi Broad Institute's Cancer Program
- http://www.kdnuggets.com/datasets/index.html is a big index of lots of available data repositories
- https://www.kaggle.com/ Kaggle competition data sets are attractive to students occasionally, but I've seen a lot of them before and don't really want to see them again. Use this only as a last resort.

I cannot guarantee the quality of any of the data sets available at these sites, but I've spent at least a little time at many of them in recent months.

2.3.5 A few tips

1. Don't use hierarchical, multi-level data. That's not what we need in 431.
2. Don't use a categorical outcome variable, or plan a logistic regression model. That's for 432, not 431.
3. Don't assume Dr. Love knows anything at all about wet lab biology work or genomics.
4. If you are planning to use data you have collected, or that you are working on as part of another course or your research work, that is probably going to work out better in 432 than 431. At a minimum, you will need to be able to convince me that the data you will provide is completely free of any restrictions (after de-identification and compliance with all HIPAA and other security standards), contains NO protected information of any kind, and can be shared freely with the general public. You will need to write a statement asserting that all of this is true for me to approve your proposal.

2.3.6 If you are storing your own data

An extremely useful link for those of you **building a spreadsheet to store data** is Karl Broman's tutorial on the subject. No one was born knowing this stuff - take a look.

2.4 Research Questions for Study 2 that worked well in the past

For Study 2, your research questions will need to fit within the confines of a regression model, where a quantitative outcome is predicted using a series of at least four predictor variables. In many cases, a key predictor will be of primary interest, with other predictors serving to "adjust" away noise and generate fairer comparisons. Here are a few examples from past classes:

- "Is the presence of cardiovascular risk factors, specifically elevated hemoglobin A1c, predictive of cognitive impairment as defined by the Digit Symbol Substitution Test in patients over the age of 60 years, after adjusting for age, education, and depression?"

- "What is the effect of thyroid dysfunction on LDL level after adjusting for age, sex, and level of physical activity in the population of patients at XXXXXXXX location who are 40 years-old and above?"
- "Do conscientiousness and openness predict more conservative or liberal attitudes about government spending and whether and how much wasteful spending exists, after accounting for age, income and professional status?"
- "Does overweight or obesity (defined by body mass index) predict insulin resistance (measured by the homeostasis model assessment of insulin resistance (HOMA-IR)) in young adults with first-time acute coronary syndrome after adjusting for age, sex, race/ethnicity and severity of (several comorbid conditions)?"

2.5 Evaluating Task A

Dr. Love will evaluate all proposals (Task A) personally, in the order in which they are received. Proposals will receive one of two grades: OK or REDO. That grade will be posted to Canvas. REDO will be accompanied with specific requests in the form of a Canvas comment that should be accomplished within a short time window (approximately 24 hours). If you materially deviate from these specifications, Dr. Love will not evaluate your proposal other than to re-specify what needs to be fixed before he will respond.

- A score of OK is worth 10/10 points for Task A, once Task B is also complete.
- You (and/or your group, if there are problems with parts 1-3) must REDO the Proposal until you reach OK. Sometimes, that's more than once.

Chapter 3

Task B (Presentation Sign-Up) Instructions

3.1 Deadline and Submission information

Task B is due at noon on 2018-10-12. Submit your Task B work by completing the Google Form linked at http://bit.ly/431-2018-project-signup-taskB.

- Please note that Task A is now due on 2018-10-15 at noon, but Task B is still due on 2018-10-12.
- All students must specify a minimum of 8 time slots, on at least two different days, when they can give their presentation.
- You will also be able to specify your two favorite time slots among those you have chosen.
- The presentation dates are 2018-12-10, 2018-12-11 and 2018-12-13.
  - University classes end December 7.
  - December 10 is one of the official University Reading Days, and December 11 and 13 are Final Exam Days.
- If you have some special problem or concern or need to give your presentation before 2018-12-10, there is a space to tell Dr. Love about that at the end of the form.

3.2 Available Time Slots

There are 55 available time slots, listed below.

3.2.1 Monday 2018-12-10 morning

Time Slot	Date	Start	Finish	**Arrival Time**
1	MON 12-10	8:00 AM	8:20 AM	**7:50 AM**
2	MON 12-10	8:25 AM	8:45 AM	**8:15 AM**
3	MON 12-10	8:50 AM	9:10 AM	**8:40 AM**
4	MON 12-10	9:15 AM	9:35 AM	**9:05 AM**
5	MON 12-10	9:40 AM	10:00 AM	**9:30 AM**
6	MON 12-10	10:15 AM	10:35 AM	**10:05 AM**
7	MON 12-10	10:40 AM	11:00 AM	**10:30 AM**

Time Slot	Date	Start	Finish	**Arrival Time**
8	MON 12-10	11:05 AM	11:25 AM	**10:55 AM**
9	MON 12-10	11:30 AM	11:50 AM	**11:20 AM**

3.2.2 Monday 2018-12-10 afternoon

Time Slot	Date	Start	Finish	**Arrival Time**
10	MON 12-10	12:35 PM	12:55 PM	**12:25 PM**
11	MON 12-10	1:00 PM	1:20 PM	**12:50 PM**
12	MON 12-10	1:25 PM	1:45 PM	**1:15 PM**
13	MON 12-10	1:50 PM	2:10 PM	**1:40 PM**
14	MON 12-10	2:15 PM	2:35 PM	**2:05 PM**
15	MON 12-10	2:40 PM	3:00 PM	**2:30 PM**
16	MON 12-10	3:15 PM	3:35 PM	**3:05 PM**
17	MON 12-10	3:40 PM	4:00 PM	**3:30 PM**
18	MON 12-10	4:05 PM	4:25 PM	**3:55 PM**
19	MON 12-10	4:30 PM	4:50 PM	**4:20 PM**
20	MON 12-10	4:55 PM	5:15 PM	**4:45 PM**

3.2.3 Tuesday 2018-12-11 morning

Time Slot	Date	Start	Finish	**Arrival Time**
21	TUE 12-11	8:00 AM	8:20 AM	**7:50 AM**
22	TUE 12-11	8:25 AM	8:45 AM	**8:15 AM**
23	TUE 12-11	8:50 AM	9:10 AM	**8:40 AM**
24	TUE 12-11	9:15 AM	9:35 AM	**9:05 AM**
25	TUE 12-11	9:40 AM	10:00 AM	**9:30 AM**
26	TUE 12-11	10:15 AM	10:35 AM	**10:05 AM**
27	TUE 12-11	10:40 AM	11:00 AM	**10:30 AM**
28	TUE 12-11	11:05 AM	11:25 AM	**10:55 AM**
29	TUE 12-11	11:30 AM	11:50 AM	**11:20 AM**

3.2.4 Tuesday 2018-12-11 afternoon

Time Slot	Date	Start	Finish	**Arrival Time**
30	TUE 12-11	12:35 PM	12:55 PM	**12:25 PM**
31	TUE 12-11	1:00 PM	1:20 PM	**12:50 PM**
32	TUE 12-11	1:25 PM	1:45 PM	**1:15 PM**
33	TUE 12-11	1:50 PM	2:10 PM	**1:40 PM**
34	TUE 12-11	2:15 PM	2:35 PM	**2:05 PM**
35	TUE 12-11	2:40 PM	3:00 PM	**2:30 PM**
36	TUE 12-11	3:15 PM	3:35 PM	**3:05 PM**

3.2.5 Thursday 2018-12-13 morning

Time Slot	Date	Start	Finish	**Arrival Time**
37	THU 12-13	8:00 AM	8:20 AM	**7:50 AM**
38	THU 12-13	8:25 AM	8:45 AM	**8:15 AM**
39	THU 12-13	8:50 AM	9:10 AM	**8:40 AM**
40	THU 12-13	9:15 AM	9:35 AM	**9:05 AM**
41	THU 12-13	9:40 AM	10:00 AM	**9:30 AM**
42	THU 12-13	10:15 AM	10:35 AM	**10:05 AM**
43	THU 12-13	10:40 AM	11:00 AM	**10:30 AM**
44	THU 12-13	11:05 AM	11:25 AM	**10:55 AM**
45	THU 12-13	11:30 AM	11:50 AM	**11:20 AM**

3.2.6 Thursday 2018-12-13 afternoon

Time Slot	Date	Start	Finish	**Arrival Time**
46	THU 12-13	12:35 PM	12:55 PM	**12:25 PM**
47	THU 12-13	1:00 PM	1:20 PM	**12:50 PM**
48	THU 12-13	1:25 PM	1:45 PM	**1:15 PM**
49	THU 12-13	1:50 PM	2:10 PM	**1:40 PM**
50	THU 12-13	2:15 PM	2:35 PM	**2:05 PM**
51	THU 12-13	2:40 PM	3:00 PM	**2:30 PM**
52	THU 12-13	3:15 PM	3:35 PM	**3:05 PM**
53	THU 12-13	3:40 PM	4:00 PM	**3:30 PM**
54	THU 12-13	4:05 PM	4:25 PM	**3:55 PM**
55	THU 12-13	4:30 PM	4:50 PM	**4:20 PM**

Chapter 4

Task C (Survey Editing) Instructions

4.1 Deadline and Submission information

Task C is due at noon on 2018-10-23. Submit your Task C work via Canvas. Please note that:

- Task D is also due at the same time.
- We do not have class on 2018-10-23 because of CWRU's Fall Break.

Dr. Love will make the Draft Survey available as soon as possible after our initial work developing the Survey is complete.

4.2 Task C has two parts.

In Task C, you need to submit a single Word document (maximum one page, 12 point font, with your name and Project Task C on the top of the Word document) containing the following two things:

4.2.1 A list of corrections and clarifications to the existing items in the Draft Survey.

This should include any typographical errors, clarifications or other edits that you wish to suggest for the items included in the Draft Survey.

- If you found no errors or items in need of clarification, write a sentence saying that.
- If you did find an issue, please be sure to specify the item number where you feel a revision is needed.

An Important Note: In addition, If you see any items in the Draft Survey that you, personally, are not comfortable answering, for whatever reason, **please indicate that to us** in your response here, and we will consider revisions appropriately.

4.2.2 A list of 0-3 new items that you want us to consider adding to the Draft Survey.

- Note that your new items *can* be but do not *need* to be anything you've previously suggested.
- Please begin with the following sentence: `I would like to submit # new items for consideration.`
  - If your number of new items to suggest is zero, then you need not write anything else here. - Should you wish to have us include 1-3 additional items, remember that nothing about sex, drugs, or performance in 431 can be asked.

For each new item you propose…

1. list the complete wording of the new item, being sure to specify the type (for instance, short answer, multiple choice, or checkbox) and the set of possible responses, as you did in Task A.
2. describe (in 2-3 complete sentences per new item) your reasons to include the item.
    - Good reasons would begin with a statement of what you intend to do. As an example of such a statement, consider `I wish to study the result of this new item as a quantitative outcome across groups established by current item #*** from the survey.` Or, perhaps, something like: `I wish to use this new item as a grouping variable to study current item #***.`
    - In either case, follow your statement with a short explanation as to why your new item's result is of interest, and is not already captured by the existing survey.

We will not consider more than 3 new items from anyone, and are eager to hold the total set of new items to 20 or less, across all 51 students. I hope that data related to each of the accepted project proposals will be found in the Draft Survey, as that is certainly the intention. So if that's the case, we are unlikely to need more than a few new items.

Chapter 5

Task D (Survey Comparison Plan) Instructions

5.1 Deadline and Submission information

Task D is now due at noon on 2018-11-02. Submit your Task D work by completing the Google Form linked at http://bit.ly/431-2018-survey-comparison-plan-taskD.

Please note that:

- Task E is also due at the same time.
- Dr. Love will provide specific information on item numbers and variable names in a document made available when the Survey itself is available, no later than noon on 2018-10-29.

5.2 Task D requires you to complete a Google Form

The Google Form for Task D is now available at http://bit.ly/431-2018-survey-comparison-plan-taskD.

In this form, you will need to specify the list of items from the Draft Survey that you plan to use in one or more of your six required analyses for Study 1 (the Class Survey.)

- The Form will ask you to specify by item number and name the items you wish to use for each of the six required Analyses for Study 1.
  - **NEW** Dr. Love will provide specific information on item numbers and variable names in a document made available when the Survey itself is available, no later than noon on 2018-10-29.
- In addition to the items you select related to each Analysis, you will also select two backup quantitative variables, and two backup factors.
- You need not do any analyses connected to the items you originally suggested, nor do you need to do analyses that mirror your original research questions.
- Items with at least 10 possible responses will be treated as quantitative. Other items will be treated as categorical (factors.)
  - For ordered categories, you can consider assigning a score to each response, then treating that score as quantitative.
  - You are permitted to categorize into a group with 2-4 levels any quantitative item you choose.
  - You are permitted to collapse any categories in an item with more than 2 categories, as you choose.
  - Some items are part of multiple-item scales. If you want to use a scale, specify each item that would go into that scale on the Task D form, as applicable.

5.3 The Six Required Analyses for Study 1

- The actual analyses you will need to do include either Analysis 1a or 1b (but not both), along with Analyses 2-6, as specified below.

Analysis	Variables needed
[1a] 2 means via paired samples	Two quantitative (outcomes)
[1b] 2 means via independent samples	One quantitative (outcome) and one categorical (2 levels)
[2] ANOVA with Tukey	One quantitative (outcome) and one categorical (3-6 levels)
[3] Regression Model	Same as either [1b] or [2], plus one quantitative (covariate)
[4] 2x2 Table	Two categorical (2 level) variables
[5] JxK Table	Two categorical variables, one with 2-6, other with 3-6 levels
[6] 2x2xJ Table	Same as [4], plus one categorical with 3-6 levels

5.3.1 Analysis 1: Comparing the Means of Two Populations

Here, you will choose either to use a paired samples design or an independent samples design.

If you're using paired samples (to do Analysis 1a), then you will specify

- Outcome A, a quantitative variable, and
- Outcome B, also a quantitative variable.

If you're using **independent** samples (to do Analysis 1b), then you will specify

- Outcome C, a quantitative variable, and
- Factor Z, a two-level categorical variable

Each level of your Factor Z must apply to a minimum of 10 subjects in the Survey.

5.3.2 Analysis 2: Comparing the Means of Three or More Populations

Here, you will complete an analysis of variance, with pre-planned Tukey HSD comparisons. You will specify:

- Outcome D, a quantitative variable (which can repeat A, B or C from before if you like)
- Factor Y, a 3-6 level categorical variable

Each level of your Factor Y must apply to a minimum of 6 subjects in the Survey.

5.3.3 Analysis 3: Regression Model with One Covariate

Here, you will use the same outcome and factor as you used in either Analysis 1 (if you used independent samples) or Analysis 2, but add a new covariate. So you will specify:

- Outcome E, which must be the same as either your Outcome C (if you're amplifying Analysis 1b) or Outcome D (if you're amplifying Analysis 2).
- Factor X, which must be the same as either Factor Z (if you're amplifying Analysis 1b) or Factor Y (if you're amplifying Analysis 2), but now you're also adding:
- Covariate G, which is to be a quantitative variable not used in Analyses 1 or 2.

5.3.4 Analysis 4: Comparing Two Population Proportions

Here, you will develop and analyze a 2x2 contingency table. You will specify:

- Factor L (which needs to have exactly 2 levels) and will be in the rows of your table, and
- Factor M (which also needs to have exactly 2 levels) and be in the columns.

Every cell in your 2x2 table needs to have at least 5 observations. You are welcome to re-use a two-level factor you've used in a previous Analysis for L or M, but must add a new factor for the other.

5.3.5 Analysis 5: A Larger Two-Way Table

Here, you will develop a contingency table analysis to describe (in the rows) a factor with 2-6 levels, and (in the columns) a factor with 3-6 levels. You will specify:

- Factor J (which must have 2-6 levels) and will be in the rows of this table, and
- Factor K (which must have 3-6 levels) and will be in the columns

You can re-use at most one of Factors L and M as Factor J, but Factor K must be new.

5.3.6 Analysis 6: Comparing Population Proportions in a 2x2xN contingency table

Here, you will amplify Analysis 4 by developing a Mantel-Haenszel analysis of a contingency table with 2 rows (re-using Factor L from Analysis 4), 2 columns (re-using Factor M from Analysis 4) and 3-6 layers (or strata) in a new factor called Factor N. You will specify:

- Factor N (which must have 3-6 levels) and can repeat Factor J or K if you like, so long as it is different from Factors L or M.

5.3.7 Backups

You will also specify

- Quantitative Variable Q: a backup quantitative variable for use as an outcome
- Quantitative Variable R: another backup quantitative variable, just in case
- Factor Variable S: a backup 2 level categorical variable for use as a group
- Factor Variable T: a backup 3-6 level categorical variable for use as a group, and

These backups are for whether the results of the Survey turn out to yield either outcome variables with no variation at all across the groups of interest (in Analyses 1-3) or tables with insufficiently populated cells (in Analysis 4-6).

5.4 Table of What You'll Specify on the Form

You will specify the following elements on your form (remember that you will either specify A and B if you choose Analysis 1a, or C and Z if you choose Analysis 1b.)

Analysis	Description	Item #	Item Name
1a	A (quantitative)	–	–
1a	B (quantitative)	–	–
1b	C (quantitative)	–	–
1b	Z (two-category)	–	–
2	D (quantitative)	–	–

Analysis	Description	Item #	Item Name
2	Y (3-6 category)	–	–
3	E (quantitative, same as C or D)	–	–
3	X (same as either Y or Z)	–	–
3	G (quantitative, not A/B/C/D)	–	–
4 & 6	L (two-category)	–	–
4 & 6	M (two-category)	–	–
5	J (2-6 category)	–	–
5	K (3-6 category)	–	–
6	N (3-6 category)	–	–
Backup	Q (quantitative)	–	–
Backup	R (quantitative)	–	–
Backup	S (two-category)	–	–
Backup	T (3-6 category)	–	–

Chapter 6

Task E (Taking the Survey) Instructions

6.1 Deadline and Submission information

Task E is due at noon on 2018-11-02. Submit your answers to the course survey via the Google Form linked at http://bit.ly/431-2018-survey. That link will go live no later than noon on 2018-10-29.

- The final item asks for your name, and the system is collecting your email address (you must be logged into Google via CWRU). These will be pruned from the survey before data sets are created.
  - You should answer all of the items. Please don't skip any items you can answer. Your colleagues need data.
  - If you want to save your work and return later, note that only the *first* item in each section of the survey must be completed for Google to let you submit your work. Once you've submitted a partially completed survey, you can return as often as you like before the deadline to finish up.
- Note that Project Task D is due at the same time.

6.2 Receiving Your Study 1 Data (early November)

Once the Survey is complete, we will post **multiple** data files, each containing some of the variables you need.

- You will need to download both files, and then *combine* and tidy to suit your needs. Combining data sets like this is a skill you'll need to master to successfully complete the Project.
- The files will be linked by the subject `id` number.

Chapter 7

Task F (Sharing Study 2 Data) Instructions

7.1 Deadline and Submission information

Task F is due at noon on the Wednesday after Thanksgiving: 2018-11-28. Submit your Task F work through Canvas.

7.2 Sharing Your Data Appropriately

Task F requires you to share your data for Study 2. The model for this Task is Jeff Leek's Guide to Data Sharing, which you should definitely read.

Specifically, you will submit the following via Canvas on time.

1. a direct link to the raw data set (without any need for me to sign up for anything) or a .csv copy of the raw data set (containing only the variables you plan to use in your actual project work) which you should call `yourname-raw.csv`
2. a single tidy .csv file with a name of your choice containing a clean, tidy data set for Study 2, along with
3. a Word or PDF file containing both
    a. a **codebook** section which describes every variable (column) and its values in your .csv file,
    b. a **study design** section which reminds (and updates) us about the source of the data and your research question.

If you must zip the raw data set, OK, but include the tidy CSV and Word/PDF files as separate documents, apart from the zip file you submit. The details for these elements of Task F follow.

7.3 The Raw Data Set

You need to show me the raw, de-identified data. What does this mean?

- The data set should be as you "received" it, **other** than the following:
    - The "raw" data set must be completely de-identified, likely by removing any columns containing identifiable information.
    - The "raw" data you post should include no protected health information, nor should it include anything you are not 100% sure you can share with Dr. Love.

- The "raw" data set must include a unique id code for each subject, and that can be generated by you if the original data contained identifiable information within its id codes.
- The "raw" data set may in fact be multiple data sets, which you will merge together to form your eventual, tidy, analytic data set.
- The "raw" data set must include all variables you will or might use in your project analyses, but you are permitted to delete any variables (columns) that you are 100% sure will NOT be used in your analyses. Note that you should include any variables that you haven't made a final decision on.
- The "raw" data set should indicate all missing values as they were originally provided to you. Do not impute missing values in the "raw" data set.
- You should not summarize the raw data in any way, nor should you delete any rows.

A direct link (without me having to sign up for anything) to an appropriate raw data file(s) is preferred, if possible. If this is not possible, then describe the original source(s) of the data carefully, and send instead a .csv file or (files) of the raw data set, called `yourname-raw.csv`[1]

7.4 The Tidy Data Set

Your tidy .csv file should include only those variables you will actually use in your analysis of Study 2. Your .csv file should include one row per subject in your data, and one column for each variable you will use. Your data are tidy if each variable you measure is in its own column, and each different observation of that variable is in its own row, identifed by the subject identification code in the left-most column, which you might call `Subj_ID` if that's helpful.

You need to provide:

1. a header row (row 1 in the spreadsheet) that contains full row names. So if you measured age at diagnosis for patients, you would head that column with the name `AgeAtDiagnosis` or `Age_at_Diagnosis` instead of something like `ADx` or another abbreviation that may be hard for another person (or you, two years from now) to understand.
2. a study identification number (I would call this variable `Subj_ID` and use consecutive integers to represent the rows in your data set) which should be the left-most variable in your tidy data.
3. a quantitative outcome with a meaningful name using no special characters other than an underscore (`_`) used to separate words, which should be the second variable in your data.
   - If you have any missing **outcome** values, **delete those rows** entirely from your tidy data set before submitting it.
4. at least four predictor variables, each with a meaningful name using no special characters other than `_` to separate words, and the predictors should be shown in columns to the right of the outcome.
   - *Continuous* variables are anything measured on a quantitative scale that could be any fractional number.
   - *Ordinal categorical* data are data that have a fixed, small (< 100) number of levels but are ordered.
   - *Nominal categorical* data are data where there are multiple categories, but they aren't ordered.
   - Categorical predictors should read into R as factors, so your categories should include letters, and not just numbers. In general, try to avoid coding nominal or ordinal categorical variables as numbers.
   - Label your categorical predictors in the way you plan to use them in your analyses.
   - *Missing data* are data that are missing and you don't know the mechanism. Missing data in the predictor variables are allowed, and you should code missing values in your tidy data set as `NA`. It is critical to report if there is a reason you know about that some of the data are missing.
   - Note that you should **not** impute any data in Project Task F. Instead, you will impute as part of your analysis and demonstrate that in Tasks G and H, as necessary.

[1] If you have to send more than one "raw" .csv data set, append numbers after each name, so you'd submit `yourname-raw1.csv`, `yourname-raw2.csv` etc.

5. any other variables you need to share with me (typically this would only include things you had to use in order to get to your final choice of outcome and predictors.) Most people will not need to share any additional variables.

I will need to be able to take your submitted tidy `.csv` file and run your eventual Markdown file (part of Task G) against it and obtain your results, so it must be completely clean. Because it is a `.csv` file, you'll have no highlighting or bolding or any other special formatting. If you have missing values, they should be indicated as `NA` in the file. If you obtain the file in R, and then write it to a .csv file, you should write the file without row numbers if you already have an identification variable. To do so, you should be able to use `write.csv(dataframeinR, "newfilename.csv", row.names = FALSE)` where you will substitute in the name of your data frame in R, and new (.csv) file name. Don't use the same name for your original data set and your tidy one.

Note If your "tidy" `.csv` file contains more than 6,000 rows after any rows with missing outcomes have been deleted, and after applying any necessary inclusion/exclusion filters, then you will be sampling from that file (to create samples of 4,000 for the "training set" and 2,000 for the "test set") as part of your analyses in Task G), but the tidy `.csv` should contain all rows in the data, up to the maximum permitted of 250,000.

7.5 The Codebook

For almost any data set, the measurements you calculate will need to be described in more detail than you will sneak into the spreadsheet. The code book contains this information. At minimum it should contain:

1. Information about the variables (including units and codes for any categorical variables) in your tidy data set
2. Information about the summary choices or transformations you made or the development of any scales from raw data

By reading the codebook, I should understand what you did to get from the raw data to your tidy data, so add any additional information you need to provide to make that clear.

7.6 The Study Design

Here is where I want you to put the information about the experimental study design you used. You can and should reuse (and edit) the information you provided as part of the Proposal in this Codebook. The material you need here consists of three parts from the proposal, updated to mirror your current plan. Specifically, you should provide:

1. Your research question describes your outcome, your key predictor and other predictors, and the population of interest. It is probably easiest to follow one of these formats[2].

- What is the effect of `*your key predictor*` on `*your outcome*` adjusting for `*your list of other predictors*` in `*your population of subjects*`?
- How effectively can `*specify your predictors*` predict `*your outcome*` in `*your population of subjects*`? or

2. A thorough description of the data collection process, with complete details as to the nature of the variables, the setting for data collection, and complete details of any apparatus you used which may affect results that **has not already been covered** in the codebook materials.
3. Specification of the subjects and methods involved.
    a. Who are the subjects under study? How many are included in your final tidy data set?
    b. When were the data gathered? By whom?
    c. What caused subjects to be included or excluded from the study?

[2]You are welcome to move the clauses around to make for a clearer question.

Chapter 8

Task G (The Portfolio) Instructions

Task G requires you to provide a written portfolio of materials, which you will also make use of in your final presentation.

The Portfolio is due at noon on 2018-12-13, via Canvas.

8.1 What Will You Submit?

There are three things you need to submit for Task G.

1. [`.csv` file] A clean, tidy data set for Study 1, which will require combining the two data sets you are provided, dealing with any missing data and any necessary combination into scales on the variables in which you are interested.
2. [`.Rmd` and `.html` files] Your document describing the six required analyses for Study 1 (as listed below), as both a Markdown file and HTML result that work with the clean and tidy data set for Study 1.
3. [`.Rmd` and `.html` files] Your document describing the nine required analyses for Study 2 (as listed below), as both a Markdown file and HTML result that work with the clean and tidy data set you submitted in Task F.

Do not submit zipped files.

8.2 Setting Up Your R Markdown Files

You will prepare separate R Markdown files for Study 1 and Study 2. In each, we would like you to include:

- a table of contents
- where each section and subsection is numbered
- code-folding, so we can hide code if we like in the HTML, but see it by default.

Do not hide any code. Do not print out anything larger than a tibble summary. Make sure that the final product (HTML) looks terrific, and that the R Markdown file contains more explanation than code, and that explanation is written in complete sentences.

You are welcome to use any R Markdown formatting approach (for instance, tabs, or a floating table of contents, or other flourishes) that accomplishes the objectives specified above, but we do want to see the results in HTML, rather than any other format.

8.3 The Six Required Analyses for Study 1

The initial work for your portfolio will include all of the code you used to merge the data sets provided to you for Study 1, then select the variables you'll actually use in your analyses, and then clean up and manage any remaining issues within those variables in your data. Following that work, the required analyses for the Project Survey that need to be in your Portfolio are:

1. A two-group comparison of population means (could use paired or independent samples)
2. An analysis of variance with Tukey HSD pairwise comparisons of population means across K subgroups, where $3 \leq K < 7$
3. A regression model to amplify the indepedent samples comparison in a or b by incorporating a quantitative covariate.
4. A 2x2 Table and resulting analyses for comparison of two population proportions in terms of relative risk, odds ratio and probability difference
5. A two-way JxK contingency table where $2 \leq J < 7$ and $3 \leq K < 7$ with an appropriate chi-square test
6. A three way 2 x 2 x J contingency table analysis whch will expand your 2x2 table from #4 and where $3 \leq J < 7$

Each analysis should be self-contained (so that I don't have to read Analysis 1 first to understand Analysis 3, for example). Present each new analysis as a subsection with an appropriate heading in the table of contents, so we can move to a new analysis efficiently. Each analysis should begin with a paragraph explaining what you are doing, specifying the items being used, and how you are using them, and then conclude with a paragraph of discussion of the key conclusions you draw from your detailed analyses, and a discussion of any limitations you can describe that apply to the results.

Missing Data: If you have missing data on any of the variables you study in Project Study 1, then you will need to do simple imputation as part of the associated Study 1 analyses. In that case, you should show the imputation process in your code, and describe explicitly any choices you made, then run the necessary analysis on both the imputed sample and the sample using complete cases alone, and compare the two results.

You should generally follow the comparison plan you outlined in Task D. If you need to make a change, please indicate that in the text setting up each analysis.

A demonstration of an appropriate analysis for each of the required pieces of Study 1 Analysis will be provided to you at https://github.com/THOMASELOVE/431-2018-project/tree/master/demo_study1.

8.4 The Nine Required Steps for Study 2

For your portfolio presentation in Study 2 (Your Data) complete these steps:

1. Identify all the variables in your tidy data set that have missing (NA) values. Delete all observations with missing outcomes (actually, this much you should have done before submitting the tidy data in Task F), and now use simple imputation to impute values for the candidate predictors with `NA`s. Use the resulting imputed data set in all subsequent work. Be sure to describe any choices you make in building your imputed data set.
    - **Note** If your data set contains more than 6,000 rows after any rows with missing outcomes have been deleted, and after applying any necessary inclusion/exclusion filters, then you will be sampling from that file[1] (to create a sample of 6,000 observations - which you will soon divide into samples of 4,000 for the "training set" and 2,000 for the "test set") here.
2. Obtain a training sample with a randomly selected 67-80% of your data[2], and have the remaining 20-33% in a test sample, properly labeled, and using `set.seed` so that the results can be replicated later. Use this training sample for Steps 3-7 below.

[1] The usual choice is to select a random sample of 6,000 observations, without replacement.

[2] The training sample should be 67% of the data if you have 6,000 rows. If you have 250 rows, 80% of the data should be in the training sample. Otherwise, anything in the range of 67-80% is OK.

3. Using the training sample, provide numerical summaries of each predictor variable and the outcome, as well as graphical summaries of the outcome variable. Your results should now show no missing values in any variable. Are there any evident problems, such as substantial skew in the outcome variable?
4. Build and interpret a scatterplot matrix to describe the associations (both numerically and graphically) between the outcome and all predictors. Use a Box-Cox plot to investigate whether a transformation of your outcome is suggested. Describe what a correlation matrix suggests about collinearity between candidate predictors.
5. Specify a "kitchen sink" linear regression model to describe the relationship between your outcome (potentially after transformation) and the main effects of each of your predictors. Assess the overall effectiveness, within your training sample, of your model, by specifying and interpreting the R^2, adjusted R^2 (especially in light of your collinearity conclusions below), the residual standard error, and the ANOVA F test. Does collinearity in the kitchen sink model have a meaningful impact? How can you tell? Specify the size, magnitude and meaning of all coefficients, and identify appropriate conclusions regarding effect sizes with 90% confidence intervals.
6. Build a second linear regression model using a subset of your four predictors, chosen by you to maximize predictive value within your training sample. Specify the method you used to obtain this new model. (Backwards stepwise elimination is a likely approach in many cases, but if that doesn't produce a new model, feel free to select two of your more interesting predictors from the kitchen sink model and run that as a new model.)
7. Compare this new (second) model to your "kitchen sink" model within your training sample using adjusted R^2, the residual standard error, AIC and BIC. Specify the complete regression equation in both models, based on the training sample. Which model appears better in these comparisons of the four summaries listed above? Produce a table to summarize your results. Does one model "win" each competition in the training sample?
8. Now, use your two regression models to predict the value of your outcome using the predictor values you observe in the test sample. Be sure to back-transform the predictions to the original units if you wound up fitting a model to a transformed outcome. Compare the two models in terms of mean squared prediction error and mean absolute prediction error in a Table, which Dr. Love will **definitely want to see** in your portfolio. Which model appears better at out-of-sample prediction according to these comparisons, and how do you know?
9. Select the better of your two models (based on the results you obtain in Steps 7 and 8) and apply it to the entire data set. Do the coefficients or summaries the model show any important changes when applied to the entire data set, and not just the training set? Plot residuals against fitted values, and also a Normal probability plot of the residuals, each of which Dr. Love **will be looking for** in your portfolio. What do you conclude about the validity of standard regression assumptions for your final model based on these two plots?

In the Study 2 work, each step should begin with at least one complete sentence explaining what you are doing, specifying the variables being used, and how you are using them, and then conclude with at least one complete sentence of discussion of the key conclusions you draw from the current step, and a discussion of any limitations you can describe that apply to the results. Present each new step as a subsection with an appropriate heading that shows up in the table of contents, so we can move to a new step efficiently in reviewing your work.

A demonstration of an appropriate analysis for each of the required steps of Study 2 Analysis will be provided to you at https://github.com/THOMASELOVE/431-2018-project/tree/master/demo_study2.

Chapter 9

Task H (Your Presentation) Instructions

Your Presentation will be held on 2018-12-10, 2018-12-11 or 2018-12-13 in Dr. Love's office (Wood WG 82-L), as determined by the Scheduling Process in Task B.

The Schedule of Project Presentations is now available at http://bit.ly/431-2018-project-schedule.

9.1 Logistics

Arrive at Dr. Love's office (Wood WG-82L on the ground floor of the Wood building at the School of Medicine) at your arrival time, which is ten minutes before your starting time. The complete schedule will be posted after the scheduling process (Task B) is complete in October.

If the door is open, please be sure that Dr. Love knows you are there. He likes to get ahead of the schedule whenever possible. If the door is closed, wait nearby so that you can hear the door when it opens and then present yourself.

You will give your final presentation in a 20-minute meeting with Dr. Love. This will involve materials from both of your studies, in a fairly regimented way, described below.

- You will need to bring a functioning laptop which you can use to show me the key results as you describe them for each of the analyses in Study 1 and in Study 2 that you wind up discussing.
- You are welcome to show me results in the context of a Powerpoint-style presentation, if you prefer to develop one, or to show me results straight from your Markdown-created HTML files in your portfolio. Whatever works for you - so long as I can see what you are talking about as you are talking, we'll be fine.
- The computers in my office will be busy while we are meeting, so I will NOT be able to pull up your portfolio or data while we are talking. You will have to be able to do that.
- It is 100% appropriate for you to ask questions before the presentation, of Dr. Love or the TAs. Please do.
- At the presentation, there will be a little time for Dr. Love to address any lingering questions, and he's eager to hear your questions at that time, too.

If you have an emergency on the day of your presentation, email Dr. Love as soon as possible.

9.2 Study 1 Presentation (6-8 minutes, total)

In Study 1, you will first select your most interesting / intriguing result out of your six main analyses and present that, in about 2 minutes. In those 2 minutes, you should be showing me the highlights of that Analysis, specifically:

a. What question were you investigating?
b. What conclusion did you draw about that question?
c. What statistical method led you to that conclusion?

I will then ask you to present the results of one of the other five main analyses, in a similar way. You will need to come prepared to present this information for any of your six Study 1 analyses at a moment's notice, as you will not know in advance which of your other five main analyses I will ask for.

9.3 Study 2 Presentation (10-12 minutes, total)

In Study 2, you will start with telling me about the most important finding of your little study in four minutes. In these 4 minutes, you will tell me:

a. What your research question was
b. Why it was interesting to you (parts 1 and 2 combined should take no more than 30 seconds)
c. What your better model has to say about the answer to your research question
    - This should include a description of the predictors that wound up in your (final) model and the direction of each of their effects on your outcome. Show me the model as you're telling me about this.
    - This should also include a sense of how well the model predicted overall (R^2 is one good choice)
    - This should also include how well the residual plots for your final model fit regression assumptions. Show me the plots as you're telling me about this.
d. Your conclusions about rational next steps to learn more from these data, or what specific new data you now wish you'd had when you started the study.

For most of the remaining time, I will ask you about your study, and try to help you think through any problems you had in obtaining or interpreting analyses. You should come prepared to share any of the nine steps of your analysis at a moment's notice, as we may want to look at any part of your work.

9.4 Final Questions (about 2 minutes)

Depending on time, I may ask you any of several questions at the end of our meeting. Some possibilities you should be prepared for…

- What percentage of your time in Study 2 did you spend obtaining, cleaning, merging and tidying data, as opposed to actually performing analyses on tidy data?
- Tell me something useful that you learned from doing the project.
- Tell me what the hardest part of doing the project was.
- What do you know now that you wished you'd known back when you were formulating the proposal, way back in Task A? What would you tell yourself if you could go back in time?