You will undertake an advanced data analysis project on a topic of your choice. The project is an opportunity to show off what you've learned about advanced data analysis with a focus on predictive modeling. It is a major component of the class, and successful completion is required to pass. Your final output for this project will be: 1) a technical report written in R Markdown; and 2) an oral presentation delivered to the class. The project must include at least one model and your interpretations of your findings. You may include other deliverables as appropriate.

Group Formation

You will work in a group of two students. Groups have been randomly assigned, assuring that at least one member of each group has had (or is currently in) Stat 231: Data Science. They are the same groups as those for the Shiny assignment.

Assignment

Your task is to use data to tell us something interesting. This project is deliberately open-ended to allow you to fully explore your creativity. There are three main rules that must be followed:

1. Your project must be centered around data. Preferably, you will work with large, complex, and/or messy data. The more challenging your data set is, the better. Two examples of things that will make your project more ambitious, and more interesting are:

   - Combining two or more data sets that are not obviously related. For example, an analysis of tweets predicting elections paper would need to combine data from Twitter that had nothing to do with politics with a data set of congressional districts, and both data sets had to be spatially-aware, since the merging was done according to the geographical location of the tweet.
   - Accessing a live data stream. There are many APIs on the Internet that allow you to access data that changes on a continual basis. This is in contrast to static data that was collected at some point in time and does not change. If your project can respond dynamically to a live data stream, then this will make it more interesting.

   Please note that neither of the above stipulations are requirements – they are merely examples of two considerations that may make your data more challenging to work with – and thus your project more ambitious.

2. Your project must tell us something! And it must include a model. Maybe you'd like to consider *data art* projects like Memo Akten's *Forms*, which involve no statistical analysis, and figure out a way to add in a model. On the other extreme are *data mining* projects like the KDD cup, which involve little visualization, but lots of modeling. Your project can be anywhere on this spectrum, but expectations may be different depending on where you are on the scale. An example of a project that doesn't tell us anything, would be something that downloads a single data source and summarizes it, with some perfunctory visualization. Make sure that your project is thought-provoking and has some underlying meaning!

3. Stronger submissions will explore topics beyond those we've seen in class. This means you will need to do some research to learn new topics. Examples of topics from beyond class include:

   - Use of a new statistical technique (e.g. hierarchical models, a different GLM, etc.)
   - Venture into a new statistical area (e.g., causal inference, spatial data analysis, covariance estimation, missing data imputation, network analysis)

- Exploring statistical computing (e.g., implement a Bayesian model with serious computational considerations)
- Use a new live data source or API (e.g. Flickr, Facebook) [Note: just downloading a new data set in a flat file does *not* count!] or use a new data format (e.g. XML, HDF5)
- Explore new data visualizations (e.g. interactive maps, heatmaps of covariance matrices, networks) (even more ambitious in a Shiny app!)

Also, domain knowledge does not count here, so your outside knowledge of a particular topic makes for a great motivation, but doesn't count as something "beyond". If you have seen a topic in an elective course, but your partner has not, you can explore it, but bear in mind you should be applying methods at a 400-level, not a 200 or 300-level.

General Rules You may discuss your project with other students, but each pair of you will have a different topic, so there is a limit to how much you can help each other. You may consult other sources, indeed, you will probably need to do some research, but of course you should credit these sources and the data set source(s) in your report. Feel free to consult with me.

1 Components

1.1 Proposals

Your initial proposal is due via your Group Github repo by the end of the day on Tuesday, October 29th. The final proposal will be due about a week later (midnight on Monday, November 4th).

Prior to writing your initial proposal you should get together with your assigned partner and brainstorm a half dozen or so ideas before coalescing around one idea. Be ambitious! Once you decide on a topic that interests both of you, think about what you would like to end up with as a final result — without worrying about how to get there. Try to visualize what your end product will look like. Will it be an interactive map? A mobile application? What type of model will you be including? Don't think about coding, or a particular data set, or what you know how to do now. If you come up with something ambitious and original, you'll be more motivated to learn new things as you go in order to accomplish your goal.

Content Your initial and revised proposals should contain the following content:

1. Team Name: Come up with a team name for your group.
2. Title: The title of your project
3. Purpose: Describe the general topic/phenomenon you want to explore, as well some carefully considered questions that you hope to address. You should make an argument *motivating* your work. Why should someone be interested in what you are doing? What do you hope people will learn from your project?
4. Data: As best you can, describe where you will find your data, and what *kind* of data it is. Will you be working with spatial data in shapefiles? Will you be accessing an API to a live data source? Be as specific as you can, listing URLs and file formats if possible.
5. Variables: List, and briefly describe, each variable that you plan to incorporate. If you can, be specific about units, scale, etc.
6. Model: What response variable will you be aiming to predict? What model are you thinking about using? Note that the model doesn't have to be at the center of the project, but you need to have one.

7. End Product: Describe what you hope to deliver as a final product. Will it be a Shiny application that will be posted on the Internet (dynamic visualizations are strongly encouraged)? Will it be a GoogleMaps mash-up? Will it be a paper that draws some statistical conclusions and uses the model to make some predictions? Will it be a combination of these or other deliverables?

An example template is provided that you can fill in.

1.2 Update I

Please respond to the following three questions via email by midnight on **Thursday, November 21st**.

1. Have you already collected and ingested all of the data that you will need in order to complete your project? If not, please estimate the percentage of the data that you have, describe any issues that you are having, and what your plan is for getting the rest of the data.
2. Do you have any questions for me?

1.3 Update II

Please respond to the following three questions via email by midnight on **Tuesday, December 3rd**.

1. What is the single biggest unresolved issue you are having? Please describe it briefly, and what your plan is for resolving this issue.
2. How far along are you on the presentation?
3. Do you have any questions for me?

1.4 Presentation

An effective oral presentation is an integral part of this project. One of the objectives of this class is to give you experience conveying the results of a technical investigation to a non-technical audience in a way that they can understand. Whether you choose to stay in academia or pursue a career in industry, the ability to communicate clearly is of paramount importance. The burden of proof is on you to convince your audience that what you are saying is true. If your audience (who may very well be less knowledgeable about statistics than you are) cannot understand your results or their interpretations, then the technical merit of your project is irrelevant.

On either Monday December 9th or Tuesday December 10th (determined randomly later), your group will make an 10-12 minute oral presentation to the class, with an additional 2 minutes for questions. You should make (good) slides. Your goal should be to convey to your audience a clear understanding of your research topic, along with a basic understanding of your project, and how well it addresses the research question you posed. You should **not** tell us everything that you did, or show a bunch of things that you tried that didn't work well. After hearing your talk, each student in the class should be able to answer:

1. What was your project about?
2. What was your data like, and what techniques did you apply to it?
3. What model was included?
4. What were your findings?

You should prepare electronic slides for your talk. PowerPoint is fine, but you might also want to consider Google Presentation, Beamer (LaTeX), or alternative, non-linear presentation software like Prezi. Demonstrating a Shiny app is appropriate if you made one for the project. Use your creativity! One thing

you should *not* do is walk us through your calculations in RStudio. If your project has any interactive elements, please demonstrate them for us!

You will need to **submit your slides via Github to the Group repo the evening *before* your presentation**. You should also bring the slides on a flash drive as a backup, or on your personal machine.

Advice There are many sources of advice for how to make a good presentation, but an excellent place to start is:

`http://techspeaking.denison.edu/`

Watch the videos on this site to identify some common mistakes. Here is some general advice:

- Budget your time. You have at most 12 minutes. If your talk runs too short or too long, it makes you seem unprepared. Rehearse your talk ahead of time (with your group) several times in order to get a better feel for your timing. Note also that you may have a tendency to talk faster during your actual talk than you will during your rehearsal. Talking faster in order to speed up is not a good strategy—you are much better off simply cutting material ahead of time. You will probably have a hard time getting through 18 slides in 12 minutes.
- Don't write too much on each slide. You don't want people to have to read your slides, because if the audience is reading your slides, then they aren't listening to you. You want your slides to provide visual cues to the points that you are making—not substitute for your spoken words. Concentrate on graphical displays and bullet-pointed lists of ideas.
- Put your problem in context. Remember that most of your audience will have little or no knowledge of your subject matter. The easiest way to lose people is to dive right into technical details that require prior domain knowledge. Spend a few minutes at the beginning of your talk introducing your audience to the most basic aspects of your topic and present some motivation for what you are studying.
- Speak loudly and clearly. Remember that you know more about your topic that anyone else in the room, so speak and act with confidence!
- Tell a story—not necessarily the whole story. It is unrealistic to expect that you can tell your audience everything that you know about your topic in 10 minutes. You should strive to convey the big ideas in a clear fashion, but not dwell on the details. Your talk will be successful if your audience is able to walk away with an understanding of what your research question was, how you addressed it, and what the implications of your findings are.

1.5 Report

Your final report and any other deliverables will be due by 5 pm on Wednesday, December 11th.

In your report, you should tell your audience about your project, why they should care about it, and what you have discovered. Your audience will be people like you—current or aspiring statisticians / data scientists. Keep in mind that this audience is extraordinarily diverse in terms of skills and abilities, so you should assume very little about what they might know. However, your audience is reasonably tech-savvy, so you need not "dumb-down" your analysis.

Your report should make it clear to me and any other student in the class what methods and techniques you have used to produce your finished product.

Content You do not need to present *all* of the R code that you wrote throughout the process of working on this project. However, the write-up should contain the *minimal* set of R code that is necessary to generate your data, carry out your wrangling, fit various models, and understand your results and findings. If you make a claim, it *must* be justified by explicit calculation. A knowledgeable reviewer should be able to compile your `.Rmd` file without modification (assuming necessary packages are installed), and verify every statement

that you have made. All of the R code necessary to produce your figures and tables *must* appear in the write-up.

Motivation Be sure to motivate your topic at the beginning of your write-up. You should try to hook the reader early on. Assume that your audience is a skeptical analyst who has stumbled across your blog but has very little time to read it. Can you give her a reason to continue reading? A cool visualization or result can help.

Format You don't need to follow a specific format in the write-up, but you should start with an introductory paragraph and finish with a conclusion. These paragraphs need not follow the formal writing style that you would use in most other classes. Here, a colloquial style that is accessible to a lay reader is appropriate.

Nevertheless, your write-up should address the following questions:

1. Why should anyone care about this?
2. What is this about? Do *not* assume that your readers have any domain knowledge! The burden of explanation as to what you are talking about is on you! For example, if your project involves phylogenetic trees, do not assume that your audience has anything other than a basic, lay understanding of genetics.
3. Where did your data come from? What kind of data was it? Is there a link to the data or some other way for the reader to follow up on your work?
4. What are your findings? What kind of statistical computations (if any) have you done to support those conclusions? Again, while the R code will show you performing the calculation, it is up to you to interpret, in English sentences, the results of these calculations. Do not forget about units, axis labels, etc.
5. What are the limitations of your work? Be clear so that others do not misinterpret your findings. To what population do your results apply? Do they generalize? Could your work be extended with more data or computational power or time to analyze? How could your study be improved? Suggesting plausible extensions don't weaken your work – they strengthen it by connecting it to future work.

Style The Markdown format is designed to be an interactive document (not dissimilar to a blog entry). Take advantage of this by including hyperlinks, figures, videos, etc. to provide context for the reader. You will need to include your references, so don't forget them! Use Markdown elements like links, lists, LaTeX, and images as needed.

Visualizations, particularly interactive ones, will be well-received. That said, do not overuse visualizations. You may be better off with one complicated but well-crafted visualization as opposed to many quick-and-dirty plots. Any plots should be well-thought out, properly labeled, informative, and visually appealing!

The R code is there to support the technical reader who wishes to dig into your work – not to substitute for written explanation. Do *not* present long unbroken chunks of R code without offering written explanations. It is rarely necessary to have more than three or four lines of R code in a single chunk.

Formatting Option I strongly recommend that you use the Statistics Comprehensive Evaluation template as a template for your write-up. This will give you practice with the template (and you can modify the title page that says it's a comps submission). I'm happy to help with this and we have other resources available that will be presented as the semester continues. You don't need to worry about this as we get started though.

1.6 Peer and Self Evaluation

It is important to reflect on group dynamics and acknowledge the contributions of all group members. A peer and self evaluation will be completed towards the end of the project time frame. This will likely be a Google form, and be due during finals week, around the time of the final exam due date. Information will be posted on Moodle and discussed in class. You should not feel pressured to give each other high marks just because you were in the same group. If you run into issues with disparate participation, please notify me immediately so I can assist.

2 Assessment Criteria

Your project will be evaluated based on the following criteria:

- Originality/Interest: Is the topic original, interesting, and substantial – or is it trite, pedantic, and trivial? How much creativity, initiative, and ambition did the group demonstrate? Is the basic question driving the project worth investigating, or is it obviously answerable without a data-based study?
- Degree of Difficulty: How challenging was the project? Were the data particularly large, complex, and/or messy? Did the data come in an obscure format? Was a scraper or API necessary to acquire the data? Was a challenging visualization or applet constructed? Were any elements from outside the coursework necessary to complete the project?
- Design: How well were the graphical elements of the project designed? Were they clunky or elegant? Was a truly original view of the data presented? Were any interactive elements usable?
- Meaning/Analysis: Did we learn anything meaningful from this project? Are the chosen analyses (including the model) appropriate for the variables/relationships under investigation, and are the assumptions underlying these analyses met? Are the analyses carried out correctly? Did the group make appropriate conclusions from the analyses, and are these conclusions justified?
- Report: How effectively does the write-up communicate the goals, procedures, and results of the study? Are the claims adequately supported? Does the writing style enhance what the group is trying to communicate? How well is it edited? Are the statistical claims justified? Are text and analyses effectively interwoven? Clear writing, correct spelling, and good grammar are important.
- Oral Presentation: How effectively does the oral presentation communicate the goals, procedures, and results of the study? Do the slides help to illustrate the points being made by the speaker without distracting the audience? Do the presenters seem to be well-rehearsed? Did they properly budget their time? Do they appear to be confident in what they are saying? Are their arguments persuasive?