**Data Science:**

**Methods for Data Analysis**

**Individual Project Guidelines and Overview**

This class requires you to complete an individual project. This can take many forms, but it should start with data acquisition, involve some data transformation, and testing a hypothesis on the data. This project is intended to let you use and show all of your statistical, programming, and data skills you have learned so far. It will be graded on the following guidelines.

* Does the project use skills covered in the class effectively? Projects should include some level of statistical reasoning covered in class.
* Is the project well written? This includes results, captions, graph labels, and tables.
* Is the code well written? The code should be human readable, well commented, and use best practices.

Projects can vary widely. For some projects, the data acquisition will be the hardest and more interesting part and for others, the statistical methods will be more interesting. I highly recommend you do not pick a project that is difficult in data acquisition AND statistical methods, unless you plan to spend plenty of time on the project.

The data used should be reasonably large enough to make statistical observations, unless otherwise approved by me (there are interesting statistical techniques for small data sets, but that is not our focus). If we must put a lower bound on the number of observations, make sure you have a data set with at least a thousand rows in a dataframe. If the data is too large to email (>10MB), please have a working subset available. These are not a hard boundaries, as many interesting data projects come from the merging/combining of multiple smaller data sets. If the data you want to use is of confidential nature (e.g. from your workplace), then please jitter the data, falsify it, or anonymize it to make it accessible.

Examples of projects:

* Personalized data collection analysis: More and more devices allow us to monitor data on ourselves. Eating habits, sleeping schedules, heart rate, etc… Nothing seems to be off limits. If you have access to a data set, it would be interesting to see if there are any interesting testable hypotheses.
* Analyze tweets using twitteR package. Do people tweet more about one subject than another? Are tweets with longer than average words also longer in total length? Of a set of hashtags, which has more images per tweet?
* Kaggle Competition data. Since there is no difficulty accessing/transforming the data, make sure that you do enough statistical testing for the results.

If you have a hard time coming up with an idea, here are a list of resources you can explore to help spark ideas:

* Reddit datasets: <http://www.reddit.com/r/datasets>
* Reddit dataisbeautiful: <http://www.reddit.com/r/dataisbeautiful/>
* R-bloggers: <http://www.r-bloggers.com/>
* K-d-nuggets: <http://www.kdnuggets.com/>
* Kaggle competitions: <https://www.kaggle.com/>
* DataKind: <http://www.datakind.org/projects/>
* Data Science for social good: <http://dssg.io/projects/>
* Data Science-stack exchange: <http://datascience.stackexchange.com/>

Project Checklist:

* Make sure you comment your code.
* Label all graphs, axes, titles, figures, tables…
* Use descriptive variable/function/class names.
* Make use of the logging function and ‘trycatch’ abilities in R where appropriate.
* Make sure to turn in references. (Sources of data, when accessed, what resources used,…)
* Make sure to have a discussion on results with future directions and takeaway points.