Term Project, Stat 196J, Fall 2017

This project is due by **midnight, Friday, December 8, 2017**. *Late projects will NOT be accepted.* You may choose from the following projects.

1. *Option 1*: The first option is a hiring interview task for a Data Scientist position at Wikimedia Foundation. This task will involve analyzing a large dataset of about 400,000 rows and 9 columns of information about internet search results.

   - Here is a link to an interesting article about their hiring process for this position: `https://blog.wikimedia.org/2017/02/02/hiring-data-scientist/`
   - Here is a link to the data analysis task: `https://github.com/wikimedia-research/Discovery-Hiring-Analyst-2016`

   Some tools and packages which will be helpful.

   (a) The data.table package or another way of dealing with large datasets

   (b) Date and time functions: as.POSIXct, as.POSIXlt, strptime()

   (c) RMarkdown for creating a reproducible report

2. *Option 2*: Create an animated graph as complex or more so than the Hans Rosling graph shown in class or the animated world gender equality graph on Kaggle using data extracted from the World Development Indicators database here `https://www.kaggle.com/worldbank/world-development-indicators`. Here are the links to the two example graphs:

   (a) Hans Rosling: `https://www.youtube.com/watch?v=jbkSRLYSojo`

   (b) World Map with Gender Equality Indicators: `https://www.kaggle.com/tentotheminus9/gender-equality`

   You must produce a reproducible report of your data extraction and graph generation using RMarkdown as well as a short summary of any interesting trends or patterns you see in your animated graph.

3. *Option 3:* You may propose a project of your choosing to me by Nov. 22, 2017. It must be of roughly equal difficulty as the options shown above. You must write up your proposal and include the data you will use and your proposed analysis. It must be done in R. You must submit an electronic copy of your data along with your report.

For all projects:

- R code must be used and have a reasonable level of commenting.
- A reproducible report must be made using RMarkdown. You will submit both a hard and soft copy of your report.
- You may use online discussion groups and resources, but you must be able to explain every line of your code.
- You will present a graph or table from your analysis to the class during the last week of classes. This project is worth 100 points and the graph presentation will be worth 7 pts.