Each student will work in a **group of 3 to 5 students** on a term project to develop a predictive modeling analysis using real-world dataset(s). All students in the same project group will, by default, get the same grade for the project. However, if a group feels that this may be unfair for their project (for example, because of greatly different workload/contribution), the group must let me know by the last class day (2nd May).

**Dates**:

1. **Project outline due Friday, 2nd March, via canvas**. One submission per group; clearly identify group members (names and eids). 2-3 pages describing the following items, each with a separate section heading: (a) problem/question(s) of interest, (b) data available (identify sources and highlight size/scope of the data), (c) some possible approaches you will consider to address the problem, (d) key uncertainties you see in the project, and (e) a short list of references. We will provide feedback within a week.
2. On **Wednesday, 8th March,** each group will do a **5** **minute project outline presentation** in class. The presentations should generally follow the items (a-d, in particular) developed in the project outline.
3. **Presentation** of project results, in-class on **Wednesday, 2nd May**, ~ 12-15 minutes per group.
4. **Written project** **report due by midnight, Monday, 7th May, via Canvas.** One submission per group. Your report should be ~ 20-25 pages (1.5 spacing) including figures, tables, and references in a single pdf file. You are encouraged to submit supplementary materials (code, data). To do that, make a folder (on UT Box, github, etc.,) and give me access to the folder/location.

**Choosing a project topic**

You will have wide latitude in selecting a topic, but here are some useful guidelines:

1. Choose an interesting problem that has associated dataset(s), analyzing which would lead to useful/actionable insights. Preferably, you will not repeat an analysis already done elsewhere or by somebody else.
2. You can start with a problem of interest (say, educational outcomes of children or spreading of the opioid epidemic) and find data. Or, you can find rich data and then form interesting questions that analyzing it could help you answer. (See below for some potential data sources.)
3. Not required, but you are encouraged to choose topics that have relatively large associated datasets (*n*x*p* ~ 100,000 or larger).
4. You are discouraged from choosing something that is part of your ongoing research.

**Some Good Data Sources**

**Rich and broad range of datasets**

Microsoft has made available a variety of datasets:

* <http://research.microsoft.com/en-us/projects/data-science-initiative/default.aspx>

A great amount and variety of public domain datasets are available on the web. See the following list by KDnuggets:

* <http://www.kdnuggets.com/datasets/index.html>

**Healthcare datasets**

* <http://www.nlm.nih.gov/hsrinfo/datasites.html>
* <http://www.dartmouthatlas.org/>
* <http://data.worldbank.org/country>
* <http://www.cdc.gov/brfss/>
* <http://www.icpsr.umich.edu/icpsrweb/SAMHDA/download>

**Government Datasets**

* <http://www.data.gov>
* <https://onthemap.ces.census.gov/>
* <https://data.austintexas.gov/> (other cities have similar datasets too)