**Survey of Data Science (DS-5320)**

## Group Data Project Preliminary Analysis

The purpose of this assignment is to complete the first cycle of your Group Data Project. To keep it simple, though, all you need to do is fill out the answers to the following questions. You may use text, tables, plots, etc. to answer the question(s).

The goal here is for you to begin working on your project, begin to develop your question, audience, and analysis, but get feedback from me before you get too far.

Please use this document to answer the following questions, convert it to a PDF, and then upload to Brightspace to turn in your work.

**Question 1**: What is question you hope to answer? Think of this as part of your ‘story’ – i.e. what is the felt need of your audience that you address?

**Question 2:** Who is your audience? This can be almost any “real” audience – i.e. do not say that I am your audience, or your classmates. Your audience may be a government or specific government agency (of any country or many countries), it may be a global body like the United Nations or World Health Organization. It may be a nonprofit like the Gates Foundation. It may be a corporation or group of corporations (like the Auto industry: Volkswagen, Ford, GM, Globally or focused on one country). It could be a group of academics (like Economists) or a professional group (like the American Medical Association – an association of doctors in the United States).

**Question 3**: What is your hypothesis? I.e. what is your hypothesized answer to the question you propose in Q1 above? I see this as a hypothesis now because you probably have not fully analyzed it enough to know if it is, in fact, the right answer.

**Question 4:** What is your data set? Please provide the following information:

* Time range (i.e. Jan 1980 – July 2019) and Time frequency (Annual, Monthly, Daily, minute, etc.)
  + Perhaps your data is event driven, like hurricane events, crimes committed, etc. – just note that.
  + Perhaps time is not a relevant variable, i.e. if you are analyzing brain scan or drug data from a large set of patients – just note the approximate dates when the data was collected.
* Main unit of observation: is it country data, city data, person-level data, company data, etc.
* How big is the data? At least provide the # of rows, and size on your computer (in GB). Also provide something like number of units (i.e. “we have 586 unique countries that provide data”)
* Provide a list of the most important fields in your data, along with a brief description (a phrase or sentence) of each. This should be approximately 10 fields (i.e. not just 2 or 3, but also not 20).   
  Note: because this is preliminary, I understand if this list changes between now and when you turn it in. Do your best to identify the important data given what you know now.

I have given you a lot of freedom to choose your data, so if one or more of the above categories do not exactly apply, simply discuss with me how to best describe your data briefly.

**Question 5:** Tell me 2 or 3 things you have already discovered that relates to your question and/or hypothesis. You should answer this at least with a few sentences of text, but a plot or table is also encouraged. This can be a screenshot from R or Python or your favorite analysis tool. You need not follow all the guidelines for presentation we’ve discussed in class because I see this as a working document – I only ask that it be readable and understandable.

Also, as before, you are not committed to even using this in your final product. Things may change.

**Question 6:** What is your plan or next steps toward the completion of this assignment? Who is doing what?