STAT 440 Project – Spring 2016
Project Guidelines

The goal of the project is to apply what you have learned in this course pertaining to managing data using SAS.

Timeline

1. A roster of your group members is due by **Friday, March 11**. Just send me one email with the list of 3-4 group members. Include all participants' University email in the CC line as a means to verify that all agree to the group.

   After that, I will create a group affiliation for you (and the team) on Compass. This will allow you to submit your proposal and your project as a single Group.

2. A proposal of your intended project is due by **Friday, March 18** to the submission folder on Compass.

   After I review the proposal, it will be evaluated in one of two ways.
   - Approved – Your group may proceed with your plans for the data and project.
   - Pending – I will provide suggestions, concerns, or needed information that must be addressed before the proposal will be approved.

3. The project is due by **Friday, May 6, 11:59pm** to the submission folder on Compass.

Guidelines

1. You are to work as a group of three or four students.

2. Select at least two raw (non-SAS) data sets to work with. These data sources might be relevant to research outside of this course, another field, or some other interest of yours. It will be helpful if you have a few questions in mind and that the data that you select would be relevant to those questions.

3. At least one of the raw data sets must be "substantially large", i.e., abiding by the following parameters: at least 3,000 records and at least 15 variables. You do not have to analyze all the variables in your data set.

4. You will apply data preparation techniques such as:
   - Reading raw data files
   - Checking data for errors
   - Validating and cleaning data
   - Creating formats and labels
   - Deriving new variables via calculations or recoding
   - Array creation

- Iterative (DO-loop) processing
- Subsetting data
- Joining, merging or appending (if multiple data sets are used)
- Conditional output

5. You will prepare a summary report regarding the data.
   a. This can be an exhaustive and thorough analysis of the data set using all available validation and cleaning techniques.
   b. This can be an initial exploratory analysis of the data using at least 5 variables that are relevant to your business or research questions such as:
      - Descriptive statistics
      - Frequency tables
      - Summary tables relating several variables, e.g., crosstabulations
   c. This is not to be a statistical analysis or modelling exercise. Keep the focus on data management and the validity, structure, and readiness of the data.

Proposal Details

A proposal of your intended project should include the following.

a. The names of the students who will be contributing to the group project.
b. A tentative title for the project.
c. Description of the data files (what they contain including number of variables and number of records). You do not necessarily have to list all the variables, but at least mention those of greatest importance.
d. Background information on the data sets, including specific citation of their source (so that I can also access it).
e. A brief statement of the business, science, or research interest you have in the data set which you hope to explore.
f. Description of data preparation that you intend to do (e.g., merging, joining, subsetting).

Simply submit a single document outlining the proposal.

Potential Sources for Your Data

If you have any questions about whether your data meets the specifications of the project given above, don't hesitate to ask. If you plan to use data from another endeavor of yours, such as a research project, be sure to gain permission from the controlling authority first.

- Data.gov -- http://www.data.gov/
- Centers for Disease Control and Prevention -- http://www.cdc.gov/datastatistics/
- StatSci.org -- http://www.statsci.org/datasets.html
- NIST's Statistical Reference Datasets (StRD) -- http://www.itl.nist.gov/div898/strd/
- U.S. Census Bureau -- http://www.census.gov/main/www/access.html
- U.S. Bureau of Labor Statistics -- http://www.bls.gov/data/

- Inter-university Consortium for Political and Social Research -- http://www.icpsr.umich.edu/icpsrweb/ICPSR/index.jsp
- Data Counts! -- http://ssdan.net/datacounts/index.html
- MathForum.org Data Library -- http://mathforum.org/workshops/sum96/data.collections/datalibrary/index.html
- StatLib Data Library (Data, Software and News from the Statistics Community) -- http://lib.stat.cmu.edu/datasets/
- Sports-Reference family of sites -- http://www.sports-reference.com/

Project Details

1. Access to the data.
   If the data is not publically accessible and cited with a link in your project report, you must also provide the actual raw data files, uploaded to Compass. Since your project proposal and related data would have already been approved, you can do this in advance of completing the project. If you find that raw data files are too large to upload, contact me for an alternate option. It is your responsibility to have a way to provide the data to me well before the deadline and unsuccessful last minute attempts may result in penalties to the grade.

2. The SAS program file.
   This should include only that executable code pertinent to the summary report.

3. A summary report that includes the following.
   a. Title of the project.
   b. List of group members.
   c. Introduction section:
      - Statement of the business, science, or research interest in preparing the data you've chosen.
      - Background information on the data source(s), including a description and links of where it originated.
   d. Methods section:
      - Description of the original data files.
      - Description of the guidelines used to validate the data.
      - Description of the issues needed to be cleaned and how it will be done (though not needing to explain the programming specifically).
      - Description of additional data preparation that you performed (e.g. merging, joining, subsetting).
      - Description of variables to be analyzed including attributes such as name and type. You do not have to list all the variables in the original sourced files, but do mention the ones you bring to SAS for the creation of the SAS data set.
   e. Results section:
      - Charts and tables pertaining to validation and cleaning.

- Write-up of the results. Point out notable information from the charts and tables.

f. Write in complete sentences and pay attention to grammar, spelling, readability and presentation. If you include a table or chart, make sure you say something about it. If you're not discussing a result, then it doesn't belong in your report.

Submit a SAS program file and the project report. You may also need to submit the data itself if it is not publicly accessible, such as through an open internet site.

Project Grading

The grading criteria for the final project are as follows:

SAS file

- 10 points – Efficacy of SAS program
- 5 points – Data management
- 5 points – Organization and writing

Summary Report

- 5 points – Data preparation techniques
- 5 points – Output and results
- 5 points – Report content (ability to address data analysis items)
- 5 points – Organization and interpretation (readability, presentation)

Maximum total points: 40