Data 602 Project Description

Course Weighting: The project component of Data 602 comprises 40% of your final grade.

Group: Each group will consist of **THREE** people. Those who are currently registered in Data 601 and have been created groups can remain in the same group for the Data 602 project. For students who are not currently registered in Data 601 will need to team up with another who is only registered in Data 602 (same section – AM students with AM students, PM students with PM students).

Data Sources: Each group will access a publically available data set. Below are some links/portals to access open data sets, some of which are substantial in size:

1. University of Calgary Data Sources for Data Science
2. Open Calgary
3. Government of Alberta's Open Data Program
4. Government of Canada's Open Data Portal
5. Kaggle's Open Data Portal
6. GitHub's Open Data Collection
7. Quandl – multitudes of financial data
8. Lending Club – data pertaining to lending
9. Google Dataset Search

Again, if you are using a certain data set for Data 601, *use the same data set for Data 602*.

Project Objective: The purpose of the 602 project is to look at investigate two different statistical analysis, each one addressing a different estimation, statistical hypotheses, or creating of a statistical prediction model. In short, two different parts to your project. Each part should apply a different statistical method/application.

Each of your two analyses should be preamble with a motivation, which may include a hypothesis. Please be reminded that any statistical hypotheses should be created/stated prior to data inspection/data visualization.

The resulting data analysis should include the following pieces:

1. Data Visualization, or 'time to play'. Depending on the type of data analysis you will be conducting, create an appropriate visualization of the data. This may be in the form a scatteplot, a series of histograms, density plots, or boxplots. Ensure you provide commentary that addresses your learning of the data from these graphs/plots. **These need to be done in R/R Studio.**

2. Condition Checking: The statistical method you are to apply to investigate 'what is happening statistically' may have certain conditions/assumptions about the data. Do these conditions/assumption appear to hold? How

can you tell? What data visualizations are needed in order to check these conditions/assumptions? What if the conditions/assumptions do not hold, what then can you do?

3. Data Analysis: In this piece, simply provide the relevant R/R Studio output. There is NO NEED to give formulas, plug the numbers into formulas to 'verify' the various statistics appearing in the R/R Studio output.

4. Inference Piece: What conclusion(s) or inference(s) can you make from your data? What inference(s) can you not make? Are there limitations to your findings?

Project Proposal (Tuesday, October 1): Prior to starting your data investigation, provide us with a two-page write up that outlines

- Purpose: What is the domain? Within said domain, what are you attempting to investigate? What are practical implications of your analysis and findings? What is/are the population(s)? What variables are of interest? How have data on these variables been observed/recorded?

- Data: What is the source of your data set? Do you have permission to use the data? Ensure you cite the course of the data and the time-period over which the data was collected. (You do not want to use data that is rather antiquated..) What types of data visualizations to you perceive will be used to visually inspect the data?

- Topic(s) to Investigate: Outline the focus of your statistical investigation. Provide context to the data to education those who are unfamiliar with the data space you are working within. Provide citations any previous research/studies completed in the domain within which you are playing.

- Statistical methods to be used: Provide a list of the types of statistical methods you will use in your analysis. For example, if you are planning on creating a confidence interval for the difference between two population proportions, are you to do so via the "bootstrap technique" or the "conventional method" (or both)?

The Project Proposal is due by Tuesday, October 1st @ in class (AM for the morning class, PM for the evening section).

Report and Final Presentation (Thursday, October 17) : Each group will give a **5 minutes** presentation of their project and findings during the class of October 17. Group projects will also be submitted in the form of a report.