**Final Project for STAT3041(2020SP)**

**Deadline**: April 30th, Thursday

**Project Objectives**:

Load ‘covid19\_dataset.Rdata’. The ‘dataset’ object contains the follow variables:

1. state
2. date
3. total\_cases: The total (cumulative) number of confirmed cases from January 21st
4. total\_deaths: The total (cumulative) number of deaths due to COVID-19 from January 21st
5. daily\_cases: The number of new confirmed cases
6. daily\_deaths: The number of new deaths due to COVID-19
7. growth\_rate: total number of cases today/total number of cases yesterday

Using this data set perform the following data analyses:

Objective 1. Identify three states with the highest cumulative confirmed cases as of April 13th. Perform Bayesian analysis to find the credible intervals for the mortality rate (i.e. probability of death when confirmed for COVID-19) in each of those three states based on the cumulative numbers of confirmed cases and deaths as of April 13th. Create at least one table or plot to summarize the results.

Hint1: Assume that the number of deaths follows a binomial distribution with n=total\_cases.

Hint2: Since no prior information is available, use the neutral (or noninformative) prior for *p*.

Objective 2. For each of those three states, find interval estimates (95% confidence intervals and credible interval) for the mean daily confirmed cases during the period of 4/7/2020-4/13/2020 using parametric bootstrapping and Bayesian analysis. For Bayesian analysis find a 95% prediction interval for the number of confirmed cases for a new day as well. Create at least one table or plot to summarize the results.

Hint1: The number of events is usually modelled by a Poisson distribution.

Hint2: For parametric bootstrapping, first compute the mean number of daily confirmed cases during 4/7/2020-4/13/2020 period, and then generate samples from a Poisson distribution with the computed mean.

Hint3: Since no prior information is available, use the neutral (or noninformative) prior for when performing Bayesian analysis.

Hint4: Obtaining data for the period of 4/7/2020-4/13/2020 can be simply done by applying filter(date<="2020-04-13",date>="2020-04-07")

Objective 3. For each of those three states, find the 95% confidence interval for the mean growth rate during the period of 3/31/2020-4/6/2020 and 4/7/2020-4/13/2020 using nonparametric bootstrapping. Create at least one table or plot to summarize the results and **compare** the mean growth rates in those two periods based on the confidence intervals.

**Common Grading Criteria**: Here is the rule for how the final project will be graded:

* Accuracy of Analysis (40 points): Analysis should be performed with correct technical details. (10 point for each objective)
* Overall quality of report (30 points): Data preparation, model fitting process, and analysis results should be clearly described in detail.
* Overall quality of programing code (30 points): Students need to submit their R code that is organized in a clear manner with proper comments.

NOTE: The R outputs included in the report need to be properly tailored – Copying and pasting unedited raw outputs or screen shots with too much redundant information will lead to 30 points reduction from the final score.

**Submission Instruction**:

Students need to submit their report (.docx or .pdf) and code (as a separate .R). The report should include properly edited R output or the screenshots to summarize and visualize the analysis results.