SML201 Test 1, Spring 2020

Test rules

During the test, you may use Rstudio, and consult the course website, the course textbook, and the course Piazza. No other aids are allowed. You may communicate with the course instructor by Zoom at https://princeton.zoom.us/j/9425325325, by email, or, only if your internet connection is unreliable, by cell phone at +1-609-375-7283. You may not communicate with anyone else regarding the test and SML201 content.

Submit the file `test1.R` containing all the answers on Gradescope at https://www.gradescope.com/courses/88811/assignments/408570/. All your answers must be in `test1.R`.

Announcements regrarding errata may be made by email during the test.

Unless you have an approved time extension, you have 100 minutes to work on the test. Submit the test on Gradescope no later than 3:40 p.m. Eastern Time. If you have an extension, *submit the test by email.*

Answer as many questions as possible. Partial credit may be given for answers where substantial progress was made toward a correct solution.

Note that the problems are not necessarily arranged in order of difficulty. Read over the test, and be strategic about which problems to work on.

Some of the questions will ask for R code as well as English and math. Below is an example of how you should answer those. Note that any English or math is commented out (using the pound sign), but R code is not commented out. You must separate the answers to the different problems using lines of pound signs.

```
################################################################################
# PROBLEM 1

sq <- function(x){
  x ** 2
}

x <- c(5, 6, 7)
sum(sq(x)) # Output: 110

# I used the sq function to compute 5^2, 6^2, and 7^2,
# summed the results up, and got 110
################################################################################
# Problem 2
# .........
```

Include the following text, and add your name to it to confirm that you have have not violated the rules for this test.

```
# I pledge my honor that I have not violated the Honor Code during this examination.
# Specifically, I have taken the no more than the amount of time alotted to me to
# write the test, I have not used any unauthorized aids, and I have not communicated
# with anyone apart from the course instructor regarding the test or the course material
# Digital siganture: <your name here>
```

Good luck!

Problem 1 (10 pts)

You can use `%in%` to check whether an element is present in a vector.

For example, `5 %in% c(20, 5, 2)` is `TRUE`, but `3 %in% c(20, 5, 2)` is `FALSE`. Write a function that takes in a number and two vectors of numbers, and returns `TRUE` if the number is present in at least one of the vectors. The function signature should be

```
present.in.at.least.one <- function(num, vec1, vec2)
```

For example, `present.in.at.least.one(5, c(1, 2, 3), c(4, 5, 6))` should return `TRUE`, but `present.in.at.least.one(15, c(1, 2, 3), c(4, 5, 6))` should return `FALSE`.

Problem 2 (15 pts)

Write a function that takes in three vectors, `nums`, `vec1`, and `vec2`, and returns `TRUE` if each entry in `nums` is present in at least one of `vec1` and `vec2`. The function signature should be

```
present.in.at.least.one.vecs <- function(nums, vec1, vec2)
```

For example, `present.in.at.least.one.vecs(c(3, 4), c(1, 2, 3), c(3, 6, 5, 4))` should return `TRUE` because both 3 and 4 are present in one of `c(1, 2, 3)` and `c(3, 6, 5, 4)`, but `present.in.at.least.one.vecs(c(5, 4), c(1, 2, 3), c(6, 7, 4))` should return `FALSE` since 5 is not present in either `c(1, 2, 3)` or `c(6, 7, 4)`.

(end of Problem 2)

For several of the questions below, you will be working with the `corona` dataset, which you can read in as follows:

```
corona <- read.csv("http://guerzhoy.princeton.edu/201s20/corona.csv", stringsAsFactors = F)
```

For the purposes of the problems below, you would count each different entry in the Country column as a separate country. So for example "China" and "Macau" count as different countries. (Even though Macau is a Special Administrative Region in China).

Note that the numbers of cases ("Confirmed", "Deaths", "Recovered") are cumulative: the value for each date is the number of e.g. confirmed cases up to and including that date.

You can assume that, since there are rows for, for example, both "British Columbia" and "Vancouver, BC" that the number for "British Columbia" do not include the number of cases in Vancouver.

If you need to make further (reasonable) assumptions about the data, state them when you are answering the questions.

Problem 3 (5 pts)

Write a function that takes in a data frame like `corona` and the name of a country, and returns the number of the different locations (provinces, states, cities, counties, etc.) in that country that are present in the data frame.

Problem 4 (15 pts)

Write a function that takes in a dataset like `corona` and the name of a country, and returns the name of the city in that country that had the largest number of confirmed cases. You can assume that the country is either the US or Canada, and the format for the location looks like "Chicago, IL" or "Orange, CA". That is, you can assume that the name of the city/county is followed by the state/province abbreviation. For example, if the answer is Los Angeles, California (appearing as "Los Angeles, CA"), your function should return "Los Angeles". (Note that we are only interested in numbers for cities/counties that are presented in the format "CITYNAME, STATE.ABBREVIATION", not for, e.g., "California".)

Problem 5 (15 points)

Write code to make one plot to display the log of the number of confirmed cases vs. the number of days since 01/21/2020, for China, the US, and Italy. Note that you need to first compute the total number of confirmed cases for each of the countries. Make sure there are appropriate labels an captions.

Submit the code we would need to recreate the plot.

Problem 6 (15 points)

Based on the data, obtain a formula for predicting the number of confirmed cases in China from the number of days since 01/21/2020. In your answer, include the code you used in order to obtain the formula. Use all the available data to obtain this formula. The formula should be the best one possible, using the tools you have from this course. Make sure that the formula outputs a prediction for the number of confirmed cases.

Include the code you used, as well as the formula you obtained, in the file you submit. Briefly explain what you did. Any text that's not R code should be included as a comment.

Problem 7 (10 points)

Assess how well the formula in Problem 6 predicts the number of confirmed cases in China. You should assess the performance using the same data that you used to obtain the formula. You should present the assessment in plain language, as much as possible. In plain language, explain any numbers that you present.

Problem 8 (5 points)

Obtain a formula for the number of *new* confirmed cases in China on day n since 01/21/2020. Show your work.

Problem 9 (10 points)

Write code to display the following plot. Note that we are not providing a dataset – it is up to you to create whatever data frames you need.