STAT 541 Final Exam-Spring 2020

This is a take-home exam; all the usual course resources are available, but you are not to discuss the exam with other students. Contact me if you have questions.

The main data sets we will be using for this exam are COVID-19 datasets from a GitHUB site used by Johns Hopkins University. We will use PROC HTTP to "web-scrape" the data, a technique used by one of our students for their project. Here is code for obtaining daily results for US states/territories on April 19 (though you will note additional non-US records for this particular data set) and storing the data in WORK.covid419; be sure the URL is read as a single string:

```
filename covid419 temp;
proc http out=covid419 url="https://raw.githubusercontent.com/CSSEGISandData/
COVID-19/master/csse_covid_19_data/
csse_covid_19_daily_reports_us/04-19-2020.csv"
method="GET";
run;

proc import out=covid419
file=covid419
replace
dbms=csv;
run;
```

1. (10 points) Repeat the code above for April 26 data; you will need to change all references to April 19 (but don't change `covid-19` in the URL!) to April 26.

2. (10 points) Chapter 2. Create two new data sets (though the April 26 data set technically doesn't need conversion) consisting only of US records.

3. (20 points) Chapter 18.
   (a) All students. Create a unique index for Province_State for the April 19 data set from Q2, then use the index and the April 26 data set from Q2 to update the Recovered column. Summarize changes.
   (b) Graduate students. Recreate the April 19 data set from Q2 (we want the modified version, not the version from Q3(a)), then update the April 19 data set from Q2 using the April 26 data set from Q2 as a transactional data set. Summarize changes.

4. (30 points) Chapter 16. The table below cross-classifies two variables from the April 26 data set in Q2: Mortality_Rate (MR) and Testing_Rate (TR). MR is rated from 1 to 3 corresponding to observed mortality rates of < 3, 3 to < 6, and 6 or more. TR is rated from 1 to 3 corresponding to testing rates of < 1000, 1000 to

< 2500, and 2500 or more. The values in the chart represent the states' response (1=Proactive, 2=Active, 3=Reactive).

	TR		
MR	1	2	3
1	2	3	3
2	1	2	3
3	1	1	2

(a) Create informats that code Mortality_Rate from 1 to 3 and code Testing_Rate from 1 to 3.

(b) In preparation for the next step, remove all records from the April 26 data in Q2 with a missing value for Mortality_Rate and/or Testing_Rate.

(c) Use the Chapter 16 method of your choice (I used the first method we learned–arrays–to assign a state response variable (StateResponse) to each observation in the data set from (b).

(d) Print Province_State, Mortality_Rate, Testing_Rate and State_Response. The printed output should include formatted values Proactive, Active, or Reactive for StateResponse as appropriate. Review the output to ensure your lookup table approach worked correctly.

5. (10 points) Chapter 17. Use PICTURE formats and directives to print the variable Last_Update in your April 19 data set from Q2 to have, e.g., the following appearance: 11:41 PM on Sunday April 19, 2020.

6. (20 points) Chapter 15. We will use the April 19 and April 26 data set from Q2 for this problem.

(a) Use PROC SQL to create index Province_State on the variable Province_State for the April 19 data.

(b) Create a new table from the April 26 data of all states with more than 100 cumulative deaths (i.e, Deaths > 100). The new table should contain only the variables Province_State, Deaths, and Recovered, though the latter two should be renamed Deaths26 and Recovered26. How many states are listed in this table?

(c) Use the index to retrospectively update the new table in (b) to add Deaths and Recovered from the April 19 data set (rename these variables Deaths19 and Recovered19). How many of these states with over 100 cumulative deaths on April 26 did not have over 100 cumulative deaths on April 19?