Stat 5080/6080 Data Technologies	Fall 2019

Homework Assignment 2 (11/5/2019)

84 Points — Due Tuesday 11/26/2019 (via Canvas by 11:59pm)

(i) (84 Points) You are asked to help me in my current role as the President of the *International Association for Statistical Computing* (IASC) with a simple question: How much activity is there in Africa with respect to computational statistics?

Background: The IASC has a *"world-wide interest in effective statistical computing and to exchange technical knowledge through international contacts and meetings between statisticians, computing professionals, organizations, institutions, governments and the general public"* (see `http://iasc-isi.org/` for more details about the IASC — new students members are always welcome!). The IASC currently has three regional sections in Europe (IASC-ERS), in Asia (IASC-ARS), and in Latin America (IASC-LARS), that was founded in 2017. Recently, the IASC was contacted by some of its members with the question whether it would be feasible to establish a new regional section in Africa. To establish a new regional section, there must be a minimum number of IASC members in that geographic region. Moreover, the IASC General Assembly (GA) must approve a new regional section. That approval likely depends on the question whether the new section has the potential to conduct typical section activities, such as organizing regional conferences, workshops, and short courses where most presenters and attendees come from this geographic region. This leads to the question whether there is currently enough high-level activity in Africa with respect to computational statistics.

Approach: There exist multiple ways to explore the activities of researchers in a geographic region. Traditionally, one might have conducted a phone or mail survey. Alternatively, one could extract information from web pages from university and research institutes. In this HW, we will use author information from two leading journals in the field of computational statistics to answer this question. Specifically, we will extract author information from authors located in Africa in the past few years and see how this relates to similar author information from authors located in Latin America from a few years ago.

You will be asked to create many tables for this HW. See how to use the *kable*

function from the *knitr & kableExtra* R packages (easier, but less powerful) or the *xtable* function from the *xtable* R package (harder, but more powerful) to export your tabular data and data frames from R into a nice printable table in your LaTeX document.

This homework is a group homework with groups of 4 or 5 members. Groups were assigned in class on Tuesday 11/5/2019. The group head has to submit the answers for the following questions on behalf of the entire group. The other group members do not have to submit answers separately. Each group member will obtain the same score.

As always, make sure to include your R part and a resulting graph if the question asks for a graph. Use *tidyverse* functionality whenever possible.

(a) (1 Point) Load all required R packages to answer this question. Show your R code.

```
> library(tidyverse)
```

(b) (15 Points) Download a list of African countries from `http://statisticstimes.com/demographics/african-countries-by-population.php`. This contains a nice table. Extract the country names and the most recent population count (from 2018) from this table and transform these numbers into a numeric. Write this information into an external csv file with just these two columns.

Repeat these steps for the five other listed continents at the bottom of the Africa page, i.e., for Asia, Europe, North America, South America, and Oceania. Further adjust the countries for North and South America: Only the United States and Canada belong to North America for the purpose of this HW. We consider Mexico and everything further to the south as South (Latin) America. You should manually check your results. Do you read in all countries correctly, do you extract the correct population counts and are those numeric, etc.?

Produce a final summary table for this question part that contains six rows (one for each continent) and three columns: The name of the continent, the number of countries (only two for North America) in that continent, and the total 2018 population for this continent. Include your R code and this final table in your HW.

(c) (24 Points) Extract author information from the *Springer* journal *Computational Statistics* (COST) for the past 5 years: Start at `https://link.springer.com/journal/volumesAndIssues/180` and extract information, beginning with Volume 30, Issue 1, March 2015, and ending with Volume 34, Issue 4, December 2019. Do not continue for 2020, even if the next issue(s) get posted while you work on this HW.

For each published article in each issue, obtain the following information and store in a data frame:

- Journal (COST, CSDA).
- Year.
- Volume.
- Issue.
- Title of the article.
- Number of authors for the article.
- Author name.
- Author affiliation.
- Author country.
- Author order (1, ..., number of authors for the article).
- Start page of article.
- End page of article.

Note that most articles have more than just one author. As this is basically a long format, we repeat similar information for each of the authors of an article.

Manually check that your results are meaningful, that you capture information for all authors, etc. Be careful with special issues, editorials, and invited/discussion papers. The list of articles in an issue may extend beyond one web page. We exclude the *Taylor & Francis* journal *Journal of Computational and Graphical Statistics* (JCGS) from this HW as they use at least three different ways to store author affiliations. Based on what I have sampled, COST seems to be consistent, but no guarantees. So, check that you do not just get NAs for some issues.

Overall, you should obtain author information for five years. Once more, check your result by sampling a few articles and checking that the information from those articles has been correctly stored in your data frame.

Repeat, now for the *Elsevier* journal *Computational Statistics & Data Analysis* (CSDA) for the past 5 years: Start at `https://www.sciencedirect.com/journal/computational-statistics-and-data-analysis/issues`, beginning with Volume 81, January 2015, and ending with Volume 140, December 2019.

Create a similar data frame as for COST. **Do not** include any of the forthcoming issues for 2020.

Write your two data frames into two external csv files, one for each journal. Produce a final summary table for this question part that contains five rows (one for each year) and five columns: year, total number of articles in COST in the year, total number of authors in COST in the year (some authors may appear in more than one article; if so, count them multiple times), total number of articles in CSDA in the year, and total number of authors in CSDA in the year (some authors may appear in more than one article; if so, count them multiple times). Include your R code and this final table in your HW.

(d) (8 Points) Likely, the country names from part (b) will not entirely match the author countries from part (c). Identify non-matching author countries from part (c) and write R code that adjusts these author countries to the format used in part (b). Update the names in your data frames and write to two new external csv files. Do not overwrite your previously created external files.

Likely problems with country names could be different versions for the same country, e.g., United States, USA, U.S.A., etc., instead of United States of America, different spellings (e.g., Viet Nam or Vietnam), use of special characters (e.g., Côte d'Ivoire), prefixes and postfixes to a country name (e.g., Republic of Moldova or just Moldova; Russian Federation or just Russia), name changes (e.g., TFYR Macedonia, Macedonia, or North Macedonia), and more.

Be careful when you work with regular expressions and what you match: There exist Niger and Nigeria (two different countries), Congo and Democratic Republic of the Congo (two different countries), Sudan and South Sudan (two different countries), Dem. People's Republic of Korea and Republic of Korea (two different countries; often referred to as North Korea and South Korea for simplicity), and many more similar examples.

Create a summary table that lists the different versions found in part (c) and to which country name from part (b) they get translated. There may be multiple versions from part (c) that all get translated to the same version from part (b). Arrange this table in alphabetical order, according to the versions from part (b). Include your R code and this table in your HW.

(e) (16 Points) Produce the following tabular summaries. Create meaningful headings in your tables and meaningful table captions.

i. Top-20 countries, based on all authors for COST for all years combined.

ii. Top-20 countries, based on all authors for CSDA for all years combined.

iii. Top-20 countries, based on all authors for both journals for all years combined.

iv. Top-20 countries, based on first author only for COST for all years combined.

v. Top-20 countries, based on first author only for CSDA for all years combined.

vi. Top-20 countries, based on first author only for both journals for all years combined.

vii. For each of the five years, count of number of first authors and number of pages for first author for Africa and Latin America for both journals combined (two columns for each continent); plus a final row that lists the totals for each column.

viii. For each of the five years, count of number of all authors and number of pages for all authors for Africa and Latin America for both journals combined (two columns for each continent); plus a final row that lists the totals for each column. If an article has two authors, each author is counted twice and the number of pages is counted twice. If there are five authors, the author count will be five and the number of pages is counted five times.

Include your R code and these eight tables in your HW.

(f) (12 Points) You have to be a devil's advocate: Groups 1, 3, and 5 should **support** the forming of a new African regional section. Groups 2 and 4 should try to **hinder** the forming of a new African regional section.

To do this from your perspective, create similar tables as in part (e), but use a different standardization, e.g., using total population in a continent,

using different weights for the number of authors and pages (e.g., for a 20-page article with 2 authors, each author only counts 1/2 or only gets 10 pages; for a 20-page article with 5 authors, each author only counts 1/5 or only gets 4 pages), etc. You want to compare Africa in 2018 and 2019 with Latin America in 2015 and 2016 (before the IASC-LARS regional section was formed). Also consider time series that show an increasing trend over the five-year period to show your point of view, e.g., that Africa is "there" or "almost there" even if it has not exactly reached the numbers for Latin America in 2015 and 2016.

Create at least three meaningful graphical summaries of three different tables that support your point of view. These might be bar charts, spine plots (these might be very useful), line charts / time series, or others. If you are familiar with choropleth maps, consider creating side-by-side choropleth maps for Africa and Latin America where one shows high counts or standardized values for most of the countries and the other shows low counts or standardized values for most of the countries.

You are not allowed to apply any of the rules for bad graphs from Stat Viz I, i.e., you cannot wiggle the baseline in stacked bar charts, you cannot graph data out of context by just comparing the top country in Africa with the top country in Latin America (and omit all other countries), etc. Also, when you create related plots, they typically have to follow the small multiple principle, in particular use identical scales and colors / intervals in maps.

Include your R code and these three (or more) graphs in your HW.

(g) (8 Points) Provide a final summary and discussion of your results from your perspective from part (f) and argue why a regional section in Africa should be formed / should not be formed. Be formal and refer to specific tables and figures from your previous parts in this summary.

If you could not create three different graphs that support your point of view, then concede! Support the forming of a regional section in Africa even if your original task was to hinder it. Or, admit that there is nothing in the data that supports the forming of a regional section in Africa even if your original task was to support it.

This summary should be between 1/2 and 1 page in length.

General Instructions

(i) Create a single pdf document, using Sweave or knitr. As this is a group project with a 6000–level student in each group, you have to use LaTeX in combination with Sweave or knitr. The group head has to receive all individual R and LaTeX components and do the final translation on his/her computer. Check early that your code is reproducible and runs on different computers. This may take some time. Sharing files among group members via box, dropbox, or a google archive may be helpful.

You only have to submit this one final pdf document to Canvas.

(ii) Include a title page that contains your group number, all names, all A–numbers, the number of the assignment, the submission date, and any other relevant information.

(iii) Start your answers to each question part on a new page. Clearly label each question part. Your answer to question part (i) should start on page 2!

(iv) Show your R code, tables, and resulting graph(s) [if any] for each question part!

(v) Reuse (and adapt) as much of the R code from the 6000–level Marvel group project for this HW as possible. There are many similarities. This is why each group has one of the 6000–level students as a group header.

(vi) Before you submit your homework, check that you follow all recommendations from Google's R Style Guide (see `http://web.stanford.edu/class/cs109l/unrestricted/resources/google-style.html`). Moreover, make sure that your R code is consistent, i.e., that you use the same type of assignments and the same type of quotes throughout your entire homework.

(vii) Give credit to external sources, such as stackoverflow or help pages. Be specific and include the full URL where you found the help (or from which help page you got the information). Consider R code from such sources as "legacy code or third–party code" that does not have to be adjusted to Google's R Style (even though it would be nice, in particular if you only used a brief code segment).

(viii) **Not following the general instructions outlined above will result in point deductions!**

(ix) For general questions related to this homework, please use the corresponding discussion board in Canvas! I will try to reply as quickly as possible. Moreover, if one of you knows an answer, please post it. It is fine to refer to web pages and R commands, but do not provide the exact R command with all required arguments or which of the suggestions from a stackoverflow web page eventually worked for you! This will be the task for each individual group!

(x) Submit your single pdf file via Canvas by the submission deadline. Late submissions will result in point deductions as outlined on the syllabus.