DS5110 HW 3 - Due Feb. 12

Kylie Ariel Bemis

2/1/2019

Instructions

Create a directory with the following structure:

- `hw3-your-name/hw3-your-name.Rmd`
- `hw3-your-name/hw3-your-name.pdf`

where `hw3-your-name.Rmd` is an R Markdown file that compiles to create `hw3-your-name.pdf`.

Do not include data in the directory. Compress the directory as `.zip`.

Your solution should include all of the code necessary to answer the problems. All of your code should run (assuming the data is available). All plots should be generated using `ggplot2`. Missing values and overplotting should be handled appropriately. Axes should be labeled clearly and accurately.

To submit your solution, create a new private post of type "Note" on Piazza, select "Individual Student(s) / Instructor(s)" and type "Instructors", select the folder "hw3", go to Insert->Insert file in the Rich Text Editor, upload your `.zip` homework solution. Title your note "[hw3 solutions] your name" and post the private note to Piazza. **Be sure to post it only to instructors**

Part A

Problem 1

Find a dataset that is personally interesting to you. It may be a publicly-available dataset, or a dataset for which you have permission to use and share results. There are many places on to find publicly-available dataset, and simply searching Google for your preferred topic plus "public dataset" may provide many hits. Here some additional resources to get you started:

- US Government datasets (https://catalog.data.gov/dataset)
- Center for Disease Control (CDC) data (https://data.cdc.gov)
- Bureau of Labor Statistics (https://www.bls.gov/data/)
- NASA datasets (https://nssdc.gsfc.nasa.gov)
- World Bank Open Data (https://data.worldbank.org)
- Kaggle Datasets (https://www.kaggle.com/datasets)

This does not have to be the same dataset you will use for your group project.

Import the dataset into R, put it into a tidy format, and print the first ten observations of the dataset.

Problem 2

Step 1: Perform exploratory data analysis on the dataset, using the techniques learned in class. Calculate summary statistics that are of interest to you and create plots using `ggplot2` that show your findings.

Step 2: Create an attractive PowerPoint or Keynote slide including your name, a description of your dataset, and your key findings, incorporating any plots and/or tables that are most relevant and interesting. Make sure you cite the source of the data!

Step 3: *Export this slide to PDF, and upload it to Piazza as a public Note titled "[mini-poster] your name - dataset name" in the "miniposter" folder, along with a brief description of the dataset by the homework due date.*

Part B

Problems 3–5 uses a subset of the DBLP database of bibliographic information on major computer science journals and proceedings, available from https://data.mendeley.com/datasets/3p9w84t5mr. The dataset has been processed to include predictions of the author's genders using the open-source Genderize API. The processed data has been made available in the form of `SQL` scripts that import the data into a MySQL database. We are primarily interested in the "general" and "authors" tables created by the "main.sql" and "authors.sql" scripts, respectively.

You have three options to load the dataset into R: (1) import the data into a MySQL database, accessed via `dbplyr`, (2) edit/convert the scripts and import the data into another RDBMS such as SQLite, which is then accessed via `dbplyr`, or (3) parse the text data in the SQL scripts into R (this is possible but difficult).

If you choose to use MySQL, the README file describes the steps to import the tables into a database, and then use `dbplyr` with the `RMySQL` package to work with the data in R.

If you choose to use another RDBMS such as SQLite (which is easier to install, and many *nix operating systems come with it installed already), you will likely need to edit or convert the scripts to be compatible.

One available conversion tool is the `mysql2sqlite` script available from https://github.com/dumblob/mysql2sqlite. This will convert a MySQL script to a SQLite script.

If you are using a POSIX compliant operating system, you could import the relevant tables into a SQLite database named `dblp.db` using the following commands in a compatible shell:

```
./mysql2sqlite main.sql | sqlite3 dblp.db
./mysql2sqlite authors.sql | sqlite3 dblp.db
```

and use `dbplyr` with the `RSQLite` package to work with the data in R.

Problem 3

Filter the data to include only the authors for whom a gender was predicted as 'male' or 'female' with a probability of 0.95 or greater, and then create a bar plot showing the total number of *distinct* male and female authors published each year. Comment on the visualization.

Problem 4

Still including only the authors for whom a gender was predicted with a probability of 0.95 or greater, create a stacked bar plot showing the *proportions* of distinct male authors vs. distinct female authors published each year. (The stacked bars for each year will sum to one.) Comment on the visualization.

Problem 5

Still including only the authors for whom a gender was predicted with a probability of 0.95 or greater, create a bar plot showing the count of papers published with (1) male first authors and (2) female first authors.

Then create a bar plot showing the count of papers published with (1) no female authors and (2) at least 1 female author. Comment on any similarities and differences between the two bar plots.