DS5110 HW 2 - Due Feb. 13

Kylie Ariel Bemis

1/30/2018

Trigger Warning: This assignment includes references to statistics about suicide. Please contact the instructor if you have difficulty completing it due to personal distress.

Instructions

Create a directory with the following structure:

- `hw2-your-name/hw2-your-name.Rmd`
- `hw2-your-name/hw2-your-name.pdf`

where `hw2-your-name.Rmd` is an R Markdown file that compiles to create `hw2-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 "hw2", go to Insert->Insert file in the Rich Text Editor, upload your `.zip` homework solution. Title your note "[hw2 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" in the "miniposter" folder, along with a brief description of the dataset by the homework due date.*

Part B

Problems 3–4 use the US Department of Education's Civil Rights Data Collection from Homework 1. It is available at https://www2.ed.gov/about/offices/list/ocr/docs/crdc-2013-14.html. Use the `read_csv()` function to import the dataset into R, handling missing data appropriately.

Problem 3

Create a bar plot showing the total number of enrolled students of each race.

Problem 4

Create a bar plot showing the number of students of each race enrolled in an Advanced Mathematics class. Comment on any similarities or differences between this distribution and the one you plotted in Problem 3.

Part C

Problems 5–7 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 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 the scripts to be compatible.

For example, to use SQLite, commands such as:

```
CHARACTER SET utf8mb4 COLLATE utf8mb4_bin
```

```
enum('M','F','-') COLLATE utf8mb4_unicode_ci
```

```
ENGINE=InnoDB DEFAULT CHARSET=utf8mb4 COLLATE=utf8mb4_unicode_ci
```

are incompatible and must be removed. Additionally, to properly escape single quotes in SQLite, you must find and replace `\'` with `''` in the scripts. To do this, you may need to use a text editor that can handle large text files, such as vim, emacs, Sublime Text, Notepad++, etc.

Alternatively, you may use an existing tool for converting MySQL scripts to SQLite scripts (see http://www.sqlite.org/cvstrac/wiki?p=ConverterTools).

After editing the scripts to be SQLite-compatible, you could then import the data into a database using:

```
sqlite3 dblp.db
.read main.sql
.read authors.sql
```

and then Ctrl+D to exit, and use `dbplyr` with the `RSQLite` package to work with the data in R.

Problem 5

Filter the data to include only the authors for whom a gender was predicted with a probability of 0.99 or greater, and then create a bar plot showing the number of *distinct* male and female authors in the dataset.

Problem 6

Again including only the authors for whom a gender was predicted with a probability of 0.99 or greater, create a stacked bar plot showing the number of distinct male and female authors published each year.

Problem 7

Still including only the authors for whom a gender was predicted with a probability of 0.99 or greater, create a stacked bar plot showing the proportions of distinct male and female authors published each year. (The stacked bars for each year will sum to one.)

Part D

Problems 3–4 uses data collected from the Virginia Transgender Health Initiative Study (THIS). It is available via the Inter-university Consortium for Political and Social Research (ICPSR), of which Northeastern University is a member, at http://www.icpsr.umich.edu/icpsrweb/ICPSR/studies/31721 or via a proxy link at http://www.icpsr.umich.edu.ezproxy.neu.edu/icpsrweb/ICPSR/studies/31721 (if you are not on a campus internet connection). You will need to create a free MyData account as well as login with your myNEU credentials to gain access to the public version of the dataset.

Download the R data (`.rda`) version of the dataset and load it into R using the `load()` function.

Problem 8

We would like to investigate how certain questions break down among trans women, trans men, and non-binary participants. However, the survey is sometimes outdated in its terminology and also includes many questioning participants who have not transitioned yet.

Transform the data to include 3 gender categories for men, women, and non-binary participants. Use the following definitions when transformaing the dataset: (1) trans women are women who were assigned-male-at-birth; (2) trans men are men who were assigned-female-at-birth; (3) combine the “Genderqueer” and “Androgynous” categories to create a single “Non-binary” category. Filter the dataset to include only participants in these categories.

Create a bar plot showing the number of participants of each of the above genders.

Then create bar plots showing the proportion of participants who have been fired or denied a job due to their transgender status and/or gender expression. The plots should be faceted by gender and show separate proportions for trans women, trans men, and non-binary participants. (Do not include missing data in the plot.)

Problem 9

Using the full dataset again, transform the dataset to have a column for `race` indicating the race of the participant. Include only the racial demographics with publicly available data (i.e., African American, Caucasian, Hispanic/Latinx, and Native American).

(Participants with two or more races may appear on multiple rows. Do not use the pre-calculated 'RACE' column in the dataset, which does not properly disambiguate multiracial participants.)

Then create bar plots showing the proportions of participants who have thought about killing themselves for African American, Caucasian, Hispanic/Latinx, and Native American demographics. (Do not include missing data in the plot.)

One of the findings reported in the National Transgender Discrimination Survey (http://www.thetaskforce.org/injustice-every-turn-report-national-transgender-discrimination-survey/) was that a staggering 41% of the respondents reported attempting suicide, compared to 1.6% in the general population. Calculate the *total* proportion of participants who have attempted suicide in the Virginia THIS survey. (Include all participants.) Is it higher or lower than the national average for trans people?

Problem 10

We would like to know if having a birth family supportive of one's gender identity and expression reduces the risk of suicide. Create bar plots showing the proportions of participants who have thought about killing themselves for each level of familial support. (Do not include participants who declined to answer.) What do you notice?