Homework 2: Programming in R Sequence

DPSS 2024

Instructions

Homework 2 will consist of two sequences – Data Analytics (DA), and Programming in R (R). **These are the instructions for the R sequence:**

- There will be 8 programming in R questions worth 1.875 points each.
- The R questions will be multiple choice or file upload. However, each R question will have a part a) and part b).
- Part a) will contain the question stem, as well as multiple-choice responses or file upload. Please copy and paste your R code into part b. In order to receive full credit, code must be provided.
- If you get the answer wrong on any of the R questions, but still include code that is close to correct, you may earn partial credit.
- You will not earn extra credit for including your code in part b) of the R questions if you receive full credit in part a).
- While you may discuss approaches with your peers, any copying of code or output will be treated as a case of academic dishonesty. All submitted work should be your own. See the syllabus for details on group collaboration.

We have a set up section to help prepare you to load the data you will need to use.

Set Up

Load the following packages and datasets to complete the programming in R questions. To find the datasets you will use throughout HW assignments in this course, please click Datasets.zip or check the assignment header in the canvas modules.

Download folder and unzip it in your computer, you do not have to unzip it in R. Once you unzip the folder, you will find 4 datasets you will use throughout this class for HW assignments, they are:

1. Traffic_Crashes-_-People.csv
2. Traffic_Crashes-_-Crashes.csv
3. wave_3_fertility.csv
4. wave_3_power.csv

Set the folder you downloaded to your working directory following the instructions from class, and load both datasets into R. An example you may use is given below, where I have saved the datasets into a folder entitled "DPSS." You can also see me load packages which will be relevant for completing the programming assignments.

```
# libraries needed
library(tidyverse)
# setting working directory
folder_path <- "/Users/josemacias/DPSS/Datasets" # use your own path here
setwd(folder_path) # passing the object into the function
# reading in the files
people <- read_csv("Traffic_Crashes_-_People.csv")
crashes <- read_csv("Traffic_Crashes_-_Crashes.csv")
fert <- read_csv("wave_3_fertility.csv")
power <- read_csv("wave_3_power.csv")
```

About the Data

- In the "**Traffic_Crashes-_-People**" dataset, each record corresponds to an occupant in a vehicle which experienced a crash. The crashes are listed in a separate Crash dataset. Some people involved in a crash may not have been an occupant in a motor vehicle, but may have been a pedestrian, bicyclist, or using another non-motor vehicle mode of transportation. Injuries reported are reported by the responding police officer. Fatalities that occur after the initial reports are typically updated in these records up to 30 days after the date of the crash.
- In the "**Traffic_Crashes-_-Crashes**" dataset, we see information about each traffic crash in Chicago for which the Chicago Police Department responded. Data are available for some police districts in 2015, but citywide data are not available until September

2017. Many crash parameters are reported by the responding police officer, including street condition data, weather condition, and posted speed limits.
- In the "**wave_3_fertility**" dataset, each record corresponds to a woman who was surveyed regarding her fertility history. She is marked by her household (FPrimary) and her number within the household (hhmid). Information on these women's interactions with the power structures around them, largely with respect to patriarchal norms, is documented in the "**wave_3_power**" dataset. The whole project documented in the above below covers three waves of surveying dating back to 2008.

Data Sources

- The traffic datasets were obtained from the City of Chicago data portal, found here.
- These two wave datasets are from the most recent wave in 2017. Most of the answers are coded as a number and can be referenced in the codebook in the dataverse link here: Harvard Dataverse.

Programming in R Questions

Question 1 ~ 2

Please follow the instructions in the HW description to download and load the dataset: "Traffic_Crashes_-_People.csv". Spend some time exploring the names of the different columns of the datasets, and how they are organized.

Among female sex individuals of age 65 and above in this dataset, what proportion are pedestrians?

Hint**:** Look at columns entitled AGE, SEX, and PERSON_TYPE. After selecting in the dataset only observations with female sex and age equal to or above 65, calculate what for what proportion PERSON_TYPE is PEDESTRIAN.

Question 3 ~ 4

Again using the dataset entitled "Traffic_Crashes_-_People.csv", answer the following.

Filter the dataset to only those with no indication of injury. How old is the oldest person in this filtered dataset? Use the INJURY_CLASSIFICATION and AGE columns to answer this question.

Question 5 ~ 6

Again using the dataset entitled "Traffic_Crashes_-_People.csv", answer the following.

How many people in the dataset are age 25 or younger and are passengers? Use PERSON_TYPE and AGE to answer this question.

Question 7 ~ 8

Begin with the "Traffic_Crashes_-_People.csv" dataset.

- Filter for a subset of observations pertaining to individuals who are age 65 or older.
- Next, Join with the crashes dataset.
- Get the year from CRASH_DATE.

How many distinct people age 65 or above were involved in a car crash during the year 2020?

Question 9

Your task is to plot the number of people who died in car crashes over time (every year) and submit an image of your final result in a lineplot, as well as your code in the following question.

- Begin with the people dataset, and keep only the observations where the variable INJURY_CLASSIFICATION indicates a **fatal injury**.
- Next, join with the crashes dataset and get the year associated with each person who got into fatal car accident.
- Use ggplot to make a lineplot which illustrates the number of people who died in car crashes every year.

Question 10 ~ 11

Note: For this question, please follow the instructions in the HW description to download and load the dataset: "wave_3_fertility.csv". Spend some time exploring the names of the different columns of the datasets, and how they are organized. For many of the questions below, make sure to refer to the codebook linked in page 1.

Among women age 40 and above who have given birth to a child in this dataset, what proportion have never given birth to a girl?

Hint*:* Look at columns entitled age, evergivenbirth, and borngirls. After selecting in the dataset only observations who have given birth and age equal to or above 40, calculate what for what proportion borngirls is zero.

Question 12 ~ 13

Again using the dataset entitled "wave_3_fertility.csv", answer the following.

How many women in the dataset are age 20 or younger and have been pregnant in the past 12 months? Use pregnantlastyear and age to answer this question.

Question 14 ~ 15

Begin with the "wave_3_fertility.csv" dataset.

- Filter to observations of women that have children (borntotal >= 0).
- Next, join with the "wave_3_power.csv" dataset along two columns, FPrimary and hhmid. These correspond to household and household member respectively.
- Count the number of rows

How many respondents with only daughter(s) believe it is better to send a son to school than a daughter (bettersontoschool = 1)? What if the respondent has only son(s)?

Question 16

Your task is to plot a histogram of the number of children born to women who can no longer have children and submit an image of your final result, as well as your code in the following question.

- Begin with the fertility dataset and keep only the observations where the variable age-menopause is not equal to -1.
- Use ggplot to make a scatterplot which shows number of children born to women who are different ages. You must properly label your graph.
- You must include a title and labels on the x and y axes and pick a fill color of your choice for the bars.