THE UNIVERSITY OF CHICAGO
HARRIS SCHOOL
OF PUBLIC POLICY

Homework 4: Programming in R Sequence

DPSS 2024

Instructions

- There will be 12 programming in R questions, most are worth 1.85 and Question 12 is worth 2 points.
- 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.

You may submit your assignment up to 3 times before the deadline. Only your most recent submission will be saved, so be careful when resubmitting. It is recommended that you write your answers down just in case you accidentally clear your previous submission.The Ed Discussion page is the best place to post questions for a quick response from the teaching assistants.

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

Set Up

For HW 4 you will use three datasets:

1. Speed_Camera_Violations.csv
2. chetty_2000.csv
3. most_states.shp
    - keep all most_states files in the same folder or .shp will not work

Load the following packages and datasets to complete the programming in R questions:

```r
# Libraries for today
library(tidyverse)
library(sf)
library(ggthemes) # this is optional and offers more themes
file_path = "your_file_path"
setwd(file_path)
speed_tickets <- #speed_tickets is the dataframe for the violations
  read_csv("Speed_Camera_Violations.csv")
chetty <- read.csv("chetty_2000.csv")
us <- st_read("most_states/most_states.shp")
```

About the Data:

Speed Camera Violations: This dataset reflects the daily volume of violations that have occurred in Children's Safety Zones for each camera. The data reflects violations that occurred from July 1, 2014 until present (we cut it off at 2021). This data may change due to occasional time lags between the capturing of a potential violation and the processing and determination of a violation. The most recent 14 days are not shown due to revised data being submitted to the City of Chicago.

The reported violations are those that have been collected by the camera and radar system and reviewed by two separate City contractors. In some instances, due to the inability the registered owner of the offending vehicle, the violation may not be issued as a citation. However, this dataset contains all violations regardless of whether a citation was issued, which provides an accurate view into the **Automated Speed Enforcement Program** violations taking place in Children's Safety Zones. Visit Children's Safety Zone Program & Automated Speed Enforcement to learn more.

The data is live here

Chetty et. al (2014): This data is from a study *Where is the land of Opportunity? The Geography of Intergenerational Mobility in the United States*. Includes socioeconomic variables and shape file data.

Programming in R Questions

Questions 1 ~ 2

Your goal is to analyze Chicago's new speed cameras policy: according to Chicago Police Department, traffic fatalities increased by 45%, with 139 deaths in 2020. As a result, the mayor decided to enforce reduced speed limits on March 1, 2021 ($35 tickets to vehicles that exceed the speed limit by 6-10 miles/h and a $100 fine for speed limit by more than 10 miles/h).

We will begin our analysis with the Speed Camera Violations dataset.

Load the speed_camera_violations dataset, saved under Week 4 HW materials.

Then, write the code to generate a new dataframe that counts the total number of speed camera violations every year. Only consider from 2018 - 2021, so you should have four rows in the data: 2018, 2019, 2020, and 2021. How many violations took place in 2021?

Select one answer from below:

- 767,235
- 800,594
- 39,601
- 3,077,604

Questions 3 ~ 4

Now, create a dataframe which shows the monthly volume of speed camera violations in Children's Safety Zones - Chicago. Again, focus your analysis only on observations pertaining to year 2018-2021. How many speed camera violations were in February, 2021, the month right before the policy took effect?

Select one answer from below:

- 103444
- 66870
- 3485
- 27289
- None of these answer choices

Questions 5 ~ 6

For the remainder of this problem set, we will try to study the impact of this policy on the daily volume of violations that have occurred in Children's Safety Zones for each camera.

Let's visualize the data. Start with the dataframe which you have created in the previous problem but before subsetting for Feburary 2021.

Please generate a line graph which depicts the following:

1. The x-axis will be labeled with the month names or numbers, including a separate label for every month. You should have 12 ticks in total on the x-axis, ranging from January to December.
2. The y-axis will be: Total Camera Violations (in thousands).
3. You will have four different lines, to reflect each year (color-coded separately).
   - Please make sure that color **is not** on a continuous scale. Each line will trace out the total monthly number of speed camera violations in that year.
4. **If you do not meet the aforementioned requirements you will not receive full credit, this needs to be a professional looking graph.**

Recall that the policy took place in March 2021, so if there was an effect of the policy on speed camera violations, you should be able to see this by looking at the 2021 line.

Submit this image, as an image file or PDF.

Questions 7 ~ 8

Research Design Background

If one were to implement a statistical model to study the impact of the policy, there are many possible approaches one might take. We will organize your data in this question so that we can estimate the following equation in a future question:

$$Y_{it} = \beta_1 (Post)_{it} + \sum_{i=1}^{N} \alpha_i (CameraID)_i + \sum_{t=1}^{12} \phi_t (month)_t + v_{it}$$

- The outcome is the **number of camera violations**, indexed by the camera ID and by day.
- The variable *Post* is a binary variable to reflect the policy. It is 0 all days before the policy takes place, and 1 all days after the policy takes place.
- The design then includes two types of fixed effects: calendar month and entity (Camera ID).

This design measures time (t) at the daily level and the treatment takes effect beginning from March 1, 2021.

Task

Before running the model, your task for this question will be to organize the dataset in a manner that is conducive for running the above regression.

1. Begin with the speed_tickets dataset.
2. Only keep observations pertaining to the speed cameras which are first observed on or before January 1, 2018. In other words, if the earliest day that a particular CAMERA ID is observed in the dataset is after January 1, 2018, then all observations associated with that camera will be dropped from analysis.

    - Hint: 0) make sure you format your date 1) use `min()`, to help you find the earliest date of a speed camera's data upload to the server 2) if you are able to identify which CAMERA ID's meet this requirement, then you can conduct a merge/filter to restrict the speeds_tickets dataset to only include observations pertaining to those cameras.

3. Next, only include the columns pertaining to CAMERA ID, violation date, and violations. Note that date should be converted to a lubridate format.
4. Reorganize the dataset in the following manner:

    - Only include observations which take place in years 2018, 2019, 2020, and 2021.
    - Construct a new column entitled "post." This is a binary treatment variable which is equal to 0 if the date is before March 1, 2021, and 1 if the date is on or after March 1, 2021.

- Construct a column containing the month, in numerical format (numbers 1 - 12) or word format (January to December).

Your final dataset should contain five columns, pertaining to: Camera ID, violations, date, post, and month.

What is the total number of rows for this dataset?

Select one answer from below:

- 297,932
- 288,316
- 140,692
- None of the above

Question 9

Finally, please run the regression model given in the previous equation. Your dependent variable is the number of violations observed by the camera on a given day. Your independent variables are post (the treatment variable) and fixed effects for month and Camera ID. Please provide your code, output and a original interpretation of the policy impact for complete credit.

If you were unable to fully complete the previous question, you may still receive partial credit if you write the code you would use to run this regression.

Question 10

For this questions and the ones here after we will new datasets. In the homework materials, there is a map of the contiguous US states (most_states.shp) from the census and a country level data from Chetty et al (2014)(chetty_2000.csv)

The first we'll load using read.csv and the second is a shapefile:

```
chetty <- read.csv("chetty_2000.csv")
# keep all most_states files in the same folder
# or .shp below will not read in.
 us <- st_read("most_states/most_states.shp")
```

Show us how to set your working directory and read in your files

Question 11

As a warm up, find out how many counties in Tennessee have a divorce rate (cs_divorced) below 12%? Hint: use columns cs_divorced and stateabbrv.

Question 12

For this question your task is to generate a map of the **average divorce rate by state** in the US

Step 1) Begin by generating a new dataframe from the county level data with the following outputs:

1) groups outcomes by the state
2) calculates the average divorce rate across counties within a given state

Step 2) Use your grouped dataframe created in step 1 as your "y" data frame and join it with the map data using the state column as the common key column. Hint: use a join that only keeps observations that have a match and drops the rest.

Step 3) Map divorce rate across the U.S. Hint: 1) consider using `geom_sf` to generate a map. 2) You may plot the rate as is or generate a quantile map using `ntile()` 3) Consider filtering out Alaska so that the map looks centered on the U.S.

Upload an image of your map.

Hint: Feel free to reference 'L9 - Spatial.R' in the 2.8 - Materials zip folder for Week 4.