Midterm Project 1

Small Business Data Set Version
MTH 3270 Data Science
Due Fri., Mar. 19

Rules

You must do your own work, and you're only allowed to speak about this project with the instructor (Grevstad).

All analyses (data wrangling, visualizations, statistical summaries, etc.) must be done using **R** (except by permission of the instructor).

The projects are **due** in **Canvas** as **pdf** file no later than **Friday, Mar. 19, 2021** at **11:59 PM**.

Instructions

The project will use the data set from the Sixth International Conference on Establishment Statistics (ICES VI) student contest focusing on the analysis/visualization of economic data from the 2007 Survey of Business Owners.

The **data set** and a **data dictionary** describing the variables in it are obtained via the links below. Save the **csv** file and read it into R using `read.csv()` (and don't forget `header = TRUE` and `stringsAsFactors = FALSE`). Check **Canvas Announcements** and/or your **email** regularly in case there are important announcements about this project.

The **data set** is here:

`https://ww2.amstat.org/meetings/ices/2021/studentcontest/track2sbo.csv`

A **data dictionary** is here:

`https://ww2.amstat.org/meetings/ices/2021/pdfs/contestdata_DataDictionary.pdf`

More **information** about the data and student contest can be found here:

`https://ww2.amstat.org/meetings/ices/2021/studentcontest.cfm`

You *might* need to do some data wrangling and tidying (which *might* involve selecting columns, adding new columns, filtering rows, grouping by a categorical variable, etc.).

Tasks

There are **two research questions**:

Q1 How does the extent to which income derived from the small business is not the primary source of income for a business owner vary by owner's sex, ethnicity, race, and veteran status and by business characteristics (e.g., size, sector, location)?

Q2 How does the business size (establishment employment, establishment payroll, establishment receipts) vary by owner's sex, ethnicity, race, and veteran status and by other business characteristics (e.g., sector, location)?

Your **tasks** are to address the **research questions** (**Q1** and **Q2**) using **both** of the following:

1. Create **visualizations** (graphical displays) pertinent to answering the research questions. The graphics must provide **context** (via titles, axis labels, legends, etc.).
2. Produce **statistical summaries** (or other statistical analyses) pertinent to answering the research questions. The summaries must be presented in **table** format.

What to Turn In

1. A **write-up** as a **pdf** file (perhaps 3-7 pages including graphs and tables) containing:
   (a) A **brief description** (e.g. 1-2 paragraphs) of any data wrangling and tidying you had to do in order to carry out tasks **1** and **2** above.
   (b) Your **graphical displays** and **statistical summary tables**.
   (c) Your **conclusions** regarding questions **Q1** and **Q2**.
2. Your **R code** with **comments** (use #) indicating **what** each chunk of code does and **why** it does it, either as an **appendix** in your **write-up pdf** or as a separate **.R file** (as produced by RStudio's script editor).

Grading

Your **grade** will be based on:

1. Your attainment of **tasks 1-2** (described above).
2. Your **write-up**, including the graphs and summary statistic tables (as described above).
3. The inclusion of and correctness of your **commented R code**.