Midterm Project 1

Racial and Ethnic Representativeness Data Sets
MTH 3270 Data Science
Due Wed., Mar. 9

Rules

You may work alone or with a partner from the class. You're only allowed to communicate about this project with the instructor (Grevstad) or your partner if you are working with one. If you work with a partner, the two of you will submit the same project and receive the same score.

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 a **pdf** file no later than **Wednesday, Mar. 9, 2021** at **11:59 PM**.

Instructions

The project will use the **Racial and Ethnic Representativeness of US Postsecondary Education Institutions** data sets from the annual Data Challenge Expo contest sponsored by the American Statistical Association:

1) **HEsegDataviz_CollegeData_4-year_v5.csv** This dataset combines public data from the Integrated Postsecondary Education Data System and the US Census Bureau's American Community Service in an index of racial and ethnic representativeness of US postsecondary education **four-year** institutions. The data link college racial composition to the racial composition of an institution's "market," defined geographically according to institutions' level, degree of selectivity, and urbanicity.

2) **HEsegDataviz_CollegeData_2-year_v5.csv** The same as HEsegDataviz_CollegeData_4-year_v5.csv, but for **two-year** institutions.

The **data sets** and a **data dictionary** (**HEsegDataviz_Dictionary.xlsx**) containing **descriptions** of the **variables** in the data sets are obtained via the link below. Save one or the other of the **csv** files containing the data 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.

`community.amstat.org/dataexpo/home`

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, recoding, etc.).

Tasks

There are **three research questions**:

Q1 Overall, to what degree do college racial and ethnic compositions differ from the racial and ethnic compositions of the institutions' geographic "markets"?

Q2 For which specific racial or ethnic groups are the discrepancies between their representations in colleges and their representations in the "markets" largest?

Q3 If colleges are *grouped* by institution level, degree of selectivity, and/or public/private/for profit status, do the discrepancies between college and "market" racial and ethnic compositions vary across groups? In other words are the discrepancies larger for some types of colleges than others? If so, for which types of colleges are the discrepancies largest?

Your **tasks** are to use the **Racial and Ethnic Representativeness of US Postsecondary Education Institutions** data sets to address the **research questions** (**Q1-Q3**) using **both** of the following:

T1 Create **visualizations** (graphical displays) satisfying the following criteria:

- The graphs must be **pertinent** to answering the **research questions** (**Q1-Q3**).
- You must have **at least one** graph addressing **each research question** (**Q1-Q3**), but you may have more than that.
- The graphs must provide **context** (via titles, axis labels, legends, etc.).
- **At least one** graph must display **three or more variables**.

T2 Produce **tables** containing **statistical summaries** (or other statistical analyses) satisfying the following criteria:

- The **statistical summaries** in the **tables** must be **pertinent** to answering the **research questions** (**Q1-Q3**).
- You must have **at least one** table addressing **each research question** (**Q1-Q3**), but it's okay for more than one research question to be addressed by the same table.

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** (at most 1-2 paragraphs) of any data **wrangling** and **tidying** you had to do in order to carry out tasks **T1** and **T2**.
   (b) Your **graphical displays** and **statistical summary tables** (**T1** and **T2**).
   (c) Your **conclusions** regarding questions **Q1-Q3**.
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 level of attainment of tasks **T1** and **T2**.
2. Your **write-up**, and in particular, the inclusion of your **graphs** and **summary statistic tables** as well as the *breadth* of your **conclusions** regarding **Q1-Q3**.
3. The inclusion of and correctness of your **commented R code**.