Midterm Project 2

Racial and Ethnic Representativeness Data Sets
MTH 3270 Data Science
Due Mon., Apr. 11

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 **Monday, April 11, 2022** 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`

Note: Because each college appears in multiple rows of the data sets (once for each of the years 2009-2017), you may pick one of the years (2017 would be a good choice), filter out those rows (using `filter()`), and do the entire project using data for just that one year.

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

Tasks

Your **tasks** are:

T1 Carry out a **multiple regression analysis** with a **minimum** of **three explanatory (X) variables** in the model. You may choose any response variable (Y) for your model, but it must be a numerical variable (*not* categorical). Likewise, you may use any explanatory (X) variables, but they too must be numerical (*not* categorical). Note that a categorical variable that's been coded using integer values is still considered to be a *categorical* variable.

- **Summarize** your fitted model by reporting the estimated model coefficients.
- **Interpret** the estimated model (coefficients).
- **Report** and **discuss** the values of at least **two** measures of **how well** the model **fits** the data (e.g. the R^2 and $\sqrt{\text{MSE}}$).

T2 Carry out a **logistic regression analyses** with a **minimum** of **two explanatory (X) variables** in the model.

For the response (Y) variable, you'll use one of the *dichotomous* (**0** or **1**) variables (your choice):

→ `forprofit`.

→ `public`.

→ `private`.

→ `selective`.

→ `more_selective`.

→ `non_selective`.

You may use any explanatory (X) variable(s), but they must be numerical (*not* categorical).

- **Summarize** your fitted model by reporting the estimated model coefficients.

T3 Carry a *machine learning* **classification** procedure (either **decision tree**, **random forest**, ***k* nearest neighbor**, or **artificial neural network** – your choice) for **predicting** the **Four-year Institution Category** (`fourcat`) using a **minimum** of **three explanatory (*X*) variables** in the model. You may use any explanatory (X) variables, but they must be numerical (*not* categorical).

Then

- **Summarize** your procedure: Indicate *which* **classification procedure** you used and *which* **explanatory variables** you used.
- **Report** the value of at least one measure of **how well** the model **predicts** (**classifies**) individuals (e.g. the *accuracy*, i.e. *correct classification rate*).
- **Provide** an **example** of a **prediction** (**classification**) using your fitted classification model.

What to Turn In

1. A **write-up** as a **pdf** file (perhaps 3-7 pages) 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**, **T2**, and **T3**.
   (b) Your **responses** addressing the **bullet items** under tasks **T1**, **T2**, and **T3** above (*seven* bullet items total).
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**, **T2**, and **T3**.
2. Your **write-up**, and in particular, the inclusion of your **responses** addressing the seven **bullet items** (as described above).
3. The inclusion of and correctness of your **commented R code**.