# Exploratory Analysis of U.S. Education Dataset

Shankar Prabhu

July 2020

The dataset we will use is linked at: https://www.kaggle.com/noriuk/us-education-datasets-unification-project?select=states_all.csv. Let's start by reading in the dataset and generating a few quick summaries of the columns.

```
data = read.csv(
  "C:/Users/shank/Documents/Data_Science_Projects/education/states_all_extended.csv"
                , header = TRUE, sep = ",")
nrow(data)
```

```
## [1] 1715
```

```
ncol(data)
```

```
## [1] 266
```

```
names(data)
```

```
##   [1] "PRIMARY_KEY"                   "STATE"
##   [3] "YEAR"                          "ENROLL"
##   [5] "TOTAL_REVENUE"                 "FEDERAL_REVENUE"
##   [7] "STATE_REVENUE"                 "LOCAL_REVENUE"
##   [9] "TOTAL_EXPENDITURE"             "INSTRUCTION_EXPENDITURE"
##  [11] "SUPPORT_SERVICES_EXPENDITURE" "OTHER_EXPENDITURE"
##  [13] "CAPITAL_OUTLAY_EXPENDITURE"   "A_A_A"
##  [15] "G01_A_A"                      "G02_A_A"
##  [17] "G03_A_A"                      "G04_A_A"
##  [19] "G05_A_A"                      "G06_A_A"
##  [21] "G07_A_A"                      "G08_A_A"
##  [23] "G09_A_A"                      "G10_A_A"
##  [25] "G11_A_A"                      "G12_A_A"
##  [27] "KG_A_A"                       "PK_A_A"
##  [29] "G01.G08_A_A"                  "G09.G12_A_A"
##  [31] "G01_AM_F"                     "G01_AM_M"
##  [33] "G01_AS_F"                     "G01_AS_M"
##  [35] "G01_BL_F"                     "G01_BL_M"
##  [37] "G01_HI_F"                     "G01_HI_M"
##  [39] "G01_HP_F"                     "G01_HP_M"
##  [41] "G01_TR_F"                     "G01_TR_M"
##  [43] "G01_WH_F"                     "G01_WH_M"
##  [45] "G02_AM_F"                     "G02_AM_M"
```

```
##  [47] "G02_AS_F"                  "G02_AS_M"
##  [49] "G02_BL_F"                  "G02_BL_M"
##  [51] "G02_HI_F"                  "G02_HI_M"
##  [53] "G02_HP_F"                  "G02_HP_M"
##  [55] "G02_TR_F"                  "G02_TR_M"
##  [57] "G02_WH_F"                  "G02_WH_M"
##  [59] "G03_AM_F"                  "G03_AM_M"
##  [61] "G03_AS_F"                  "G03_AS_M"
##  [63] "G03_BL_F"                  "G03_BL_M"
##  [65] "G03_HI_F"                  "G03_HI_M"
##  [67] "G03_HP_F"                  "G03_HP_M"
##  [69] "G03_TR_F"                  "G03_TR_M"
##  [71] "G03_WH_F"                  "G03_WH_M"
##  [73] "G04_AM_F"                  "G04_AM_M"
##  [75] "G04_AS_F"                  "G04_AS_M"
##  [77] "G04_BL_F"                  "G04_BL_M"
##  [79] "G04_HI_F"                  "G04_HI_M"
##  [81] "G04_HP_F"                  "G04_HP_M"
##  [83] "G04_TR_F"                  "G04_TR_M"
##  [85] "G04_WH_F"                  "G04_WH_M"
##  [87] "G05_AM_F"                  "G05_AM_M"
##  [89] "G05_AS_F"                  "G05_AS_M"
##  [91] "G05_BL_F"                  "G05_BL_M"
##  [93] "G05_HI_F"                  "G05_HI_M"
##  [95] "G05_HP_F"                  "G05_HP_M"
##  [97] "G05_TR_F"                  "G05_TR_M"
##  [99] "G05_WH_F"                  "G05_WH_M"
## [101] "G06_AM_F"                  "G06_AM_M"
## [103] "G06_AS_F"                  "G06_AS_M"
## [105] "G06_BL_F"                  "G06_BL_M"
## [107] "G06_HI_F"                  "G06_HI_M"
## [109] "G06_HP_F"                  "G06_HP_M"
## [111] "G06_TR_F"                  "G06_TR_M"
## [113] "G06_WH_F"                  "G06_WH_M"
## [115] "G07_AM_F"                  "G07_AM_M"
## [117] "G07_AS_F"                  "G07_AS_M"
## [119] "G07_BL_F"                  "G07_BL_M"
## [121] "G07_HI_F"                  "G07_HI_M"
## [123] "G07_HP_F"                  "G07_HP_M"
## [125] "G07_TR_F"                  "G07_TR_M"
## [127] "G07_WH_F"                  "G07_WH_M"
## [129] "G08_AM_F"                  "G08_AM_M"
## [131] "G08_AS_F"                  "G08_AS_M"
## [133] "G08_BL_F"                  "G08_BL_M"
## [135] "G08_HI_F"                  "G08_HI_M"
## [137] "G08_HP_F"                  "G08_HP_M"
## [139] "G08_TR_F"                  "G08_TR_M"
## [141] "G08_WH_F"                  "G08_WH_M"
## [143] "G09_AM_F"                  "G09_AM_M"
## [145] "G09_AS_F"                  "G09_AS_M"
## [147] "G09_BL_F"                  "G09_BL_M"
## [149] "G09_HI_F"                  "G09_HI_M"
## [151] "G09_HP_F"                  "G09_HP_M"
## [153] "G09_TR_F"                  "G09_TR_M"
```

```
## [155] "G09_WH_F"                 "G09_WH_M"
## [157] "G10_AM_F"                 "G10_AM_M"
## [159] "G10_AS_F"                 "G10_AS_M"
## [161] "G10_BL_F"                 "G10_BL_M"
## [163] "G10_HI_F"                 "G10_HI_M"
## [165] "G10_HP_F"                 "G10_HP_M"
## [167] "G10_TR_F"                 "G10_TR_M"
## [169] "G10_WH_F"                 "G10_WH_M"
## [171] "G11_AM_F"                 "G11_AM_M"
## [173] "G11_AS_F"                 "G11_AS_M"
## [175] "G11_BL_F"                 "G11_BL_M"
## [177] "G11_HI_F"                 "G11_HI_M"
## [179] "G11_HP_F"                 "G11_HP_M"
## [181] "G11_TR_F"                 "G11_TR_M"
## [183] "G11_WH_F"                 "G11_WH_M"
## [185] "G12_AM_F"                 "G12_AM_M"
## [187] "G12_AS_F"                 "G12_AS_M"
## [189] "G12_BL_F"                 "G12_BL_M"
## [191] "G12_HI_F"                 "G12_HI_M"
## [193] "G12_HP_F"                 "G12_HP_M"
## [195] "G12_TR_F"                 "G12_TR_M"
## [197] "G12_WH_F"                 "G12_WH_M"
## [199] "KG_AM_F"                  "KG_AM_M"
## [201] "KG_AS_F"                  "KG_AS_M"
## [203] "KG_BL_F"                  "KG_BL_M"
## [205] "KG_HI_F"                  "KG_HI_M"
## [207] "KG_HP_F"                  "KG_HP_M"
## [209] "KG_TR_F"                  "KG_TR_M"
## [211] "KG_WH_F"                  "KG_WH_M"
## [213] "PK_AM_F"                  "PK_AM_M"
## [215] "PK_AS_F"                  "PK_AS_M"
## [217] "PK_BL_F"                  "PK_BL_M"
## [219] "PK_HI_F"                  "PK_HI_M"
## [221] "PK_HP_F"                  "PK_HP_M"
## [223] "PK_TR_F"                  "PK_TR_M"
## [225] "PK_WH_F"                  "PK_WH_M"
## [227] "G04_A_A_READING"          "G04_A_A_MATHEMATICS"
## [229] "G04_A_M_READING"          "G04_A_M_MATHEMATICS"
## [231] "G04_A_F_READING"          "G04_A_F_MATHEMATICS"
## [233] "G04_WH_A_READING"         "G04_WH_A_MATHEMATICS"
## [235] "G04_BL_A_READING"         "G04_BL_A_MATHEMATICS"
## [237] "G04_HI_A_READING"         "G04_HI_A_MATHEMATICS"
## [239] "G04_AS_A_READING"         "G04_AS_A_MATHEMATICS"
## [241] "G04_AM_A_READING"         "G04_AM_A_MATHEMATICS"
## [243] "G04_HP_A_READING"         "G04_HP_A_MATHEMATICS"
## [245] "G04_TR_A_READING"         "G04_TR_A_MATHEMATICS"
## [247] "G08_A_A_READING"          "G08_A_A_MATHEMATICS"
## [249] "G08_A_M_READING"          "G08_A_M_MATHEMATICS"
## [251] "G08_A_F_READING"          "G08_A_F_MATHEMATICS"
## [253] "G08_WH_A_READING"         "G08_WH_A_MATHEMATICS"
## [255] "G08_BL_A_READING"         "G08_BL_A_MATHEMATICS"
## [257] "G08_HI_A_READING"         "G08_HI_A_MATHEMATICS"
## [259] "G08_AS_A_READING"         "G08_AS_A_MATHEMATICS"
## [261] "G08_AM_A_READING"         "G08_AM_A_MATHEMATICS"
```

```
## [263] "G08_HP_A_READING"             "G08_HP_A_MATHEMATICS"
## [265] "G08_TR_A_READING"             "G08_TR_A_MATHEMATICS"
```

```
min(data$YEAR)
```

```
## [1] 1986
```

```
max(data$YEA)
```

```
## [1] 2019
```

```
unique(data$STATE)
```

```
##  [1] ALABAMA              ALASKA               ARIZONA
##  [4] ARKANSAS             CALIFORNIA           COLORADO
##  [7] CONNECTICUT          DELAWARE             DISTRICT_OF_COLUMBIA
## [10] FLORIDA              GEORGIA              HAWAII
## [13] IDAHO                ILLINOIS             INDIANA
## [16] IOWA                 KANSAS               KENTUCKY
## [19] LOUISIANA            MAINE                MARYLAND
## [22] MASSACHUSETTS        MICHIGAN             MINNESOTA
## [25] MISSISSIPPI          MISSOURI             MONTANA
## [28] NEBRASKA             NEVADA               NEW_HAMPSHIRE
## [31] NEW_JERSEY           NEW_MEXICO           NEW_YORK
## [34] NORTH_CAROLINA       NORTH_DAKOTA         OHIO
## [37] OKLAHOMA             OREGON               PENNSYLVANIA
## [40] RHODE_ISLAND         SOUTH_CAROLINA       SOUTH_DAKOTA
## [43] TENNESSEE            TEXAS                UTAH
## [46] VERMONT              VIRGINIA             WASHINGTON
## [49] WEST_VIRGINIA        WISCONSIN            WYOMING
## [52] DODEA                NATIONAL
## 53 Levels: ALABAMA ALASKA ARIZONA ARKANSAS CALIFORNIA COLORADO ... WYOMING
```

Overall, looks like we have education data from 1986 to 2019 (~33 years). The data is organized by (state, year) pairs, and we have 53 unique values for state (Department of Defense, D.C., etc.).

There are three sections for the data: funding/spending, enrollment demographics, testing demographics.

Funding/Spending includes columns like "FEDERAL_REVENUE" or "INSTRUCTION_EXPENDITURE", and will be helpful in understanding education finances over time.

Enrollment demographics are organized into columns with three parts to their name (ex: "GO2_AS_F"). The first part refers to the grade level (ex. GO2 is grade 2), the second part refers to the race (ex: AS means Asian, there are 7 different racial categories), and the third part is gender (ex. F for female). If one of these parts is "A" it refers to all students, so "A_A_A" means all students enrolled in that state for some year. These columns will show how demograhpics have changed in U.S. education in the past 30 years.

Testing demographics use a similar categorization system as enrollment demograhpics, but it also adds a fourth part to refer to the average "READING" or "MATHEMATICS" test score for that particular group (on the NAEP exam). These columns will help us understanding student performance on standardized tests through time.

**As a result, we will break down the analysis into these three sections of the data, and then bring them together for some final conclusions.**