Tidying and Joining Data Project

Ken Wood

8/11/2024

In Part Three of the project, you'll take the information presented in two data frames and create a single, tidy data set that contains all the variables you need and is ready for analysis.

The file state_education_and_income.csv lists the proportion of adults (25 and older) who had earned a Bachelor's degree by 2019 and the median income in 2019. This data is provided for most states in the U.S., as well as the District of Columbia and Puerto Rico. The file state_poverty_and_population.csv gives the poverty rate and population of each state in 2019. Run the following code chunk to load the tidyverse and view these data sets:

```r
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become error
```

```r
library(ellipsis)
library(utf8)
edu <- read.csv("state_education_and_income.csv", check.names = FALSE)
pop <- read.csv("state_poverty_and_population.csv", check.names = FALSE)
head(edu)
```

```
##    Measurement  Alabama  Alaska  Arizona Arkansas California Colorado
## 1  CollegeRate     0.25     0.3     0.29     0.23       0.34     0.41
## 2 MedianIncome 51771.00 77203.0 62027.00 49020.00   80423.00 77104.00
##   Connecticut Delaware District of Columbia Florida  Georgia   Hawaii    Idaho
## 1        0.39     0.32                 0.59     0.3     0.31     0.33     0.28
## 2    78920.00 70348.00             90395.00 59198.0 61950.00 83734.00 60830.00
##   Illinois  Indiana     Iowa   Kansas Kentucky Louisiana    Maine Maryland
## 1     0.35     0.26     0.29     0.33     0.24      0.24     0.32      0.4
## 2 69212.00 57617.00 61807.00 62028.00 52256.00  51108.00 58824.00  86644.0
##   Massachusetts Michigan Minnesota Mississippi Missouri  Montana Nebraska
## 1          0.44     0.29      0.36        0.22     0.29     0.32     0.32
## 2      85700.00 59522.00  74529.00    45928.00 57375.00 57248.00 63290.00
##     Nevada New Hampshire New Jersey New Mexico New York North Carolina
## 1     0.25          0.37        0.4       0.27     0.37           0.31
## 2 63268.00      78571.00    85786.0   52021.00 72038.00       57388.00
##   North Dakota     Ohio Oklahoma   Oregon Pennsylvania Rhode Island
## 1          0.3     0.28     0.26     0.34         0.31         0.34
## 2      67402.0 58704.00 54447.00 66955.00     63455.00     70383.00
##   South Carolina South Dakota Tennessee   Texas     Utah  Vermont Virginia
## 1           0.28         0.29      0.27     0.3     0.34     0.38     0.39
## 2       56360.00     60414.00  56047.00 64044.0 75705.00 63293.00 76471.00
##   Washington West Virginia Wisconsin  Wyoming Puerto Rico
## 1       0.36          0.21       0.3     0.27        0.26
## 2   78674.00      48659.00   64177.0 66152.00          NA
```

```r
head(pop)
```

```
##   Measurement   Alabama   Alaska   Arizona Arkansas California  Colorado
## 1 PovertyRate      15.6     10.2      13.5       16       11.8       9.4
## 2  Population 4903185.0 731545.0 7278717.0  3017804 39512223.0 5758736.0
##   Connecticut Delaware District of Columbia    Florida    Georgia  Hawaii
## 1         9.9     11.2                 14.1       12.7       13.5       9
## 2   3565287.0 973764.0             705749.0 21477737.0 10617423.0 1415872
##     Idaho   Illinois   Indiana     Iowa     Kansas Kentucky Louisiana     Maine
## 1      11       11.4      11.9       11       11.3       16      18.8      10.9
## 2 1787065 12671821.0 6732219.0 3155070 2913314.0  4467673 4648794.0 1344212.0
##    Maryland Massachusetts  Michigan Minnesota Mississippi  Missouri   Montana
## 1       9.1           9.5      12.9       8.9        19.5      12.9      12.6
## 2 6045680.0     6892503.0 9986857.0 5639632.0   2976149.0 6137428.0 1068778.0
##    Nebraska    Nevada New Hampshire New Jersey New Mexico   New York
## 1       9.9      12.7           7.5        9.1       17.5       13.1
## 2 3080156.0 1359711.0     8882190.0  2096829.0 19453561.0 10488084.0
##   North Carolina North Dakota    Ohio  Oklahoma     Oregon Pennsylvania
## 1           13.6         10.5      13      15.1       11.5           12
## 2       762062.0   11689100.0 3956971 4217737.0 12801989.0      1059361
##   Rhode Island South Carolina South Dakota  Tennessee     Texas     Utah
## 1         11.6           13.9         11.9       13.8      13.6      8.8
## 2    5148714.0       884659.0    6829174.0 28995881.0 3205958.0 623989.0
##     Vermont  Virginia Washington West Virginia Wisconsin   Wyoming
## 1      10.1       9.9        9.8          16.2      10.4       9.9
## 2 8535519.0 7614893.0  1792147.0     5822434.0  578759.0 3193694.0
```

Question 1

For the edu data set, the college completion rate and median income are provided on the rows and each state is listed on a column. Create a tidy version of this data set that has each state listed on a different row and has the college completion rate and the median income in separate columns.

```r
edu1 <-  pivot_longer(data = edu,
                      cols = -Measurement,
                      names_to = "State")

edu2 <- pivot_wider(data = edu1,
                     names_from = Measurement,
                     values_from = value)

head(edu2)  # Dataframe edu2 is the desired result.
```

```
## # A tibble: 6 x 3
##   State      CollegeRate MedianIncome
##   <chr>            <dbl>        <dbl>
## 1 Alabama           0.25        51771
## 2 Alaska            0.3         77203
## 3 Arizona           0.29        62027
## 4 Arkansas          0.23        49020
## 5 California        0.34        80423
## 6 Colorado          0.41        77104
```

Question 2

For the pop data set, the poverty rate and population are provided on the rows and each state is listed on a column. Create a tidy version of this data set that has each state listed on a different row and the poverty rate and the population in separate columns.

```
pop1 <-  pivot_longer(data = pop,
                      cols = -Measurement,
                      names_to = "State")

pop2 <- pivot_wider(data = pop1,
                     names_from = Measurement,
                     values_from = value)  # Dataframe pop2 is the desired result.

head(pop2)
```

```
## # A tibble: 6 x 3
##   State      PovertyRate Population
##   <chr>            <dbl>      <dbl>
## 1 Alabama           15.6    4903185
## 2 Alaska            10.2     731545
## 3 Arizona           13.5    7278717
## 4 Arkansas          16      3017804
## 5 California        11.8   39512223
## 6 Colorado           9.4    5758736
```

Question 3

Once both the edu and pop data sets are tidy, join the two data sets to create a *single* data set that displays the college completion percentage, median income, poverty rate, and population for each state. Keep all rows and all columns from each of the two data frames you're joining.

```
joinedDF <- full_join(x=edu2, y=pop2, by="State")
head(joinedDF)
```

```
## # A tibble: 6 x 5
##   State      CollegeRate MedianIncome PovertyRate Population
##   <chr>            <dbl>        <dbl>       <dbl>      <dbl>
## 1 Alabama           0.25        51771        15.6    4903185
## 2 Alaska            0.3         77203        10.2     731545
## 3 Arizona           0.29        62027        13.5    7278717
## 4 Arkansas          0.23        49020        16      3017804
## 5 California        0.34        80423        11.8   39512223
## 6 Colorado          0.41        77104         9.4    5758736
```

Question 4

When you joined the two data frames, you created a data frame that contains some missing values. Check which variable(s) contain missing data, then fill in any missing data with the **average** value of the variable.

```r
# Values for `MedianIncome`, `PovertyRate` and `Population` are missing for Puerto Rico.

avg_med_income = mean(joinedDF$MedianIncome, na.rm=TRUE)
joinedDF[is.na(joinedDF$MedianIncome), 'MedianIncome' ] <- avg_med_income

avg_pov_rate = mean(joinedDF$PovertyRate, na.rm=TRUE)
joinedDF[is.na(joinedDF$PovertyRate), 'PovertyRate' ] <- avg_pov_rate

avg_pop = mean(joinedDF$Population, na.rm=TRUE)
joinedDF[is.na(joinedDF$Population), 'Population' ] <- avg_pop

joinedDF
```

```
## # A tibble: 52 x 5
##    State                CollegeRate MedianIncome PovertyRate Population
##    <chr>                      <dbl>        <dbl>       <dbl>      <dbl>
##  1 Alabama                     0.25        51771        15.6    4903185
##  2 Alaska                      0.3         77203        10.2     731545
##  3 Arizona                     0.29        62027        13.5    7278717
##  4 Arkansas                    0.23        49020        16      3017804
##  5 California                  0.34        80423        11.8   39512223
##  6 Colorado                    0.41        77104         9.4    5758736
##  7 Connecticut                 0.39        78920         9.9    3565287
##  8 Delaware                    0.32        70348        11.2     973764
##  9 District of Columbia        0.59        90395        14.1     705749
## 10 Florida                     0.3         59198        12.7   21477737
## # i 42 more rows
```