

Data Wizards (Group 4) Project 1 R Notebook

Di Chen

Mai Castellano

Tyler Kussee

Spencer Hutchison

Introduction

The American Community Survey is an ongoing survey that provides vital information on a yearly basis about our nation and its people. The information from the survey generates data that helps inform how trillions of dollars in federal funds are distributed.

With the data from the American Community Survey, the Data Wizards have come up with the following 3 questions they would like to answer:

- Does higher education mean higher income accounting for sex and age?
- Can we predict household income based on education, occupation, and race/ethnicity?
- Is there a relationship between race/ethnicity and home ownership?

In particular, we'll be focusing on the first question to get started.

Obtain the Data

In the case of the ACS dataset, we'll pull it via an API. We've already declared our key earlier, so now we just need to pull our dataset with the variables and filtering we'd like. In this case we'll want to pull our Age (AGEP), Education Level (SCHL), sex (SEX), and Total Person's Income (PINCP) while filtering to the year 2022 and the State of OR.

Scrub the data

Since we've already pulled the data, now we need to just clean up the data first. Let's begin by assigning the column names more descriptive names:

```
##Rename column names with a mutate.
puma_data <- puma_data %>% rename(Age = AGEP,
                                   EducationLevel = SCHL,
                                   Income = PINCP,
                                   Sex = SEX,
                                   PersonNumber = SPORDER)
```

We can remove the the ST, Serial Number, and housing weight columns for now, since we don't really need that information. I'll keep PersonNumber in for now, but we can remove it if not needed.

```
puma_data <- puma_data %>%
  select(-ST, -SERIALNO, -WGTP, -PWGTP)
```

Next I'll rearrange the columns:

```
puma_data <- puma_data %>%
  select(EducationLevel, Income, Age, Sex, PersonNumber)
```

Now we should work on the variable codes to convert them to their actual descriptions to make the content easier to peruse. We'll start with Education Level and convert it from it's characters to the actual descriptions.

```
puma_data <- puma_data %>%
  mutate(
    EducationLevel = case_when(
      EducationLevel == "bb" ~ "(less than 3 years old)",
      EducationLevel == "01" ~ "No schooling completed",
      EducationLevel == "02" ~ "Nursery School / Preschool",
      EducationLevel == "03" ~ "Kindergarten",
      EducationLevel == "04" ~ "Grade 1",
      EducationLevel == "05" ~ "Grade 2",
      EducationLevel == "06" ~ "Grade 3",
      EducationLevel == "07" ~ "Grade 4",
      EducationLevel == "08" ~ "Grade 5",
      EducationLevel == "09" ~ "Grade 6",
      EducationLevel == "10" ~ "Grade 7",
      EducationLevel == "11" ~ "Grade 8",
      EducationLevel == "12" ~ "Grade 9",
      EducationLevel == "13" ~ "Grade 10",
      EducationLevel == "14" ~ "Grade 11",
      EducationLevel == "15" ~ "Grade 12 - no diploma",
      EducationLevel == "16" ~ "Regular high school diploma",
      EducationLevel == "17" ~ "GED or alternative credential",
      EducationLevel == "18" ~ "Some college, but less than 1 year",
      EducationLevel == "19" ~ "1 or more years of college credit, no degree",
      EducationLevel == "20" ~ "Associates Degree",
      EducationLevel == "21" ~ "Bachelors Degree",
      EducationLevel == "22" ~ "Masters Degree",
      EducationLevel == "23" ~ "Professional Degree beyond a Bachelors Degree",
      EducationLevel == "24" ~ "Doctorate Degree",
      TRUE ~ as.character(EducationLevel)
    )
  )
```

Now lets identify the genders in the dataset:

```
puma_data <- puma_data %>%
  mutate(
    Sex = case_when(
      Sex == 1 ~ "Male",
      Sex == 2 ~ "Female",
      TRUE ~ as.character(Sex)
    )
  )
```

With the data mostly cleaned up, we can now explore the data.

Explore the data

Reference on Variables for now

Model the data

Interpret the data