A Regression Analysis of the Gender Pay Gap

Maria-Cristiana Gîrjău

Revised on 2019-09-20

Data wrangling

```
ACSSampleRaw <- read_csv("data/acs230_3k.csv")
```

```
# setting factor levels
sex_levels <- c(M = "1",
                "F" = "2")

citizenship_levels <- c("US born" = "1",
                        "territories born" = "2",
                        "aborad born" = "3",
                        naturalized = "4",
                        "not a citizen" = "5")

race_levels <- c(white = "1",
                 black = "2",
                 native = "3",
                 native = "4",
                 native = "5",
                 asian = "6",
                 native = "7",
                 other = "8",
                 multiple = "9")


military_levels <- c(active = "1",
                     past = "2",
                     training = "3",
                     never = "4")

married_levels <- c(married = "1",
                    widowed = "2",
                    divorced = "3",
                    separated = "4",
                    single = "5")

children_age_levels <- c("under six" = "1",
                         "above six" = "2",
                         both = "3",
                         none = "4")

education_levels <- c(none = "1",
                      preschool = "2",
                      kindergarten = "3",
                      "grade school" = "4",
```

```
                      "grade school" = "5",
                      "grade school" = "6",
                      "grade school" = "7",
                      "grade school" = "8",
                      "grade school" = "9",
                      "grade school" = "10",
                      "grade school" = "11",
                      "some high school" = "12",
                      "some high school" = "13",
                      "some high school" = "14",
                      "some high school" = "15",
                      "high school diploma" = "16",
                      "high school diploma" = "17",
                      "less than one year of college" = "18",
                      "more than one year of college" = "19",
                      "associate degree" = "20",
                      "bachelor's degree" = "21",
                      "master's degree" = "22",
                      "professional degree" = "23",
                      "doctoral degree" = "24")

degree_levels <- c("Agriculture" = "11",
                   "Environmental" = "13",
                   "Architecture" = "14",
                   "Ethnic Studies" = "15",
                   "Media and Journalism" = "19",
                   "Communication" = "20",
                   "Computer Science and IT" = "21",
                   "Cosmetology and Gastronomy" = "22",
                   "Education" = "23",
                   "Engineering" = "24",
                   "Engineering" = "25",
                   "Languages" = "26",
                   "Consumer Sciences" = "29",
                   "Law and Policy" = "32",
                   "English Literature" = "33",
                   "Liberal Arts" = "34",
                   "Library Science" = "35",
                   "Biological Sciences" = "36",
                   "Mathematics and Statistics" = "37",
                   "Military" = "38",
                   "Interdisciplinary" = "40",
                   "Fitness" = "41",
                   "Philosophy" = "48",
                   "Theology" = "49",
                   "Physical Sciences" = "50",
                   "Physical Sciences" = "51",
                   "Psychology" = "52",
                   "Law and Policy" = "53",
                   "Law and Policy" = "54",
                   "Social Sciences" = "55",
                   "Construction" = "56",
                   "Construction" = "57",
```

```
                    "Transportation" = "59",
                    "Arts" = "60",
                    "Medicine" = "61",
                    "Business and Finance" = "62",
                    "History" = "64")


employment_levels <- c("employed working" = "1",
                       "employed not working" = "2",
                       unemployed = "3",
                       "military working" = "4",
                       "military not working" = "5",
                       "not in labor force" = "6")

worker_class_levels <- c("Private for-profit" = 1,
                         "Private not-for-profit" = 2,
                         "Local government" = 3,
                         "State government" = 4,
                         "Federal government" = 5,
                         "Self-employed" = 6,
                         "Self-employed" = 7,
                         "Family business" = 8,
                         "Never worked" = 9)

region_levels <- c(Northeast = "1",
                   Midwest = "2",
                   South = "3",
                   West = "4",
                   "Puerto Rico" = "9")

state_levels <- c(AL = "1",
                  AK = "2",
                  AZ = "4",
                  AR = "5",
                  CA = "6",
                  CO = "8",
                  CT = "9",
                  DE = "10",
                  DC = "11",
                  FL = "12",
                  GA = "13",
                  HI = "15",
                  ID = "16",
                  IL = "17",
                  IN = "18",
                  IA = "19",
                  KS = "20",
                  KY = "21",
                  LA = "22",
                  ME = "23",
                  MD = "24",
                  MA = "25",
                  MI = "26",
```

```
                    MN = "27",
                    MS = "28",
                    MO = "29",
                    MT = "30",
                    NE = "31",
                    NV = "32",
                    NH = "33",
                    NJ = "34",
                    NM = "35",
                    NY = "36",
                    NC = "37",
                    ND = "38",
                    OH = "39",
                    OK = "40",
                    OR = "41",
                    PA = "42",
                    RI = "44",
                    SC = "45",
                    SD = "46",
                    TN = "47",
                    TX = "48",
                    UT = "49",
                    VT = "50",
                    VA = "51",
                    WA = "53",
                    WV = "54",
                    WI = "55",
                    WY = "56",
                    "Puerto Rico" = "72")

industry_levels <- c("Agriculture, Forestry, Fishing and Hunting" = "11",
                     "Mining, Quarrying, and Oil and Gas Extraction" = "21",
                     "Utilities" = "22",
                     "Construction" = "23",
                     "Manufacturing" = "31",
                     "Manufacturing" = "32",
                     "Manufacturing" = "33",
                     "Manufacturing" = "3M",
                     "Wholesale Trade" = "42",
                     "Retail Trade" = "44",
                     "Retail Trade" = "45",
                     "Transportation and Warehousing" = "48",
                     "Transportation and Warehousing" = "49",
                     "Information" = "51",
                     "Finance and Insurance" = "52",
                     "Real Estate and Rental and Leasing" = "53",
                     "Professional, Scientific, and Technical Services" = "54",
                     "Management of Companies and Enterprises" = "55",
                     "Administrative and Support and Waste Management
                     and Remediation Services" = "56",
                     "Educational Services" = "61",
                     "Health Care and Social Assistance" = "62",
                     "Arts, Entertainment, and Recreation" = "71",
```

```
                    "Accommodation and Food Services" = "72",
                    "Other Services (except Public Administration)" = "81",
                    "Public Administration" = "92")

# expanding sample (?)
```

```
# Wrangling the data
ACSSample <- ACSSampleRaw %>%

  mutate(ADJINC.x = ADJINC.x / 10^6, # adding decimal point to ADJINC
         WAGP = WAGP * ADJINC.x) %>% # adjusting dollar amounts for inflation

  # selecting which variables to keep
  select(SEX, AGEP, CIT, RAC1P, MIL, DIS, # general demographics
         MAR, PAOC, NRC, FER, # family and household
         SCHL, FOD1P, FOD2P, SCIENGP, # educational background
         ESR, COW, WKW, WKHP, NAICSP, # employment
         WAGP, # income
         REGION.x, ST.x) %>% # location

  # renaming the variables
  rename(sex = SEX,
         age = AGEP,
         citizenship = CIT,
         race = RAC1P,
         military = MIL,
         disabled = DIS,
         married = MAR,
         children_age = PAOC,
         children_no = NRC,
         gave_birth = FER,
         education = SCHL,
         degree_1 = FOD1P,
         degree_2 = FOD2P,
         stem_degree = SCIENGP,
         employment = ESR,
         worker_class = COW,
         weeks_worked = WKW,
         hours_worked = WKHP,
         industry = NAICSP,
         wage_income = WAGP,
         region = REGION.x,
         state = ST.x) %>%

  # converting inputs to an appropriate data type
  mutate(sex = as.factor(sex) %>%
           fct_recode(!!!sex_levels),
         citizenship = as.factor(citizenship) %>%
           fct_recode(!!!citizenship_levels),
         race = as.factor(race) %>%
           fct_recode(!!!race_levels),
         military = as.factor(military) %>%
           fct_recode(!!!military_levels),
```

```
        married = as.factor(married) %>%
          fct_recode(!!!married_levels),
        children_age = as.factor(children_age) %>%
          fct_recode(!!!children_age_levels),
        education = as.factor(education) %>%
          fct_recode(!!!education_levels),
        employment = as.factor(employment) %>%
          fct_recode(!!!employment_levels),
        region = as.factor(region) %>%
          fct_recode(!!!region_levels),
        state = as.factor(state) %>%
          fct_recode(!!!state_levels),
        gave_birth = ifelse(gave_birth == 1, TRUE, FALSE),
        stem_degree = ifelse(stem_degree == 1, TRUE, FALSE),
        disabled = ifelse(disabled == 1, TRUE, FALSE),
        # collapsing industry codes into broad NAICS sectors
        # (only taking the first two digits of the NAICS code,
        # which represent the broader industry sectors)
        industry = substr(industry, start = 1, stop = 2) %>%
          as.factor() %>%
          fct_recode(!!!industry_levels),
        # collapsing education codes into broader fields
        # (only taking the first two digits of each code,
        # which represent the broader field)
        degree_1 = substr(degree_1, start = 1, stop = 2) %>%
          as.factor() %>%
          fct_recode(!!!degree_levels),
        degree_2 = substr(degree_2, start = 1, stop = 2) %>%
          as.factor() %>%
          fct_recode(!!!degree_levels),
        # merging the two degree variables
        # taking care of NAs and repeated values
        degree = ifelse(is.na(degree_1) | is.na(degree_2),
                        as.character(degree_1),
                        ifelse(as.character(degree_1) == as.character(degree_2),
                               as.character(degree_1),
                               paste(degree_1, degree_2, sep = " and ")))) %>%

  # filtering the individuals to keep
  filter(!(is.na(wage_income)) & wage_income > 0, # salary income is positive
        employment %in% c("employed working",
                          "employed not working",
                          "military working"), # employed and/or working
        age >= 18) %>% # only people over 18

  # removing merged variables and those used for filtering
  select(-employment)  %>%

  # dropping unused factor levels
  mutate_if(is.factor, fct_drop)
```

Data exploration

```
# some functions for repetitive tasks
render_table <- function(data, title = NULL) {
  data %>%
      kable(booktabs = TRUE, caption = title) %>%
      kable_styling(latex_options = "striped")
}
```

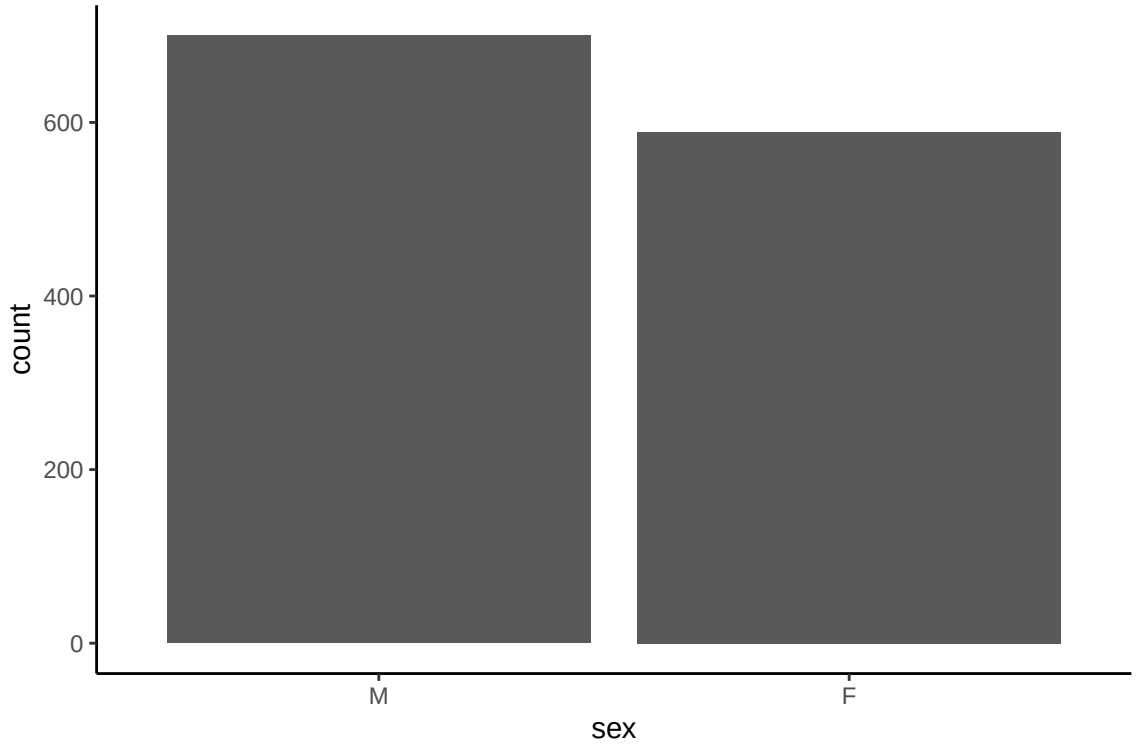
Univariate data exploration

```
# must make graphics nicer - colors, axis labels, titles, captions

# sex ---------------------------------------------------------------------
ACSSample %>%
  group_by(sex) %>%
  tally() %>%
  render_table()
```

sex	n
M	700
F	589

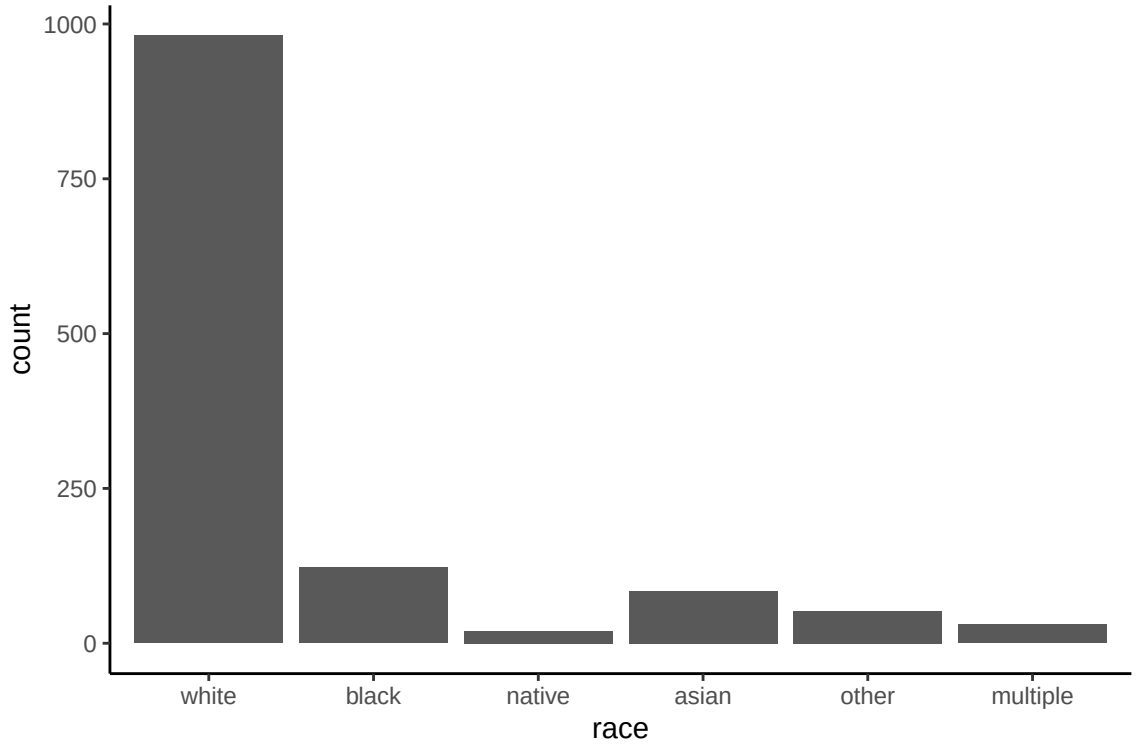
```
ggplot(ACSSample, aes(x = sex)) +
  geom_bar()
```



```r
# race ---------------------------------------------------------------------
ACSSample %>%
  group_by(race) %>%
  tally() %>%
  render_table()
```

race	n
white	981
black	122
native	20
asian	84
other	52
multiple	30

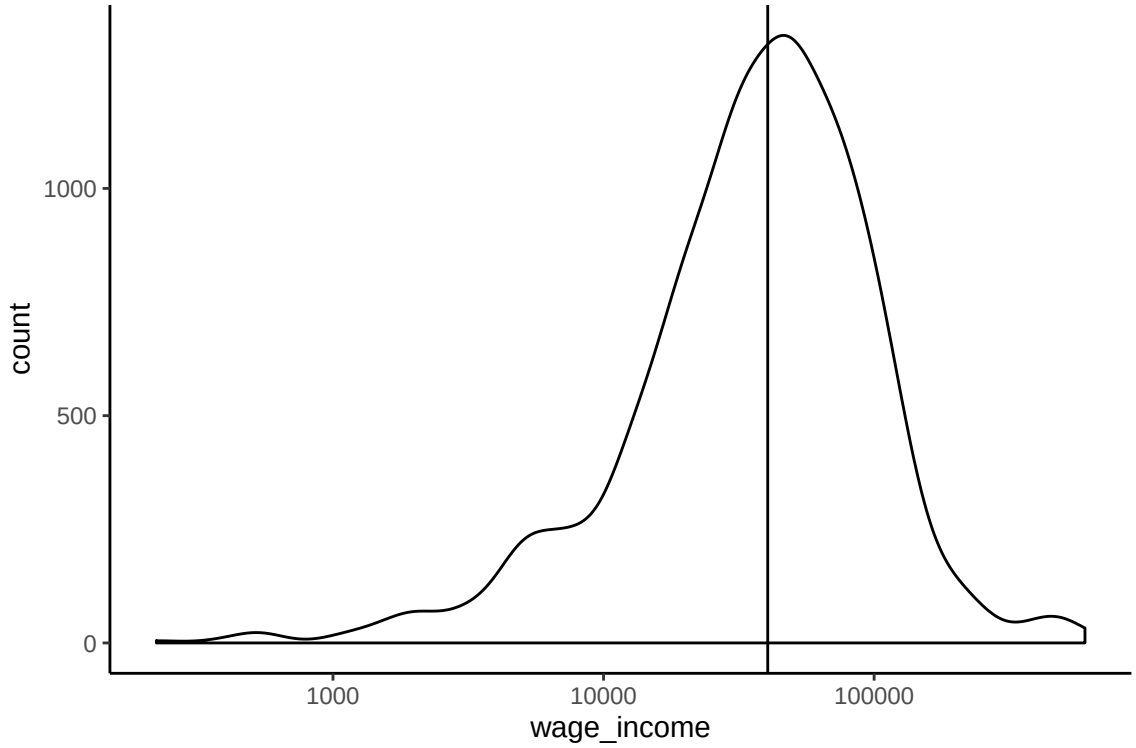
```r
ggplot(ACSSample, aes(x = race)) +
  geom_bar()
```




```r
# income -------------------------------------------------------------------
ACSSample %>%
  favstats(~wage_income, data = .) %>%
  render_table()
```

min	Q1	median	Q3	max	mean	sd	n	missing
222.4616	20223.78	40447.56	70783.23	601657.5	54908.94	62257.96	1289	0

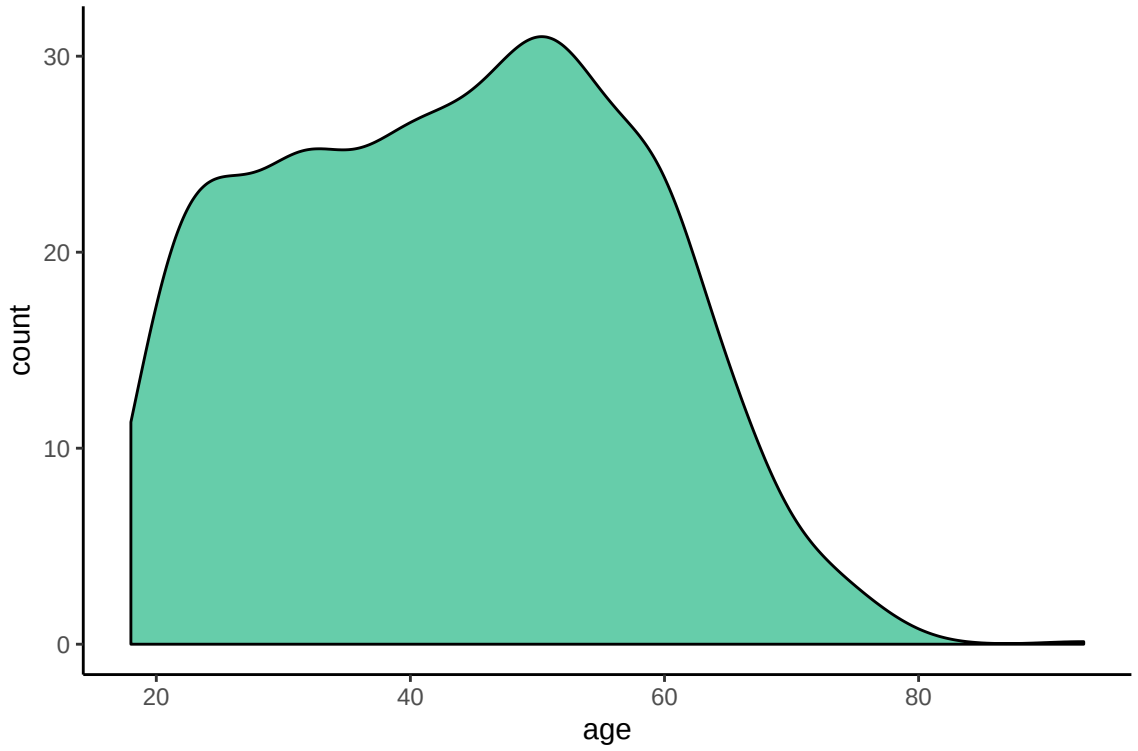
```
ggplot(ACSSample, aes(x = wage_income, y = ..count..)) +
  geom_density() +
  scale_x_log10() +
  geom_vline(xintercept = median(ACSSample$wage_income))
```




```
# age ---------------------------------------------------------------------
ACSSample %>%
  favstats(~age, data = .) %>%
  render_table()
```

min	Q1	median	Q3	max	mean	sd	n	missing
18	31	44	54	93	43.45617	14.28057	1289	0

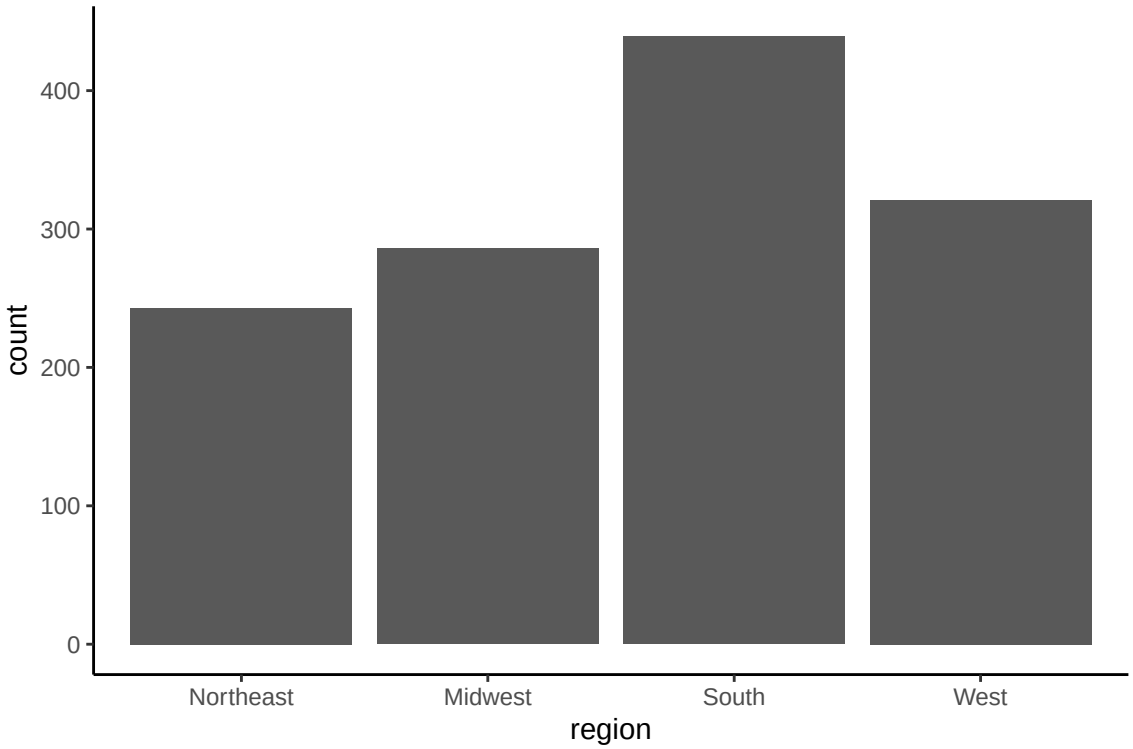
```
ggplot(ACSSample, aes(x = age, y = ..count..)) +
  geom_density(fill = "aquamarine3")
```




```
# region ------------------------------------------------------------------
ACSSample %>%
  group_by(region) %>%
  tally() %>%
  render_table()
```

region	n
Northeast	243
Midwest	286
South	439
West	321

```
ggplot(ACSSample, aes(x = region)) +
  geom_bar()
```



Data analysis

Assessment

Current questions