A Regression Analysis of the Gender Pay Gap

Maria-Cristiana Gîrjău

Revised on 2019-09-20

Data wrangling

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

```
# Wrangling the data
acs_sample <- acs_sample_raw_small %>%

  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)) %>%

  # 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

# must drop unused levels after filtering
```

Data exploration

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

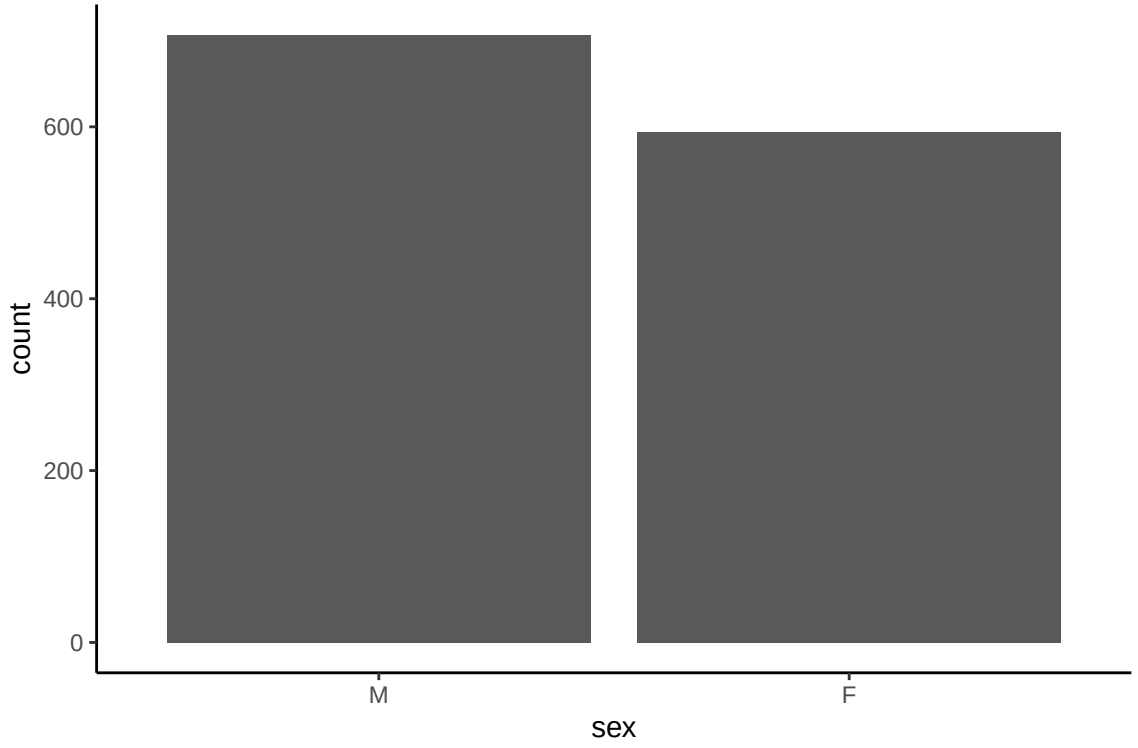
Univariate data exploration

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

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

sex	n
M	707
F	594

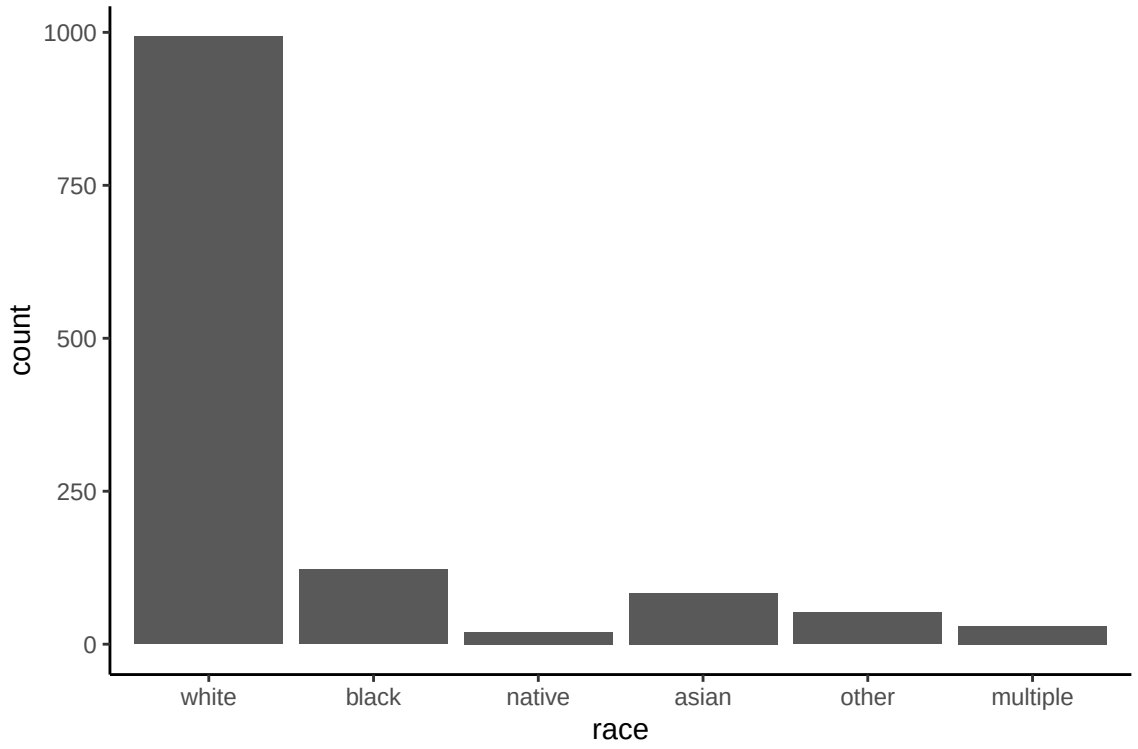
```
ggplot(acs_sample, aes(x = sex)) +
  geom_bar()
```




```
# race ------------------------------------------------------------------
acs_sample %>%
  group_by(race) %>%
  tally() %>%
  render_table()
```

race	n
white	993
black	122
native	20
asian	84
other	52
multiple	30

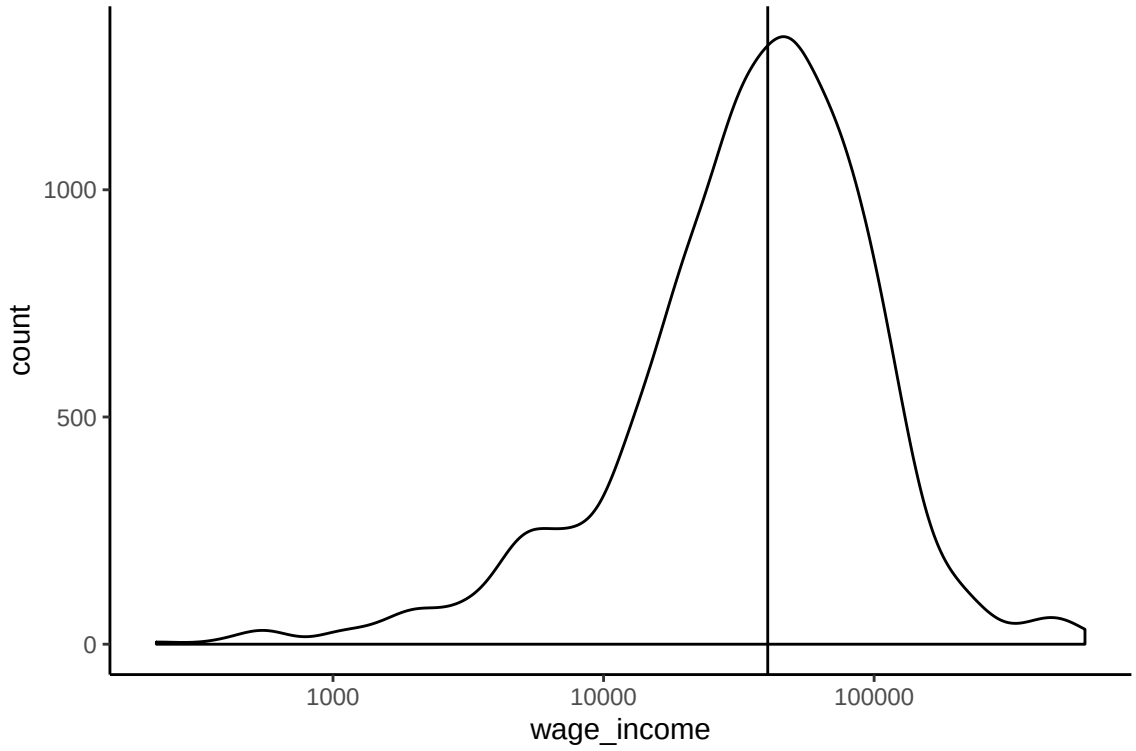
```
ggplot(acs_sample, aes(x = race)) +
  geom_bar()
```




```
# income ------------------------------------------------------------------
acs_sample %>%
  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	54437.45	62163.78	1301	0

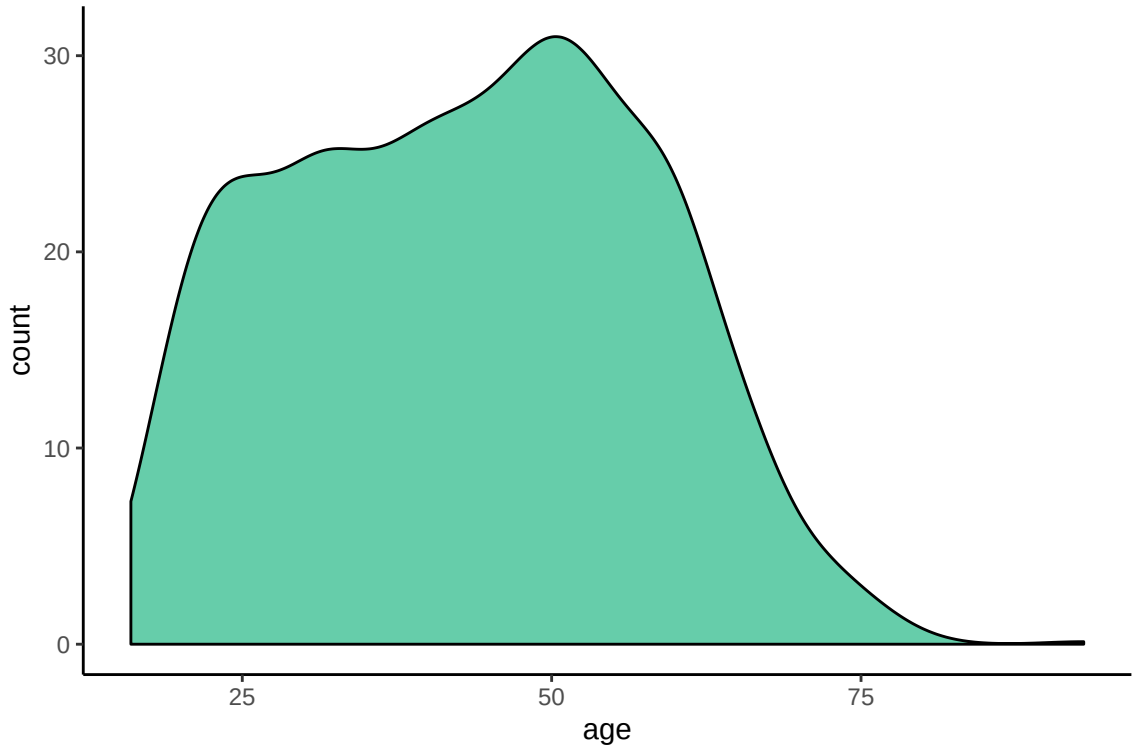
```
ggplot(acs_sample, aes(x = wage_income, y = ..count..)) +
  geom_density() +
  scale_x_log10() +
  geom_vline(xintercept = median(acs_sample$wage_income))
```




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

min	Q1	median	Q3	max	mean	sd	n	missing
16	31	44	54	93	43.20984	14.44218	1301	0

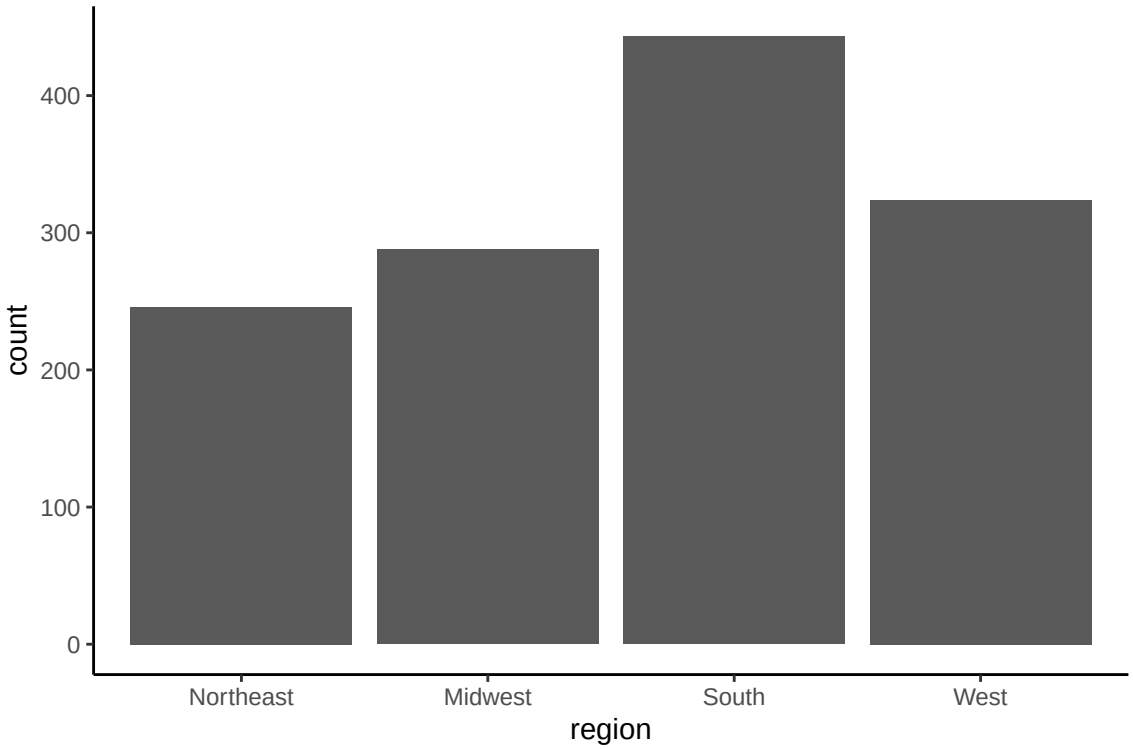
```
ggplot(acs_sample, aes(x = age, y = ..count..)) +
  geom_density(fill = "aquamarine3")
```




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

region	n
Northeast	246
Midwest	288
South	443
West	324

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



Data analysis

Assessment

Current questions