ProjectProposal

Zachary Palmore

r sys.date()

Data Preparation

```
# Packages
require(reshape2)
require(tidyverse)
require(readr)
require(DT)


# Load Data
income_2018_bystate  <-  read_csv("https://raw.githubusercontent.com/palmorezm/msdsdata606/master/Proje

# Considered adding years to see the trends over time
# income_2014_bystate <-  read_csv("ACSST5Y2014.S2001.csv")
# income_2010_bystate <-  read_csv("ACSST5Y2010.S2001.csv")
income_2018_bystate <- income_2018_bystate[2:52,] %>%
  select(c(
    # General Reference Variables
    GEO_ID,
           NAME,
                S2001_C01_001E,
                             S2001_C01_001M,

    # Mean Earnings for FULL-TIME Workers (age 16+)
           S2001_C03_002E,
           S2001_C03_002M,
           S2001_C05_002E,
           S2001_C05_002M,

           # Male Variables
           S2001_C03_016E,
           S2001_C03_016M,
           S2001_C03_017E,
           S2001_C03_017M,
           S2001_C03_018E,
           S2001_C03_018M,
           S2001_C03_019E,
           S2001_C03_019M,
           S2001_C03_020E,
           S2001_C03_020M,
```

```
          # Female Variables
          S2001_C05_016E,
          S2001_C05_016M,
          S2001_C05_017E,
          S2001_C05_017M,
          S2001_C05_018E,
          S2001_C05_018M,
          S2001_C05_019E,
          S2001_C05_019M,
          S2001_C05_020E,
          S2001_C05_020M
          )) %>%
rename(
# Location Information
  GEO_ID = GEO_ID,
          State = NAME,

   # General Statistics on Male and Female Earnings aged 16+
        TotalPop = S2001_C01_001E,
        TotalPop_moe = S2001_C01_001M,
         M_Earnings = S2001_C03_002E,
         M_Earnings_moe =  S2001_C03_002M,
         F_Earnings = S2001_C05_002E,
         F_Earnings_moe =  S2001_C05_002M,

 # Male Earnings by Level of Education
       M_LTHS = S2001_C03_016E,
         M_HS = S2001_C03_017E,
         M_AS = S2001_C03_018E,
         M_BS = S2001_C03_019E,
         M_MS = S2001_C03_020E,

      # Male Margins of Error by Education
       M_LTHS_moe = S2001_C03_016M,
         M_HS_moe = S2001_C03_017M,
         M_AS_moe = S2001_C03_018M,
         M_BS_moe = S2001_C03_019M,
         M_MS_moe = S2001_C03_020M,

# Female Earnings by Level of Education
       F_LTHS = S2001_C05_016E,
         F_HS = S2001_C05_017E,
         F_AS = S2001_C05_018E,
         F_BS = S2001_C05_019E,
         F_MS = S2001_C05_020E,

      # Female Margins of Error by Education
        F_LTHS_moe = S2001_C05_016M,
         F_HS_moe = S2001_C05_017M,
         F_AS_moe = S2001_C05_018M,
         F_BS_moe = S2001_C05_019M,
         F_MS_moe = S2001_C05_020M,
       )
```

```r
# Converting data types
income_2018_bystate[,3:28] <- lapply(income_2018_bystate[,3:28], as.numeric)
income_2018_bystate$GEO_ID <- as.factor(income_2018_bystate$GEO_ID)
income_2018_bystate<- as.data.frame(income_2018_bystate)

# Checking for missing values - there should be none
sum(is.na(income_2018_bystate))

# Creating subsets of the data to isolate variables of interest

# excluding geo_id and moe for summary purposes
pop_income_2018 <- income_2018_bystate[,c(2,3,5,7)]
male_income_2018 <- income_2018_bystate[, c(2,seq(9, 18, 2))]
female_income_2018 <- income_2018_bystate[, c(2,seq(19, 28, 2))]

# These male and female stats can also be recombined
mf_income_2018 <- cbind(female_income_2018, male_income_2018[2:6])

# Calculate observed differences for the entire study
pop_obs <- pop_income_2018 %>%
  mutate(Obs_diff = M_Earnings - F_Earnings) %>%
  mutate(Pmf = F_Earnings/M_Earnings)

# Create a table with the highs and lows of states
# Alternatively tail could be used:
# min(tail(sort(pop_obs$Obs_diff),5))
top5 <- pop_obs %>%
  filter(Obs_diff >= min(head(sort(pop_obs$Obs_diff, decreasing=TRUE), 5)))
top5$Gap <- as.factor("Wide")
low5 <- pop_obs %>%
  filter(Obs_diff <= max(head(sort(pop_obs$Obs_diff, decreasing=FALSE), 5)))
low5$Gap <- as.factor("Narrow")
hilo_obs <- rbind(top5,low5)

# Calculating the observed differences of sex
mf_obs <- mf_income_2018 %>%
  mutate(obs_diff_lths = M_LTHS - F_LTHS) %>%
  mutate(obs_diff_hs = M_HS - F_HS) %>%
  mutate(obs_diff_as = M_AS - F_AS) %>%
  mutate(obs_diff_bs = M_BS - F_BS) %>%
  mutate(obs_diff_ms = M_MS - F_MS)

# Give each variable its own row in education
mf_stateobs <- melt(mf_obs)
  mf_stateobs  <- mf_stateobs %>%
    rename(Cagetory = variable,
           Observation = value)
mf_state_obsdiffs <- melt(mf_obs[,12:16])
  mf_state_obsdiffs <- mf_state_obsdiffs %>%
    rename(Observation = variable,
           Difference = value)
mf_earnings_byedu <- melt(mf_obs[,2:11])
# summarizing the education earning using means
```

```r
mf_earnings_byedu <- mf_earnings_byedu %>%
  group_by(variable) %>%
  summarise(AvgEarning = mean(value)) %>%
  rename(Education = variable)
# Add in the variable of sex for later comparisons
mf_earnings_byedu$Sex <- c("Female","Female","Female","Female","Female","Male","Male","Male","Male","Ma
# Rename the education observations to be descriptive of
# the entire data set and remove the male/female bounds of edu
mf_earnings_byedu$Education <-
  c("LTHS","HS","AS","BS","MS","LTHS","HS","AS","BS","MS")
# Variables chosen to describe the data are not mutually exclusive
# For example 'BS' is not just for Bachelors of Science.
# That category includes all those individuals that
# attained a bachelors degree on the ACS in 2018
# and as another example 'AS' contains those with 'some college'
# from the ACS in 2018. These variables were only used for ease
# in describing the variables visually
```

Research question

You should phrase your research question in a way that matches up with the scope of inference your dataset allows for.

For people age twenty-five and older, does the level of education attained have an affect on the average annual earnings of males and females across the United States in 2018? If so, is there a difference by state?

Cases

What are the cases, and how many are there?

Each case is a state's median dollar amount of income earned for the variables of male or female in 28 variable types as calculated by the Census Bureau. These variable types represent the highest education level attained by the individuals surveyed. There are exclusions.

The margin of error is given for each median dollar amount and listed immediately following each of the state's observations with the characters "moe" in the column name. This could be useful to comprehend variability calculations later. There is also a column with the unique geographic identifier of each state. This may be used for mapping the presence or absence of a wage gap by state. The total population is also a summed value for each state calculated by the Census Bureau. There are 51 rows in the given data set. One row for each of the 50 states and the District of Columbia.

Data collection

Describe the method of data collection.

This data comes from the American Community Survey (ACS) of the United States Census Bureau. They collect a random sample from the American population every year with no household ever receiving the survey more than once every five years. Find out more about their methodology here: https://www.census.gov/programs-surveys/acs/methodology.html

Type of study

What type of study is this (observational/experiment)?

This is an observational study.

Data Source

If you collected the data, state self-collected. If not, provide a citation/link.

U.S. Census Bureau. (2019). Earnings in the past 12 months (in 2019 Inflation-adjusted dollars). TableID: S2001. Retrieved from https://data.census.gov.

URL: https://data.census.gov/cedsci/table?q=Income%20%28Households,%20Families,%20Individuals%29&t=Earnings%20%28Individuals%29%3AIncome%20and%20Earnings%3AIncome%20and%20Poverty&g=0100000US.04000.001&tid=ACSST1Y2019.S2001&moe=true&tp=true&hidePreview=true

Dependent Variable

What is the response variable? Is it quantitative or qualitative?

The response or dependent variable is average annual earnings. It is quantitative.

Independent Variable

You should have two independent variables, one quantitative and one qualitative.

The qualitative independent variables are that of sex, education, and state.

Relevant summary statistics

Provide summary statistics for each the variables. Also include appropriate visualizations related to your research question (e.g. scatter plot, boxplots, etc). This step requires the use of R, hence a code chunk is provided below. Insert more code chunks as needed.

First, we will determine the presence or absence of a difference in the entire sample from the ACS. Then we compute the observed differences between males and females average earnings in each state to examine the patterns among the states. With that, we can begin to answer the question, is there a difference in earnings of male and female Americans in 2018?

We also compute the observed differences in education groups for both males and females. This will be reviewed across the states for patterns as well. In this, we will look for a reason to investigate individual earnings by sex further, to find if there truly is a difference in the earnings of an individual by their education level and if the earnings are higher or lower for females when compared to the average male earnings. We begin with the first states observed differences.

```
# Here we review the data as a whole for ACS. Since all states have
# larger incomes for males we can find the proportion of a females
# income to that of a males across the states. The mean dollar amounts
# are listed in the "Obs_diff" column.
head(pop_obs[,c(1,5,6)])
```

```
##        State Obs_diff       Pmf
## 1    Alabama    11936 0.6736658
## 2     Alaska    13266 0.7064720
## 3    Arizona     8336 0.7672808
## 4   Arkansas     8730 0.7404026
## 5 California    10268 0.7452488
## 6   Colorado    11259 0.7322537
```

We can find the mean, median, and other important statistics through this summary. We can also compute the range and variance of the proportion of average female earnings to average male earnings.

```
summary(pop_obs[,c(5,6)])
```

```
##     Obs_diff          Pmf
##  Min.   : 6916   Min.   :0.5476
##  1st Qu.: 9753   1st Qu.:0.6769
##  Median :11918   Median :0.7081
##  Mean   :11787   Mean   :0.7068
##  3rd Qu.:13303   3rd Qu.:0.7354
##  Max.   :18369   Max.   :0.8402
```

```
var(pop_obs$Pmf)
```

```
## [1] 0.00273604
```

```
pop_rng <- range(pop_obs$Obs_diff)[2] - range(pop_obs$Obs_diff)[1]
pop_rng
```

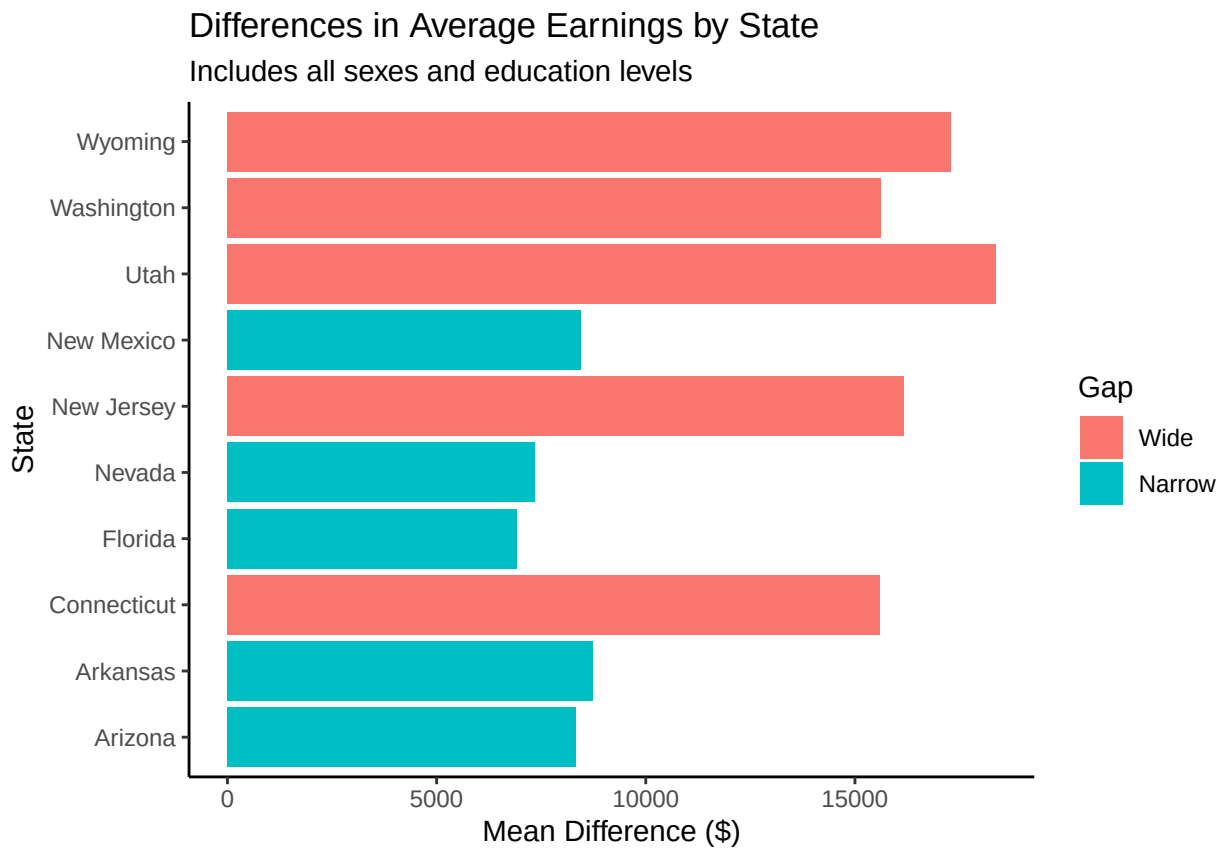
```
## [1] 11453
```

There appears to be a difference as the mean and median proportion are both near .71. The mean observed difference in dollars is $11,787 while the median is $11,918. This indicates that, across the U.S. females earned about 71% percent of the average male earnings in 2018. As a note, this is not adjusted for relative proportions of individuals and assumes about the same number of males as females are employed in the workforce.

Additionally, the data from ACS has a large sample size (approximately 2.143 million) and is thought to be a good estimate representative of the population within their margins of error. Those margins were provided in the 'income_2018_bystate' data frame. Approximate normality is reinforced in the small difference between the mean and median of the variables.

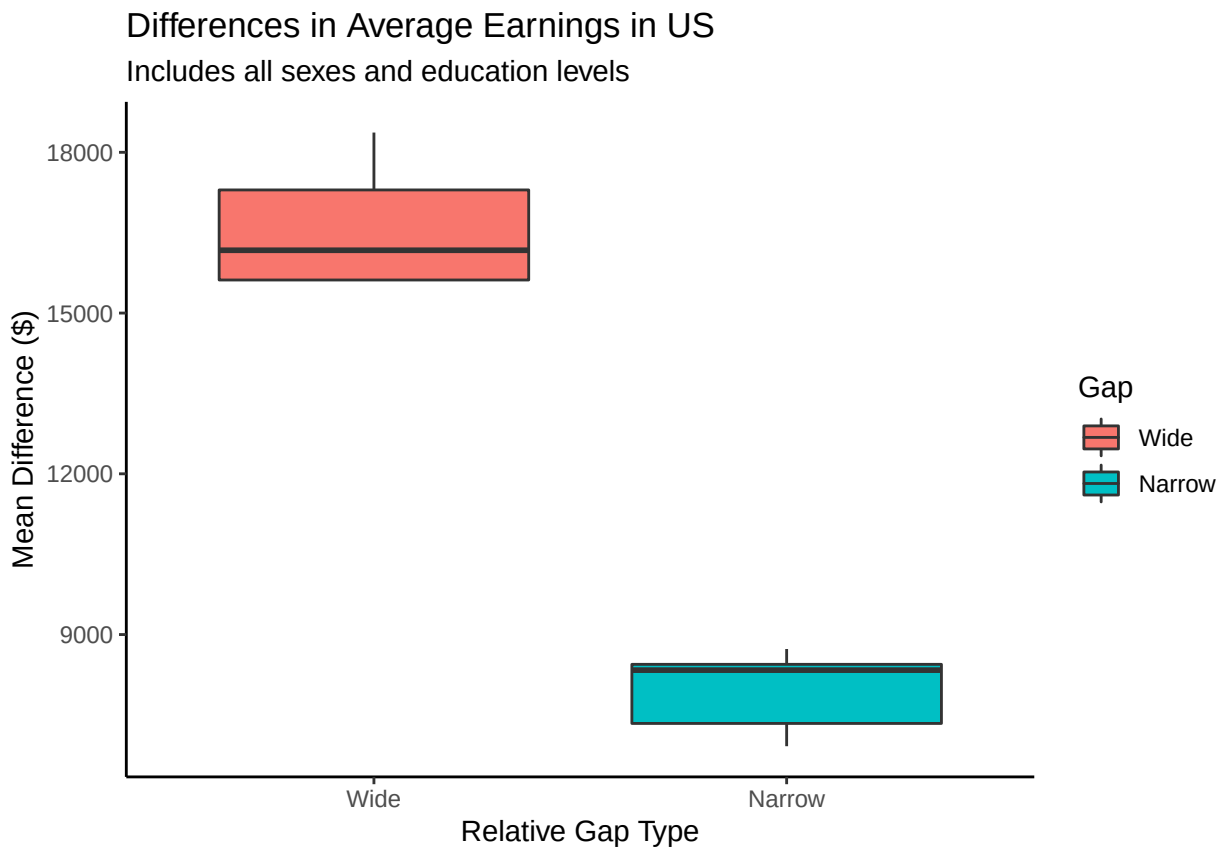
The differences in individual earnings by state can be shown in the following chart. Only the states with highest 5 and lowest 5 differences in earnings were selected for clarity. The states with the highest differences can be thought of as having a wider earnings gap. The reverse applies for the lowest where their earnings gap can be thought of as narrower relative to the rest of the states.

```
ggplot(hilo_obs, aes(x = State, y = Obs_diff, fill = Gap)) +
  labs(x = "State", y = "Mean Difference ($)", title = "Differences in Average Earnings by State", subt
  geom_col() +
  theme_classic() +
  coord_flip()
```



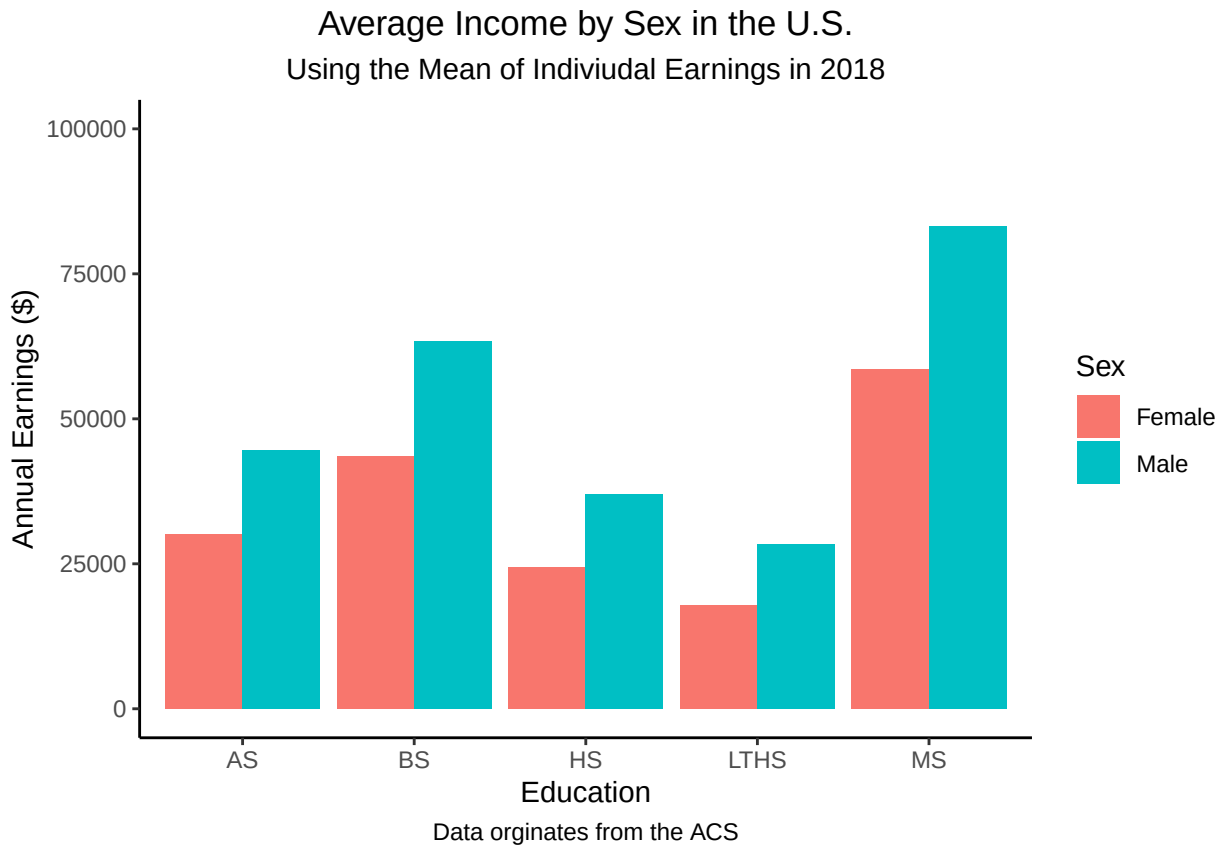
It may help to also see the differences plotted as a boxplot where the distributions means are several thousands of dollars apart. Those with the widest gap have been labeled in red and narrowest in blue.

```
ggplot(hilo_obs, aes(x = Gap, y = Obs_diff, fill = Gap)) +
  labs(x = "Relative Gap Type", y = "Mean Difference ($)", title = "Differences in Average Earnings in U
  geom_boxplot() + theme_classic()
```



Next, we observe the differences in the earnings of individuals by education and sex. We want to find out if there appears to be any differences in earnings by the level of education attained for Americans in 2018. The observations of education use abbreviated acronyms of science degrees to describe the various groups of education. This is for visualizing purposes only and is not representative of the full scope of education data collected in the ACS.

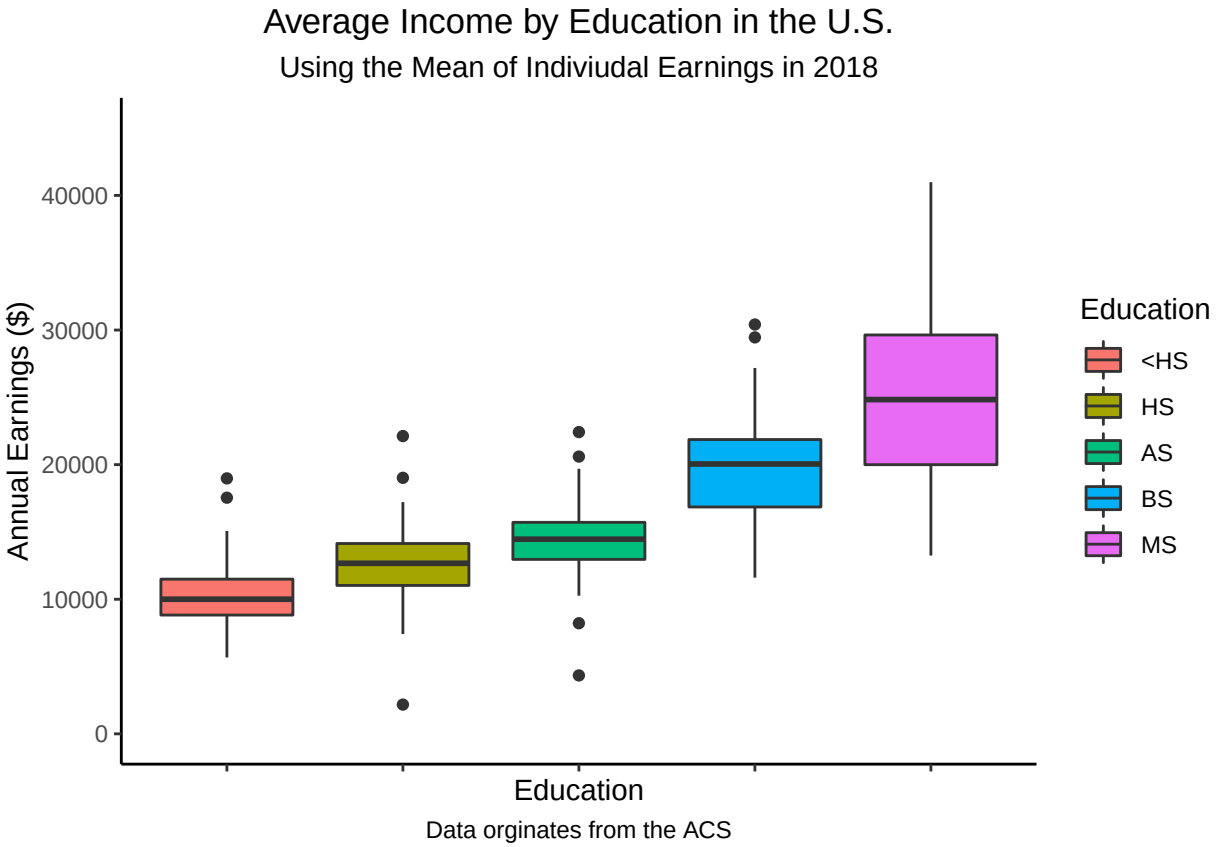
```
ggplot(mf_earnings_byedu, aes(x = Education, y = AvgEarning, fill = Sex)) + geom_col(position = "dodge"
  labs(x = "Education", y = "Annual Earnings ($)", title = "Average Income by Sex in the U.S.", subtitl
  theme_classic() +
  theme(plot.caption = element_text(hjust = 0.5),
        plot.title = element_text(hjust = 0.5),
        plot.subtitle = element_text(hjust = 0.5))
```



We can see that in every mean of every education type, the male earnings bar in blue is higher than the bars of female earnings in red. A boxplot adds some detail to this.

```
ggplot(mf_state_obsdiffs, aes(x = Observation, y = Difference)) + geom_boxplot(aes(fill = Observation))
            scale_fill_discrete(name = "Education", labels = c("<HS", "HS", "AS", "BS", "MS")) +
  labs(x = "Education",
       y = "Annual Earnings ($)",
       title = "Average Income by Education in the U.S.",
       subtitle = "Using the Mean of Indiviudal Earnings in 2018",
       caption = "Data orginates from the ACS") +

  theme_classic() +
  theme(axis.text.x = element_blank(),
        plot.caption = element_text(hjust = 0.5),
        plot.title = element_text(hjust = 0.5),
        plot.subtitle = element_text(hjust = 0.5))
```



As the level of education increases, so too do the average annual earnings of individuals. This boxplot also shows that the variation of individuals earnings may increase with higher levels if education attainment too. Further research is needed to test these hypotheses and compare across the states.