Module 1

Day 1

Goals for Today

Economics:

- Learn about using ACS data
- Discover a relationship between wages and education

Goals for Today

R:

- Create an R markdown document for our analysis
- Practice reading data into R
- Practice loading packages
- Learn the select and mutate dplyr verbs
- Make line and scatter plots using ggplot

The American Community Survey

(From Census ACS Information)

- The ACS is a nationwide survey that collects and produces information on social, economic, housing, and demographic characteristics about our nation's population every year
- The Census Bureau contacts over 3.5 million households across the country to participate in the ACS.
- James Madison ensured that the Constitution gave Congress the authority to collect additional information beyond the population count in order to "enable future legislators to adapt the public measures to the particular circumstances of the community."
- After the 2000 Census, the long form Census became the ACS, and this survey continues to collect long-form-type information each year.

The ACS Data We Will Use

I have prepared a 0.01% random sample of the 2016 ACS. Spend some time looking through the codebook and try and answer the following questions:

- How many variables are in our sample?
- Which variables are continuous? Which are categorical?
- What is a potential difference between wage and total income? Why might we restrict ourselves to one or the other?
- What does the EMPSTAT variable tell us? Why will this be important later on?

Reading in Data

First we need to read the into R.

- Use the `setwd()` function to set the working directory
- Now use the `read_csv()` function to read in the file acs_2016_educ.csv
- What are the dimensions of our data? What function should we use?
- Use the `head()` function to look at the first few rows.

```
## Parsed with column specification:
## cols(
##   EDUC = col_integer(),
##   INCTOT = col_double(),
##   INCWAGE = col_double()
## )
```

dplyr

Why dplyr?

- dplyr provides a grammar of data manipulation
- It is fast, expressive, and agnostic about the format of your data
- Now try the `glimpse()` function from the dplyr library. What's the difference?

select()

- We can use the select function to select certain columns from our data.
- The arguments to select are first the data frame followed by the columns you wish to keep.

```
select(acs_2016_educ, EDUC)
select(acs_2016_educ, -EDUC)
```

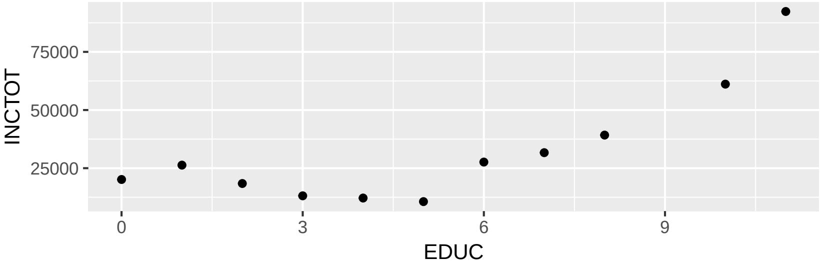
- Use the select function to create a new data frame called acs_small that only has the columns EDUC and INCTOT.

ggplot2

- Now that we have some data we are ready to start plotting
- Let's make a simple scatter plot

Plotting

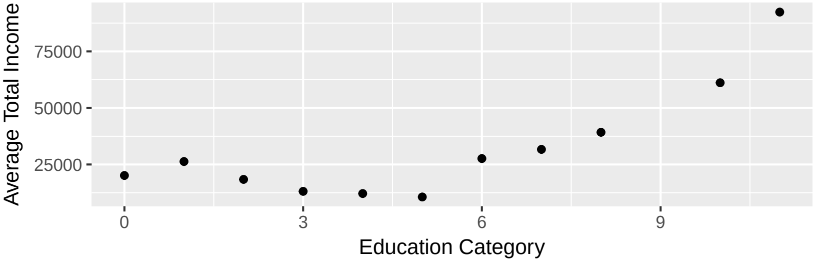
```
ggplot(data = acs_small,
       aes(x = EDUC, y = INCTOT)) +
    geom_point()
```



What's a major problem with this chart?

Plotting

```r
ggplot(data = acs_small) +
  geom_point(aes(x = EDUC, y = INCTOT)) +
  labs(x = "Education Category",
       y = "Average Total Income")
```



Plotting

- What does this graph tell us about returns to education?
- Why might this chart not give us the full picture?

ggplot2 Layering

How does the ggplot function work? By adding layers.

- Give ggplot() an input data set

```r
ggplot(data = plot_data...)
```

- Tell it which columns you want as your x and y variables

```r
aes(x = Years_Educ, y = Average_income)
```

- Tell it what type of shape you want

```r
+ geom_line()
```

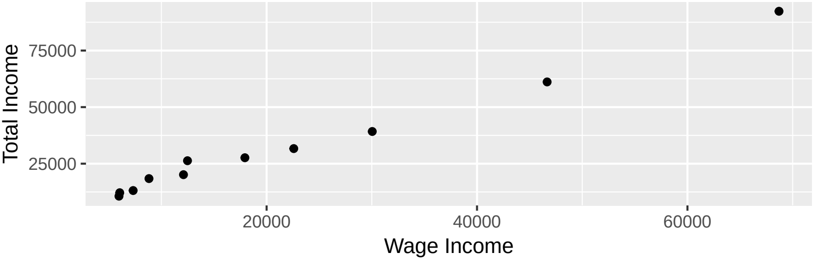
What happens if we comment out our geom layer?

More Plotting

- Now please create a chart that shows average wages on the x axis and average income on the y axis.
- Be sure to label your axes.
- What would you expect this to look like?
- What steps do you need to take to prepare the chart?

More Plotting

```r
ggplot(data = acs_2016_educ,
       aes(x = INCWAGE, y = INCTOT)) +
    geom_point() +
    labs(x = "Wage Income",
         y = "Total Income")
```



Age and Income

- Now read in the file acs_2016_age.csv
- What variables are we dealing with?
- What functions can we use to look at the data?
- Let's make a plot showing the gap between wage and total income.

mutate()

First we need to mutate our data frame:

- dplyr is great because things are named intuitively.
- What do you think mutate() does?
- What do you think the arguments for mutate() are? (what were the arguments for select()?)

mutate()

For some example data frame named `df` with columns named `column1` and `column2` we could create a new column using the following code:

```r
mutated_df <- mutate(df, column3 = column1 + column2)
```

Now try adding a new column to the acs_2016_age dataframe that is the difference between total and wage income.

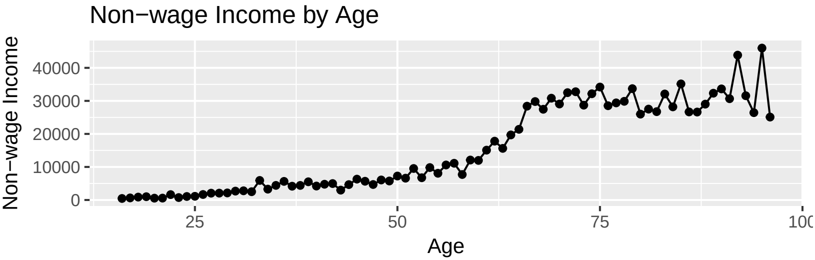
Age and Non-wage Income

Now make a plot:

- age on the x-axis
- nonwage_income on the y-axis
- appropriate axis labels and a title
- both points and a line

Age and Non-wage Income

```
ggplot(acs_2016_age,
       aes(x = AGE, y = nonwage_income)) +
  geom_point() +
  geom_line() +
  labs(x = "Age",
       y = "Non-wage Income",
       title = "Non-wage Income by Age")
```



- Why is this variable “noisy”?
- What’s happening around age 65?

Raw ACS Data

- So far we have only been dealing with group averages of variables
- Let's go back to the microdata and start to get a sense of the data
- Please read in the data file acs_2016_sample.csv
- How many observations are in this dataset? How many variables?

Raw ACS Data

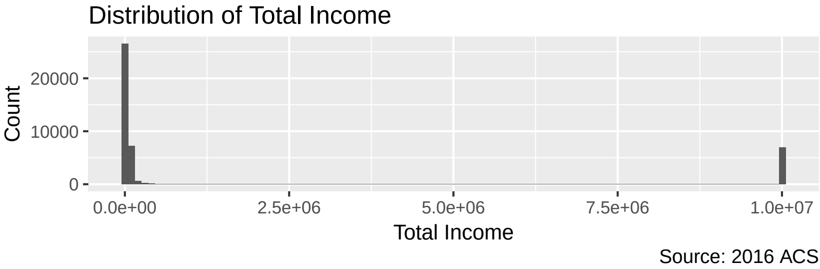
Let's make a histogram of the raw ACS income variable:

- What does a histogram show?
- What aesthetics does geom_histogram use?
- What other arguments does geom_histogram use?

Please make an appropriately labeled histogram of total income with axes labels, a title, and a caption

Raw ACS Income

```
ggplot(acs_2016_sample, aes(x = INCTOT)) +
  geom_histogram(bins = 100) +
  labs(x = "Total Income",
       y = "Count",
       title = "Distribution of Total Income",
       caption = "Source: 2016 ACS")
```



- Why does the distribution have a lot of mass at 10 million?

Raw ACS Income

- We need to remove the observations that are missing INCTOT (coded as 9999999)
- The appropriate dplyr verb is filter()
- What are the arguments to filter?
- What logical operator could we use?

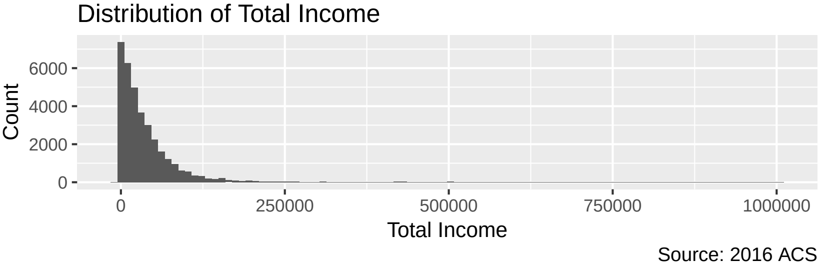
Logical Operators

- ==
- !=
- <, >
- <=, >=

```
acs_2016_sample <-
  filter(acs_2016_sample, INCTOT != 9999999)
```

Cleaned ACS Income

```r
ggplot(acs_2016_sample, aes(x = INCTOT)) +
  geom_histogram(bins = 100) +
  labs(x = "Total Income",
       y = "Count",
       title = "Distribution of Total Income",
       caption = "Source: 2016 ACS")
```


Distribution of Total Income

Source: 2016 ACS

Cleaned ACS Income

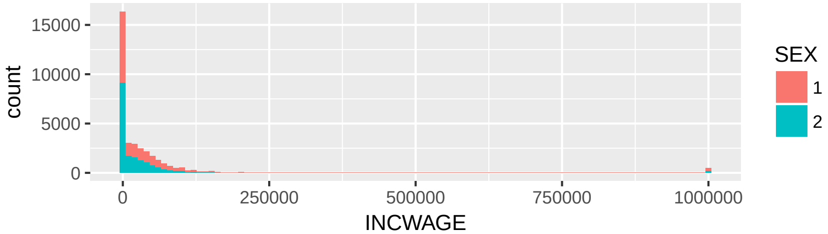
- Clean up the INCWAGE variable as well.
- Please created a new data frame called acs_filtered that only has people who make more than $100,000 from wages.
- What are the dimensions of this dataframe? How many observations did we filter out?
- Make a histogram of their total income

Color as an Axis

- Just like we can map the x and y aesthetics in ggplot2 we can map color and fill
- First we need to turn the SEX variable into a categorical variable using mutate() and factor()

```r
acs_2016_sample <-
  mutate(acs_2016_sample, SEX = factor(SEX))

ggplot(acs_2016_sample,
       aes(x = INCWAGE, fill = SEX)) +
  geom_histogram(bins = 100)
```



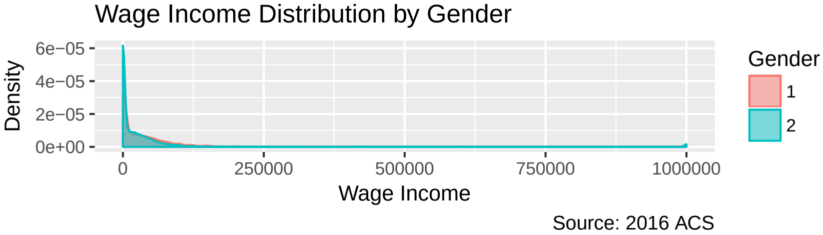
Color as an Axis

We can improve this chart

- Add axis labels and a title (be sure to label the fill axis)
- Change the alpha value to 0.5 (What's the alpha value?)
- Change the geom type from histogram to density (Why do we want to do this?)

Color as an Axis

```r
ggplot(acs_2016_sample,
       aes(x = INCWAGE, fill = SEX, color = SEX)) +
  geom_density(alpha = 0.5)+
  labs(x = "Wage Income",
       y = "Density",
       fill = "Gender",
       color = "Gender",
       title = "Wage Income Distribution by Gender",
       caption = "Source: 2016 ACS")
```



Wage Income Distribution by Gender

Source: 2016 ACS

Recoding Variables

- We want to recode the SEX variable so that instead of 1,2 it is “Male”, “Female”
- We can do this using the `ifelse()` function.
- Take a look at what `ifelse()` does:

```r
ifelse(1 == 1, "yes", "no")
```

```
## [1] "yes"
```

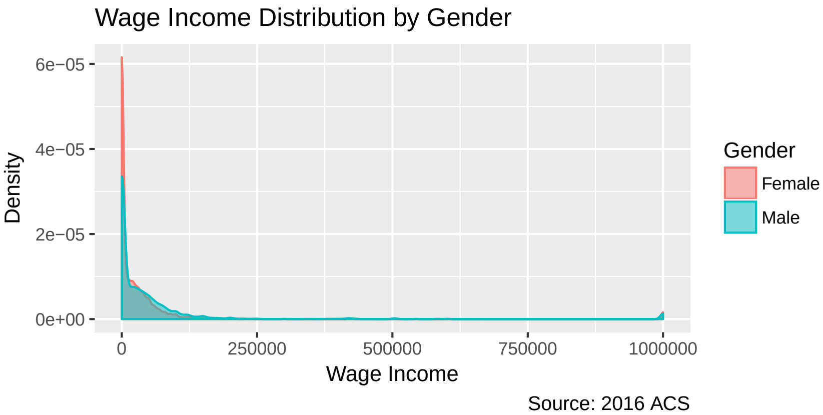
```r
ifelse(1 == 2, "yes", "no")
```

```
## [1] "no"
```

- How can we use `ifelse()` to recode the SEX variable?

Recoding Variables

```r
acs_2016_sample <-
  mutate(acs_2016_sample,
         SEX = ifelse(SEX == 1, "Male", "Female"))
```



Recap

Dplyr

- select()
- mutate(), ifelse()
- filter()

ggplot2

- aes()
- x, y, color, fill
- geom_line(), geom_point(), geom_histogram()