Module 1 Day 1

```
# load libraries
library()
library()
```

The American Community Survey

Use the `setwd()` function to set the working directory:

```
setwd("")
```

Now use the `read_csv()` function to read in the file acs_2016_educ.csv

```
acs_2016_educ <- read_csv("")
```

What are the dimensions of our data? What function should we use?

Use the `head()` function to look at the first few rows.

```
head()
```

dplyr

Load the library at the top of your program using `library(dplyr)`

Now try the `glimpse()` function from the dplyr library. What's the difference?

```
glimpse()
```

select()

The arguments to select are first the data frame followed by the columns you wish to keep. What does the following code return?

```
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

Try loading the ggplot2 library. Let's make a simple scatter plot:

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

What's a major problem with this chart?

Let's add axis labels

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

What does this graph tell us about returns to education?

Why might this chart not give us the full picture?

What happens if we comment out our geom layer?

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?

```r
ggplot(data = acs_2016_educ,
       aes(x = , y = )) +
    geom_  +
    labs(x = "",
         y = "")
```

Age and Income

Now read in the file acs_2016_age.csv

```r
acs_2016_age <- read_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. First we need to mutate our data frame. Try adding a new column to the acs_2016_age dataframe that is the difference between total and 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

```r
ggplot(,
       aes(x = , y = )) +
  geom_ +
  geom_ +
  labs(x = "",
       y = "",
       title = "")
```

Why is this variable "noisy"? What's happening around age 65?

Raw ACS Data

Please read in the data file acs_2016_sample.csv. How many observations are in this dataset? How many variables?

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

```
ggplot(acs_2016_sample, aes(x = )) +
  geom_histogram(bins = ) +
  labs(x = "",
       y = "",
       title = "",
       caption = "")
```

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

filter()

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

```
acs_2016_sample <-
  filter( , )
```

Now remake the same histogram but with the filtered data:

Clean up the INCWAGE variable as well:

```
acs_2016_sample <-
  filter( , )
```

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 thier total income

```
acs_filtered <-
  filter( , )

dim(acs_filtered)
```

```
ggplot(acs_2016_sample, aes(x = )) +
  geom_histogram(bins = ) +
  labs(x = "",
       y = "",
       title = "",
       caption = "")
```

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()

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

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

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?)

```
ggplot(acs_2016_sample,
       aes(x = , fill = , color = )) +
  geom_density(alpha = )+
  labs(x = "",
       y = "",
       fill = "",
       color = "",
       title = "",
       caption = "")
```

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.

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