Icebreaker

What's one thing we would never guess about you?

SOC245: Visualizing Data
Precept 2: Central Tendency and Spread

Chris Felton and Vikram Ramaswamy

Freshman Scholars Institute
Princeton University

July 12, 2022

Recall from Lecture

Recall from Lecture

- Central tendency: mean vs. median
  - How are each defined? What different information do they provide?

Recall from Lecture

- Central tendency: mean vs. median
  - How are each defined? What different information do they provide?
- Frequency distributions

R scripts

- On Friday, we plotted the life expentancy of different countries against the per-capita income from the gapminder dataset.

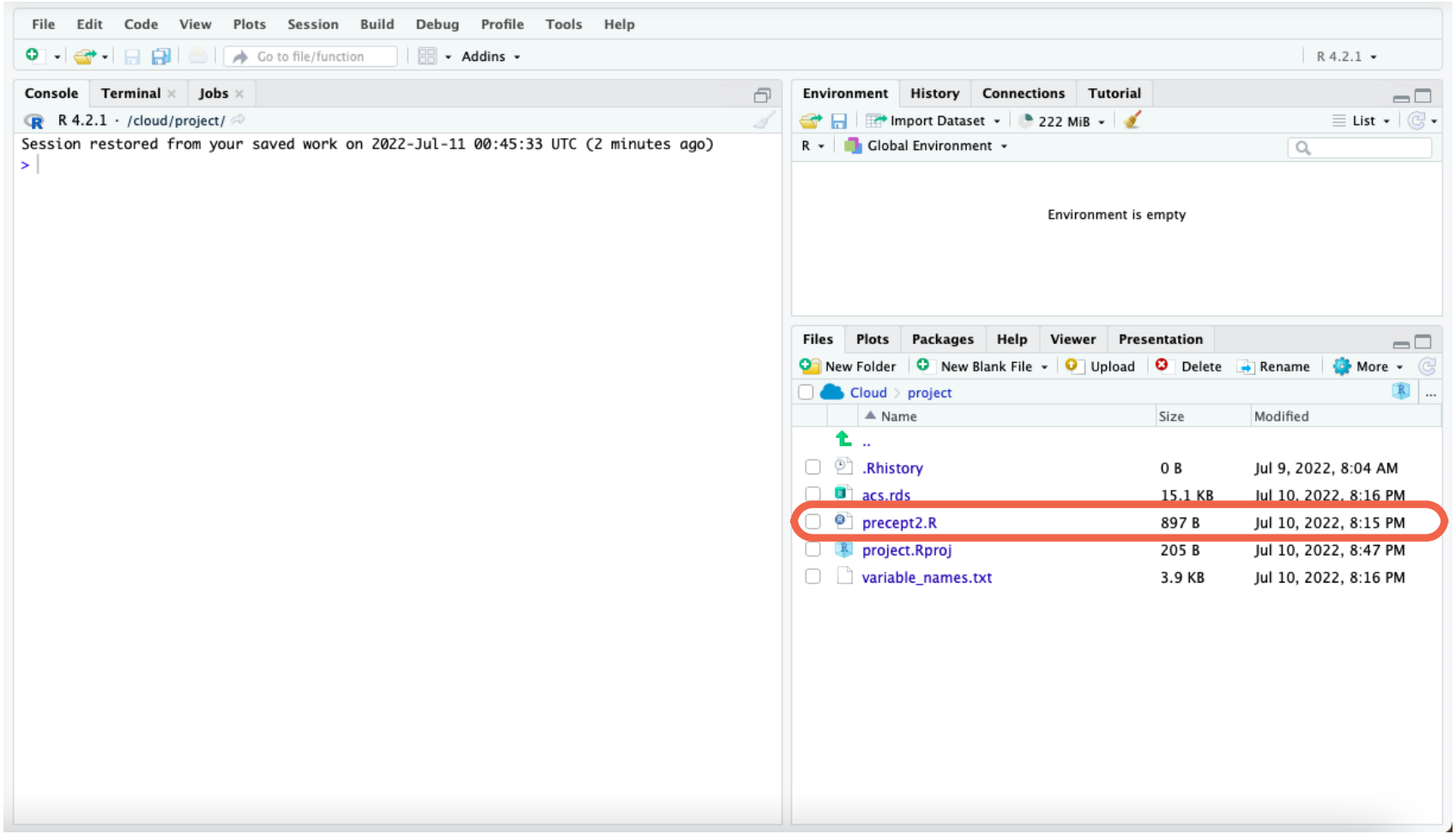
R scripts

- On Friday, we plotted the life expentancy of different countries against the per-capita income from the gapminder dataset.
- What if we want to rerun this today?

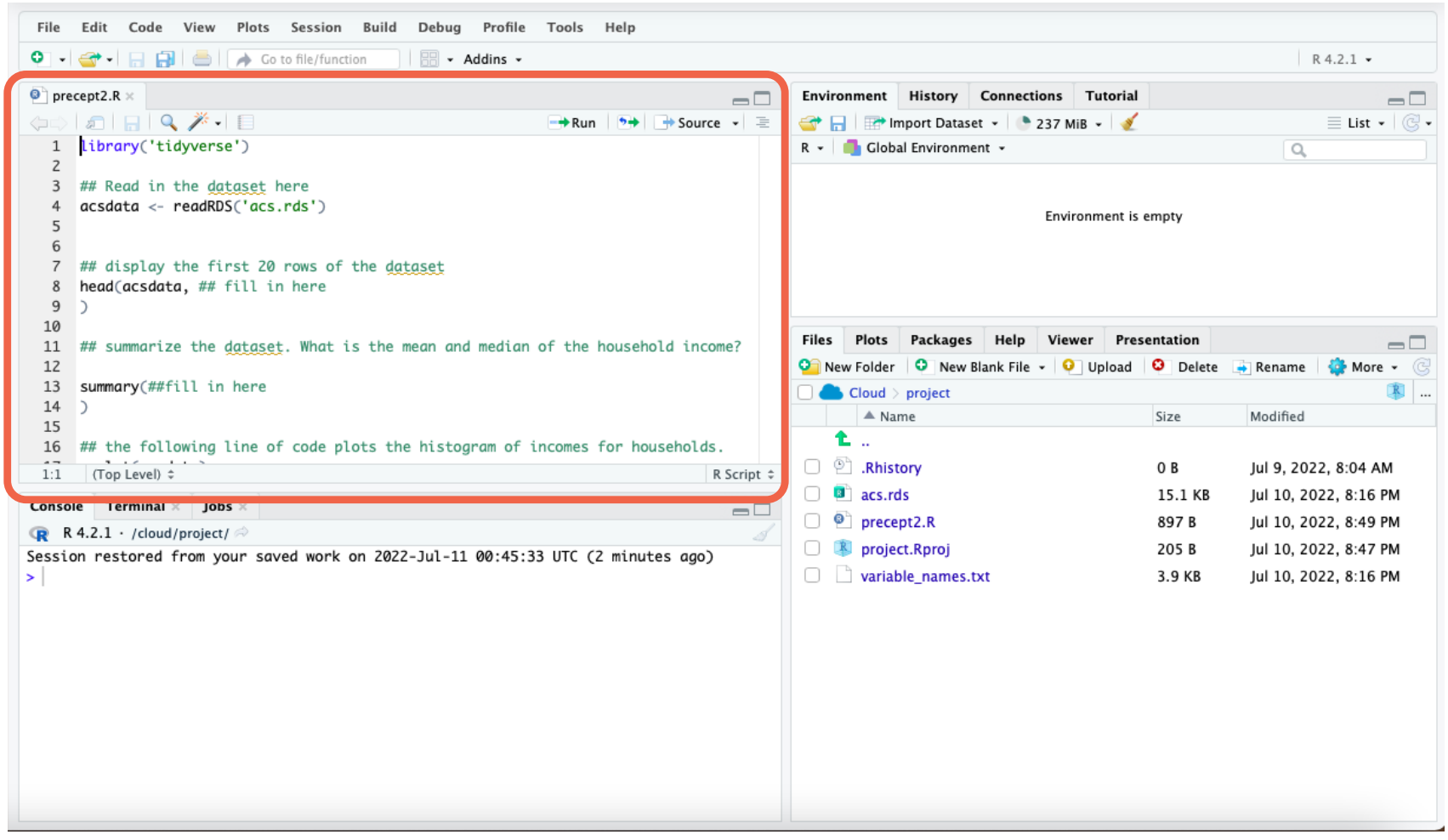
R scripts

- On Friday, we plotted the life expentancy of different countries against the per-capita income from the gapminder dataset.
- What if we want to rerun this today?
- R scripts let us keep a record of the code, letting us save code.

R Script


File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
R 4.2.1
Console Terminal Jobs
R 4.2.1 · /cloud/project/
Session restored from your saved work on 2022-Jul-11 00:45:33 UTC (2 minutes ago)
>
Environment History Connections Tutorial
Import Dataset
222 MiB
List
R
Global Environment
Environment is empty
Files Plots Packages Help Viewer Presentation
New Folder
New Blank File
Upload
Delete
Rename
More
Cloud
project
Name
Size
Modified
..
.Rhistory 0 B Jul 9, 2022, 8:04 AM
acs.rds 15.1 KB Jul 10, 2022, 8:16 PM
precept2.R 897 B Jul 10, 2022, 8:15 PM
project.Rproj 205 B Jul 10, 2022, 8:47 PM
variable_names.txt 3.9 KB Jul 10, 2022, 8:16 PM

R scripts


File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
R 4.2.1
precept2.R
Run
Source
1 library('tidyverse')
2
3 ## Read in the dataset here
4 acsdata <- readRDS('acs.rds')
5
6
7 ## display the first 20 rows of the dataset
8 head(acsdata, ## fill in here
9)
10
11 ## summarize the dataset. What is the mean and median of the household income?
12
13 summary(##fill in here
14)
15
16 ## the following line of code plots the histogram of incomes for households.
1:1
(Top Level)
R Script
Console
Terminal
Jobs
R 4.2.1 · /cloud/project/
Session restored from your saved work on 2022-Jul-11 00:45:33 UTC (2 minutes ago)
>
Environment
History
Connections
Tutorial
Import Dataset
237 MiB
List
R
Global Environment
Environment is empty
Files
Plots
Packages
Help
Viewer
Presentation
New Folder
New Blank File
Upload
Delete
Rename
More
Cloud
project
Name
Size
Modified
..
.Rhistory 0 B Jul 9, 2022, 8:04 AM
acs.rds 15.1 KB Jul 10, 2022, 8:16 PM
precept2.R 897 B Jul 10, 2022, 8:49 PM
project.Rproj 205 B Jul 10, 2022, 8:47 PM
variable_names.txt 3.9 KB Jul 10, 2022, 8:16 PM

R scripts

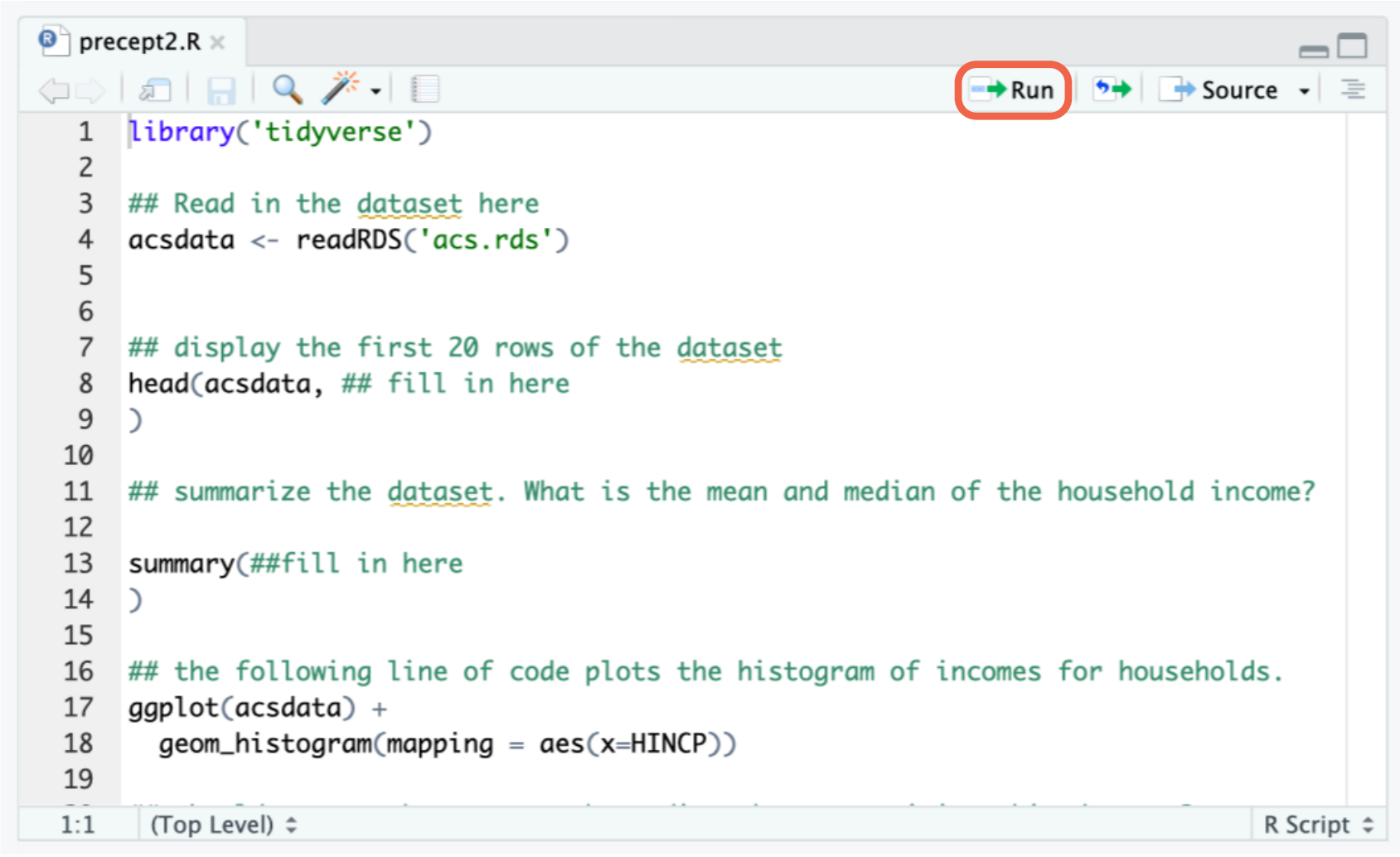
precept2.R

Run | Source

```r
library('tidyverse')

## Read in the dataset here
acsdata <- readRDS('acs.rds')


## display the first 20 rows of the dataset
head(acsdata, ## fill in here
)

## summarize the dataset. What is the mean and median of the household income?

summary(##fill in here
)

## the following line of code plots the histogram of incomes for households.
ggplot(acsdata) +
  geom_histogram(mapping = aes(x=HINCP))

```

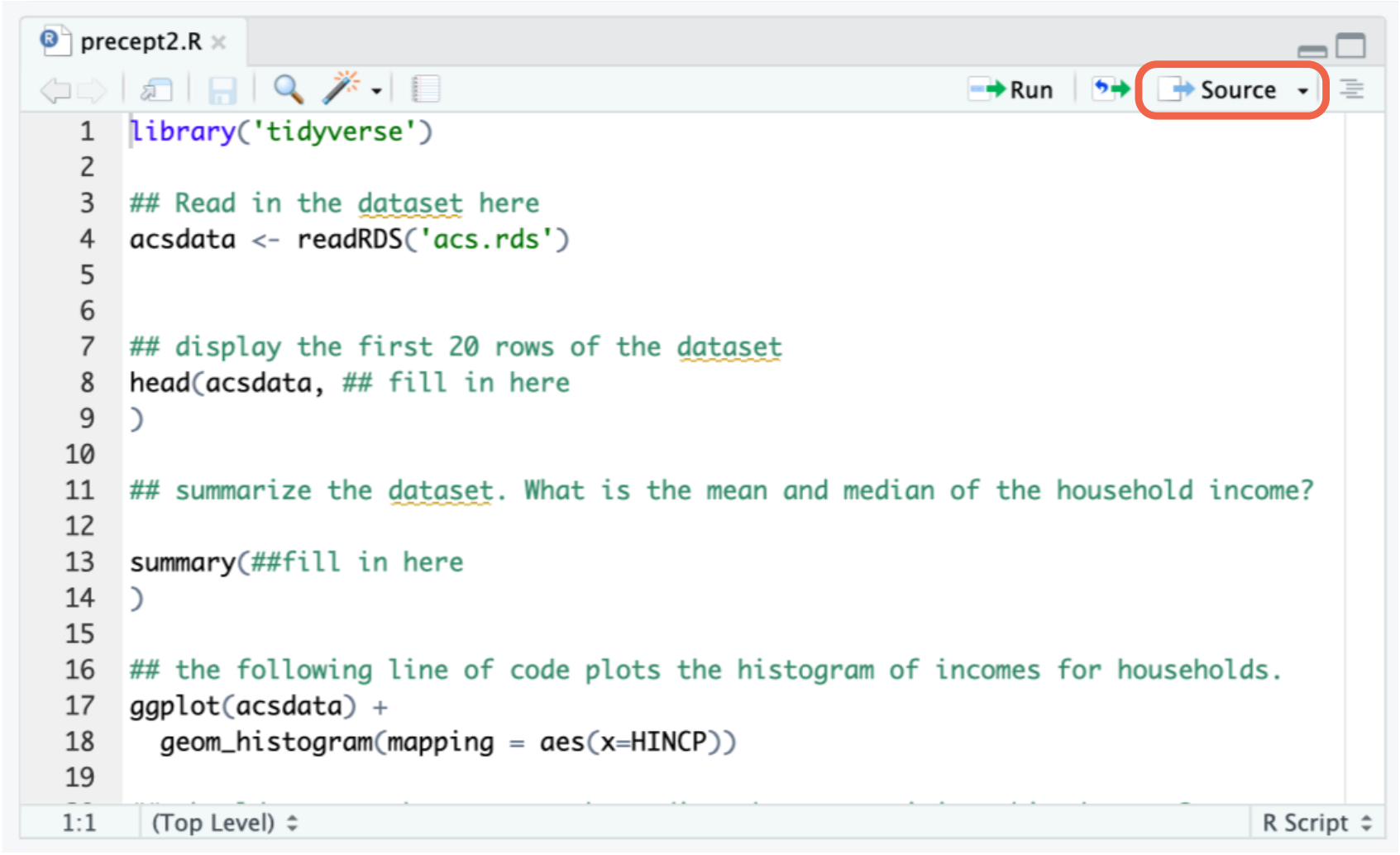
1:1 (Top Level) R Script

R scripts



To run the line of code where the cursor is, click `Run` (or use `Cmd + Enter`)

R scripts



To run the entire script, click Source (or use Cmd + Shift + Enter)

R scripts

precept2.R

Run Source

```
1  library('tidyverse')
2
3  ## Read in the dataset here
4  acsdata <- readRDS('acs.rds')
5
6
7  ## display the first 20 rows of the dataset
8  head(acsdata, ## fill in here
9  )
10
11 ## summarize the dataset. What is the mean and median of the household income?
12
13 summary(##fill in here
14 )
15
16 ## the following line of code plots the histogram of incomes for households.
17 ggplot(acsdata) +
18   geom_histogram(mapping = aes(x=HINCP))
19
```

1:1 (Top Level) R Script

Lines starting with a # are **comments**. These are typically used to explain the code.

R scripts

precept2.R

Run Source

```
1  library('tidyverse')
2
3  ## Read in the dataset here
4  acsdata <- readRDS('acs.rds')
5
6
7  ## display the first 20 rows of the dataset
8  head(acsdata, ## fill in here
9  )
10
11 ## summarize the dataset. What is the mean and median of the household income?
12
13 summary(##fill in here
14 )
15
16 ## the following line of code plots the histogram of incomes for households.
17 ggplot(acsdata) +
18   geom_histogram(mapping = aes(x=HINCP))
19
```

1:1 (Top Level) R Script

Lines starting with a # are **comments**. These are typically used to explain the code.

R scripts

precept2.R

Run Source

```r
library('tidyverse')

## Read in the dataset here
acsdata <- readRDS('acs.rds')


## display the first 20 rows of the dataset
head(acsdata, ## fill in here
)

## summarize the dataset. What is the mean and median of the household income?

summary(##fill in here
)

## the following line of code plots the histogram of incomes for households.
ggplot(acsdata) +
  geom_histogram(mapping = aes(x=HINCP))

```

1:1 (Top Level) R Script

For the first 2 assignments, you'll submit an R script along with any plots and a text file with answers.

Question for today

What is the income of a typical person from Mercer county?

Question for today

What is the income of a typical person from Mercer county?

(Mercer county is where we are right now!)

Data

- Working with the American Community Survey (ACS) dataset (`https://data.census.gov/mdat/#/`)
- Contains infomation about household incomes, race, gender, etc.
- We're using the 2019 data from just Mercer county.

Loading and Viewing a dataset

Do you remember the command to read in a dataset?

Loading and Viewing a dataset

Do you remember the command to read in a dataset?

```
acsdata <- readRDS("/cloud/project/acs.rds")
```

Loading and Viewing a dataset

Do you remember the command to read in a dataset?

```
acsdata <- readRDS("/cloud/project/acs.rds")
```

How about the command to view the top rows of a dataset?

Loading and Viewing a dataset

Do you remember the command to read in a dataset?

```
acsdata <- readRDS("/cloud/project/acs.rds")
```

How about the command to view the top rows of a dataset?

```
head(acsdata)
```

You try!

- In `precept2.R`, fill in the command to read and view the first few observations from the `acs` dataset

You try!

- In `precept2.R`, fill in the command to read and view the first few observations from the `acs` dataset
- Look up `?head`. Can you figure out how to change `head` to display the first 20 observations?

Summarizing the data

Try running this block of code:

```
summary(acsdata)
```

Summarizing the data

Try running this block of code:

```
summary(acsdata)
```

`summary` **allows you to quickly look at a table showing the central tendency and spread of the data**

Summarizing the data

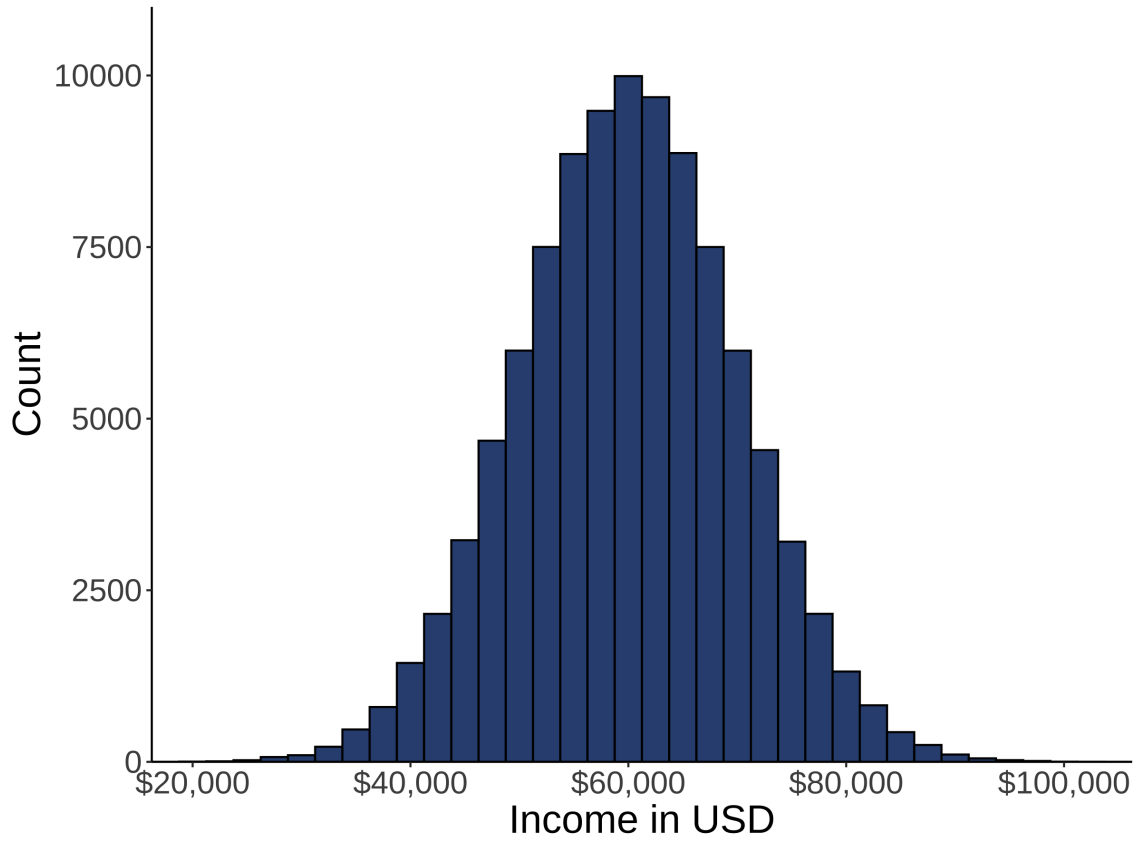
Try running this block of code:

```
summary(acsdata)
```

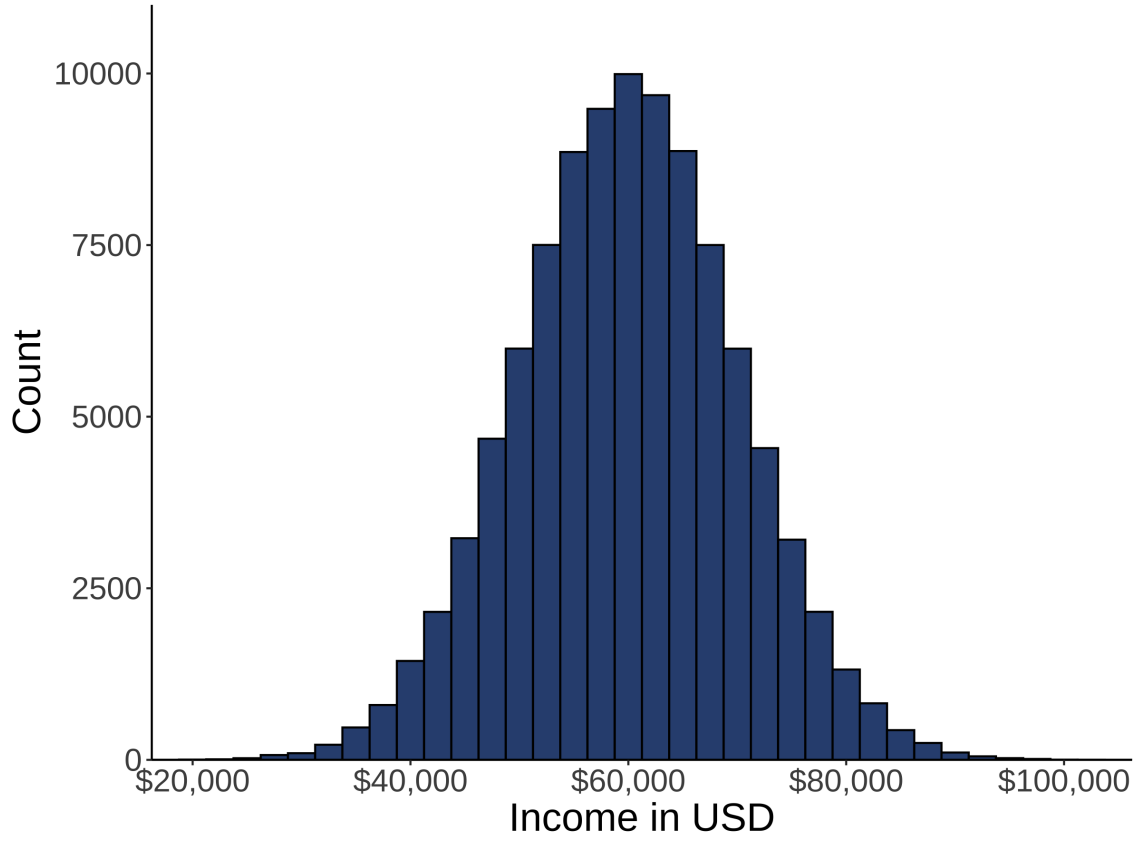
`summary` **allows you to quickly look at a table showing the central tendency and spread of the data**

The mean and median are very different for this dataset. What do we expect the *distribution* to look like?

Recall: Frequency distributions

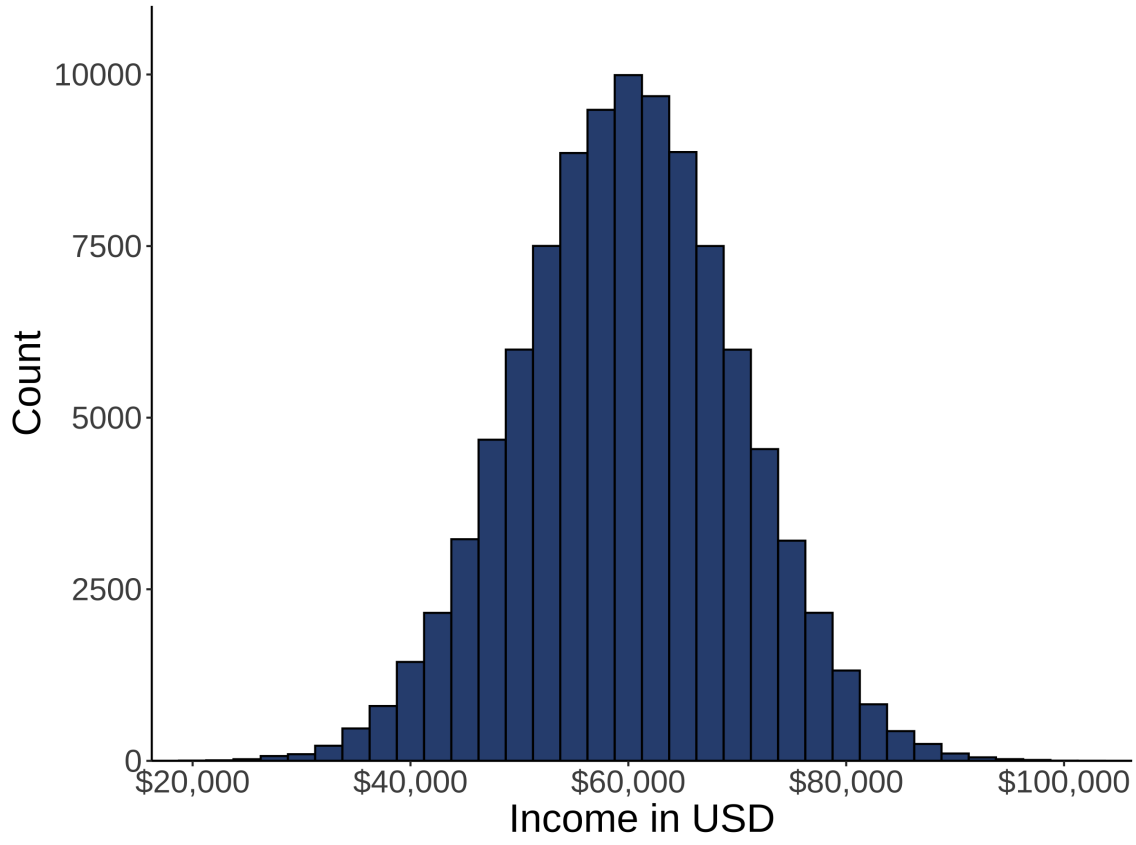


Recall: Frequency distributions



- Each bar covers a range of values $2,500 wide.
- Height of bar is number of people whose income falls in that range.

Recall: Frequency distributions



How can we plot this?

Introduction to `ggplot2`

- Going to be using `ggplot2` to make most of our visualizations within this course
- One of the packages within `tidyverse`
- Need to *install* and *load* this package.

Introduction to `ggplot2`

- Going to be using `ggplot2` to make most of our visualizations within this course
- One of the packages within `tidyverse`
- Need to *install* and *load* this package.
- **Install**: Similar to installing apps on your phone. Syntax: `install.packages("tidyverse")`
  - Only need to run once
  - We've already done this for you!

Introduction to `ggplot2`

- Going to be using `ggplot2` to make most of our visualizations within this course
- One of the packages within `tidyverse`
- Need to *install* and *load* this package.
- **Install**: Similar to installing apps on your phone. Syntax: `install.packages("tidyverse")`
  - Only need to run once
  - We've already done this for you!
- **Load**: Tells the `R` compiler that you will be using functions from this package. Syntax: `library("tidyverse")`
  - Need to run this each time. Think of difference between installing a lightbulb and switching on a lightbulb.

Introduction to ggplot2

- Basic syntax for a histogram:

```
ggplot(data=acsdata) +
    geom_histogram(mapping = aes(x=HINCP))
```

- Note: the package name is ggplot2, but the command is (confusingly) ggplot

Introduction to ggplot2

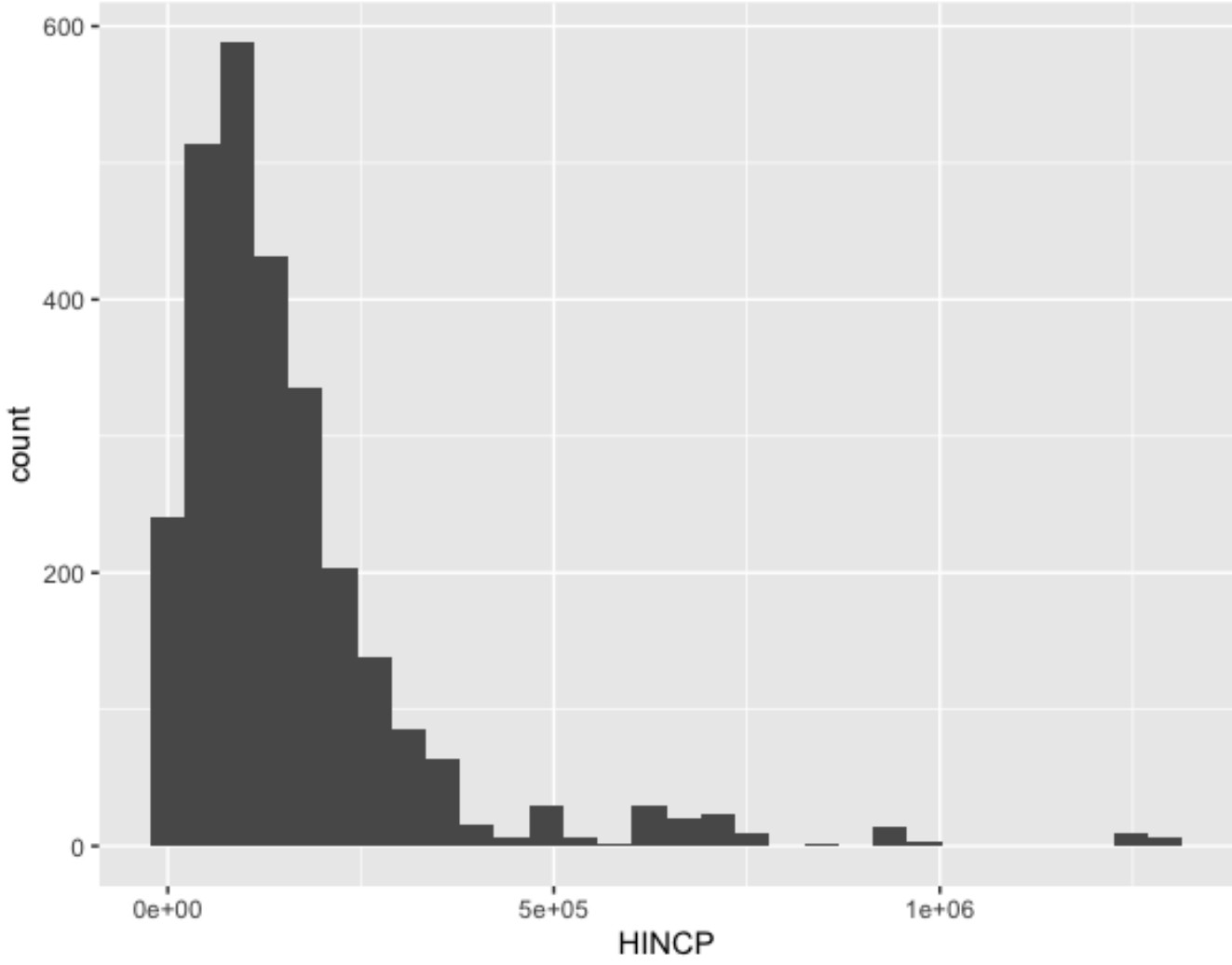
- Basic syntax for a histogram:

```
ggplot(data=acsdata) +
    geom_histogram(mapping = aes(x=HINCP))
```

- Note: the package name is ggplot2, but the command is (confusingly) ggplot

What does this output?

Introduction to ggplot2



You try!

Try summarizing the data and plotting a simple histogram.

Introduction to ggplot2

```
ggplot(data=acsdata) +
    geom_histogram(mapping = aes(x=HINCP))
```

Let's break this down.

- `ggplot()`: This creates an empty plot that we can add layers to.
- `ggplot(data=acsdata)`: says that we're going to be using this data for the rest of this plot, so we don't need to specify it again.

Introduction to ggplot2

```
ggplot(data=acsdata) +
    geom_histogram(mapping = aes(x=HINCP))
```

Let's break this down.

- `geom_histogram` adds a histogram to the plot.
  - lots of `geom` functions you can use! Common ones: `geom_point` (for a scatter plot), `geom_boxplot` (for a box plot), `geom_vline`, `geom_hline` (for vertical and horizontal lines), etc.

Introduction to ggplot2

```
ggplot(data=acsdata) +
    geom_histogram(mapping = aes(x=HINCP))
```

Let's break this down.

- `mapping=aes(x=HINCP)` : what the graph is going to contain; i.e, the x-axis contains buckets with the household income.
- `mapping=aes()` maps the data to the plot. Try running the previous command, but with

```
geom_histogram(mapping=aes(x=HINCP, color="blue"))
```

Introduction to ggplot2

```
ggplot(data=acsdata) +
    geom_histogram(mapping = aes(x=HINCP))
```

Let's break this down.

- `mapping=aes(x=HINCP)` : what the graph is going to contain; i.e, the x-axis contains buckets with the household income.
- `mapping=aes()` maps the data to the plot. Try running the previous command, but with

```
geom_histogram(mapping=aes(x=HINCP, color="blue"))
```

- What if instead we run:

```
geom_histogram(mapping=aes(x=HINCP), color="blue")
```

Plotting a histogram: binwidths

```
‘stat_bin()’ using ‘bins = 30’.  Pick better value
with ‘binwidth’
```

You might have seen this warning when plotting the histogram.

Plotting a histogram: binwidths

`‘stat_bin()’ using ‘bins = 30’. Pick better value with ‘binwidth’`

You might have seen this warning when plotting the histogram.

- A bin size that's too small looks very noisy
- However a bin size that's too large can hide important trends.

You try!

Try varying the bin size by adding `binwidth` to the parameters of `geom_histogram`. What's a good size to use?

ggplot2: Grammar of graphics

Designed to mimic human thought when creating graphics, breaking up a visualization. Plots consist of

- `data`: What data we want to visualize

ggplot2: Grammar of graphics

Designed to mimic human thought when creating graphics, breaking up a visualization. Plots consist of

- `data`: What data we want to visualize
- `mappings`: How the data maps to the plot

ggplot2: Grammar of graphics

Designed to mimic human thought when creating graphics, breaking up a visualization. Plots consist of

- `data`: What data we want to visualize
- `mappings`: How the data maps to the plot
  - `geom`: What type of plot we want (histogram / scatter plot / ...)

ggplot2: Grammar of graphics

Designed to mimic human thought when creating graphics, breaking up a visualization. Plots consist of

- `data`: What data we want to visualize
- `mappings`: How the data maps to the plot
  - `geom`: What type of plot we want (histogram / scatter plot / ...)
  - `stat`: Statistical transformations, like the size of the bins in the histogram

ggplot2: Grammar of graphics

Designed to mimic human thought when creating graphics, breaking up a visualization. Plots consist of

- `data`: What data we want to visualize
- `mappings`: How the data maps to the plot
  - `geom`: What type of plot we want (histogram / scatter plot / ...)
  - `stat`: Statistical transformations, like the size of the bins in the histogram
  - `coord`: a coordinate system (We're going to stick to the Cartesian coordinate system, but other options are available)

ggplot2: Grammar of graphics

Designed to mimic human thought when creating graphics, breaking up a visualization. Plots consist of

- `data`: What data we want to visualize
- `mappings`: How the data maps to the plot
  - `geom`: What type of plot we want (histogram / scatter plot / ...)
  - `stat`: Statistical transformations, like the size of the bins in the histogram
  - `coord`: a coordinate system (We're going to stick to the Cartesian coordinate system, but other options are available)
  - `scale`: Mapping values in the data space to the coordinate space. Uses color, shape, size, etc.

ggplot2: Grammar of graphics

Designed to mimic human thought when creating graphics, breaking up a visualization. Plots consist of

- `data`: What data we want to visualize
- `mappings`: How the data maps to the plot
  - `geom`: What type of plot we want (histogram / scatter plot / ...)
  - `stat`: Statistical transformations, like the size of the bins in the histogram
  - `coord`: a coordinate system (We're going to stick to the Cartesian coordinate system, but other options are available)
  - `scale`: Mapping values in the data space to the coordinate space. Uses color, shape, size, etc.
  - `facet`: Breaks up data into subsets.

ggplot2: Grammar of graphics

Designed to mimic human thought when creating graphics, breaking up a visualization. Plots consist of

- `data`: What data we want to visualize
- `mappings`: How the data maps to the plot
  - `geom`: What type of plot we want (histogram / scatter plot / ...)
  - `stat`: Statistical transformations, like the size of the bins in the histogram
  - `coord`: a coordinate system (We're going to stick to the Cartesian coordinate system, but other options are available)
  - `scale`: Mapping values in the data space to the coordinate space. Uses color, shape, size, etc.
  - `facet`: Breaks up data into subsets.
  - `theme`: stylistic changes like font size.

Adding median values to a histogram

We can add additional lines to our plot using by adding layers to our graph. Suppose `med` is the median of the household incomes:

```
ggplot(data=acsdata) +
    geom_histogram(mapping = aes(x=HINCP)) +
    geom_vline(mapping = aes(xintercept=med))
```

You try!

Add a median line to your plot. How would you change the color of the line?

Changing axis labels

The current axis labels aren't that easy to understand.

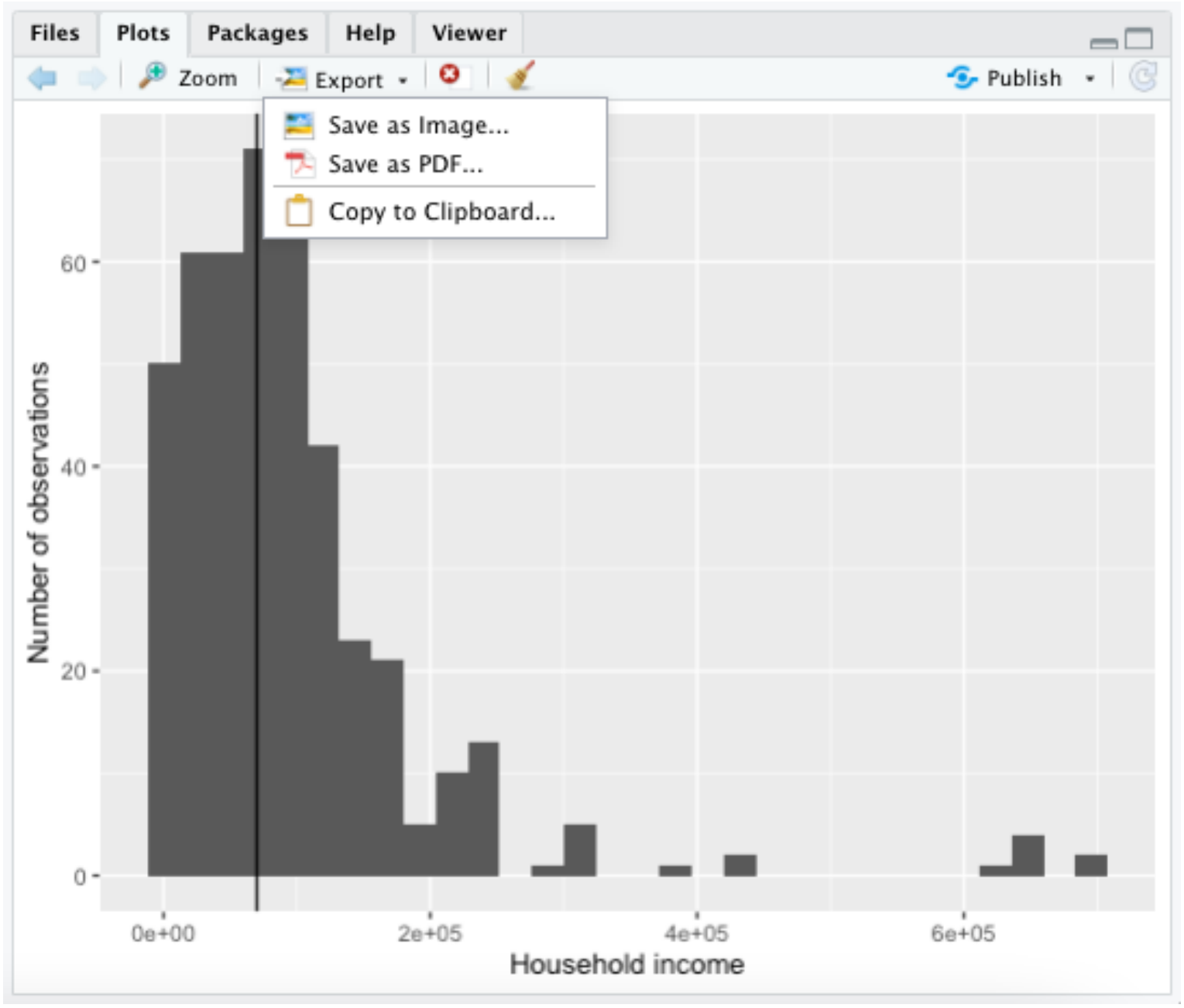
Changing axis labels

The current axis labels aren't that easy to understand.

- We can add more layers to our plot to change these!

```
ggplot(data=acsdata) +
    geom_histogram(mapping = aes(x=HINCP)) +
    geom_vline(mapping = aes(xintercept=med)) +
    xlab("Household income") +
    ylab("Number of observations")
```

Saving a plot



- This is what the income of all the people from Mercer county look like.
- What if we wanted to plot the income of just Black people in Mercer county?

Subsetting a dataset

- Want to get all rows in the dataset that correspond to Black individuals

Subsetting a dataset

- Want to get all rows in the dataset that correspond to Black individuals
- Type in `?filter`

Subsetting a dataset

- Want to get all rows in the dataset that correspond to Black individuals
- Type in `?filter`

```
filter(acsdata, RACBLK==1)
```

Subsetting a dataset

```
filter(acsdata, RACBLK==1)
```

- Syntax: dataset, followed by **condition**
- **Condition**: what we want to subset the dataset by.
  - `RACBLK==1` checks if the value of RACBLK is exactly 1
- What condition would filter out all Asian and Asian Americans in the dataset?

Subsetting a dataset

```
filter(acsdata, RACBLK==1)
```

- Syntax: dataset, followed by **condition**
- **Condition**: what we want to subset the dataset by.
  - `RACBLK==1` checks if the value of RACBLK is exactly 1
- What condition would filter out all Asian and Asian Americans in the dataset?
- How about if you wanted to find all people who are atleast 25 years old?

Subsetting a dataset

```
filter(acsdata, RACBLK==1)
```

- Syntax: dataset, followed by **condition**
- **Condition**: what we want to subset the dataset by.
  - `RACBLK==1` checks if the value of RACBLK is exactly 1
- What condition would filter out all Asian and Asian Americans in the dataset?
- How about if you wanted to find all people who are atleast 25 years old?

Note: `filter` is part of the '`dplyr`' package, which is one of the packages in `tidyverse`!

Putting it together

Make a histogram for household incomes of Black inviduals.
What's a good bin width to use?

What we did and what's next

- **What we learned**:
  - Summarizing data using `summary`
  - Plotting a histogram
  - Finding subsets of the dataset

What we did and what's next

- **What we learned**:
  - Summarizing data using `summary`
  - Plotting a histogram
  - Finding subsets of the dataset
- **Next precept**:
  - More about plotting histograms, learning about the spread of the distribution
  - Counts and proportions

What we did and what's next

- **What we learned**:
  - Summarizing data using `summary`
  - Plotting a histogram
  - Finding subsets of the dataset
- **Next precept**:
  - More about plotting histograms, learning about the spread of the distribution
  - Counts and proportions
- **Homework 1 will be available on Canvas today. Due on Thursday July 14th at 1:29 pm**
  - Goal for homework: You get to practice what you learned today with a different subset of the ACS dataset.