R06: ggplot2, Part 1

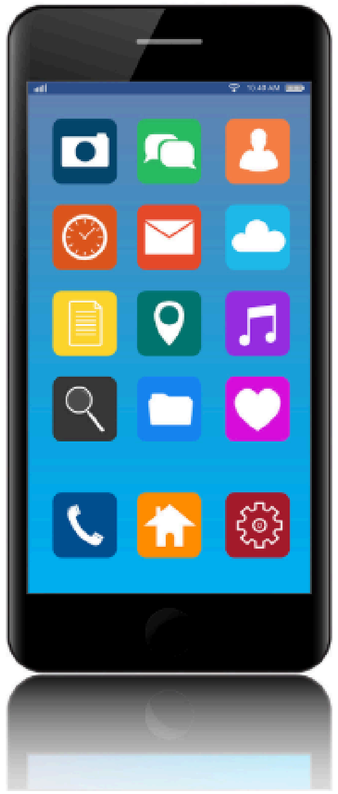
Meike Niederhausen and Nicky Wakim

2024-10-16

From last time: R Packages

A good analogy for R packages is that they are like apps you can download onto a mobile phone:

R: A new phone



R Packages: Apps you can download

ModernDive Figure 1.4

From last time: Install the packages listed below

- `knitr`
  - this might actually already be installed
  - check your packages list
- `tidyverse`
  - this is actually a bundle of packages
  - *Warning: it will take a while to install!!!*
  - see more info at https://tidyverse.tidyverse.org/
- `rstatix`
  - for summary statistics of a dataset
- `janitor`
  - for cleaning and exploring data
- `ggridges`
  - for creating ridgeline plots
- `devtools`
  - used to create R packages
  - for our purposes, needed to install some packages
- `oi_biostat_data`
  - this package is on github
  - **see the next slide for directions on how to install `oi_biostat_data`**
- `here`
  - More info in slides ahead

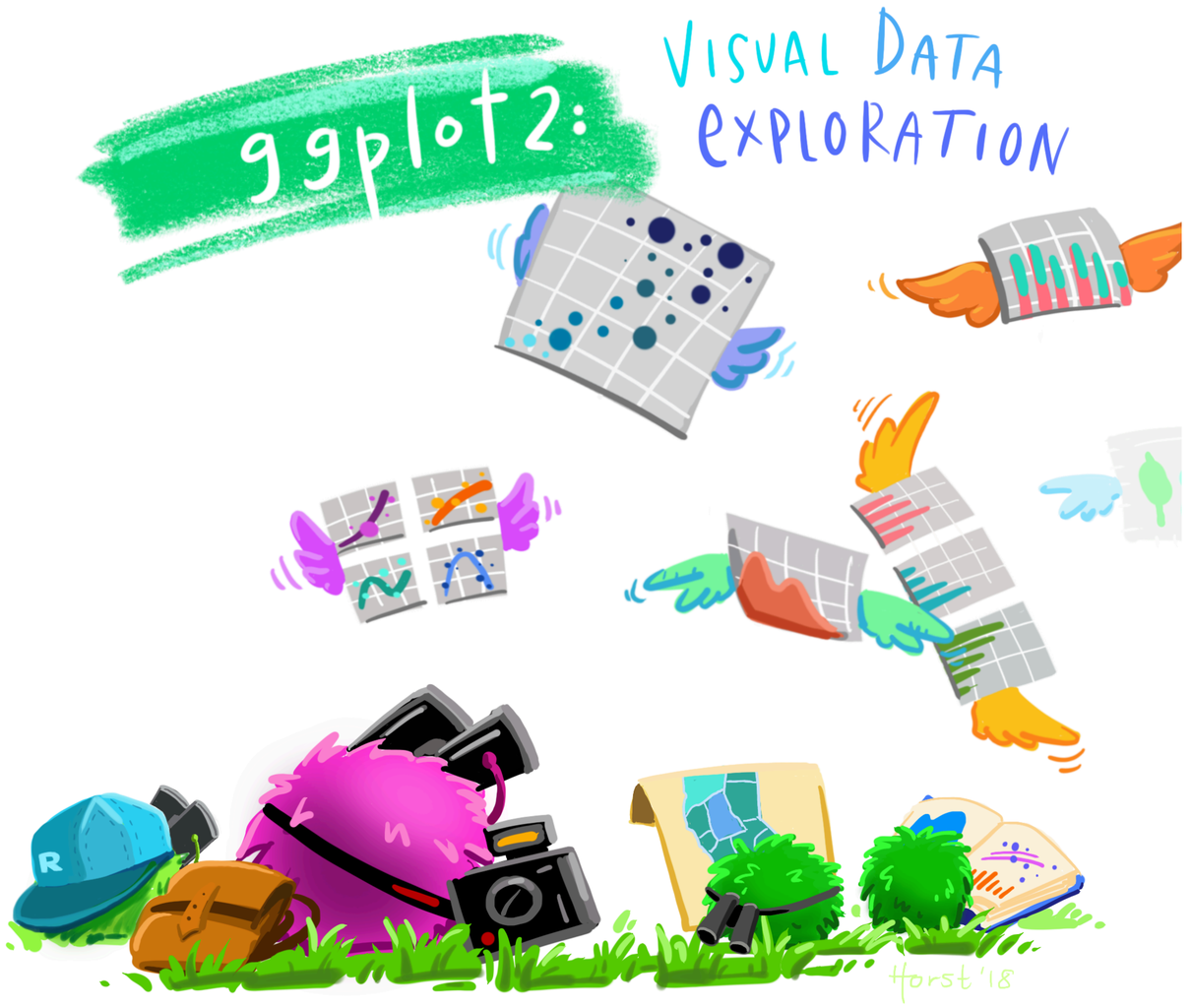
From last time: Load packages with `library()` command

- Tip: **at the top of your Qmd file,** create a chunk that loads all of the R packages you want to use in that file.
- Use the `library()` command to load each required package.
  - Packages need to be reloaded *every* time you open Rstudio.
  - `library()` commands to load needed packages *must* be in the Qmd file

```
# run these every time you open Rstudio
library(tidyverse) # contains ggplot2
library(oibiostat)
library(ggridges)
library(janitor)
library(rstatix)
library(knitr)
library(gtsummary) # NEW!!
```

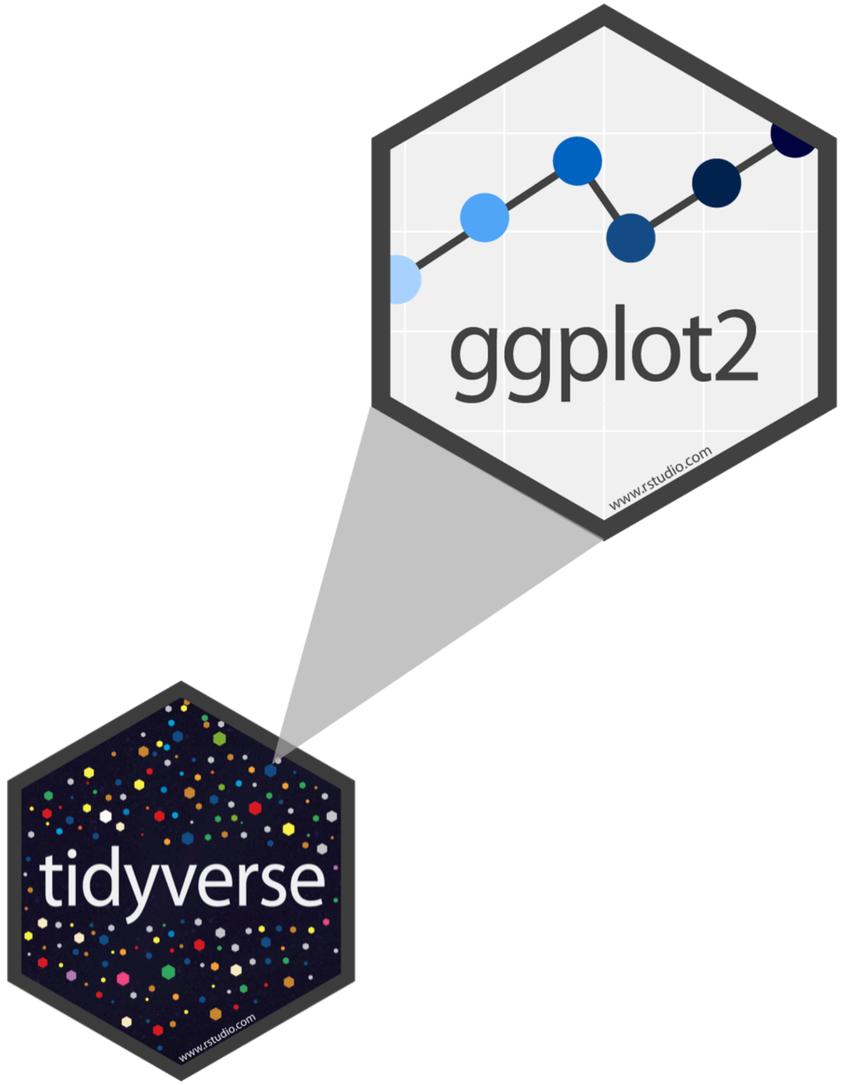
- You can check whether a package has been loaded or not
  - by looking at the Packages tab and
  - seeing whether it has been checked off or not

Introduction to `ggplot2`



Artwork by @allison_horst

`ggplot2` in tidyverse



- We talked about this in our review notes
  - I want to revisit it: always helps to have more examples!
  - This example is closer to the multivariable work we'll do in this class!

- **ggplot2** is tidyverse's data visualization package

- The gg in "ggplot2" stands for Grammar of Graphics

- It is inspired by the book **Grammar of Graphics** by Leland Wilkinson

Case Study Description

- In the US, individuals with developmental disabilities typically receive services and support from state governments.
  - California allocates funds to developmentally disabled residents through the *Department of Developmental Services (DDS)*
  - Recipients of DDS funds are referred to as "consumers."

- Dataset `dds.discr`
  - Sample of 1,000 DDS consumers (out of a total of ~ 250,000)
  - Data include age, sex assigned at birth, race/ethnicity, and annual DDS financial support per consumer

- For now, we are going to explore these data with R
- See Section 1.7.1 in the textbook for more details

Load `dds.discr` dataset from `oibiostat` package

- The textbook's datasets are in the R package `oibiostat`
- Make sure the `oibiostat` package is installed before running the code below.
- Load the `oibiostat` package and the dataset `dds.discr`

The code below needs to be run *every time* you restart R or render a Qmd file:

```
1 library(oibiostat)
2 data("dds.discr")
```

- After loading the dataset `dds.discr` using `data("dds.discr")`, you will see `dds.discr` in the Data list of the Environment window.

glimpse()

- We previously used the base R structure command `str()` to get information about variable types in a dataset (in R03: R basics part 2)
- Use `glimpse()` from the `tidyverse` package (technically it's from the `dplyr` package) to get information about variable types.
- `glimpse()` tends to have nicer output for `tibbles` than `str()`

```
1 library(tidyverse)
2 glimpse(dds.discr)  # from tidyverse package (dplyr)
```

```
Rows: 1,000
Columns: 6
$ id           <int> 10210, 10409, 10486, 10538, 10568, 10690, 10711, 10778, 1…
$ age.cohort   <fct> 13-17, 22-50, 0-5, 18-21, 13-17, 13-17, 13-17, 13-17, 13-…
$ age          <int> 17, 37, 3, 19, 13, 15, 13, 17, 14, 13, 13, 14, 15, 17, 20…
$ gender       <fct> Female, Male, Male, Female, Male, Female, Female, Male, F…
$ expenditures <int> 2113, 41924, 1454, 6400, 4412, 4566, 3915, 3873, 5021, 28…
$ ethnicity    <fct> White not Hispanic, White not Hispanic, Hispanic, Hispani…
```

Some things to note on this dataset

```
1 glimpse(dds.discr)  # from tidyverse package (dplyr)
```

```
Rows: 1,000
Columns: 6
$ id           <int> 10210, 10409, 10486, 10538, 10568, 10690, 10711, 10778, 1…
$ age.cohort   <fct> 13-17, 22-50, 0-5, 18-21, 13-17, 13-17, 13-17, 13-17, 13-…
$ age          <int> 17, 37, 3, 19, 13, 15, 13, 17, 14, 13, 13, 14, 15, 17, 20…
$ gender       <fct> Female, Male, Male, Female, Male, Female, Female, Male, F…
$ expenditures <int> 2113, 41924, 1454, 6400, 4412, 4566, 3915, 3873, 5021, 28…
$ ethnicity    <fct> White not Hispanic, White not Hispanic, Hispanic, Hispani…
```

- This happens in older datasets (and honestly some newer ones): gender and sex get conflated
  - I try to catch these issues before sharing datasets with you, but when we load datasets directly from the `oibiostat` package, I can't make these changes
  - If you are unfamiliar with the differences, please see this NIH site on sex and gender
- Also, race and ethnicity can be mislabeled
  - "White not hispanic" combines race and ethnicity
  - If you are unfamiliar with the differences, please see this APA site on race and ethnicity

`rename()`: one of the first things I usually do

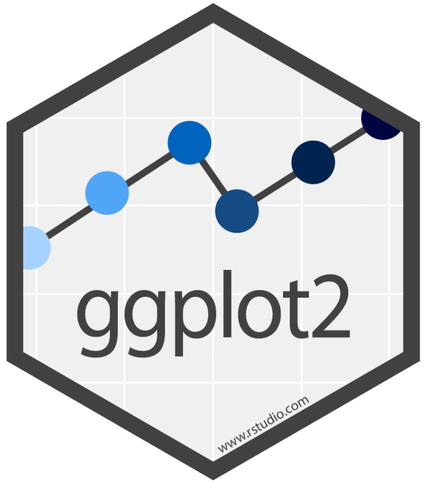
- I want to rename the variable, gender, to sex and rename ethnicity to r_e (race and ethnicity)

```
dds.discr1 = dds.discr %>%
  rename(sex_MF = gender,
         r_e = ethnicity)

glimpse(dds.discr1)
```

```
Rows: 1,000
Columns: 6
$ id           <int> 10210, 10409, 10486, 10538, 10568, 10690, 10711, 10778, 1…
$ age.cohort   <fct> 13-17, 22-50, 0-5, 18-21, 13-17, 13-17, 13-17, 13-17, 13-…
$ age          <int> 17, 37, 3, 19, 13, 15, 13, 17, 14, 13, 13, 14, 15, 17, 20…
$ sex_MF       <fct> Female, Male, Male, Female, Male, Female, Female, Male, F…
$ expenditures <int> 2113, 41924, 1454, 6400, 4412, 4566, 3915, 3873, 5021, 28…
$ r_e          <fct> White not Hispanic, White not Hispanic, Hispanic, Hispani…
```

Visualize numerical variables with `ggplot2`



ggplot

Artwork by @allison_horst

Basics of a ggplot

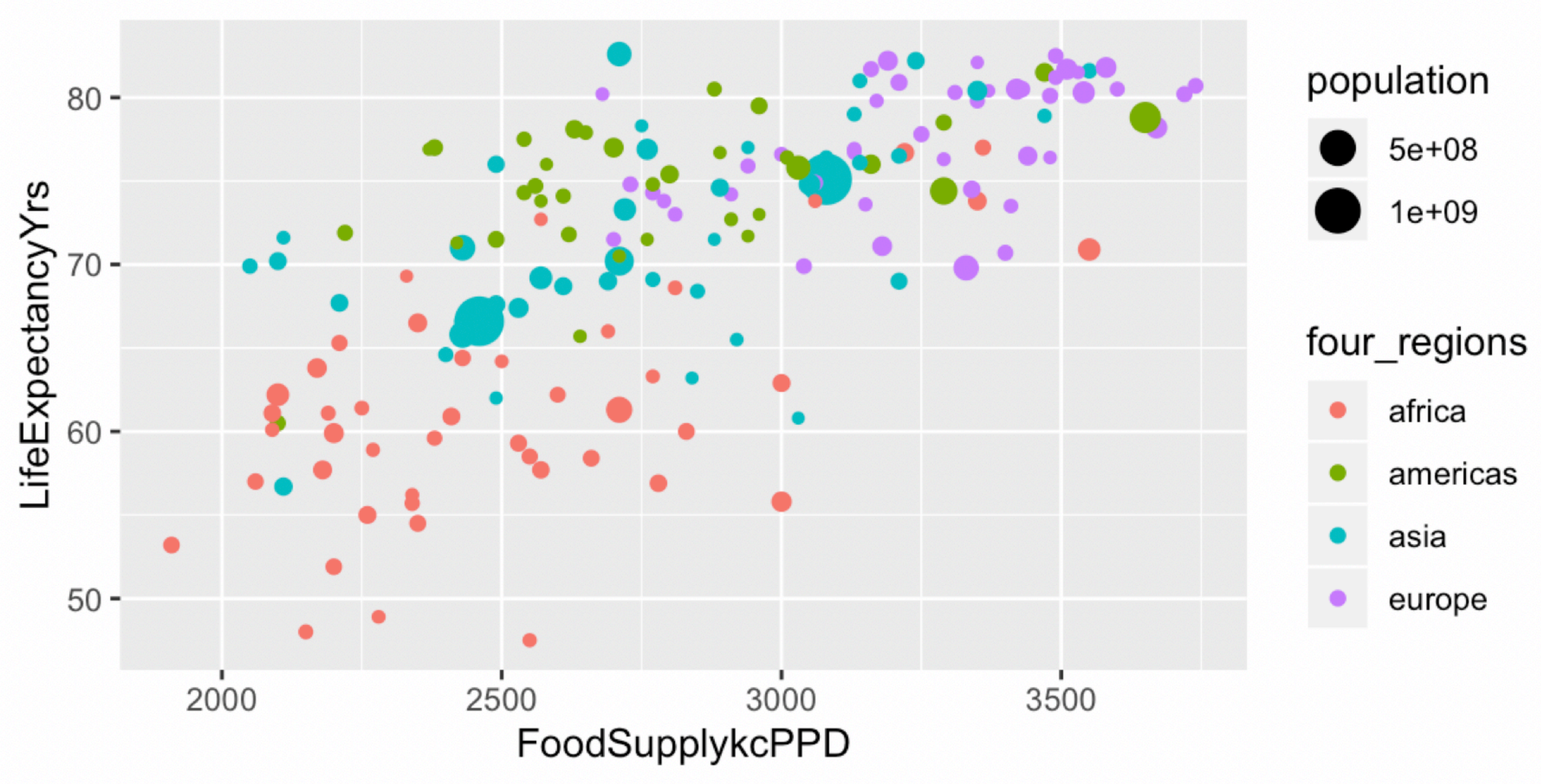
Function

Dataset

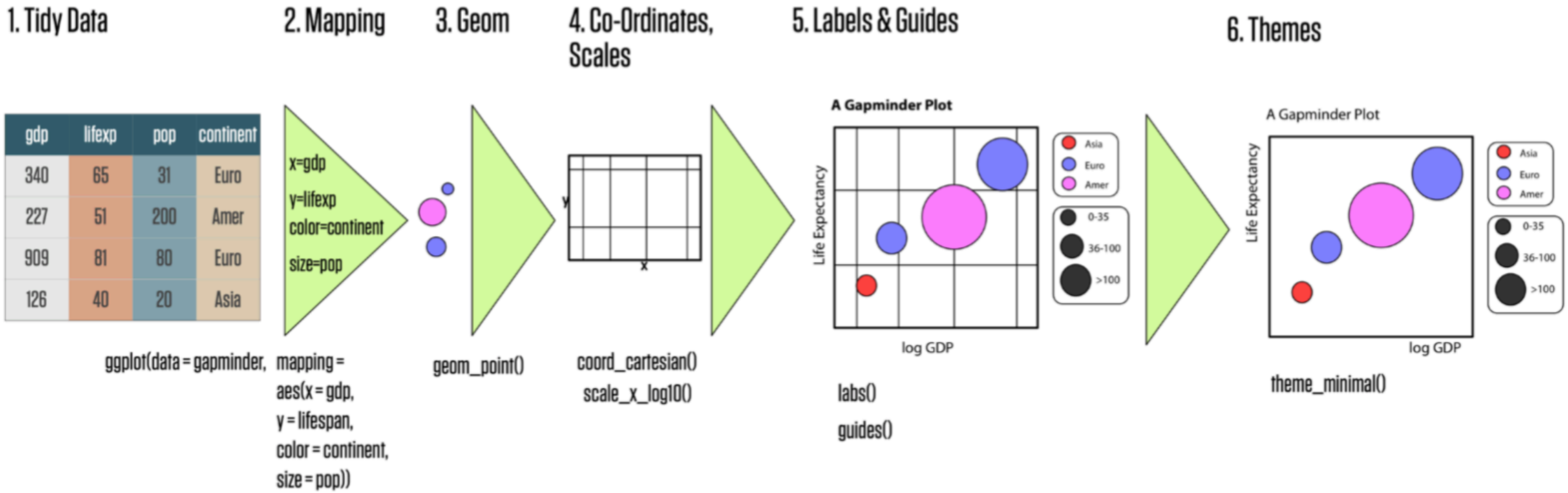
```
ggplot(data = gapminder2011,
       aes(x = FoodSupplykcPPD, y = LifeExpectancyYrs,
           color = four_regions, size = population)) +
       geom_point()
```

Which variables to plot

What kind of plot to make



Grammar of ggplot2

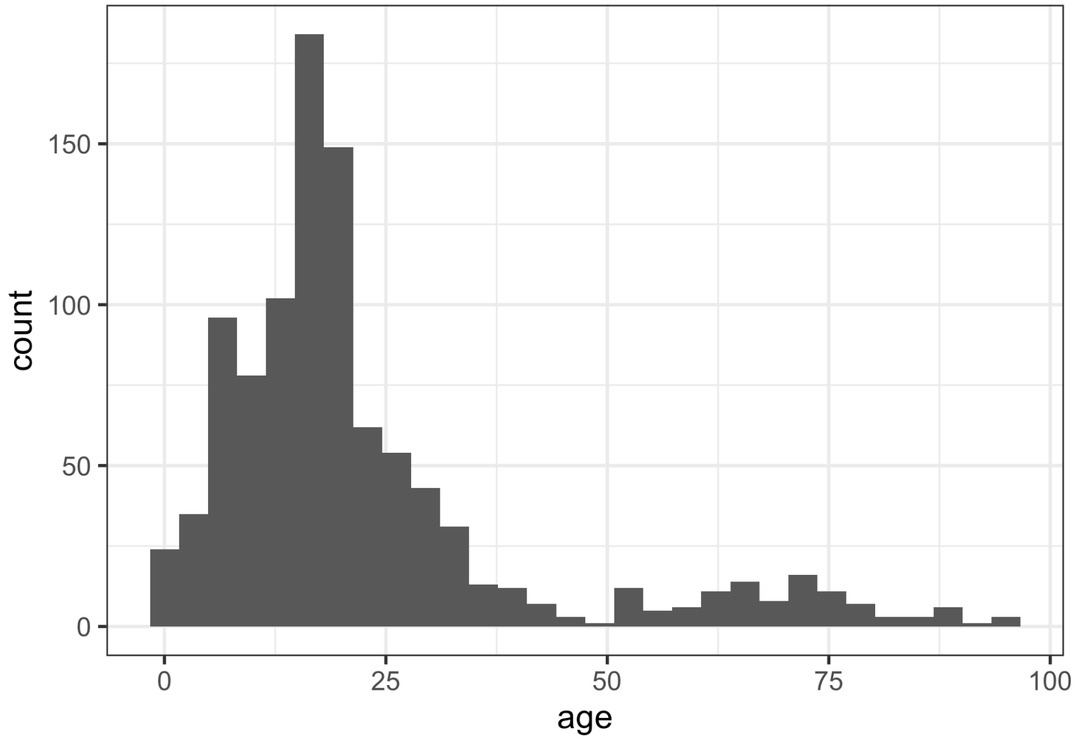


Kieran Healy

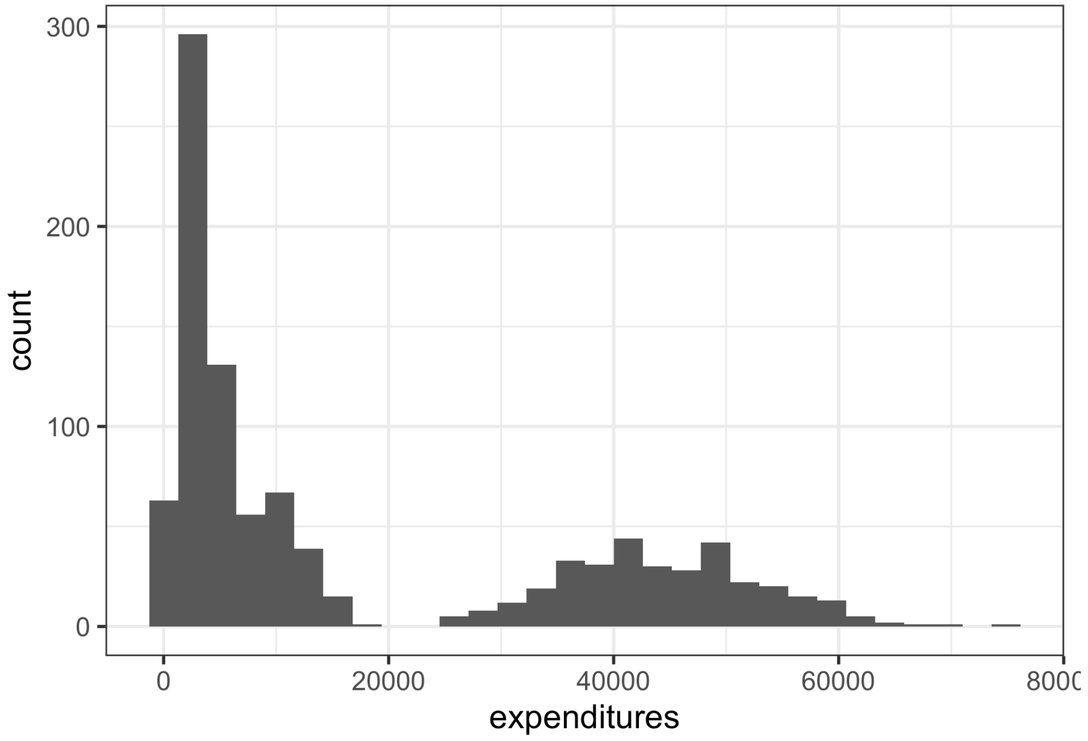
Histograms

What is being measured on the vertical axes?

```
1 ggplot(data = dds.discr1,
2        aes(x = age)) +
3   geom_histogram()
```

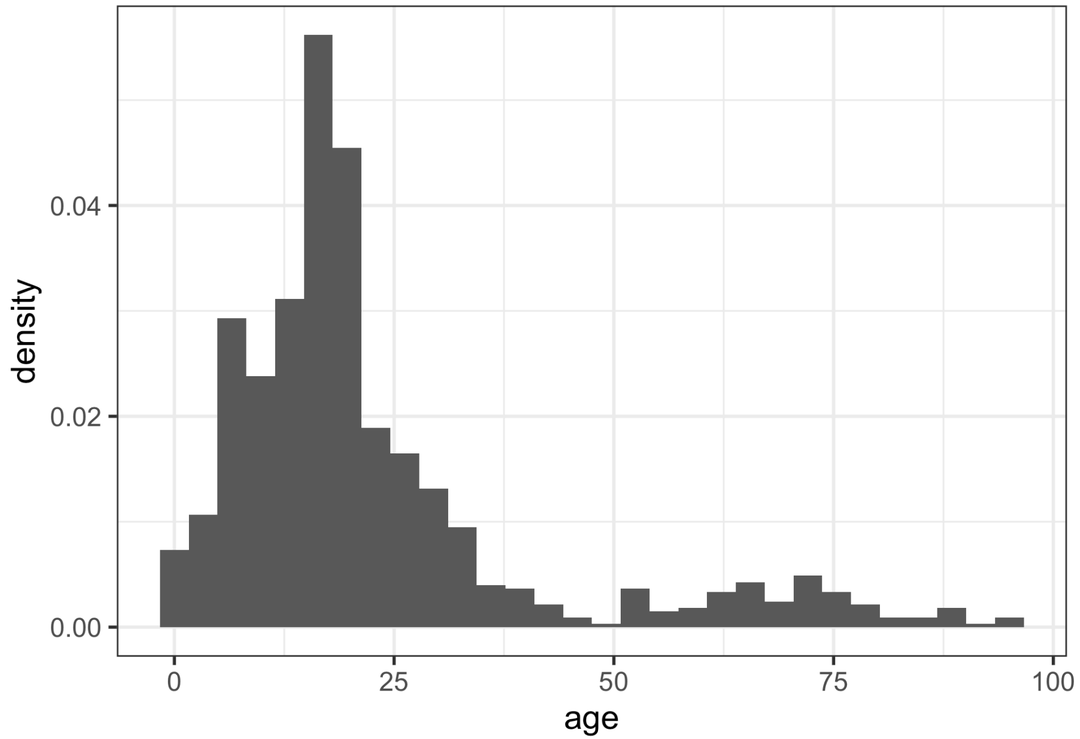



```
1 ggplot(data = dds.discr1,
2        aes(x = expenditures)) +
3   geom_histogram()
```

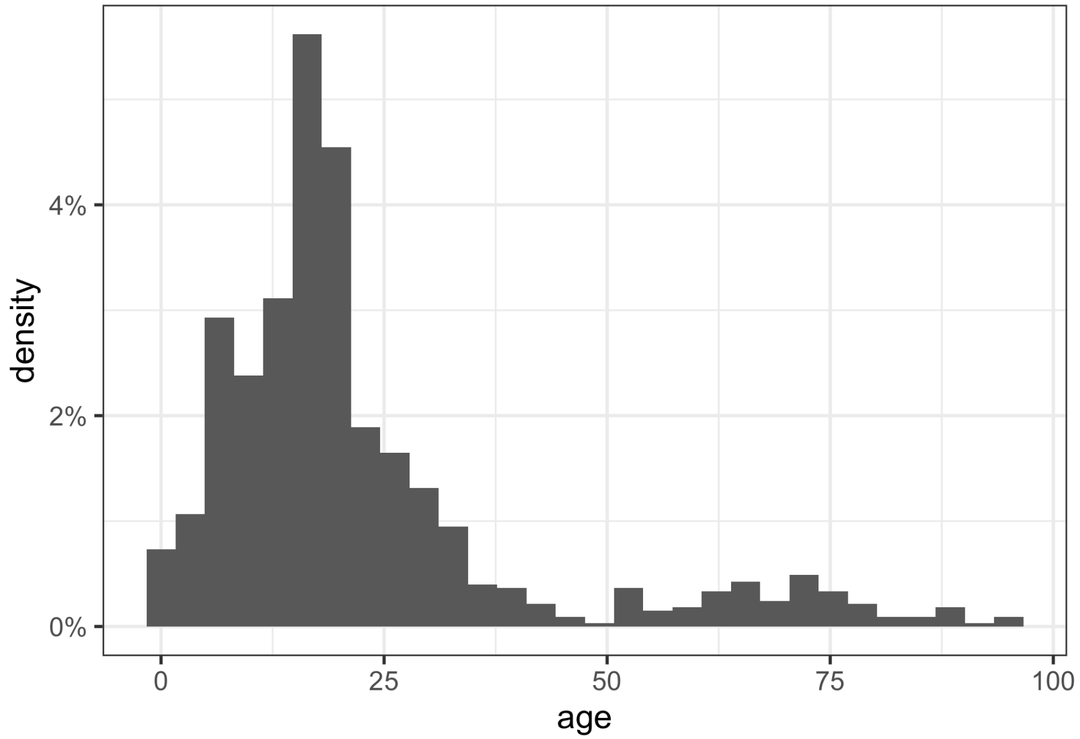


Histograms showing proportions

```
ggplot(data = dds.discr1,
       aes(x = age)) +
  geom_histogram(
    aes(y = stat(density)))
```

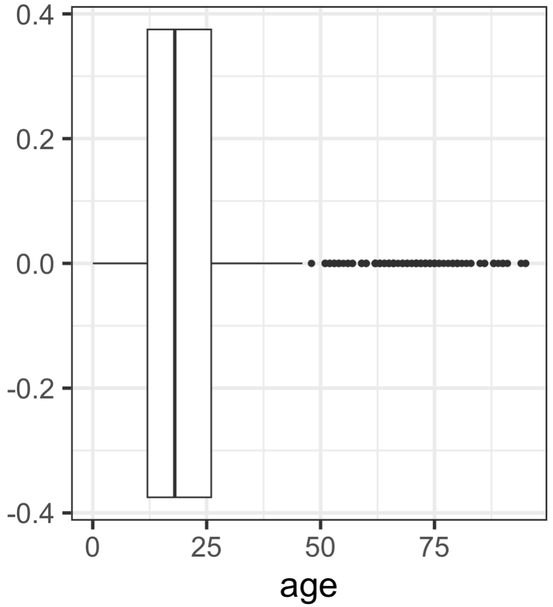



```
ggplot(data = dds.discr1,
       aes(x = age)) +
  geom_histogram(
    aes(y = stat(density))) +
  scale_y_continuous(labels =
      scales::percent_format())
```

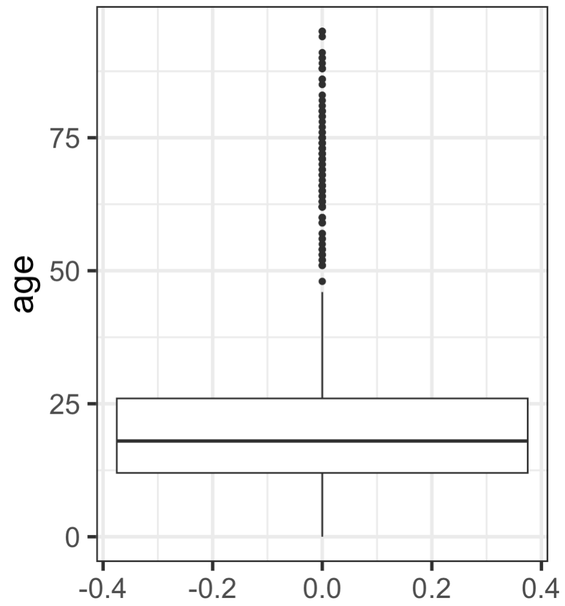


Boxplots

```
ggplot(data = dds.discr1,
       aes(x = age)) +
  geom_boxplot()
```




```
ggplot(data = dds.discr1,
       aes(y = age)) +
  geom_boxplot()
```

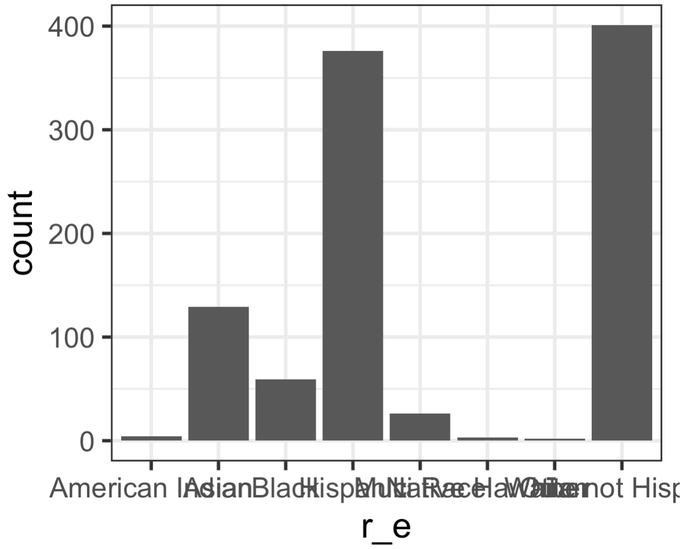


Categorical data

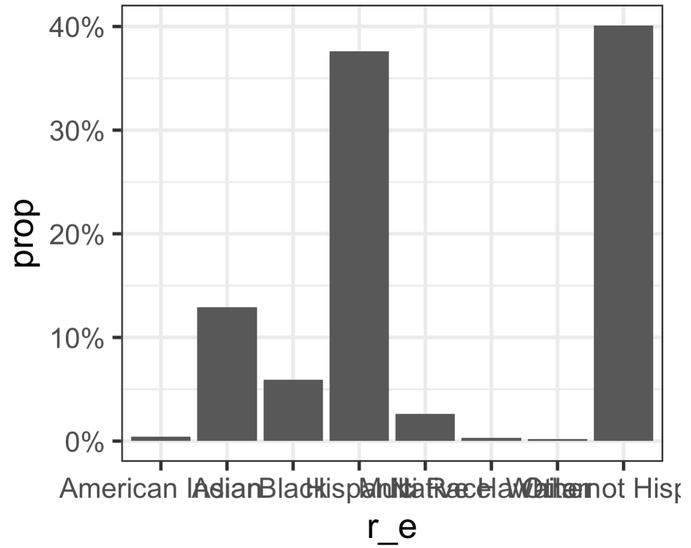
Barplots

Counts (below) vs. percentages (right)

```
ggplot(data = dds.discr1,
       aes(x = r_e)) +
  geom_bar()
```



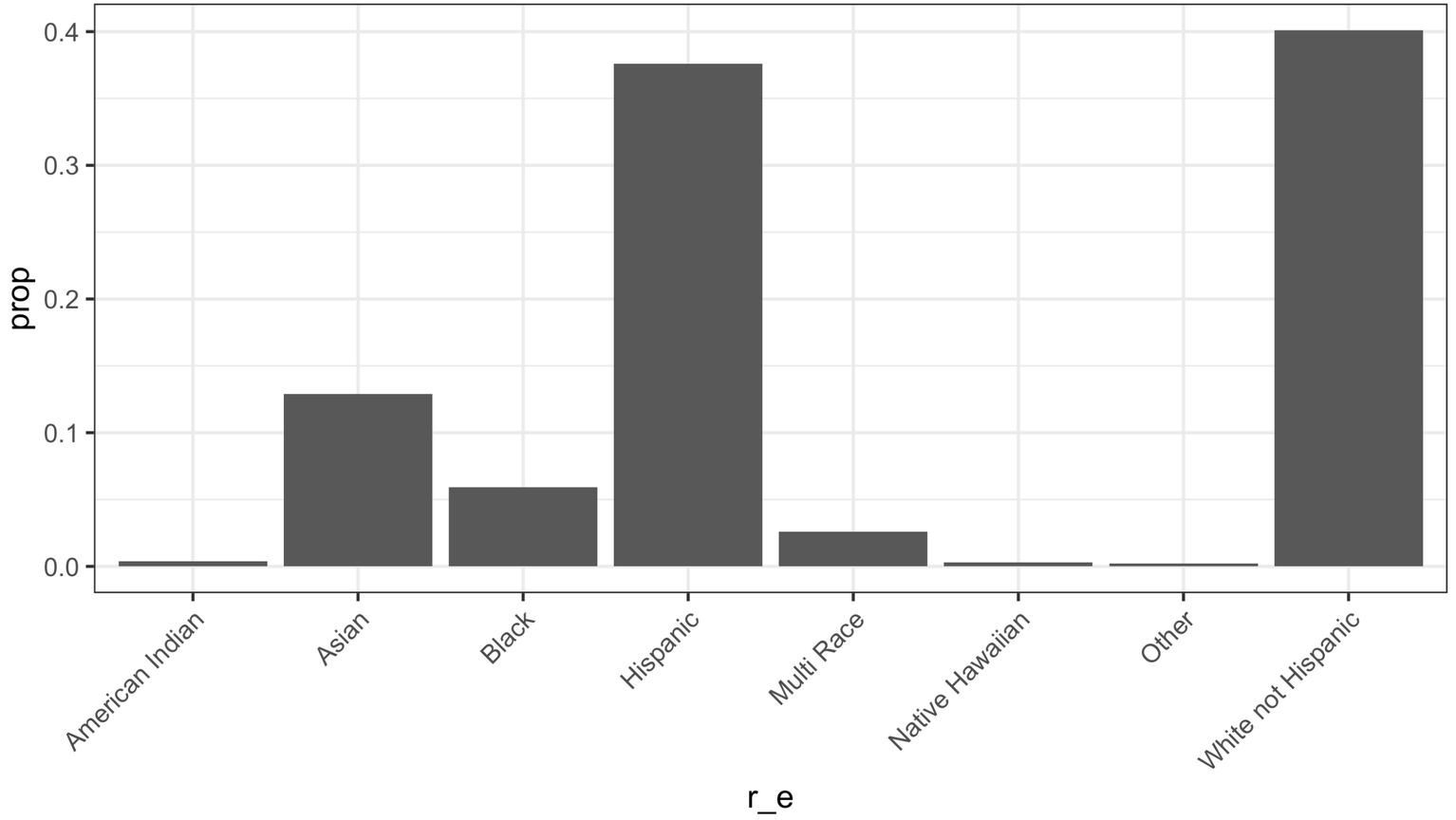

```
ggplot(data = dds.discr1,
       aes(x = r_e)) +
  geom_bar(aes(y = stat(prop),
                group = 1)) +
  scale_y_continuous(labels =
    scales::percent_format())
```



Adding more to plots!

Tilt text so we can read it!

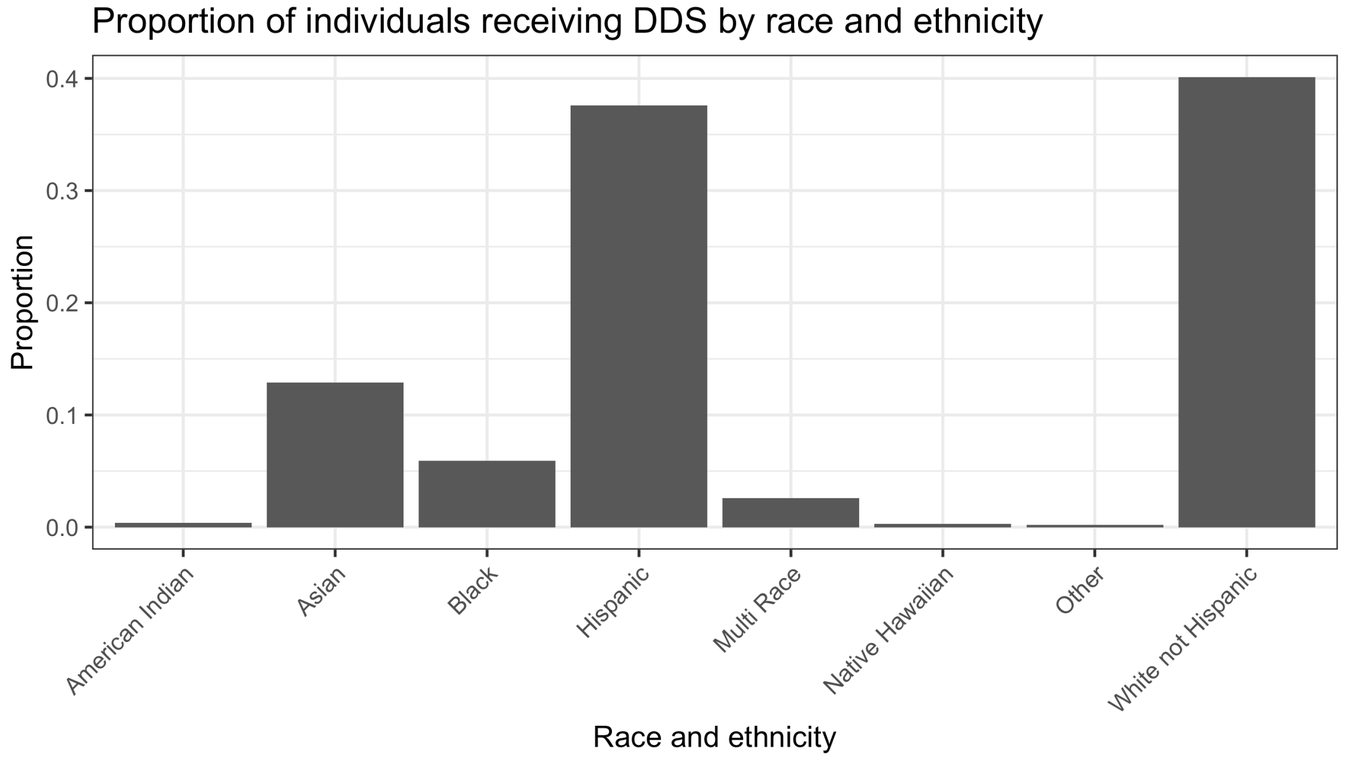
```
ggplot(data = dds.discr1, aes(x = r_e)) +
  geom_bar(aes(y = stat(prop), group = 1)) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Adding more to plots!

We can change labels!

```
ggplot(data = dds.discr1, aes(x = r_e)) +
  geom_bar(aes(y = stat(prop), group = 1)) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(x = "Race and ethnicity", y = "Proportion",
       title = "Proportion of individuals receiving DDS by race and ethnicity")
```



Adding more to plots!

Increase text size so we can read it!

```
ggplot(data = dds.discr1, aes(x = r_e)) +
  geom_bar(aes(y = stat(prop), group = 1)) +
  theme(axis.text.x = element_text(angle = 60, hjust = 1, size = 35),
        axis.title = element_text(size = 35))
```