DAY 3: DATA VISUALIZATION

BSTA 511/611, OHSU

Meike Niederhausen, PhD

2023-10-04

GOALS FOR TODAY

- **Exploratory Data Analysis (EDA)** (Sections 1.4, 1.5, 1.6, 1.7.1)
  - Data visualization with ggplot
    - numerical & categorical variables, and relationships between variables
  - Summarizing numerical data
  - Frequency (two-way) tables
- Some **data wrangling** techniques along the way

Artwork by @allison_horst

INTERNATIONAL DAY OF WOMEN IN STATISTICS AND DATA SCIENCE

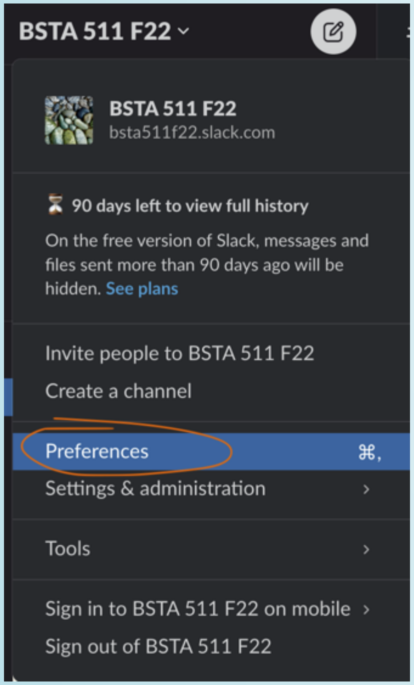
Tuesday, October 10, 2022

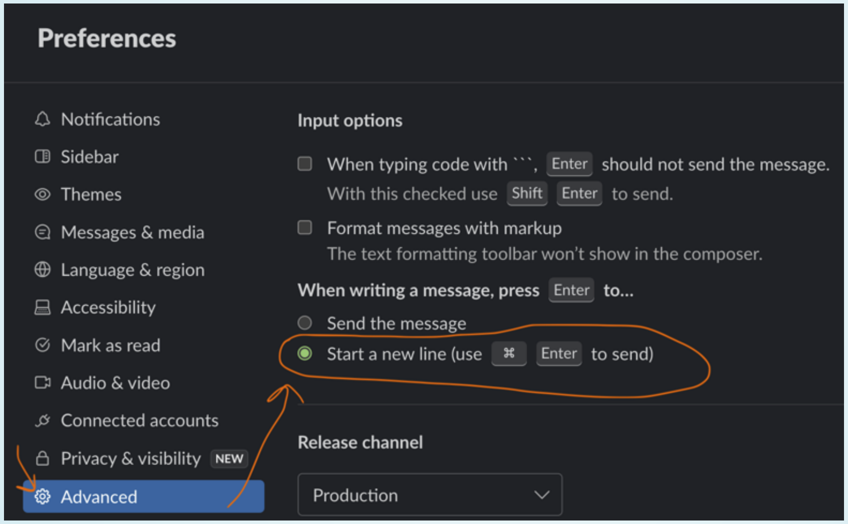
12 am - 11:59 pm UTC (5pm 10/9 to 4:59 pm 10/10 here)

International Day of Women in Statistics and Data Science

MIMI'S TIP OF THE DAY: SENDING MESSAGES IN SLACK

Are you frustrated that Slack sends a message when you press Enter? You can change that!



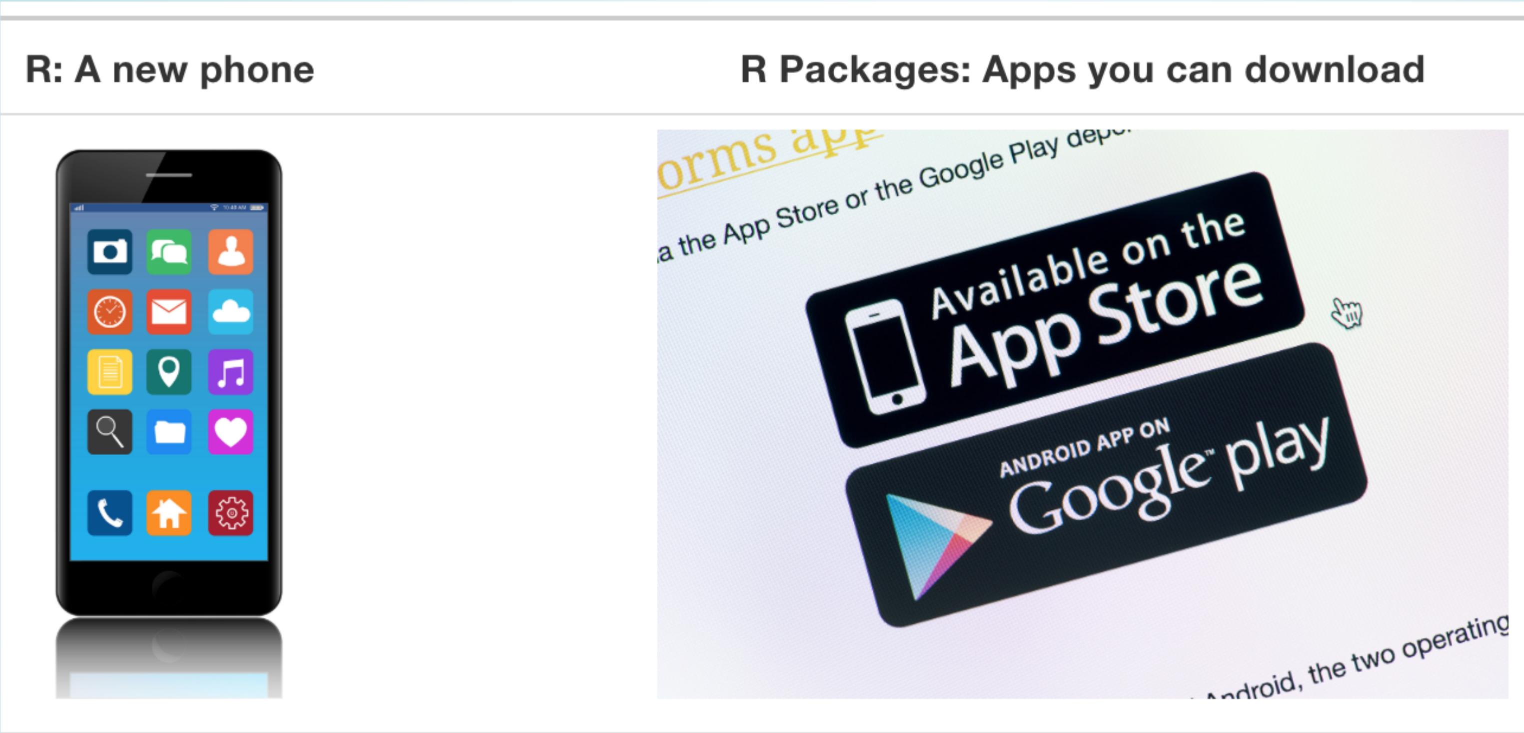


RECAP OF LAST TIME

- (1.3) **Data collection principles**
  - Population vs. sample
  - Sampling methods
  - Experiments vs. Observational studies
- (1.2) **Intro to Data**
  - Data types
    - Numerical: discrete (integer in R), continuous (double or numeric in R)
    - Categorical: ordinal, nominal
      - character or factor in R
  - How are data stored in R? data frames, tibbles
  - Working with data in R: `dim()`, `nrow()`, `ncol()`, `names()`, `str()`, `summary()`, `head()`, `tail()`, `$`
- (1.4) **Summarizing numerical data**
  - `mean()`, `median()`, `sd()`, `quantile()`

R PACKAGES

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



ModernDive Figure 1.4

FROM LAST TIME: INSTALL THE PACAKGES 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`**

DIRECTIONS FOR INSTALLING PACKAGE `oibiostat`

- The textbook's datasets are in the R package `oibiostat`
- Explanation of code below
  - Installation of `oibiostat` package requires first installing `devtools` package
  - The code `devtools::install_github()` tells R to use the command `install_github()` from the `devtools` package without loading the entire package and all of its commands (which `library(devtools)` would do).

```
install.packages("devtools")
devtools::install_github("OI-Biostat/oi_biostat_data", force = TRUE)
```

- After running the code above, put `#` in front of the commands so that RStudio doesn't evaluate them when rendering.
- Now load the `oibiostat` package
  - the code below needs to be run *every time* you restart R or render a Qmd file

```
library(oibiostat)
```

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

```r
# run these every time you open Rstudio
library(tidyverse)
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

CASE STUDY: DISCRIMINATION IN DEVELOPMENTAL DISABILITY SUPPORT (SECTION 1.7.1)

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, gender, race/ethnicity, and annual DDS financial support per consumer**
- **Previous research**
  - Researchers examined expenditures on consumers by ethnicity
  - Found that the mean annual expenditure on Hispanics was less than that on White non-Hispanics.
- Result: an allegation of ethnic discrimination was brought against the California DDS.
- **Question: Are the data sufficient evidence of ethnic discrimination?**
- 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

```
library(oibiostat)
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.

GETTING TO KNOW THE DATASET

```
dim(dds.discr)
```

```
[1] 1000    6
```

```
names(dds.discr)
```

```
[1] "id"              "age.cohort"    "age"             "gender"        "expenditures"
[6] "ethnicity"
```

```
length(unique(dds.discr$id)) # How many unique id's are there?
```

```
[1] 1000
```

str() STRUCTURE

- We previously used the base R structure command `str()` to get information about variable types in a dataset.
- Note this dataset is a `tibble` instead of a `data.frame`

```
str(dds.discr)        # base R
```

```
tibble [1,000 × 6] (S3: tbl_df/tbl/data.frame)
 $ id          : int [1:1000] 10210 10409 10486 10538 10568 10690 10711 10778
10820 10823 ...
 $ age.cohort  : Factor w/ 6 levels "0-5","6-12","13-17",..: 3 5 1 4 3 3 3 3 3 3
...
 $ age         : int [1:1000] 17 37 3 19 13 15 13 17 14 13 ...
 $ gender      : Factor w/ 2 levels "Female","Male": 1 2 2 1 2 1 1 2 1 2 ...
 $ expenditures: int [1:1000] 2113 41924 1454 6400 4412 4566 3915 3873 5021 2887
...
 $ ethnicity   : Factor w/ 8 levels "American Indian",..: 8 8 4 4 8 4 8 3 8 4 ...
 - attr(*, "spec")=
  .. cols(
  ..   ID = col_integer(),
  ..   `Age Cohort` = col_character(),
  ..   Age = col_integer(),
```

glimpse()

New: `glimpse()`

- 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()`

```r
library(tidyverse)
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…
```

summary()

- We previously used the base R structure command `summary()` to get summary information about variables

```
summary(dds.discr)        # base R
```

```
       id          age.cohort       age            gender      expenditures  
 Min.   :10210   0-5   : 82   Min.   : 0.0   Female:503   Min.   :  222  
 1st Qu.:31809   6-12  :175   1st Qu.:12.0   Male  :497   1st Qu.: 2899  
 Median :55384   13-17:212   Median :18.0                Median : 7026  
 Mean   :54663   18-21:199   Mean   :22.8                Mean   :18066  
 3rd Qu.:76135   22-50:226   3rd Qu.:26.0                3rd Qu.:37713  
 Max.   :99898   51+   :106   Max.   :95.0                Max.   :75098  
                                                                         
              ethnicity  
 White not Hispanic:401  
 Hispanic          :376  
 Asian             :129  
 Black             : 59  
 Multi Race        : 26  
 American Indian   :  4  
```

`tbl_summary()`: SUMMARY TABLE

- **New:** Use `tbl_summary()` from the `gtsummary` package to get summary information

```
# library(gtsummary)
tbl_summary(dds.discr)
```

Characteristic	N = 1,000[1]
id	55,385 (31,809, 76,135)
age.cohort	
0-5	82 (8.2%)
6-12	175 (18%)
13-17	212 (21%)
18-21	199 (20%)
22-50	226 (23%)
51+	106 (11%)
age	18 (12, 26)
gender	
Female	503 (50%)
Male	497 (50%)
expenditures	7,026 (2,899, 37,713)
ethnicity	
American Indian	4 (0.4%)
Asian	129 (13%)
Black	59 (5.9%)
Hispanic	376 (38%)
Multi Race	26 (2.6%)
Native Hawaiian	3 (0.3%)
Other	2 (0.2%)
White not Hispanic	401 (40%)

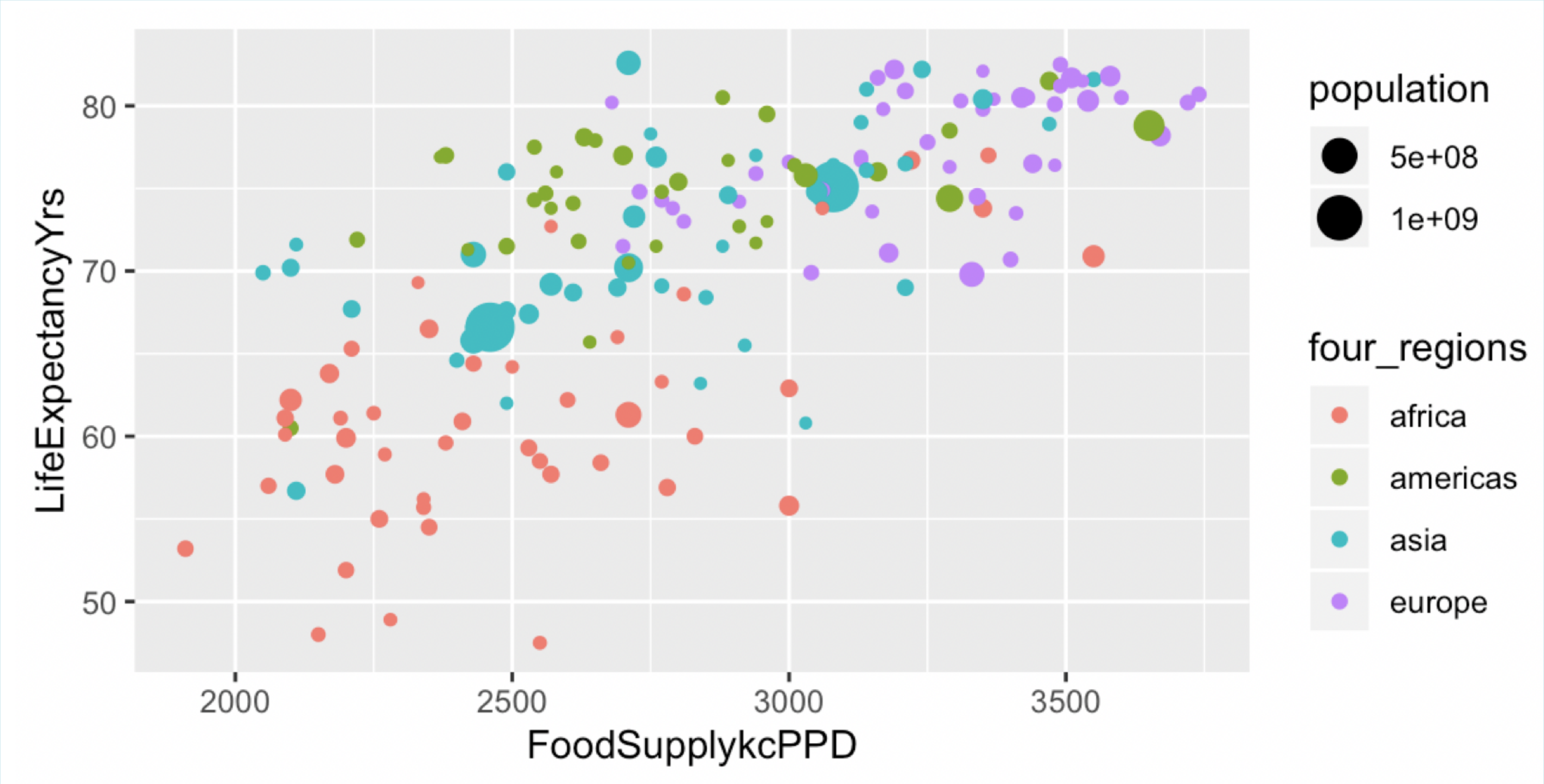
[1] Median (IQR); n (%)

VISUALIZE NUMERICAL VARIABLES WITH ggplot

ggplot

Artwork by @allison_horst

WHAT DATA (VARIABLES) ARE INCLUDED IN THE PLOT BELOW?



BASICS OF A GGPLOT

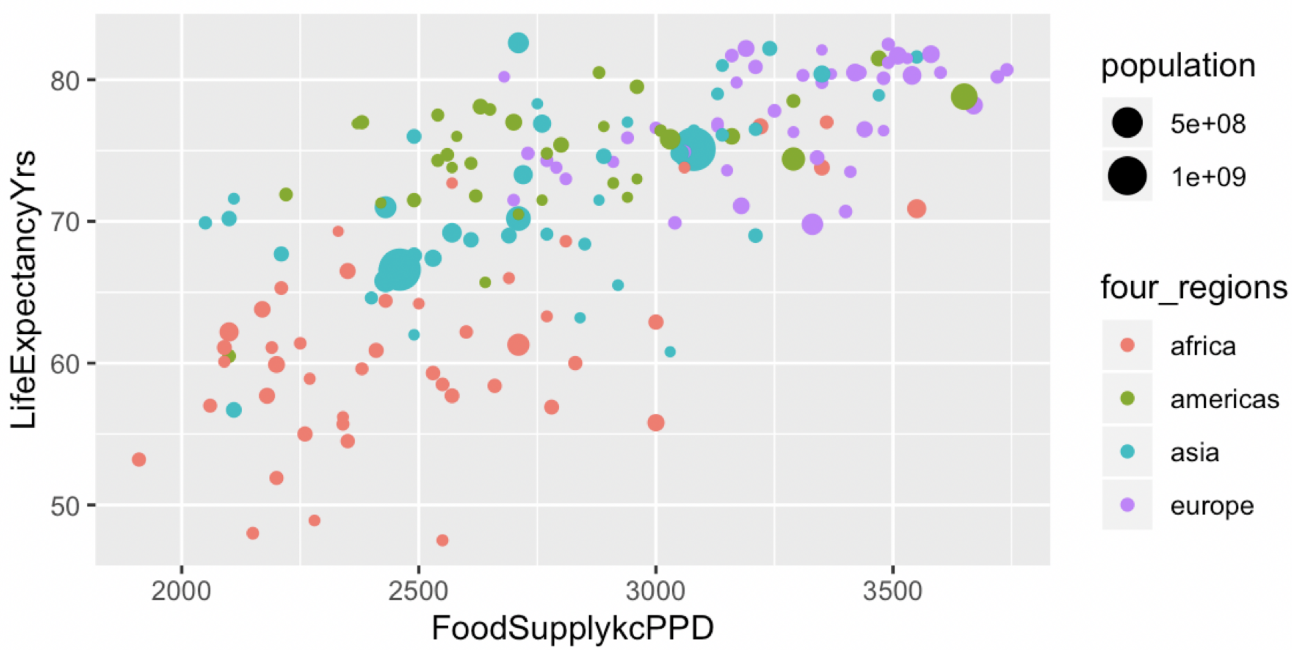
Function

Dataset

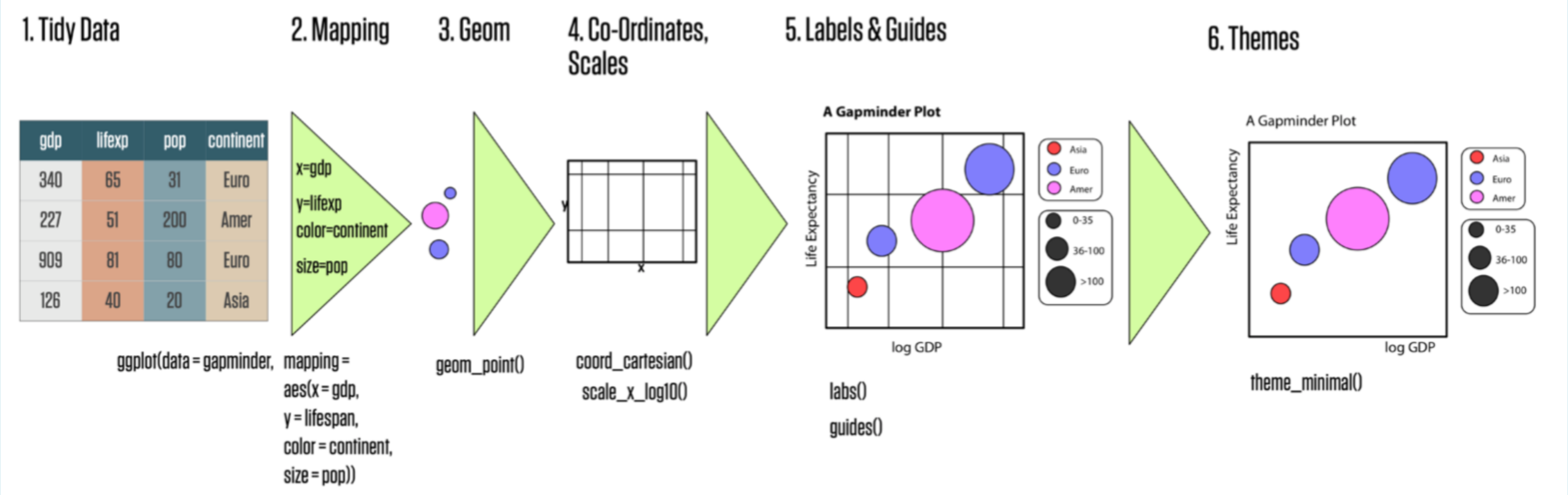
```
ggplot(data = gapminder2011,
       aes(x = FoodSupplykcPPD, y = LifeExpectancyYrs,
           color = four_regions, size = population)) +
       geom_point()
```

Which variables to plot

What kind of plot to make



GRAMMAR OF GGPLOT2



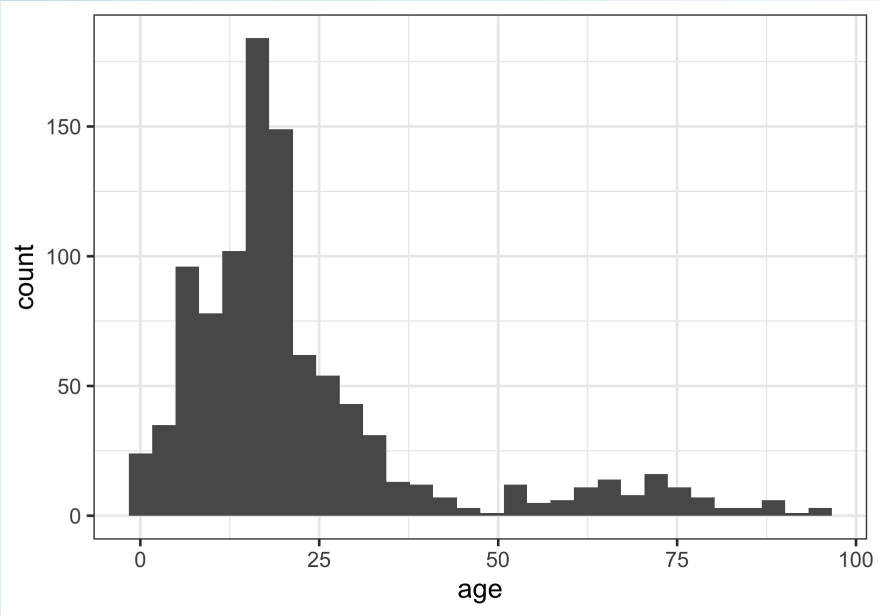
Kieran Healy

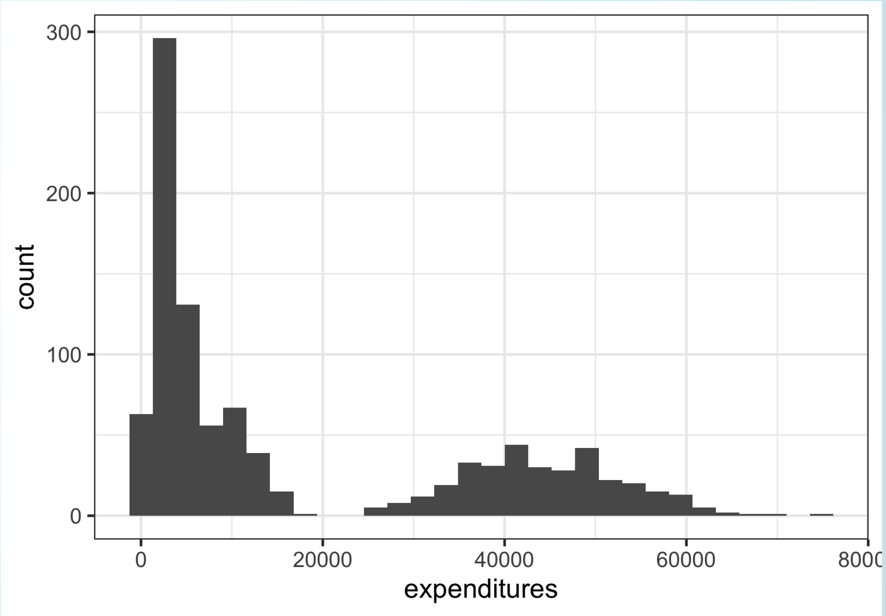
HISTOGRAMS

What is being measured on the vertical axes?

```r
ggplot(data = dds.discr,
       aes(x = age)) +
  geom_histogram()
```

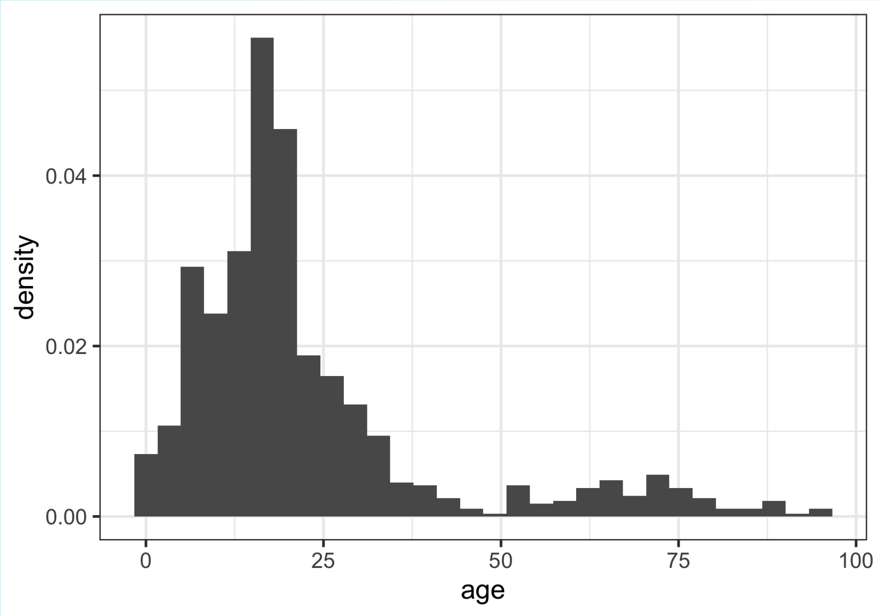
```r
ggplot(data = dds.discr,
       aes(x = expenditures)) +
  geom_histogram()
```



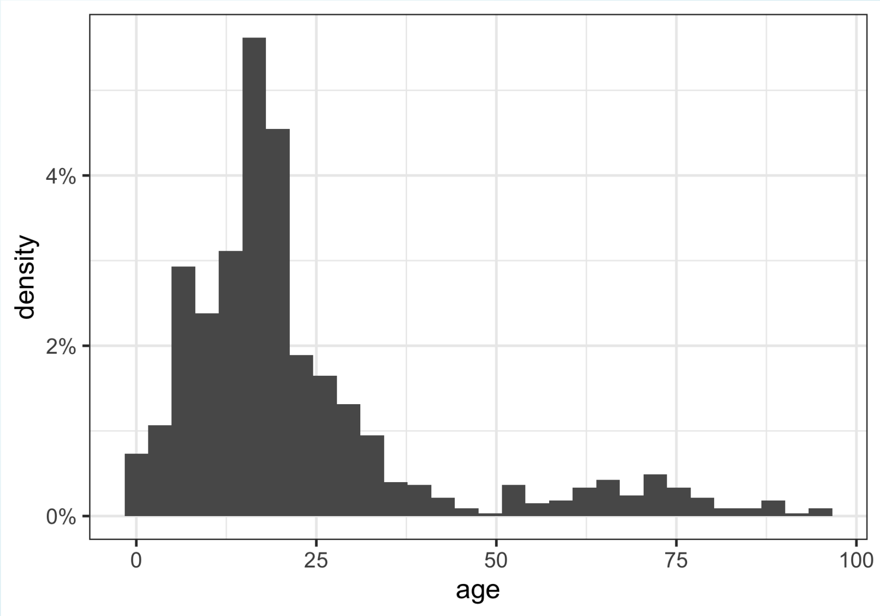


HISTOGRAMS SHOWING PROPORTIONS

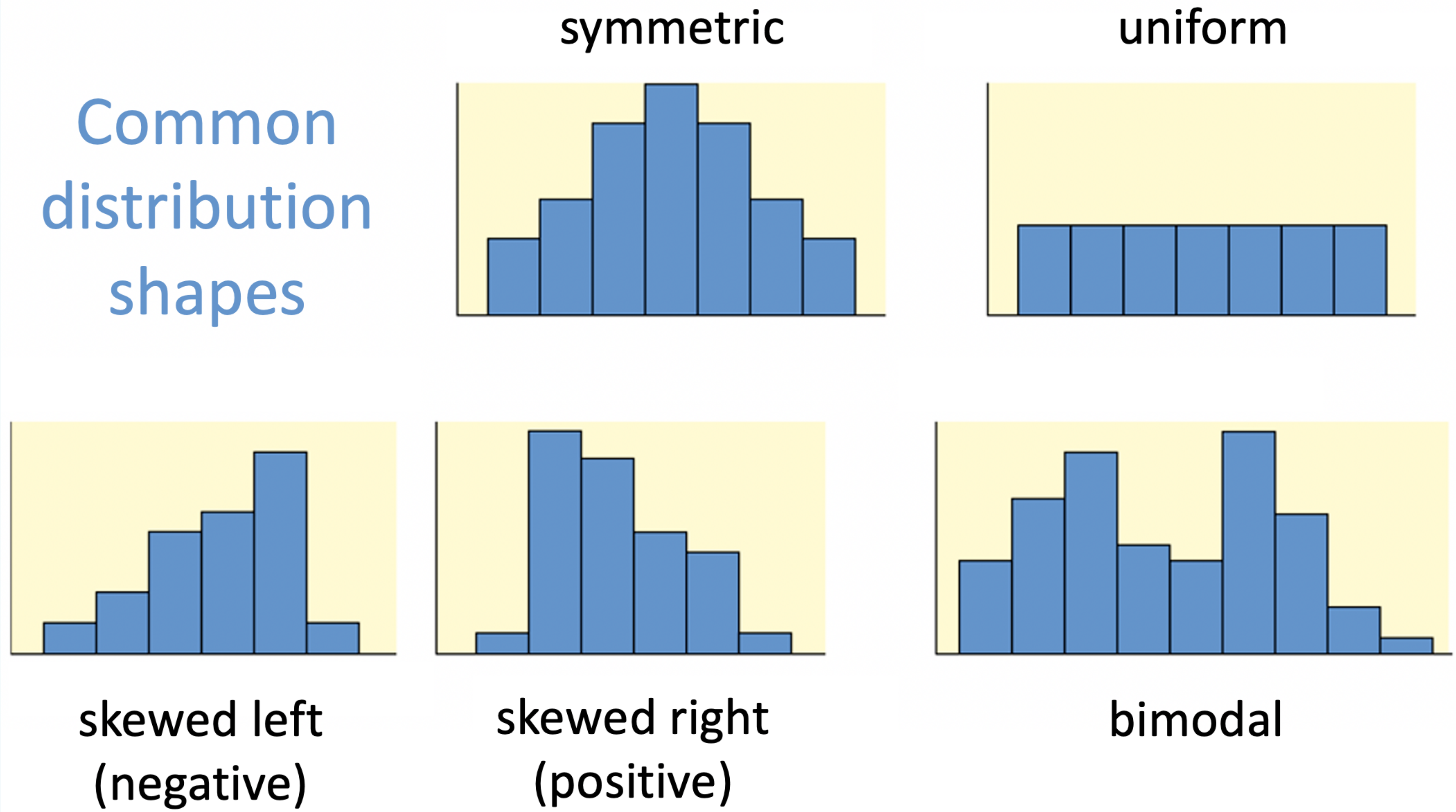
```
ggplot(data = dds.discr,
          aes(x = age)) +
  geom_histogram(
    aes(y = stat(density)))
```




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



DISTRIBUTION SHAPES

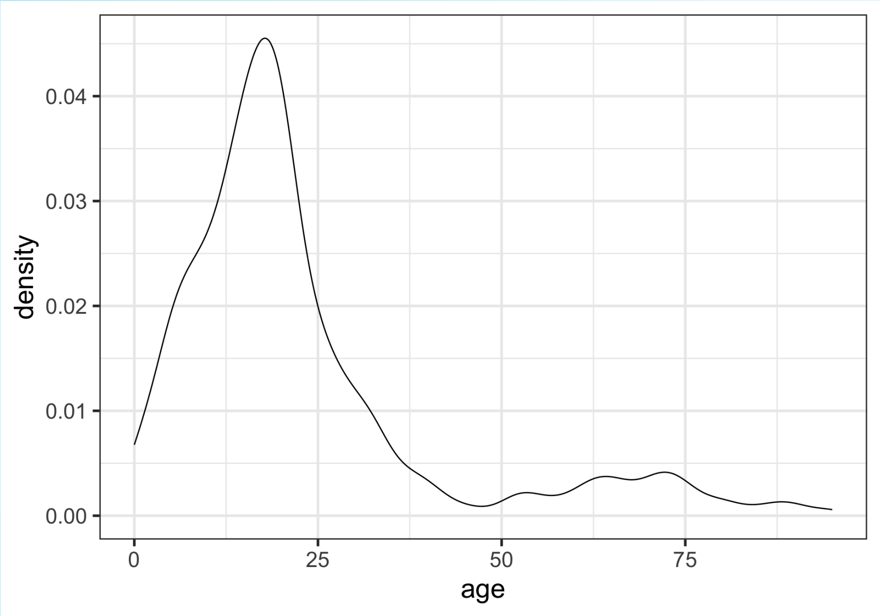


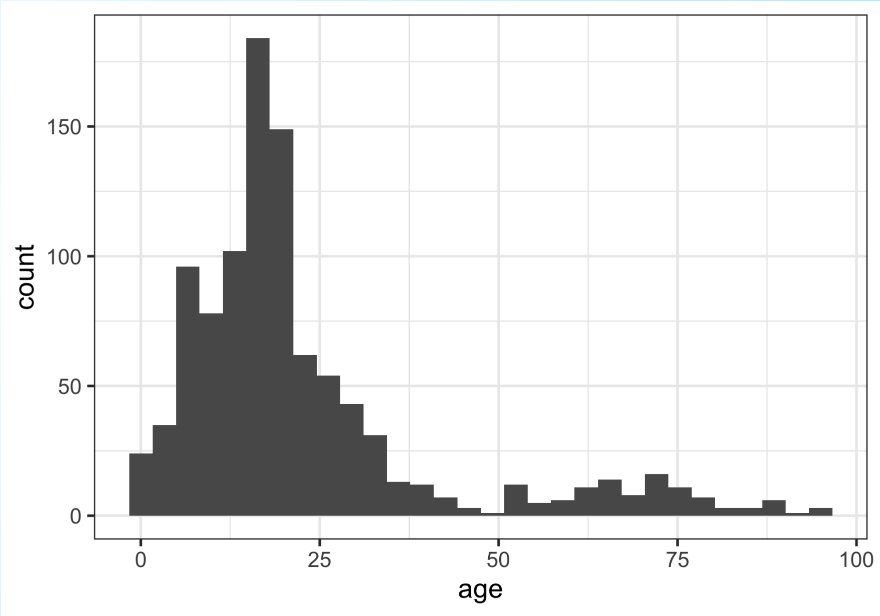
DENSITY PLOTS

What is being measured on the vertical axes?

```
ggplot(data = dds.discr,
       aes(x = age)) +
  geom_density()
```

```
ggplot(data = dds.discr,
       aes(x = age)) +
  geom_histogram()
```



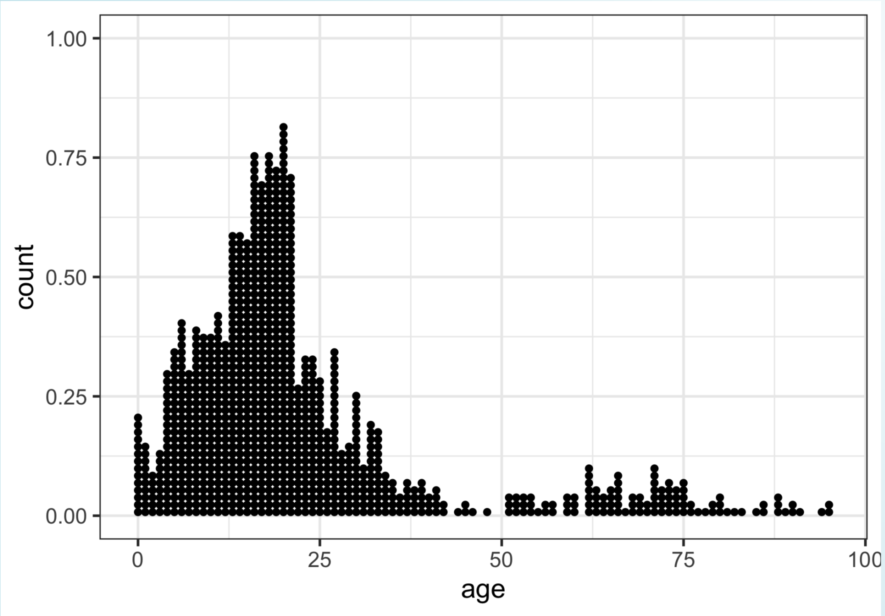


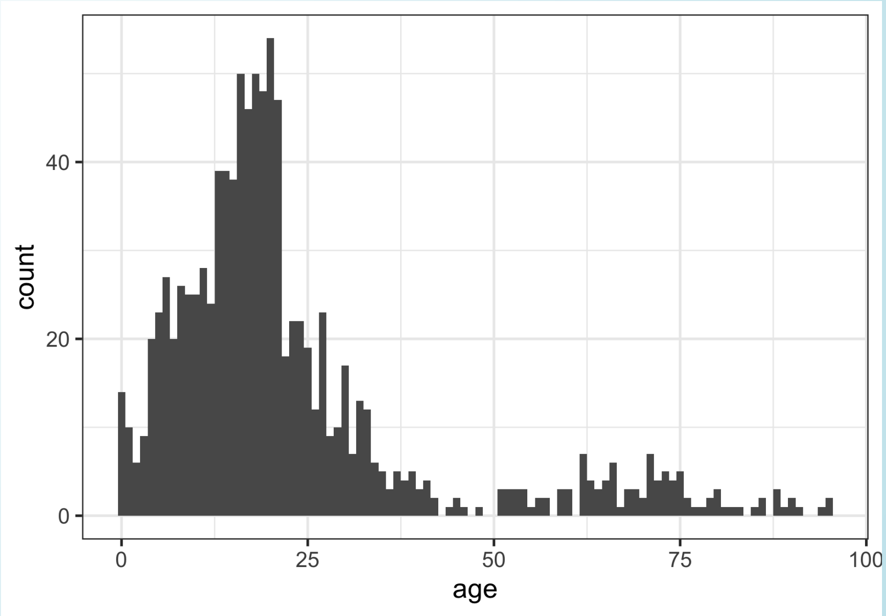
DOT PLOTS

- Better for smaller samples
- What is being measured on the vertical axes?

```
ggplot(data = dds.discr,
       aes(x = age)) +
  geom_dotplot(binwidth =1)
```

```
ggplot(data = dds.discr,
       aes(x = age)) +
  geom_histogram(binwidth =1)
```

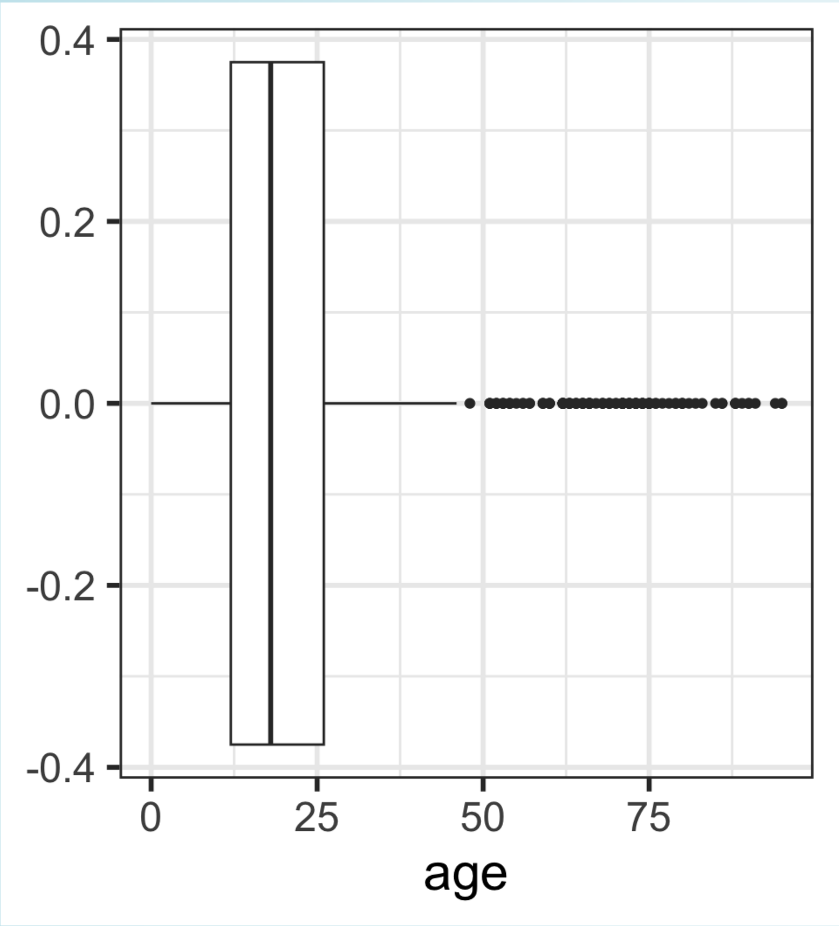


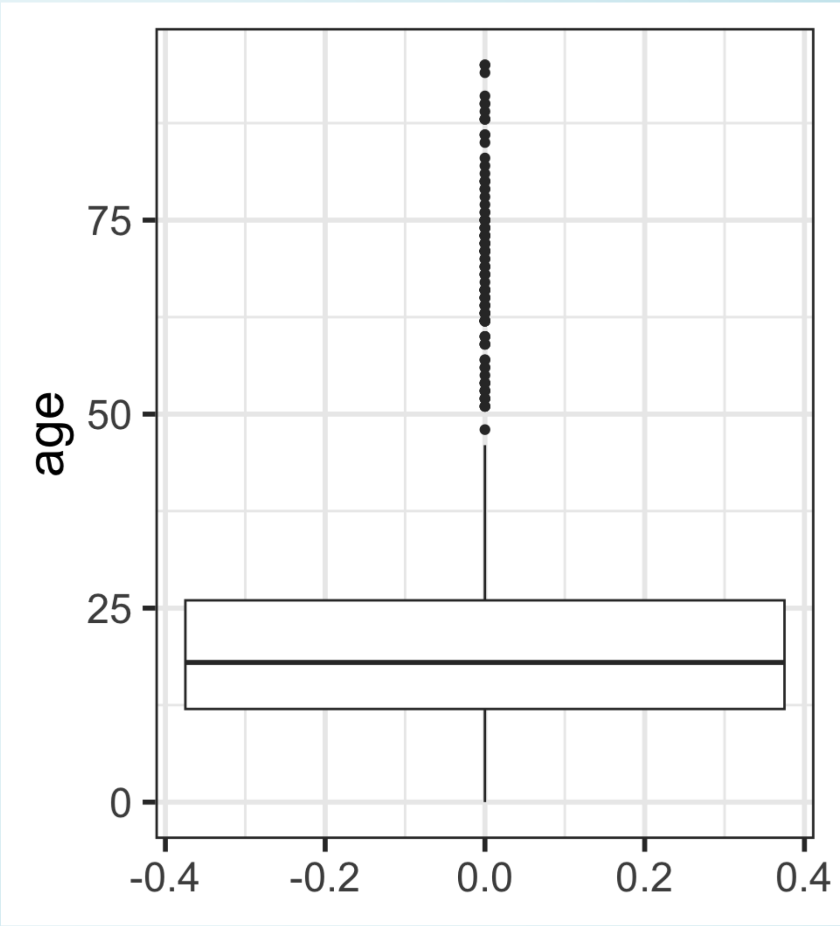


BOXPLOTS

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

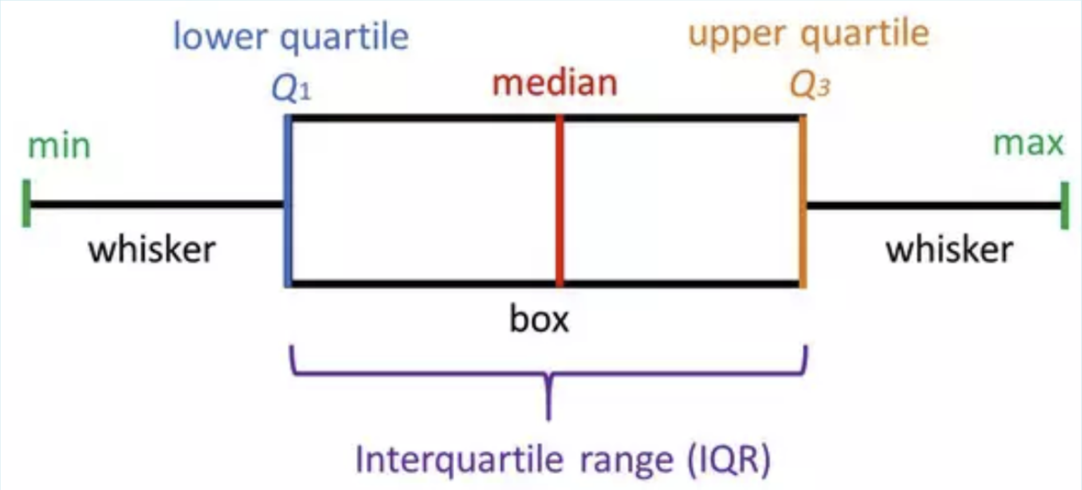
```
ggplot(data = dds.discr,
       aes(y = age)) +
  geom_boxplot()
```



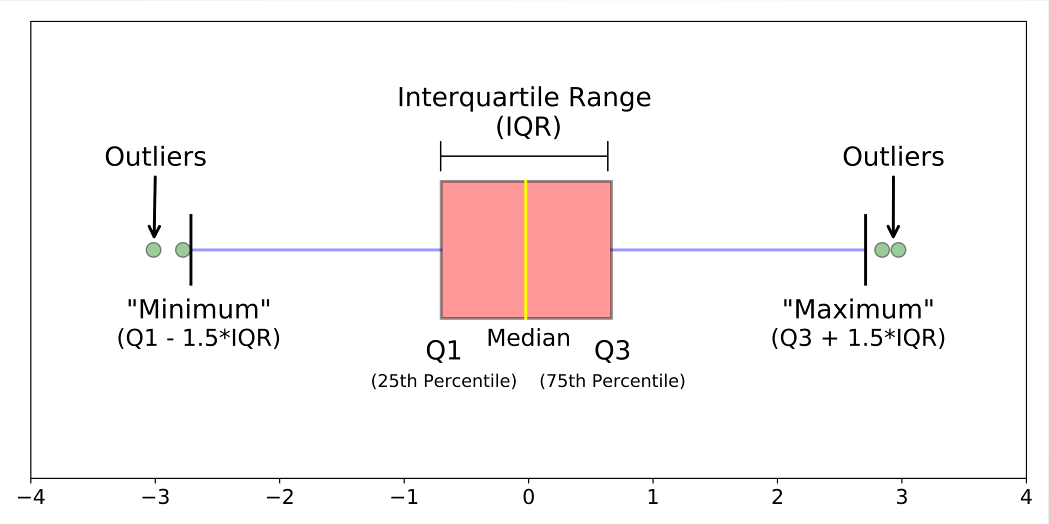


BOXPLOTS: 5 NUMBER SUMMARY VISUALIZATION

No outliers:



With outliers:



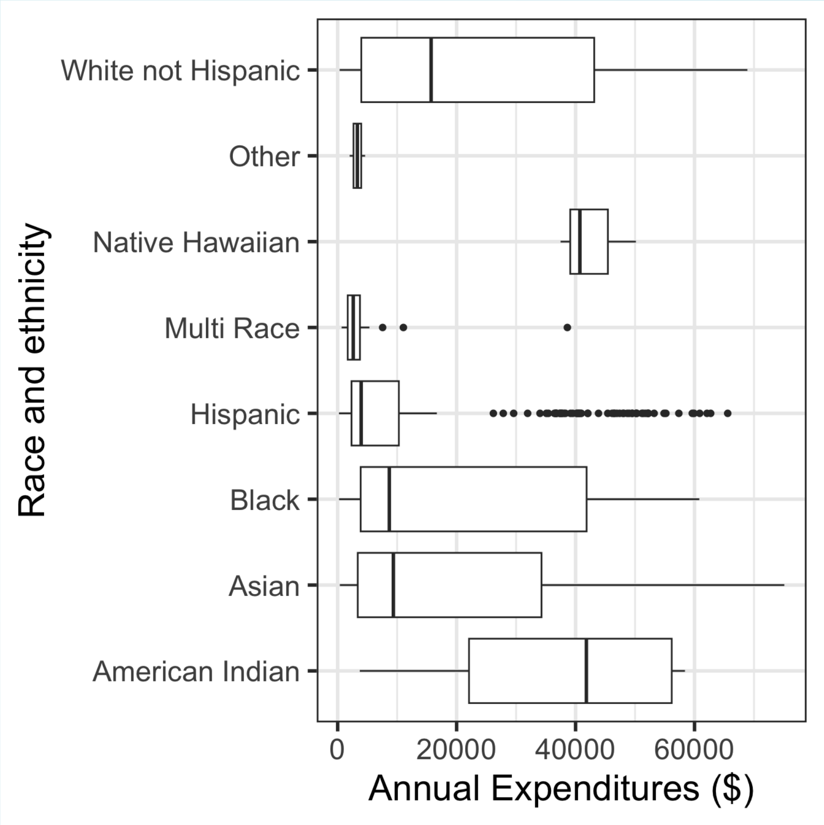
VISUALIZING RELATIONSHIPS BETWEEN NUMERICAL AND CATEGORICAL VARIABLES (1.6.3)

SIDE-BY-SIDE BOXPLOTS

```r
ggplot(data = dds.discr,
       aes(x = expenditures,
           y = ethnicity)) +
  geom_boxplot() +
  labs(x = "Annual Expenditures ($)",
       y = "Race and ethnicity")
```

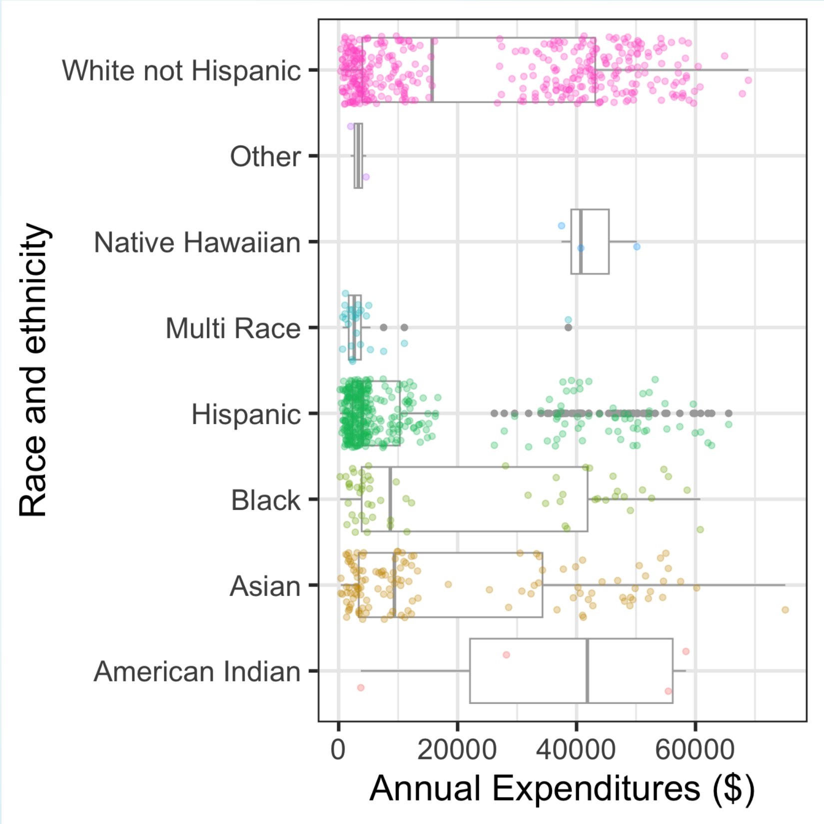
Can you determine the following using boxplots?

- distribution shape
- sample size



SIDE-BY-SIDE BOXPLOTS WITH DATA POINTS

```
ggplot(data = dds.discr,
       aes(x = expenditures,
           y = ethnicity)) +
  geom_boxplot(color="darkgrey") +
  labs(x = "Annual Expenditures ($)",
       y = "Race and ethnicity") +
  geom_jitter(
    aes(color = ethnicity),
    alpha = 0.3,
    show.legend = FALSE,
    position = position_jitter(
      height = 0.4))
```

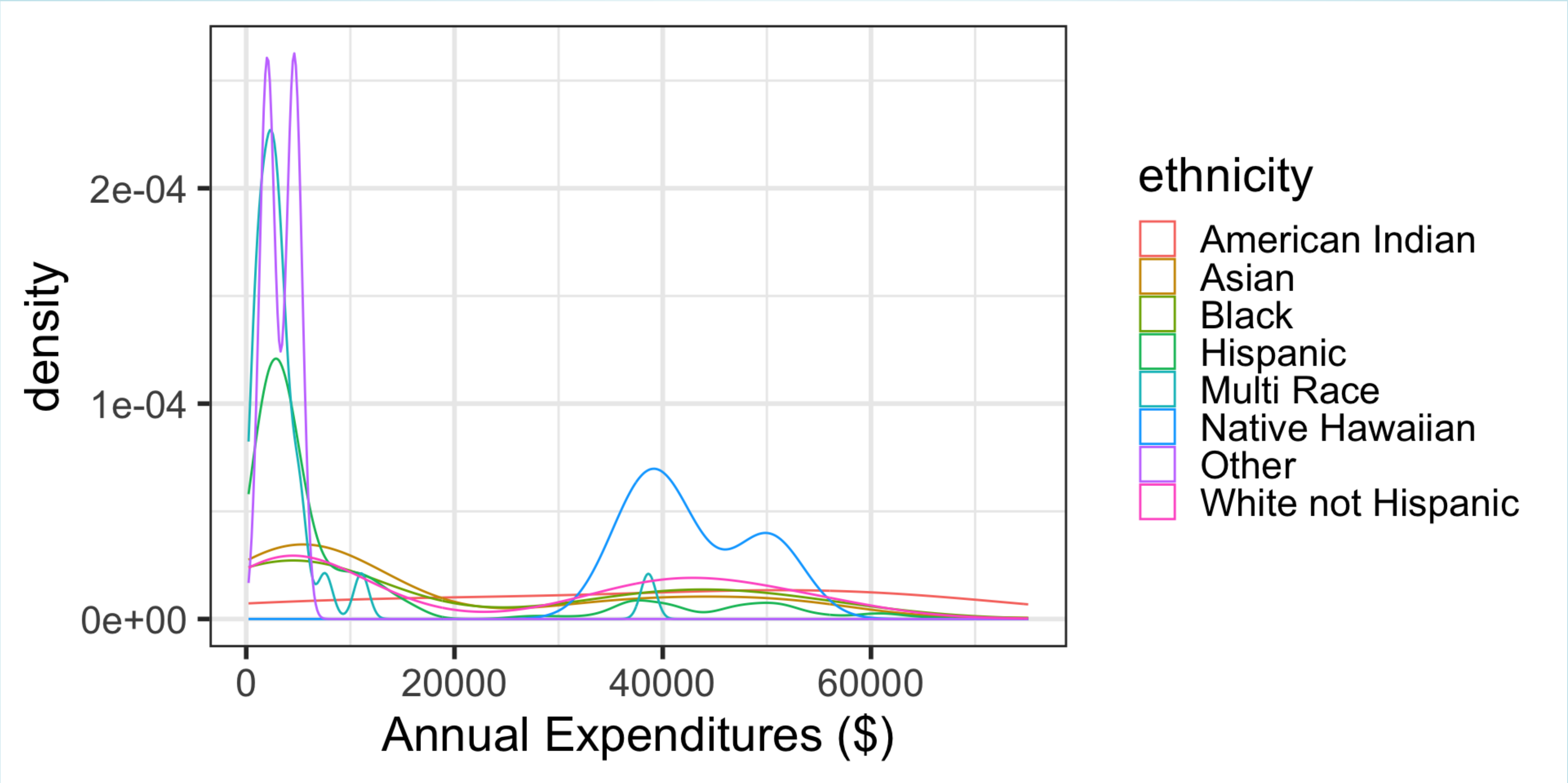


Can you determine the following using boxplots?

- distribution shape
- sample size

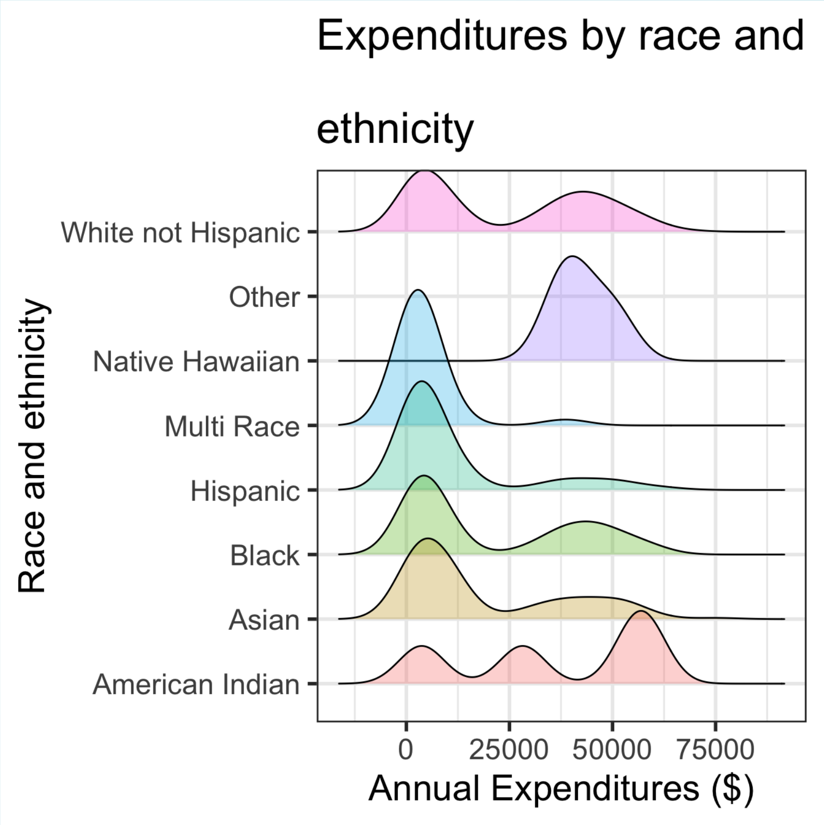
DENSITY PLOTS BY GROUP

```
ggplot(data = dds.discr,
       aes(x = expenditures,
           color = ethnicity)) +
  geom_density() +
  labs(x = "Annual Expenditures ($)")
```



RIDGELINE PLOT

```r
# library(ggridges)
ggplot(data = dds.discr,
       aes(x = expenditures,
           y = ethnicity,
           fill = ethnicity)
       ) +
  geom_density_ridges(
    alpha = 0.3,
    show.legend = FALSE) +
  labs(x = "Annual Expenditures ($)",
       y = "Race and ethnicity",
       title =
"Expenditures by race and
       \nethnicity")
```

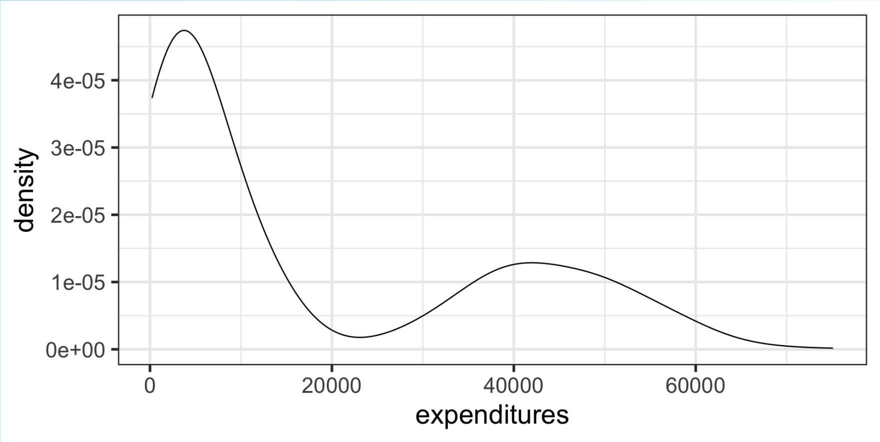


TRANSFORMING DATA (14.5)

- We sometimes apply a transformation to highly skewed data to make it more symmetric
- Log transformations are often used for skewed right data

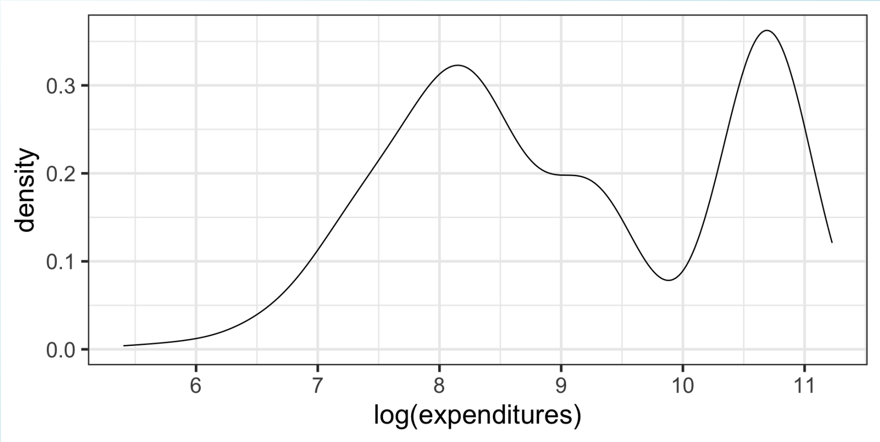
x = expenditures

```
ggplot(data = dds.discr,
       aes(x = expenditures)) +
  geom_density()
```



x = log(expenditures)

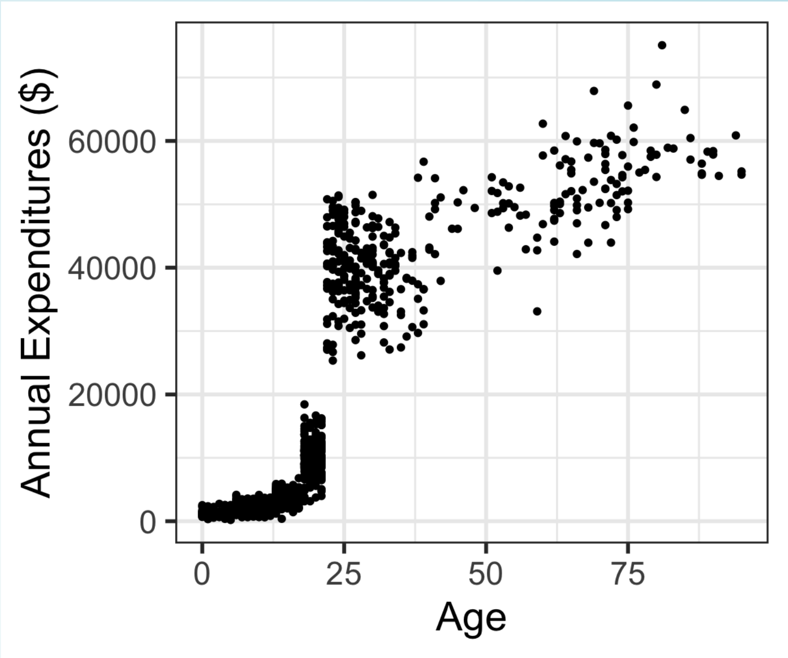
```
ggplot(data = dds.discr,
       aes(x = log(expenditures))) +
  geom_density()
```



RELATIONSHIPS BETWEEN TWO NUMERICAL VARIABLES (1.6.1)

SCATTERPLOTS

```r
ggplot(data = dds.discr,
       aes(x = age,
           y = expenditures)) +
  geom_point() +
  labs(x = "Age",
       y = "Annual Expenditures ($)")
```



Response vs. explanatory variables (Section 1.2.3)

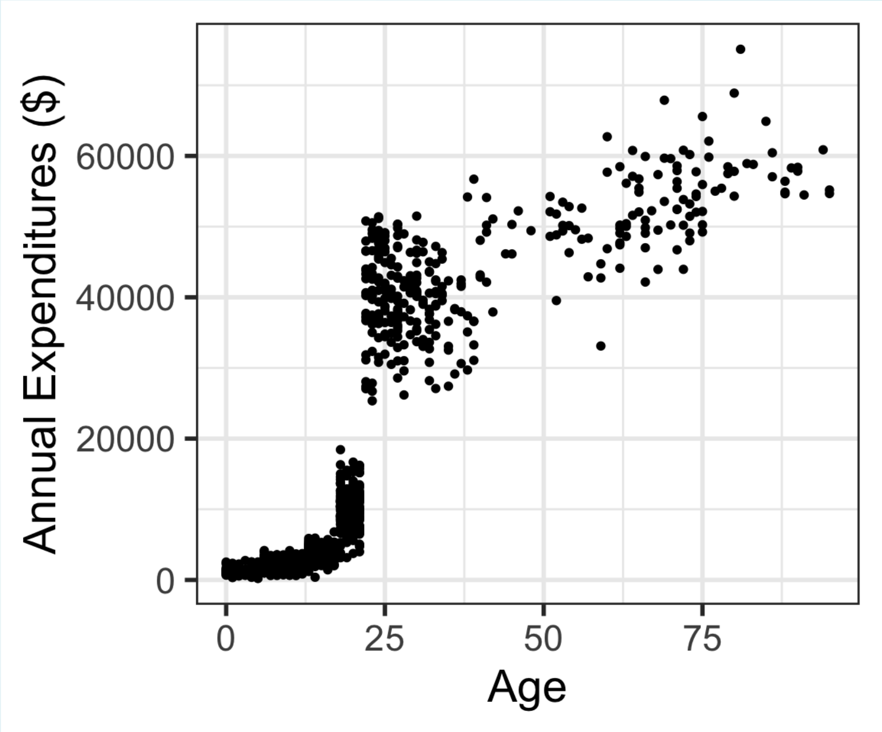
- A **response variable** measures the outcome of interest in a study
- A study will typically examine whether the values of a response variable differ as values of an **explanatory variable** change

Describe the association between the variables

DESCRIBING ASSOCIATIONS BETWEEN 2 NUMERICAL VARIABLES

Two variables x and y are

- **positively associated** if y increases as x increases.
- **negatively associated** if y decreases as x increases.
- If there is no association between the variables, then we say they are **uncorrelated** or **independent**.



- The term "association" is a very general term.
  - Can be used for numerical or categorical variables
  - Not specifically referring to linear associations

(PEARSON) CORRELATION COEFFICIENT R

- $r = -1$ indicates a **perfect negative linear relationship**: As one variable increases, the value of the other variable tends to go down, following a *straight line*.
- $r = 0$ indicates **no linear relationship**: The values of both variables go up/down independently of each other.
- $r = 1$ indicates a **perfect positive linear relationship**: As the value of one variable goes up, the value of the other variable tends to go up as well in a linear fashion.
- The closer r is to ±1, the stronger the linear association.

(PEARSON) CORRELATION COEFFICIENT (R): FORMULA

The (Peasron) correlation coefficient of variables x and y can be computed using the formula

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \Big(\frac{x_i - \bar{x}}{s_x}\Big)\Big(\frac{y_i - \bar{y}}{s_y}\Big)$$

where

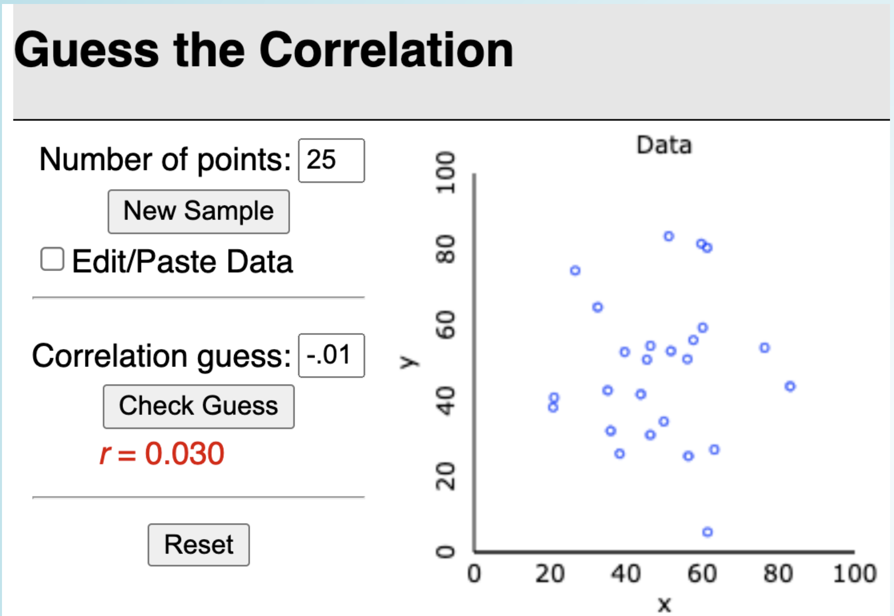
- $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ are the n paired values of the variables x and y
- s_x and s_y are the sample standard deviations of the variables x and y, respectively

```
cor(dds.discr$age, dds.discr$expenditures)
```

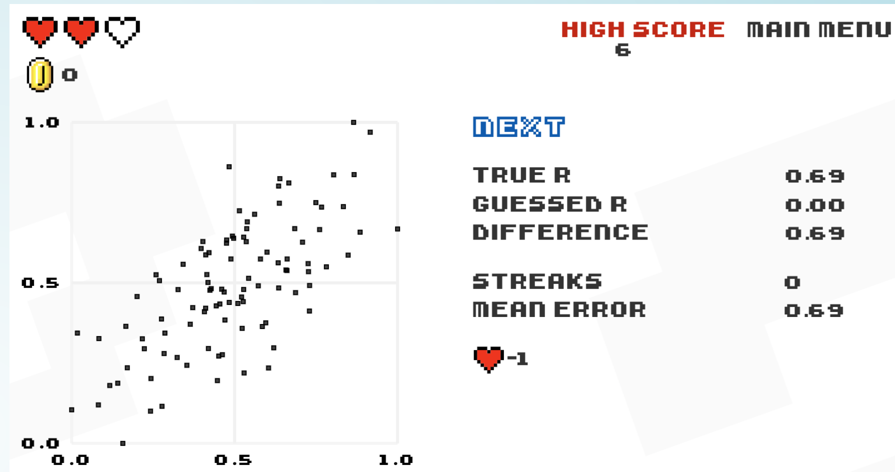
```
[1] 0.8432422
```

GUESS THE CORRELATION GAME!

Rossman & Chance's applet



Or, for the Atari-like experience



http://guessthecorrelation.com/

Tracks performance of guess vs. actual, error vs. actual, and error vs. trial

http://www.rossmanchance.com/applets/GuessCorrelation.html

SCATTERPLOTS WITH COLOR-CODED DOTS

Describe the association between the variables

```
ggplot(data = dds.discr,
       aes(x = age, y = expenditures,
           color = ethnicity)) +
  geom_point(alpha = .5) +
  labs(x = "Age", y = "Annual Expenditures ($)") +
  theme(legend.position = "bottom")
```

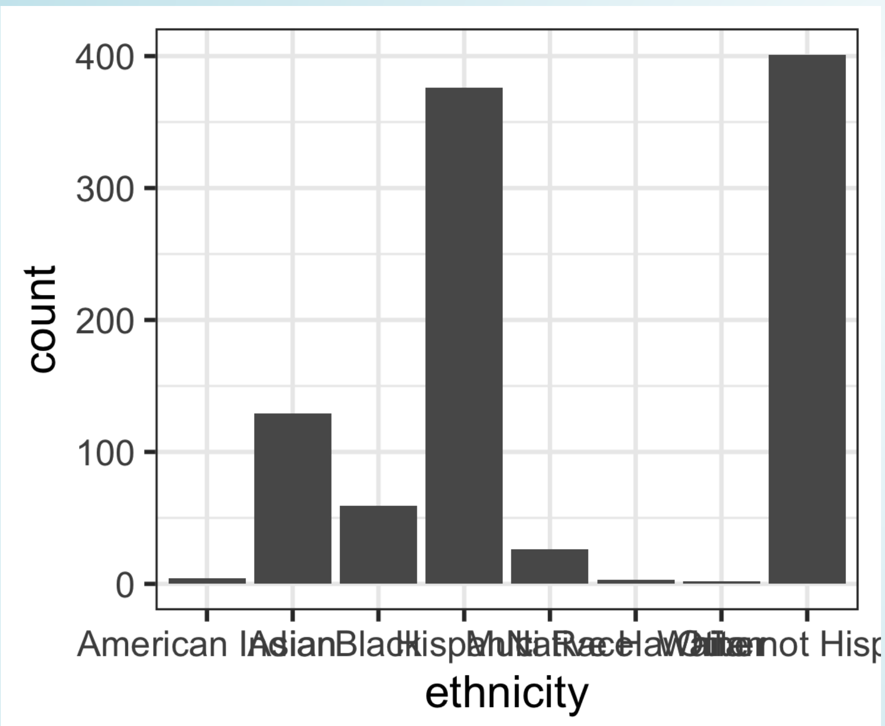
CATEGORICAL DATA (1.5) AND RELATIONSHIPS BETWEEN TWO CATEGORICAL VARIABLES (1.6.2)

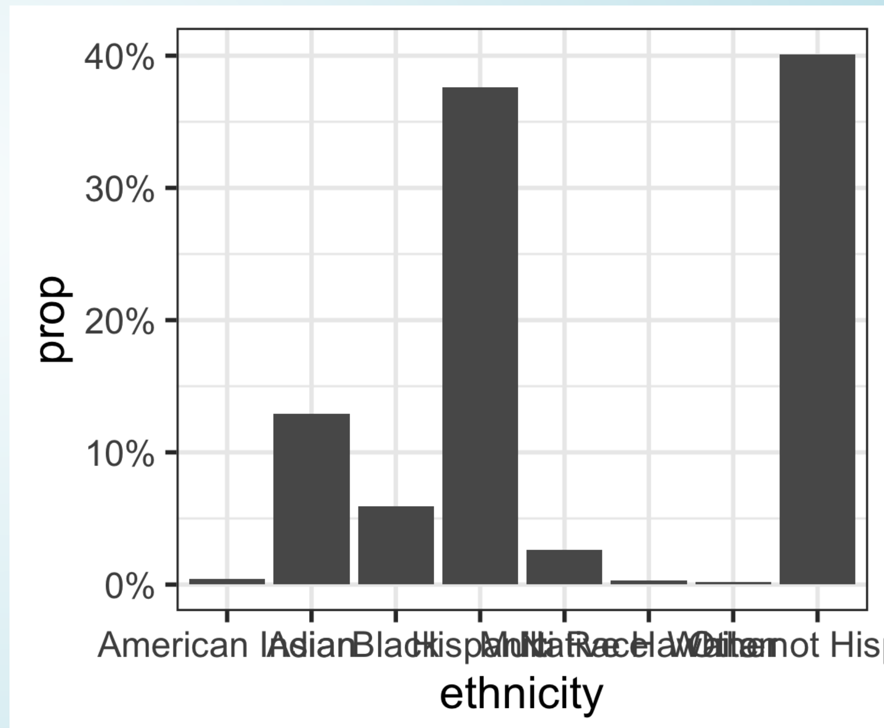
BARPLOTS

Counts (below) vs. percentages (right)

```
ggplot(data = dds.discr,
       aes(x = ethnicity)) +
  geom_bar()
```

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

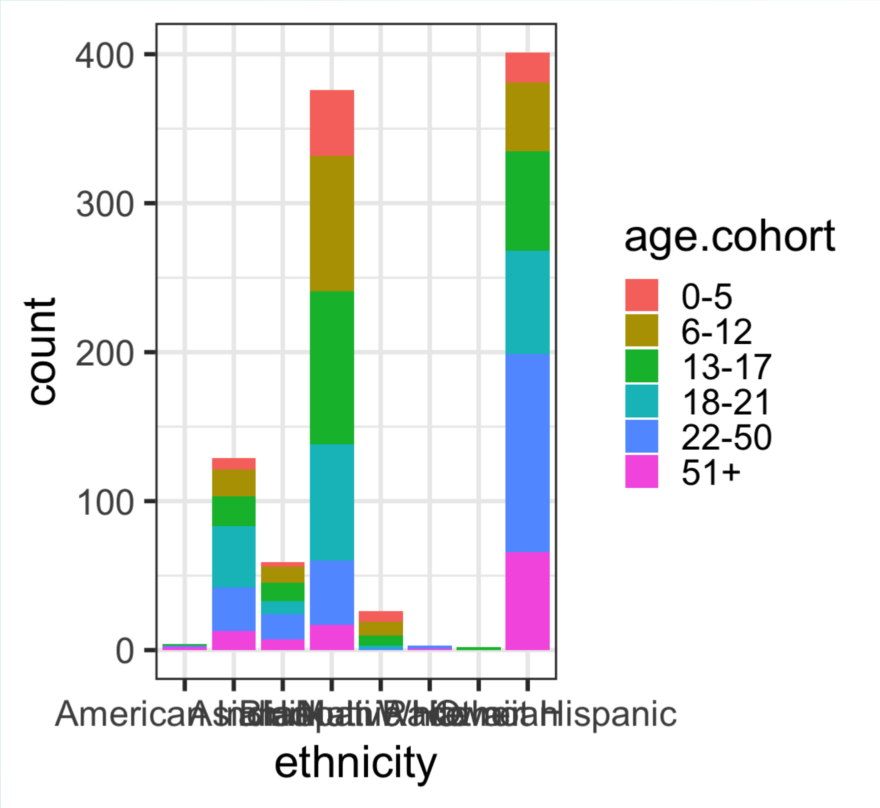


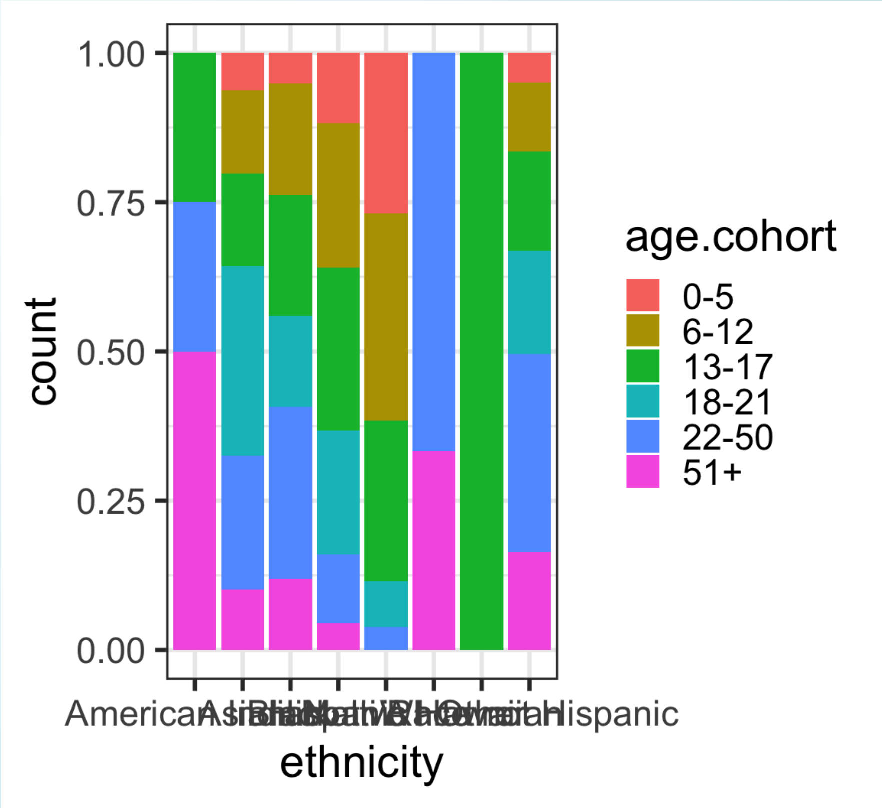


BARPLOTS WITH 2 VARIABLES: SEGMENTED BAR PLOTS

```r
ggplot(data = dds.discr,
       aes(x = ethnicity,
           fill = age.cohort)) +
  geom_bar()
```

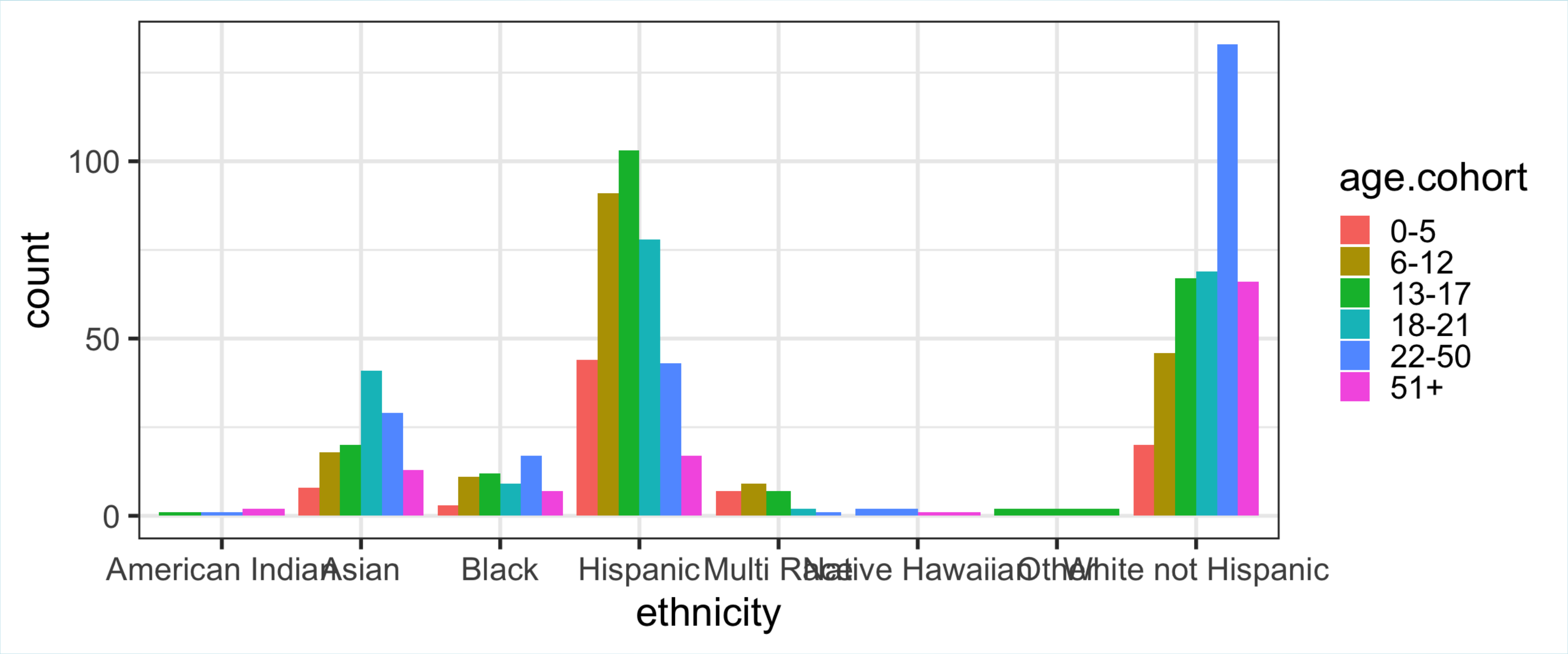
```r
ggplot(data = dds.discr,
       aes(x = ethnicity,
           fill = age.cohort)) +
  geom_bar(position = "fill")
```





BARPLOTS WITH 2 VARIABLES: SIDE-BY-SIDE BAR PLOTS

```
ggplot(data = dds.discr,
       aes(x = ethnicity,
           fill = age.cohort)) +
  geom_bar(position = "dodge")
```



SUMMARIZING CATEGORICAL DATA AND SOME DATA WRANGLING

dplyr

magrittr

janitor

FREQUENCY TABLES: count()

- count is from the dplyr package
- the output is a long tibble, and not a “nice” table

```
dds.discr %>% count(ethnicity)
```

```
# A tibble: 8 × 2
  ethnicity              n
  <fct>              <int>
1 American Indian        4
2 Asian                129
3 Black                 59
4 Hispanic             376
5 Multi Race            26
6 Native Hawaiian        3
7 Other                  2
8 White not Hispanic   401
```

```
dds.discr %>%
  count(ethnicity, age.cohort)
```

```
# A tibble: 35 × 3
   ethnicity       age.cohort     n
   <fct>           <fct>      <int>
 1 American Indian 13-17          1
 2 American Indian 22-50          1
 3 American Indian 51+            2
 4 Asian           0-5            8
 5 Asian           6-12          18
 6 Asian           13-17         20
 7 Asian           18-21         41
 8 Asian           22-50         29
 9 Asian           51+           13
10 Black           0-5            3
# ℹ 25 more rows
```

HOW TO USE THE PIPE %>%

The pipe operator %>% strings together commands to be performed sequentially

```
dds.discr %>% head(n=3)       # pronounce %>% as "then"
```

```
# A tibble: 3 × 6
     id age.cohort   age gender expenditures ethnicity
  <int> <fct>      <int> <fct>         <int> <fct>
1 10210 13-17         17 Female         2113 White not Hispanic
2 10409 22-50         37 Male          41924 White not Hispanic
3 10486 0-5            3 Male           1454 Hispanic
```

- Always *first list the tibble* that the commands are being applied to
- Can use **multiple pipes** to run multiple commands in sequence
  - What does the following code do?

```
dds.discr %>% head(n=3) %>% summary()
```

FREQUENCY TABLES: `janitor` PACKAGE'S `tabyl` FUNCTION

```r
# default table
dds.discr %>%
  tabyl(ethnicity)
```

ethnicity	n	percent
American Indian	4	0.004
Asian	129	0.129
Black	59	0.059
Hispanic	376	0.376
Multi Race	26	0.026
Native Hawaiian	3	0.003
Other	2	0.002
White not Hispanic	401	0.401

`adorn_` your table!

```r
dds.discr %>%
  tabyl(ethnicity) %>%
  adorn_totals("row") %>%
  adorn_pct_formatting(digits=2)
```

ethnicity	n	percent
American Indian	4	0.40%
Asian	129	12.90%
Black	59	5.90%
Hispanic	376	37.60%
Multi Race	26	2.60%
Native Hawaiian	3	0.30%
Other	2	0.20%
White not Hispanic	401	40.10%
Total	1000	100.00%

RELATIVE FREQUENCY TABLE

- A **relative frequency** table shows **proportions (or percentages)** instead of counts
- To the right I removed (deselected) the counts column (n) to create a relative frequency table

```
dds.discr %>%
  tabyl(ethnicity) %>%
  adorn_totals("row") %>%
  adorn_pct_formatting(digits=2) %>%
  select(-n)
```

```
           ethnicity percent
     American Indian   0.40%
               Asian  12.90%
               Black   5.90%
            Hispanic  37.60%
          Multi Race   2.60%
     Native Hawaiian   0.30%
               Other   0.20%
  White not Hispanic  40.10%
               Total 100.00%
```

CONTINGENCY TABLES (TWO-WAY TABLES)

- **Contingency tables** summarize data for two categorical variables
  - with each value in the table representing the number of times a particular combination of outcomes occurs
- **Row & column totals** are sometimes called **marginal totals**

```r
dds.discr %>%
  tabyl(ethnicity, gender) %>%
  adorn_totals(c("row", "col"))
```

```
          ethnicity Female Male Total
    American Indian      3    1     4
              Asian     61   68   129
              Black     26   33    59
           Hispanic    192  184   376
         Multi Race     13   13    26
    Native Hawaiian      2    1     3
              Other      1    1     2
 White not Hispanic    205  196   401
              Total    503  497  1000
```

CONTINGENCY TABLES WITH PERCENTAGES

```
dds.discr %>%
  tabyl(ethnicity, age.cohort) %>%
  adorn_totals(c("row")) %>%
  adorn_percentages("row") %>%
  adorn_pct_formatting(digits=0) %>%
  adorn_ns()
```

ethnicity	0-5	6-12	13-17	18-21	22-50	51+
American Indian	0% (0)	0% (0)	25% (1)	0% (0)	25% (1)	50% (2)
Asian	6% (8)	14% (18)	16% (20)	32% (41)	22% (29)	10% (13)
Black	5% (3)	19% (11)	20% (12)	15% (9)	29% (17)	12% (7)
Hispanic	12% (44)	24% (91)	27% (103)	21% (78)	11% (43)	5% (17)
Multi Race	27% (7)	35% (9)	27% (7)	8% (2)	4% (1)	0% (0)
Native Hawaiian	0% (0)	0% (0)	0% (0)	0% (0)	67% (2)	33% (1)
Other	0% (0)	0% (0)	100% (2)	0% (0)	0% (0)	0% (0)
White not Hispanic	5% (20)	11% (46)	17% (67)	17% (69)	33% (133)	16% (66)
Total	8% (82)	18% (175)	21% (212)	20% (199)	23% (226)	11% (106)

SUMMARIZING NUMERIC DATA

MEAN ANNUAL DDS EXPENDITURES BY RACE/ETHNICITY

```r
mean(dds.discr$expenditures)
```

```
[1] 18065.79
```

```r
dds.discr %>%
  summarize(
    ave = mean(expenditures),
    SD = sd(expenditures),
    med = median(expenditures))
```

```
# A tibble: 1 × 3
     ave     SD   med
   <dbl>  <dbl> <dbl>
1 18066. 19543.  7026
```

```r
dds.discr %>%
  group_by(ethnicity) %>%
  summarize(
    ave = mean(expenditures),
    SD = sd(expenditures),
    med = median(expenditures))
```

```
# A tibble: 8 × 4
  ethnicity              ave     SD   med
  <fct>                <dbl>  <dbl> <dbl>
1 American Indian     36438. 25694. 41818.
2 Asian               18392. 19209.  9369
3 Black               20885. 20549.  8687
4 Hispanic            11066. 15630.  3952
5 Multi Race           4457.  7332.  2622
6 Native Hawaiian     42782.  6576. 40727
7 Other                3316.  1836.  3316.
8 White not Hispanic 24698. 20604. 15718
```

get_summary_stats() FROM rstatix PACKAGE

```r
dds.discr %>% get_summary_stats()
```

```
# A tibble: 3 × 13
  variable         n   min   max median     q1     q3   iqr     mad    mean      sd
  <fct>        <dbl> <dbl> <dbl>  <dbl>  <dbl>  <dbl> <dbl>   <dbl>   <dbl>   <dbl>
1 id            1000 10210 99898 55384. 31809. 76135. 44326 3.27e4 5.47e4 2.56e4
2 age           1000     0    95     18     12     26    14 1.04e1 2.28e1 1.85e1
3 expenditures  1000   222 75098   7026   2899. 37713. 34814 7.76e3 1.81e4 1.95e4
# ℹ 2 more variables: se <dbl>, ci <dbl>
```

```r
dds.discr %>%
  group_by(ethnicity) %>%
  get_summary_stats(expenditures, type = "common")
```

```
# A tibble: 8 × 11
  ethnicity variable      n   min   max median    iqr   mean     sd     se     ci
  <fct>     <fct>     <dbl> <dbl> <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>
1 American… expendi…      4  3726 58392 41818. 34085. 36438. 25694. 12847. 40885.
2 Asian     expendi…    129   374 75098   9369  30892 18392. 19209.  1691.  3346.
3 Black     expendi…     59   240 60808   8687  37987 20885. 20549.  2675.  5355.
4 Hispanic  expendi…    376   222 65581   3952   7961. 11066. 15630.   806.  1585.
5 Multi Ra… expendi…     26   669 38619   2622   2060.  4457.  7332.  1438.  2962.
6 Native H… expendi…      3 37479 50141  40727   6331 42782.  6576.  3797. 16337.
7 Other     expendi…      2  2018  4615  3316.  1298.  3316.  1836.  1298. 16499.
8 White no… expendi…    401   340 68890  15718  39157 24698. 20604.  1029.  2023.
```

HOW TO FORCE ALL OUTPUT TO BE SHOWN? (1/2)

Use `kable()` from the `knitr` package.

```
dds.discr %>% get_summary_stats() %>% kable()
```

variable	n	min	max	median	q1	q3	iqr	
id	1000	10210	99898	55384.5	31808.75	76134.75	44326	327
age	1000	0	95	18.0	12.00	26.00	14	
expenditures	1000	222	75098	7026.0	2898.75	37712.75	34814	77

HOW TO FORCE ALL OUTPUT TO BE SHOWN? `knitr` (2/2)

Use `kable()` from the `knitr` package.

```
dds.discr %>%
  group_by(ethnicity) %>%
  get_summary_stats(expenditures, type = "common") %>%
  kable()
```

ethnicity	variable	n	min	max	median	iqr	mean
American Indian	expenditures	4	3726	58392	41817.5	34085.25	36438.250
Asian	expenditures	129	374	75098	9369.0	30892.00	18392.372
Black	expenditures	59	240	60808	8687.0	37987.00	20884.593
Hispanic	expenditures	376	222	65581	3952.0	7961.25	11065.569
Multi Race	expenditures	26	669	38619	2622.0	2059.75	4456.731
Native Hawaiian	expenditures	3	37479	50141	40727.0	6331.00	42782.333
Other	expenditures	2	2018	4615	3316.5	1298.50	3316.500

ethnicity	variable	n	min	max	median	iqr	mean
White not Hispanic	expenditures	401	340	68890	15718.0	39157.00	24697.549

BACK TO RESEARCH QUESTION

CASE STUDY: DISCRIMINATION IN DEVELOPMENTAL DISABILITY SUPPORT (1.7.1)

- **Previous research**
  - Researchers examined DDS expenditures for developmentally disabled residents by ethnicity
  - Found that the mean annual expenditures on Hispanics was less than that on White non-Hispanics.
- **Result**: an allegation of ethnic discrimination was brought against the California DDS.
- **Question: Are the data sufficient evidence of ethnic discrimination?**