N741 - Data Wrangling - learning dplyr and more tidyverse packages

Melinda Higgins

2/1/2022

Load the Davis dataset from the carData package

```
# load the carData package
# we'll work with the Davis dataset
# which is a part of this package

library(carData)
data(Davis)
```

Take a quick look at the Davis dataset

Load the `tibble` package and use the `glimpse()` function to take a quick peek at the `Davis` dataset.

The `Davis` dataset has 200 rows and 5 columns. The subjects were men and women engaged in regular exercise. There are some missing data. The 5 variables are:

- sex
  - a factor variable with 2 levels for F, female and M, male
- weight
  - a numeric variable for **measured** weight in kg
- height
  - a numeric variable for **measured** height in cm
- repwt
  - a numeric variable for **self-reported** weight in kg
- repht
  - a numeric variable for **self-reported** height in cm

To learn more about this dataset run `help(Davis, package = "carData")`.

```
library(tibble)
glimpse(Davis)
```

```
## Rows: 200
## Columns: 5
## $ sex    <fct> M, F, F, M, F, M, M, M, M, M, M, F, F, F, F, F, M, F, M, F, M, ~
## $ weight <int> 77, 58, 53, 68, 59, 76, 76, 69, 71, 65, 70, 166, 51, 64, 52, 65~
## $ height <int> 182, 161, 161, 177, 157, 170, 167, 186, 178, 171, 175, 57, 161,~
## $ repwt  <int> 77, 51, 54, 70, 59, 76, 77, 73, 71, 64, 75, 56, 52, 64, 57, 66,~
## $ repht  <int> 180, 159, 158, 175, 155, 165, 165, 180, 175, 170, 174, 163, 158~
```

Compute BMI from measured `height` and `weight`

The equation for BMI is

$$BMI = \frac{weight(kg)}{[height(m)]^2}$$

To compute BMI we need to:

1. convert `height` in cm to m
2. then compute BMI

So, let's add 2 new variables to our dataset using the `mutate()` function from the `dplyr` package.

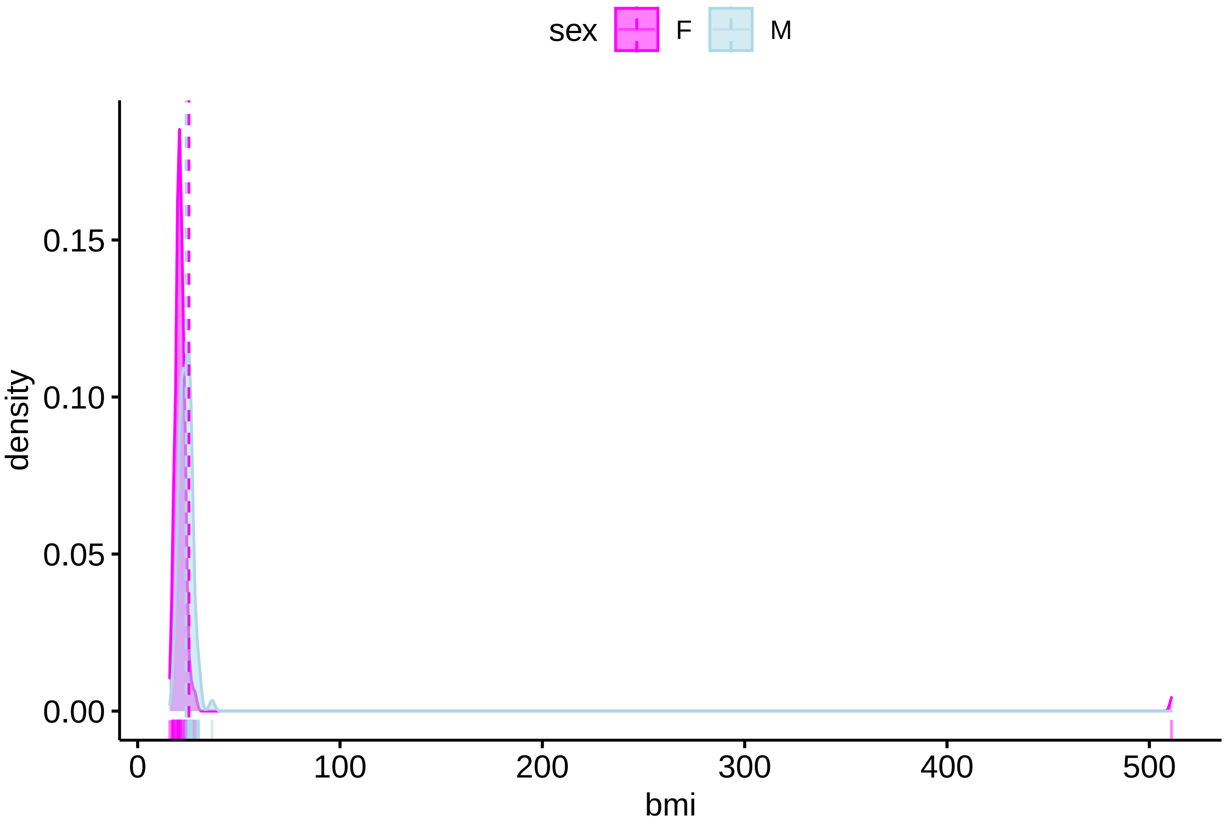
```
# load dplyr package
library(dplyr)

Davis2 <- Davis %>%
  mutate(height_m = height/100) %>%
  mutate(bmi = weight / ((height_m)^2))
```

BMI histograms by `sex`

Let's try out the `ggpubr` package and use the `ggdensity()` function.

Learn more at https://rpkgs.datanovia.com/ggpubr/.

```
library(ggpubr)
ggdensity(Davis2, x = "bmi",
   add = "mean", rug = TRUE,
   color = "sex", fill = "sex",
   palette = c("magenta", "light blue"))
```



Well this looks odd. I'm guessing there is an outlier somewhere. We could open the data and look at it in a viewer, but let's try to do it with code.

The `dplyr` package also has an `arrange()` function. So, let's sort the data and see if we can spot which case has the really large BMI value.

The default is to arrange (or sort) the rows in ascending order. But we want to see the largest value, so we'll add `desc()` to get the descending sorted order.

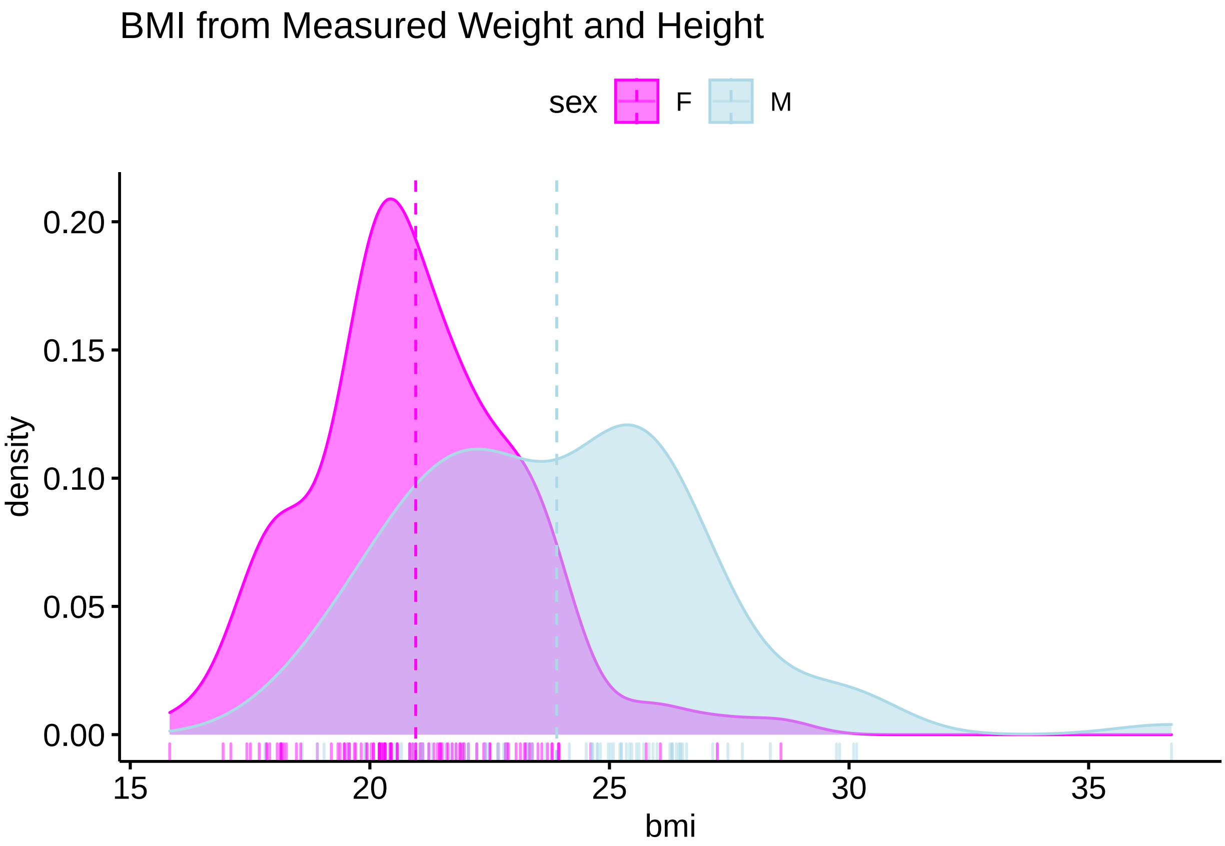
```
Davis2 %>%
  arrange(desc(bmi)) %>%
  head()
```

	sex	weight	height	repwt	repht	height_m	bmi
12	F	166	57	56	163	0.57	510.92644
21	M	119	180	124	178	1.80	36.72840
30	M	101	183	100	180	1.83	30.15916
97	M	103	185	101	182	1.85	30.09496
54	M	102	185	107	185	1.85	29.80278
192	M	89	173	86	173	1.73	29.73704

If I had to guess, it looks like the measured height and weight were flipped for case 12. But for now let's filter this case out and remake our plot.

So, we'll use the `filter()` function also from `dplyr` package.

```
Davis3 <- Davis2 %>%
  filter(bmi < 50)

ggdensity(Davis3, x = "bmi",
   add = "mean", rug = TRUE,
   color = "sex", fill = "sex",
   palette = c("magenta", "light blue")) +
  ggtitle("BMI from Measured Weight and Height")
```



Get summary statistics of weight, height and bmi

Let's get the mean for weight, height and bmi

```
Davis3 %>%
  summarise(across(c(weight, height, bmi),
            ~ mean(.x, na.rm = TRUE))
  )
```

weight	height	bmi
65.29648	170.5879	22.25761

Add `group_by()` to get the means by `sex`

```
Davis3 %>%
  group_by(sex) %>%
  summarise(across(c(weight, height, bmi),
            ~ mean(.x, na.rm = TRUE))
  )
```

sex	weight	height	bmi
F	56.89189	164.7027	20.95632
M	75.89773	178.0114	23.89901

This is easier with `get_summary_stats()` from `rstatix` package

Let's try this again and get more stats.

First use the `select()` function from `dplyr` and then get the summary stats.

```
library(rstatix)

Davis3 %>%
  select(weight, height, bmi) %>%
  get_summary_stats()
```

variable	n	min	max	median	q1	q3	iqr	mad	mean	sd	se	ci
bmi	199	15.822	36.728	21.799	20.223	23.936	3.713	2.552	22.258	3.009	0.213	0.421
height	199	148.000	197.000	170.000	164.000	177.500	13.500	10.378	170.588	8.949	0.634	1.251
weight	199	39.000	119.000	63.000	55.000	73.500	18.500	11.861	65.296	13.343	0.946	1.865

Let's just get mean and sd (standard deviation) and add `group_by()` to get the stats by `sex`. *NOTE: Add `sex` to the `select()` step.*

```
Davis3 %>%
  group_by(sex) %>%
  select(sex, weight, height, bmi) %>%
  get_summary_stats(type = "mean_sd")
```

sex	variable	n	mean	sd
F	bmi	111	20.956	2.176
F	height	111	164.703	5.683
F	weight	111	56.892	6.891
M	bmi	88	23.899	3.120
M	height	88	178.011	6.441
M	weight	88	75.898	11.890

Compare measured vs self-report heights and weights by sex

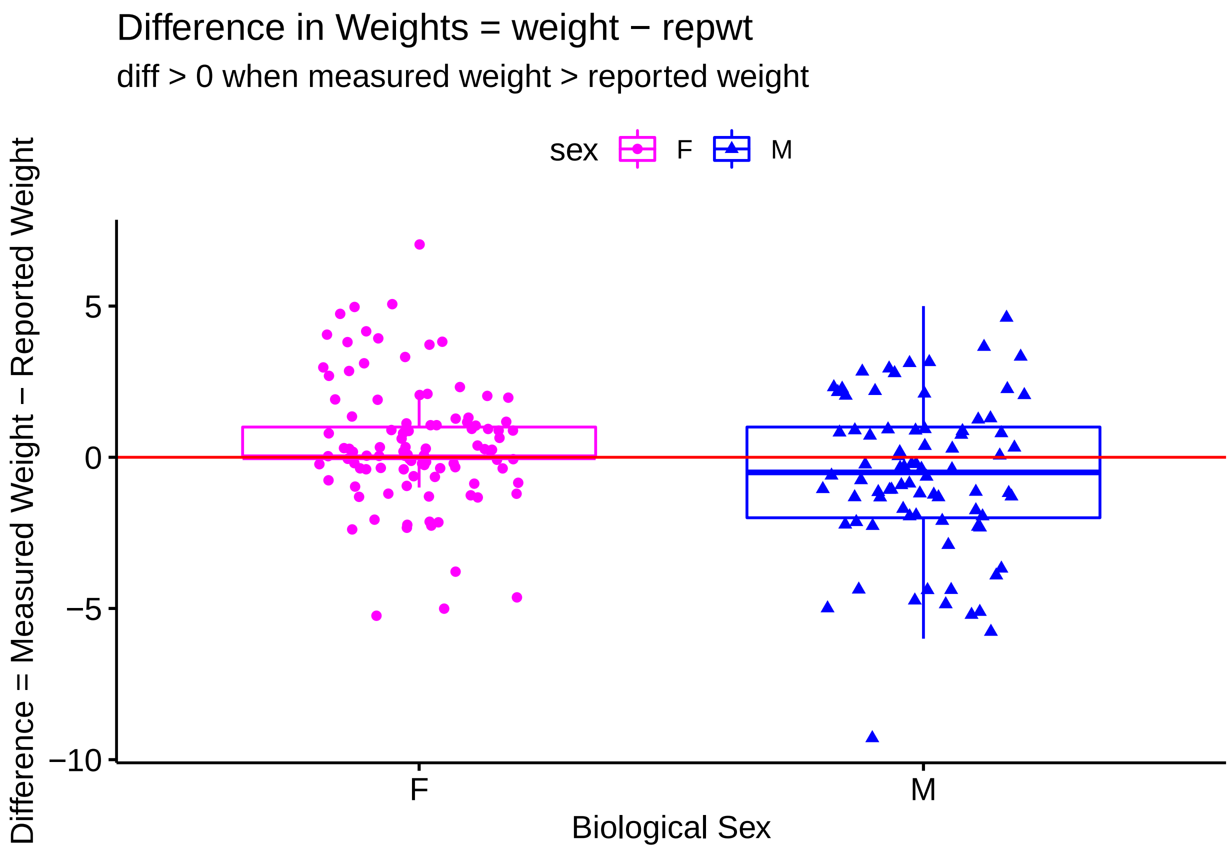
I've often heard a saying that "women weigh less and men are taller on paper". But let's take a look at the discrepancies between the directly measures `height` and `weight` to the self-reports `repwt` and `repht` - overall and by `sex`.

```
Davis3 <- Davis3 %>%
  mutate(diff_wt_repwt = weight - repwt) %>%
  mutate(diff_ht_repht = height - repht)
```

Now that we've computed these differences, let's look at these differences by sex. Differences < 0 indicate that the self-reported `repwt` or `repht` were larger than the measured `weight` or `height`.

Let's keep using the `ggpubr` package and try the `ggboxplot()` function and add a reference line, a title and a subtitle using functions from `ggplot2` package which is loaded with `ggpubr`. We'll also clean up the x-axis and y-axis labels.

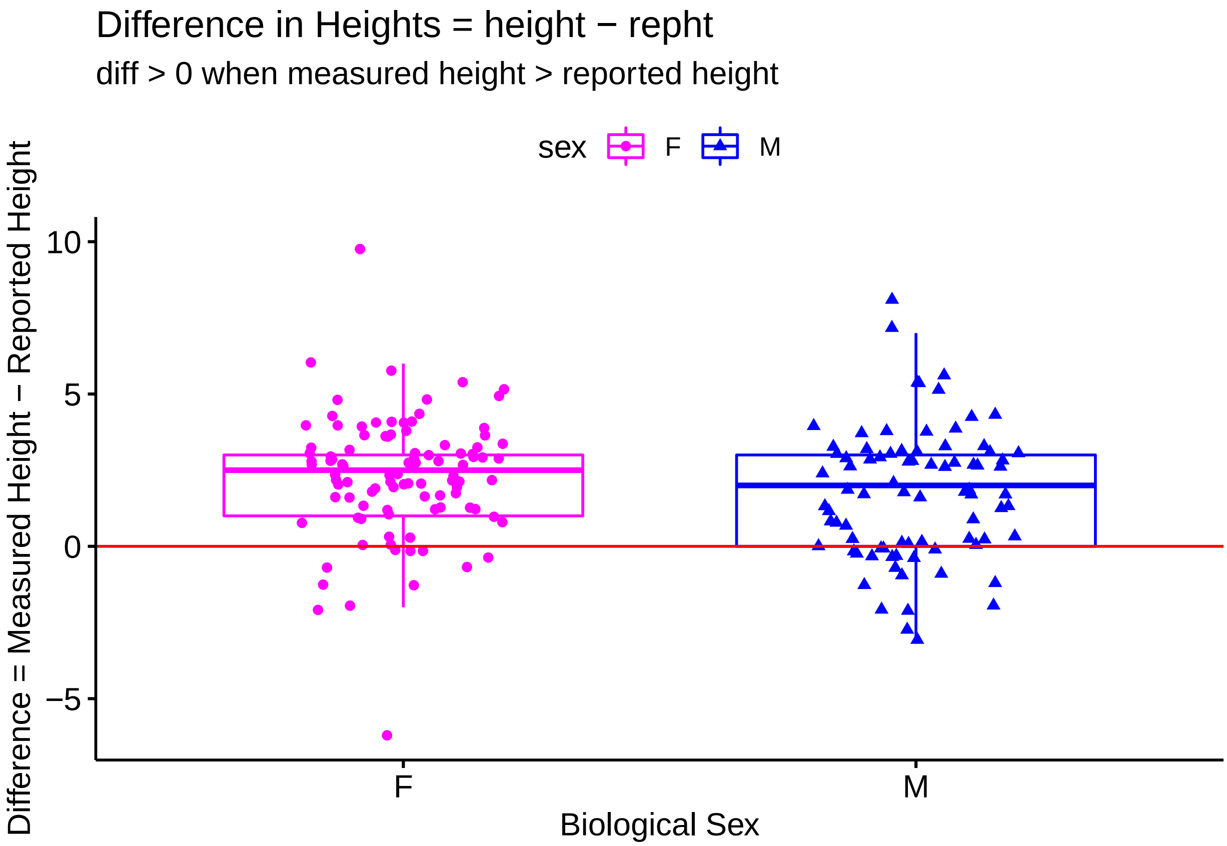
```
ggboxplot(Davis3, x = "sex", y = "diff_wt_repwt",
                color = "sex",
                palette =c("magenta", "blue"),
                add = "jitter", shape = "sex") +
  geom_hline(yintercept = 0, color = "red") +
  labs(title = "Difference in Weights = weight - repwt",
       subtitle = "diff > 0 when measured weight > reported weight") +
  xlab("Biological Sex") +
  ylab("Difference = Measured Weight - Reported Weight")
```



It looks like for females, their actual weights are larger than their self-reported weights.

Let's take a look at the differences in the heights.

```
ggboxplot(Davis3, x = "sex", y = "diff_ht_repht",
                color = "sex",
                palette =c("magenta", "blue"),
                add = "jitter", shape = "sex") +
  geom_hline(yintercept = 0, color = "red") +
  labs(title = "Difference in Heights = height - repht",
       subtitle = "diff > 0 when measured height > reported height")  +
  xlab("Biological Sex") +
  ylab("Difference = Measured Height - Reported Height")
```



From this plot it looks like the measured heights are higher than the self-reported heights for both females and males.

Get summary stats of these differences by `sex`.

```
Davis3 %>%
  group_by(sex) %>%
  select(sex, diff_wt_repwt, diff_ht_repht) %>%
  get_summary_stats(type = "mean_sd")
```

sex	variable	n	mean	sd
F	diff_ht_repht	100	2.330	2.000
F	diff_wt_repwt	100	0.480	2.062
M	diff_ht_repht	82	1.756	2.146

sex	variable	n	mean	sd
M	diff_wt_repwt	82	-0.585	2.489