Data Visualization

Part II

Table of contents

1 Agenda

We are going to **visually** analyze two datasets and see if we can tell stories from the visuals.

```
Registered S3 method overwritten by 'printr':
  method                from     
  knit_print.data.frame rmarkdown
```

2 Setting up

Let's first load the `ggplot2` package:

```r
if (!require(ggplot2)) {
  install.packages("ggplot2") # install if not already installed
}
```

```
Loading required package: ggplot2
```

```r
library (ggplot2)
```

3 Titanic Survival

Credit: John Parrot/Stocktrek Images

3.1 Load the dataset

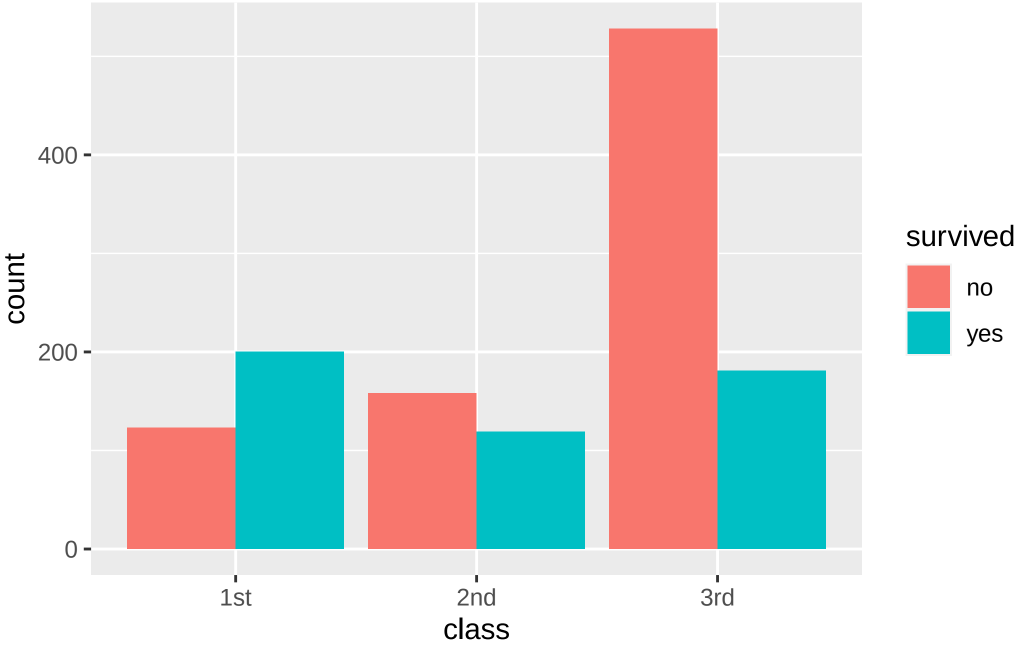
```r
titanic = read.csv("https://raw.githubusercontent.com/ahmedmoustafa/datasets/main/titanic/titanic.csv")
head(titanic)
```

name	survived	sex	age	class
Allen, Miss. Elisabeth Walton	yes	female	29.0000	1st
Allison, Master. Hudson Trevor	yes	male	0.9167	1st
Allison, Miss. Helen Loraine	no	female	2.0000	1st
Allison, Mr. Hudson Joshua Crei	no	male	30.0000	1st
Allison, Mrs. Hudson J C (Bessi	no	female	25.0000	1st
Anderson, Mr. Harry	yes	male	48.0000	1st

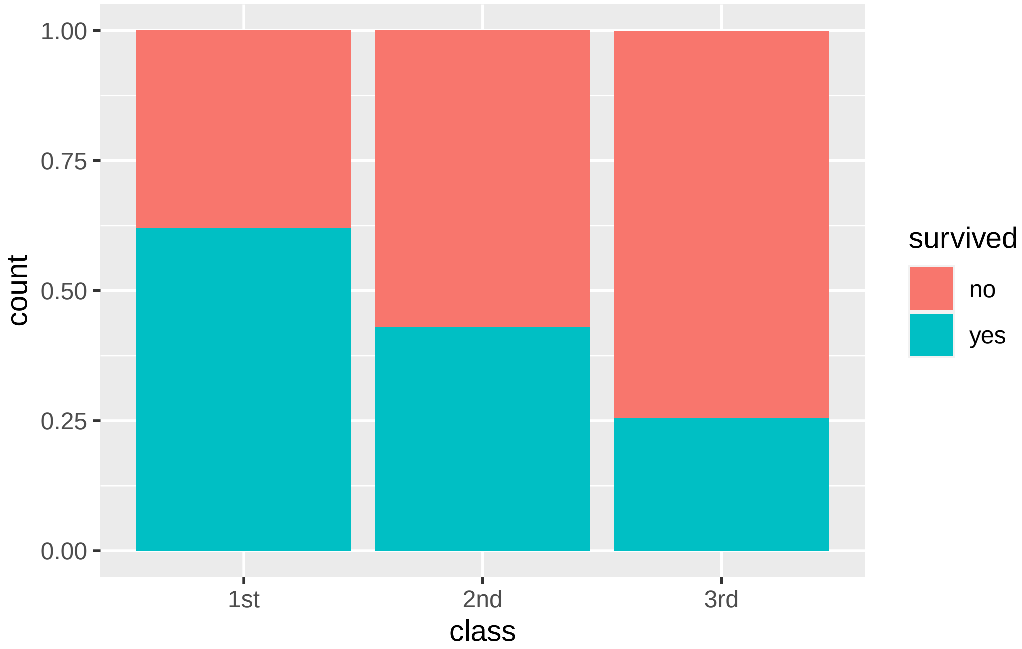
name	survived	sex	age	class

3.2 Survival by class

```r
ggplot(titanic) +
  geom_bar(aes(x = class, fill = survived), position = "dodge")
```

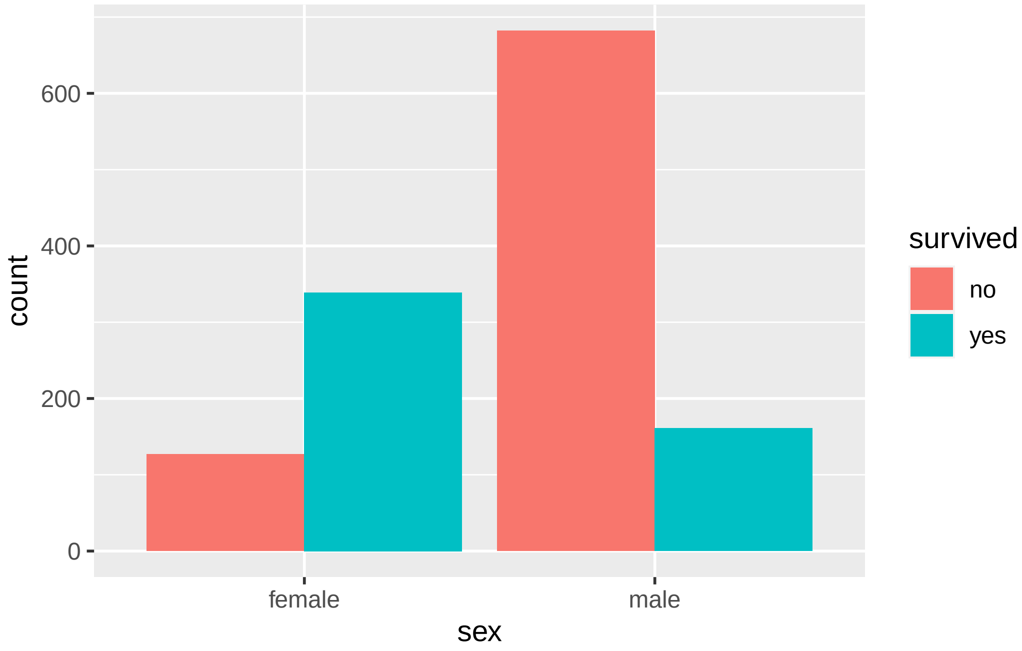


```r
ggplot(titanic) +
  geom_bar(aes(x = class, fill = survived), position = "fill")
```

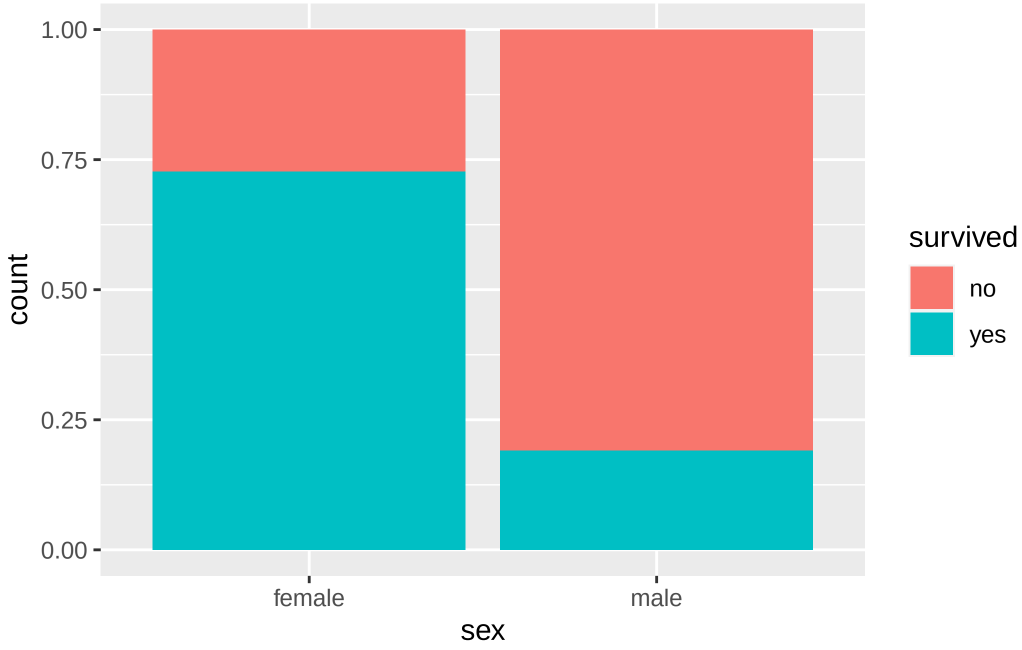


3.3 Survival by sex

```r
ggplot(titanic) +
  geom_bar(aes(x = sex, fill = survived), position = "dodge")
```



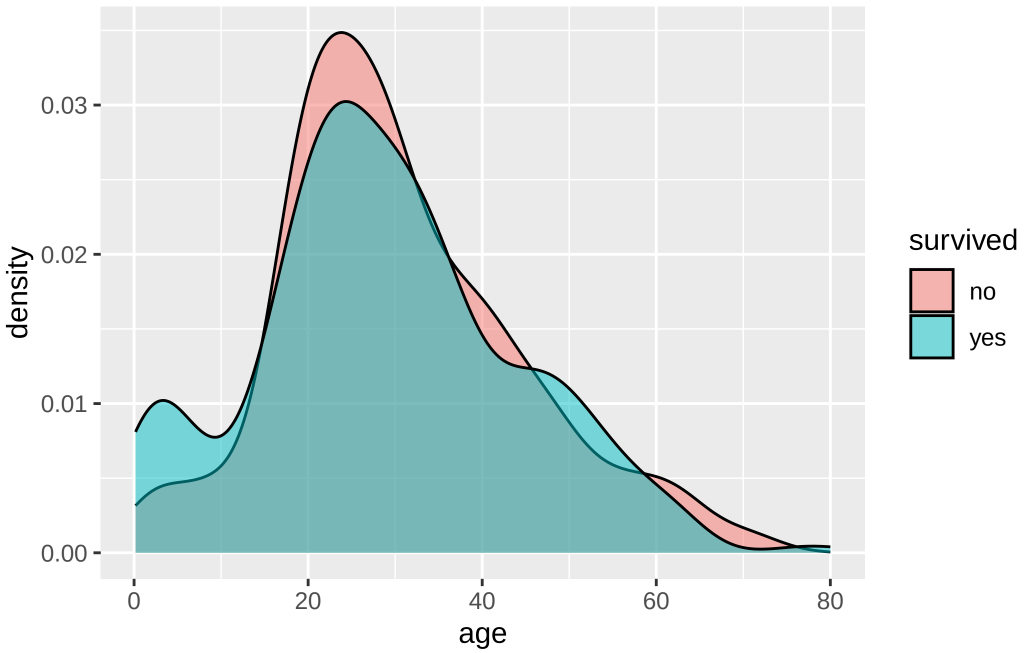
```r
ggplot(titanic) +
  geom_bar(aes(x = sex, fill = survived), position = "fill")
```



3.4 Survival by age

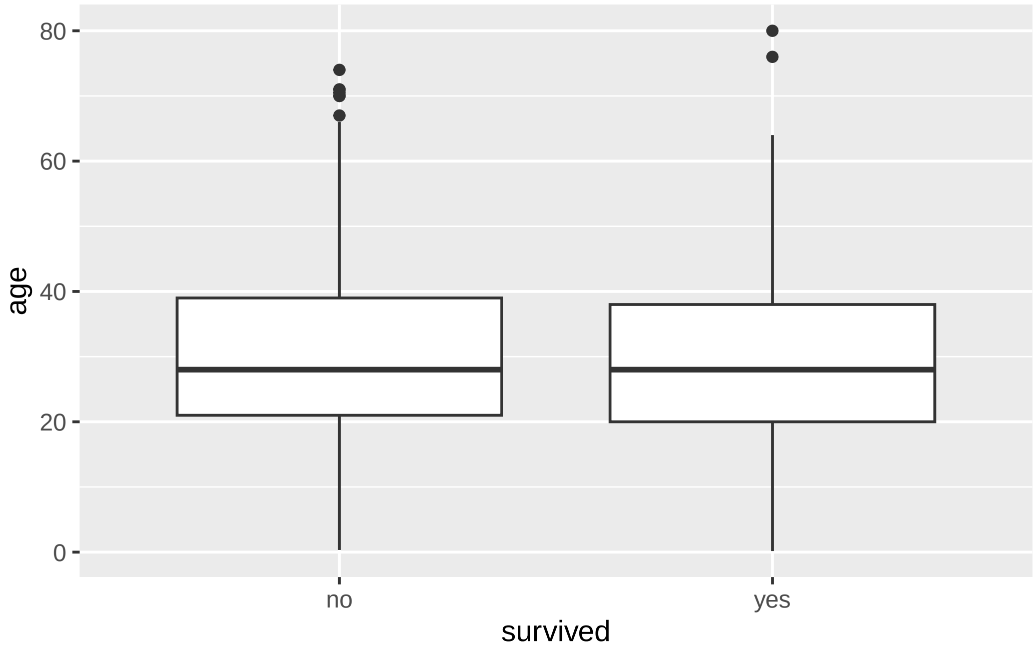
```r
ggplot(titanic) +
  geom_density(aes(x = age, fill = survived), alpha = 0.5)
```

```
Warning: Removed 263 rows containing non-finite values (`stat_density()`).
```



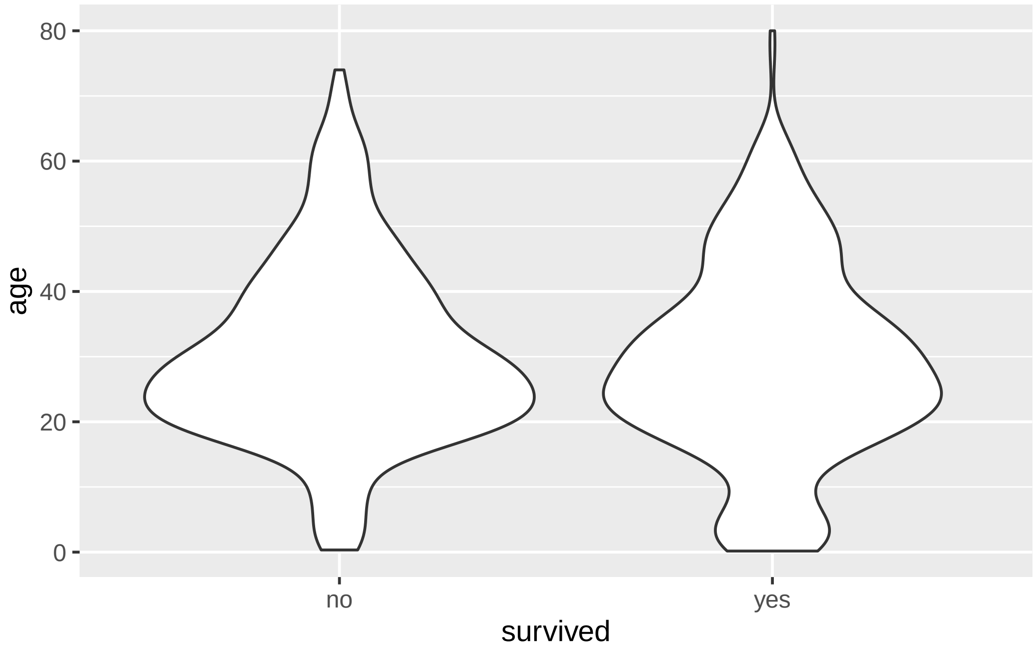
```r
ggplot(titanic) +
  geom_boxplot(aes(x = survived, y = age))
```

```
Warning: Removed 263 rows containing non-finite values (`stat_boxplot()`).
```



```
ggplot(titanic) +
  geom_violin(aes(x = survived, y = age))
```

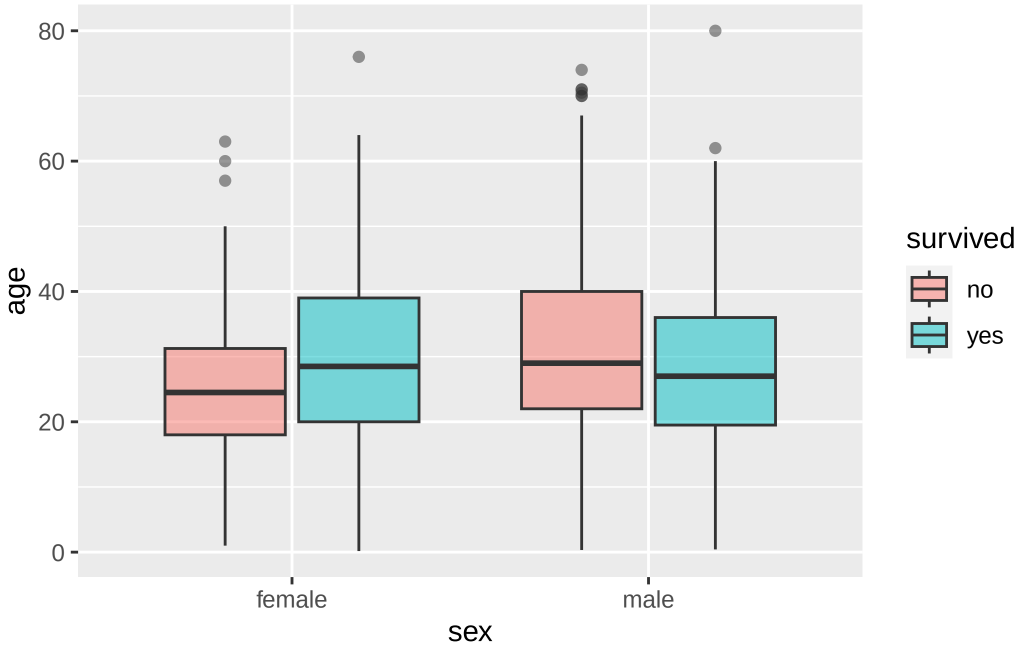
```
Warning: Removed 263 rows containing non-finite values (`stat_ydensity()`).
```



3.5 Survival by age & sex

```r
ggplot(titanic) +
  geom_boxplot(aes(x = sex, y = age, fill = survived), alpha = 0.5)
```

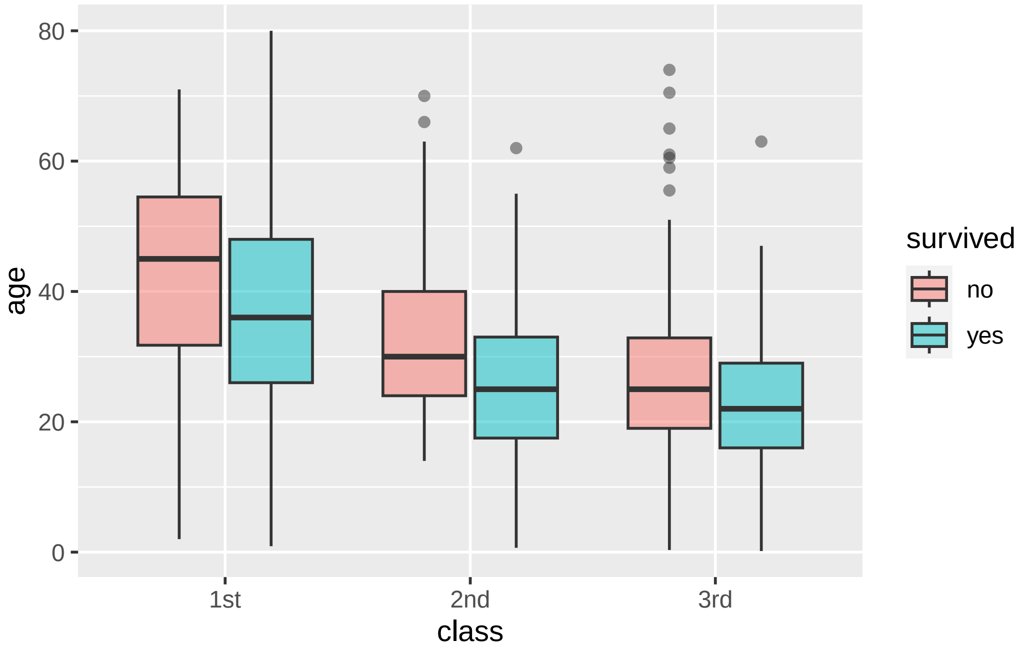
```
Warning: Removed 263 rows containing non-finite values (`stat_boxplot()`).
```



3.6 Survival by class & age

```r
ggplot(titanic) +
  geom_boxplot(aes(x = class, y = age, fill = survived), alpha = 0.5)
```

```
Warning: Removed 263 rows containing non-finite values (`stat_boxplot()`).
```


80
60
40
20
0
age
1st
2nd
3rd
class
survived
no
yes

4 Smoking and Pregnancy

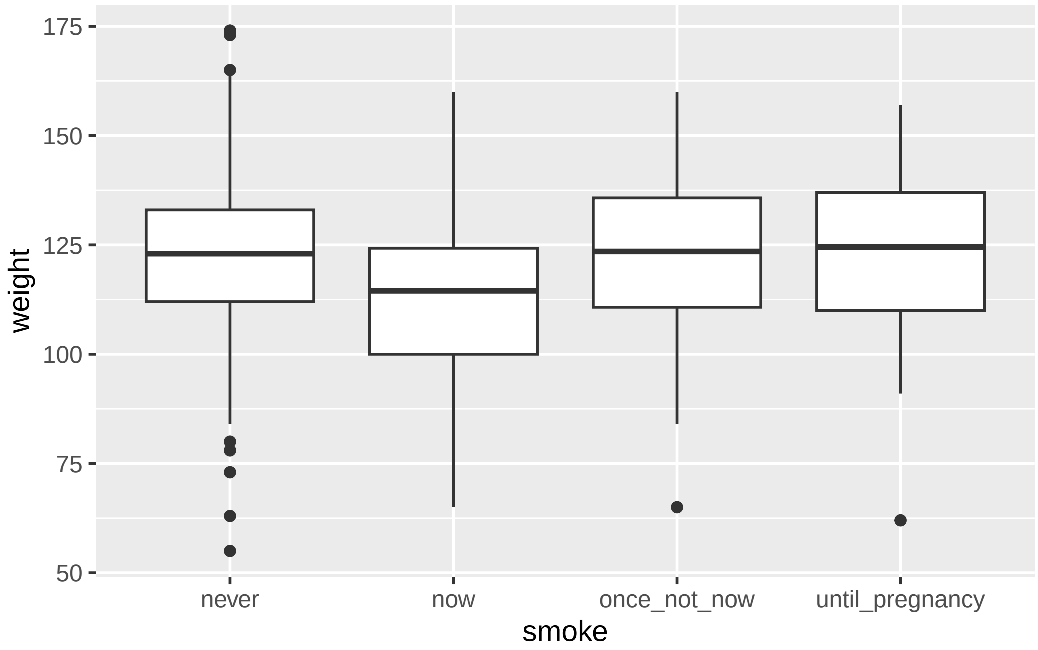
4.1 Load the dataset

```
smoking =
  ↪  read.csv("https://raw.githubusercontent.com/ahmedmoustafa/datasets/main/smoking/smokin
head(smoking)
```

id	date	gestation	weight	parity	mom.race	mom.age	mom.edu	mom.height	mom.weight	dad.race	dad.age	dad.edu	dad.height	dad.weight	marital	income	smoke	quit.time	cigs
15	1411	284	120	1	asian	27	5	62	100	asian	31	5	65	110	1	1	never	0	0
20	1499	282	113	2	white	33	5	64	135	white	38	5	70	148	1	4	never	0	0
100	1673	286	136	4	white	25	2	62	93	white	28	2	64	130	1	4	until_pregnancy	2	2
129	1562	245	132	2	black	23	1	65	140	black	23	4	71	192	1	2	never	0	0
142	1408	289	120	3	white	25	4	62	125	white	26	1	70	180	0	2	never	0	0
171	1593	282	144	4	white	32	2	64	124	white	36	1	74	185	1	2	now	1	1

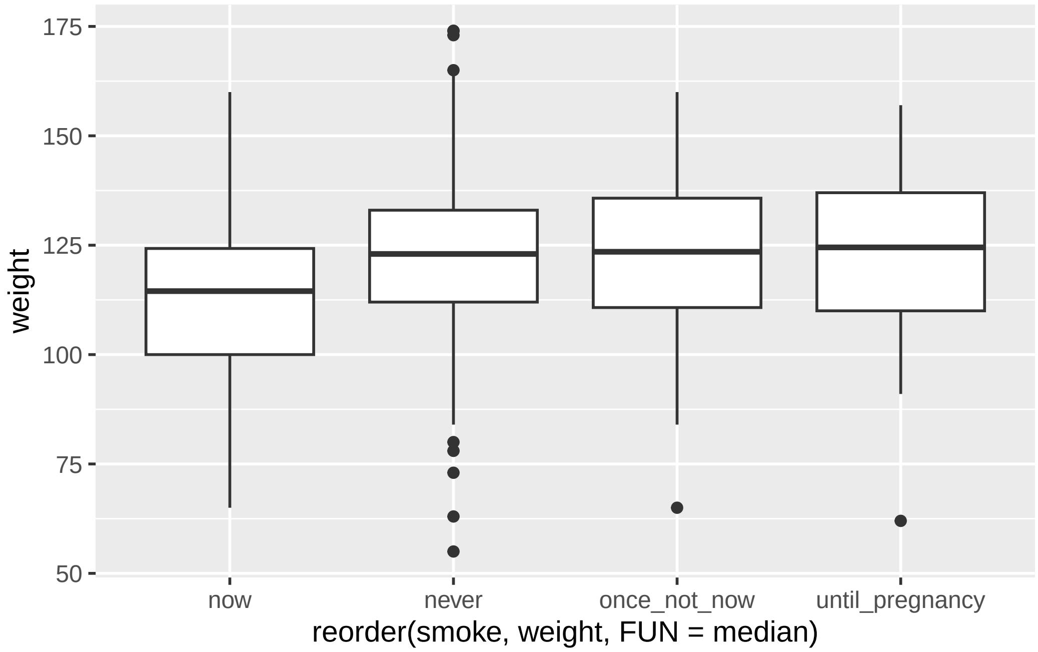
4.2 Mom's smoking and baby's weight

```r
ggplot(smoking) +
  geom_boxplot(aes(x = smoke, y = weight))
```



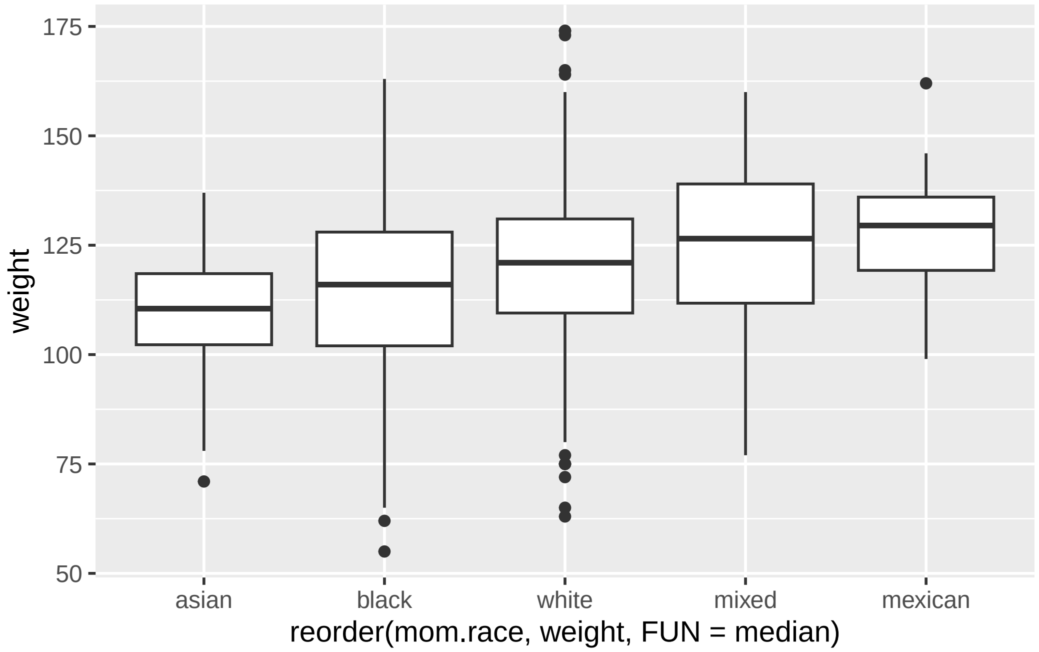
4.3 Mom's smoking and baby's weight with reordered x-axis

```r
ggplot(smoking) +
  geom_boxplot(aes(x = reorder(smoke, weight, FUN = median), y =
  ↪  weight))
```



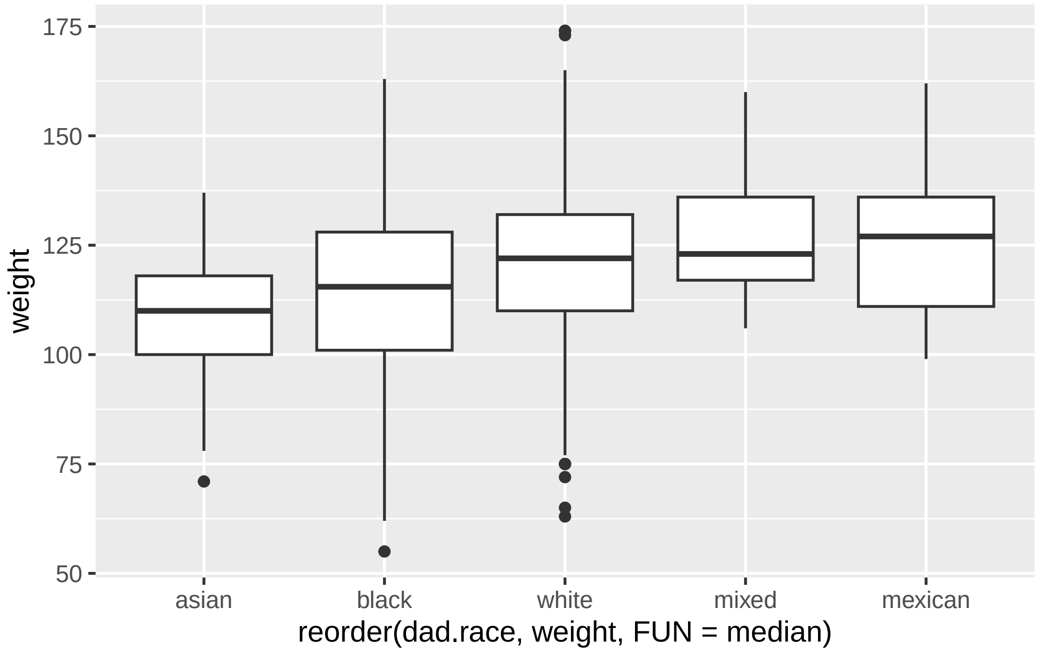
4.4 Mom's race and baby's weight

```r
ggplot(smoking) +
  geom_boxplot(aes(x = reorder(mom.race, weight, FUN = median), y =
    weight))
```



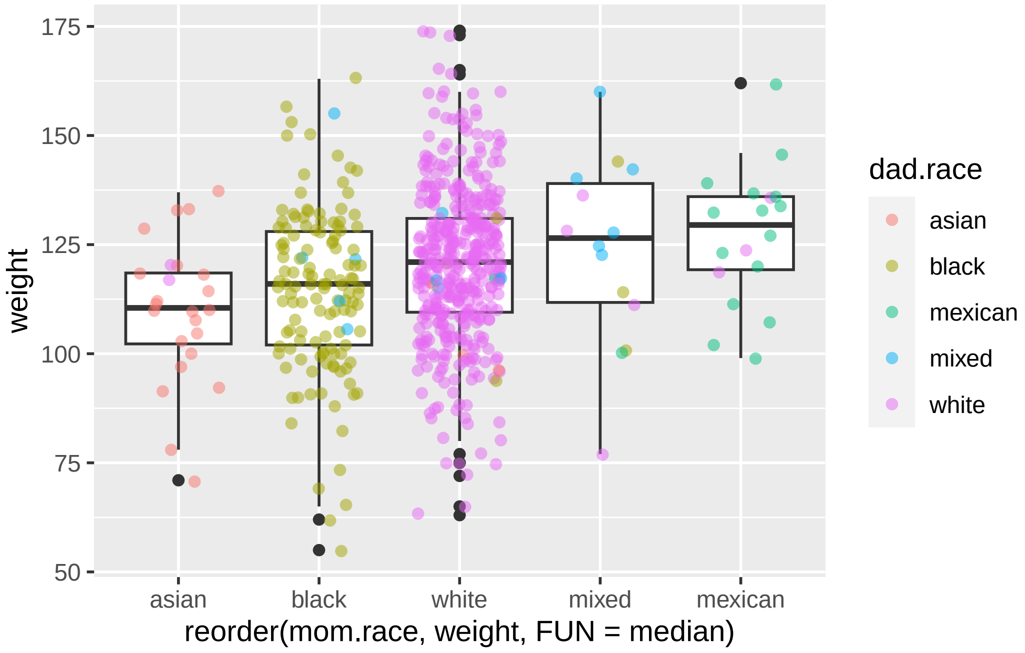
4.5 Dad's race and baby's weight

```
ggplot(smoking) +
  geom_boxplot(aes(x = reorder(dad.race, weight, FUN = median), y =
   ↪   weight))
```



4.6 Mom's race and baby's weight and dad's race

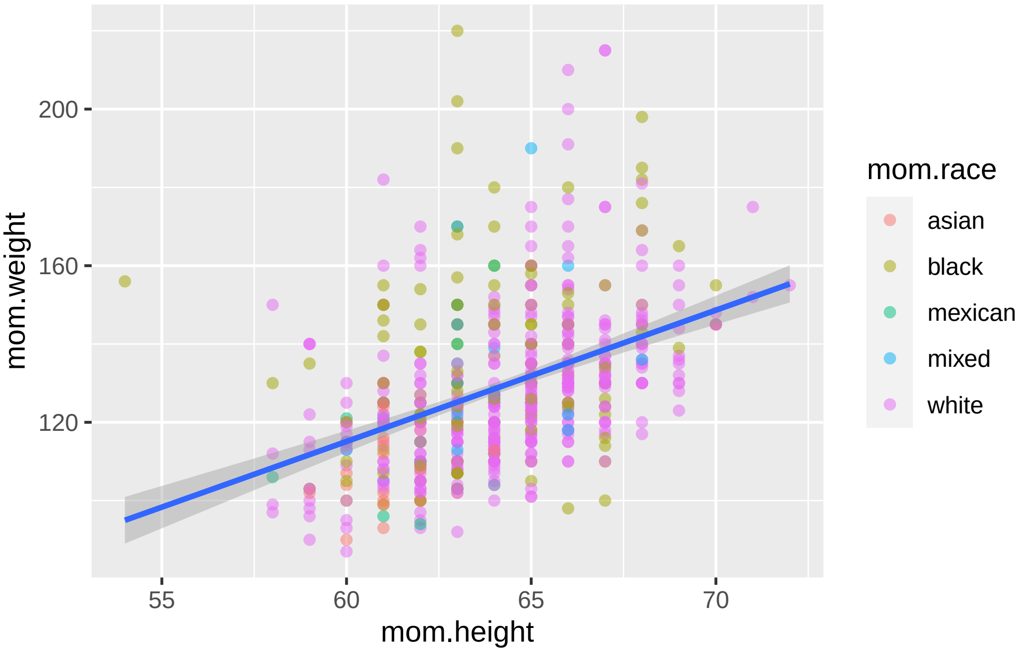
```
ggplot(smoking) +
  geom_boxplot(aes(x = reorder(mom.race, weight, FUN = median), y =
  ↪  weight)) +
  geom_jitter(aes(x = reorder(mom.race, weight, FUN = median), y =
  ↪  weight, color = dad.race), alpha = 0.5, width = 0.3)
```



4.7 Mom's height and moms's weight

```
ggplot(smoking) +
  geom_point(aes(x = mom.height, y = mom.weight, color = mom.race),
   ↪  alpha = 0.5) +
  geom_smooth(aes(x = mom.height, y = mom.weight), method = "lm")
```

```
`geom_smooth()` using formula = 'y ~ x'
```



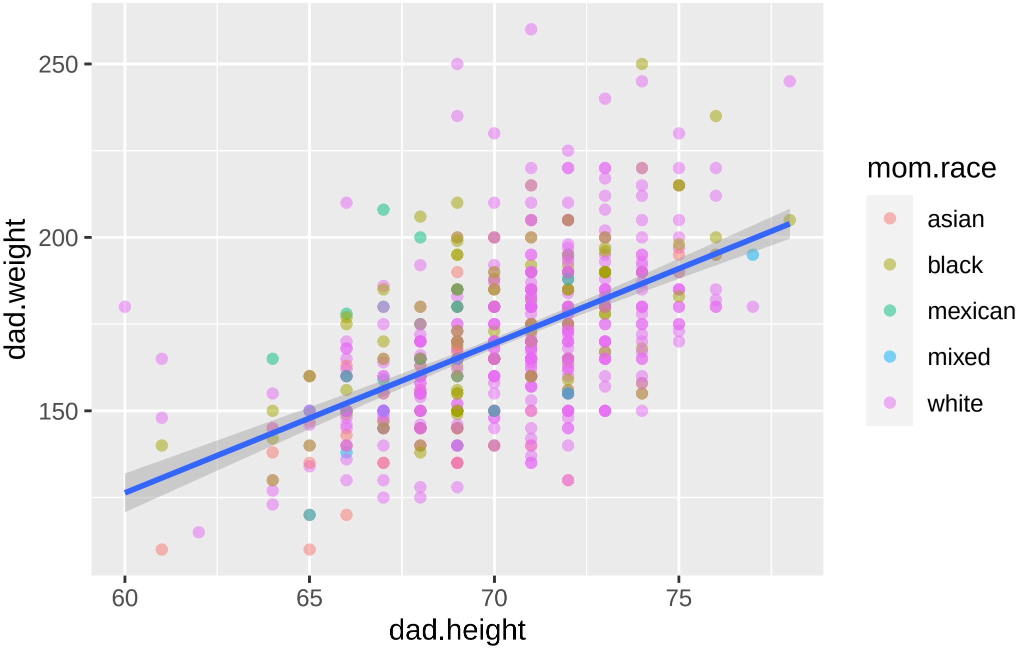
```r
model = lm (data = smoking, formula = mom.weight ~ mom.height)
summary(model)
```

```
Call:
lm(formula = mom.weight ~ mom.height, data = smoking)

Residuals:
    Min      1Q  Median      3Q     Max 
-38.579 -11.933  -3.515   7.276  94.839 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -86.174     18.671  -4.615 4.79e-06 ***
mom.height     3.354      0.291  11.526  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 18.55 on 608 degrees of freedom
Multiple R-squared:  0.1793,    Adjusted R-squared:  0.178 
F-statistic: 132.8 on 1 and 608 DF,  p-value: < 2.2e-16
```

4.8 Dad's height and dad's weight

```r
ggplot(smoking) +
  geom_point(aes(x = dad.height, y = dad.weight, color = mom.race),
  ↪  alpha = 0.5) +
  geom_smooth(aes(x = dad.height, y = dad.weight), method = "lm")
```

```
`geom_smooth()` using formula = 'y ~ x'
```



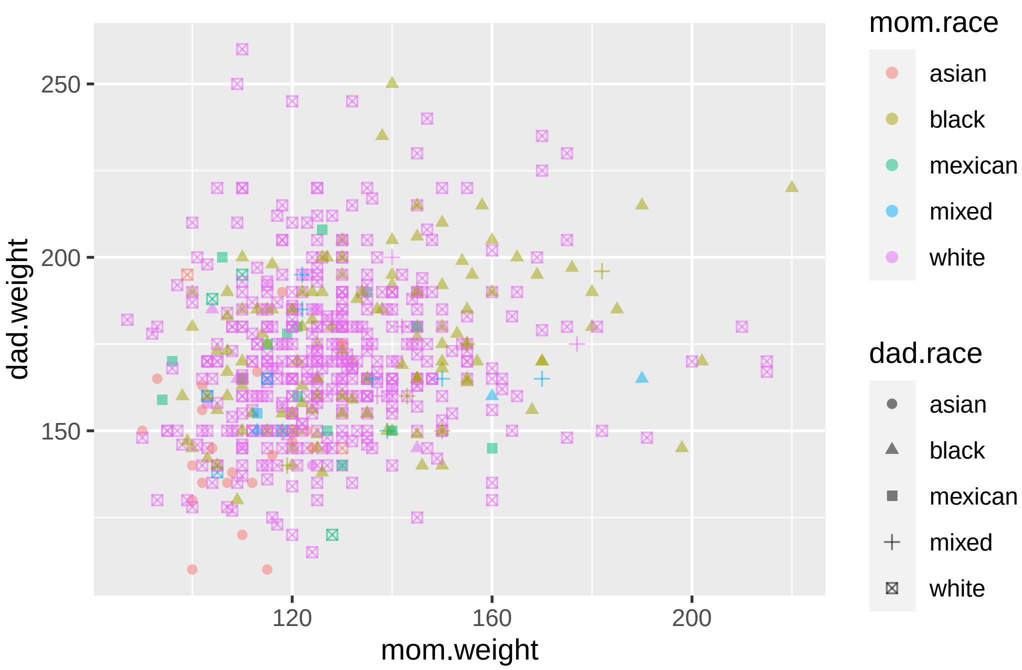
```r
model = lm (data = smoking, formula = dad.weight ~ dad.height)
summary(model)
```

```
Call:
lm(formula = dad.weight ~ dad.height, data = smoking)

Residuals:
    Min      1Q  Median      3Q     Max 
-48.067 -13.067  -1.825  10.554  86.243 
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -132.2898    18.8057  -7.035  5.4e-12 ***
dad.height     4.3105     0.2674  16.120  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 19.02 on 608 degrees of freedom
Multiple R-squared:  0.2994,    Adjusted R-squared:  0.2983
F-statistic: 259.9 on 1 and 608 DF,  p-value: < 2.2e-16
```

4.9 Mom's weight and dad's weight

```r
ggplot(smoking) +
  geom_point(aes(x = mom.weight, y = dad.weight, color = mom.race, shape
   ↪  = dad.race), alpha = 0.5)
```

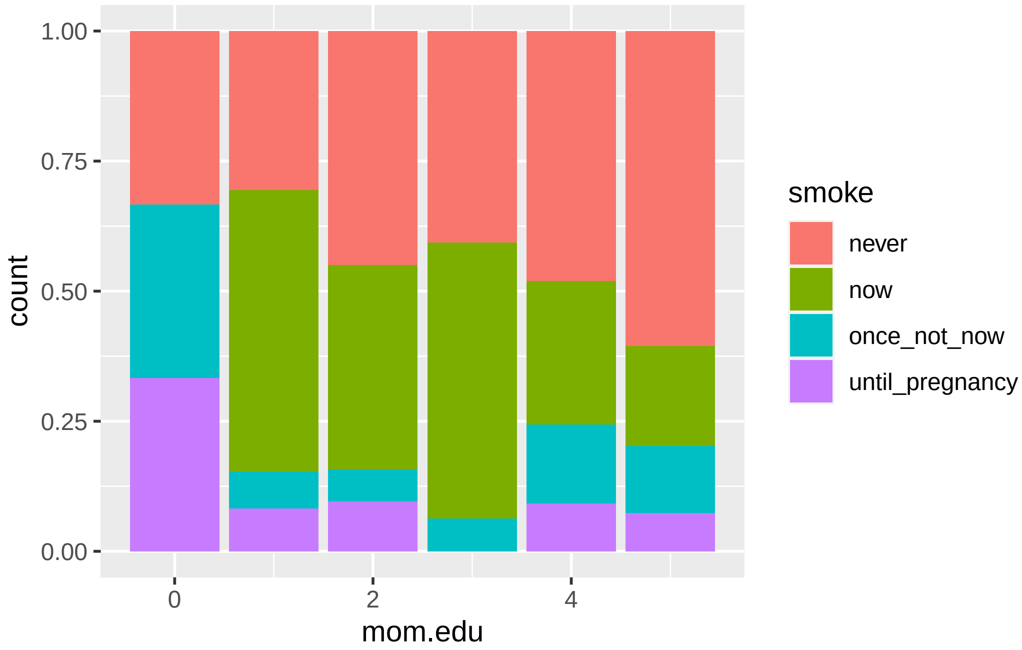


```r
model = lm (data = smoking, formula = dad.weight ~ mom.weight)
summary(model)
```

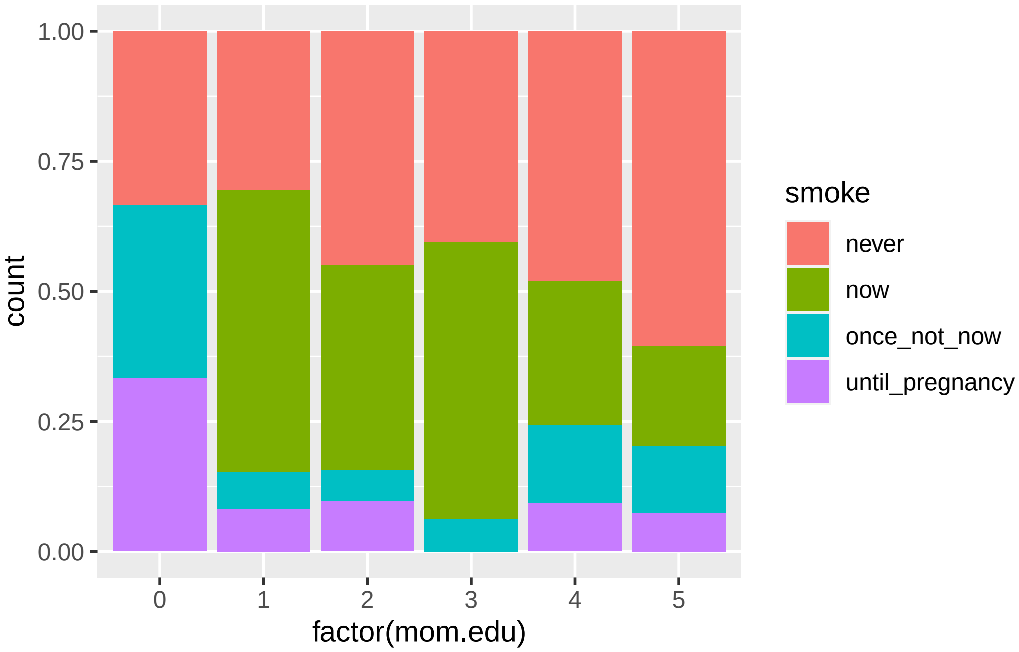
```
Call:
lm(formula = dad.weight ~ mom.weight, data = smoking)

Residuals:
    Min      1Q  Median      3Q     Max 
-57.481 -15.817  -2.051  14.097  93.646 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 141.54810    5.74900  24.621  < 2e-16 ***
mom.weight    0.22551    0.04407   5.117 4.16e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 22.25 on 608 degrees of freedom
Multiple R-squared:  0.04129,	Adjusted R-squared:  0.03972 
F-statistic: 26.19 on 1 and 608 DF,  p-value: 4.159e-07
```

4.10 Mom's smoking and mom's education

```
ggplot(smoking) +
  geom_bar(aes(x = mom.edu, fill = smoke), position = "fill")
```

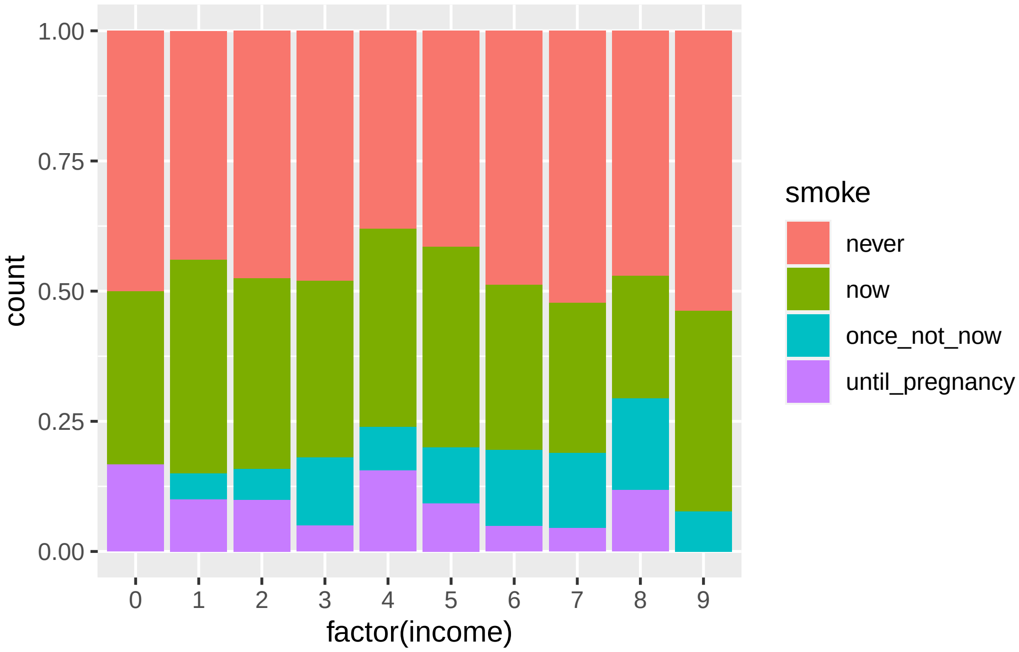


```
ggplot(smoking) +
  geom_bar(aes(x = factor(mom.edu), fill = smoke), position = "fill")
```



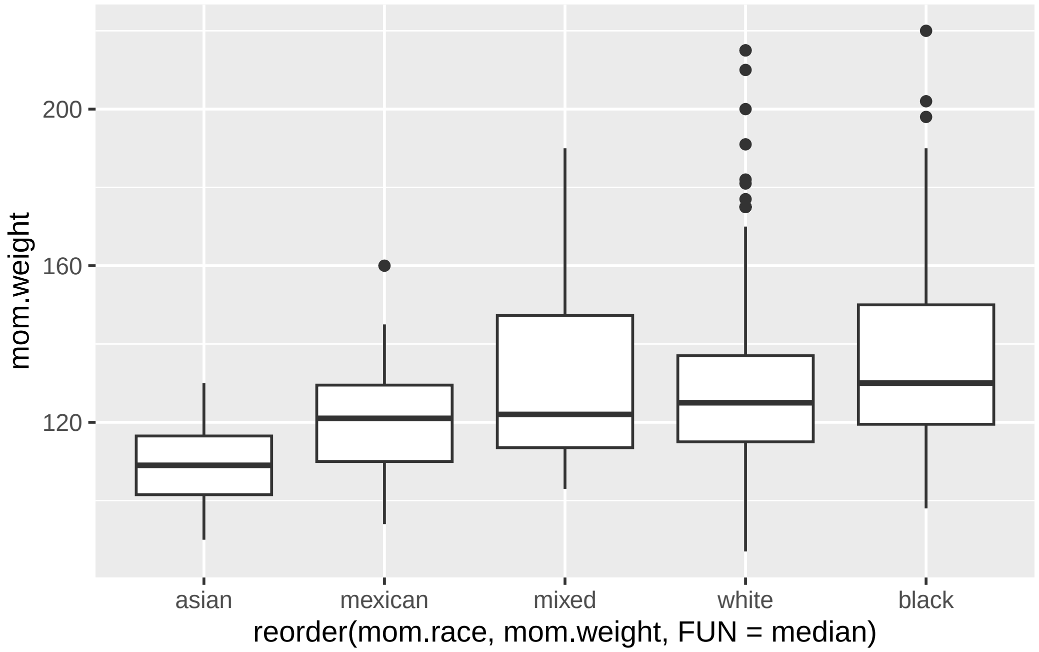
4.11 Mom's smoking and the family's income

```
ggplot(smoking) +
  geom_bar(aes(x = factor(income), fill = smoke), position = "fill")
```



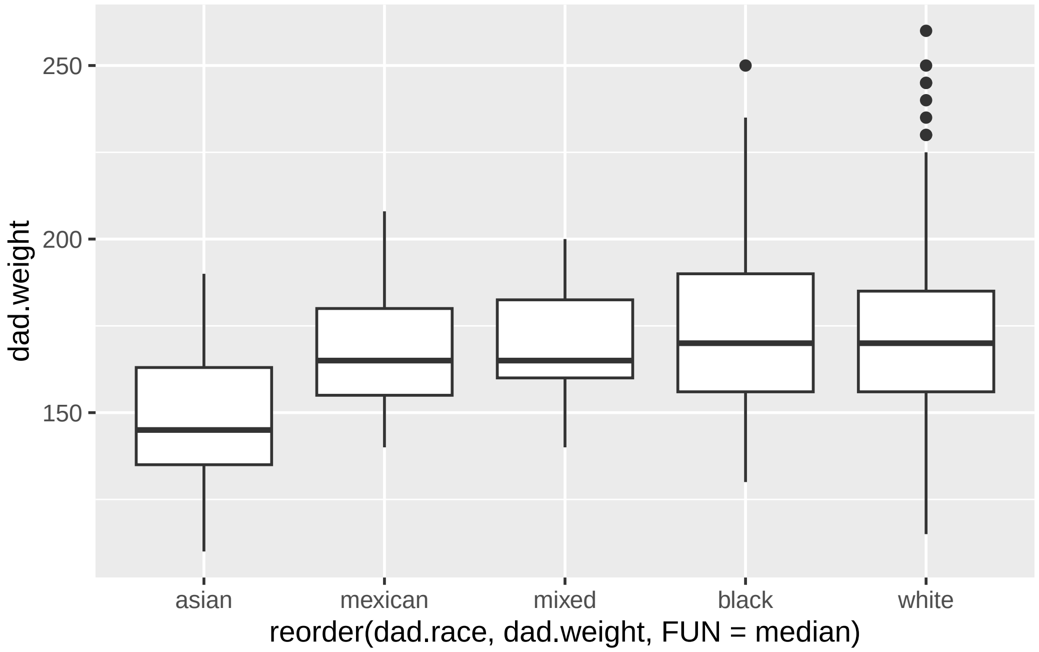
4.12 Mom's race and mom's weight

```r
ggplot(smoking) +
  geom_boxplot(aes(x = reorder(mom.race, mom.weight, FUN = median), y = mom.weight))
```



4.13 Dad's race and dad's weight

```
ggplot(smoking) +
  geom_boxplot(aes(x = reorder(dad.race, dad.weight, FUN = median), y =
   ↪  dad.weight))
```


250
200
150
dad.weight
asian
mexican
mixed
black
white
reorder(dad.race, dad.weight, FUN = median)

5 The End

The End.
gettyimages
Credit: Nora Carol Photography
656357474