

# Data Visualization

## Part II

### Table of contents

<b>1</b>	<b>Agenda</b>	<b>2</b>
<b>2</b>	<b>Setting up</b>	<b>2</b>
<b>3</b>	<b>Titanic Survival</b>	<b>3</b>
3.1	Load and explore the dataset . . . . .	3
3.2	Survival by class . . . . .	5
3.3	Survival by sex . . . . .	6
3.4	Survival by age . . . . .	8
3.5	Survival by age & sex . . . . .	10
3.6	Survival by class & age . . . . .	11
<b>4</b>	<b>Smoking and Pregnancy</b>	<b>13</b>
4.1	Load and explore the dataset . . . . .	13
4.2	Mom's smoking and baby's weight . . . . .	15
4.3	Mom's smoking and baby's weight with reordered x-axis . . . . .	16
4.4	Mom's race and baby's weight . . . . .	16
4.5	Dad's race and baby's weight . . . . .	17
4.6	Mom's race and baby's weight and dad's race . . . . .	18
4.7	Mom's height and moms's weight . . . . .	19
4.8	Dad's height and dad's weight . . . . .	21
4.9	Mom's weight and dad's weight . . . . .	22
4.10	Mom's smoking and mom's education . . . . .	23
4.11	Mom's smoking and the family's income . . . . .	25
4.12	Mom's race and mom's weight . . . . .	25
4.13	Dad's race and dad's weight . . . . .	26
<b>5</b>	<b>The End</b>	<b>28</b>

# 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:

```
1 if (!require(ggplot2)) {  
2   install.packages("ggplot2") # install if not already installed  
3 }
```

Loading required package: `ggplot2`

```
1 library (ggplot2)
```

### 3 Titanic Survival



#### 3.1 Load and explore the dataset

```
1 titanic = read.csv
  ↪ ("https://raw.githubusercontent.com/ahmedmoustafa/datasets/main/titanic/titanic.csv")
2 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

name	survived	sex	age	class
Anderson, Mr. Harry	yes	male	48.0000	1st

```

1 titanic$survived = factor(titanic$survived)
2 titanic$sex = factor(titanic$sex)
3 titanic$class = factor(titanic$class, levels = c("1st", "2nd", "3rd"))
4 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

```

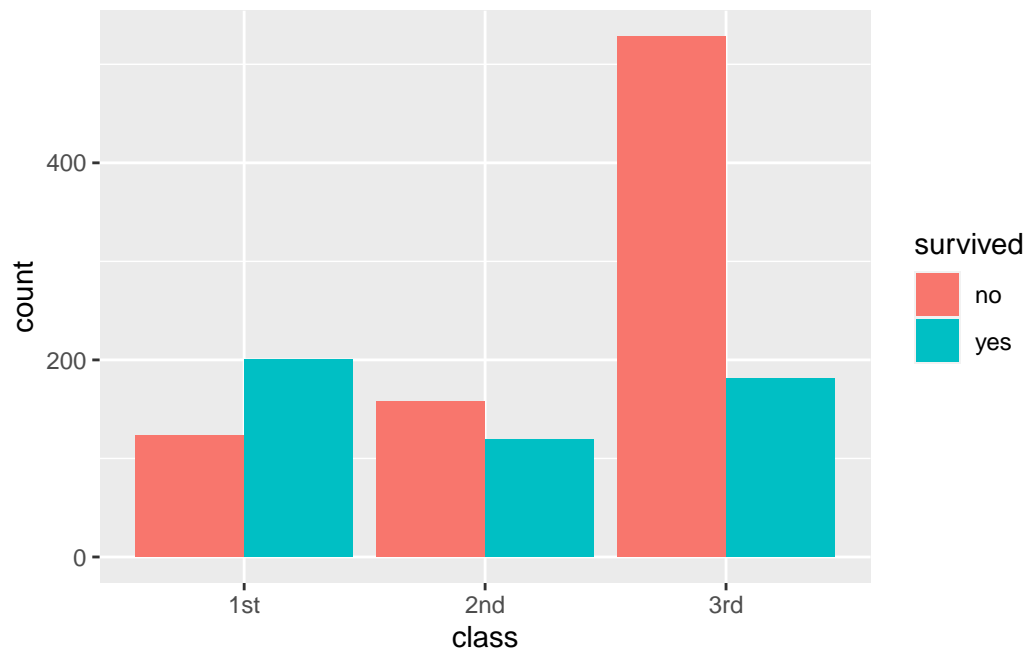
1 summary(titanic)

```

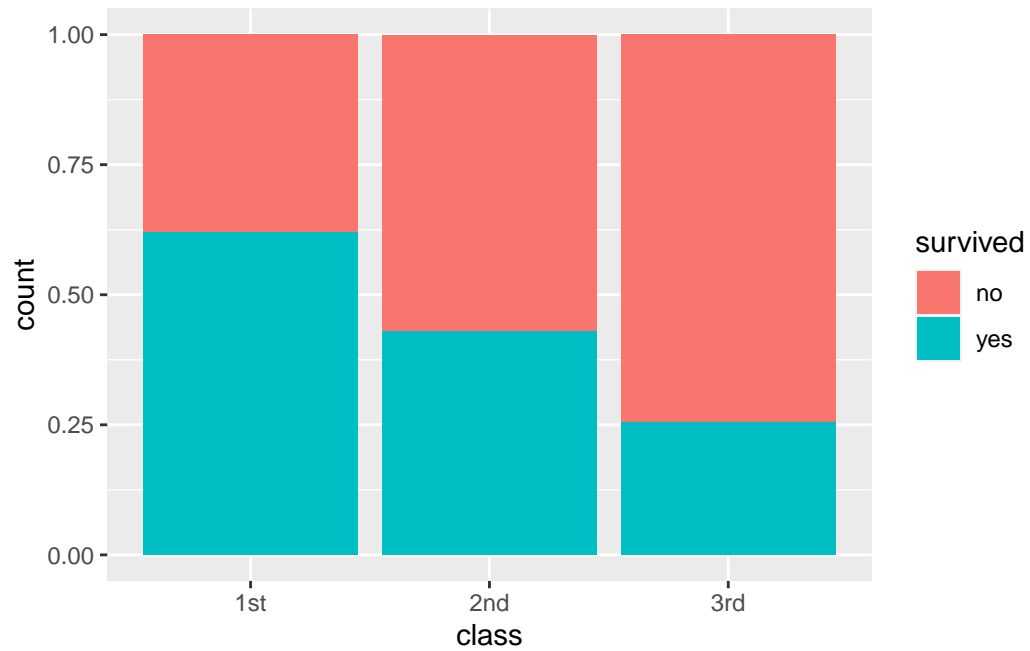
name	survived	sex	age	class
Length:1309	no :809	female:466	Min. : 0.1667	1st:323
Class :character	yes:500	male :843	1st Qu.:21.0000	2nd:277
Mode :character	NA	NA	Median :28.0000	3rd:709
NA	NA	NA	Mean :29.8811	NA
NA	NA	NA	3rd Qu.:39.0000	NA
NA	NA	NA	Max. :80.0000	NA
NA	NA	NA	NA's :263	NA

### 3.2 Survival by class

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

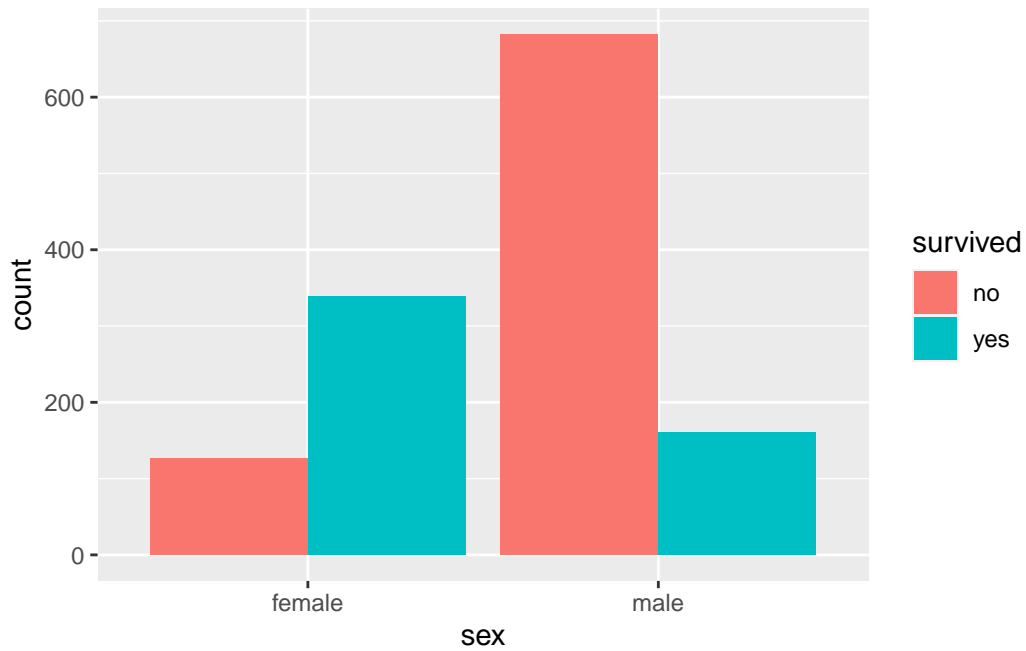


```
1 ggplot(titanic) +  
2   geom_bar(aes(x = class, fill = survived), position = "fill")
```

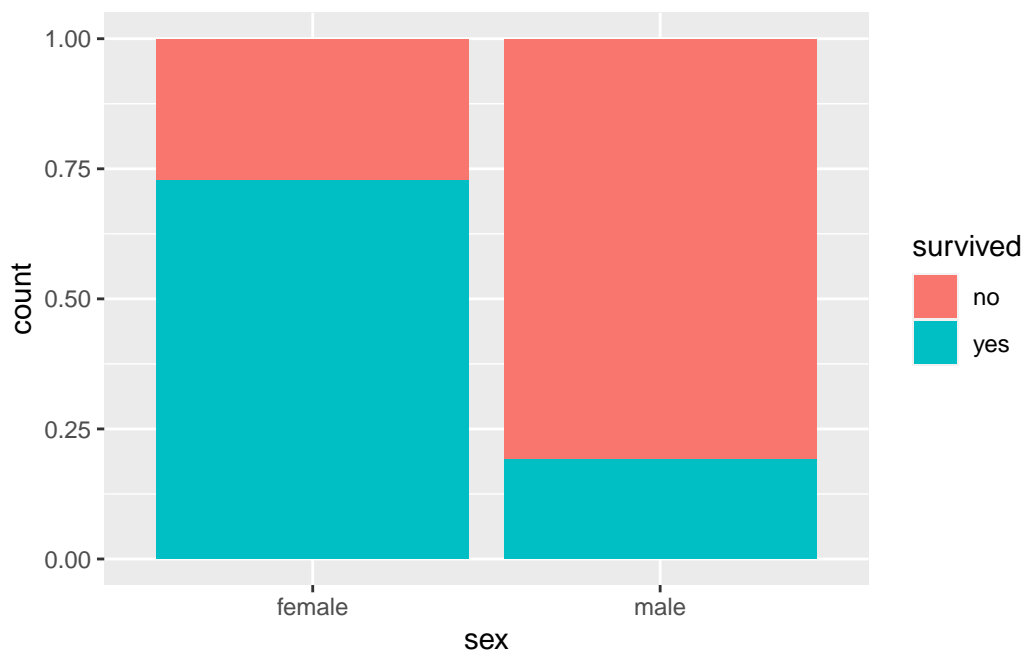


### 3.3 Survival by sex

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



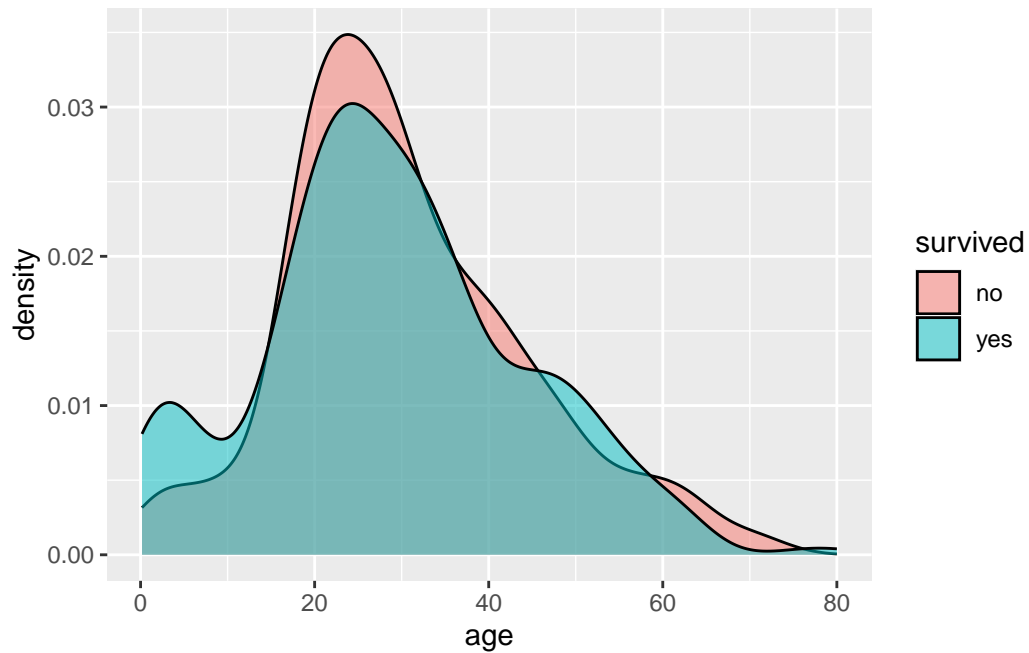
```
1 ggplot(titanic) +  
2   geom_bar(aes(x = sex, fill = survived), position = "fill")
```



### 3.4 Survival by age

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

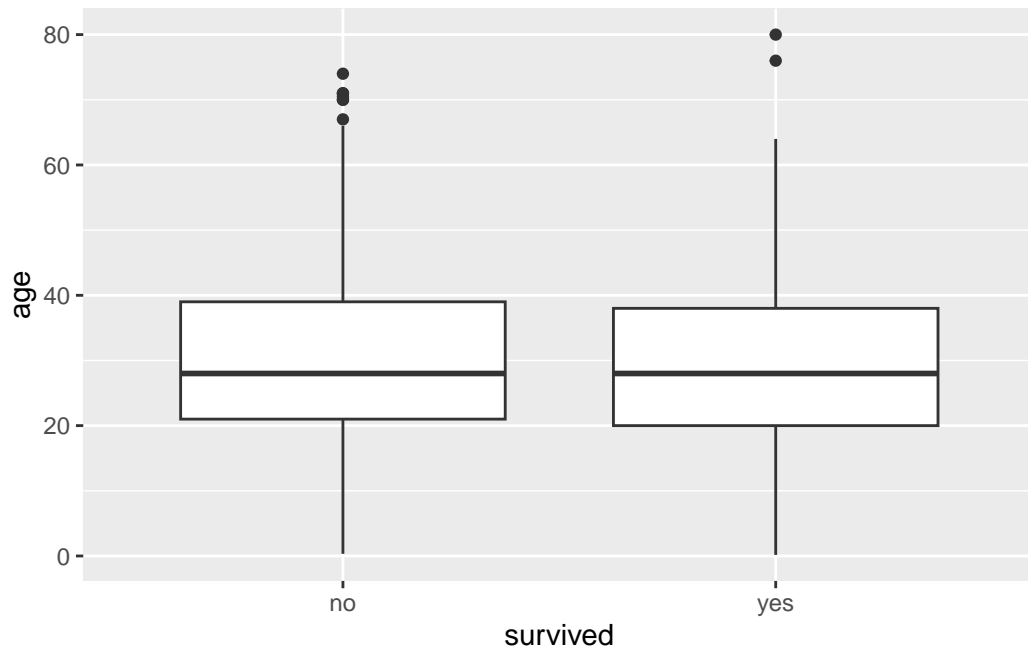
Warning: Removed 263 rows containing non-finite values (`stat\_density()`).



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

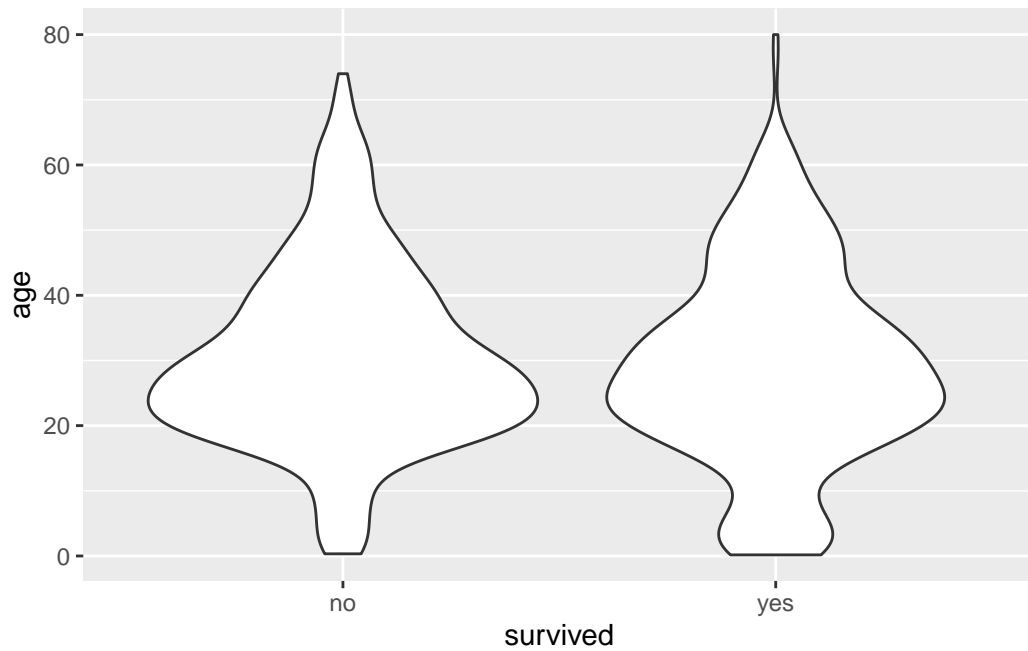
Warning: Removed 263 rows containing non-finite values (`stat\_boxplot()`).





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

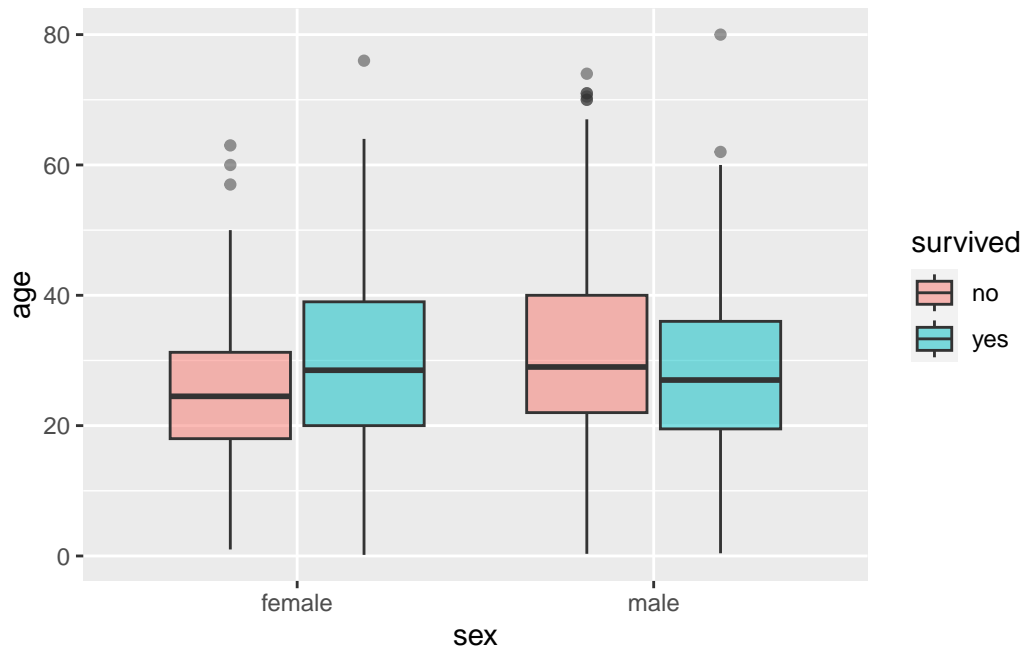
Warning: Removed 263 rows containing non-finite values (`stat\_ydensity()`).



### 3.5 Survival by age & sex

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

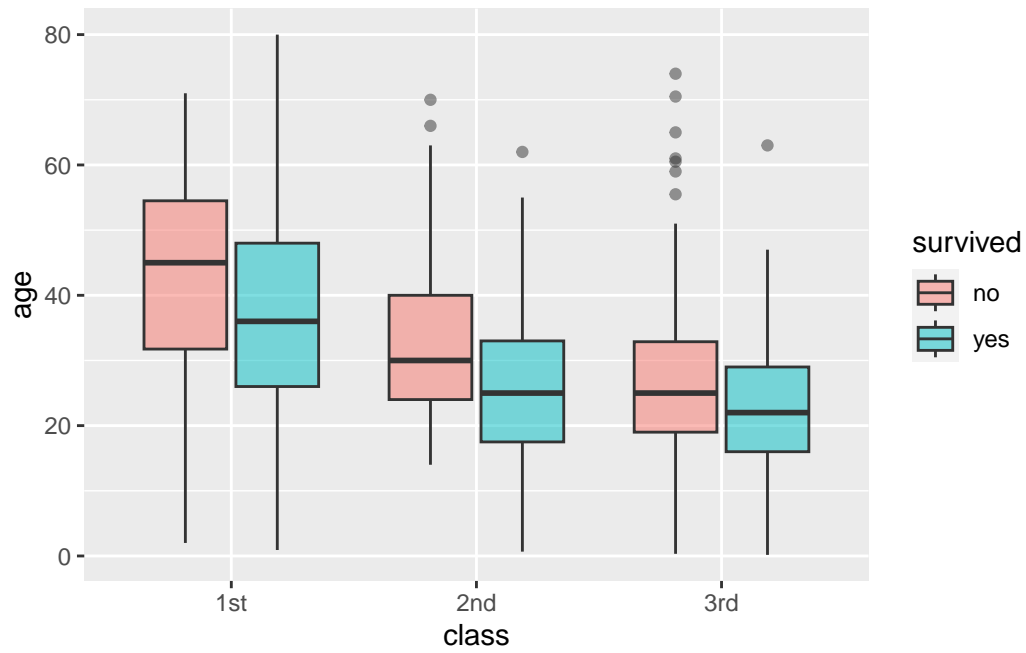
Warning: Removed 263 rows containing non-finite values (`stat\_boxplot()`).



### 3.6 Survival by class & age

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

Warning: Removed 263 rows containing non-finite values (`stat\_boxplot()`).



## 4 Smoking and Pregnancy



### 4.1 Load and explore the dataset

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

id	date	gestaw	weigh	parity	nom	mm	age	med	dm	height	black	age	et	height	weight	at	smoke	quit	time
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

---

```

1 smoking$parity = factor(smoking$parity)
2 smoking$mom.race = factor(smoking$mom.race)
3 smoking$mom.edu = factor(smoking$mom.edu)
4 smoking$dad.race = factor(smoking$dad.race)
5 smoking$dad.edu = factor(smoking$dad.edu)
6 smoking$marital = factor(smoking$marital)
7 smoking$income = factor(smoking$income)
8 smoking$smoke = factor(smoking$smoke)
9 smoking$quit.time = factor(smoking$quit.time)
10 smoking$cigs = factor(smoking$cigs)
11 head(smoking)

```

---

id	date	gestaw	weigh	parity	mom	racemom	agemom	medmom	height	height	dad	age	height	weight	marital	smoke	quit.time	cigs
15	141284	120	1	asian	27	5	62	100	asian	31	5	65	110	1	1	never	0	0
20	149282	113	2	white	33	5	64	135	white	38	5	70	148	1	4	never	0	0
100	167286	136	4	white	25	2	62	93	white	28	2	64	130	1	4	until_pregnancy		
129	156245	132	2	black	23	1	65	140	black	23	4	71	192	1	2	never	0	0
142	140289	120	3	white	25	4	62	125	white	26	1	70	180	0	2	never	0	0
171	159282	144	4	white	32	2	64	124	white	36	1	74	185	1	2	now	1	1

---

```

1 summary(smoking)

```

---

id	date	gestaw	weigh	parity	mom	racemom	agemom	medmom	height	height	dad	age	height	weight	marital	smoke	quit.time	cigs
Min.	Min.	Min.	Min.	1	asian	Min.0:	Min.	Min.	asian	Min.2	Min.	Min.0:	2	never	0	0		
:	:	1350	148.0	:152:	:	15.00	:54.0:	:	:	18.00	189	:60.00	110.0	:101:	282	:282:	282	
15		55.0	24				87.0	25										
1st	1st	1st	1st	0	black	1st	1st	1st	black	1st	5	1st	1st	1:600	now	1	5	
Qu.:	Qu.:	Qu.:	Qu.:	Qu.:	Qu.:	Qu.:	Qu.:	Qu.:	Qu.:	Qu.:	Qu.:	Qu.:	Qu.:	Qu.:	Qu.:	Qu.:	Qu.:	
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
Median	Median	Median	Median	Median	mexican	Median	Median	Median	mexican	Median	Median	Median	Median	Median	3	once	2	1
:6907	:1574	:280.0	:120.0	:117	18	:27.00	:64.0:	:125.07	:30.00	131	:71.00	170.0	:100:		52	75		
															60			
Mean	Mean	Mean	Mean	3:	mixed	Mean	Mean	Mean	mixed	Mean	Mean	Mean	Mean	3:	7:	until_pregnancy:		
:6090	:1559	:278.8	:119.82	:	:	:27.53	:64.1:	:128.9	:30.49	1	:70.27	170.0	90	52	15	73		
					14					15								

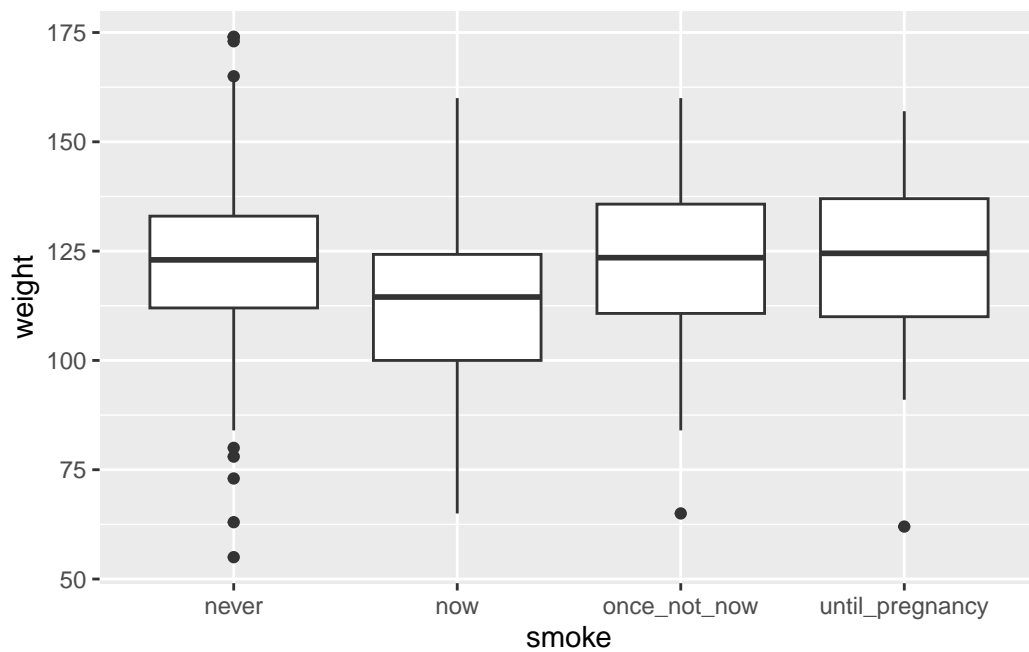
id	date	gesta	weight	parity	mom_race	mom_age	med	height	height	age	dad_race	dad_height	weight	smoke	quit	cigs
3rd	3rd	3rd	3rd	4:	white	3rd	4:15	2	3rd	3rd	white	3rd	3:	3rd	3rd	5: 4: NA 4: 3:
Qu.:	792.0	650.2	38.0	131.0	423	Qu.:	31.00	Qu.:	60.0	140.0	Qu.:	30.00	Qu.:	72.0	185.0	71 13 40
Max	Max	Max	Max	5:	NA	Max	5:10	Max	Max	NA	Max	0:	Max	Max	NA	5: NA 7: 6:
:926	317	143	38.0	174.0		:43.00	:72.0	:220.0	:53.00	2	:78.00	260.0	65			13 17
NA	NA	NA	NA	(Other):	NA	NA	NA	NA	NA	NA	(Other):	NA	NA	NA	(Other):	(Other):
				34							2			83		19 22

## 4.2 Mom's smoking and baby's weight

```

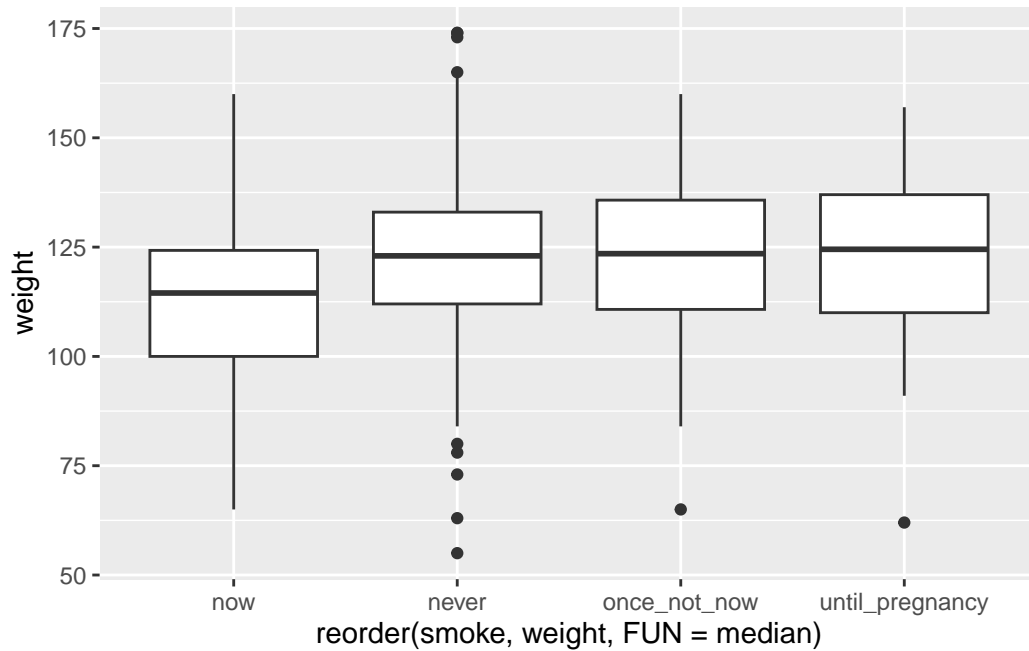
1 ggplot(smoking) +
2   geom_boxplot(aes(x = smoke, y = weight))

```



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

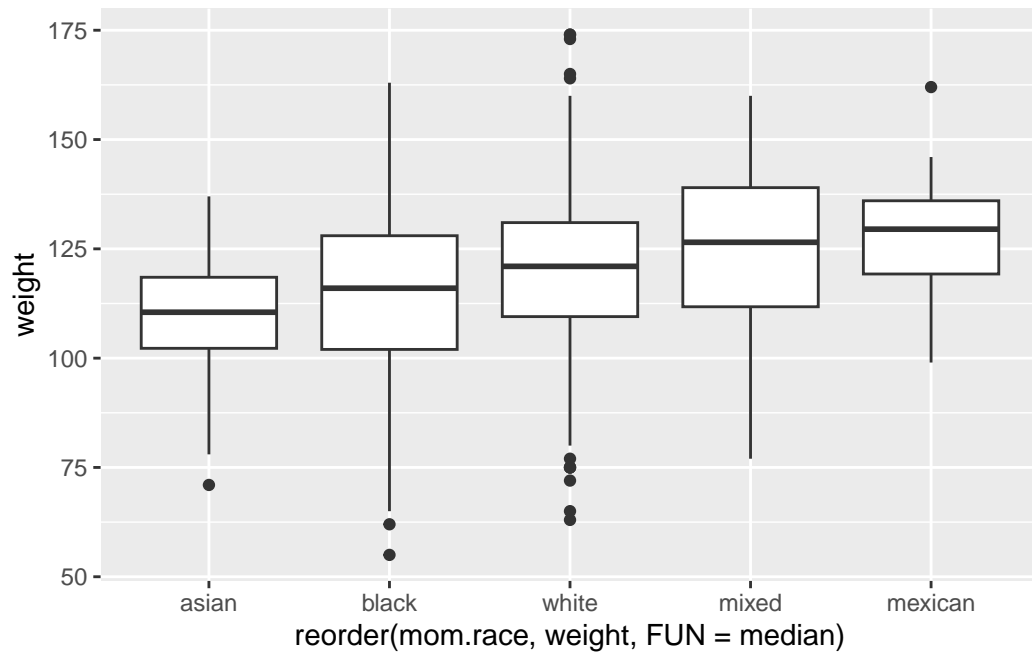
```
1 ggplot(smoking) +  
2   geom_boxplot(aes(x = reorder(smoke, weight, FUN = median), y =  
   ↪ weight))
```



### 4.4 Mom's race and baby's weight

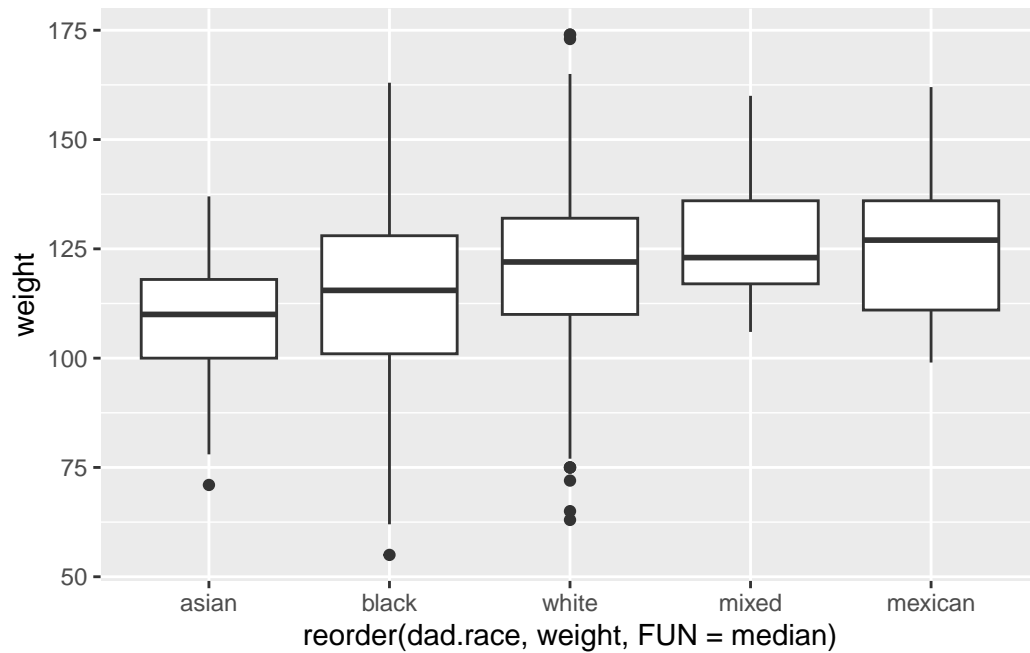
```
1 ggplot(smoking) +  
2   geom_boxplot(aes(x = reorder(mom.race, weight, FUN = median), y =  
   ↪ weight))
```





#### 4.5 Dad's race and baby's weight

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

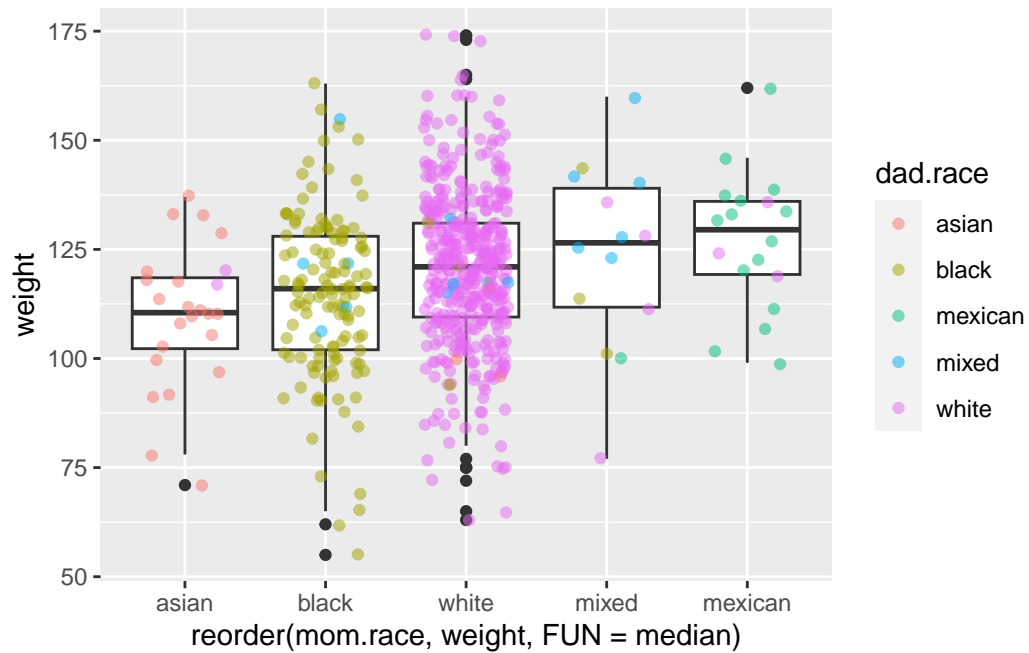


#### 4.6 Mom's race and baby's weight and dad's race

```

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

```



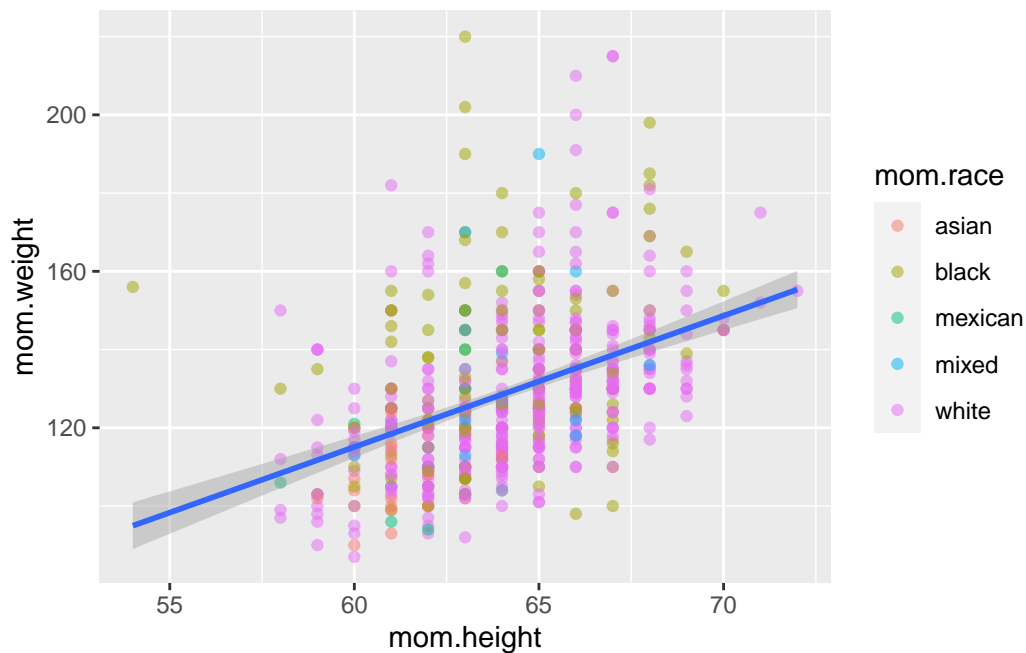
#### 4.7 Mom's height and moms's weight

```

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

```

`geom\_smooth()` using formula = 'y ~ x'



```
1 model = lm (data = smoking, formula = mom.weight ~ mom.height)
2 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

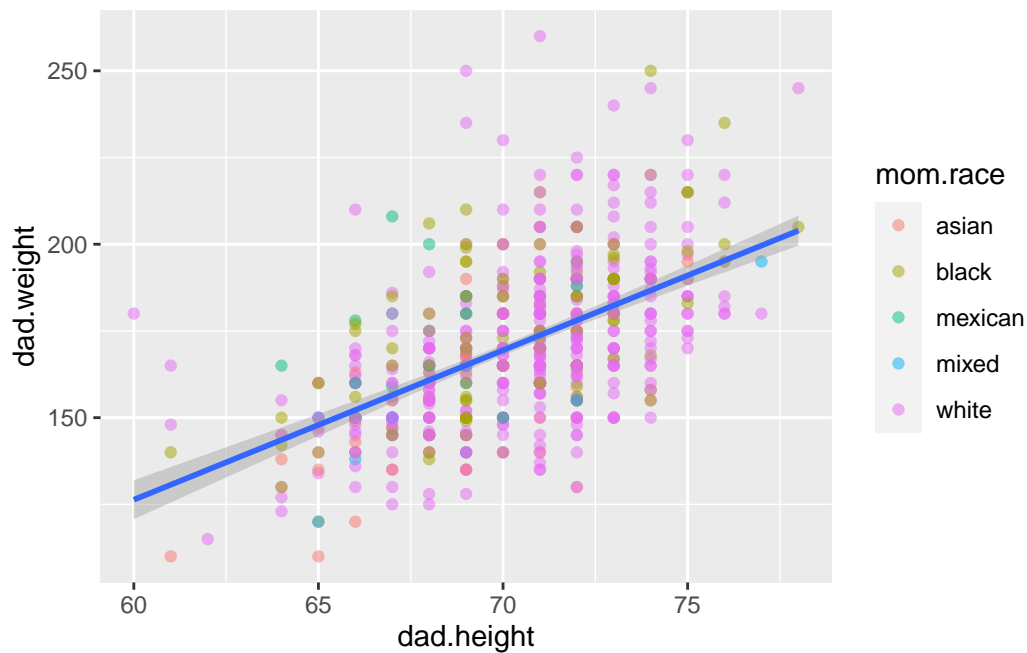
Multiple R-squared: 0.1793, Adjusted R-squared: 0.178

F-statistic: 132.8 on 1 and 608 DF, p-value: < 2.2e-16

## 4.8 Dad's height and dad's weight

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

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



```
1 model = lm (data = smoking, formula = dad.weight ~ dad.height)  
2 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

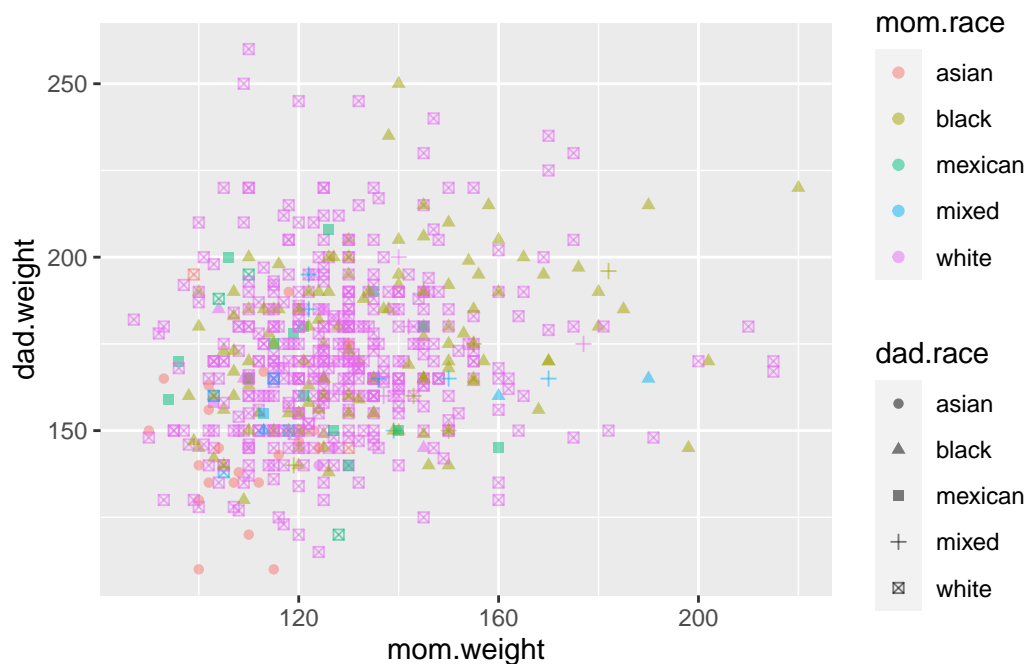
Residual standard error: 19.02 on 608 degrees of freedom

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

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



```
1 model = lm (data = smoking, formula = dad.weight ~ mom.weight)  
2 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

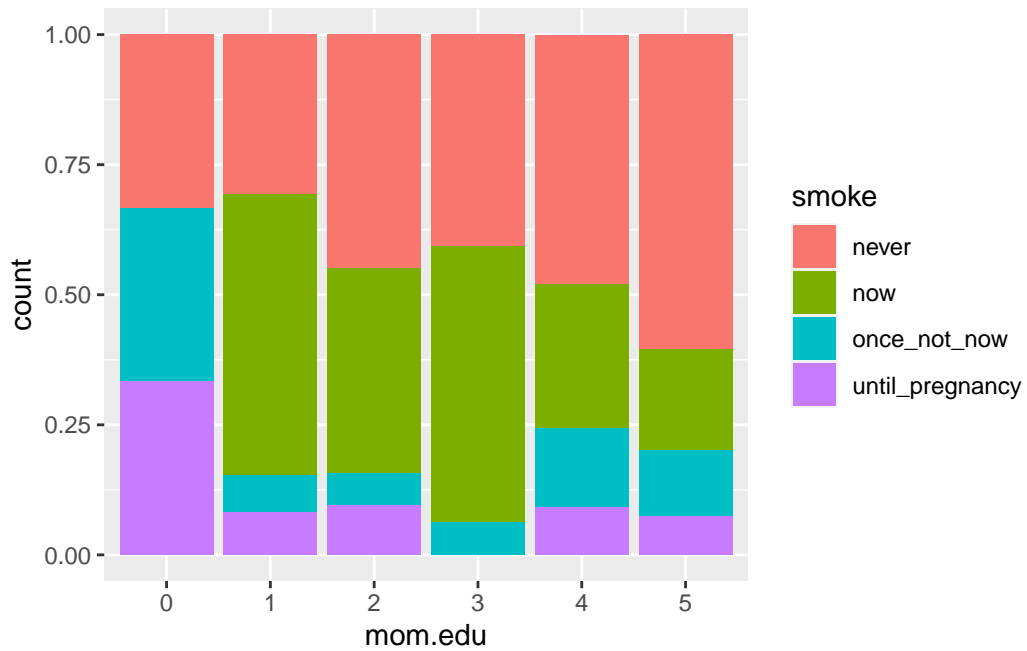
```

#### 4.10 Mom's smoking and mom's education

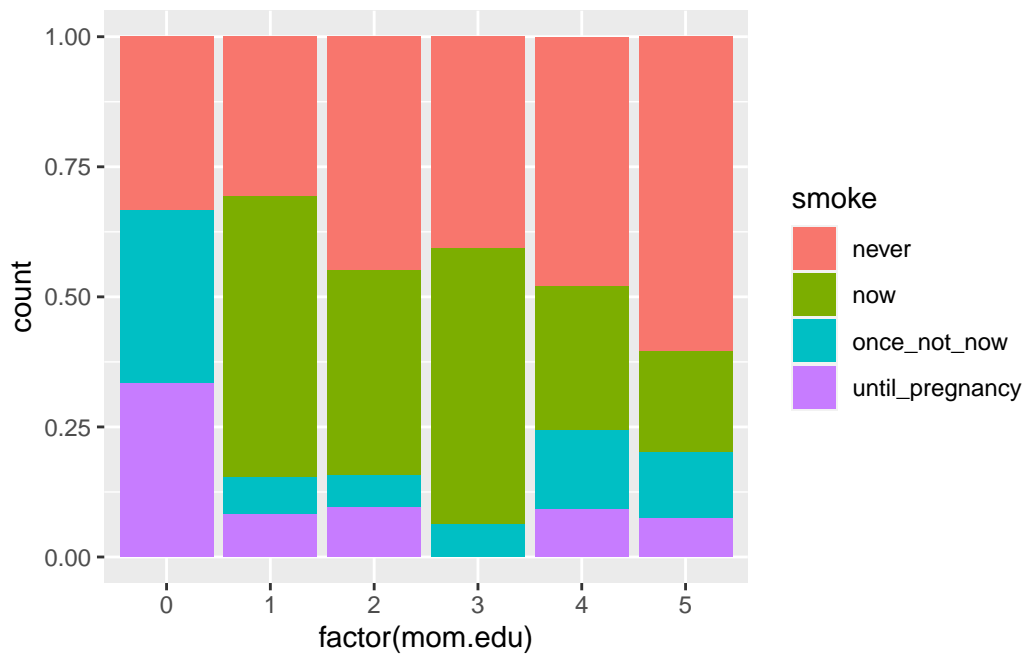
```

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

```



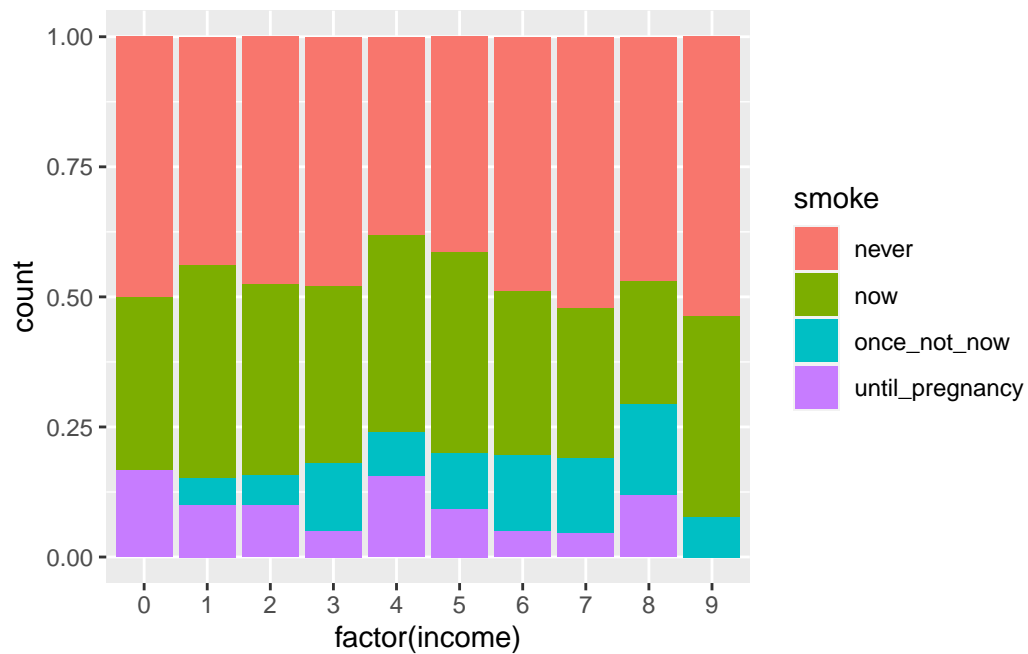
```
1 ggplot(smoking) +
2   geom_bar(aes(x = factor(mom.edu), fill = smoke), position = "fill")
```





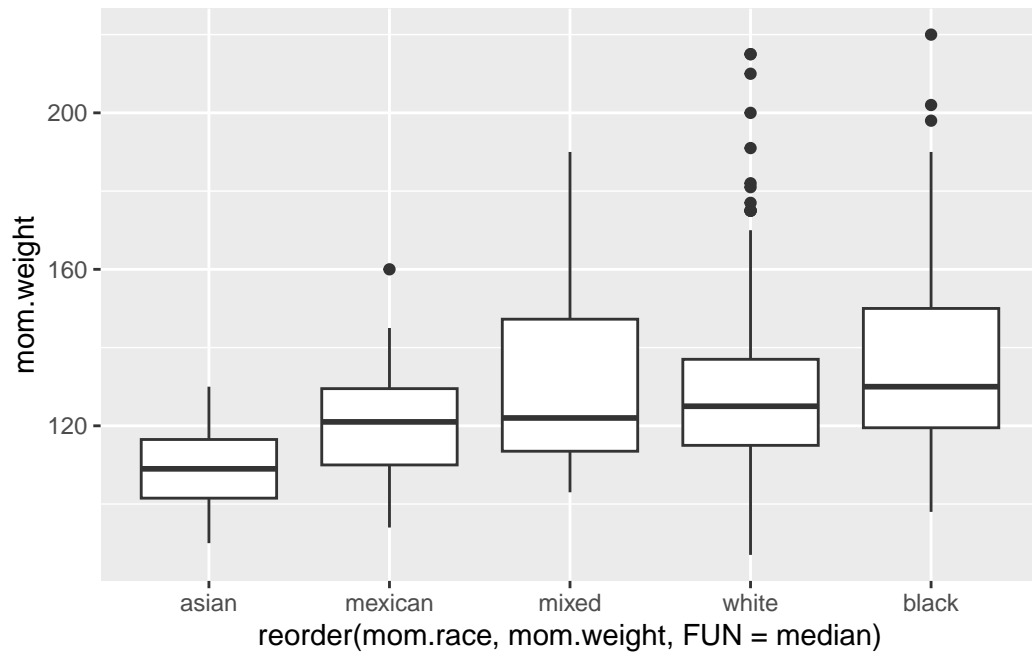
### 4.11 Mom's smoking and the family's income

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



### 4.12 Mom's race and mom's weight

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

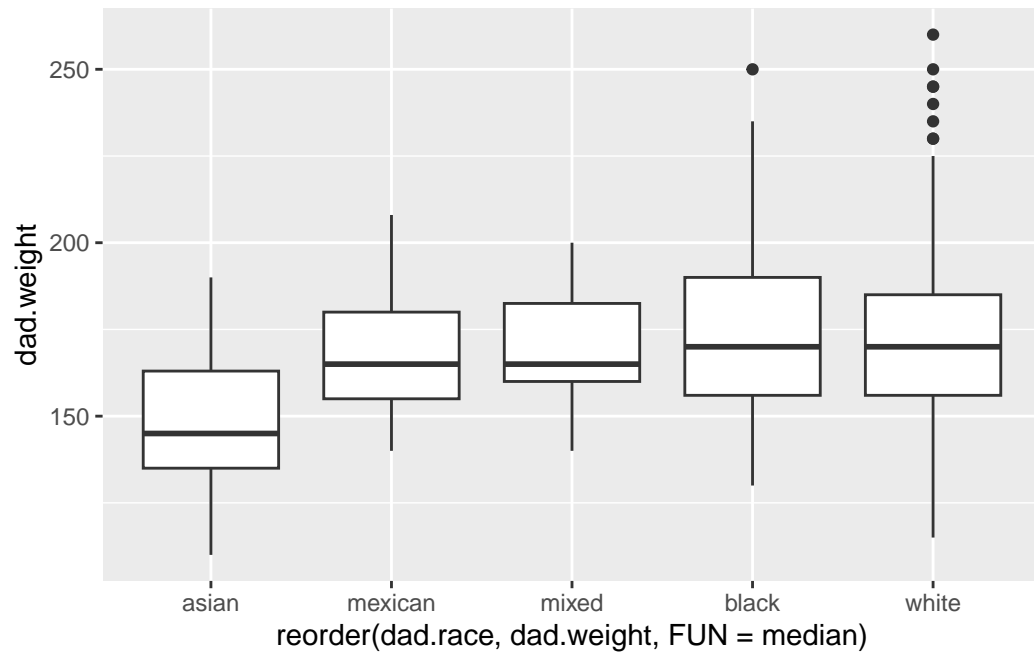


### 4.13 Dad's race and dad's weight

```

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

```



## 5 The End

