# STAT 224 Lecture/Activities on Friday, 10/01

Yibi Huang

# The NC Births Data

The NC Births data came from a random sample of 1000 birth records released by the State of North Carolina in 2004. The variables include:

- `weight`: weight of the baby at birth in pounds.
- `gender`: gender of the baby, female or male.
- `habit` : status of the mother as a nonsmoker or a smoker.
- `marital`: whether mother is married or not married at birth.
- `whitemom`: whether mom is white or not white.
- `fage`: father's age in years.
- `mage`: mother's age in years.
- `gained`: weight gained by mother during pregnancy in pounds.

# Loading Data

You can download the data at

https://www.openintro.org/stat/data/csv/ncbirths.csv

Please save it at **the same folder** as this RMD file.
Then you can easily **change the working directory** by

[Session] – [Set Working Directory] – [To Source File Location]

Run the command below to load the data

```
nc = read.csv("ncbirths.csv")
```

# Data Summary by Group

```
library(mosaic)
favstats(weight ~ gender + habit, data=nc)
      gender.habit  min   Q1 median   Q3   max  mean    sd   n missing
1 female.nonsmoker 1.00 6.31   7.13 7.81 11.63 6.939 1.511 445       0
2   male.nonsmoker 1.38 6.69   7.50 8.38 11.75 7.357 1.499 428       0
3    female.smoker 2.19 6.00   6.88 7.44  8.38 6.675 1.078  57       0
4      male.smoker 1.69 6.25   7.31 8.13  9.19 6.956 1.593  69       0
```

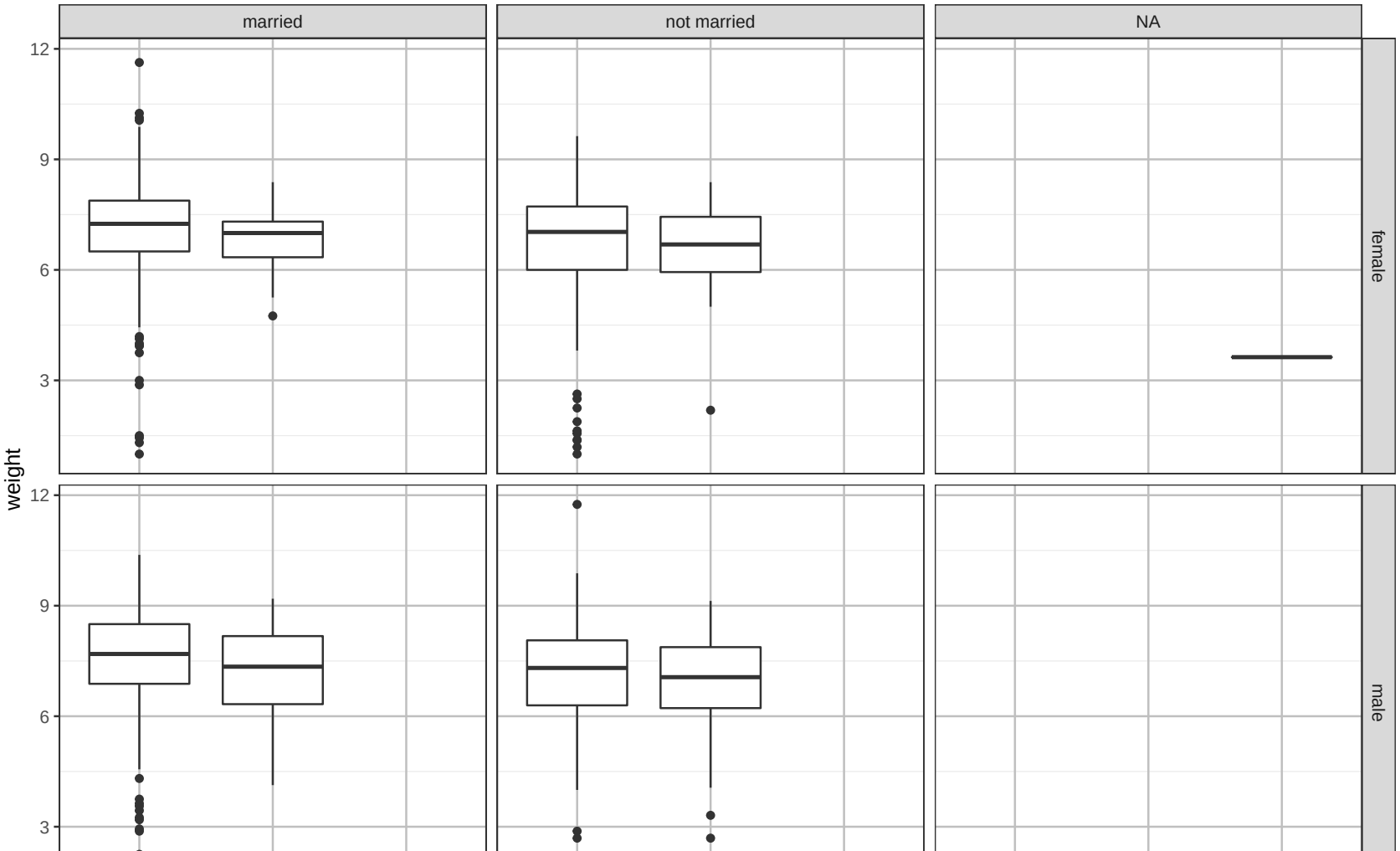
- Describe the effect of `gender` on `weight` controlling for mom's smoking status (`habit`)
- Describe the effect of the mom's smoking `habit` on `weight` controlling for baby's gender

# On Your Own

**Q1**: Find appropriate data summary and use it to describe the effect of mother's marital status (`marital`) on `weight` controlling for the `gender` of baby

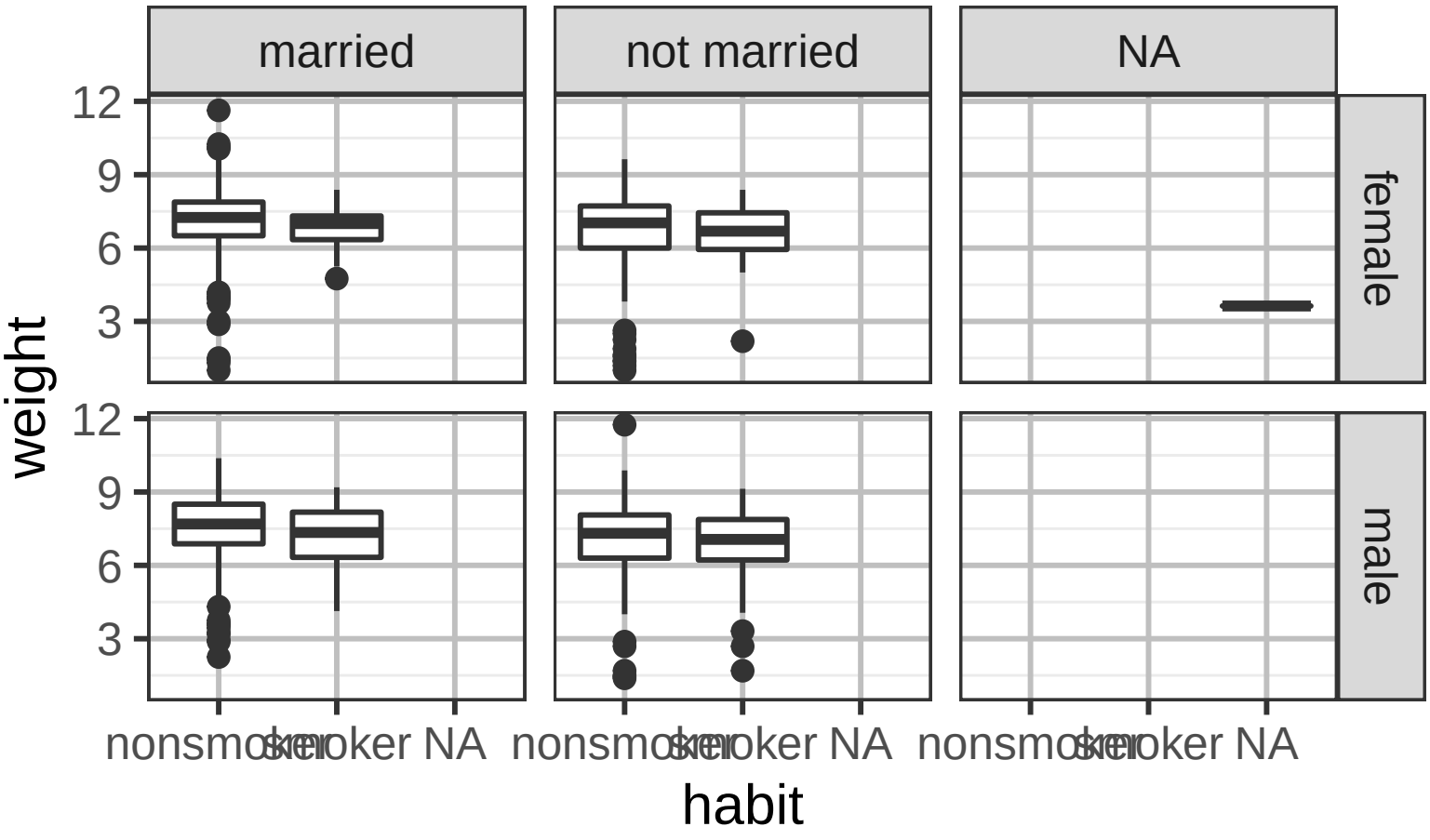
# Side-By-Side Boxplots w/ Facet

```
ggplot(nc, aes(x=habit, y=weight)) +
  geom_boxplot() +
  facet_grid(gender ~ marital)
```



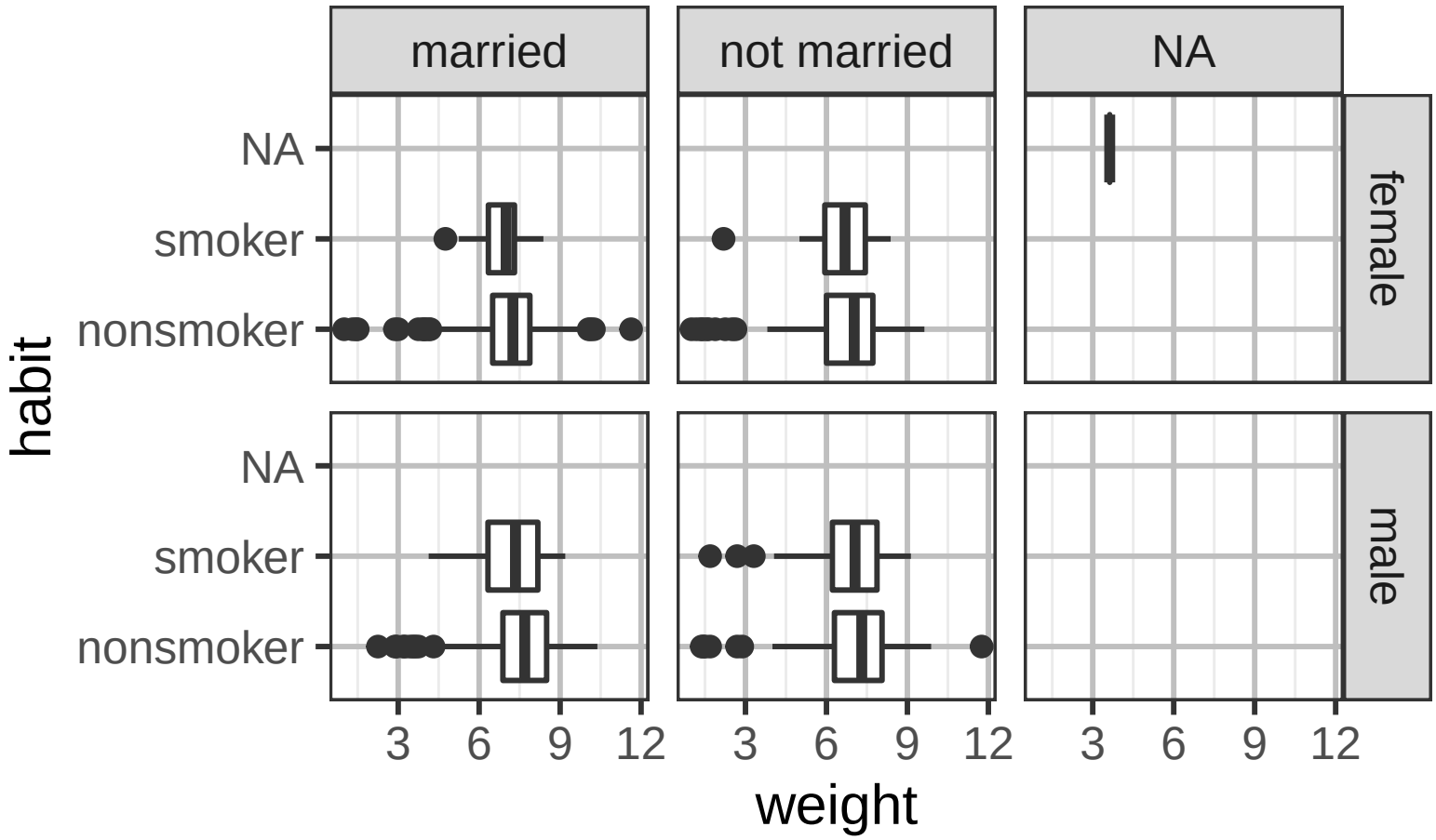
## Same Plot Resized:

```
ggplot(nc, aes(x=habit, y=weight)) +
  geom_boxplot() +
  facet_grid(gender ~ marital)
```



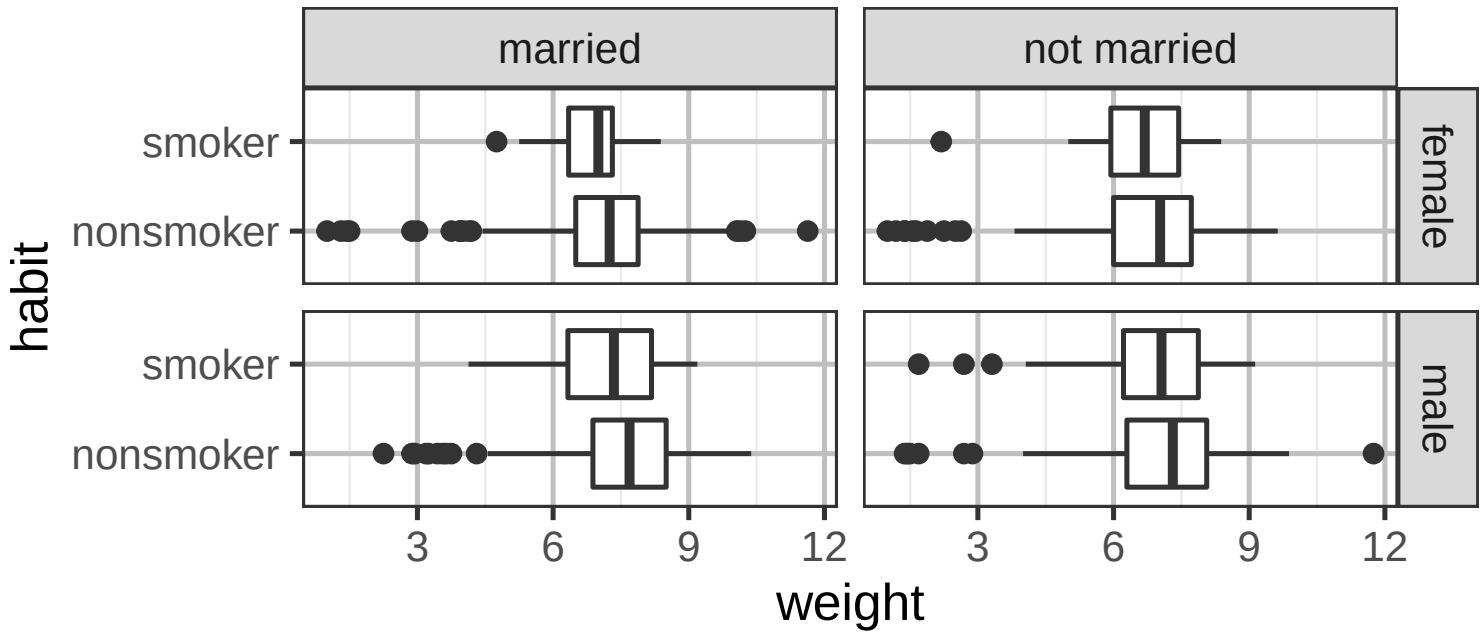
# Same Boxplots but Horizontal

```
ggplot(nc, aes(y=habit, x=weight)) +
  geom_boxplot() +
  facet_grid(gender ~ marital)
```



# Same Plot w/ Missing Values Removed

```
ggplot(subset(nc, !is.na(habit)), aes(y=habit, x=weight)) +
  geom_boxplot() +
  facet_grid(gender ~ marital)
```




Based on the plot, describe the effect of smoking on baby's birth weight, controlling for `gender` and mom's marital status.

Answer #1: For babies of the same gender and same marital status of their moms, babies of smoking moms have a lower median birth weight than those of nonsmoking moms.

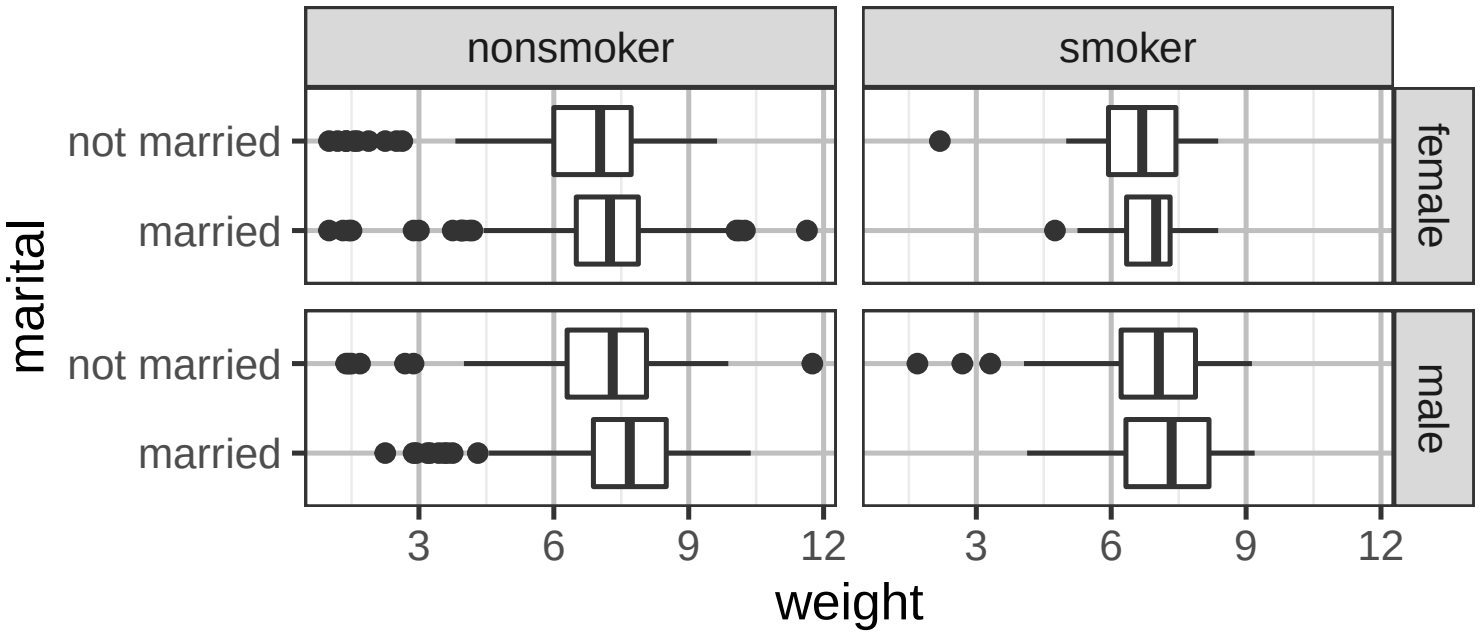
Answer #1: For babies of the same gender and same marital status of their moms, babies of smoking moms have a lower median birth weight than those of nonsmoking moms.

Answer #2: Babies of smoking moms have a lower median birth weight than those of nonsmoking moms, controlling/adjusted for the gender of the baby and the mom’s marital status

# On Your Own

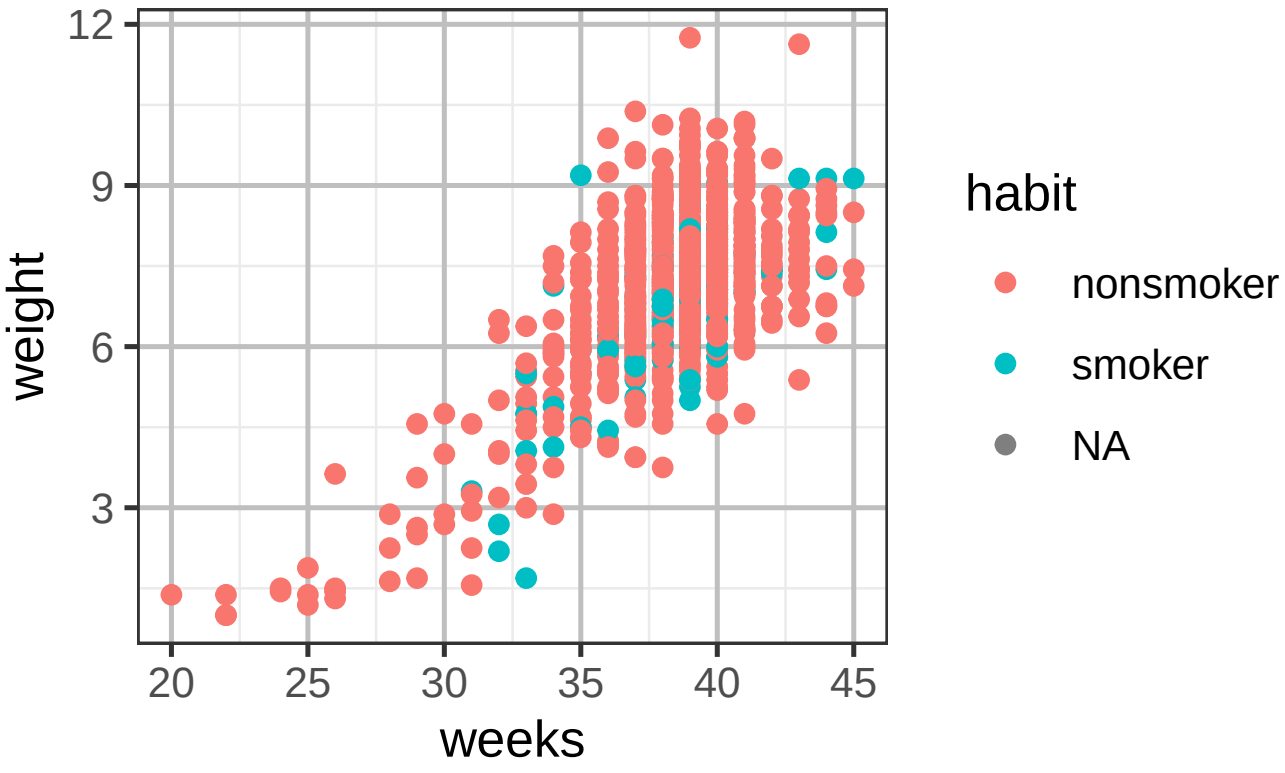
**Q2**: Find appropriate boxplots of the data and use them to describe the effect of mother's marital status (`marital`) on `weight` controlling for the `gender` of baby and mother's smoking `habit`.

```
ggplot(subset(nc, !is.na(habit)), aes(y=marital, x=weight)) +
  geom_boxplot() +
  facet_grid(gender ~ habit)
```



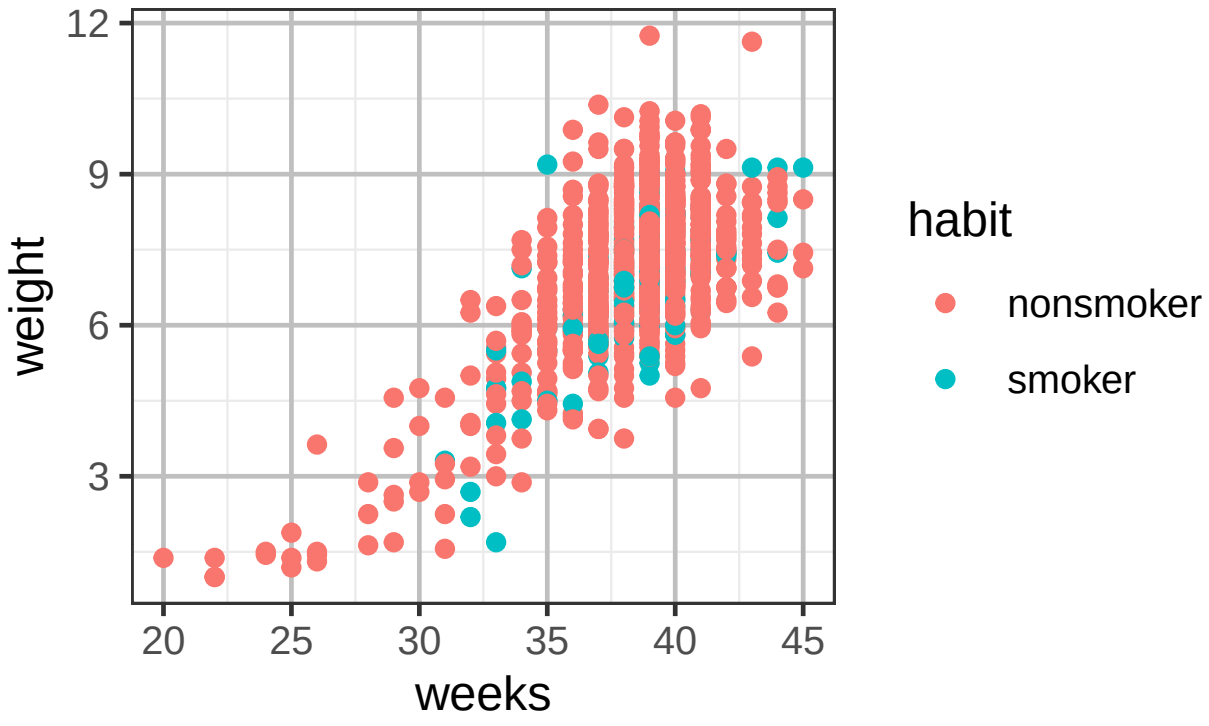
# Coded Scatter Plot

```
ggplot(nc, aes(x=weeks, y=weight, color=habit)) +
  geom_point()
Warning: Removed 2 rows containing missing values (geom_point).
```



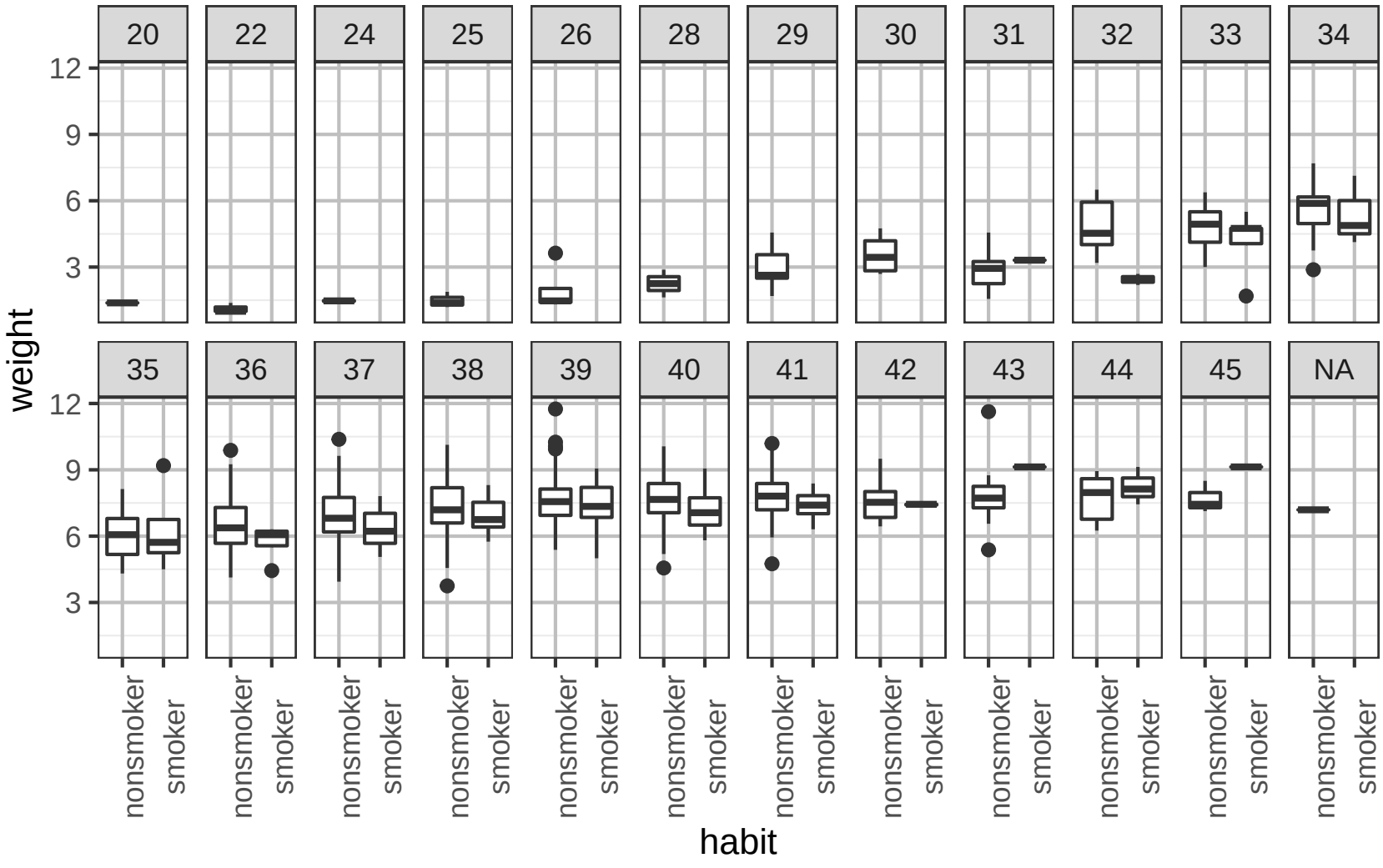
Add `warning=FALSE` inside `{r}` to get rid of the warning message in the knitted output. The missing value is also removed.

```
nc = subset(nc, !is.na(habit))
ggplot(nc, aes(x=weeks, y=weight, color=habit)) + geom_point()
```




Hard to see whether mom's smoking `habit` has any effect on birth weights after accounting for the length of pregnancy (`weeks`).

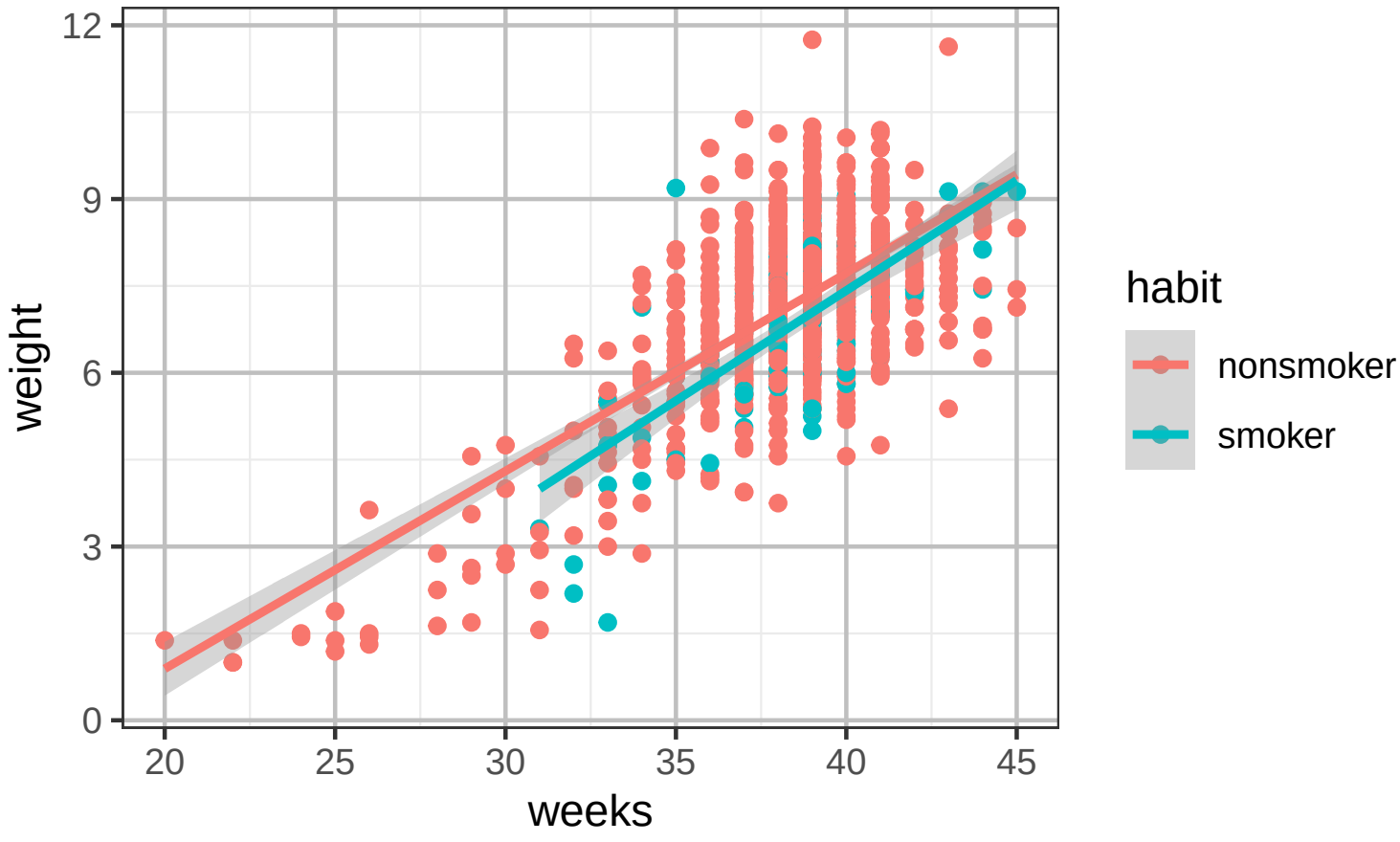
```
ggplot(nc, aes(x=habit, y=weight)) +
  geom_boxplot() +
  facet_wrap(~weeks, nrow=2) +
  theme(axis.text.x = element_text(angle = 90))
```



- can ignore week 20-30 where there were no smokers
- For 32 to 42 weeks of pregnancies, the median birth weight were slightly lower for those born to smoking moms, comparing babies w/ the same weeks of pregnancy.
- For 31, 43, 44, 45 weeks, smoking group has higher median birth weights, but those groups had only a few observations

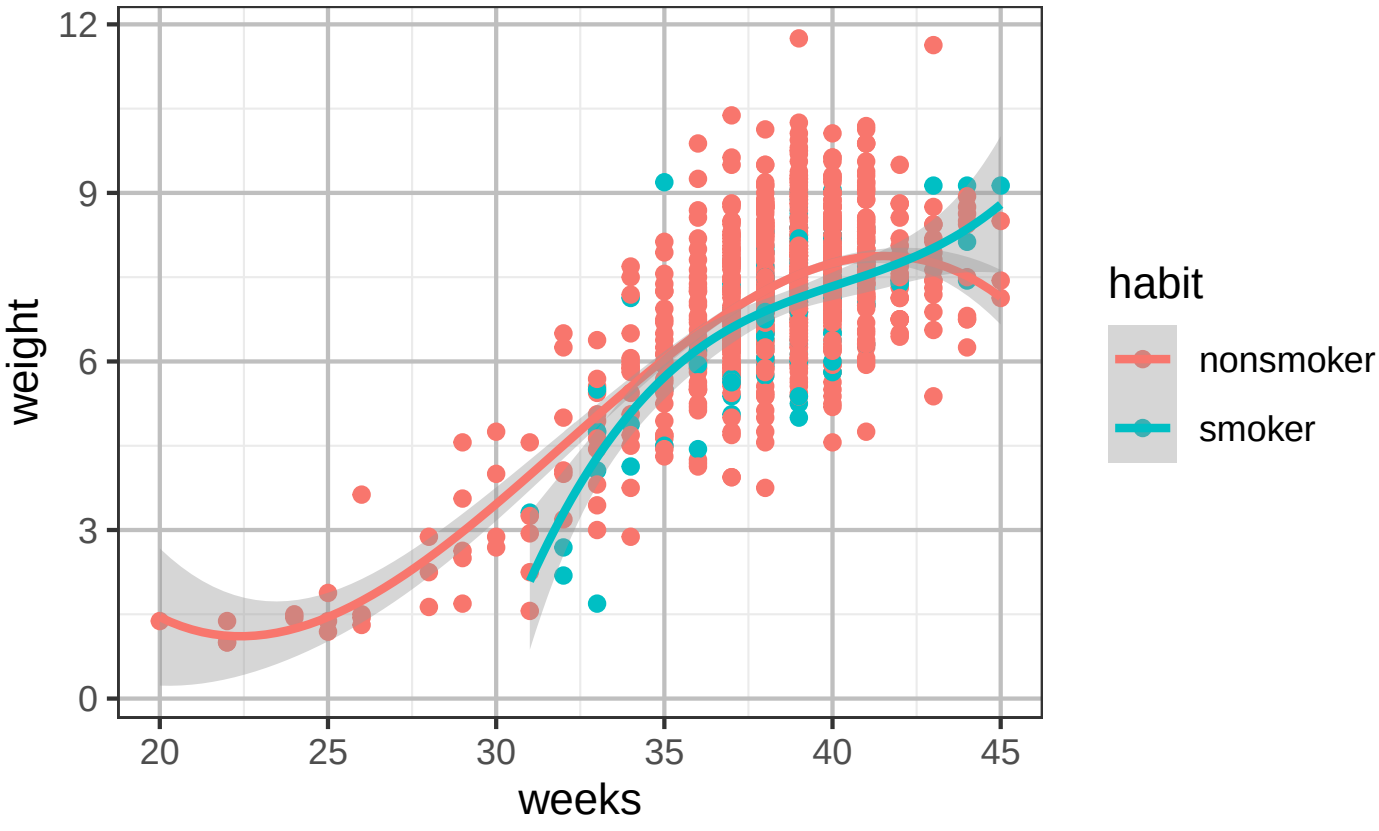
```
xtabs(~habit + weeks, data=nc)
            weeks
habit         20  22  24  25  26  28  29  30  31  32  33  34
  nonsmoker    1   3   2   3   4   3   5   4   5   6  11  16
  smoker       0   0   0   0   0   0   0   0   1   2   5   3
            weeks
habit         35  36  37  38  39  40  41  42  43  44  45
  nonsmoker  28  42  93 155 201 150  89  22  16  10   3
  smoker       4   4  12  23  36  18  10   3   1   3   1
```

```
ggplot(subset(nc, !is.na(habit)), aes(x=weeks, y=weight, color=habit))
  geom_point() +
  geom_smooth(method='lm', formula='y~x')
```




Birth weights seem to increase with the length of pregnancy (`weeks`) in a **nonlinear** manner.

```
ggplot(nc, aes(x=weeks, y=weight, color=habit)) +
  geom_point() +
  geom_smooth(method='lm', formula='y~x+I(x^2)+I(x^3)')
```




Describe the effect of mother's smoking `habit` on birth weights of babies, after adjusting for the length of pregnancy (`weeks`)

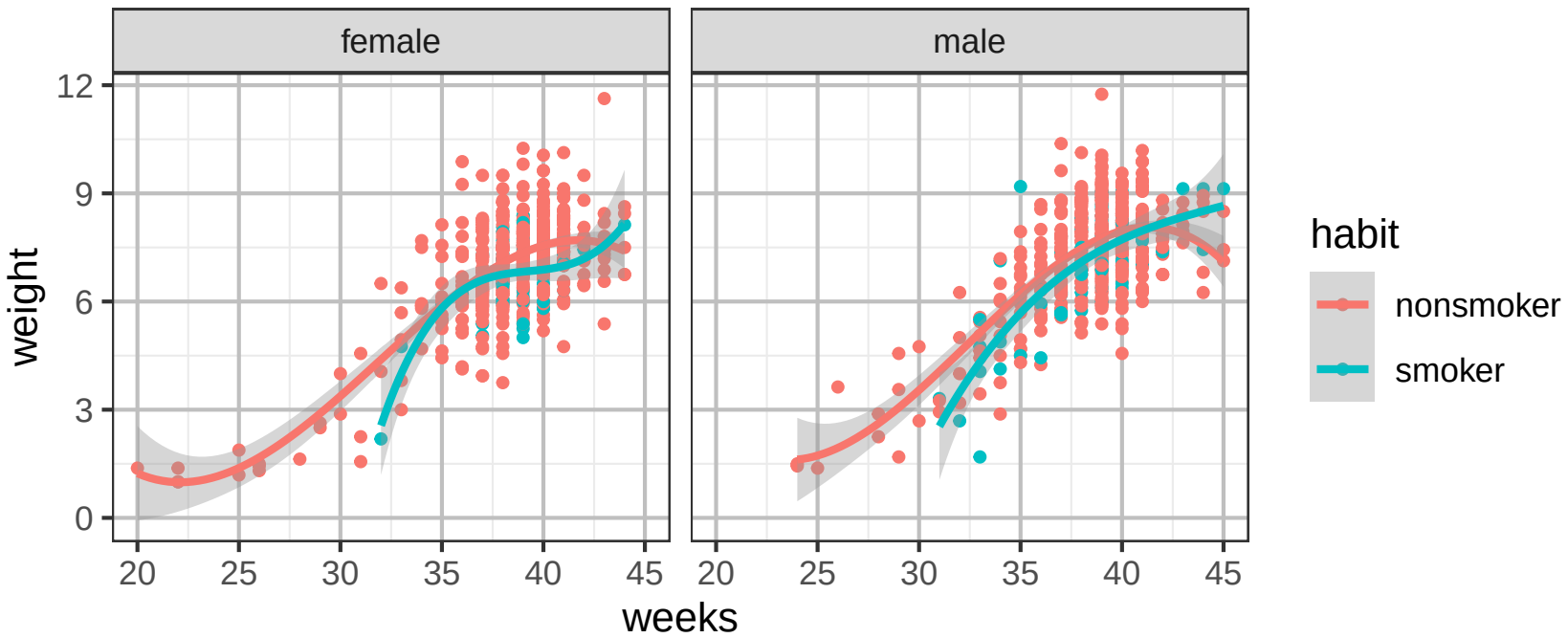
Answer #1: Comparing babies with the same of pregnancy, babies of smoking moms have lower birth weights on average than those of nonsmoking moms.

Answer #1: Comparing babies with the same of pregnancy, babies of smoking moms have lower birth weights on average than those of nonsmoking moms.

Answer #2: Babies of smoking moms have a lower mean birth weight than those of nonsmoking moms, controlling/adjusted for the length of pregnancy

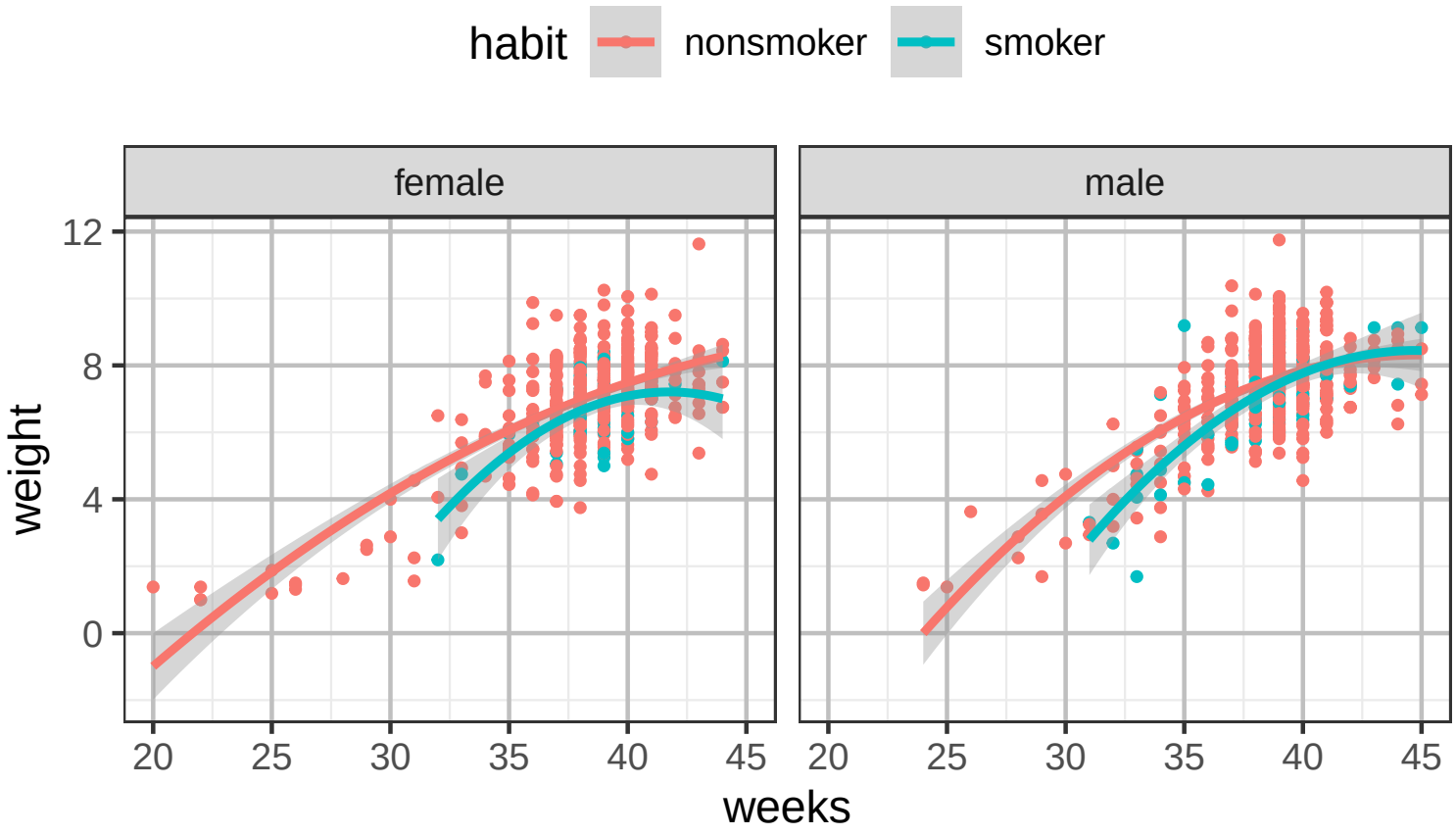
# Controlling for Gender As Well . . .

```
ggplot(nc, aes(x=weeks, y=weight, color=habit)) +
  geom_point(size = 1) +
  geom_smooth(method='lm', formula='y~x+I(x^2)+I(x^3)')+
  facet_wrap(~gender)
```



One can move the legend to the top.

```
ggplot(nc, aes(x=weeks, y=weight, color=habit)) +
  geom_point(size=0.8) + facet_wrap(~gender) +
  geom_smooth(method='lm', formula='y~x+I(x^2)')+
  theme(legend.position="top")
```

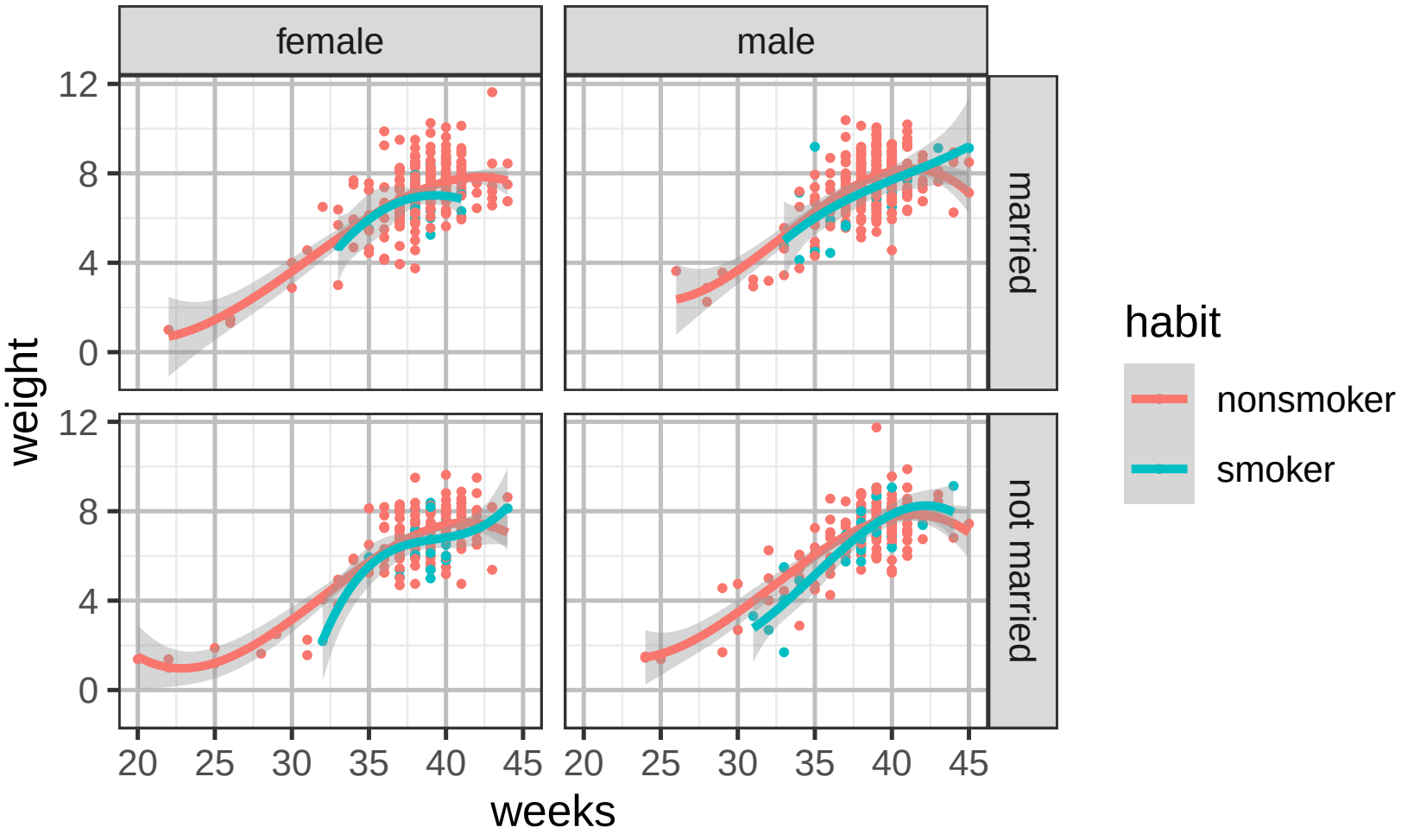


Describe the effect of mother's smoking `habit` on birth weights of babies, after adjusting for the length of pregnancy (`weeks`) and the gender of the baby.

# On Your Own

**Q3:** Make an appropriate plot of the data and use it to describe the effect of mother's `marital` status on birth weights of babies, after adjusting for the length of pregnancy (`weeks`) and the gender of the baby.

```
ggplot(nc, aes(x=weeks, y=weight, color=habit)) +
  geom_point(size=0.5) +
  geom_smooth(method='lm', formula='y~x+I(x^2)+I(x^3)')+
  facet_grid(marital ~ gender)
```



What's the effect of `habit` on `weight` after adjusting for `weeks`, `gender`, and `marital`?

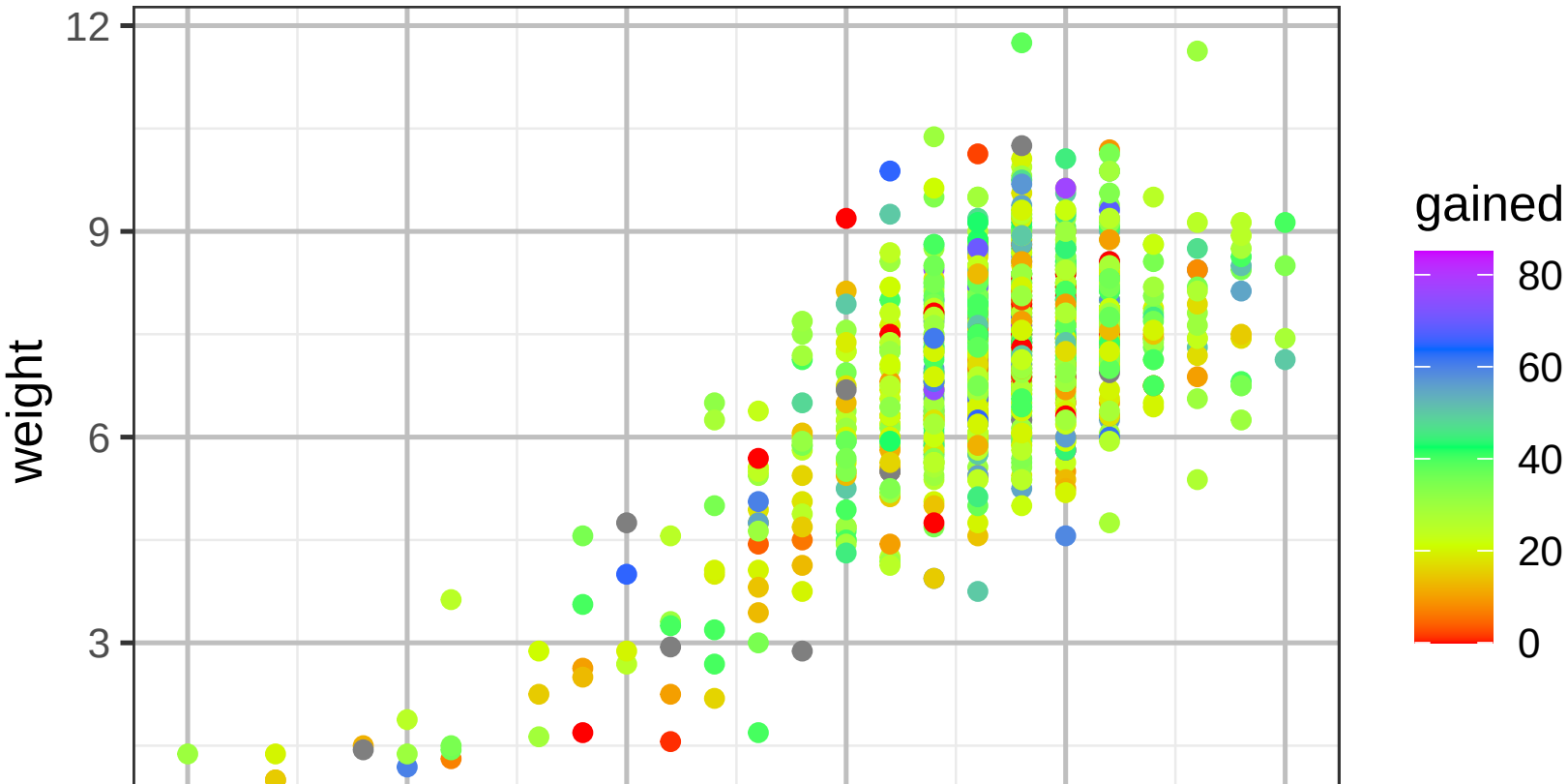
# On Your Own

**Q4:** Make an appropriate plot of the data and use it to describe the effect of `gender` on `weight` after adjusting for `weeks`, `habit`, and `marital`?

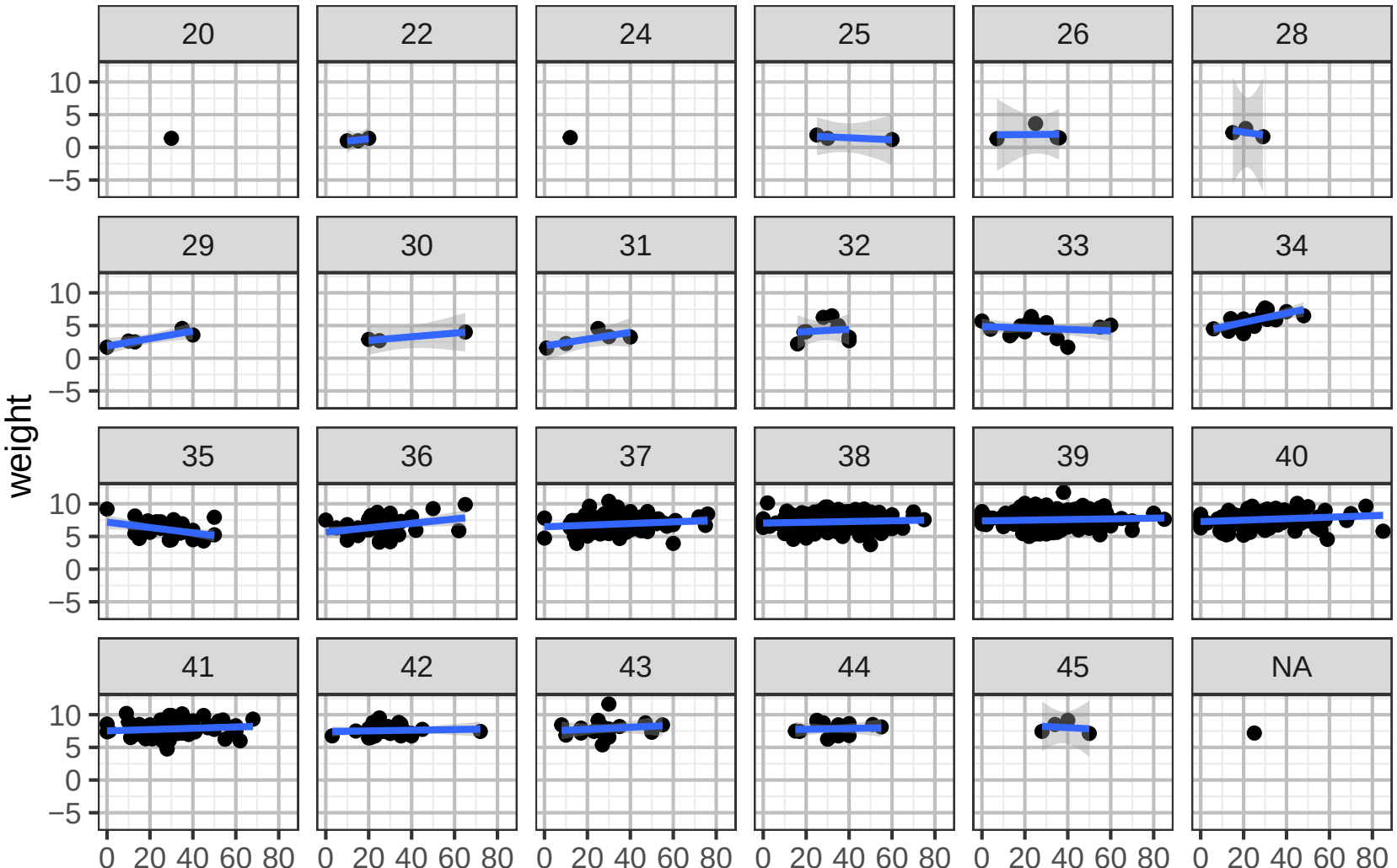
**Q5:** Make another plot of the data and use it to describe the effect of `marital` on `weight` after adjusting for `weeks`, `gender`, and `habit`?

The variable `gained` is mother's weight gain during pregnancy in pounds. Let's see if `gained` has any effect on baby's birth weight, after adjusting for `weeks`.

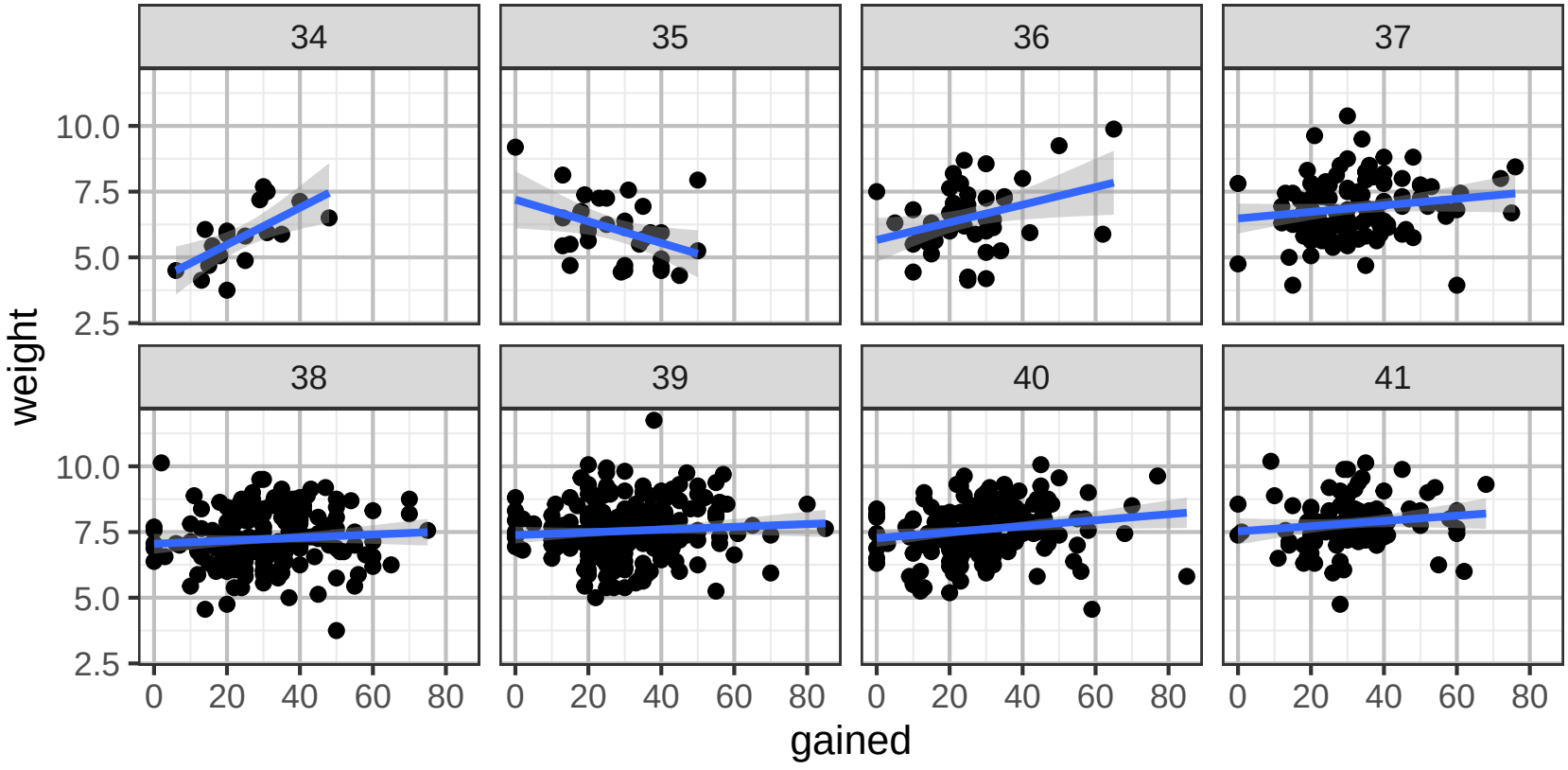
```
ggplot(nc, aes(x=weeks, y=weight, color=gained)) +
  geom_point() +
  scale_color_gradientn(colours = rainbow(5))
```



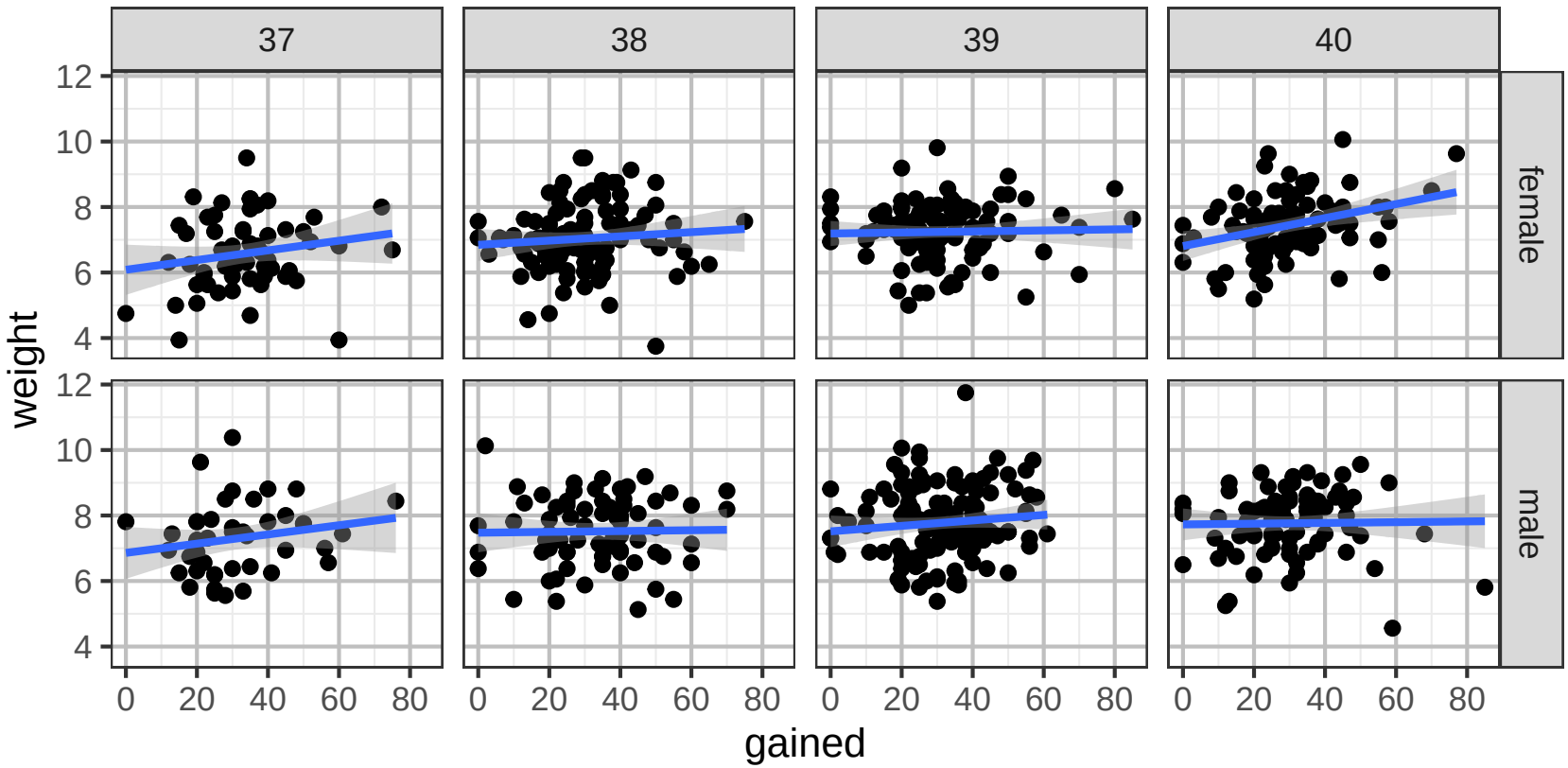
```
ggplot(nc, aes(x=gained, y=weight)) +
  geom_point() + geom_smooth(method='lm') +
  facet_wrap(~weeks, nrow=4)
`geom_smooth()` using formula 'y ~ x'
```



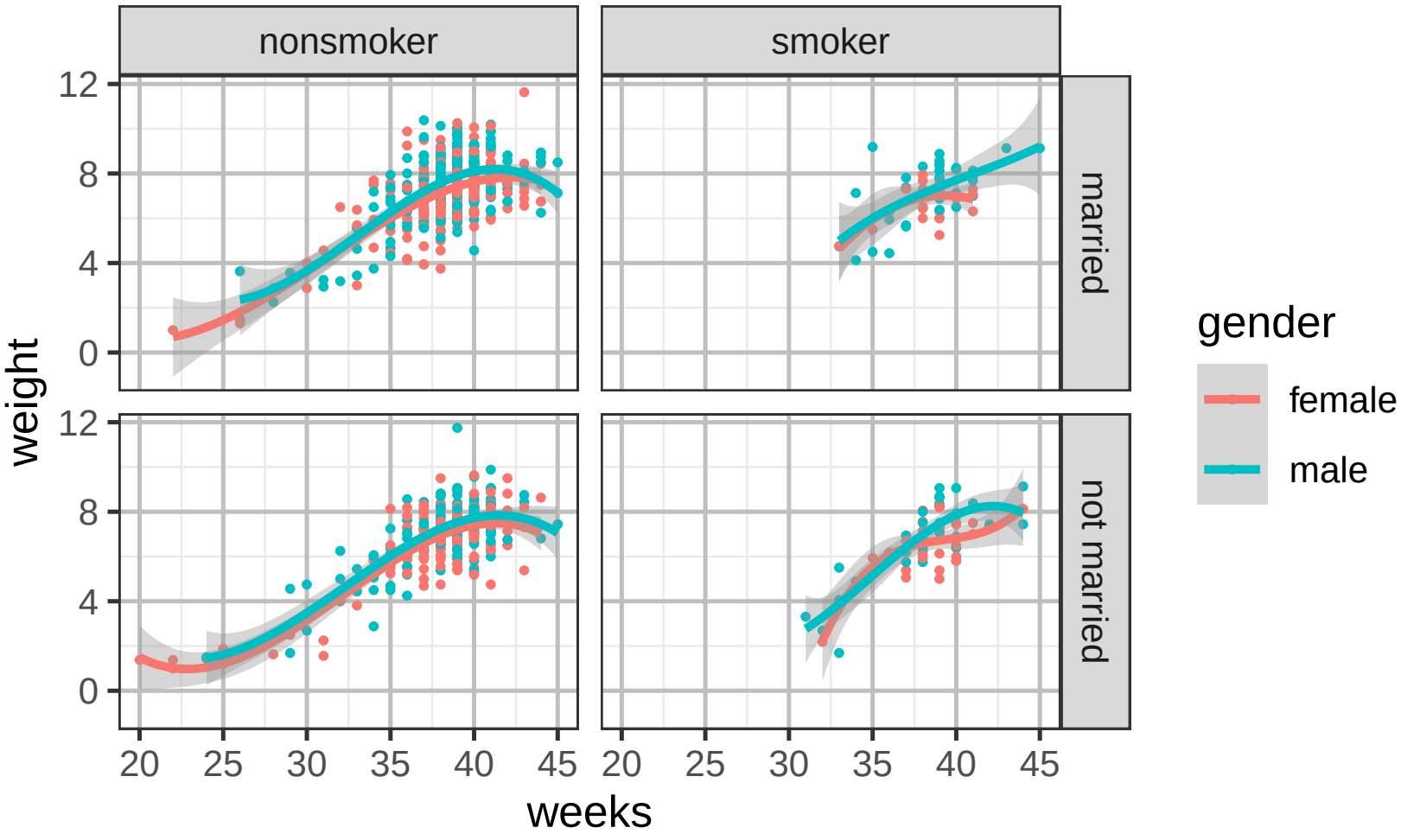
```
ggplot(subset(nc, weeks > 33 & weeks < 42), aes(x=gained, y=weight)) +
  geom_point() + geom_smooth(method='lm') +
  facet_wrap(~weeks, nrow=2)
`geom_smooth()` using formula 'y ~ x'
```



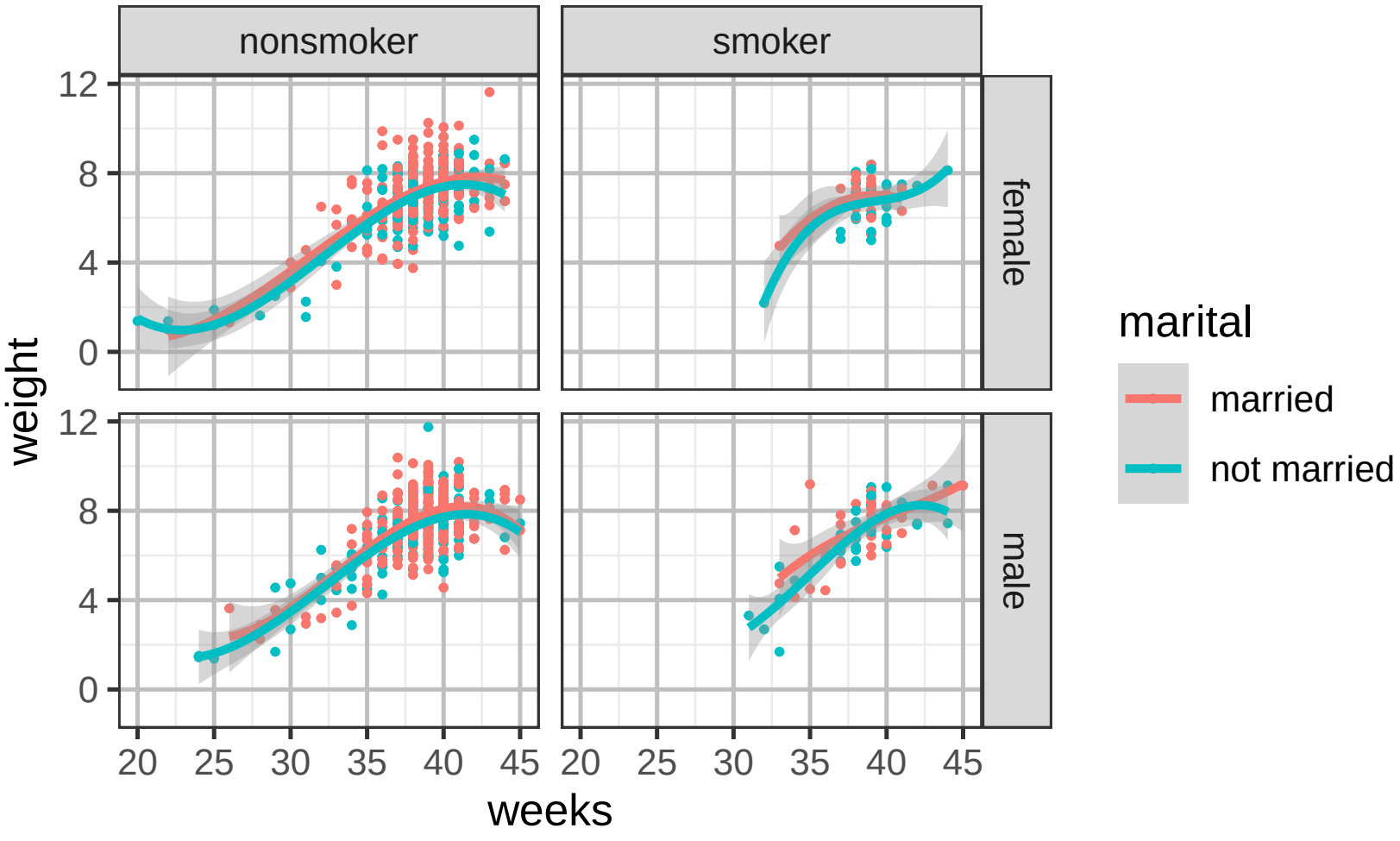
```
ggplot(subset(nc, weeks > 36 & weeks < 41), aes(x=gained, y=weight)) +
  geom_point() + geom_smooth(method='lm') +
  facet_grid(gender~weeks)
`geom_smooth()` using formula 'y ~ x'
```



```
ggplot(nc, aes(x=weeks, y=weight, color=gender)) +
  geom_point(size=0.5) +
  geom_smooth(method='lm', formula='y~x+I(x^2)+I(x^3)')+
  facet_grid(marital ~ habit)
```



```
ggplot(nc, aes(x=weeks, y=weight, color=marital)) +
  geom_point(size=0.5) +
  geom_smooth(method='lm', formula='y~x+I(x^2)+I(x^3)')+
  facet_grid(gender ~ habit)
```



```
xtabs(~ marital + weeks, data=subset(nc,gender=="female" & habit=="smok
              weeks
marital        32 33 35 36 37 38 39 40 41 42 44
  married       0  1  1  0  1  8  9  3  3  0  0
  not married   1  0  1  1  3  7  8  6  2  1  1
xtabs(~ marital + weeks, data=subset(nc,gender=="male" & habit=="smoker
              weeks
marital        31 32 33 34 35 36 37 38 39 40 41 42 43 44 45
  married       0  0  1  2  2  3  4  1 11  4  4  0  1  0  1
  not married   1  1  3  1  0  0  4  7  8  5  1  2  0  2  0
xtabs(~ marital + weeks, data=subset(nc,gender=="male" & habit=="nonsmo
              weeks
marital        24 25 26 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43
  married       0  0  1  2  1  0  2  1  4  3 11  9 26 46 79 47 22  8  3
  not married   2  1  0  0  2  2  0  3  2  6  5 10 11 19 35 30 20  4  2
xtabs(~ marital + weeks, data=subset(nc,gender=="female" & habit=="nons
              weeks
marital        20 22 25 26 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42
  married       0  1  0  3  0  0  2  1  1  3  5  7 11 32 64 64 45 27  3
  not married   1  2  2  0  1  2  0  2  1  2  2  5 12 24 26 23 28 20  7
```