Lab 4 - Data Visualization

Solutions

Last Updated 09-21-2018

Introduction

This lab uses the `NCbirths` data set. Use the following code chunk to read the data into R. Need help? Look at lesson 03 - importing data.

```
NCbirths <- read.csv("data/NCbirths.csv", header=TRUE)
```

You will use a combination of base and `ggplot2` graphics for this lab. Your answers can be for your eyes only (exploratory). They do not have to be pretty.

The code chunk below loads the `ggplot2` package for you, and also sets some code chunk options (using `opts_chunk` from the `knitr` package) to make your knitted report output more readable. I encourage you to play around with these options to learn how they work.

```
library(ggplot2)
library(knitr)
opts_chunk$set(warning=FALSE, message=FALSE, fig.height=4, fig.width=5, fig.align='center')
```

Univariate

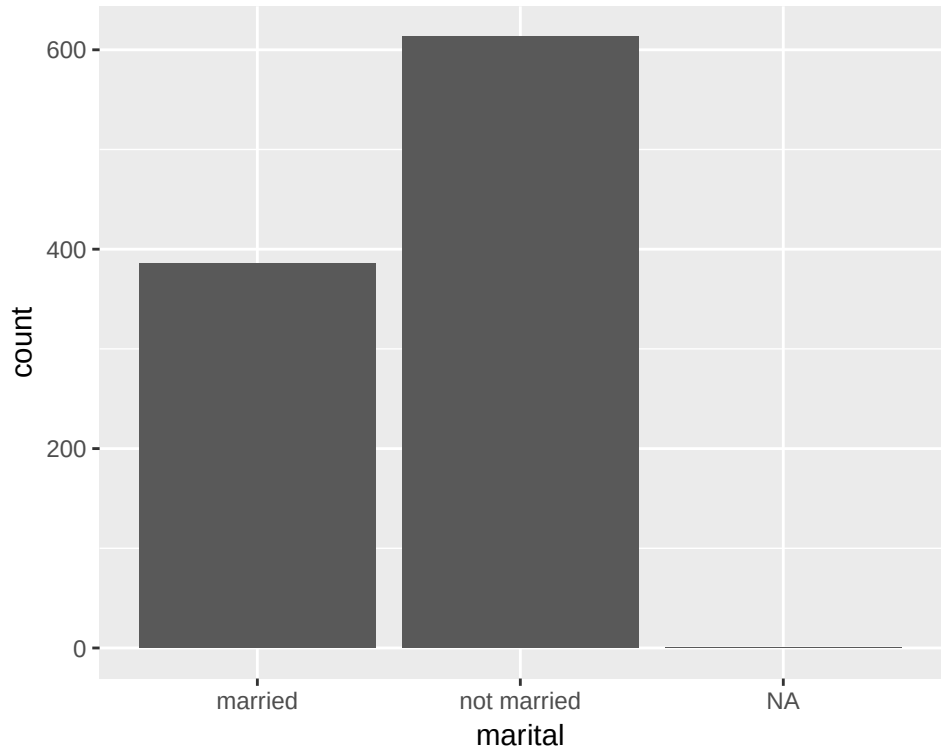
1. A table of marital status (`marital`)

```
table(NCbirths$marital)
```

```
##
##     married not married
##         386         613
```

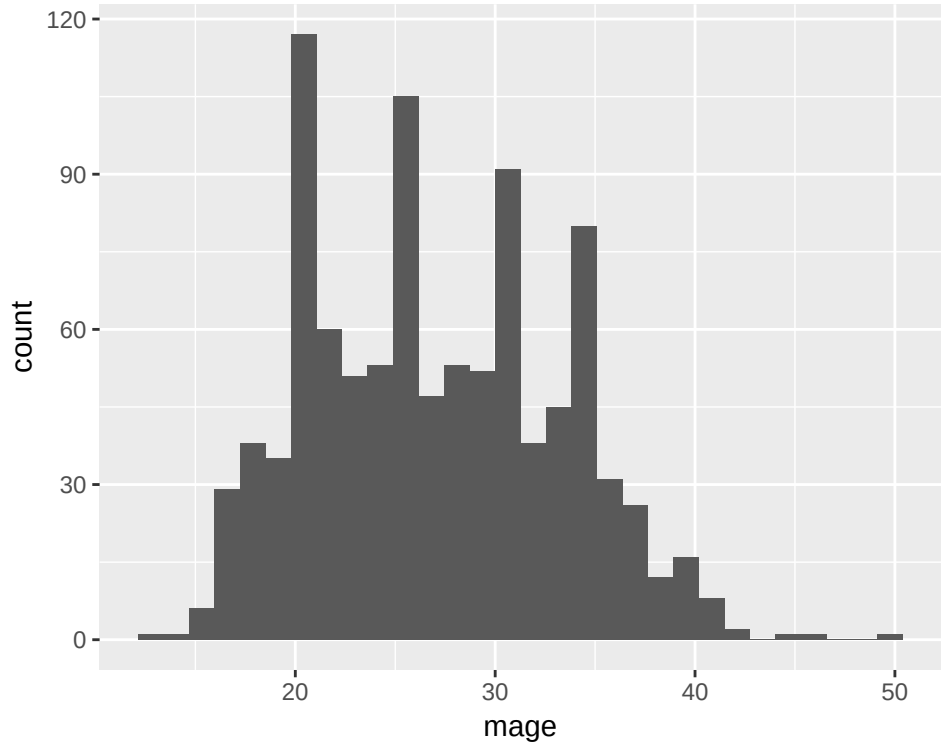
2. A barchart of maritial status

```
ggplot(NCbirths, aes(x=marital)) + geom_bar()
```



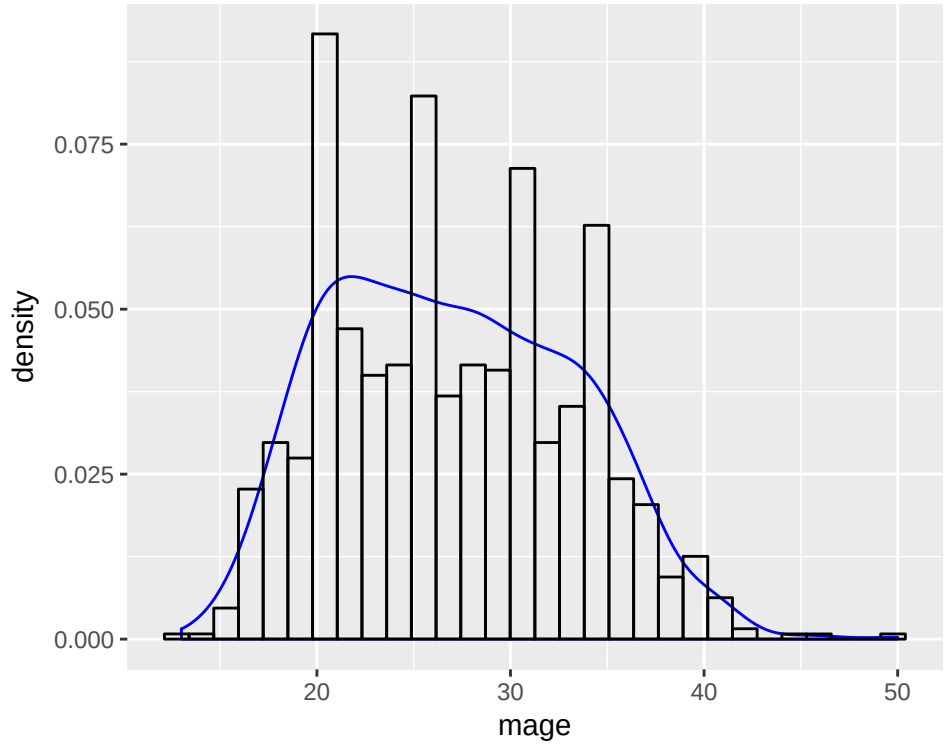
3. A histogram of mothers age (`mage`)

```
ggplot(NCbirths, aes(x=mage)) + geom_histogram()
```



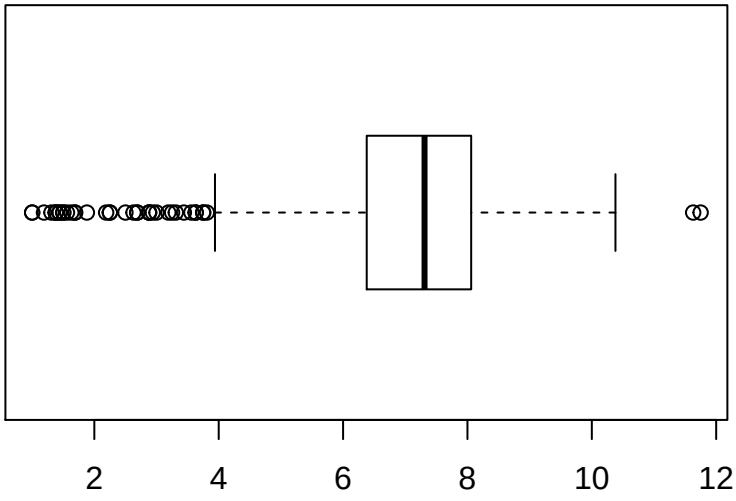
a. Now add an overlaid density plot in a different color. Don't forget to adjust the `aes` on the histogram.

```
ggplot(NCbirths, aes(x=mage)) + geom_density(col="blue") +
  geom_histogram(aes(y=..density..), colour="black", fill=NA)
```



4. A horizontal boxplot of weight of the baby (`weight`)

```
boxplot(NCbirths$weight, horizontal=TRUE)
```



Bivariate

5. Create a two-way frequency table of maturity status (`mature`) against smoking `habit`

```
table(NCbirths$mature, NCbirths$habit)
```

```
##
##                nonsmoker smoker
##   mature mom         121     11
##   younger mom        752    115
```

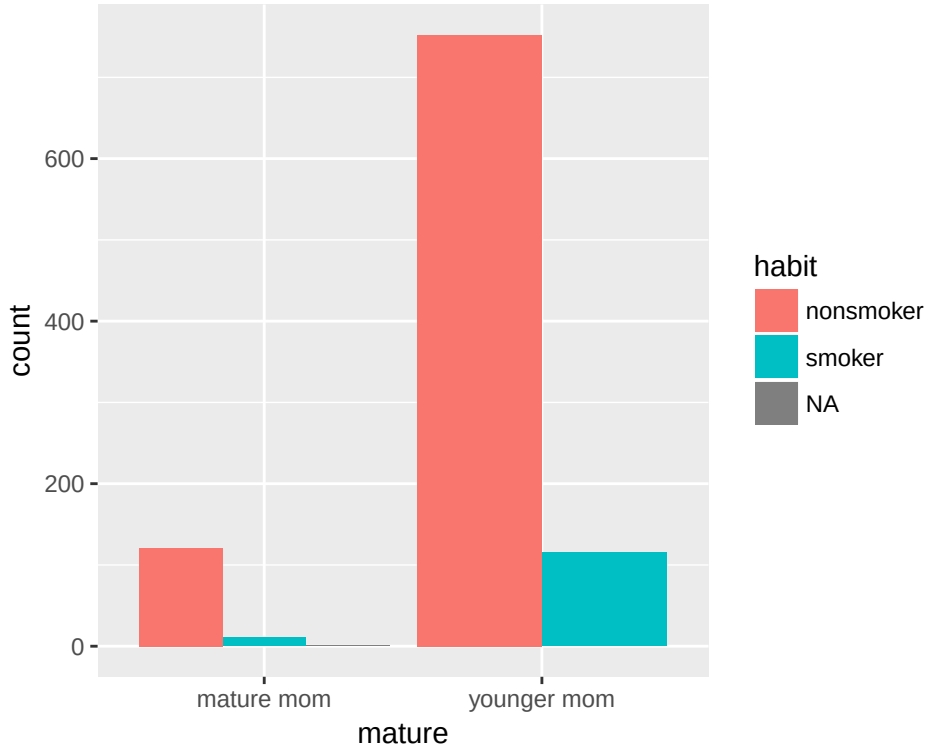
6. Create a proportion table of smoking habit *within* maturity status. Round to 3 digits.

```
round(prop.table(table(NCbirths$mature, NCbirths$habit), margin=1),3)
```

```
##
##                nonsmoker smoker
##   mature mom       0.917  0.083
##   younger mom      0.867  0.133
```

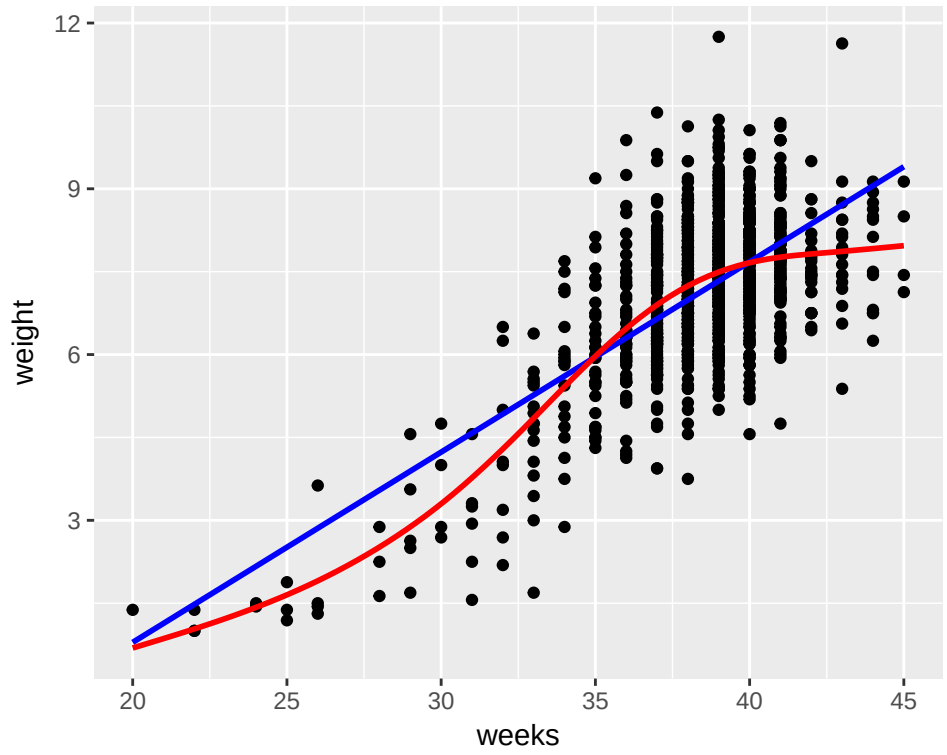
7. Create a grouped barchart that reflects the proportions you calculated above. Think carefully which variable goes on the `x` axis, and which one is used for the `fill`

```
ggplot(NCbirths, aes(x=mature, fill=habit)) + geom_bar(position="dodge")
```



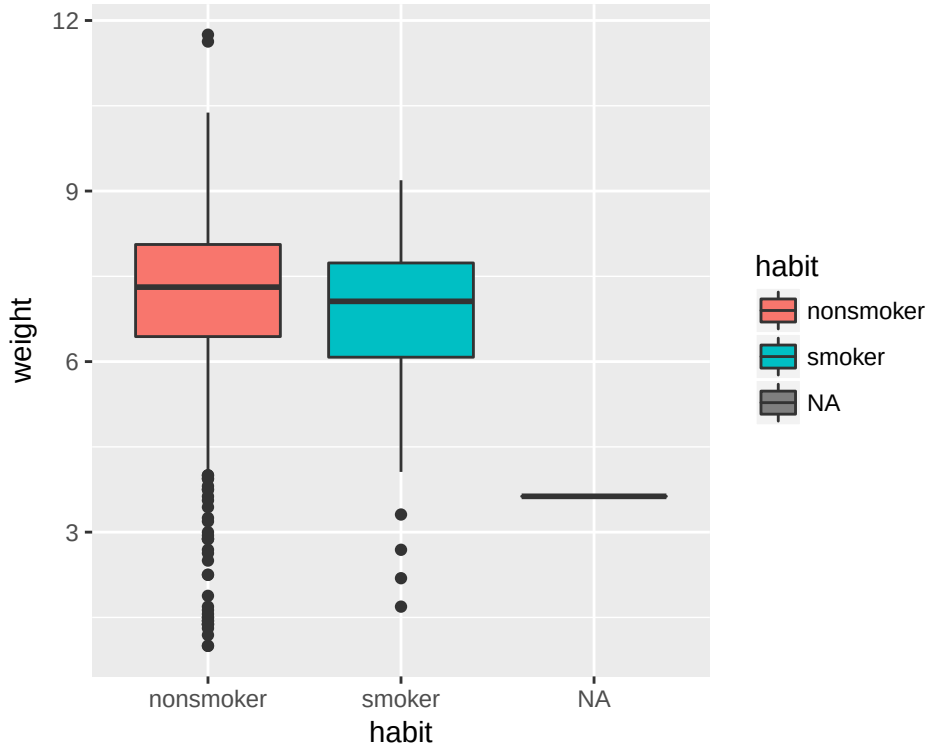
8. A scatterplot of length of pregnancy in `weeks` and the babies `weight`.
    a. Include a smoother line in brown, and a best fit linear model line in purple

```
ggplot(NCbirths, aes(x=weeks, y=weight)) + geom_point() +
  geom_smooth(se=FALSE, method="lm", color="blue") +
  geom_smooth(se=FALSE, color="red")
```



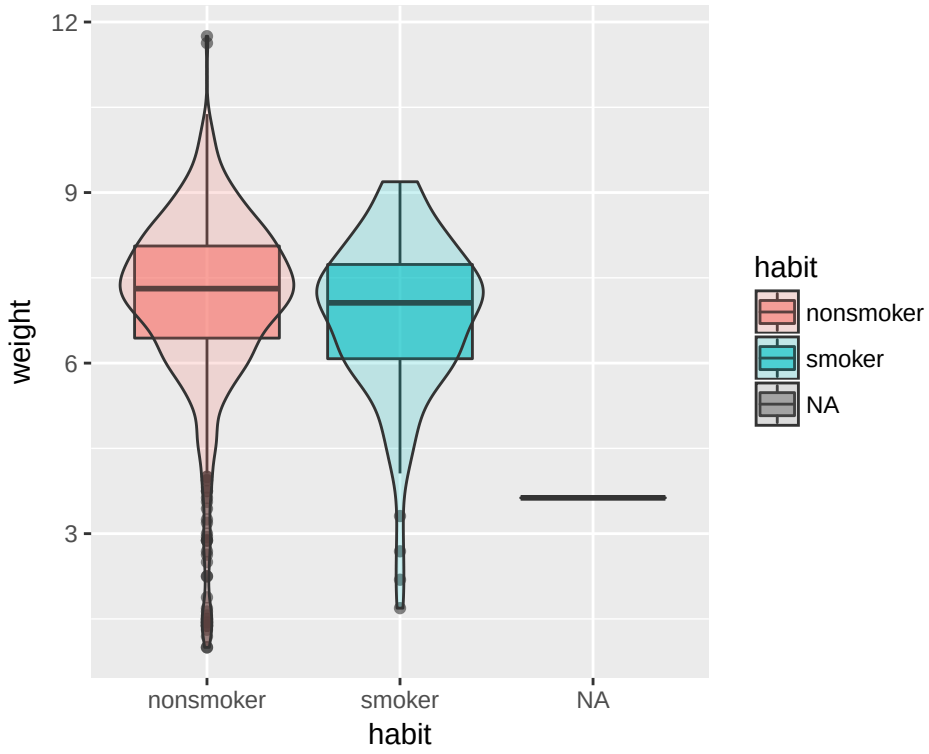
9. Grouped boxplots of baby `weight` by mothers smoking `habit`. Make sure you `fill` the boxes by `habit` as well.

```
ggplot(NCbirths, aes(x=habit, y=weight, fill=habit)) + geom_boxplot()
```



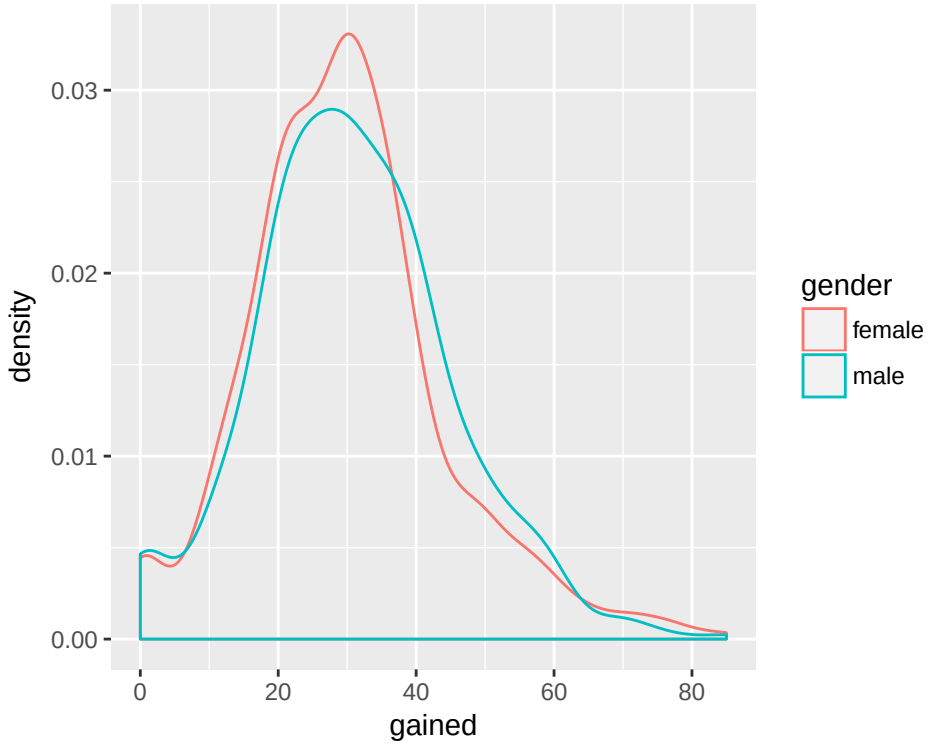
10. Replicate the same plot as above, but overlay a violin plot and change the transparency of both violin and boxplot layers.

```
ggplot(NCbirths, aes(x=habit, y=weight, fill=habit)) +
       geom_boxplot(alpha=.6) + geom_violin(alpha=.2)
```



11. Overlaid density plots of weight `gained` by babies `gender`. No fill, color only.

```
ggplot(NCbirths, aes(x=gained, col=gender)) + geom_density()
```

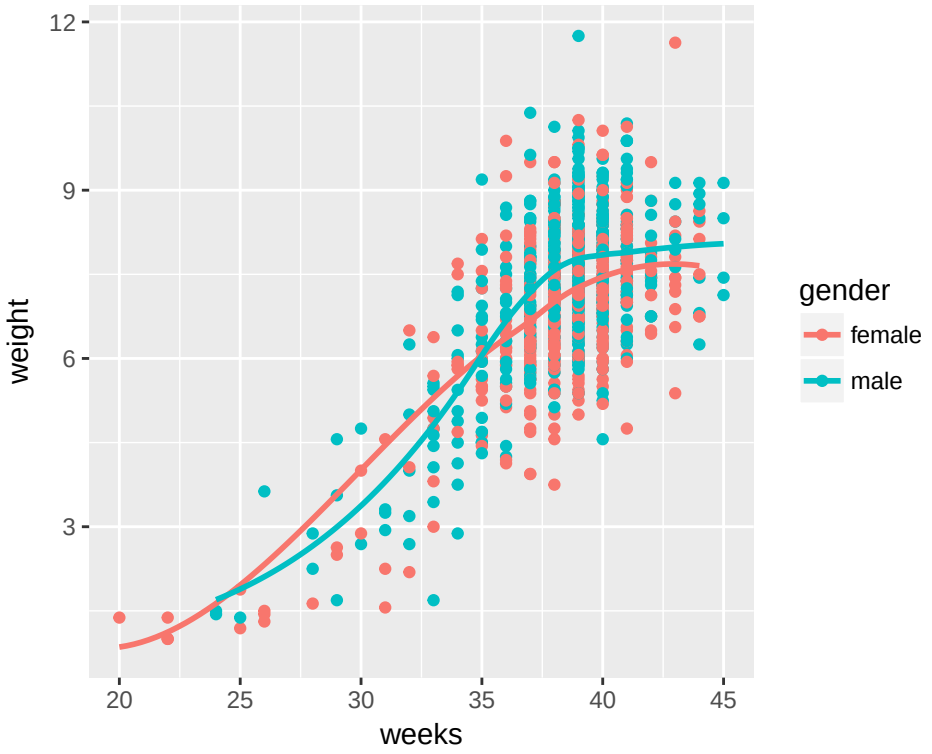


Bonus optional questions.

Looking for more? Want to start learning how to clean up your plots to make them professional looking? Go check out the full data viz tutorial and try the following problems.

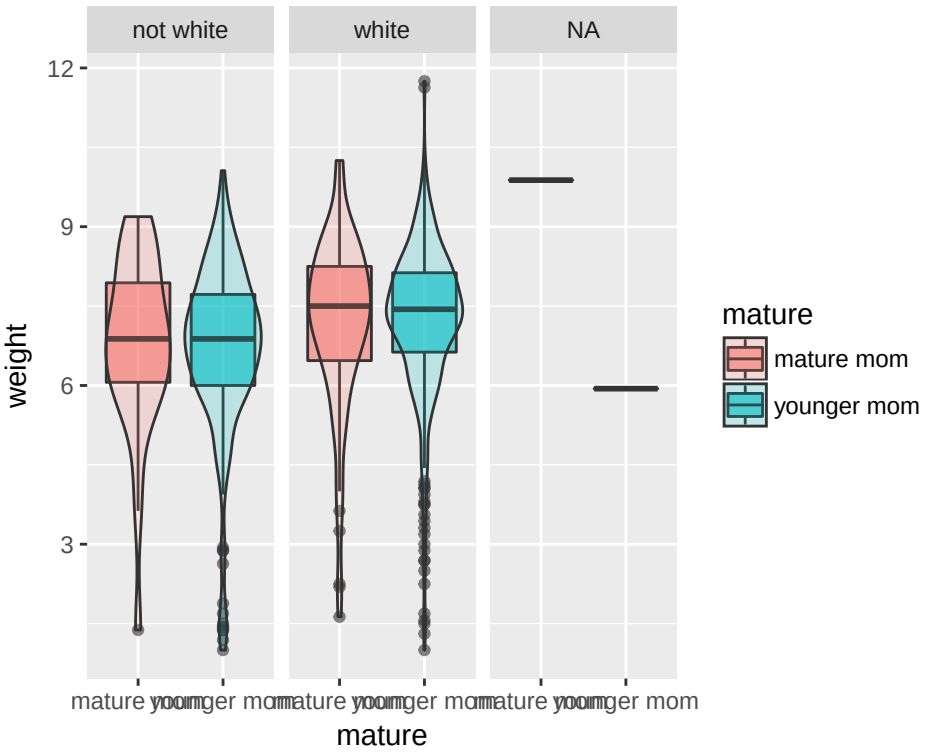
1. A scatterplot of length of pregnancy in `weeks` and the babies `weight`, color the points and lowess lines by `gender`.

```
ggplot(NCbirths, aes(x=weeks, y=weight, col=gender)) + geom_point() + geom_smooth(se=FALSE)
```



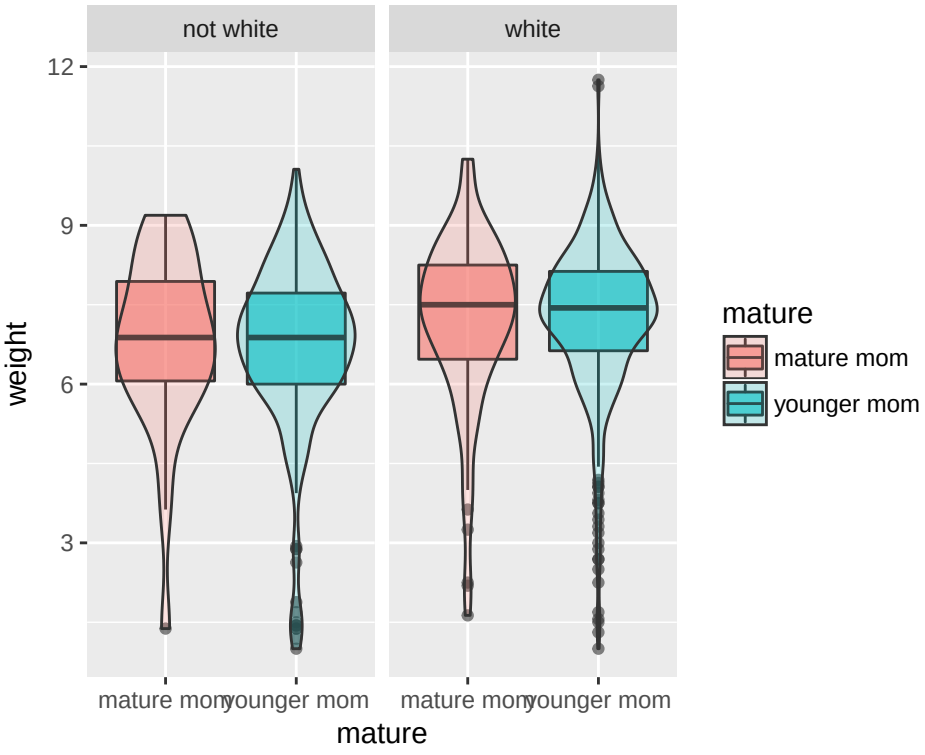
2. Use `facet_wrap` to create side by side boxplots with overlaid violins (using good trasparency options) of `weight` by `mature`, paneled by `whitemom`.

```
ggplot(NCbirths, aes(x=mature, y=weight, fill=mature)) +
        geom_boxplot(alpha=.6) + geom_violin(alpha=.2) + facet_wrap(~whitemom)
```



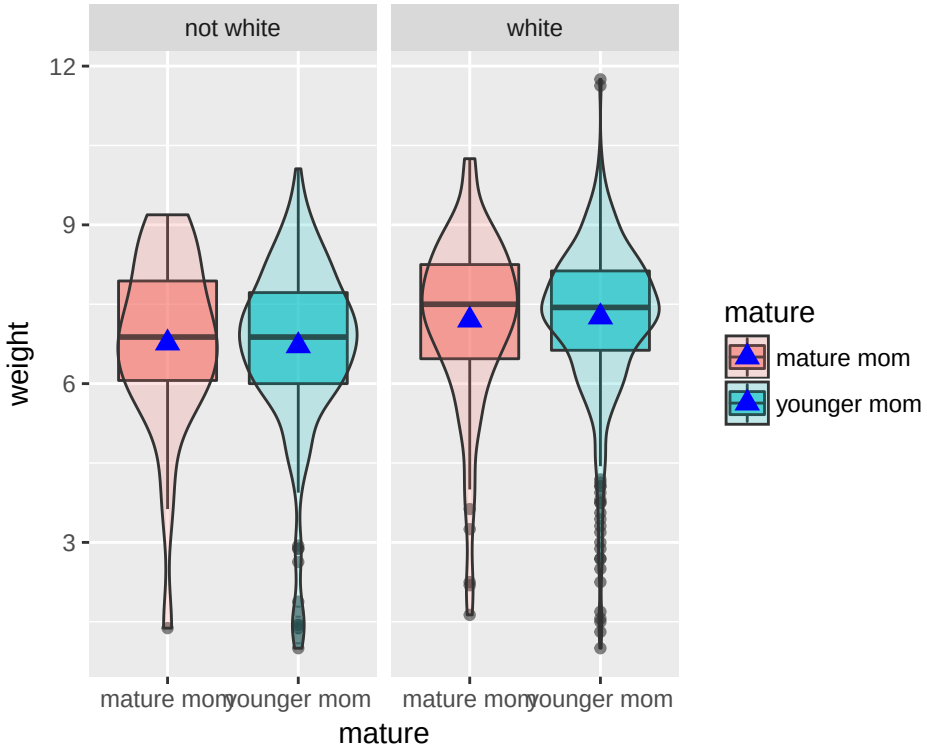
3. Remove the NA group from the previous plot and recreate the plot.

```
library(dplyr)
plot.data <- NCbirths %>% select(mature, weight, whitemom) %>% na.omit()
ggplot(plot.data, aes(x=mature, y=weight, fill=mature)) +
        geom_boxplot(alpha=.6) + geom_violin(alpha=.2) + facet_wrap(~whitemom)
```



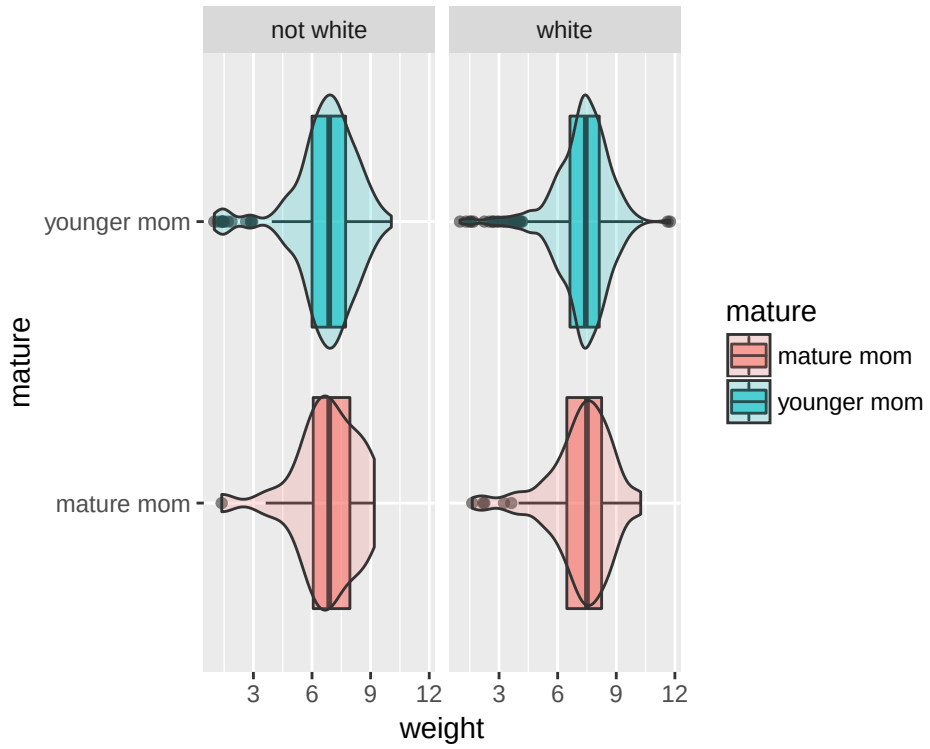
4. Now add points for the means to the above plot.

```
ggplot(plot.data, aes(x=mature, y=weight, fill=mature)) +
        geom_boxplot(alpha=.6) + geom_violin(alpha=.2) + facet_wrap(~whitemom) +
        stat_summary(fun.y="mean", geom="point", size=3, pch=17,color="blue")
```



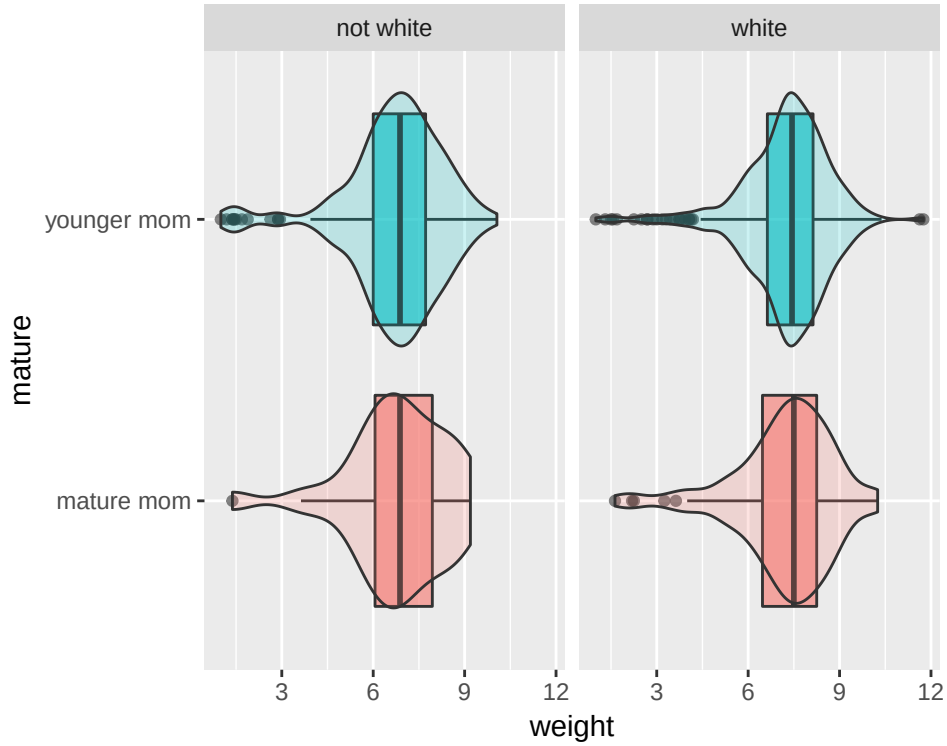
5. Use `coord_flip()` to make the boxes horizontal

```
ggplot(plot.data, aes(x=mature, y=weight, fill=mature)) +
        geom_boxplot(alpha=.6) + geom_violin(alpha=.2) + facet_wrap(~whitemom) +
        coord_flip()
```



6. Go to the R Cookbook and learn how to remove that legend.

```
ggplot(plot.data, aes(x=mature, y=weight, fill=mature)) +
        geom_boxplot(alpha=.6) + geom_violin(alpha=.2) + facet_wrap(~whitemom) +
        coord_flip() + scale_fill_discrete(guide=FALSE)
```



7. Apply a `theme_minimal()` or `theme_bw()` layer to remove the grey background from the plot.

```
ggplot(plot.data, aes(x=mature, y=weight, fill=mature)) +
        geom_boxplot(alpha=.6) + geom_violin(alpha=.2) + facet_wrap(~whitemom) +
        coord_flip() + scale_fill_discrete(guide=FALSE) + theme_bw()
```