Graphics with with ggplot2

What is ggplot

- ggplot2 is an R package for producing statistical, or data, graphics.
- Under the tidyverse family of packages
- ggplot2 has an underlying grammar, based on the Grammar of Graphics
- compose graphs by combining independent components.

How ggplot works

- ggplot2 divides plot into three different fundamental parts:
  - Plot = data + Aesthetics + Geometry
- The principal components of every plot can be defined as follow:
  - **data** is a data frame
  - **Aesthetics** is used to indicate x and y variables. It can also be used to control the color, the size or the shape of points, the height of bars, etc. . . ..
  - **Geometry** defines the type of graphics (histogram, box plot, line plot, density plot, dot plot, ...)

Install ggplot2

```
# Installation
install.packages("ggplot2")
# Loading
library(ggplot2)
```

Load the data

```
setwd("where/your/work/folder/is/")
```

```
library(readr)
bw_df <- read_csv("birth_weight.csv")
```

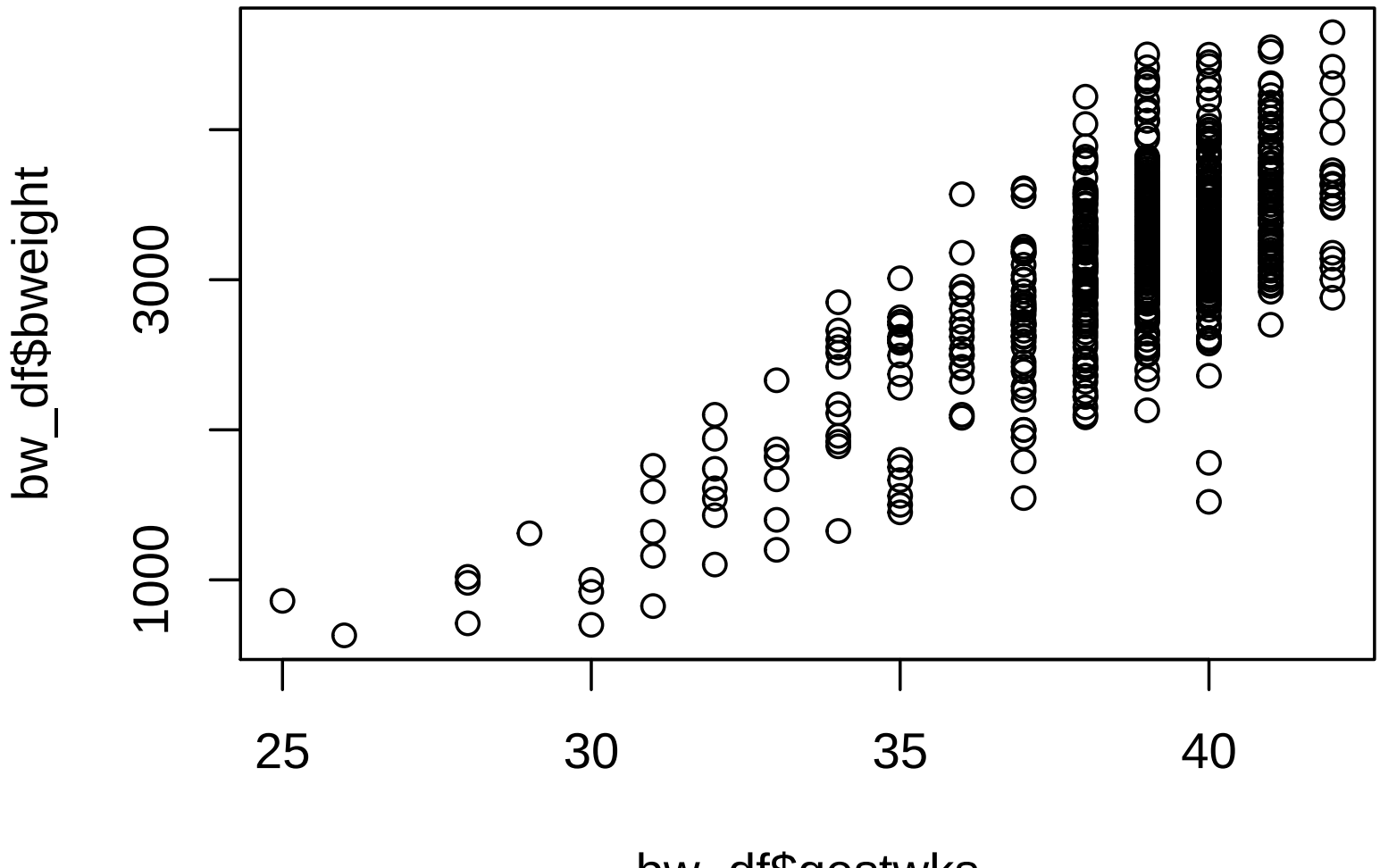
```
##
## -- Column specification --------------------------------
## cols(
##   id = col_double(),
##   matage = col_double(),
##   ht = col_double(),
##   gestwks = col_double(),
##   sex = col_character(),
##   bweight = col_double(),
##   ethnic = col_double(),
##   agegrp = col_character()
## )
```

Why?

- To have an understanding of you data we normally conduct exploratory data analysis (EDA) which can be graphical or numerical
- Primarily EDA is for seeing what the data can tell us before the formal modelling or hypothesis testing task
- Typical graphical techniques used in EDA are:
  - Scatter plots,
  - Box plots,
  - Bar plots

Scatter with base R

```
plot(bw_df$gestwks, bw_df$bweight)
```



Elements of grammar of graphics

- Data: variables mapped to aesthetic (aes function) features of the graph.
- Geoms: objects/shapes on the graph.
- Stats: stastical transformations that summarize data,(e.g mean, confidence intervals)
- Scales: define which aesthetic values are mapped to data values. Legends and axes display these mappings.
- Coordiante systems: define the plane on which data are mapped on the graphic.
- Faceting: splits the data into subsets to create multiple variations of the same graph (paneling).

Aesthetic mappings and aes

- Aesthetics are the visually perceivable components of the graph.
- Map variables to aesthetics using the aes function, such as:
  - which variables appear on the x-axis and y-axis.
  - a classification variable to colors
  - a numeric variable to the size of graphical objects

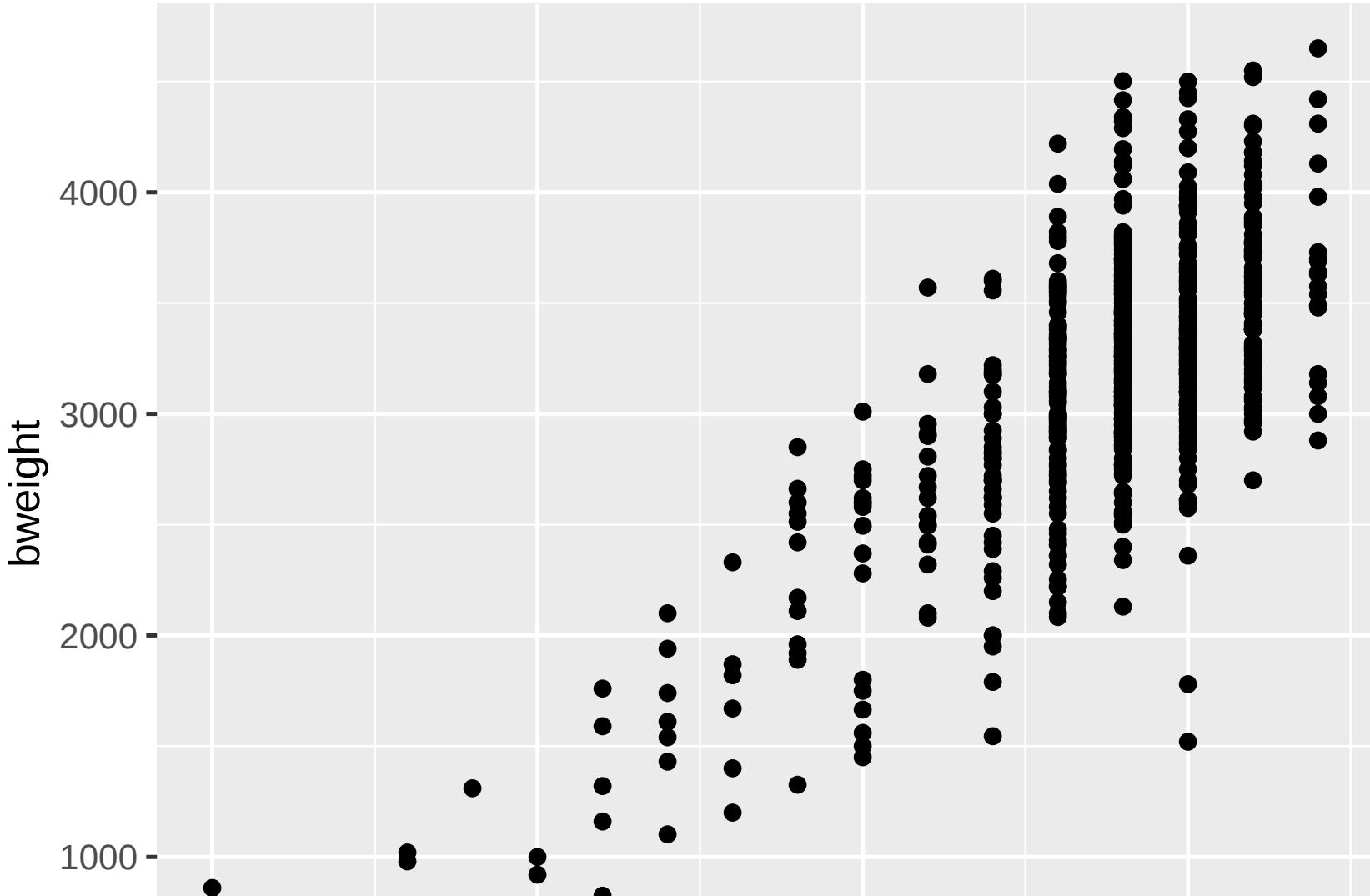
ggplot() template

```
ggplot(data = DATA) +
  GEOM_FUNCTION(mapping = aes(AESTHETIC MAPPINGS))
```

```
ggplot(data = troops) +
  geom_path(mapping = aes(x = longitude,
                          y = latitude,
                          color = direction,
                          size = survivors))
```

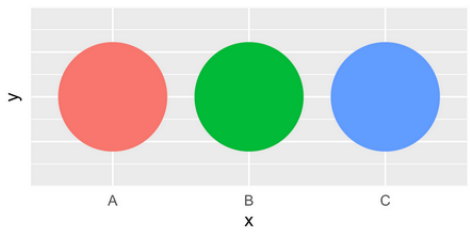
Scatter plot with ggplot2

```
# declare data and x and y aesthetics, but no shapes yet
ggplot(data = bw_df) + geom_point(mapping = aes(x = gestwks
    y = bweight))
```

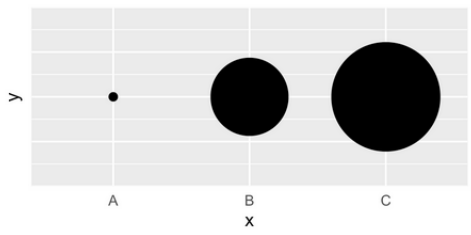


Aesthetics

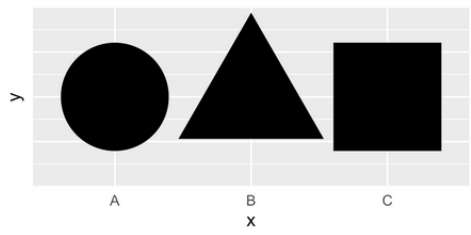
`color` (discrete)



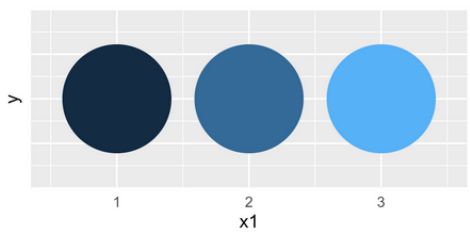
`size`



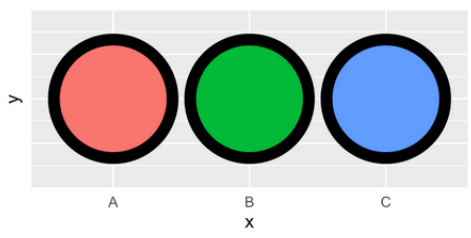
`shape`



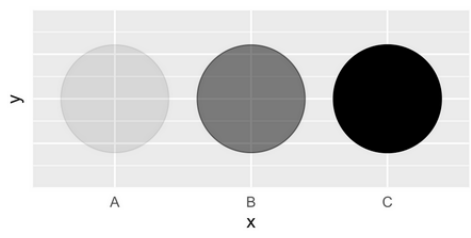
`color` (continuous)



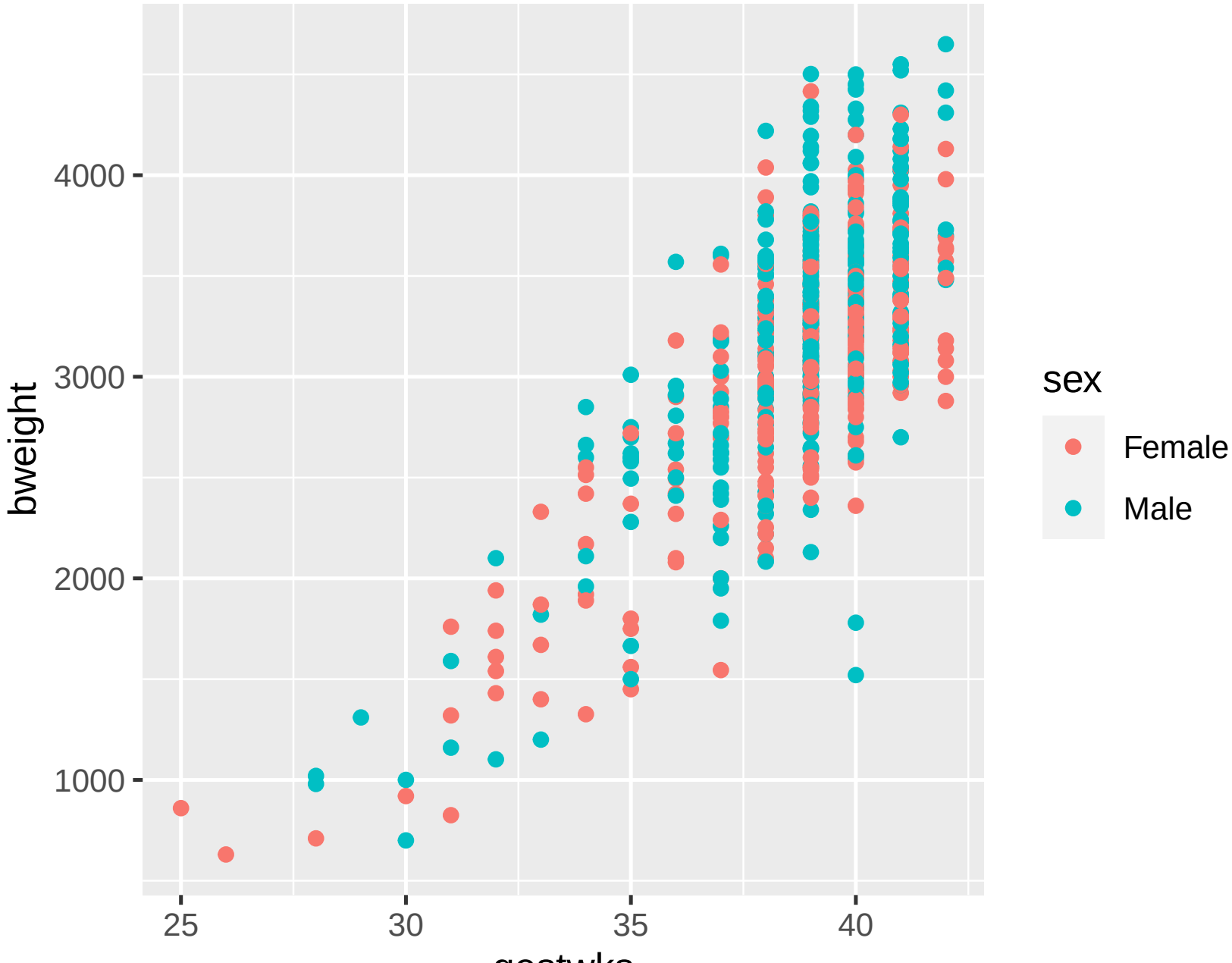
`fill`



`alpha`

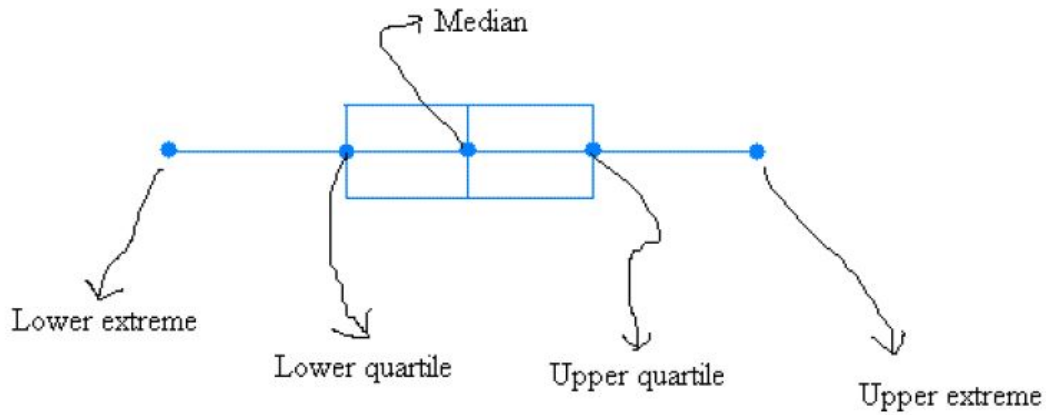


```
ggplot(data = bw_df) + geom_point(mapping = aes(x = gestwks
    y = bweight, color = sex))
```

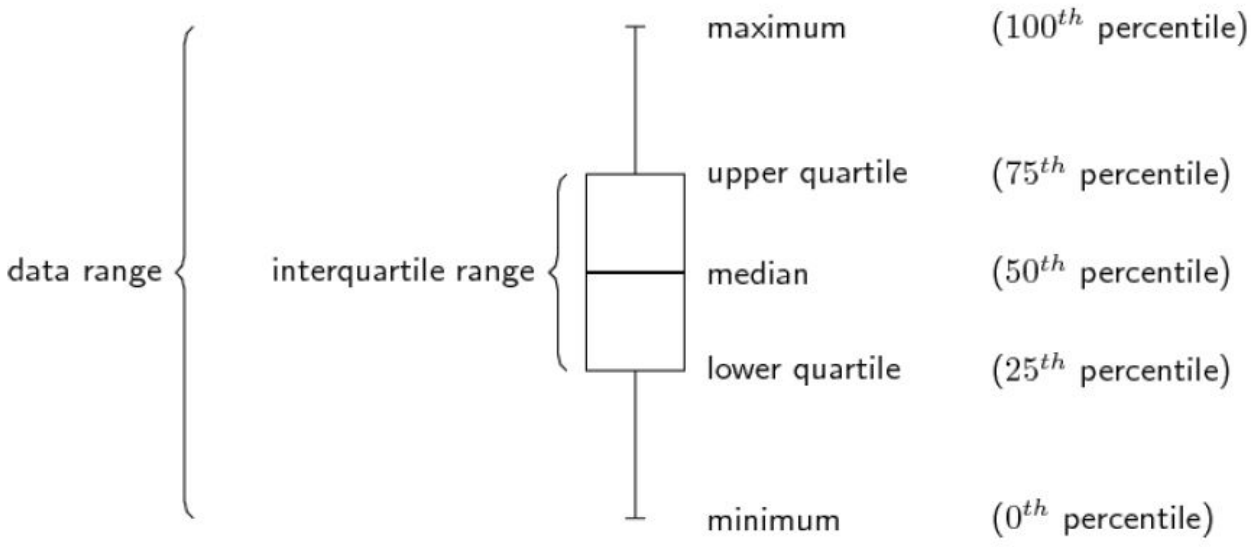


Boxplot

- Provides a standardized way of displaying the distribution of data.
- It attempts to provide a visual shape of the data distribution.
- This is based on some summary measures: min, 1^{st} quartile, median, 3^{rd} quartile, and max.
- Range, IQR, Outliers= $3^{*}IQR$ above 3^{rd} or below 1^{st} quartiles.



Definition

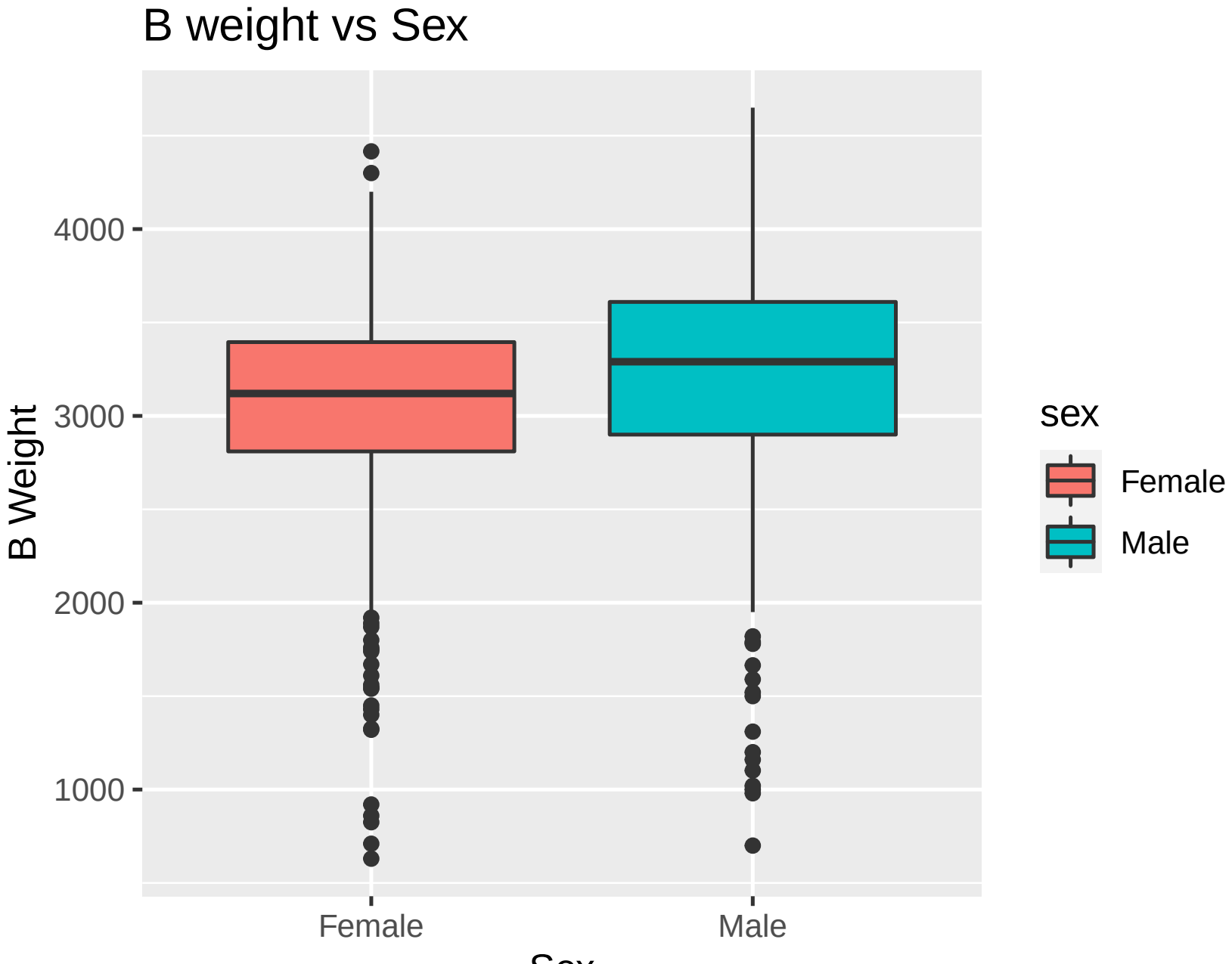

data range
interquartile range
maximum
(100^{th} percentile)
upper quartile
(75^{th} percentile)
median
(50^{th} percentile)
lower quartile
(25^{th} percentile)
minimum
(0^{th} percentile)

Lets do a Box plot?

- A box plot of bweight vs sex

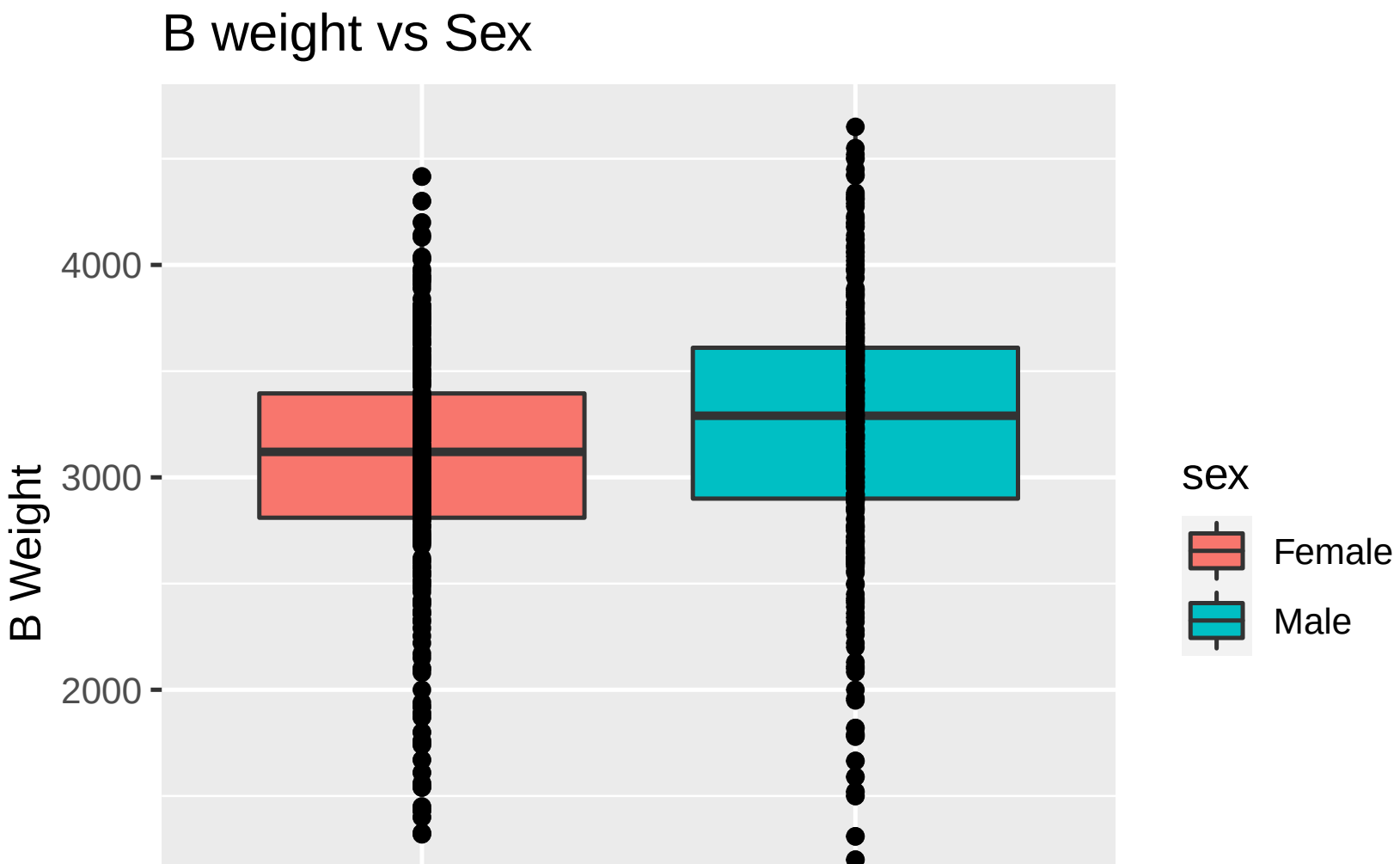
```
ggplot(data = bw_df)
```

```
ggplot(data = bw_df) + geom_boxplot(aes(y = bweight, x = se
    fill = sex)) + ylab("B Weight") + xlab("Sex") + ggtitle
```



Box plot and add scatter

```
ggplot(data = bw_df) + geom_boxplot(aes(y = bweight, x = se
    fill = sex)) + # geom_jitter(aes(y=bweight, x=sex))+
geom_point(aes(y = bweight, x = sex)) + ylab("B Weight") +
    ggtitle("B weight vs Sex")
```

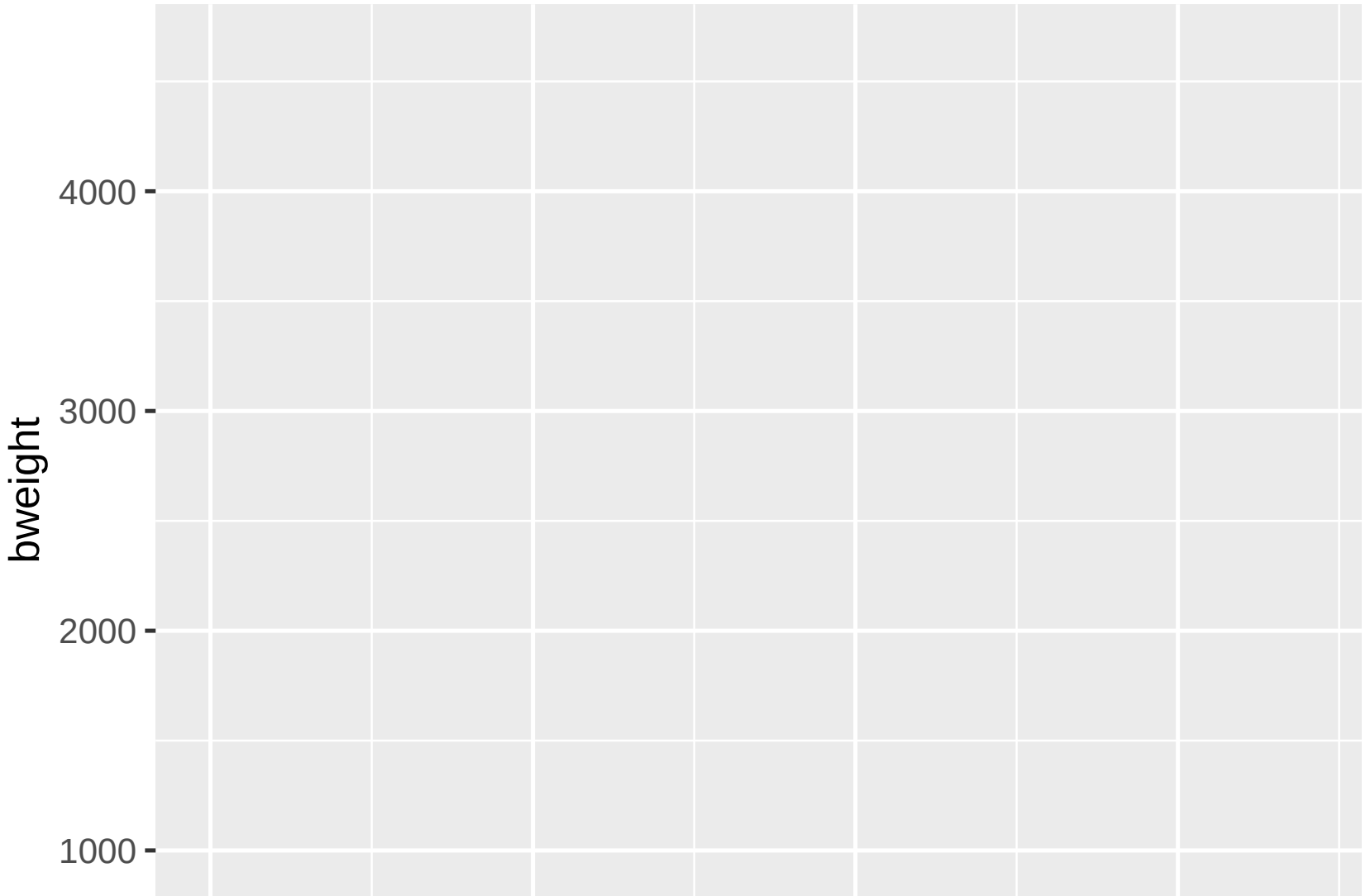


Putting it all together

We can build a plot sequentially
to see how each grammatical layer
changes the appearance

Start with data and aesthetics

```
# Start with data and aesthetics
ggplot(data = bw_df, mapping = aes(x = gestwks, y = bweight
    color = sex))
```



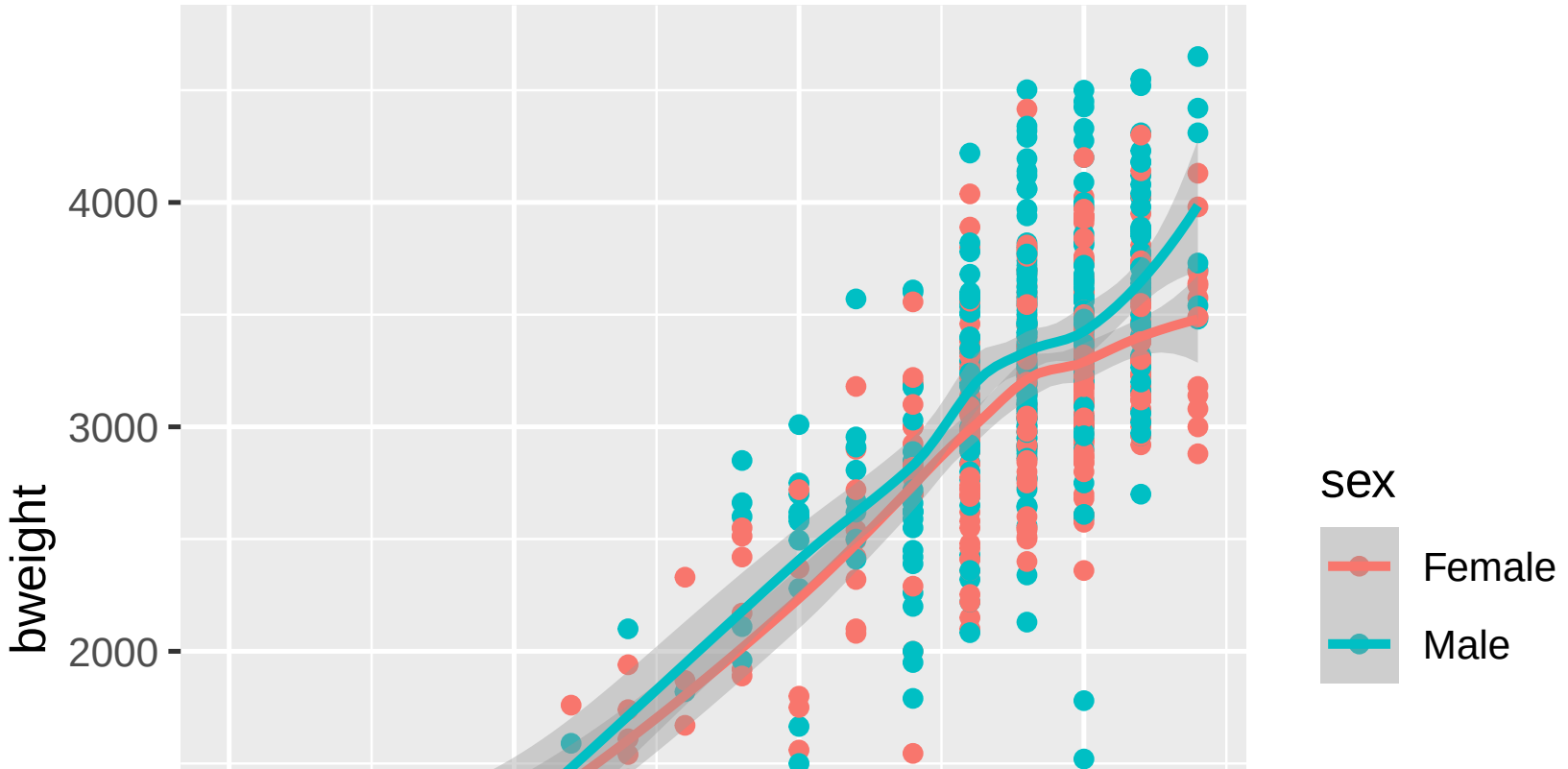
Add a point geom

```
# Start with data and aesthetics
ggplot(data = bw_df, mapping = aes(x = gestwks, y = bweight
    color = sex)) + # Add a point geom
geom_point()
```

Add a smooth geom

```
# Start with data and aesthetics
ggplot(data = bw_df, mapping = aes(x = gestwks, y = bweight
    color = sex)) + # Add a point geom
geom_point() + ## Add a smooth geom
geom_smooth()
```

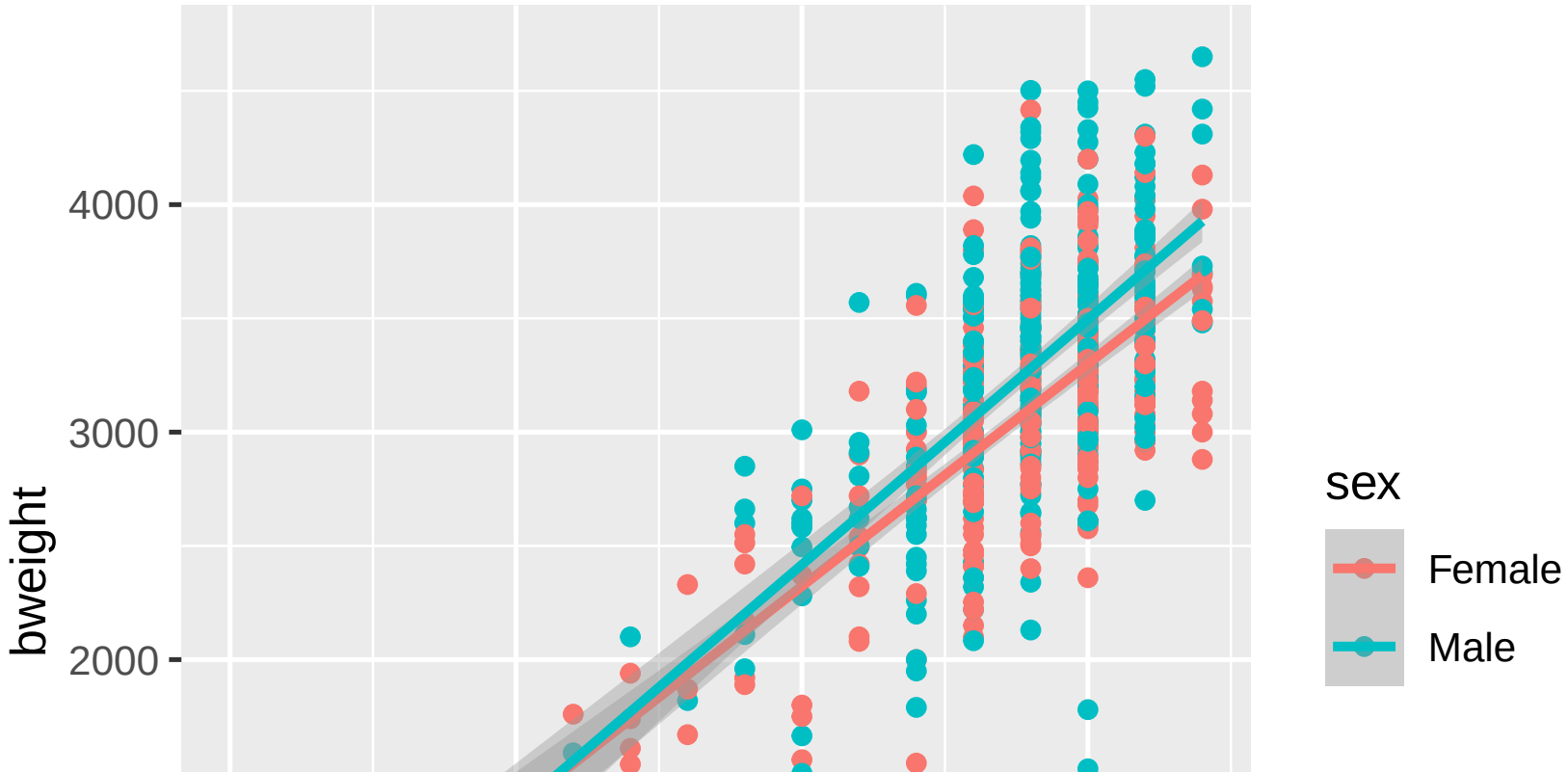
```
## `geom_smooth()` using method = 'loess' and formula 'y ~
```



Make the smooth geom straight

```
# Start with data and aesthetics
ggplot(data = bw_df, mapping = aes(x = gestwks, y = bweight
    color = sex)) + # Add a point geom
geom_point() + ## Add a smooth geom geom_smooth() + Make it
geom_smooth(method = "lm")
```

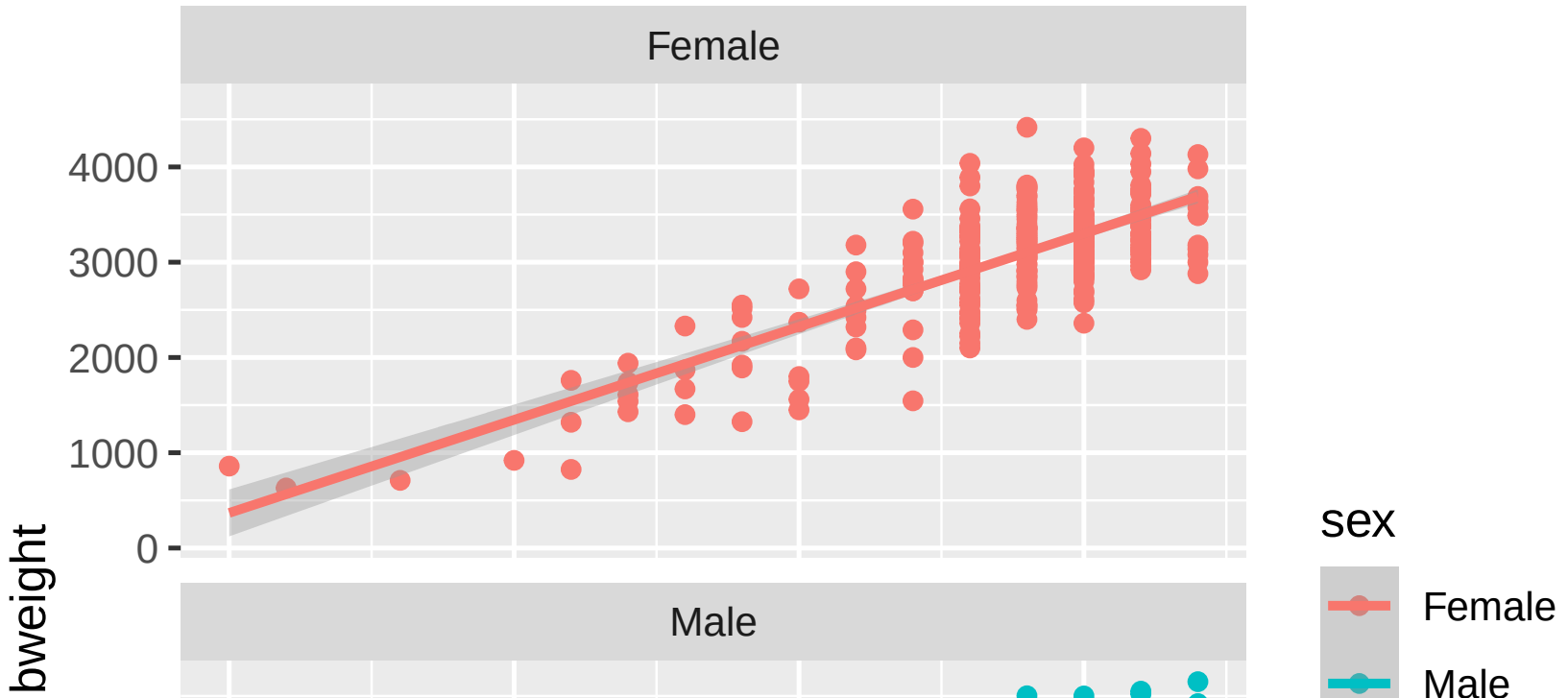
```
## `geom_smooth()` using formula 'y ~ x'
```



Facet by sex

```
# Start with data and aesthetics
ggplot(data = bw_df, mapping = aes(x = gestwks, y = bweight
    color = sex)) + # Add a point geom
geom_point() + ## Add a smooth geom geom_smooth() + Make it
geom_smooth(method = "lm") + # Facet by sex
facet_wrap(vars(sex), ncol = 1)
```

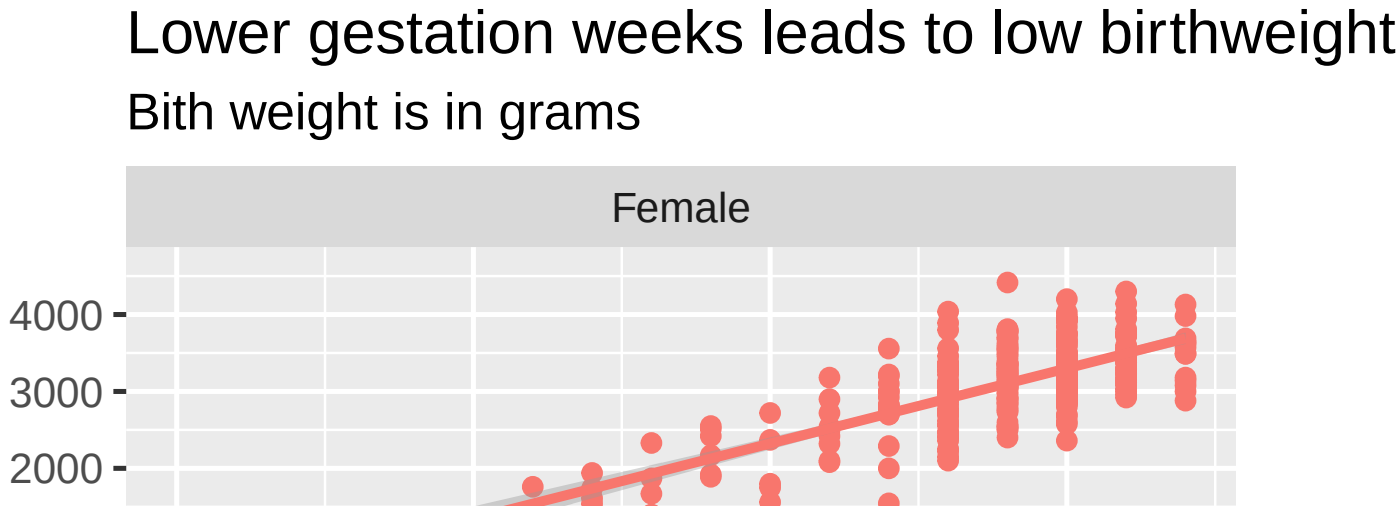
```
## `geom_smooth()` using formula 'y ~ x'
```



add labels

```
# Start with data and aesthetics
ggplot(data = bw_df, mapping = aes(x = gestwks, y = bweight
    color = sex)) + # Add a point geom
geom_point() + ## Add a smooth geom geom_smooth() + Make it
geom_smooth(method = "lm") + # Facet by sex
facet_wrap(vars(sex), ncol = 1) + ## add labels
labs(x = "Gestation weeks", y = "Birthweight", color = "Sex
    title = "Lower gestation weeks leads to low birthweight
    subtitle = "Bith weight is in grams", caption = "Is the
```

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



Themes

https://ggplot2.tidyverse.org/reference/ggtheme.html

```r
# Start with data and aesthetics
ggplot(data = bw_df, mapping = aes(x = gestwks, y = bweight
    color = sex)) + # Add a point geom
geom_point() + ## Add a smooth geom geom_smooth() + Make it
geom_smooth(method = "lm") + # Facet by sex
facet_wrap(vars(sex), ncol = 1) + ## add labels
labs(x = "Gestation weeks", y = "Birthweight", color = "Sex
    title = "Lower gestation weeks leads to low birthweight
    subtitle = "Bith weight is in grams", caption = "Is the
    # Add a theme
theme_bw()
```

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