Problem Set 02: Data Visualization

Your Name

Last modified on August 24, 2024 11:10:34 Eastern Daylight Time

This problem set will use the `ggplot2` package to generate graphics. "The Grammar of Graphics," is the theoretical basis for the `ggplot2` package. Much like how we construct sentences in any language by using a linguistic grammar (nouns, verbs, etc.), the grammar of graphics allows us to specify the components of a statistical graphic.

> **Note**
>
> In short, the grammar tells us that:
> A statistical graphic is a `mapping` of `data` variables to `aes`thetic attributes of `geom`etric objects.
> A graphic can be broken into three **essential** components:
>
> - `data`: the data-set comprised of variables that we plot
> - `geom`: the type of `geom`etric objects visible in a plot (points, lines, bars, etc.)
> - `aes`: aesthetic attributes of the geometric object that one perceives on a graphic. For example, x/y position, color, shape, and size. Each assigned aesthetic attribute can be mapped to a variable in our data-set.

Getting Set up

> **Directions**
>
> Type complete sentences to answer all questions in the Quarto document. Round all numeric answers you report to four decimal places. Use inline `R` code to report all numeric answers (i.e. do not hard code your numeric answers).
> Remember to save your work as you go along. Click the floppy disk (save current document) button in the upper left hand corner of the Quarto source panel.

Once you have opened the document:

- Change the author field to your First and Last name (example, `author: "John Smith"`).

`R` Packages

`R` Packages are like apps on a cell phone - they are tools for accomplishing common tasks. `R` is an open-source programming language, meaning that people can contribute packages that make our lives easier, and we can use them for free. For this problem set, the following `R` packages will be used:

- `dplyr`: for data wrangling
- `ggplot2`: for data visualization
- `readr`: for reading in data

The above packages are already installed on Appalachian's R Studio Server. **Every time** you open a new `R` session you need to **load (open)** any packages you want to use. Loading a package is done with the `library()` function.

R Code

```
library(dplyr)
library(ggplot2)
library(readr)
```

Remember, "running code means" telling `R` "do this". You tell `R` to do something by passing code through the console. You can run existing code many ways:

- re-typing code out directly in the console (most laborious method)
- copying and pasting existing code into the console and hitting enter (easier method)
- click on the green triangle in the code chunk (easiest method 1)
- highlight the code and select `Control-Enter` on a PC or `Command-Return` on a Mac (easiest method 2)

The Data

Today, we will practice data visualization using data on births from the state of North Carolina. The code below reads a `*.CSV` file from a URL into the object `nc`.

R Code

```
url <- "https://docs.google.com/spreadsheets/d/e/2PACX-1vTm2WZwNBoQdZhMgot7urbtu8eG7tzAq
if(!dir.exists("./data/"))
{dir.create("./data/")}
if(!file.exists("./data/nc.csv")){
    download.file(url, destfile = "./data/nc.csv")}
nc <- read_csv("./data/nc.csv")
```

The data set that displays in your Environment is a large **data frame**. Each observation or **case** is a birth of a single child.

The workspace area in the upper right hand corner of the R Studio window should now list a data set called `nc` with 800 observations (rows or cases) and 13 variables (columns).

How to Look at Data in `R`

Take a Glimpse

You can see the **dimensions of this data frame (# of rows and columns)**, the names of the variables, the variable types and the first few observations using the `glimpse` function.

R Code

```
glimpse(nc)
```

```
Rows: 800
Columns: 13
$ fage           <dbl> 19, 21, 18, 17, 20, 30, 21, 14, 16, 20, 18, 20, 20, 26,~
$ mage           <dbl> 15, 15, 15, 15, 16, 16, 16, 16, 16, 17, 17, 17, 17, 17,~
$ mature         <chr> "younger mom", "younger mom", "younger mom", "younger m~
$ weeks          <dbl> 37, 41, 37, 35, 37, 45, 38, 40, 24, 40, 37, 40, 39, 38,~
$ premie         <chr> "full term", "full term", "full term", "premie", "full ~
$ visits         <dbl> 11, 6, 12, 5, 13, 9, 15, 12, 5, 8, 10, 17, 9, 11, 10, 1~
$ marital        <chr> "married", "married", "married", "married", "married", ~
$ gained         <dbl> 38, 34, 76, 15, 52, 28, 75, 9, 12, 20, 39, 38, 36, 30, ~
$ weight         <dbl> 6.63, 8.00, 8.44, 4.69, 6.94, 7.44, 7.56, 5.81, 1.50, 8~
$ lowbirthweight <chr> "not low", "not low", "not low", "low", "not low", "not~
$ gender         <chr> "female", "male", "male", "male", "female", "male", "fe~
$ habit          <chr> "nonsmoker", "nonsmoker", "nonsmoker", "nonsmoker", "no~
```

```
$ whitemom      <chr> "white", "white", "not white", "not white", "white", "w~
```

We can see that there are 800 observations and 13 variables in this data set. It is good practice to see if `R` is treating variables as factors `<fct>`; as numbers `<int>` or `<dbl>` (basically numbers with decimals); or as characters (i.e. text) `<chr>`. The variable names are `fage`, `mage`, `mature`, etc. The output from `glimpse(nc)` tells us that six of the variables are numbers with decimals (`<dbl>`). The other seven variables are character (`<chr>`).

Problem 1

What type of variable is `R` considering the variable `habit` to be? What variable type is `visits`? (answer with text)

Problem 1 Answers

```
# Type your code and comments inside the code chunk
```

- Delete this and put your text answer here.

The Data Viewer

By clicking on the name `nc` in the *Environment* pane (upper right window), the data is displayed in the Source pane (upper left window) in the *Data Viewer*. `R` has stored these data in a kind of spreadsheet called a *data frame*. Each row represents a different birth: the first entry or column in each row is simply the row number, the rest are the different variables that were recorded for each birth. You can close the data viewer by clicking on the `x` in the appropriate tab.

Instructions

It is a good idea to try render your document from time to time as you go along. Go ahead, and make sure your document is rendering, and that your html file includes Exercise headers, text, and code. Note that rendering automatically saves your `*.qmd` file too.

Types of Graphs

Three types of graphs are explored in this problem set:

- scatterplots

- boxplots
- histograms

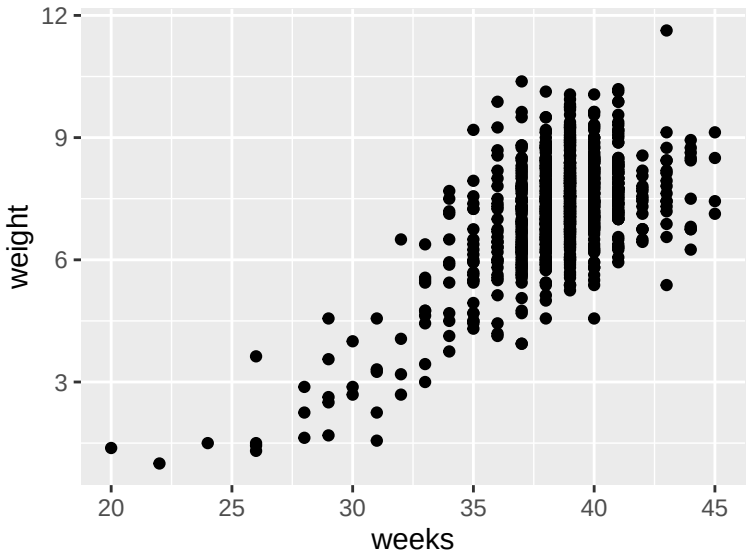
Scatterplots

Scatterplots allow you to investigate the relationship between two **numerical** variables. While you may already be familiar with this type of plot, let's view it through the lens of the Grammar of Graphics. Specifically, we will graphically investigate the relationship between the following two numerical variables in the `nc` data frame:

- `weeks`: length of a pregnancy on the horizontal "x" axis and
- `weight`: birth weight of a baby in pounds on the vertical "y" axis

R Code

```
ggplot(data = nc, aes(x = weeks, y = weight)) +
  geom_point()
```



Let's view this plot through the grammar of graphics. Within the `ggplot()` function call, we specified:

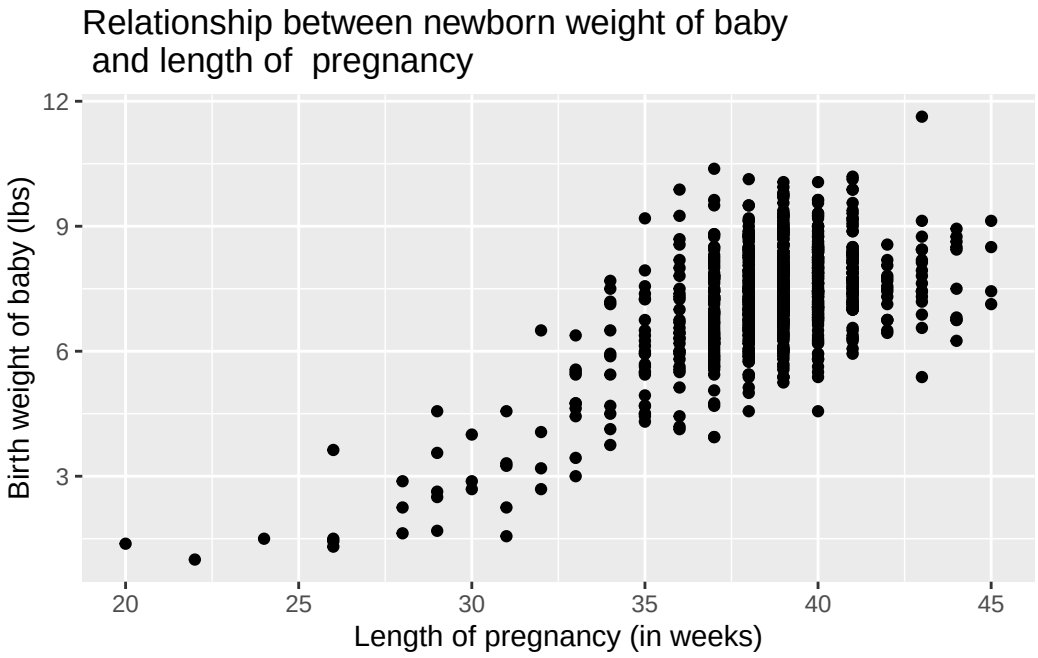
- The data frame to be `nc` by setting `data = nc`
- The `aes`thetic `mapping` by setting `aes(x = weeks, y = weight)`
- The variable `weeks` maps to the `x`-position `aes`thetic
- The variable `weight` maps to the `y`-position `aes`thetic

We also add a layer to the `ggplot()` function call using the `+` sign. The layer in question specifies the `geom`etric object as `point`s using `geom_point()`.

Finally, we can also add axes labels and a title to the plot as shown below. Again we add a new layer, this time a `labs` or labels layer.

R Code

```
ggplot(data = nc, aes(x = weeks, y = weight)) +
  geom_point() +
  labs(x = "Length of pregnancy (in weeks)",
       y = "Birth weight of baby (lbs)",
       title = "Relationship between newborn weight of baby \n and length of  pregnancy"
```



Problem 2

Is there a positive or negative relationship between `weight` and `weeks`? (text only to answer)

Problem 2 Answers

- Delete this and put your text answer here.

Problem 3

Make a graph showing `weeks` again on the x axis and the variable `gained` on the y axis (the amount of weight a mother gained during pregnancy). Include axis labels with measurement units, and a title. (code only to answer)

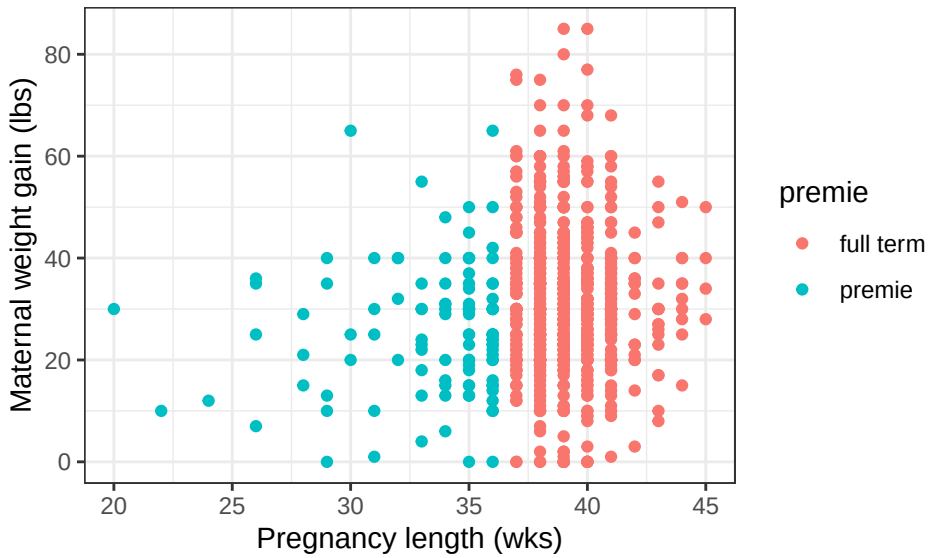
Problem 3 Answers

```
# Type your code and comments inside the code chunk
```

Problem 4

Study the code below, and the resulting graphical output. Note that I added a new argument of `color = premie` **inside** the `aes`thetic mapping. The variable `premie` indicates whether a birth was early (premie) or went full term. Please answer with text:
A. What did adding the argument `color = premie` accomplish?
B. How many **variables** are now displayed on this plot?
C. What appears to (roughly) be the pregnancy length cutoff for classifying a newborn as a "premie" versus a "full term".

```
ggplot(data = nc, aes(x = weeks, y = gained, color = premie))+
  geom_point() +
  labs(x = "Pregnancy length (wks)", y = "Maternal weight gain (lbs)") +
  theme_bw()
```



Problem 4 Answers

A. Delete this and put your text answer here.
B. Delete this and put your text answer here.
C. Delete this and put your text answer here.

```
nc %>%
  group_by(premie) %>%
  summarize(Max = max(weeks), Min = min(weeks))
```

```
# A tibble: 2 x 3
  premie      Max   Min
  <chr>     <dbl> <dbl>
1 full term    45    37
2 premie       36    20
```

Problem 5

Make a new scatterplot that shows a mothers age on the x axis (variable called `mage`) and birth weight of newborns on the y axis (`weight`). Color the points on the plot based on the gender of the resulting baby (variable called `gender`). Does there appear to be any strong relationship between a mother's age and the weight of her newborn? (code and text to answer)

Problem 5 Answers

```
# Type your code and comments inside the code chunk
```

- Delete this and put your text answer here.

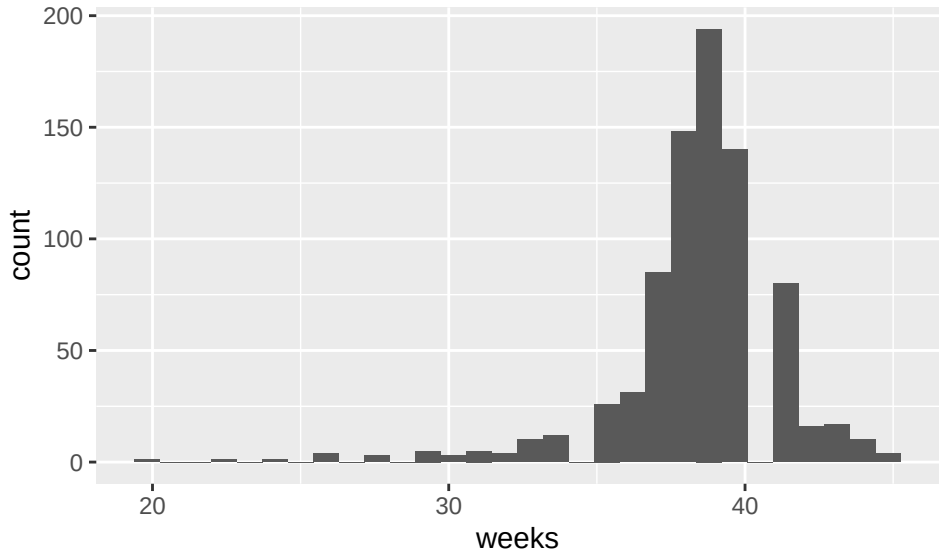
Histograms

Histograms are useful plots for showing how many elements of a **single numerical** variable fall in specified bins. This is a very useful way to get a sense of the **distribution** of your data. Histograms are often one of the first steps in exploring data visually.

For instance, to look at the distribution of pregnancy duration (variable called `weeks`), consider the following code:

R Code

```
ggplot(data = nc, aes(x = weeks))+
  geom_histogram()
```



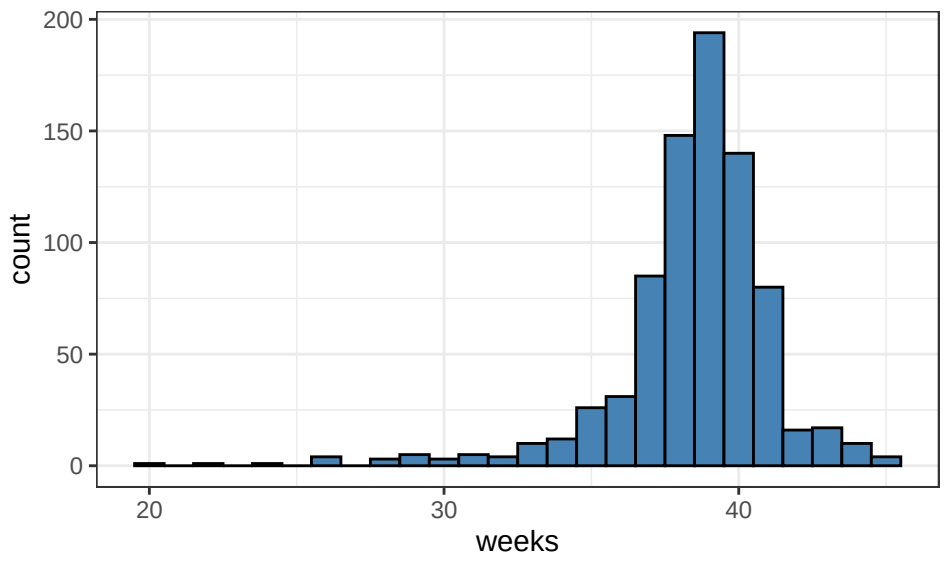
A few things to note here:

- There is only one variable being mapped in `aes()`: the single numerical variable `weeks`. You don't need to compute the `y-aes`thetic: `R` calculates it automatically.
- We set the geometric object as `geom_histogram()`
- The warning message encourages us to specify the number of bins on the histogram, as `R` chose 30 for us.

We can change the binwidth (and thus the number of bins), and the colors as shown next.

R Code

```
ggplot(data = nc, aes(x = weeks))+
  geom_histogram(binwidth = 1, color = "black", fill = "steelblue") +
  theme_bw()
```



Note that none of these arguments went inside the `aes`thetic `mapping` argument as they do not specifically represent mappings of variables.

Problem 6

Inspect the histogram of the `weeks` variable. Answer each of the following with **text**.
A. The y axis is labeled **count**. What is specifically being counted in this case? Hint: think about what each case is in this data set.
B. What appears to be roughly the average length of pregnancies in weeks?
C. If we changed the binwidth to 100, how many bins would there be? Roughly how many cases would be in each bin?

Problem 6 Answers

A. Delete this and put your text answer here.
B. Delete this and put your text answer here.

```
# Type your code and comments inside the code chunk
```

C. Delete this and put your text answer here.

Problem 7

Make a histogram of the birth `weight` of newborns (which is in lbs), include a descriptive title and axis labels. Make the bins lightblue with a blue border and a binwidth of 1. (code only to answer)

Problem 7 Answers

```
# Type your code and comments inside the code chunk
```

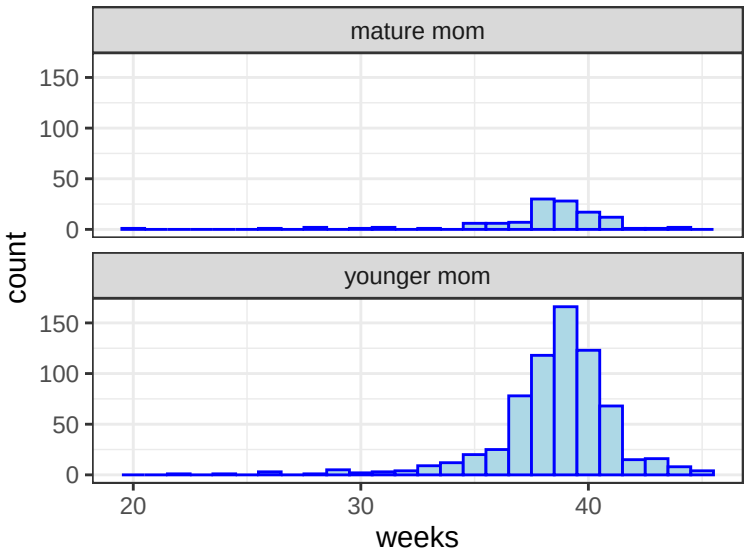
Faceting

Faceting is used to create small multiples of the same plot over a different categorical variable. By default, all of the small multiples will have the same vertical axis.

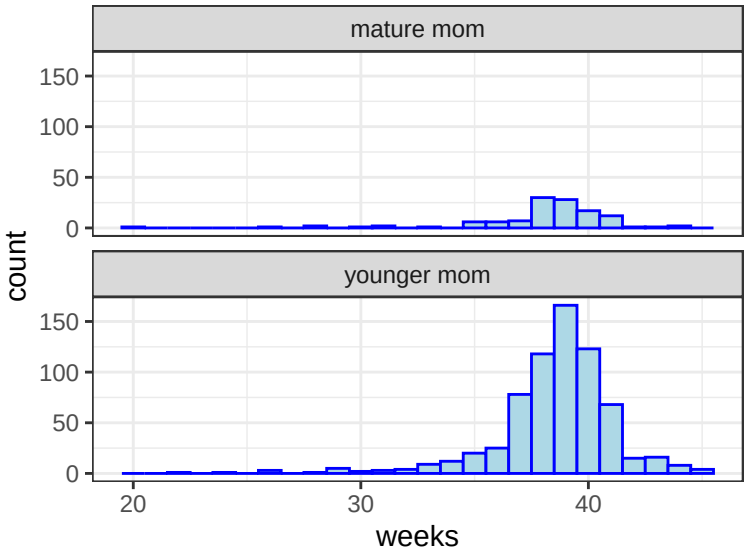
For example, suppose we are interested in looking at whether pregnancy length varies by the maturity status of a mother (column name `mature`). This is what is meant by "the distribution of one variable over another variable": `weeks` is one variable and `mature` is the other variable. In order to look at histograms of `weeks` for older and more mature mothers, add a plot layer using `facet_wrap(~ mature, ncol = 1)`. The `ncol = 1` argument tells `R` to stack the two histograms into one column.

R Code

```
ggplot(data = nc, aes(x = weeks)) +
  geom_histogram(binwidth = 1, color = "blue", fill = "lightblue") +
  facet_wrap(~ mature, ncol = 1) +
  theme_bw()
```



```
# Or
ggplot(data = nc, aes(x = weeks)) +
  geom_histogram(binwidth = 1, color = "blue", fill = "lightblue") +
  facet_wrap(facets = vars(mature), ncol = 1) +
  theme_bw()
```



Problem 8

Make a histogram of newborn birth `weight` split by `gender` of the child. Set the binwidth to 0.5. Which gender appears to have a slightly larger average birth weight? **Extra Credit:** Have the bins of the female histogram a different color than the bins of the male histogram. Outline the bins of both histograms in black. (code and text to answer)

Problem 8 Answers

```
# Type your code and comments inside the code chunk
```

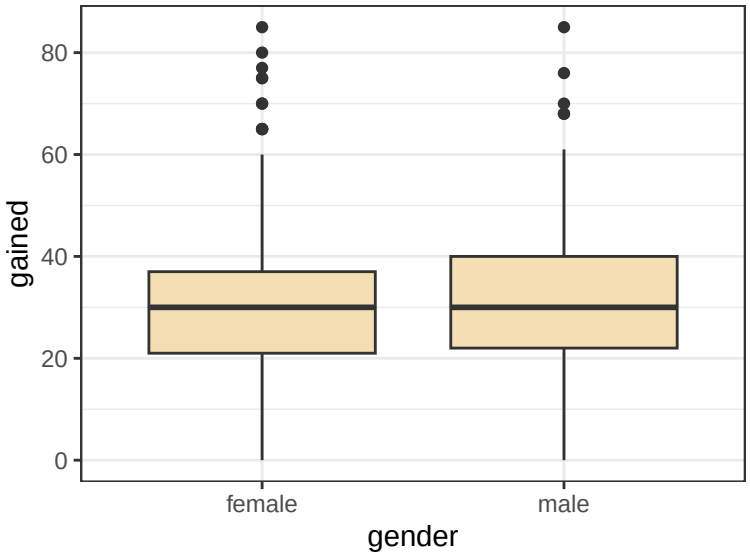
- Delete this and put your text answer here.

Boxplots

While histograms can help to show the distribution of data, boxplots have much more flexibility and can provide even more information in a single graph. The y `aes`thetic is the numeric variable you want to include in the boxplot, and the x `aes`thetic is a grouping variable. For instance, below `gender` is the `aes`thetic `mapping` for x, and `gained` is the `aes`thetic `mapping` for y. This creates a boxplot of the weight gained for mothers that had male and female newborns. Note that the `fill` argument is not necessary, but sets a color for the boxplots.

R Code

```
ggplot(data = nc, aes(x = gender, y = gained)) +
  geom_boxplot(fill = "wheat") +
  theme_bw()
```



Review

For review, these are the different parts of the boxplot: ’

- The bottom of the “box” portion represents the 25th percentile (1st quartile)
- The horizontal line in the “box” shows the median (50th percentile, 2nd quartile)
- The top of the “box” represents the 75th percentile (3rd quartile)
- The height of each “box”, i.e. the value of the 3rd quartile minus the value of the 1st quartile, is called the interquartile range (IQR). It is a measure of spread of the middle 50% of values. Longer boxes indicating more variability.

- The "whiskers" extending out from the bottoms and tops of the boxes represent points less than the 25th percentile and greater than the 75th percentiles respectively. They extend out **no more than** 1.5 x IQR units away from either end of the boxes. The length of these whiskers show how the data outside the middle 50% of values vary. Longer whiskers indicate more variability.
- The dots represent values falling outside the whiskers or outliers. The definition of an outlier is somewhat arbitrary and not absolute. In this case, they are defined by the length of the whiskers, which are no more than 1.5 x IQR units long.

Problem 9

Make a boxplot of the weight `gained` by moms, split by the maturity status of the mothers (`mature`). Include axis labels and a title on your plot. Is the **median** weight gain during pregnancy larger for younger or older moms? (text and code)

Problem 9 Answers

```
# Type your code and comments inside the code chunk
```

- Delete this and put your text answer here.

Problem 10

Make a boxplot of pregnancy duration in `weeks` by smoking `habit`. Is the duration of pregnancy more **variable** for smokers or non-smokers? (i.e. which group has the greater spread for the variable `weeks`? Make sure to specify how you are measuring spread in your answer).

Problem 10 Answers

```
# Type your code and comments inside the code chunk
```

- Delete this and put your text answer here.

More Practice

For the following, determine which type of plot to use, **make the plot** and answer any questions with **text**. There is a table at the end of this document that can help you determine

which plot to use given the question/types of variables.

Problem 11

Using a data visualization, visually assess: Is the variable for father's age (`fage`) symmetrical, or does it have a skew?

Problem 11 Answers

```
# Type your code and comments inside the code chunk
```

- Delete this and put your text answer here.

Problem 12

Using a data visualization, visually assess: (in this sample) is the median birth `weight` of babies greater for white or non-white mothers (variable called `whitemom`)?

Problem 12 Answers

```
# Type your code and comments inside the code chunk
```

- Delete this and put your text answer here.

Problem 13

Using a data visualization, visually assess: (in this sample) as a mother's age (`mage`) increases, does the duration of pregnancy (`weeks`) appear to decrease?

Problem 13 Answers

```
# Type your code and comments inside the code chunk
```

- Delete this and put your text answer here.

Data Visualization Table

This table is a great resource for thinking about how to visualize data.

TABLE 3.5: Summary of 5NG

	Named graph	Shows	Geometric object	Notes
1	Scatterplot	Relationship between 2 numerical variables	`geom_point()`	
2	Linegraph	Relationship between 2 numerical variables	`geom_line()`	Used when there is a sequential order to x-variable e.g. time
3	Histogram	Distribution of 1 numerical variable	`geom_histogram()`	Facetted histogram shows distribution of 1 numerical variable split by 1 categorical variable
4	Boxplot	Distribution of 1 numerical variable split by 1 categorical variable	`geom_boxplot()`	
5	Barplot	Distribution of 1 categorical variable	`geom_bar()` when counts are not pre-counted	Stacked & dodged barplots show distribution of 2 categorical variables
			`geom_col()` when counts are pre-counted	

Table 3.5 from Modern Dive

Turning in Your Work

You will need to make sure you commit and push all of your changes to the github education repository where you obtained the lab.

Tip

- Make sure you **render a final copy with all your changes** and work.
- Look at your final html file to make sure it contains the work you expect and is formatted properly.

Logging out of the Server

There are many statistics classes and students using the Server. To keep the server running as fast as possible, it is best to sign out when you are done. To do so, follow all the same steps for closing Quarto document:

> Tip
>
> - Save all your work.
> - Click on the orange button in the far right corner of the screen to quit `R`
> - Choose **don't save** for the **Workspace image**
> - When the browser refreshes, you can click on the sign out next to your name in the top right.
> - You are signed out.

```
sessionInfo()
```

```
R version 4.4.1 (2024-06-14)
Platform: x86_64-redhat-linux-gnu
Running under: Red Hat Enterprise Linux 9.4 (Plow)

Matrix products: default
BLAS/LAPACK: FlexiBLAS OPENBLAS-OPENMP;  LAPACK version 3.9.0

locale:
 [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
 [3] LC_TIME=en_US.UTF-8        LC_COLLATE=en_US.UTF-8
 [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
 [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
 [9] LC_ADDRESS=C               LC_TELEPHONE=C
[11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C

time zone: America/New_York
tzcode source: system (glibc)

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
[1] readr_2.1.5        nycflights13_1.0.2 dplyr_1.1.4        ggplot2_3.5.1
[5] knitr_1.48
```

```
loaded via a namespace (and not attached):
 [1] bit_4.0.5          gtable_0.3.5       jsonlite_1.8.8     compiler_4.4.1
 [5] crayon_1.5.3       tinytex_0.52       tidyselect_1.2.1   parallel_4.4.1
 [9] scales_1.3.0       yaml_2.3.10        fastmap_1.2.0      R6_2.5.1
[13] labeling_0.4.3     generics_0.1.3     tibble_3.2.1       munsell_0.5.1
[17] pillar_1.9.0       tzdb_0.4.0         rlang_1.1.4        utf8_1.2.4
[21] xfun_0.47          bit64_4.0.5        cli_3.6.3          withr_3.0.1
[25] magrittr_2.0.3     digest_0.6.36      grid_4.4.1         vroom_1.6.5
[29] rstudioapi_0.16.0 hms_1.1.3           lifecycle_1.0.4    vctrs_0.6.5
[33] evaluate_0.24.0    glue_1.7.0         farver_2.1.2       fansi_1.0.6
[37] colorspace_2.1-1   rmarkdown_2.28     tools_4.4.1        pkgconfig_2.0.3
[41] htmltools_0.5.8.1
```