# Problem Set 02: Data Visualization

## Contents



In this problem set we will use the `ggplot2` package to generate graphics. The "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.

In short, the grammar tells us that:

> A statistical graphic is a `mapping` of `data` variables to `aes`thetic attributes of `geom`etric objects.

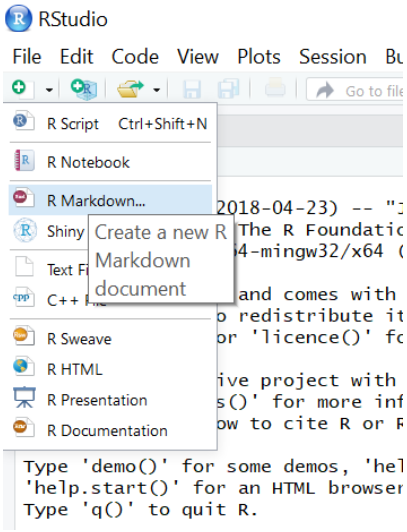
We can break a graphic into the following three **essential** components:

- `data`: the data-set comprised of variables that we plot
- `geom`: this refers to our type of `geom`etric objects we see in our plot (points, lines, bars, etc.)
- `aes`: aesthetic attributes of the geometric object that we can perceive 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

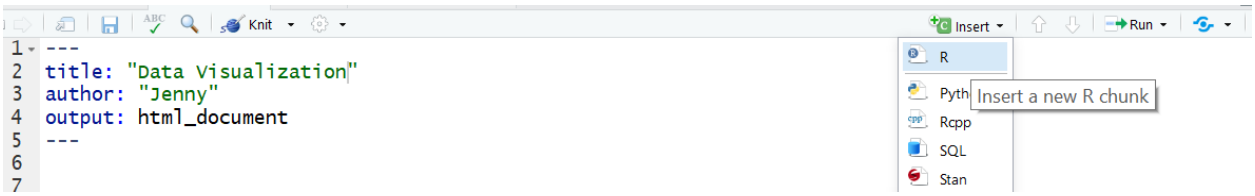
Go ahead and launch RStudio, and open a new R Markdown file. Recall, to open a new R Markdown (.Rmd) file:

- Click on the little green plus on the upper left of the window.
- Select R Markdown, as in the image below.




Once you have opened the document:

- Change the title at the top to "Data Visualization". Be sure to keep the quotation marks.
- Add an author line, following the example below. You need quotation marks!
- Delete any extra code (everything from line 6 down).




Finally, save your new R Markdown document:

- Click File > Save As…
- Browse to the folder where you want to store the R Markdown document and its output.
- Name the file something informative, e.g., `PS01_lastname_firstname` (fill in your first name and last name). Make sure to keep the `.Rmd` file extension.

You will hand in a **knitted pdf file** as your problem set. It is OK if your lab report includes the example code from the lab, as well as your Exercises. Just be sure to **make a header to label each Exercise**. Please type your code to answer the questions in a code chunk (gray part), under the exercise headers and type (**BRIEF**) answers to any interpretation questions in the white part under the headers.

### R Packages

For this problem set we will use the following R packages:

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

Copy, paste and run the following in a code chunk (see the figure above if you forget how to insert a code chunk).

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

Remember, "running code means" telling R "do this". You tell R to do something by passing it 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 hit `Control-Enter` on 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. Copy, paste and run the code below to load the data.

```
nc <- read_csv("https://docs.google.com/spreadsheets/d/e/2PACX-1vTm2WZwNBoQdZhMgot7urbtu8eG7tzAq-60ZJsQ
```

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). Each observation or **case** is a birth of a single child.

## 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. Copy, paste, and run the following in a new code chunk.

```
glimpse(nc)
```

We can see that there are 1000 observations and 13 variables in this data set. The variable names are `fage`, `mage`, `mature`, etc. This output also tells us that some variables are numbers…some specifically integers `<int>`, others are numbers with decimals `<dbl>`. Some of the variables are factors `<fct>`. It is a 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>`.

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

#### The data viewer

You can view the data by clicking on the name `nc` in the *Environment* pane (upper right window). This will bring up an alternative display of the data set in the *Data Viewer* (upper left window). 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 upper left hand corner.

It is a good idea to try kitting your document from time to time as you go along! Go ahead, and make sure your document is knitting, and that your knitted file includes Exercise headers, text, and code. Note that knitting automatically saves your Rmd file too!

## Types of Graphs

We will explore three different types of graphs in this problem set.

- Scatterplots
- Boxplots
- Histograms

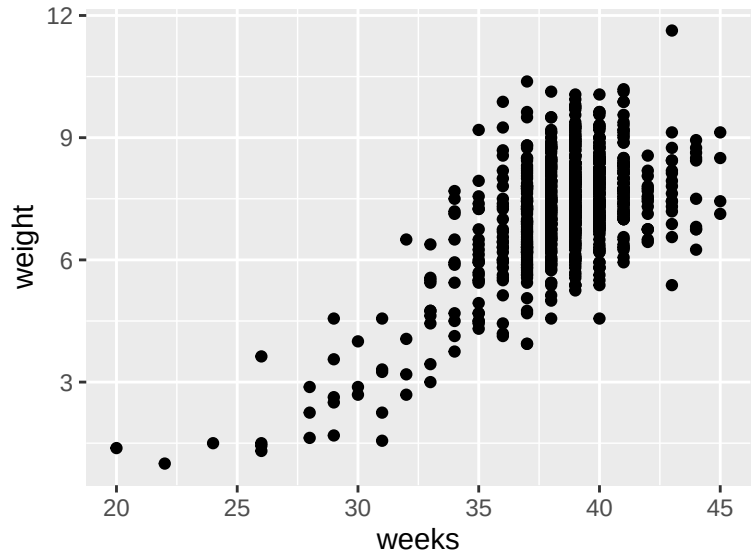
### Scatterplots

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

```
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 here as `points`, by specifying `geom_point()`.

Finally, we can also add axis labels and a title to the plot like so. Again we add a new layer, this time a `labs` or labels layer.

```
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 pregnancy duration and newborn weight")
```

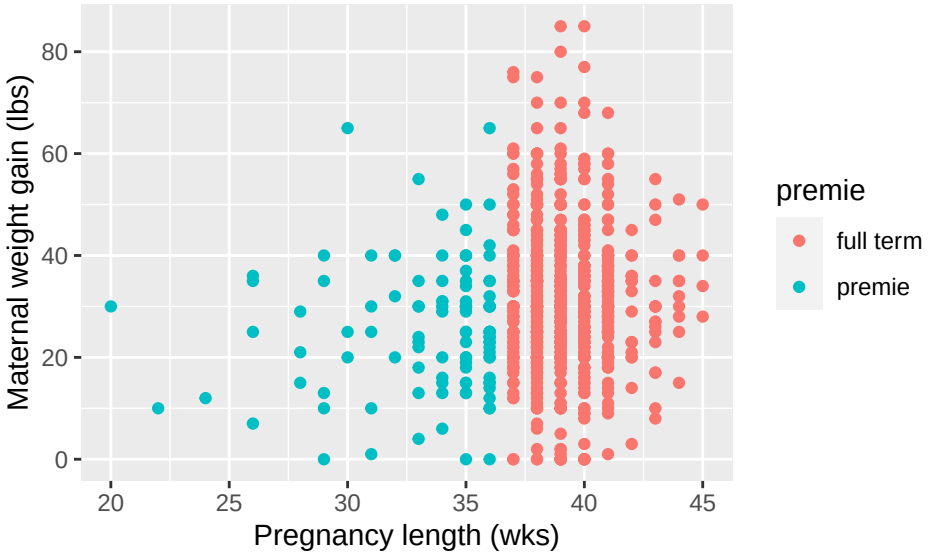
1. Is there a positive or negative relationship between these variables? (Text only to answer)
2. Make a graph showing `weeks` again on the x axis and the variable `gained` on the y axis (the amount of weight a mother gained during pregnancy). Include axis labels with measurement units, and a title. (R code only to answer)
3. 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)")
```




1. 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? (R code and text to answer)

Make sure your document is knitting, and that your knitted file includes Exercise headers, text, and code. Note that knitting automatically saves your Rmd file too!

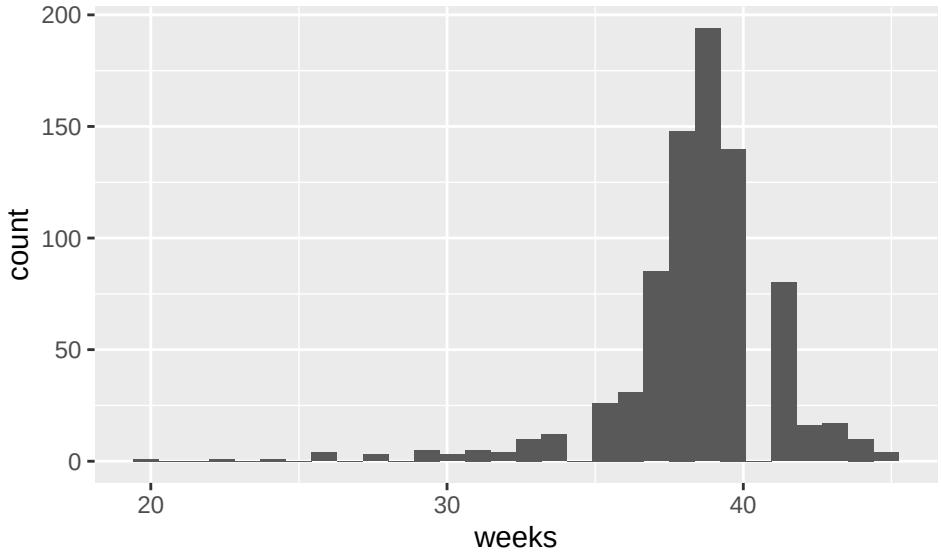
### Histograms

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`), copy, paste and run the following in a new code chunk:

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

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

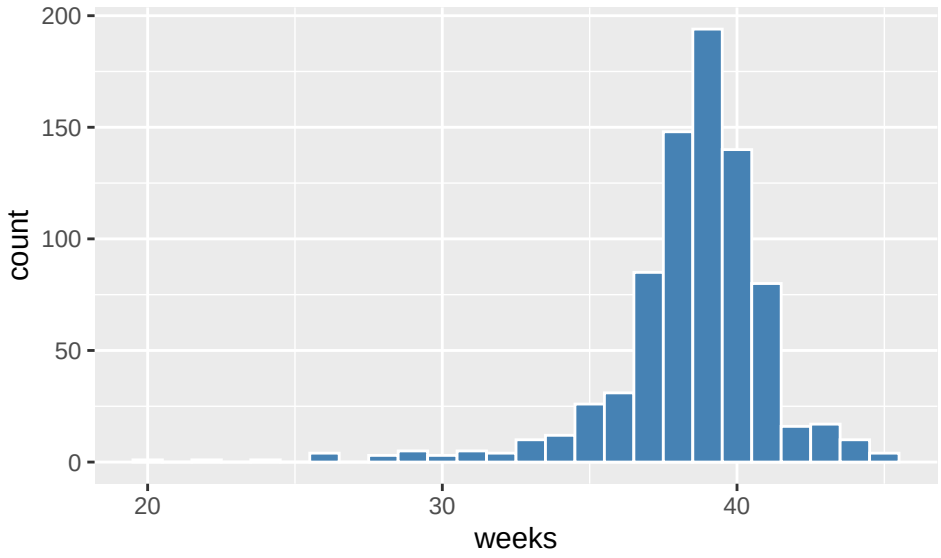



A few things to note here:

- There is only one variable being mapped in `aes()`: the single numerical variable `weeks`. You don't need to compute the `y`-aesthetic: 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), as well as the colors like so.

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




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

1. 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?

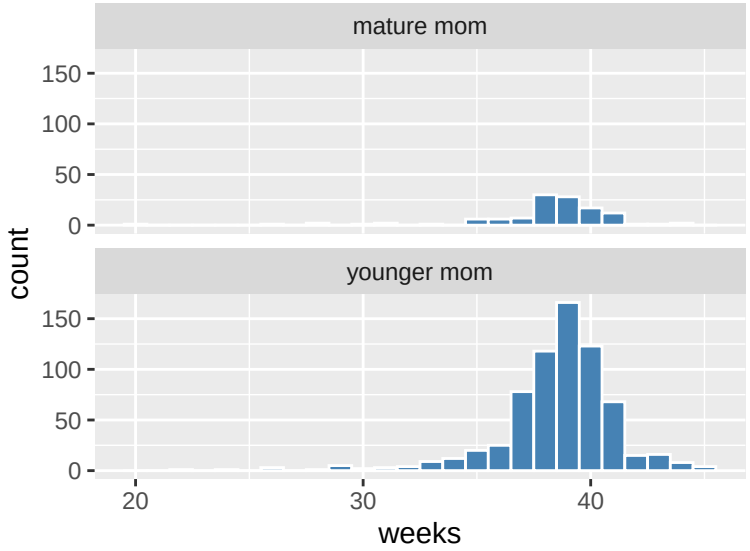
2. Make a histogram of the birth `weight` of newborns (which is in lbs), including a title and axis labels. (code only to answer)

#### Faceting

Faceting is used when we'd like 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 were interested in looking at whether pregnancy length varied 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, we add a plot layer `facet_wrap(~ mature, ncol = 1)`. The `ncol = 1` argument just tells R to stack the two histograms into one column.

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



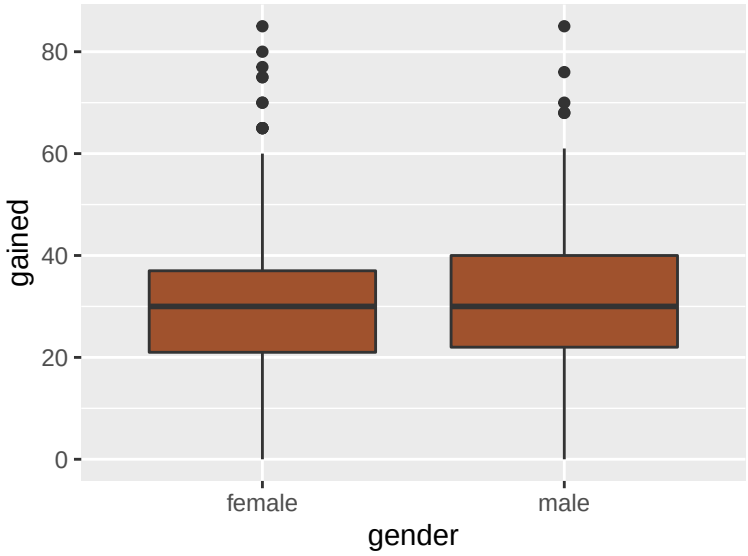

1. Make a histogram of newborn birth `weight` split by `gender` of the child. Set the binwidth to 0.5. Which gender appears to have a slightly larger average birth weight? (Code and text to answer)

Make sure your document is knitting, and that your knitted file includes Exercise headers, text, and code. Note that knitting automatically saves your Rmd file too!

### 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 `aesthetic` is the numeric variable you want to include in the boxplot, and the x `aesthetic` is a grouping variable. For instance, below we set `gender` as the `aesthetic` `mapping` for x, and `gained` as the `aesthetic` `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.

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




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.

8. 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)

9. 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`?). (Code and text to answer)

Make sure your document is knitting, and that your knitted file includes Exercise headers, text, and code. Note that knitting automatically saves your Rmd file too!

## Independent Practice

For the following, you need to determine which type of plot to use, **make the plot**, and answer any questions with **text**. There is a table below that can help you determine which plot to use, given the question/types of variables.

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

2. 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`)?

3. 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?

---

**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 http://moderndive.netlify.com/index.html