So you want to make a plot?

The plotting steps vary by dataset and project.

But you should consider four things:

1. What data do you have?
2. What do you want to know about the data?
3. What visualization methods should you use?
4. What do you see and does it make sense?
Start here

What data do you have?

None or next to nothing

Best data ever

What do you want to know about your data?

Don’t know yet

(Specific question)

New questions arise

What do you see and does it make sense?

Find related data

Bar chart
Pie chart
Line plot
Scatterplot
Treemap
etc

What visualization methods should you use?

Explore different dimension

The iterative data exploration process by Nathan Yau
What data do you have?

How many variables?
1. One variable
2. Two variables
3. Three or more

What type of variables?
4. Quantitative, qualitative, time
What do you want to know about your data?

- Part-to-whole analysis
- Ranking analysis
- Deviation analysis
- Times series (trends in time)
- Distribution analysis
- Correlation analysis
- Multivariate analysis
What do you do see?

- Systematic variation
- Increasing patterns
- Decreasing patterns
- Atypical or outliers
- Noise?
Visualization is simply mapping data to geometry and color
Visualization is simply mapping data to geometry and color.
Title of this Graph

A description of the data or something worth highlighting to set the stage

Source: Somewhere reputable
Visual Cues
Encoding data with shapes, colors, and sizes. Which cues you choose depends on your data and your goals
Coordinate System
Mapping data requires a system of coordinates: cartesian, polar, etc
Scale
Increments that make sense can increase readability as well as shift focus
Title of this Graph
A description of the data or something worth highlighting to set the stage

Context
If your audience is unfamiliar with the data, it’s your job to clarify what values represent and explain how people should read your plot

Source: Somewhere reputable
Title of this Graph

A description of the data or something worth highlighting to set the stage

Source: Somewhere reputable
Chart Elements
Manufacturers are posting jobs, not filling them
Change since June 2009, seasonally adjusted

SOURCE: BUREAU OF LABOR STATISTICS
Manufacturers are posting jobs, not filling them
Change since June 2009, seasonally adjusted

Source: Bureau of Labor Statistics
Manufacturers are posting jobs, not filling them
Change since June 2009, seasonally adjusted
Manufacturers are posting jobs, not filling them
Change since June 2009, seasonally adjusted

Openings
Hires

tick marks

SOURCE: BUREAU OF LABOR STATISTICS
Manufacturers are posting jobs, not filling them
Change since June 2009, seasonally adjusted
Manufacturers are posting jobs, not filling them
Change since June 2009, seasonally adjusted
Another example
New Yorkers pay up for a shorter commute
Median monthly NYC rent in 2015 vs. commute time by subway

Rental prices are based on FiveThirtyEight's analysis of 2015 listings on StreetEasy. Commutes are calculated as the average time from the subway station nearest a home to the nearest 42nd Street and Chambers Street stations. Commutes without at least 10 listings are excluded.
New Yorkers pay up for a shorter commute

Median monthly NYC rent in 2015 vs. commute time by subway

Rental prices are based on FiveThirtyEight's analysis of 2015 listings on StreetEasy. Commutes are calculated as the average time from the subway station nearest a home to the nearest 42nd Street and Chambers Street stations. Commutes without at least 10 listings are excluded.
New Yorkers pay up for a shorter commute

Median monthly NYC rent in 2015 vs. commute time by subway

Rental prices are based on FiveThirtyEight’s analysis of 2015 listings on StreetEasy. Commutes are calculated as the average time from the subway station nearest a home to the nearest 42nd Street and Chambers Street stations. Commutes without at least 10 listings are excluded.
New Yorkers pay up for a shorter commute
Median monthly NYC rent in 2015 vs. commute time by subway

Rental prices are based on FiveThirtyEight's analysis of 2015 listings on StreetEasy. Commutes are calculated as the average time from the subway station nearest a home to the nearest 42nd Street and Chambers Street stations. Commutes without at least 10 listings are excluded.
New Yorkers pay up for a shorter commute
Median monthly NYC rent in 2015 vs. commute time by subway

Rental prices are based on FiveThirtyEight’s analysis of 2015 listings on StreetEasy. Commutes are calculated as the average time from the subway station nearest a home to the nearest 42nd Street and Chambers Street stations. Commutes without at least 10 listings are excluded.
Grammar of Graphics with “ggplot2”
The Grammar of Graphics
About the grammar of graphics

The Grammar of Graphics is Wilkinson’s attempt to define a theoretical framework for graphics.

Grammar: formal system of rules for generating graphics:

- Some rules are mathematic
- Some rules are aesthetic (i.e. visual)
Aesthetics ≠ Beauty

Aesthetics (GG): attributes of the geometric objects
Meaning of aesthetic in the Grammar of Graphics

Aesthetics: pertaining to sense perception

Aisthesthai = perceive

**GG aesthetic attributes**: visual properties that affect the way observations are displayed
About the grammar of graphics

Three stages of graphic creation

**Specification:** link data to graphic objects

**Assembly:** put everything together

**Display:** render of a graphic
R package

ggplot2
Resources

Documentation: [http://docs.ggplot2.org](http://docs.ggplot2.org)

Book: *ggplot2: Elegant Graphics for Data Analysis* by Hadley Wickham

Book: *R Graphics Cookbook* by Winston Chang

RStudio ggplot2 cheat sheet

ggplot2 book
ggplot2 book
R package “ggplot2”

Remember to install ggplot2 (just once)

install.packages("ggplot2")

library(ggplot2)

?ggplot
About ggplot2

“ggplot2” is an R package for producing statistical graphics.

It provides a framework based on Leland Wilkinson’s Grammar of Graphics.

“ggplot2” provides beautiful plots while taking care of fiddly details like legends, axes, colors.
About ggplot2

Default appearance of plots carefully chosen

Designed with visual perception in mind

Inclusion of some components, like legends, are automated

Great flexibility for annotating, editing, and embedding output
base graphics

\begin{figure}
\centering
\includegraphics[width=0.4\textwidth]{base_graphics}
\caption{Plot generated by base graphics package.}
\end{figure}

\begin{figure}
\centering
\includegraphics[width=0.4\textwidth]{ggplot2}
\caption{Plot generated by ggplot2 package.}
\end{figure}
About ggplot2

“ggplot2” is the name of the package (don’t forget the 2)

The gg in ggplot2 stands for Grammar of Graphics

Inspired in the Grammar of Graphics by Lee Wilkinson

`ggplot()` is the main function in “ggplot2”
ggplot2 philosophy:
Describe a wide range of graphics with a compact syntax and independent components
What is a Statistical Graphic?
## Data set mtcars

<table>
<thead>
<tr>
<th></th>
<th>mpg</th>
<th>hp</th>
<th>cyl</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mazda RX4</td>
<td>21.0</td>
<td>110</td>
<td>6</td>
</tr>
<tr>
<td>Mazda RX4 Wag</td>
<td>21.0</td>
<td>110</td>
<td>6</td>
</tr>
<tr>
<td>Datsun 710</td>
<td>22.8</td>
<td>93</td>
<td>4</td>
</tr>
<tr>
<td>Hornet 4 Drive</td>
<td>21.4</td>
<td>110</td>
<td>6</td>
</tr>
<tr>
<td>Hornet Sportabout</td>
<td>18.7</td>
<td>175</td>
<td>8</td>
</tr>
<tr>
<td>Valiant</td>
<td>18.1</td>
<td>105</td>
<td>6</td>
</tr>
<tr>
<td>Duster 360</td>
<td>14.3</td>
<td>245</td>
<td>8</td>
</tr>
<tr>
<td>Merc 240D</td>
<td>24.4</td>
<td>62</td>
<td>4</td>
</tr>
<tr>
<td>Merc 230</td>
<td>22.8</td>
<td>95</td>
<td>4</td>
</tr>
<tr>
<td>Merc 280</td>
<td>19.2</td>
<td>123</td>
<td>6</td>
</tr>
</tbody>
</table>
Miles per gallon –vs– Horsepower
Elements to draw the chart “manually”

Coordinate system
x and y axes
Axis tick marks
Axis labels, and title
Points (of a given size and color)
Regression line (and ribbon)
Legend
A statistical graphic is ...

A mapping from data to aesthetic attributes (color, shape, size) of geometric objects (points, lines, bars)

A plot may also contain statistical transformations of the data

A plot is drawn on a specific coordinate system

Sometime faceting can be used to get the same plot for different subsets of the dataset
Example
<table>
<thead>
<tr>
<th>name</th>
<th>gender</th>
<th>height</th>
<th>weight</th>
<th>jedi</th>
<th>species</th>
<th>weapon</th>
</tr>
</thead>
<tbody>
<tr>
<td>Luke Skywalker</td>
<td>male</td>
<td>1.72</td>
<td>77</td>
<td>jedi</td>
<td>human</td>
<td>lightsaber</td>
</tr>
<tr>
<td>Leia Skywalker</td>
<td>female</td>
<td>1.5</td>
<td>49</td>
<td>no_jedi</td>
<td>human</td>
<td>blaster</td>
</tr>
<tr>
<td>Obi-Wan Kenobi</td>
<td>male</td>
<td>1.82</td>
<td>77</td>
<td>jedi</td>
<td>human</td>
<td>lightsaber</td>
</tr>
<tr>
<td>Han Solo</td>
<td>male</td>
<td>1.8</td>
<td>80</td>
<td>no_jedi</td>
<td>human</td>
<td>blaster</td>
</tr>
<tr>
<td>R2-D2</td>
<td>male</td>
<td>0.96</td>
<td>32</td>
<td>no_jedi</td>
<td>droid</td>
<td>unarmed</td>
</tr>
<tr>
<td>C-3PO</td>
<td>male</td>
<td>1.67</td>
<td>75</td>
<td>no_jedi</td>
<td>droid</td>
<td>unarmed</td>
</tr>
<tr>
<td>Yoda</td>
<td>male</td>
<td>0.66</td>
<td>17</td>
<td>jedi</td>
<td>yoda</td>
<td>lightsaber</td>
</tr>
<tr>
<td>Chewbacca</td>
<td>male</td>
<td>2.28</td>
<td>112</td>
<td>no_jedi</td>
<td>wookiee</td>
<td>bowcaster</td>
</tr>
</tbody>
</table>

Let’s use these variables to make a scatterplot
How does it work?
1. Dataset

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
</tr>
</thead>
</table>

2. Which variables

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
</tr>
</thead>
</table>

3. Which Geometric objects

- points
- abcd
- text
- lines
- bars

4. Which Aesthetic attributes

- x = A
- y = C
- color = E
- size = default
- shape = default
Building a scatterplot

**Dataset:** starwars

**Variables:** height, weight, jedi

**Geoms:** points

**Aesthetic** (perceptive attributes):

- X-axis: height
- Y-axis: weight
- Color: jedi
Scatterplot with ggplot2

```r
ggplot(data = starwars) +
  geom_point(aes(x = height, y = weight, color = jedi))
```

**ggplot()** initializes a “ggplot” object
You specify the data set (data frame) with **data**

**geom_point()** indicates the type of geometric object
You use **aes()** to map aesthetic attributes to variables:
  X-position: height
  Y-position: weight
  Color: jedi
Automated things in ggplot2

- Axis labels
- Legends (positions, labels, symbols)
- Choice of colors for points
- Background color (i.e. gray)
- Grid lines (major and minor)
- Axis tick marks

You can always override the default settings (this is the tricky part in ggplot2)
### Mapping

#### data values

<table>
<thead>
<tr>
<th>height</th>
<th>weight</th>
<th>jedi</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.72</td>
<td>77</td>
<td>jedi</td>
</tr>
<tr>
<td>1.50</td>
<td>49</td>
<td>no_jedi</td>
</tr>
<tr>
<td>1.82</td>
<td>77</td>
<td>jedi</td>
</tr>
<tr>
<td>1.80</td>
<td>80</td>
<td>no_jedi</td>
</tr>
<tr>
<td>0.96</td>
<td>32</td>
<td>no_jedi</td>
</tr>
<tr>
<td>1.67</td>
<td>75</td>
<td>no_jedi</td>
</tr>
<tr>
<td>0.66</td>
<td>17</td>
<td>jedi</td>
</tr>
<tr>
<td>2.28</td>
<td>112</td>
<td>no_jedi</td>
</tr>
</tbody>
</table>

#### aesthetic attributes

<table>
<thead>
<tr>
<th>x</th>
<th>y</th>
<th>color</th>
</tr>
</thead>
<tbody>
<tr>
<td>x_1</td>
<td>y_1</td>
<td>#F8766D</td>
</tr>
<tr>
<td>x_2</td>
<td>y_2</td>
<td>#00BFC4</td>
</tr>
<tr>
<td>x_3</td>
<td>y_3</td>
<td>#F8766D</td>
</tr>
<tr>
<td>x_4</td>
<td>y_4</td>
<td>#00BFC4</td>
</tr>
<tr>
<td>x_5</td>
<td>y_5</td>
<td>#00BFC4</td>
</tr>
<tr>
<td>x_6</td>
<td>y_6</td>
<td>#00BFC4</td>
</tr>
<tr>
<td>x_7</td>
<td>y_7</td>
<td>#F8766D</td>
</tr>
<tr>
<td>x_8</td>
<td>y_8</td>
<td>#00BFC4</td>
</tr>
</tbody>
</table>

These values are meaningful to us, but not to the computer. They need to be converted from data units to physical units that the computer can display.
Main elements

A graphic is a mapping from data to aesthetic attributes (color, shape, size) of geometric objects (points, lines, bars, etc)

`ggplot(data, ...)`

`aes()`

`geom_objects()`
How does ggplot2 work?

Plots are created piece-by-piece

Plot components added with + operator

Aesthetic attributes mapped to data values

Computation of scales for aesthetic attributes
The data MUST BE in a data frame!
Always ask

What is the data set of interest?

What variables (columns) will be used to make the plot?

What graphic shapes (geoms) will be used to display the data?

What features of the shapes will be used to represent the data values?
Warning

ggplot2 comes with the function \texttt{qplot()} (i.e. quick plot)

Avoid using it!

As Karthik Ram says: “you’ll end up unlearning and relearning a good bit”