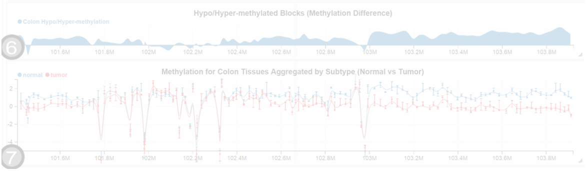


Introduction to Data Science: Basic Plotting

Héctor Corrada Bravo

University of Maryland, College Park, USA

2019-08-01

Data Visualization

We will spend a good amount of time in the course discussing data visualization.

It serves many important roles in data analysis.

We use it to gain understanding of dataset characteristics throughout analyses and it is a key element of communicating insights we have derived from data analyses with our target audience.

Grammar of Graphics (ggplot)

In this section, we will introduce basic functionality of the `ggplot` package (available in both R and python) to start our discussion of visualization throughout the course.

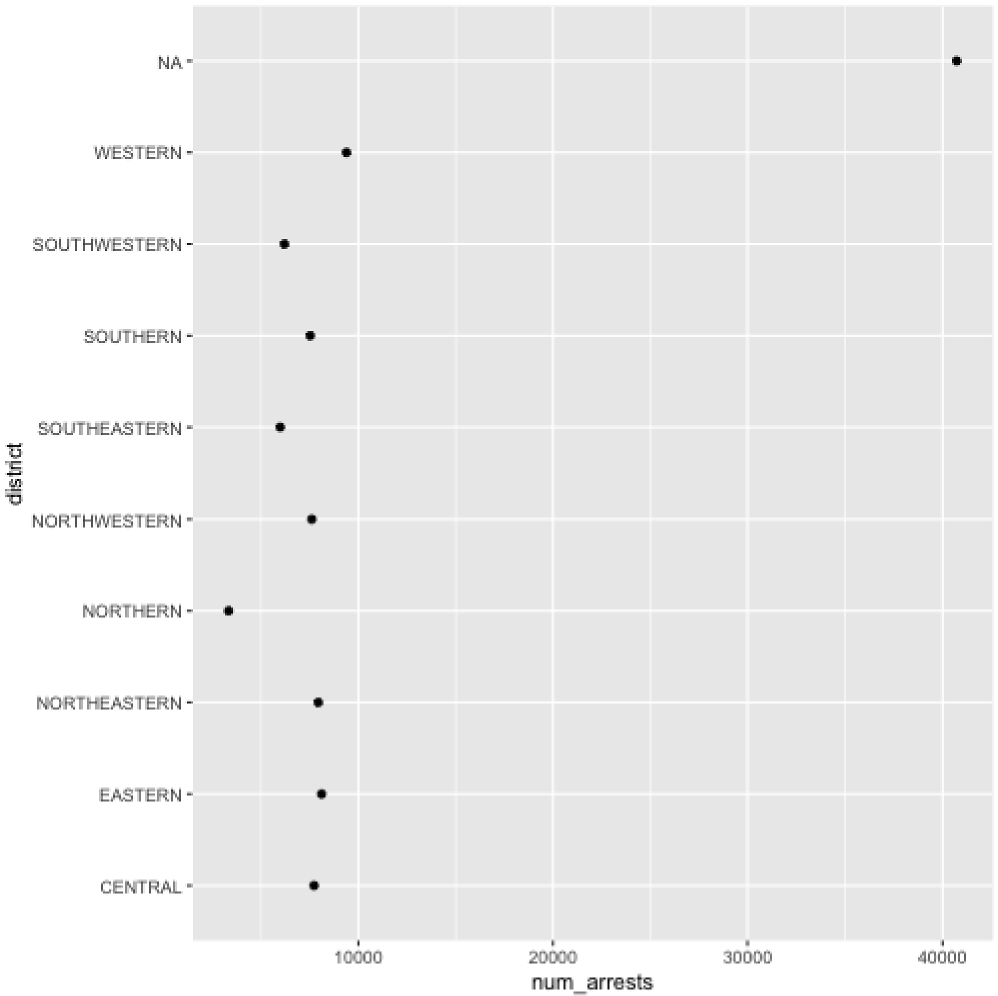
The `ggplot` package is designed around the Entity-Attribute data model.

Also, it can be included as part of data frame operation pipelines.

Grammar of Graphics (ggplot)

Let's create a *dot plot* of the number of arrests per district in our dataset:

```
arrest_tab %>%
  group_by(district) %>%
  summarize(num_arrests=n()) %>%
  ggplot(mapping=aes(y=district,
                     x=num_arrests)
    geom_point()
```



Grammar of Graphics (ggplot)

The `ggplot` design is very elegant, takes some thinking to get used to, but is extremely powerful.

The central premise is to characterize the building pieces behind `ggplot` plots as follows:

1. The **data** that goes into a plot, a data frame of entities and attributes
2. The **mapping** between data attributes and graphical (aesthetic) characteristics
3. The *geometric* representation of these graphical characteristics

Grammar of Graphics (ggplot)

So in our example we can fill in these three parts as follows:

1) **Data**: We pass a data frame to the `ggplot` function with the `%>%` operator at the end of the group_by-summarize pipeline.

2) **Mapping**: Here we map the `num_arrests` attribute to the `x` position in the plot and the `district` attribute to the `y` position in the plot. Every `ggplot` will contain one or more `aes` calls.

3) **Geometry**: Here we choose points as the *geometric* representations of our chosen graphical characteristics using the `geom_point` function.

Grammar of Graphics (ggplot)

In general, the `ggplot` call will have the following structure:

```
<data_frame> %>%

  ggplot(mapping=aes(<graphical_characteristic>=<attribute>)) +

    geom_<representation>()
```

Plot Construction Details

Mappings

Argument	Definition
`x`	position along x axis
`y`	position along y axis
`color`	color
`shape`	shape (applicable to e.g., points)
`size`	size
`label`	string used as label (applicable to text)

Plot Construction Details

Representations

Function	Representation
`geom_point`	points
`geom_bar`	rectangles
`geom_text`	strings
`geom_smooth`	smoothed line (advanced)
`geom_hex`	hexagonal binning

Plot Construction Details

We can include multiple geometric representations in a single plot, for example points and text, by adding (+) multiple `geom_<representation>` functions.

Also, we can include mappings inside a `geom_` call to map characteristics to attributes strictly for that specific representation.

For example `geom_point(mapping=aes(color=<attribute>))` maps color to some attribute only for the point representation specified by that call. Mappings given in the `ggplot` call apply to *all* representations added to the plot.

Frequently Used Plots

We will look comprehensively at data visualization in more detail later in the course, but for now will list a few common plots we use in data analysis and how they are created using `ggplot`.

Let's switch data frame to the `mpg` dataset for our examples:

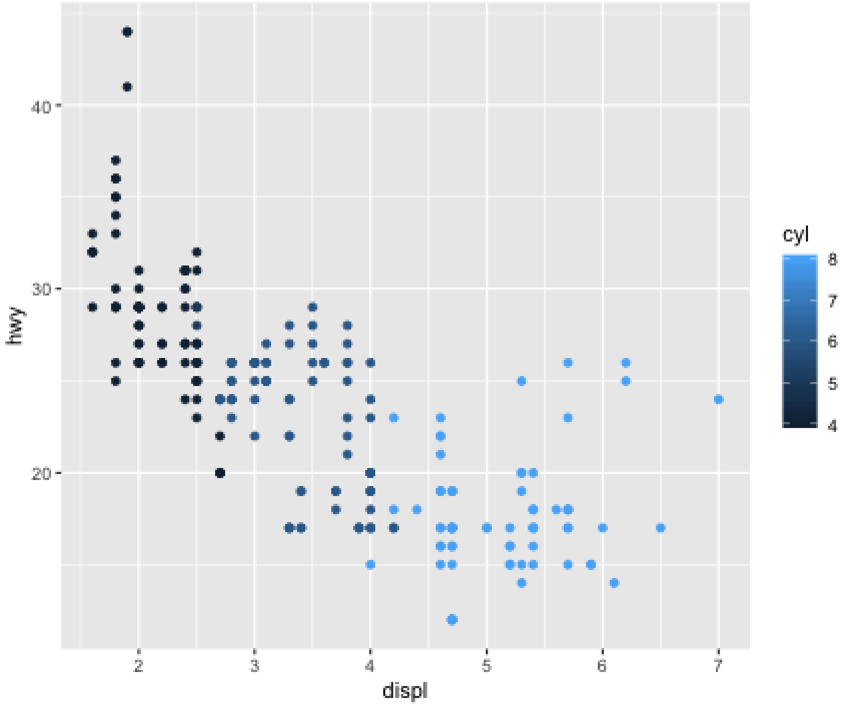
```
mpg
```

```
## # A tibble: 234 x 11
##    manufacturer model displ  year   cyl trans drv     cty   hwy fl    class
##    <chr>        <chr> <dbl> <int> <int> <chr> <chr> <int> <int> <chr> <chr>
##  1 audi         a4      1.8  1999     4 auto… f        18    29 p     comp…
```

Scatter plot

Used to visualize the relationship between two attributes.

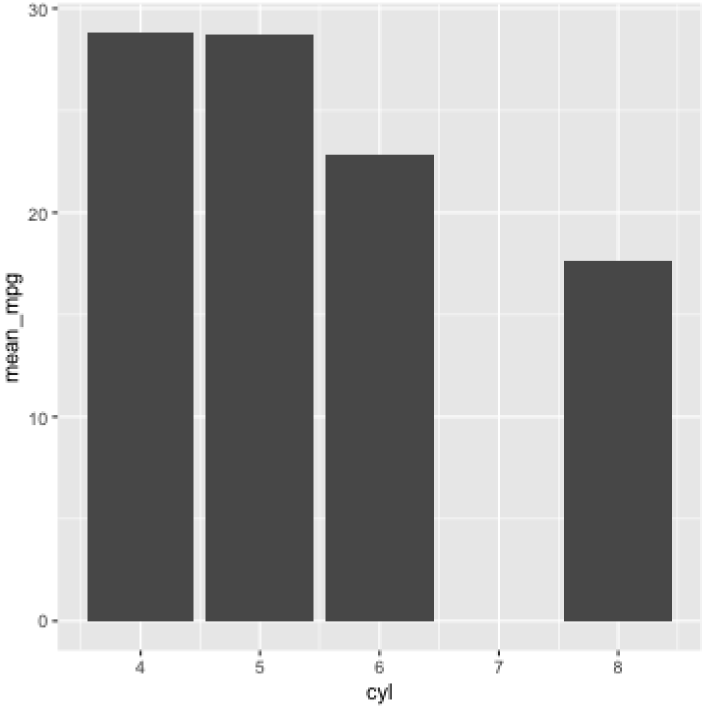
```
mpg %>%
  ggplot(mapping=aes(x=displ, y=hwy)) +
    geom_point(mapping=aes(color=cyl))
```



Bar graph

Used to visualize the relationship between a continuous variable to a categorical (or discrete) attribute

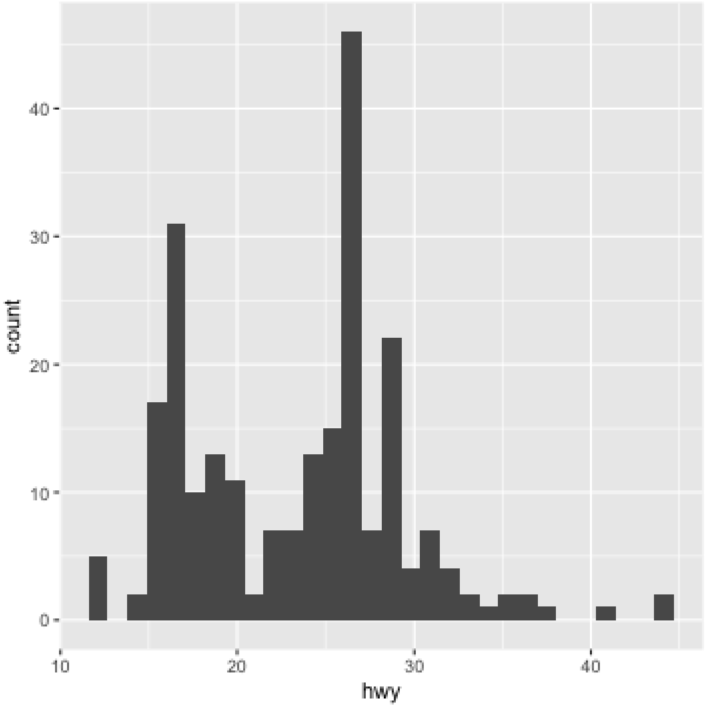
```
mpg %>%
  group_by(cyl) %>%
  summarize(mean_mpg=mean(hwy)) %>%
  ggplot(mapping=aes(x=cyl, y=mean_mpg)) +
    geom_bar(stat="identity")
```



Histogram

Used to visualize the distribution of the values of a numeric attribute

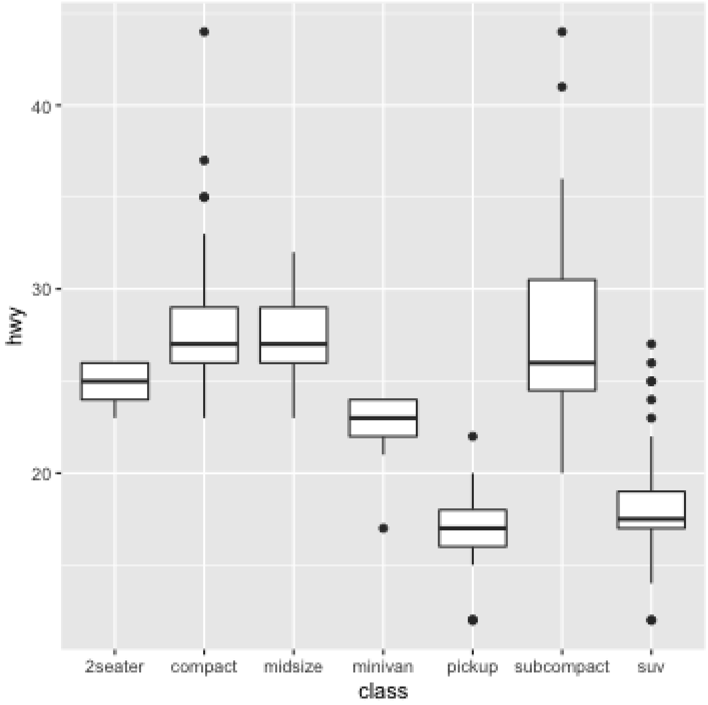
```
mpg %>%

  ggplot(mapping=aes(x=hwy)) +

    geom_histogram()
```



Boxplot

Used to visualize the distribution of a numeric attribute based on a categorical attribute

```
mpg %>%
  ggplot(mapping=aes(x=class, y=hwy)) +
    geom_boxplot()
```


hwy
2seater
compact
midsize
minivan
pickup
subcompact
suv
class