Data Visualization with ggplot2

Anthony Chau

UCI Center for Statistical Consulting

7/24/2021 (updated: 2021-07-28)

What is ggplot2?

- `ggplot2` is a R package for creating statistical and data graphics
- `ggplot2`'s approach to graphics is based on The Grammar of Graphics
- Mature package
- Powerful and extensible

Grammar of Graphics

- Big idea: a visualization is constructed from many independent components
- We put together different components to create our desired visualization
- Components of a plot:
  - Data
  - Aesthetic mappings
  - Geometric objects
  - Scales
  - Facet specification
  - Statistical Transformation
  - Coordinate System

US Midwest Demographics

- Let's use ggplot2 on a dataset containing demographics information for the US Midwest from the 2000 Census

```
midwest ← ggplot2::midwest
midwest[1, c(1,2,3,4,5)]
#> # A tibble: 1 x 5
#>     PID county state  area poptotal
#>   <int> <chr>  <chr> <dbl>    <int>
#> 1   561 ADAMS  IL    0.052    66090
dim(midwest)
#> [1] 437  28
colnames(midwest)
#>  [1] "PID"                  "county"               "state"
#>  [4] "area"                 "poptotal"             "popdensity"
#>  [7] "popwhite"             "popblack"             "popamerindian"
#> [10] "popasian"             "popother"             "percwhite"
#> [13] "percblack"            "percamerindan"        "percasian"
#> [16] "percother"            "popadults"            "perchsd"
#> [19] "percollege"           "percprof"             "poppovertyknown"
#> [22] "percpovertyknown"     "percbelowpoverty"     "percchildbelowpovert"
#> [25] "percadultpoverty"     "percelderlypoverty"   "inmetro"
#> [28] "category"
```

Scatterplots

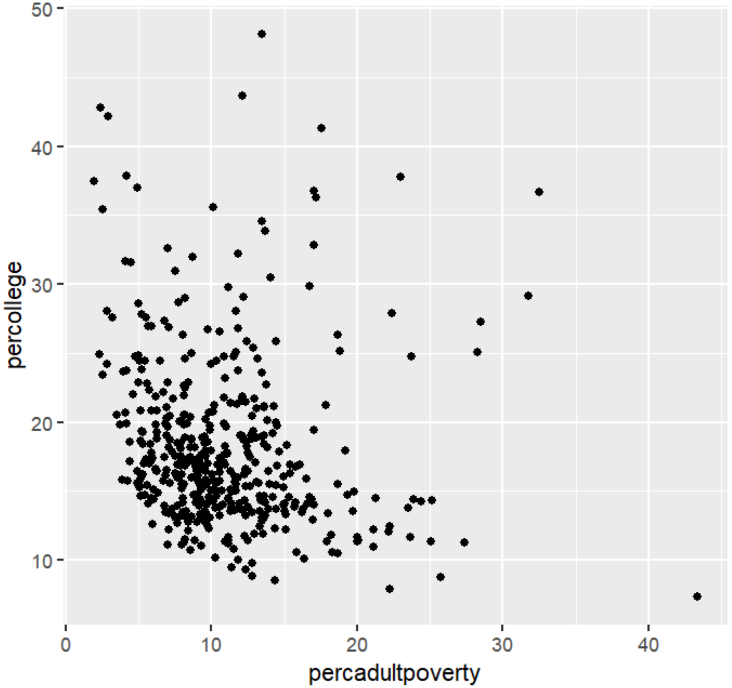
- A scatterplot visualizes the relationship between two quantitative variables
- The data points are commonly represented as points.

Scatterplot example

- Suppose we wanted to know the relationship between the percent of people below poverty line and the percent of people college educated
- We can use a scatterplot to visualize the relationship since both variables are quantitative

Scatterplot example

```
ggplot(midwest) +
  aes(x = percadultpoverty,
      y = percollege) +
  geom_point()
```



Recap

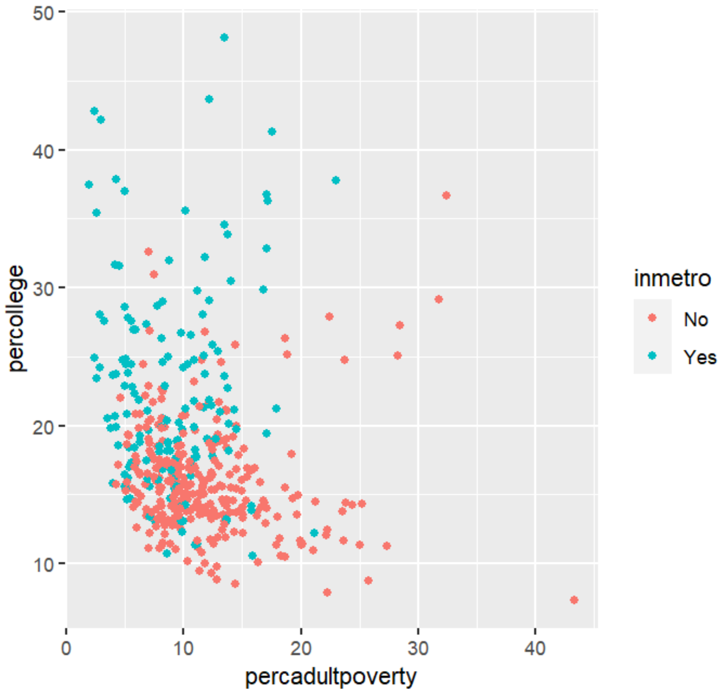
- Intialize a plot with the `ggplot` function
- Specify our data source
- *Aesthetic* properties: choose which variables to use for x and y position
- *Geometric* object (geom): Specify the type of plot
- We used the *point geom* `geom_point()` which produces a scatterplot

Extending the plot

- Suppose we wanted to see how the relationship between the percent of people below poverty line and the percent of people college educated depends on if someone lives in a metropolitian area
- We can display this visually by assigning another aesthetic element to your desired variable
- Let's use the color aesthetic and map it to the `inmetro` variable
- List of common aesthetics: color, shape, size, line type, line size, transparency

Scatterplot example

```
midwest$inmetro <-
  factor(midwest$inmetro,
         levels = c("0", "1"),
         labels = c("No", "Yes"))
ggplot(midwest) +
  aes(x = percadultpoverty,
      y = percollege,
      color = inmetro) +
  geom_point()
```

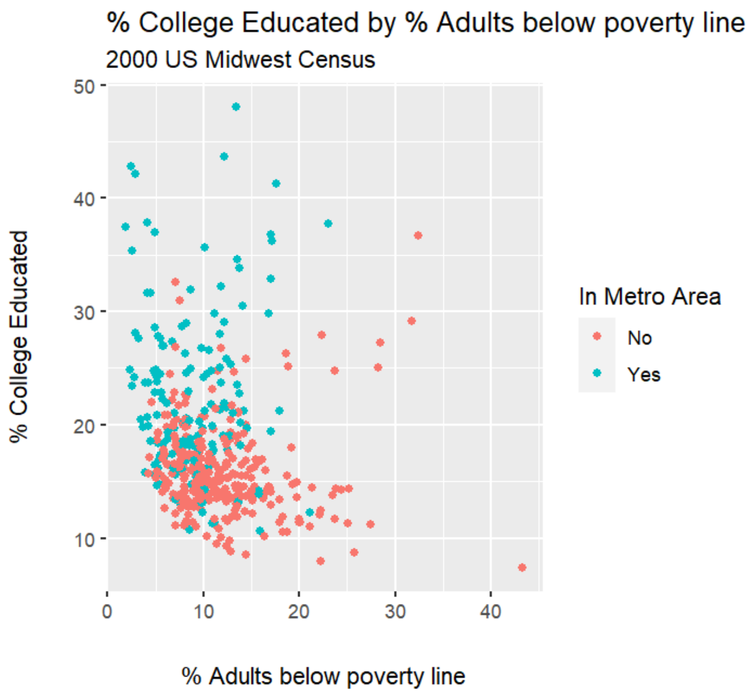


Improving the scatterplot

- Let's make some adjustment to the formatting of the plot
- List of changes:
    - Add descriptive plot title
    - Make axis labels more descriptives
    - Change legend title
    - Add spacing between axis and axis labels

Scatterplot example

```
ggplot(midwest) +
  aes(x = percadultpoverty,
      y = percollege,
      color = inmetro) +
  geom_point() +
  labs(title = "% College Educated by
       subtitle = "2000 US Midwest Ce
       color = "In Metro Area",
       x = "% Adults below poverty li
       y = "% College Educated") +
    theme(
      axis.title.x =
        element_text(
          margin = margin(t = 20, r =
                          b = 0, l =
      axis.title.y =
        element_text(
          margin = margin(t = 0, r = 2
                          b = 0, l =
    )
```



Histograms

- A histogram can be used to view the distribution for a single quantitative variable
- The histogram geom (`geom_histogram`) in `ggplot2` displays a histogram

Histograms

- A histogram divides your data into equal sized bins and draws rectangular bars to represent each bin.
- For example, suppose you want to visualize the percentage of high school graduates. The histogram can have ten bins with the width of each bin being 10%
- The height of a bar represents how many times a value in a bin occurs
- The binwidth in a histogram is important for interpretation

Histograms

- Binwidth = 10

Histograms

- Binwidth = 20

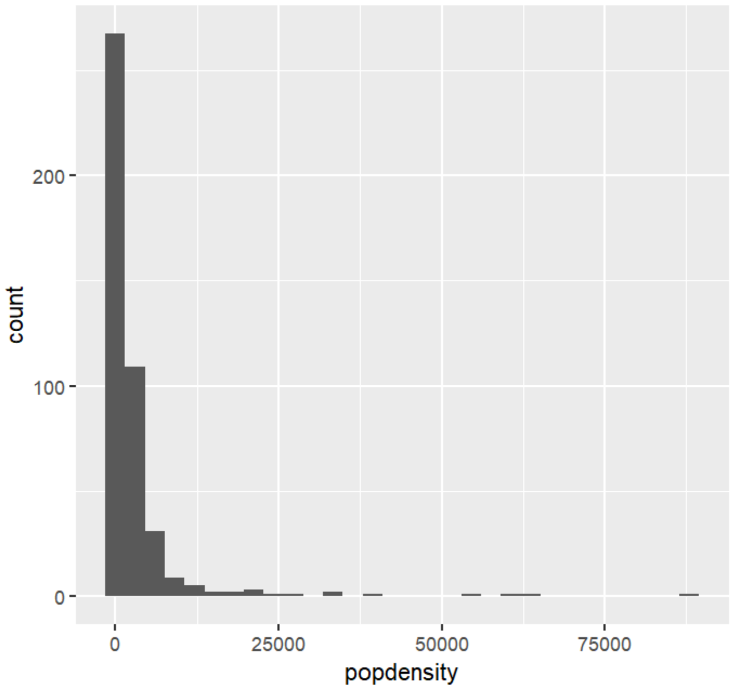
Histogram example

- Suppose we want to know the distribution of population density for all counties
- We can use a histogram to visualize this distribution
- `ggplot2` will choose a default binwidth for you. Usually, you should change the binwidth

Histogram example

```
ggplot(midwest) +
  aes(x = popdensity) +
  geom_histogram()
```

```
#> `stat_bin()` using `bins = 30`. Pick better v
```

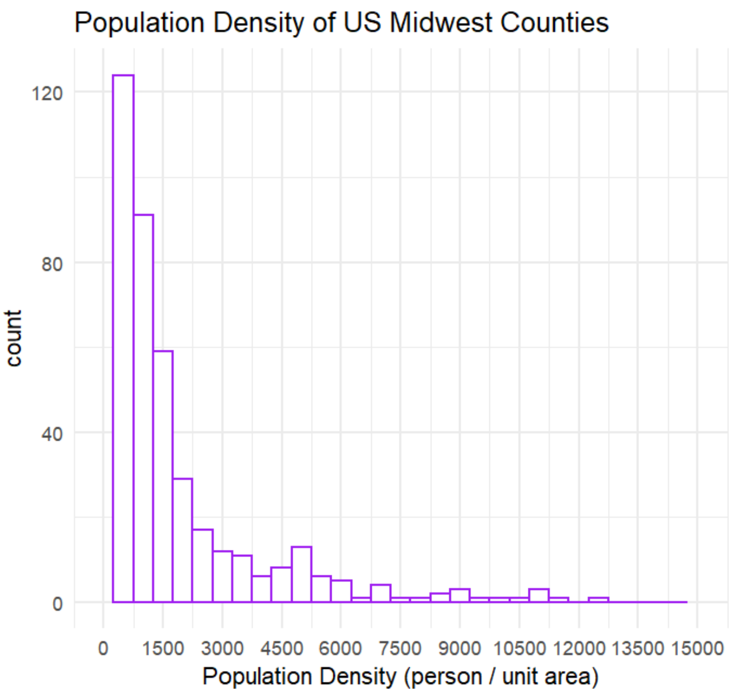


Improving the histogram

- The x-axis label can be changed to be more descriptive
- We can tinker with the binwidth (how wide each individual bar is) to display our desired plot
- Increase the number of x-axis *breaks* (positions on the axis that are marked)
- Adjust the x-axis *limits* to only include the majority of the data
- Make the bars hollow and have an outline color for the bars

Improving the histogram

```
ggplot(midwest) +
  aes(x = popdensity) +
  geom_histogram(binwidth = 500,
                 fill = "white",
                 color = "purple") +
  scale_x_continuous(
    breaks = seq(0, 15000, 1500),
    limits = c(0, 15000)) +
  labs(title = "Population Density of
       x = "Population Density (person
  theme_minimal()
```



Barplots

- A barplot can be used to view the distribution of a categorical variable
- A barplot draws rectangular bars where the height represents some numerical quantity.
- This "numerical quantity" can be how many times a category occurs or the value of another quantitative variable for each category
- Barplots = categorical variables
- Histograms = quantitative variables

Barplots

- Height of bar = Count of each unique category
  - Number of students enrolled at different universities
  - Categorical variable: "enrolled_university"
  - Count how many times "UCI", "UCLA", "USC" occur in the `enrolled_university` variable
- Height of bar = Value of another quantitative variable
  - Mean age of students at different universities
  - No counting - display the value of a quantitative variable (`mean_age`) for each category

Barplots

- The barplot displays how many people of each sex

Barplots

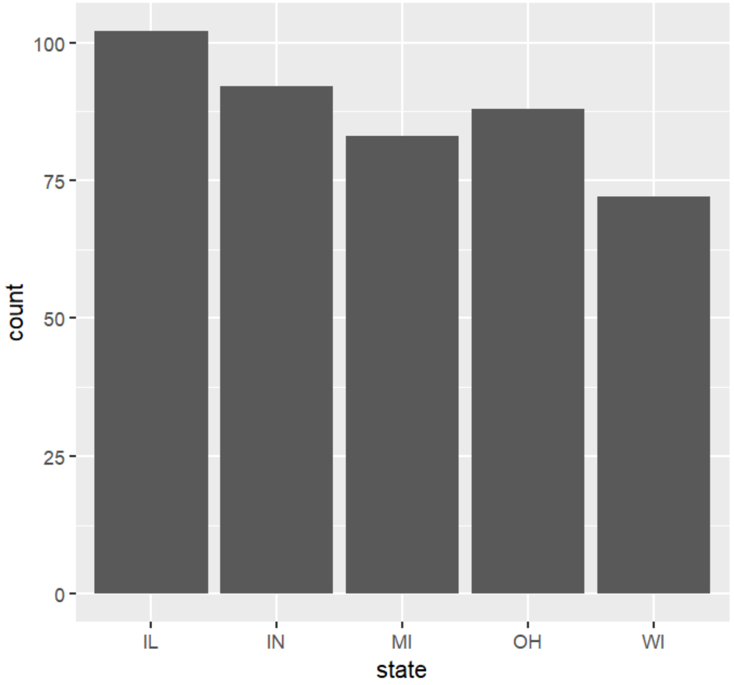
- The barplot displays the number of stores for each coffee chain

Barplot example

- Now, let's see how many counties are in each state
- Since state is categorical, we can use a barplot to display how many counties are in each state
- The bar geom (`geom_bar`) and the col geom (`geom_col`) in `ggplot2` displays a barplot
- The difference between `geom_bar` and `geom_col` is that `geom_bar` makes the height of the bar proportional to the number of each case in each group and `geom_col` makes the heights of the bars represent values in the data
- Knowing the above point, we need to use `geom_bar`

Barplot example

```
ggplot(midwest) +
  aes(x = state) +
  geom_bar()
```



Improving the barplot

- Order the barplot by count
- Flip coordinate system
- Make the bars hollow and have an outline color for the bars

Improving the barplot

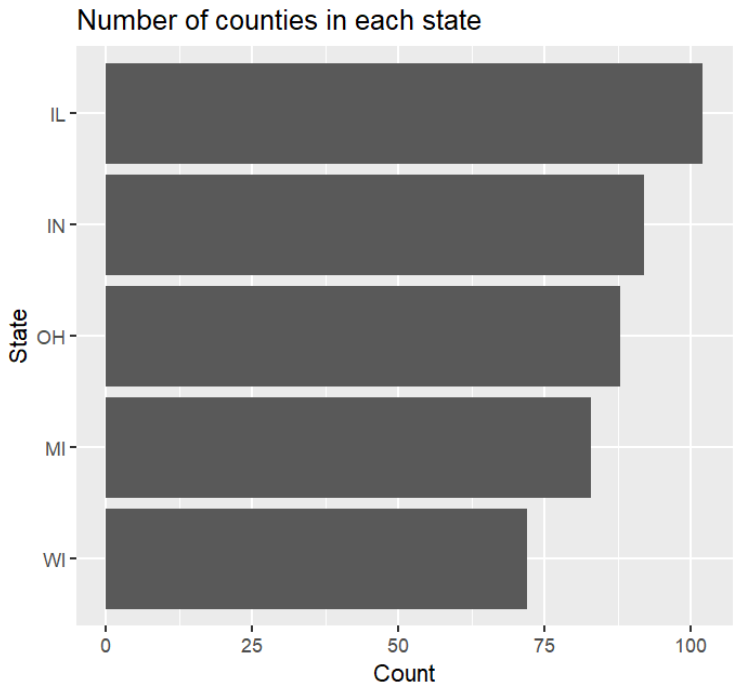
```
# count number of counties in each st
county_count <-
  as.data.frame(
    table(midwest$state)
    )

county_count

# rename column
colnames(county_count)[colnames(county

ggplot(county_count) +
  aes(x = reorder(state, Freq),
      y = Freq) +
  geom_col() +
  coord_flip() +
  labs(title = "Number of counties in
       x = "State",
       y = "Count")
```

```
#>   Var1 Freq
#> 1   IL  102
#> 2   IN   92
#> 3   MI   83
#> 4   OH   88
#> 5   WI   72
```



Boxplots

- A boxplot can be used to view the distribution of a quantitative variable by a categorical variable
- A box with "whiskers" are drawn to show the distribution for each category
- Often, we can read off the first quartile (Q1), median, and third quartile (Q3) from the boxplot.
- The **Interquartile range** (IQR) is the third quartile minus the first quartile: Q3-Q1
- The IQR is a measure of dispersion. A large IQR indicates that is greater spread

Boxplots

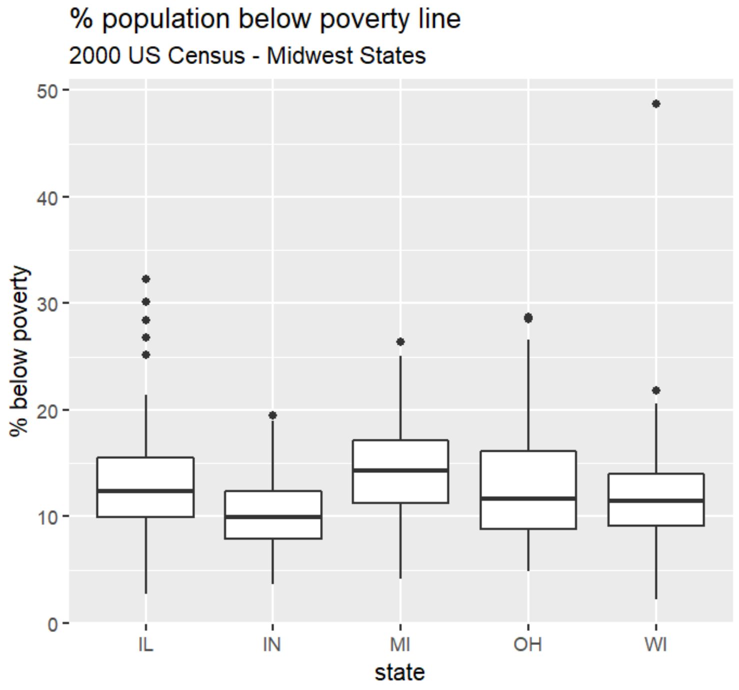
- Top "Whisker": Q3 + 1.5 * IQR
- Top line of box: First Quartile
- Bold line: Median
- Bottom line of box: Third Quartile
- Bottom "Whisker": Q1 - 1.5 * IQR
- Additional points: outliers

Boxplot example

- Now, let's see the distribution of the percent of people below the poverty line for each state
- A boxplot can be used to see how the percentage of people below the poverty line varies for each state
- The boxplot geom (`geom_boxplot`) in `ggplot2` displays a boxplot

Boxplot example

```
ggplot(midwest) +
  aes(x = state,
      y = percbelowpoverty) +
  geom_boxplot() +
  labs(title = "% population below pov
       subtitle = "2000 US Census - Mi
       y = "% below poverty")
```



Summary

- `ggplot2` builds a plot by combining multiple components
- Plots need a **data** source as well as variables to map to the x and y positions
- **Aesthetics** are visual properties of the plot - what you can see on the plot. Common aesthetics are the x position, y position, color, shape, and size.
- The selection of a **geom** determines the type of plot
- The **scale** can be modified to change how the data maps over to aesthetic properties. There is one scale for each aesthetic property.
- The **coord** can be modified to change the position of objects relative to the plane of the plot

Summary

- The **theme** controls the non-data components of the plot. Some examples: titles, labels, fonts, background, gridlines. See `?theme`.
- **Facets** specify how to display subsets of your data
- **Statistics** are transformations of your data that can be drawn on top of the data