R Graphics

Shirley Liao

6/15/2017

Starting At The End

My goal: by the end of the workshop you will be able to reproduce this graphic from the Economist:

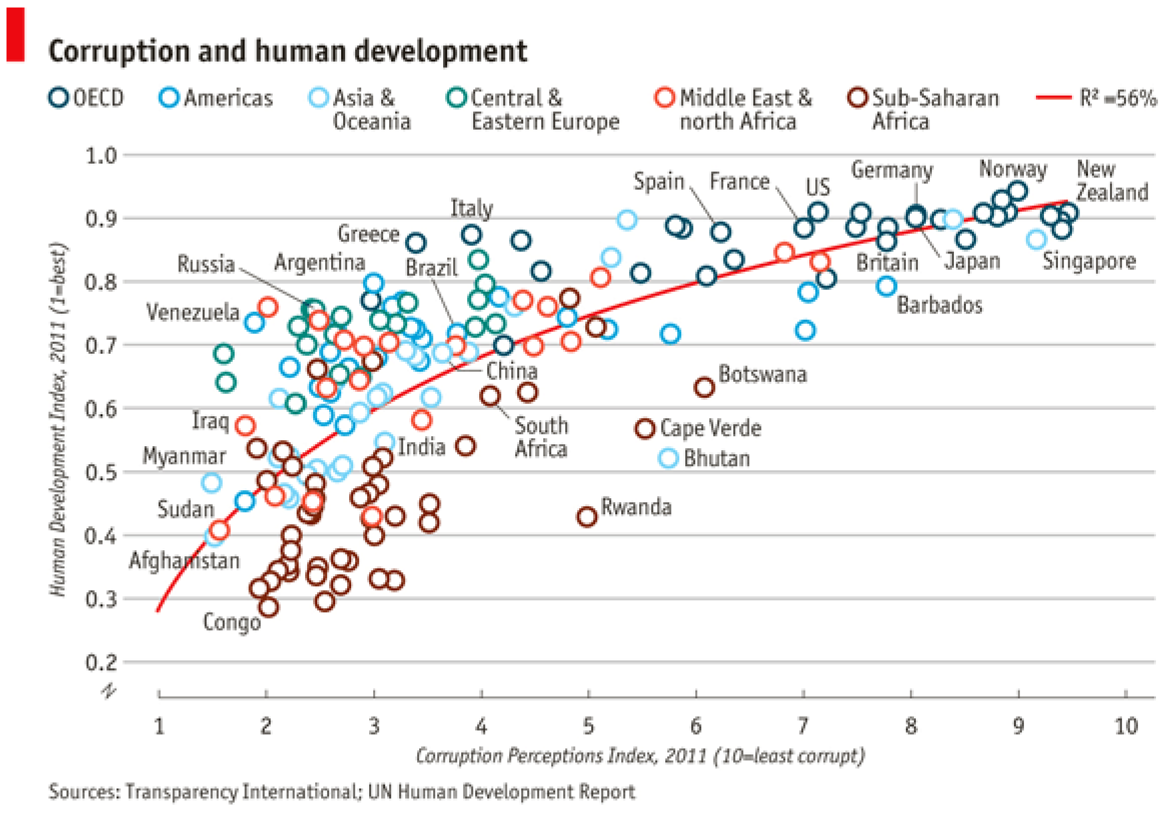


Figure 1:

Why ggplot2?

Advantages of ggplot2:

- consistent underlying grammar of graphics (Wilkinson, 2005)
- plot specification at a high level of abstraction
- very flexible
- theme system for polishing plot appearance
- mature and complete graphics system
- many users, active mailing list

That said, there are some things you cannot (or should not) do with ggplot2:

- 3-dimensional graphics (see the rgl package)
- Graph-theory type graphs (nodes/edges layout; see the igraph package)

- Interactive graphics (see the ggvis package)

What Is The Grammar Of Graphics?

The basic idea: independently specify plot building blocks and combine them to create just about any kind of graphical display you want. Building blocks of a graph include:

- data
- aesthetic mapping
- geometric object
- statistical transformations
- scales
- coordinate system
- position adjustments
- faceting

We will use the babyNames.csv dataset we imported yesterday:

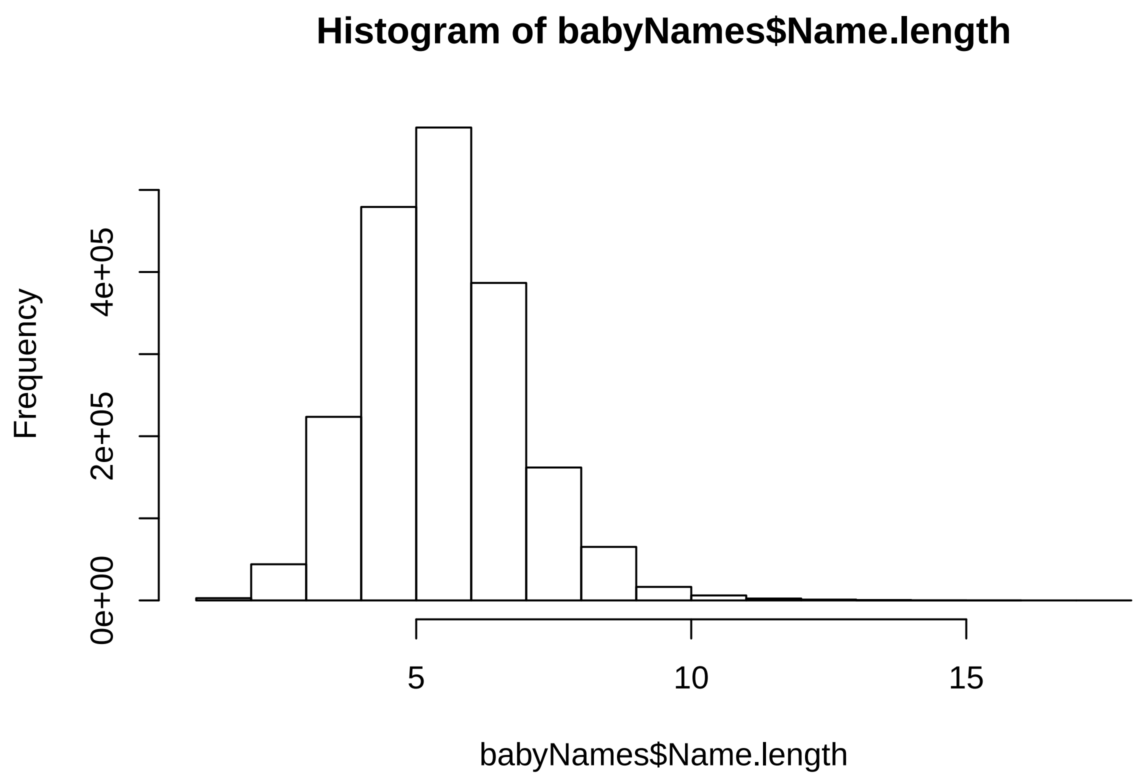
ggplot2 versus basic graphics

Compared to base graphics, ggplot2:

- is more verbose for simple / canned graphics
- is less verbose for complex / custom graphics
- does not have methods (data should always be in a data.frame)
- uses a different system for adding plot elements

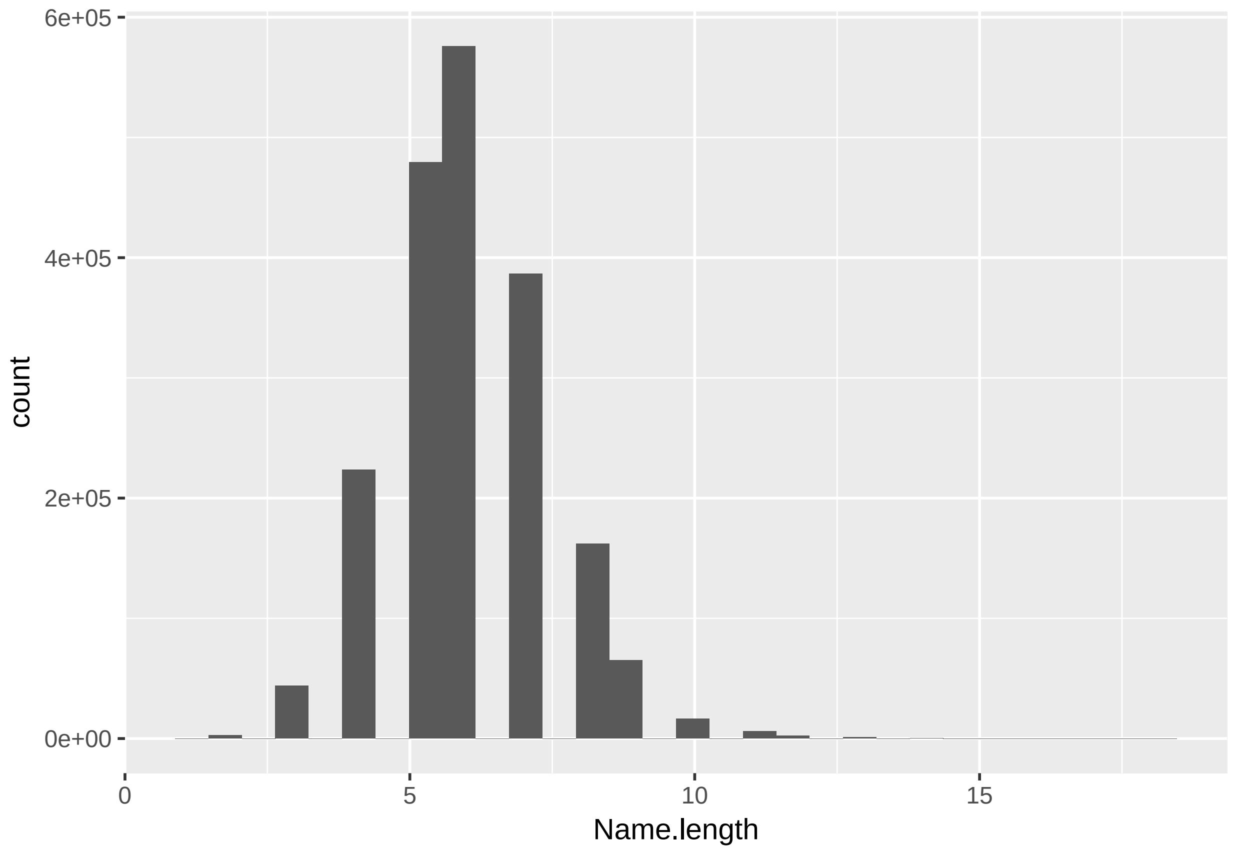
Simple graphs

```r
hist(babyNames$Name.length)
```



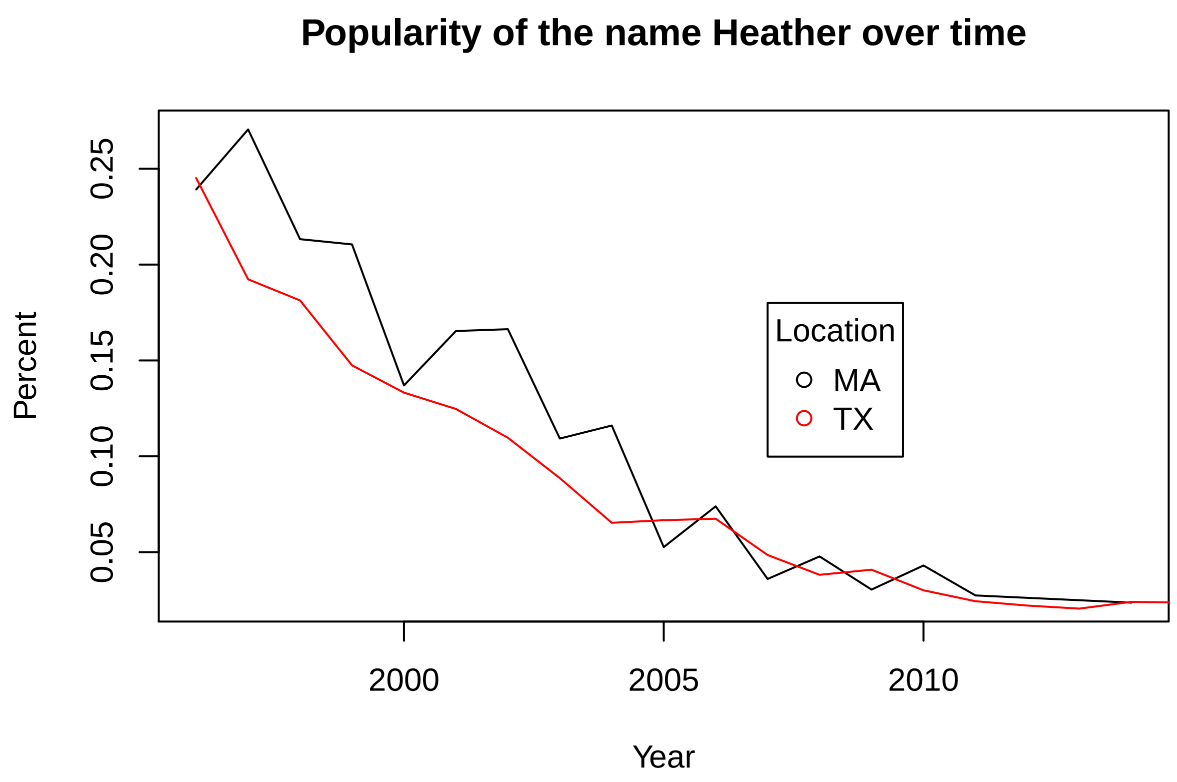
```r
#install.packages("ggplot2")
library(ggplot2)
ggplot(babyNames, aes(x = Name.length)) +
  geom_histogram()
```

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

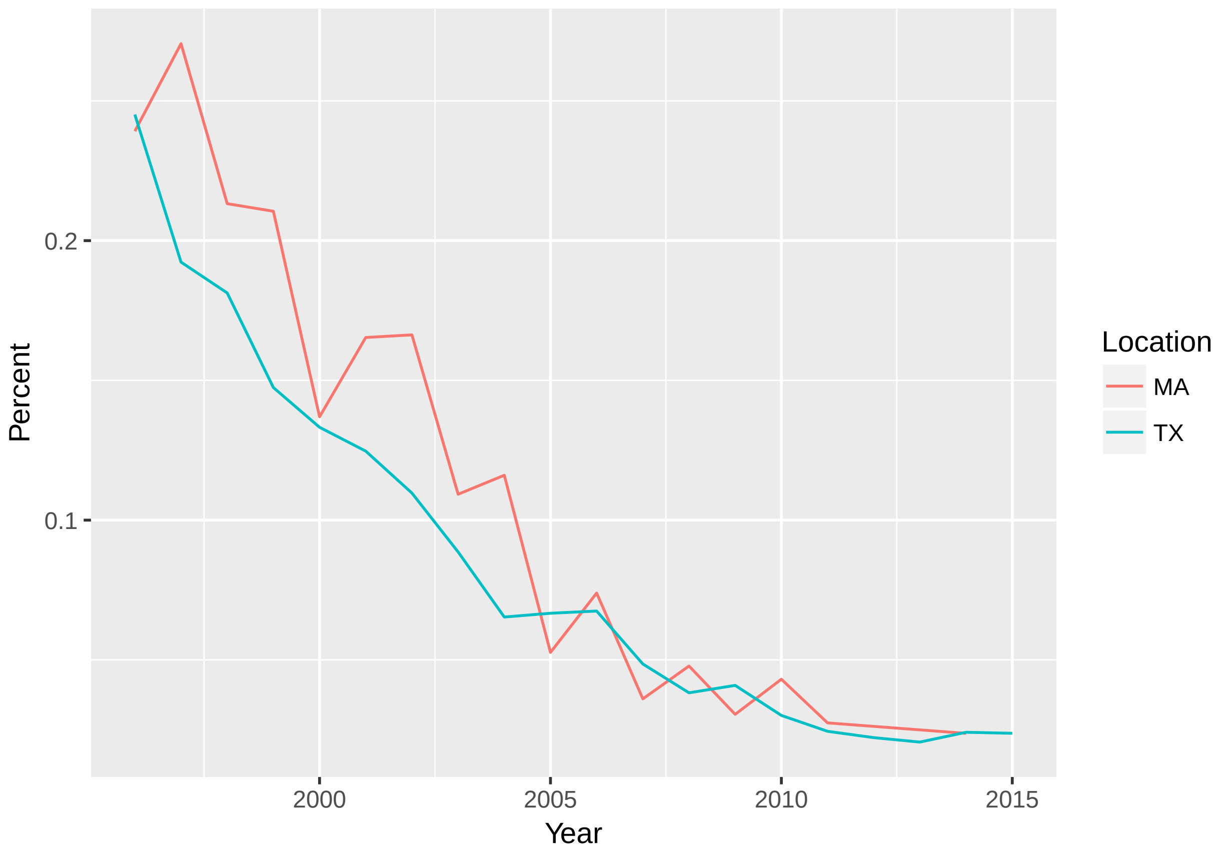


Complicated graphs

```
babyNames = babyNames[order(babyNames$Year),]
plot(Percent ~ Year, type="l",
     data=subset(babyNames, Location == "MA" & Name=="heather"))
points(Percent ~ Year, type="l",col="red",
       data=subset(babyNames, Location == "TX" & Name =="heather"))
title("Popularity of the name Heather over time")
legend(2007, 0.18,
       c("MA", "TX"), title="Location",
       col=c("black", "red"),
       pch=c(1, 1))
```



```
ggplot(subset(babyNames, Location %in% c("MA", "TX") & Name == "heather"),
       aes(x=Year,
           y=Percent,
           color=Location))+
  geom_line()
```



Geometric Objects And Aesthetics

Aesthetic Mapping

In ggplot land aesthetic means "something you can see". Examples include:

- position (i.e., on the x and y axes)
- color ("outside" color)
- fill ("inside" color)
- shape (of points)
- linetype
- size

Each type of geom accepts only a subset of all aesthetics–refer to the geom help pages to see what mappings each geom accepts. Aesthetic mappings are set with the aes() function.

Geometic Objects (geom)

Geometric objects are the actual marks we put on a plot. Examples include:

- points (geom_point, for scatter plots, dot plots, etc)
- lines (geom_line, for time series, trend lines, etc)
- boxplot (geom_boxplot, for, well, boxplots!)

A plot must have at least one geom; there is no upper limit. You can add a geom to a plot using the + operator

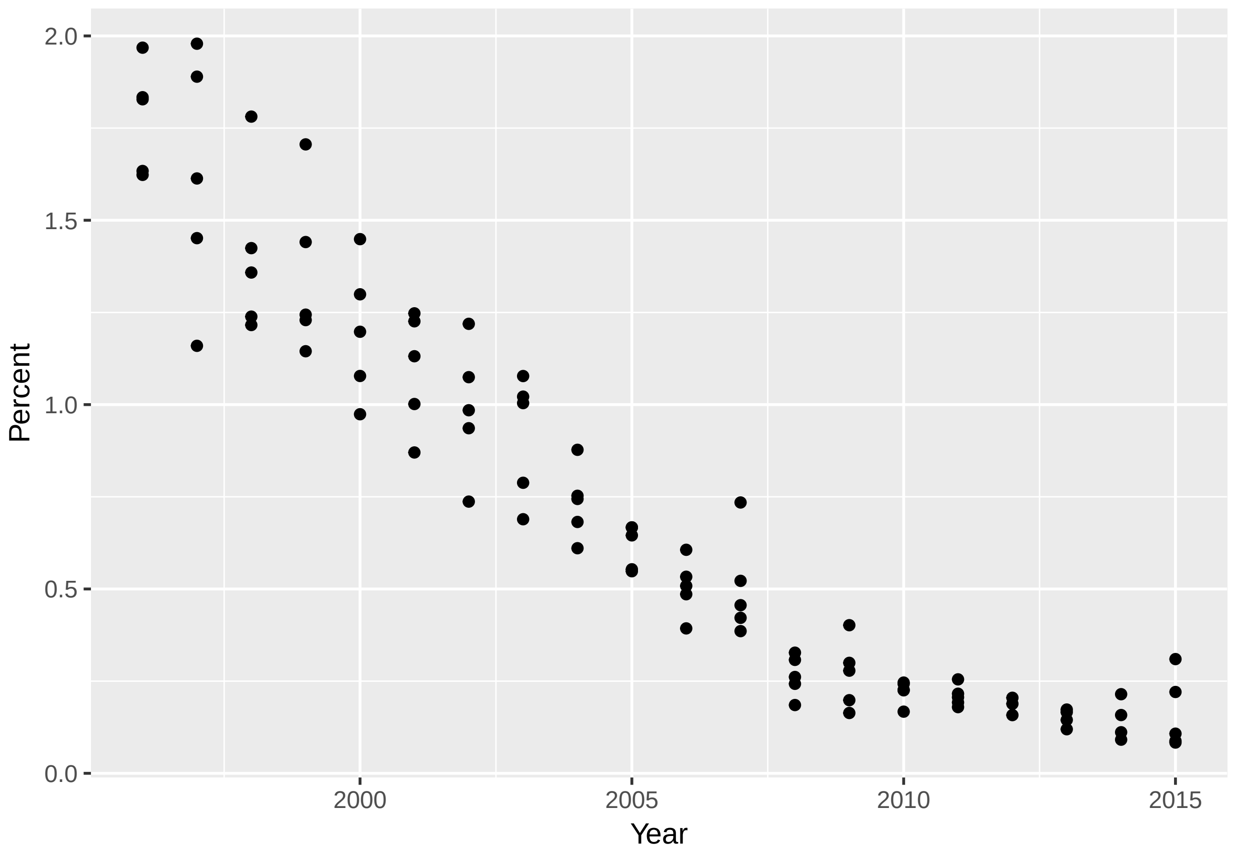
You can get a list of available geometric objects using the code below:

```
help.search("geom_", package = "ggplot2")
```

Example 1: Making a scatterplot

Now that we know about geometric objects and aesthetic mapping, we can make a ggplot. geom_point requires mappings for x and y, all others are optional.

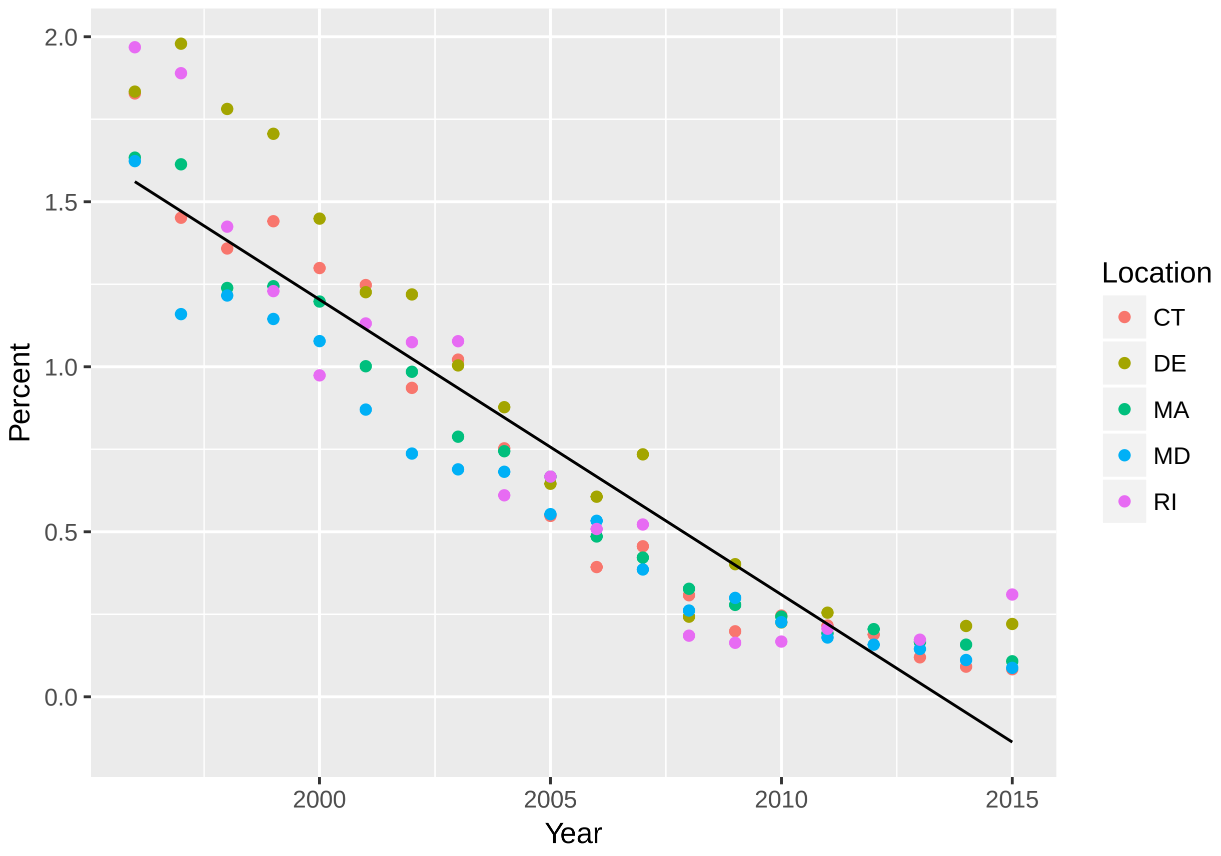
```
hp2001Q1 <- subset(babyNames, Name == "jessica" & Location %in% c("MA","DE","CT","RI","MD"))
ggplot(hp2001Q1,
       aes(y = Percent, x = Year)) +
  geom_point()
```



Example 2: Scatterplot with regression line

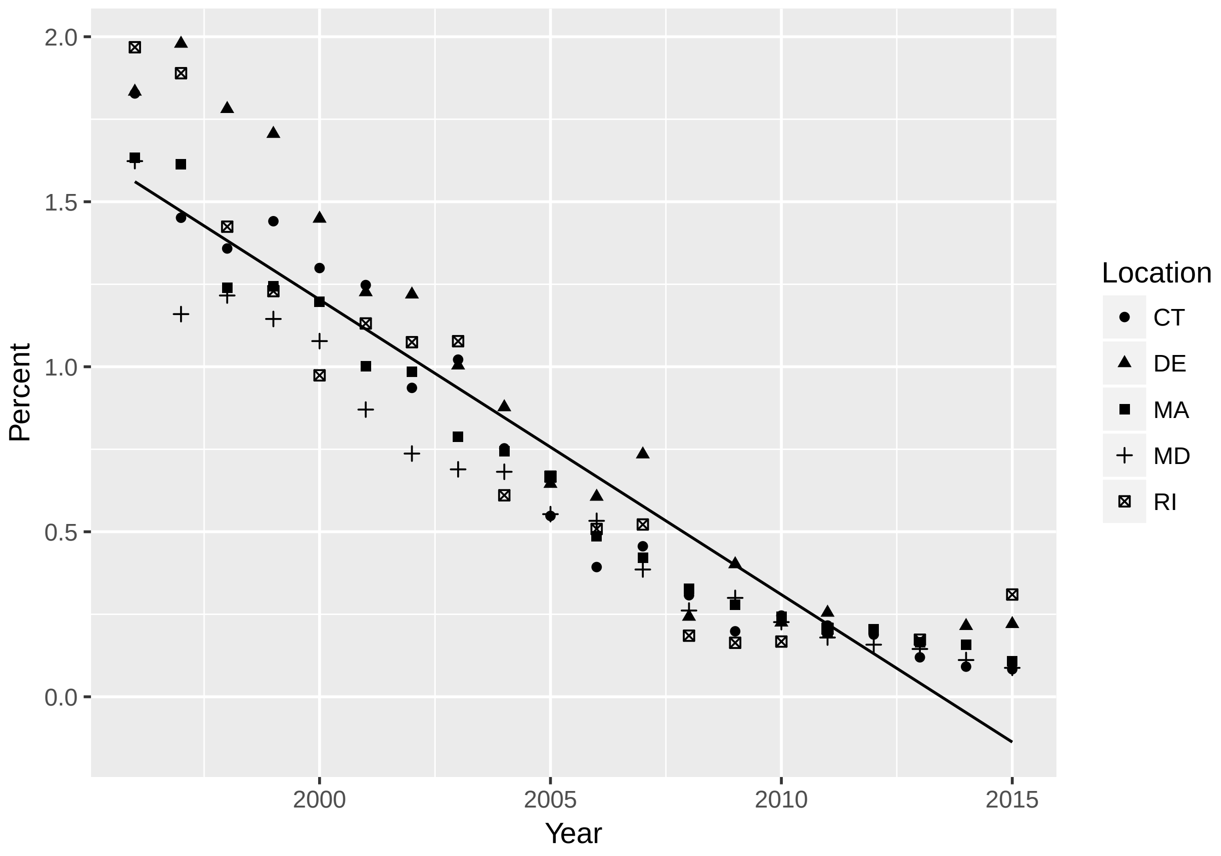
A plot constructed with ggplot can have more than one geom. In that case the mappings established in the ggplot() call are plot defaults that can be added to or overridden. Our plot could use a regression line:

```
hp2001Q1$pred.SC <- predict(lm(Percent ~ Year, data = hp2001Q1))

ggplot(hp2001Q1, aes(x = Year, y = Percent)) + geom_point(aes(color = Location)) +
  geom_line(aes(y = pred.SC))
```



Instead of using color change to symbolize different states we could use shapes instead:

```
hp2001Q1$pred.SC <- predict(lm(Percent ~ Year, data = hp2001Q1))

ggplot(hp2001Q1, aes(x = Year, y = Percent)) + geom_point(aes(shape = Location)) +
  geom_line(aes(y = pred.SC))
```



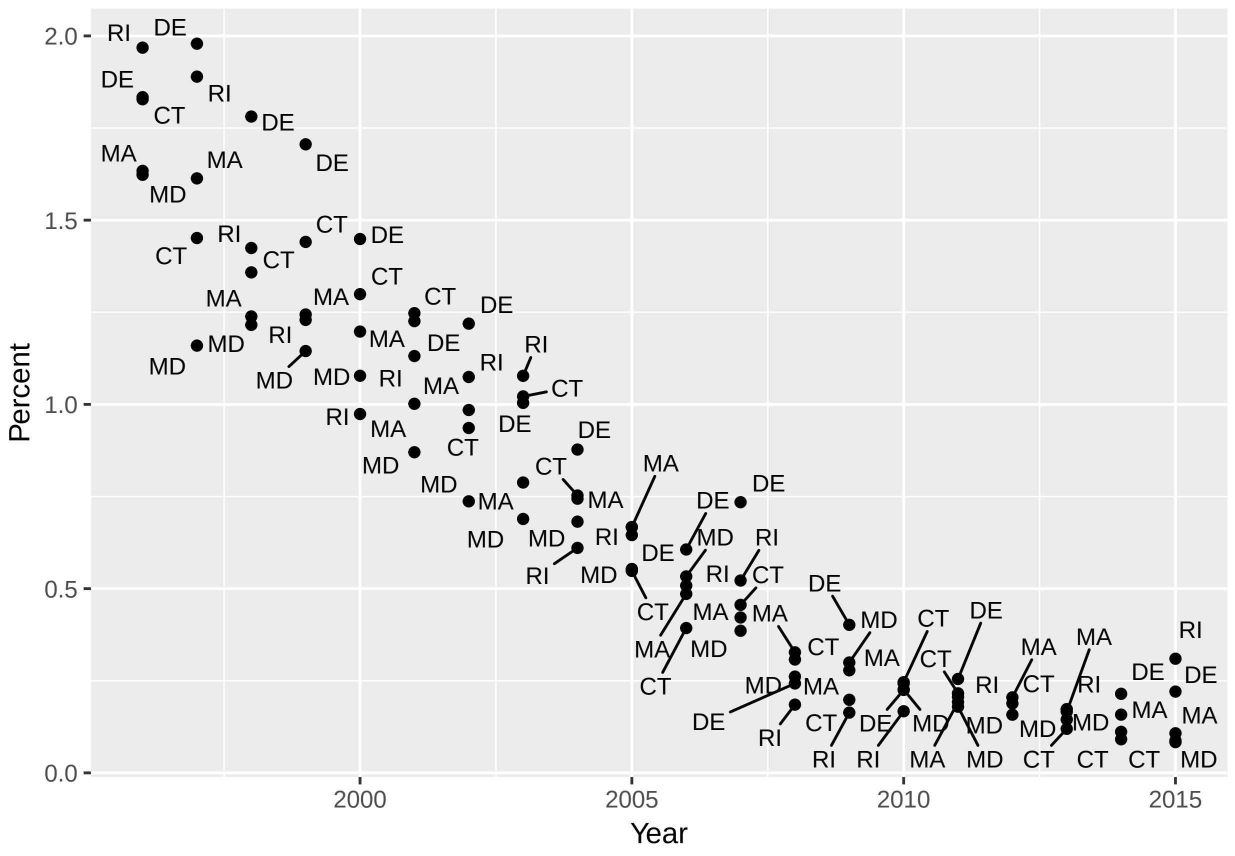
Or we could label the points:

```
install.packages("ggrepel",repos="https://cloud.r-project.org")
```

```
## Installing package into '/Users/shirleyliao/Library/R/3.3/library'
## (as 'lib' is unspecified)
```

```
##
## The downloaded binary packages are in
##  /var/folders/96/7xw011zd5vs0f_80xf76pwrr0000gn/T//RtmpedkxbO/downloaded_packages
```

```
library("ggrepel")
ggplot(hp2001Q1, aes(x = Year, y = Percent)) +
  geom_point() +
  geom_text_repel(aes(label=Location), size = 3)
```

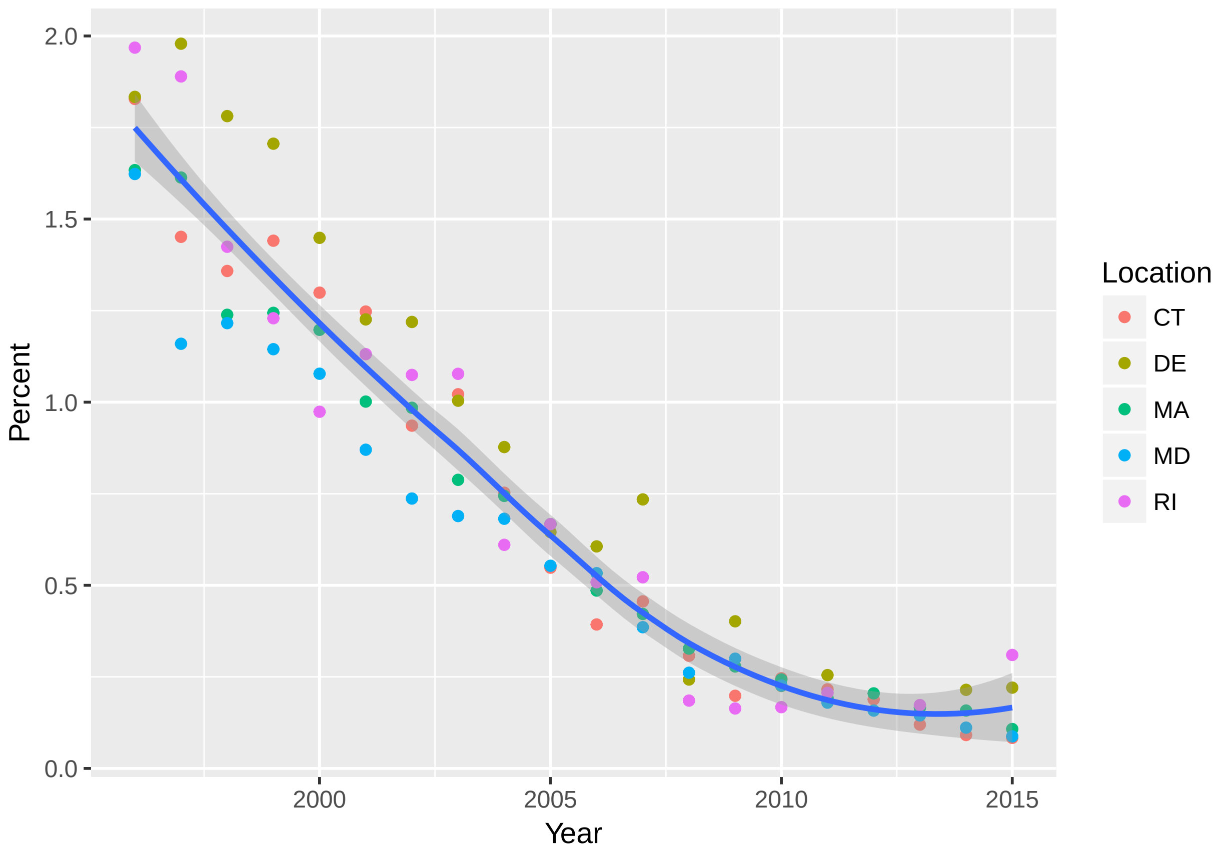


Example 3: Smoothers

Not all geometric objects are simple shapes–the smooth geom includes a line and a ribbon.

```
ggplot(hp2001Q1, aes(x = Year, y = Percent)) +
  geom_point(aes(color = Location)) +
  geom_smooth()
```

```
## `geom_smooth()` using method = 'loess'
```



NOTE! aes() should only hold aesthetic specifications which change along values of a covariate. "Constant" specifications must be written outside the aes() function:

```
ggplot(hp2001Q1, aes(x = Year, y = Percent)) +
  geom_point(aes(size = 2),# incorrect! 2 is not a variable
             color="red") # this is fine -- all points red

#should be:
ggplot(hp2001Q1, aes(x = Year, y = Percent)) +
  geom_point(size = 2, color="red")
```

Exercise 1

Read in the EconomistData.csv.

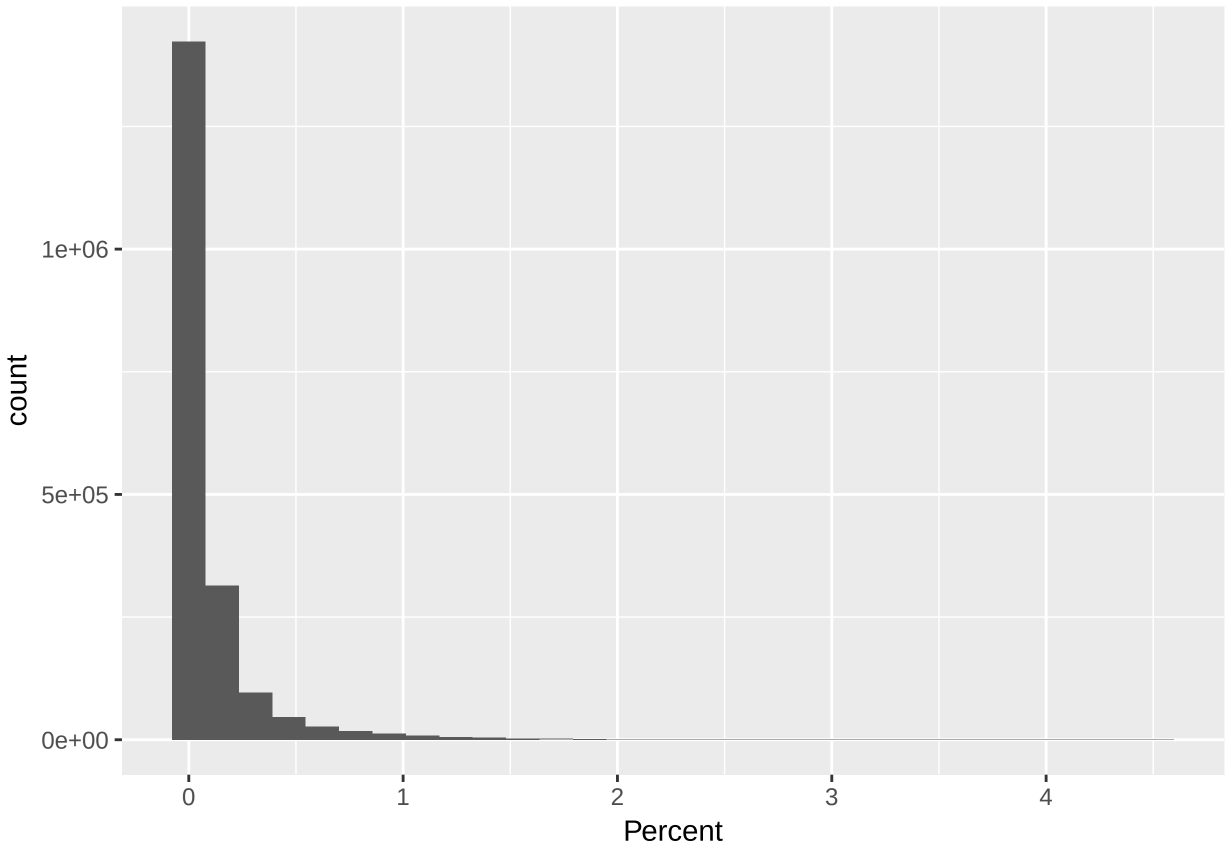
1. Create a scatter plot with CPI on the x axis and HDI on the y axis.
2. Color the points blue.
3. Map the color of the the points to Region.
4. Make the points bigger by setting size to 2
5. Map the size of the points to HDI.Rank
6. Overlay a smoothing line on top of the scatter plot using geom_smooth.
7. Overlay a smoothing line on top of the scatter plot using geom_smooth, but use a linear model for the predictions. Hint: see ?stat_smooth.
8. Overlay a smoothing line on top of the scatter plot using geom_line. Hint: change the statistical transformation.

Histograms

Let's create a histogram of name popularity (counted as percent of all names) in the year 2000:

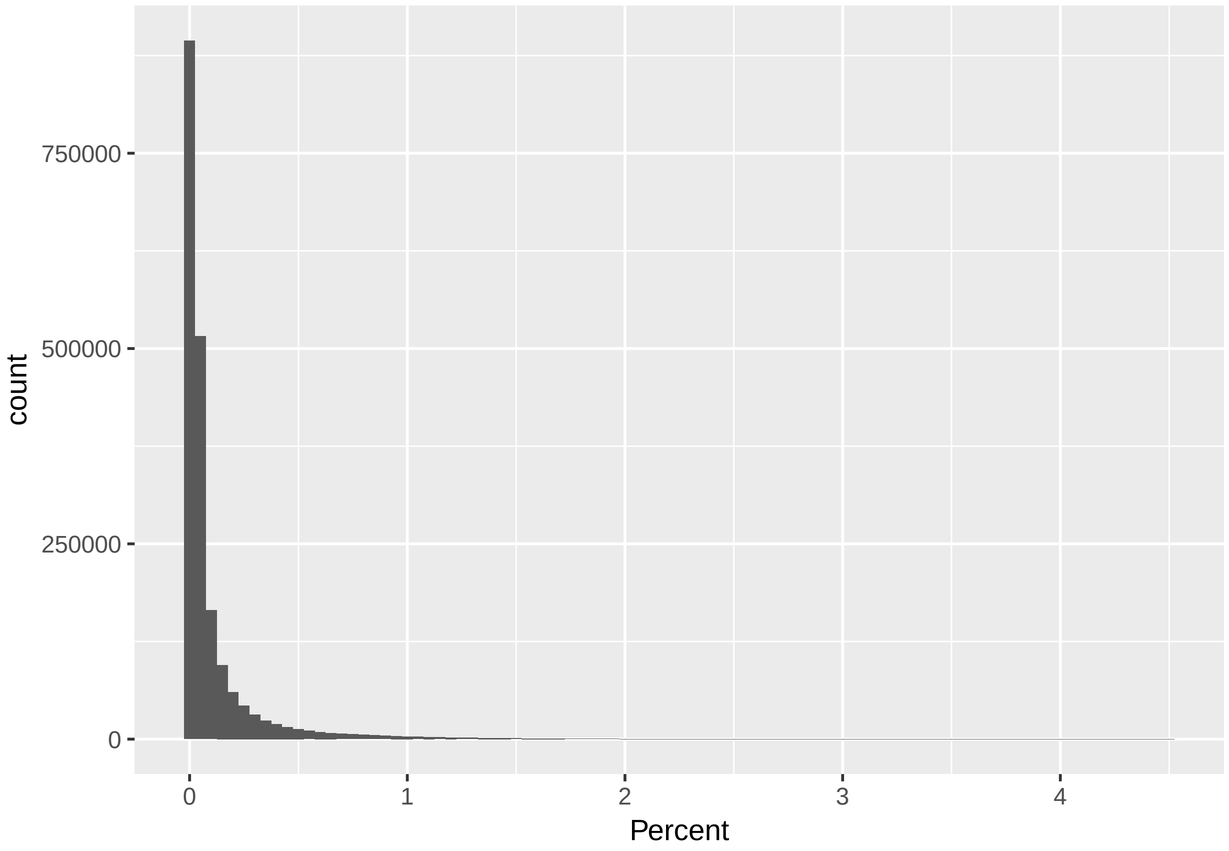
```r
ggplot(subset(babyNames,Year=2000), aes(x = Percent))  + geom_histogram()
```

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



Binwidth is specified by passing a "stat" argument:

```r
ggplot(subset(babyNames,Year=2000), aes(x = Percent))  + geom_histogram(stat = "bin", binwidth=0.05)
```



Scales

Scales: Controlling Aesthetic Mapping

Aesthetic mapping (i.e., with aes()) only says that a variable should be mapped to an aesthetic. It doesn't say how that should happen. For example, when mapping a variable to shape with aes(shape = x) you don't say what shapes should be used. Similarly, aes(color = z) doesn't say what colors should be used. Describing what colors/shapes/sizes etc. to use is done by modifying the corresponding scale. In ggplot2 scales include

- position
- color and fill
- size
- shape
- line type

Scales are modified with a series of functions using a scale_ *naming scheme. Try typing scale* to see a list of scale modification functions.

Common Scale Arguments

The following arguments are common to most scales in ggplot2:

- name
- the first argument gives the axis or legend title
- limits
- the minimum and maximum of the scale

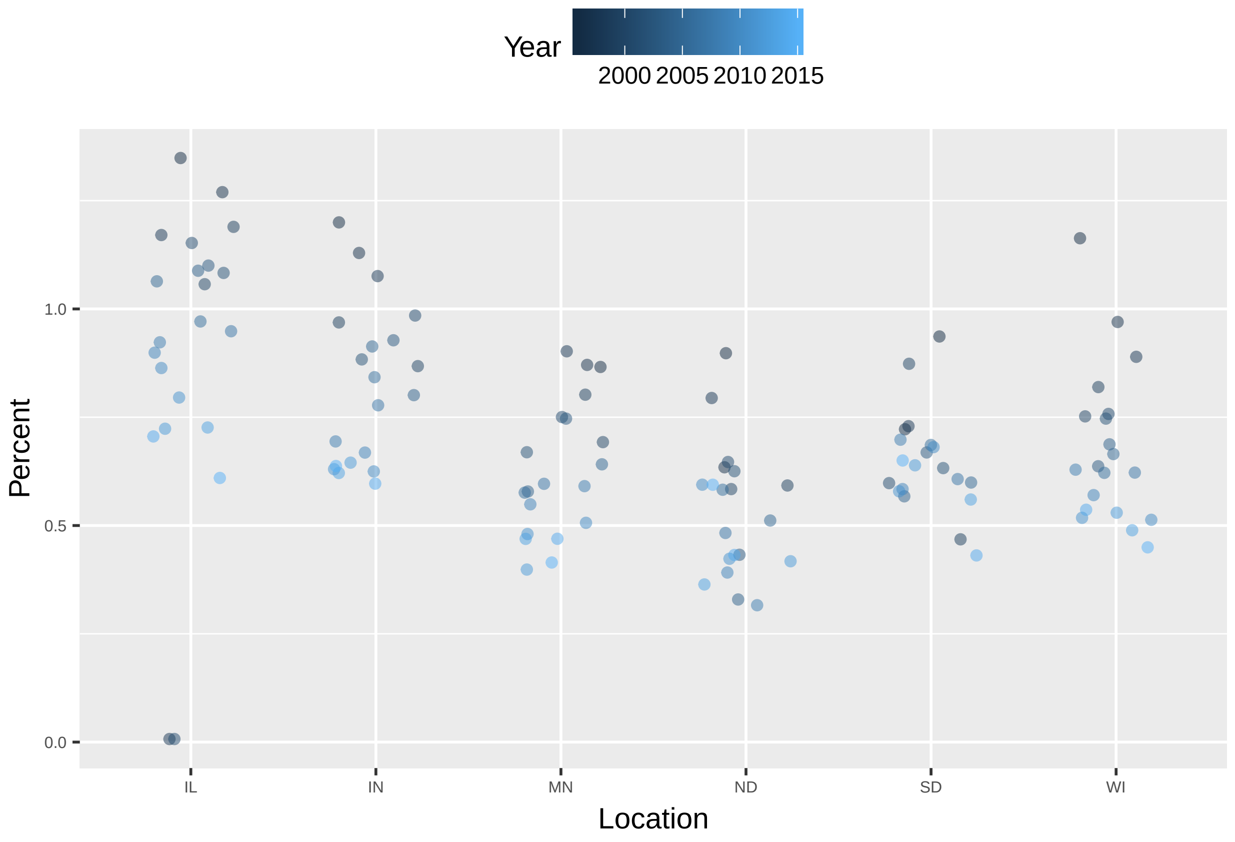
- breaks
- the points along the scale where labels should appear
- labels
- the labels that appear at each break

Specific scale functions may have additional arguments; for example, the scale_color_continuous function has arguments low and high for setting the colors at the low and high end of the scale.

Scale Modification Examples

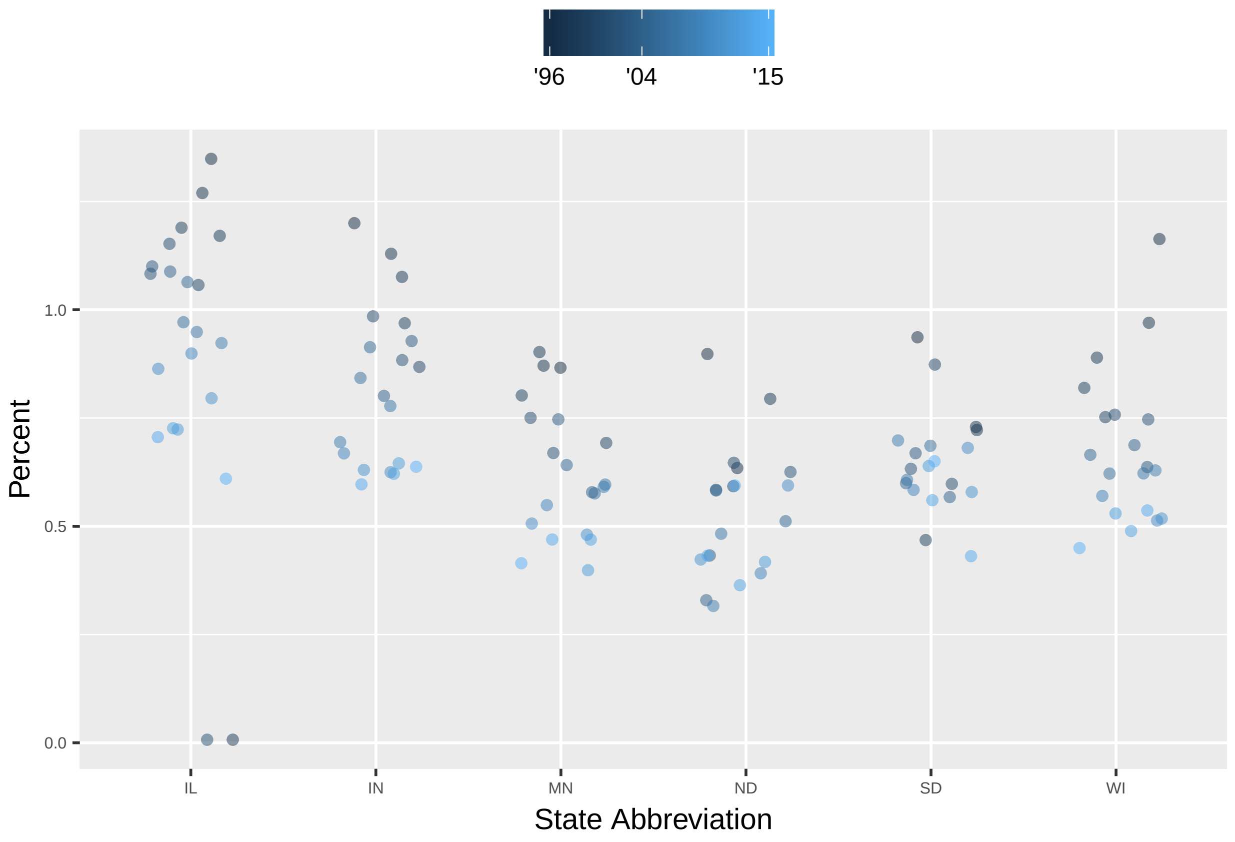
Start by constructing a dotplot showing the distribution of the popularity of the name "David" by Location and Year between 2000 and 2005 in the Midwest.

```
p1 = ggplot(subset(babyNames,Name=="david" & Location %in% c("MN","ND","SD","IN","WI","IL")), #I'm goin
             aes(x = Location,
                 y = Percent)) +
        theme(legend.position="top",
              axis.text=element_text(size = 6))
(p2 <- p1 + geom_point(aes(color = Year),
                       alpha = 0.5,
                       size = 1.5,
                       position = position_jitter(width = 0.25, height = 0)))
```



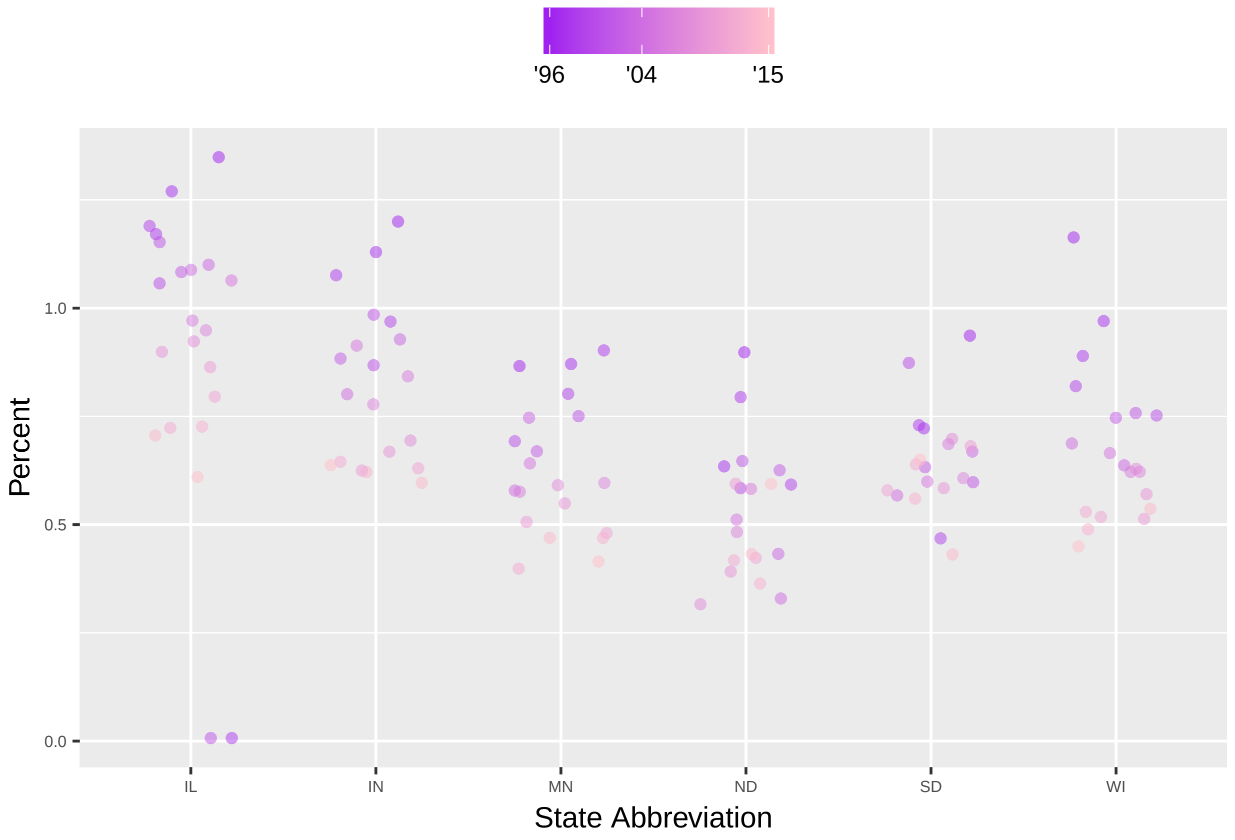
Now we might want to modify the legend and x axis label:

```
p2 + scale_x_discrete(name="State Abbreviation") + #changes our X axis label
  scale_color_continuous(name="", #changes the look of our legend
                         breaks = c(1996, 2004, 2015),
                         labels = c("'96", "'04", "'15"))
```



Let's change the "low" and "high" colors of the points:

```
p2 + scale_x_discrete(name="State Abbreviation") + #changes our X axis label
  scale_color_continuous(name="", #changes the look of our legend
                         breaks = c(1996, 2004, 2015),
                         labels = c("'96", "'04", "'15"),
                          low = "purple", high = "pink")
```



Note that in RStudio you can type scale_ followed by TAB to get the whole list of available scales.

Exercise 2

1. Create a scatter plot with CPI on the x axis and HDI on the y axis. Color the points to indicate region.
2. Modify the x, y, and color scales so that they have more easily-understood names (e.g., spell out “Human development Index” instead of “HDI”).
3. Modify the color scale to use specific values of your choosing. Hint: see ?scale_color_manual.

Faceting

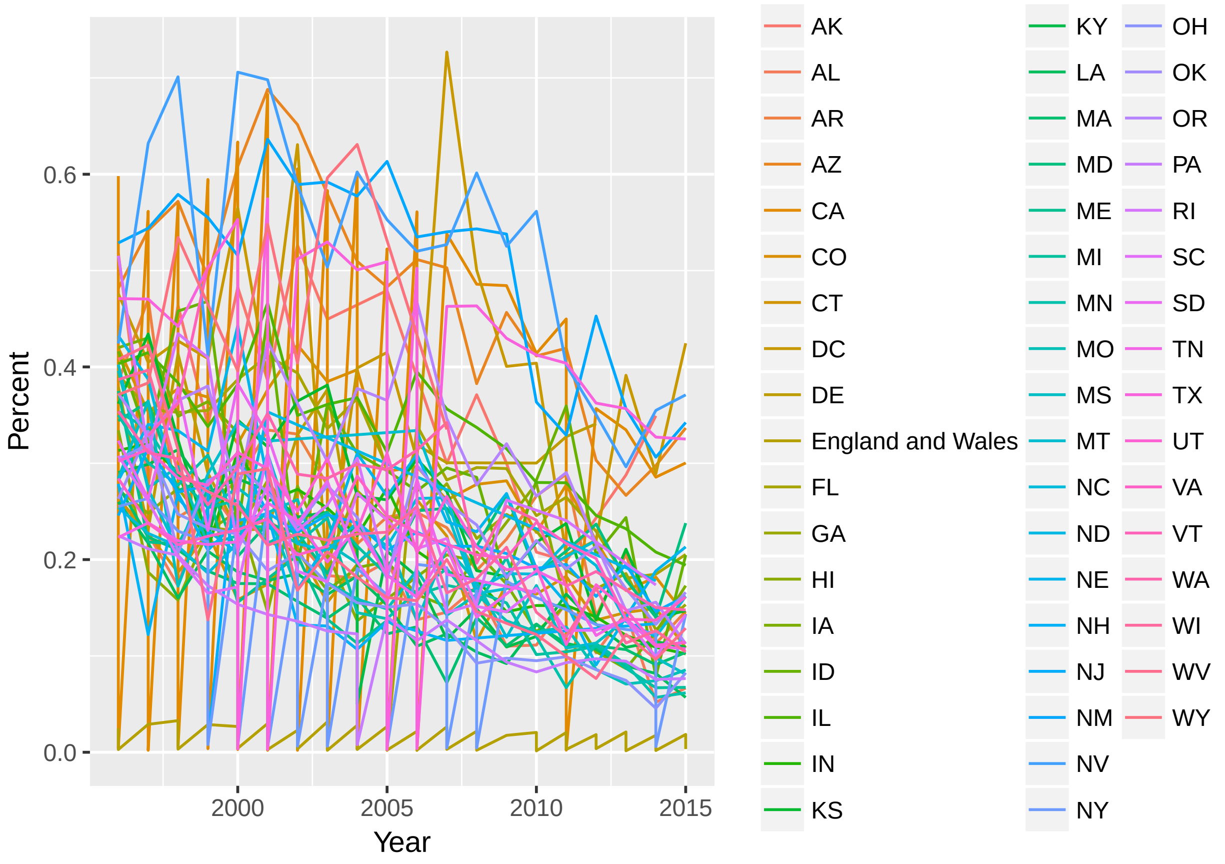
Faceting is the idea of creating separate graphs for subsets of data ggplot2 offers two functions for creating small graphs:

1. facet_wrap(): define subsets as the levels of a single grouping variable
2. facet_grid(): define subsets as the crossing of two grouping variables

Facilitates comparison among plots, not just of geoms within a plot

Let’s create a rather messy plot of the popularity of the name “andrea” over time:

```
ggplot(subset(babyNames,Name=="andrea"), aes(x = Year, y = Percent)) + geom_line(aes(color = Location))
```

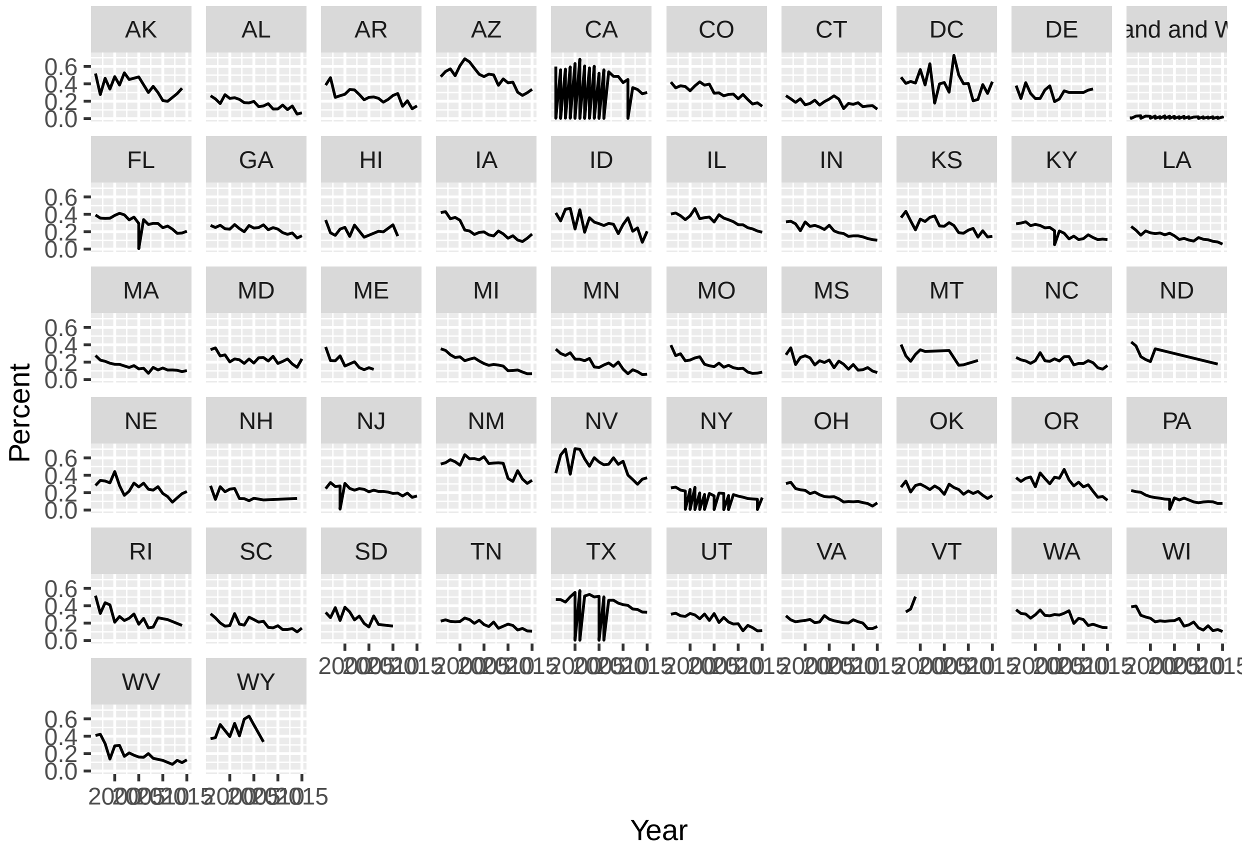


There are two problems here–there are too many states to distinguish each one by color, and the lines obscure one another.

Faceting to the rescue

We can remedy the deficiencies of the previous plot by faceting by state rather than mapping state to color.

```
ggplot(subset(babyNames,Name=="andrea"), aes(x = Year, y = Percent)) + geom_line() + facet_wrap(~Locati
```

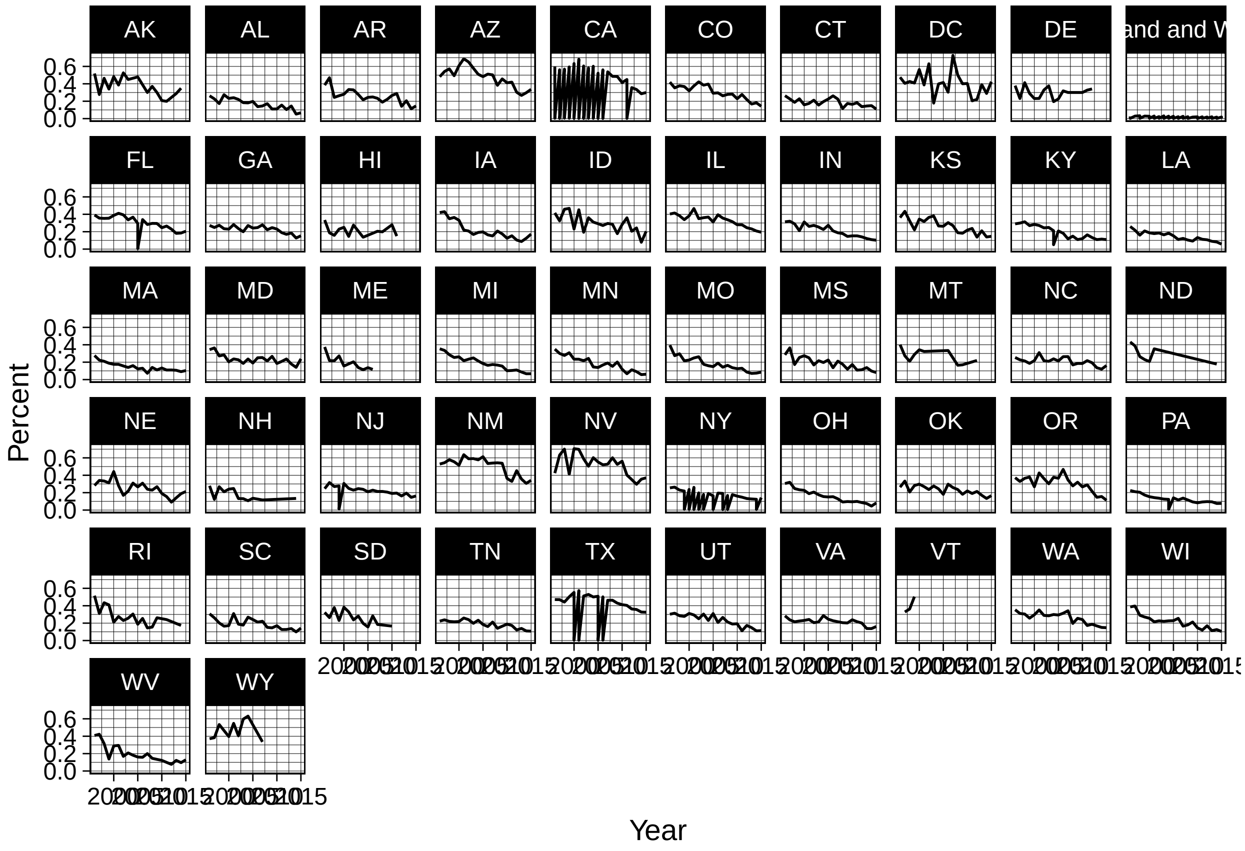


Themes

The ggplot2 theme system handles non-data plot elements such as

- Axis labels
- Plot background
- Facet label backround
- Legend appearance
- Built-in themes include: theme_gray() (default), theme_bw() and theme_classc()

```
ggplot(subset(babyNames,Name=="andrea"), aes(x = Year, y = Percent)) + geom_line() + facet_wrap(~Locati
```

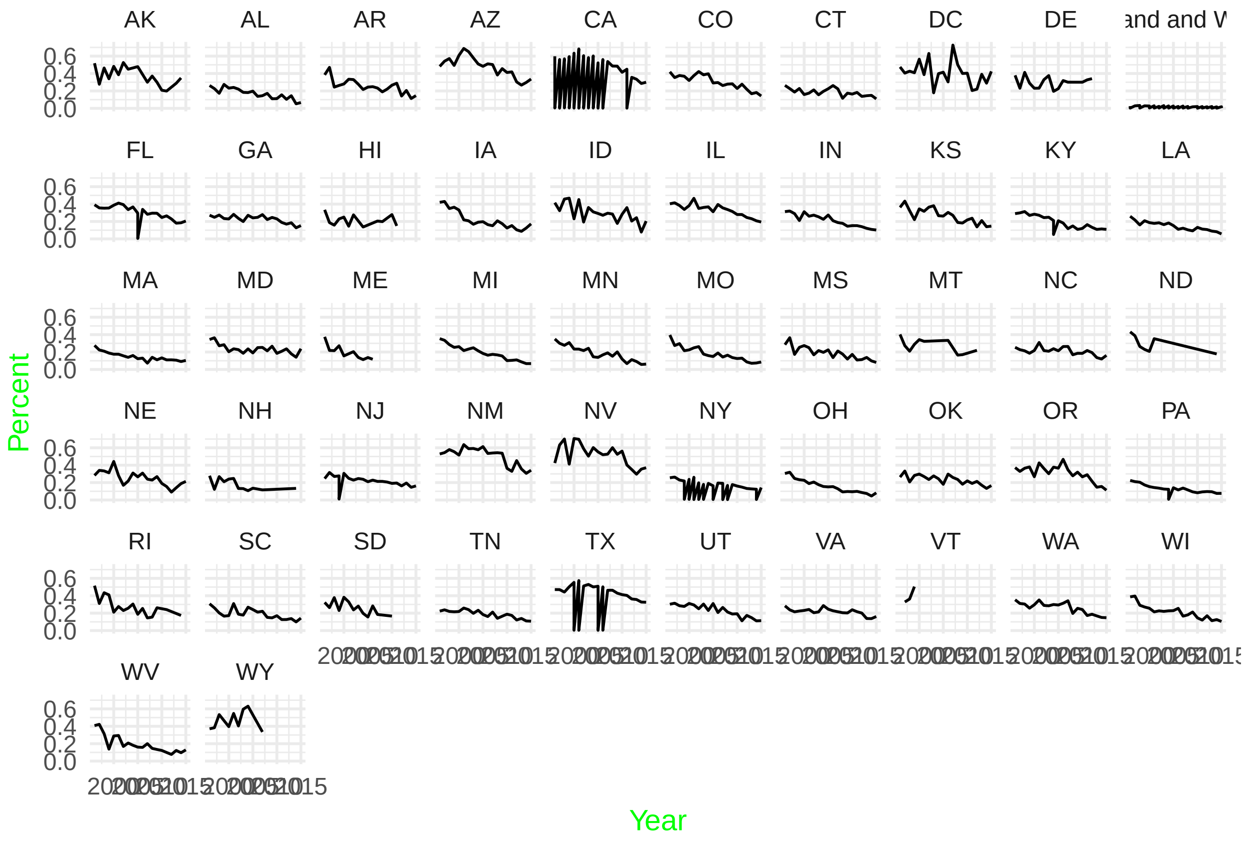


Which slightly changes the appearence of the boxes.

Overriding theme defaults

Specific theme elements can be overridden using theme(). For example:

```
ggplot(subset(babyNames,Name=="andrea"), aes(x = Year, y = Percent)) + geom_line() + facet_wrap(~Locati
  theme(text = element_text(color = "green"))
```

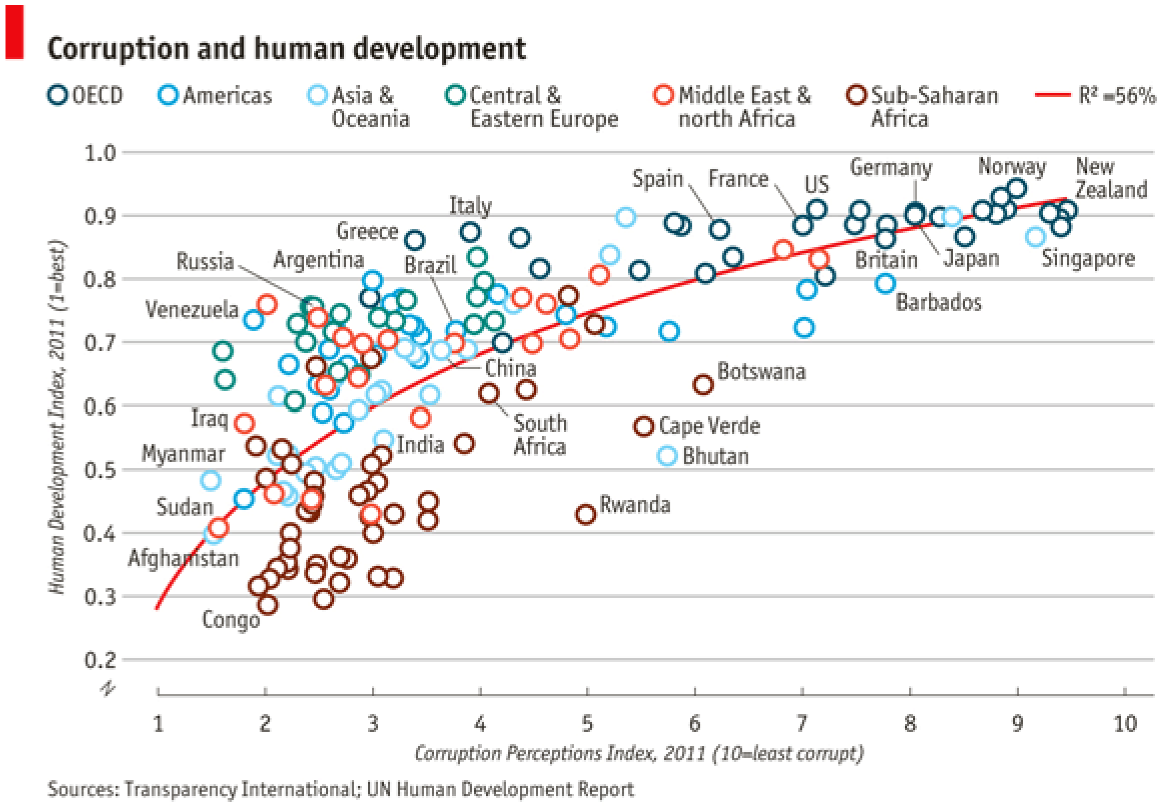


Final challenge: Recreating the Economist graph!

Building off of the graphics you created in the previous exercises, put the finishing touches to make it as close as possible to the original economist graph.

Hint: do this in steps!

1. Create a scatterplot
2. Add a trend line
3. Change the point shape to open circle
4. Change the order and labels of Region 5.Label select points
5. Fix up the tick marks and labels
6. Move color legend to the top
7. Title, label axes, remove legend title
8. Theme the graph with no vertical guides
9. Add model R2 (hard)
10. Add sources note (hard)
11. Final touches to make it perfect (use image editor for this)



Corruption and human development
OECD
Americas
Asia & Oceania
Central & Eastern Europe
Middle East & north Africa
Sub-Saharan Africa
R² =56%
Human Development Index, 2011 (1=best)
Corruption Perceptions Index, 2011 (10=least corrupt)
Spain
France
US
Germany
Norway
New Zealand
Italy
Greece
Russia
Argentina
Brazil
Venezuela
Britain
Japan
Singapore
Barbados
China
Botswana
Iraq
South Africa
Cape Verde
Bhutan
Myanmar
India
Sudan
Rwanda
Afghanistan
Congo
Sources: Transparency International; UN Human Development Report

Figure 2: