Big Data in Economics

Lecture 1: Introduction

Grant McDermott
University of Oregon | EC 510

Table of contents

Prologue

Introductions

Course

https://github.com/uo-ec510-2020-spring

You'll soon receive access to a linked GitHub Classroom environment, which is how you'll receive and submit assignments, etc.

Us

Grant McDermott (instructor)

Assistant Professor of Economics

https://grantmcdermott.com

grantmcd@uoregon.edu

Garrett Stanford (GE)

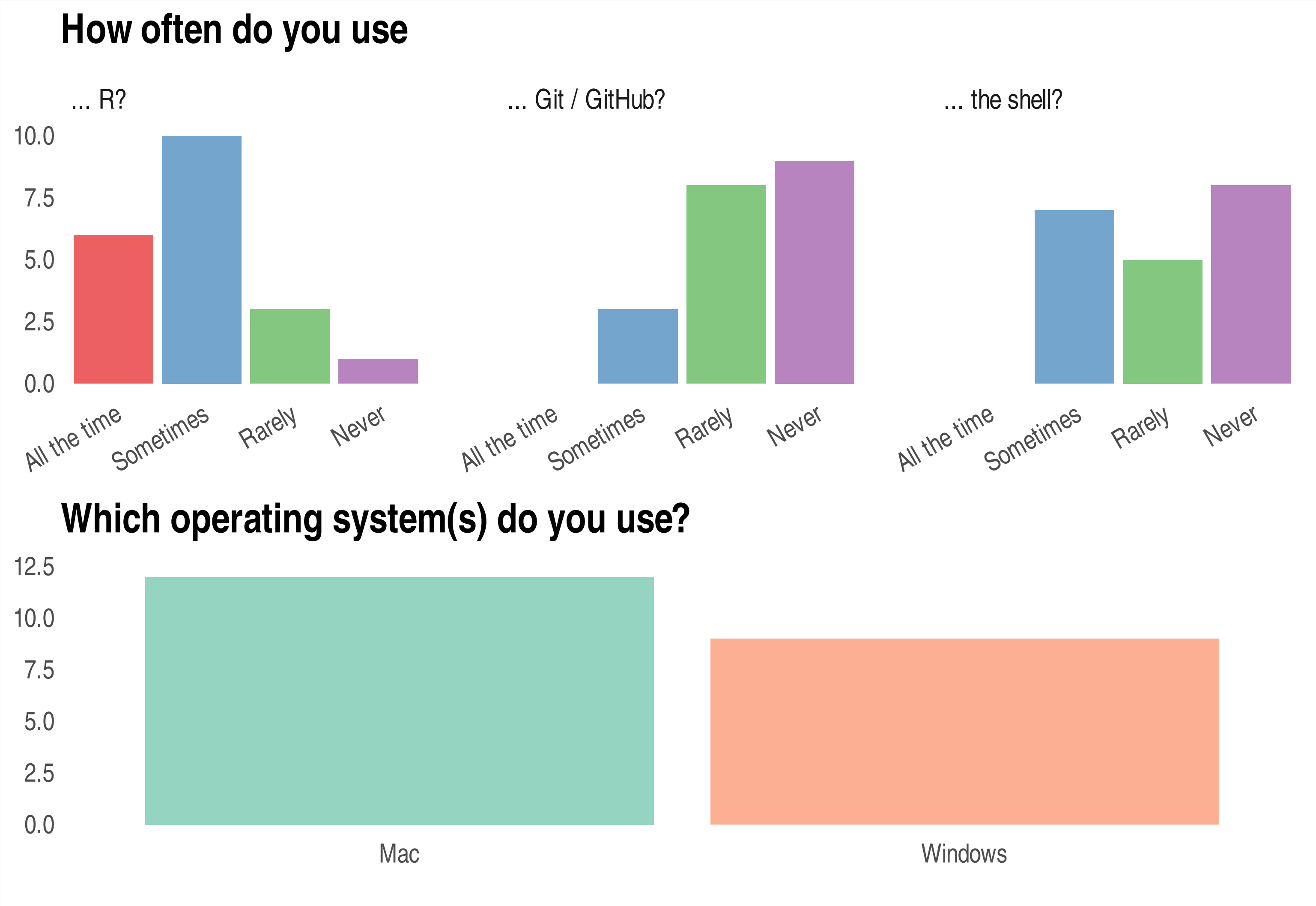
Doctoral student in economics

gos@uoregon.edu

You

Summary of our survey...

Survey


How often do you use
... R?
... Git / GitHub?
... the shell?
10.0
7.5
5.0
2.5
0.0
All the time
Sometimes
Rarely
Never
Which operating system(s) do you use?
12.5
Mac
Windows

A quick note on remote instruction

This is not only a new course (albeit derived from my EC 607 class), but also the first time most (all?) of us have experimented with remote instruction.

I'm going to try Zoom to start, but may switch to precorded lectures depending on how things go.

Regardless, I have some ideas on how we'll proceed (e.g. randomly select some of you to turn your videos on for visual feedback and cues), but we'll very much be experimenting as we go.

- The lecture notes will be very detailed, though.

Motivation

Why this course?

As data get bigger, economists need to look beyond their traditional empirical tools.

- How do you run a regression on data that won't even fit into memory?
- How do you make the most of the computational resources at your disposal?
- How do you acquire more resources?
- How do you write efficient, fault-tolerant code that scales and reproduces in multiple environments?
- Etc.

Big data are everywhere

- Cellphones
- Social media
- Networks
- Remote sensing (i.e. satellites)
- Transaction data (credit cards, etc.)
- Finance
- Transportation
- Administrative data
- etc.

Moreover, economists are increasingly working with "traditional" big data sources that, until recently, were more closely identified with other fields (genomics, climate, etc.)

Why this course? (redux)

You are going to learn some of the key tools and frameworks for dealing with big data.

- We're going to start small, learning how to write efficient code on our local machines and then scale up.

> *"By the end of the quarter, you will have connected to cloud-based based services and high-performance computing clusters, queried petabyte-sized databases, and run distributed code across a network of computers. More importantly, you will have a better understanding of how computers work, what tools are at your disposal for tackling big data problems, and how to meaningfully integrate them into your everyday workflow"*

Caveat: We will not be doing much, if any, machine-learning in this class. For that, you should take Ed Rubin's EC 524 class (part of the same MSc sequence).

Expectations

Part of the new Masters in Applied Economics program here at UO.

- Prereqs: MATH 253; EC 311; EC 313; one from EC 320, EC 423.
- For undergrads, I have generally only admitted students that received an "A" for EC 421.

Formal prereqs aside, my basic working assumption is that you have a solid understanding of R (data wrangling, econometrics, etc.) that we can build on quickly.

- We will, however, spend some time laying proper foundations for the latter sections.

We'll get to software in a few slides, but the lecture notes (a) will be posted ahead of class and (b) will contain a list of SW requirements for that lecture. My expectation is that you will have completed the SW requirements *before* class.

You, at the end of this course

Syllabus highlights

(Read the full document here.)

Grading

EC 410

Component	Weight
4 × HW assignments (25% each)	100%

EC 510

Component	Weight
4 × HW assignments (20% each)	80%
1 × OSS contribution	20%

- HW assignments are to be completed in **groups of 2**. If you'd like to pair up, let me know soon. Else, I'm going to assign groups at random.
- For the EC 510 students, I encourage you to make an OSS contribution to LOST.

PS — I may award a class participation bonus (2.5%) at my discretion.

Lecture outline

Foundations

- Introduction: Motivation, software installation, and data visualization
- Version control with Git(Hub)
- Learning to love the shell
- R language basics (*Optional*)

Data wrangling, I/O, and acquistion

- Data cleaning and wrangling: (1) Tidyverse
- Data cleaning and wrangling: (2) data.table
- Big data I/O
- Webscraping: (1) Server-side and CSS
- Webscraping: (2) Client-side and APIs

Lecture outline (cont.)

Programming

- Functions in R: (1) Introductory concepts
- Functions in R: (2) Advanced concepts
- Parallel programming

Cloud resources and distributed computation

- Docker
- Cloud computing with Google Compute Engine
- High performance computing (Talapas cluster)
- Databases: SQL(ite) and BigQuery
- Spark

Lecture outline (cont.)

Other potential topics (time permitting)

- Regression tools for big data problems
- Google Earth Engine
- Networks
- Deep learning
- Automation and workflow
- Rcpp (i.e. integrating C++ with R)

Getting started

Software installation and registration

1. Download R.

2. Download RStudio.

3. Download Git.

4. Create an account on GitHub and register for a student/educator discount.

    - You will soon receive an invitation to the quarter-specific course org. on GitHub, as well as GitHub classroom, which is how we'll disseminate and submit assignments, receive feedback and grading, etc.

If you had trouble completing any of these steps, please let me know ASAP.

- My go-to place for installation guidance and troubleshooting is Jenny Bryan's http://happygitwithr.com.

Some OS-specific extras

I'll detail further software requirements as and when the need arises.[1] However, to help smooth some software installation issues further down the road, please also do the following (depending on your OS):

- **Windows:** Install Rtools. I also recommend that you install Chocolately.
- **Mac:** Install Homebrew. I also recommend that you configure/open your C++ toolchain (see here.)
- **Linux:** None (you should be good to go).

[1] **Important:** I expect you to have fulfilled any SW requirements well before class starts. Don't wait until the last minute, otherwise we won't be able to help troubleshoot.

Checklist

☑ Do you have the most recent version of R?

```
version$version.string
```

```
## [1] "R version 3.6.3 (2020-02-29)"
```

☑ Do you have the most recent version of RStudio? (The preview version is fine.)

```
RStudio.Version()$version
## Requires an interactive session but should return something like "[1] ‘1.3.92
```

☑ Have you updated all of your R packages?

```
update.packages(ask = FALSE, checkBuilt = TRUE)
```

Checklist (cont.)

Open up the shell.

- Windows users, make sure that you installed a Bash-compatible version of the shell. If you installed Git for Windows, then you should be good to go.

☑ Which version of Git have you installed?

```
git --version
```

```
## git version 2.25.1
```

☑ Did you introduce yourself to Git? (Substitute in your details.)

```
git config --global user.name 'Grant McDermott'
git config --global user.email 'grantmcd@uoregon.edu'
git config --global --list
```

☑ Did you register an account in GitHub?

Checklist (cont.)

We will make sure that everything is working properly with your R and GitHub setup next lecture.

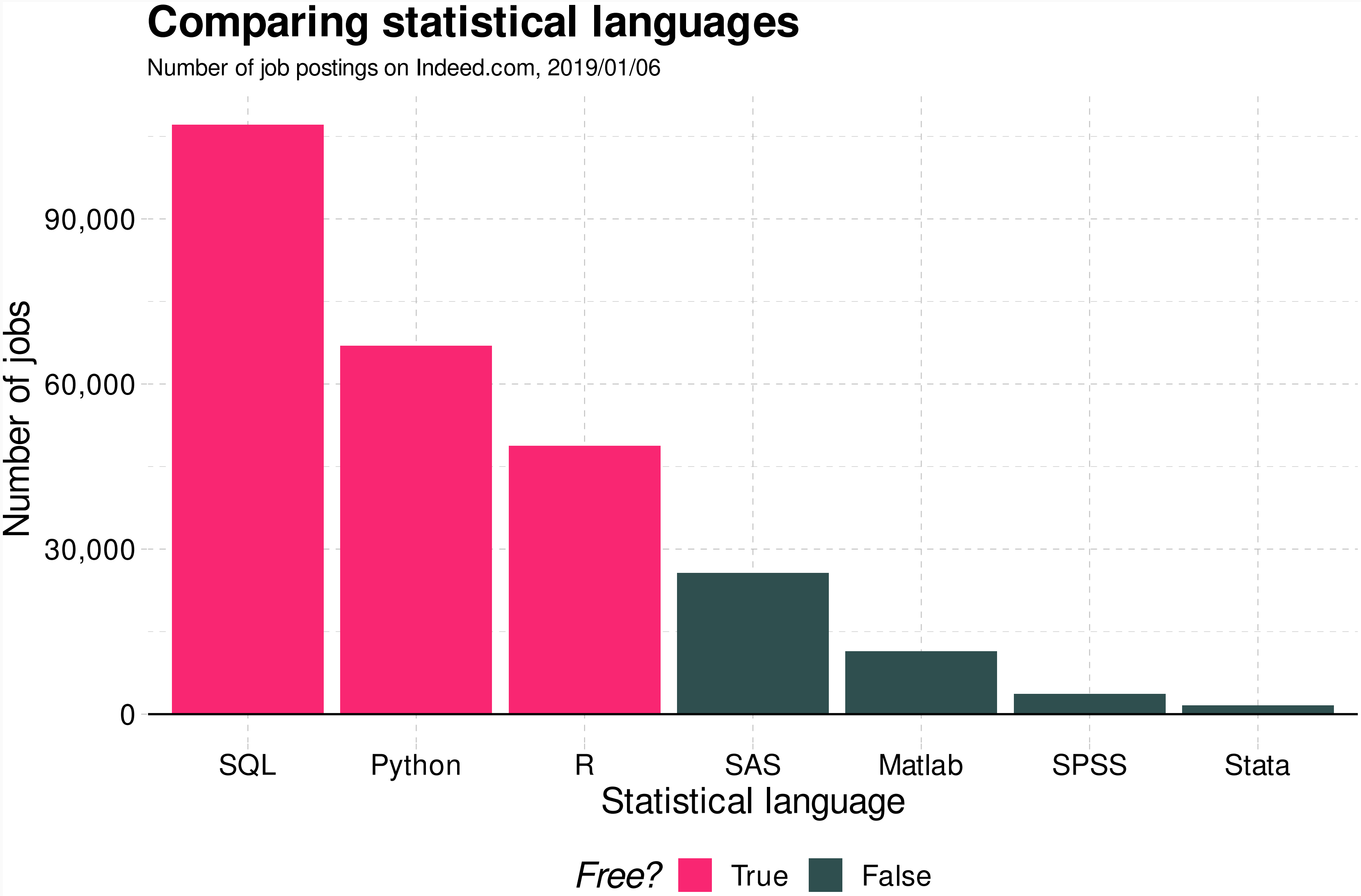
For the rest of today's lecture, I want to go over some very basic R concepts.

PS — Just so you know where we're headed: We'll return to these R concepts (and delve much deeper) next week after a brief, but important detour to the lands of Git(Hub) and the shell.

- Don't worry, it will all make sense. You'll see.

R for data science

Why R and RStudio? (cont.)



Why R and RStudio? (cont.)

Data science positivism

- Alongside Python, R has become the *de facto* language for data science.
  - See: *The Impressive Growth of R*, *The Popularity of Data Science Software*
- Open-source (free!) with a global user-base spanning academia and industry.
  - "Be a profit center, not a cost center." (source)

Bridge to applied economics and other tools

- Already has all of the statistics and econometrics support, and is amazingly adaptable as a “glue” language to other programming languages and APIs.
- The RStudio IDE and ecosystem allow for further, seamless integration.

Path dependency

- It's also the language that I know best.
- (Learning multiple languages is a good idea, though.)

Some R basics

1. Everything is an object.
2. Everything has a name.
3. You do things using functions.
4. Functions come pre-written in packages (i.e. "libraries"), although you can — and should — write your own functions too.

Points 1. and 2. can be summarised as an object-orientated programming (OOP) approach.

- This may sound super abstract now, but we'll see *lots* of examples over the coming weeks that will make things clear.

R vs Stata

If you're coming from Stata, some additional things worth emphasizing:

- Multiple objects (e.g. data frames) can exist happily in the same workspace.
  - No more `keep`, `preserve`, `restore` hackery. (Though, props to Stata 16.)
  - This is a direct consequence of the OOP approach.
- You will load packages at the start of every new R session. Make peace with this.
  - "Base" R comes with tons of useful in-built functions. It also provides all the tools necessary for you to write your own functions.
  - However, many of R's best data science functions and tools come from external packages written by other users.
- R easily and infinitely parallelizes. For free.
  - Compare the cost of a Stata/MP license, nevermind the fact that you effectively pay per core...
- You don't need to `tset or xtset` your data. (Although you can too.)

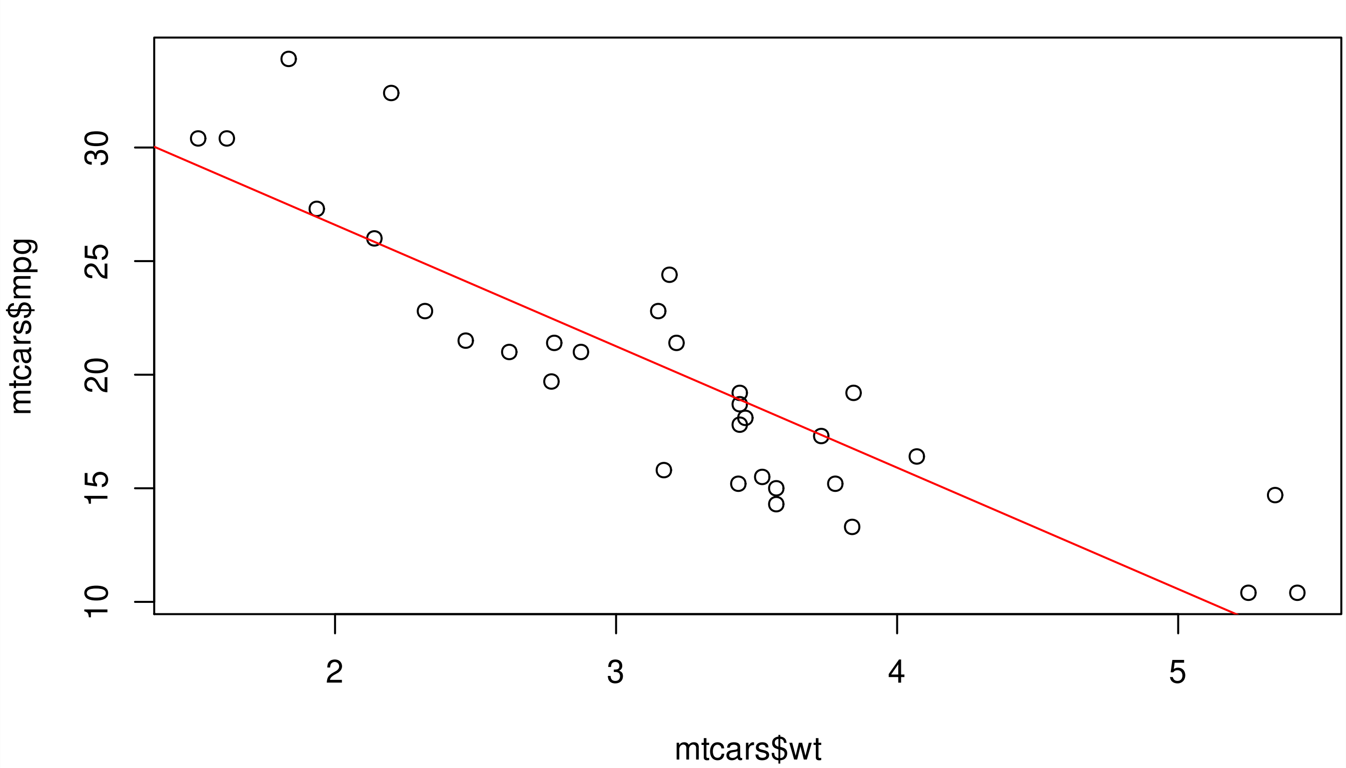
R code example (linear regression)

```
fit = lm(mpg ~ wt, data = mtcars)
summary(fit)
```

```
##
## Call:
## lm(formula = mpg ~ wt, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5432 -2.3647 -0.1252  1.4096  6.8727
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  37.2851     1.8776  19.858  < 2e-16 ***
## wt           -5.3445     0.5591  -9.559 1.29e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.046 on 30 degrees of freedom
## Multiple R-squared:  0.7528,    Adjusted R-squared:  0.7446
## F-statistic: 91.38 on 1 and 30 DF,  p-value: 1.294e-10
```

Base R plot

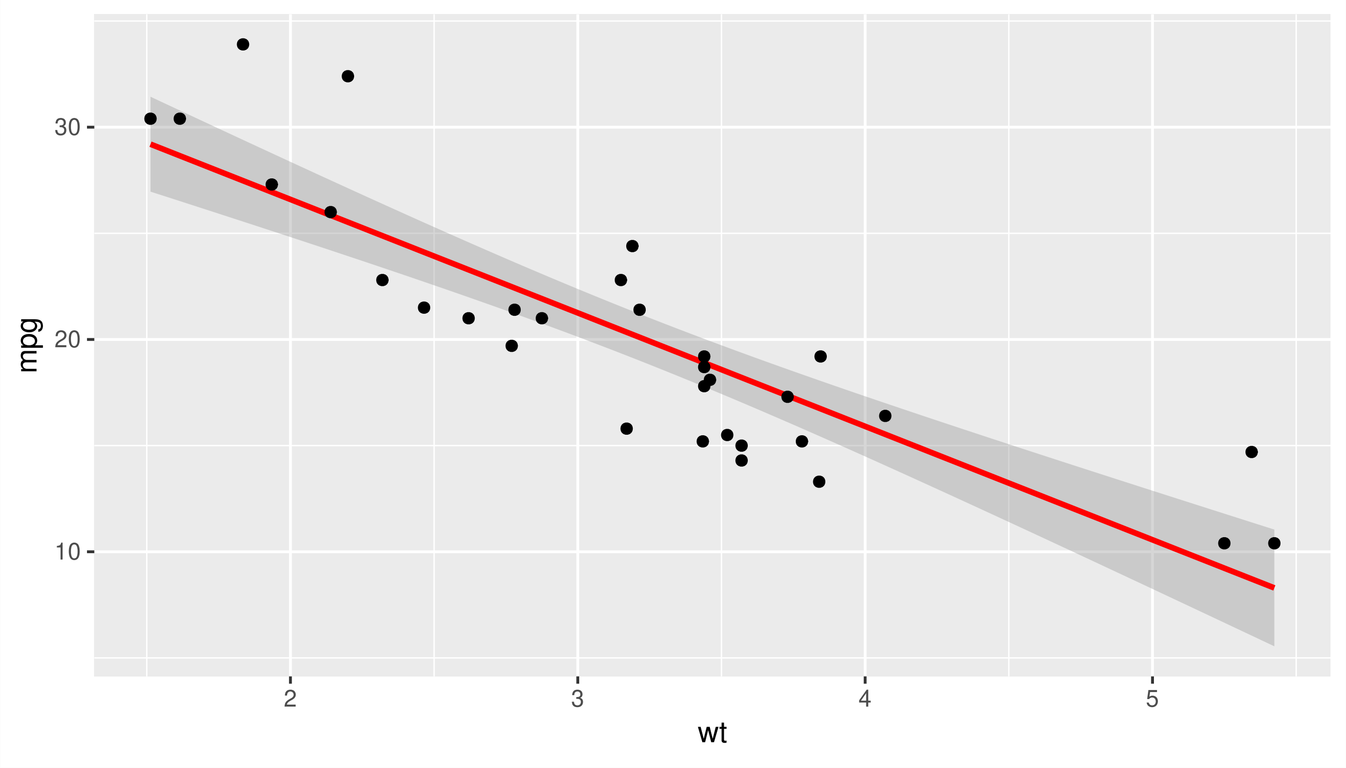
```
par(mar = c(4, 4, 1, .1)) ## Just for nice plot margins on this slide deck
plot(mtcars$wt, mtcars$mpg)
abline(fit, col = "red")
```



ggplot2

```r
library(ggplot2)
ggplot(data = mtcars, aes(x = wt, y = mpg)) +
  geom_smooth(method = "lm", col = "red") +
  geom_point()
```

```
## `geom_smooth()` using formula 'y ~ x'
```



More ggplot2

Install and load

Open up your laptops. For the remainder of this first lecture, we're going continue playing around with ggplot2 (i.e. livecoding).

If you don't have them already, install the `ggplot2` and `gapminder` packages via either:

- **Console:** Enter `install.packages(c("ggplot2", "gapminder"), dependencies=T)`.
- **RStudio:** Click the "Packages" tab in the bottom-right window pane. Then click "Install" and search for these two packages.

Install and load (cont.)

Once the packages are installed, load them into your R session with the `library()` function.

```
library(ggplot2)
library(gapminder) ## We're just using this package for the gapminder data
```

Notice too that you don't need quotes around the package names any more. Reason: R now recognises these packages as defined objects with given names. ("Everything in R is an object and everything has a name.")

PS — A convenient way to combine the package installation and loading steps is with the pacman package's `p_load()` function. If you run `pacman::p_load(ggplot, gapminder)` it will first look to see whether it needs to install either package before loading them. Clever.

- We'll get to this next week, but if you want to run a function from an (installed) package without loading it, you can use the `PACKAGE::package_function()` syntax.

Brief aside: The gapminder dataset

Because we're going to be plotting the gapminder dataset, it is helpful to know that it contains panel data on life expectancy, population size, and GDP per capita for 142 countries since the 1950s.

```
gapminder
```

```
## # A tibble: 1,704 x 6
##    country     continent  year lifeExp      pop gdpPercap
##    <fct>       <fct>     <int>   <dbl>    <int>     <dbl>
##  1 Afghanistan Asia       1952    28.8  8425333      779.
##  2 Afghanistan Asia       1957    30.3  9240934      821.
##  3 Afghanistan Asia       1962    32.0 10267083      853.
##  4 Afghanistan Asia       1967    34.0 11537966      836.
##  5 Afghanistan Asia       1972    36.1 13079460      740.
##  6 Afghanistan Asia       1977    38.4 14880372      786.
##  7 Afghanistan Asia       1982    39.9 12881816      978.
##  8 Afghanistan Asia       1987    40.8 13867957      852.
##  9 Afghanistan Asia       1992    41.7 16317921      649.
## 10 Afghanistan Asia       1997    41.8 22227415      635.
## # … with 1,694 more rows
```

Elements of ggplot2

Hadley Wickham's ggplot2 is one of the most popular packages in the entire R canon.

- It also happens to be built upon some deep visualization theory: i.e. Leland Wilkinson's *The Grammar of Graphics*.

There's a lot to say about ggplot2's implementation of this "grammar of graphics" approach, but the three key elements are:

1. Your plot ("the visualization") is linked to your variables ("the data") through various **aesthetic mappings**.

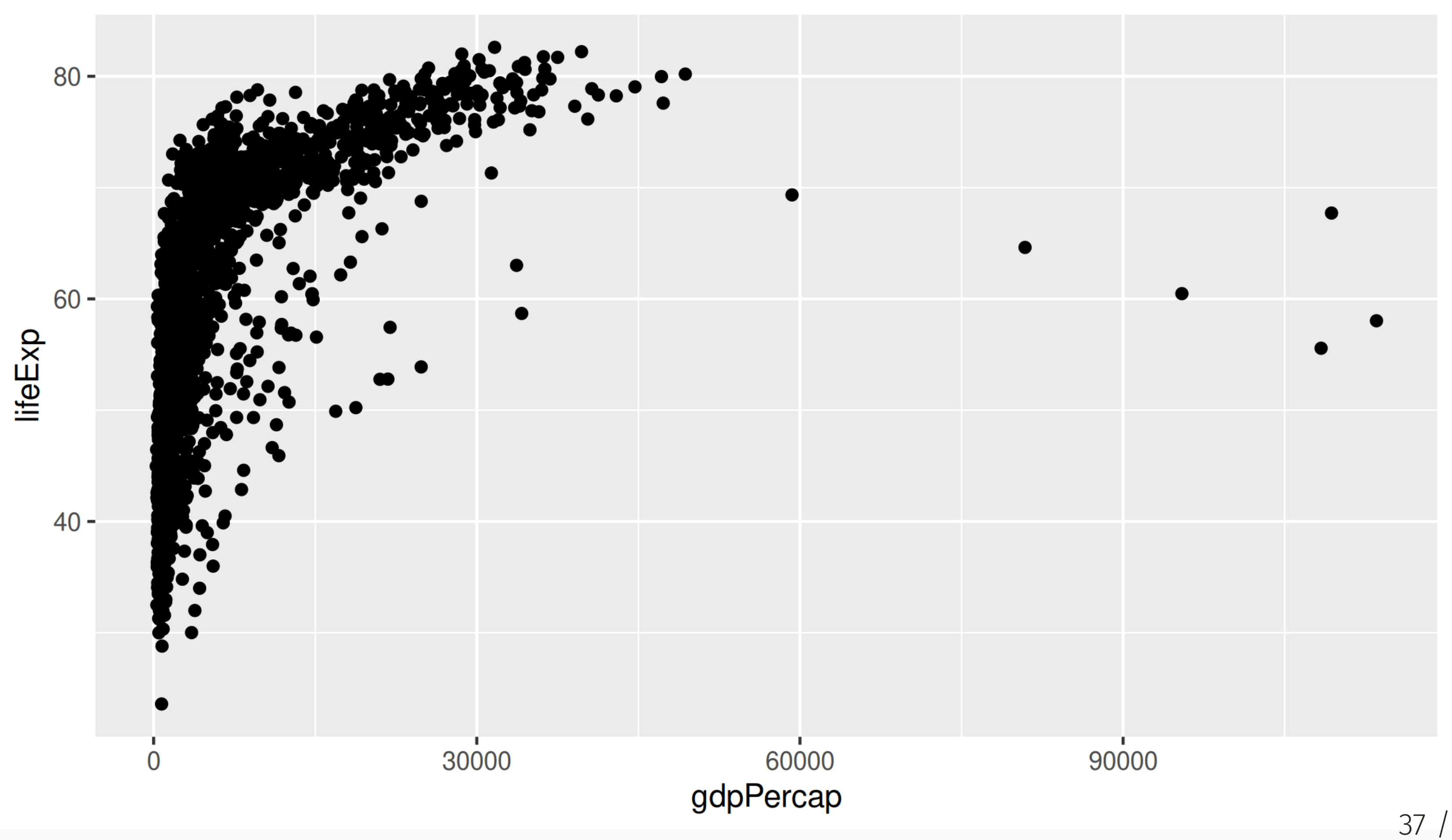
2. Once the aesthetic mappings are defined, you can represent your data in different ways by choosing different **geoms** (i.e. "geometric objects" like points, lines or bars).

3. You build your plot in **layers**.

That's kind of abstract. Let's review each element in turn with some actual plots.

1. Aesthetic mappings

```
ggplot(data = gapminder, mapping = aes(x = gdpPercap, y = lifeExp)) +
  geom_point()
```



1. Aesthetic mappings (cont.)

```
ggplot(data = gapminder, mapping = aes(x = gdpPercap, y = lifeExp)) +
  geom_point()
```

Focus on the top line, which contains the initialising `ggplot()` function call. This function accepts various arguments, including:

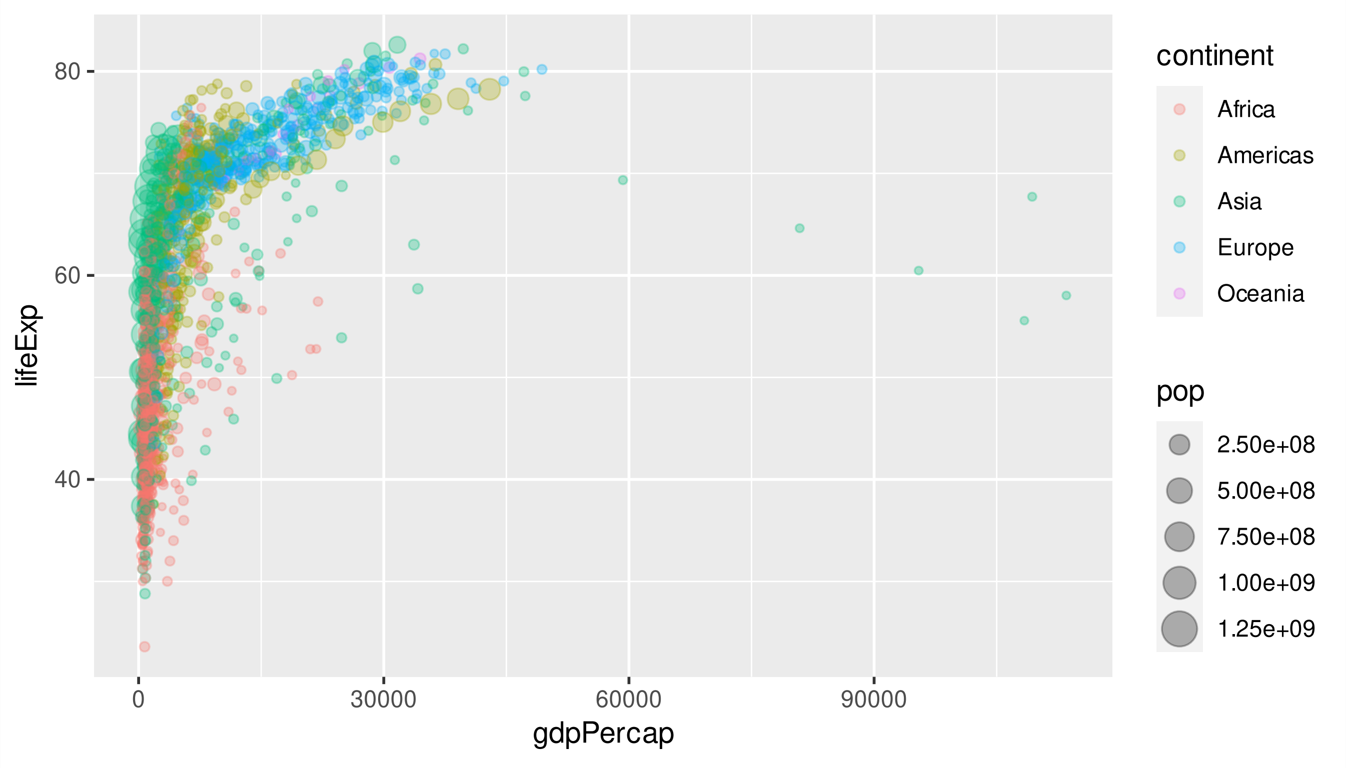
- Where the data come from (i.e. `data = gapminder`).
- What the aesthetic mappings are (i.e. `mapping = aes(x = gdpPercap, y = lifeExp)`).

The aesthetic mappings here are pretty simple: They just define an x-axis (GDP per capita) and a y-axis (life expecancy).

- To get a sense of the power and flexibility that comes with this approach, however, consider what happens if we add more aesthetics to the plot call...

1. Aesthetic mappings (cont.)

```
ggplot(data = gapminder, aes(x = gdpPercap, y = lifeExp, size = pop, col = contine
  geom_point(alpha = 0.3) ## "alpha" controls transparency. Takes a value between
```

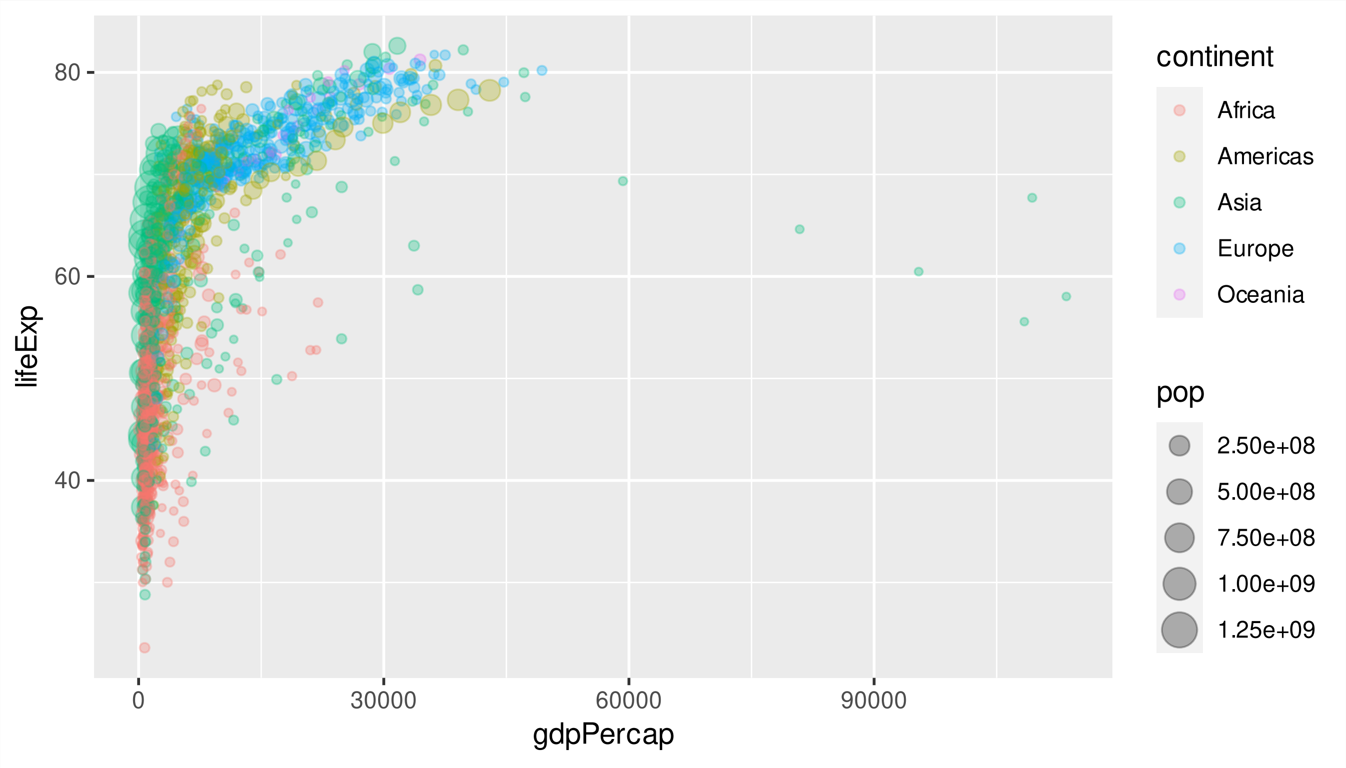


Note that I've dropped the "mapping =" part of the ggplot call. Most people just start with "aes(...)", since `ggplot2` knows the order of the arguments.

1. Aesthetic mappings (cont.)

We can specify aesthetic mappings in the geom layer too.

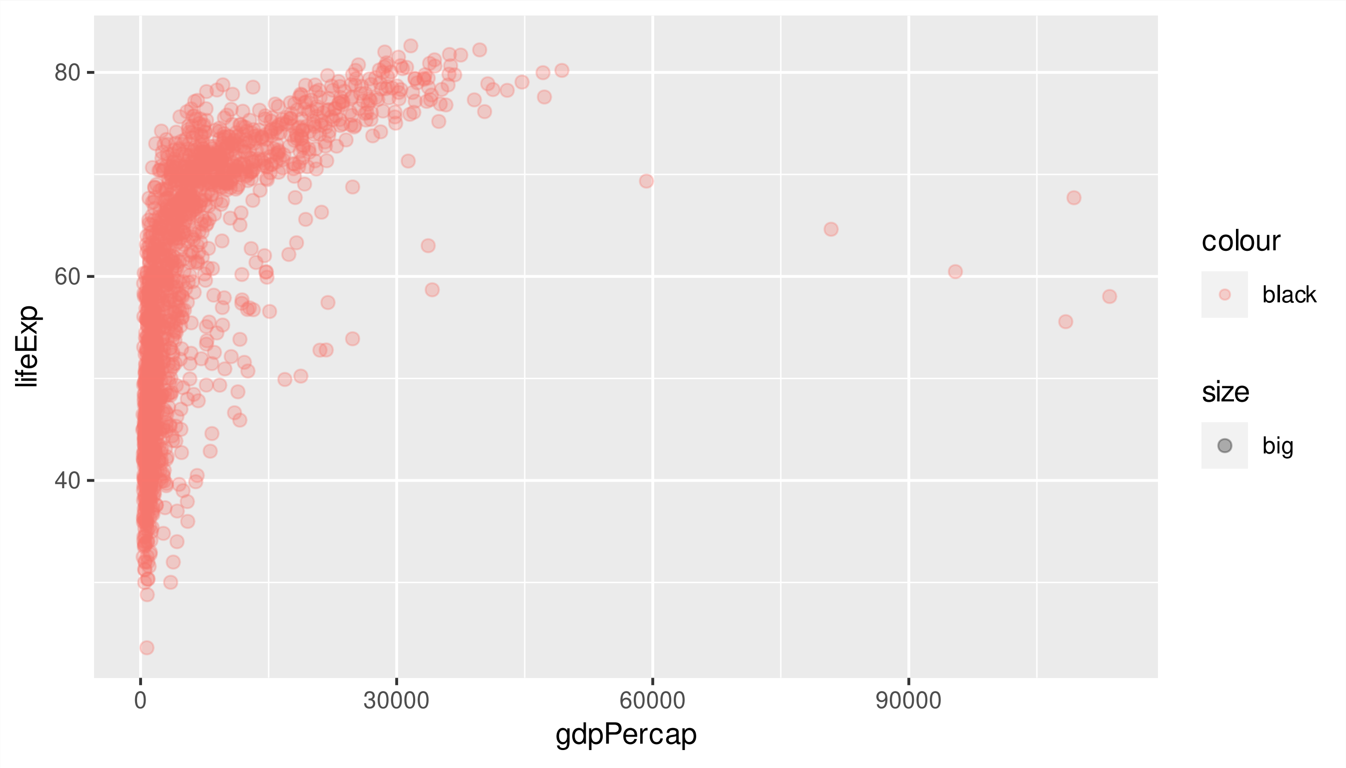
```
ggplot(data = gapminder, aes(x = gdpPercap, y = lifeExp)) + ## Applicable to all g
  geom_point(aes(size = pop, col = continent), alpha = 0.3) ## Applicable to this
```



1. Aesthetic mappings (cont.)

Oops. What went wrong here?

```
ggplot(data = gapminder, aes(x = gdpPercap, y = lifeExp)) +
  geom_point(aes(size = "big", col="black"), alpha = 0.3)
```

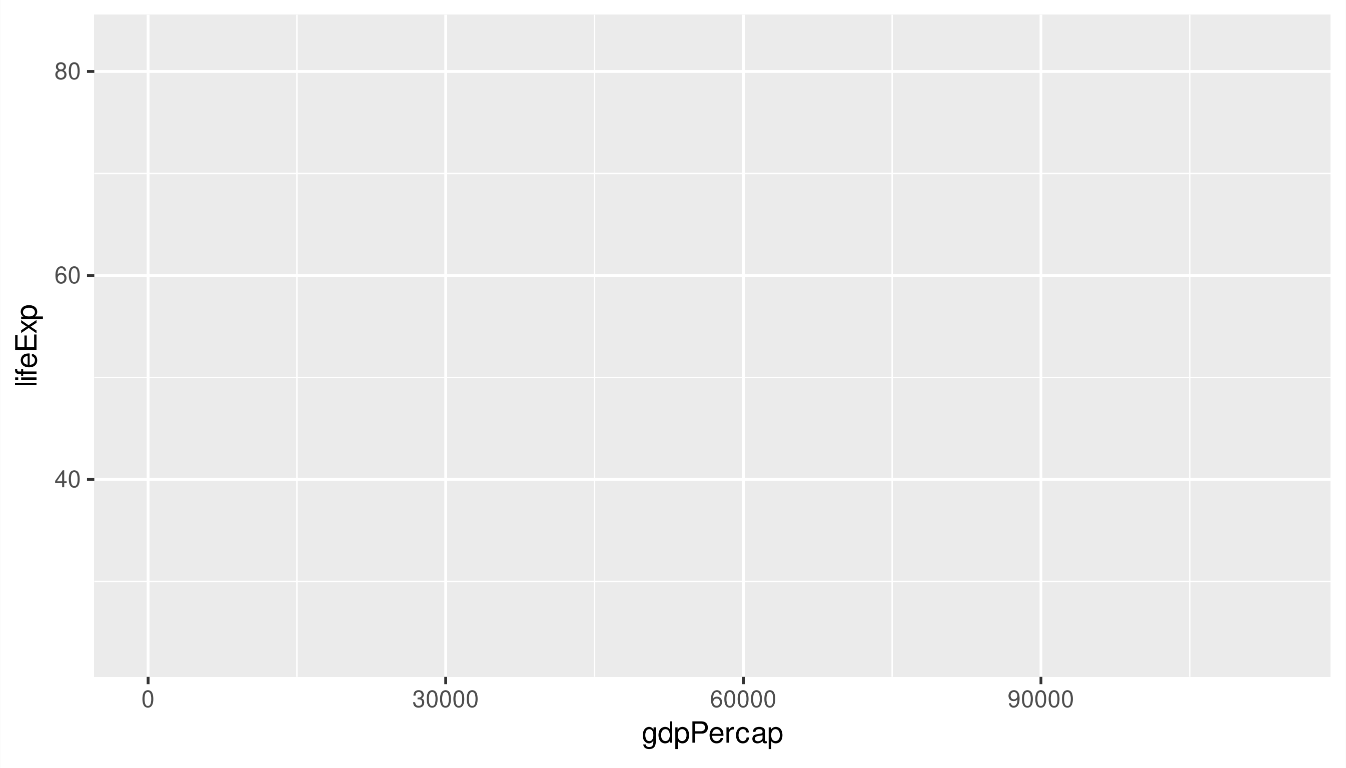


Answer: Aesthetics must be mapped to variables, not descriptions!

1. Aesthetic mappings (cont.)

At this point, instead of repeating the same ggplot2 call every time, it will prove convenient to define an intermediate plot object that we can re-use.

```
p = ggplot(data = gapminder, aes(x = gdpPercap, y = lifeExp))
p
```



2. Geoms

Once your variable relationships have been defined by the aesthetic mappings, you can invoke and combine different geoms to generate different visulaizations.

```
p +
  geom_point(alpha = 0.3)  +
  geom_smooth(method = "loess")
```

```
## `geom_smooth()` using formula 'y ~ x'
```

2. Geoms (cont.)

Aesthetics can be applied differentially across geoms.

```
p +
  geom_point(aes(size = pop, col = continent), alpha = 0.3)  +
  geom_smooth(method = "loess")
```

```
## `geom_smooth()` using formula 'y ~ x'
```

2. Geoms (cont.)

The previous plot provides a good illustration of the power (or effect) that comes from assigning aesthetic mappings "globally" vs in the individual geom layers.

- Compare: What happens if you run the below code chunk?

```
ggplot(data = gapminder, aes(x = gdpPercap, y = lifeExp, size = pop, col = contine
  geom_point(alpha = 0.3)  +
  geom_smooth(method = "loess")
```

2. Geoms (cont.)

Similarly, note that some geoms only accept a subset of mappings. E.g. `geom_density()` doesn't know what to do with the "y" aesthetic mapping.

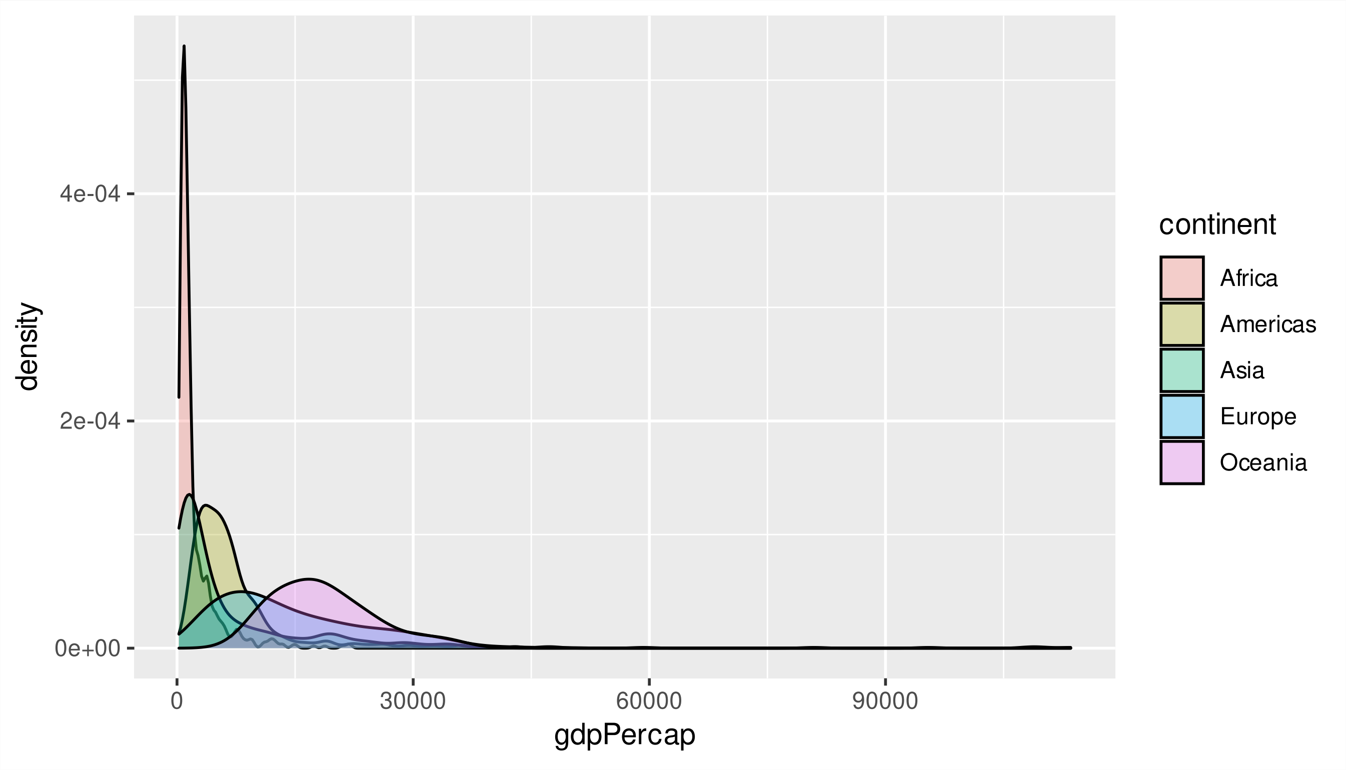
```
p + geom_density()
```

```
## Error: geom_density requires the following missing aesthetics: y
```

2. Geoms (cont.)

We can fix that by being more careful about how we build the plot.

```
ggplot(data = gapminder) + ## i.e. No "global" aesthetic mappings"
  geom_density(aes(x = gdpPercap, fill = continent), alpha=0.3)
```



3. Build your plot in layers

We've already seen how we can chain (or "layer") consecutive plot elements using the `+` connector.

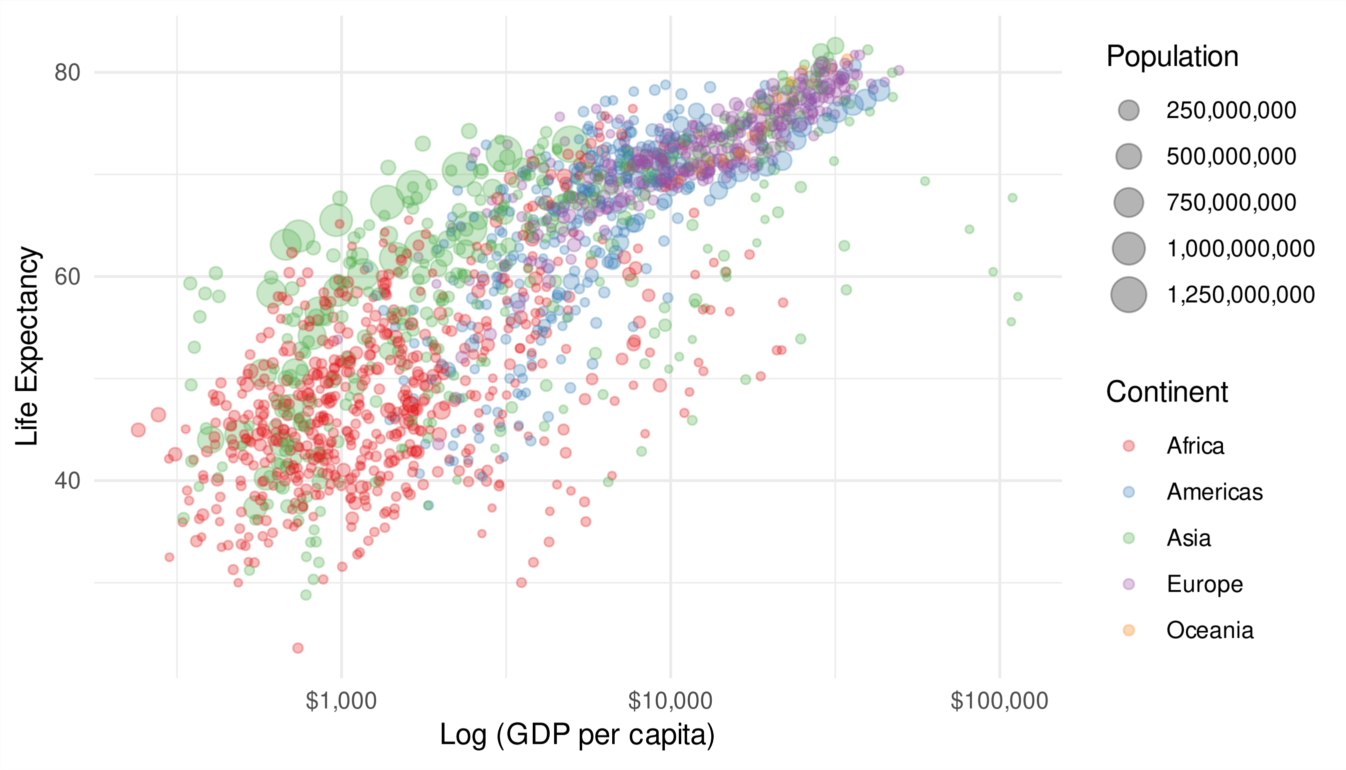
- The fact that we can create and then re-use an intermediate plot object (e.g. "p") is testament to this.

But it bears repeating: You can build out some truly impressive complexity and transformation of your visualization through this simple layering process.

- You don't have to transform your original data; ggplot2 takes care of all of that.
- For example (see next slide for figure).

```
p2 =
  p +
  geom_point(aes(size = pop, col = continent), alpha = 0.3) +
  scale_color_brewer(name = "Continent", palette = "Set1") + ## Different colour s
  scale_size(name = "Population", labels = scales::comma) + ## Different point (i.
  scale_x_log10(labels = scales::dollar) + ## Switch to logarithmic scale on x-axi
  labs(x = "Log (GDP per capita)", y = "Life Expectancy") + ## Better axis titles
  theme_minimal() ## Try a minimal (b&w) plot theme
```

3. Build your plot in layers (cont.)



What else?

We have barely scratched the surface of ggplot2's functionality... let alone talked about the entire ecosystem of packages that has been built around it.

- Here's are two quick additional examples to whet your appetite

Note that you will need to install and load some additional packages if you want to recreate the next two figures on your own machine. A quick way to do this:

```
if (!require("pacman")) install.packages("pacman")
```

```
## Loading required package: pacman
```

```
pacman::p_load(hrbrthemes, gganimate)
```

What else? (cont.)

Simple extension: Use an external package theme.

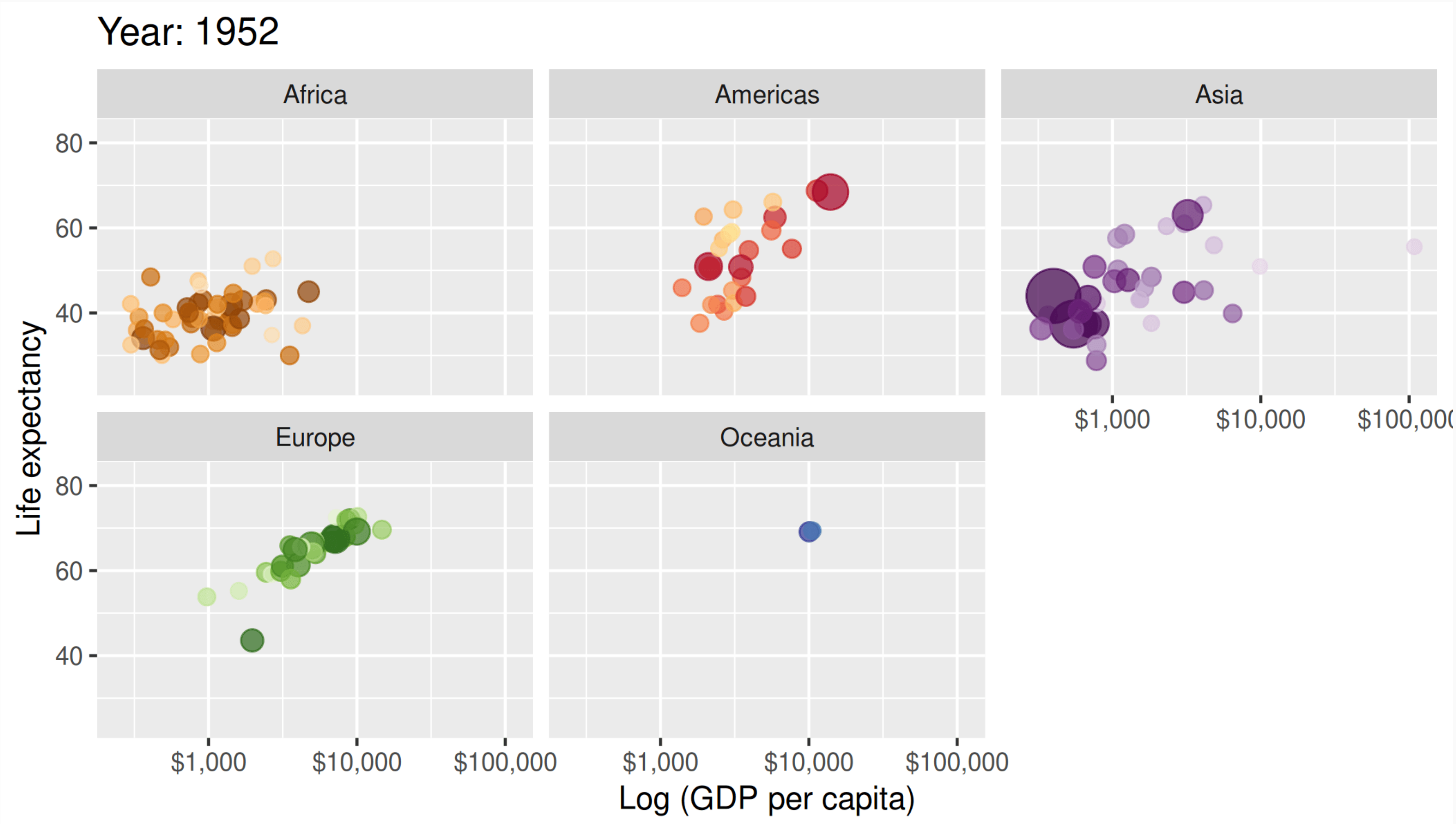
```
# library(hrbrthemes)
p2 + theme_modern_rc() + geom_point(aes(size = pop, col = continent), alpha = 0.2)
```

What else? (cont.)

Elaborate extension: Animation! (See the next slide for the resulting GIF.)

```
# library(gganimate)
ggplot(gapminder, aes(gdpPercap, lifeExp, size = pop, colour = country)) +
  geom_point(alpha = 0.7, show.legend = FALSE) +
  scale_colour_manual(values = country_colors) +
  scale_size(range = c(2, 12)) +
  scale_x_log10() +
  facet_wrap(~continent) +
  # Here comes the gganimate specific bits
  labs(title = 'Year: {frame_time}', x = 'GDP per capita', y = 'life expectancy')
  transition_time(year) +
  ease_aes('linear')
```

What else? (cont.)



Note that this animated plot provides a much more intuitive understanding of the underlying data. Just as Hans Rosling intended.

What else? (cont.)

There's a lot more to say, but I think we'll stop now for today's lecture.

We also haven't touched on ggplot2's relationship to "tidy" data.

- It actually forms part of a suite of packages collectively known as the tidyverse.
- We will get back to this in Lecture 5.

Rest assured, you will be using ggplot2 throughout the rest of this course and developing your skills along the way.

- Your very first assignment (coming up) is a chance specifically to hone some of those skills.

In the meantime, I want you to do some reading and practice on your own. Pick either of the following (or choose among the litany of online resources) and work through their examples:

- Chapter 3 of *R for Data Science* by Hadley Wickham and Garett Grolemund.
- *Data Visualization: A Practical Guide* by Kieran Healy.
- Designing ggplots by Malcom Barrett.

Next lecture: Deep dive into Git(Hub).