Quantitative Methods and Applications

Econ 5100 – Claus C Pörtner

Introductions–me

Danish (little country north of Germany)

PhD - University of Copenhagen

Research:

- Development economics
- Household and population economics
- Labor economics

Introductions–You

Background (work / study)

Where from

“Getting to know you” form

Purpose

- Set you up for success in the rest of the MSBA
- Basic statistics
- Regression analysis
- Proficient in R

Statistical tools

- Hypothesis testing
- Simple / multiple regression models with continuous dependent variables
- Model diagnostics
- Modeling choices
- Resampling methods / bootstrapping
- A bit of Bayesian if time permits

Software

Canvas

R / RStudio

You can use Mac, Windows, Linux or VLAB

Why R?

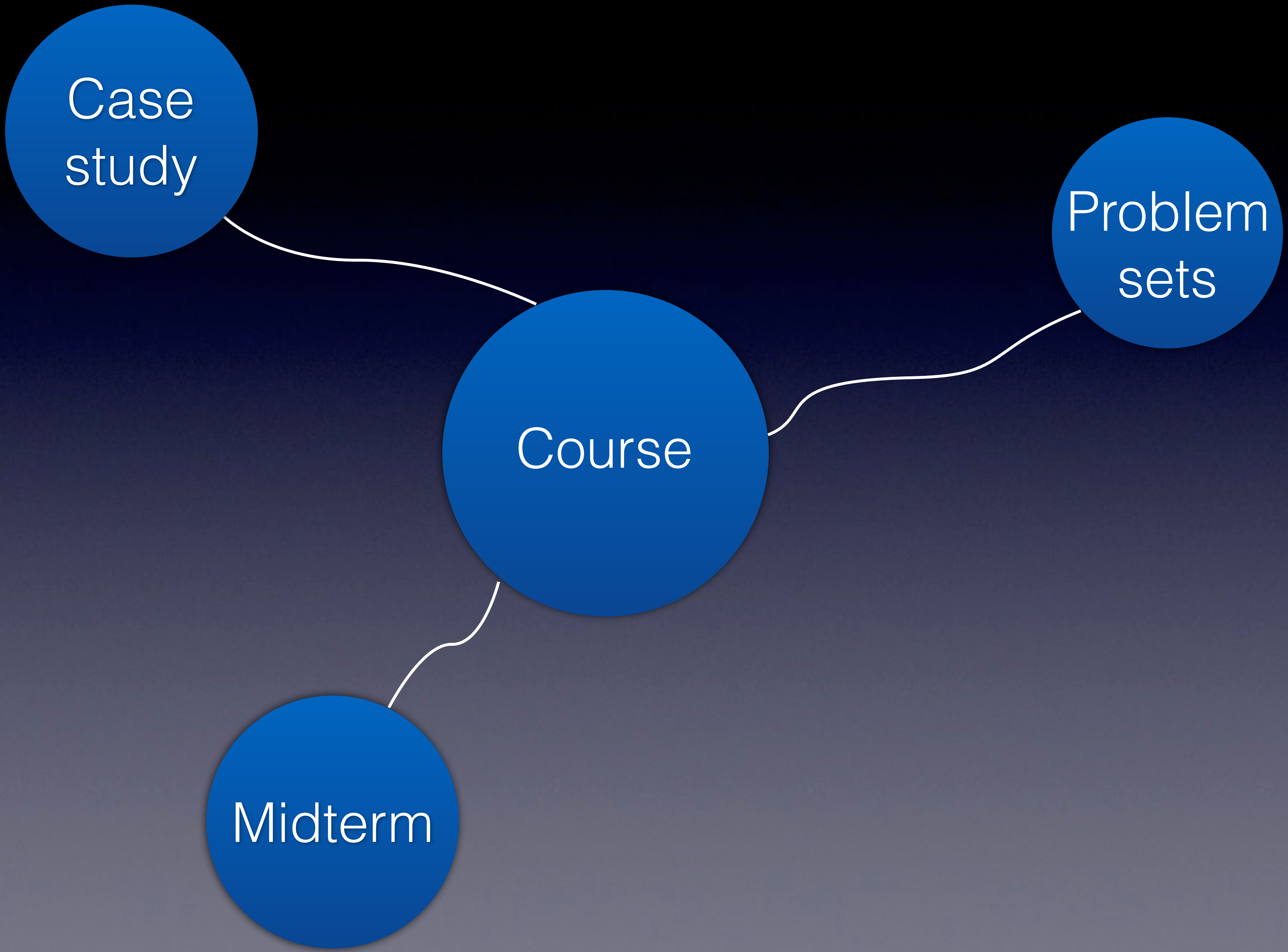
Pros	**Cons**
Easy to correct/modify	Confusing at first
Log	Memory hog
Replication	
Internships/jobs	
OS independent	
Expandable	

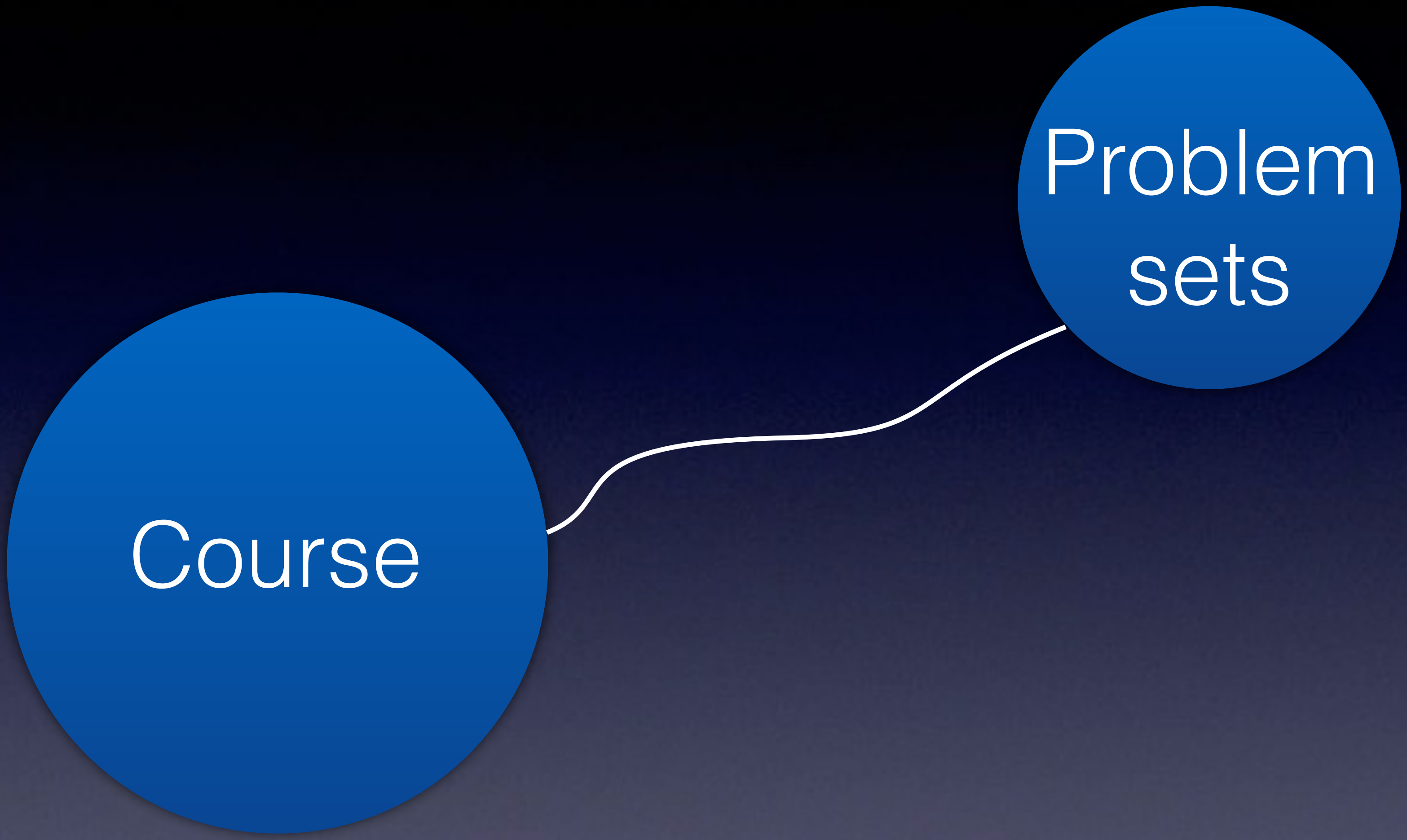
Learning R

Class demonstrations / practice

RStudio as GUI

Syllabus


Case study
Problem sets
Course
Midterm



Course

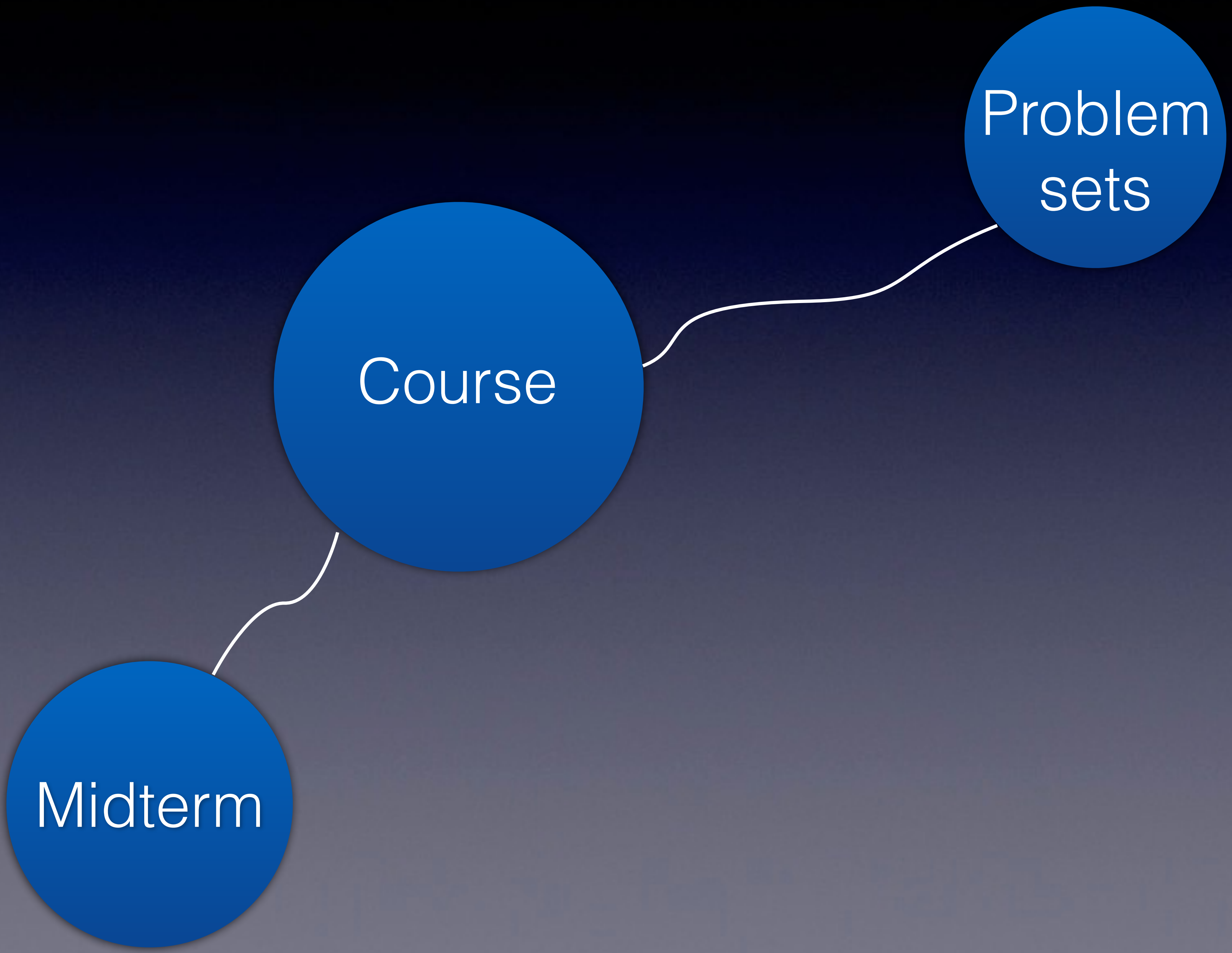
Problem sets

Problem Sets

TK% of grade

Done on Canvas

Mostly using R



Problem sets

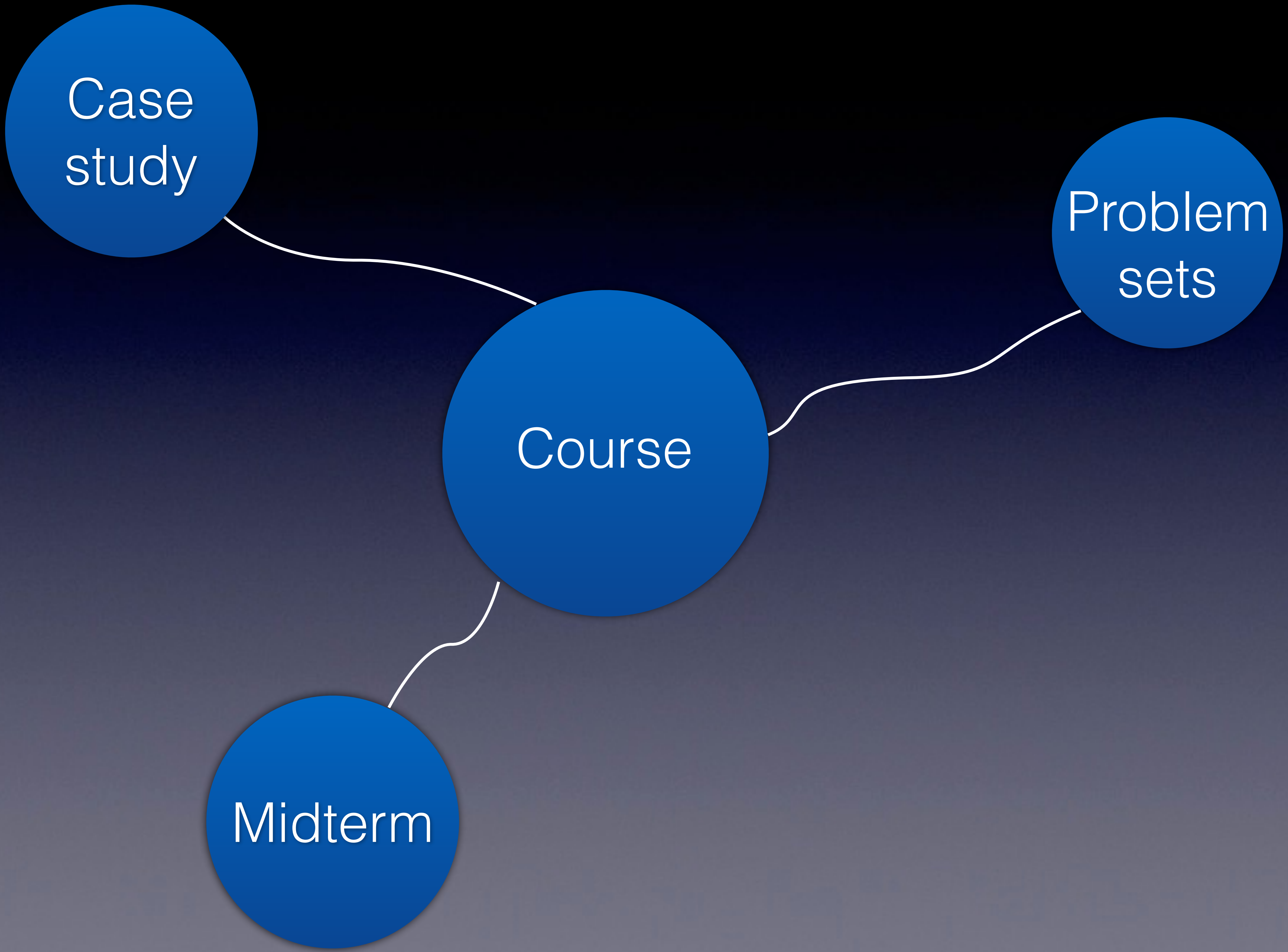
Course

Midterm

Midterm

Covers regression analysis:

TK% of grade



Case study

Problem sets

Course

Midterm

Case Study

Your chance to work through a statistical project

Group project

TK% of grade

Best Practices

Style

Code for clarity

- Use clear file and object names that are easy to read
  - snake_case (my preferred)
  - camelCase
  - period.case (easy to confuse with R commands)
- Break up lines–R does not care (much) about line breaks
- Minimize nested commands
- Indent code
- Do this–even if in a hurry!

Naming stuff - I (surprisingly hard)

Descriptive file names–even if long

✓

```
01_load_sales_data.R

02_eda_sales_data.R
```

✗

```
data.R

analysis.R
```

Naming stuff - II
(surprisingly hard)

✓

`campaign_may` (dataframe ~ noun)

`get_elasticity` (function ~ verb)

`campaign_exposure` (variable ~ noun)

RStudio autosuggests matches, no reason to type the whole thing every time

Syntax

- Spacing – makesiteasytoread
- Place spaces around all operators (`=, +, -, /, <-, <=`)
- Put a space after comma
  - Good: `double_mean <- mean(initial_data * 2, na.rm = TRUE)`
  - Bad: `double_mean<-mean(initial_data*2,na.rm=TRUE)`
- Extra spacing is ok to improve alignment

Syntax

- Use <- for assignment (not = because that is used for setting function attributes)

- Indent within curly brackets and have last curly bracket on own line:

```
If (y == 0) {
       log(x)
    } else {
       y ^ x
    }
```

Documentation and versioning

Always Write so Others Can Use/Understand Your Code!

- This includes your future self!
- Include lots of comments–even if it is obvious to you now
  - Good: `# Reduce all columns to mean to create a bar graph ----`
  - Bad: #column means

More on Comments

- But, avoid obvious comments like

  ```
  // if country code is US
  if (country_code == 'US') {…
  ```

- Group code and use a comment to describe what is going on in the group
- Comments for "why"

Separate into Sections

- Use: `# Heading ----`
  - Easier to read
  - RStudio makes it easy to jump around
  - Keyboard short-cut if "----" included

Why Versioning?

- Project history
  - You can roll back mistakes easily
  - Commit messages serve as documentation
- Branching: try new stuff without breaking the old
- Sync between your computers
- Sharing / collaboration

How?

- My favorite is Git on GitHub.com
- Free educational account
- RStudio plays nicely with GitHub

Project / File Organization

Be Consistent!

My set-up for **EVERY** research project:

```
wage_elasticity
  |— code
  |— data
  |— figures
  |— paper
  |— presentations
  |— raw_data
  |— tables
  |— read_me.md
  |— wage_elasticity.Rproj
```

Split!

- More files, rather than one big
- Option: number the files

```
code
   |— 01_load_data.R
   |— 02_regress.R
   |— 03_elasticities.R
   |— 04_experience.R
   |— functions.R
```

RStudio Project

- Use this for every data analysis project
- Keep everything there
- Always relative paths (project sets home dir)
- Remember: use "/" for separating path components
  - Example `~/projects/socs/wage_elasticity/`

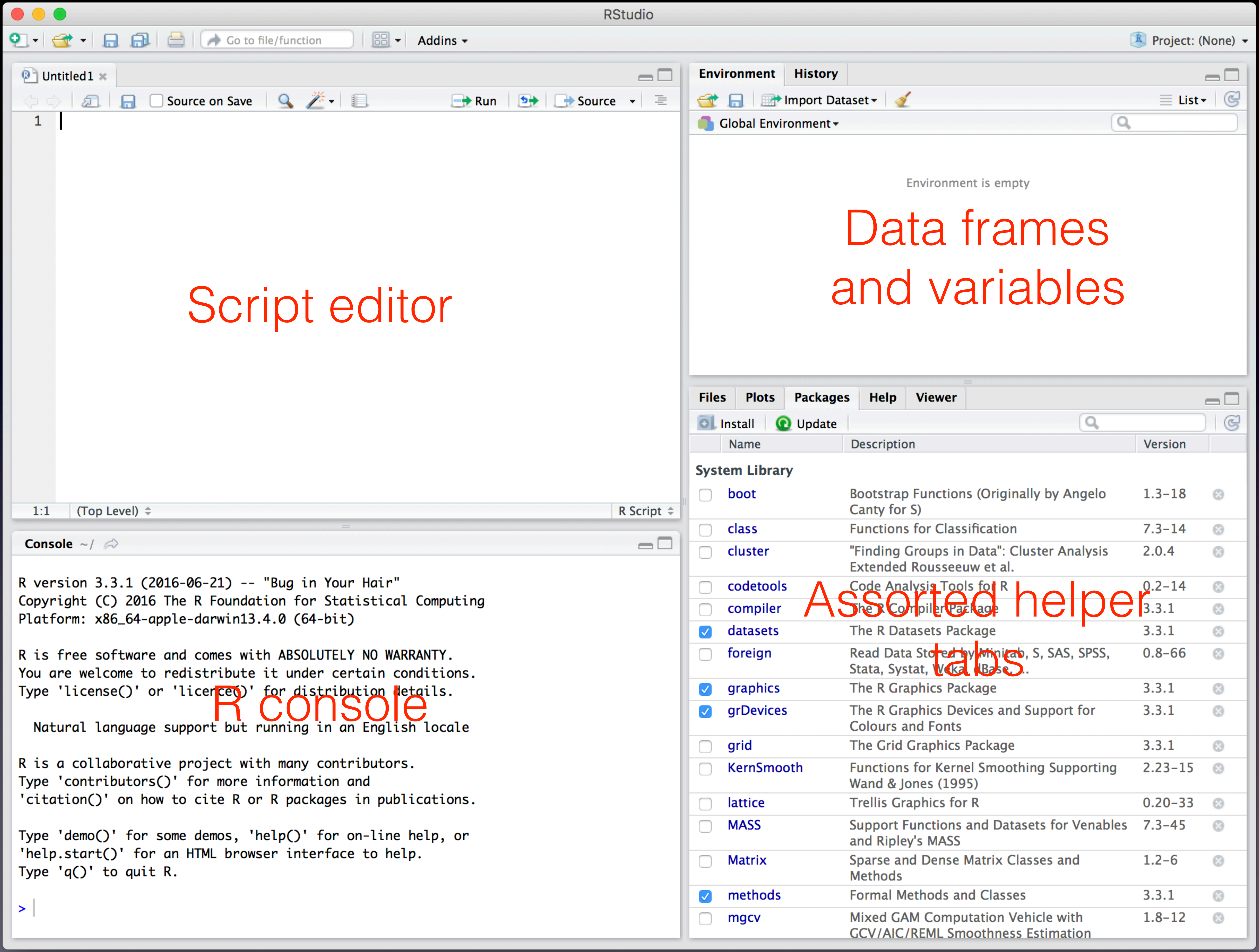
Intro R / RStudio

How to start R Studio

Find the R Studio application and double-click

Demonstration using desktop.seattleu.edu

Use "Seattle University Virtual Desktop"



Script editor

Data frames and variables

R console

Assorted helper tabs

```
R version 3.3.1 (2016-06-21) -- "Bug in Your Hair"
Copyright (C) 2016 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin13.4.0 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> 
```

System Library

	Name	Description	Version
□	boot	Bootstrap Functions (Originally by Angelo Canty for S)	1.3-18
□	class	Functions for Classification	7.3-14
□	cluster	"Finding Groups in Data": Cluster Analysis Extended Rousseeuw et al.	2.0.4
□	codetools	Code Analysis Tools for R	0.2-14
□	compiler	The R Compiler Package	3.3.1
☑	datasets	The R Datasets Package	3.3.1
□	foreign	Read Data Stored by Minitab, S, SAS, SPSS, Stata, Systat, Weka, dBase, ...	0.8-66
☑	graphics	The R Graphics Package	3.3.1
☑	grDevices	The R Graphics Devices and Support for Colours and Fonts	3.3.1
□	grid	The Grid Graphics Package	3.3.1
□	KernSmooth	Functions for Kernel Smoothing Supporting Wand & Jones (1995)	2.23-15
□	lattice	Trellis Graphics for R	0.20-33
□	MASS	Support Functions and Datasets for Venables and Ripley's MASS	7.3-45
□	Matrix	Sparse and Dense Matrix Classes and Methods	1.2-6
☑	methods	Formal Methods and Classes	3.3.1
□	mgcv	Mixed GAM Computation Vehicle with GCV/AIC/REML Smoothness Estimation	1.8-12

Script editor

- Where you write your program
- Make sure you add comments
- Others should be able to run it and understand what the program does and why

Data frames

- All data frames and variables show up here
- You can see variables within frames here

R console

- Actual R
- Great place to try stuff before adding to script
- File manipulation

Assorted helper

- Packages:
  - A major advantage of R is its extensibility
  - You can install packages here, update them, and make them active
- Figures show here as well
- The place to go for help files

Where is my data?

Create a directory for the class:
econ_5100

Download data there and keep R files there

If using desktop.seattleu.edu:
save econ_5100 under P drive

How to get data into R

Method 1

Set your working directory. For example,

```
setwd("~/econ_5100")
```

use file tab in RStudio–look under "More" or use project

```
Alumni <- read.csv("Alumni.csv")
```

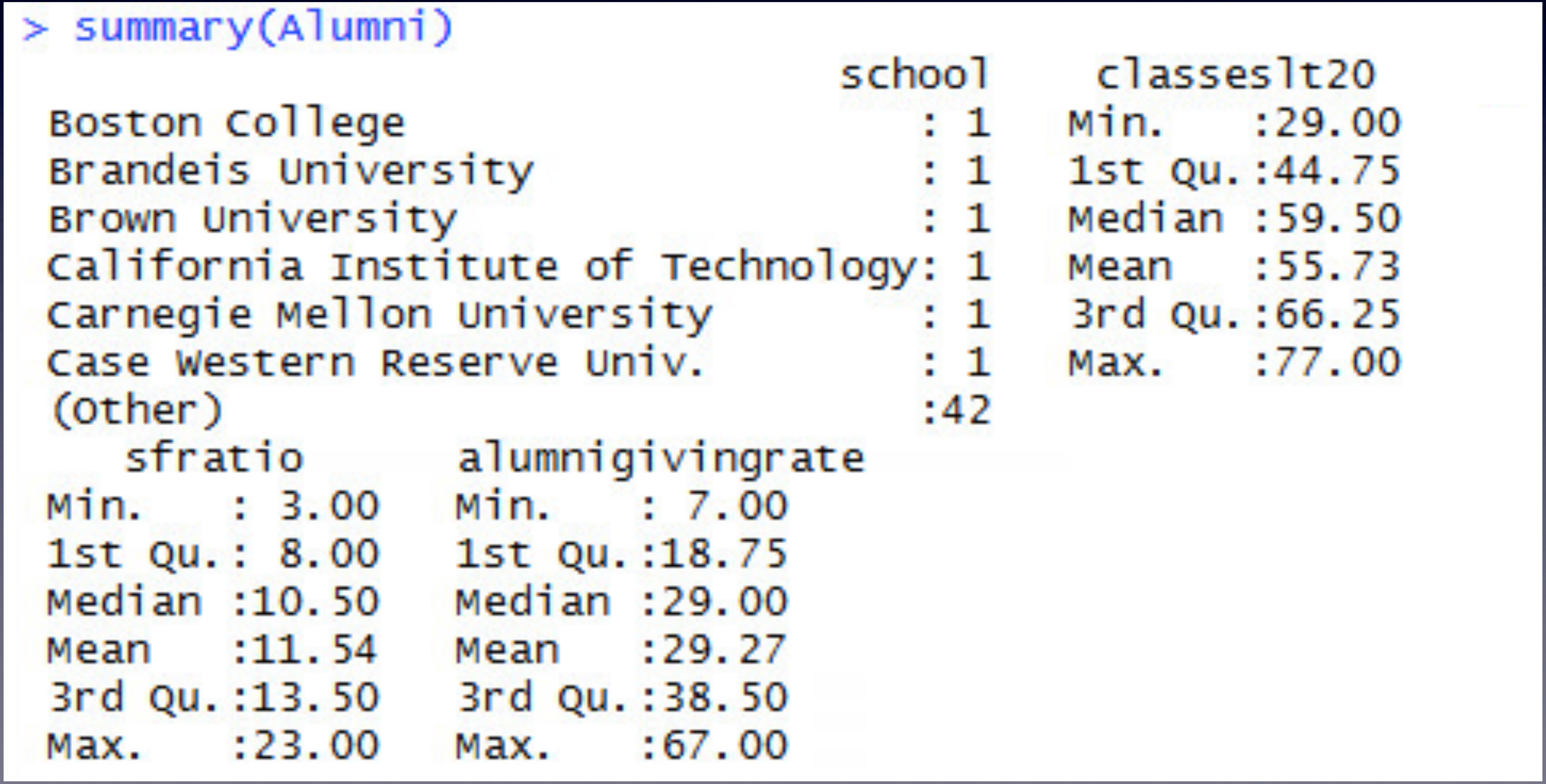
Method 2 (bad)

Write full file name (first Mac, second Desktop):

```
Alumni <- read.csv(“~/econ_5100/Alumni.csv”)
Alumni <- read.csv(“P:/econ_5100/Alumni.csv”)
```

Descriptive stats

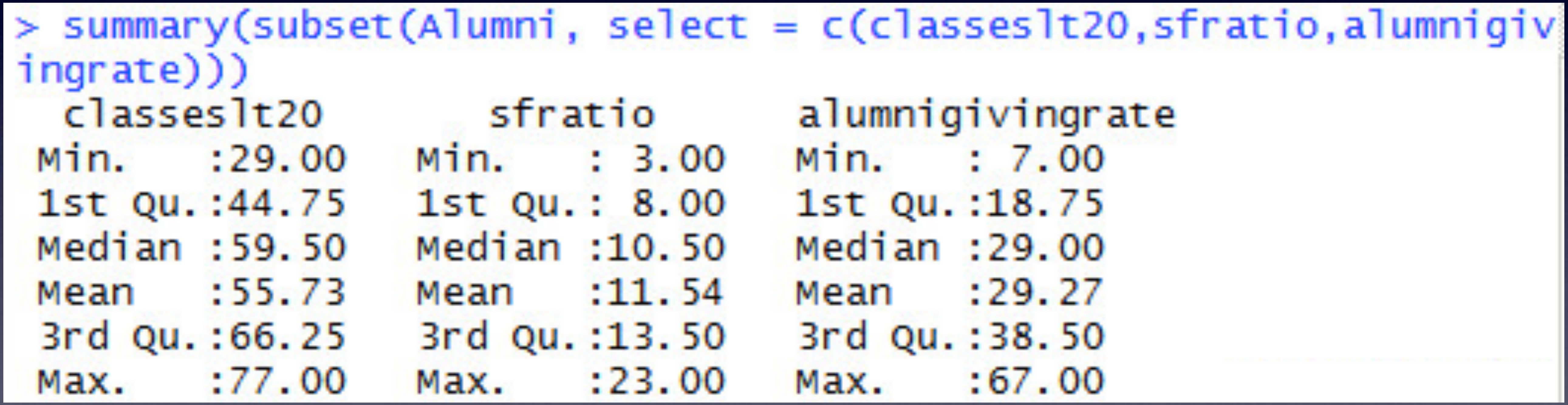
Basic summary of all variables: `summary(Alumni)`



```
> summary(Alumni)
                                  school      classeslt20   
 Boston College                     : 1   Min.   :29.00  
 Brandeis University                : 1   1st Qu.:44.75  
 Brown University                   : 1   Median :59.50  
 California Institute of Technology: 1   Mean   :55.73  
 Carnegie Mellon University         : 1   3rd Qu.:66.25  
 Case Western Reserve Univ.         : 1   Max.   :77.00  
 (Other)                            :42                  
    sfratio      alumnigivingrate
 Min.   : 3.00   Min.   : 7.00   
 1st Qu.: 8.00   1st Qu.:18.75   
 Median :10.50   Median :29.00   
 Mean   :11.54   Mean   :29.27   
 3rd Qu.:13.50   3rd Qu.:38.50   
 Max.   :23.00   Max.   :67.00   
```

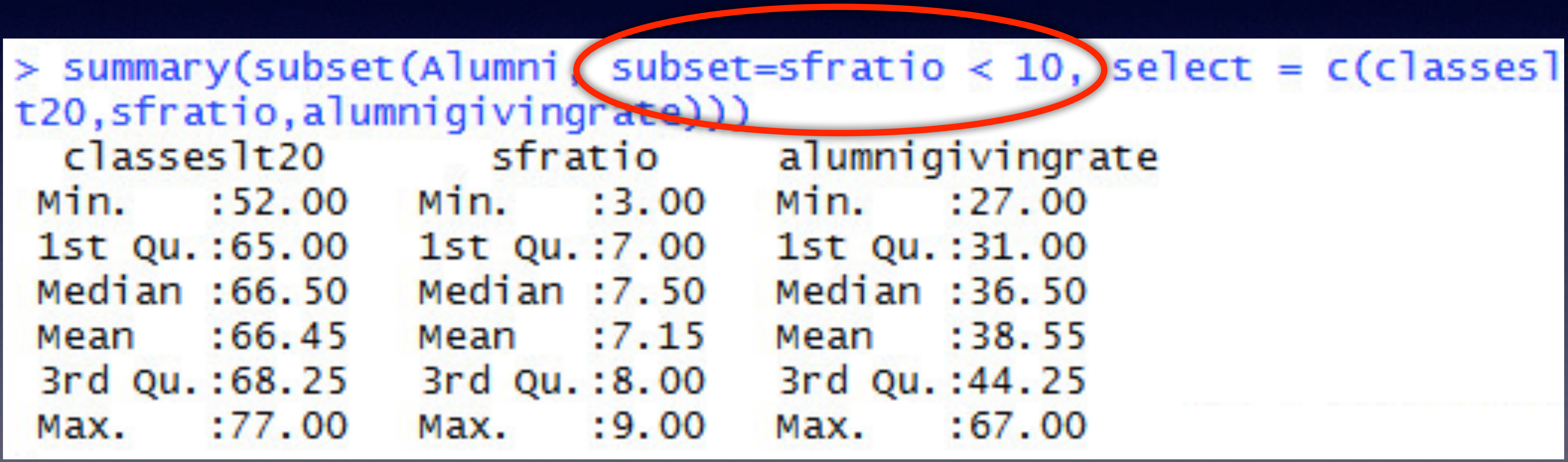
Only some vars?

“subset” is one option



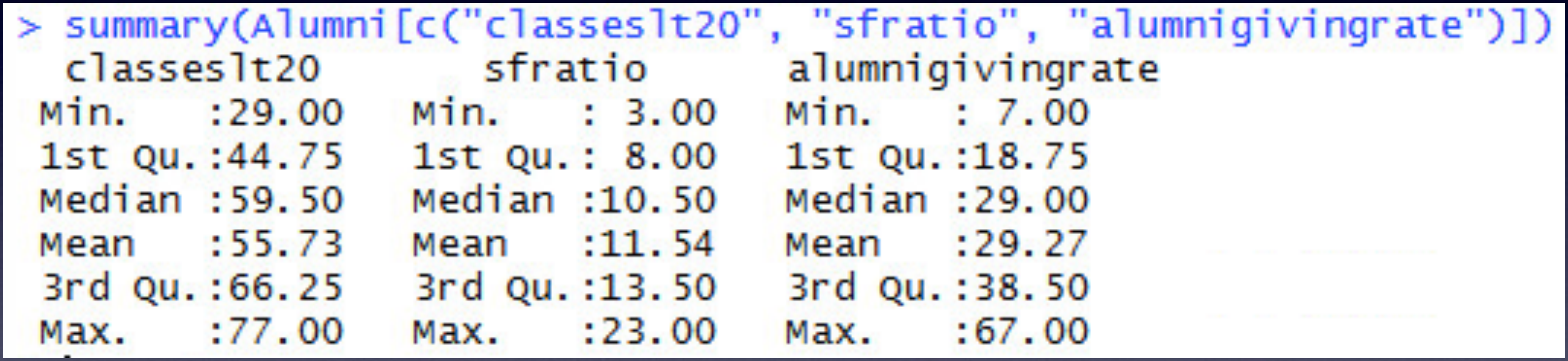
```
> summary(subset(Alumni, select = c(classeslt20,sfratio,alumnigiv
ingrate)))
  classeslt20        sfratio       alumnigivingrate
 Min.   :29.00   Min.   : 3.00   Min.   : 7.00   
 1st Qu.:44.75   1st Qu.: 8.00   1st Qu.:18.75   
 Median :59.50   Median :10.50   Median :29.00   
 Mean   :55.73   Mean   :11.54   Mean   :29.27   
 3rd Qu.:66.25   3rd Qu.:13.50   3rd Qu.:38.50   
 Max.   :77.00   Max.   :23.00   Max.   :67.00   
```

Can also do conditions



```
> summary(subset(Alumni, subset=sfratio < 10, select = c(classesl
t20,sfratio,alumnigivingrate)))
  classeslt20        sfratio      alumnigivingrate
 Min.    :52.00   Min.    :3.00   Min.    :27.00  
 1st Qu.:65.00   1st Qu.:7.00   1st Qu.:31.00  
 Median :66.50   Median :7.50   Median :36.50  
 Mean   :66.45   Mean   :7.15   Mean   :38.55  
 3rd Qu.:68.25   3rd Qu.:8.00   3rd Qu.:44.25  
 Max.    :77.00   Max.    :9.00   Max.    :67.00  
```

An alternative



```
> summary(Alumni[c("classeslt20", "sfratio", "alumnigivingrate")])
  classeslt20        sfratio       alumnigivingrate
 Min.   :29.00   Min.   : 3.00   Min.   : 7.00   
 1st Qu.:44.75   1st Qu.: 8.00   1st Qu.:18.75   
 Median :59.50   Median :10.50   Median :29.00   
 Mean   :55.73   Mean   :11.54   Mean   :29.27   
 3rd Qu.:66.25   3rd Qu.:13.50   3rd Qu.:38.50   
 Max.   :77.00   Max.   :23.00   Max.   :67.00   
```

Use `names(Alumni)` or click on the data frame to see variable names

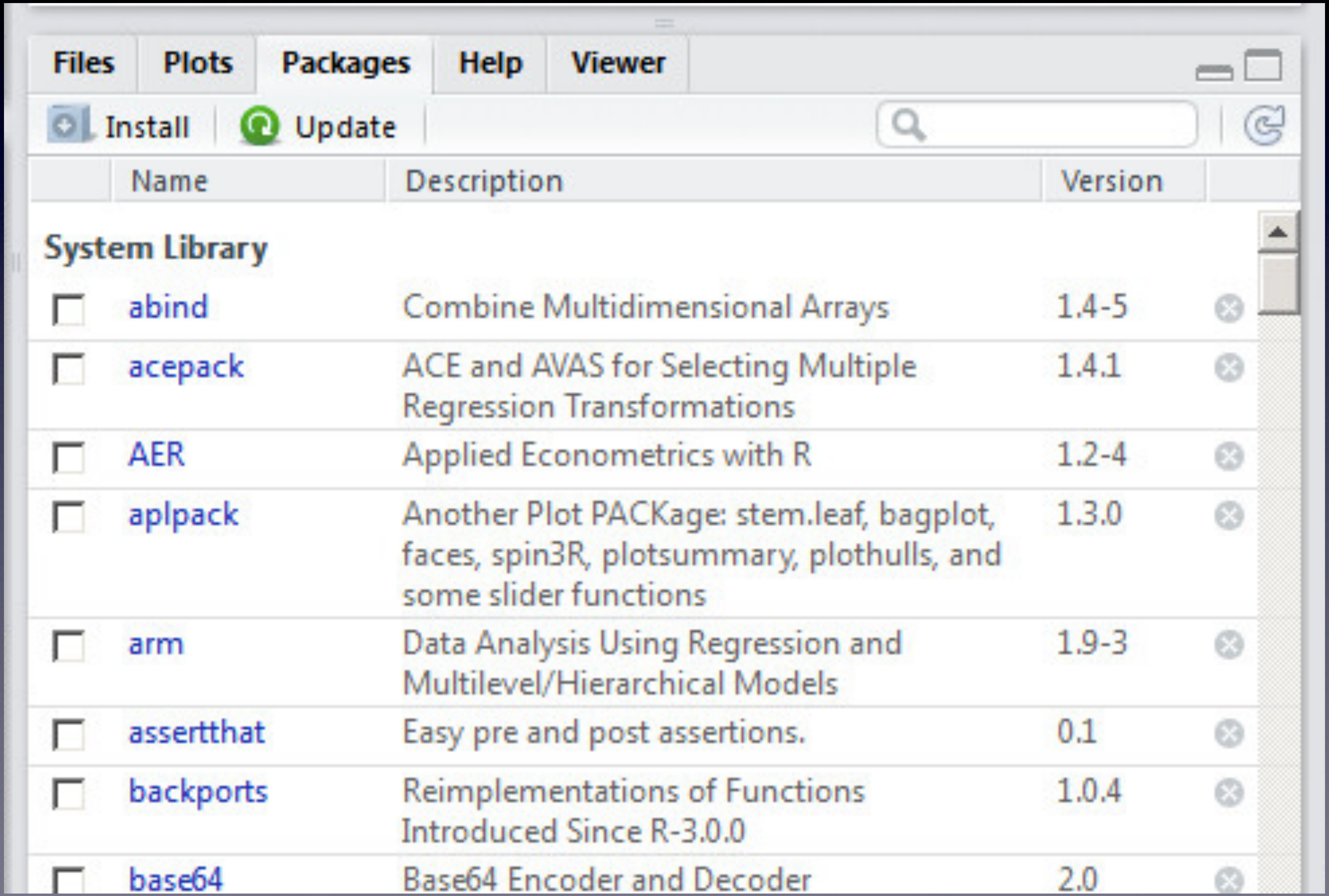
Descriptive analysis

```
summary(myData)
```

- Mean: `mean(myData$myVar)`
- Standard deviation: `sd(myData$myVar)`
- Minimum: `min(myData$myVar)`
- Maximum: `max(myData$myVar)`
- Median: `median(myData$myVar)`

Either write these in console or script

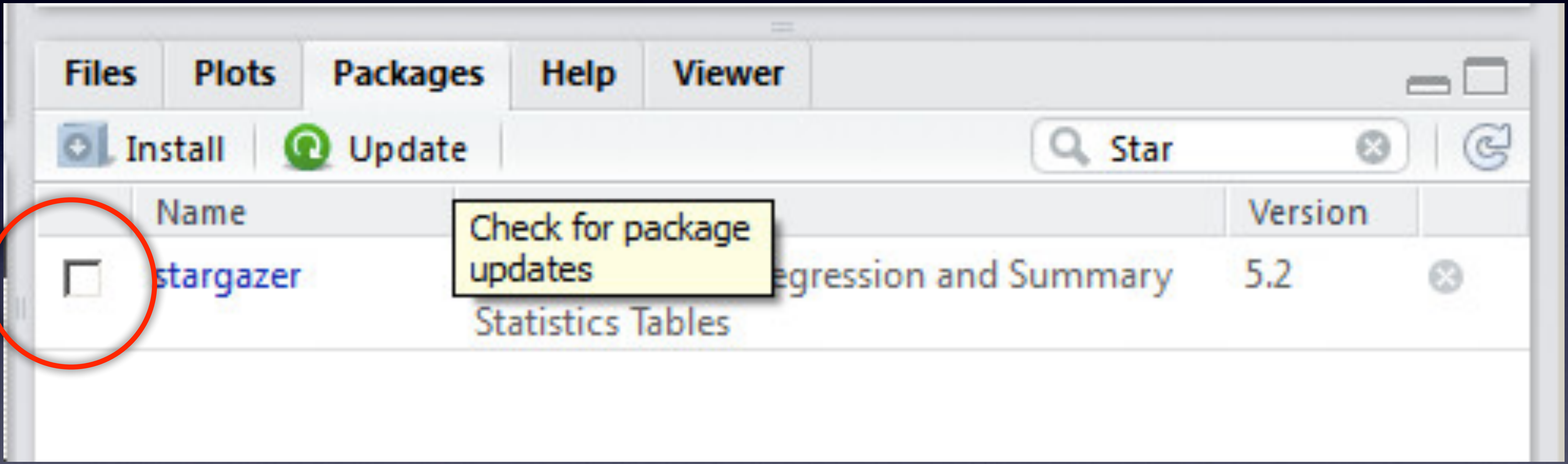
Packages in R



System Library

	Name	Description	Version
□	abind	Combine Multidimensional Arrays	1.4-5
□	acepack	ACE and AVAS for Selecting Multiple Regression Transformations	1.4.1
□	AER	Applied Econometrics with R	1.2-4
□	aplpack	Another Plot PACKage: stem.leaf, bagplot, faces, spin3R, plotsummary, plothulls, and some slider functions	1.3.0
□	arm	Data Analysis Using Regression and Multilevel/Hierarchical Models	1.9-3
□	assertthat	Easy pre and post assertions.	0.1
□	backports	Reimplementations of Functions Introduced Since R-3.0.0	1.0.4
□	base64	Base64 Encoder and Decoder	2.0

Stargazer



Once found, click checkbox or add to script as

```
library(stargazer)
```

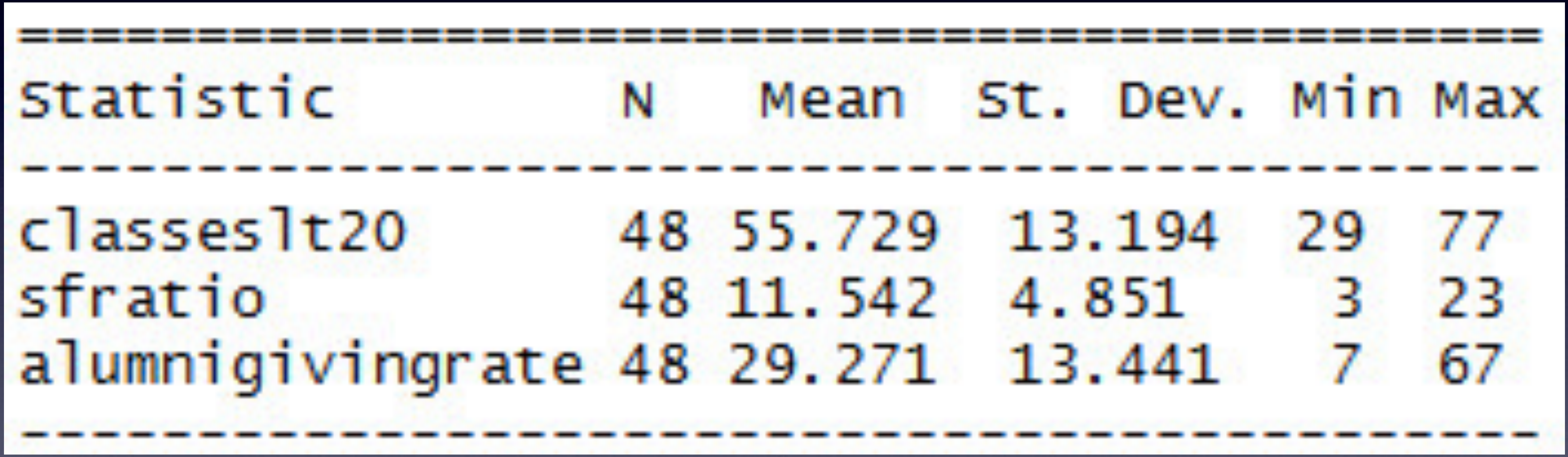
Making things pretty

Many options, but “stargazer” is easy

Simple first step–prints to console

```
stargazer( # in text format and with var labels
    Alumni[ c("classeslt20", "sfratio", "alumnigivingrate")],
    type = "text"
   )
```

First try

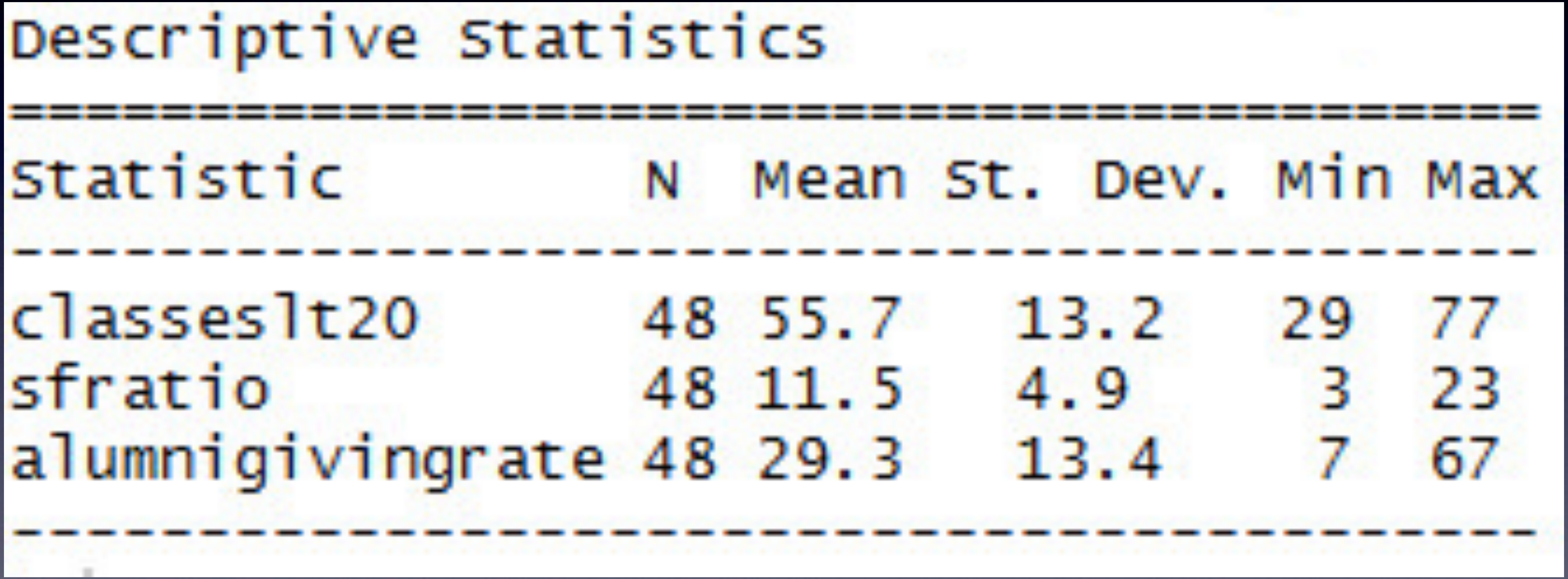


```
===========================================
Statistic        N   Mean  St. Dev. Min Max
-------------------------------------------
classeslt20      48 55.729  13.194  29  77
sfratio          48 11.542  4.851    3  23
alumnigivingrate 48 29.271  13.441   7  67
-------------------------------------------
```

Making it nicer

```
stargazer( # in text format and with var labels
    Alumni[ c("classeslt20", "sfratio", "alumnigivingrate")],
    type = “text",
    title = "Descriptive statistics",
    digits = 1 # number of digits after the point
   )
```

Second try

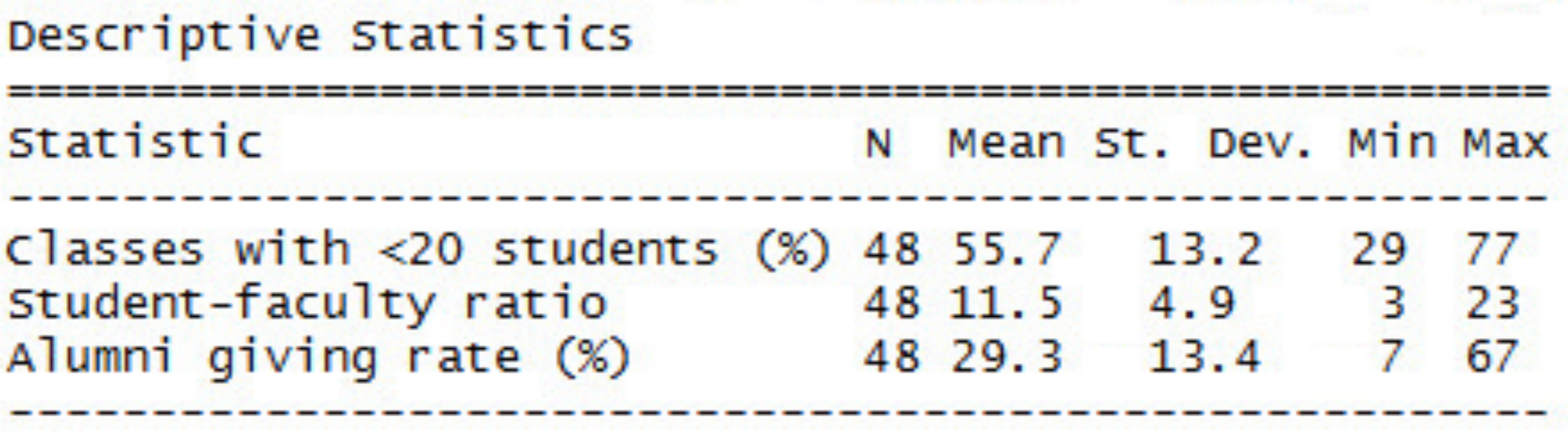


```
Descriptive Statistics
=============================================
Statistic        N  Mean St. Dev. Min Max
---------------------------------------------
classeslt20      48 55.7   13.2    29  77
sfratio          48 11.5   4.9      3  23
alumnigivingrate 48 29.3   13.4     7  67
---------------------------------------------
```

Adding labels

```
stargazer( # in text format and with var labels
    Alumni[ c("classeslt20", "sfratio", "alumnigivingrate")],
    type = “text",
    title = "Descriptive statistics",
    digits = 1 # number of digits after the point,
    covariate.labels = c(“Classes with <20 students (%)”,
     “Student-faculty ratio”, “Alumni giving rate (%)”)
   )
```

Third try



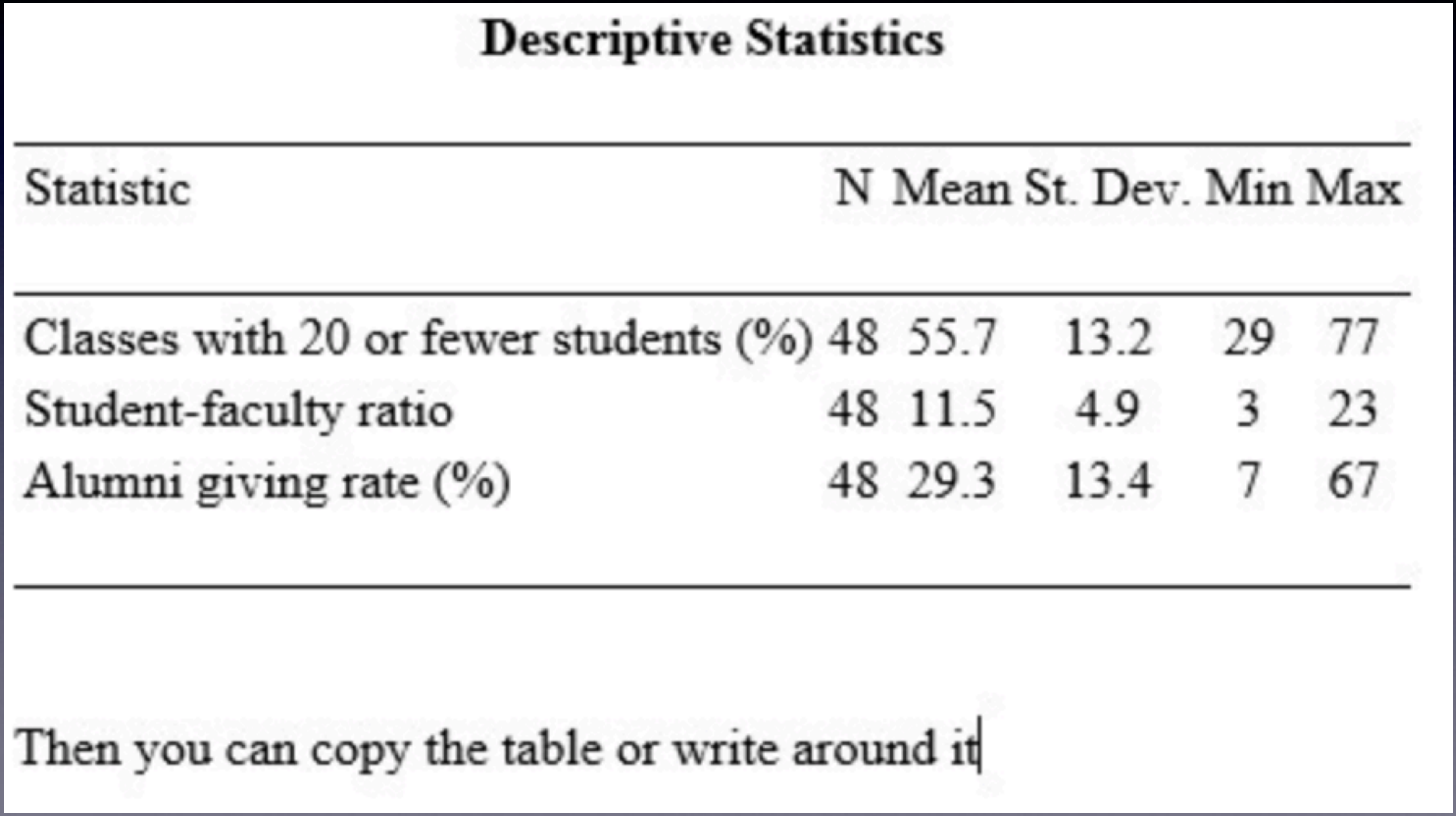
```
Descriptive Statistics
======================================================
Statistic                     N  Mean St. Dev. Min Max
------------------------------------------------------
Classes with <20 students (%) 48 55.7   13.2   29  77
Student-faculty ratio         48 11.5   4.9     3  23
Alumni giving rate (%)        48 29.3   13.4    7  67
------------------------------------------------------
```

HTML to Word

```
stargazer( # in text format and with var labels
    Alumni[ c("classeslt20", "sfratio", "alumnigivingrate")],
    type = “html",
    title = "Descriptive statistics",
    digits = 1 # number of digits after the point,
    covariate.labels = c(“Classes with 20 or fewer students (%)”,
      “Student-faculty ratio”, “Alumni giving rate (%)”),
      out = “desStat.htm” # saved to your working directory
   )
```

Use “Open file” in Word and edit as needed
Save as Word document

Pretty!!



Descriptive Statistics

Statistic	N	Mean	St. Dev.	Min	Max
Classes with 20 or fewer students (%)	48	55.7	13.2	29	77
Student-faculty ratio	48	11.5	4.9	3	23
Alumni giving rate (%)	48	29.3	13.4	7	67

Then you can copy the table or write around it

For Next Monday

- Read Chapters 8 through 12 in Keller (on Canvas)
- Install R / RStudio on your own computer (instructions on Canvas)
- Work through "R for Data Science", part I (Explore)