STOR 390: Introduction to Data Science

Spring 2017

Iain Carmichael (iain@unc.edu)

THE UNIVERSITY
of NORTH CAROLINA
at CHAPEL HILL

“All models are wrong, but some models are useful”

– George Box

“All models are wrong, but some models are useful”

– George Box

Model

A simplified description, especially a mathematical one, of a system or process, to assist calculations and predictions - New Oxford American Dictionary

Model

A simplified description, especially a mathematical one, of a system or process, to assist calculations and predictions - New Oxford American Dictionary

or

“an abstract representation of some process, be it a baseball game, an oil company’s supply chain, a foreign government’s actions or a movie theater’s attendance” - Weapons of Math Destruction

Newton’s three laws of motion are a simple model of the universe

$F = ma$

Newton’s three laws of motion are a simple model of the universe

F = ma

Special/General relativity

Newton’s three laws of motion are a simple model of the universe

F = ma

Special/General relativity

Vast majority of physics applications use Newtonian mechanics

Some people are introverts, some people are extroverts

Places people into two categories (or maybe on a continuum)

https://www.ted.com/talks/susan_cain_the_power_of_introverts

Some people are introverts, some people are extroverts

Places people into two categories (or maybe on a continuum)

Fails to capture a lot about you

https://www.ted.com/talks/susan_cain_the_power_of_introverts

Some people are introverts, some people are extroverts

Places people into two categories (or maybe on a continuum)

Fails to capture a lot about you

Helpful for understanding how people operate

https://www.ted.com/talks/susan_cain_the_power_of_introverts

Relationship advice…

“Absence makes the heart grow fonder”

Relationship advice…

“Absence makes the heart grow fonder”

or is it

“Out of sight, out of mind”

Relationship advice…

“Absence makes the heart grow fonder”

or is it

“Out of sight, out of mind”

Self driving cars use a lot of models

Where is the car on the road?

Where are other cars it going?

http://www.theverge.com/2014/5/28/5756852/googles-self-driving-car-isnt-a-car-its-the-future

Self driving cars use a lot of models

Where is the car on the road?

Where are other cars it going?

What is an object I should avoid?

http://www.theverge.com/2014/5/28/5756852/googles-self-driving-car-isnt-a-car-its-the-future

Self driving cars use a lot of models

Where is the car on the road?

Where are other cars it going?

What is an object I should avoid?

Is that a stop sign?

http://www.theverge.com/2014/5/28/5756852/googles-self-driving-car-isnt-a-car-its-the-future

What is data science?

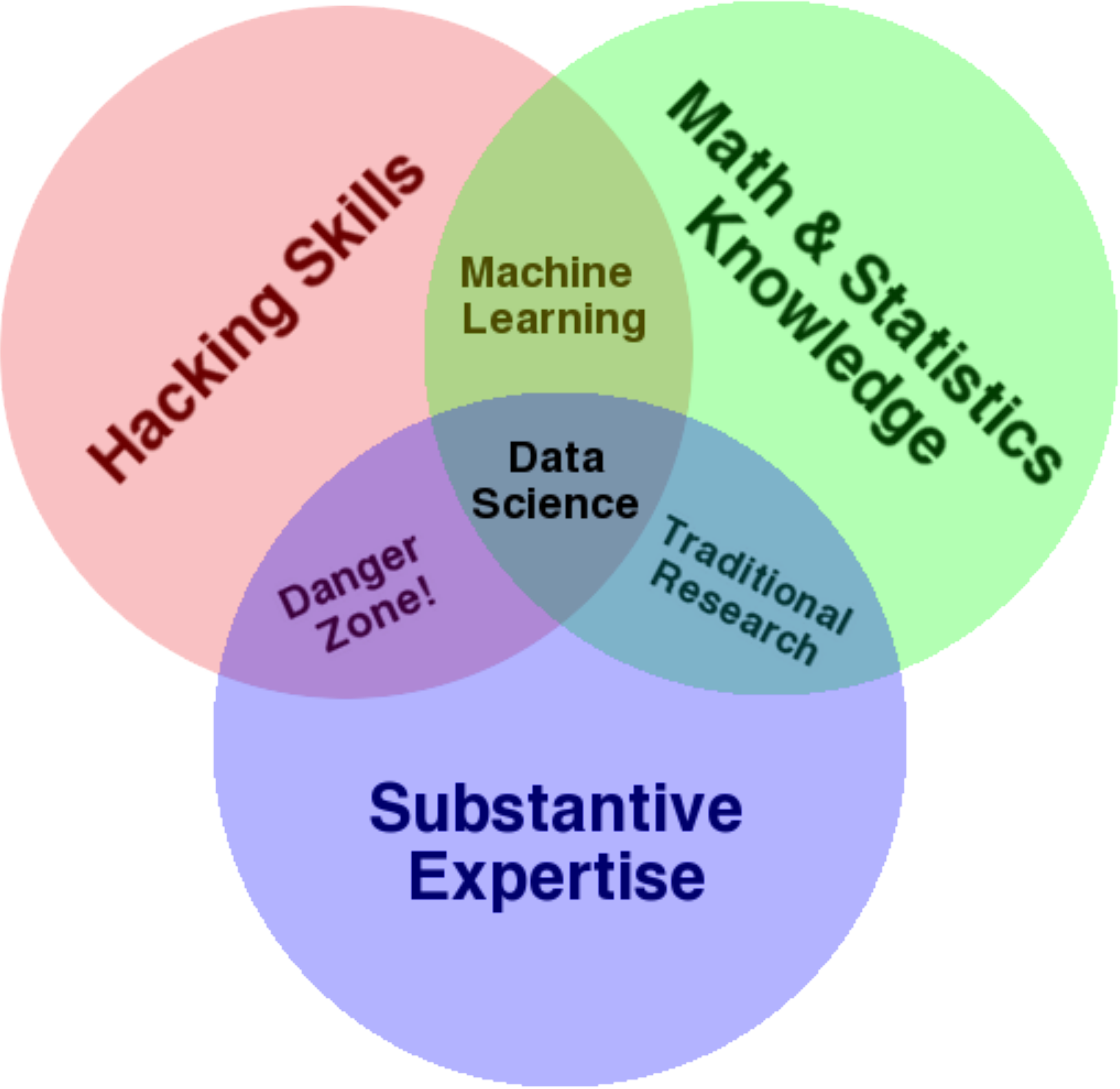
Brian Caffo, Jeff Leek, Roger Peng

@jtleek

www.jtleek.com

https://www.coursera.org/learn/data-science-course/lecture/X4Z9T/what-is-data-science

The data science Venn Diagram



http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

Lots of buzz words

https://www.linkedin.com/pulse/putting-science-back-data-paul-dalen

Big Data can mean a lot of things

Lots of observations

Big Data can mean a lot of things

Lots of observations

Lots of variables

Big Data can mean a lot of things

Lots of observations

Lots of variables

Non-standard data

- Text, images, networks

Big Data can mean a lot of things

Lots of observations

Lots of variables

Non-standard data

- Text, images, networks

or that someone is trying to impress you…

Big data = data is ubiquitous

Neuroscience

Ecommerce

Cars

Finance

Medicine

Journalism

Big data = data is ubiquitous

Neuroscience

Ecommerce

Cars

Finance

Medicine

Journalism

Where is data absent?

Is this always a good thing?

Use data to **understand** something

What customers are interested in my product?

Who will respond to this cancer treatment?

Use data to **understand** something

What customers are interested in my product?

Who will respond to this cancer treatment?

“Classical” science, now applied to many areas

Use data to **understand** something

What customers are interested in my product?

Who will respond to this cancer treatment?

“Classical” science, now applied to many areas

New and interesting

- problems
- datasets
- algorithms

Use data to **do** something

Facebook can do facial recognition

Write an algorithm to beat the stock market

Program a computer to beat humans at Go

Use data to **do** something

Facebook can do facial recognition

Write an algorithm to beat the stock market

Program a computer to beat humans at Go

More like “engineering”

Use data to **do** something

Facebook can do facial recognition

Write an algorithm to beat the stock market

Program a computer to beat humans at Go

More like “engineering”

Same algorithms, different goals

Course Information

STOR 390: Introduction to Data Science

- TuTh: 5:00 - 6:15 pm
- Greenlaw 101

Instructor: Iain Carmichael (iain@unc.edu)

Teaching Assistants:

- Varun Goel (varung@live.unc.edu)
- Brendan Brown (bb@live.unc.edu)

Website

https://idc9.github.io/stor390/

Iain Carmichael

BA in Math and Physics from Cornell

PhD candidate in Statistics

Gamalon Machine Intelligence

Research

- networks, probability and high-dimensional statistics
- neuroscience and law

Brendan Brown

PhD student in statistics

2+ years experience in data science for the UNC system office

- visualization
- presentation
- forecasting, modeling, with large datasets

Varun Goel

PhD candidate in Geography

Data Scientist at Indian School of Business, Hyderabad - Involved in informing agricultural public policy through data science

Current Research

Spatial Statistics, GIS, Disease ecology,

Population Health

Waitlist…

The waitlist is very long

Sign up at: https://stat-or.unc.edu/waitlist/

I do not control the waitlist

Course organization

Homework: 35%

- ~ 4 data analyses

Labs: 35%

- Start in class, due the next class

Class participation: 15%

Final project: 15%

Extra Credit: up to 5%

Group work for homework and final project

The instructor will assign teams

Final grade will be adjusted by peer ratings

As a last resort a team may fire an uncooperative member

Final Project

Novel data analysis

- get a data set
- Analyze it
- Write a blog post

In a team

Goals and learning objective

core R programming skills

statistical and programming best practices

communication, problem solving, teamwork

literate programming

- R Markdown

Goals and learning objective

core R programming skills

statistical and programming best practices

communication, problem solving, teamwork

literate programming

- R Markdown

Goals and learning objective

core R programming skills

statistical and programming best practices

communication, problem solving, teamwork

literate programming

- R Markdown

Goals and learning objective

core R programming skills

statistical and programming best practices

communication, problem solving, teamwork

literate programming

- R Markdown

Topics (see syllabus)

Visualization with ggplot2

Data manipulation dplyr

R Markdown

Programming e.g. functions, loops, if/else, comments

Tidy Data, relational data, data import

Reproducibility

Strings/regular expressions

EDA

Classification, clustering, regression

Web scraping

Text data and Natural Language Processing

Additional topics (if time permits)

interactive graphics with shiny

effective visualization for communication

date/time data

github

GIS data

data privacy/ethics

Google is your best friend as a programmer

Lots of resources on the course website
https://idc9.github.io/stor390/course_info/references.html

stack**overflow**

http://lmsotfy.com/

Install R and R Studio

https://idc9.github.io/stor390/notes/getting_started/getting_started.html

http://rhrv.r-forge.r-project.org/

https://www.rstudio.com/about/trademark/

R vs. Python

Better to be really good an one then mediocre at both

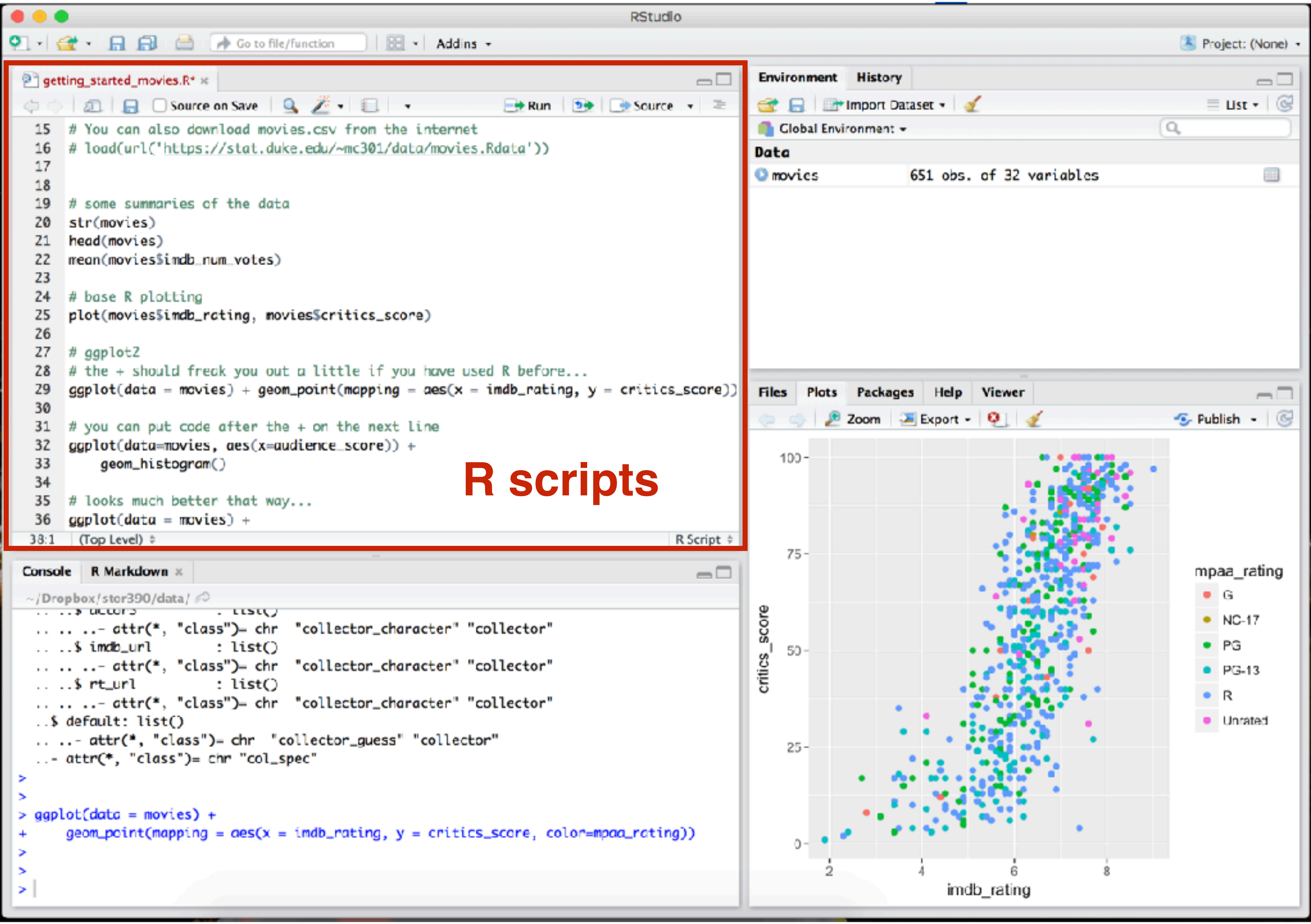
Both and pluses and minuses

Class survey: how will this information be used against you?

Please fill out this survey: https://goo.gl/forms/S0pvK0IrQRTqN2jW2

I may use major/year information to make teams

This data will not be released outside the class

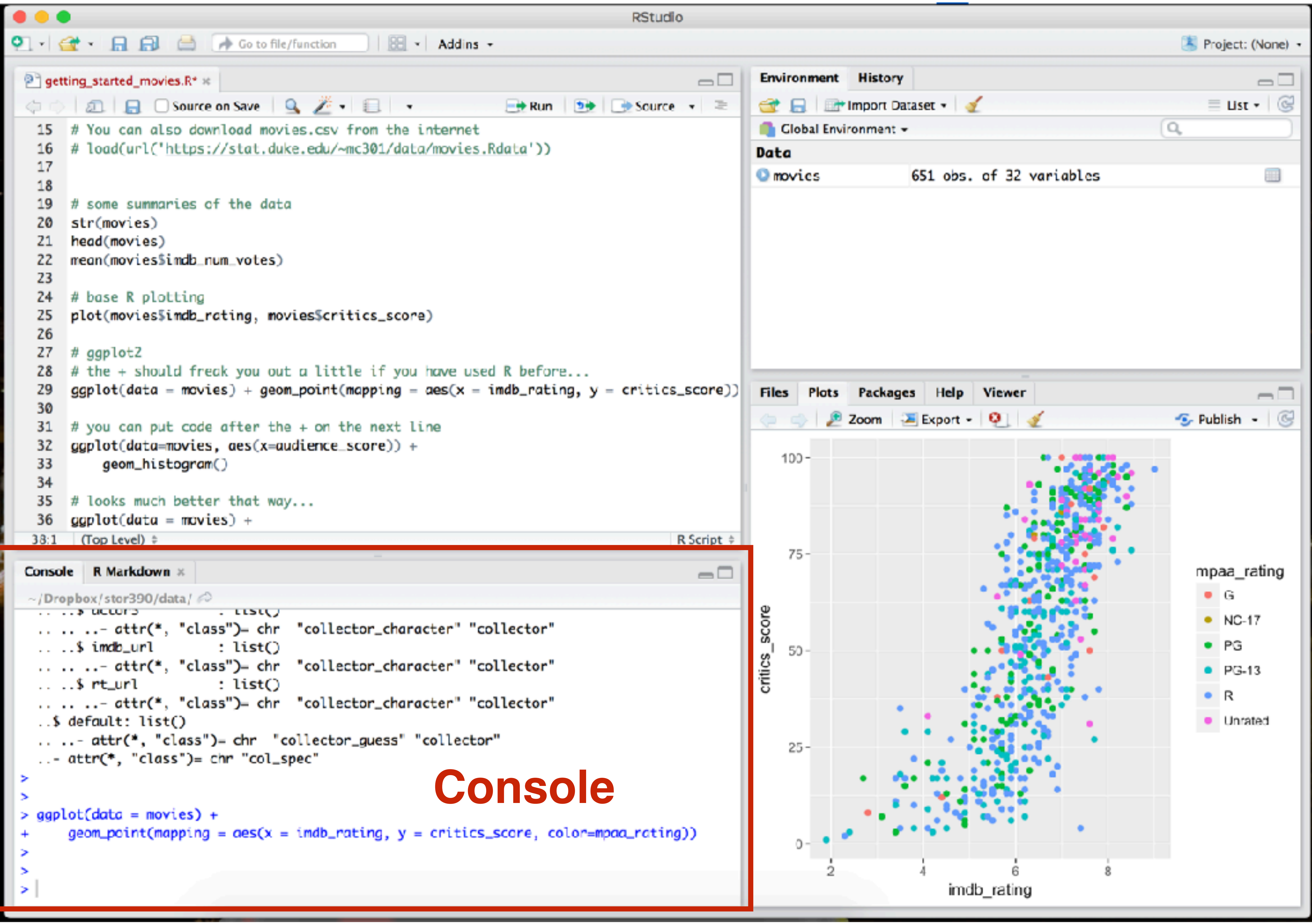


R scripts

```
# You can also download movies.csv from the internet
# load(url('https://stat.duke.edu/~mc301/data/movies.Rdata'))


# some summaries of the data
str(movies)
head(movies)
mean(movies$imdb_num_votes)

# base R plotting
plot(movies$imdb_rating, movies$critics_score)

# ggplot2
# the + should freak you out a little if you have used R before...
ggplot(data = movies) + geom_point(mapping = aes(x = imdb_rating, y = critics_score))

# you can put code after the + on the next line
ggplot(data=movies, aes(x=audience_score)) +
    geom_histogram()

# looks much better that way...
ggplot(data = movies) +
```

movies   651 obs. of 32 variables

```
  .. ..- attr(*, "class")= chr  "collector_character" "collector"
  .. ..$ imdb_url        : list()
  .. .. ..- attr(*, "class")= chr  "collector_character" "collector"
  .. ..$ rt_url          : list()
  .. .. ..- attr(*, "class")= chr  "collector_character" "collector"
  ..$ default: list()
  .. ..- attr(*, "class")= chr  "collector_guess" "collector"
  ..- attr(*, "class")= chr "col_spec"
>
>
> ggplot(data = movies) +
+     geom_point(mapping = aes(x = imdb_rating, y = critics_score, color=mpaa_rating))
>
>
>
```

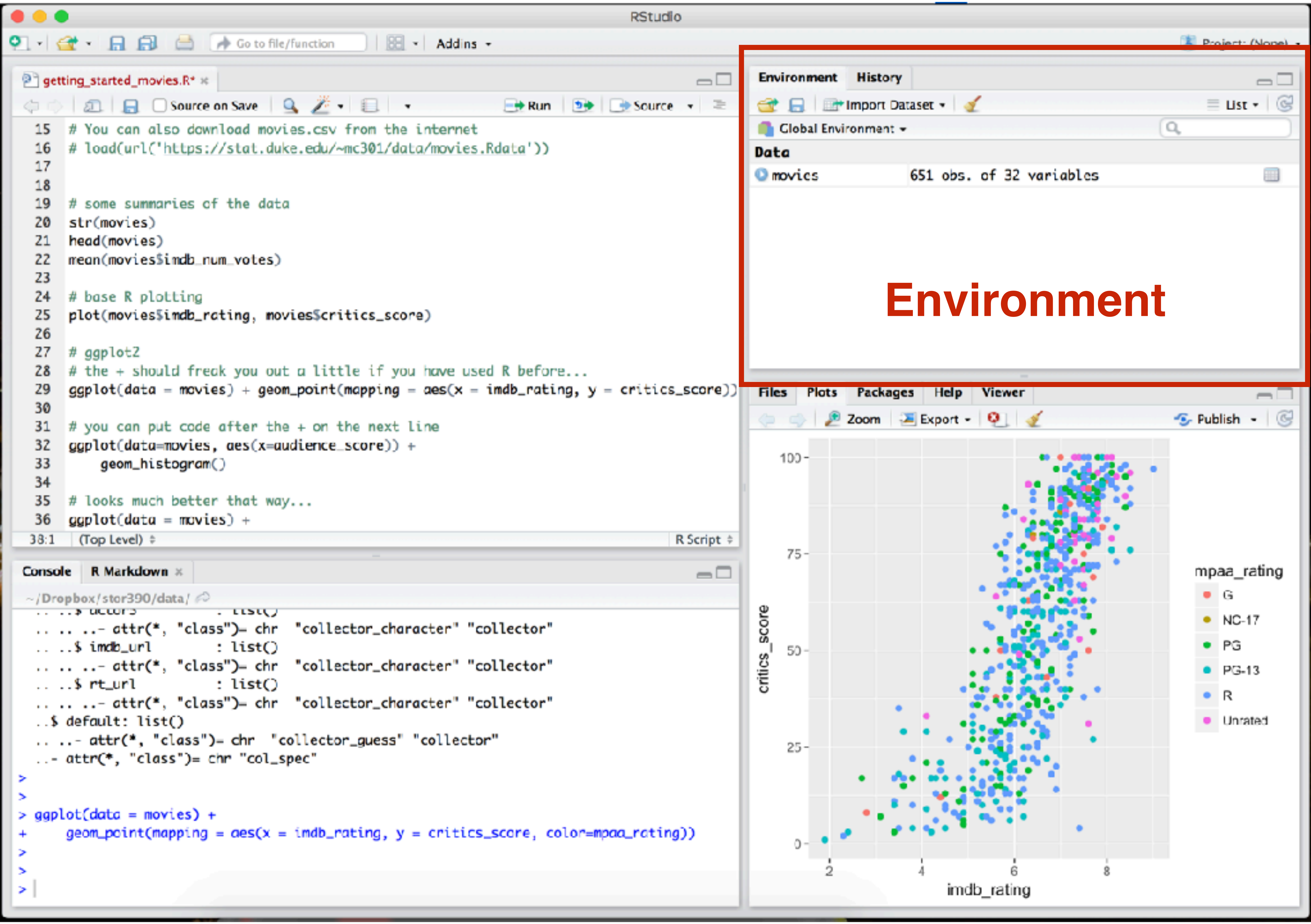


movies 651 obs. of 32 variables

```
# You can also download movies.csv from the internet
# load(url('https://stat.duke.edu/~mc301/data/movies.Rdata'))


# some summaries of the data
str(movies)
head(movies)
mean(movies$imdb_num_votes)

# base R plotting
plot(movies$imdb_rating, movies$critics_score)

# ggplot2
# the + should freak you out a little if you have used R before...
ggplot(data = movies) + geom_point(mapping = aes(x = imdb_rating, y = critics_score))

# you can put code after the + on the next line
ggplot(data=movies, aes(x=audience_score)) +
    geom_histogram()

# looks much better that way...
ggplot(data = movies) +
```

```
  .. ..$ actors           : list()
  .. .. ..- attr(*, "class")= chr  "collector_character" "collector"
  .. ..$ imdb_url         : list()
  .. .. ..- attr(*, "class")= chr  "collector_character" "collector"
  .. ..$ rt_url           : list()
  .. .. ..- attr(*, "class")= chr  "collector_character" "collector"
  ..$ default: list()
  .. ..- attr(*, "class")= chr  "collector_guess" "collector"
  ..- attr(*, "class")= chr "col_spec"
>
>
> ggplot(data = movies) +
+     geom_point(mapping = aes(x = imdb_rating, y = critics_score, color=mpaa_rating))
>
>
>
```

Console



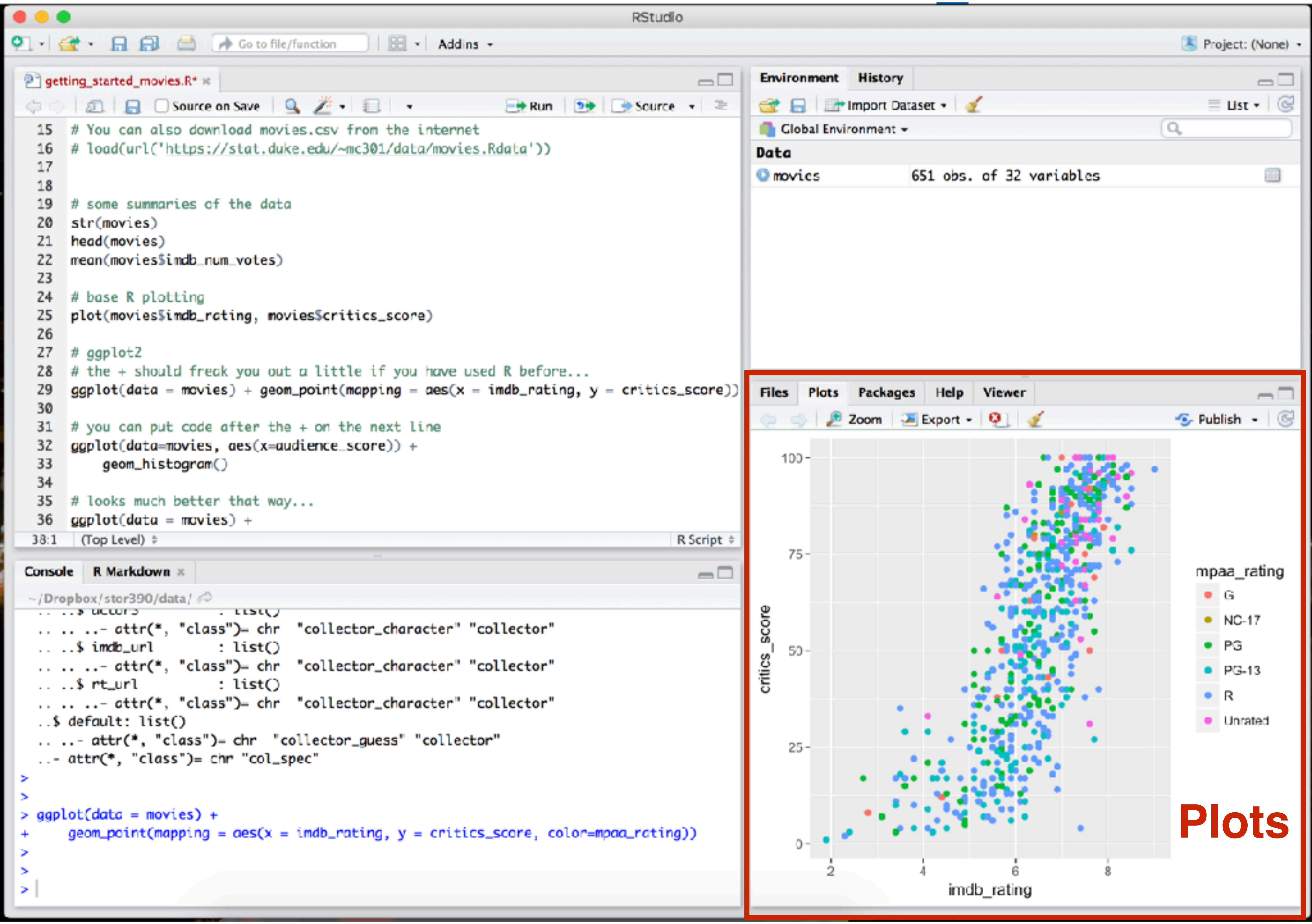
```
# You can also download movies.csv from the internet
# load(url('https://stat.duke.edu/~mc301/data/movies.Rdata'))


# some summaries of the data
str(movies)
head(movies)
mean(movies$imdb_num_votes)

# base R plotting
plot(movies$imdb_rating, movies$critics_score)

# ggplot2
# the + should freak you out a little if you have used R before...
ggplot(data = movies) + geom_point(mapping = aes(x = imdb_rating, y = critics_score))

# you can put code after the + on the next line
ggplot(data=movies, aes(x=audience_score)) +
    geom_histogram()

# looks much better that way...
ggplot(data = movies) +
```

```
  .. .. ..- attr(*, "class")= chr  "collector_character" "collector"
  .. ..$ imdb_url         : list()
  .. .. ..- attr(*, "class")= chr  "collector_character" "collector"
  .. ..$ rt_url           : list()
  .. .. ..- attr(*, "class")= chr  "collector_character" "collector"
  ..$ default: list()
  .. ..- attr(*, "class")= chr  "collector_guess" "collector"
  ..- attr(*, "class")= chr "col_spec"
>
>
> ggplot(data = movies) +
+     geom_point(mapping = aes(x = imdb_rating, y = critics_score, color=mpaa_rating))
>
>
> 
```

Data

movies	651 obs. of 32 variables

Environment



```
# You can also download movies.csv from the internet
# load(url('https://stat.duke.edu/~mc301/data/movies.Rdata'))


# some summaries of the data
str(movies)
head(movies)
mean(movies$imdb_num_votes)

# base R plotting
plot(movies$imdb_rating, movies$critics_score)

# ggplot2
# the + should freak you out a little if you have used R before...
ggplot(data = movies) + geom_point(mapping = aes(x = imdb_rating, y = critics_score))

# you can put code after the + on the next line
ggplot(data=movies, aes(x=audience_score)) +
    geom_histogram()

# looks much better that way...
ggplot(data = movies) +
```

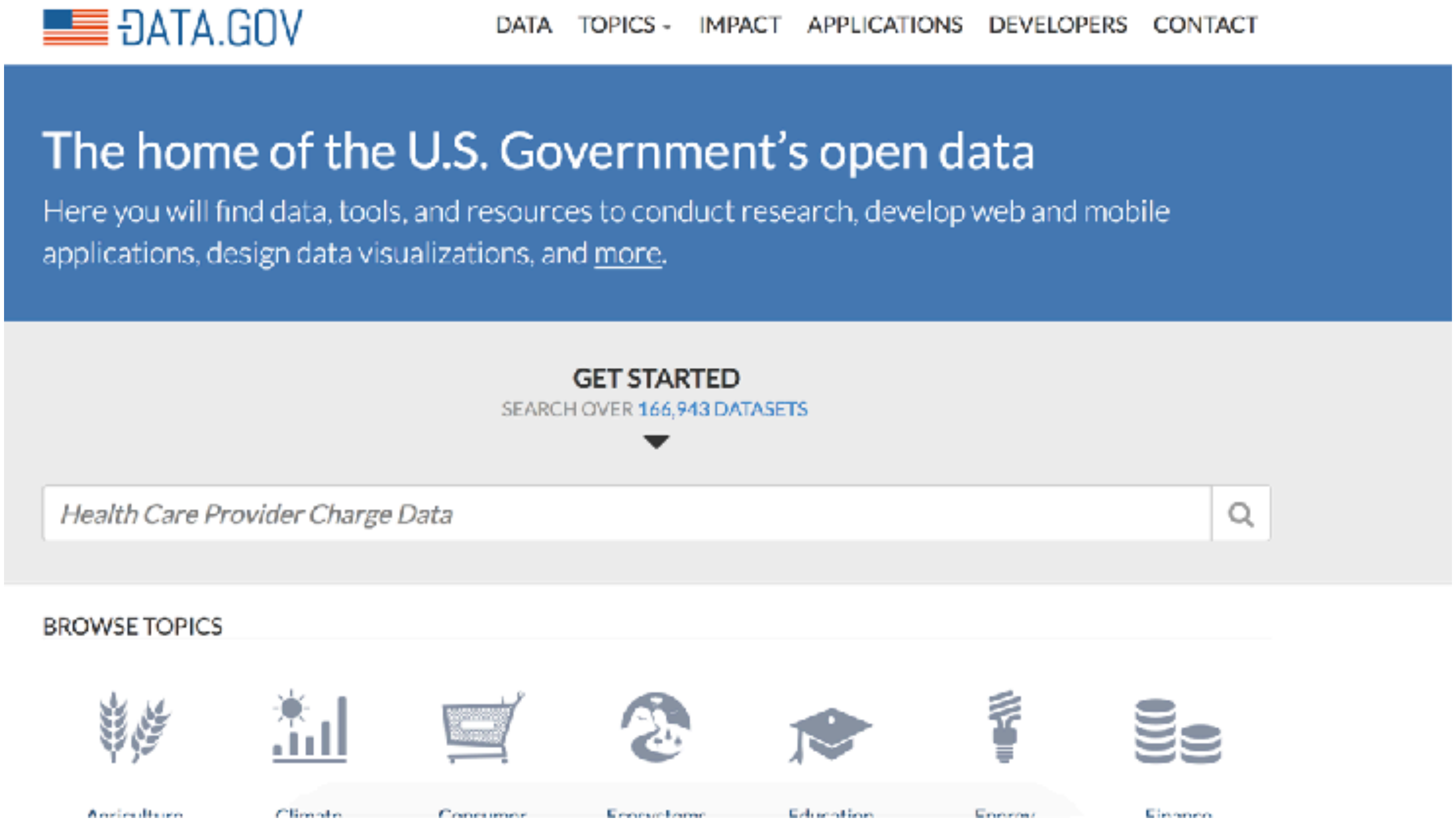
Data

movies   651 obs. of 32 variables

```
  .. ..$ actors          : list()
  .. .. ..- attr(*, "class")= chr  "collector_character" "collector"
  .. ..$ imdb_url        : list()
  .. .. ..- attr(*, "class")= chr  "collector_character" "collector"
  .. ..$ rt_url          : list()
  .. .. ..- attr(*, "class")= chr  "collector_character" "collector"
  ..$ default: list()
  .. ..- attr(*, "class")= chr  "collector_guess" "collector"
  ..- attr(*, "class")= chr "col_spec"
>
>
> ggplot(data = movies) +
+     geom_point(mapping = aes(x = imdb_rating, y = critics_score, color=mpaa_rating))
>
>
> |
```

Plots

First lab: data.gov



DATA.GOV

DATA TOPICS - IMPACT APPLICATIONS DEVELOPERS CONTACT

The home of the U.S. Government's open data

Here you will find data, tools, and resources to conduct research, develop web and mobile applications, design data visualizations, and more.

GET STARTED

SEARCH OVER 166,943 DATASETS

Health Care Provider Charge Data

BROWSE TOPICS

Agriculture Climate Consumer Ecosystems Education Energy Finance

Write code for humans, not computers

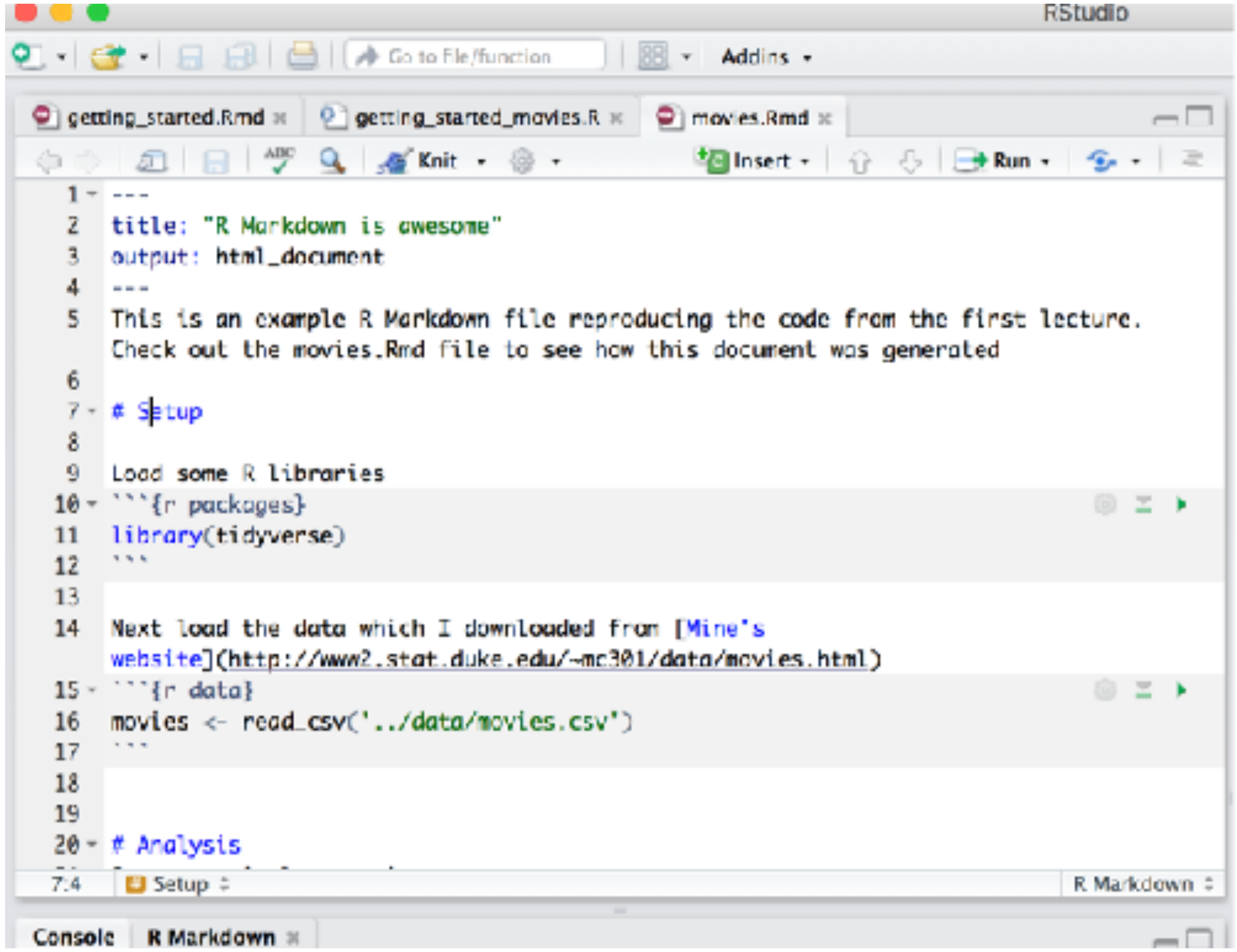
literate programming

- http://brandonrose.org/clustering

- https://cran.r-project.org/web/packages/tidytext/vignettes/tidytext.html

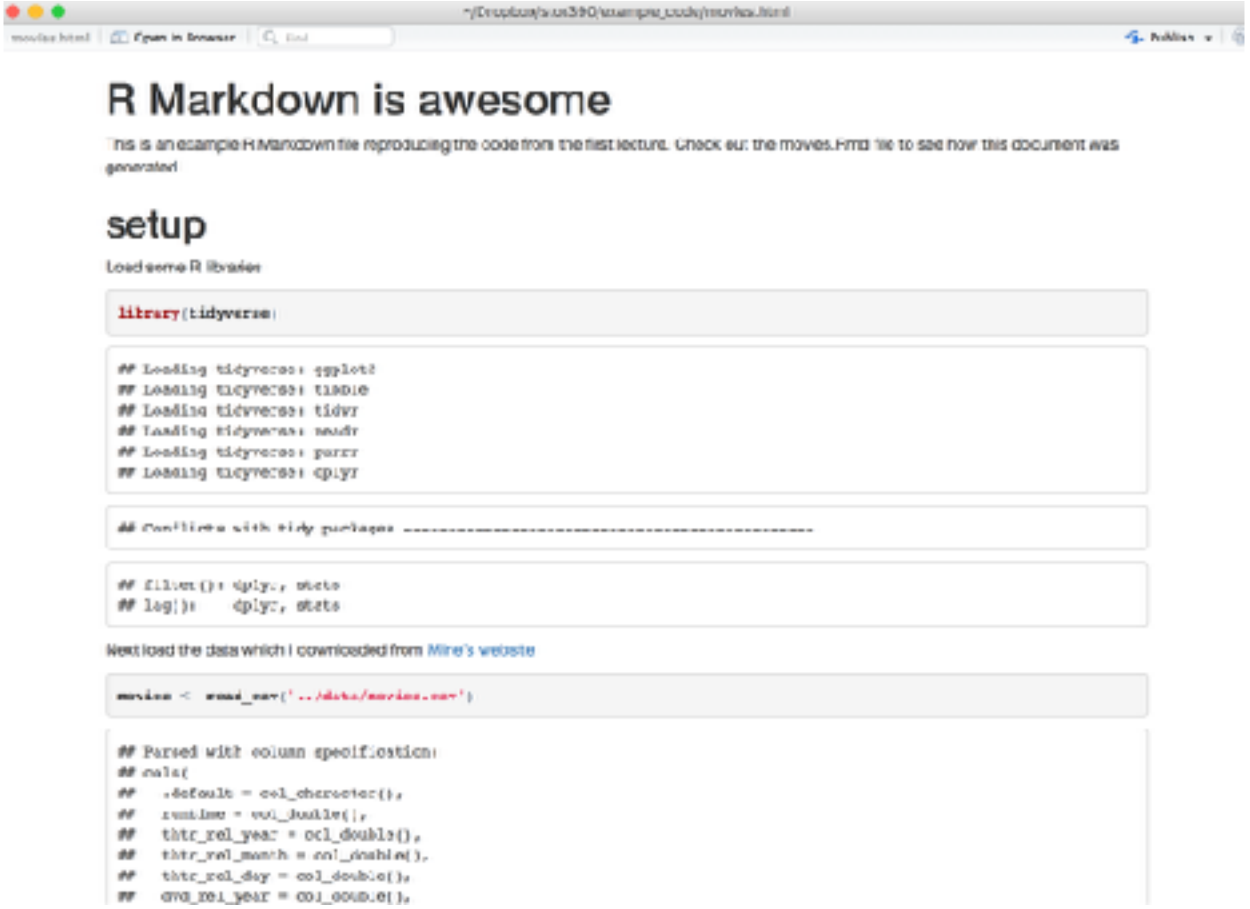
- https://github.com/idc9/brain-networks/blob/master/explore_igraph/EDA.ipynb

R Markdown is awesome



~~~
---
title: "R Markdown is awesome"
output: html_document
---
This is an example R Markdown file reproducing the code from the first lecture.
Check out the movies.Rmd file to see how this document was generated

# Setup

Load some R libraries
```{r packages}
library(tidyverse)
~~~

Next load the data which I downloaded from [Mine's
website](http://www2.stat.duke.edu/~mc301/data/movies.html)
```{r data}
movies <- read_csv('../data/movies.csv')
```

Analysis
```


movies.html
Open in Browser
Find
Publish

R Markdown is awesome

This is an example R Markdown file reproducing the code from the first lecture. Check out the movies.Rmd file to see how this document was generated

setup

Load some R libraries

```
library(tidyverse)
```

```
## Loading tidyverse: ggplot2
## Loading tidyverse: tibble
## Loading tidyverse: tidyr
## Loading tidyverse: readr
## Loading tidyverse: purrr
## Loading tidyverse: dplyr
```

```
## Conflicts with tidy packages ----------------------------------------------
```

```
## filter(): dplyr, stats
## lag():    dplyr, stats
```

Next load the data which I downloaded from Mine's website

```
movies <- read_csv('../data/movies.csv')
```

```
## Parsed with column specification:
## cols(
##   .default = col_character(),
##   runtime = col_double(),
##   thtr_rel_year = col_double(),
##   thtr_rel_month = col_double(),
##   thtr_rel_day = col_double(),
##   dvd_rel_year = col_double(),
```
```