

Topic 1: Introduction to R Programming

ISOM3390: Business Programming in R

Basics – What is R?

R is a dialect of the **S language**.

- A long past but a short history
- S developed primarily by John Chambers at Bell Labs in 70s

R is a language and environment for statistical computing and graphics, providing an **interactive** environment for data analysis.

Data analysis work performed with R can also be saved as scripts that can be easily executed at any moment.

R is **open source** and cross-platform.

- There is a large, growing, and active community of R users who contribute add-on functionalities and create interesting features.

Functionalities are shipped in form of **packages**.

- Currently, the **CRAN (Comprehensive R Archive Network)** package repository features 13000+ available packages.

Installing R and the R Console

We can download R freely from the [CRAN](#).

Interactive data analysis usually occurs on the R console. It executes commands as we type them.

RStudio: An Integrated Development Environment for R

RStudio includes an editor with many R specific features, a console to execute your code, and other useful panes.

Instructions on how to install RStudio are [here](#).

Once RStudio is installed, we can simply start RStudio rather than R since that program automatically starts R.

The R Ecosystem

The functionality provided by a fresh install of R is only a small fraction of what is possible, which we refer to as base R.

The extra functionality comes from add-ons available from users. There are currently thousands of these available from **CRAN**.

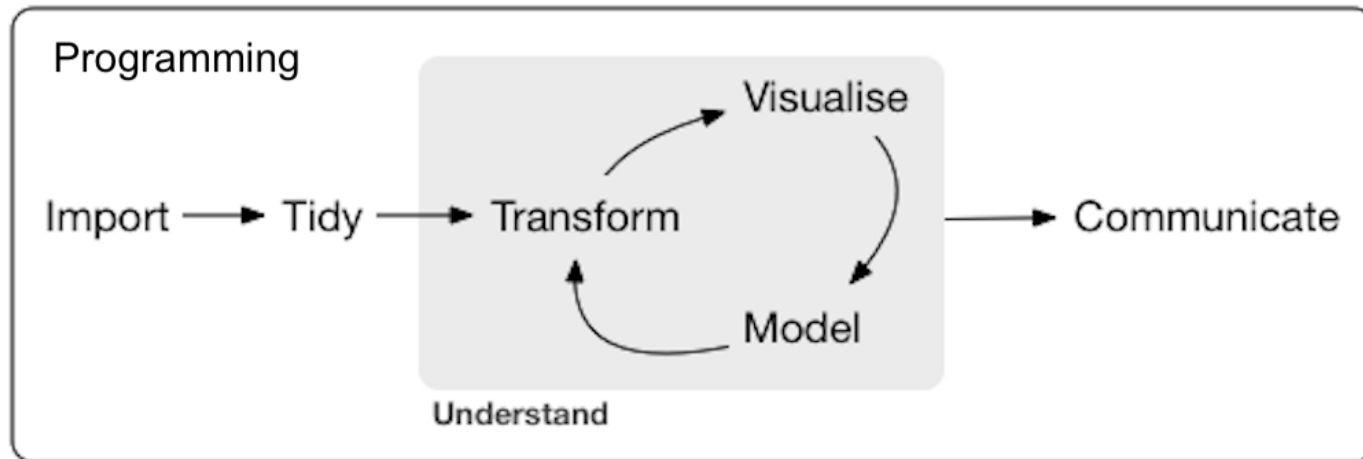
R makes it very easy to install packages from within R.

```
install.packages("tidyverse")
```

Once a package is installed, it remains installed and can then be loaded into an R sessions using the **library** function:

```
library("tidyverse")
```

What You will Learn?



Example: Sentiment Analysis

Blade Runner (1982)

Even better than Harrison Ford's *Raiders* movies. This one is a keeper. Replicants with feeling—that's a twist. The visual effects and fine photography draw you in. It's like you are really there in the metropolis of the future. Not a pretty picture of what's coming our way, but a great movie nonetheless.

My rating: 9



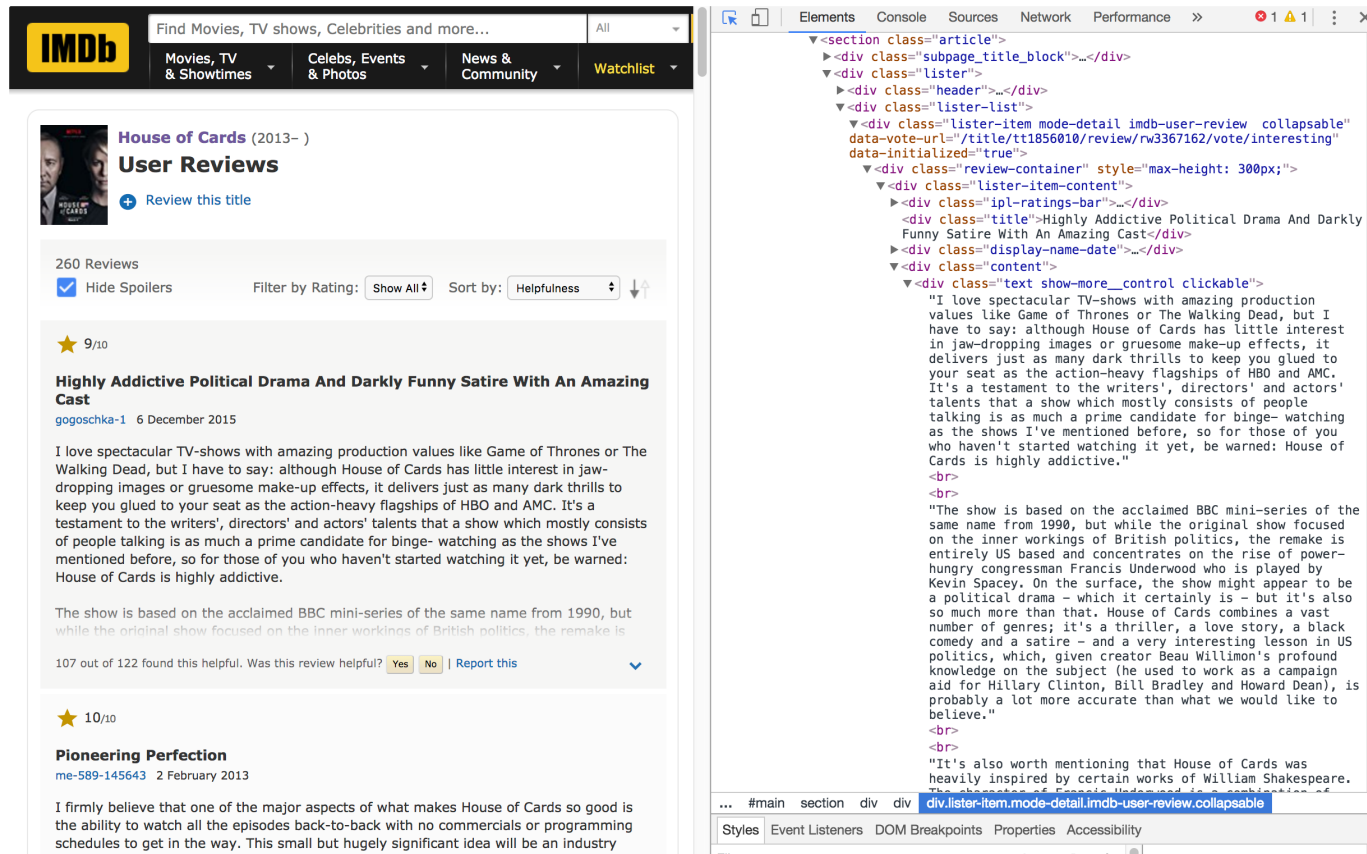
Moneyball (2011)

You might think I would like this movie, given my love of sports and analytics, but I had a hard time caring about the Oakland Athletics or the main character. Not Brad Pitt again. That guy sure keeps busy. I suppose the movie is a fair rendering of a true story. After all, you can't have the As winning the World Series when they didn't. But I was left with an empty feeling at the end. I think the story would have been better told from the point of view of the nerdy analyst. Give Jonah Hill more time in front of the camera and see what he can do. Maybe you could even work in a love interest. Some of my students asked me what I thought of the movie. I said it was OK.

My rating: 4



IMDb Review Data



The image displays the IMDb website interface for the TV show "House of Cards (2013-)" under the "User Reviews" section. The page shows 260 reviews, with filters for "Hide Spoilers" and sorting by "Helpfulness". Two reviews are visible: one by "gogoschka-1" dated 6 December 2015 with a 9/10 rating, and another by "me-589-145643" dated 2 February 2013 with a 10/10 rating.

The right side of the image shows the browser's developer tools with the DOM tree expanded. The selected node is `div.list-item.mode-detail.imdb-user-review.collapsible`, which contains the review text and rating information.

```
<div class="list-item mode-detail imdb-user-review collapsible" data-vote-url="/title/tt1856010/review/rw3367162/vote/interesting" data-initialized="true">
  <div class="review-container" style="max-height: 300px;">
    <div class="list-item-content">
      <div class="ipl-ratings-bar">...</div>
      <div class="title">Highly Addictive Political Drama And Darkly Funny Satire With An Amazing Cast</div>
      <div class="display-name-date">...</div>
      <div class="content">
        <div class="text show-more_control clickable">
          "I love spectacular TV-shows with amazing production values like Game of Thrones or The Walking Dead, but I have to say: although House of Cards has little interest in jaw-dropping images or gruesome make-up effects, it delivers just as many dark thrills to keep you glued to your seat as the action-heavy flagships of HBO and AMC. It's a testament to the writers', directors' and actors' talents that a show which mostly consists of people talking is as much a prime candidate for binge- watching as the shows I've mentioned before, so for those of you who haven't started watching it yet, be warned: House of Cards is highly addictive."
          <br>
          "The show is based on the acclaimed BBC mini-series of the same name from 1990, but while the original show focused on the inner workings of British politics, the remake is entirely US based and concentrates on the rise of power-hungry congressman Francis Underwood who is played by Kevin Spacey. On the surface, the show might appear to be a political drama – which it certainly is – but it's also so much more than that. House of Cards combines a vast number of genres; it's a thriller, a love story, a black comedy and a satire – and a very interesting lesson in US politics, which, given creator Beau Willimon's profound knowledge on the subject (he used to work as a campaign aid for Hillary Clinton, Bill Bradley and Howard Dean), is probably a lot more accurate than what we would like to believe."
          <br>
          "It's also worth mentioning that House of Cards was heavily inspired by certain works of William Shakespeare. The character of Francis Underwood is a combination of..."
        </div>
      </div>
    </div>
  </div>
</div>
```

To scrape the webpage, we need to locate specific nodes and extract the textual data of our interest.

Dynamic Web Scraping

The content is updated dynamically based on our interactions and inputs, and is determined only when the page is fully rendered.

We need a **Web browser automation tool** that opens a browser of our choice and "drives" it to perform tasks as a human being would, such as:

- Clicking buttons
- Entering information in forms
- Searching for specific information on the web pages

in a programmatical way.

Importing Data

```
## # A tibble: 273 x 4
##   revid usrid   rating review
##   <int> <chr>     <int> <chr>
## 1     1  ur157940~      9 I love spectacular TV-shows with amazing produc~
## 2     2  ur402241~     10 I firmly believe that one of the major aspects ~
## 3     3  ur6323884    10 When I thought about Francis Urquhart of the or~
## 4     4  ur236546~      8 Fans of David Fincher and Kevin Spacey have bee~
## 5     5  ur304446~      9 Just watched the first episode. It is outstandi~
## 6     6  ur5367771    10 "I was so looking forward to this, having been ~
## 7     7  ur569366~      4 "Show was great first couple seasons, but as wi~
## 8     8  ur205527~      7 "'House of Cards' much of the time was one of t~
## 9     9  ur485601~      7 OK (with spoilers).. so S5 is now over.. and wh~
## 10    10  ur483837~      3 Unexpexted failure all of the 3rd season. Total~
## 11    11  ur3290734      9 From the start, Kevin Spacey captivates & impre~
## 12    12  ur589825~      3 It's the worst season ever! Such a disappointme~
## 13    13  ur9710713      6 "After a couple of great season, the show is to~
## 14    14  ur0005383      2 What happened to series three? Even though the ~
## 15    15  ur528401~      1 I am struggling through the third season of Hou~
## 16    16  ur220310~      8 "As I was describing the show to a friend, I co~
## 17    17  ur302286~      7 I have to say Im not really into this show anym~
## 18    18  ur378228~      6 This show was fun to watch for the first 2 seas~
## 19    19  ur9096041    10 "I've seen only the first season. (DVDs of the ~
## 20    20  ur113447~      2 We started watching Season 1 and by Episode 6 w~
## # ... with 253 more rows
```

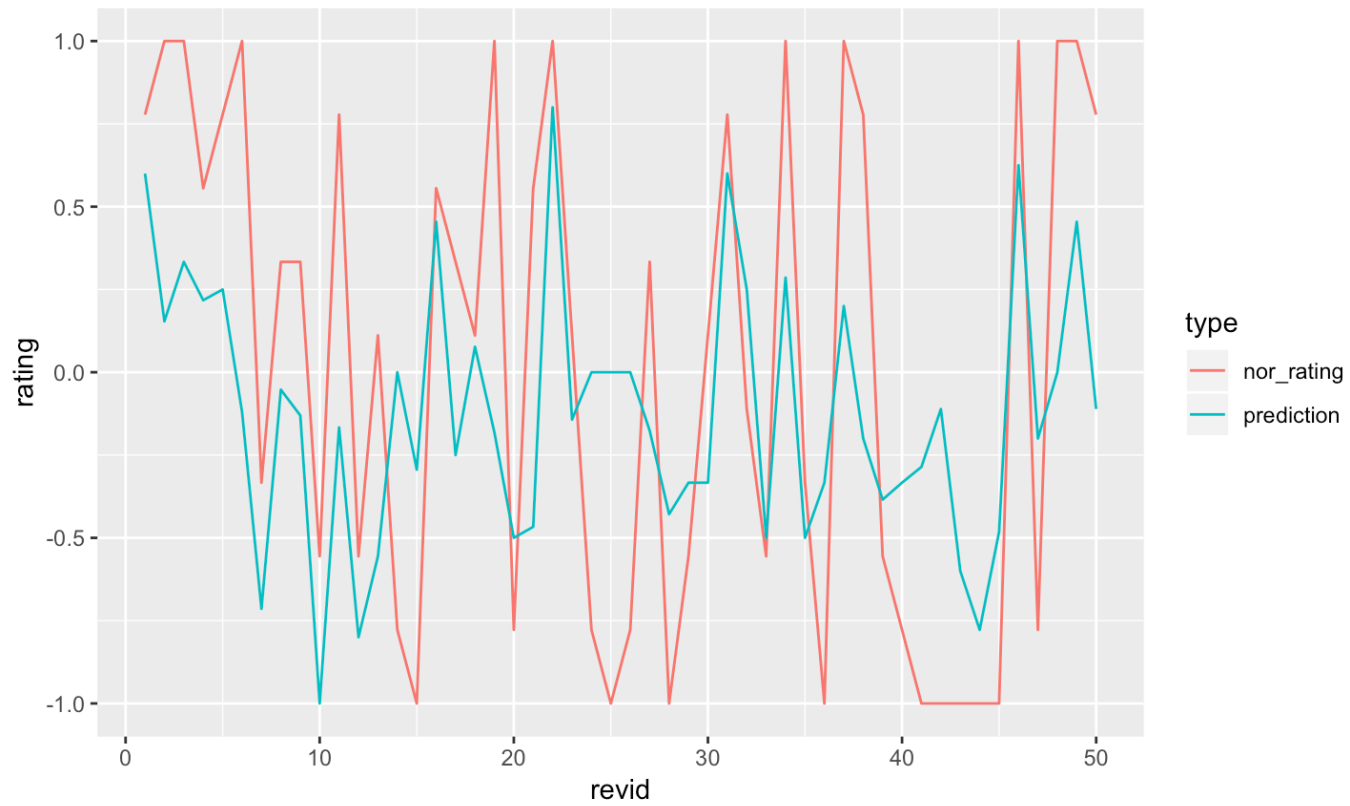
Data Wrangling

```
## # A tibble: 3,570 x 5
##   revid usrid      rating word      sentiment
##   <int> <chr>      <int> <chr>      <chr>
## 1     1 1 ur15794099      9 love      positive
## 2     1 1 ur15794099      9 spectacular positive
## 3     1 1 ur15794099      9 amazing    positive
## 4     1 1 ur15794099      9 dead       negative
## 5     1 1 ur15794099      9 gruesome   negative
## 6     1 1 ur15794099      9 dark       negative
## 7     1 1 ur15794099      9 thrills    positive
## 8     1 1 ur15794099      9 talents    positive
## 9     1 1 ur15794099      9 warned     negative
## 10    1 1 ur15794099      9 acclaimed  positive
## 11    1 1 ur15794099      9 love       positive
## 12    1 1 ur15794099      9 profound   positive
## 13    1 1 ur15794099      9 accurate   positive
## 14    1 1 ur15794099      9 worth      positive
## 15    1 1 ur15794099      9 evil       negative
## 16    1 1 ur15794099      9 famous     positive
## 17    1 1 ur15794099      9 fun        positive
## 18    1 1 ur15794099      9 charming   positive
## 19    1 1 ur15794099      9 deadly     negative
## 20    1 1 ur15794099      9 perfect    positive
## # ... with 3,550 more rows
```

Predicting Sentiments of Reviews

```
## # A tibble: 270 x 5
##   revid usrid      negative positive prediction
##   <int> <chr>      <dbl>    <dbl>    <dbl>
## 1     1    1 ur15794099      8     32     0.6
## 2     2    2 ur40224167     11     15    0.154
## 3     3    3 ur6323884      2      4    0.333
## 4     4    4 ur23654692      9     14    0.217
## 5     5    5 ur30444670      3      5    0.25
## 6     6    6 ur5367771     14     11   -0.12
## 7     7    7 ur56936628      6      1   -0.714
## 8     8    8 ur20552756     20     18   -0.0526
## 9     9    9 ur48560127     13     10   -0.130
## 10    10   10 ur48383725      9      0    -1
## 11    11   11 ur3290734      7      5   -0.167
## 12    12   12 ur58982585      9      1   -0.8
## 13    13   13 ur9710713      7      2   -0.556
## 14    14   14 ur0005383      5      5    0
## 15    15   15 ur52840124     11      6   -0.294
## 16    16   16 ur22031037      3      8    0.455
## 17    17   17 ur30228681      5      3   -0.25
## 18    18   18 ur37822881      6      7    0.0769
## 19    19   19 ur9096041     13      9   -0.182
## 20    20   20 ur11344732     12      4   -0.5
## # ... with 250 more rows
```

Visualizing for Comparison

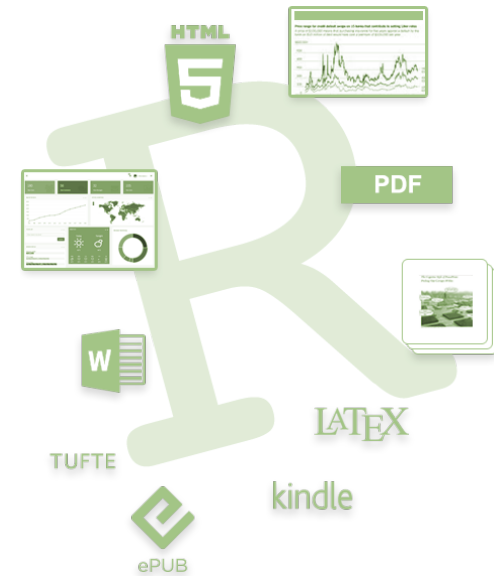


R Markdown

R Markdown provides a productive notebook interface and an unified authoring framework that weave together narrative text and code to produce elegantly formatted output.

R Markdown documents are fully reproducible and support dozens of output formats, like PDFs, Word files, slideshows, and more.

An R Markdown file is a plain text file that has the extension `.Rmd`.



An R Markdown Example

```
---  
title: "Diamond Sizes"  
output:  
  pdf_document: default  
  html_document: default  
---
```

```
```{r setup, include = FALSE}  
library(ggplot2)
library(dplyr)
```

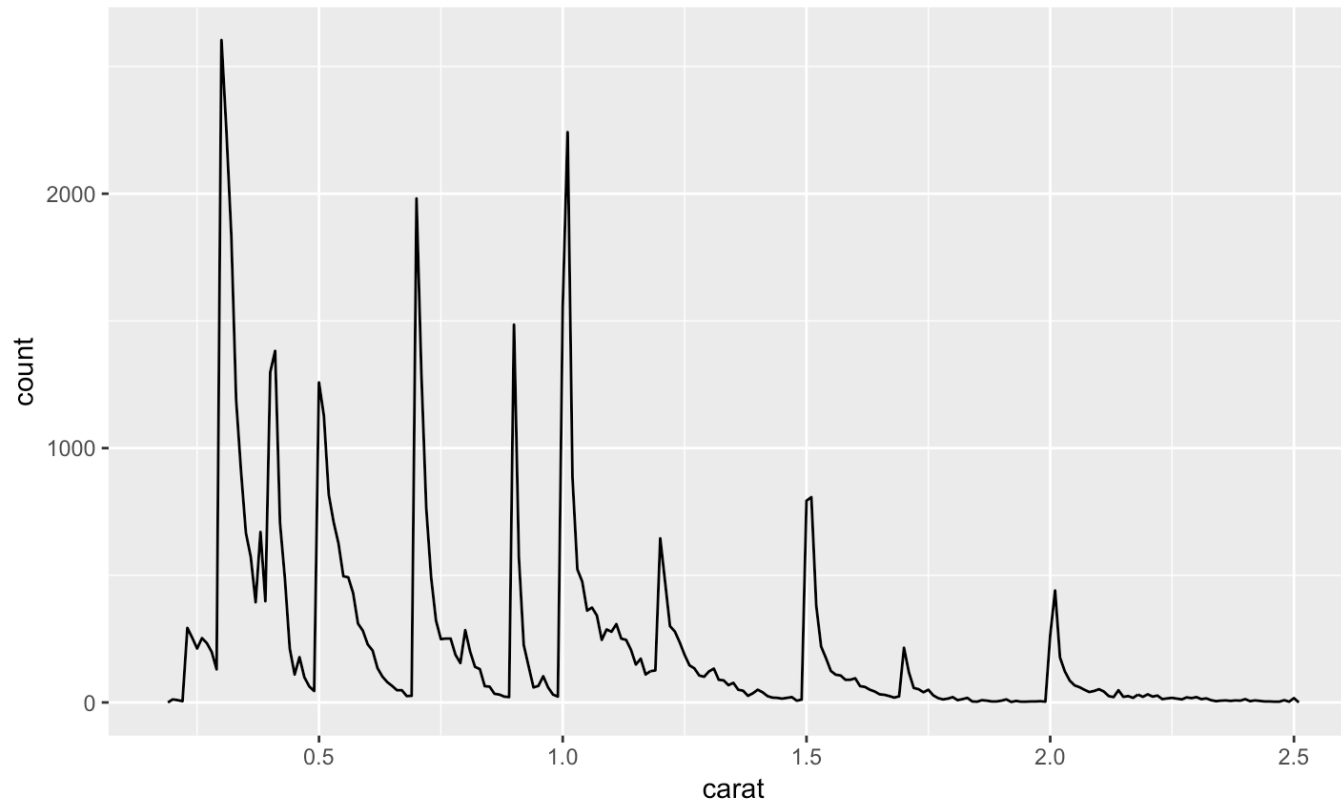
```
smaller <- diamonds %>%
 filter(carat <= 2.5)
```
```

We have data about `nrow(diamonds)` diamonds. Only `nrow(diamonds) - nrow(smaller)` are larger than 2.5 carats. The distribution of the remainder is shown below:

```
```{r, echo = FALSE}  
smaller %>%
 ggplot(aes(carat)) +
 geom_freqpoly(binwidth = 0.01)
```
```

An R Markdown Example, Continued

We have data about 53940 diamonds. Only 126 are larger than 2.5 carats. The distribution of the remainder is shown below:



R Markdown Basics

An R Markdown file contains three important types of content:

- An (optional) YAML header surrounded by `---`.
- Chunks of R code surrounded by `````.
- Text mixed with simple text formatting like `#` heading and `_italics_`.

Check the [cheet sheet](#) for more information.

Click "Knit" or press `Cmd/Ctrl + Shift + K` to produce a complete report containing all text, code, and results.

Course Mechanics

Two lectures a week: concepts, methods, examples, etc.

Lab to try stuff out and get fast feedback (10%)

Homework assignments do longer and more complex things (40%)

An in-class midterm quiz (Mar. 26) (25%)

Final group project (25%) (10% out of 25% will be based on your team-mates' assessment of your contribution to the project)

Assignments, class notes, grading policies, useful links on Canvas