Intro to R, Markdown and the Tidyverse

Jameson Watts, Ph.D.

Agenda

1. The R programming environment
2. RMarkdown
3. Intro to Tidyverse

Strong suggestion: download some cheat sheets

The R programming environment

What is R?

- R is a software environment for statistical computing and graphics.
- It is also a programming language
  - Variables (of different types)
  - Structures
  - Functions
  - Syntax

Using RStudio

- *Console:* where you can execute code
- *Source file:* create and save scripts, which get sent to the console
- *Comments:* document what you are doing
- *Environment:* a snapshot of the current state of affairs
- *Help:* lookup how to do stuff
- *Packages:* bundles of code written by others

Exercise

Open RStudio and:

1. Create a new R script
2. Create a comment that explains what you are doing
3. Compute the solution to 2*(28-7)
4. Run the script

Solution

```
# Math is fun!
2*(28-7)
```

```
## [1] 42
```

Exercise

1. Edit your R script to assign the solution to 2*(28-7) to a variable
2. Multiply your variable by 42 and then take the square root
3. Display the answer

Hint: use `sqrt()`

Solution

```
# Math is fun!
s <- 2*(28-7)
s <- s*42
sqrt(s)
```

```
## [1] 42
```

Functions

Input → Function → Output

```
# Exponentiate a number and assign to a variable
s <- exp(42)
s
```

```
## [1] 1.739275e+18
```

```
# Take the log of this variable and display it
log(s)
```

```
## [1] 42
```

Packages

- A package is a collection of functions, documentation, and sometimes data
- There are a number of packages that are part of base R
- You can install other packages from CRAN
- Not all packages are created equal

Note: The functions **exp()** and **log()** are part of Base R

Excercise

1. Install the tidyverse package
2. load tidyverse into memory

Solution

```r
# install.packages('tidyverse')
library(tidyverse)
```

```
## ── Attaching packages ──────────────────────────────────────────────────────────────────────────
```

```
## ✔ ggplot2 3.2.1     ✔ purrr   0.3.2
## ✔ tibble  2.1.3     ✔ dplyr   0.8.3
## ✔ tidyr   0.8.3     ✔ stringr 1.4.0
## ✔ readr   1.3.1     ✔ forcats 0.4.0
```

```
## ── Conflicts ───────────────────────────────────────────────────────────────────────────────────
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
```

RMarkdown

The what and why of RMarkdown

RMarkdown is a file format that allows you to embed your code directly within your report.

Why is this important?

Preamble

To begin, you need to specify they type of document you want and set some options…

…you need to write some YAML

```
---
 title: "Intro to R, Markdown and the Tidyverse"
 author: "Jameson Watts, Ph.D."
 output:
  ioslides_presentation:
    widescreen: yes
---

{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
```

…but luckily this is all pretty much done for you when you create a new RMarkdown file in RStudio.

Formatting

```
Plain text

# header 1
## header 2
### header 3 etc.

an equation $ y = mx + b $

image: ![caption]("pathtofile.png")

url: [link text](http://address.com)

> block quote

*italics*

**bold**
```

Formatting cont'd

```
superscript^2^

- unordered list
- next item
  + some subitems
  + subitem 2

1. unordered list
2. next item
  + some subitems
  + subitem 2

Table header | Column 2 header
-------------|--------------
cell 1       | cell 2
cell 3       | cell 4
```

Code Evaluation

Chunk it

```{r eval=FALSE, cache=FALSE}

# here is a comment
my_var <- 42
my_var

```

Inline

My favorite variable is: `r my_var`

Code Chunk Options

- eval
- echo
- warning
- error
- message
- cache
- fig.width
- fig.height

…and so much more can be found here

Exercise: Reproduce the output of this slide

Some rad markdown

Note: Using **markdown** is an **easy** way to marry analysis with report writing.

The advantages are:

1. Simple, output-agnostic formatting
2. Reproducible results
3. Easy for others to follow your analysis

Example analysis:

```
my_var <- exp(log(42))
```

Note[2]*:* It is also cool because I can embed R code inline. For instance, the value of my_var is 42.

Intro to the Tidyverse

Philosophy

An ‘opinionated’ collection of R packages designed for data science. All packages share an underlying design philosophy, grammar, and data structures.

- readr – importing data
- dplyr – manipulating data
- tidyr – cleaning data
- ggplot2 – visualizing data

…more are being added regularly

Importing data (pair up and follow along)

1. Goto jamesonwatts.github.io, scroll down to the teaching section and click on the GSMDS Class link
2. Find the resources folder and download the wine-data.csv file
3. Create a new R code chunk in your RMD file and add the following:

```
library(tidyverse)
wine <- read_csv("../resources/winemag-data.csv")
```

Data Frames

```
wine <- read_csv("../resources/winemag-data.csv")
```

```
## Parsed with column specification:
## cols(
##   X1 = col_double(),
##   country = col_character(),
##   description = col_character(),
##   designation = col_character(),
##   points = col_double(),
##   price = col_double(),
##   province = col_character(),
##   region_1 = col_character(),
##   region_2 = col_character(),
##   taster_name = col_character(),
##   taster_twitter_handle = col_character(),
##   title = col_character(),
##   variety = col_character(),
##   winery = col_character()
## )
```

Rectangular data

- All columns are variables
- All rows are observations
- Each cell is a value

```
head(wine,5)
```

```
## # A tibble: 5 x 14
##      X1 country description designation points price province region_1
##   <dbl> <chr>   <chr>       <chr>        <dbl> <dbl> <chr>    <chr>
## 1     0 Italy   Aromas inc… Vulkà Bian…     87    NA Sicily … Etna
## 2     1 Portug… This is ri… Avidagos        87    15 Douro    <NA>
## 3     2 US      Tart and s… <NA>            87    14 Oregon   Willame…
## 4     3 US      Pineapple … Reserve La…     87    13 Michigan Lake Mi…
## 5     4 US      Much like … Vintner's …     87    65 Oregon   Willame…
## # … with 6 more variables: region_2 <chr>, taster_name <chr>,
## #   taster_twitter_handle <chr>, title <chr>, variety <chr>, winery <chr>
```

Note: you can also open data files within RStudio just like in Excel

Let's take a glimpse instead

```
glimpse(wine)
```

```
## Observations: 129,971
## Variables: 14
## $ X1                    <dbl> 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, …
## $ country               <chr> "Italy", "Portugal", "US", "US", "US", "Sp…
## $ description           <chr> "Aromas include tropical fruit, broom, bri…
## $ designation           <chr> "Vulkà Bianco", "Avidagos", NA, "Reserve L…
## $ points                <dbl> 87, 87, 87, 87, 87, 87, 87, 87, 87, 87, 87…
## $ price                 <dbl> NA, 15, 14, 13, 65, 15, 16, 24, 12, 27, 19…
## $ province              <chr> "Sicily & Sardinia", "Douro", "Oregon", "M…
## $ region_1              <chr> "Etna", NA, "Willamette Valley", "Lake Mic…
## $ region_2              <chr> NA, NA, "Willamette Valley", NA, "Willamet…
## $ taster_name           <chr> "Kerin O'Keefe", "Roger Voss", "Paul Gregu…
## $ taster_twitter_handle <chr> "@kerinokeefe", "@vossroger", "@paulgwine …
## $ title                 <chr> "Nicosia 2013 Vulkà Bianco  (Etna)", "Quin…
## $ variety               <chr> "White Blend", "Portuguese Red", "Pinot Gr…
## $ winery                <chr> "Nicosia", "Quinta dos Avidagos", "Rainsto…
```

Or a skim

```
#install.packages("skimr")
library(skimr)
skim(wine)
```

```
## Skim summary statistics
##  n obs: 129971
##  n variables: 14
##
## ── Variable type:character ──────────────────────────────────────────────────────────────────────────────────────
##               variable missing complete      n min max empty n_unique
##                country      63   129908 129971   2  22     0       43
##            description       0   129971 129971  20 829     0   119955
##            designation   37465    92506 129971   1  95     0    37979
##               province      63   129908 129971   3  31     0      425
##               region_1   21247   108724 129971   3  50     0     1229
##               region_2   79460    50511 129971   4  17     0       17
##            taster_name   26244   103727 129971  10  18     0       19
##  taster_twitter_handle   31213    98758 129971   6  16     0       15
##                  title       0   129971 129971  12 136     0   118840
##                variety       1   129970 129971   4  35     0      707
##                 winery       0   129971 129971   1  54     0    16757
##
## ── Variable type:numeric ────────────────────────────────────────────────────────────────────────────────────────
##  variable missing complete      n     mean       sd p0     p25   p50
##    points       0   129971 129971    88.45     3.04 80    86      88
##     price    8996   120975 129971    35.36    41.02  4    17      25
##        X1       0   129971 129971 64985    37519.54  0 32492.5 64985
##      p75   p100     hist
##     91      100 ▂▂▅▇▅▂▂▂
##     42     3300 ▇▂▂▂▂▂▂▂
##  97477.5 129970 ▇▇▇▇▇▇▇▇
```

28/37

The Pipe

Starting with a dataset, you can use the pipe operator to perform operations in series.

For instance, I can filter the dataset to only show wines from Oregon

```
wine %>%
  filter(province=="Oregon")
```

```
## # A tibble: 5,373 x 14
##       X1 country description designation points price province region_1
##    <dbl> <chr>   <chr>       <chr>        <dbl> <dbl> <chr>    <chr>
##  1     2 US      Tart and s… <NA>            87    14 Oregon   Willame…
##  2     4 US      Much like … Vintner's …     87    65 Oregon   Willame…
##  3    21 US      A sleek mi… <NA>            87    20 Oregon   Oregon
##  4    35 US      As with ma… Hyland          86    50 Oregon   McMinnv…
##  5    41 US      A stiff, t… <NA>            86    22 Oregon   Willame…
##  6    78 US      Some rosés… Rosé of         86    25 Oregon   Eola-Am…
##  7   173 US      This wine … <NA>            91    38 Oregon   Willame…
##  8   233 US      There is a… Reserve         85    28 Oregon   Willame…
##  9   248 US      This seems… Estate Sin…     85    45 Oregon   Willame…
## 10   251 US      Spicy and … Papillon E…     85    22 Oregon   Willame…
## # … with 5,363 more rows, and 6 more variables: region_2 <chr>,
## #   taster_name <chr>, taster_twitter_handle <chr>, title <chr>,
## #   variety <chr>, winery <chr>
```

Multiple Filters

…or I can filter the data so that I only have wines from Oregon that are over $100

```
wine %>%
  filter(province=="Oregon") %>%
  filter(price > 100)
```

```
## # A tibble: 37 x 14
##       X1 country description designation points price province region_1
##    <dbl> <chr>   <chr>       <chr>        <dbl> <dbl> <chr>    <chr>
##  1   778 US      The Winder… Winderlea …     92   105 Oregon   Dundee …
##  2  1082 US      Lighter in… The Tribe …     94   120 Oregon   Walla W…
##  3  5218 US      Just one b… Select Who…     92   150 Oregon   McMinnv…
##  4 16333 US      From selec… Pas de Nom      94   125 Oregon   Willame…
##  5 16531 US      This is a … <NA>            96   240 Oregon   Walla W…
##  6 16535 US      The aromas… Sur Echala…     95   120 Oregon   Walla W…
##  7 18000 US      Focused an… Abetina         94   105 Oregon   Willame…
##  8 33791 US      Focused an… Abetina         94   105 Oregon   Willame…
##  9 33811 US      Multiple v… Récolte Gr…     93   125 Oregon   Dundee …
## 10 34617 US      Just two b… Olson Esta…     91   125 Oregon   Dundee …
## # … with 27 more rows, and 6 more variables: region_2 <chr>,
## #   taster_name <chr>, taster_twitter_handle <chr>, title <chr>,
## #   variety <chr>, winery <chr>
```

Exercise

Find only the wines with variety “Pinot Gris” that cost less than $10

Answer

```
wine %>%
  filter(variety=="Pinot Gris") %>%
  filter(price<10)
```

```
## # A tibble: 7 x 14
##       X1 country description designation points price province region_1
##    <dbl> <chr>   <chr>       <chr>        <dbl> <dbl> <chr>    <chr>
## 1  27397 US      An enjoyab… <NA>            86     9 Washing… Columbi…
## 2  28038 Moldova There's a … Golden Lan…     86     9 Moldova  <NA>
## 3  35035 Moldova This Moldo… <NA>            86     9 Moldova  <NA>
## 4  62157 US      Firm and f… <NA>            90     9 Washing… Rattles…
## 5  64951 Argent… Flat on th… <NA>            83     9 Mendoza… Mendoza
## 6  70018 US      Light and … <NA>            84     9 Washing… Yakima …
## 7 117199 Romania Aromas of … Vine in Fl…     88     9 Dealu M… <NA>
## # … with 6 more variables: region_2 <chr>, taster_name <chr>,
## #   taster_twitter_handle <chr>, title <chr>, variety <chr>, winery <chr>
```

why do I put quotes around “Pinot Gris” but not around the price?

Combining functions

You can also combine functions using the pipe operator.

```
wine %>%
  filter(variety=="Chardonnay") %>%
  filter(province=="Oregon") %>%
  arrange(desc(points), price)
```

```
## # A tibble: 498 x 14
##        X1 country description designation points price province region_1
##     <dbl> <chr>   <chr>       <chr>        <dbl> <dbl> <chr>    <chr>
##  1 102489 US      Even if wi… Estate          96    27 Oregon   Dundee …
##  2  47900 US      This sensa… Aurora Vin…     96    60 Oregon   Willame…
##  3  47902 US      This is th… Récolte Gr…     96   125 Oregon   Dundee …
##  4  31421 US      As fabulou… Shea Viney…     95    35 Oregon   Willame…
##  5  78307 US      Here is an… Shea Viney…     95    35 Oregon   Willame…
##  6  48377 US      Stunning i… <NA>            95    42 Oregon   Willame…
##  7  47917 US      Sourced fr… Original V…     95    45 Oregon   Dundee …
##  8  95058 US      One large … Essence         95    45 Oregon   Ribbon …
##  9  76390 US      This conve… <NA>            95    49 Oregon   Willame…
## 10  31413 US      If anyone … Hyland Vin…     95    50 Oregon   McMinnv…
## # … with 488 more rows, and 6 more variables: region_2 <chr>,
## #   taster_name <chr>, taster_twitter_handle <chr>, title <chr>,
## #   variety <chr>, winery <chr>
```

Filtered summary of select variables

Let's summarize the data from the last slide by combining filter with skim and select

```
wine %>%
  filter(variety=="Chardonnay") %>%
  filter(province=="Oregon") %>%
  select(price, points) %>%
  skim()
```

```
## Skim summary statistics
##  n obs: 498
##  n variables: 2
##
## ── Variable type:numeric ──────────────────────────────────────────────────────────────────────────────────────
##  variable missing complete   n  mean    sd p0 p25 p50   p75 p100     hist
##    points       0      498 498 89.72  2.92 80  88  90 92      96 ▁▁▁▂▅▅▅▅▂▁
##     price       0      498 498 34.94 18.92  7  22  30 44.25  125 ▅▅▂▂▁▁▁▁▁▁
```

Long exercise

1. Gather into 4 teams
2. Assign one person to “drive”
3. Create a new RMD file using the default HTML output format
4. Suggest some Oregon wines in the region of…

- Eola-Amity Hills
- Dundee Hills
- Chehalem Mountains
- Umpqua Valley

Email me the resulting html file with your marked up analysis. Don’t forget to justify your recommendations!

Summary

- R is a programming language (in all its glory and ugliness)
- RMarkdown makes beautiful, reproducible, data science documents
- The Tidyverse is a philosophy (and set of packages) for data science in R

Bonus

###Beautiful tables with kable

```
wine %>%
  filter(variety=="Chardonnay") %>%
  filter(province=="Oregon") %>%
  arrange(desc(points), price) %>%
  select(points, price, province, region_1, title) %>%
  head(5) %>%
  knitr::kable(padding=0)
```

points	price	province	region_1	title
96	27	Oregon	Dundee Hills	The Eyrie Vineyards 2014 Estate Chardonnay (Dundee Hills)
96	60	Oregon	Willamette Valley	Ponzi 2012 Aurora Vineyard Chardonnay (Willamette Valley)
96	125	Oregon	Dundee Hills	Domaine Serene 2011 Récolte Grand Cru Chardonnay (Dundee Hills)
95	35	Oregon	Willamette Valley	Shea 2014 Shea Vineyard Chardonnay (Willamette Valley)
95	35	Oregon	Willamette Valley	Shea 2013 Shea Vineyard Chardonnay (Willamette Valley)