CT5102: Programming for Data Analytics

Lecture 11: R Markdown

Dr. Jim Duggan,

School of Engineering & Informatics

National University of Ireland Galway.
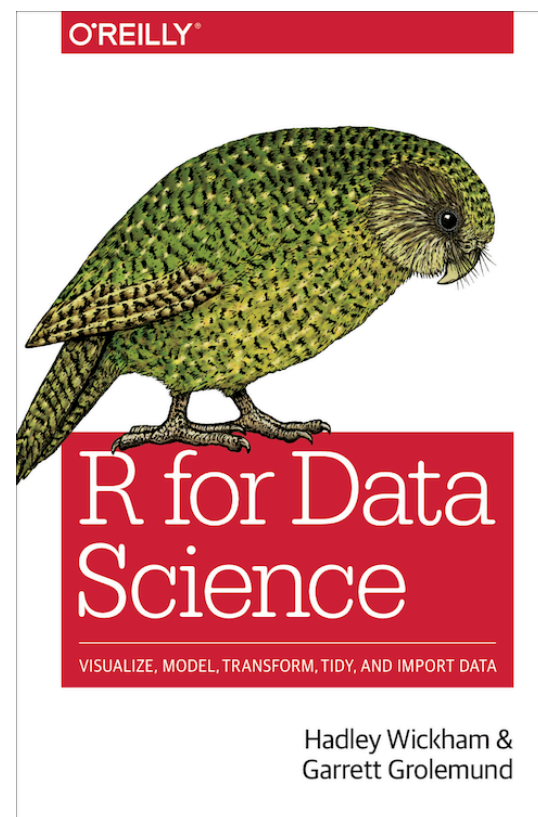
https://github.com/JimDuggan/EDAR

https://twitter.com/_jimduggan

R Markdown
(Wickham and Grolemund)

- R Markdown provides a unified authoring framework for data science, combining your code, results and prose commentary

- R Markdown documents are fully reproducible and support many output formats (pdfs, slideshows, and more).

Why use R Markdown?

- **For communicating to decision makers**, who want to focus on conclusions, not the code behind the analysis.

- **For collaborating with other data scientists**, who are interested in your conclusions, and how you reached them

- As an environment in which to do data science, where you capture not only what you did, but what you were thinking

NUI Galway
OÉ Gaillimh

R Markdown, loaded by R Studio

- 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_

Example

```
---
title: "Diamond Sizes"
date: 2017-08-25
output:
  html_document: default
---
```

Here is an example of using **R Markdown**.

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

```{r, echo=FALSE}
smaller <- diamonds %>%
  filter(carat <= 2.5)
```

We have data about `r nrow(diamonds)` diamonds in our data set.
Only **`r nrow(diamonds) - nrow(smaller)`**
are larger than 2.5 carats.

The distribution of the remainder is show below:

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

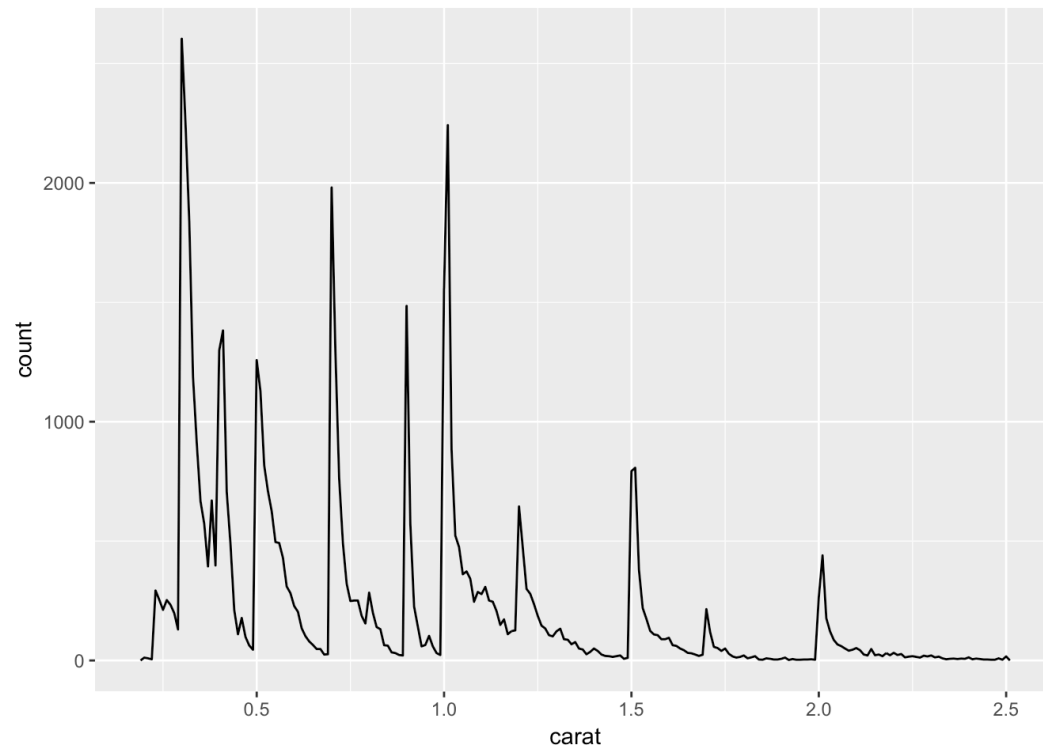https://yihui.name/knitr/options/
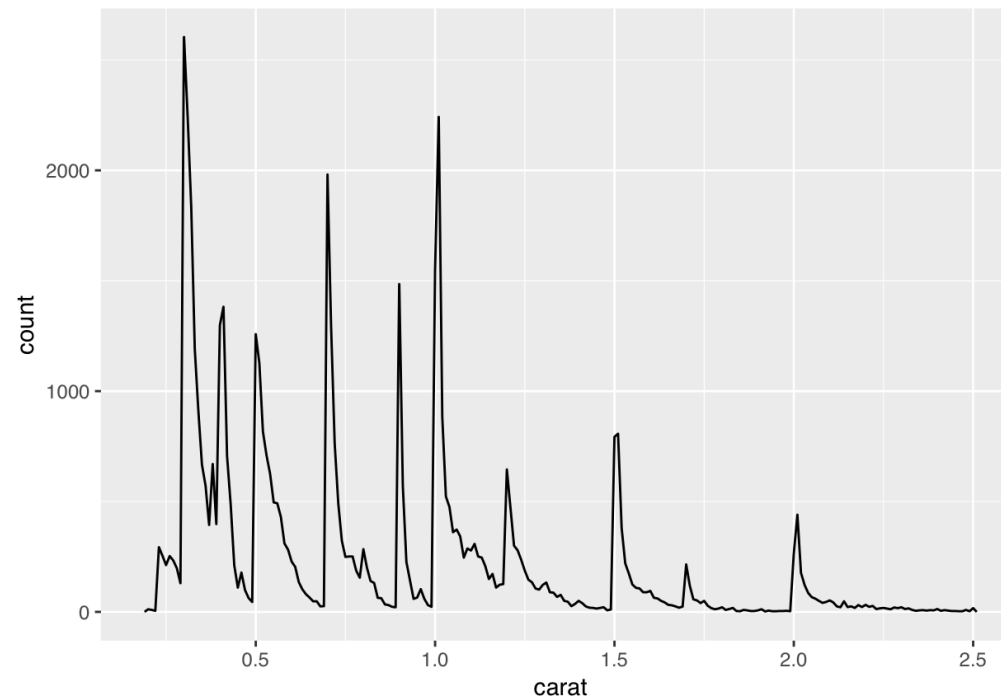
"knit" to HTML

Diamond Sizes

2017-08-25

Here is an example of using **R Markdown**.

We have data about 53940 diamonds in our data set. Only **126** are larger than 2.5 carats.

The distribution of the remainder is show below:

"knit" to PDF

Text Formatting with Markdown

- Text Formatting
  - *italic* or _italic_
  - **bold** or _bold_
  - `code`
- Headings
  - # First Level header
  - ## Second Level header
  - ### Third Level header
- Lists
  * Bulleted list item 1
  1 Numbered list item 1

- Links
  [linked phrase]http://example.com
- Images
  ![caption text](path/to/imp.png)

Inserting Chunks (Cmd/Ctrl-Alt-I)

This table summarizes what types of output each option suppresses…

Option	Run code	Show Code	Output	Plots	Messages	Warnings
eval = FALSE	X		X	X	X	X
include = FALSE		X	X	X	X	X
echo = FALSE		X				
results= "hide"			X			
fig.show="hide"				X		
message=FALSE					X	
Warning=FALSE						X

Table

```
---
title: "Table Test"
output: html_document
---

```{r}
mtcars[1:5,1:10]
```

```{r}
knitr::kable(
  mtcars[1:5,1:10],
  caption="A knitr kable"
)
```
```

Table Test

```
mtcars[1:5,1:10]
```

```
##                    mpg cyl disp  hp drat    wt  qsec vs am gear
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3
```

```
knitr::kable(
  mtcars[1:5,1:10],
  caption="A knitr kable"
)
```

A knitr kable

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear
Mazda RX4	21.0	6	160	110	3.90	2.620	16.46	0	1	4
Mazda RX4 Wag	21.0	6	160	110	3.90	2.875	17.02	0	1	4
Datsun 710	22.8	4	108	93	3.85	2.320	18.61	1	1	4
Hornet 4 Drive	21.4	6	258	110	3.08	3.215	19.44	1	0	3
Hornet Sportabout	18.7	8	360	175	3.15	3.440	17.02	0	0	3

Inline Code

There are `r nrow(mtcars)` records in *mtcars*

Table Test

```
mtcars[1:5,1:10]
```

```
##                    mpg cyl disp  hp drat    wt  qsec vs am gear
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3
```

```
knitr::kable(
  mtcars[1:5,1:10],
  caption="A knitr kable"
)
```

A knitr kable

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear
Mazda RX4	21.0	6	160	110	3.90	2.620	16.46	0	1	4
Mazda RX4 Wag	21.0	6	160	110	3.90	2.875	17.02	0	1	4
Datsun 710	22.8	4	108	93	3.85	2.320	18.61	1	1	4
Hornet 4 Drive	21.4	6	258	110	3.08	3.215	19.44	1	0	3
Hornet Sportabout	18.7	8	360	175	3.15	3.440	17.02	0	0	3

There are 32 records in *mtcars*

YAML Header

- "Yet Another Markup Language"

- Useful features
  - Parameters
  - Bibliographies

Example

```
---
title: "Parameter Test"
bibliography: ref.bib
params:
  my_class: suv
  my_time: !r lubridate::now()

output:
  html_document: default
  pdf_document: default
---

The time is now `r params$my_time`

The reference is [@paper1]
```

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

class <- mpg %>% filter(class == params$my_class)

```
````

````
```{r, message=FALSE}
ggplot(class,aes(x=displ,y=hwy))+
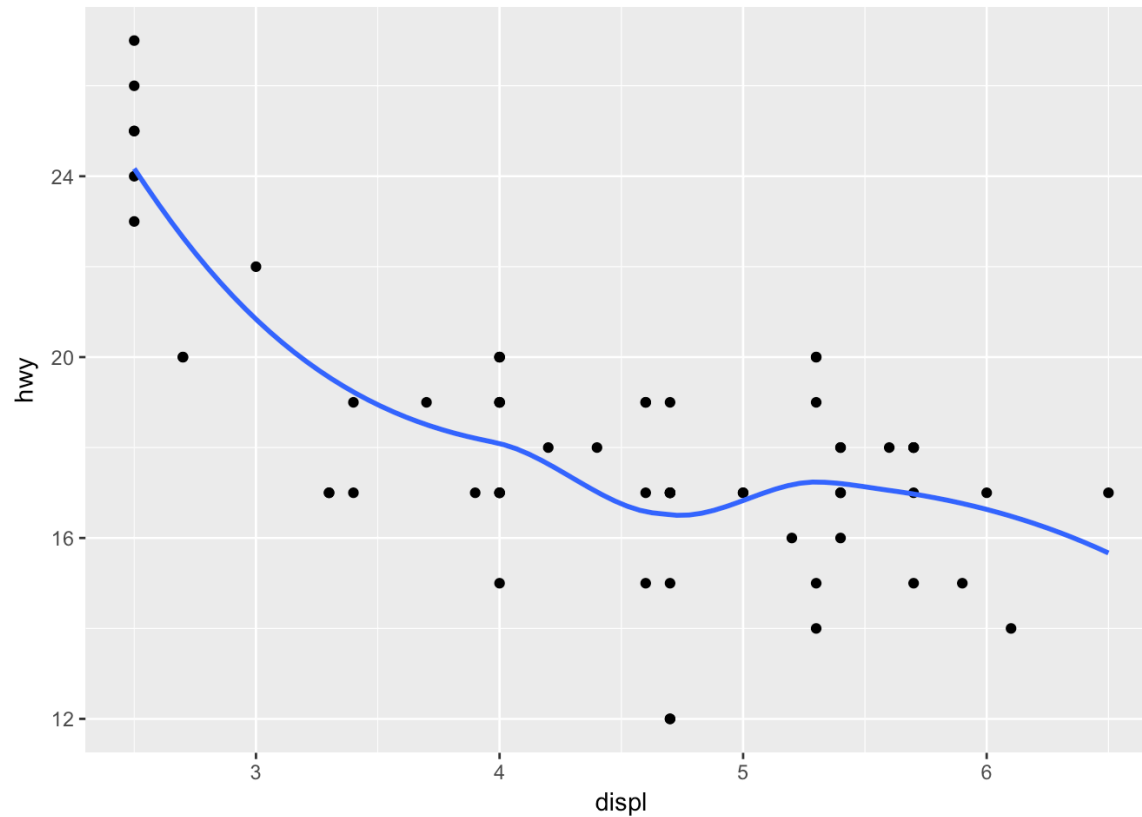  geom_point()+
  geom_smooth(se=F)

```
````

Parameter Test

The time is now 2017-11-15 19:41:58

The reference is (Koppeschaar et al. 2017)

```
ggplot(class,aes(x=displ,y=hwy))+
   geom_point()+
   geom_smooth(se=F)
```

Koppeschaar, E. Carl, Vittoria Colizza, Caroline Guerrisi, Clément Turbelin, Jim Duggan, John W. Edmunds, Charlotte Kjelsø, et al. 2017. "Influenzanet: Citizens Among 10 Countries Collaborating to Monitor Influenza in Europe." *JMIR Public Health Surveill* 3 (3): e66. http://publichealth.jmir.org/2017/3/e66/.

Challenge 11.1

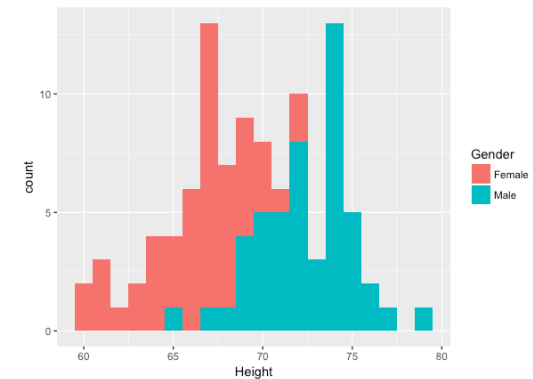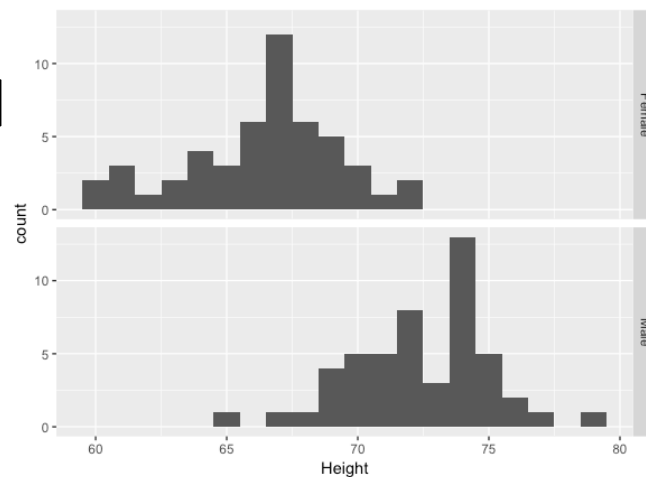http://www.math.hope.edu/swanson/data/heights.txt

- Create a tibble

- Plot histograms

- Perform statistical test on difference in category

- Design an outline for a short report

- Implement report using R Markup, and create PDF

```
> sd
# A tibble: 2 x 5
  Gender  Mean Median    SDev Number
  <fctr> <dbl>  <dbl>   <dbl>  <int>
1 Female 66.40   67.0 2.920721     50
2   Male 72.42   72.5 2.665655     50
```



```
Welch Two Sample t-test

data:  males and females
t = 10.765, df = 97.193, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 4.910132 7.129868
sample estimates:
mean of x mean of y
    72.42     66.40
```