GG606

Intro & Data Vis

Background

- Not a lot of data skills/courses
- Various ways to teach
  - Teach language
  - Teach stats

Why

- Why grad courses?
- Thesis vs project
- Pedagogy

Introductions

- Why are you in GG606
- Undergrad degree
- Research
- Experience with R, other languages

Caveats

- Not a ‘Learn R’ course per se
  - Use R to discuss concepts
  - Pros + Cons to R
- Based around *R4DS* book
  - Also the IDS book from UBC
  - You will bring new data most weeks

Structure of Grad Courses

- Topics & Discussions
- Lectures & Assignments
- Professional Development
- Skill & Practicum

R & RStudio

- Use R & RStudio example platforms
  - R language
  - RStudio integrated development environment IDE
  - posit.cloud

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins | Project: (None)

code.R

Source on Save | Run | Source

```
1  library(ggplot2)
2
3  summary(diamonds)
4
5  ggplot(diamonds, aes(depth, fill = cut, colour = cut)) +
6    geom_density(alpha = 0.5) +
7    xlim(55, 70)
8
```

8:1 (Top Level) R Script

Console C:/Users/moragasp/

```
 Max.   :79.00   Max.   :95.00   Max.   :18823   Max.   :10.740

       y                z
 Min.   : 0.000   Min.   : 0.000
 1st Qu.: 4.720   1st Qu.: 2.910
 Median : 5.710   Median : 3.530
 Mean   : 5.735   Mean   : 3.539
 3rd Qu.: 6.540   3rd Qu.: 4.040
 Max.   :58.900   Max.   :31.800

>
> ggplot(diamonds, aes(depth, fill = cut, colour = cut)) +
+   geom_density(alpha = 0.5) +
+   xlim(55, 70)
Warning message:
Removed 45 rows containing non-finite values (stat_density).
>
```

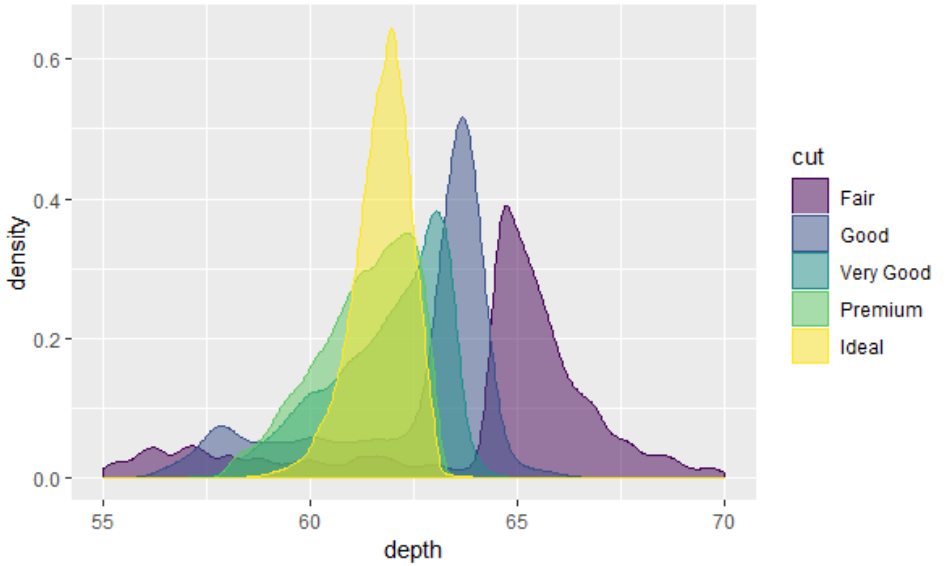
Environment History Connections

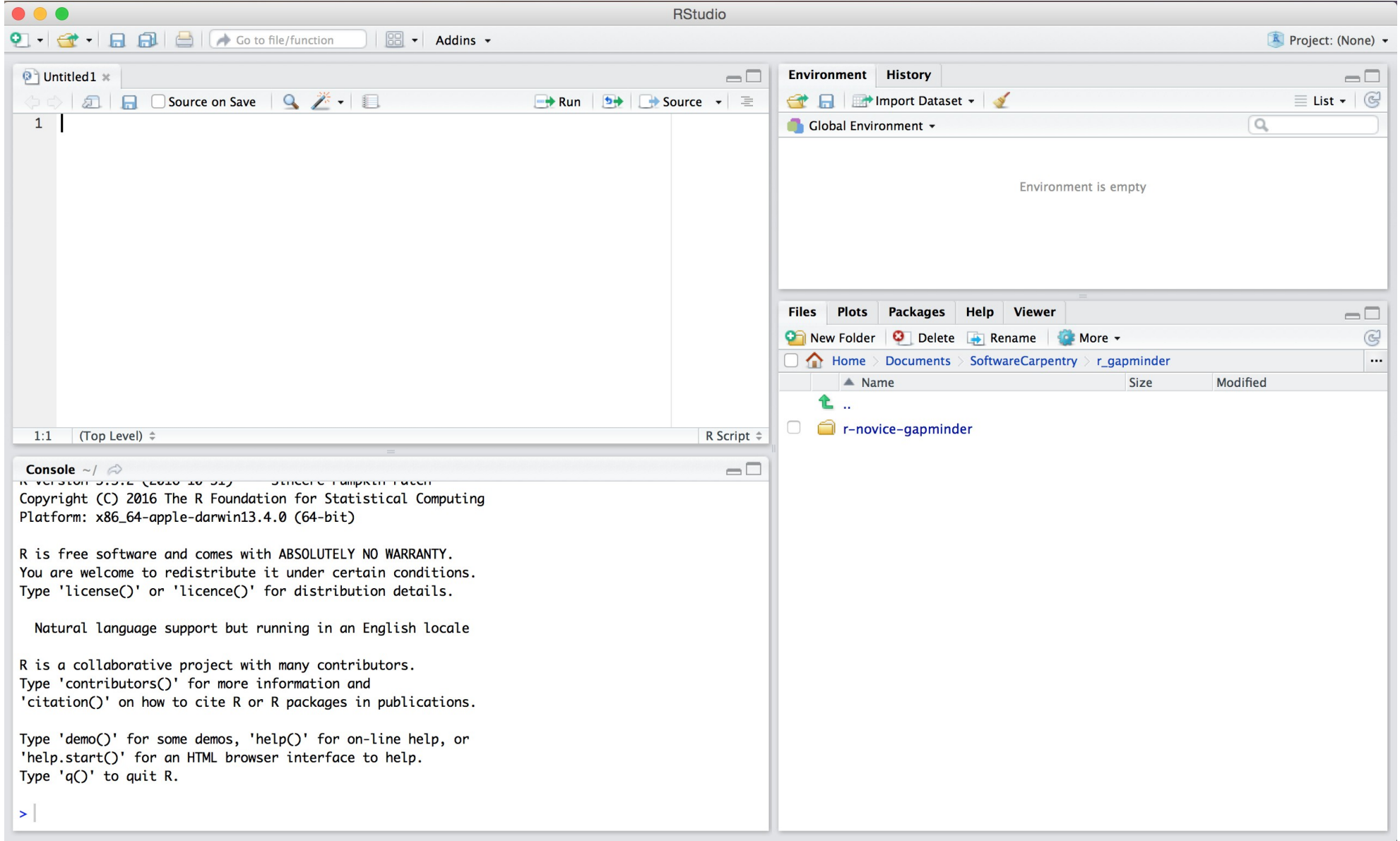
To Console | To Source

```
library(ggplot2)
summary(diamonds)
ggplot(diamonds, aes(depth, fill = cut, colour = cut)) +
geom_density(alpha = 0.5) +
xlim(55, 70)
```

Files Plots Packages Help Viewer

Zoom | Export | Publish

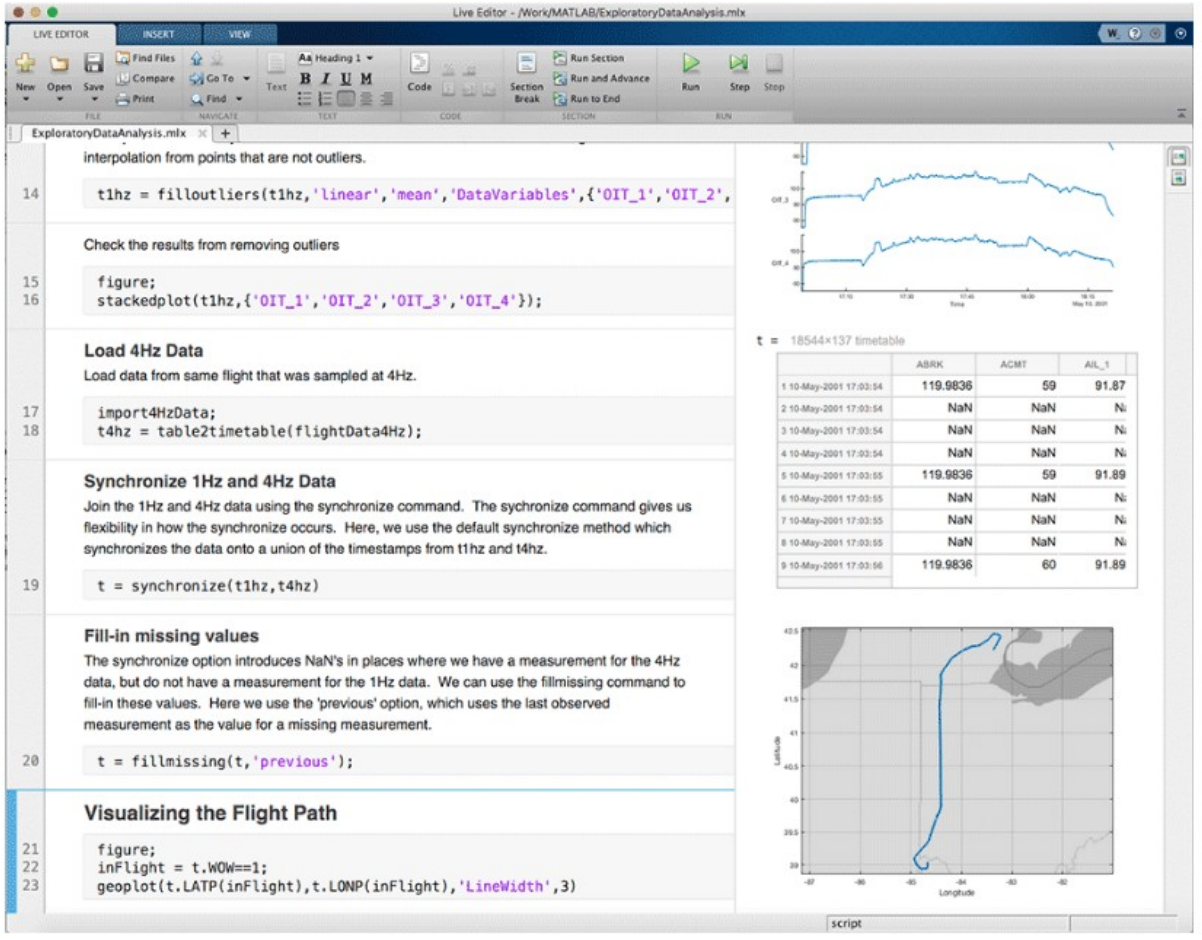


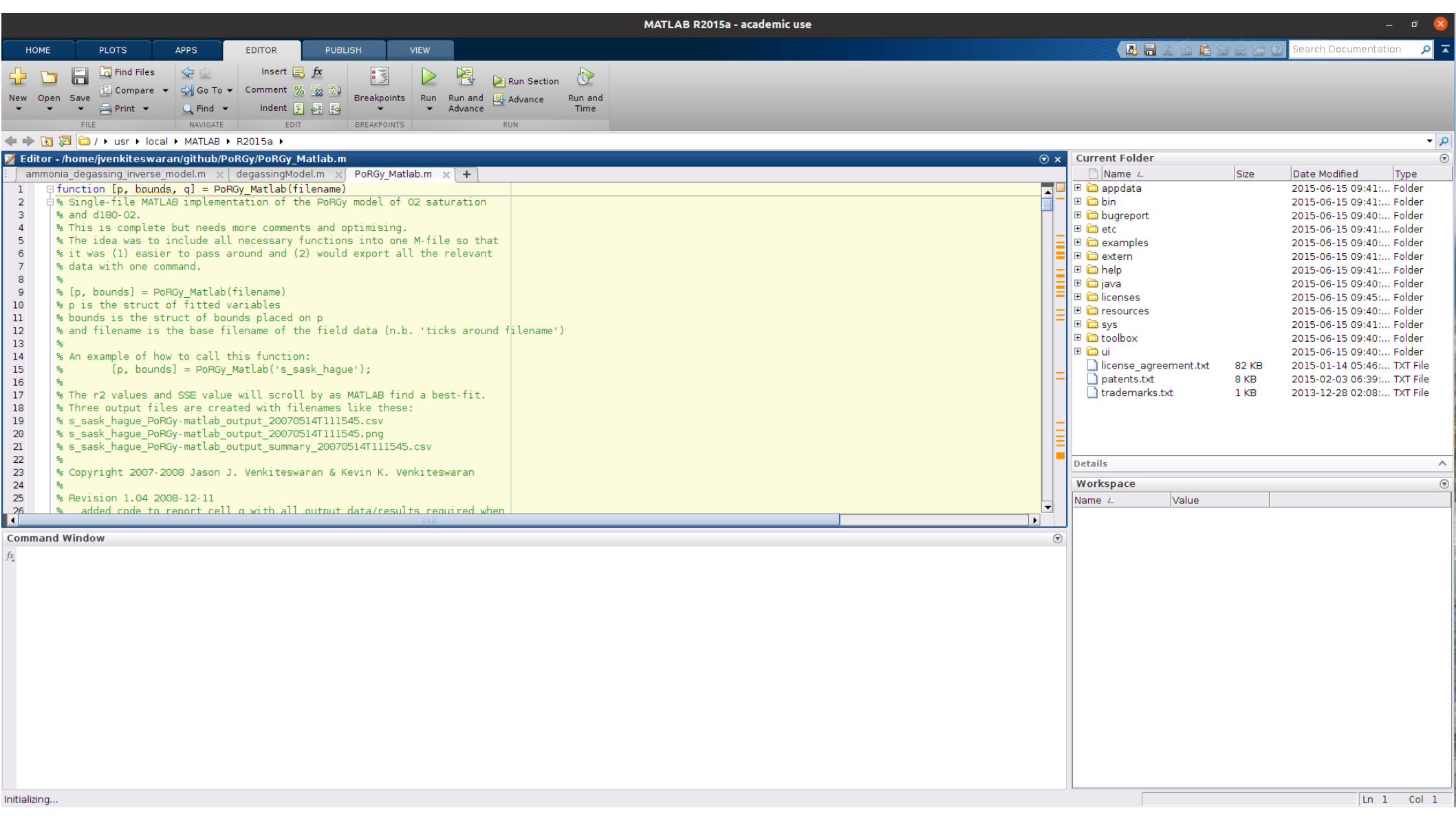

RStudio
Go to file/function
Addins
Project: (None)
Untitled1
Source on Save
Run
Source
1
1:1
(Top Level)
R Script
Console ~/
Copyright (C) 2016 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin13.4.0 (64-bit)
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome 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.
>
Environment
History
Import Dataset
List
Global Environment
Environment is empty
Files
Plots
Packages
Help
Viewer
New Folder
Delete
Rename
More
Home
Documents
SoftwareCarpentry
r_gapminder
Name
Size
Modified
..
r-novice-gapminder

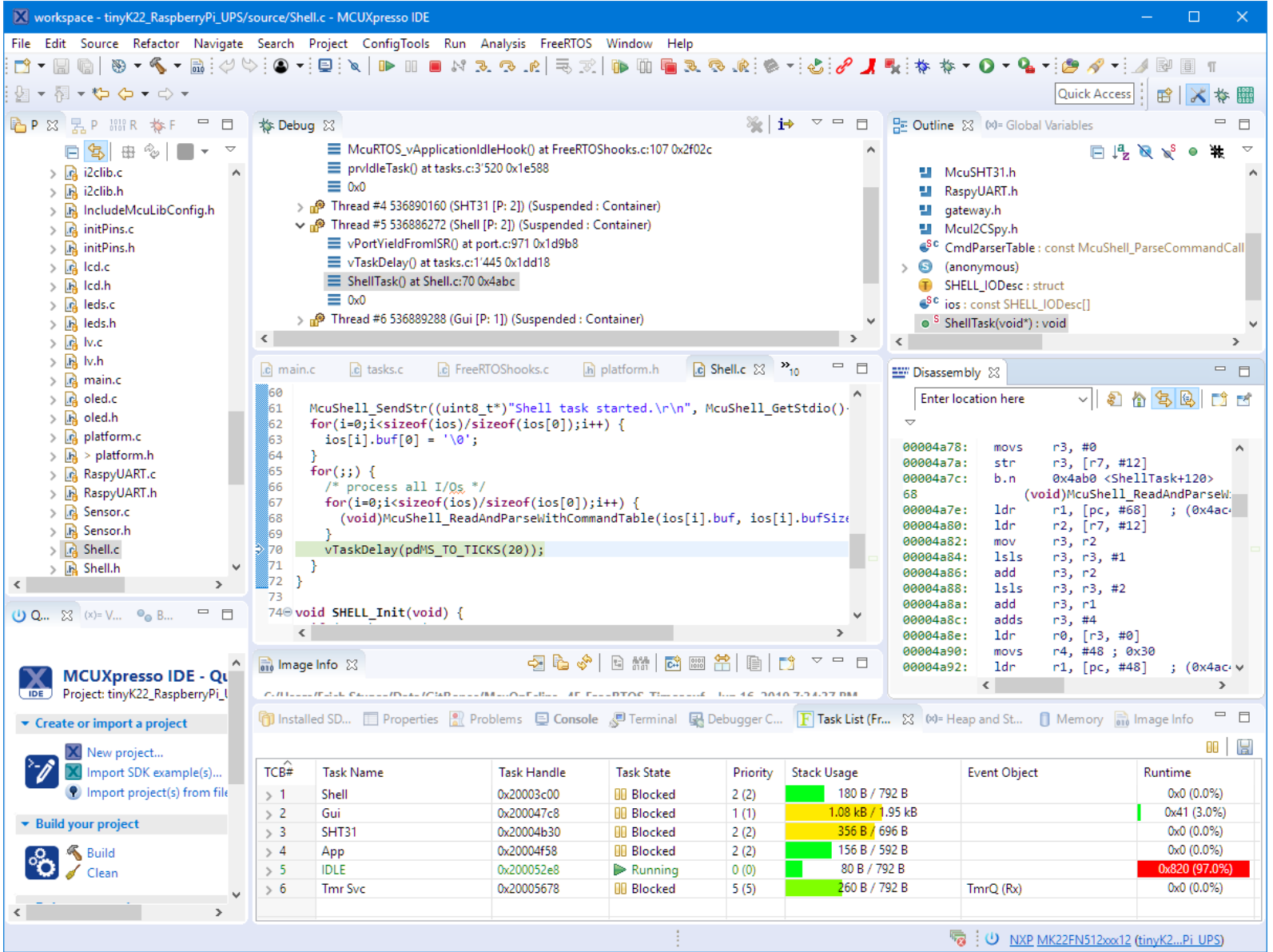
Why Use MATLAB for Data Science?

Exploratory Data Analysis

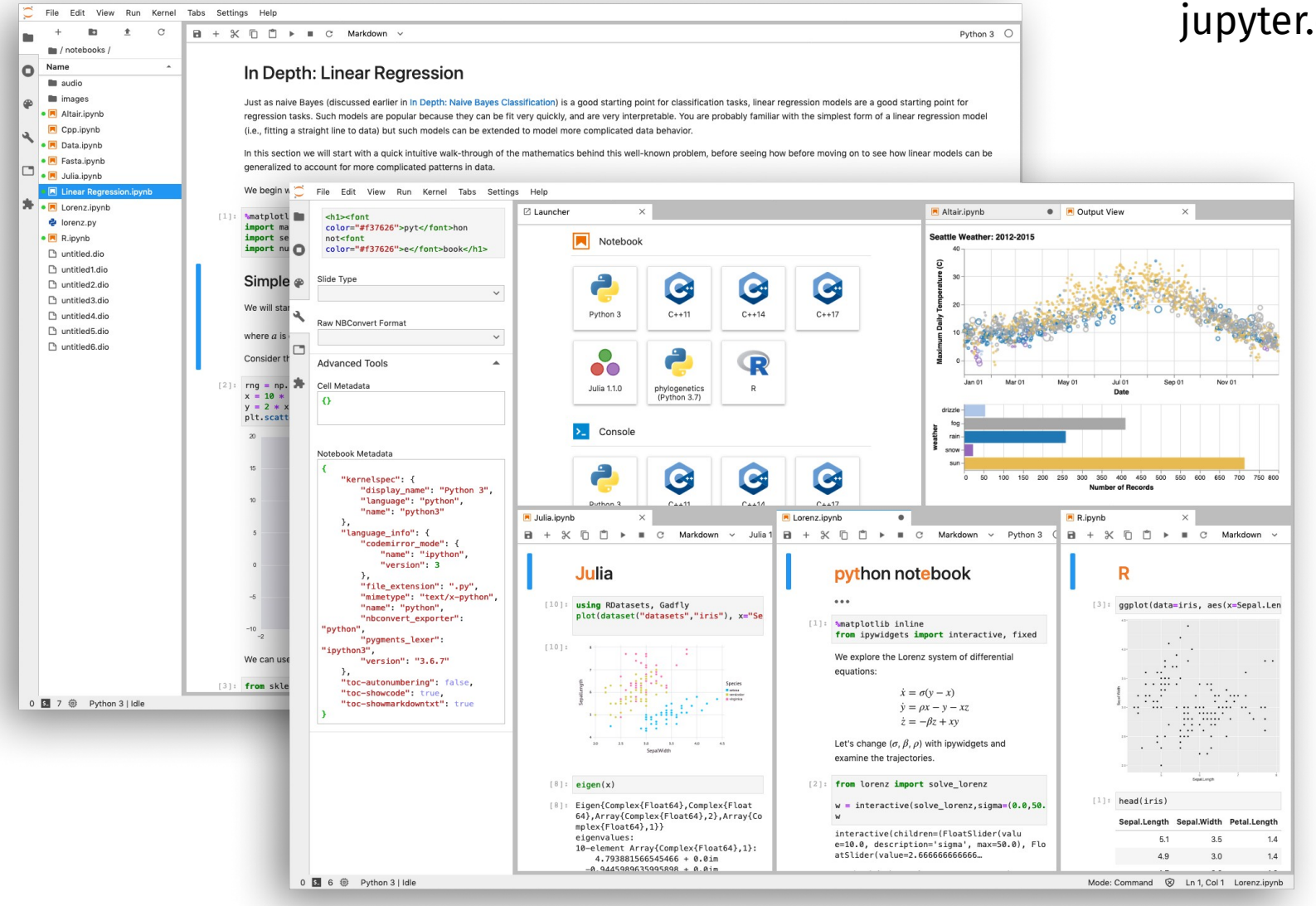
Spend less time preprocessing data. From time-series sensor data to images to text, MATLAB datatypes significantly reduce the time required to preprocess data. High-level functions make it easy to synchronize disparate time series, replace outliers with interpolated values, filter noisy signals, split raw text into words, and much more. Quickly visualize your data to understand trends and identify data quality issues with plots and the Live Editor.




MATLAB R2015a - academic use
HOME
PLOTS
APPS
EDITOR
PUBLISH
VIEW
Search Documentation
New
Open
Save
Find Files
Compare
Print
FILE
Go To
Find
NAVIGATE
Insert
Comment
Indent
EDIT
Breakpoints
BREAKPOINTS
Run
Run and Advance
Run Section
Advance
Run and Time
RUN
/ usr local MATLAB R2015a
Editor - /home/jvenkiteswaran/github/PoRGy/PoRGy_Matlab.m
ammonia_degassing_inverse_model.m
degassingModel.m
PoRGy_Matlab.m
function [p, bounds, q] = PoRGy_Matlab(filename)
% Single-file MATLAB implementation of the PoRGy model of O2 saturation
% and d18O-O2.
% This is complete but needs more comments and optimising.
% The idea was to include all necessary functions into one M-file so that
% it was (1) easier to pass around and (2) would export all the relevant
% data with one command.
%
% [p, bounds] = PoRGy_Matlab(filename)
% p is the struct of fitted variables
% bounds is the struct of bounds placed on p
% and filename is the base filename of the field data (n.b. 'ticks around filename')
%
% An example of how to call this function:
%   [p, bounds] = PoRGy_Matlab('s_sask_hague');
%
% The r2 values and SSE value will scroll by as MATLAB find a best-fit.
% Three output files are created with filenames like these:
% s_sask_hague_PoRGy-matlab_output_20070514T111545.csv
% s_sask_hague_PoRGy-matlab_output_20070514T111545.png
% s_sask_hague_PoRGy-matlab_output_summary_20070514T111545.csv
%
% Copyright 2007-2008 Jason J. Venkiteswaran & Kevin K. Venkiteswaran
%
% Revision 1.04 2008-12-11
%   added code to report cell q with all output data/results required when
Current Folder
Name
Size
Date Modified
Type
appdata 2015-06-15 09:41:... Folder
bin 2015-06-15 09:41:... Folder
bugreport 2015-06-15 09:40:... Folder
etc 2015-06-15 09:41:... Folder
examples 2015-06-15 09:40:... Folder
extern 2015-06-15 09:41:... Folder
help 2015-06-15 09:41:... Folder
java 2015-06-15 09:40:... Folder
licenses 2015-06-15 09:45:... Folder
resources 2015-06-15 09:40:... Folder
sys 2015-06-15 09:41:... Folder
toolbox 2015-06-15 09:40:... Folder
ui 2015-06-15 09:40:... Folder
license_agreement.txt 82 KB 2015-01-14 05:46:... TXT File
patents.txt 8 KB 2015-02-03 06:39:... TXT File
trademarks.txt 1 KB 2013-12-28 02:08:... TXT File
Details
Workspace
Name
Value
Command Window
Initializing...
Ln 1 Col 1


workspace - tinyK22_RaspberryPi_UPS/source/Shell.c - MCUXpresso IDE
File Edit Source Refactor Navigate Search Project ConfigTools Run Analysis FreeRTOS Window Help
Quick Access
i2clib.c
i2clib.h
IncludeMcuLibConfig.h
initPins.c
initPins.h
lcd.c
lcd.h
leds.c
leds.h
lv.c
lv.h
main.c
oled.c
oled.h
platform.c
platform.h
RaspyUART.c
RaspyUART.h
Sensor.c
Sensor.h
Shell.c
Shell.h
Debug
McuRTOS_vApplicationIdleHook() at FreeRTOShooks.c:107 0x2f02c
prvIdleTask() at tasks.c:3'520 0x1e588
0x0
Thread #4 536890160 (SHT31 [P: 2]) (Suspended : Container)
Thread #5 536886272 (Shell [P: 2]) (Suspended : Container)
vPortYieldFromISR() at port.c:971 0x1d9b8
vTaskDelay() at tasks.c:1'445 0x1dd18
ShellTask() at Shell.c:70 0x4abc
0x0
Thread #6 536889288 (Gui [P: 1]) (Suspended : Container)
Outline
Global Variables
McuSHT31.h
RaspyUART.h
gateway.h
Mcul2CSpy.h
CmdParserTable : const McuShell_ParseCommandCall
(anonymous)
SHELL_IODesc : struct
ios : const SHELL_IODesc[]
ShellTask(void*) : void
main.c
tasks.c
FreeRTOShooks.c
platform.h
Shell.c
60
61 McuShell_SendStr((uint8_t*)"Shell task started.\r\n", McuShell_GetStdio()
62 for(i=0;i<sizeof(ios)/sizeof(ios[0]);i++) {
63 ios[i].buf[0] = '\0';
64 }
65 for(;;) {
66 /* process all I/Os */
67 for(i=0;i<sizeof(ios)/sizeof(ios[0]);i++) {
68 (void)McuShell_ReadAndParseWithCommandTable(ios[i].buf, ios[i].bufSize
69 }
70 vTaskDelay(pdMS_TO_TICKS(20));
71 }
72 }
73
74 void SHELL_Init(void) {
Disassembly
Enter location here
00004a78: movs r3, #0
00004a7a: str r3, [r7, #12]
00004a7c: b.n 0x4ab0 <ShellTask+120>
68 (void)McuShell_ReadAndParseW
00004a7e: ldr r1, [pc, #68] ; (0x4ac
00004a80: ldr r2, [r7, #12]
00004a82: mov r3, r2
00004a84: lsls r3, r3, #1
00004a86: add r3, r2
00004a88: lsls r3, r3, #2
00004a8a: add r3, r1
00004a8c: adds r3, #4
00004a8e: ldr r0, [r3, #0]
00004a90: movs r4, #48 ; 0x30
00004a92: ldr r1, [pc, #48] ; (0x4ac
Image Info
MCUXpresso IDE - Qu
Project: tinyK22_RaspberryPi_
Create or import a project
New project...
Import SDK example(s)...
Import project(s) from file
Build your project
Build
Clean
Installed SD...
Properties
Problems
Console
Terminal
Debugger C...
Task List (Fr...
Heap and St...
Memory
Image Info
TCB#
Task Name
Task Handle
Task State
Priority
Stack Usage
Event Object
Runtime
1 Shell 0x20003c00 Blocked 2 (2) 180 B / 792 B 0x0 (0.0%)
2 Gui 0x200047c8 Blocked 1 (1) 1.08 kB / 1.95 kB 0x41 (3.0%)
3 SHT31 0x20004b30 Blocked 2 (2) 356 B / 696 B 0x0 (0.0%)
4 App 0x20004f58 Blocked 2 (2) 156 B / 592 B 0x0 (0.0%)
5 IDLE 0x200052e8 Running 0 (0) 80 B / 792 B 0x820 (97.0%)
6 Tmr Svc 0x20005678 Blocked 5 (5) 260 B / 792 B TmrQ (Rx) 0x0 (0.0%)
NXP MK22FN512xxx12 (tinyK2...Pi_UPS)

jupyter.org/try



R

- S (John Chambers et al.) at Bell Labs, 1976
  - Implemented as fortran libs
- Rewritten in C, 1988 as v3, v4 in 1998
- R, 1993 (Ross Ihaka & Robert Gentleman)
- 1995 GNU GPL, paper 1996
- ‘packages’ to expand capabilities

R + ‘tidyverse’

- Hadley Wickham PhD thesis 2008
- ‘ggplot’ 2005/7, Grammar of Graphics (Leland Wilkinson)
- ‘plyr’ 2008, tools ‘split-apply-combine’
- ‘tidy data’
- *opinionated*

Journal of Statistical Software

MMMMMM YYYY, Volume VV, Issue II. *http://www.jstatsoft.org/*

Tidy Data

Hadley Wickham
RStudio

Abstract

A huge amount of effort is spent cleaning data to get it ready for analysis, but there has been little research on how to make data cleaning as easy and effective as possible. This paper tackles a small, but important, component of data cleaning: data tidying. Tidy datasets are easy to manipulate, model and visualise, and have a specific structure: each variable is a column, each observation is a row, and each type of observational unit is a table. This framework makes it easy to tidy messy datasets because only a small set of tools are needed to deal with a wide range of un-tidy datasets. This structure also makes it easier to develop tidy tools for data analysis, tools that both input and output tidy datasets. The advantages of a consistent data structure and matching tools are demonstrated with a case study free from mundane data manipulation chores.

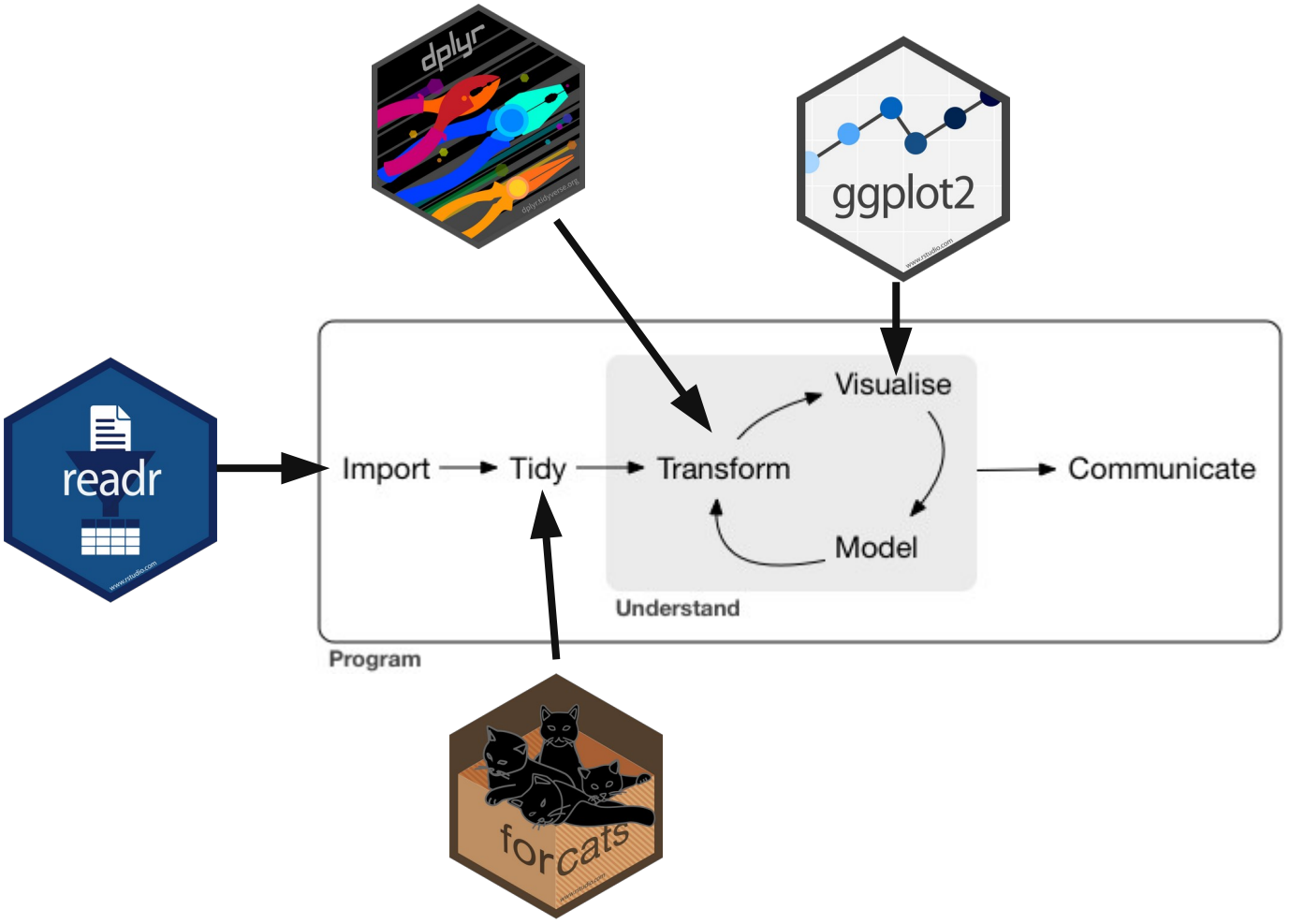
Tidy vs Messy Data

- each variable is a column
- each observation (or case) is a row
- *Give an example where you didn't follow this in the past*

variables

observations

values


dplyr
ggplot2
readr
Import
Tidy
Transform
Visualise
Model
Communicate
Understand
Program
forcats


tidyverse
www.rstudio.com

tidyverse.org

Break?

Putting Little Things Together

- RStudio IDE
- Project Management
- R Intro

RStudio IDE

- Default 4 pane layout

RStudio Project

- Click the “File” menu button, then “New Project”.
- Click “New Directory”.
- Click “Empty Project”.
- Type in the name of the directory to store your project, e.g. “my_project”.
- *If available, select the checkbox for “Create a git repository.”*
- Click the “Create Project” button.

R Intro 1

- Calculator
- Functions
- Comparisons
- Assignment
- *R-Scripts*

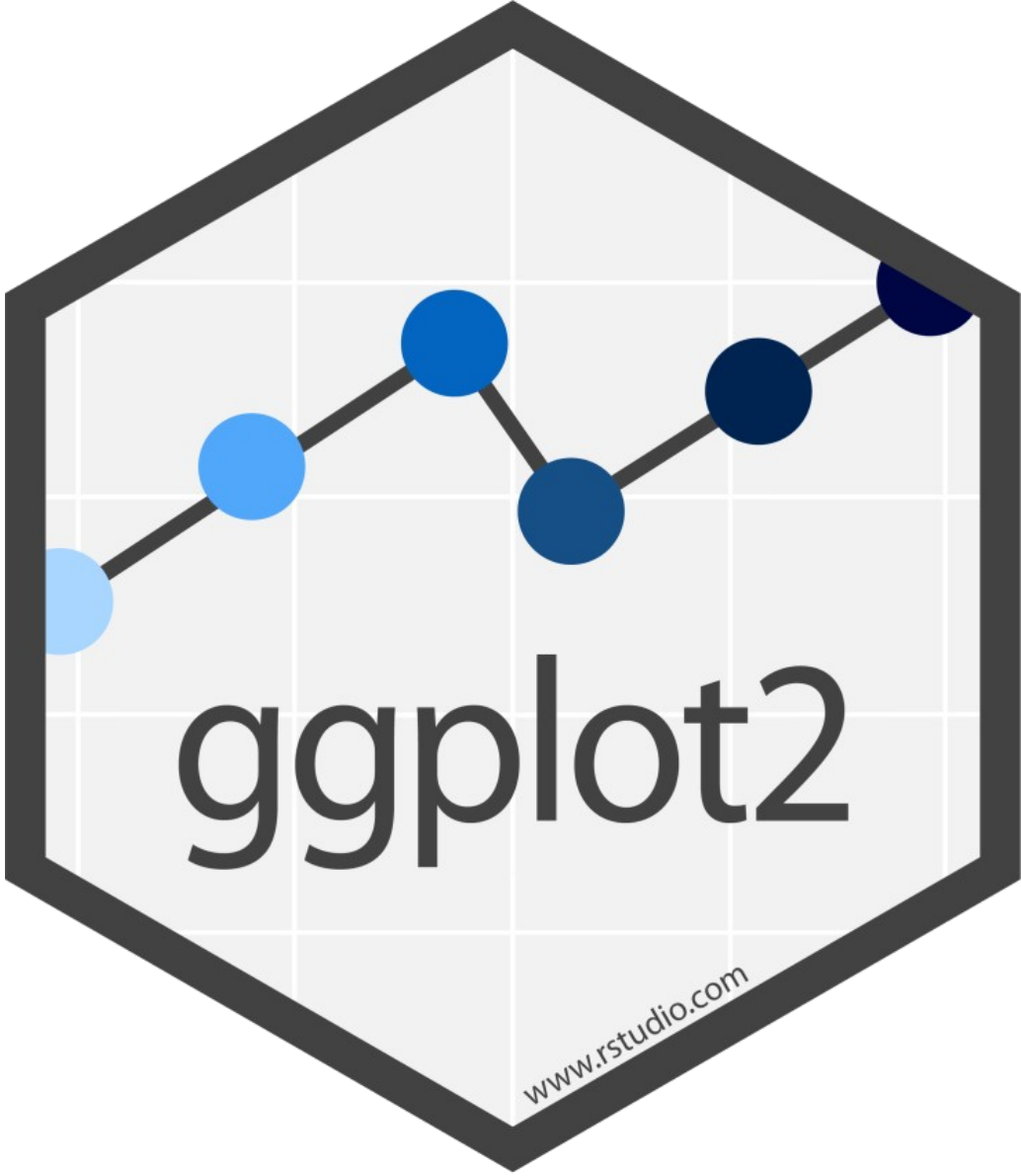
- `1+2`
- `sin(1)`
- `1 == 1`
- `x <- 3`
- File, New File, R Script

R Intro 2

- Write commands in an R Script and save it
- periods.between.words
- underscores_between_words
- camelCaseForWords
- *Do you want to create a variable if that name is already in use?*

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

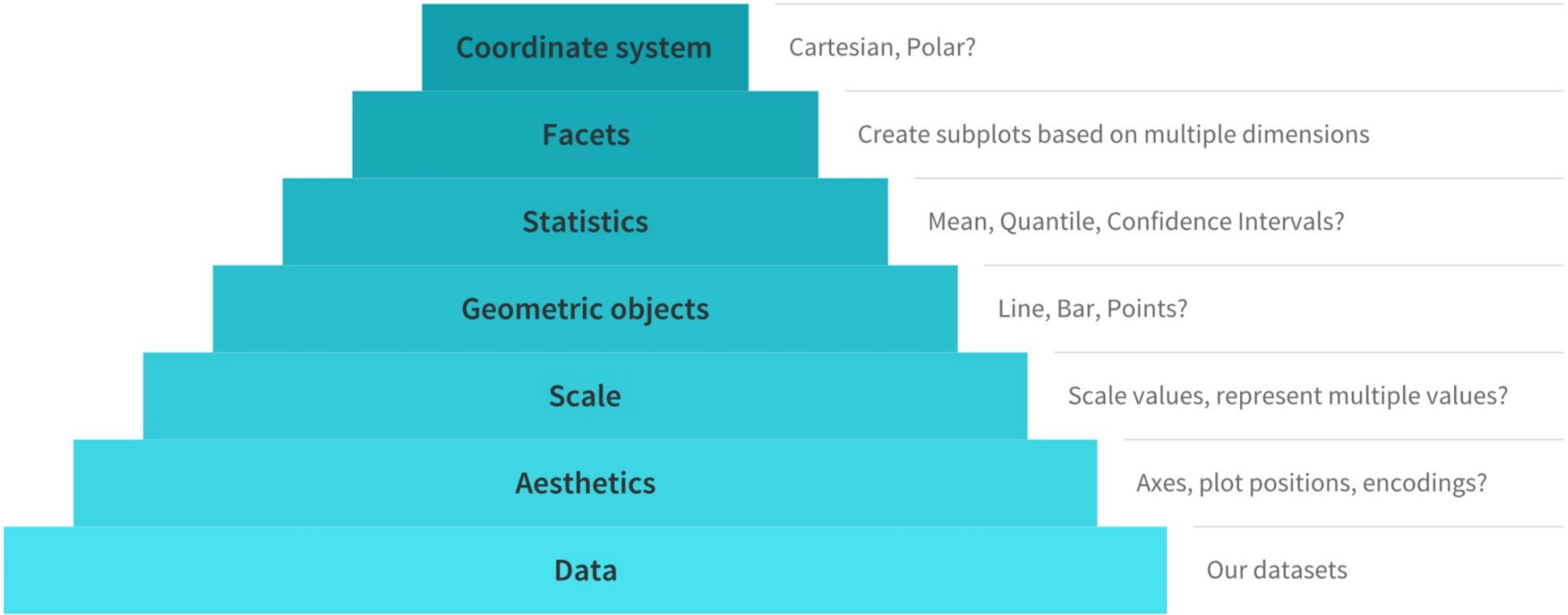

readr
www.rstudio.com


ggplot2
www.rstudio.com

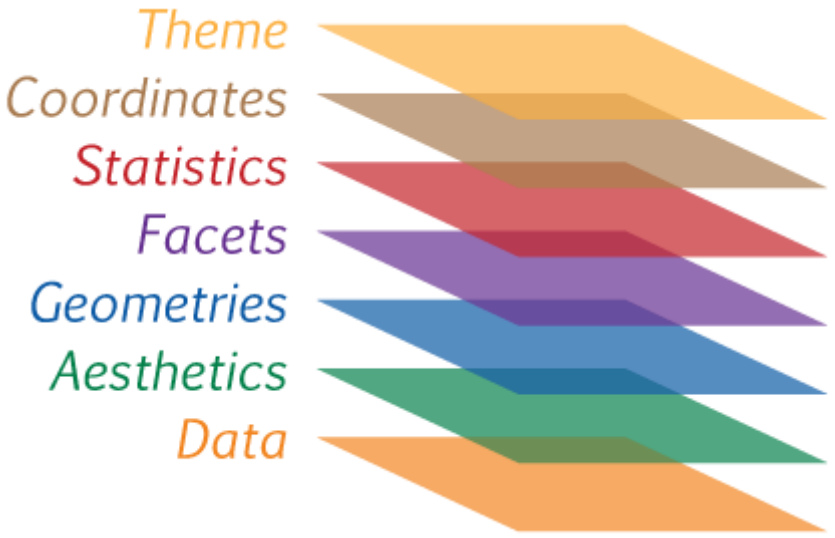
ggplot2

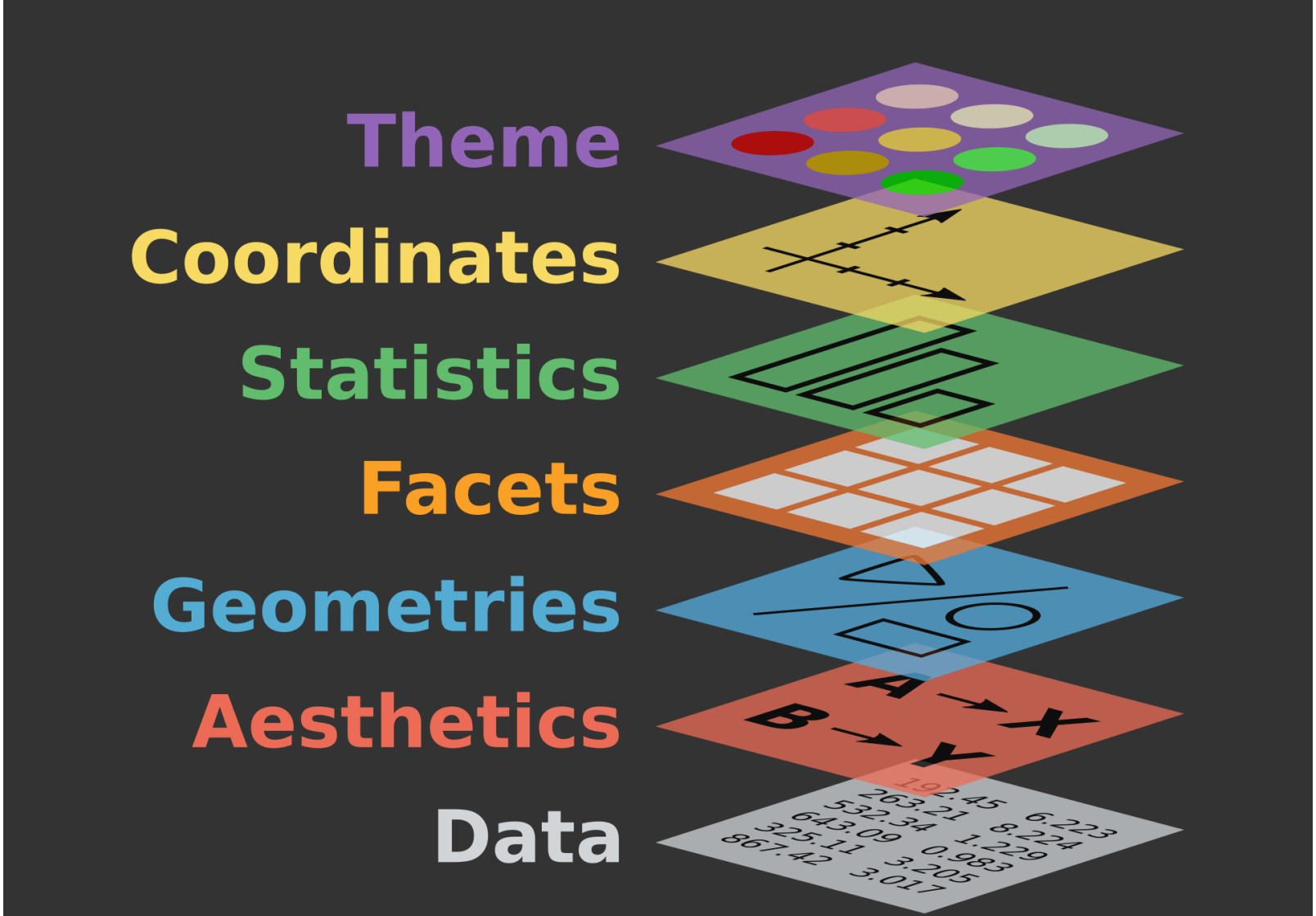
- *grammar is about structure*
  - `data`, variables mapped to aesthetics
  - `layer` (geom elements, stat transforms)
  - `mapping` (aesthetics)
  - `scale` (data space to aes space, colour, size, shape)
  - `coord` (Cartesian, polar, map)
  - `facet` (small multiples, latticing, trellising)
  - `theme` (details)
- *What does this not do?*

ggplot2-book.org



The major elements of the grammar of graphics. Source: Towards Data Science


Theme
Coordinates
Statistics
Facets
Geometries
Aesthetics
Data


Theme
Coordinates
Statistics
Facets
Geometries
Aesthetics
Data
A → X
B → Y
192.45 6.223
263.21 8.224
532.34 1.229
643.09 0.983
325.11 3.205
867.42 3.017

Start

- `library(tidyverse)`
- `mpg`
- `ggplot(mpg) + geom_point(aes(x = displ, y = hwy))`
- *Discuss*

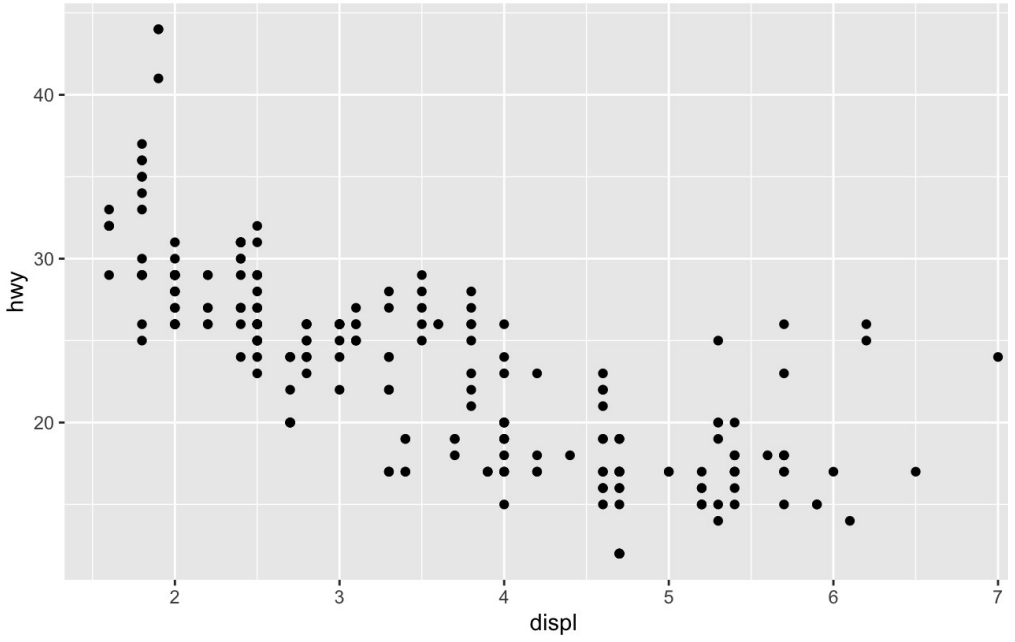
Start

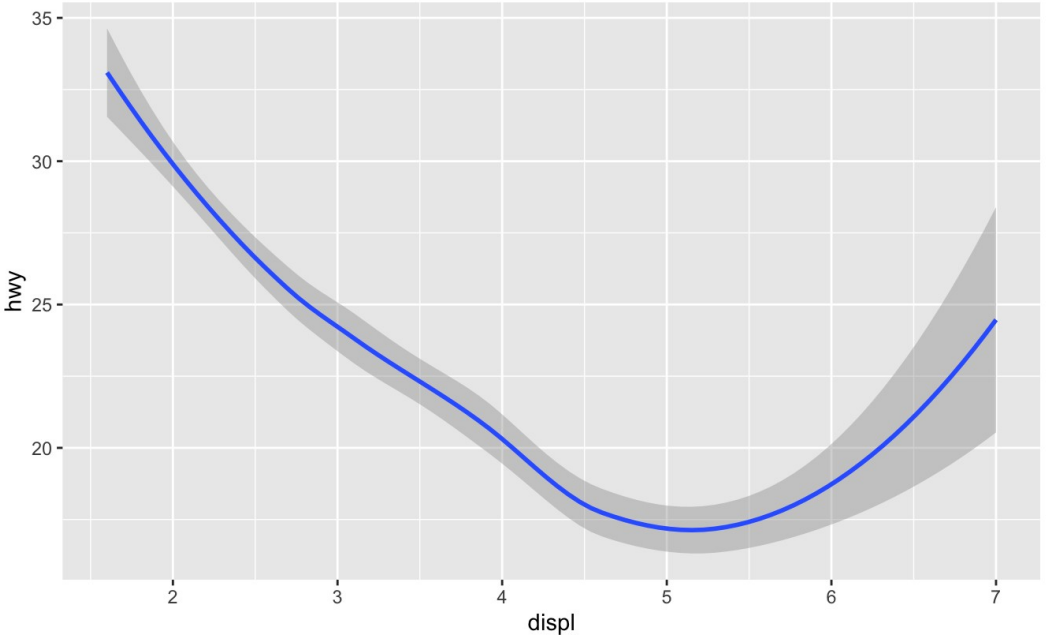
- `library(tidyverse)`
- `mpg`
- `ggplot(mpg) + geom_point(aes(x = displ, y = hwy))`
- *Discuss*
- `colour, size, shape, alpha`
- Inside vs outside the `aes()`

Facets

- Splitting plot by a variable
- ```
  ggplot(data = mpg) +
    geom_point(mapping = aes(x = displ, y = hwy)) +
    facet_wrap(~ class, nrow = 2)
  ```
- ```
  ggplot(data = mpg) +
    geom_point(mapping = aes(x = displ, y = hwy)) +
    facet_grid(drv ~ cyl)
  ```
- *Discuss*

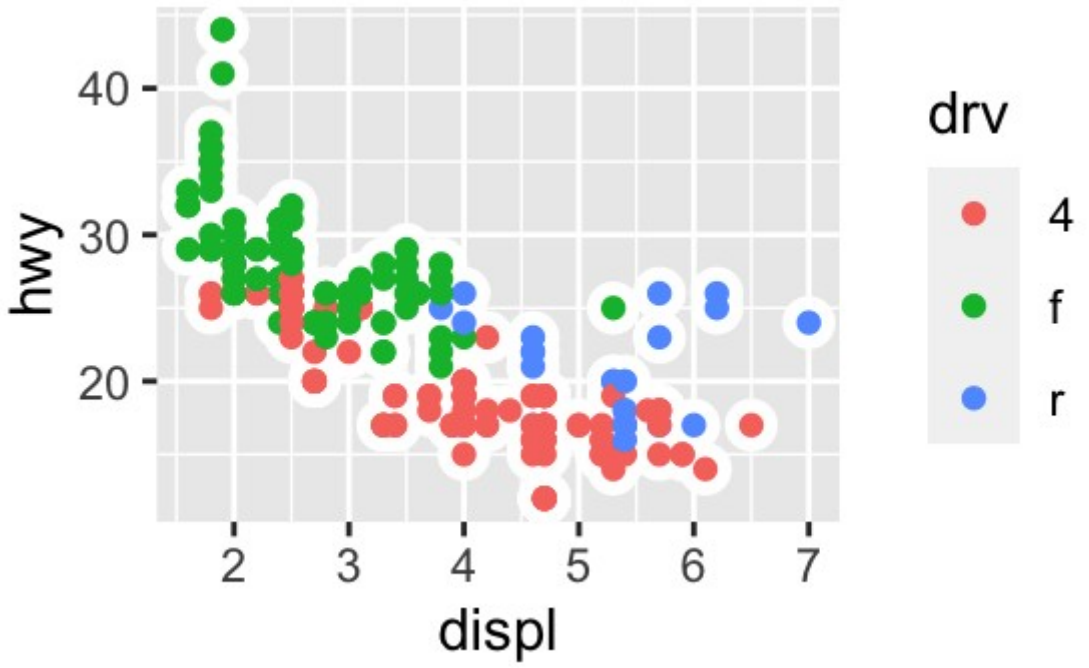
Other geom





Other geoms

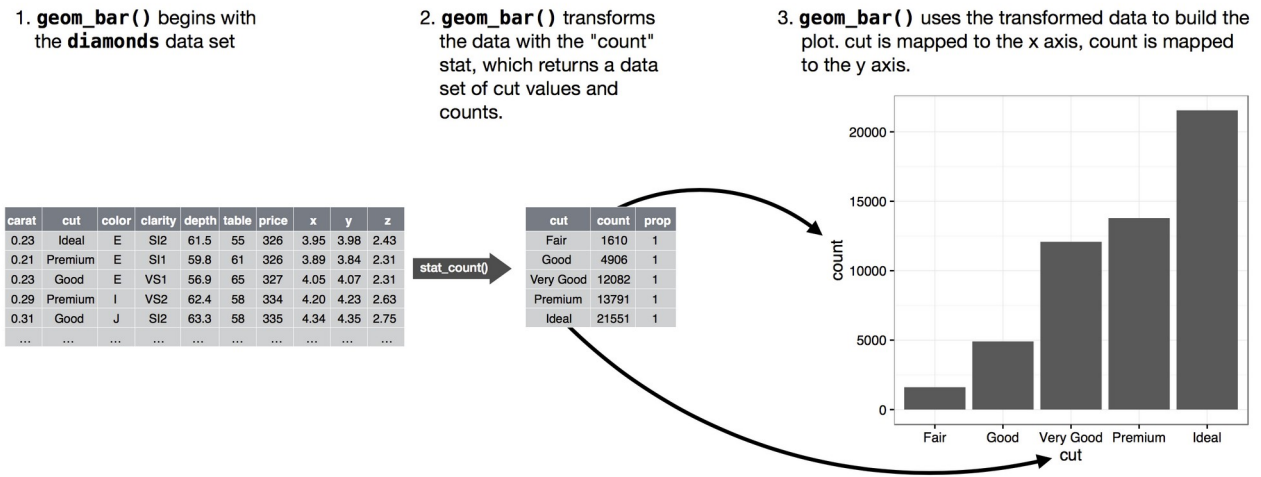
- ```
  ggplot(data = mpg) +
    geom_point(mapping = aes(x = displ, y = hwy)) +
    geom_smooth(mapping = aes(x = displ, y = hwy))
  ```
- ```
  ggplot(data = mpg, mapping = aes(x = displ, y = hwy)) +
    geom_point(mapping = aes(colour = class)) +
    geom_smooth()
  ```
- *Discuss*


hwy
40
30
20
2
3
4
5
6
7
displ
drv
4
f
r

Break?

Stats

- `ggplot(diamonds) + geom_bar(aes(cut))`



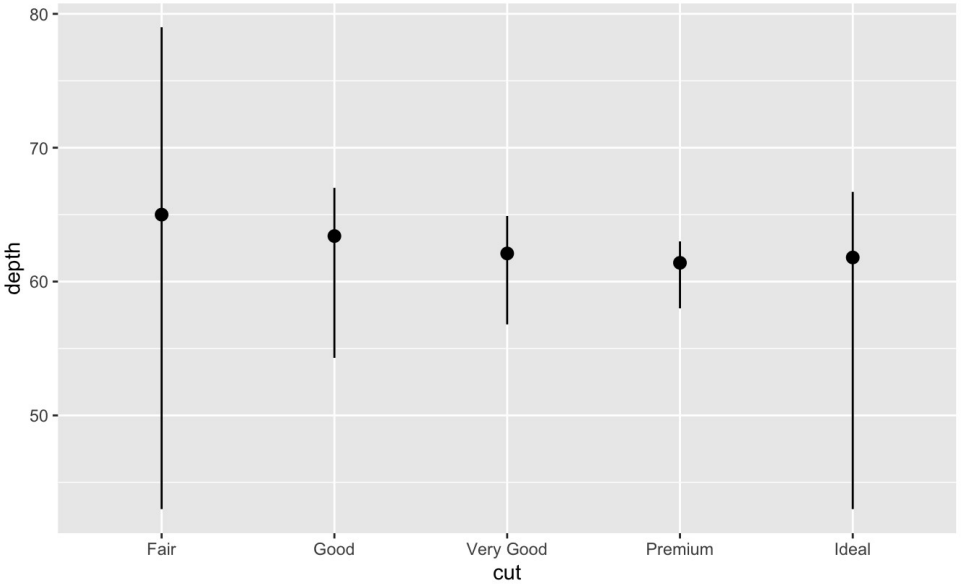
carat	cut	color	clarity	depth	table	price	x	y	z
0.23	Ideal	E	SI2	61.5	55	326	3.95	3.98	2.43
0.21	Premium	E	SI1	59.8	61	326	3.89	3.84	2.31
0.23	Good	E	VS1	56.9	65	327	4.05	4.07	2.31
0.29	Premium	I	VS2	62.4	58	334	4.20	4.23	2.63
0.31	Good	J	SI2	63.3	58	335	4.34	4.35	2.75
...	...	...	...	...	...	...	...	...	...

cut	count	prop
Fair	1610	1
Good	4906	1
Very Good	12082	1
Premium	13791	1
Ideal	21551	1

`?geom_bar`

Stats

```
ggplot(data = diamonds) +
  stat_summary(
    mapping = aes(x = cut, y = depth),
    fun.min = min,
    fun.max = max,
    fun = median
  )
```



Position

- Colour vs fill
- ```
  ggplot(data = diamonds) +
    geom_bar(aes(cut, X = clarity))
  ```
  ```

Position

- Colour vs fill
- `ggplot(data = diamonds) +`
  `geom_bar(aes(cut, x = clarity))`
- `identity`
  special for bars, stacked, overlapping, proportion, side-by-side

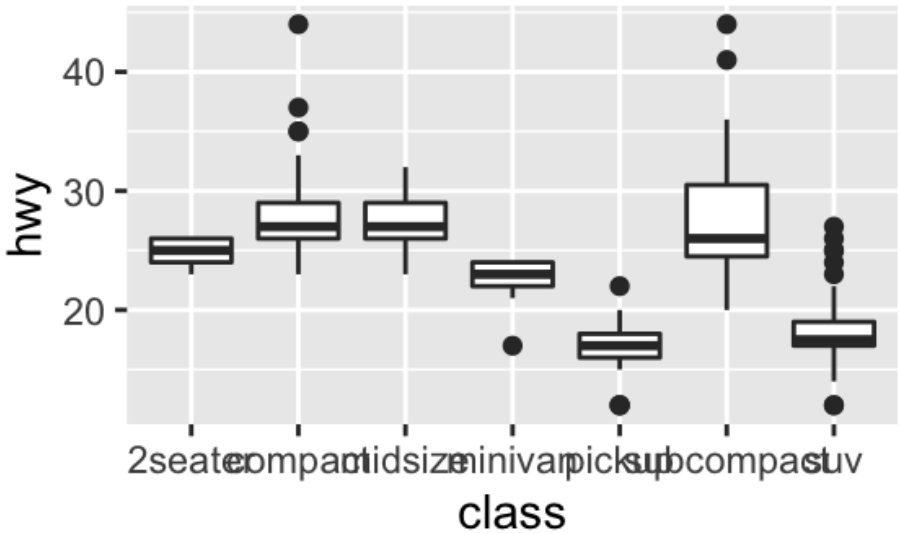
Position

- `ggplot(mpg, aes(x = cty, y = hwy)) + geom_point()`
- 234 rows in `mpg`, where are they?
- `position = "jitter"`
  or
  `geom_jitter()`
- *Discuss how to improve figure*

`?position_`

Coords

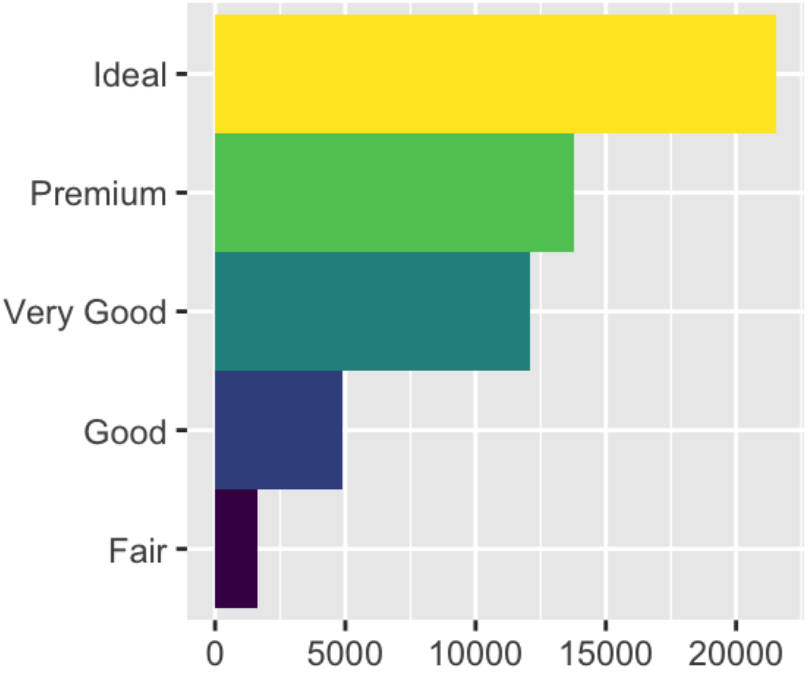
- *Which coordinate system?*
- ```
  ggplot(mpg, aes(class, hwy)) +
     geom_boxplot() +
     coord_flip()
  ```
- *What will this do? Useful?*


  ```

Coords

- When would other coordinate systems be useful?
- `?coord_polar`

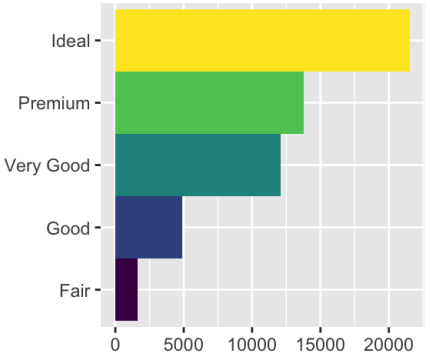
Coords


Ideal
Premium
Very Good
Good
Fair
0
5000
10000
15000
20000

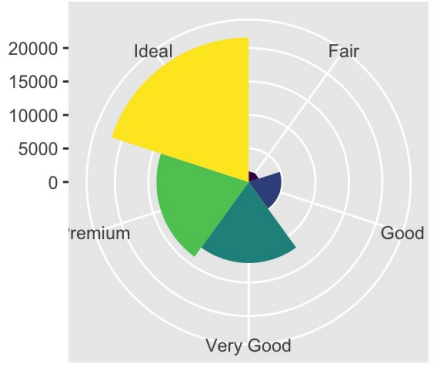

20000
15000
10000
5000
0
Ideal
Fair
remium
Good
Very Good

Coords

- ```
  bar ← ggplot(diamonds) +
    geom_bar(aes(x = cut, fill = cut),
      show.legend = FALSE,
      width = 1) +
    theme(aspect.ratio = 1) +
    labs(x = NULL, y = NULL)
  ```




- ```
  bar + coord_flip()
  ```
- ```
  bar + coord_polar()
  ```


  ```

Summary

- ggplot(data = <DATA>) +
  <GEOM_FUNCTION>(
    mapping = aes(<MAPPINGS>),
    stat = <STAT>,
    position = <POSITION>
-) +
  <COORDINATE_FUNCTION> +
  <FACET_FUNCTION>

Homework

- `install.packages("palmerpenguins")`

```
> library(palmerpenguins)
> penguins
# A tibble: 344 x 8
   species island bill_length_mm bill_depth_mm flipper_length_… body_mass_g
   <fct>   <fct>           <dbl>         <dbl>             <int>       <int>
 1 Adelie  Torge…           39.1          18.7               181        3750
 2 Adelie  Torge…           39.5          17.4               186        3800
 3 Adelie  Torge…           40.3          18                 195        3250
 4 Adelie  Torge…           NA            NA                  NA          NA
 5 Adelie  Torge…           36.7          19.3               193        3450
 6 Adelie  Torge…           39.3          20.6               190        3650
 7 Adelie  Torge…           38.9          17.8               181        3625
 8 Adelie  Torge…           39.2          19.6               195        4675
 9 Adelie  Torge…           34.1          18.1               193        3475
10 Adelie  Torge…           42            20.2               190        4250
# … with 334 more rows, and 2 more variables: sex <fct>, year <int>
> 
```

Homework

- `install.packages("palmerpenguins")`
- Make some interesting plots with new `geom_` and `theme_`
- ggplot2.tidyverse.org
- Try out `ggsave`
- pangaea.de & www.frdr-dfdr.ca
- allisonhorst.github.io/palmerpenguins/