Better Analytics with R and RStudio

•••

Mar 1, 2020

Overview

1. Why use R (vs Excel)?
2. What is R?
3. What is RStudio?
4. Data processing tutorial + practice
5. Plotting intro + practice
6. Advanced plotting examples
7. Shiny demo (time permitting)

Challenges with Using Excel

Readability

Interpreting someone else’s Excel files can be a challenge.

Particularly when they contain complex functions that are difficult to reverse engineer.

Error Prone

Formula and calculation errors can be hard to identify.

Version Control

Version control features and lacking or under utilized.

Notes to yourself or other users is limited as well.

Figuring Out Someone’s Excel Sheet

WHAT ARE THE ODDS THAT YOU MADE THIS COMPLICATED SPREAD-SHEET WITHOUT ANY CRITICAL ERRORS?
Dilbert.com @ScottAdamsSays
DOES IT MATTER, AS LONG AS IT GIVES ME THE ANSWER I WANT?
1-7-16 © 2016 Scott Adams, Inc. /Dist. by Universal Uclick
IT SHOULD.
BUT ASK YOURSELF IF IT DOES.

Excel Functions are Very Hard to Parse

Who has seen an abomination like this thing?

```
=IF($F$10=”Y”,(IF($C$9=0,”$0.00?,IF($C$9<113,’Revenues & Costs’!$B$3,IF($C$9<281,
(‘Revenues & Costs’!$C$3),IF($C$9<451,($C$9*’Revenues & Costs’!$D$3),IF($C$9<851,
($C$9*’Revenues & Costs’!$E$3),IF($C$9<1201,($C$9*’Revenues & Costs’!$F$3),
IF($C$9>1201,($C$9*’Revenues & Costs’!$G$3)))))))))*2, IF($C$9=0,”$0.00?,
IF($C$9<113,’Revenues & Costs’!$B$3,IF($C$9<281,(‘Revenues & Costs’!$C$3),IF($C$9<451,
($C$9*’Revenues & Costs’!$D$3),IF($C$9<851,($C$9*’Revenues & Costs’!$E$3),IF($C$9<1201,
($C$9*’Revenues & Costs’!$F$3),IF($C$9>1201,($C$9*’Revenues & Costs’!$G$3)))))))))+
IF($C$11=1,0,($C$11*10)*$C$12*’Revenues & Costs’!$B$5)+($F$7*’Revenues & Costs’!$B$6)
```

Mistakes Can be Easy to Miss, and Expensive

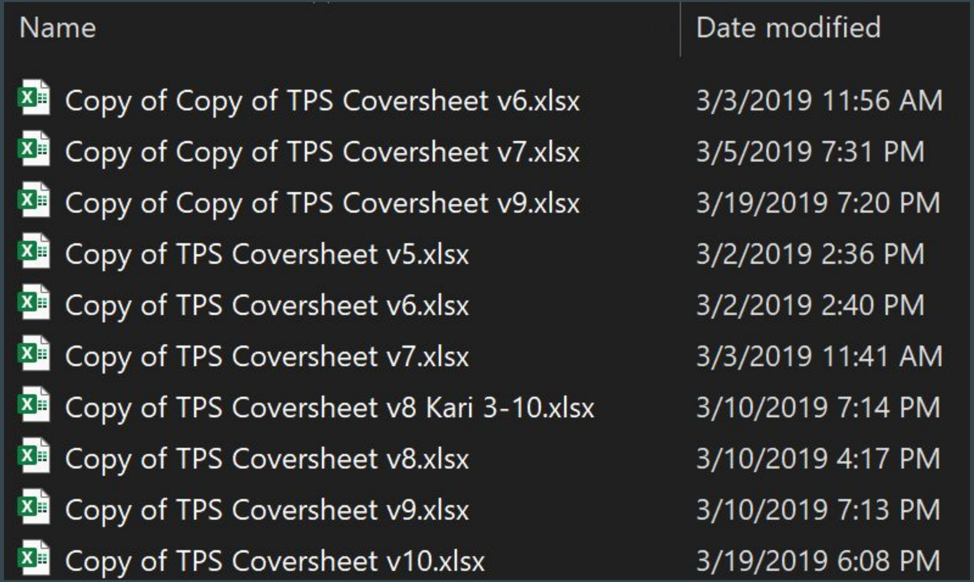
- Fidelity investments missed a minus sign and made a $2.6B calculation error, forcing the cancelation of dividend payments to customers.
- The 2012 London Olympics oversold thousands of tickets due to an single cell being off.
- Barclay's used an excel sheet as a list of contracts to buy, but hid rather than deleted the ones they didn't want to buy. The court ruled they had to pay for them anyway.
- JP Morgan was hit with $6.5B in losses and fines when a copy/paste error underestimated the downside of a synthetic portfolio.
- AstraZeneca mistakenly released highly confidential finance information to external analysts due to template error.
- Harvard economists produced pro-austerity research with significant excel calculation errors that affected budget decisions for several countries from 2010-2012.

Excel’s Version Control is Lacking

Version control depends on external tools.

Email chains, and file_ver_2.xlsx don’t make for reliable version controls.

No consistent or standard way to leave notes or guidance to future users (yourself included).



Name	Date modified
Copy of Copy of TPS Coversheet v6.xlsx	3/3/2019 11:56 AM
Copy of Copy of TPS Coversheet v7.xlsx	3/5/2019 7:31 PM
Copy of Copy of TPS Coversheet v9.xlsx	3/19/2019 7:20 PM
Copy of TPS Coversheet v5.xlsx	3/2/2019 2:36 PM
Copy of TPS Coversheet v6.xlsx	3/2/2019 2:40 PM
Copy of TPS Coversheet v7.xlsx	3/3/2019 11:41 AM
Copy of TPS Coversheet v8 Kari 3-10.xlsx	3/10/2019 7:14 PM
Copy of TPS Coversheet v8.xlsx	3/10/2019 4:17 PM
Copy of TPS Coversheet v9.xlsx	3/10/2019 7:13 PM
Copy of TPS Coversheet v10.xlsx	3/19/2019 6:08 PM

Opening a Spreadsheet from Last Year

There must be a better way!

What is R? What is RStudio?

R is:

Free, open source software

Built specifically with statistics and data analytics in mind

The most robust data visualization solution

Thousands of packages to help with general to weirdly specific tasks

Strong community of users

RStudio is:

Front end user interface for R

Easily import data from a variety of sources

Built in help tools

Frameworks for reporting, sharing and visualizing data and analytics

Free for individuals with enterprise support options

But I can’t write code!

Actually, You Already Do

Function	Excel Formula	R Code
Column Sum	=SUM(column)	sum(dataframe$column)
If Statement	=IF(test, true, false)	ifelse(test, true, false)
Min/Max	=MIN(data)/=MAX(data)	min(data)/max(data)
Count	=COUNT(data)	count(data)
Combine Tables	=VLOOKUP(table, source, target, exact)	left_join(table1, table2, key)

Any Excel formula will have an equivalent in R, but the reverse is not the case.

The “Tidy Data” Structure

R uses what is called “Tidy Data”

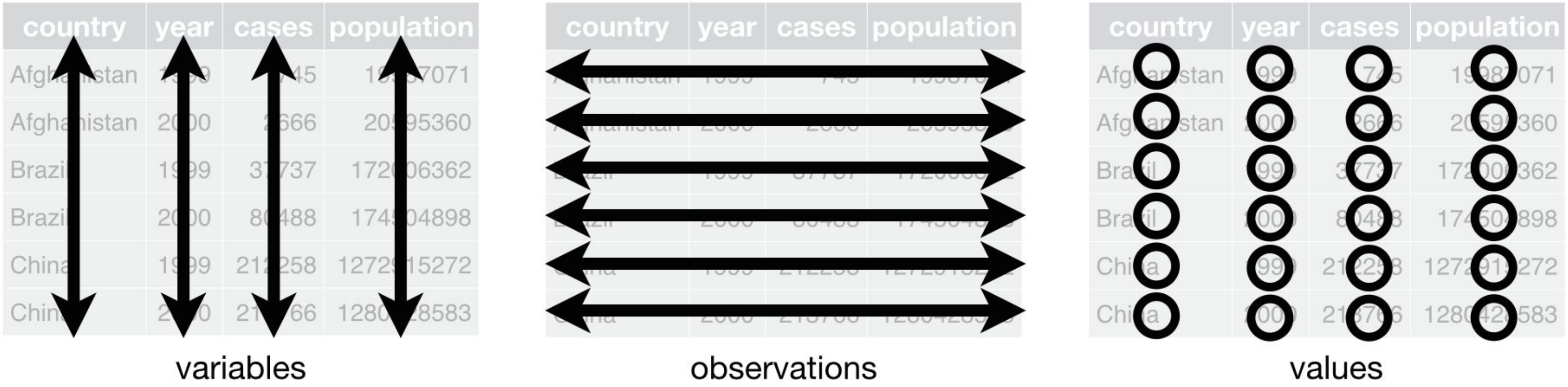
A common data structure is required

Universal and easily shareable

Standardized data formatting (for everything other than Excel)

Requirements:

1. Each variable must have its own column.
2. Each observation must have its own row.
3. Each value must have its own cell.



country	year	cases	population
Afghanistan	1999	745	19987071
Afghanistan	2000	2666	20595360
Brazil	1999	37737	172006362
Brazil	2000	80488	174504898
China	1999	212258	1272915272
China	2000	213766	1280428583

Tidy Data

R 101

Basic R Practice Problems

1. Load the video game sales data into R.
2. Look at the head and tail of the dataframe.
3. What is the mean sales in North America? Which game did the best and the worst?
4. What are the different platforms in this dataset?
5. How many unique publishers are there?

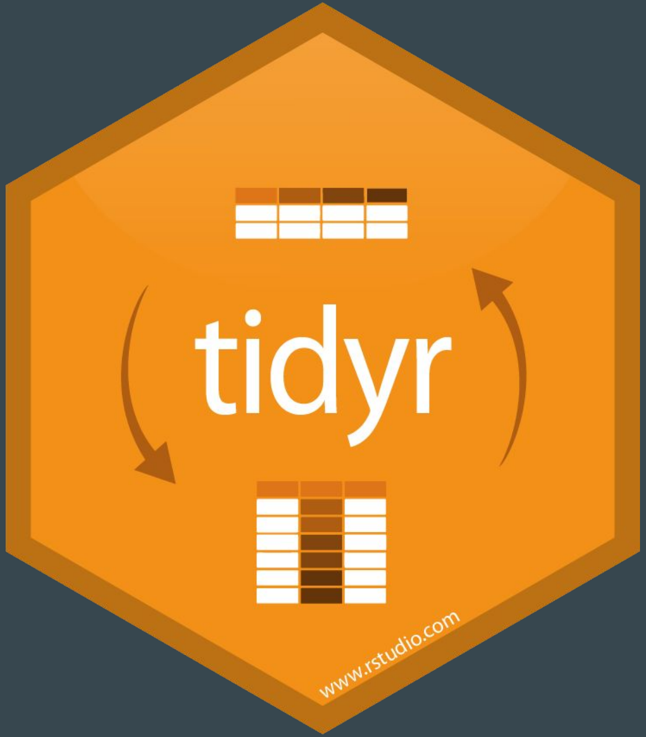
The Tidyverse

- Group of R packages for data science
- Developed by Hadley Wickam and RStudio
- Very clear, concise syntax

1. Import
2. Clean
3. Manipulate
4. Model
5. Visualize
6. Report

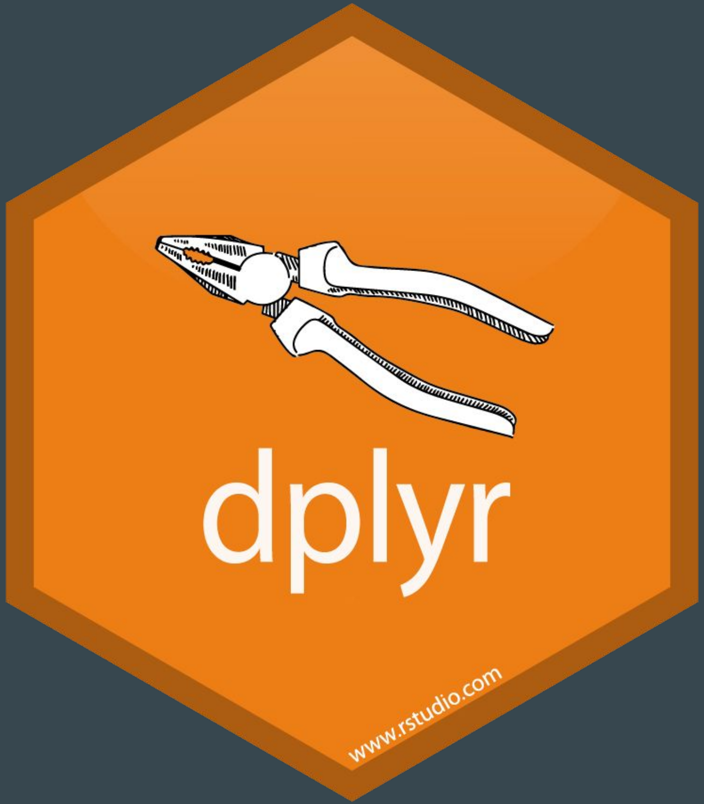
tidyr - The Data Cleanup Maestro

1. Part of the tidyverse
2. Clean up messy data
3. Spread out or gather up columns
4. Handle missing data
5. Split cells into multiple columns or rows



tidyr demo

dplyr - The Data Manipulation Workhorse



1. Part of the tidyverse
2. Group, aggregate and cluster
3. Filter, slice or sample
4. Summarize
5. Create new variables
6. Join or append multiple datasets

dplyr - The Pipe Operator aka %>%

1. Improves overall readability and usability
2. Automatically loaded with tidyr or dplyr
3. Executes functions sequentially
4. Output of a function is the input of the next function
5. Eliminates nested functions

Nested:

```
funC(funB(funA(data)))
```

Piped:

```
data %>%
   funA() %>%
   funB() %>%
   funC()
```

dplyr example

>filter() selects a portion of the dataset

>group_by() groups the data

>summarize() performs operations on grouped data

>mutate() creates new columns

```
sales_data %>%
    filter(Date >= “2020-01-01”) %>%
    group_by(GSC, Region) %>%
    summarize(Tot_Sales = sum(Sales),
              Tot_Quant = sum(Quantity)) %>%
    mutate(Net_Price = Tot_Sales / Tot_Quant)
```

dplyr demo

dplyr and tidyr practice

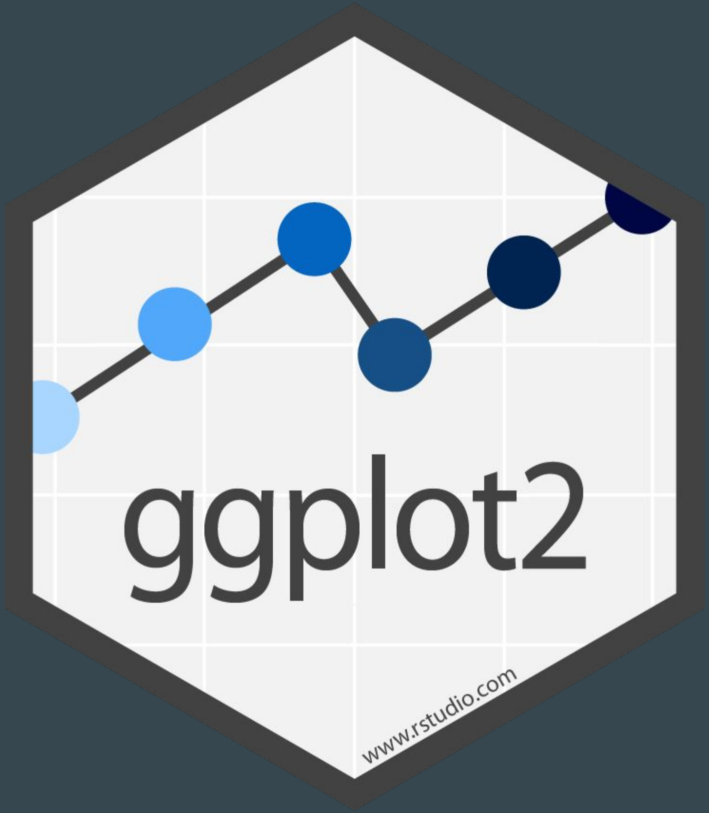
1. What are total sales by Year? By Year and Region?
2. Which Genre has the highest sales?
3. Since 2010, which Publisher has the highest sales?
4. Which Publisher has been the most successful in each Region?

Challenge:
Find the games that were released on more than one platform. On how many platforms was it released? On which did it sell the best?

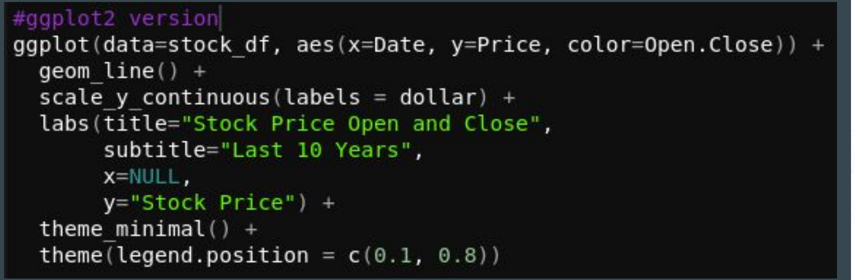
Hint: look into arrange(), slice() and n_distinct() functions

ggplot2 - ‘Grammar of Graphics’

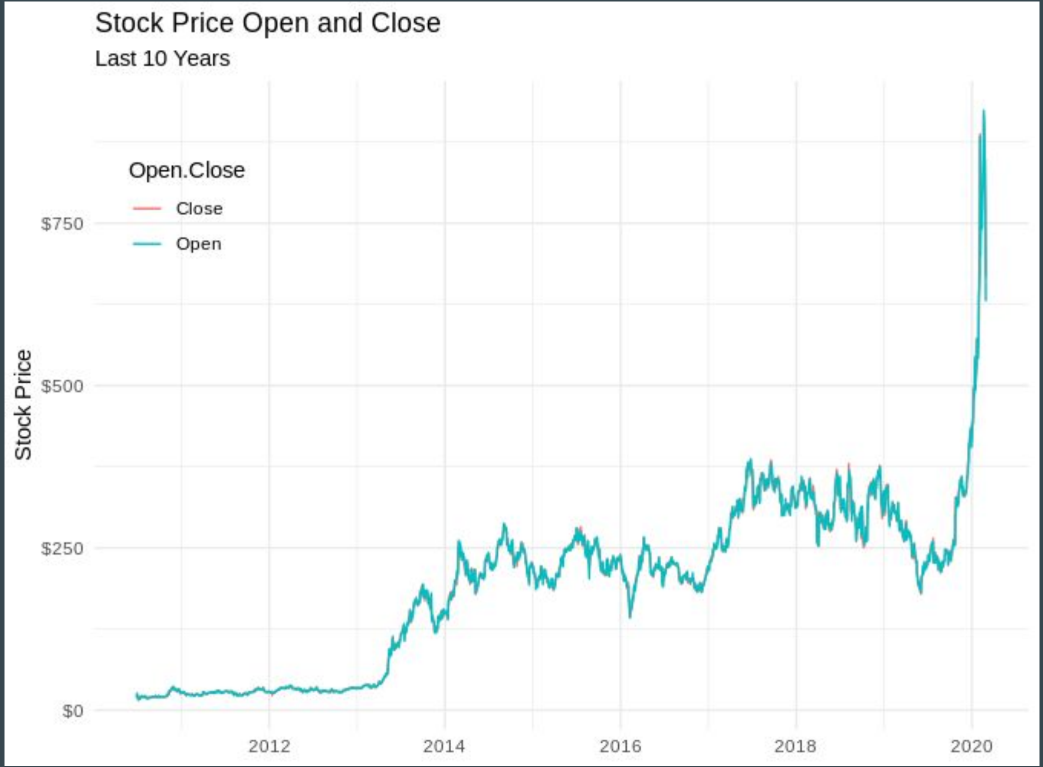
1. Easily plot tidy data
2. Highly customizable plots
3. Huge array of available visualizations
4. Popular for professional data visualizations
5. “Opinionated”



ggplot2 example



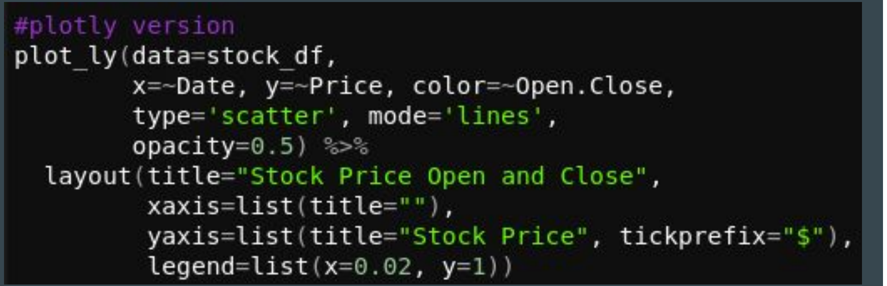
```r
#ggplot2 version
ggplot(data=stock_df, aes(x=Date, y=Price, color=Open.Close)) +
  geom_line() +
  scale_y_continuous(labels = dollar) +
  labs(title="Stock Price Open and Close",
       subtitle="Last 10 Years",
       x=NULL,
       y="Stock Price") +
  theme_minimal() +
  theme(legend.position = c(0.1, 0.8))
```



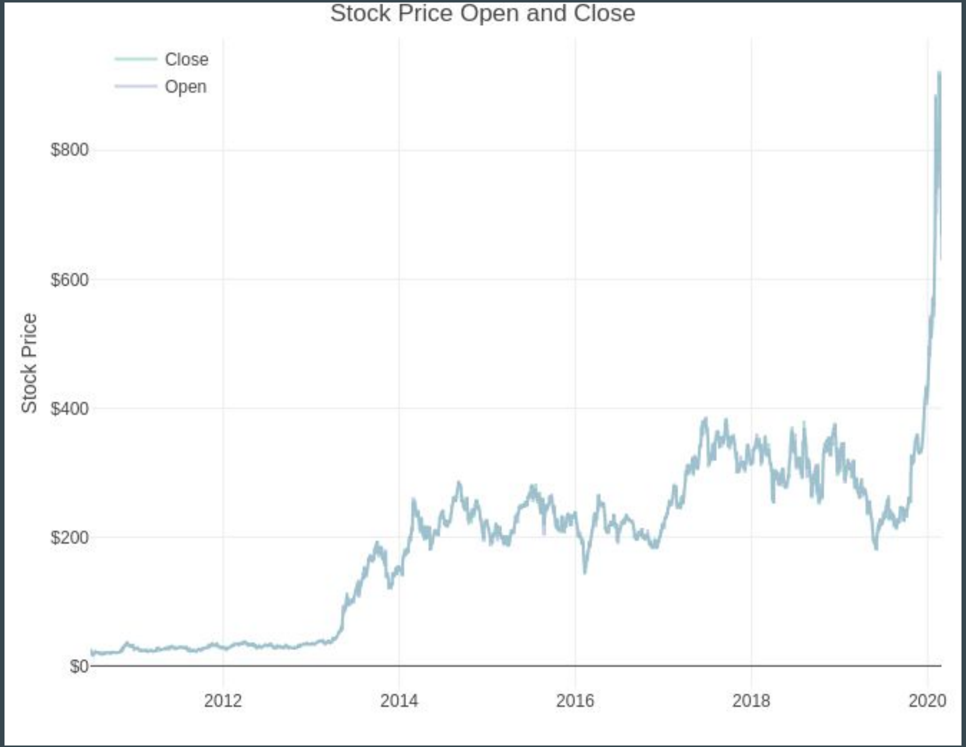
Plotly - Interactive Visualizations

1. Open source data visualization
2. Straightforward syntax
3. Interactive elements
4. Great for dashboards, exploratory analysis
5. Cross platform

Plotly Example



```
#plotly version
plot_ly(data=stock_df,
        x=~Date, y=~Price, color=~Open.Close,
        type='scatter', mode='lines',
        opacity=0.5) %>%
  layout(title="Stock Price Open and Close",
         xaxis=list(title=""),
         yaxis=list(title="Stock Price", tickprefix="$"),
         legend=list(x=0.02, y=1))
```



Plotly and ggplot2 Practice

1. With the ‘diamonds’ dataset, make a scatter plot of carat vs price and colored by clarity
2. With ‘diamonds’ make a box plot carat on y, cut on x. (Break out by color for extra practice)
3. Add axis labels and titles to one of the previous charts. (You can use the stock charts above as a reference)
4. With ‘diamonds’ make any chart you’d like, but incorporate some sort of faceting.

Challenge:
Convert the ‘UCBAdmissions’ dataset to a data frame (like with Titanic). Create a visualization that shows the number of students admitted or rejected by gender and department.

Advanced Plotting Demos

- Pareto Plot
- Parallel Coordinates / Sankey Plot
- Funnel Diagram
- Treemap
- Waterfall chart
- Sunburst

Additional Resources

R for Data Science: https://r4ds.had.co.nz/

ggplot2 reference: https://ggplot2.tidyverse.org/

ggplot2 extensions: http://www.ggplot2-exts.org/gallery/

The R graph gallery: https://www.r-graph-gallery.com/index.html

Plotly: https://plot.ly/r/

Shiny: https://shiny.rstudio.com/