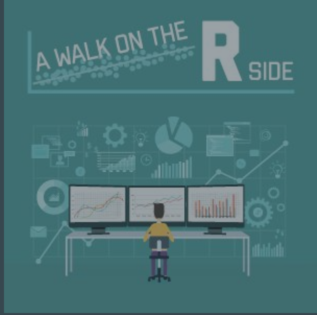
A Walk on the R Side

an introduction to R for data analysis

GW Libraries/STEMWorks Workshop
November 2017

go.gwu.edu/gwlibrworkshop



Agenda

- About R and RStudio
- Hands-on:
  - variables
  - logical expressions
  - values, vectors, and data frames
  - R Studio projects
  - reading in data
  - exploring data
  - data wrangling: cleaning and reshaping
  - data visualization
  - data analysis
  - functions
- Resources for further learning

Acknowledgments

software carpentry

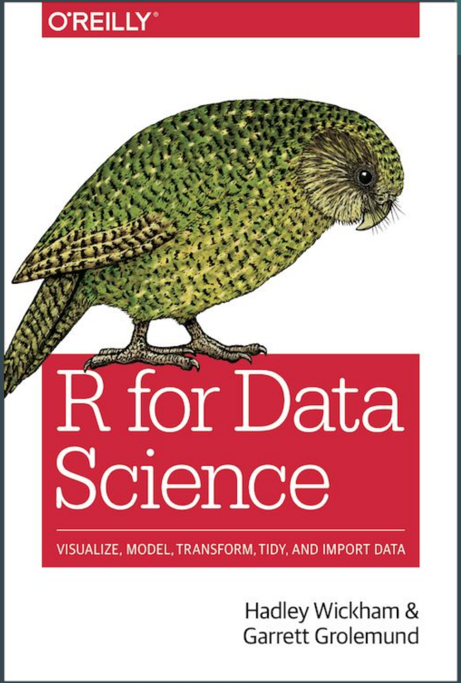
Teaching basic lab skills for research computing

DATA CARPENTRY

BUILDING COMMUNITIES TEACHING UNIVERSAL DATA LITERACY

R Tutorial

An R Introduction to Statistics



O'REILLY

R for Data Science

VISUALIZE, MODEL, TRANSFORM, TIDY, AND IMPORT DATA

Hadley Wickham & Garrett Grolemund

Workshop Housekeeping

Ask questions!

Respect every question and person asking the question

Help each other out!

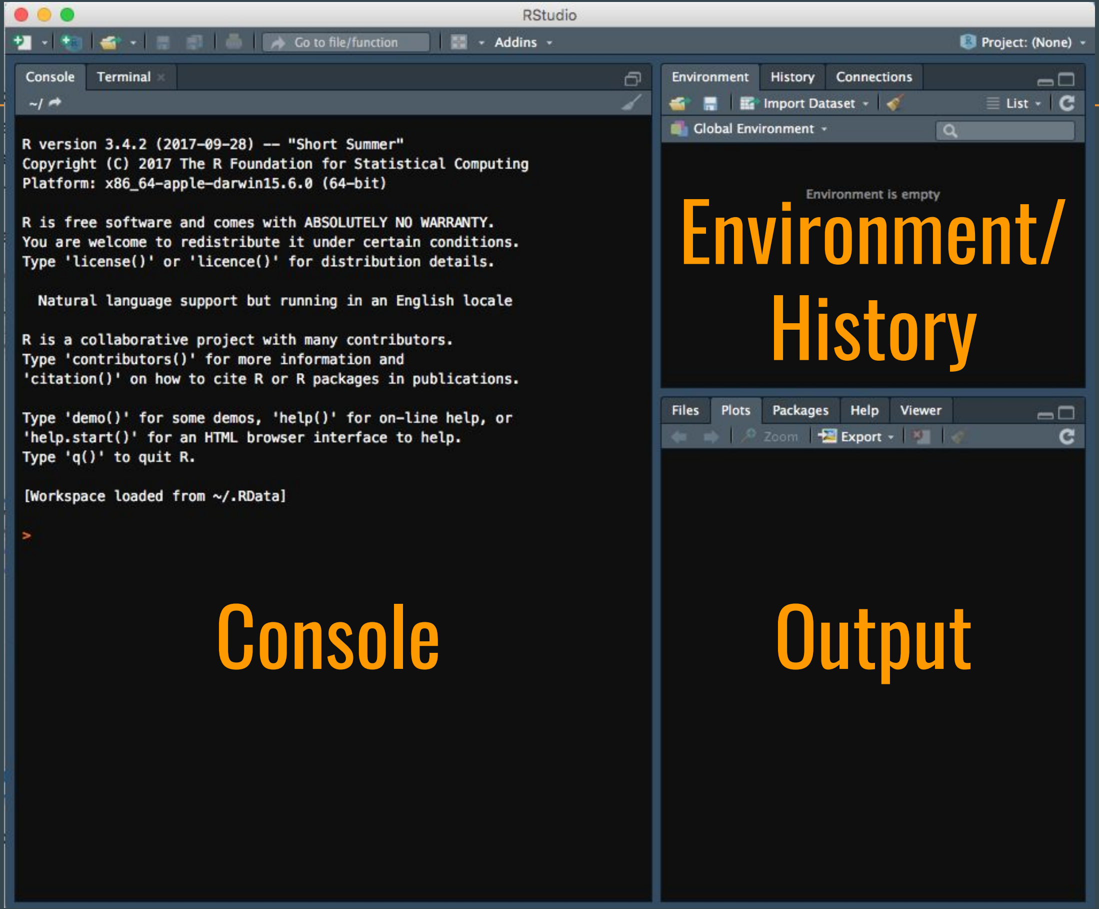
If something is confusing in the workshop,
it probably needs improvement; let us know.

Stay as long as you like

About R

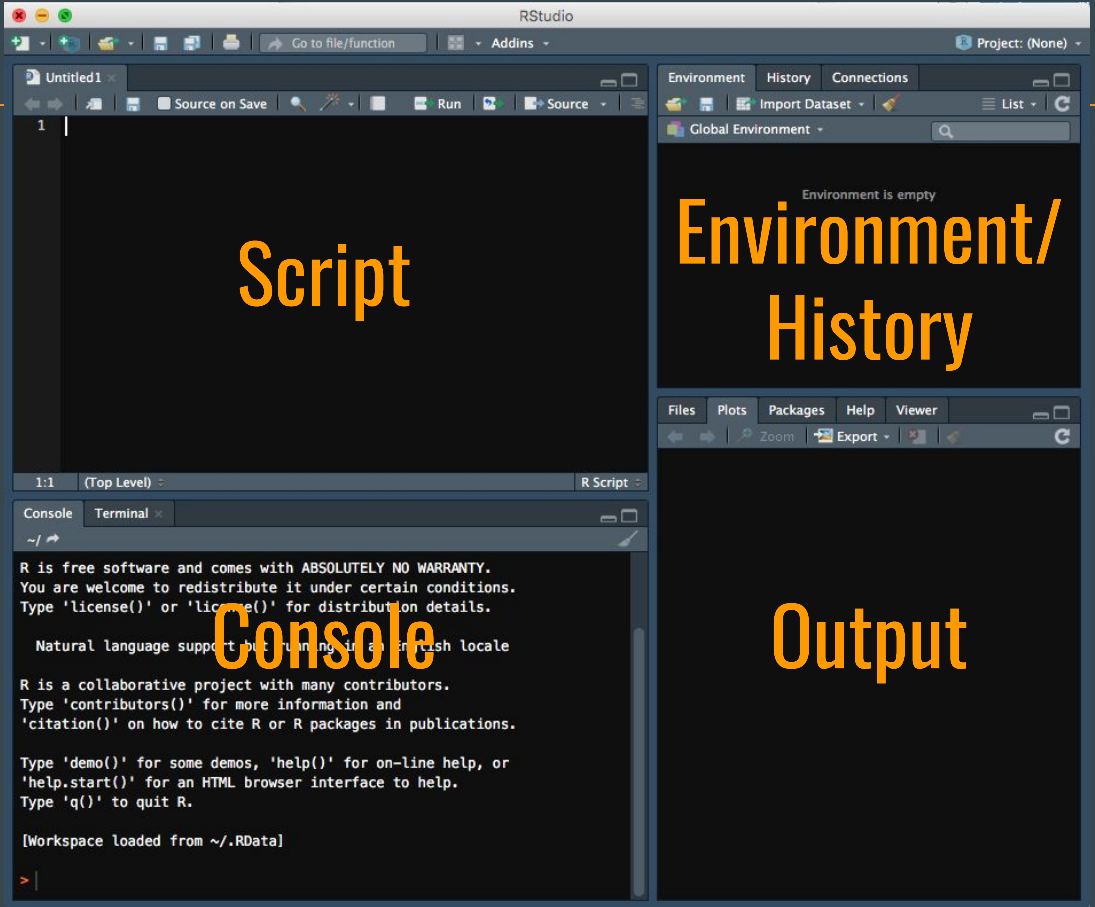
- Open source
- For statistical computing and graphics
- CRAN
    - R packages
    - R events
    - R journal
    - ...

R Studio



```
R version 3.4.2 (2017-09-28) -- "Short Summer"
Copyright (C) 2017 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin15.6.0 (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.

[Workspace loaded from ~/.RData]

>
```

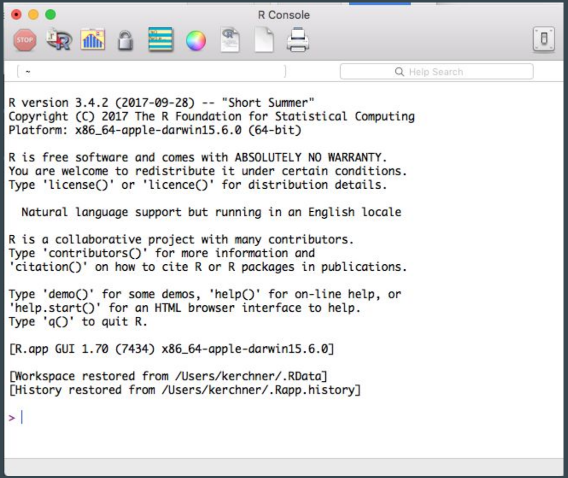
R Studio



```
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.

[Workspace loaded from ~/.RData]

>
```

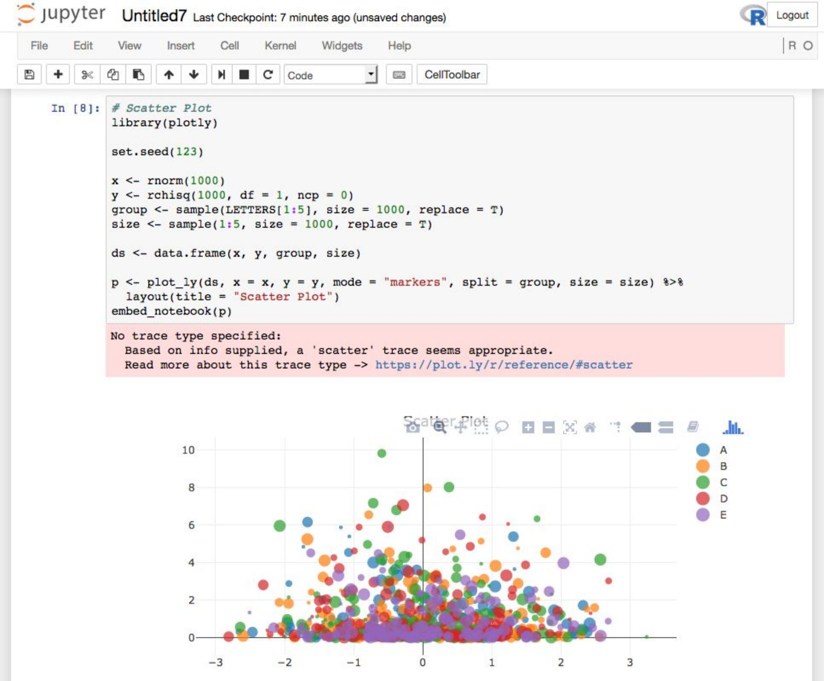
Other Ways To Use R

Plain R console



```
R version 3.4.2 (2017-09-28) -- "Short Summer"
Copyright (C) 2017 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin15.6.0 (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.

[R.app GUI 1.70 (7434) x86_64-apple-darwin15.6.0]

[Workspace restored from /Users/kerchner/.RData]
[History restored from /Users/kerchner/.Rapp.history]

> |
```

Jupyter Notebook (e.g. in Anaconda)



```
In [8]: # Scatter Plot
        library(plotly)

        set.seed(123)

        x <- rnorm(1000)
        y <- rchisq(1000, df = 1, ncp = 0)
        group <- sample(LETTERS[1:5], size = 1000, replace = T)
        size <- sample(1:5, size = 1000, replace = T)

        ds <- data.frame(x, y, group, size)

        p <- plot_ly(ds, x = x, y = y, mode = "markers", split = group, size = size) %>%
          layout(title = "Scatter Plot")
        embed_notebook(p)
```

```
No trace type specified:
  Based on info supplied, a 'scatter' trace seems appropriate.
  Read more about this trace type -> https://plot.ly/r/reference/#scatter
```

Let's tRy it!

Variables

- Try using R as a "calculator" in the Console
  - Try some mathematical functions, too
- Create some variables
  - variable naming
  - `<-` for assigning values to variables (Option - on Mac, Alt - on Win)
  - numeric, character, logical
  - Watch the Environment pane!
  - `class()`
  - Coercion w/ `as.integer, as.character, as.logical, as...`

Logical Expressions

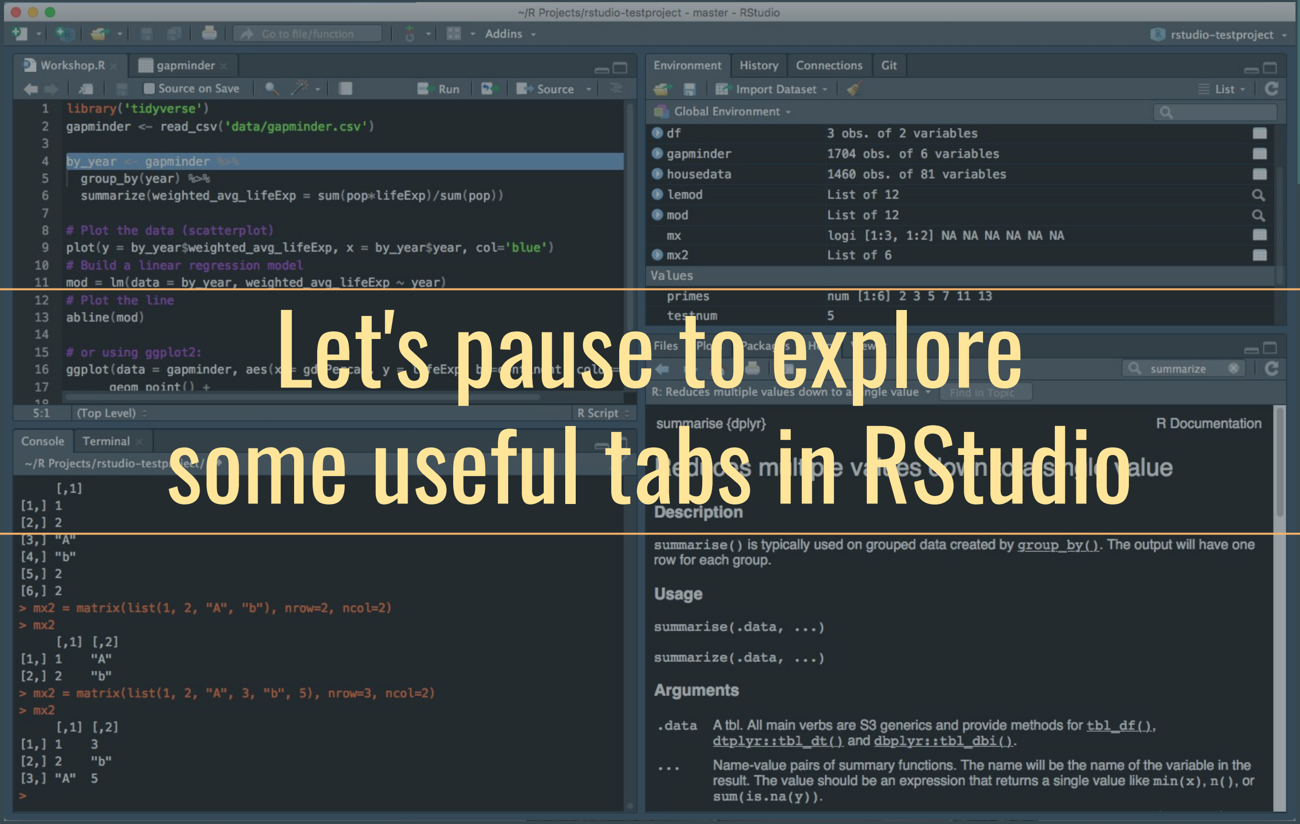
- Operators include:
  ==, <, >, ! (not), & (and), | (or), etc.

Vectors

Vectors

- A vector is
  - A sequence of data elements (components) all of the same type.
- Create vectors with c()

Let's pause to explore some useful tabs in RStudio



```
library('tidyverse')
gapminder <- read_csv('data/gapminder.csv')

by_year <- gapminder %>%
  group_by(year) %>%
  summarize(weighted_avg_lifeExp = sum(pop*lifeExp)/sum(pop))

# Plot the data (scatterplot)
plot(y = by_year$weighted_avg_lifeExp, x = by_year$year, col='blue')
# Build a linear regression model
mod = lm(data = by_year, weighted_avg_lifeExp ~ year)
# Plot the line
abline(mod)

# or using ggplot2:
ggplot(data = gapminder, aes(x = gdpPercap, y = lifeExp
      geom_point() +
```

```
     [,1]
[1,] 1
[2,] 2
[3,] "A"
[4,] "b"
[5,] 2
[6,] 2
> mx2 = matrix(list(1, 2, "A", "b"), nrow=2, ncol=2)
> mx2
     [,1] [,2]
[1,] 1    "A"
[2,] 2    "b"
> mx2 = matrix(list(1, 2, "A", 3, "b", 5), nrow=3, ncol=2)
> mx2
     [,1] [,2]
[1,] 1    3
[2,] 2    "b"
[3,] "A"  5
>
```

```
df          3 obs. of 2 variables
gapminder   1704 obs. of 6 variables
housedata   1460 obs. of 81 variables
lemod       List of 12
mod         List of 12
mx          logi [1:3, 1:2] NA NA NA NA NA NA
mx2         List of 6
Values
primes      num [1:6] 2 3 5 7 11 13
testnum     5
```

summarise {dplyr} R Documentation

Reduces multiple values down to a single value

Description

summarise() is typically used on grouped data created by group_by(). The output will have one row for each group.

Usage

```
summarise(.data, ...)

summarize(.data, ...)
```

Arguments

.data A tbl. All main verbs are S3 generics and provide methods for tbl_df(), dtplyr::tbl_dt() and dbplyr::tbl_dbi().

... Name-value pairs of summary functions. The name will be the name of the variable in the result. The value should be an expression that returns a single value like min(x), n(), or sum(is.na(y)).

Data Frames

Data Frames

- A data.frame stores a data table
- Comprised of vectors of equal length. Vectors become columns.
- Columns and rows can have names.
- `tibble` (from the tibble package) has some advantages over `data.frame`

A WALK ON THE R SIDE

To summarize...

Value

10.2

Vector

1	10.2
2	11.3
3	11.5
4	12.0

Data Frame

	time	temp	boiling
1	51	10.2	FALSE
2	58	11.3	FALSE
3	63	11.5	FALSE
4	70	12.0	TRUE

A brief word on `list` and `matrix`

Projects in RStudio

Projects in RStudio

Recommendations:

- Use [Github for] **version control**!
- Create folders to keep things organized

It's time to import some data!

Data Importing

- Prepare data as "tidy"
  - rectangular
  - one table per file
- Formats: CSV, TSV, Excel, Fixed-Width, JSON... and with the right packages: Stata, SPSS, SAS... (using **`foreign`**)
- A word about "big data"

R Packages

flickr.com/photos/philocrites/405250216

Installing and loading R packages

- install.packages('mypackage')
- library('mypackage') -- or check the box on the Packages tab in RStudio

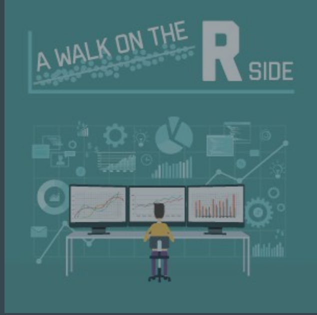
Tidyverse Core Packages

tidyverse.org

- ggplot2 - graphics
- dplyr - data manipulation
- tidyr - tidying data
- readr - reading in data
- tibble - modern data frame
- purrr

Other often-used R packages

Basic stats functions, like ANOVA	↔	MASS
Mapping	↔	tmap, tmaptools, leaflet
Analyzing 2D and 3D shapes	↔	geomorph
Genomic data	↔	bioconductor
Cluster analyses	↔	cluster
Time series data	↔	forecast
Text mining	↔	qdap, sentimentr, tidytext
Interactive web visualizations	↔	shiny



Exploring Data

- head, tail
- subsetting
- slicing and dicing

Data Wrangling

flickr.com/photos/thewomensmuseum/3607075017

Data Transformation using the dplyr package

- filter()
- arrange()
- select()
- mutate()
- summarize()
- group_by()
- ...

You will want to use a "pipe": %>%

(shortcut: control-shift-M)

flickr.com/photos/jonasb/3151160389

Data Tidying

- gather()
- spread()
- separate()
- unite()

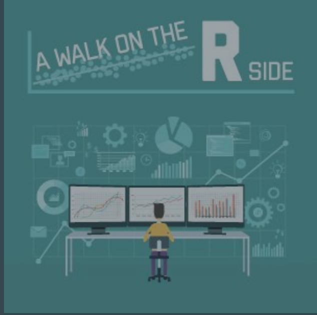
Joining

"Merges" tables together

- left_join()
- right_join()
- ...

Analyzing Data

Data Visualization



Data Visualization

- 3 main packages:
  - "base R"
  - lattice
  - ggplot2

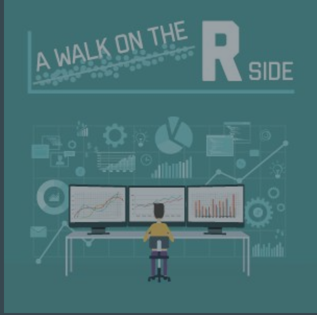
Functions

Some Handy R Links

Tutorials

- Software Carpentry:
  - http://swcarpentry.github.io/r-novice-inflammation
  - http://swcarpentry.github.io/r-novice-gapminder
- Data Carpentry:
  - http://datacarpentry.github.io/R-ecology-lesson/
  - http://www.datacarpentry.org/R-genomics/
- Lynda.com lynda.it.gwu.edu - 3 video courses (~12 hours)
- r-tutor.com/r-introduction
  r-tutor.com/elementary-statistics
- R for Data Science http://r4ds.had.co.nz

Classes at GW that teach R



- PSC 2012
  Visualizing and Modeling Politics
  Prof. Eric Lawrence
  currently scheduled next for Fall 2018

- Others?

Reference Links

- r-project.org
- R search engine: rseek.org
- rstudio.com
    - Cheat Sheets
- stackoverflow.com

Thanks!

Dan Kerchner	kerchner@gwu.edu
Dr. Kes Schroer	schroerk@gwu.edu
Vishwesh Haldevanekar	vishwesh_s_h@gwu.edu

These slides: go.gwu.edu/gwlibrworkshop

Statistics Appointments with Vishwesh:
calendly.com/vishwesh_s_h