RSCH 630 - Research Methods

Final Exam

Your Name

Due 17:00 6/7/2016

Abstract

"The purpose of this entire course has been to introduce techniques that can help you during your research. And while it's easy to focus on the mathematical methods, you're likely to spend more time on the grunt-work tasks, such as getting and cleaning data, building tables and figures, numbering equations and citing sources. On the first day of class I stated that a lot of effort has gone into developing tools that make these tasks simpler - and in some cases almost trivial. It would be wrong for me to simply make such a statement without showing you how to take advantage of these tools. Therefore, the purpose of this exam is to have each of you build (and use) a data science toolkit that links these tools together. R/RStudio will comprise the core of this toolkit. However, even if you don't use R, this toolkit will allow you to add in almost any other language you want. Building the toolkit will take some effort up-front and will introduce you to several new concepts, but once it's up and running the workflow becomes almost seamless. Alright, let's get started."

Installing Resources

Step 1 - Create a GitHub Account

- Go to GitHub and sign up for an account
  - Choose a username
  - Add an email address (don't use your afit.edu address as your primary, you can add it as a secondary address later)
  - Enter a password
  - Before moving forward check your inbox for an email from GitHub to verify your address.
- GitHub is a version control tool for ASCII or UTF-8 encoded files
  - Organizes code, text files, images, etc. into repositories 'repos'

  - Easy to collaborate on projects with other users any where in the world without
  - No more emailing files, DVD's, USB drives - It's like dropbox on steroids
  - It can be used on AF NIPR machines (probably not all, but haven't found one yet that couldn't)
  - GitHub has a learning curve, but once you get the hang of it you don't go back
  - Outside of AFIT/AF, GitHub is the standard - some employers prefer GitHub profiles over resumes

Step 2 - Fork the rsch-630/spring-2016 repository

- Click the round GitHub 'Octet' logo near the top to ensure you're not inside a repository
- In the search box type `rsch-630` and click search
- Click the `users` link toward to access the rsch-630 organization account
- Click the `spring-2016` link to access the repository
- Near the top of the page are three buttons `watch`, `star`, and `fork`. Click the `fork` button.
- You should now see the `your.username/spring-2016` repository under your account
- This forked copy belongs to you - your changes have no affect on the `rsch-630/spring-2016`
- Changes can only be made to my repo if you submit a pull request and I accept it.

Step 3 - Install R

- R version 3.3.0 (2016-05-03) can be installed from https://cloud.r-project.org/
- Select the version for your OS
- Follow the prompts for a default installation

Step 4 - Install a Bunch of R Packages

- These are some basic packages needed for
  - Creating markdown documents
  - Developing shiny apps
  - Create instant tables
  - Cite references
- Copy and paste the following lines of code, one at a time
  - `install.packages(c('installr','devtools','rmarkdown'))`
  - `install.packages(c('shiny','RJSONIO','xtable'))`
  - `install.packages(c('DT','knitcitations','RefManageR'))`
- Any packages that are dependencies of this packages will also be installed
- Note: the `installr` package is specific to Windows machines and is not available for Mac, Linux, Unix machines

Step 5 - Install RStudio

- On Windows machines this is really easy - just run `installr::install.rstudio()` inside of R
- For non-Windows machines, go to https://www.rstudio.com/products/rstudio/download/ to install the most recent version
- **NOTE:** RStudio is an IDE
  - You can work on files written in different languages side-by-side
  - You can mix languages together in a single file

Step 6 Install and Configure Git

- Installing Git on Windows machines is again easy - Just run `installr::install.git()`
- For non Windows machines install Git from https://git-scm.com/
- Search your machine for the Git Bash application - once found open it
- In Git Bash type the following (The quotes are for emphasis don't use them)
  - git config –global user.name "your-GitHub-username"
  - git config –global user.email "your-primary-GitHub-email-address"
- **DON'T** close Git Bash - you'll need it again soon

Step 7 Connect Git to RStudio

- In RStudio, open the `Tools` menu and select `Global Options`
- In the `Global Options` window select the Git/SVN tab
- In `Git Executable` window, select browse and find `git.exe`
  - In Windows machines, `git.exe` can usually be found at **C:/Program Files/Git/bin/git.exe**
  - I'm less familiar with other OS's, but the file path should be similar
  - Once `git.exe` has been located select ok - **BUT DON'T CLOSE THE GLOBAL OPTIONS WINDOW**
- While still at the `Git/SVN` menu, press the small 'Create RSA Key' button
- If prompted for a password, ignore it and select 'create'
- When the Key appears, close the window and select `View Public Key`
  - Copy the Key to your clipboard and return to Github.com
  - At the top of the page select the arrow beside your avatar image and choose `Settings`
  - Select `SSH and GPG Keys`

  - Select new `SSH Key`
  - Type a name for the Key and paste the public key into the window, save and close.
  - Linking GitHub and RStudio with a SSH key-pair makes the connection ultra-secure and will greatly speed-up your workflow

Step 8 Create a new RProject

- On GitHub, return to the `your.GitHub.username/spring-2016` repository
- On the repository page select the green 'Clone or Download' button
- A window with a url address (https://...) should appear
- Click the button next to the window to copy the address to your clipboard
- Return to RStudio - at the top-right of the screen find the project dropdown - it should now read 'Project (none)'
- Select the dropdown arrow and choose 'New Project', then 'Version Control', then 'Git'
- Three windows should be visible, in the top window paste the address for your repo that you copied from GitHub
- In the bottom window, select browse to choose the folder you want the files from your repo to be 'cloned' into
- Select OK, a new instance of R should appear - any files you were working on before creating the project are still there and will be visible when you close the project.
- Clicking the 'Files' tab in the lower-right pane should show the files that are in your repo at GitHub.com and on your machine in the folder you chose
- Open and change any of the files. Upon saving - the modified file should be listed under the 'Git' tab in the upper-right pane.
- Click the checkbox next to the file and press 'Commit'
- Type a brief commit message about the changes made and press ok

- Close the window and select 'Push'
  - Enter your GitHub user.name and password
  - If the push was successful the update to the file should be reflected on the repo page at GitHub.com

Step 9 Install & Connect Rtools (Windows Only)

- On Windows machines run `installr::install.rtools()`
- Similar to what we did for Git, we need to make sure that R can find Rtools
  - Go to start and do a file search for 'environment'
  - Options should appear to modify the system environment variables or the account environment variables
  - If you are have admin right on the machine you are using - choose to modify the system environment variables
  - One of these environment variables will be called 'path' choose to edit this one
  - We need to add entries onto the front of this list so that two binaries folders can be located
  - On most machines, Rtools is installed into the C:/ by default
  - Under this scenario the following should be added to the system `path` C:\Rtools\bin;C:\Rtools\gcc-4.6.3\bin;
  - If Rtools is installed in a different location, change these entries accordingly to find the `bin` and `gcc-4.6.3\\bin` folders.

Step 10 Install a LaTeXdistribution for creating PDF's

- On Windows machines run `installr::install.miktex()`
- On Mac machines, install the full version of MacTeX from http://tug.org/mactex/

- On Linux/Unix machines, install the full version of TeX Live http://tug.org/texlive/

Congratulations! You've installed and integrated R/RStudio Data Science Toolkit

- Now that you have it, learn how to use it
- Below is a working example that highlights several R packages for
  - Retreiving data
  - Cleaning and modifying data
  - Automatically creating tables
  - Citing sources and creating a bibliography
  - Generating plots

Working Example

Scraping data from Data.gov using the `httr` package

```
getPackage('httr', repo = 'CRAN')

url <- 'https://data.cms.gov/resource/ehrv-m9r6.json'

hosp.data <- content(GET(url))

DATA <- matrix(NA, nrow = 1000, ncol = length(hosp.data[[1]]))

for(i in 1:1000) {

dats <- data.frame(unlist(hosp.data[[i]]),stringsAsFactors = F)
DATA[i,] <- t(data.frame(dats,stringsAsFactors = F))
}
```

Processing The Data using the `data.table` package

```
getPackage('data.table',repo = 'CRAN')

DATA <- as.data.frame(DATA, stringsAsFactors = F)

colnames(DATA) <- rownames(dats)

DATA[c(1:3,12)] <- lapply(X = DATA[c(1:3,12)], FUN = {unlist ; as.numeric} )
#DATA[c(1:3,12)] <- lapply(X = DATA[c(1:3,12)], FUN = as.numeric)
dats <- data.table::as.data.table(DATA)
setkey(dats, provider_state)

covered = dats[,j = sum(average_covered_charges), by = provider_state]
payment = dats[,j = sum(average_medicare_payments,average_medicare_payments_2), by = provider_s
discharge = dats[,j = sum(total_discharges), by = provider_state]

sums <- covered[payment[discharge]]

colnames(sums) <- c('Provider State','Covered Charges','Medicare Payments','Total Discharges')
```

Inserting tables automatically with the `xtable` package

```
getPackage('xtable', repo = 'CRAN')

xsums <- xtable(sums,
                caption = 'Extracranial Procedures Medical Payment Information')
print(xsums,
      comment = F,
      include.rownames = F,
      caption.placement = 'top',
      table.placement = 'h',
      type = switch(output, 'html' = 'html', 'pdf' = 'latex'))
```

Citing sources and creating bibliographies with the `knitcitations` package

```
getPackage('knitcitations',repo = 'CRAN')
knitcitations::cite_options(cite.style = "numeric", citation_format = 'pandoc')

cleanbib()
bib  <- read.bibtex('resources/rsch-630.bib')
bort <- citet(bib[author='bort'])
```

```
write.bibtex(file = 'resources/bibliography.bib')
```

"... if GitHub goes down, the software development world practically stops." [1]

References

[1] J. Bort, "Why $2 billion startup GitHub is apparently in crisis, again," *Business Insider*. Feb-2016.

Table 1: Extracranial Procedures Medical Payment Information

Provider State	Covered Charges	Medicare Payments	Total Discharges
AK	34805.13	14815.73	23.00
AL	737022.05	237893.71	879.00
AR	439412.21	176920.66	652.00
AZ	802640.65	316425.33	606.00
CA	3758360.36	1022881.47	1715.00
CO	352522.07	124718.52	255.00
CT	324480.09	220795.03	384.00
DC	142109.63	52814.37	47.00
DE	71899.56	38542.59	144.00
FL	3453570.67	896814.74	2847.00
GA	817751.40	335019.92	1015.00
HI	27809.95	17828.91	24.00
IA	308569.52	157788.85	518.00
ID	115049.71	72258.57	113.00
IL	1775623.21	663680.82	1240.00
IN	967990.18	421914.42	1078.00
KS	413753.56	175982.37	594.00
KY	402152.95	209508.20	728.00
LA	854004.03	286390.00	743.00
MA	571138.89	390889.57	847.00
MD	202294.89	359042.13	706.00
ME	43441.97	37621.12	205.00
MI	892884.86	519228.72	1325.00
MN	351564.46	188878.79	445.00
MO	783848.97	342104.18	1030.00
MS	412137.08	170357.37	588.00
MT	90997.88	58351.72	174.00
NC	571979.55	322140.07	1141.00
ND	86702.08	70695.29	178.00
NE	214240.29	111884.31	318.00
NH	193600.00	100773.88	216.00
NJ	1088217.96	415506.82	819.00
NM	227310.18	79553.73	128.00
NV	522703.23	137947.50	311.00
NY	1069991.17	694620.36	1336.00
OH	1352733.33	564024.31	1186.00
OK	239155.88	117278.91	463.00
OR	295619.25	154057.80	354.00
PA	1459743.61	536073.30	1003.00
RI	74741.41	43908.66	76.00
SC	643419.21	228318.54	721.00
SD	87347.69	49616.82	156.00
TN	845356.01	352225.32	986.00
TX	2378767.25	900469.38	2759.00
UT	123586.15	69379.32	109.00
VA	120592.55	65796.73	161.00
VT	17934.35	17226.83	57.00