Data Wrangling in R

http://sisbid.github.io/Module1/
Preliminaries
<table>
<thead>
<tr>
<th>Course Info</th>
</tr>
</thead>
<tbody>
<tr>
<td>Course name</td>
</tr>
<tr>
<td>Instructors</td>
</tr>
<tr>
<td>Course website</td>
</tr>
<tr>
<td>Goals</td>
</tr>
<tr>
<td>Pre-reqs</td>
</tr>
</tbody>
</table>
How many people feel about data wrangling
How we feel about data wrangling
About us

(John)
Welcome to the Leek group

Welcome to the Leek group in the Data Science Lab and the Department of Biostatistics at the Johns Hopkins Bloomberg School of Health. We are a group of researchers, educators, and data scientists using data to solve problems in molecular biology, human health, meta-research, education, and anything else we think could be useful for the world. We produce data tools and code that you can use for your projects as well. We teach online open classes so you can learn how to use data too. If you think any of this sounds cool consider joining us in working to make the world a better place. If you just want to keep up with everything we are working on, follow Jeff on Twitter https://twitter.com/jtleek.
Classes

Introduction to R
Faculty Office Hour

Short Courses

Neurohacking
Imaging in R (aka Neurohacking 2.0)
Building R Packages

https://johnmuschelli.com/
The 3 ‘Times’ of a Project

Posted on May 15, 2020 by strictlystat

During a conversation with Sean Kross about projects, particularly data science projects, I tried to explain how things can go right and wrong with a project. I was explaining things with respect to being the data scientist on academic projects, but I think these issues are cross-cutting so figured I’d post them here.

I thought back to when projects did not go well or someone was left frustrated or angry during or at the end of the interaction. To me, the issues usually come down to the 3 “time”s of a project: time, timeline, and timeliness.
About us

(Andrew)
Andrew Jaffe

- Lead Investigator at the Lieber Institute for Brain Development
- Associate Professor at Johns Hopkins University (Mental Health, Biostats, Psychiatry, and Human Genetics)
- Run academic data science team
- My research focuses on molecular correlations of psychiatric brain disorders like schizophrenia, bipolar disorder, and major depression
Overview

The Jaffe Lab is led by Andrew E Jaffe.

The lab is associated with the Lieber Institute for Brain Development and the Departments of Mental Health and Biostatistics at Johns Hopkins Bloomberg School of Public Health. We are also part of the Center for Computational Biology at Johns Hopkins University.

Research Interests

We are a computational biology and genomics lab within the Lieber Institute for Brain Development (LIBD). We are interested in better understanding and characterizing genomics signatures in the human brain, including DNA methylation and gene expression.

Contact

Email: @andrewejaffe
http://www.aejaffe.com
Why this class
This is Jeff's awesome website

![[](https://media.giphy.com/media/drXGoW1iudhKw/giphy.gif)](https://media.giphy.com/media/drXGoW1iudhKw/giphy.gif)
library(bigrquery)
set_service_token("file.json")

con <- dbConnect(
  bigquery(),
  project = "project_name",
  dataset = "dataset_name"
)

unique_elements = con %>%
  tbl("dataset1") %>%
  count()

unique_elements
Running job 'job_id.US'...
Complete
Billed: 32.51 MB
Downloading 10 rows in 1 pages.
# Source: lazy query [?? x 2]
# Database: BigQueryConnection

  n
<int>
1 3700675
library(httr)
library(dplyr)

username = 'janeeverydaydoe'

url_git = 'https://api.github.com/

api_response = GET(url = paste0(url_git, 'users/', username, '/repos'))

content(api_response)[[1]]
---
title: "How does your BMI measure up?"
output: flexdashboard::flex_dashboard
runtime: shiny
---

Inputs (.sidebar)
-------------------------------------
```
library(flexdashboard); library(NHANES); library(plotly);library(dplyr)
sliderInput("height", "Height in inches",0,100,72)
sliderInput("weight", "Weight in pounds",0,500,100)
sliderInput("age", "Age in years",0,120,50)
```

Column
-------------------------------------

### Chart 1
```
nhanes = sample_n(NHANES,100)
renderPlotly({
  df = data.frame(bmi = c(nhanes$BMI,input$weight*0.45/(input$height*0.025)^2),
               age = c(nhanes$Age,input$age),
               who = c(rep("nhanes",100),"you"))
  ggplotly(ggplot(df) +
    geom_point(aes(x=age,y=bmi,color=who)) +
    scale_x_continuous(limits=c(0,90)) +
    scale_y_continuous(limits=c(0,60)) +
    theme_minimal()
  )
})
```

---

---

---

---

---

---

---

---

---
But also...
Genomic signatures to guide the use of chemotherapeutics

Anil Potti¹,², Holly K Dressman¹,², Andrea Bild¹,³, Richard F Riedel¹,², Gina Chan⁴, Robyn Sayer⁴, Janiel Cragun⁴, Hope Cottrill⁴, Michael J Kelley², Rebecca Petersen⁵, David Harpole⁵, Jeffrey Marks⁵, Andrew Berchuck¹,⁶, Geoffrey S Ginsburg¹,², Phillip Febbo¹,²,³, Johnathan Lancaster⁴ & Joseph R Nevins¹,²,³

Using in vitro drug sensitivity data coupled with Affymetrix microarray data, we developed gene expression signatures that predict sensitivity to individual chemotherapeutic drugs. Each signature was validated with response data from an independent set of cell line studies. We further show that many of these signatures can accurately predict clinical response in individuals treated with these drugs. Notably, signatures developed to predict response to individual agents, when combined, could also predict response to multidrug regimens. Finally, we integrated the chemotherapy response signatures with signatures of oncogenic pathway deregulation to identify new therapeutic...
When is Reproducibility an Ethical Issue? Genomics, Personalized Medicine, and Human Error

Keith A. Baggerly
Bioinformatics and Computational Biology
UT M. D. Anderson Cancer Center
kabagg@mdanderson.org

BIRS Workshop, Aug 14, 2013

http://www.birs.ca/events/2013/5-day-workshops/13w5083/videos/watch/201308141121-Baggerly.mp4
NORTH CAROLINA
DURHAM COUNTY

Richard Aiken, Jean K. Carroll, Executors of the Estate of Harold G. Carroll, Jean K. Carroll, Individually, Peggy Cox, as Administratrix of the Estate of Paul F. Cox, Peggy Cox, Individually, Helene L. Fligel, Jason Gannon, as Personal Representative of the Estate of Jennifer L. Gannon, John Haddock, as Executor of the Estate of Karen Heath, Walter Jacobs, as Executor of the Estate of Juliet J. Jacobs, Walter Jacobs, Individually, Polly Johnson, as Executor of the Estate of Malcolm W. Johnson, and Polly Johnson, Individually,

Plaintiffs

VS.

COMPLAINT
(JURY TRIAL DEMANDED)

http://www.dukechronicle.com/articles/2015/05/03/duke-lawsuit-involving-cancer-patients-linked-anil-potti-settled
Doesn’t seem that important....
```r
> load("~/Documents/Work/workingpapers/openreview/data/processed-data-may11.rda")
> dim(dat)
[1] 730 15
> summary(glm(dat$correct ~ dat$study_type + dat$study_id,family="binomial"))

Call:
glm(formula = dat$correct ~ dat$study_type + dat$study_id, family = "binomial")

Deviance Residuals:
           Min         1Q     Median         3Q        Max
-1.6173 -1.4259   0.7941   0.9478   1.1431

Coefficients: (1 not defined because of singularities)
               Estimate Std. Error   z value Pr(>|z|)
(Intercept)    0.5675      0.1475     3.847  0.000120
dat$study_type 0.4250      0.2182     1.948  0.051458
```
ON THE ONE HAND...
Most of the attention is on the last step
This course is about all the steps that come before
They are critical for getting things rights
The garden of forking paths: Why multiple comparisons can be a problem, even when there is no “fishing expedition” or “p-hacking” and the research hypothesis was posited ahead of time*

Andrew Gelman† and Eric Loken‡
14 Nov 2013

“I thought of a labyrinth of labyrinths, of one sinuous spreading labyrinth that would encompass the past and the future ... I felt myself to be, for an unknown period of time, an abstract perceiver of the world.” — Borges (1941)
Explorable Multiverse Analyses

We focus our analysis on task completion times, reported in Figures 3 and 4. Dots indicate sample means, while error bars are 95% confidence intervals computed on log-transformed data using the t-distribution method. Strictly speaking, all we can assert about each interval is that it comes from a procedure designed to capture the

![Figure 3. Average task completion time (geometric mean) for each condition. Error bars are 95% t-based CI.](image)

![Figure 4. Perceptually-driven ranking of visualizations depending on the correlation sign (neg/pos), as a function of correlation value (r) and overall (right column).](image)

Pierre Dragicevic (Inria), Yvonne Jansen (CNRS - Sorbonne Université), Abhraneel Sarma (University of Michigan)
Matthew Kay (University of Michigan), Fanny Chevalier (University of Toronto)

With **explorable multiverse analysis reports**, readers of research papers can explore alternative analysis options by interacting with the paper itself. This new approach to statistical reporting draws from two recent ideas: **multiverse analysis**, a philosophy of statistical reporting where paper authors report the outcomes of many different statistical analyses in order to show how fragile or robust their findings are; and **explorable explanations**, narratives that can be read as normal explanations but where the reader can also become active by dynamically changing some elements of the explanation.

https://explorablemultiverse.github.io/examples/frequentist/
And so we data wrangle
Herein lies the dirty secret about most data scientists' work -- it's more data munging than deep learning. The best minds of my generation are deleting commas from log files, and that makes me sad. A Ph.D. is a terrible thing to waste.
For Big-Data Scientists, ‘Janitor Work’ Is Key Hurdle to Insights

By STEVE LOHR   AUG. 17, 2014

What you wished data looked like
What it actually looks like

http://healthdesignchallenge.com/
What it actually looks like

https://dev.twitter.com/docs/api/1/get/blocks/blocking

Example Request

```
GET https://api.twitter.com/1/blocks/blocking.json?cursor=-1&include_entities=true
```

```json
{
    "previous_cursor": 0,
    "previous_cursor_str": "0",
    "next_cursor": 0,
    "users": [],
    "profile_sidebar_background_color": "CO8B97",
    "profile_sidebar_border_color": "CO8B97",
    "name": "Javier Heady \r",
    "profile_sidebar_fill_color": "DDEEF4",
    "profile_background_tile": false,
    "location": null,
    "created_at": "Thu Mar 01 00:16:47 +0000 2012",
    "profile_image_url": "http://a0.twimg.com/sticky/default_profile_images/default_profile_1_normal.png",
    "profile_link_color": "0084B4",
    "follow_request_sent": false,
    "contributors_enabled": false,
    "default_profile": true,
    "url": null,
    "favourites_count": 0,
}```
What it actually looks like

---

**ALLERGIES**

**Last Updated:** 01 Dec 2011 @ 0851

**Allergy Name:** TRIMETHOPRIM

**Date Entered:** 09 Mar 2011

**Allergy Type:** DRUG

**Drug Class:** ANTI-INFECTIVES, OTHER

**Observed/Historical:** HISTORICAL

**Comments:** The reaction to this allergy was MILD (NO SQUELAE)

---

**MEDICATION HISTORIC**

**Last Updated:** 11 Apr 2011 @ 1737

**Medication:** AMLODIPINE BESYLATE 10MG TAB

**Instruction:** TAKE ONE TABLET BY MOUTH TAKE ON GRAPEFRUIT JUICE

**Status:** Active

**Refills Remaining:** 3

**Last Filled On:** 20 Aug 2010

**Initially Ordered On:** 13 Aug 2010

**Quantity:** 45

**Days Supply:** 90

**Pharmacy:** DAYTON

**Prescription Number:** 2718953
I'm seeking TRUE, crazy spreadsheet stories. Happy to get the actual sheet or just a description of the crazy. Also: I can keep a secret.
### Enron North America - West Gas

#### November 9, 2001

#### ENA - West Gas Contacts

**Houston Office**
- Barry Tycholiz: (713) 853-1587
- Kim Ward: (713) 853-0685
- Stephanie Miller: (713) 853-1688
- Philip Polsky: (713) 853-5181

**Regional Offices**
- Mark Whitt: (303) 575-6473 Denver
- Paul Lucci: (303) 575-6474 Denver
- Tyrell Harrison: (303) 575-6478 Denver
- Dave Fuller: (503) 464-3732 Portland

#### Forward Prices (US$/MMBtu)

**NYMEX**

<table>
<thead>
<tr>
<th>Settle</th>
<th>Cash</th>
<th>Dec-01</th>
<th>Dec-01 to Mar-02</th>
<th>Apr-02 to Oct-02</th>
<th>Nov-02 to Mar-03</th>
<th>One Year Strip*</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Cash</td>
<td>2.960</td>
<td>3.088</td>
<td>3.166</td>
<td>3.051</td>
<td>3.165</td>
</tr>
<tr>
<td></td>
<td>ROM</td>
<td>1.960</td>
<td>2.088</td>
<td>2.166</td>
<td>2.051</td>
<td>2.165</td>
</tr>
</tbody>
</table>

#### IF CIG Rocky Mountains

<table>
<thead>
<tr>
<th>Contract Month</th>
<th>Fixed Price</th>
<th>Basis</th>
</tr>
</thead>
<tbody>
<tr>
<td>BID</td>
<td>OFFER</td>
<td>BID</td>
</tr>
<tr>
<td>3.165</td>
<td>3.401</td>
<td>3.165</td>
</tr>
</tbody>
</table>

#### IF EL Paso Permian

<table>
<thead>
<tr>
<th>Contract Month</th>
<th>Fixed Price</th>
<th>Basis</th>
</tr>
</thead>
<tbody>
<tr>
<td>BID</td>
<td>OFFER</td>
<td>BID</td>
</tr>
<tr>
<td>3.020</td>
<td>3.301</td>
<td>3.020</td>
</tr>
</tbody>
</table>

#### IF EL Paso San Juan

<table>
<thead>
<tr>
<th>Contract Month</th>
<th>Fixed Price</th>
<th>Basis</th>
</tr>
</thead>
<tbody>
<tr>
<td>BID</td>
<td>OFFER</td>
<td>BID</td>
</tr>
<tr>
<td>3.020</td>
<td>3.301</td>
<td>3.020</td>
</tr>
</tbody>
</table>

#### IF NWPL Canadian Border (Sumas)

<table>
<thead>
<tr>
<th>Contract Month</th>
<th>Fixed Price</th>
<th>Basis</th>
</tr>
</thead>
<tbody>
<tr>
<td>BID</td>
<td>OFFER</td>
<td>BID</td>
</tr>
<tr>
<td>2.980</td>
<td>3.261</td>
<td>2.980</td>
</tr>
</tbody>
</table>

#### IF PEPL TX-OK

<table>
<thead>
<tr>
<th>Contract Month</th>
<th>Fixed Price</th>
<th>Basis</th>
</tr>
</thead>
<tbody>
<tr>
<td>BID</td>
<td>OFFER</td>
<td>BID</td>
</tr>
<tr>
<td>3.041</td>
<td>3.301</td>
<td>3.041</td>
</tr>
</tbody>
</table>

---

*Slide from Jenny Bryan

Today's updates on #otherpeoplesdata:
Where you wish data was
Where data actually is

Where data actually is

https://dev.twitter.com/docs/api/1/get/blocks/blocking
Raw & processed data
“Data are values of qualitative or quantitative variables, belonging to a set of items.”
“Data are values of qualitative or quantitative variables, belonging to a set of items.”

Set of items: Sometimes called the population; the set of objects you are interested in
“Data are values of qualitative or quantitative variables, belonging to a set of items.”

Variables: A measurement or characteristic of an item
“Data are values of **qualitative** or **quantitative** variables, belonging to a set of items.”

**Qualitative:** Country of origin, sex, treatment

**Quantitative:** Height, weight, blood pressure
Data sharing
1. The raw data.

2. A tidy data set

3. A code book describing each variable and its values in the tidy data set.

4. An explicit and exact recipe you used to go from 1 -> 2,3
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.
A tidy data set

- One variable per column
- One observation per row
- One table per "kind" of variable
- Linking indicators for columns

Reference: http://brianknaus.com/software/srtoolbox/s_4_1_sequence80.txt
anything doesn’t make sense.

Files:

1 Demographics: tab 1 is schizophrenia patients, tab 2 is controls. 
A. Cohort: $M = \text{Mannheim (Germany)}, C = \text{Cologne (Germany)}, H= \text{Hopkins}$. We had a few of our own patients so we included them too. 
B. patient identification number 
C. Age at time of CSF collection 
D. Gender 
E. BMI 
F. Ethnicity (mostly Caucasian) 
G. Diagnosis: DSM/ICD-10 diagnosis 
H. Group: control, schizophrenia, or prodromal. I don’t think we have enough power to run them as three groups so I combined prodromal and controls. I am not sure if this was ok. Is it appropriate to do a t-test for $SZ$? 
I. Medication: mostly untreated 
J. Education more or less than 13 years 
K. current smoking status: yes or no
Recipe

R/Python Code

Input raw data -> output tidy

No parameters
Step 1 - take the raw file, run version 3.1.2 of summarize software with parameters a=1, b=2, c=3
Step 2 - run the software separately for each sample
Step 3 - take column three of outputfile.out for each sample and that is the corresponding row in the output data
Step 1 - take the raw file, run version 3.1.2 of summarize software with parameters a=1, b=2, c=3
Step 2 - run the software separately for each sample
Step 3 - take column three of output file.out for each sample and that is the corresponding row in the output data.
<table>
<thead>
<tr>
<th>When...</th>
<th>Be sure to...</th>
<th>So Do this...</th>
<th>Avoid this...</th>
<th>Why?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naming variables</td>
<td>Use meaningful variable names</td>
<td>&quot;AgeAtDiagnosis&quot;</td>
<td>&quot;ADx&quot;</td>
<td>&quot;ADx&quot; is an unclear and uninformative abbreviation</td>
</tr>
<tr>
<td>Naming variables</td>
<td>Avoid spacing in column headers</td>
<td>&quot;AgeAtDiagnosis&quot;</td>
<td>&quot;Age At Diagnosis&quot;</td>
<td>Spacing in variable names makes the analyst's life more difficult</td>
</tr>
<tr>
<td>Naming variables</td>
<td>Use consistent capitalization</td>
<td>&quot;AgeAtDiagnosis&quot;</td>
<td>Using both 'AgeAtDiagnosis' and 'ageAtDiagnosis'</td>
<td>Using consistent column names across tables/spreadsheets simplifies any merging the statistician may have to do.</td>
</tr>
<tr>
<td>Naming variables</td>
<td>Avoid using separators, but if it's necessary, use an underscore ('_')</td>
<td>'IGF1' (or 'IGF_1')</td>
<td>'IGF.1', 'IGF-1', 'IGF/1', '1GF/1', '1GF/1'</td>
<td>Separators (commas, periods, hyphens, slashes, spaces etc.) often have different meanings in coding languages than they do in text. Avoiding them avoids error.</td>
</tr>
<tr>
<td>Coding variables</td>
<td>Avoid unnecessary spaces</td>
<td>'male'</td>
<td>'male '</td>
<td>That extra space after 'male' makes it different from 'male' without a space.</td>
</tr>
<tr>
<td>Coding variables</td>
<td>Be consistent!</td>
<td>'male'</td>
<td>'Male', 'male', and 'M',</td>
<td>In the eyes of the statistician, 'Male', 'male', and 'M' could be incorrectly perceived as three different values.</td>
</tr>
<tr>
<td>Coding variables</td>
<td>Be careful of spelling errors</td>
<td>'male'</td>
<td>'maale'</td>
<td>That extra 'a' makes these two different categories.</td>
</tr>
<tr>
<td>Coding date and time</td>
<td>Use ISO 8601 coding</td>
<td>'YYYY-MM-DD'</td>
<td>'MM/DD/YYYY' and 'Month Day, Year'</td>
<td>Consistency simplifies the analyst's life, and YYYY-MM-DD will not be misconstrued if opened in Excel.</td>
</tr>
<tr>
<td>Coding missing data</td>
<td>Not leave any cells blank and use a consistent value</td>
<td>'NA'</td>
<td>'0', '0', red-highlighted blank cells, ' ', ' ', ' ', ...</td>
<td>Each cell should be filled with a consistent value. Pick a way to denote missingness (ideally 'NA') and stick with it. Avoid using numbers or punctuation to denote missing data.</td>
</tr>
<tr>
<td>Entering data</td>
<td>Stick to text and numbers</td>
<td>Convey all information with direct text/numerical entry</td>
<td>Using cell highlighting or font color to convey information</td>
<td>Your analyst may not use the same platform for analysis as you used for data entry, so avoiding font color and cell highlighting will minimize issues.</td>
</tr>
<tr>
<td>Generating an Excel file</td>
<td>Save the data in an appropriate format</td>
<td>Use one worksheet per table and save as CSV or text files</td>
<td>Multiple worksheets</td>
<td>Statisticians require this format to import your data onto other platforms.</td>
</tr>
<tr>
<td>Entering Data</td>
<td>Avoid entering unnecessary lines of text at the start</td>
<td>Start your first row with variable names</td>
<td>Adding lines of text</td>
<td>This violates the rules of tidy data and makes processing more difficult. Include this information in the &quot;Code book&quot; instead.</td>
</tr>
<tr>
<td>Opening files in Excel</td>
<td>Know and avoid its pitfalls</td>
<td>Consistently include one value per cell and be careful of date and time data.</td>
<td>Using macros, splitting cells, and merging cells</td>
<td>These formats are not amenable to data analysis on other platforms.</td>
</tr>
</tbody>
</table>

Rules for Tidy Spreadsheets

1. Be consistent
2. Choose good names for things
3. Write dates as YYYY-MM-DD
4. No empty cells
5. Put just one thing in a cell
6. Don't use font color or highlighting as data
7. Save the data as plain text files

Organize thyself
"File organization and naming are powerful weapons against chaos."
- Jenny Bryan
<table>
<thead>
<tr>
<th>Allergies</th>
<th>Medication History</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Name</strong>: TRIMETHOPRIM</td>
<td><strong>Medication</strong>: APLIDIPINE BESYLATE 10MG TAB</td>
</tr>
<tr>
<td><strong>Location</strong>: DAYT29</td>
<td><strong>Instructions</strong>: TAKE ONE TABLET BY MOUTH TAKE ONE-HALF TABLET FOR CHILDREN</td>
</tr>
<tr>
<td><strong>Date Entered</strong>: 09 Mar 2011</td>
<td><strong>Status</strong>: Active</td>
</tr>
<tr>
<td><strong>Reaction</strong>:</td>
<td><strong>Refills Remaining</strong>: 3</td>
</tr>
<tr>
<td><strong>Type</strong>: DRUG</td>
<td><strong>Last Filled On</strong>: 20 Aug 2010</td>
</tr>
<tr>
<td><strong>Drug Class</strong>: ANTI-INFECTIVES,OTHER</td>
<td><strong>Initially Ordered On</strong>: 13 Aug 2010</td>
</tr>
<tr>
<td><strong>Observed/Historical</strong>: HISTORICAL</td>
<td><strong>Quantity</strong>: 45</td>
</tr>
<tr>
<td><strong>Comments</strong>: The reaction to this allergy was MILD (NO SQUILAE)</td>
<td><strong>Days Supply</strong>: 90</td>
</tr>
<tr>
<td></td>
<td><strong>Pharmacy</strong>: DAYTON</td>
</tr>
<tr>
<td><strong>Name</strong>: TRAMADOL</td>
<td><strong>Prescription Number</strong>: 2718953</td>
</tr>
<tr>
<td><strong>Location</strong>: DAYT29</td>
<td><strong>Medication</strong>: IBUPROFEN 600MG TAB</td>
</tr>
<tr>
<td><strong>Date Entered</strong>: 09 Mar 2011</td>
<td><strong>Instructions</strong>: TAKE ONE TABLET BY MOUTH FOUR TIMES A DAY WITH FOOD</td>
</tr>
<tr>
<td><strong>Reaction</strong>: URINARY RETENTION</td>
<td><strong>Status</strong>: Active</td>
</tr>
<tr>
<td><strong>Type</strong>: DRUG</td>
<td><strong>Refills Remaining</strong>: 3</td>
</tr>
<tr>
<td><strong>Drug Class</strong>: NON-OPIOID ANALGESICS</td>
<td><strong>Last Filled On</strong>: 20 Aug 2010</td>
</tr>
<tr>
<td><strong>Observed/Historical</strong>: HISTORICAL</td>
<td><strong>Initially Ordered On</strong>: 01 Jul 2010</td>
</tr>
<tr>
<td><strong>Comments</strong>: gradually worsening difficulty emptying bladder</td>
<td><strong>Quantity</strong>: 360</td>
</tr>
</tbody>
</table>
Processed data

- Processed data should be named so it is easy to see which script generated the data.
- The processing script - processed data mapping should occur in the README
- Processed data should be tidy
Raw scripts

- May be less commented (but comments help you!)
- May be multiple versions
- May include analyses that are later discarded
Final scripts

- Clearly commented
  - Small comments liberally - what, when, why, how
  - Bigger commented blocks for whole sections
- Include processing details
This is the README file for my_first_project

Last updated: 02-Mar-2018

The folders in this project are:

- **data** - is the folder where you can find all the collected data.
- **figures** - is where you can find all the plots, data pictures, and other images.
- **code** - is where you can find code files for collecting, cleaning up, or analyzing data.
- **products** - is where you can find reports, presentations, or products

Data on crime is obtained from International Crime Data collected between 2015-2018 and is publicly available. Data on happiness is collected from the Survey of International Happiness.

Contributors:

- Jane Everyday Doe, jane.everyday.doe@gmail.com
- John Everyday Doe, john.everyday.doe@gmail.com

Cite: Doe, J, and Doe, J, Sample Analysis Using Sample Data, Working Paper, 2018
Just no


https://xkcd.com/1459/
key principles of file naming for data science projects:

- Machine readable
- Human readable
- Be nicely ordered

Source: Jenny Bryan
<table>
<thead>
<tr>
<th>Bad Naming</th>
<th>Good Naming</th>
</tr>
</thead>
<tbody>
<tr>
<td>2013 my report.md</td>
<td>2013_my_report.md</td>
</tr>
<tr>
<td>malik's_report.md</td>
<td>maliks_report.md</td>
</tr>
<tr>
<td>01_zoë_report.md</td>
<td>01_zoe_report.md</td>
</tr>
<tr>
<td>AdamHooverReport.md</td>
<td>adam-hoover-report.md</td>
</tr>
<tr>
<td>executivereportpepsiv1.md</td>
<td>executive_report_pepsi_v1.md</td>
</tr>
<tr>
<td>Year</td>
<td>Month</td>
</tr>
<tr>
<td>-------</td>
<td>-------</td>
</tr>
<tr>
<td>2018</td>
<td>jan</td>
</tr>
<tr>
<td>2017</td>
<td>mar</td>
</tr>
<tr>
<td>2016</td>
<td>may</td>
</tr>
<tr>
<td>2017</td>
<td>jan</td>
</tr>
<tr>
<td>2015</td>
<td>oct</td>
</tr>
<tr>
<td>2015</td>
<td>oct</td>
</tr>
</tbody>
</table>
Which one is better?

analysis.R

or

2017-exploratory_analysis_crime.R?
Which one is better?

05-21-2017-analysis-cust001.R
or
2017-05-21-analysis-cust001.R?
Structure of a filename

processed_pvalue_data_from_pubmed_oct24.rda
What did I do to this data

processed_pvalue_data_from_pubmed_oct24.rda
What kind of data is this?

processed_pvalue_data_from_pubmed_oct24.rda
Where did it come from?

`processed_pvalue_data_from_pubmed_oct24.rda`
When did I get it?

processed_pvalue_data_from_pubmed_oct24.rda
Underscores/slashes not dots/whitespace

processed_pvalue_data_from_pubmed_oct24.rda
Consistency is the main rule

processed_pvalue_data_from_pubmed_oct24.rda
raw_pvalue_data_from_pubmed_oct24.rda
Your closest collaborator is you six months ago, but you don’t reply to emails

- Karl Broman

(http://kbroman.org/Tools4RR/assets/lectures/06_org_eda.pdf)
Step 1: slow down and document.
Step 2: have sympathy for your future self.
Step 3: have a system.

- Karl Broman

(http://kbroman.org/Tools4RR/assets/lectures/06_org_eda.pdf)
*squints at the files I was sent*

#otherpeoplesdata
R + RStudio
The R Project for Statistical Computing

Getting Started

R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. To download R, please choose your preferred CRAN mirror.

If you have questions about R like how to download and install the software, or what the license terms are, please read our answers to frequently asked questions before you send an email.

News

- The R Journal Volume 7/1 is available.
- R version 3.2.1 (World-Famous Astronaut) has been released on 2015-06-18.
- R version 3.1.3 (Smooth Sidewalk) has been released on 2015-03-09.
- useR! 2015, will take place at the University of Aalborg, Denmark, June 30 - July 3, 2015.
- useR! 2014, took place at the University of California, Los Angeles, USA June 30 - July 3, 2014.
Welcome to RStudio
Software, education, and services for the R community

Powerful IDE for R
RSstudio IDE is a powerful and productive user interface for R. It's free and open source, and works great on Windows, Mac, and Linux.

R training and education
We've got hands-on courses for beginners and even R experts. Customize an on-site training or enroll in one of our public workshops.

Open source R packages
Our developers and expert trainers are the authors of several popular R packages, including ggplot2, plyr, lubridate, and others.
Manual Line Breaks
End a line with two or more spaces:
Roses are red,
Violets are blue.

Links
Use a plain http address or add a link to a phrase:
http://example.com
[linked phrase](http://example.com)

Images
Images on the web or local files in the same directory:
![](http://example.com/logo.png)
![](figures/img.png)

Blockquotes
A friend once said:
> It's always better to give
> than to receive.

R Code Blocks
R code will be evaluated and printed
```r
summary(rnorm(5))
```
Some useful commands

**Cmd + Enter**

**Ctrl + Enter**

**Ctrl + 1**

**Ctrl + 2**
R packages
Available CRAN Packages By Name

A3: Accurate, Adaptable, and Accessible Error Metrics for Predictive Models
Access to Abbyy Optical Character Recognition (OCR) API
Tools for Approximate Bayesian Computation (ABC)
Computed ABC Analysis
Data Only: Tools for Approximate Bayesian Computation (ABC)
ABCDE_FBA: A-Biologist-Can-Do-Everything of Flux Balance Analysis with this package
Implementation of Artificial Bee Colony (ABC) Optimization
Tools for ABC Analyses
The Analysis of Biological Data
Load Gap-Free Axon ABF2 Files
Combine Multidimensional Arrays
Data Modelling with Additive Bayesian Networks
Abundant regression and high-dimensional principal fitted components
A Package to Processes Accelerometer Data
Functions for Processing Minute-to-Minute Accelerometer Data
Creation and evaluation of Acceptance Sampling Plans
ACC & LMA Graph Plotting
Bayesian Accrual Prediction
Data Quality Visualization Tools for Partially Accruing Data
Categorical data analysis with complete or missing responses
ace() and avas() for selecting regression transformations
install.packages("devtools")
install.packages("dplyr")
### Bioconductor version 3.1 (Release)

Autocomplete biocViews search:

- **Software** (1024)
  - AssayDomain (345)
  - BiologicalQuestion (313)
  - Infrastructure (211)
  - ResearchField (225)
  - StatisticalMethod (293)
  - Technology (645)
  - WorkflowStep (525)
- AnnotationData (883)
- ExperimentData (241)

### Packages found under Software:

<table>
<thead>
<tr>
<th>Package</th>
<th>Maintainer</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>a4</strong></td>
<td>Tobias Verbeke, Willem Ligttenberg</td>
<td>Automated Affymetrix Array Analysis Umbrella Package</td>
</tr>
<tr>
<td><strong>a4Base</strong></td>
<td>Tobias Verbeke, Willem Ligttenberg</td>
<td>Automated Affymetrix Array Analysis Base Package</td>
</tr>
<tr>
<td><strong>a4Classif</strong></td>
<td>Tobias Verbeke, Willem Ligttenberg</td>
<td>Automated Affymetrix Array Analysis Classification Package</td>
</tr>
<tr>
<td><strong>a4Core</strong></td>
<td>Tobias Verbeke, Willem Ligttenberg</td>
<td>Automated Affymetrix Array Analysis Core Package</td>
</tr>
<tr>
<td><strong>a4Preproc</strong></td>
<td>Tobias Verbeke, Willem Ligttenberg</td>
<td>Automated Affymetrix Array Analysis Preprocessing Package</td>
</tr>
<tr>
<td><strong>a4Reporting</strong></td>
<td>Tobias Verbeke, Willem Ligttenberg</td>
<td>Automated Affymetrix Array Analysis Reporting Package</td>
</tr>
<tr>
<td><strong>ABarray</strong></td>
<td>Yongming Andrew Sun</td>
<td>Microarray QA and statistical data analysis for Applied Biosystems Genome Survey Microarray (AB1700) gene expression data.</td>
</tr>
<tr>
<td><strong>ABSSeq</strong></td>
<td>Wentao Yang</td>
<td>ABSSeq: a new RNA-Seq analysis method based on absolute expression differences and generalized Poisson model</td>
</tr>
<tr>
<td><strong>aCGH</strong></td>
<td>Peter Dimitrov</td>
<td>Classes and functions for Array Comparative Genomic Hybridization data.</td>
</tr>
</tbody>
</table>
Surrogate Variable Analysis

Bioconductor version: Release 3.1

The sva package contains functions for removing batch effects and other unwanted variation in high-throughput experiment. Specifically, the sva package contains functions for the identifying and building surrogate variables for high-dimensional data sets. Surrogate variables are covariates constructed directly from high-dimensional data (like gene expression/RNA sequencing/methylation/brain imaging data) that can be used in subsequent analyses to adjust for unknown, unmodeled, or latent sources of noise. The sva package can be used to remove artifacts in three ways: (1) identifying and estimating surrogate variables for unknown sources of variation in high-throughput experiments (Leek and Storey 2007 PLoS Genetics, 2008 PNAS), (2) directly removing known batch effects using ComBat (Johnson et al. 2007 Biostatistics) and (3) removing batch effects with known control probes (Leek 2014 bioXiv). Removing batch effects and using surrogate variables in differential expression analysis have been shown to reduce dependence, stabilize error rate estimates, and improve reproducibility, see (Leek and Storey 2007 PLoS Genetics, 2008 PNAS or Leek et al. 2011 Nat. Reviews Genetics).

Author: Jeffrey T. Leek <jtleek at gmail.com>, W. Evan Johnson <wej at bu.edu>, Hilary S. Parker <hparker at jhsp.h.edu>, Elana J. Fertig <ejfertig at jhmi.edu>, Andrew E. Jaffe <ajaffe at jhsp.h.edu>, John D. Storey <jstorey at princeton.edu>

Maintainer: Jeffrey T. Leek <jtleek at gmail.com>, John D. Storey <jstorey at princeton.edu>, W. Evan Johnson <wej at bu.edu>
Responsiveness

Surrogate Variable Analysis

Bioconductor version: Release 3.1

The sva package contains functions for removing batch effects and other unwanted variation in high-throughput experiment. Specifically, the sva package contains functions for the identifying and building surrogate variables for high-dimensional data sets. Surrogate variables are covariates constructed directly from high-dimensional data (like gene expression/RNA sequencing/methylation/brain imaging data) that can be used in subsequent analyses to adjust for unknown, unmodeled, or latent sources of noise. The sva package can be used to remove artifacts in three ways: (1) identifying and estimating surrogate variables for unknown sources of variation in high-throughput experiments (Leek and Storey 2007 PLoS Genetics, 2008 PNAS, (2) directly removing known batch effects using ComBat (Johnson et al. 2007 Biostatistics) and (3) removing batch effects with known control probes (Leek 2014 bioRxiv). Removing batch effects and using surrogate variables in differential expression analysis have been shown to reduce dependence, stabilize error rate estimates, and improve reproducibility, see (Leek and Storey 2007 PLoS Genetics, 2008 PNAS or Leek et al. 2011 Nat. Reviews Genetics).

Author: Jeffrey T. Leek <jtleek at gmail.com>, W. Evan Johnson <wej at bu.edu>, Hilary S. Parker <hparker at jhsph.edu>, Elana J. Fertig <ejfertig at jhmi.edu>, Andrew E. Jaffe <ajaffe at jhsph.edu>, John D. Storey <jstorey at princeton.edu>

Maintainer: Jeffrey T. Leek <jtleek at gmail.com>, John D. Storey <jstorey at princeton.edu>, W. Evan Johnson <wej at bu.edu>
Surrogate Variable Analysis

Bioconductor version: Release (3.1)

The sva package contains functions for removing batch effects and other unwanted variation in high-throughput experiment. Specifically, the sva package contains functions for identifying and building surrogate variables for high-dimensional data sets. Surrogate variables are covariates constructed directly from high-dimensional data (like gene expression/RNA sequencing/methylation/brain imaging data) that can be used in subsequent analyses to adjust for unknown, unmodeled, or latent sources of noise. The sva package can be used to remove artifacts in three ways: (1) identifying and estimating surrogate variables for unknown sources of variation in high-throughput experiments (Leek and Storey 2007 PLoS Genetics, 2008 PNAS, (2) directly removing known batch effects using ComBat (Johnson et al. 2007 Biostatistics) and (3) removing batch effects with known control probes (Leek 2014 bioXiv).

Removing batch effects and using surrogate variables in differential expression analysis have been shown to reduce dependence, stabilize error rate estimates, and improve reproducibility, see (Leek and Storey 2007 PLoS Genetics, 2008 PNAS or Leek et al. 2011 Nat. Reviews Genetics).

Author: Jeffrey T. Leek <jtleek at gmail.com>, W. Evan Johnson <wej at bu.edu>, Hilary S. Parker <hparker at jnhsph.edu>, Elana J. Fertig <ejfertig at jhmi.edu>, Andrew E. Jaffo <ajaffo at jhsphs.edu>, John D. Storey <jstorey at princeton.edu>

Maintainer: Jeffrey T. Leek <jtleek at gmail.com>, John D. Storey <jstorey at princeton.edu>, W. Evan Johnson <wej at bu.edu>
After version 3.6 of R
install.packages("BiocManager")
BiocManager:::install(c("sva"))

Before version 3.6 of R
source("http://bioconductor.org/biocLite.R")
biocLite("sva")
Convert statistical analysis objects from R into tidy format

latest commit ec5c0bd980

- Merge pull request #51 from zeehio/master
- Overhaul of how augmenting works across many objects. In particular t...
- Add a `tidy` method for x,y,z lists
- Changed `rowwise_df_tidiers` to allow the original data to be saved a...
- Added `gam` to README. Removed rolnames from glmnet output. Few typo ...
- Update cran comments.
- Update cran comments.
- Merge pull request #51 from zeehio/master
Other people like it
People have been working on it
install.packages("remotes")
library(remotes)
install_github("tidymodels/broom")
Average trustworthiness

Bioconductor > R > GitHub
rstudio.cloud tour

https://rstudio.cloud/spaces/77416/projects