Hackathon 2019

- discussion
    - + - initial
        - AI option sounds really hard
            - researchers want specific things
            - could do some basic stuff
            - might even set up system for machine learning
            - go-to visualizations
                - histograms
            - can spit out lots of information
            - information page for each variable
            - work better on smaller datasets
            - size limits for data files
        - want to play with flex
        - like summary table idea
        - data checking component
            - automated
            - ranges
            - histograms
            -
        - reseach component
            - more detailed examination
            - requires more sophisticated evaluaion
        - for large datasets too overwhelming
            - need different approaches
            - can make choice based on number of variables
        - can do both
            - don't want too many junk figrues
        - as AI research problem
        - datasest level- info
        - pick plots based on correlations
        - not data integration or ranking!
            - future hackathon - tools for integrating data
        - interactive part has processing power issues - can't handle huge data
            - shiny may charge for large use
        - could run shiny app locally
            - but may have usability issues
        - shiny info
            - requires an engine
            - can purchase server access - they host
            - can set up locally on server - but may be difficult
            - performance limits
            - stability issues
            - even if go to something else
            - graphs can be complex - groups
            - to run locally need to know how to run a Shiny object - localhost

      - could wrap in a function
  - want to learn enough about what data will work
    - interactive is handy
  - could make decisions about where to put it after alpha-testing
  - users
    - identifying data for team study - use server
    - template - for folks with only a little R
    - comfortable in R
  - hearing support for both multigraph and shiny approaches
  - raw tabular vs metadata
    - see how far tabular can take you
    - then add metadata elements if you can't do otherwise
  - metadata uses?
    - issues of different types of metadata
    - dates - want as POSIXct
    - assume dates are already in appropriate classes
      - maybe not
      - may want to coerce some
    - app could support casts to different classes
    - want to know more about it
    -
- Second
  - research timeline
  - what is typical data
    - can do stats on PASTA
  - ingestion issues
  - start with single rectangular data table
    -
  - for summary data - can set line for what to do or not
  - Wishlist
    - download
    - import
      - with metadata
      - different data models
    - GUI to generate dynamic plots
    - Interactive maps
    - prioritize search results
    - R templates
    - summary table
      - specific datasets
      - across datasets
      - table vs graph
    - data specific
      - graph static
      - dynamic
  - multiple data tables within datapackage

  - also CSVY files - YAML header
- General Group 2
  - what
    - General title of dataset, DOI etc. as header, links etc.
    - units
      - in metadata
    - variable types
    - range - domain - levels
      - min max median mode - R summary
    - number of valid observations
    - correlation matrix for numerical data
      - matrix
      - combined matrix or plot
      - limited by number of pairs
      - correlogram
    - histogram of numerical variables
    - frequency histogram of categorical variables
    - bivariate frequency crosstabulations for categorical
    - boxes by categories for numerical
    - missing - NA
      - how many
      - complete cases
      - graph of where NAs are
    - scatter or line
      - X is date variables
        - check for strings that are really dates
      - bivariate numerical
      - need to be many on each page
    - lattice plots of each numerical variables by each categorical variable
    - location maps - if have lat lon
      - hard - issue semantics
    - appendix of R code used for plots
  - how
    - NA plot
    - visually -dense plots
    - correlogram - correlation plot
    - need table of contents
    - try to keep basic information for a variable on a place
  - if
    - target around 10-20 pages
      - brief summary of overall
      - 2 pages per variable
    - do not do box plots if levels > n
    - correlation matrix limits
- Next Steps

- moved everything to new GITHUB repository
- still not there on data ingestion - really hard
    - pass not just data but metadata
    - based on DOI
    - getting list object to migrate?
        - also subsetting to individual data frame with out it being overwritten
    - might put ISSUE and add BRANCH with DEV example
    - Jason will write into issue
- testing of report
    - stress testing and debugging with a wider array of data
    - some graphics and reports need modification
    - JP will write in as issue
- challenges with nested KNITR for report
    - An will input issue
- changes to GUI broke things
    - need to go through and clean up code for GUI
    - "hard core scrub"
    - spliting things up made it harder to work on.... May bring back together
    - need a development branch to the GUI
    - Kathy & Li will work on
- Some additional functionality needs to be added to the report
    - still need to add in datetime (not just date)
    - general formatting of document
    - An will take charge of
    - still additional work on bringing in metajam
- bring in static report features to GUI
    - full-boat report button
    - datapackage summary page?
        - so far have been working at single data table level
        - may want to stay at that level
        - dataone metadata has stuff on entire data frame
    - Sheila has made issue
- summary output 1 variable at a time displays
    - could be useful for clicking on in external web sites
    - would require adding some web services
    - do we need an additional summary tab at variable level?
    - can support both ways
    - create R markdown report
- length of time to create report is a problem
    - don't want to freeze GUI
    - performance issues
    - could do eager report generation and caching
        - requires trigger when data updated
    - or "lazy" - generate report only when needed

- conventions?
    - formatting etc.
    - depends on what people will be doing after the hackathon
    - add some comments to functions
        - what does function do
        - ROxygen is really useful for this
        - some ROxygen comments have already been added
        - some R files contain multiple functions - do we want to do this?
            - prefer one function per file
    - split functions into single files - except for internal functions within other functions
    - naming conventions
        - summary functions
        - plot functions
        - need to make sure that calls ALSO get renamed :)
        - Colin will work on....
- testing - if report works, individual plots will work
    - start a tests file
    - each function should have one
    - feel free to write your own unit tests
    - Colin will work on
    - TRAVIS is set up for building package and installing and testing
        - can run once a day
        - or everytime new content is pushed to repository
        - can track build by clicking on badge
    - there is also one for PC - see issue that Jason is setting up
- timeline
    - Colin will work on in next week - by next Friday
        - want functioning product by then
    - Colin can assign tasks to individuals
        - but he doesn't want to be intermediary between people
        - An will also communicate needs for help on functions
            - report group
    - someone needs to spend time on downloading
        - Jason
            - try to get done by Monday
        - also work on streamlining
        - Kathy also willing to test
    - Explanatory text
        - Sylvia will work on
    - next meeting?
        - 4 pm EDT on Thursday June 20
- Long view
    - fame and fortune
    - getting integrated into web sites

- Hackathon Review
  - would have like to have come better prepared
  - id skillsets and get presentations
  - GIT was helpful
  - more on
    - package development
    - git
    - shiny
  - but stimulus of hackathon brings out best
  - learn-do-teach helps
    - working in pairs helped
  - first day figuring it out was very useful
    - ground rules - work towards concensus
    - was it too much planning an design?
      - about right
    - also helped figure out how we each communicated
    - one room worked well - even in subgroups
  - group size about right - 8 to 10