Introductory Data Science Course Project: Retrieving Data from an API and Conducting an EDA

Justin Post[1]
jbpost2@ncsu.edu

[1] *Department of Statistics, NC State University*
jbpost2.github.io

Purpose & Goals

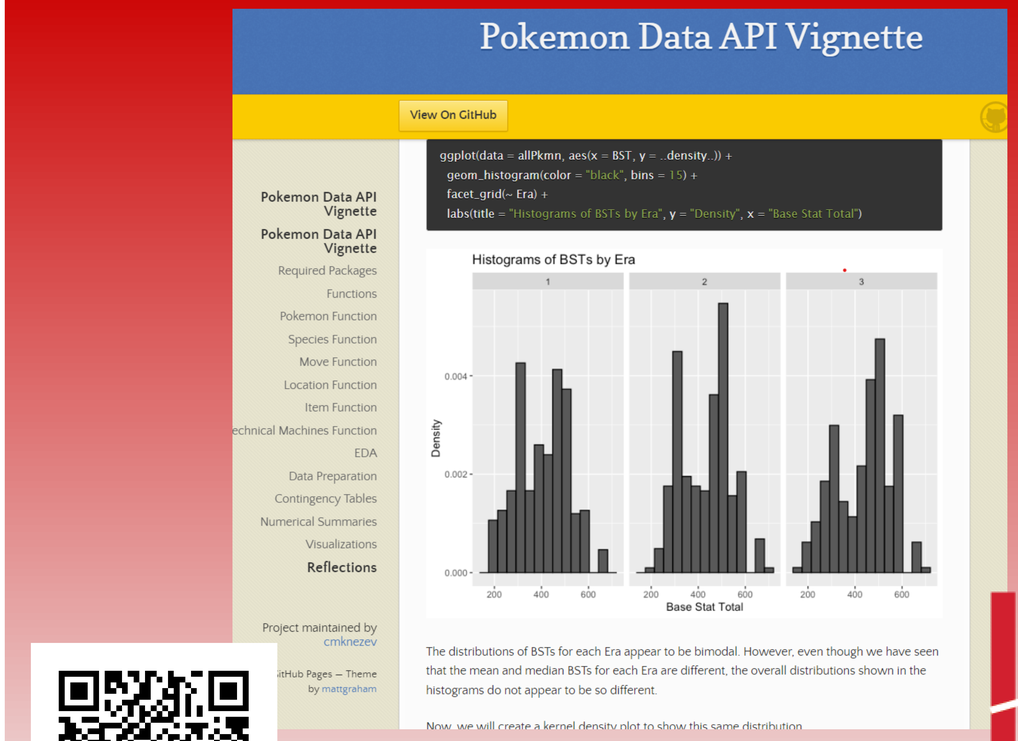
Creating meaningful projects in a data science course can be a time consuming task! With the given example project instructions students:

- must **conceptualize how R data is stored and how to manipulate it** into a useful form
- must be able to **write custom R functions** meant for another user
- may seek out and **find a data source meaningful to them**
- are forced to **think about the type(s) of data** they are downloading and **how they can be summarized** to meet the Exploratory Data Analysis (EDA) requirements
- **use good programming practices** in a larger project setting
- **communicate their code and results**
- (advanced version) create a sleek website to show off their work through RMarkdown and github pages

Project Requirements

- `R` & `RStudio`
- `RMarkdown` to easily create an HTML document with code & output embedded
- `dplyr` (or Base R) for common data manipulation tasks
- `ggplot2` (or Base R) for summarizing the data
- Ability to write custom `R` functions
- `Github` (optional) - for easy creation of a web site to share their work
- Basic lesson on APIs and handling JSON data

Allow students to show off their R skills and communication ability while investigating data that is meaningful to them!



Application Program Interfaces (APIs)

API - think of as a protocol for passing information between computers

- Build URLs to request specific data: `https://api.polygon.io/v2/aggs/ticker/`AAPL`/range/1/day/`2023-01-09`/`2023-01-09`?apiKey=*`
- `httr::GET()` for contacting the API via the URL
- Process `content` element using `rawToChar()`
- Use `jsonlite::fromJSON()` to turn results into lists!:
  - `httr::GET(URL)$content %>% rawToChar() %>% jsonlite::fromJSON()`

Project Instructions

Create a vignette (long form description of how a problem was solved) that provides a narrative for using custom functions to contact an API, parse, and return well-structured data. They then use those functions to obtain data from the API and do some exploratory data analysis.

- Vignette written in R Markdown
- Functions return well-formatted data frames. Requirement:
  - Query at least five different *end points*
  - Not the entire API!
- EDA conducted on resulting data. Requirements:
  - Contingency tables
  - Summary statistics (means, sds, etc.) at levels of categorical variables
  - Bar plot, histogram, box plot, and scatter plot
- Narrative through document explaining process and results
- (Optional) Upload to `github` and use *GitHub pages* to render a sleek looking web page

Resources

Material below available at go.ncsu.edu/uscots2023 (or use the QR code)

- Example course notes available (including API querying)
- Two versions of project instructions
  - Basic version with no github
  - Advanced version that uses github pages to render a web page
  - Simple point rubric included
- Example course projects (thank you to these former students!):
  - spoonacular API (wine and food)
  - College Scorecard API
  - polygon.io Stock Data API
  - Open Movie Database API
  - NASA API
  - Pokemon API

NC STATE UNIVERSITY
DEPARTMENT OF STATISTICS