Sourcing Data for Viz

Nick Sun

May 4, 2019

CFB Data

College football is a pretty popular American past time. The first American football game was played on November 6, 1869 right down to road from where I grew up. The sport has evolved greatly in the paste 160 years and there is now a plethora of analytics related to football of all levels, especially college. As a fan of college football, particularly the Rutgers Scarlet Knights and Oregon State Beavers, I have tons of football related question to ask. Some example visualization questions I want to investigate are:

1. How has scoring in college football changed over time? Do teams on average score more points? Have some parts of the country increased their points per game more than others (this question is partially motivated by midwest Big Ten teams having a reputation of sticking to old school leatherhead football techniques)?
2. How has the passing game developed over time? Are quarterbacks throwing more bombs than in years past?
3. How have defensive turnovers changed over time? Are there some conferences that play better defense than others (partially motivated by Big 12 schools having a reputation for playing Pro Bowl aka flag football style defense)?

I will be using a public API (api.collegefootballdata.com) to get my data. The API is well documented by a developer team that also has an active presence on the subreddit r/cfbanalysis. After playing around with `httr` and `jsonlite`, I found that `jsonlite::fromJSON` can pull the data I need with very little hairpulling.

The endpoints I will be examining will concern play-by-play data. This data has 16 columns, including important variables such as:

- play_type (Rush, Pass, etc.)
- yards_gained (yards is in integers)
- yard_line
- period (1 through 4)
- play_text (actual text taken from ESPN play-by-play describing which player did what action)
- important identifying information for teams involved, who is on offense, which conferences are involved, etc.

The size of the data will depend upon the number of teams and years analyzed. From my investigaions, the data for one team over 10 seasons will be approximately 10MB so the data for all of D1 college football over the same time period will be approximately 1 GB.

An example of play-by-play data is provided below:

```
url <- "https://api.collegefootballdata.com/plays?seasonType=regular&year=2018&week=1&team=Rutgers"

jsontest <- fromJSON(url)

cfbdataexample <- head(jsontest, n = 10)
write.csv(cfbdataexample, "cfbdataexample.csv")

cfbdataexample <- read.csv("cfbdataexample.csv")
pander(cfbdataexample[,1:10])
```

Table 1: Table continues below

X	id	offense	offense_conference	defense
1	4.01e+17	Rutgers	Big Ten	Texas State
2	4.01e+17	Rutgers	Big Ten	Texas State
3	4.01e+17	Rutgers	Big Ten	Texas State
4	4.01e+17	Texas State	Sun Belt	Rutgers
5	4.01e+17	Rutgers	Big Ten	Texas State
6	4.01e+17	Rutgers	Big Ten	Texas State
7	4.01e+17	Rutgers	Big Ten	Texas State
8	4.01e+17	Rutgers	Big Ten	Texas State
9	4.01e+17	Rutgers	Big Ten	Texas State
10	4.01e+17	Rutgers	Big Ten	Texas State

defense_conference	offense_score	defense_score	drive_id	period
Sun Belt	0	0	4.01e+09	1
Sun Belt	0	0	4.01e+09	1
Sun Belt	0	0	4.01e+09	1
Big Ten	0	0	4.01e+09	1
Sun Belt	7	0	4.01e+09	1
Sun Belt	0	0	4.01e+09	1
Sun Belt	0	0	4.01e+09	1
Sun Belt	0	0	4.01e+09	1
Sun Belt	0	0	4.01e+09	1
Sun Belt	0	0	4.01e+09	1

Wildfire Data

The West Coast region of the United States has experienced some devastating wildfires this past year. I had to drive up the I5 as I was moving to Corvallis right after the roads reopened and I got my first look at the wake of some of these massive fires. A dataset found on Kaggle (https://www.kaggle.com/rtatman/188-million-us-wildfires) has information on 188 million US wildfires.

With this available data, we can ask questions like:

1. How has the frequency of wildfires changed over time?
2. How has the total acreage burned changed year over year?
3. How has the cost of fighting wildfires changed year over year?

It would also be interesting to create a chlorepleth map of the United States with number of wildfires as a fill aesthetic.

This data is located within a SQLite database. The schema is based around one table called `fires` and has several dozen columns, including:

- several columns related to location
- FIRE_CODE, a code to compile cost information for wilfires
- FIRE_SIZE in acres
- Cause
- several columns related to discovery date and time
- several columns related to reporting agency, ownership agency, responding agency, etc.

- A Shape column which is an S3 blob containing specific information for GIS maps (this will be dropped since it is difficult to read in as a data.frame)

I plan to pull out the data using the `RSQLite` package. The SQLite database is approximately 175 MB, however if we pull out the main table as a data.frame, it is approximately 1 gigabyte (1.8 million rows with 39 variables).

Here I piped the database connection into the `dplyr::collect()` and created a dataframe from there. The first few rows and columns of the resulting data.frame are provided below:

```
examplefires <- read.csv("examplefires.csv")
pander(examplefires[,8:15])
```

Table 3: Table continues below

NWCG_REPORTING_UNIT_ID	NWCG_REPORTING_UNIT_NAME
USCAPNF	Plumas National Forest
USCAENF	Eldorado National Forest
USCAENF	Eldorado National Forest
USCAENF	Eldorado National Forest
USCAENF	Eldorado National Forest
USCAENF	Eldorado National Forest
USCAENF	Eldorado National Forest
USCASHF	Shasta-Trinity National Forest
USCASHF	Shasta-Trinity National Forest
USCAENF	Eldorado National Forest

Table 4: Table continues below

SOURCE_REPORTING_UNIT	SOURCE_REPORTING_UNIT_NAME	LOCAL_FIRE_REPORT_ID
511	Plumas National Forest	1
503	Eldorado National Forest	13
503	Eldorado National Forest	27
503	Eldorado National Forest	43
503	Eldorado National Forest	44
503	Eldorado National Forest	54
503	Eldorado National Forest	58
514	Shasta-Trinity National Forest	3
514	Shasta-Trinity National Forest	5
503	Eldorado National Forest	61

LOCAL_INCIDENT_ID	FIRE_CODE	FIRE_NAME
PNF-47	BJ8K	FOUNTAIN
13	AAC0	PIGEON
021	A32W	SLACK
6	NA	DEER
7	NA	STEVENOT
8	NA	HIDDEN
9	NA	FORK
02	BK5X	SLATE

LOCAL_INCIDENT_ID	FIRE_CODE	FIRE_NAME
03	BLPQ	SHASTA
10	NA	TANGLEFOOT

Motor Vehicle Accidents

Bodily injury, including injury from motor vehicle crashes, is the leading cause of death among Americans aged 1-44 (source: https://www.cdc.gov/injury/wisqars/overview/key_data.html). Analysis of this data could lead to creating better policies and engineering practices, leading to an overall safer country. This is espeically interesting to me since this summer I'll be working with the Department of Transportation on analyzing accident and safety policy data so this is an opportunity to really work with some data that I might encounter.

Data from traffic accidents for the past three years can be found on the the data.gov website (https://catalog.data.gov/dataset/motor-vehicle-crashes-vehicle-information-beginning-2009). This CSV is almost 300 MB (almost 1.6 million rows) and contains variables such as:

- Year
- Vehicle body type (sedan, SUV, truck, etc.)
- Vehicle make
- Vehicle year
- State of Registration
- Event type (animal, motor vehicle collision, collision with fixed object, submersion, collision with pedestrian, etc.)
- Action taken prior to accident

Some possible questions and visualizations I would like to explore are:

- Distribution of accident types over time
- Distribution of vehicles involved in accidents (perhaps some makes or types are more accident prone than others?)
- Most common actions prior to motor vehicle accidents, or perhaps which accidents those action are most related to
- Chloropleth map of states with the most traffic accidents of a particular type

A small snip of this data is included below.

Table 6: Table continues below

Case.Vehicle.ID	Vehicle.Body.Type	Registration.Class
13364180	SUBURBAN	PASSENGER OR SUBURBAN
13364181	4 DOOR SEDAN	PASSENGER OR SUBURBAN
13364182	4 DOOR SEDAN	PASSENGER OR SUBURBAN
13364283	4 DOOR SEDAN	PASSENGER OR SUBURBAN
13364291	SUBURBAN	PASSENGER OR SUBURBAN
13364292	4 DOOR SEDAN	PASSENGER OR SUBURBAN
13364304	PICKUP TRUCK	PASSENGER OR SUBURBAN
13364305	4 DOOR SEDAN	PASSENGER OR SUBURBAN
13364306	SUBURBAN	PASSENGER OR SUBURBAN
13364307	SUBURBAN	PASSENGER OR SUBURBAN

Table 7: Table continues below

Action.Prior.to.Accident	Type. . . Axles.of.Truck.or.Bus	Direction.of.Travel
Going Straight Ahead	Not Entered	East
Merging	Not Entered	South
Going Straight Ahead	Not Entered	South
Going Straight Ahead	Not Entered	North
Changing Lanes	Not Entered	West
Going Straight Ahead	Not Entered	West
Going Straight Ahead	Not Entered	East
Going Straight Ahead	Not Entered	North
Going Straight Ahead	Not Entered	North
Going Straight Ahead	Not Entered	North

Fuel.Type	Vehicle.Year
Gas	2015
Gas	2007
Gas	2009
Gas	2007
Gas	2012
Gas	2013
Gas	2014
Gas	2007
Gas	2004
Gas	2011