Data Documentation

Nicolas Restrepo

Contents

Here, I will keep track of some cool datasets to use in the course.

UN Votes

This is a dataset that provides information about how countries have voted in the UN National Assembly. The data comes neatly bundled in a package that you can get by running `install.packages("unvotes")`.

The package contains three datasets:

1) `un_votes` contains all the decisions for each resolution and each country.

```
# Load the packages
library(unvotes)
library(tidyverse)

# Eagle eye view of the data
glimpse(un_votes)
```

```
## Rows: 869,937
## Columns: 4
## $ rcid         <dbl> 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, ~
```

```
## $ country      <chr> "United States", "Canada", "Cuba", "Haiti", "Dominican Re~
## $ country_code <chr> "US", "CA", "CU", "HT", "DO", "MX", "GT", "HN", "SV", "NI~
## $ vote         <fct> yes, no, yes, yes, yes, yes, yes, yes, yes, yes, yes, yes~
```

2) `un_roll_calls` includes information about each roll call: what the vote was about, when it happened, and what was being discussed:

```
glimpse(un_roll_calls)
```

```
## Rows: 6,202
## Columns: 9
## $ rcid          <int> 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18,~
## $ session       <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,~
## $ importantvote <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ date          <date> 1946-01-01, 1946-01-02, 1946-01-04, 1946-01-04, 1946-01~
## $ unres         <chr> "R/1/66", "R/1/79", "R/1/98", "R/1/107", "R/1/295", "R/1~
## $ amend         <int> 1, 0, 0, 0, 1, 1, 0, 1, 0, 1, 1, 1, 1, 1, 1, 1, 0, 0, 1,~
## $ para          <int> 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0,~
## $ short         <chr> "AMENDMENTS, RULES OF PROCEDURE", "SECURITY COUNCIL ELEC~
## $ descr         <chr> "TO ADOPT A CUBAN AMENDMENT TO THE UK PROPOSAL REFERRING~
```

3) `un_roll_call_issues` relates each vote to broader themes.

```
# View data
glimpse(un_roll_call_issues)
```

```
## Rows: 5,745
## Columns: 3
## $ rcid       <int> 77, 9001, 9002, 9003, 9004, 9005, 9006, 128, 129, 130, 131,~
## $ short_name <chr> "me", "me", "me", "me", "me", "me", "me", "me", "me", "me",~
## $ issue      <fct> Palestinian conflict, Palestinian conflict, Palestinian con~
```

```
# How many votes per issue?
un_roll_call_issues %>%
  count(issue, sort = T)
```

```
## # A tibble: 6 x 2
##   issue                                  n
##   <fct>                              <int>
## 1 Arms control and disarmament        1092
## 2 Palestinian conflict                1061
## 3 Human rights                        1015
## 4 Colonialism                          957
## 5 Nuclear weapons and nuclear material  855
## 6 Economic development                  765
```

Causact

The `causact` package is designed to help with the construction and use of causal models. However, it also contains some interesting datasets. Let's highlight a couple.

As usual, you can get this package by running `install.packages("causact")`

Baseball

`baseballData` contains the final scores of 12,145 baseball games played in 2010.

```
library(causact)
glimpse(baseballData)
```

```
## Rows: 12,145
## Columns: 5
## $ Date         <int> 20100405, 20100405, 20100405, 20100405, 20100405, 2010040~
## $ Home         <fct> ANA, CHA, KCA, OAK, TEX, ARI, ATL, CIN, HOU, MIL, NYN, PI~
## $ Visitor      <fct> MIN, CLE, DET, SEA, TOR, SDN, CHN, SLN, SFN, COL, FLO, LA~
## $ HomeScore    <int> 6, 6, 4, 3, 5, 6, 16, 6, 2, 3, 7, 11, 1, 3, 4, 2, 4, 3, 0~
## $ VisitorScore <int> 3, 0, 8, 5, 4, 3, 5, 11, 5, 5, 1, 5, 11, 5, 6, 1, 3, 6, 3~
```

Corruption

If you are more into international relations and development, `corruptDF` contains information about the human development index and the corruption perception index for different countries.

```
glimpse(corruptDF)
```

```
## Rows: 174
## Columns: 7
## $ country     <chr> "Afghanistan", "Albania", "Algeria", "Angola", "Argentina"~
## $ region      <chr> "Asia Pacific", "East EU Cemt Asia", "MENA", "SSA", "Ameri~
## $ countryCode <chr> "AFG", "ALB", "DZA", "AGO", "ARG", "ARM", "AUS", "AUT", "A~
## $ regionCode  <chr> "AP", "ECA", "MENA", "SSA", "AME", "ECA", "AP", "WE/EU", "~
## $ population  <int> 35530081, 2873457, 41318142, 29784193, 44271041, 2930450, ~
## $ CPI2017     <int> 15, 38, 33, 19, 39, 35, 77, 75, 31, 65, 36, 28, 68, 44, 75~
## $ HDI2017     <dbl> 0.498, 0.785, 0.754, 0.581, 0.825, 0.755, 0.939, 0.908, 0.~
```

NYC Tickets

`ticketsDF` contains observations of the number of tickets written by NYC precincts each day between 2014 and 2015.

```
glimpse(ticketsDF)
```

```
## Rows: 55,167
## Columns: 4
## $ precinct      <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,~
## $ date          <date> 2014-01-01, 2014-01-02, 2014-01-03, 2014-01-04, 2014-01~
## $ month_year    <chr> "01-2014", "01-2014", "01-2014", "01-2014", "01-2014", "~
## $ daily_tickets <int> 17, 67, 17, 42, 54, 109, 115, 119, 154, 95, 38, 37, 91, ~
```

Malaria

If you are interested in public health, this could be of interest. In the folder, you will find some datasets compiled from the `malariaAtlas` package. There are three datasets here. The first contais the incidence of Malaria for each country each year.

You can find more information here.

```
df <- read_csv("malaria_incidence.csv")
glimpse(df)
```

```
## Rows: 508
## Columns: 4
## $ Entity                                                                    <chr> ~
## $ Code                                                                      <chr> ~
## $ Year                                                                      <dbl> ~
## $ `Incidence of malaria (per 1,000 population at risk) (per 1,000 population at risk)` <dbl> ~
```

The second dataset contains the deaths of malaria standardized across all ages for a given country-year.

```
df <- read_csv("malaria_deaths_standard.csv")
glimpse(df)
```

```
## Rows: 6,156
## Columns: 4
## $ Entity                                                                          <chr> ~
## $ Code                                                                            <chr> ~
## $ Year                                                                            <dbl> ~
## $ `Deaths - Malaria - Sex: Both - Age: Age-standardized (Rate) (per 100,000 people)` <dbl> ~
```

Finally, we have a dataset of malaria deaths but among specific age groups.

```
df <- read_csv("malaria_deaths_age_specific.csv")
glimpse(df)
```

```
## Rows: 30,780
## Columns: 6
## $ ...1      <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 1~
## $ entity    <chr> "Afghanistan", "Afghanistan", "Afghanistan", "Afghanistan", ~
## $ code      <chr> "AFG", "AFG", "AFG", "AFG", "AFG", "AFG", "AFG", "AFG", "AFG~
## $ year      <dbl> 1990, 1991, 1992, 1993, 1994, 1995, 1996, 1997, 1998, 1999, ~
## $ age_group <chr> "Under 5", "Under 5", "Under 5", "Under 5", "Under 5", "Unde~
## $ deaths    <dbl> 184.6064, 191.6582, 197.1402, 207.3578, 226.2094, 236.3280, ~
```

NFL Stats (2000-2017)

If you are interested in Football, we have a a dataset of player stats in the NFL from 2000 to 2017.

You can find more information here.

```
df <- read_csv("nfl_stats.csv")
glimpse(df)
```

```
## Rows: 81,525
## Columns: 23
## $ ...1         <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17~
## $ name         <chr> "Duce Staley", "Lamar Smith", "Tiki Barber", "Stephen Dav~
## $ team         <chr> "PHI", "MIA", "NYG", "WAS", "IND", "BAL", "NYJ", "MIN", "~
## $ game_year    <dbl> 2000, 2000, 2000, 2000, 2000, 2000, 2000, 2000, 2000, 200~
## $ game_week    <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ~
## $ rush_att     <dbl> 26, 27, 13, 23, 28, 27, 30, 14, 15, 10, 20, 13, 23, 14, 2~
## $ rush_yds     <dbl> 201, 145, 144, 133, 124, 119, 110, 109, 88, 87, 84, 80, 7~
## $ rush_avg     <dbl> 7.7, 5.4, 11.1, 5.8, 4.4, 4.4, 3.7, 7.8, 5.9, 8.7, 4.2, 6~
## $ rush_tds     <dbl> 1, 1, 2, 1, 1, 0, 1, 0, 0, 1, 0, 0, 1, 1, 0, 0, 1, 1, 3, ~
## $ rush_fumbles <dbl> 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 2, 0, 1, 1, ~
## $ rec          <dbl> 4, 1, 3, 4, 6, 4, 6, 2, 2, NA, 4, 3, 1, 4, 1, 1, 1, NA, N~
## $ rec_yds      <dbl> 61, 12, 25, 37, 40, 32, 34, 3, 20, NA, 29, 10, -2, 100, 1~
## $ rec_avg      <dbl> 15.3, 12.0, 8.3, 9.3, 6.7, 8.0, 5.7, 1.5, 10.0, NA, 7.3, ~
## $ rec_tds      <dbl> 0, 0, 0, 0, 1, 0, 1, 0, 0, NA, 0, 0, 0, 1, 0, 0, 0, NA, N~
## $ rec_fumbles  <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, NA, 0, 0, 0, 0, 0, 0, 0, NA, N~
## $ pass_att     <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, 41, NA, NA, NA, NA, N~
## $ pass_yds     <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, 290, NA, NA, NA, NA, ~
## $ pass_tds     <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, 2, NA, NA, NA, NA, NA~
## $ int          <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, 0, NA, NA, NA, NA, NA~
## $ sck          <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, 2, NA, NA, NA, NA, NA~
## $ pass_fumbles <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, 0, NA, NA, NA, NA, NA~
```

```
## $ rate         <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, 102.7, NA, NA, NA, NA~
## $ position     <chr> "RB", "RB", "RB", "RB", "RB", "RB", "RB", "RB", "RB", "QB~
```

Maybe you are interested in how the salaries for different positions have evolved across the years. We have you covered.

```
df <- read_csv("nfl_salaries.csv")
glimpse(df)
```

```
## Rows: 800
## Columns: 11
## $ year                <dbl> 2011, 2011, 2011, 2011, 2011, 2011, 2011, 2011, 20~
## $ Cornerback          <dbl> 11265916, 11000000, 10000000, 10000000, 10000000, ~
## $ `Defensive Lineman` <dbl> 17818000, 16200000, 12476000, 11904706, 11762782, ~
## $ Linebacker          <dbl> 16420000, 15623000, 11825000, 10083333, 10020000, ~
## $ `Offensive Lineman` <dbl> 15960000, 12800000, 11767500, 10358200, 10000000, ~
## $ Quarterback         <dbl> 17228125, 16000000, 14400000, 14100000, 13510000, ~
## $ `Running Back`      <dbl> 12955000, 10873833, 9479000, 7700000, 7500000, 703~
## $ Safety              <dbl> 8871428, 8787500, 8282500, 8000000, 7804333, 76527~
## $ `Special Teamer`    <dbl> 4300000, 3725000, 3556176, 3500000, 3250000, 32250~
## $ `Tight End`         <dbl> 8734375, 8591000, 8290000, 7723333, 6974666, 61333~
## $ `Wide Receiver`     <dbl> 16250000, 14175000, 11424000, 11415000, 10800000, ~
```

California Fires

Maybe you are interested in climate change. Then, the number of California fires across years, and the damaged they have caused, might be of interest.

You can find more information here.

```
df <- read_csv("fires.csv")
glimpse(df)
```

```
## Rows: 83
## Columns: 4
## $ YEAR              <dbl> 1933, 1934, 1935, 1936, 1937, 1938, 1939, 1940, 1941~
## $ `NUMBER OF FIRES` <dbl> 1994, 2338, 1447, 3805, 2907, 4150, 2491, 4497, 5460~
## $ `ACRES BURNED`    <dbl> 129210, 363052, 127262, 756696, 71312, 221061, 51362~
## $ `DOLLAR DAMAGE`   <dbl> 318636, 563710, 165543, 1877147, 151584, 404225, 847~
```

TV & Movie Ratings

We also have a dataset that contains TV & movie ratings taken from IMDB. The dataset covers titles released between 2000 and 2019.

You can find more information here.

```
df <- read_csv("tv_ratings.csv")
glimpse(df)
```

```
## Rows: 2,266
## Columns: 7
## $ titleId      <chr> "tt2879552", "tt3148266", "tt3148266", "tt3148266", "tt31~
## $ seasonNumber <dbl> 1, 1, 2, 3, 4, 1, 2, 1, 2, 3, 4, 5, 6, 7, 8, 1, 1, 1, 1, ~
## $ title        <chr> "11.22.63", "12 Monkeys", "12 Monkeys", "12 Monkeys", "12~
## $ date         <date> 2016-03-10, 2015-02-27, 2016-05-30, 2017-05-19, 2018-06-~
## $ av_rating    <dbl> 8.4890, 8.3407, 8.8196, 9.0369, 9.1363, 8.4370, 7.5089, 8~
## $ share        <dbl> 0.51, 0.46, 0.25, 0.19, 0.38, 2.38, 2.19, 6.67, 7.13, 5.8~
```

```
## $ genres       <chr> "Drama,Mystery,Sci-Fi", "Adventure,Drama,Mystery", "Adven~
```

Incarceration Trends

Here's a data about the people incarcerated at the county level from 1970 to 2015. This is a big dataset so it might take a while to load.

You can find more information here.

```
df <- read_csv("incarceration_trends.csv")
glimpse(df)
```

```
## Rows: 1,327,797
## Columns: 9
## $ year              <dbl> 1970, 1971, 1972, 1973, 1974, 1975, 1976, 1977, 1978~
## $ state             <chr> "AL", "AL", "AL", "AL", "AL", "AL", "AL", "AL", "AL"~
## $ county_name       <chr> "Autauga County", "Autauga County", "Autauga County"~
## $ urbanicity        <chr> "small/mid", "small/mid", "small/mid", "small/mid", ~
## $ region            <chr> "South", "South", "South", "South", "South", "South"~
## $ division          <chr> "East South Central", "East South Central", "East So~
## $ pop_category      <chr> "Total", "Total", "Total", "Total", "Total", "Total"~
## $ population        <dbl> 14154, 14765, 15939, 16906, 17578, 18007, 18476, 190~
## $ prison_population <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
```

Grand Slams Timeline

If you like tennis, this could be interesting. Here's data about participation in different Grand Slams. Every observation is a player and we have information about how they did in the tournament.

You can find more information here.

```
df <- read_csv("grand_slam_timeline.csv")

glimpse(df)
```

```
## Rows: 12,605
## Columns: 5
## $ player     <chr> "Margaret Court", "Billie Jean Moffitt King", "Maria Bueno"~
## $ year       <dbl> 1968, 1968, 1968, 1968, 1968, 1968, 1968, 1968, 1968, 1968,~
## $ tournament <chr> "Australian Open", "Australian Open", "Australian Open", "A~
## $ outcome    <chr> "Finalist", "Won", "Absent", "Absent", "Absent", "Semi-fina~
## $ gender     <chr> "Female", "Female", "Female", "Female", "Female", "Female",~
```

Emplyoment by Gender

This is an interesting dataset about patterns of emplyoment, across gender, for different occupations. The dataset includes information about the type of occupational group and the share of employees who are either female or male. There is also information about the earnings for each group.

```
df <- read_csv("gender_employment.csv")

glimpse(df)
```

```
## Rows: 2,088
## Columns: 12
## $ year                  <dbl> 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013, ~
## $ occupation            <chr> "Chief executives", "General and operations mana~
## $ major_category        <chr> "Management, Business, and Financial", "Manageme~
```

```
## $ minor_category        <chr> "Management", "Management", "Management", "Manag~
## $ total_workers         <dbl> 1024259, 977284, 14815, 43015, 754514, 44198, 10~
## $ workers_male          <dbl> 782400, 681627, 8375, 17775, 440078, 16141, 7287~
## $ workers_female        <dbl> 241859, 295657, 6440, 25240, 314436, 28057, 3683~
## $ percent_female        <dbl> 23.6, 30.3, 43.5, 58.7, 41.7, 63.5, 33.6, 27.5, ~
## $ total_earnings        <dbl> 120254, 73557, 67155, 61371, 78455, 74114, 62187~
## $ total_earnings_male   <dbl> 126142, 81041, 71530, 75190, 91998, 90071, 66579~
## $ total_earnings_female <dbl> 95921, 60759, 65325, 55860, 65040, 66052, 55079,~
## $ wage_percent_of_male  <dbl> 76.04208, 74.97316, 91.32532, 74.29179, 70.69719~
```

You can find more information here.

Soccer Women's World Cup

This one is close to my heart; it contains information about the outcomes of all World Cup games in the women's game.

You can see more information about this dataset here.

There is *free* play-to-play data for the last World Cup here, if you are interested in more detailed analysis.

```
df <- read_csv("wwc_outcomes.csv")
glimpse(df)
```

```
## Rows: 568
## Columns: 7
## $ year           <dbl> 1991, 1991, 1991, 1991, 1991, 1991, 1991, 1991, 1991, 1~
## $ team           <chr> "CHN", "NOR", "DEN", "NZL", "JPN", "BRA", "GER", "NGA",~
## $ score          <dbl> 4, 0, 3, 0, 0, 1, 4, 0, 2, 3, 0, 5, 4, 0, 2, 2, 0, 8, 1~
## $ round          <chr> "Group", "Group", "Group", "Group", "Group", "Group", "~
## $ yearly_game_id <dbl> 1, 1, 2, 2, 3, 3, 4, 4, 5, 5, 6, 6, 7, 7, 8, 8, 9, 9, 1~
## $ team_num       <dbl> 1, 2, 1, 2, 1, 2, 1, 2, 1, 2, 1, 2, 1, 2, 1, 2, 1, 2, 1~
## $ win_status     <chr> "Won", "Lost", "Won", "Lost", "Lost", "Won", "Won", "Lo~
```

UFO Sightings

Just a wild resource. More that 80,000 recorded UFO sightings including the length of the encounter and what shape the supposed flying object was. An incredible testament to human creativity. The dataset is big so I am not including it in the repo directly. Instead, I am going to show you how to load it directly from Github.

You can find more information here.

```
ufo_sightings <- readr::read_csv("https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/

glimpse(ufo_sightings)
```

```
## Rows: 80,332
## Columns: 11
## $ date_time                  <chr> "10/10/1949 20:30", "10/10/1949 21:00", "10~
## $ city_area                  <chr> "san marcos", "lackland afb", "chester (uk/~
## $ state                      <chr> "tx", "tx", NA, "tx", "hi", "tn", NA, "ct",~
## $ country                    <chr> "us", NA, "gb", "us", "us", "us", "gb", "us~
## $ ufo_shape                  <chr> "cylinder", "light", "circle", "circle", "l~
## $ encounter_length           <dbl> 2700, 7200, 20, 20, 900, 300, 180, 1200, 18~
## $ described_encounter_length <chr> "45 minutes", "1-2 hrs", "20 seconds", "1/2~
## $ description                <chr> "This event took place in early fall around~
## $ date_documented            <chr> "4/27/2004", "12/16/2005", "1/21/2008", "1/~
```

```
## $ latitude                   <dbl> 29.88306, 29.38421, 53.20000, 28.97833, 21.~
## $ longitude                  <dbl> -97.941111, -98.581082, -2.916667, -96.6458~
```

Nobel Prize Winners

Here's a dataset that compiles all Nobel prize winners. It contains information about their discipline, alma mater, what the recognition was for, and much more.

You can find more information here.

```
df <- read_csv("nobel_winners.csv")
glimpse(df)
```

```
## Rows: 969
## Columns: 18
## $ prize_year           <dbl> 1901, 1901, 1901, 1901, 1901, 1901, 1902, 1902, 1~
## $ category             <chr> "Chemistry", "Literature", "Medicine", "Peace", "~
## $ prize                <chr> "The Nobel Prize in Chemistry 1901", "The Nobel P~
## $ motivation           <chr> "\"in recognition of the extraordinary services h~
## $ prize_share          <chr> "1/1", "1/1", "1/1", "1/2", "1/2", "1/1", "1/1", ~
## $ laureate_id          <dbl> 160, 569, 293, 462, 463, 1, 161, 571, 294, 464, 4~
## $ laureate_type        <chr> "Individual", "Individual", "Individual", "Indivi~
## $ full_name            <chr> "Jacobus Henricus van 't Hoff", "Sully Prudhomme"~
## $ birth_date           <date> 1852-08-30, 1839-03-16, 1854-03-15, 1828-05-08, ~
## $ birth_city           <chr> "Rotterdam", "Paris", "Hansdorf (Lawice)", "Genev~
## $ birth_country        <chr> "Netherlands", "France", "Prussia (Poland)", "Swi~
## $ gender               <chr> "Male", "Male", "Male", "Male", "Male", "Male", "~
## $ organization_name    <chr> "Berlin University", NA, "Marburg University", NA~
## $ organization_city    <chr> "Berlin", NA, "Marburg", NA, NA, "Munich", "Berli~
## $ organization_country <chr> "Germany", NA, "Germany", NA, NA, "Germany", "Ger~
## $ death_date           <date> 1911-03-01, 1907-09-07, 1917-03-31, 1910-10-30, ~
## $ death_city           <chr> "Berlin", "Châtenay", "Marburg", "Heiden", "Paris~
## $ death_country        <chr> "Germany", "France", "Germany", "Switzerland", "F~
```

Rap artists

In 2020, the BBC asked about 100 critics, artists, and music industry folks about their top 5 hip hop tracks. This dataset contains the aggregated results.

You can find more information here.

```
df <- read_csv("rap_artists.csv")
glimpse(df)
```

```
## Rows: 311
## Columns: 12
## $ ID     <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, ~
## $ title  <chr> "Juicy", "Fight The Power", "Shook Ones (Part II)", "The Messag~
## $ artist <chr> "The Notorious B.I.G.", "Public Enemy", "Mobb Deep", "Grandmast~
## $ year   <dbl> 1994, 1989, 1995, 1982, 1992, 1993, 1993, 1992, 1994, 1995, 201~
## $ gender <chr> "male", "male", "male", "male", "male", "male", "male", "male",~
## $ points <dbl> 140, 100, 94, 90, 84, 62, 50, 48, 46, 42, 38, 36, 36, 34, 32, 3~
## $ n      <dbl> 18, 11, 13, 14, 14, 10, 7, 6, 7, 6, 5, 5, 4, 6, 5, 5, 4, 5, 5, ~
## $ n1     <dbl> 9, 7, 4, 5, 2, 3, 2, 3, 1, 2, 2, 1, 2, 1, 1, 0, 2, 2, 1, 1, 1, ~
## $ n2     <dbl> 3, 3, 5, 3, 4, 1, 2, 2, 3, 1, 0, 1, 2, 0, 1, 3, 1, 0, 1, 1, 1, ~
## $ n3     <dbl> 3, 1, 1, 1, 2, 1, 2, 0, 1, 1, 3, 3, 0, 2, 2, 1, 0, 1, 1, 1, 0, ~
## $ n4     <dbl> 1, 0, 1, 0, 4, 4, 0, 0, 1, 2, 0, 0, 0, 3, 0, 0, 1, 0, 1, 0, 2, ~
```

```
## $ n5     <dbl> 2, 0, 2, 5, 2, 1, 1, 1, 1, 0, 0, 0, 0, 0, 1, 1, 0, 2, 1, 2, 1, ~
```

The NCAA Women's Basketball Tournament

This dataset contains detailed information about how each of the teams that has qualified to the tournament since 1982 has done.

More information here.

```
df <- read_csv("wncaa.csv")
glimpse(df)
```

```
## Rows: 2,092
## Columns: 19
## $ year              <dbl> 1982, 1982, 1982, 1982, 1982, 1982, 1982, 1982, 1982~
## $ school            <chr> "Arizona St.", "Auburn", "Cheyney", "Clemson", "Drak~
## $ seed              <dbl> 4, 7, 2, 5, 4, 6, 5, 8, 7, 7, 4, 8, 2, 1, 1, 2, 3, 6~
## $ conference        <chr> "Western Collegiate", "Southeastern", "Independent",~
## $ conf_w            <dbl> NA, NA, NA, 6, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
## $ conf_l            <dbl> NA, NA, NA, 3, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
## $ conf_percent      <dbl> NA, NA, NA, 66.7, NA, NA, NA, NA, NA, NA, NA, NA, NA~
## $ conf_place        <chr> "-", "-", "-", "4th", "-", "-", "-", "-", "-", "-", ~
## $ reg_w             <dbl> 23, 24, 24, 20, 26, 19, 21, 14, 21, 28, 24, 17, 22, ~
## $ reg_l             <dbl> 6, 4, 2, 11, 6, 7, 8, 10, 8, 7, 5, 13, 7, 5, 1, 6, 4~
## $ reg_percent       <dbl> 79.3, 85.7, 92.3, 64.5, 81.3, 73.1, 72.4, 58.3, 72.4~
## $ how_qual          <chr> "at-large", "at-large", "at-large", "at-large", "aut~
## $ x1st_game_at_home <chr> "Y", "N", "Y", "N", "Y", "N", "N", "N", "N", "N", "Y~
## $ tourney_w         <dbl> 1, 0, 4, 0, 2, 0, 0, 0, 0, 0, 2, 0, 2, 1, 5, 3, 1, 1~
## $ tourney_l         <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1~
## $ tourney_finish    <chr> "RSF", "1st", "N2nd", "1st", "RF", "1st", "1st", "1s~
## $ full_w            <dbl> 24, 24, 28, 20, 28, 19, 21, 14, 21, 28, 26, 17, 24, ~
## $ full_l            <dbl> 7, 5, 3, 12, 7, 8, 9, 11, 9, 8, 6, 14, 8, 6, 1, 7, 5~
## $ full_percent      <dbl> 77.4, 82.8, 90.3, 62.5, 80.0, 70.4, 70.0, 56.0, 70.0~
```

Tour de France

Here's some cool data about the biggest cycling competition in the world. We have two datasets `tour_winners` and `tour_stages`. The former gives details about each tour overall winner. The latter gives information about each specific stage: its length, origin, destination, and winner.

You can find more information here

```
df <- read_csv("tour_winners.csv")
glimpse(df)
```

```
## Rows: 106
## Columns: 19
## $ edition       <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 1~
## $ start_date    <date> 1903-07-01, 1904-07-02, 1905-07-09, 1906-07-04, 1907-07~
## $ winner_name   <chr> "Maurice Garin", "Henri Cornet", "Louis Trousselier", "R~
## $ winner_team   <chr> "La Française", "Conte", "Peugeot-Wolber", "Peugeot-Wolb~
## $ distance      <dbl> 2428, 2428, 2994, 4637, 4488, 4497, 4498, 4734, 5343, 52~
## $ time_overall  <dbl> 94.55389, 96.09861, NA, NA, NA, NA, NA, NA, NA, NA, 197.~
## $ time_margin   <dbl> 2.98916667, 2.27055556, NA, NA, NA, NA, NA, NA, NA, NA, ~
## $ stage_wins    <dbl> 3, 1, 5, 5, 2, 5, 6, 4, 2, 3, 1, 1, 1, 4, 2, 0, 3, 4, 4,~
## $ stages_led    <dbl> 6, 3, 10, 12, 5, 13, 13, 3, 13, 13, 8, 15, 2, 14, 14, 3,~
## $ height        <dbl> 1.62, NA, NA, NA, NA, NA, 1.78, NA, NA, NA, NA, NA, NA, ~
```

```
## $ weight        <dbl> 60, NA, NA, NA, NA, NA, 88, NA, NA, NA, NA, NA, NA, NA, ~
## $ age           <dbl> 32, 19, 24, 27, 24, 25, 22, 22, 26, 23, 23, 24, 33, 30, ~
## $ born          <date> 1871-03-03, 1884-08-04, 1881-06-29, 1879-06-05, 1882-10~
## $ died          <date> 1957-02-19, 1941-03-18, 1939-04-24, 1907-01-25, 1917-12~
## $ full_name     <chr> NA, NA, NA, NA, "Lucien Georges Mazan", "Lucien Georges ~
## $ nickname      <chr> "The Little Chimney-sweep", "Le rigolo (The joker)", "Le~
## $ birth_town    <chr> "Arvier", "Desvres", "Paris", "Moret-sur-Loing", "Plessé~
## $ birth_country <chr> "Italy", "France", "France", "France", "France", "France~
## $ nationality   <chr> " France", " France", " France", " France", " France", "~
```

```
df <- read_csv("tour_stages.csv")
glimpse(df)
```

```
## Rows: 2,236
## Columns: 8
## $ Stage          <chr> "1", "2", "3", "4", "5", "6", "7", "8", "9", "10", "11"~
## $ Date           <date> 2017-07-01, 2017-07-02, 2017-07-03, 2017-07-04, 2017-0~
## $ Distance       <dbl> 14.0, 203.5, 212.5, 207.5, 160.5, 216.0, 213.5, 187.5, ~
## $ Origin         <chr> "Düsseldorf", "Düsseldorf", "Verviers", "Mondorf-les-Ba~
## $ Destination    <chr> "Düsseldorf", "Liège", "Longwy", "Vittel", "La Planche ~
## $ Type           <chr> "Individual time trial", "Flat stage", "Medium mountain~
## $ Winner         <chr> "Geraint Thomas", "Marcel Kittel", "Peter Sagan", "Arna~
## $ Winner_Country <chr> "GBR", "GER", "SVK", "FRA", "ITA", "GER", "GER", "FRA",~
```

Government Spending on Kids

Here's one for the folks interested in public policy. Here's information about government spending on policy directed at children. The dataset includes the specific policy on which the money was spent, the raw amount spent, the amount adjusted for inflation, and the amount spent per child.

More information here

```
df <- read_csv("kids.csv")
glimpse(df)
```

```
## Rows: 23,460
## Columns: 6
## $ state            <chr> "Alabama", "Alaska", "Arizona", "Arkansas", "Californ~
## $ variable         <chr> "PK12ed", "PK12ed", "PK12ed", "PK12ed", "PK12ed", "PK~
## $ year             <dbl> 1997, 1997, 1997, 1997, 1997, 1997, 1997, 1997, 1997,~
## $ raw              <dbl> 3271969, 1042311, 3388165, 1960613, 28708364, 3332994~
## $ inf_adj          <dbl> 4665308.5, 1486170.0, 4830985.5, 2795523.0, 40933568.~
## $ inf_adj_perchild <dbl> 3.929449, 7.548493, 3.706679, 3.891275, 4.282325, 4.3~
```

Video Game Trends

This dataset documents the amount of players that video games every month, and it records how much these numbers have changed across months.

For more information, you can click here.

```
df <- read_csv("video_games.csv")
glimpse(df)
```

```
## Rows: 83,631
## Columns: 7
## $ gamename      <chr> "Counter-Strike: Global Offensive", "Dota 2", "PLAYERUNK~
```

```
## $ year          <dbl> 2021, 2021, 2021, 2021, 2021, 2021, 2021, 2021, 2021, 20~
## $ month         <chr> "February", "February", "February", "February", "Februar~
## $ avg           <dbl> 741013.24, 404832.13, 198957.52, 120982.64, 117742.27, 1~
## $ gain          <dbl> -2196.42, -27839.52, -2289.67, 49215.90, -24374.98, 1808~
## $ peak          <dbl> 1123485, 651615, 447390, 196799, 224276, 133620, 146438,~
## $ avg_peak_perc <chr> "65.9567%", "62.1275%", "44.4707%", "61.4752%", "52.4988~
```

Mario Kart Records

This one is fun: it is a dataset that records Mario Kart records. It gives you information about the racer, the track, the speed, and how long the record held. There are two complementary datasets here. One is called `world_records` and it holds information about the records themselves. `drivers` gives more information about the players that set the records.

You can go here for more information.

```
df <- read_csv("world_records.csv")
glimpse(df)
```

```
## Rows: 2,334
## Columns: 9
## $ track           <chr> "Luigi Raceway", "Luigi Raceway", "Luigi Raceway", "Lu~
## $ type            <chr> "Three Lap", "Three Lap", "Three Lap", "Three Lap", "T~
## $ shortcut        <chr> "No", "No", "No", "No", "No", "No", "No", "No", "No", ~
## $ player          <chr> "Salam", "Booth", "Salam", "Salam", "Gregg G", "Rocky ~
## $ system_played   <chr> "NTSC", "NTSC", "NTSC", "NTSC", "NTSC", "NTSC", "NTSC"~
## $ date            <date> 1997-02-15, 1997-02-16, 1997-02-16, 1997-02-28, 1997-~
## $ time_period     <chr> "2M 12.99S", "2M 9.99S", "2M 8.99S", "2M 6.99S", "2M 4~
## $ time            <dbl> 132.99, 129.99, 128.99, 126.99, 124.51, 122.89, 122.87~
## $ record_duration <dbl> 1, 0, 12, 7, 54, 0, 0, 27, 0, 64, 3, 0, 90, 132, 1, 74~
```

```
df <- read_csv("drivers.csv")
glimpse(df)
```

```
## Rows: 2,250
## Columns: 6
## $ position <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1~
## $ player   <chr> "Penev", "Penev", "Penev", "Penev", "Penev", "Penev", "Penev"~
## $ total    <dbl> 344, 344, 344, 344, 344, 344, 344, 344, 344, 344, 344, 344, 3~
## $ year     <dbl> 1997, 1998, 1999, 2000, 2001, 2002, 2003, 2004, 2005, 2006, 2~
## $ records  <dbl> NA, 181, 126, 14, 5, 11, 2, 5, NA, NA, NA, NA, NA, NA, NA, NA~
## $ nation   <chr> "Australia", "Australia", "Australia", "Australia", "Australi~
```

Olympic Medals

Here's a comprehensive dataset of athletes who have won medals in the Olympic games. Given that the dataset is quite heavy, I will show you how to load it directly.

For more information go here.

```
# Read in data
df  <- read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/2021/2021-07-
glimpse(df)
```

```
## Rows: 271,116
## Columns: 15
## $ id     <dbl> 1, 2, 3, 4, 5, 5, 5, 5, 5, 5, 6, 6, 6, 6, 6, 6, 6, 6, 7, 7, 7, ~
```

```
## $ name   <chr> "A Dijiang", "A Lamusi", "Gunnar Nielsen Aaby", "Edgar Lindenau~
## $ sex    <chr> "M", "M", "M", "M", "F", "F", "F", "F", "F", "F", "M", "M", "M"~
## $ age    <dbl> 24, 23, 24, 34, 21, 21, 25, 25, 27, 27, 31, 31, 31, 31, 33, 33,~
## $ height <dbl> 180, 170, NA, NA, 185, 185, 185, 185, 185, 185, 188, 188, 188, ~
## $ weight <dbl> 80, 60, NA, NA, 82, 82, 82, 82, 82, 82, 75, 75, 75, 75, 75, 75,~
## $ team   <chr> "China", "China", "Denmark", "Denmark/Sweden", "Netherlands", "~
## $ noc    <chr> "CHN", "CHN", "DEN", "DEN", "NED", "NED", "NED", "NED", "NED", ~
## $ games  <chr> "1992 Summer", "2012 Summer", "1920 Summer", "1900 Summer", "19~
## $ year   <dbl> 1992, 2012, 1920, 1900, 1988, 1988, 1992, 1992, 1994, 1994, 199~
## $ season <chr> "Summer", "Summer", "Summer", "Summer", "Winter", "Winter", "Wi~
## $ city   <chr> "Barcelona", "London", "Antwerpen", "Paris", "Calgary", "Calgar~
## $ sport  <chr> "Basketball", "Judo", "Football", "Tug-Of-War", "Speed Skating"~
## $ event  <chr> "Basketball Men's Basketball", "Judo Men's Extra-Lightweight", ~
## $ medal  <chr> NA, NA, NA, "Gold", NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
```

College Sports Budgets

This might be one might peak your interest. It is a detailed dataset about how colleges spend their sports budgets, and how much they get back in revenue. This one is also quite big so we will read it directly.

Here's more information about the data.

```
df  <- readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/2022/
glimpse(df)
```

```
## Rows: 132,327
## Columns: 28
## $ year                 <dbl> 2015, 2015, 2015, 2015, 2015, 2015, 2015, 2015, 2~
## $ unitid               <dbl> 100654, 100654, 100654, 100654, 100654, 100654, 1~
## $ institution_name     <chr> "Alabama A & M University", "Alabama A & M Univer~
## $ city_txt             <chr> "Normal", "Normal", "Normal", "Normal", "Normal",~
## $ state_cd             <chr> "AL", "AL", "AL", "AL", "AL", "AL", "AL", "AL", "~
## $ zip_text             <chr> "35762", "35762", "35762", "35762", "35762", "357~
## $ classification_code  <dbl> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 1, 1, 1, 1, 1, 1, 1~
## $ classification_name  <chr> "NCAA Division I-FCS", "NCAA Division I-FCS", "NC~
## $ classification_other <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
## $ ef_male_count        <dbl> 1923, 1923, 1923, 1923, 1923, 1923, 1923, 1923, 1~
## $ ef_female_count      <dbl> 2300, 2300, 2300, 2300, 2300, 2300, 2300, 2300, 2~
## $ ef_total_count       <dbl> 4223, 4223, 4223, 4223, 4223, 4223, 4223, 4223, 4~
## $ sector_cd            <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1~
## $ sector_name          <chr> "Public, 4-year or above", "Public, 4-year or abo~
## $ sportscode           <dbl> 1, 2, 3, 7, 8, 15, 16, 22, 26, 33, 1, 2, 3, 8, 12~
## $ partic_men           <dbl> 31, 19, 61, 99, 9, NA, NA, 7, NA, NA, 32, 13, NA,~
## $ partic_women         <dbl> NA, 16, 46, NA, NA, 21, 25, 10, 16, 9, NA, 20, 68~
## $ partic_coed_men      <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
## $ partic_coed_women    <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
## $ sum_partic_men       <dbl> 31, 19, 61, 99, 9, 0, 0, 7, 0, 0, 32, 13, 0, 10, ~
## $ sum_partic_women     <dbl> 0, 16, 46, 0, 0, 21, 25, 10, 16, 9, 0, 20, 68, 7,~
## $ rev_men              <dbl> 345592, 1211095, 183333, 2808949, 78270, NA, NA, ~
## $ rev_women            <dbl> NA, 748833, 315574, NA, NA, 410717, 298164, 13114~
## $ total_rev_menwomen   <dbl> 345592, 1959928, 498907, 2808949, 78270, 410717, ~
## $ exp_men              <dbl> 397818, 817868, 246949, 3059353, 83913, NA, NA, 9~
## $ exp_women            <dbl> NA, 742460, 251184, NA, NA, 432648, 340259, 11388~
## $ total_exp_menwomen   <dbl> 397818, 1560328, 498133, 3059353, 83913, 432648, ~
## $ sports               <chr> "Baseball", "Basketball", "All Track Combined", "~
```

Twitch Data

This is an interesting dataset. Streaming is now one of the preferred forms of entertainment and Twitch has become the natural home for this genre. This dataset was uploaded to kaggle and can be found here. It contains a lot of information about Twitch's top streamers:

```
twitch_data <- read_csv("../Data/twitchdata-update.csv")
glimpse(twitch_data)
```

```
## Rows: 1,000
## Columns: 11
## $ Channel                <chr> "xQcOW", "summit1g", "Gaules", "ESL_CSGO", "Tfu~
## $ `Watch time(Minutes)`  <dbl> 6196161750, 6091677300, 5644590915, 3970318140,~
## $ `Stream time(minutes)` <dbl> 215250, 211845, 515280, 517740, 123660, 82260, ~
## $ `Peak viewers`         <dbl> 222720, 310998, 387315, 300575, 285644, 263720,~
## $ `Average viewers`      <dbl> 27716, 25610, 10976, 7714, 29602, 42414, 24181,~
## $ Followers              <dbl> 3246298, 5310163, 1767635, 3944850, 8938903, 15~
## $ `Followers gained`     <dbl> 1734810, 1370184, 1023779, 703986, 2068424, 554~
## $ `Views gained`         <dbl> 93036735, 89705964, 102611607, 106546942, 78998~
## $ Partnered              <lgl> TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, TRUE,~
## $ Mature                 <lgl> FALSE, FALSE, TRUE, FALSE, FALSE, FALSE, FALSE,~
## $ Language               <chr> "English", "English", "Portuguese", "English", ~
```

World Cup Shootouts

This is one of my favorite datasets. It contains data for every penalty kick in a World Cup shootout from 1982 to 2018. It was built and uploaded to kaggle by Pablo L. Landeros. You can find it here.

```
penalties <- read_csv("../Data/WorldCupShootouts.csv")
glimpse(penalties)
```

```
## Rows: 304
## Columns: 9
## $ Game_id        <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2~
## $ Team           <chr> "FRA", "GER", "FRA", "GER", "FRA", "GER", "FRA", "GER",~
## $ Zone           <dbl> 7, 9, 6, 2, 9, 4, 8, 3, 9, 9, 7, 9, 4, 2, 6, 6, 8, 9, 4~
## $ Foot           <chr> "R", "R", "R", "R", "R", "R", "L", "R", "R", "R", "R", ~
## $ Keeper         <chr> "R", "C", "L", "C", "L", "L", "L", "R", "L", "C", "L", ~
## $ OnTarget       <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1~
## $ Goal           <dbl> 1, 1, 1, 1, 1, 0, 0, 1, 1, 1, 0, 1, 0, 1, 1, 1, 1, 1, 1~
## $ Penalty_Number <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 1, 2, 3, 4, 5, 6~
## $ Elimination    <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0~
```

English Premier League

This is a dataset containing the results for the 2015/2016 English Premier League season. It was a truly unexpected season and you should look it up. I got the dataset from an excellent tutorial by Milad Kharratzadeh. You can find it here. This is what the data looks like:

```
pl_data <- read_csv("../Data/premier_league.csv")

glimpse(pl_data)
```

```
## Rows: 380
## Columns: 4
## $ home_team  <chr> "Bournemouth", "Chelsea", "Everton", "Leicester", "Man Unit~
## $ away_team  <chr> "Aston Villa", "Swansea", "Watford", "Sunderland", "Tottenh~
```

```
## $ score_diff <dbl> -1, 0, 0, 2, 1, -2, -2, 0, -1, -3, -1, -3, -2, 2, 0, 0, -1,~
## $ result     <chr> "aw", "d", "d", "hw", "hw", "aw", "aw", "d", "aw", "aw", "a~
```

Each row represents a match. We have the home team, away team, the score (in terms of difference), and the categorical result. That last one is `aw` when the away team wins, `d` when there's a draw, and `hw` when the home team takes the points.