ICD Tidy Tuesday
Nov. 19, 2020

1 Data

Data from the week of 2020-10-20: Great American Beer Festival. Code to download data below:

```
beer_awards <- readr::read_csv("https://raw.githubusercontent.com/
    rfordatascience/tidytuesday/master/data/2020/2020-10-20/beer_awards.csv"
    )
```

2 Cleaning

There are a few entry errors in the `state` column. Identify and fix them. After cleaning, there are 50 unique `states`, but that includes `DC`. Which state has no beers at all?

3 Exploration

There are a *lot* of categories, really fine ones. I categorized the majority of beer types (297/515) into one major category each, and removed malts and wines. Open my GitHub gist and add the code to your script. Now `beer_major` contains each beer for which I identified a category (3753/4970), with a major category column and dummy variables for each type.

3.1 Pick your beer

Pick one or a handful of beers to explore. Don't just pick the first one so we can compare. Here is some code to pick a random beer style:

```
X <- # Insert your own code to identify how many unique beer types there
    are.
unique(beer_major$major_category)[runif(1, 1, X)]
```

3.2 Visualize

Visualize the number of beers produced in each state. Some ideas for presentations:

- A bar chart
  - Try splitting each bar into `Gold`, `Silver`, `Bronze`
  - Try placing the sub-bars next to each other.
- Color a US map based on number of beers.
  - Try package `usmap`

Which state has the most of your beer[1]? Try controlling for population. **Hint 1:** Here is a population CSV. Try calculating over-18 population only. **Hint 2:** Here is how I created the population data frame.

Fifty states is a lot to compare. Try grouping by some factor. `state.region` is built in to R. Some other ideas: population stratus, presidential vote, area.

Now visualize against your regions.

3.3 Tests

Perform some statistical tests to determine if more of your beers are produced in one region. Test whether one region does better than another (e.g. `Gold = 1, Silver = 2, Bronze = 3`).

[1] It's probably California, right?

4 Population

```
population <- read_csv("https://raw.githubusercontent.com/jakevdp/data-USstates/master/state-population.csv") %>%
rename(
 # Rename first column, using ‘‘ because ’/’ is a special character
 state = ‘state/region‘
) %>%
filter(
 # Most recent year
 year == 2013,
 # States only + DC
 state %in% c(state.abb, "DC")
) %>%
select(-year) %>%
# Pivot wider so we can subtract under18 from total
pivot_wider(state, names_from = "ages", values_from = population) %>%
mutate(
 over18 = total - under18
)
```