ERHS 535 Project Report

Daniel Dean, Jessica Nunez , Erin Walls, Chaoyu Zhai

12/13/2019

Research Question:

We set out to explore which countries are most frequently referenced in the iconic Jeopardy! quiz show over its current 35-season run. In addition, we were interested in relationships with indicators like GDP and land area, as well as which areas seem to attract the highest premiums.

Data Processing:

Our dataset was drawn from Github user `jwolle`, who hosts a compilation of questions and answers asked in every season of Jeopardy! since its September 10, 1984 premiere, along with peripheral data including the monetary reward assigned to correct answers, comments from host Alex Trebek, whether the question was a Daily Double, and more. This trove of data, which we read in with the `readr` function `read_tsv`, is already nicely formatted and adheres to Hadley Wickham's Tidy Data philosophy for organization and information density, making the initial import an easy step. However, obtaining a viable list of countries proved more challenging. To account for various ways of referring to a given country, we used a list of up to eight synonyms for countries in the package `rworldmap`, which contains country border polygons with associated data, and which we used for downstream mapping. These were organized by three-digit United Nations ISO3 codes, which provided a backbone for the rest of the analysis (e.g. "The United States" and "U.S.A" share the code "usa"). Because restricting matches to , we used the Hadley Wickham-authored `rvest` package to scrape the HTML table from the Wikipedia entry on Demonyms (adjectives associated with a given country and nationality), which we matched to existing country names for additional nuance. We used the `pivot_longer` function from `tidyr` to collapse the resulting table into two paired columns, one giving ISO3 codes and one giving the alternative country names. From there, the `stringr` package's `str_extract_all` function was used to pull matches from our list of country names/adjectives, with any adjacent alphanumerical character excluded to avoid false matches like "India" in "Indiana." Speaking of India, a final challenge was accounting for context-dependent false positives, with the most emblematic example being the common use of "Indians" to refer to Native Americans; category (e.g. "Early American History") provided some context, but this remains a source of ambiguity, especially in cases like "Georgia O'Keefe," which returns a match for the country Georgia. From here, we were able to match the ISO3 codes to `rworldmap's` data and polygon geometry for mapping and a head-start on analysis. We made liberal use of the `janitor` package's `clean_names` function, which automatically reformats input column names to uniform lower-case words separated by underscore in this process.

Mapping:

Given the nature of our question, we used mapping extensively, settling on `rworldmap` for the most part, with help from the `sf` package's `st_as_sf` function, which converts a variety of spatial vector data into the universal 'simple format' for downstream mapping with tools like `leaflet`. Because our combined jeopardy and geographic data became quite large over the 35-season run, `select`, `summarize`, and `distinct` from `dplyr` were very helpful in picking relevant rows, creating tallies, averages, etc., and removing redundancies prior to mapping. The `ggplot`, `leaflet`, and `plotly` packages were used to create interactive maps based on `rworldmap` polygons and our data.

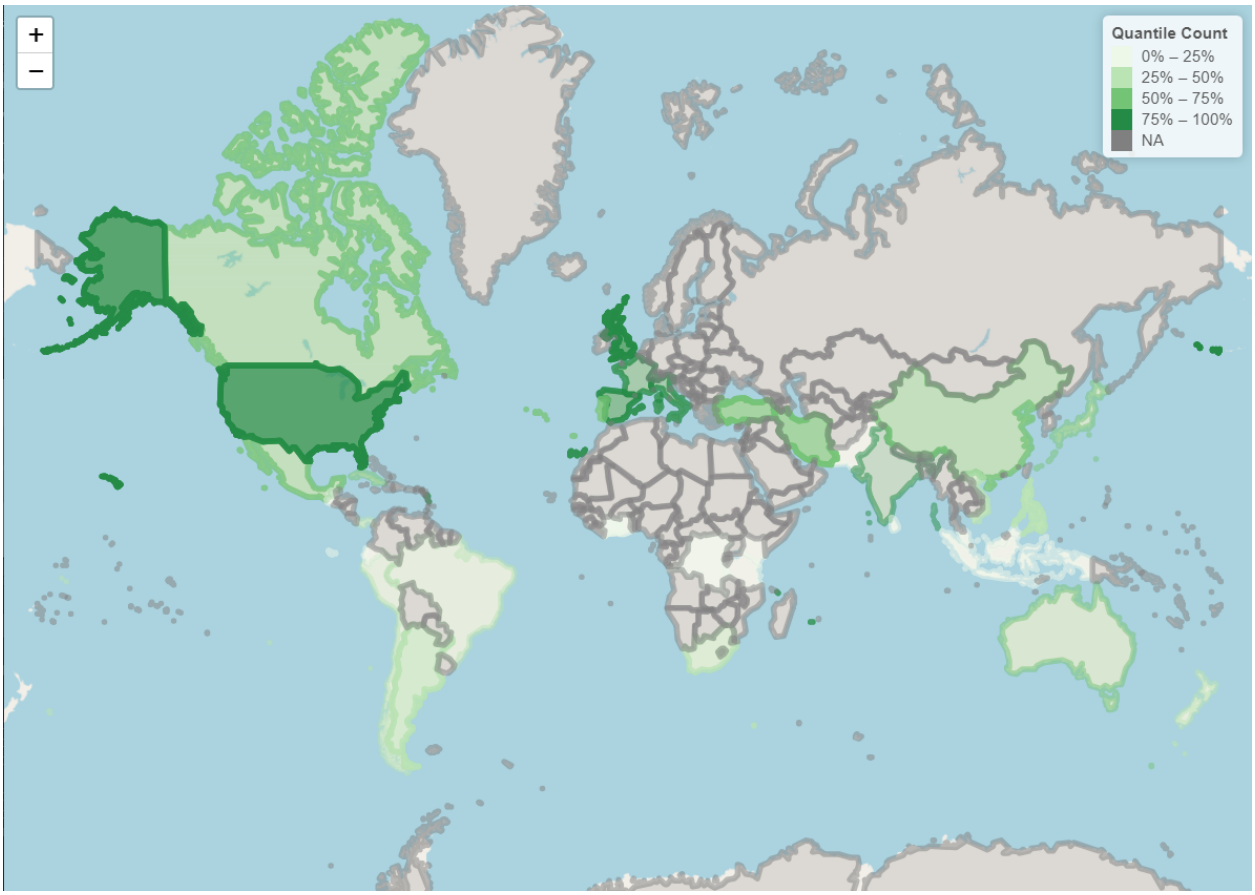


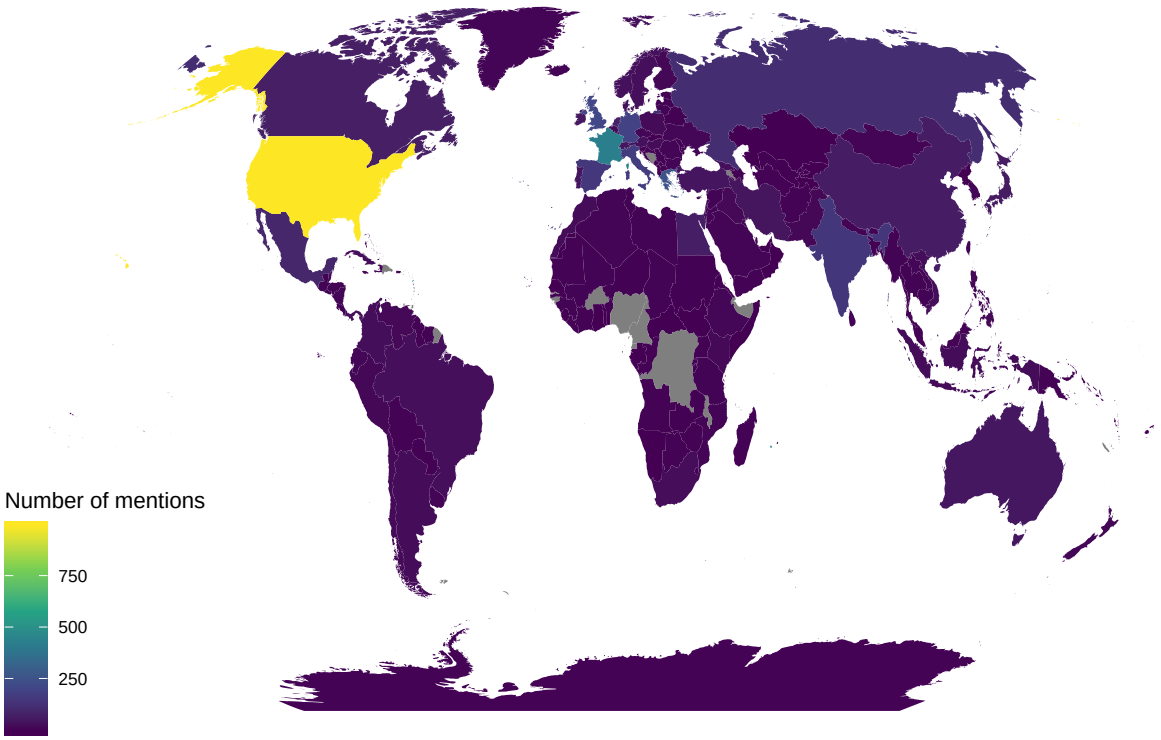
Figure 1: Screenshot of Leaflet Plot.

Frequency of Mentions by Country

Analyses:

Our question of which countries are mentioned most frequently in Jeopardy! was relatively simple to answer. Rather unsurprisingly, the United States was by far the most common, at 16,395 cumulative mentions, well ahead of runner-up France (6,847), the UK (3,844), and the other, mostly European, leaders. Despite the undoubtedly high number of remaining false positives, India remained fairly high in the ranking even after a relatively aggressive category-based filtering at around 3,000 references. We created a leaflet map of cumulative references to each ISO3-coded country (unfortunately we discovered this excluded some historical countries, notably the USSR) for self-guided viewing, along with other variables of interest. We used a relatively straightforward statistical approach in exploring the relationship between Jeopardy! references and two factors, GDP and land area. In both cases, the data was summarized upstream of mapping to a year/iso3 level, and we then used the base R `stats` function `lm` to fit a simple linear model with the format `count ~ <gdp or landarea>`; these proved to be uniformly significant on inspection with the `Anova` function from `car`. Unsurprisingly, we found a significant positive correlation between GDP and references, with or without the US; when the US was included, it almost single-handedly influenced the slope of the trend with its annual references. While we had speculated the popularity of relatively small European countries like Italy and Britain, (as well as some Caribbean Islands) could create an inverse relationship between land area and references, there was a relatively consistent positive trend, which appeared largely controlled by Russia's enormous landmass, as well as relatively large countries like the US, Canada, Brazil, and Australia. To explore patterns in the distribution of country mentions within daily double answers and questions, the `plotly` package was utilized after creating a `ggplot` object to create an interactive visualitzation. Once again, the United States was by far the most frequently mentioned country in daily doubles at 989 mentions throughout all seasons of Jeopardy!. France was the runner up again with 424 mentions; after that, number of mentions dropped off dramatically, with many countries being mentioned less than 50 times.This distribution pattern seems to follow a distribution similar to the overall number of mentions as described above.

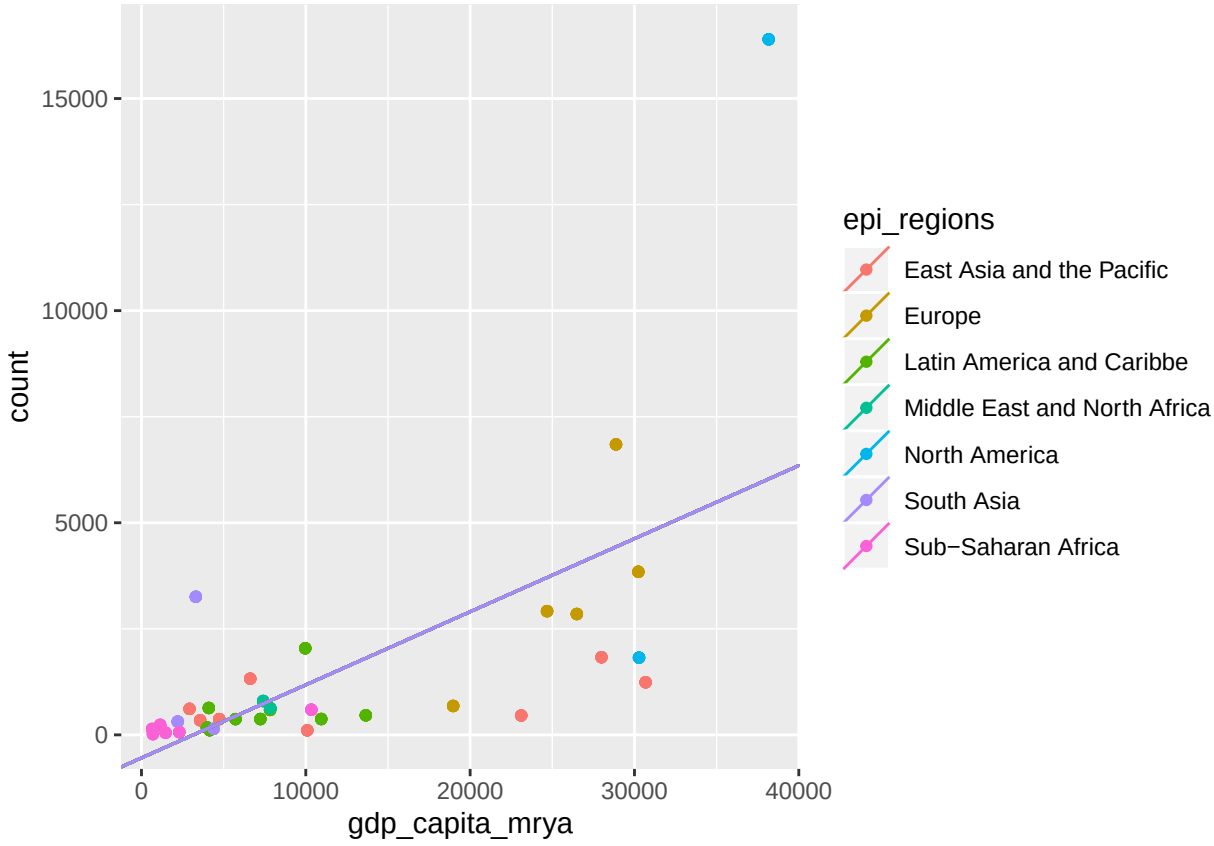
Frequency of Mentions in Daily Doubles by Country



All-Year GDP Trends

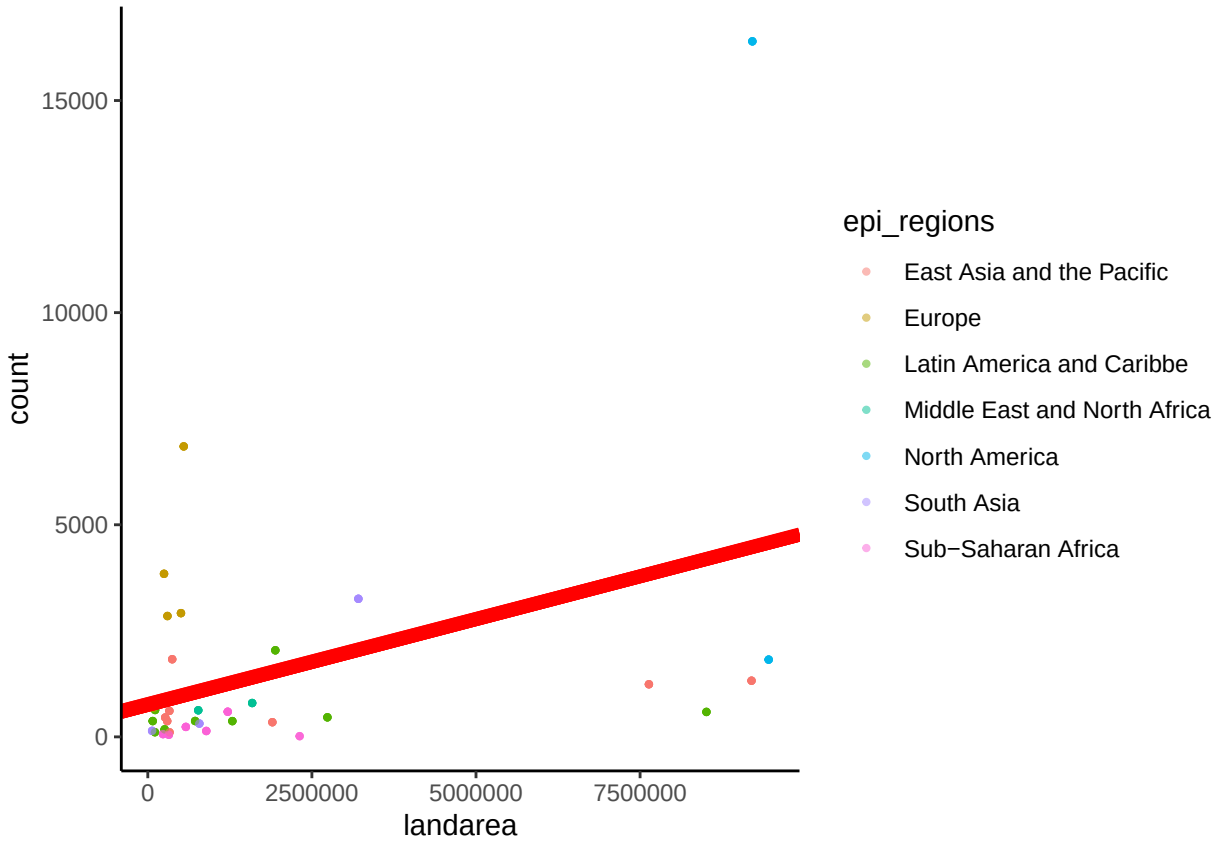
```
## Parsed with column specification:
## cols(
##   round = col_double(),
##   value = col_double(),
##   daily_double = col_character(),
##   category = col_character(),
##   comments = col_character(),
##   answer = col_character(),
##   question = col_character(),
##   air_date = col_date(format = ""),
##   notes = col_character(),
##   type = col_character(),
##   country = col_character(),
##   iso3 = col_character(),
##   count = col_double()
## )
```

```
## Joining, by = "ISO3V10"
```



All-year land area and mentions correlation:

```
## Joining, by = "ISO3V10"
```


15000
10000
5000
0
count
0
2500000
5000000
7500000
landarea
epi_regions
East Asia and the Pacific
Europe
Latin America and Caribbe
Middle East and North Africa
North America
South Asia
Sub−Saharan Africa