DATA 606 Data Project Proposal

Kevin Havis

Data Preparation

Our data comes from FiveThirtyEight's International Food Association's 2014 World Cup. Food World Cup survey in 2014. The data contains survey responses of U.S. adults, asking them about their general familiarity and preference for the native foods from different countries. Each respondent rated each country's native food based on how much they enjoy it. Respondents also provided some demographic information on themselves.

```
source("load_data.R")
```

Research questions

I plan to answer the below questions using chi-squared and difference in means tests; chi-squared for multinomial categorical tests (such as different age groups), and proportion tests for binomial tests (such as gender).

1. Do better educated U.S. adults prefer different cuisines?
2. Do different generations (age groups) have different preferences of cuisines?
3. Do male and female U.S. adults have different preferences of cuisine?

Cases

The cases in this dataset are 1,373 U.S. adults who responded to the survey.

Data collection

The data was collected by sending a voluntary online survey to readers of FiveThirtyEight and monitoring responses.

Type of study

This is an observational study.

Data Source

The survey and data were collected by FiveThirtyEight and referenced in their article The FiveThirtyEight International Food Association's 2014 World Cup

Describe your variables

The main variables of interest;

- `knowledge_cuisines`: Indicates how knowledgeable the respondent finds themselves from `Novice`, `Intermediate`, and `Advanced`
- `interest_cuisines` Indicates how interest in foreign cuisines the respondent finds themselves, rated 1-5 with 5 being the most interested
- `gender` The respondents gender, male or female
- `age` The age of the respondent, binned as 18-29, 30-44, 45-60, and 60+
- `household_income` The household income of the respondent, binned into 0-24k, 25k-49k, 50k-99k, 100k-150k, and 150k+
- `education` The highest level of education the respondent has obtained, described as "Less than a high school degree", "high school degree", "Some college or Associate degree", "bachelor degree", or "graduate degree"
- `location` The geographic area of the U.S. the respondent lives in, including "West South Central", "Pacific", "New England", "East North Central", "South Atlantic", "Mountain","Middle Atlantic", "East South Central", and "West North Central"
- `country` Countries the respondents were asked about.
- `country_rating` The rating each respondent gave to the food native to the `country`. The question was presented as "Please rate how much you like the traditional cuisine of X", which respondents were directed to respond to per the below rubric.

Value	Description
5	I love this country's traditional cuisine. I think it's one of the best in the world.
4	I like this country's traditional cuisine. I think it's considerably above average.
3	I'm OK with this county's traditional cuisine. I think it's about average.
2	I dislike this country's traditional cuisine. I think it's considerably below average.
1	I hate this country's traditional cuisine. I think it's one of the worst in the world.
N/A	I'm unfamiliar with this country's traditional cuisine.

Relevant summary statistics

Country ratings We will need to check if we have the right data to answer these questions.

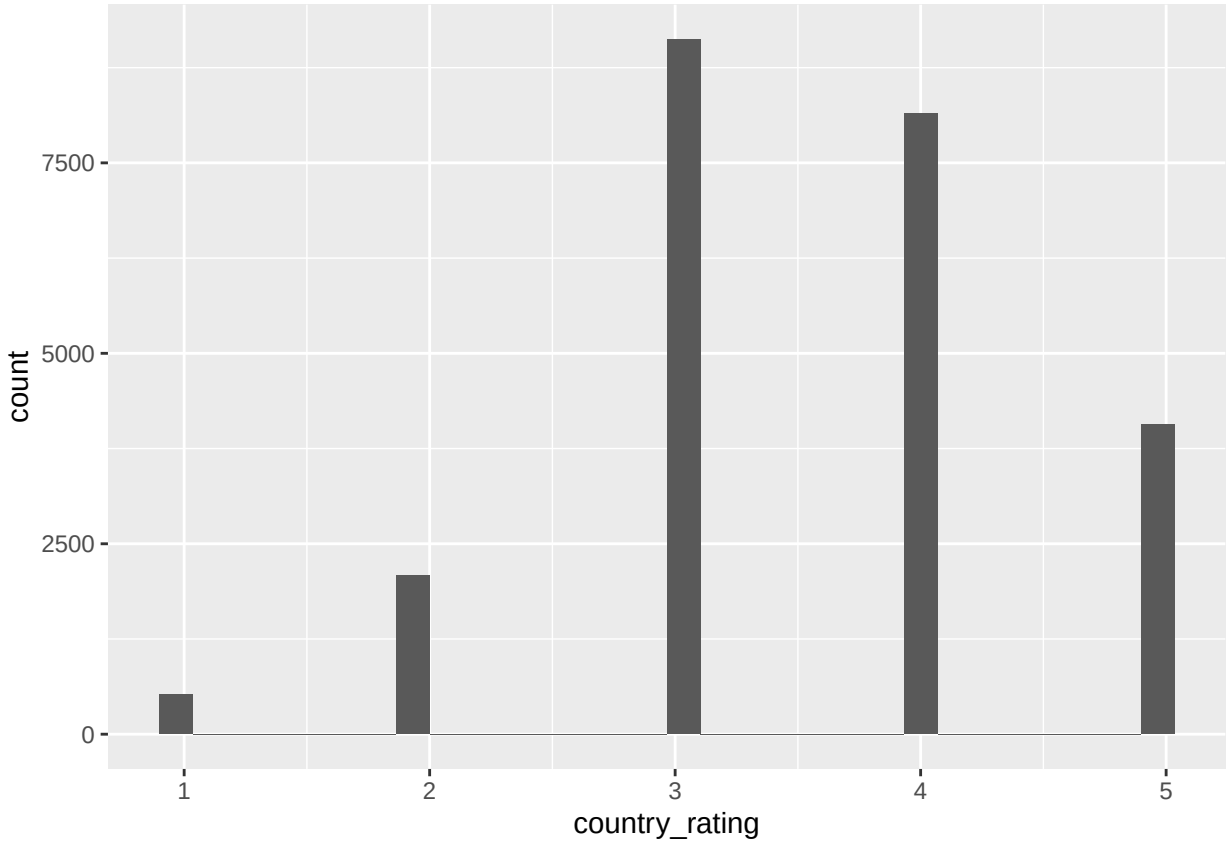
```
# convert countries to long format
long_data <- data |>
  pivot_longer(4:43, names_to = "country", values_to = "country_rating")
```

The rating data is relatively well distributed across countries. Ivory Coast, Cameroon, and Ghana have very low non-na proportions, so may need to remove them from analysis.

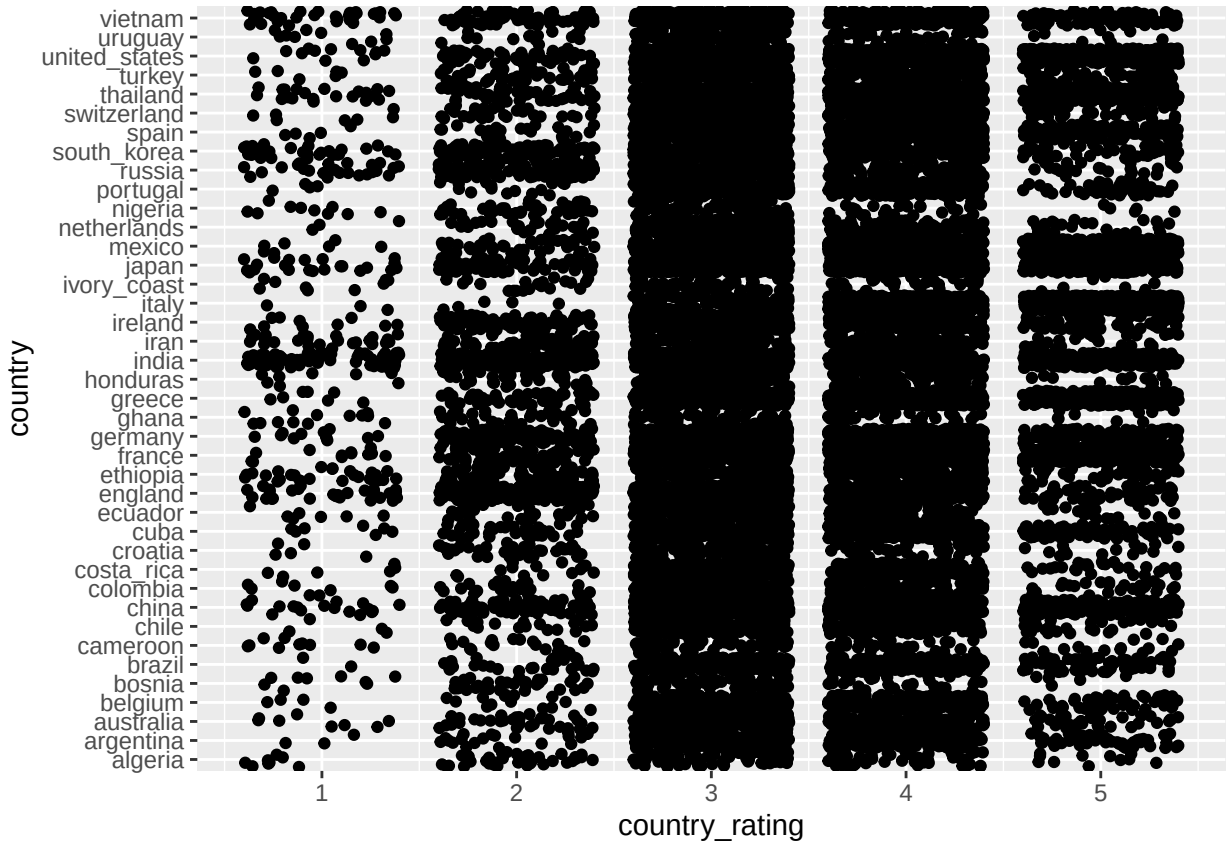
```
ggplot(long_data, aes(x=country_rating)) +
  geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 30977 rows containing non-finite outside the scale range
## (‘stat_bin()‘).
```

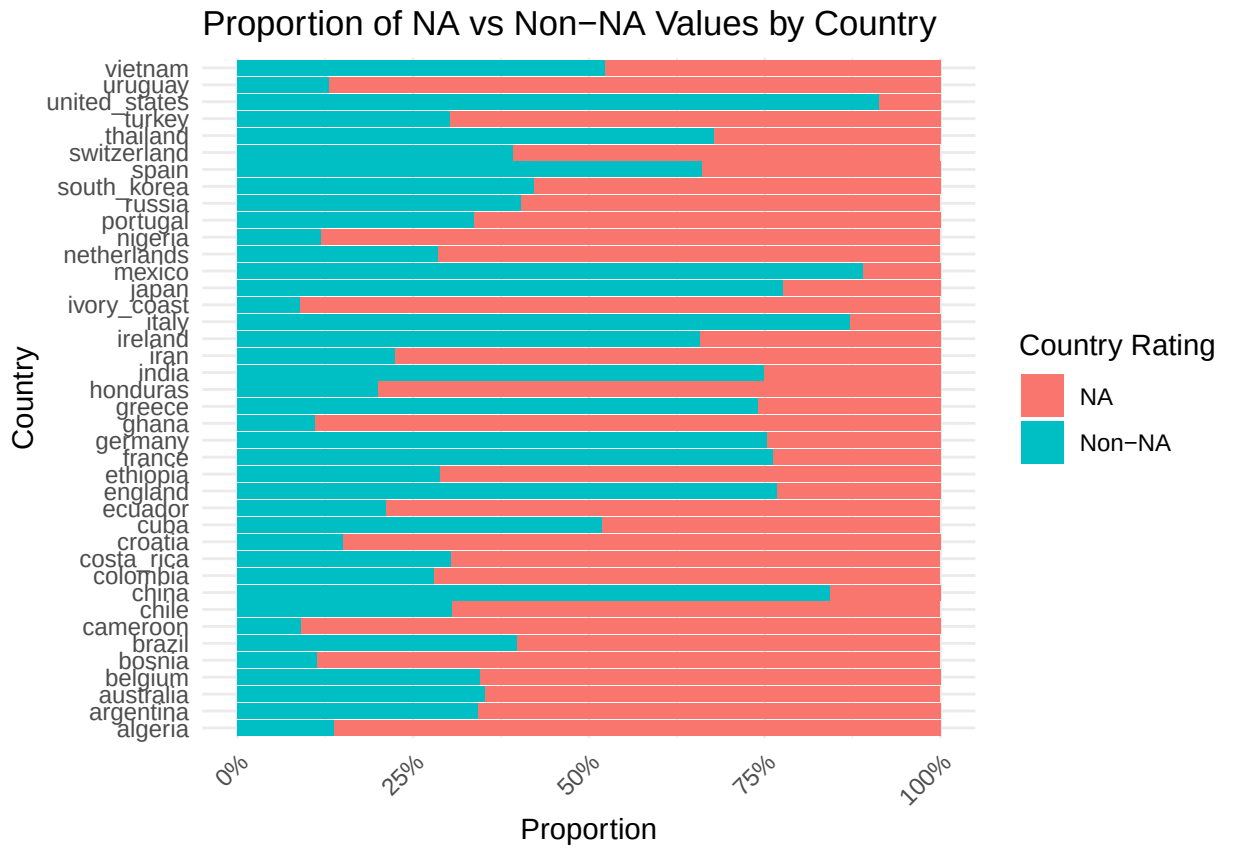



```
# Show high level distribution of ratings and countries
ggplot(long_data, aes(x=country, y = country_rating)) +
  geom_jitter() + coord_flip()
## Warning: Removed 30977 rows containing missing values or values outside the scale range
## (`geom_point()`).
```




```r
data_prop <- long_data |>
  mutate(is_na = ifelse(is.na(country_rating), "NA", "Non-NA")) |>
  count(country, is_na) |>
  group_by(country) |>
  mutate(prop = n / sum(n))

# Plot proportions using ggplot
ggplot(data_prop, aes(x = country, y = prop, fill = is_na)) +
  geom_bar(stat = "identity", position = "fill") +
  scale_y_continuous(labels = scales::percent) +
  labs(
    title = "Proportion of NA vs Non-NA Values by Country",
    x = "Country",
    y = "Proportion",
    fill = "Country Rating"
  ) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  coord_flip()
```




```
# Check proportions of countries with high nulls
data_prop |>
  filter(country %in%(c('ivory_coast', 'ghana', 'cameroon')))
```

```
## # A tibble: 6 x 4
## # Groups:   country [3]
##   country     is_na      n   prop
##   <chr>       <chr>  <int>  <dbl>
## 1 cameroon    NA      1249 0.910
## 2 cameroon    Non-NA   124 0.0903
## 3 ghana       NA      1221 0.889
## 4 ghana       Non-NA   152 0.111
## 5 ivory_coast NA      1250 0.910
## 6 ivory_coast Non-NA   123 0.0896
```

Respondent data Respondent data is relatively well distributed as well, with the exception of income; there's a decent gap in the data so we will not use this for our analysis.

```
display_prop_for_col <- function(x_col) {
  long_data |>
  group_by({{ x_col }}) |>
  summarize(prop = n() / nrow(long_data))
}
```

Gender Gender data has nulls that could either represent non-binary responses or simply gaps in the analysis. Unfortunately, since we cannot be sure of either, we will exclude nulls from our analysis.

```
display_prop_for_col(gender)
```

```
## # A tibble: 3 x 2
##   gender     prop
##   <chr>     <dbl>
## 1 ""       0.0976
## 2 "Female" 0.481
## 3 "Male"   0.422
```

Cuisine Knowledge Cuisine knowledge and interest are skewed, but usable. The analysis would definitely need to account for group size if we were to use these dimensions.

```
display_prop_for_col(knowledge_cuisines)
```

```
## # A tibble: 4 x 2
##   knowledge_cuisines   prop
##   <chr>               <dbl>
## 1 Advanced           0.138
## 2 Expert             0.0197
## 3 Intermediate       0.444
## 4 Novice             0.398
```

```
display_prop_for_col(interest_cuisines)
```

```
## # A tibble: 4 x 2
##   interest_cuisines   prop
##               <dbl>  <dbl>
## 1                 1 0.0612
## 2                 2 0.146
## 3                 3 0.474
## 4                 4 0.319
```

Age Age has a good assortment, and could easily approximated to generations. We are unfortunately missing about 10% of ages, which we cannot easily impute, so we will discard.

```
display_prop_for_col(age)
```

```
## # A tibble: 5 x 2
##   age       prop
##   <chr>    <dbl>
## 1 ""      0.0976
## 2 "18-29" 0.191
## 3 "30-44" 0.224
## 4 "45-60" 0.251
## 5 "> 60"  0.237
```

Income As mentioned, 30% of income data is missing, so we are not planning to use in the analysis.

```
display_prop_for_col(household_income)
```

```
## # A tibble: 6 x 2
##   household_income        prop
##   <chr>                  <dbl>
## 1 ""                     0.305
## 2 "$0 - $24,999"         0.101
## 3 "$100,000 - $149,999" 0.117
## 4 "$150,000+"            0.0910
## 5 "$25,000 - $49,999"    0.153
## 6 "$50,000 - $99,999"    0.233
```

Location Location of the respondents should be usable, and we may explore that if we have time.

```
display_prop_for_col(location)
```

```
## # A tibble: 10 x 2
##    location               prop
##    <chr>                 <dbl>
##  1 ""                    0.105
##  2 "East North Central" 0.137
##  3 "East South Central" 0.0291
##  4 "Middle Atlantic"    0.124
##  5 "Mountain"           0.0743
##  6 "New England"        0.0532
##  7 "Pacific"            0.157
##  8 "South Atlantic"     0.146
##  9 "West North Central" 0.0779
## 10 "West South Central" 0.0976
```