STA 199 Project Proposal

Maggie Brooks, Allison Bunker, Jonah Cagley, Jacqueline Wright

03-22-21

```
netflix <- read.csv("~/R/final_project_proposal/data/netflix_titles.csv")
glimpse(netflix)
```

```
## Rows: 7,787
## Columns: 12
## $ show_id      <chr> "s1", "s2", "s3", "s4", "s5", "s6", "s7", "s8", "s9", ...
## $ type         <chr> "TV Show", "Movie", "Movie", "Movie", "Movie", "TV Sho...
## $ title        <chr> "3%", "7:19", "23:59", "9", "21", "46", "122", "187", ...
## $ director     <chr> "", "Jorge Michel Grau", "Gilbert Chan", "Shane Acker"...
## $ cast         <chr> "João Miguel, Bianca Comparato, Michel Gomes, Rodolfo ...
## $ country      <chr> "Brazil", "Mexico", "Singapore", "United States", "Uni...
## $ date_added   <chr> "August 14, 2020", "December 23, 2016", "December 20, ...
## $ release_year <int> 2020, 2016, 2011, 2009, 2008, 2016, 2019, 1997, 2019, ...
## $ rating       <chr> "TV-MA", "TV-MA", "R", "PG-13", "PG-13", "TV-MA", "TV-...
## $ duration     <chr> "4 Seasons", "93 min", "78 min", "80 min", "123 min", ...
## $ listed_in    <chr> "International TV Shows, TV Dramas, TV Sci-Fi & Fantas...
## $ description  <chr> "In a future where the elite inhabit an island paradis...
```

The source of this data is Kaggle but it was originally collected from Flixable, a third party Netflix search engine, in 2019. It contains 12 columns with 7787 rows. Some relevant variables that it contains are the title of the show or movie, identifier as a show or movie, country it was produced in, release year, and rating. From this data set we can analyze what Netflix content is available in different countries and how this has changed over time. We can hypothesize that the ratio of tv shows to movies on Netflix has increased over time and varies by country. We can also compare this data set to IMDB ratings or Rotten Tomatoes to see if movie or tv show reviews play a role in the changing shift from primarily movies to primarily tv shows.

```
tested <- read.csv("~/R/final_project_proposal/data/tested_worldwide.csv")
glimpse(tested)
```

```
## Rows: 27,641
## Columns: 12
## $ Date             <chr> "2020-01-16", "2020-01-17", "2020-01-18", "2020-01...
## $ Country_Region   <chr> "Iceland", "Iceland", "Iceland", "South Korea", "U...
## $ Province_State   <chr> "All States", "All States", "All States", "All Sta...
## $ positive         <int> 3, 4, 7, 1, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, NA, NA, ...
## $ active           <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ hospitalized     <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ hospitalizedCurr <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ recovered        <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ death            <int> NA, NA, NA, NA, 0, 0, 0, 0, 0, 0, NA, 0, 0, 0, NA,...
## $ total_tested     <dbl> NA, NA, NA, 4, 0, 0, 0, 0, 0, 0, 27, 0, 0, 0, NA, ...
## $ daily_tested     <int> NA, NA, NA, NA, NA, NA, NA, 0, 0, 0, 5, 0, 0, 0, N...
## $ daily_positive   <int> NA, 1, 3, NA, NA, NA, NA, 0, 0, 0, 0, 0, 0, 0, NA,...
```

```r
vaccinations <- read.csv("~/R/final_project_proposal/data/country_vaccinations.csv")
glimpse(vaccinations)
```

```
## Rows: 7,488
## Columns: 15
## $ country                             <chr> "Afghanistan", "Afghanistan", "...
## $ iso_code                            <chr> "AFG", "AFG", "AFG", "AFG", "AF...
## $ date                                <chr> "2021-02-22", "2021-02-23", "20...
## $ total_vaccinations                  <dbl> 0, NA, NA, NA, NA, NA, 8200, NA...
## $ people_vaccinated                   <dbl> 0, NA, NA, NA, NA, NA, 8200, NA...
## $ people_fully_vaccinated             <dbl> NA, NA, NA, NA, NA, NA, NA, NA,...
## $ daily_vaccinations_raw              <dbl> NA, NA, NA, NA, NA, NA, NA, NA,...
## $ daily_vaccinations                  <dbl> NA, 1367, 1367, 1367, 1367, 136...
## $ total_vaccinations_per_hundred      <dbl> 0.00, NA, NA, NA, NA, NA, 0.02,...
## $ people_vaccinated_per_hundred       <dbl> 0.00, NA, NA, NA, NA, NA, 0.02,...
## $ people_fully_vaccinated_per_hundred <dbl> NA, NA, NA, NA, NA, NA, NA, NA,...
## $ daily_vaccinations_per_million      <dbl> NA, 35, 35, 35, 35, 35, 35, 41,...
## $ vaccines                            <chr> "Oxford/AstraZeneca", "Oxford/A...
## $ source_name                         <chr> "Government of Afghanistan", "G...
## $ source_website                      <chr> "http://www.xinhuanet.com/engli...
```

The source of this data is Kaggle, but it is (actively) being updated and merged from Our World in Data GitHub repository for COVID-19. It contains 15 columns of data. From this data we can analyze geographical trends in vaccination rates, identify common factors that lead to an increase or decrease in vaccination rates, and track how vaccination rates are changing in real time. We question how the vaccination rate trends will change over time, and what factors are the best indicators of the vaccination rates. We hypothesize that countries in Asia, which is the continent with the highest GDP worldwide, will have the highest percent vaccination rate because this is where the virus originated from (most time to develop and research vaccines). We can also compare this dataset to national financial trends and the COVID testing rates by country worldwide.

```r
happiness <- read.csv("~/R/final_project_proposal/data/DataPanelWHR2021C2.csv")
glimpse(happiness)
```

```
## Rows: 1,949
## Columns: 11
## $ Country.name                     <chr> "Afghanistan", "Afghanistan", "Afg...
## $ year                             <int> 2008, 2009, 2010, 2011, 2012, 2013...
## $ Life.Ladder                      <chr> "3,724", "4,402", "4,758", "3,832"...
## $ Log.GDP.per.capita               <chr> "7,370", "7,540", "7,647", "7,620"...
## $ Social.support                   <chr> "0,451", "0,552", "0,539", "0,521"...
## $ Healthy.life.expectancy.at.birth <chr> "50,800", "51,200", "51,600", "51,...
## $ Freedom.to.make.life.choices     <chr> "0,718", "0,679", "0,600", "0,496"...
## $ Generosity                       <chr> "0,168", "0,190", "0,121", "0,162"...
## $ Perceptions.of.corruption        <chr> "0,882", "0,850", "0,707", "0,731"...
## $ Positive.affect                  <chr> "0,518", "0,584", "0,618", "0,611"...
## $ Negative.affect                  <chr> "0,258", "0,237", "0,275", "0,267"...
```

This data source is from Kaggle and has 11 columns of data. The data is based on various surveys in 155 countries starting in 2012. For this dataset, we think it would be interesting to evaluate trends in happiness varying by location and time. There are clear external factors that are affecting this data (such as COVID-19, political events, war, weather etc). I hypothesize that for a given country, if there is any major event such as war, happiness will be lower. I also expect the overall happiness around the world to be lower in 2020 and 2021 due to COVID. I am curious to evaluate countries based on varying forms of government and see how this effects general happiness.