Project Proposal

due October 11, 2021 by 11:59 PM

Arthi Vaidyanathan and Denise Shkurovich: YAY STATS

10/11/2021

Load Packages

```
library(tidyverse)
```

Load Data

```
food <- readr::read_csv("data/Food_Supply_kcal_Data.csv")
```

Introduction and Data, including Research Questions

We are interested in looking at the relationship between COVID-19 outcomes and nutrition worldwide. The USDA Center for Nutrition Policy and Promotion suggests a dietary intake which consists of 30% grains, 40% vegetable, 10% fruits, and 20% proteins (dietaryguidlines.gov). Previous studies demonstrate an increased mortality in patients infected with COVID-19 which have chronic inflammatory diseases such as obesity, diabetes, and hypertension. The prevalence of these chronic inflammatory diseases are known to be correlated with an individual's diet (Onishi 2020). Furthermore, previous studies show that maintaining a healthy diet can decrease risk of severe infection by promoting the immune system (Messina et al. 2020, Iddir et. al 2020). Adequate protein consumption is essential for antibody production and poor nutrient consumption has been shown to increase inflammation and oxidative stress (Iddir et. al 2020). We are ultimately interested in seeing if countries that tend to consume similar diets to those suggested by the USDA show increased rates of recovery from COVID-19 controlling for income and vaccination levels.

This dataset, "COVID-19 Healthy Diet Dataset" comes from Kaggle. The dataset provides energy intake (kcal) as percentages of total diet by food group. In addition, it provides percentages of obesity and undernourished individuals. Finally it provides data for total confirmed COVID-19 cases, recovered COVID cases, COVID deaths, and active COVID cases for 170 countries. The food supply quantities in addition to the prevalence of obesity and undernourishment in the populations were obtained from the Food and Agricultural Organization of the United Nations, the population count was taken from the Population Reference Bureau, and the Johns Hopkins Center for Systems Science and Engineering was used for COVID-19 data.

Glimpse

```
glimpse(food)
```

```
## Rows: 170
## Columns: 32
## $ Country                               <chr> "Afghanistan", "Albania", "Algeria", "A~
## $ `Alcoholic Beverages`          <dbl> 0.0000, 0.9120, 0.0896, 1.9388, 2.3041,~
## $ `Animal Products`              <dbl> 4.7774, 16.0930, 6.0326, 4.6927, 15.367~
## $ `Animal fats`                  <dbl> 0.8504, 1.0591, 0.1941, 0.2644, 1.5429,~
## $ `Aquatic Products, Other`      <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ `Cereals - Excluding Beer`     <dbl> 37.1186, 16.2107, 25.0112, 18.3521, 13.~
## $ Eggs                           <dbl> 0.1501, 0.8091, 0.4181, 0.0441, 0.2057,~
## $ `Fish, Seafood`                <dbl> 0.0000, 0.1471, 0.1195, 0.8372, 1.7280,~
## $ `Fruits - Excluding Wine`      <dbl> 1.4757, 3.8982, 3.1805, 2.3133, 3.6824,~
## $ Meat                           <dbl> 1.2006, 3.8688, 1.2543, 2.9302, 7.0356,~
## $ `Milk - Excluding Butter`      <dbl> 2.4512, 9.9441, 3.9869, 0.5067, 4.6904,~
## $ Miscellaneous                  <dbl> 0.0250, 0.0588, 0.1045, 0.0661, 0.3086,~
## $ Offals                         <dbl> 0.1251, 0.2648, 0.0597, 0.1102, 0.1646,~
## $ Oilcrops                       <dbl> 0.1751, 1.0886, 0.2688, 1.0795, 0.5966,~
## $ Pulses                         <dbl> 0.5003, 0.8091, 1.0900, 1.4981, 0.4526,~
## $ Spices                         <dbl> 0.1001, 0.0000, 0.1195, 0.0000, 0.3497,~
## $ `Starchy Roots`                <dbl> 0.3252, 1.2651, 1.9262, 12.6239, 0.8434~
## $ Stimulants                     <dbl> 0.0750, 0.2501, 0.1493, 0.0441, 0.4937,~
## $ `Sugar Crops`                  <dbl> 0.0000, 0.0000, 0.0000, 0.0000, 0.0000,~
## $ `Sugar & Sweeteners`           <dbl> 2.2261, 3.4422, 3.9869, 2.7539, 5.8218,~
## $ Treenuts                       <dbl> 0.1251, 0.3972, 0.2240, 0.0000, 0.0823,~
## $ `Vegetal Products`             <dbl> 45.2476, 33.9070, 43.9749, 45.3184, 34.~
## $ `Vegetable Oils`               <dbl> 2.3012, 2.8244, 5.7638, 4.2741, 4.6904,~
## $ Vegetables                     <dbl> 0.7504, 2.7508, 2.0457, 0.3525, 1.2960,~
## $ Obesity                        <dbl> 4.5, 22.3, 26.6, 6.8, 19.1, 28.5, 20.9,~
## $ Undernourished                 <chr> "29.8", "6.2", "3.9", "25", NA, "4.6", ~
## $ Confirmed                      <dbl> 0.142134196, 2.967300916, 0.244897085, ~
## $ Deaths                         <dbl> 0.0061857789, 0.0509513742, 0.006558153~
## $ Recovered                      <dbl> 0.123373921, 1.792635659, 0.167572198, ~
## $ Active                         <dbl> 0.0125744965, 1.1237138830, 0.070766733~
## $ Population                     <dbl> 38928000, 2838000, 44357000, 32522000, ~
## $ `Unit (all except Population)` <chr> "%", "%", "%", "%", "%", "%", "%", "%",~
```

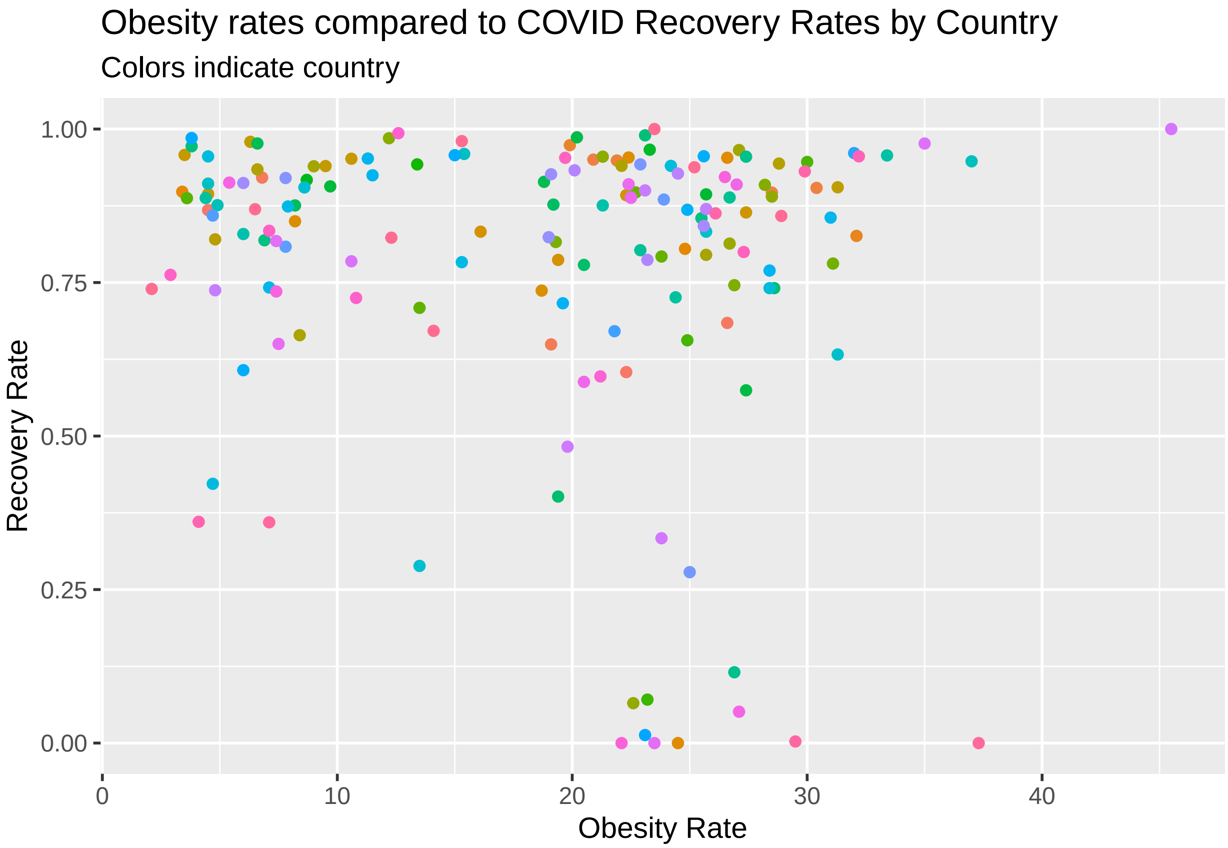
Data Analysis Plan

The outcome variable will be the percentage of recovered COVID cases over the percentage of total COVID cases, also known as the percentage of COVID cases in that country that recover. The predictor variable is how closely the diet of the country matches up with the USDA Center for Nutrition Policy and Promotion recommended values. Comparison groups include each country and countries grouped by income level and vaccination rates.

Statistical methods that will be useful in answering the question include classifying countries by both vaccination rates and income and then generating means and standard deviations in order to compare the different groups. We will also use statistics to create a similarity value between each of the countries/groups nutrition data and the recommended values. We will do this by taking the average of the absolute values of differences between the given values for each food group and the recommended values. Therefore, the countries/groups with caloric distributions closest to the recommended values will have values closer to 0. This value will help us support our hypothesized answer that countries/groups with dietary distributions closest to the recommended distribution will have higher percentages of COVID cases that recover by providing a single numerical value to quantify the adherence of the country to the recommended diet. The standard deviations as well as t-tests performed between the calculated values for the groups will help us determine if our results are significant or potentially due to chance variation.

```r
food %>%
mutate(recovery=Recovered/Confirmed)%>%
ggplot(aes(x=Obesity, y=recovery, color=Country))+geom_point()+ theme(legend.position = "none")+labs(ti
```

```
## Warning: Removed 7 rows containing missing values (geom_point).
```

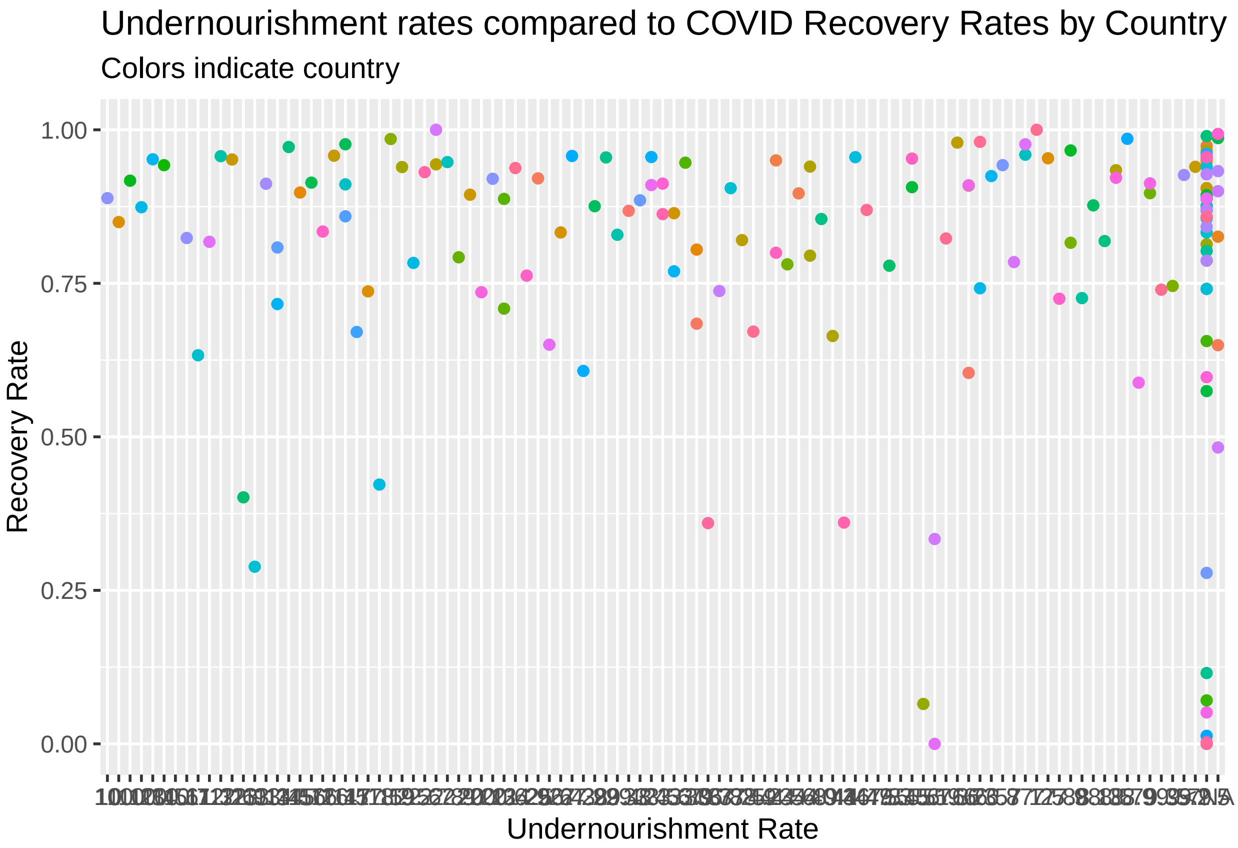


Obesity rates compared to COVID Recovery Rates by Country

Colors indicate country

```
food %>%
mutate(recovery=Recovered/Confirmed)%>%
ggplot(aes(x=Undernourished, y=recovery, color=Country))+geom_point()+ theme(legend.position = "none")+
```

```
## Warning: Removed 6 rows containing missing values (geom_point).
```



Undernourishment rates compared to COVID Recovery Rates by Country

Colors indicate country