Mosier520Milestone2

Kaylynn Mosier

2024-02-11

Final Project Step 1

Introduction

- Food insecurity is an epidemic that affects billions of people throughout the world. Increases in food prices are not being met with appropriate increases in wages. Federal funding for programs that may reduce food insecurity seems to get slimmer every year. Other risk factors such as race, household income, and geographic location may increase a person's chance of being affected by food insecurity. By using data science to identify circumstances that affect food insecurity, we can learn to act more proactively when multiple risk factors are present to reduce food insecurity and improve the overall health of the population.

Research Questions

1. What major factors effect food insecurity?
2. Is food insecurity rate effected by education? (USDA Dataset)
    - Do those with higher education experience a lower rate of food insecurity? (USDA Dataset)
3. Does GDP have an effect on food insecurity? (GDP Dataset)
4. Are food insecurity statistics and statistics reporting undernourishment aligned? (Food Insecurity and Undernourishment Datasets)
5. How have rates of food insecurity changed over time? (Food Insecurity Dataset)

Approach

- I plan to use multiple datasets to focus on these research questions. For questions specific to the United States, I will use a large dataset provided by the USDA. I will be able to filter this dataset based on demographic and geographic factors that may give insight to variables that effect food insecurity. Additional datasets that focus on percent of food insecurity and percent of the population that is undernourished will be used to investigate global trends. I also plan to look at how GDP effects food insecurity.

How your approach addresses (fully or partially) the problem

- This approach will hopefully bring to light risk factors for food insecurity specific to the US. It may also be able to highlight GDP as an indicator of global food insecurity. Finally, it will highlight the differences between individuals reporting as food insecure and individuals reporting as undernourished.

Data

1. USDA Household Food Insecurity Report 2022-
    - US Census Bureau. (2022). Current Population Survey, December 2022: Food Security Supplement Machine-readable data file [Dataset; Web]. https://www.census.gov/data/datasets/time-series/demo/cps/cps-supp_cps-repwgt/cps-food-security.html
    - See also documentation for data at: cpsdec22.pdf (census.gov)
    - This data is collected every year by the US Census Bureau for the Economic Research Service (ERS) and US Department of Agriculture (USDA) with intent to "research the range and severity of food insecurity as experience in U.S. households." Data was collected from approximately 54,000 households and reported in December of 2022. Responses were collected from these households once a month for four consecutive months and then again the following year during the same time period. This is an expansive data set with the power to highlight many possible factors affecting food insecurity. In addition to questions about food insecurity, demographic and geographic responses are represented which may further highlight relationships. There are 507 variables captured in this data set, but I will limit my focus to variables pertaining to demographic information, geographic information, and food scarcity. There are a large number of questions about food scarcity that result in a final food scarcity status divided by adults ad children, I will likely focus on these variables.
2. Prevalence of Undernourishment (% of Population)
    - World Bank. "Prevalence of Undernourishment (% of Population). World Development Indicators, The World Bank Group, Prevalence of undernourishment (% of population) | Data (worldbank.org)). 9 February 2024.
    - https://data.worldbank.org/indicator/SN.ITK.DEFC.ZS?view=chart
    - This dataset is located at the World Data Bank but was initially compiled by the Food and Agriculture Organization of the United Nations with intent to capture various aspects of food insecurity. This data began being compiled in 2011 and includes backdated information to 2001 and was last updated on December 18, 2023. There are columns containing country name, country code, indicator name, indicator code, and years from 1960-2021. Columns between 1960-2000 are left blank because data was not reported during that time. There are other cells left blank that will need to be dealt with in the data cleaning step.
3. Prevalence of Moderate to Severe Food Insecurity in the Population (%)
    - World Bank. "Prevalence of Moderate or Severe Food Insecurity in the Population (%)." World Development Indicators, The World Bank Group, Prevalence of moderate or severe food insecurity in the population (%) | Data (worldbank.org). 9 February 2024.
    - https://data.worldbank.org/indicator/SN.ITK.MSFI.ZS?view=chart
    - This dataset is located at the World Data Bank but was initially compiled by the Food and Agriculture Organization of the United Nations with intent to capture various aspects of food insecurity. This data began being compiled in 2015 and was last updated on December 18, 2023. There are columns containing country name, country code, indicator name, indicator code, and years from 1960-2021. Columns between 1960-2014 are left blank because data was not reported during that time. There are other cells left blank that will need to be dealt with in the data cleaning step.
4. GDP (current US)
    - World Bank. "GDP (current US)." World Development Indicators, The World Bank Group, GDP (current US$) | Data (worldbank.org). 11 February 2024.
    - https://data.worldbank.org/indicator/NY.GDP.MKTP.CD
    - This dataset is location at the World Data Bank and reports the GDP by country and year. GDP (Gross Domestic Product), is essentially the total output and income within a certain economy. It is often representative of the overall state of an economy. The site does not list the original

intention of this dataset, but it can be assumed that it was meant to communicate changes in GDP by country over time. In the dataset, there are a lot of blanks through the file that will need to be removed or converted during data cleaning. The columns of the dataset are country name, country code, indicator name, indicator code, and years from 1960-2022.

After looking at datasets 2-4, it will probably be advantageous to combine these into one larger dataset before indepth evaluation. These datasets can all be filtered by country and country code which will allow me to merge them into one dataset.

Required packages

- ggplot2
- dplyr
- Readxl
- reshape 2
- magrittr
- There may be additional required packaged added later

Plots and tables needed

- Scatter plots of variables to identify relationships
- Histograms of variables to check normality
- Boxplots of variables to identify outliers
- Correlation tables
- Covariance tables

Questions for Future Steps

- What is the relationship between my variables? I need to dig further into them before I can truly answer this question.
- How will I account for missing data? I am considering focusing on data between 2015-2021 because it seems to have the most complete information. But I may lose some insights by doing this.

Final Project Step 2

Import & Clean Data

I plan to clean the undernourishedData, foodInsecureData and gdpData in the same way, because I want all datasets to be have as similar paramaters as possible. I will remove the years 1960-2014 because the majority of these years are blank in the undernourishedData and foodInsecureData. Finally with years 2015-2021 selected, I will remove countries that do not have data for every year. Finally, I reshaped the data so that years was it's own column.

Load and clean undernourishment dataset.

```
library(readxl)
library(magrittr)  # Provides pipe function
library(tidyr)
```

```
##
## Attaching package: 'tidyr'

## The following object is masked from 'package:magrittr':
##
##     extract
```

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
undernourishedData <- read_excel('UndernourishedData.xlsx', col_names = TRUE)

# Remove years 1960-2014 and 2022
# Remove contries that do not have full data for selected years
undernourishedData <- subset(undernourishedData, select = -c(5:59, 67)) %>%
    drop_na('2015', '2016', '2017', '2018', '2019', '2020', '2021')

# Remove indicator code column, it is redundant
# Remove indicator name column
# Remove country code column, it is redundant
undernourishedData <- subset(undernourishedData, select = -c(2:4))

# Change column names
undernourishedData <- undernourishedData %>%
  rename(CountryName = 'Country Name')

# Reshape data to have a column for years
library(reshape2)
```

```
##
## Attaching package: 'reshape2'

## The following object is masked from 'package:tidyr':
##
##     smiths
```

```
undernourishedData <- melt(undernourishedData, id.vars = c('CountryName'),
                           variable.name='Year',
                           value.name='UndernourishedPercentPop')

head(undernourishedData)
```

```
##                   CountryName Year UndernourishedPercentPop
## 1 Africa Eastern and Southern 2015                  22.27844
## 2                 Afghanistan 2015                  21.30000
## 3  Africa Western and Central 2015                  11.49919
## 4                      Angola 2015                  13.50000
## 5                     Albania 2015                   4.30000
## 6                  Arab World 2015                  11.23944
```

```
foodInsecureData <- read_excel('ModtoSevereFoodInsecureData.xlsx', col_names = TRUE)

# Remove years 1960-2014 and 2022
# Remove contries that do not have full data for selected years
foodInsecureData <- subset(foodInsecureData, select = -c(5:59, 67)) %>%
    drop_na('2015', '2016', '2017', '2018', '2019', '2020', '2021')

# Remove indicator code column, it is redundant
# Remove indicator name
# Remove country code column, it is redundant
foodInsecureData <- subset(foodInsecureData, select = -c(2:4))

# Change column names
foodInsecureData <- foodInsecureData %>%
    rename(CountryName = 'Country Name')

# Reshape data to have a column for years
foodInsecureData <- melt(foodInsecureData, id.vars = c('CountryName'),
                         variable.name='Year',
                         value.name='FoodInsecurityPercentPop')

head(foodInsecureData)
```

Load and clean food insecurity dataset

```
##   CountryName Year FoodInsecurityPercentPop
## 1 Afghanistan 2015                     45.1
## 2     Albania 2015                     38.8
## 3   Argentina 2015                     19.2
## 4   Australia 2015                     10.8
## 5     Austria 2015                      5.5
## 6  Azerbaijan 2015                      5.9
```

```
gdpData <- read_excel('GDPData.xlsx', col_names=TRUE)

# Remove years 1960-2014 and 2022
# Remove contries that do not have full data for selected years
gdpData <- subset(gdpData, select = -c(5:59, 67)) %>%
    drop_na('2015', '2016', '2017', '2018', '2019', '2020', '2021')
```

```r
# Remove indicator code column, it is redundant
# Remove indicator name column
# Remove country code column, it is redundant
gdpData <- subset(gdpData, select = -c(2:4))

# Change column names
gdpData <- gdpData %>%
    rename(CountryName = 'Country Name')

# Reshape data to have a column for years
gdpData <- melt(gdpData, id.vars = c('CountryName'), variable.name='Year', value.name='GDP')

head(gdpData)
```

Load and clean GDP dataset

```
##                   CountryName Year          GDP
## 1                       Aruba 2015   2962907263
## 2 Africa Eastern and Southern 2015 932513471557
## 3                 Afghanistan 2015  19134221745
## 4  Africa Western and Central 2015 769263195357
## 5                      Angola 2015  90496420626
## 6                     Albania 2015  11386853113
```

```r
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v forcats   1.0.0     v readr     2.1.4
## v ggplot2   3.4.4     v stringr   1.5.1
## v lubridate 1.9.3     v tibble    3.2.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x tidyr::extract()   masks magrittr::extract()
## x dplyr::filter()    masks stats::filter()
## x dplyr::lag()       masks stats::lag()
## x purrr::set_names() masks magrittr::set_names()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```r
# Join three datasets into one by CountryName and Year
df_list <- list(foodInsecureData, undernourishedData, gdpData)
combinedCountryData <- df_list %>% reduce(full_join, by=c('CountryName', 'Year'))

# Drop any cells that contain NA after join
combinedCountryData <- drop_na(combinedCountryData)
head(combinedCountryData)
```

```
##   CountryName Year FoodInsecurityPercentPop UndernourishedPercentPop
## 1 Afghanistan 2015                     45.1                     21.3
## 2     Albania 2015                     38.8                      4.3
## 3   Argentina 2015                     19.2                      2.7
## 4   Australia 2015                     10.8                      2.5
## 5     Austria 2015                      5.5                      2.5
```

```
## 6  Azerbaijan 2015                    5.9                    2.5
##            GDP
## 1 1.913422e+10
## 2 1.138685e+10
## 3 5.947493e+11
## 4 1.351769e+12
## 5 3.819711e+11
## 6 5.307624e+10
```

The census data is difficult to read the way it is, I will limit my focus to only the columns listed below. I've included a short description of each column, but a more detailed description can be found at cpsdec22.pdf (census.gov)

- HRFS12M1 : Categorical - Food Security rating (1: Food Secure 2: Low food security 3: Very low food security)
- HRFS12MC : Categorical - Food Security rating in children (1: Food Secure 2: Low food security 3: Very low food security)
- HEFAMINC : Categorical - Family income
- GESTFIPS : Categorical - State code
- PEMARITL: Categorical - Martial status
- PEEDUCA : Categorical - Education level
- PTDTRACE : Categorical - Race
- PEMLR : Categorical - Employment status
- PRHRUSL : Categorical - Usual hours worked
- PRMJIND1 : Categorical - Major industry of employment

```r
censusData <- read.csv('USDA2022.csv', header=TRUE)
# Remove all columns except the ones listed above
censusData <- subset(censusData, select = c('HRFS12M1', 'HRFS12MC', 'HEFAMINC',
                                            'GESTFIPS', 'PEMARITL', 'PEEDUCA',
                                            'PTDTRACE', 'PEMLR', 'PRHRUSL', 'PRMJIND1'))

# Change column names
library(dplyr)
censusData <- censusData %>%
    rename('FoodSecurity'=HRFS12M1, 'FoodSecurityChildren'=HRFS12MC,
           'Income'=HEFAMINC, 'State'=GESTFIPS, 'MaritialStatus'=PEMARITL,
           'Education'=PEEDUCA,'Race'=PTDTRACE, 'EmploymentStatus'=PEMLR,
           'HoursWorked'=PRHRUSL,'Industry'=PRMJIND1)
```

```r
# Removing negative values from FoodSecurity and FoodSecuirtyChildren
# These values code for NAs
censusData <- censusData[censusData$FoodSecurity>0 & censusData$FoodSecurityChildren>0, ]

# MaritialStatus, Education, EmploymentStatus, HoursWorked, and Industry contain -1 values
# I cannot find what these values mean- it is not listed in the documentation
# This leads me to believe these are meant to be NAs, so I will remove them.
```

```
censusData <- censusData[censusData$MaritialStatus>0 &
                           censusData$Education>0 &
                           censusData$EmploymentStatus>0 &
                           censusData$HoursWorked>0 &
                           censusData$Industry>0, ]
head(censusData)
```

```
##    FoodSecurity FoodSecurityChildren Income State MaritialStatus Education Race
## 14            1                    1     15     1              1        40    1
## 15            1                    1     15     1              1        42    1
## 41            2                    2      9     1              4        39    1
## 48            1                    1     15     1              1        32    1
## 74            1                    1     16     1              1        44    1
## 75            1                    1     16     1              1        44    1
##    EmploymentStatus HoursWorked Industry
## 14                1           6        5
## 15                1           4       10
## 41                1           4       12
## 48                1           4        9
## 74                1           4        8
## 75                1           4        9
```

What is not self-evident?

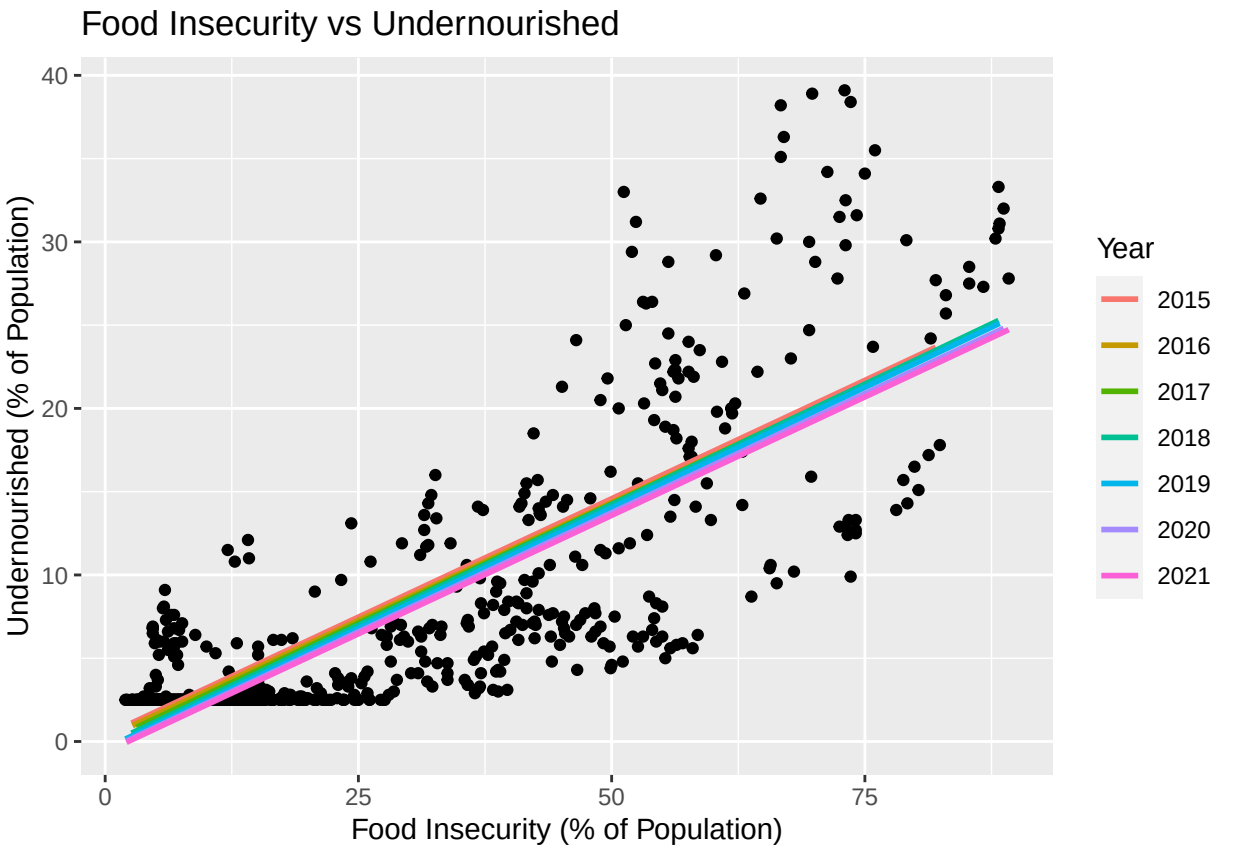
Before evaluating the data, it's difficult to identify relationships between variables. In the combinedCountryData, I am unable to see if there is a correlation between GDP, food insecurity, and undernourishment. I am also unable to see how these variables change through time. In the censusData, I am unable to see relationships between food scarcity and the other variables such as income, race, marital status and education level. I am also unable to see if there is a difference between food scarcity in adults versus children.

What are different ways you could look at this data? / How do you plan to slice and dice the data?

I could view the combinedCountryData by country and year for each indicator (GDP, food scarcity, and undernourishment). This may allow me to see relationships that are not immediately evident. Separating the data by year may reveal trends through time. Separating the data by country may reveal relationships that are hidden when data is viewed as a whole. Plotting histograms of variables may be helpful in early decision making. Additionally, correlation tables may also be helpful.

```
# Plot FoodInsecurity vs Undernourished to check for relationship
InsecurityUndernourishedPlot <-
    ggplot(combinedCountryData, aes(x=FoodInsecurityPercentPop, y=UndernourishedPercentPop)) +
    geom_point() +
    geom_smooth(method=glm, se=FALSE, aes(color=Year)) +
    labs(x='Food Insecurity (% of Population)', y = 'Undernourished (% of Population)',
         title = 'Food Insecurity vs Undernourished')

InsecurityUndernourishedPlot
```

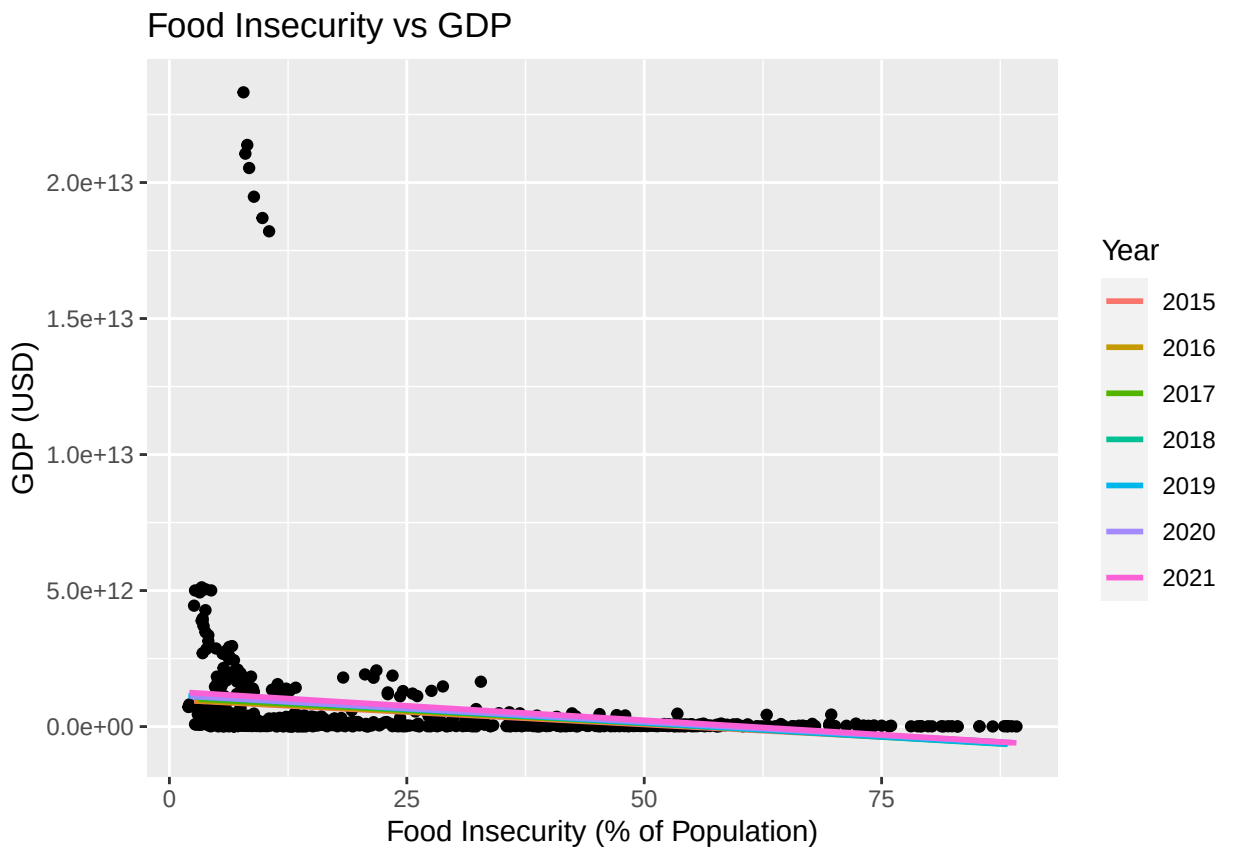
```
## `geom_smooth()` using formula = 'y ~ x'
```



- The plot above shows the relationship between the percent of population with food insecurity and the percent of population that is undernourished. I have also added a line of best fit that represents each year. There appears to be a positive relationships that is consistent throughout the years.

```
# Plot FoodInsecurity vs GDP to check for relationship
InsecurityGDPPlot <-
    ggplot(combinedCountryData, aes(x=FoodInsecurityPercentPop, y=GDP)) +
    geom_point() +
    geom_smooth(method=glm, se=FALSE, aes(color=Year)) +
    labs(x='Food Insecurity (% of Population)', y = 'GDP (USD)',
         title = 'Food Insecurity vs GDP')

InsecurityGDPPlot
```

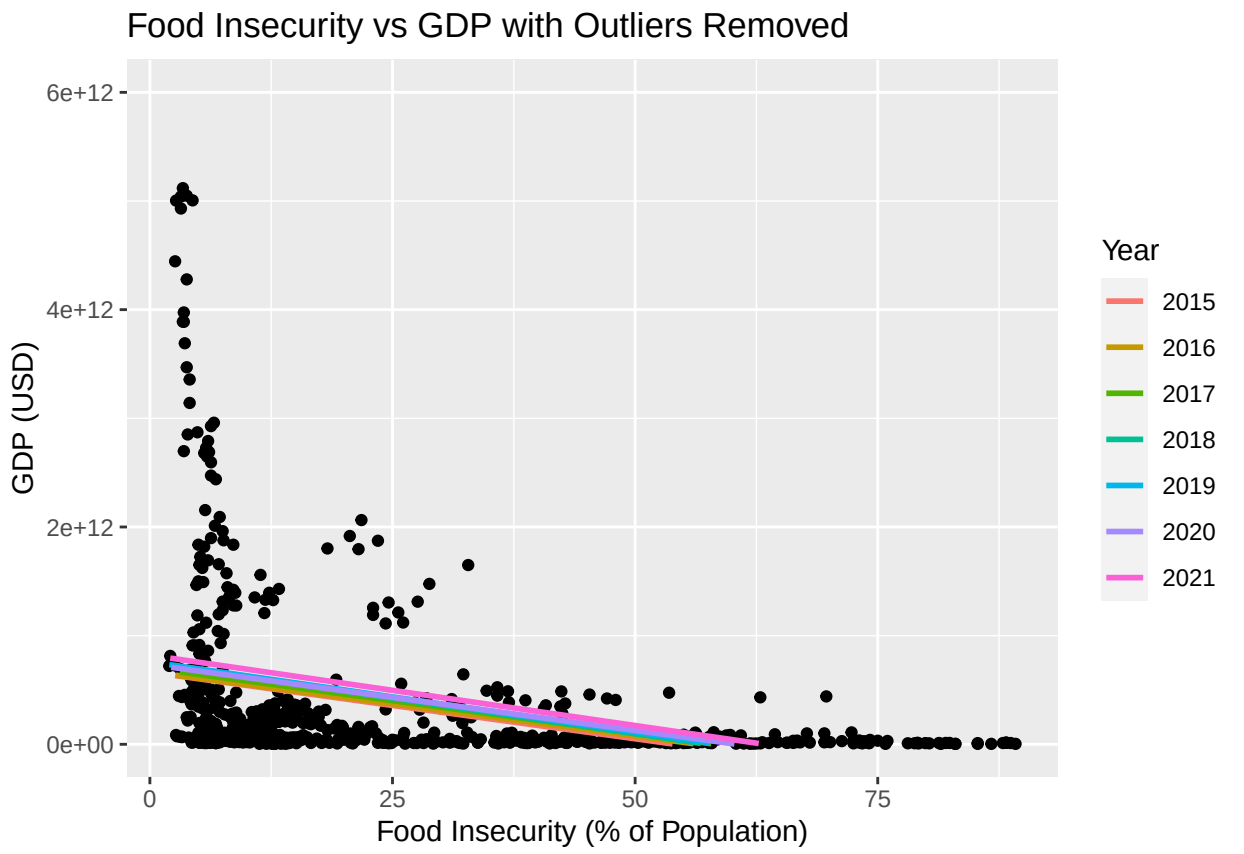
```
## `geom_smooth()` using formula = 'y ~ x'
```



- There are some obvious outliers in this dataset. Likely there are countries with GDP that are much higher than others. It may be necessary to remove these countries. If removing these countries is not an option, it may be beneficial to plot them separately. Even with outliers, it appears there is a negative relationship between GDP and percent of the population that is food insecure.

```
# Plot FoodInsecurity vs GDP with outliers removed
InsecurityGDPPlot2 <-
    ggplot(combinedCountryData, aes(x=FoodInsecurityPercentPop, y=GDP)) +
    geom_point() +
    scale_y_continuous(limits = c(0, 6.0e+12)) +
    geom_smooth(method=glm, se=FALSE, aes(color=Year)) +
    labs(x='Food Insecurity (% of Population)', y = 'GDP (USD)',
         title = 'Food Insecurity vs GDP with Outliers Removed')

InsecurityGDPPlot2
```

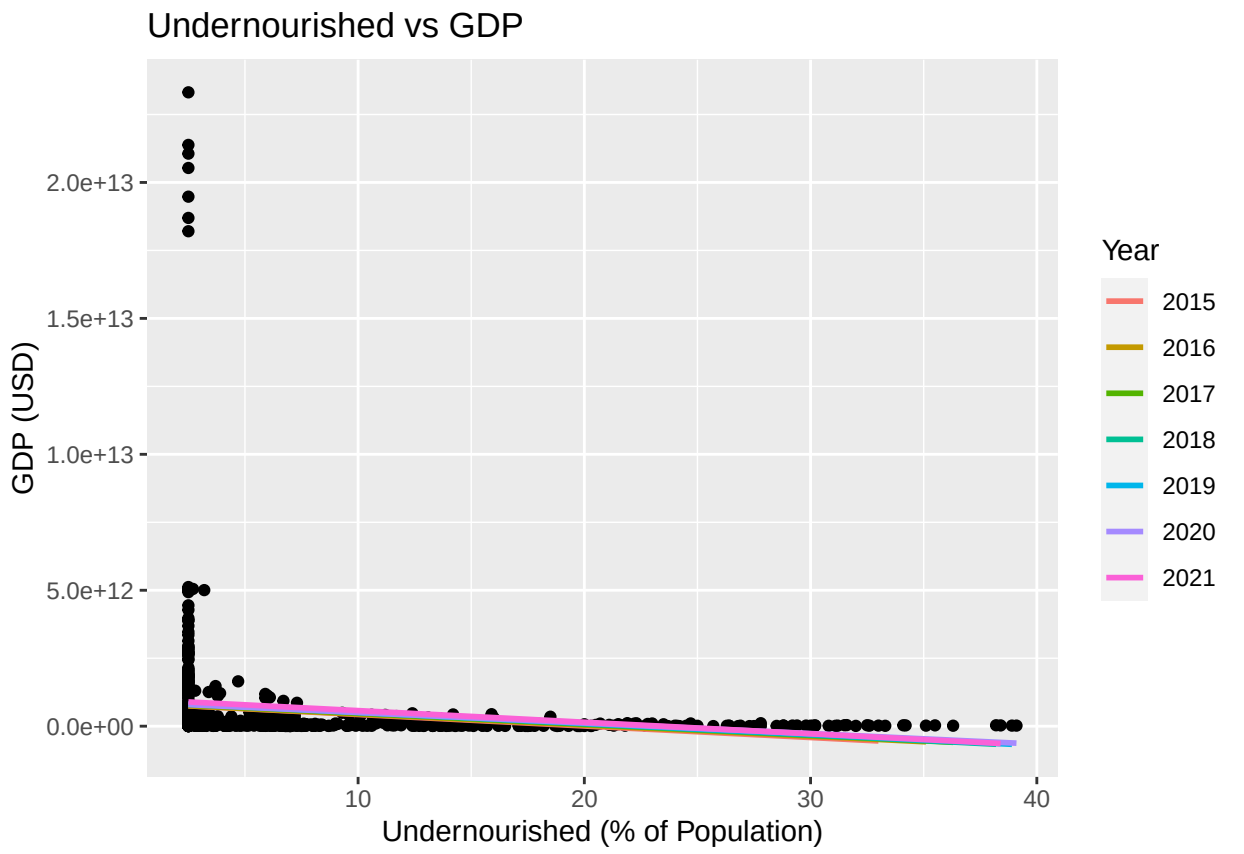
```
## `geom_smooth()` using formula = 'y ~ x'

## Warning: Removed 7 rows containing non-finite values (`stat_smooth()`).

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

## Warning: Removed 191 rows containing missing values (`geom_smooth()`).
```



- Removing these outliers from the plot has confirmed the negative relationship. A linear model may not be the best fit for this data.

```
# Plot Undernourished vs GDP to check for relationship
UndernourishedGDPPlot <-
    ggplot(combinedCountryData, aes(x=UndernourishedPercentPop, y=GDP)) +
    geom_point() +
    geom_smooth(method=glm, se=FALSE, aes(color=Year)) +
    labs(x='Undernourished (% of Population)', y = 'GDP (USD)',
         title = 'Undernourished vs GDP')

UndernourishedGDPPlot
```

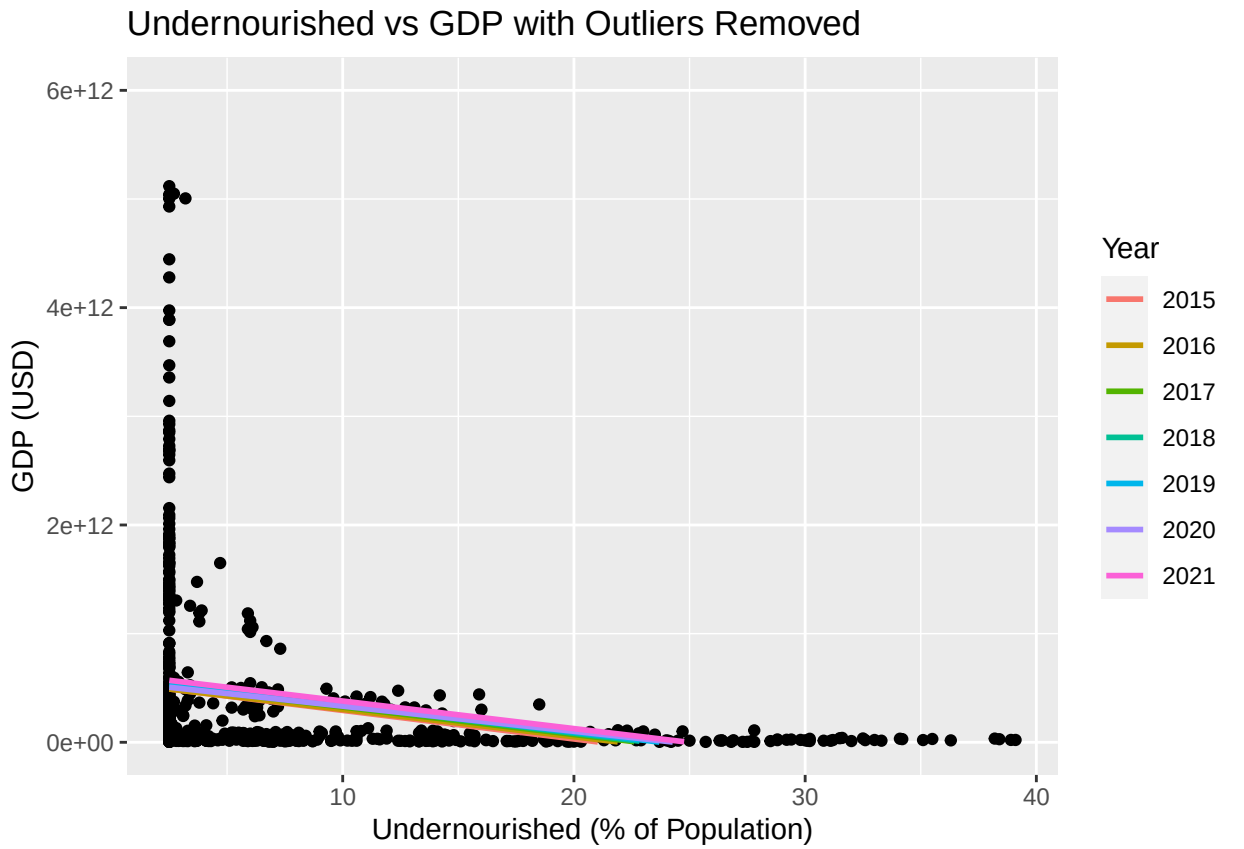
```
## `geom_smooth()` using formula = 'y ~ x'
```



- Plotting percent of population vs GDP does not reveal much. There may be a slight negative relationship, but the scale of the graph needs to be adjusted before further evaluation. There seem to be a lot of countries that have a very low rate of undernourishment. It is likely a good idea to set a range of values to focus on.

```
# Plot Undernourished vs GDP with outliers removed
UndernourishedGDPPlot2 <-
    ggplot(combinedCountryData, aes(x=UndernourishedPercentPop, y=GDP)) +
    geom_point() +
    scale_y_continuous(limits = c(0, 6.0e+12)) +
    geom_smooth(method=glm, se=FALSE, aes(color=Year)) +
    labs(x='Undernourished (% of Population)', y = 'GDP (USD)',
         title = 'Undernourished vs GDP with Outliers Removed')

UndernourishedGDPPlot2
```

```
## `geom_smooth()` using formula = 'y ~ x'

## Warning: Removed 7 rows containing non-finite values (`stat_smooth()`).

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

## Warning: Removed 224 rows containing missing values (`geom_smooth()`).
```



- Removing some of the major outliers highlights a negative relationship between GDP and percent of the population that is undernourished. A linear model may not be the best to view this data.
- Investigating GDP vs Undernourished and GDP vs FoodInsecure shows countries with higher GDP have a lower rate of both undernourishment and food insecurity.

I could view the censusData by state to test for differences in geographic region, by income or education to see their effect on food security. I could also see if there is a certain industry that produces more households with food scarcity. There are other combinations that I could look at. I will start by seeing which variables have the highest correlation with food scarcity.

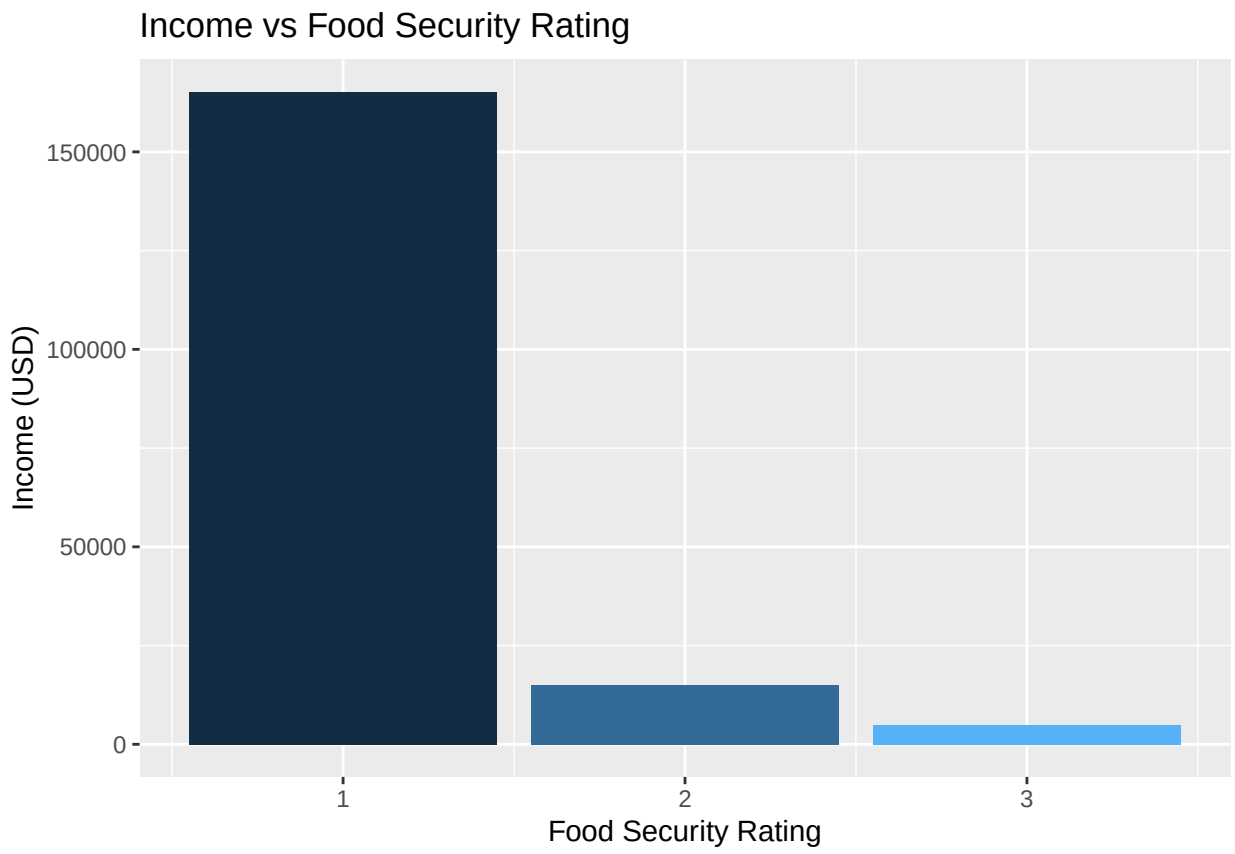
```
# Correlation table of variables in censusData
cor(censusData)
```

```
##                     FoodSecurity FoodSecurityChildren       Income
## FoodSecurity          1.000000000           0.72523147 -0.328463501
## FoodSecurityChildren  0.725231474           1.00000000 -0.230784080
## Income               -0.328463501          -0.23078408  1.000000000
## State                 0.009129273           0.01205838  0.004840009
## MaritialStatus        0.188338552           0.14173805 -0.301991239
## Education            -0.195137362          -0.14171535  0.394947511
## Race                  0.043174502           0.04394543  0.008945281
## EmploymentStatus      0.010317766           0.01198271 -0.026910520
## HoursWorked          -0.044640660          -0.02669407  0.081324553
## Industry             -0.036799428          -0.03319330  0.063869033
```

```
##                             State MaritialStatus    Education         Race
## FoodSecurity          0.0091292726     0.18833855 -0.195137362  0.043174502
## FoodSecurityChildren  0.0120583801     0.14173805 -0.141715354  0.043945434
## Income                0.0048400093    -0.30199124  0.394947511  0.008945281
## State                 1.0000000000    -0.01786092  0.001046977 -0.077729334
## MaritialStatus       -0.0178609220     1.00000000 -0.360204381  0.031069833
## Education             0.0010469773    -0.36020438  1.000000000  0.033175575
## Race                 -0.0777293344     0.03106983  0.033175575  1.000000000
## EmploymentStatus      0.0008137071     0.03010807 -0.038317092  0.011536948
## HoursWorked          -0.0007405119    -0.21528116  0.163575570 -0.016869076
## Industry             -0.0016083607     0.01282975  0.200951281  0.042401037
##                     EmploymentStatus   HoursWorked     Industry
## FoodSecurity              0.0103177658 -0.0446406600 -0.036799428
## FoodSecurityChildren      0.0119827108 -0.0266940685 -0.033193299
## Income                   -0.0269105196  0.0813245531  0.063869033
## State                     0.0008137071 -0.0007405119 -0.001608361
## MaritialStatus            0.0301080662 -0.2152811643  0.012829745
## Education                -0.0383170920  0.1635755698  0.200951281
## Race                      0.0115369476 -0.0168690763  0.042401037
## EmploymentStatus          1.0000000000 -0.0301437854 -0.002901272
## HoursWorked              -0.0301437854  1.0000000000 -0.132237829
## Industry                 -0.0029012718 -0.1322378295  1.000000000
```

- The correlation table looks intimidating because there are so many variables, but I will focus on the first two columns because those columns cover both food security variables I am interested in. There is a high correlation between food security and food security in children. This is not surprising because a house hold reporting food insecurity will have children that also report food insecurity. There is a moderate negative relationship between food security and income. This indicates as food insecurity increases, income decreases. In other words, as food security decreases, income decreases. From this data, we would expect people with a lower income to have a higher rate of food insecurity. The same relationship is seen between food security of children and income. When looking at this data, it's important to remember that for the food security variables, a value of 1 means the household has very food security while a value of 3 means the household has very low food security. It may be interesting to further investigate how income effects the two food security variables.

```
# Graph showing foodSecurity vs Income separated by foodSecurity rating
foodSecurityIncome <- ggplot(censusData, aes(x=FoodSecurity, y=Income, fill=FoodSecurity)) +
    geom_bar(stat='identity') +
    labs(x='Food Security Rating', y='Income (USD)', title='Income vs Food Security Rating') +
    theme(legend.position='none')

foodSecurityIncome
```



- This bar graph confirms the relationship between food security rating and income. As income decreases, food security decreases.

How could you summarize your data to answer key questions?

Correlation tables would be the most effective way to highlight relationships between the variables (research questions 1-3). A scatter plot showing food scarcity and undernourishment vs some key variable may highlight the relationship between the two variables (research question 4). A plot of food insecurity vs time will highlight the change in food insecurity rates over time (research question 5). It would also be interesting to plot this by country, or even just a few countries of interest.

What types of plots and tables will help you illustrate the findings to your questions?

I have outlined this above in detail. Correlation tables, histograms, scatterplots, and bar graphs will illustrate my findings.

Do you plan on incorporating any machine learning techniques to answer your questions? Explain.

I don't think machine learning techniques are necessary in this case. I do not plan to make predictions from this data, rather I just want to reveal trends that may not be immediately evident. This can be done without machine learning.

What questions do you have now that will lead to further analysis or additional steps?

I have outlined my outstanding questions above. My questions revolve around relationships between the variables.

1) What variables have the largest effect on food insecurity? Is education one of these variables?
2) Does GDP have an effect on food insecurity? What about undernourishment?
3) Are rates of food insecurity and undernourishment aligned? What does it mean if they aren't?
4) How have rates of food insecurity changed over time?

Final Project Step 3

Introduction

Throughout the world, food scarcity is an epidemic of global proportions. It may seem that in our industrialized economy, full of fast food and cheap snacks, that food is readily accessible to everyone. However, that's not the case. Inflation, price gouging, and inadequate wages cause people to choose between their next meal and paying rent nearly every day. It is essential that we gain a better understanding of circumstances that lead to food insecurity. By leveraging data science, we can identify risk factors that lead to food insecurity and learn to act more proactively when multiple risk factors are present to reduce food insecurity and improve the overall health of the population.

At the beginning of my research, I set out to gain a better understanding of the data around food insecurity. I wanted to understand what factors affect food insecurity to better mitigate their effects when possible. In preliminary research, I found that gross domestic product (GDP) can often be indicative of a country's overall health. Countries with a higher GDP would be expected to be 'healthier' and therefore have lower rates of food insecurity. In this preliminary research, I also made the realization that rates of food insecurity can be recorded in different ways. Another important metric that could be a measure of food insecurity is undernourishment. I wanted to know if countries that reported higher rates of undernourishment also reported higher rates of food insecurity. In other words, do these values align with each other or could we be misrepresenting food insecurity rates?

Analysis- Dataset 1

I was lucky to find four very in-depth data sets to help me answer these questions. One dataset reported food insecurity rates by country for numerous years, another reported undernourishment rates by country and year, a third reported GDP by country and year. Through data wrangling, I was able to combine these three datasets into one cohesive dataset that held rates of undernourishment, food insecurity and GDP and could be filtered by year and country if necessary.

Using this dataset, I plotted a scatter plot of food insecurity versus undernourishment to understand how these values relate to each other. Adding a line of best fit for each year that data was collected revealed a positive relationship; as food insecurity rates increased, rates of undernourishment also increased. This shows there likely is a correlation between rates of food insecurity and rates of undernourishment in a population. From this dataset, I also wanted to understand the relationship between GDP and food insecurity rates. Again, I plotted a scatter plot of the variables and added a line of best fit for each year data was collected. This scatter plot showed a negative relationship between GDP and food insecurity; as GDP decreased, food insecurity rates increased. This means countries with a lower GDP are expected to experience higher rates of food insecurity. Finally, I plotted a third scatter plot of GDP versus rates of undernourishment. Adding a line of best fit revealed a negative relationship between GDP and rates of undernourishment; as GDP decreased, rates of undernourishment increased.

Analysis- Dataset 2

I also came across the USDA Household Food Insecurity Report for 2022. This report contains relevant information about food insecurity rates in the United States and includes demographic and geographic information that is key in highlighting risk factors for food insecurity. This is a very large dataset that contains information far beyond the scope of this analysis. I chose to focus on a subset of the data set that contained two different measures of food security; one measure was for adults and the other for children. I also chose variables I thought may have an effect on food scarcity ratings. These included income, state of residence, marital status, education level, race, employment status, hours worked in the last pay period, and the industry the respondent is employed by.

I began my study of this dataset by running a correlation analysis to better understand what factors had the highest correlation with food security. This table revealed a high correlation between food security in adults and food security in children. This is not a surprise, households that struggle to feed adults will likely also struggle to feed the children in that home. This calculation also revealed a moderate negative relationship between food security and income indicating that as food security decreases, income decreases. From this, we can conclude individuals with a lower household income have higher rates of food insecurity. Education and marital status also both had a slight correlation with food security ratings. Education had a weak negative relationship indicating that individuals with higher education have higher food security. Marital status has a positive relationship with food security. This value seemed misleading at first, until I referenced the codebook. Marital status is a categorical variable that stores values 1-6. One means the respondent is married with the spouse present, values two through six code for households that are likely single income. The value from the correlation table tells us that single income households are more likely to struggle with food insecurity.

Implications and Limitations

This analysis has revealed there are some strong indicators for food scarcity. We can likely use GDP as a marker for countries at risk for high food insecurity rates. Countries with a low GDP should be watched closely to ensure rates of undernourishment and food insecurity do not rise too high. In the United States, low-income households are at a higher risk of facing food insecurity. Households with a single income are also at a higher risk of facing food shortages. We also saw that those with higher education are at a lower risk struggling with food insecurity. To me, marital status and education are both contributing factors to household income. Households with two working adults will often have a higher income than households with one working adult. Individuals with more education are often employed in higher paying jobs. From this analysis, I conclude that household income has the largest effect on food insecurity in the United States.

This analysis is by no means an exhaustive research into food scarcity. There is a nearly endless supply of data on the topic, much of which is outside the scope of this project. Even the datasets I chose to work with were greatly shortened. Analyzing more variables from the USDA dataset may very well negate the conclusions I came to. I chose not to employ any data modeling in this analysis. Further research could attempt to predict food insecurity using income, education level and marital status as predictor variables. It is also worth noting that although I used GDP as a measure of national health, there is rigorous debate as to whether GDP is an appropriate measure of national health. The use of this data is subject to scrutiny but in this case, I think it's appropriate.

Food scarcity and undernourishment may be difficult topics, but we should not shy away from them. I believe our role as data scientists is to investigate difficult topics, draw conclusions from them, and use our knowledge for the good of others. This analysis, and others like it, can be used as support for programs that allow more access to food to those at the highest risk. Allowing this continued support helps not only those who directly benefit from it, but our society as a whole.