# EDA on Food Security

Kaylynn Mosier

2024-02-11

## 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.

## 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 countries 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
```

*Load and clean food insecurity dataset*

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

# Remove years 1960-2014 and 2022
# Remove countries that do not have full data for selected years
```

```r
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)
```

```
##   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
```

### *Load and clean GDP dataset*

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

# Remove years 1960-2014 and 2022
# Remove countries 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')

# 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)
```

```
##                    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 ──
## ✔ forcats   1.0.0     ✔ readr     2.1.5
## ✔ ggplot2   3.5.1     ✔ stringr   1.5.1
## ✔ lubridate 1.9.3     ✔ tibble    3.2.1
## ✔ purrr     1.0.2
## ── Conflicts ──────────────────────────────────────────
tidyverse_conflicts() ──
## ✖ tidyr::extract()   masks magrittr::extract()
## ✖ dplyr::filter()    masks stats::filter()
## ✖ dplyr::lag()       masks stats::lag()
## ✖ 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)

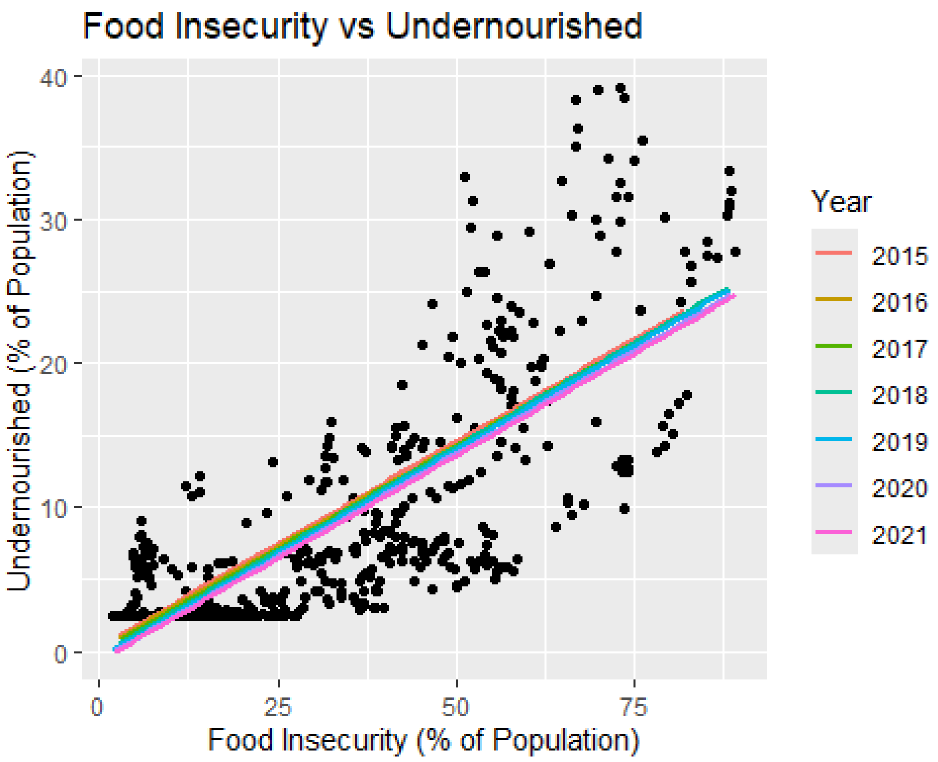
- HRFS12M1 : Categorical - Food Security rating (1: Food Secure 2: Low food security 3: Very low food security)
- HRFS12MC : Categorical - Food Security rating in children (1: Food Secure 2: Low food security 3: Very low food security)
- HEFAMINC : Categorical - Family income
- GESTFIPS : Categorical - 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

```
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)

# 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
```

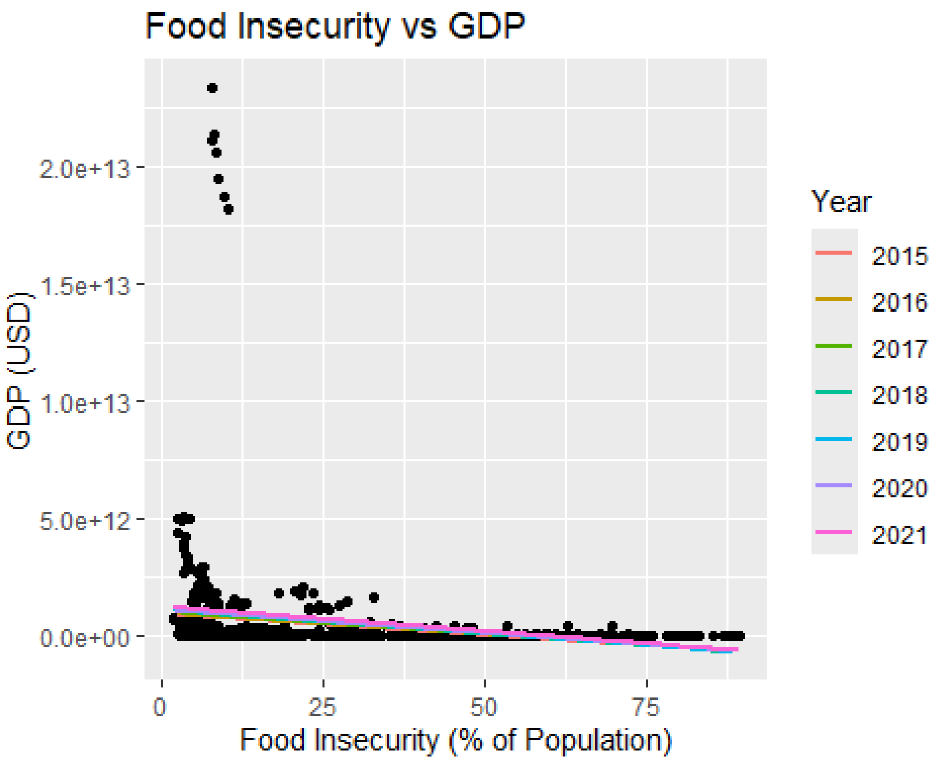
```
# Plot FoodInsecurity vs Undernourished to check for relationship
InsecurityUndernourishedPlot <-
    ggplot(combinedCountryData, aes(x=FoodInsecurityPercentPop,
y=UndernourishedPercentPop)) +
    geom_point() +
    geom_smooth(method=glm, se=FALSE, aes(color=Year)) +
    labs(x='Food Insecurity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

InsecurityUndernourishedPlot
```

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



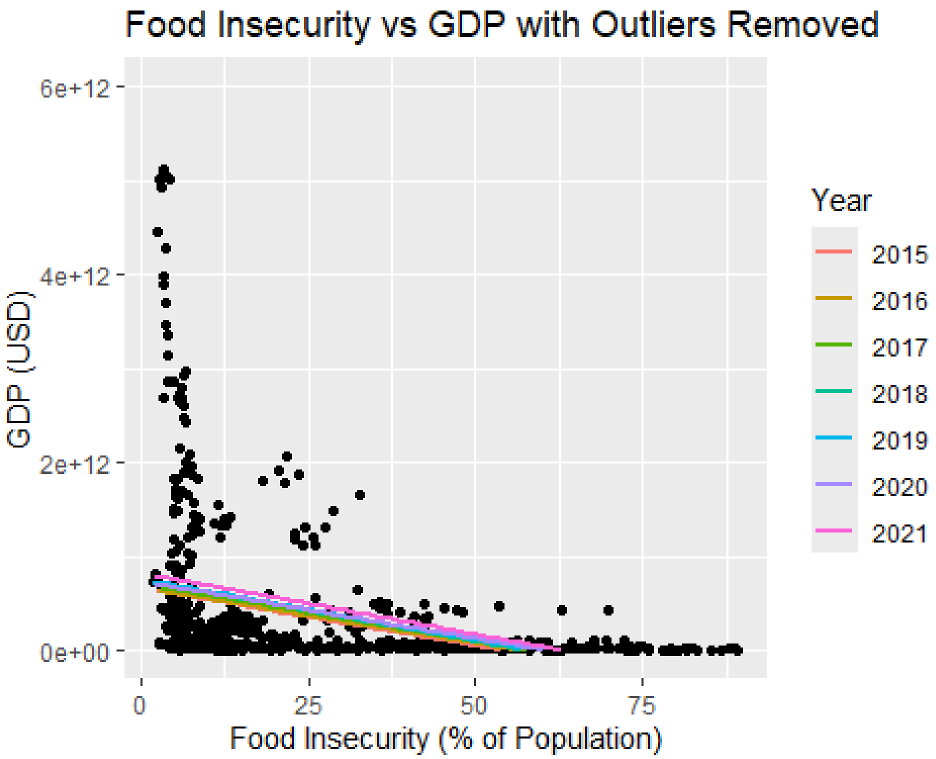
- The plot above shows the relationship between the percent of population with food insecurity and the percent of population that is undernourished. I have also added a line of best fit that represents each year. There appears to be a positive relationships that is consistent throughout the years.

```
# 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'
```



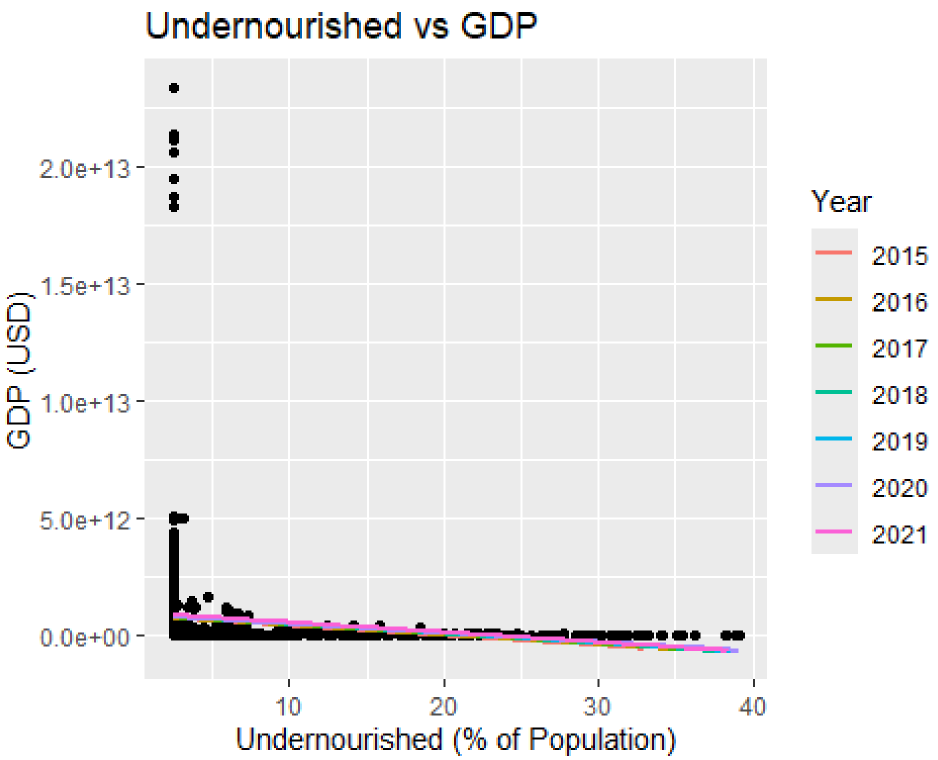
- There are some obvious outliers in this dataset. Likely there are countries with GDP that are much higher than others. It may be necessary to remove these countries. If removing these countries is not an option, it may be beneficial to plot them separately. Even with outliers, it appears there is a 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 outside the scale range
## (`stat_smooth()`).

## Warning: Removed 7 rows containing missing values or values outside the
scale range
## (`geom_point()`).
## Warning: Removed 191 rows containing missing values or values outside the
scale range
## (`geom_smooth()`).
```



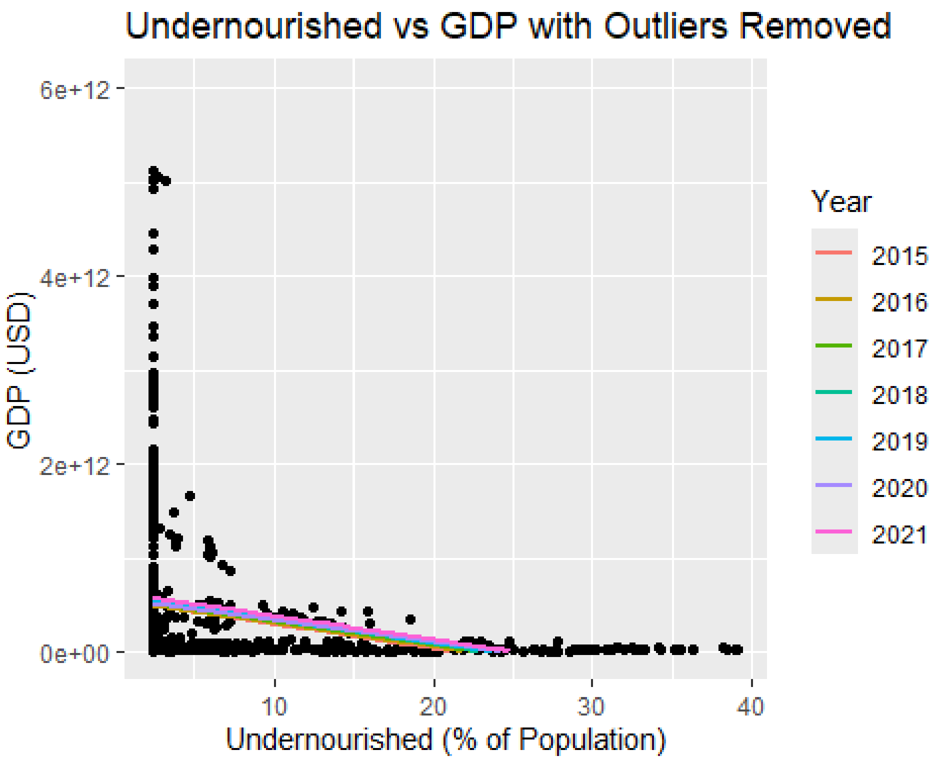
- Removing these outliers from the plot has confirmed the negative relationship. A linear model may not be the best fit for this data.

```
# 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 outside the scale range
## (`stat_smooth()`).

## Warning: Removed 7 rows containing missing values or values outside the
scale range
## (`geom_point()`).

## Warning: Removed 224 rows containing missing values or values outside the
scale range
## (`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.

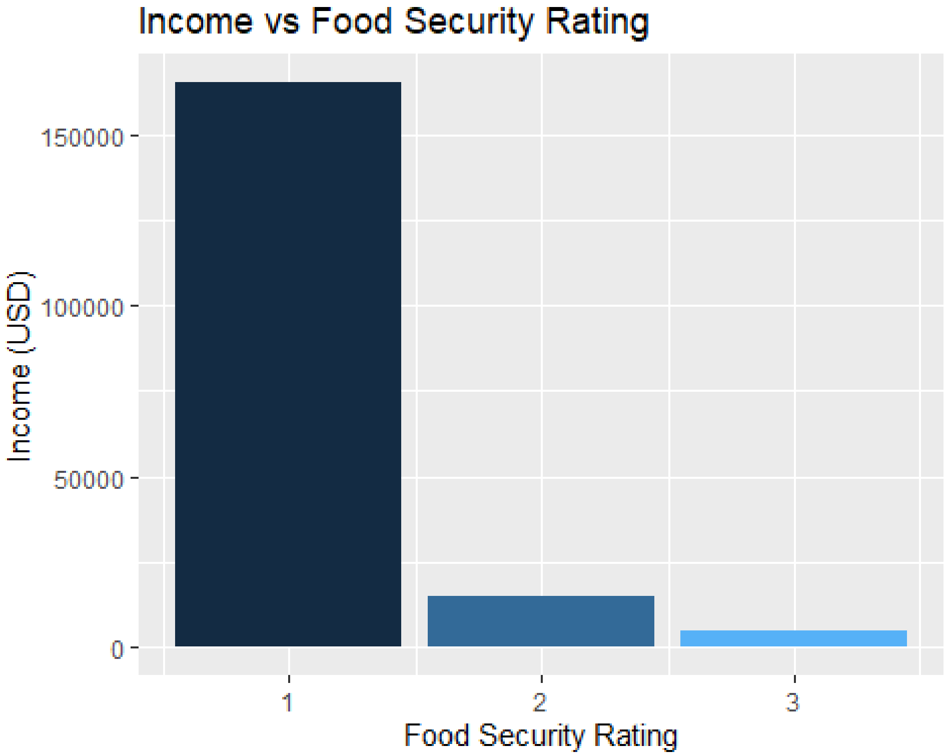
- Investigating GDP vs Undernourished and GDP vs FoodInsecure shows countries with higher GDP have a lower rate of both undernourishment and food insecurity.

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.

```r
# 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.

## 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.