Final Presentation

Arthi Vaidyanathan and Denise Shkurovich

11/16/2021

Introduction and Data:

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 decreased rates of mortality from COVID-19 and how it/if it is related to income levels. We also explored each nutritional group in isolation as well as alcohol consumption and fat consumption to analyze if there was a significant correlation with COVID recovery rates across countries.

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 and all data was last updated in February of 2021. The Food and Agricultural Organization of the United Nations as well as the Johns Hopkins Center for Systems Science and Engineering did not specify the methodology used to collect their data. In addition, data on the income group classification of each country was obtained from the World Bank.

Works Cited:

Iddir M., et al. Strengthening the Immune System and Reducing Inflammation and Oxidative Stress through Diet and Nutrition: Considerations during the COVID-19 Crisis. Nutrients. 2020;12(6):1562.

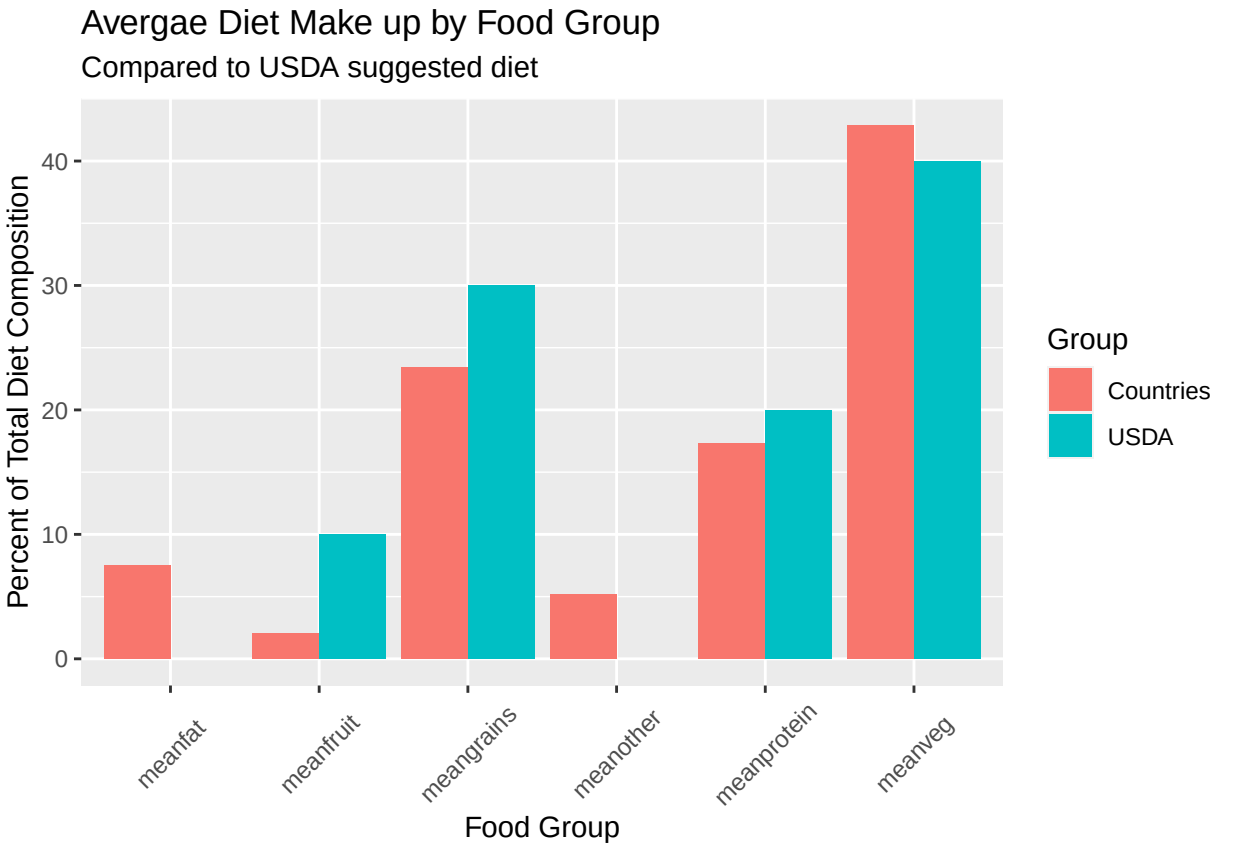
Messina G., et al. Functional Role of Dietary Intervention to Improve the Outcome of COVID-19: A Hypothesis of Work. International Journal of Molecular Sciences. 2020; 21(9):3104

Onishi J., et al. Can Dietary Fatty Acids Affect the COVID-19 Infection Outcome in Vulnerable Populations? mBio. 11(4).

Exploratory Data Analysis

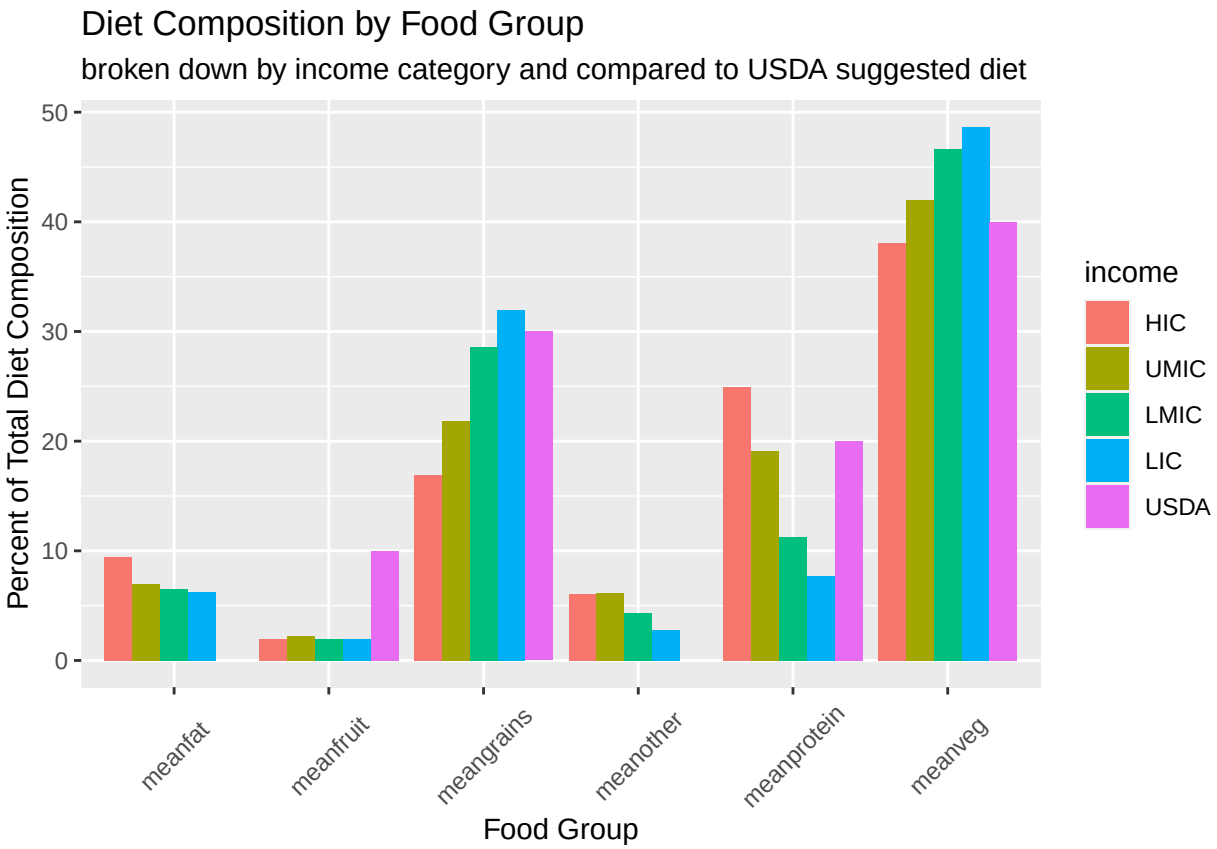
Food Group Visualizations

Our dataset included more specific dietary composition than the USDA values. Therefore, to be able to compare them, we grouped the datasets into broader dietary categories including grains, vegetables, fruits, fats, proteins, and a general miscellaneous category. We did so by reading the descriptions of what composes each variable in the dataset.



We were then curious to see if the mean percentages of the dietary categories of countries generally aligned with the USDA values. This helps inform us if we are making a valid comparison between the data and USDA values for the duration of the project. In order to do so we calculated the means of each of the values and added an additional column with the USDA values before plotting the bar graph. As shown, the values for grains, proteins, and vegetables are comparable to the USDA values indicating that our grouping of variables likely successfully captured the majority of dietary groups belonging to the broader categories. The value for fruits is significantly lower indicating that either the dataset does not accurately capture fruit consumption or countries on average eat significantly less fruit than recommended by the USDA. Furthermore, while USDA does not provide a value for suggested fat, it is a nutritional group significantly represented in our data set.

One of the major factors to consider in both dietary consumption as well as COVID-19 outcomes is income. Since some foods such as grains are typically cheaper than others such as proteins, income could play a significant role as a potential confounding factor. In order to account for this discrepancy we created a new income variable and classified each country according to their World Bank income status. HIC refers to high income countries, UMIC refers to upper middle income countries, LMIC refers to lower middle income countries, and LIC refers to low income countries.



In this graph, we summarized the break down of diet by income country in comparison to the recommended values from USDA. This visualization shows that low income countries seem to be consuming less than the recommended amount of protein and more grains than recommended. We can also see that higher income countries consume more protein and less grains than lower income countries. This graph suggests that in our analysis we should in fact include income level as a possible predictor of COVID outcomes since countries seem to show trends in their nutrition in most categories by income level.

Food Group Data Analysis

Below, we use ANOVA tests to see if the means for each of the dietary categories differ between at least one of the income groups.

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## income        3   5104  1701.4   86.92 <2e-16 ***
## Residuals   166   3249    19.6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Using the overall F test, we identified a significant difference in percent of diet that comes from grains among income groups. Our p value is «0.001.

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## income        3   2617   872.4   67.49 <2e-16 ***
## Residuals   166   2146    12.9
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Using the overall F test, we identified a significant difference in percent of diet that comes from vegetables

among income groups. Our p value is «0.001.

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## income        3    2.8  0.9269   0.458  0.712
## Residuals   166  336.1  2.0250
```

There are no significant differences in percent of diet that comes from fruit among income groups. Our p value is great than 0.05.

```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## income        3    288   95.99    14.3 2.48e-08 ***
## Residuals   166   1114    6.71
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Using the overall F test, we identified a significant difference in percent of diet that comes from fats among income groups. Our p value is «0.001.

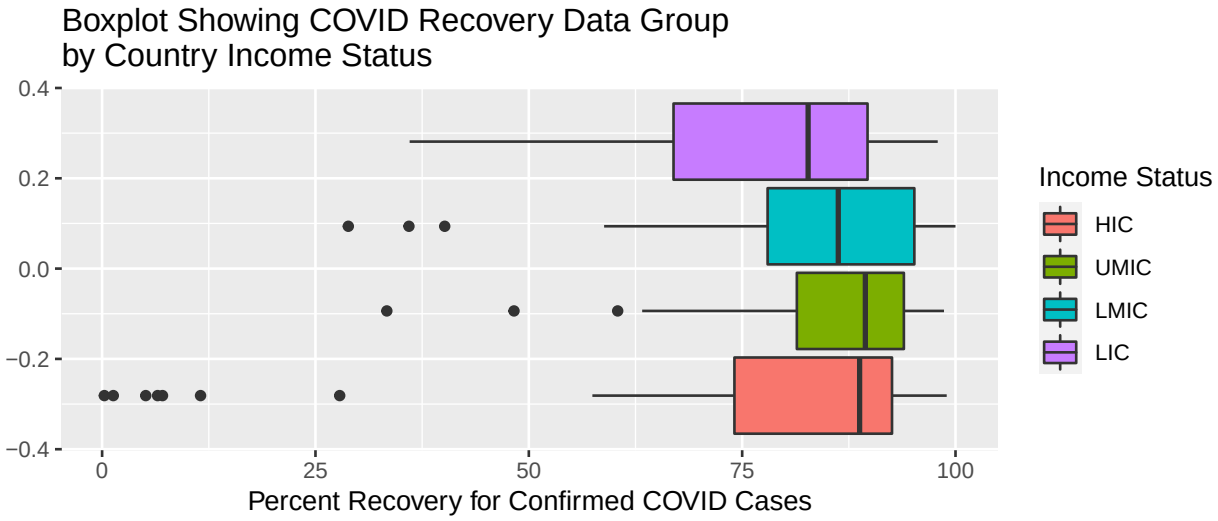
```
##              Df Sum Sq Mean Sq F value Pr(>F)
## income        3   6911    2304   67.69 <2e-16 ***
## Residuals   166   5649      34
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Using the overall F test, we identified a significant difference in percent of diet that comes from proteins among income groups. Our p value is «0.001.

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## income        3  98.03   32.68   32.79 <2e-16 ***
## Residuals   166 165.44    1.00
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Using the overall F test, we identified a significant difference in percent of diet that comes from alcohol/stimulants among income groups. Our p value is «0.001.

Recovery Rate Visualizations



In order to further our analysis, we created a variable that quantifies the COVID recovery rate for each country for comparison. We did this by dividing the percentage people in a country that recovered from

COVID over the percentage of total people in that country that were confirmed COVID cases and multiplying by 100 to obtain a rate of what percentage of people in that country that had COVID were able to recover. We also explored if income had an effect on percent recovery for COVID. High income countries hypothetically might have more resources to be able to provide better care for their citizens and improve their recovery rate. However, according to the boxplot, countries in the 4 income groups seem to have comparable recovery rates with the median for UMIC actually being the highest by a slight margin. Furthermore, the HIC group has numerous outliers near around 0.3 which might be important to consider. This seemed unusual to us due to the nature of disease and the resources the countries have. We speculate that more serious cases might be reported at disproportionately higher rates or there might potentially be other sources of error in the data set that we will later discuss.

Recovery Rate Analysis

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## income        3   3235  1078.5   2.546  0.058 .
## Residuals   156  66074   423.6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Based on the graphs, we wanted to see whether the mean recovery rate were significantly different between any of the income categories. ANOVA depends on variation between groups/within groups. From our graphs we can see that there is a decent amount of variation in recovery rate within each income category. This is probably due to the fact that these categories are very broad and there is different medical care/amount of COVID cases in each country even within income groups. However, the results of the ANOVA are not significant so therefore we cannot reject the null hypothesis that there are no significant differences due to income in recovery rate from COVID-19.

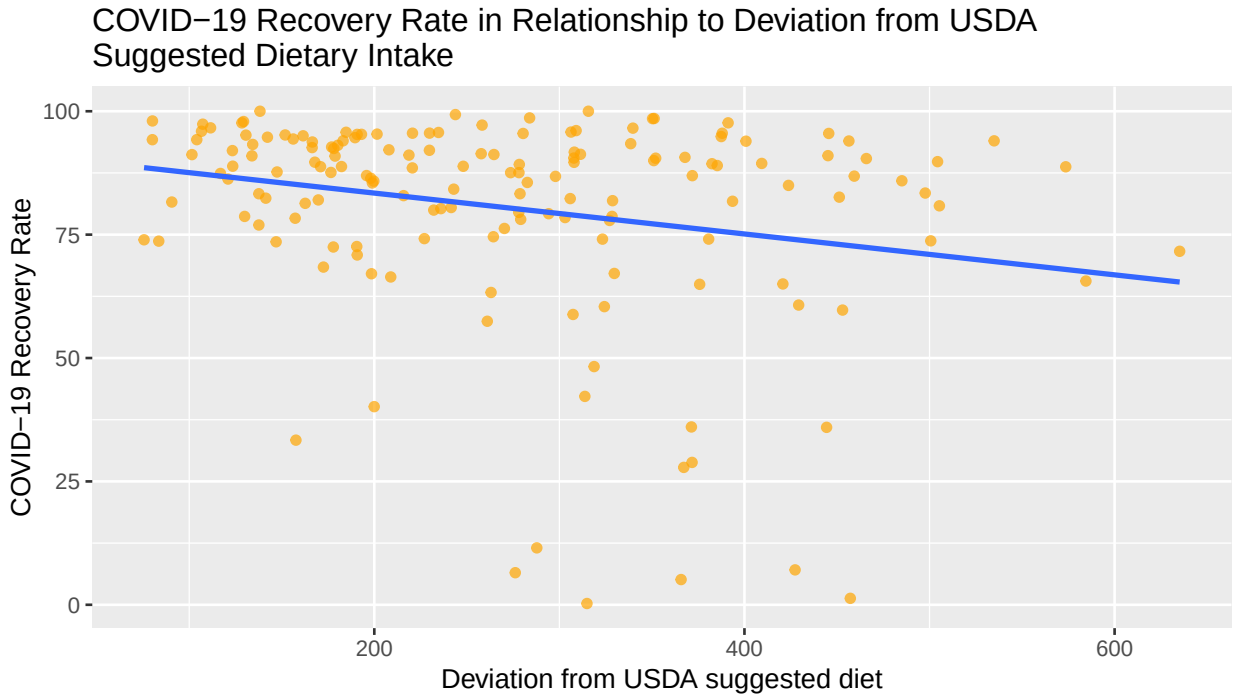
Dietary Deviation Variable Creation

We created a variable to quantify how much the average diet in each country differs from the suggested USDA value (This variable will be referred to as dietary deviation throughout the rest of this project). We did so by subtracting the mean value from each food group from the USDA value and then squaring that value. We then summed up this value from each food group to create total dietary deviation variable. We squared them to account for negative values if the average diet was less than the USDA value for a certain category as well as increase the spread of deviation values for better comparison and visualization. We filtered the dietary deviation category to remove 1 country (Iceland) that lacked accurate data in one of the food categories according to the dataset and therefore resulted in a falsely high deviation value.

Methodology and Results

Linear model of correlation between dietary deviation and COVID recovery rate

We used a linear regression model to see how COVID recovery rate is related to dietary deviation. We used a linear regression because both variables are continuous and quantitative.



Here we compared the relationship between deviation from USDA suggested intake with recovery rates from COVID-19. In the graph, we see that there is a negative correlation between recovery rates and deviation from USDA suggested dietary intake. This suggests that as ones dietary intake begins to differ from the USDA suggested intake, they on average seem to have a lower change of recovering from COVID-19.

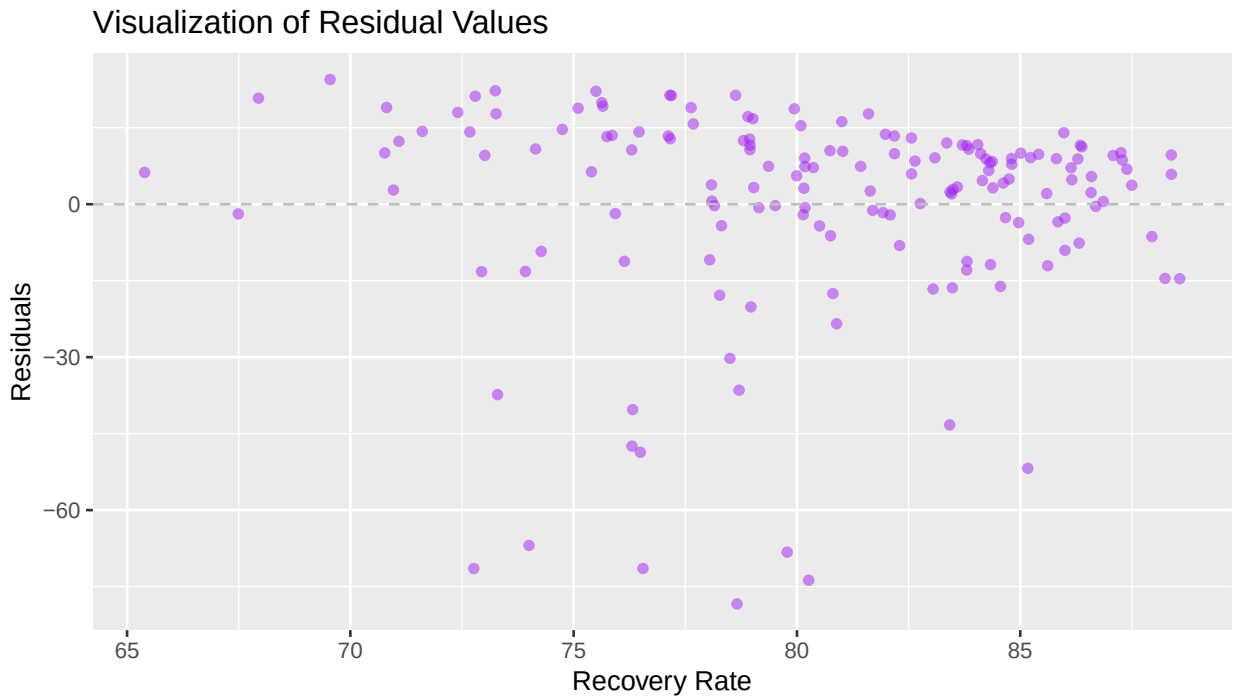
```
## # A tibble: 2 x 5
##   term          estimate std.error statistic  p.value
##   <chr>            <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)    91.7       4.00       22.9  5.43e-52
## 2 deviation_tot  -0.0414    0.0135     -3.07 2.54e- 3
```

The adjusted R^2 value for our model is:

```
## [1] 0.05055607
```

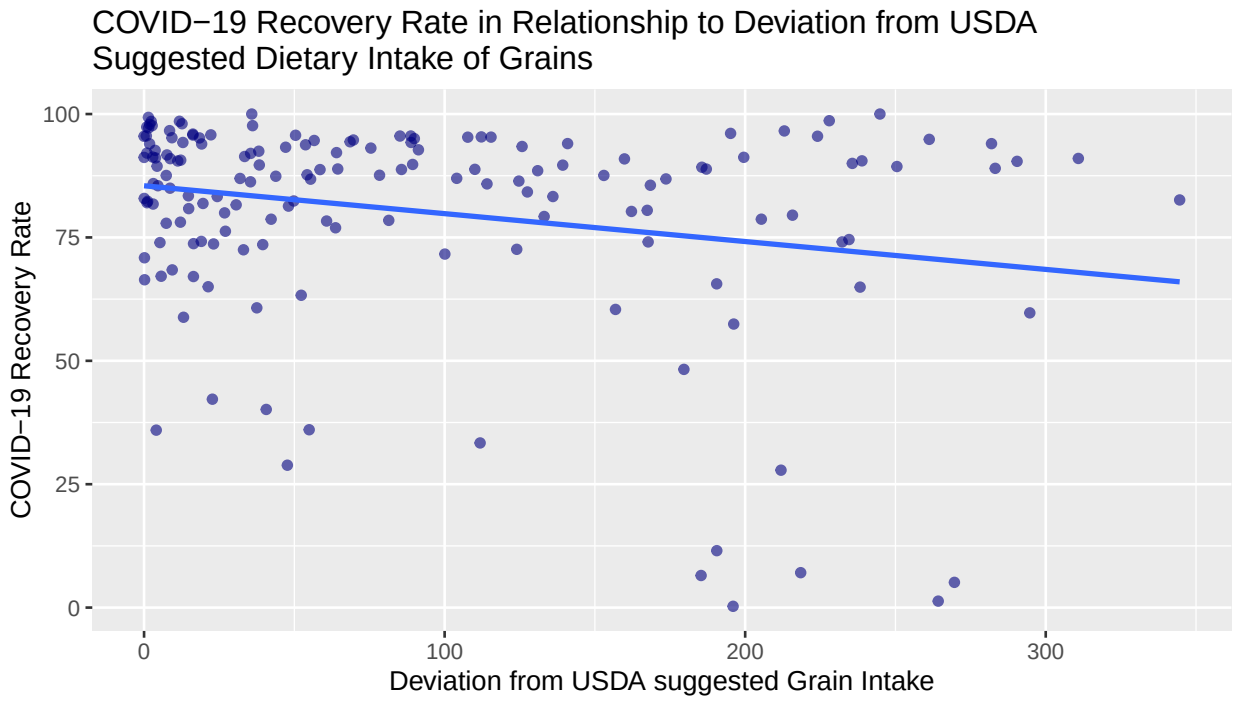
This linear regression model shows that there is a significant relationship between recovery rate and deviation rate since our p value is less than the alpha level of 0.05. For every integer increase in deviation rate, the chances of recovering are on average 0.041 lower.

However, the adjusted r squared value is quite low. It indicates that our model only accounts for 5% of the variance in the data.



In this model, the residuals do not seem to be even distributed around 0. The graph of the model along the the adjusted r square value, suggest that we are perhaps some other predictors of the outcome that we are not including in our model.

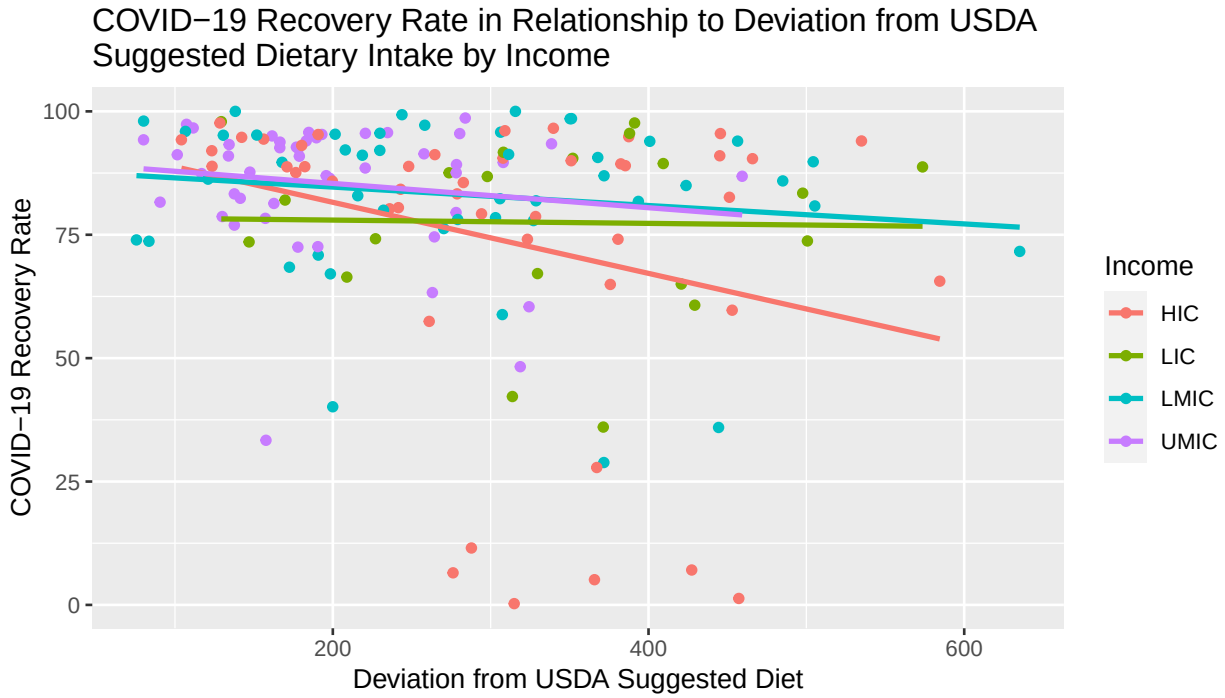
To further explore the various dietary factors in isolation, we ran linear regression models comparing each of the USDA category deviation values as well as Fat and Alcohol consumption levels to COVID recovery rates. The only regression model that reached the level of statistical significance was that of grain deviation as shown below.



```
## # A tibble: 2 x 5
##   term             estimate std.error statistic  p.value
##   <chr>               <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)       85.5       2.28       37.4  3.84e-80
## 2 deviation_grains  -0.0565    0.0183     -3.09 2.38e- 3
```

According to this model, as the deviation value for grains increases by 1 unit, COVID recovery rates decrease by 0.056%. This is a slightly stronger relationship than that of total deviation and suggests that grain consumption could play a more consequential role in COVID recovery rate than the other food groups explored.

Linear model of correlation between dietary deviation and COVID recovery rate accounting for income



To further analyze this relationship, we compared by income level. Since the majority of food groups included in the USDA guidelines showed significant differences across income groups as demonstrated by the earlier ANOVAs, we added income as a predictor to the model in an attempt to account for the variance in our model. We can see that while lower middle income and upper middle income countries have a slighty negative relationship, lower income countries almost show no relationship between deviation rate and recovery rate. In contrast, high income countries show the strongest correlation between deviation rate and recovery rate.

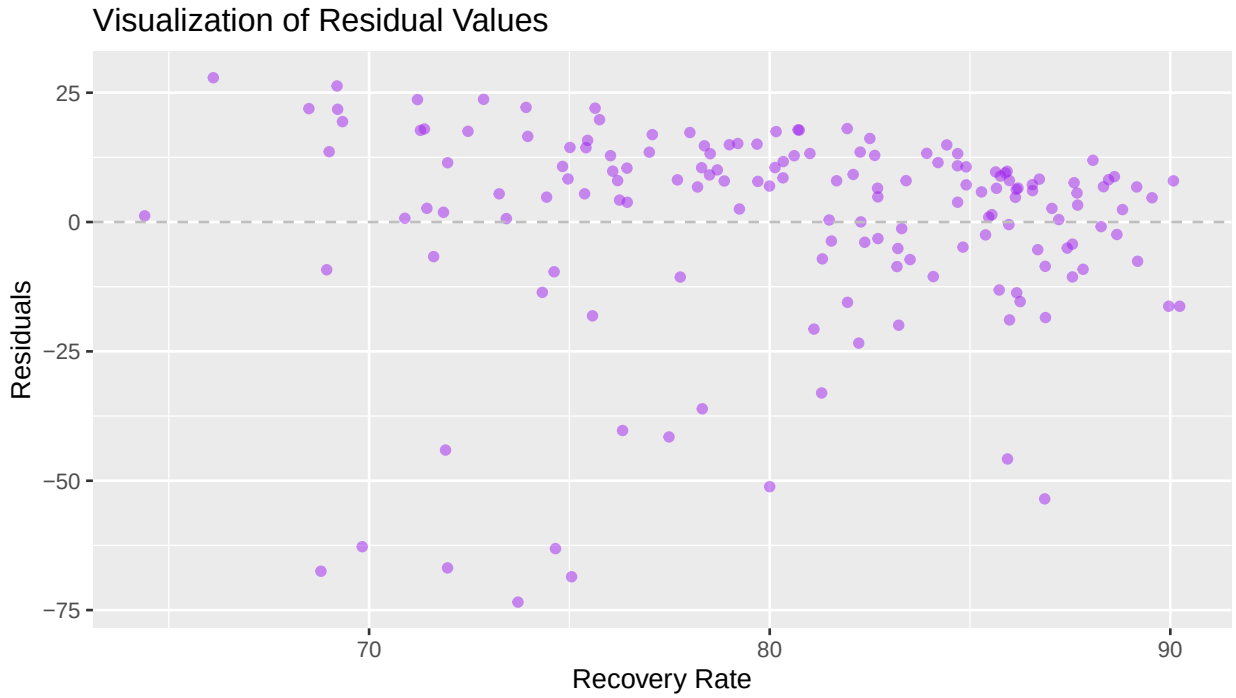
```
## # A tibble: 8 x 5
##   term                      estimate std.error statistic  p.value
##   <chr>                        <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)               96.0       8.38      11.5   2.95e-22
## 2 deviation_tot             -0.0720    0.0258    -2.79  5.88e- 3
## 3 incomeLIC                -17.3      16.0       -1.08  2.82e- 1
## 4 incomeLMIC                -7.59     11.1       -0.685 4.94e- 1
## 5 incomeUMIC                -5.65     11.8       -0.479 6.33e- 1
## 6 deviation_tot:incomeLIC    0.0687    0.0462     1.49  1.39e- 1
## 7 deviation_tot:incomeLMIC   0.0534    0.0350     1.53  1.29e- 1
## 8 deviation_tot:incomeUMIC   0.0474    0.0466     1.02  3.11e- 1
```

The adjusted R^2 value for our model is:

```
## [1] 0.06169532
```

This model shows that the slope for the reference group, HIC countries, has a value of -.07 meaning that in HIC as diet deviates from the USDA value by 1 unit, COVID recovery rate decreases by 0.07%. This value is significant with a p-value of 0.006. However, none of the other terms for income intercepts or interactions reach the level of statistical significance and therefore we fail to reject the null hypothesis that the value for the intercepts or slopes for the other income groups is the same as that of the reference group.

Furthermore, this model has a slightly larger adjusted r squared value, (6.17% compared to 5.05%) indicated that it accounts for slightly more variance in the data than the first model that did not include income.



This residual graph also indicates that the model that includes income as well as dietary deviation is a slightly better predictor of COVID outcomes than a model that solely relies on dietary deviation due to the minimally improved distribution. We still see residuals clustering along the x-axis indicating that the are further predictors we did not account for. Furthermore, none of the individual dietary factors in isolation showed statistically significant differences in effect on COVID recovery by income.

Discussion and Conclusion

The linear regression model does in fact show a significant relationship between COVID-19 recovery rate and dietary deviation from USDA suggested values, both overall and specifically for grains. While the value of that correlation is small, with a decrease of 0.04% in recovery rate per change in deviation dietary value, the burden of disease for COVID is massive and ongoing and even a small change in recovery rate has significant implications for human health. As shown, the linear regression model accounts for a very small percent of the variation in the data. Although income seems to mediate certain aspects of diet composition itself, it did not have a significant effect on the relationship between diet deviation and COVID recovery rate as shown by the linear regression model. In the linear model accounting for income however, the HIC countries did seem to show a stronger negative relationship between dietary deviation and COVID recovery rate which was potentially modulated due to the few HIC countries that have a significantly lower recovery rate than the rest of the dataset. Further studies should look into those values for those countries to see if they are true representations of recovery rate or the result of a source of error in reporting.

Furthermore, these food group categories are very broad. The majority of the studies that compare diets to immune system function look at composition of specific proteins, amino acids, or vitamins rather than groups as broad as carbs, fruits, or vegetables. Moreover, even within a country, each individual consumes a very distinct diet. Therefore, it is difficult to make predictions or understand relationships when there is such variation across individuals and communities in a population. We also used broad classifications for income groups that do not account for how wealth was distributed within the country as well as other government policies/funding related to healthcare and nutrition.

However, this data might shed light on the fact that the USDA recommended diet is too broad to make targeted conclusions about the impact of diet on health. Future studies should look at diet in a more controlled environment where it is possible to track individual food consumption and its effects on their health and COVID outcomes.

Another limitation of this study is that we did not control for vaccination rate, age, or immunocompromised status. Fully vaccinated individuals have been shown to be on average 13.1 times less likely to die from COVID-19 (CDC 2020). Therefore, future studies should control for vaccination rates, age, and general health across each country and condition on them when creating other linear regression models.

Furthermore, because the pandemic is still ongoing, rates of COVID-19 infection, recovery, and deaths are still changing. The last time that the data was updated was in February of 2022. These rates should be updated with the latest data to have the most accurate results.

Works Cited: CDC. (2020, March 28). COVID Data Tracker. Centers for Disease Control and Prevention. https://covid.cdc.gov/covid-data-tracker