Covid-19 and Nutrition: Correlation Connections

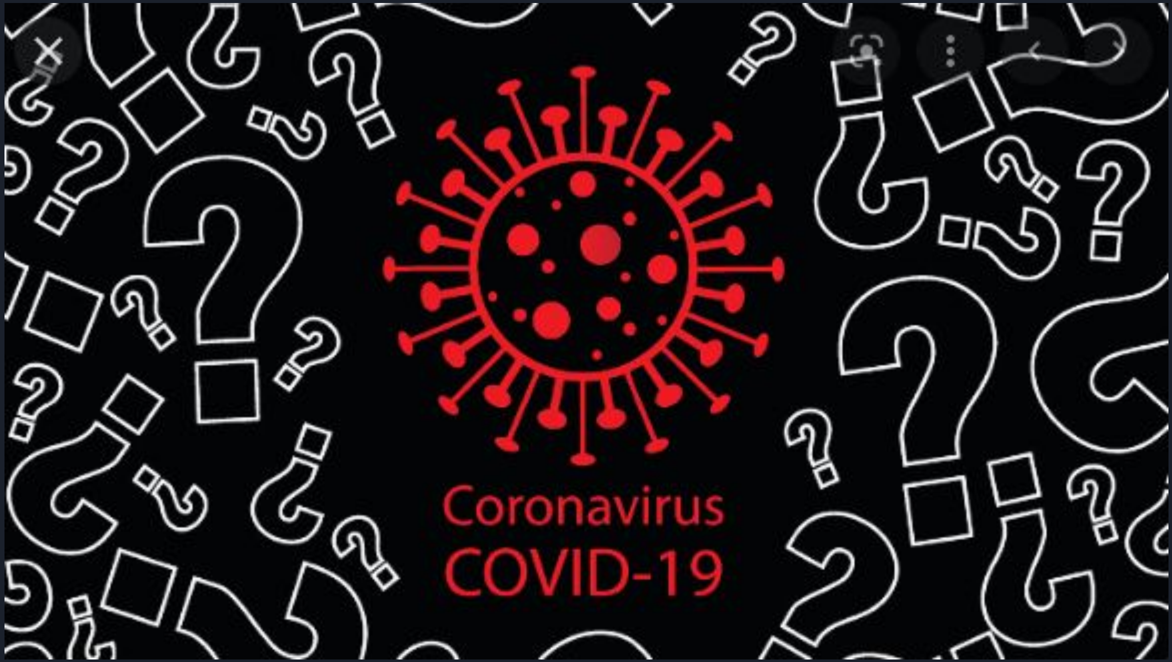
By:

Rebecca Gibbs
Bekka Ross
Witt Cordell
Sterling Miller

Topic and why it was selected

- Worldwide there have been 521 million cases of COVID-19 and 6.26 million deaths since first reports of the disease in December 2019
- It will be years before the short term and long impacts of the pandemic are understood
    - Health impacts to those infected
    - Financial/economic impacts
    - Societal impacts

Why Nutrition Data?



Data Sources

- COVID-19 Cases and Deaths by Country: https://covid19.who.int/WHO-COVID-19-global-table-data.csv
- COVID-19 Vaccination Dataset: https://ourworldindata.org/covid-vaccinations
- Nutritional Dataset by Country: https://www.kaggle.com/datasets/mariaren/covid19-healthy-diet-dataset?select=Food_Supply_Quantity_kg_Data.csv

Data Exploration Phase

- Searched public health sites and Kaggle for Covid related datasets
- Dataset originally chosen contained too many null values - scrapped and chose new dataset
- Covid datasets included basic COVID-19 info (cases/deaths/etc) by country
- Nutritional dataset (below) - % of food intake by category & obesity/undernourished rates

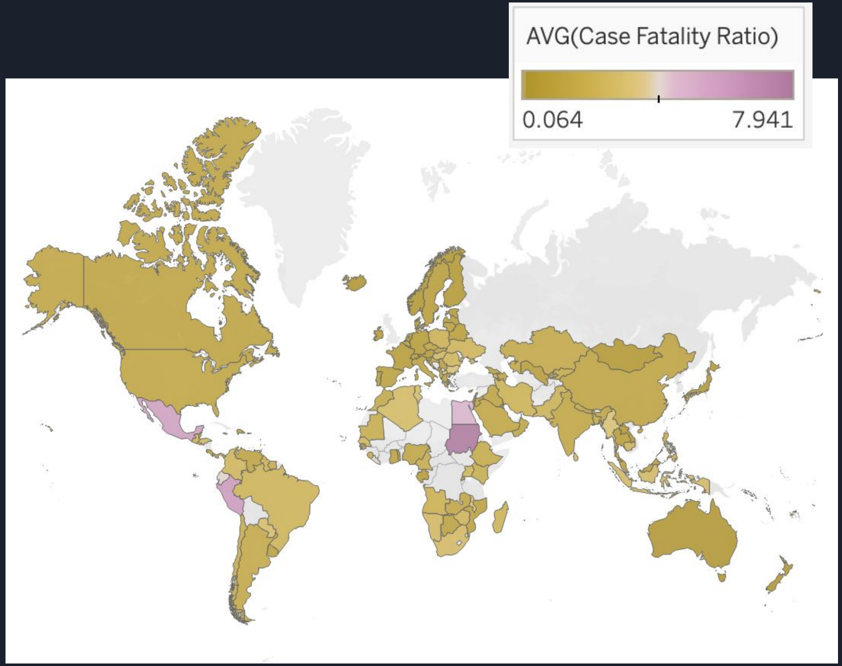
Country	Alcoholic Beverages	Animal fats	Animal Products	Aquatic Products, Other	Cereals - Excluding Beer	Eggs	Fish, Seafood	Fruits - Excluding Wine	Meat	Milk - Excluding Butter	Sugar & Sweeteners	Treenuts	Vegetable Oils	Vegetables	Vegetal Products	Obesity	Undernourished
Afghanistan	0.0014	0.1973	9.4341	0	24.8097	0.2099	0.035	5.3495	1.202	7.5828	1.3489	0.077	0.5345	6.7642	40.5645	4.5	29.8
Albania	1.6719	0.1357	18.7684	0	5.7817	0.5815	0.2126	6.7861	1.8845	15.7213	1.5367	0.1515	0.3261	11.7753	31.2304	22.3	6.2
Algeria	0.2711	0.0282	9.6334	0	13.6816	0.5277	0.2416	6.3801	1.1305	7.6189	1.8342	0.1152	1.031	11.6484	40.3651	26.6	3.9
Angola	5.8087	0.056	4.9278	0	9.1085	0.0587	1.7707	6.0005	2.0571	0.8311	1.8495	0.0061	0.6463	2.3041	45.0722	6.8	25
Antigua and Barbuda	3.5764	0.0087	16.6613	0	5.996	0.2274	4.1489	10.7451	5.6888	6.3663	3.8749	0.0253	0.8102	5.4495	33.3233	19.1	NA
Argentina	4.2672	0.2234	19.3454	0	8.4102	0.9979	0.4693	6.0435	7.0421	10.2328	3.0536	0.02	0.9541	4.3503	30.6559	28.5	4.6
Armenia	0.4014	0.1833	13.564	0	7.2982	0.5783	0.2896	6.0989	2.2675	9.9407	2.6579	0.1108	0.4705	16.7019	36.4358	20.9	4.3
Australia	5.5436	0.3143	21.4175	0.0033	5.4979	0.4428	1.4264	4.1883	6.7049	12.1018	2.5364	0.3176	1.2798	5.1406	28.5806	30.4	<2.5

Data Cleaning

- Used Pandas for most of data cleaning
  - Joined Covid & Vaccination datasets
  - Dropped columns
  - Removed null values
  - Calculated “case fatality ratio” column
- Used SQLAlchemy to import datasets into PostgreSQL as tables
- Used SQL code to join covid and nutritional datasets

COUNTRY	WHO Region	Case_Fatality_Ratio
Albania	Europe	1.269711106
Algeria	Africa	2.586518363
Angola	Africa	1.913642773
Argentina	Americas	1.414397903
Armenia	Europe	2.039018833
Australia	Western Paci	0.119539245
Austria	Europe	0.468346951

Analysis Phase



- **Goal:**
  - Predict likelihood of fatality from COVID-19 based on nutritional and COVID-19 record data.

Approach:
  - We opted for a multilinear regression for our analysis.
  - We aimed to produce fatality probability predictions for every country provided in the utilized data
  - We utilized R^2 value metric to determine the model's validity and accuracy
  - We utilized each input factor's p-value to determine its significance in the model's predictions and the calculations' results.

The Model

(Details)

Database Connectivity

Build Database Connection

```python
# Import Module to Communicate with PostgreSQL
import psycopg2 as pg

# Import Password Protector
from getpass import getpass
passw = getpass('Enter your Password')

# Build Engine for Connection
engine = pg.connect(database="Final_Project",  user="postgres", host="localhost", port="5432", password=passw)

dataframe = pd.read_sql('SELECT covid.*, nutrition."Alcoholic Beverages",nutrition."Animal Products",nutrition."Cereals - Excluding Beer",nutrition."Eggs",nutrition."Fish, Sea
```

Model Choice

Instantiate, Fit, & Evaluate the Model

```python
# Instantiate the Model
lr_model = LinearRegression()
```

Training / Testing Settings

Create Training & Testing Splits

```python
# Split Data into Training & Testing (Default: 75%/25% Split)
X_train, X_test, y_train, y_test = train_test_split(X, y, random_state=615, train_size=None)

# | Change the split % by editting the "train_size parameter to your training split percentage" |
    # Example: train_size = 0.80 results in an 80%/20% split
```

```python
# Preview the shapes of the Split Datasets
print(X_train.shape)
print(X_test.shape)
print(y_train.shape)
print(y_test.shape)
```

```
(98, 34)
(33, 34)
(98,)
(33,)
```

Model Accuracy, Model Validity, & Feature Significance

Model Effectiveness & Feature Weights

+ Code

```python
import statsmodels.api as sma
import numpy as np
```

```python
X2 = sma.add_constant(X)

est = sma.OLS(y,X2)
est2 = est.fit()

print(est2.summary())

# NOTE: R-squared represents our accuracy through accountability for variance
# while P>|t| values represent statistical significance of each input factor
# toward the model.
```

Results

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     Case_Fatality_Ratio   R-squared:                       0.603
Model:                             OLS   Adj. R-squared:                  0.463
Method:                  Least Squares   F-statistic:                     4.296
Date:                 Mon, 23 May 2022   Prob (F-statistic):           1.00e-08
Time:                         21:08:10   Log-Likelihood:                -148.97
No. Observations:                  131   AIC:                             367.9
Df Residuals:                       96   BIC:                             468.6
Df Model:                           34
Covariance Type:             nonrobust
=====================================================================================================================================
                                                              coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------------------------------------------------------
const                                                     214.3324   1162.016      0.184      0.854   -2092.251    2520.915
Cases - cumulative total                                -1.027e-07   6.07e-08     -1.693      0.094   -2.23e-07    1.77e-08
Cases - cumulative total per 100000 population           -4.48e-05    1.2e-05     -3.724      0.000   -6.87e-05   -2.09e-05
Cases - newly reported in last 7 days                    -4.47e-07   5.41e-06     -0.083      0.934   -1.12e-05    1.03e-05
Cases - newly reported in last 7 days per 100000 population 1.011e-06   0.001      0.001      0.999      -0.002       0.002
Cases - newly reported in last 24 hours                   2.56e-05   3.21e-05      0.798      0.427    -3.8e-05    8.92e-05
Deaths - cumulative total                                7.022e-06   3.15e-06      2.226      0.028     7.6e-07    1.33e-05
Deaths - cumulative total per 100000 population             0.0048      0.001      4.814      0.000       0.003       0.007
Deaths - newly reported in last 7 days                     -0.0012      0.001     -0.861      0.392      -0.004       0.002
Deaths - newly reported in last 7 days per 100000 population 0.0217     0.126      0.172      0.864      -0.229       0.272
Deaths - newly reported in last 24 hours                   -0.0023      0.007     -0.334      0.739      -0.016       0.012
...
Notes:
```

https://public.tableau.com/app/profile/witt.cordell/viz/FinalProject_Eda/Project_EDA#1

- Our model proved functional with database connectivity and fully integrated data cleaning and preprocessing.
- Our accuracy metrics determined a 60% accuracy based on initial input factors included, upon model revision the accuracy reduced to 30%
- Our p-value significance calculations proved that very few input factors from the original input factors were statistically significant toward the calculations that we were solving for in our model.
- In the end, our model was able to predict potentiality for fatalities from COVID-19 cases per country via a metric that indicates how influential the input factors were in determining whether or not a fatality was probable or not.

Recommendations and what we would have done differently

1. Devote more time to data sourcing
2. Consider input factors that relate more conveniently to individuals' health data
3. More time for model revision and accuracy improvement
4. Develop more nuanced data for each country's data point
5. More time dedicated to deriving our correlation

Thank You All!