Jordan Whitaker

BAN 530

Final Report

Analysis of COVID-19 in Developed Countries

**Executive Summary**

The COVID-19 pandemic has devastated the world. Some of the most well developed countries in the world, countries with access to advanced science and technology, have been disproportionally affected by this novel virus. This report analyzed a subset of these developed countries and sought to understand what might have led to these differences in outcomes.

To analyze the differences in outcomes three research questions were developed. These research questions have been discussed in the news-media and have been critical focus points for international public health experts. They are: *Does an increase in testing result in higher case counts? Do lockdowns, or tighter government measures, reduce case counts? What underlying health issues (comorbidities) can make a person more susceptible to contracting COVID-19?*

Before in-depth analysis was done, an objective determination had to be made to ensure that there was statistical evidence to pursue these research questions. This evidence was found in a statistical hypothesis test. The result of the hypothesis test determined that there was a statistical difference in at least one of the means of COVID-19 cases among the developed countries. Having found enough evidence to continue with the analysis, a variety of plots were created to visualize the data. The visualizations aided in determining what relationships exist within the dataset. These visual representations of critical data elements and their relationships provided more evidence to continue seeking answers to the research questions.

The next phase of analysis sought to understand what data elements are responsible for an increase in new COVID-19 cases. Predictive modeling was applied to the dataset, specifically; linear regression, polynomial regression, and machine learning models. After evaluating the performance of each model, the neural network model was the most capable of predicting new COVID-19 cases, with an R-Squared value of 0.96. The model was able to objectively assist in answering the research questions. The model concluded that an increase in testing will lead to an increase of newly reported cases. It also found that government interventions via lockdowns and other measures do, in fact lead to a decrease in new cases. And lastly, countries with high numbers of smokers and cardiovascular issues are likely to see an increase in new cases. Overall, the analysis in this report was able to successfully answer the research questions and the model that was created may be able to provide public health experts with more information.

It is likely that the long term solution to COVID-19 will be a vaccine. The development and distribution of a vaccine is a tremendous logistical challenge and was illustrated through an optimization case study. This case study underscored one of the major hurdles that pharmaceutical companies and governments will face: demand is greater than the current supply. This fact will cause governments to prioritize its citizens and ensure that the appropriate people will be the first recipients of the COVID-19 vaccine.

**1.0 Introduction**

This analysis will showcase the knowledge gained throughout the Master of Science in Business Analytics program via three major analytical domains: descriptive, predictive, and prescriptive. These themes will be used to provide insight to why there are varying outcomes of the COVID-19 pandemic in developed countries and challenges that these countries may face when administering a long term solution.

In this report, a developed country is defined as a country with a Human Development Index (HDI), greater than 0.915. The HDI metric is a criterion for assessing the state of development in a country by not focusing solely on the economic growth. HDI has been used to create an even scale when comparing countries with different economic situations. The United Nations Development Program defines the HDI as “a summary measure of average achievement in key dimensions of human development: a long and healthy life, being knowledgeable and have a decent standard of living. The HDI is the geometric mean of normalized indices for each of the three dimensions” (United Nations, 2020). To understand how these countries were affected so differently by COVID-19, research questions were established to analyze the data.

**1.1 Research Questions**

There are many possibilities for why developed countries were affected differently by COVID-19. This report will focus on three key features of COVID-19: COVID-19 testing, stringency index, and comorbidities of the country’s population. These features have headlined most global public health conversations and have been at the forefront of heated debate. The research questions that have been generated from these features are: *Does an increase in testing result in higher case counts? Do lockdowns, or tighter government measures, reduce case counts? What underlying health issues (comorbidities) can make a population more susceptible to contracting COVID-19?* These research questions will be addressed using a variety of statistical methods.

* 1. **Data Overview**

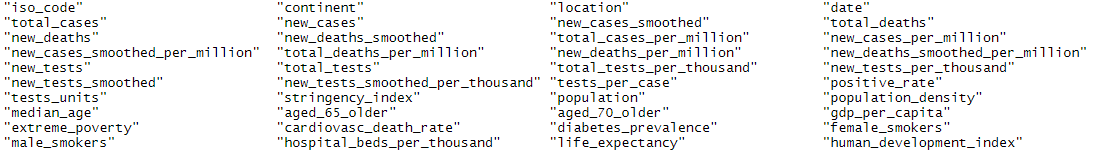
 The provided dataset is a comprehensive record of 211 countries dating from December 31, 2019 to October 18, 2020. There are 50,350 observations and 41 variables in the dataset. To review all variables in the dataset, refer to *Figure 1.2.1*.

Figure .2.1 Variables that have been provided in the COVID-19 data set

The scope of this analysis will focus on countries with an HDI > 0.915. The application of this constraint refines the dataset to 4,728 observations, and involves the following sixteen countries: Australia, Belgium, Canada, Finland, Germany, Hong Kong, Iceland, Ireland, Lichtenstein, Netherlands, New Zealand, Norway, Singapore, Switzerland, United Kingdom, and United States. With a defined scope of countries, further data analysis was required to account for missingness and other data inconsistencies. The dataset indicates that there were varying levels of success with reporting case counts and other health metrics. Upon further examination of the scoped dataset, it was determined that Lichtenstein and Hong Kong had inconsistencies in reporting data, which resulted in those countries being removed from the dataset. To assist in answering the research questions, a derived variable (Percent of Population Infected) was created. This variable took the new cases and divided that number by the country’s population and was then multiplied by 100.

**2.0 Descriptive Analytics**

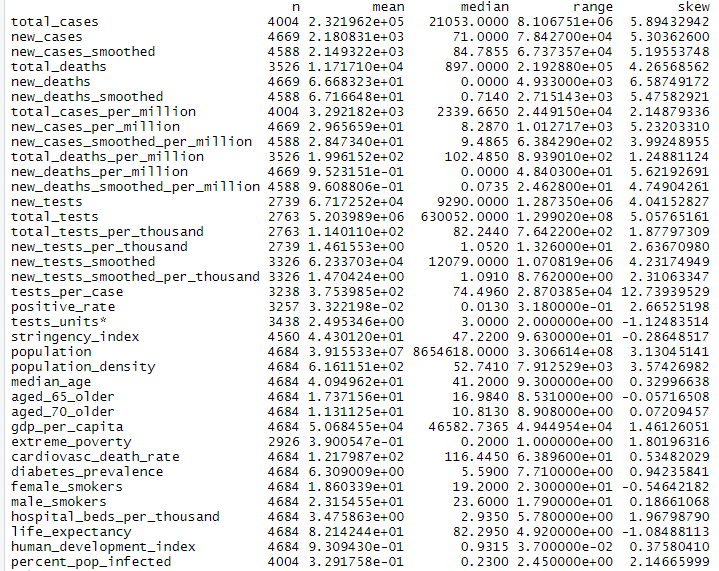
 To examine the proposed research questions, graphical representations of the data was used to showcase the current state of cases within each country, case count trends, infection rates, health metrics, and other relevant data figures. The chart in *Figure 2.1* provides a statistical overview of the dataset.

Figure .1 Descriptive statistics output for all variables

The statistical chart in *Figure 2.1* shows that many of the variables have a heavy positive skew. A positive skew indicates that the mean of the variable will be greater than the median, which indicates that there is high level of variability and potentially extreme outliers in the data.

The identification of a heavy positive skew throughout the dataset underscores the earlier mentioned points of how developed countries were disproportionately affected by COVID-19.

A snapshot of current case counts as of October 18, 2020 is depicted in *Figure 2.2*. The figure is a boxplot, which provides an overview of cases through a statistical lens of quartiles. The line in the middle of each plot shows the median value for each country’s current case count. This boxplot visualization has been plotted on a y-axis with a Log2 scale. Utilizing the Log2 statistical scaling technique was important as it provides a normalizing aspect to the data, which aided in dealing with the high variability. The United States has a significant number of cases compared to the other developed countries, and countries like New Zealand and Iceland have much fewer cases.

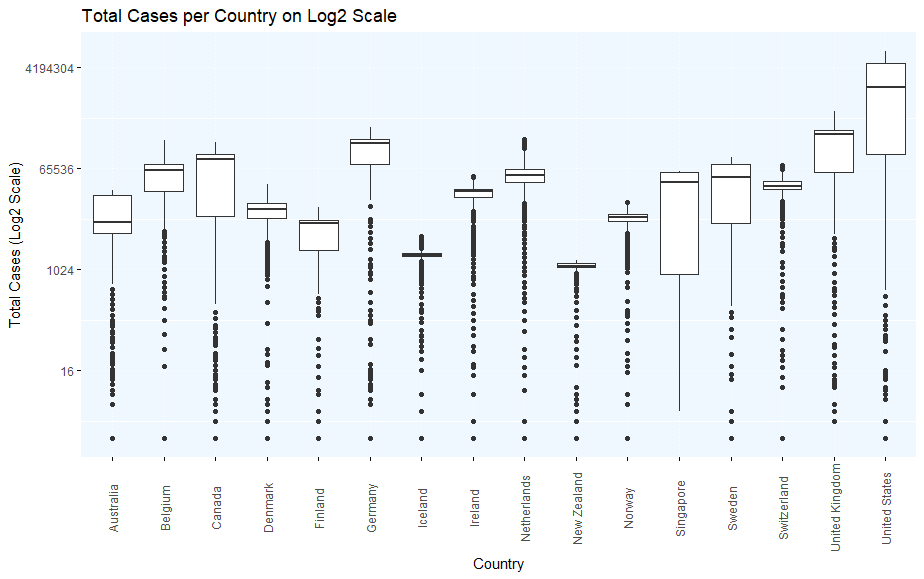


Figure 2.2 Box plot on Log2 scale of total COVID-19 cases per country

It was critical to develop an objective framework to assess the data and to determine if there was evidence to continue with analysis. The framework to assist in the process was a hypothesis test. The two components of a hypothesis test are the null hypothesis, expressed as *H0* and the alternative hypothesis, expressed as *Ha*. *Figure 2.3* shows the expression of the hypothesis test.

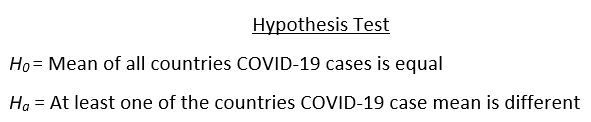


Figure 2.3 Expression of hypothesis test

To test this hypothesis, an ANOVA test was conducted. The output of the ANOVA test can be viewed in *Figure 2.4*. Based on the output of the ANOVA test, the alternative hypothesis was accepted, therefore at least one country’s mean was different. There was enough statistical evidence to continue with this analysis and to seek answers to the research questions.

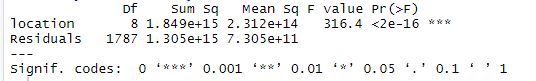


Figure 2.4 ANOVA Output

The first research question that was evaluated was the testing efforts. Administration of COVID-19 tests has been a critical component of understanding the spread of infection and the number of persons infected within a country. *Figure 2.5* examines these efforts by comparing the total cases per country to the total tests that have been administered. *Figure 2.5* indicates that there was a relationship between the number of tests that a country has done and the number of cases within that country.

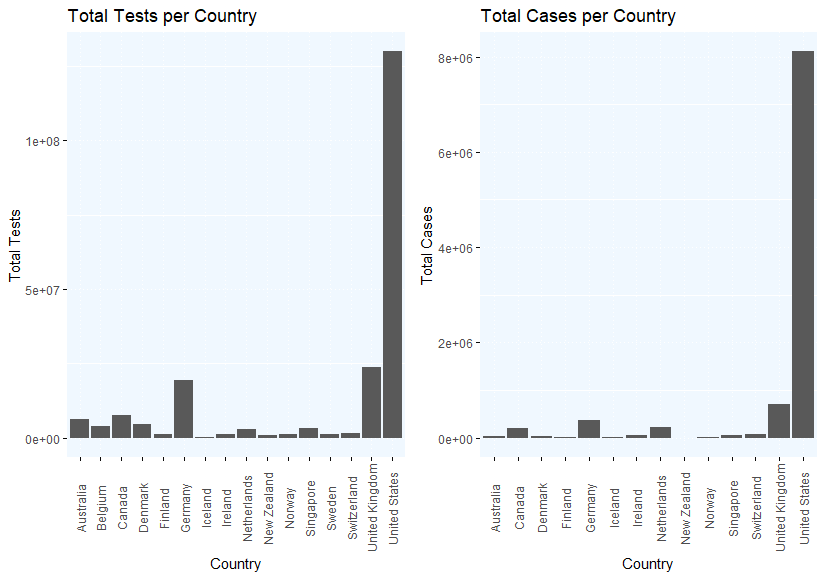


Figure 2.5 Comparing total COVID-19 tests per country to the total number of COVID-19 cases per country

Many developed countries have been fearful of fully closing due to the potential unintended economic consequences. The stringency index is the metric that defines how strict a country is, based on nine response indicators including school closures, workplace closures, and travel bans, rescaled to a value from 0 to 100 (100 = strictest response). *Figure 2.6* examines the relationship between the stringency index of a country and the number of new cases being reported.

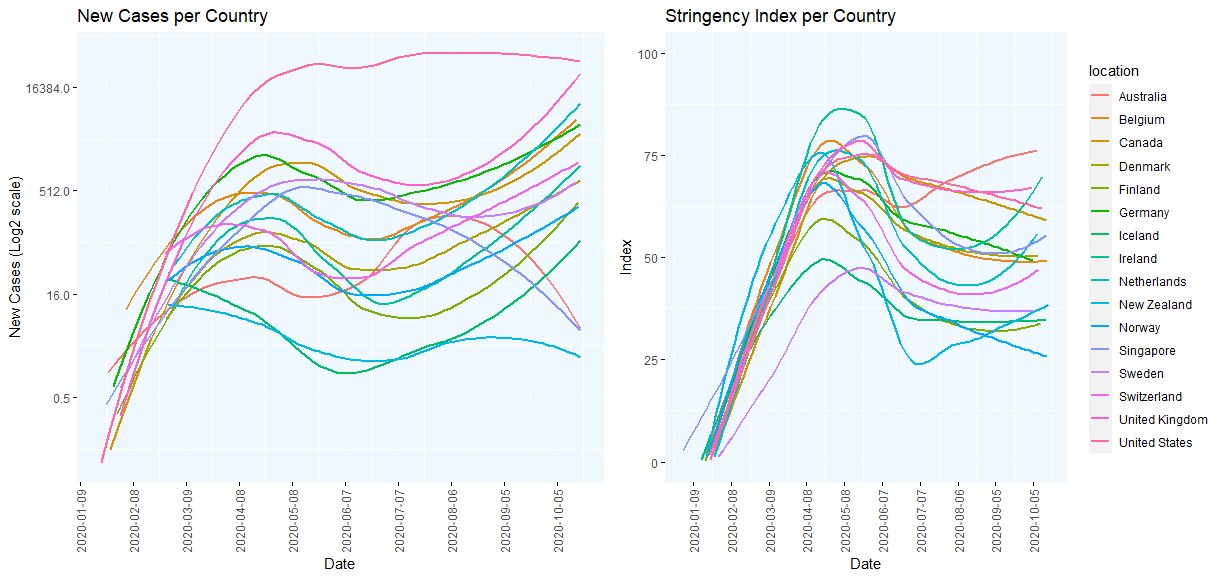
When examining *Figure 2.6*, there was a clear lag between the number of new cases and the stringency index. This indicates that there was a reactive response by these countries instead of a proactive one. This could be due to unintended consequences of shutting down the country. There also seems to be an inverse relationship between the number of new cases and the stringency index (as the stringency index goes up, new cases go down).

Figure 2.6 New Case trends compared to the Stringency Index, broken down by country

The last grouping of features that were explored were some of the health characteristics of the developed countries. These characteristics included the combined share of male and female smokers, death rate from cardiovascular disease (annual number of deaths per 100,000 people as of 2017) and diabetes prevalence (percentage of population aged 20 to 79 as of 2017).

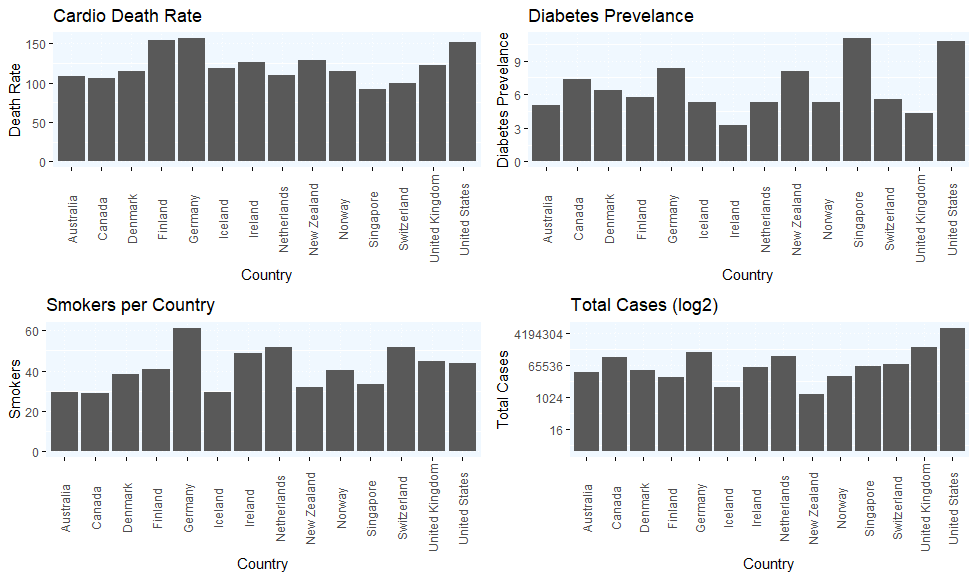
These health characteristics were compared visually to the number of total cases in their respective countries. This comparison can be viewed in *Figure 2.7.*

Figure 2.7 An overview of comorbidities V COVID-19 cases per country

*Figure 2.7* depicts how the rate of comorbidities seem to be related to an increase in COVID-19 cases. This can be seen in countries like Germany. Germany has high rates of cardiovascular deaths, and high rates of smokers, which may make that population more susceptible to contracting COVID-19. While countries like New Zealand and Iceland have a relatively healthy population and have had fewer COVID-19 cases. The rates of comorbidities within the countries seems to indicate that they are related to the number of COVID-19 cases within a given country. With an understanding of how the data looks and what potential relationships exist, the next phase of analysis was predictive.

**3.0 Predictive Analytics**

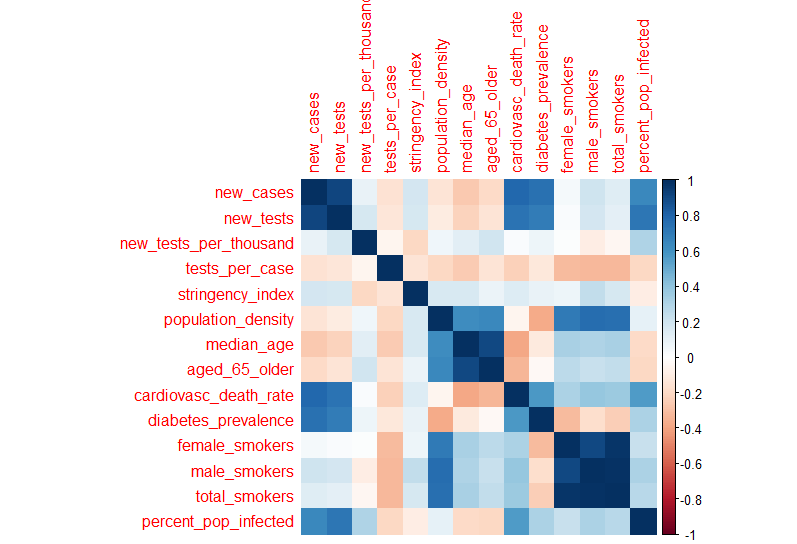
Predictive modeling is an analytical technique that uses past data to predict future outcomes. In the context of this report, the analysis was focused on predicting new COVID-19 cases. To enable the predictive model, it was required to prune the dataset of redundant variables to ensure that the model will not overfit the data. A model that is overfit will contain more parameters that can be justified by the model, therefore resulting in an artificially accurate model. After pruning variables, the predictors that remained were: new cases, new tests, new tests per thousand, tests per case, stringency index, population density, median age, aged 65 or older, cardiovascular death rate, diabetes prevalence, total smokers (sum of female and male smokers), percent population infected. The relationship between these fields is illustrated in *Figure 3.1*.

Figure 3.1 Correlation Matrix of Predictors

After feature selection, another data cleansing technique was applied to the data to assist the model. This was a systematic row-wise deletion of rows with missing data. The dataset now contains 1,796 observations. With feature selection and data cleansing completed, the next task was to implement a cross validation column. The implementation of a cross validation column ensures that there is no inherent bias during model development and training. To establish a validation column, it was decided that a 60 percent, 20 percent, 20 percent split should be implemented. That means that 60 percent of the data will be used for model training, 20 percent for model validation, and 20 percent for model testing. The results of the testing data will be the unbiased results of the models. Once the validation column has been established, model development can begin.

There were a variety of models that were chosen and applied to the data. These models include a linear regression model, polynomial regression model, random forest, boosted random forest, and a neural network. After applying all the chosen models to the data, they each performed relatively well with R-Square values above 0.90. However, as depicted in *Figure 3.2*, the model that performed the best, with an R-Square of 0.96 was the neural network.

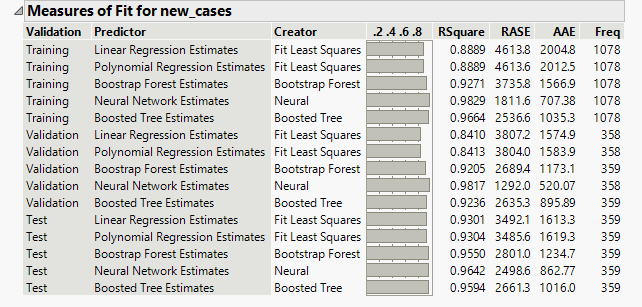
Neural networks are complex algorithms that have been designed to mimic the synapsis of the human brain. To mimic the biology of a human brain, the neural network creates a collection of connected units, referred to as nodes or artificial neurons. In this analysis, two layers and all three activation functions were leveraged to create a robust neural network. With a champion model having been identified, further analysis was done to investigate which predictors were significant to the model. In *Figure 3.3*, the variable importance output shows that the variable that had the greatest effect on the model was the percent of the population that is infected, which had an effect of 0.67. Other variables that had a significant effect on the model were the cardiovascular death rate (0.31), the stringency index (0.18), new tests (0.13), and diabetes prevalence (0.13).

Figure 3.2 Model Comparison Output

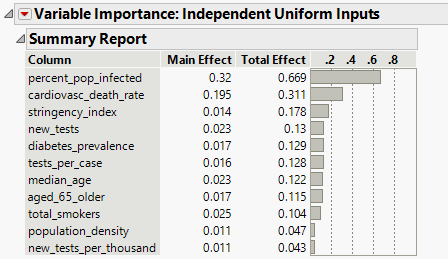


Figure 3.3 Variable Importance Output

To determine if the champion model can objectively assist in answering the research questions, the prediction profile was analyzed. Upon examining the prediction profiler (*Figure 3.4)*, there was a positive linear relationship between new cases and the number of new tests, which suggests that an increase in testing results in an increase number of new cases. When it comes to increased government intervention (stringency index), the prediction profiler indicated that a stringency index > 63 will result in a decrease in new cases. And lastly, the comorbidities that influence new cases are the total smokers and cardiovascular death rate of a country. Both comorbidity predictors have a positive linear relationship with new cases.

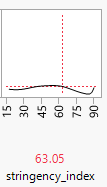
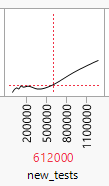
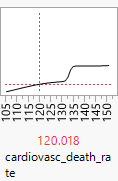
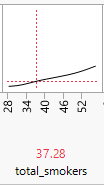


Figure 3.4 Predictors with a significant effect on the prediction of new COVID-19 cases

**4.0 Prescriptive Analytics (Case Study)**

Scientists across the globe have been working to develop a vaccine for the novel coronavirus, COVID-19. Traditionally vaccine development has been a multi-year process plagued by red tape and bureaucracy. However, due to advancements in science, global collaboration, and government intervention, a vaccine has been developed in record time and will soon be distributed to the public. This case study will examine the logistic challenges that will be faced by pharmaceutical companies as they begin distributing their vaccine. The objective of this case study is to minimize the cost associated with distributing the vaccine to the public. To derive an optimal solution for minimizing costs, it is critical that there is enough capacity (supply) to meet the demand. There are also a variety of defined variables and constraints that must be adhered to when solving for the optimal solution. The variables, constraints, and formulation can be seen below:

1, if the Warehouse in location *i* is open

0, otherwise. *i* {A, F, M, R, S}

Wi =

1, if the Distribution Center in location *j* is open

0, otherwise. *j* { B, F, M, A, V, R, H }

Dj  =

Variables Definition

Xij Number of vials transported from Warehouse *i* to Distribution Center *j*

*i* ϵ {A, F, M, R, S} and *j* ϵ {B, F, M, A, V, R, H}

Yja  Number of vials transported from Distribution Center *j* to Administration Location *a*

*j* ϵ {B, F, M, A, V, R, H} and *a* ϵ {L, N, C, MD, M, H, O, S, D, J}

Cij Cost of shipping from Warehouse *i* to Distribution Center *j*

*i* ϵ {A, F, M, R, S} and *j* ϵ {B, F, M, A, V, R, H}

Cja Cost of shipping from Distribution Center *j* to Administration Location *a*

*j* ϵ {B, F, M, A, V, R, H} and *a* ϵ {L, N, C, MD, M, H, O, S, D, J}

Fi Fixed Cost of opening a Warehouse in location *i, i* ϵ {A, F, M, R, S}

Fj Fixed Cost of opening a Distribution Center in location *j, j* ϵ {B, F, M, A, V, R, H}

Fv Fixed Cost of disposing of unused vials

Dv Demand for vials

Vi Variable Cost of Warehouse production

Objective Function

Obj: **Min:** Wi Fi + ΣCij + ΣXij Vi + Dj Fj + ΣCja + Fv Yja - Dv

Constraints

XAB **+** XAF + XAM + XAA + XAV + XAR + XAH <= 250,000Wa

XFB **+** XFB + XFA + XFA + XFV + XFR + XFH <= 250,000Wf

XMB **+** XMF + XMM + XMA + XMV + XMR + XMH <= 250,000Wm

XRB **+** XRF + XRM + XRA + XRV + XRR + XRH <= 250,000Wr

XSB **+** XSF + XSM + XSA + XSV + XSR + XSH <= 250,000Ws

XBL **+** XBN + XBC + XBMD + XBM + XBH + XBO + XBS + XBD + XBJ <= 4,800,000Db

XFL **+** XFN + XFC + XFMD + XFM + XFH + XFO + XFS + XFD + XFJ <= 4,800,000DF

XML **+** XMN + XMC + XMMD + XMM + XMH + XMO + XMS + XMD + XMJ <= 4,800,000Dm

XAL **+** XAN + XAC + XAMD + XAM + XAH + XAO + XAS + XAD + XAJ <= 4,800,000Da

XVL **+** XVN + XVC + XVMD + XVM + XVH + XVO + XVS + XVD + XVJ <= 4,800,000Dv

XRL **+** XRN + XRC + XRMD + XRM + XRH + XRO + XRS + XRD + XRJ <= 4,800,000DR

XHL **+** XHN + XHC + XHMD + XHM + XHH + XHO + XHS + XHD + XHJ <= 4,800,000DH

XBL **+** XFL + XML + XAL + XVL + XRL + XHL >= 361,796

XBN **+** XFN + XMN + XAN + XVN + XRN + XHN >= 315,967

XBC **+** XFC + XMC + XAC + XVC + XRC + XHC >= 218,106

XBMD **+** XFMD + XMMD + XAMD + XVMD + XRMD + XHMD >= 245,066

XBM **+** XFM + XMM + XAM + XVM + XRM + XHM >= 200,343

XBH **+** XFH + XMH+ XAH + XVH + XRH + XHH >= 197,602

XBO **+** XFO + XMO+ XAO + XVO + XRO + XHO >= 74,570

XBS **+** XFS + XMS+ XAS + XVS + XRS + XHS >= 35,611

XBD **+** XFD + XMD+ XAD + XVD + XRD + XHD >= 128,869

XBJ **+** XFJ + XMJ+ XAJ + XVJ + XRJ + XHJ >= 30,111

Wia +Wif + Wim + Wir + Wis >= 1

Djb +Djf + Djm + Dja + Djv +Djr + Djh >= 1

ΣXij >= Dv

With the variables, constraints, and formulation defined, the next steps were to determine the total capacity for the manufacturing facilities and to calculate the predicted demand for the chosen administration locations of the vaccine. In the provided case study, the known capacity of all manufacturing facilities totals 1,250,000. Using forecasting techniques and historical time series data, the total predicted demand for the administration locations in the case study is 1,807,943. This indicates that the demand for the vaccine is greater than the supply of vaccine. This will break the constraint ΣXij >= Dv which will lead to an infeasible solution.

This case study highlights the major challenge of distributing the COVID-19 vaccine; demand is much greater than the current available supply. It will be up to individual governments to prioritize its citizens and determine who will be inoculated first. Within the United States, the Centers for Disease Control has advised that healthcare professionals and residents of nursing homes should be the first recipients of the vaccine.

**5.0 Conclusion**

The focus of this report was to identify a subset of developed countries and attempt to understand why these countries had varying outcomes of COVID-19 cases. This subset of countries was defined as countries with a HDI > 0.915. To understand what might have led to these differences in outcomes, three research questions were established: *Does an increase in testing result in higher case counts? Do lockdowns, or tighter government measures reduce case counts? What underlying health issues (comorbidities) can make a population more susceptible to contracting COVID-19?* After in-depth analysis via descriptive and predictive analytics, answers to the research questions were found. Via machine learning, specifically a neural network model, it was determined that an increase in testing does lead to an increase in new COVID-19 cases, also stricter government interventions via lockdowns or other means can lead to a decrease in the number of new COVID-19 cases, and lastly, the comorbidities smoking and cardiovascular disease, can lead to an increase in COVID-19 cases. The analysis in this report was able to successfully answer the research questions and potentially assist international public health experts in stemming the spread of the COVID-19 virus.

A year has passed since the first reported case of COVID-19. In this past year the scientific community, governments, and citizens of the world have endured a once in a century event. When analyzing the data and generating this report, it was important to recognize the tremendous suffering that this pandemic has caused. The magnitude of this event is far-reaching and will have lasting effects on social, economic, and scientific communities.