Predicting Countries' COVID-19 Vaccination Sucess

Diego G. Davila Margaret Gardner Joelle Bagautdinova

Due before midnight, May 1st

1 Executive Summary

1.1 Background and Goals

The events of the devastating COVID-19 pandemic have been felt in all areas of human civilization. With the rapid development of mRNA vaccines to combat the spread of the virus, and reduce associated mortality and morbidity, societies have been slowly returning to a more normal way of life. However, this has not been uniform across countries. While in certain nations, vaccination campaigns have been largely successful, others have not fared as well and have faced a protracted pandemic. The efficacy of a nation's COVID-19 vaccination campaign is thus imperative to the well-being of that nation's populace, its global and domestic economy, and the prevention of the emergence of novel viral variants. Understanding the country-level factors that have led some nations to effectively vaccinate their population, and others to struggle with this process is, naturally, critical to tailoring international efforts to increase vaccination, and reduce the harm caused by the COVID-19 pandemic. Furthermore, the progression of Global Warming increases the likelihood of COVID-19-like pandemics in the future. Developing models to predict nations' vaccination effectiveness is of the utmost importance, and could guide the focus of future global resource allocation.

Given the above, the goals of this project are: 1. To elucidate the country-level variables that are predictive of a nation's vaccination success. 2. To develop models that can predict a nation's vaccination success given social, political, and economic variables, as well as features of disease toll prior to vaccine availability.

1.2 Data Description

To address the goals of our project, we have combined the following data sets to form an array of predictors:

- The World Sustainability Dataset: A data set describing several measures of a nation's social, political, economic, and climate-related characterization. This data set was originally aggregated for the TrueCue Women+Data Hackathon, and was released as publicly available after the event.
- The COVID vaccines: A data set containing several vaccination measures, including cumulative vaccinations, the total number of vaccinated people and the number of vaccine doses administered per individual. The data set is updated daily based on Our World in Data GitHub repository for covid-19 but for the purpose of this study, data includes vaccine rates up until March 17, 2022.
- COVID cases: A dataset containing daily reports of cumulative and new cases and deaths in each country. Country populations in 2021 are added to this data set to normalize absolute counts relative to population ("per hundred" variables) in addition to the existing cumulative cases and deaths.

1.2.1 Vaccination Success Index

The success of a nation's vaccination campaign reflected in many variables, including number of vaccinated, days to vaccinating percentages of the population, and daily vaccinations. Describing, quantifying, and predicting such a multifaceted concept can be a challenge. To address this, we have developed a quantified measure of vaccination success that takes into account several such factors pertaining to national vaccination campaigns. We have termed this measure the Vaccination Success Index (VSI). Demonstrated in more detail further, the VSI was developed by applying Principal Component Analysis (PCA) to various measures of world nations' vaccination campaigns. VSI serves as our primary outcome measure.

1.3 Results

Our analyses have yielded several useful insights that may help inform global policy decisions as international resources are allocated to help struggling nations increase vaccination success.

Linear Regression models based on economic factors, disease burden factors, and political factors, as well as data-driven LASSO and Regsubsets models, have converged on the following variables as significantly predictive of vaccination success:

- World Region
  - Policy Recommendation: Aid should focus on increasing vaccinations in struggling regions (i.e. Sub-Saharan Africa)
- Life Expectancy (Prior to Pandemic)
  - Policy Recommendation: Life expectancy often serves as an indicator of a country's overall medical infrastructure. Medical infrastructure resources should be the focus of aid, more so than economic relief alone.
- Women in Business and Law
  - Policy Recommendation: Initiatives to promote womens' equality and educational attainment may also have secondary benefits to vaccination and similar public-health campaigns and thus should not be overlooked.
- Regime Type
  - Policy Recommendation: Mutual aid relationships should be facilitated between nations of varying regime types.
- Number of Internet Users
  - Policy Recommendation: Internet access should be made widely available as part of aid packages. This may help to provide adequate information regarding vaccinations, and help coordinate a nation's vaccination logistics (i.e. online signups and messaging).
- Government Expenditure as Percent GDP
  - Policy Recommendation: Governments that spend more relative to GDP may be preemptively investing more in public health services and infrastructure required for rapid, broad-scale vaccination campaigns. Countries should be willing to invest in the public services that will be necessary in times of crisis.
- World Bank Income Class
  - Policy Recommendation: High-Income nations had more successful vaccination campaigns and should provide aid to lower-income countries to reduce global disease burden.

Furthermore, we have developed more robust models that predict VSI with varying degrees of accuracy (Linear Regression & LASSO, Trees & Random Forest, and Artificial Neural Networks). While predictive accuracy was moderate for each model individually, as well as an ensemble model that takes the weighted averages the predictions of all models, we have achieved a level of predictive ability that could help forecast vaccination success in subsequent COVID-19 vaccination campaigns and future pandemics.

1.4 Conclusion & Limitations

We have developed a robust vaccination index capturing a country's past vaccination success and our models highlighted several political, economic, social and health-related factors that have a crucial influence on a country's ability to vaccinate its populace. However, this analysis comes with certain limitations. First, the dataset is constrained to countries for which a sufficient amount of data was present to compute both VSI and summary political, economic, social and health-related factors. While this was possible for a substantial number of countries (n = 99), some countries had to be dropped (n = 126 from the vaccination data, n = 42 from the world sustainability data) including large countries such as China, which did not sufficiently report on vaccination data. By obtaining more complete information, future studies may further improve the completeness of this dataset and thereby increase the accuracy of VSI predictions. Moreover, in addition to the factors found to influence VSI in this study, it is likely that other factors influence a country's vaccination success, including more psychosocial and individual characteristics related to beliefs around vaccinations; however, it's likely that some of these latent factors may be captured in our data as they relate to educational attainment, regime type, and life expectancy when interpreted as a broad indicator of a country's healthcare and social safety net.

In conclusion, we were able to quantitatively define the success of each country's vaccination campaign against COVID-19 and to identify key, modifiable factors that may contribute. While the modest predictive ability of our final model on unseen data is one of several limitations to consider in this work, the COVID-19 crisis has presented a multifaceted and evolving challenge to governments across the globe. Therefore, our analysis represents an important contribution to objectively assessing each country's response and identifying factors that should be considered in addressing future global health crises.

2 Data Preprocessing, EDA & Summary Measures

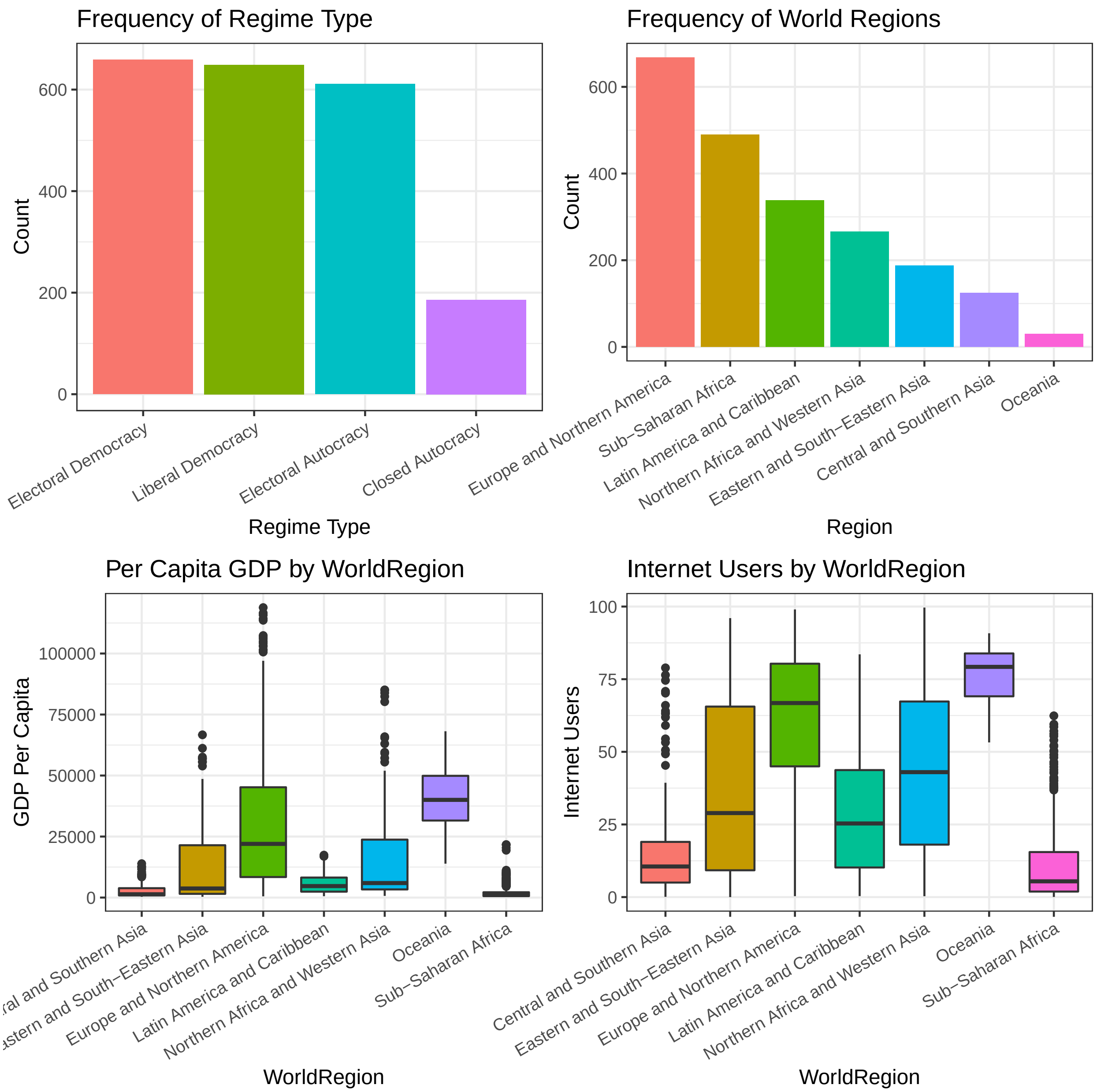
Before starting to do more involved EDA, we will preprocess our data to ensure we have workable datasets to begin our exploration with. At the completion of preprocessing, we will have two longitudinal dataframes: one containing WSD predictors and one containing COVID case, death, and vaccination data. These dataframes will subsequently be summarized to capture the most important features of each variable by country, enabling us to merge the dataframes and model on these synthesized per-country metrics.

2.1 WSD Preprocessing & EDA

First, we process the World Sustainability Dataset, cleaning missing elements, removing columns that have less than 90% complete rate, and removing rows with NAs in these remaining columns.

After cleaning, we are left with 141 countries, which is a workable amount. Next, we rename some columns to make them readable, turn strings into factors where appropriate, and export our processed data to "WorldSustainabilityData_Processed.csv".

We'll do some exploratory plotting for a few key predictors across all years of reported data.



Further visualizations can be found in the Appendix under `Additional EDA`.

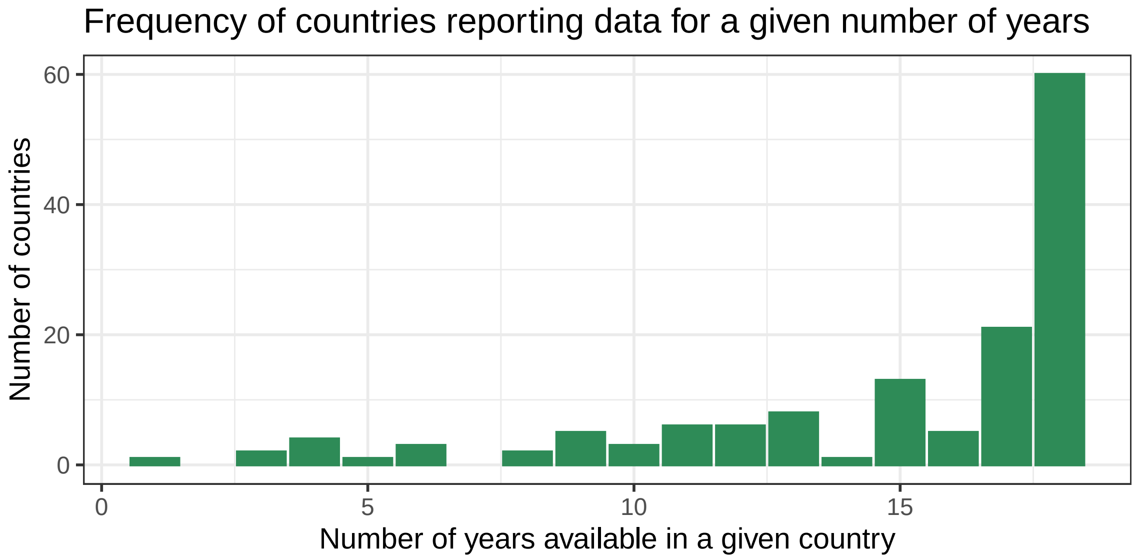
2.2 WSD Summary Measures

Next, we'll try several different methods of summarizing this longitudinal WSD data into a single "timepoint". We selected the latest data reported for each country (restricted to 2016 or later) and calculated: the average of each parameter over time (2001 to 2018); the average of the last 3 years of data (2016-2018); and the relative and absolute change in each parameter over time (2001-2018). In order to maximize completeness and recency of the dataset while minimizing colinearity, we decided to use each country's latest data and data averaged over 17 years in our models below. Code for unused summary dataframes can be found in the appendix.

First we extract just the last year of reported data for each country, restricted to 2016 or later. As seen below, the majority of countries have 2018 data, though we are able to preserve an additional 23 countries by reaching back to 2016 data.

```
##
## 2016 2017 2018
##    6   17  109
```

As seen below, there is a great deal of variability in how many years each country has data reported in the WSD dataset:



This is a limitation that should be kept in mind when interpreting the results of our model and the predictive power of averaged variables in particular, since an averaged GDP for one country might reflect 17 years of data while the averaged GDP of another may only reflect 3 years. A full graphical representation of years reported by country can be found in the Appendix.

Next we'll merge the latest data and averaged data into one wide WSD dataframe and remove any variables that are highly correlated. Plots displaying the correlations between the original WSD variables are displayed in the Appendix.

2.3 COVID Preprocessing & EDA

Next, we will read in COVID cases/deaths and vaccination data sets. We remove countries with too many NAs and generate new variables indicating cases and deaths per hundred, to account for the total population in each country. Finally, we'll merge the case, death, and vaccination info into a single dataframe.

We will start by visualizing some of the most relevant variables and patterns in the cases, deaths and vaccinations data. Further plots can also be found in the Appendix. To view current vaccination status nationally, we bin countries by their latest achieved vaccination rate (per hundred to account for their total populations):

```
## PhantomJS not found. You can install it with webshot::install_phantomjs(). If it is installed, please
```

As we can see, vaccination rates are currently still very variable across the world.

Let's now pick a few countries to examine the evolution of the number of cases with respect to important vaccination milestones, such as when vaccination start and when progressively greater percentages of the population have been vaccinated. Keep in mind that some countries have not yet reached some of these milestones as of today and therefore, a different number of milestones may be present for each country represented below. Plots relating COVID-19 deaths with vaccination milestones can be found in the Appendix.

Based on the plots above, we can see that countries had variable vaccination start dates and that each has reach a different number of vaccination milestones. For instance, Cuba started vaccinating relatively late (May 2021) but had 90% of its population vaccinated by January 2022. In the US, vaccinations started in December 2020 and have reached a rate of 70% in December 2021. It is unclear from these graphs whether vaccinations have direct consequences on the number of cases and/or deaths.

2.4 COVID Summary Measures

In order to come up with meaningful summary predictors and dependent variables that are compatible with our WSD, we'll next summaries our data on covid cases, deaths, and vaccinations. For our purposes in this project, we'll treat cases and deaths prior to each country's vaccine roll-out as independent variables (predictors) alongside the WSD. Vaccination data will be synthesized using PCA to create a the Vaccination Success Index (VSI), a composite score which we will use to indicate how well a country was able to vaccinate its citizens.

First, we create a dataframe with the latest COVID data from each country.

```
##         Min.      1st Qu.       Median         Mean      3rd Qu.         Max.
## "2022-01-07" "2022-03-01" "2022-03-02" "2022-03-02" "2022-03-06" "2022-03-17"
```

Every country has some data from March 17, 2022, though some countries' last reported vaccination data was from as early as 1.9×10^4.Next, we'll calculate predictors from COVID case and death rates prior to the day each country reported starting to vaccinate.

Finally we'll merge these COVID predictors into a single dataframe and remove strongly colinear predictors. Correlation plots can be found in the Appendix.

2.5 Computing Vaccination Success Index

Since defining a country's success in administering vaccines is multifaceted, we'll construct the VSI on several summary metrics in addition to final vaccination counts. For instance, it is likely meaningful to account for how long it took each country to reach certain milestones (such as first starting to administer vaccines, administering X doses to each citizen, or vaccinating X% of the population) as well its peak rate of vaccination (the maximum rate of daily or weekly vaccinations) in determining whether a country did a "good job" in administering COVID vaccines to its populace.

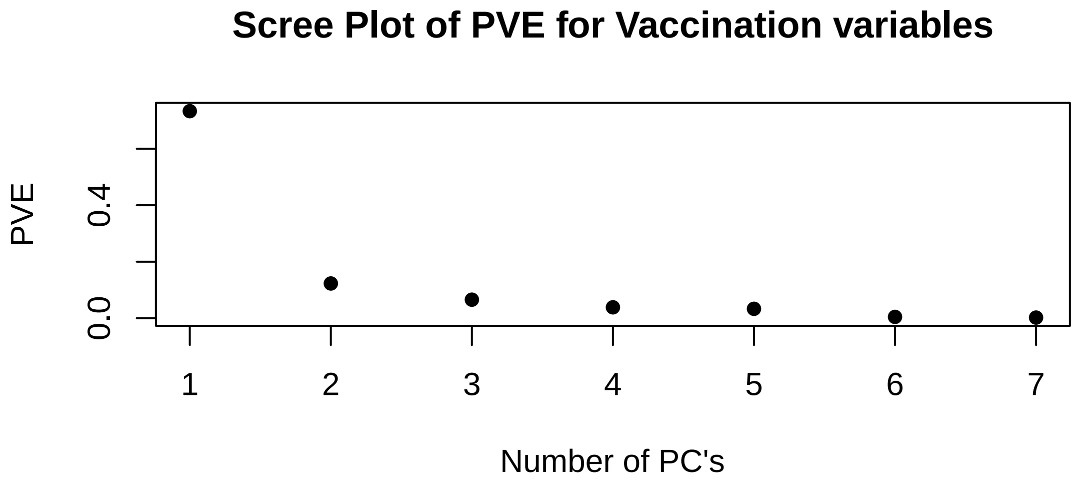
```
## 'mutate_if()' ignored the following grouping variables:
## Column 'country'
```

Now let's merge all these vaccination dfs into one dataframe so we can create a VSI to predict!

After combining the vaccination variables, we'll pair down to keep only key variables that will be meaningful in denoting a country's vaccination success. Variables were selected to be meaningful and non-redundant, and to have high completion rates.

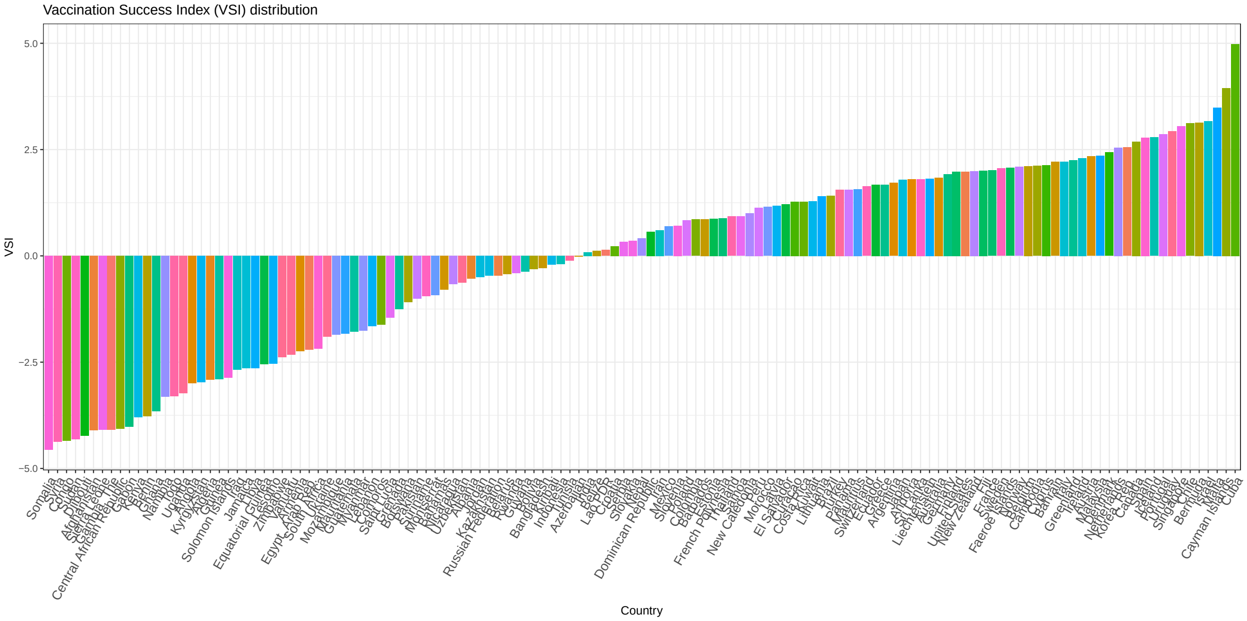
Next we'll perform PCA to create a score to summarize these important vaccination metrics into a continuous rating, the VSI.

```
##                                              PC1     PC2    PC3    PC4     PC5
## days.to.start_rel                          0.260 -0.7951  0.400  0.362 -0.0654
## days.to.10pct_fromstart                    0.353  0.3810  0.549 -0.128 -0.6362
## daily.vac.per.mil_6moavg                  -0.359 -0.4294 -0.309 -0.425 -0.6120
## people_vaccinated_per_hundred_last        -0.421  0.0934  0.209  0.420 -0.0364
## daily.vac.per.mil_max                     -0.368 -0.1051  0.613 -0.579  0.3764
## total_vaccinations_per_hundred_last       -0.428  0.0799  0.108  0.223 -0.2588
## people_fully_vaccinated_per_hundred_last -0.428  0.1099  0.113  0.332 -0.0798
##                                               PC6     PC7
## days.to.start_rel                          0.0523  0.0482
## days.to.10pct_fromstart                   -0.0869  0.0271
## daily.vac.per.mil_6moavg                  -0.1900 -0.0220
## people_vaccinated_per_hundred_last        -0.5098 -0.5765
## daily.vac.per.mil_max                      0.0172  0.0137
## total_vaccinations_per_hundred_last        0.8082 -0.1715
## people_fully_vaccinated_per_hundred_last -0.2008  0.7965
```

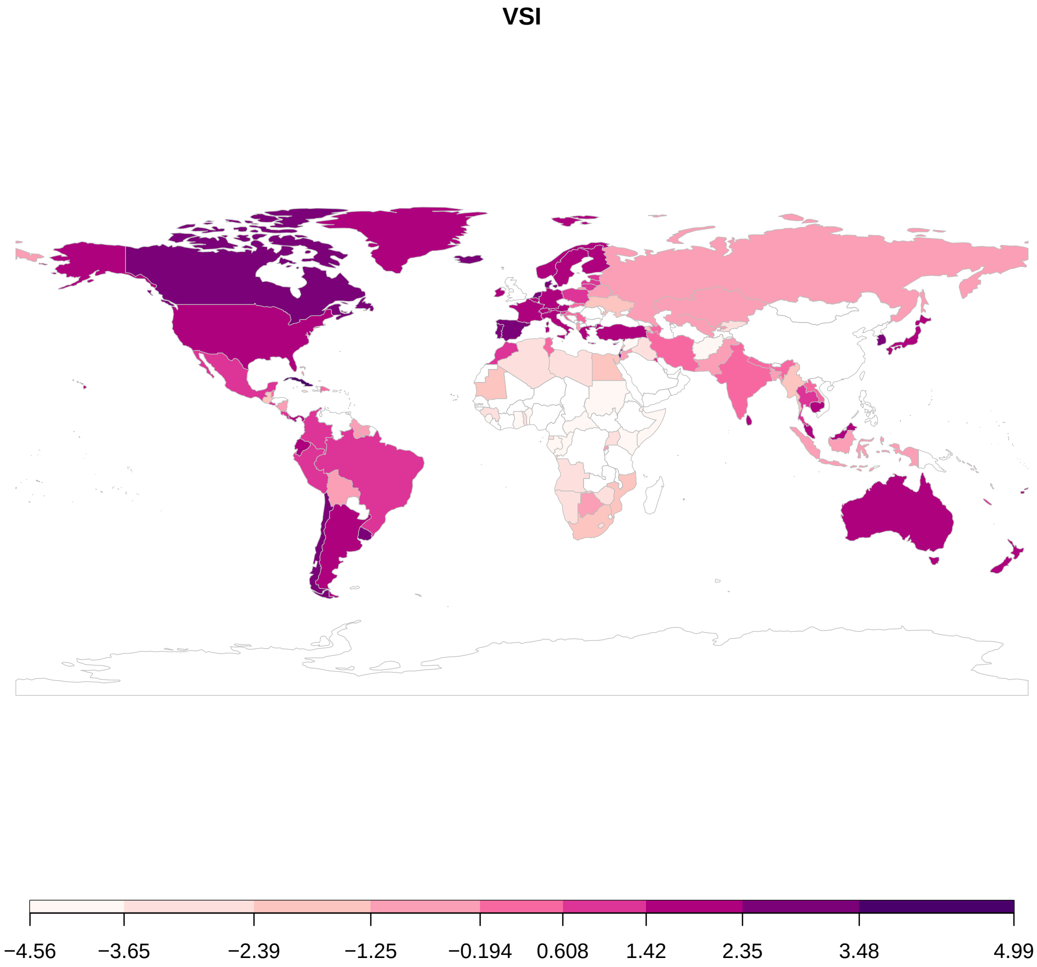


PC1 explains 73% of the variance and is therefore an excellent index indicating how well a country did at vaccinating their population. Based on the relative loadings of each variable, we'll take the inverse of PC1 so that a higher metric indicates a more successful vaccination campaign. We are calling it the "Vaccination Success Index", i.e. `VSI`.

Let's see how the VSI loadings are distributed across countries:



```
## 133 codes from your data successfully matched countries in the map
## 0 codes from your data failed to match with a country code in the map
## 110 codes from the map weren't represented in your data
```



Based on the distribution plot, we can see that Somalia appears to be the country with the lowest VSI, and Cuba appears to have the highest VSI. The world map furthermore indicates that generally, North and South America, Europe, Australia and Japan are some regions with higher VSI scores, whereas regions including Africa, Russia and some Asian countries seem to have lower VSI scores. Note that countries without any color are not included in this dataset as they did not report sufficient vaccination data required for the computation of VSI; notable absences include China and the UK.

3 Compiling the Final Dataset

We'll write our final, cleaned dataset (`final.df`) by joining our WSD predictors, prevaccination COVID case and death predictors, and the VSI.

Upon merging, we are left with data on 99 unique countries.

4 Modeling

In this section, we are using socioeconomic factors and pre-vaccination COVID cases and deaths to predict Vaccination Success Index (VSI). Note that VSI corresponds the inverse signed PC1 scores and (based on the relative loadings of each raw vaccination value) indicates how well a country did at getting people vaccinated, i.e. a positive score indicates a good vaccination performance, while a negative score indicates a poor vaccination performance.

4.1 Data splitting

We'll start by subsetting our dataset of 99 countries into `data.train`(85%) and `data.val` (15%). These splits were chosen because our dataset has a relatively small number of predictors and our most robust modeling methods have means of estimating testing error (Random Forest's OOB error, Regsubsets Cp, Neural Networks' internal data.val split). `data.val` will be reserved to validate our final ensemble model.

Let's now see how various socioeconomic, sustainability and political factors may be associated with VSI. To develop the most accurate prediction, we will compare a range of methods going from supervised to unsupervised:

- Linear models
- LASSO model selection
- Regsubsets
- Random forests
- Neural networks

5 Linear models

In the first approach, we are creating several linear models by selecting variables based on domain knowledge. Each model is focused on specific themes that may be of importance for predicting a country's `VSI` score. The following themes will be covered: + Economic factors + "COVID burden" factors, i.e. how heavily a country suffered in terms of number of cases and deaths prior to vaccination start + Political and country development factors + Full model: we combine the most relevant factors into a comprehensive model

5.0.1 Thematic linear models

All linear models are compiled using backward selection. Summaries of the intermediate models are ommited for brevity.

5.0.1.1 Economic factors Using domain knowledge, we explore the relationship between measures capturing the quality of a nation's economy over the last reported 17 years and vaccination success.

Selected variables:

- `ExportGoodsServices.GDP_avg` - part of the GDP resulting from exporting goods, averaged over 17 years
- `FinalConsumptionExpenditure.GDP_avg` - apart of the GDP resulting from consumption/expenses of the country, averaged over 17 years
- `GDP.Current_avg` - current overall GDP, averaged over 17 years
- `GDP.PerCapita.Current_avg` - current GDP per capita, averaged over 17 years
- `ConsumerPriceInflation_avg` - how much a country was affected by price inflation, averaged over 17 years

We further refine our model by backward selection.

Economic Model

Dependent variable:

VSI

(1)

(2)

(3)

(4)

ExportGoodsServices.GDP_avg

0.004

(0.009)

FinalConsumptionExpenditure.GDP_avg

-0.004

-0.007

(0.015)

(0.014)

GDP.Current_avg

0.000

0.000

0.000

(0.000)

(0.000)

(0.000)

GDP.PerCapita.Current_avg

0.0001***

0.0001***

0.0001***

0.0001***

(0.00001)

(0.00001)

(0.00001)

(0.00001)

ConsumerPriceInflation_avg

-0.086**

-0.088**

-0.083**

-0.083**

(0.042)

(0.041)

(0.040)

(0.040)

Constant

-0.239

0.203

-0.400

-0.400

(1.540)

(1.260)

(0.354)

(0.352)

Observations

84

84

84

84

R2

0.402

0.400

0.398

0.397

Adjusted R2

0.363

0.369

0.375

0.382

Residual Std. Error

1.670 (df = 78)

1.660 (df = 79)

1.650 (df = 80)

1.640 (df = 81)

F Statistic

10.500*** (df = 5; 78)

13.100*** (df = 4; 79)

17.600*** (df = 3; 80)

26.700*** (df = 2; 81)

Note:

p<0.1; ***p<0.05;*** p<0.01

GDP per capita and consumer price inflation have a significant effect on Vaccination Success Index at the 0.05 level. More specifically, the model indicates that an increase in the per-capita GDP of a nation averaged over the last 17 years corresponds to an increase in the success of that country's vaccination campaign. Conversely, an increase in consumer price inflation averaged over the last 17 years corresponds to a decrease in a country's vaccination success.

5.0.1.2 "COVID burden" factors Let's now build a model focused on COVID related factors. Note that all variables correspond to COVID cases and deaths reported prior to a country's given vaccination start date.

Selected variables:

- `monthly.new.cases.per.100_max` - max monthly case rate reached prior to vaccination start
- `monthly.new.deaths.per.100_max` - max monthly death rate reached prior to vaccination start
- `cumulative_total_deaths_per_hundred_prevac` - cumulative number of deaths on last report prior to vaccinations beginning in each country, normalized by population

Again, we refine our model by backward selection.

Disease Burden Model

Dependent variable:

VSI

(1)

(2)

(3)

monthly.new.cases.per.100_max

1.360***

1.110***

1.260***

(0.484)

(0.333)

(0.287)

monthly.new.deaths.per.100_max

-25.600

(36.300)

cumulative_total_deaths_per_hundred_prevac

8.770

4.160

(8.060)

(4.690)

Constant

-0.942***

-0.923***

-0.856***

(0.309)

(0.307)

(0.297)

Observations

84

84

84

R2

0.204

0.200

0.192

Adjusted R2

0.175

0.180

0.182

Residual Std. Error

1.900 (df = 80)

1.890 (df = 81)

1.890 (df = 82)

F Statistic

6.850*** (df = 3; 80)

10.100*** (df = 2; 81)

19.500*** (df = 1; 82)

Note:

p<0.1; ***p<0.05;*** p<0.01

The peak number of monthly new cases per hundred prior to vaccination start has a significant effect on `VSI`. Therefore, vaccination success seems influenced by a country's peak number of cases prior to starting its vaccination campaign, where a higher number of cases leads to a better vaccination score. This could be explained by the fact that people may have a stronger vaccination incentive when COVID cases are more frequent around them.

5.0.1.3 Political and country development factors Let's build a model focused on political factors as well as indices reflecting the development of a country.

Selected variables:

- `WorldRegion` - part of the world (more specific than continents)

- `Electricity.Access_avg` - reported electricity access, may influence the infrastructure available for vaccination, averaged over 17 years
- `GDP.PerCapita.Current_avg` - GDP per capita, averaged over 17 years
- `UrbanPopulation.Prop_avg` - how urban is the population, averaged over 17 years
- `LifeExpectancy_avg` - life expectancy, averaged over 17 years
- `CompulsoryEducationDurationYears_latest` - latest level of mandated education
- `WomenInBusinessLawIndex_avg` - proportion of women working in law and business fields, averaged over 17 years
- `WorldBankIncomeClass_avg` - categorical classifying as low, lower-middle, upper-middle, and high-income countries, averaged over 17 years
- `RegimeType_avg` - categorical classifying as Closed Autocracy, Electoral Autocracy, Electoral Democracy, Liberal Democracy
- `IndividualsUsingInternet_latest` - latest internet usage

```
## Anova Table (Type II tests)
##
## Response: VSI
##                                         Sum Sq Df F value  Pr(>F)
## WorldRegion                               25.8  6    5.68 9.7e-05 ***
## Electricity.Access_avg                     1.8  1    2.36  0.1299
## GDP.PerCapita.Current_avg                  0.0  1    0.04  0.8394
## UrbanPopulation.Prop_avg                   0.1  1    0.18  0.6757
## LifeExpenctancy_avg                       14.9  1   19.72 3.8e-05 ***
## CompulsoryEducationDurationYears_latest    2.9  1    3.80  0.0559 .
## WomenInBusinessLawIndex_avg                7.7  1   10.12  0.0023 **
## WorldBankIncomeClass_avg                   0.6  3    0.26  0.8557
## RegimeType_avg                             8.1  3    3.56  0.0193 *
## IndividualsUsingInternet_latest            0.8  1    1.02  0.3167
## Residuals                                 46.2 61
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We refine our model by backward selection.

```
## Anova Table (Type II tests)
##
## Response: VSI
##                             Sum Sq Df F value  Pr(>F)
## WorldRegion                   23.5  6    5.07 0.00023 ***
## LifeExpenctancy_avg           26.8  1   34.84 1.2e-07 ***
## WomenInBusinessLawIndex_avg   11.8  1   15.34 0.00021 ***
## RegimeType_avg                12.8  3    5.53 0.00185 **
## Residuals                     53.2 69
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

World region and 17-year averages of life expectancy, women in business and law professions, and regime type have a significant effect on VSI. More specifically, it appears that compared to the reference world region (Central and Southern Asia), all other regions are associated with lower VSI except Eastern and South-Eastern Asia, who have a higher VSI. Higher life expectancy and more women employed in business or law are associated with higher VSI. Finally, compared to the reference regime type (closed autocracy, which includes countries such as Angola, Morocco, Jordan, etc.), other more liberal regime types are associated with lower VSI scores. This likely reflects the imposition of vaccination mandates - which often face opposition in countries with robust legal precedents for health autonomy - in more authoritarian countries.

5.0.1.4 Full model We will now use all relevant variables found in the economic, COVID burden and political/development models and feeding them into one overall model, with the goal to capture a more diverse set of factors that may impact VSI the strongest.

Selected variables:

- `GDP.PerCapita.Current_avg` - current GDP per capita, averaged over 17 years
- `cumulative_total_cases_per_hundred_prevac` - last reported total number of cases prior to vaccination start
- `WorldRegion` - part of the world (more specific than continents)
- `LifeExpenctancy_avg` - life expectancy, averaged over 17 years
- `WomenInBusinessLawIndex_avg` - proportion of women working in law and business fields, averaged over 17 years
- `RegimeType_avg` - categorical classifying as Closed Autocracy, Electoral Autocracy, Electoral Democracy, Liberal Democracy

```
## Anova Table (Type II tests)
##
## Response: VSI
##                               Sum Sq Df F value  Pr(>F)
## GDP.PerCapita.Current_avg        0.9  1    1.15 0.28734
## ConsumerPriceInflation_avg       1.1  1    1.37 0.24601
## monthly.new.cases.per.100_max    0.1  1    0.07 0.79022
## WorldRegion                     22.1  6    4.79 0.00042 ***
## LifeExpenctancy_avg             16.3  1   21.22 1.9e-05 ***
## WomenInBusinessLawIndex_avg     11.5  1   14.92 0.00026 ***
## RegimeType_avg                  10.9  3    4.74 0.00470 **
## Residuals                       50.7 66
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Now we refine our model by backward selection.

Full Model

Dependent variable:

VSI

WorldRegionEastern and South-Eastern Asia

0.131

(0.612)

WorldRegionEurope and Northern America

-1.970***

(0.509)

WorldRegionLatin America and Caribbean

-1.060**

(0.470)

WorldRegionNorthern Africa and Western Asia

-1.500***

(0.487)

WorldRegionOceania

-2.370**

(1.020)

WorldRegionSub-Saharan Africa

-1.820***

(0.564)

LifeExpenctancy_avg

0.166***

(0.028)

WomenInBusinessLawIndex_avg

0.038***

(0.010)

RegimeType_avgElectoral Autocracy

-1.330***

(0.463)

RegimeType_avgElectoral Democracy

-1.410***

(0.527)

RegimeType_avgLiberal Democracy

-0.482

(0.605)

Constant

-12.200***

(2.010)

Observations

81

R2

0.850

Adjusted R2

0.826

Residual Std. Error

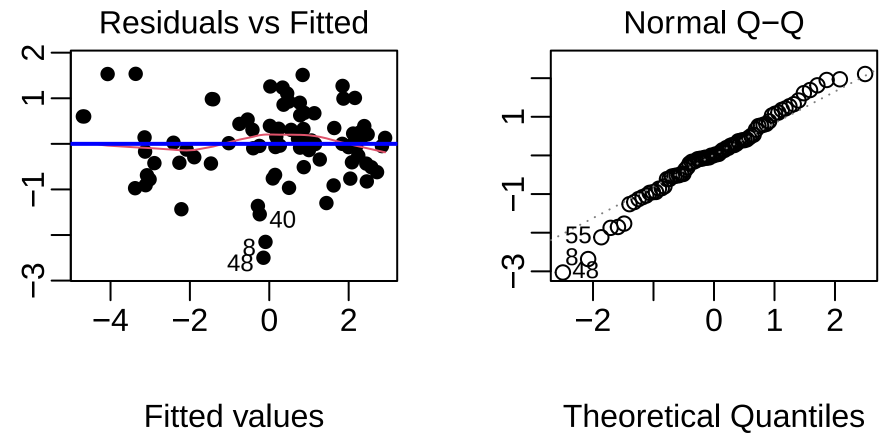
0.878 (df = 69)

F Statistic

35.400*** (df = 11; 69)

Note:

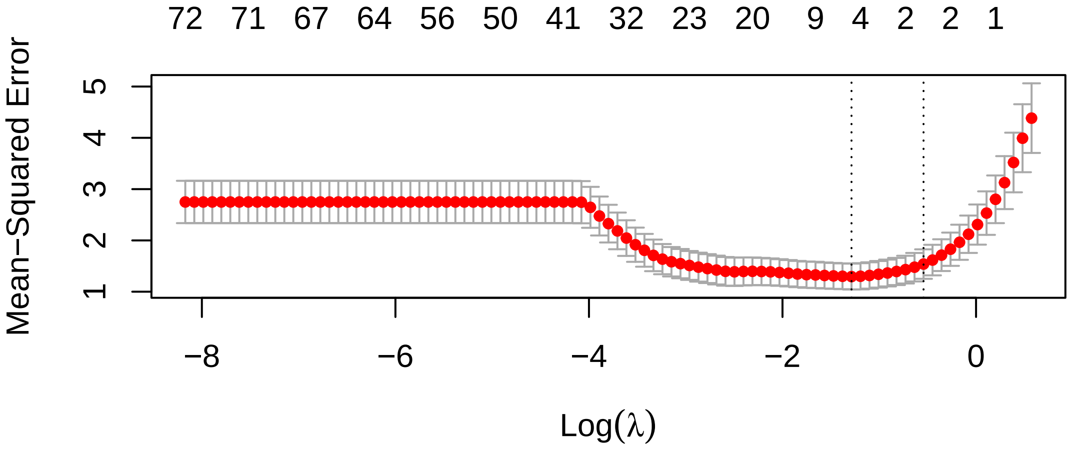
p<0.1; ***p<0.05;*** p<0.01



When using all relevant variables found in prior models, it turns out that world region, life expectancy, the number of women employed in business and law, and regime type have a significant effect on `VSI`. Diagnostic plots show that the model sufficiently fulfills linear model assumptions of normality and homoscedasticity.

5.1 LASSO model selection

Let's now try a slightly more data-driven approach, i.e. using LASSO to select the most relevant variables.



[1]

0.581

```
## Note: Using an external vector in selections is ambiguous.
## i Use ‘all_of(var.1se)‘ instead of ‘var.1se‘ to silence this message.
## i See <https://tidyselect.r-lib.org/reference/faq-external-vector.html>.
## This message is displayed once per session.
```

LASSO Model

Dependent variable:

VSI

IndividualsUsingInternet_avg

0.019**

(0.008)

LifeExpenctancy_avg

0.171***

(0.023)

Constant

-13.000***

(1.470)

Observations

75

R2

0.750

Adjusted R2

0.743

Residual Std. Error

1.060 (df = 72)

F Statistic

108.000*** (df = 2; 72)

Note:

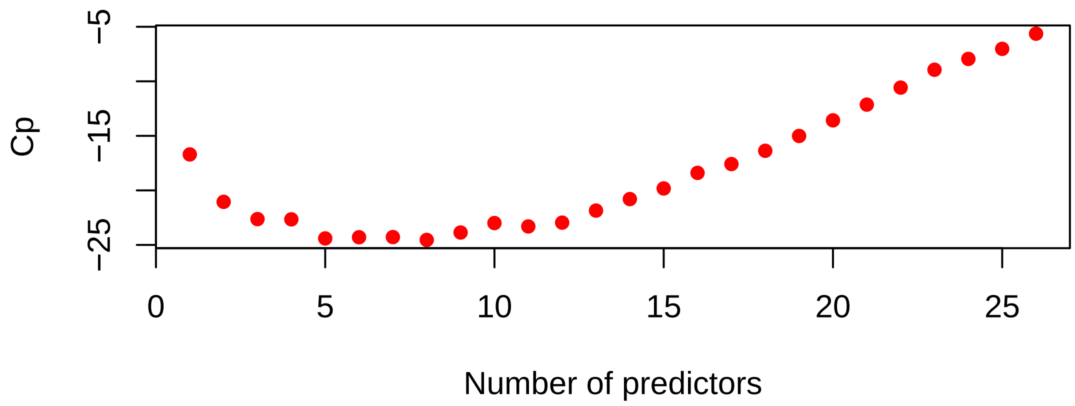
p<0.1; ***p<0.05;*** p<0.01

LASSO results using the `coef.1se` on 75 complete countries suggest again that life expectancy is an important predictor of `VSI`. In addition, the LASSO approach selected the 17 year average number of individuals using internet. When feeding these into a linear model, both have a significant effect on `VSI`.

5.2 Regsubsets

Next we'll run regsubsets to optimize a model by minimizing Cp. This does return a warning that regsubsets still finds our data to be colinear, so we re-run correlation reduction with 95% cutoff on `data.train` to whether we can build a more optimal model. While this alternate model does better in modeling VSI (see `Comparing Regsubsets Models` in the Appendix), we felt this improvement in performance was not worth the lack of comparability with LASSO, RF and other models built on the full `data.train` set of predictors and thus proceed with the following model.

```
## Reordering variables and trying again:
```



```
## Anova Table (Type II tests)
##
## Response: VSI
##                                             Sum Sq Df F value    Pr(>F)    
## WorldRegion                                   24.7  6    5.66   9.5e-05 ***
## WorldBankIncomeClass_latest                    7.4  3    3.38    0.0236 *  
## WorldBankIncomeClass_avg                       3.1  3    1.42    0.2451    
## Govt.FinalConsumptionExpenditure.GDP_latest    5.5  1    7.53    0.0079 ** 
## WomenInBusinessLawIndex_avg                    8.3  1   11.35    0.0013 ** 
## LifeExpectancy_avg                            23.0  1   31.61   4.6e-07 ***
## cumulative_total_cases_prevac                  1.6  1    2.19    0.1442    
## weekly_new_deaths_per_hundred_prevac           0.7  1    0.95    0.3326    
## Residuals                                     45.9 63                      
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

Now let’s run some backward selection on this model:

```
## Anova Table (Type II tests)
##
## Response: VSI
##                                             Sum Sq Df F value    Pr(>F)    
## WorldRegion                                   30.0  6    6.76   1.2e-05 ***
## WorldBankIncomeClass_latest                    8.4  3    3.76   0.01463 *  
## Govt.FinalConsumptionExpenditure.GDP_latest    4.6  1    6.26   0.01476 *  
## WomenInBusinessLawIndex_avg                    9.3  1   12.57   0.00071 ***
## LifeExpectancy_avg                            20.8  1   28.04   1.4e-06 ***
## Residuals                                     50.4 68                      
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

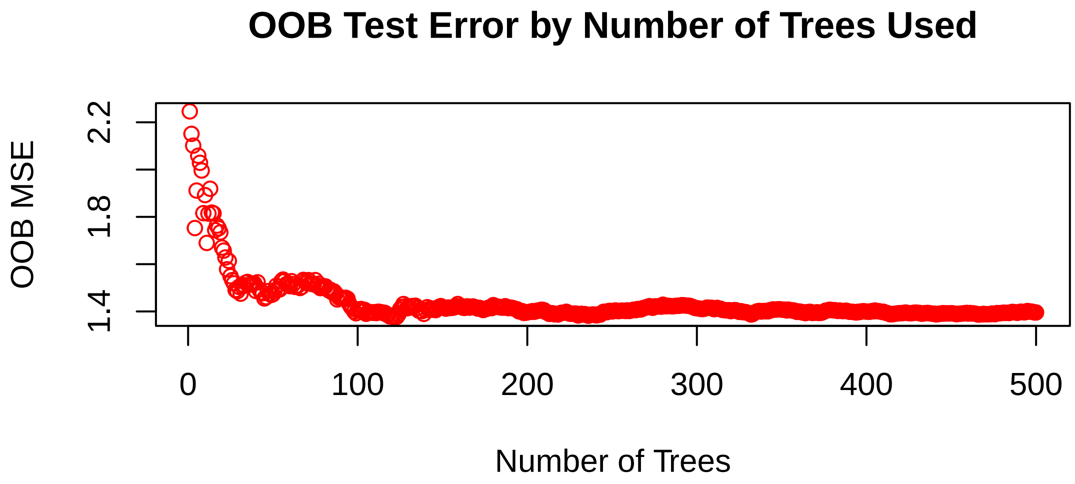
Reducing to only significantly predictive variables, we find they are fairly similar to those identified by our full linear model. This is reassuring to have our data-driven model identify similar predictors to our hypothesis-driven models.

6 Random Forest

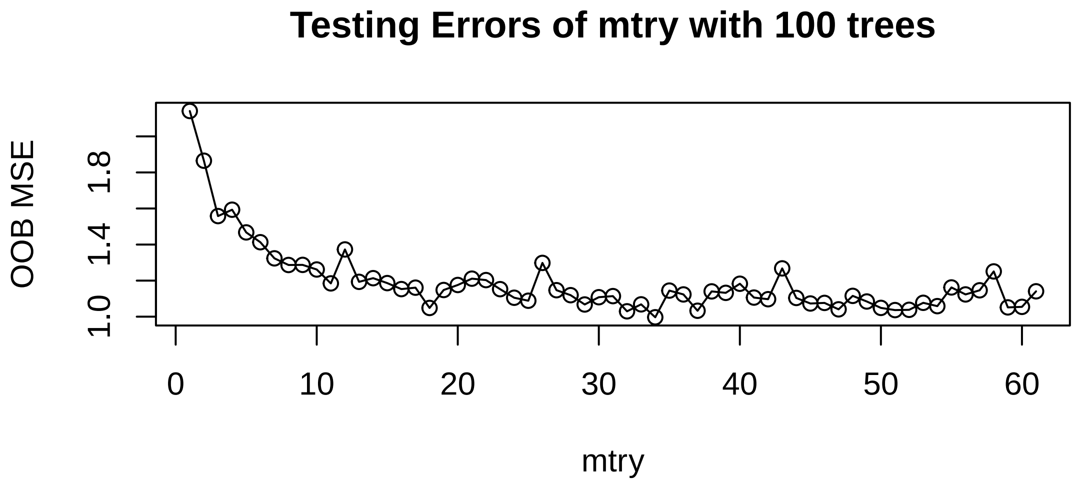
Now, we will try to use tree based methods, specifically Random Forest, to see if we can predict a country's vaccination success based on social and economic measures prior to the pandemic, as well as COVID case data prior to the start of vaccination campaigns.

While an advantage of using a single tree is that we can see which variables are diving our prediction, Random Forest may provide better results at the cost of interpretability. Accordingly, now we will tune and train a Random Forest model. For those interested, an example of a single regression tree can be found in the Appendix under `Regression Tree`.

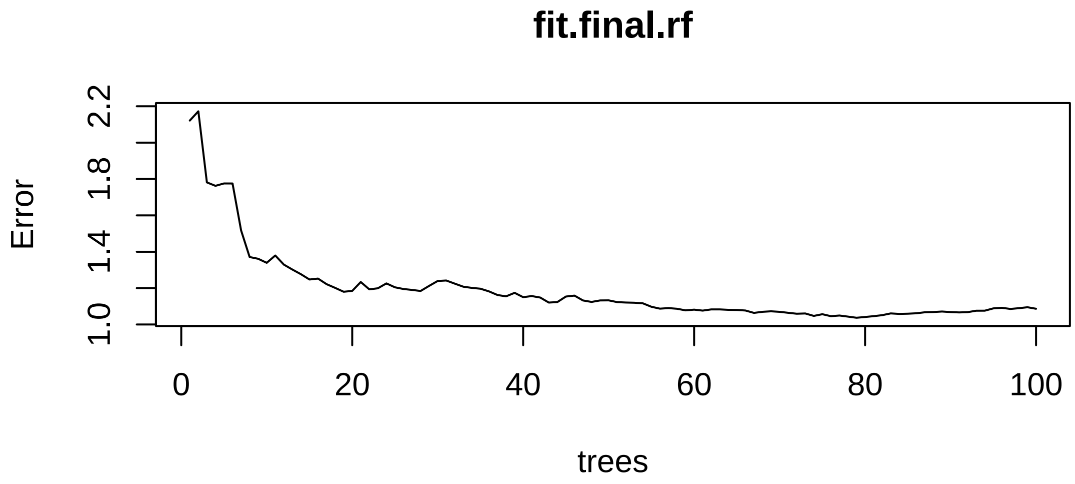
Let us tune the parameters mtry (number of variables to sample) and ntree (number of trees to use). First, we will tune ntree.



Based on out of bag error, it seems that 100 trees is sufficient. Now, let's tune mtry.



An mtry of around 30 seems to provide good results, so we will use this as our parameter. Now, we train the final Random Forest model based on our tuned parameters.



The final Random Forest model, now that it's been tuned, is performing with an Out-Of-Bag Testing error of 1.087. While our final testing error is still somewhat high, Random Forest can achieve a somewhat accurate estimate of a nation's COVID vaccination success, given social and economic factors prior to the pandemic, and COVID case data prior to the start of vaccination campaigns.

7 Neural Networks

Next, we'll run neural nets. Though the data has high completion to begin with (only 1.53% missing in `df.pred`), neural nets cannot take NA inputs. Because the benefits of NN arise in large datasets and our data is relatively small to begin with, we'll impute missing values prior to building, training and fitting our network.

```
##    NRMSE      PFC
## 1.89e-10 1.36e-01
```

Based on the very low normalized root mean squared error (NRMSE $= 1.895 \times 10^{-10}$) of numeric variables and proportion of falsely classified entries (PFC $= 0.136$) estimated for categorical variables, we can feel confident in using these imputed values to train our neural network.

We'll now proceed to build and compile our model:

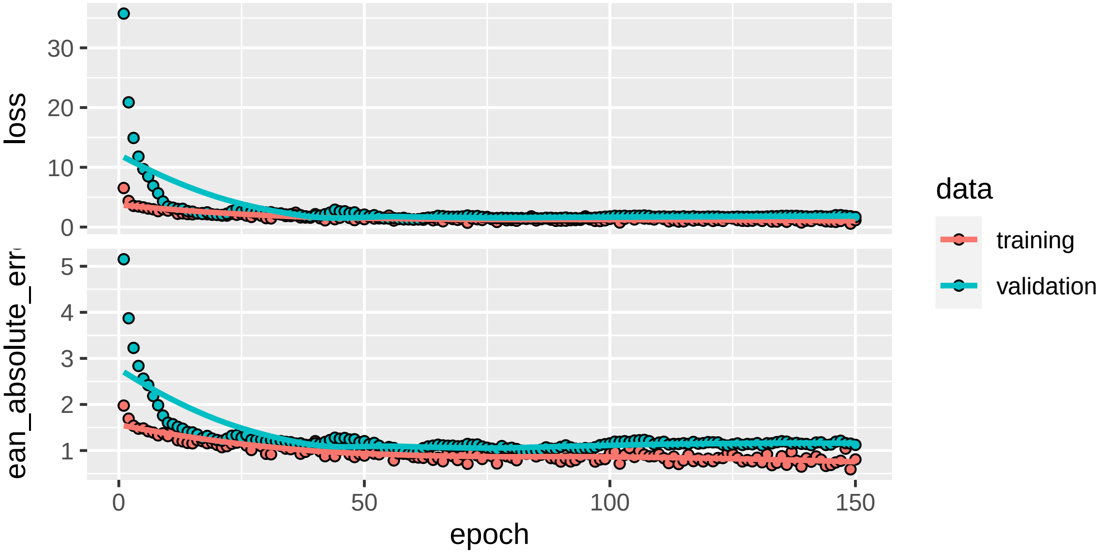
```
## Loaded Tensorflow version 2.8.0
```

```
## Model: "sequential"
## ________________________________________________________________________________
##  Layer (type)                       Output Shape                     Param #
## ================================================================================
##  batch_normalization (BatchNormaliz  (None, 73)                      292
##  ation)
##
##  dense_2 (Dense)                    (None, 32)                       2368
##
##  dropout_1 (Dropout)                (None, 32)                       0
```

```
##
##  dense_1 (Dense)                    (None, 12)                      396
##
##  dropout (Dropout)                  (None, 12)                      0
##
##  dense (Dense)                      (None, 1)                       13
##
## ================================================================================
## Total params: 3,069
## Trainable params: 2,923
## Non-trainable params: 146
## ________________________________________________________________________________
```

Finally we fit, outputting MSE for both the training data and an internal validation sample (15% of `data.train` that's been cleaned, imputed and normalized for NN input).

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



The lowest MSE calculated on the internal validation sample is 1.238 at the 60. Now that we've defined our model, we can refit it without the internal validation split and using 60 epochs.

```
## Model: "sequential_1"
## ________________________________________________________________________________
##  Layer (type)                       Output Shape                    Param #
## ================================================================================
##  batch_normalization_1 (BatchNormal  (None, 73)                     292
##  ization)
##
##  dense_5 (Dense)                    (None, 32)                      2368
##
##  dropout_3 (Dropout)                (None, 32)                      0
##
##  dense_4 (Dense)                    (None, 12)                      396
##
```

```
##  dropout_2 (Dropout)                (None, 12)                      0
##
##  dense_3 (Dense)                    (None, 1)                       13
##
## ================================================================================
## Total params: 3,069
## Trainable params: 2,923
## Non-trainable params: 146
## ________________________________________________________________________________
```

The MSE of this model is now 1.556, slightly larger than the training loss on the original (MSE at epoch 65 = 1.259).

8 Ensemble

Now we'll attempt to create the best model by bagging those created above, averaging the predicted VSI values from each model. Each country in the training data's VSI is estimated by each model and scores are compiled in the dataframe below. An ensemble score is then estimated by averaging across the models.

```
##       VSI full.lm  lasso     rf regsub     V1 en.pred
## 9   1.839   1.839  2.466  2.024  1.839  2.099  2.0533
## 74  1.136   1.066  0.269     NA  0.583  1.192  0.7775
## 76  2.861   1.866  1.507  2.311  1.620  1.423  1.7455
## 55  0.141   1.432 -2.027     NA  0.209  0.352 -0.0086
## 72 -1.006  -1.015 -1.676 -0.893 -1.162 -1.000 -1.1492
## 54  1.285   0.018  0.736  1.450  0.179  0.351  0.5467
```

```
## [1] 0.458
```

Hurrah! Our training MSE from our ensemble method, 0.458, is actually pretty good! We can reduce this even further by weighting the contributions of each model based on their individual training MSEs, displayed below:

```
##           model train.mse
## 1       full.lm     0.626
## 2         LASSO     1.102
## 3  RandomForest     0.171
## 4        RegSub     0.566
## 5            NN     0.684
## 6 full.ensemble     0.458
```

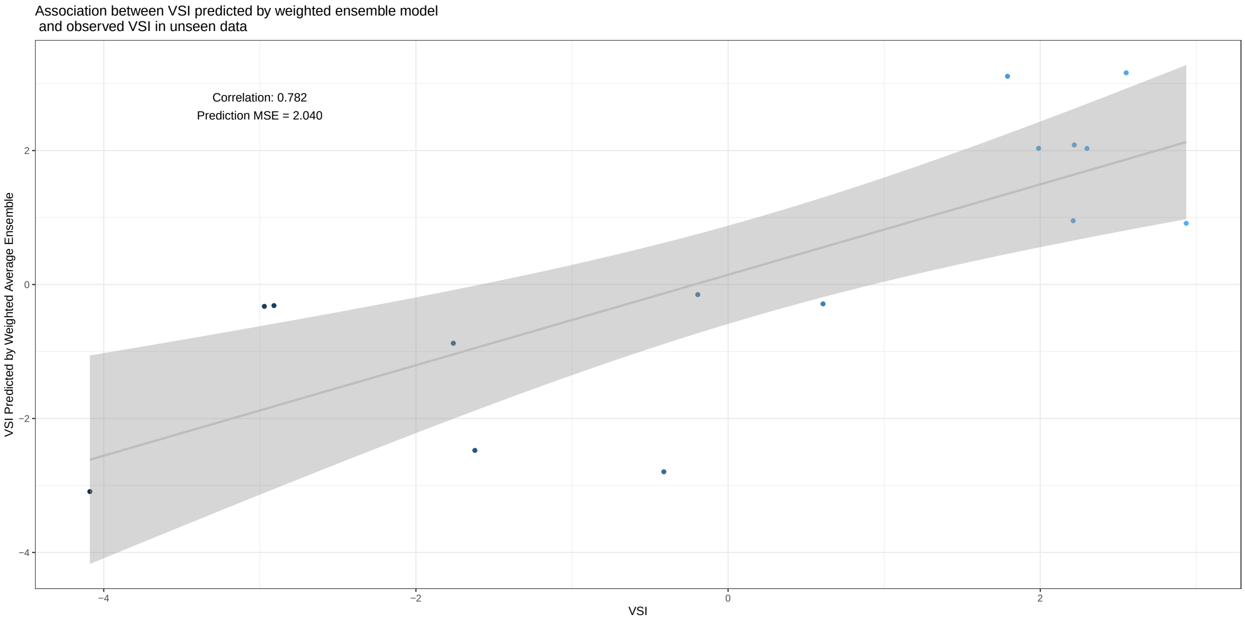
As we can see, the RandomForest MSE is lowest while LASSO's MSE is highest. By weighting the contributions of each model to our final prediction accordingly, we obtain a training MSE of 0.324.

We'll conclude by running our weighted ensemble method on the validation data we split out at the beginning to see how well this model predicts VSI scores based on data from unseen countries.

```
##       VSI full.lm   lasso     rf regsub     V1 weighted.pred
## 1   0.608  -0.529  0.0579 -0.445  0.208 -0.376        -0.289
## 2  -2.971   0.751 -0.8299 -1.080  1.016 -0.567        -0.326
## 6  -2.909  -0.545  0.4464 -0.550  0.187 -0.385        -0.314
## 12  2.211   0.402  1.2984  1.310  0.235  1.150         0.952
## 36 -4.089  -3.051 -2.6254 -3.328 -3.190 -2.528        -3.091
## 43 -0.194   0.126 -0.9307 -0.473  0.154  0.549        -0.151
## [1] 2.04
```

Unfortunately the MSE from applying our weighted ensemble method to unseen validation data (2.04) is much larger than that of our training data (0.324). This suggests that the ensemble model and/or one of the underlying models bagged to create it suffered from overfitting.

```
## ‘geom_smooth()‘ using formula ’y ~ x’
```



9 Conclusions

In summary, we were able to create a variety of models predicting our VSI score. The MSE of each model's predictions on the training data is shown here:

```
##                model         train.mse
## 1            full.lm 0.626470711796047
## 2              LASSO  1.10247014138825
## 3       RandomForest 0.170612648153805
## 4             RegSub 0.566104279987661
## 5                 NN 0.683601748727079
## 6      full.ensemble 0.458311711054015
## 7 weighted.ensemble 0.324177892641474
```

Given that our dataset was relatively small, a training MSE of around 0.5 in predicting VSI (which ranges almost 10 pts from -4.353 to 3.165) may be considered very good performance; this was notably achieved by hypothesis-driven linear modeling, Random Forests, Regsubsets, and our weighted and unweighted ensemble models. While our final ensemble model did not perform as well on our validation data - suggesting that it and/or one of the underlying models were overfitted - we hope the methodology explored here can be improved with application to similar, more complete datasets in the future.

Beyond prediction, the linear models provide a number of insights into country-level variables that are significantly related to vaccination success. Each factor, as well as policy recommendations to address it, are noted above in the Executive Summary. However, it is worth reiterating some key findings here.

Firstly, both LASSO and Regsubsets identified similar sets of predictors to those in our hypothesis-driven thematic linear models. This is reassuring to have our data-driven model identify similar predictors to our hypothesis-driven models and validates our choices in predictors. Our hypothesis-driven linear model also had very good MSE in training relative to even purely data-driven models.

Similarly, every one of our interpretable models (full linear model, LASSO, regsubsets and decision tree) identified averaged life expectancy as a significant predictor of VSI. Life expectancy likely captures a wide range of latent predictors including healthcare access, infrastructure, public health education, and potentially attitudes and social supports that may be beneficial both in physical health outcomes and encouraging vaccination compliance.

The relationship between life expectancy and VSI is plotted here:

```
## ‘geom_smooth()‘ using formula ’y ~ x’
```

Furthermore, none of the COVID-19 death or case metrics prior to vaccination onset were found to be predictive when other, non-COVID variables were included. This is surprising, considering that one would expect COVID cases and mortality to have a strong affect on a country’s ability, national investment, and individual willingness to undertake large-scale vaccination campaigns. Our analysis is significant in the fact that it indicates non-disease specific factors are actually more relevant to vaccination against COVID-19 than the burden of the disease itself.

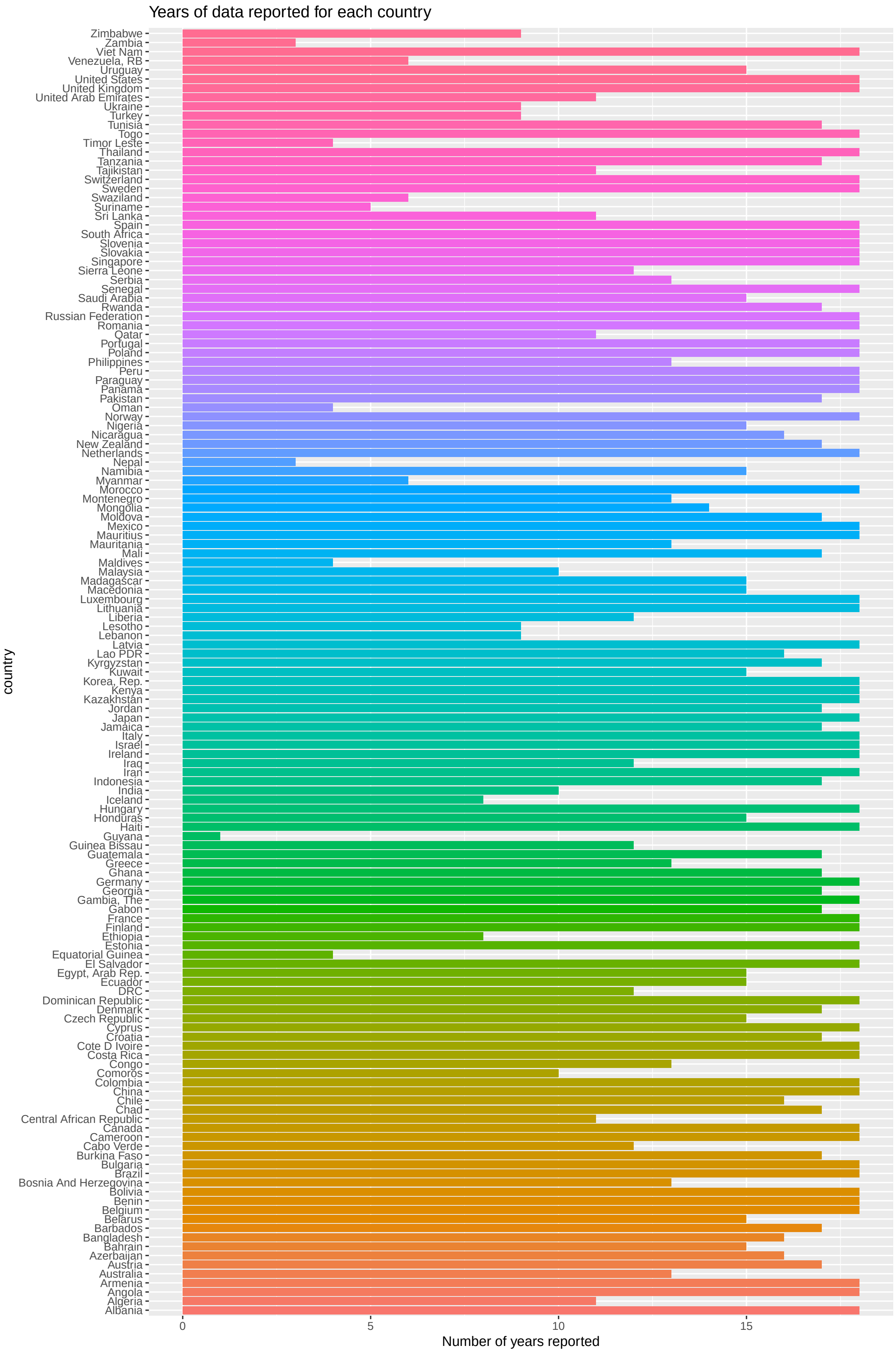
We hope that these analyses may be useful in shaping future discussions about the varying strengths of each nation’s COVID-19 vaccination campaigns and in global response to health crises more broadly.

10 Appendix

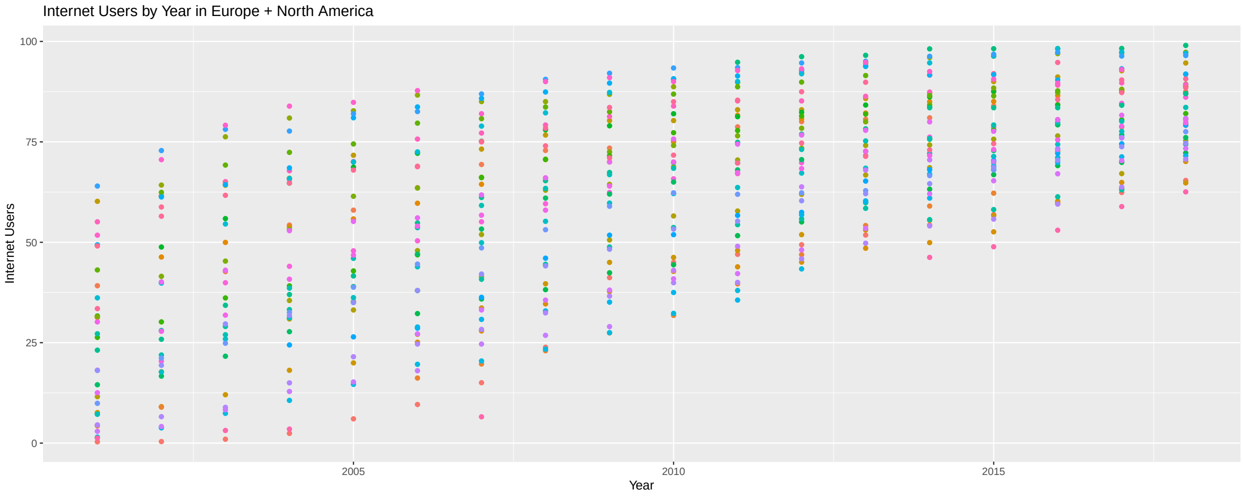
10.1 Additional EDA

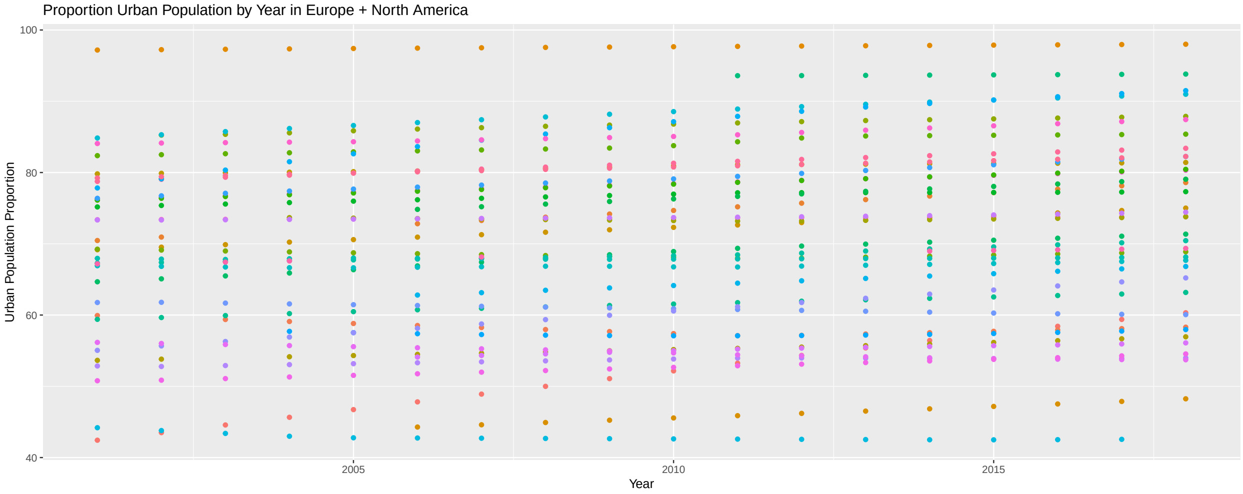
10.1.1 WSD

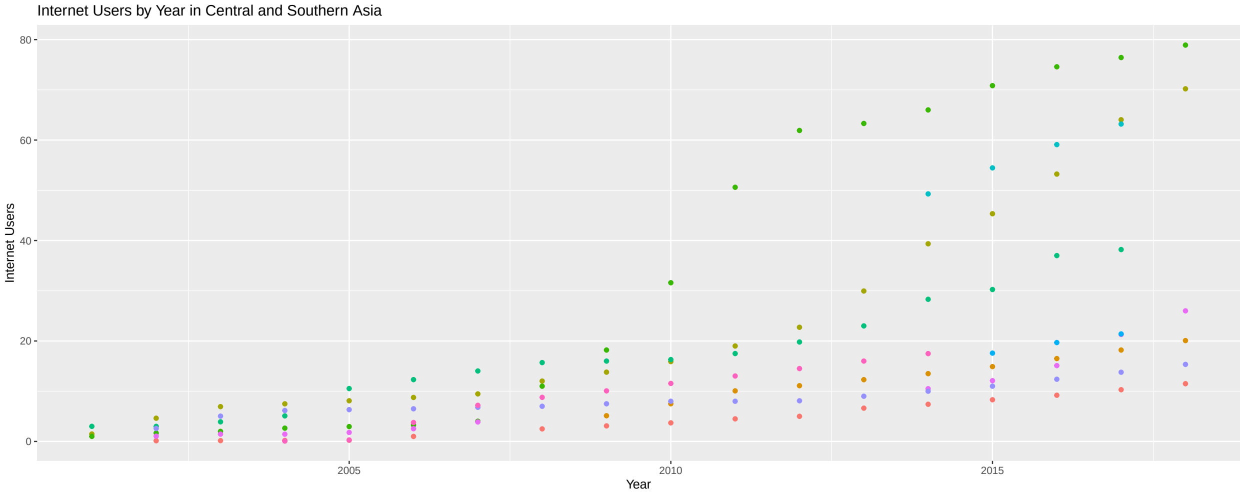
Detailed graph of the number of years with data reported for each country:

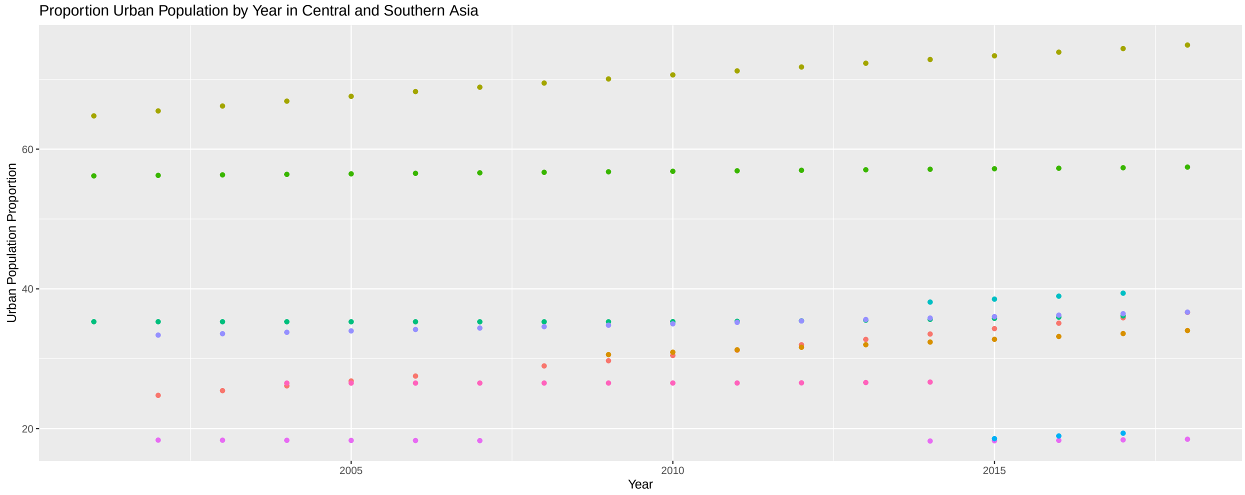

Years of data reported for each country
Zimbabwe
Zambia
Viet Nam
Venezuela, RB
Uruguay
United States
United Kingdom
United Arab Emirates
Ukraine
Turkey
Tunisia
Togo
Timor Leste
Thailand
Tanzania
Tajikistan
Switzerland
Sweden
Swaziland
Suriname
Sri Lanka
Spain
South Africa
Slovenia
Slovakia
Singapore
Sierra Leone
Serbia
Senegal
Saudi Arabia
Rwanda
Russian Federation
Romania
Qatar
Portugal
Poland
Philippines
Peru
Paraguay
Panama
Pakistan
Oman
Norway
Nigeria
Nicaragua
New Zealand
Netherlands
Nepal
Namibia
Myanmar
Morocco
Montenegro
Mongolia
Moldova
Mexico
Mauritius
Mauritania
Mali
Maldives
Malaysia
Madagascar
Macedonia
Luxembourg
Lithuania
Liberia
Lesotho
Lebanon
Latvia
Lao PDR
Kyrgyzstan
Kuwait
Korea, Rep.
Kenya
Kazakhstan
Jordan
Japan
Jamaica
Italy
Israel
Ireland
Iraq
Iran
Indonesia
India
Iceland
Hungary
Honduras
Haiti
Guyana
Guinea Bissau
Guatemala
Greece
Ghana
Germany
Georgia
Gambia, The
Gabon
France
Finland
Ethiopia
Estonia
Equatorial Guinea
El Salvador
Egypt, Arab Rep.
Ecuador
DRC
Dominican Republic
Denmark
Czech Republic
Cyprus
Croatia
Cote D Ivoire
Costa Rica
Congo
Comoros
Colombia
China
Chile
Chad
Central African Republic
Canada
Cameroon
Cabo Verde
Burkina Faso
Bulgaria
Brazil
Bosnia And Herzegovina
Bolivia
Benin
Belgium
Belarus
Barbados
Bangladesh
Bahrain
Azerbaijan
Austria
Australia
Armenia
Angola
Algeria
Albania
country
0
5
10
15
Number of years reported

Plotting change in key predictors over the 17-year data period.

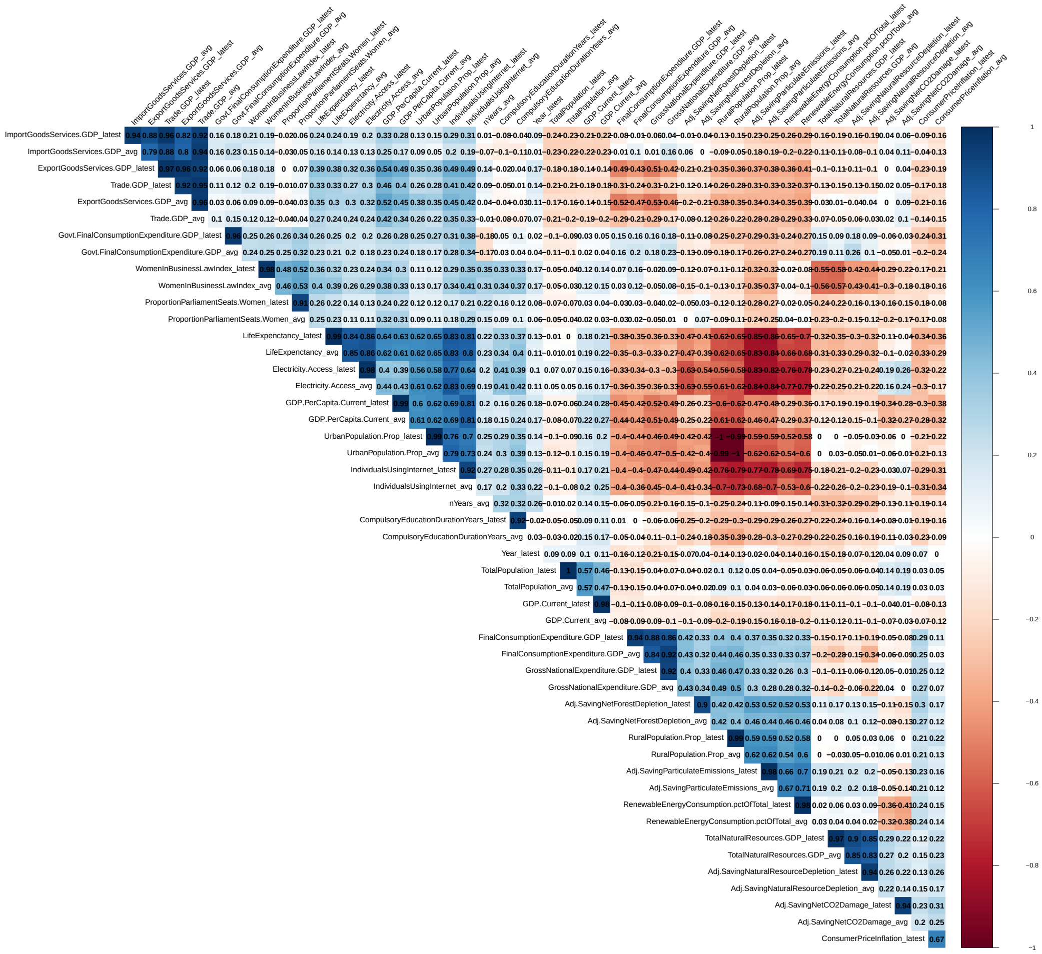

Internet Users by Year in Europe + North America


Proportion Urban Population by Year in Europe + North America


Internet Users by Year in Central and Southern Asia



Plotting correlations between variables



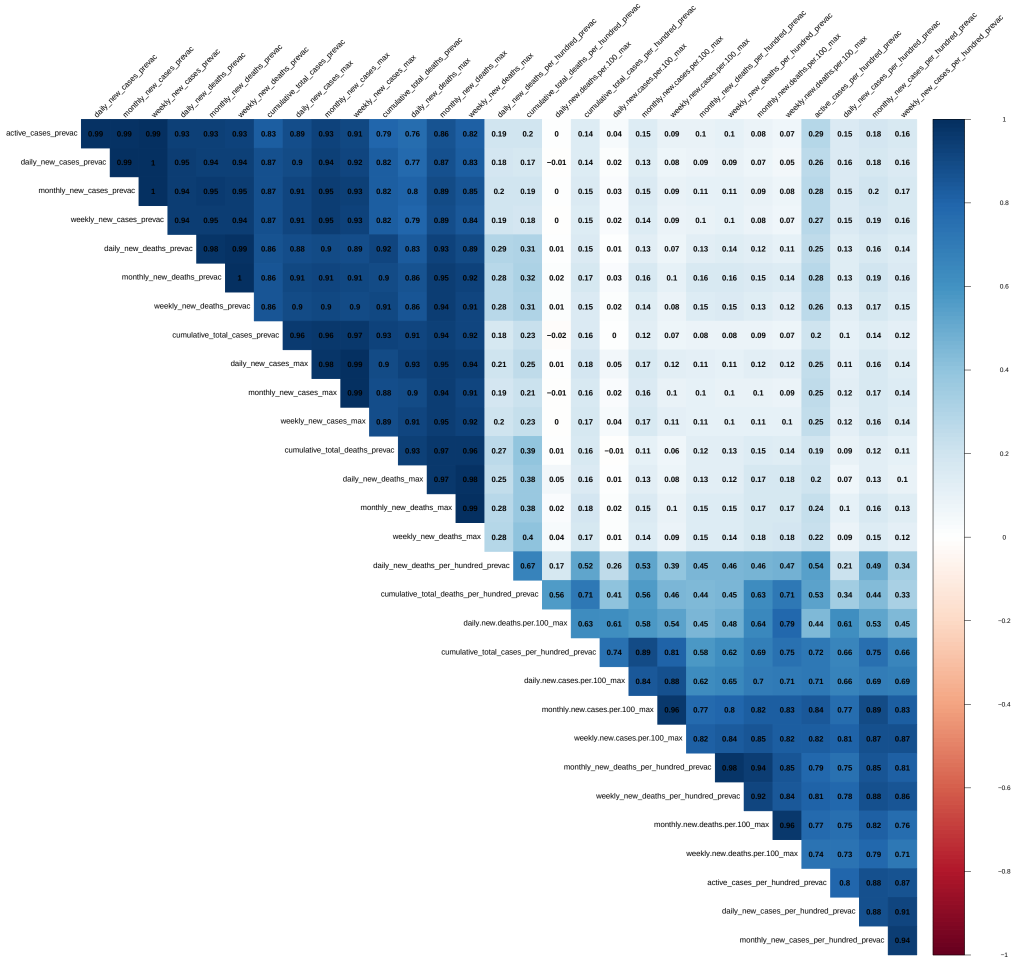
10.1.2 COVID

Below, we display plots of the cumulative an monthly number of cases worldwide.

We also plot the number of cumulative and monthly new deaths.

Plots visualizing the a wide range of vaccination rates over time, with some countries reaching high levels of vaccination at a very fast rate such as Gibraltar and Cuba, some countries showing more progressively increasing rates (e.g., European countries) and some countries having low vaccination rates up until today.

Plotting correlations between variables



10.2 Unused Dataframes

Below is code to generate dataframes that where in the end not used for modeling - may be useful for EDA or post-hoc analyses:

10.2.1 WSD Summary Dataframes

10.2.2 COVID Summary Dataframes

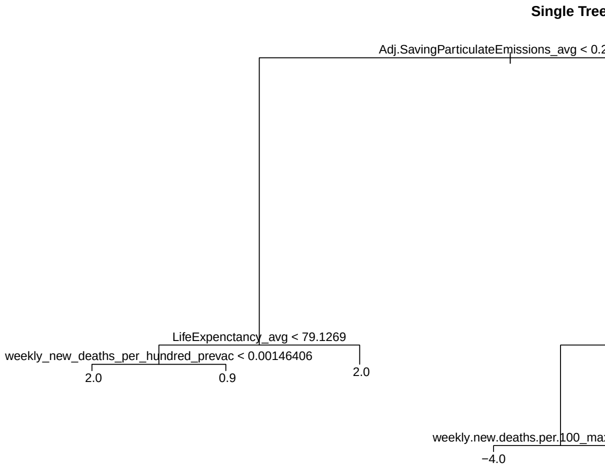
10.2.3 Integrated Summary Dataframes

Here's code to subdivide the train/test dfs based on what type of WSD predictors they include (i.e. so you can easily train a model only on the WSD 17-yr averages, plus static measures and covid cases/deaths).

10.3 Comparing Regsubsets Models

Code initially used to identify discrepancies in country names between COVID and WSD dataframes.

10.4 Regression Tree


Single Tree
Adj.SavingParticulateEmissions_avg < 0.2
LifeExpenctancy_avg < 79.1269
weekly_new_deaths_per_hundred_prevac < 0.00146406
2.0
0.9
2.0
weekly.new.deaths.per.100_max
−4.0

Single Regression Tree to Predict VSI. Created for illustrative purposes.