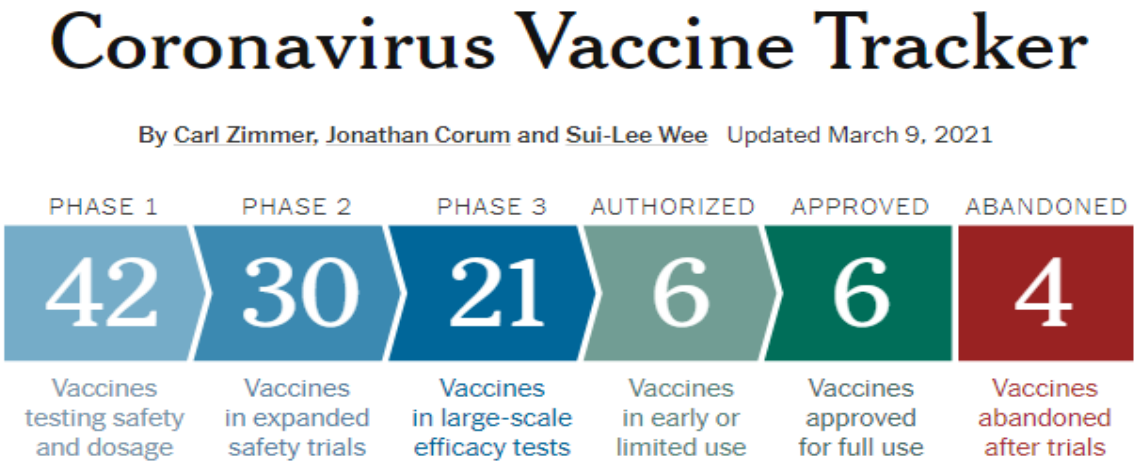
Contents

Introduction

Wondering about COVID-19 vaccination process?

Vaccines typically require years of research and testing before reaching the clinic, but in 2020, scientists embarked on a race to produce safe and effective coronavirus vaccines in record time. Researchers are currently testing 75 vaccines in clinical trials on humans, and 21 have reached the final stages of testing. At least 78 preclinical vaccines are under active investigation in animals according to The New York Times. **6 vaccines** are approved for full use as of today (3rd March 2021)



The arrival of coronavirus vaccines is beginning to have an impact on everyday life, with millions of newly inoculated people eagerly anticipating a return to long-postponed activities and visits with sorely missed relatives and friends.

1. Which countries have rolled out COVID-19 vaccines?
2. How is the daily vaccination trend in developing countries?
3. Which countries are ahead on vaccination process?

The answers aren't simple. In the meantime, the asymmetric nature of the rollout — with many elderly and health-care workers receiving shots first, while millions of others await their turns — is shifting relationships in families and in society more broadly. Grandparents who once hunkered down at home, most vulnerable to a virus that preys on the elderly, are likely to be better protected than younger relatives who are waiting to be vaccinated.

Datasets:

1. `Country_vaccinations.csv` file contains country-by-country data on global COVID-19 vaccinations.
   - ✓ Data is collected daily from Our World in Data GitHub Repository. This dataset only relies on figures that are verifiable based on public official sources.

   - **`location`**: name of the country (or region within a country).
   - **`iso_code`**: ISO 3166-1 alpha-3 – three-letter country codes.
   - **`date`**: date of the observation.
   - **`total_vaccinations`**: total number of doses administered.
   - **`total_vaccinations_per_hundred`**: `total_vaccinations` per 100 people in the total population of the country.
   - **`daily_vaccinations_raw`**: daily change in the total number of doses administered.
   - **`daily_vaccinations`**: new doses administered per day (7-day smoothed).
   - **`daily_vaccinations_per_million`**: `daily_vaccinations` per 1,000,000 people in the total population of the country.
   - **`people_vaccinated`**: total number of people who received at least one vaccine dose.
   - **`people_vaccinated_per_hundred`**: `people_vaccinated` per 100 people in the total population of the country.
   - **`people_fully_vaccinated`**: total number of people who received all doses prescribed by the vaccination protocol.
   - **`people_fully_vaccinated_per_hundred`**: `people_fully_vaccinated` per 100

2. Country_profile_variables.csv file contains the indicator variables of all countries.
   - ✓ Data is extracted from UNData website which in turn is collected from more than 20 international statistical sources compiled regularly by the Statistics Division and the Population Division of the United Nations, the statistical services of the United Nations, specialized agencies and other international organizations and institutions.
     - **`General Information`**
     - **`Economic Indicators`**
     - **`Social Indicators`**
     - **`Environmental & Infrastructure Indicators`**

Methodology:

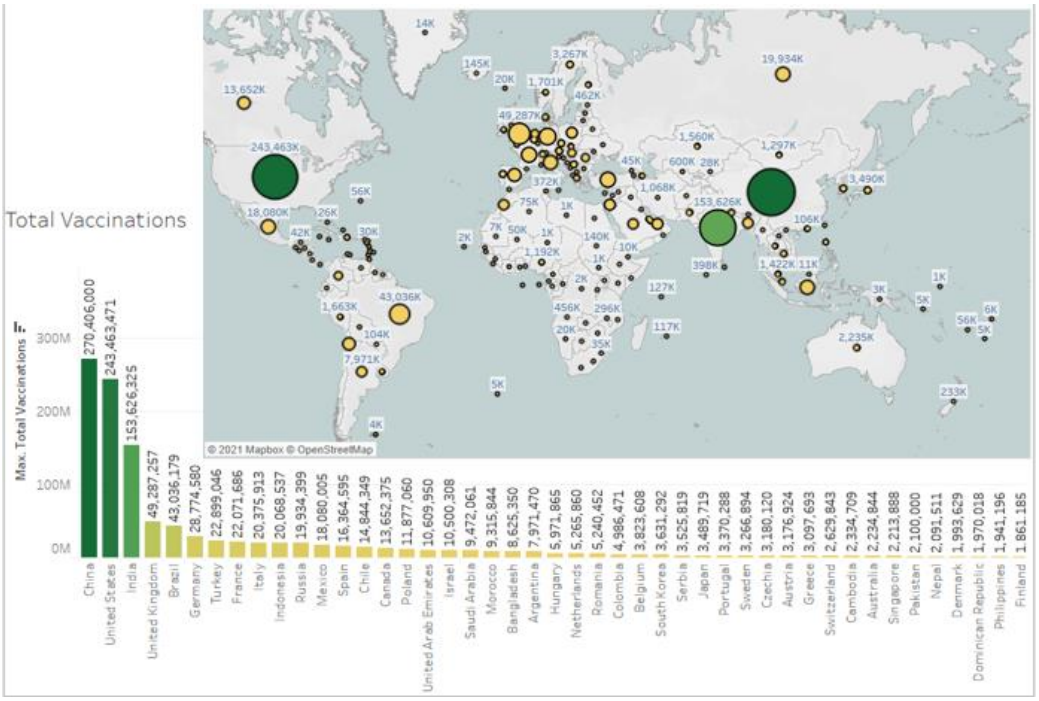
Q1: Use descriptive analysis and visualize data after the data wrangling.

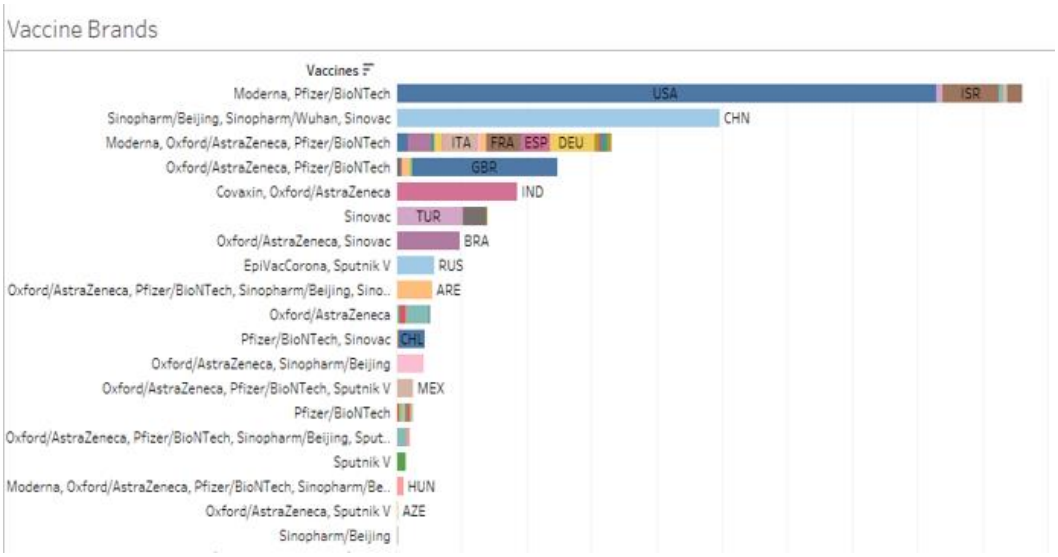
Q2: Preform multivariate regression analysis with the responses Y1 = daily_vaccinations, and Y2 total_vaccinations, and the predictors: Economic Indicators

Q3: Perform cluster analysis to segregate countries into distinct groups based on vaccination progress.

Multivariate Analysis & Interpretations

1. Which countries have rolled out COVID-19 vaccines?





Conclusion: Pictures are self-explanatory indicating the facts that China is leading country in administering Sinopharm, Sinovac vaccine brands followed by United States administering Medorna, Oxford, Pfizer, followed by India administering Covaxin, Oxford and so on.

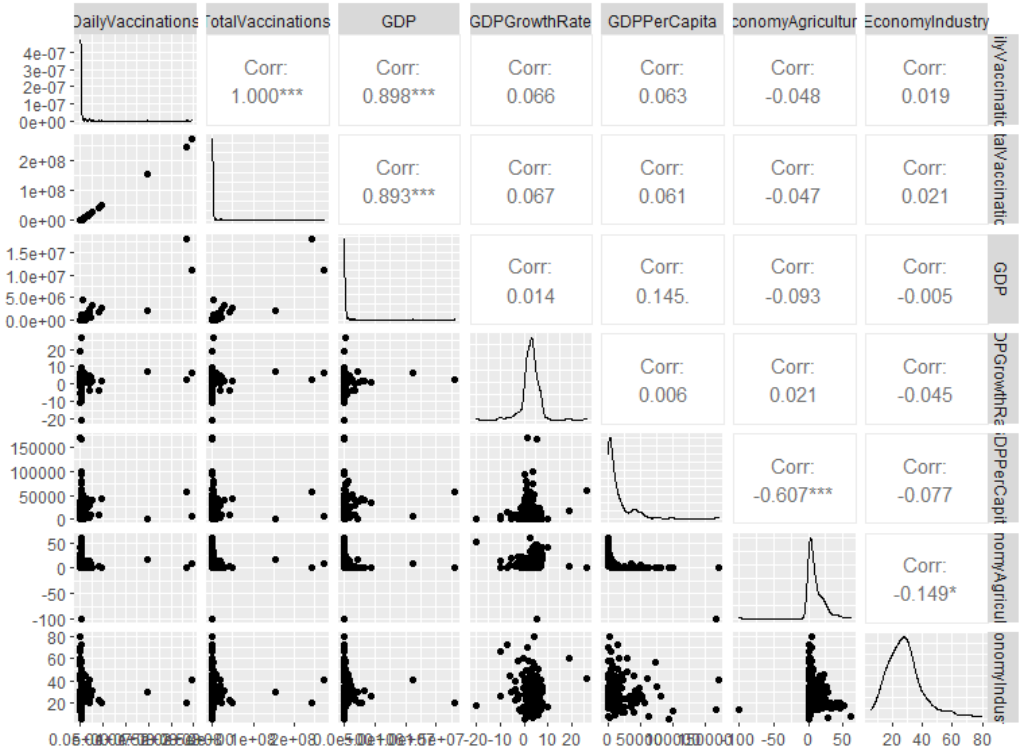
2. How is the daily vaccination trend in developing countries?

I have merged Daily and Total vaccinations from Dataset1 with Economic Indicators such as GDP, GDP Growth Rate, GDP Per Capita, Economy Agriculture, Economy Industry, Economy Services from Dataset2 into one and create new dataset for this analysis. So, let's find out good **Economy** predictors of Daily/Total Vaccinations.

Since we are using multivariate linear regression model for this analysis, the following assumptions should be met. For each value of the independent variable, the distribution of the dependent variable must be normal. The variance of the distribution of the dependent variable should be constant for all values of the independent variable. The relationship between the dependent variable and the independent variables should be linear, and all observations should be independent. So, the assumptions are: normality; linearity; independence; homoscedasticity. In other words, the residuals of a good model should be normally and randomly distributed i.e. the unknown does not depend on X ("homoscedasticity")

Let's profile our data to find the skewness and kurtosis of all selected variables to check normality.

```
> print(profiling_num(economy))
              variable          mean      std_dev variation_coef     p_01     p_05      p_25      p_50       p_75
1  DailyVaccinations 6.124395e+06 2.818056e+07      4.6013622 807.000 2637.250 41317.500 314775.50 1650043.750
2  TotalVaccinations 6.498206e+06 3.003616e+07      4.6222233 737.500 2832.250 46274.750 344799.00 1840945.500
3                GDP 4.062921e+05 1.675233e+06      4.1232237 240.750  765.000  6534.000  32062.00  194564.500
4      GDPGrowthRate 2.750568e+00 4.064705e+00      1.4777693  -9.975   -2.950     1.200      2.90       4.425
5       GDPPerCapita 1.658589e+04 2.504812e+04      1.5102063 369.425  618.100  2396.125   6581.65   19106.975
6 EconomyAgriculture 9.741477e+00 1.385285e+01      1.4220485   0.075    0.500     2.300      6.55      15.400
7    EconomyIndustry 2.838182e+01 1.291001e+01      0.4548689   7.175   12.100    19.675     26.85      33.325
8    EconomyServices 6.131591e+01 1.486160e+01      0.2423776  25.600   37.325    51.800     61.30      72.100
        p_95        p_99   skewness  kurtosis       iqr      range_98       range_80
1 19102849.500 1.693552e+08  7.3285960 58.553783 1608726.250  [807, 169355162.25]      [9599, 9492513]
2 20145381.000 1.760856e+08  7.4011661 59.864185 1794670.750 [737.5, 176085611.5] [10480.5, 9986184.5]
3  1607753.750 6.076921e+06  8.3398332 80.544062  188030.500 [240.75, 6076921.25]   [1406.5, 662004.5]
4        6.925 1.187500e+01 -0.1264740 15.747502       3.225      [-9.975, 11.875]       [-0.05, 6.5]
5    57168.750 1.165883e+05  3.2965946 17.449378   16710.850 [369.425, 116588.25]   [895.75, 44140.05]
6       32.100 4.712500e+01 -1.7918444 24.582300      13.100       [0.075, 47.125]          [0.7, 25.2]
7       54.975 7.077500e+01  1.2061302  5.168050      13.650       [7.175, 70.775]         [14.6, 44.3]
8       85.975 8.980000e+01 -0.2406456  2.870273      20.300         [25.6, 89.8]        [41.5, 79.85]
```

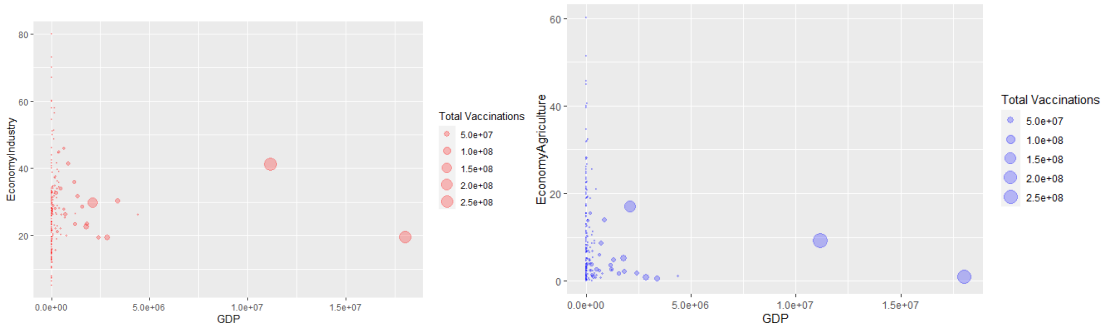


Above chart shows distribution (normal/skewed) of all variables and correlation among variables.

Daily, Total vaccinations, GDP, GDP Per Capita are completely right skewed, while GDP Growth Rate and Economy Industry are fairly normally distributed.

Daily vaccinations are highly positively correlated with Total vaccinations, GDP and highly negatively correlated with Economy Agriculture.

Let's look closely at Economy Industry and Economy Agriculture correlation with GDP to check independence of variables. We see an outlier in Economy Agriculter which is removed the below plots and for multivariate regression analysis.



Fit multivariate regression model on cleaned dataset.

```
lm.multi <- lm(cbind(DailyVaccinations,TotalVaccinations) ~ GDP + GDPGrowthRate
        + GDPPerCapita + EconomyAgriculture + EconomyIndustry + EconomyServices
        , data=economy)
```

```
Response DailyVaccinations :

Call:
lm(formula = DailyVaccinations ~ GDP + GDPGrowthRate + GDPPerCapita +
    EconomyAgriculture + EconomyIndustry + EconomyServices, data = economy)

Residuals:
      Min        1Q    Median        3Q       Max
-61345781  -2383113   -677645   1102630 111713158

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         1.616e+07  1.560e+07   1.036   0.3017
GDP                 1.532e+01  5.660e-01  27.076   <2e-16 ***
GDPGrowthRate       3.633e+05  2.302e+05   1.578   0.1164
GDPPerCapita       -8.584e+01  4.851e+01  -1.769   0.0786 .
EconomyAgriculture -1.359e+05  1.427e+05  -0.952   0.3422
EconomyIndustry    -1.214e+05  1.647e+05  -0.737   0.4621
EconomyServices    -1.805e+05  1.617e+05  -1.116   0.2658
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12330000 on 169 degrees of freedom
Multiple R-squared:  0.8151,    Adjusted R-squared:  0.8085
F-statistic: 124.2 on 6 and 169 DF,  p-value: < 2.2e-16
```

Estimates: The intercept tells us that when all the features are at 0

Standard Error is the standard error of our estimate, which allows us to construct marginal confidence intervals for the estimate of that particular feature.

t-value which tells us about how far our estimated parameter is from a hypothesized 0 value, scaled by the standard deviation of the estimate.

p-value is for the individual coefficient. If this probability is sufficiently low, we can reject the null hypothesis that this coefficient is 0.

Residual Standard Error: 123300000 on 169 degrees of freedom is Essentially standard deviation of residuals / errors of your regression model. It gives the standard deviation of the residuals, and tells us about how large the prediction error is in-sample or on the training data.

Multiple R-Squared: 0.81 is the Percent of the variance of Y variance is explained by our model.

Adjusted R-Squared: 0.8 is same as multiple R-Squared but takes into account the number of samples and variables you're using.

F-Statistic: 124.4 on 6 and 169 DF is a global test to check if your model has at least one significant variable taking into account number of variables and observations used.

Using the estimated residuals, check that your linear model satisfies the assumptions of a traditional linear regression model. Include the plots that you used and the relevant interpretations. We make a few assumptions when we use linear regression to model the relationship between a response and a predictor. These assumptions are essentially conditions that should be met before we draw inferences regarding the model estimates or before we use a model to make a prediction.

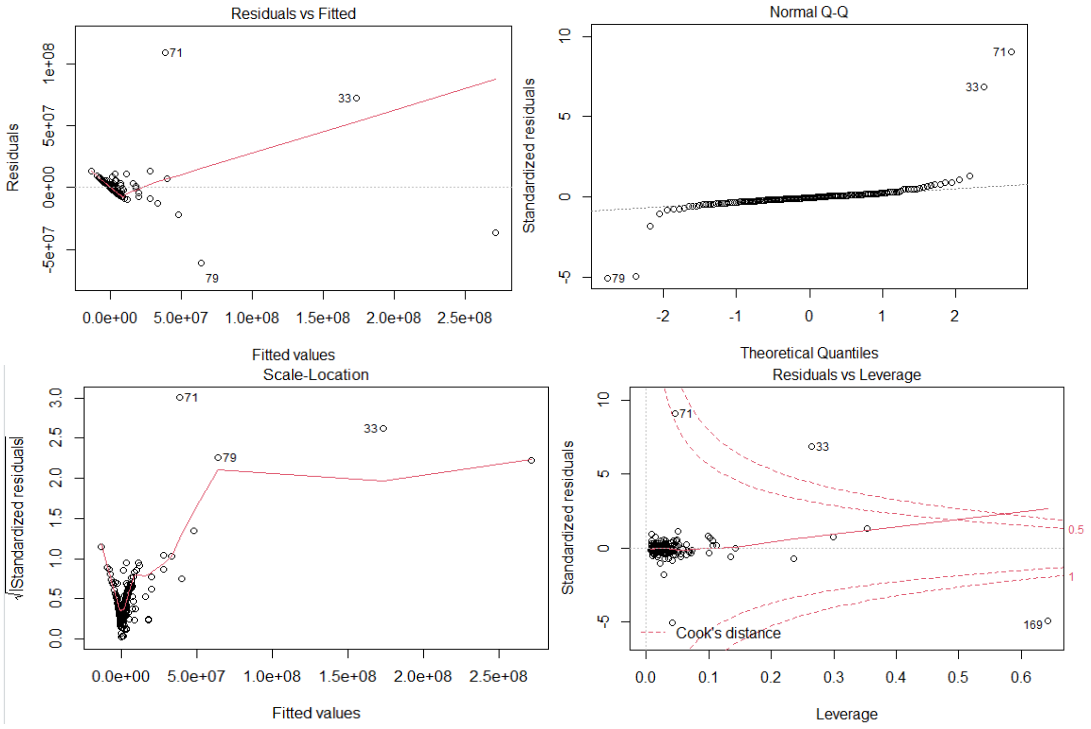
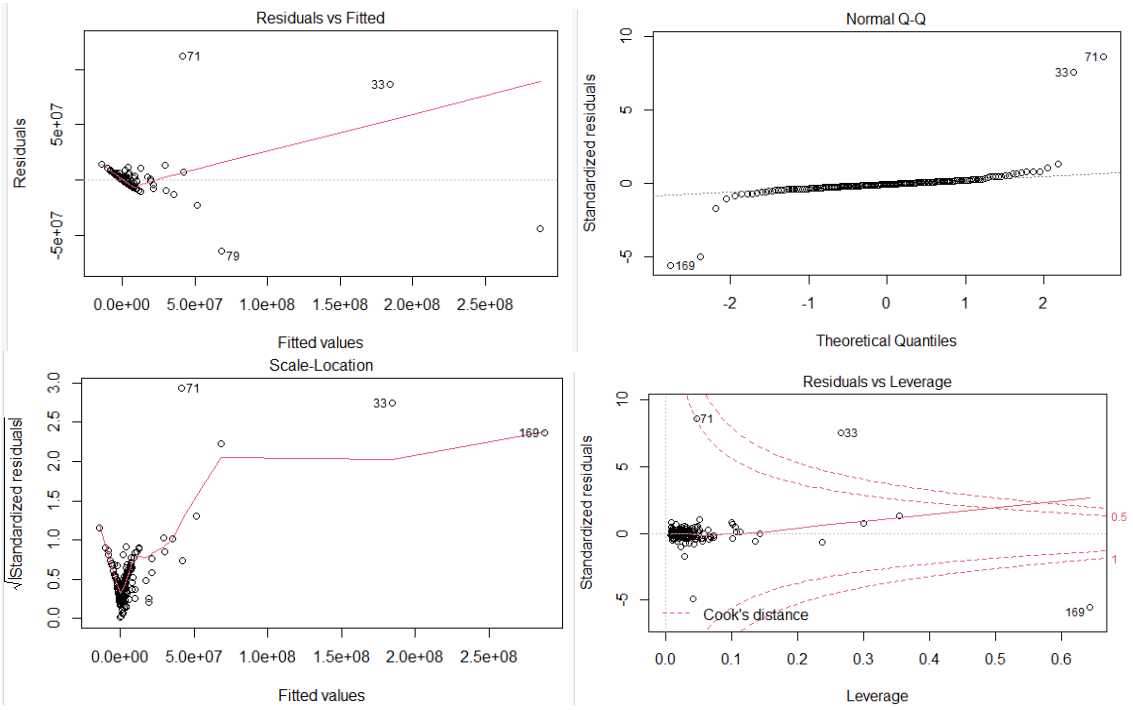


Figure: Diagnostics on model. DailyVaccinations ~ GDP + GDPGrowthRate + GDPPerCapita + EconomyAgriculture + EconomyIndustry + EconomyServices

```
Response TotalVaccinations :

Call:
lm(formula = TotalVaccinations ~ GDP + GDPGrowthRate + GDPPerCapita +
    EconomyAgriculture + EconomyIndustry + EconomyServices, data = economy)

Residuals:
      Min        1Q    Median        3Q       Max
-64531161  -2992687   -793192   1749769 112385627

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         3.218e+09  1.935e+09   1.663   0.0981 .
GDP                 1.624e+01  6.115e-01  26.553   <2e-16 ***
GDPGrowthRate       3.827e+05  2.487e+05   1.538   0.1258
GDPPerCapita       -8.974e+01  5.244e+01  -1.711   0.0889 .
EconomyAgriculture -3.213e+07  1.933e+07  -1.662   0.0984 .
EconomyIndustry    -3.214e+07  1.935e+07  -1.661   0.0986 .
EconomyServices    -3.220e+07  1.935e+07  -1.664   0.0979 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13320000 on 168 degrees of freedom
Multiple R-squared:  0.8113,    Adjusted R-squared:  0.8045
F-statistic: 120.3 on 6 and 168 DF,  p-value: < 2.2e-16
```

Diagnostic Plots



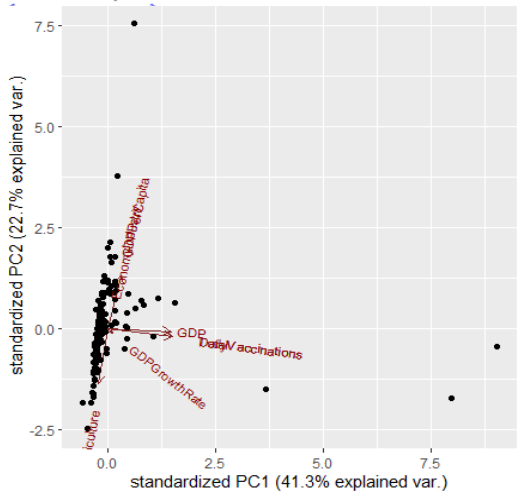

```
> n<-nrow(Resid)
> hat.Sigma <- t(Resid)%*%Resid/n
> hat.Sigma
                   DailyVaccinations TotalVaccinations
DailyVaccinations       1.442481e+14      1.563407e+14
TotalVaccinations       1.563407e+14      1.702374e+14
> sqrt(diag(hat.Sigma))
DailyVaccinations TotalVaccinations
         12010331          13047507
```

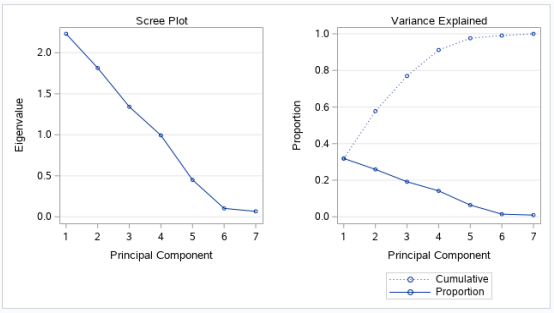
Conclusion: Residual plots shows two outliers exists in the data and multivariate model may not be very good fit, this may be due to lack of normality of all variables.

3. Which countries are ahead on vaccination process?

Let's first perform Principal Component Analysis (PCA) on the same dataset we created in previous analysis to find low dimensional representation of dataset that contains the most interesting or important features with majority of variance. For more information on PCA analysis please visit https://en.wikipedia.org/wiki/Principal_component_analysis

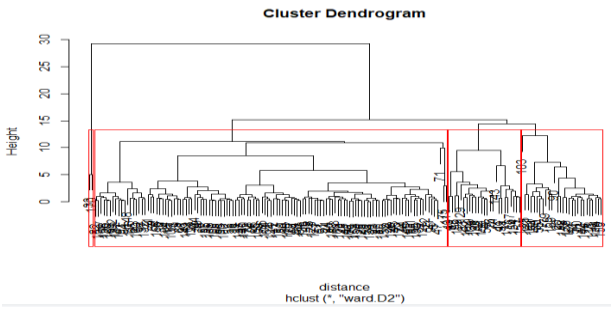
```
> summary(economy.pca)
Importance of components:
                          PC1    PC2    PC3    PC4     PC5     PC6     PC7
Standard deviation     1.7007 1.2607 1.0326 0.9834 0.59533 0.36045 0.02083
Proportion of Variance 0.4132 0.2271 0.1523 0.1382 0.05063 0.01856 0.00006
Cumulative Proportion  0.4132 0.6403 0.7926 0.9307 0.98138 0.99994 1.00000
```





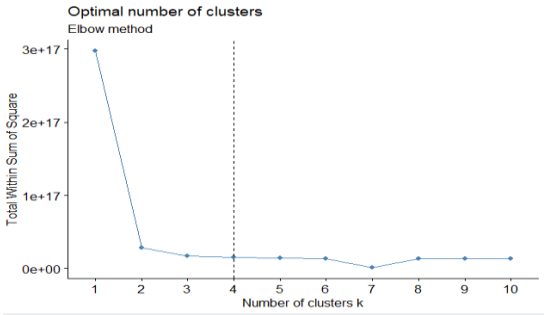
We obtained 7 principal components, which you call PC1-7. Each of these explains a percentage of the total variation in the dataset. That is to say: PC1 explains 41% of the total variance, which means that nearly 40% of the information in the dataset (7 variables) can be encapsulated by just that one Principal Component. PC2 explains 22% of the variance. So, by knowing the position of a sample in relation to just PC1 and PC2, you can get a very accurate view on where it stands in relation to other samples, as just PC1 and PC2 can explain 64% of the variance. From scree plot we see that all variables are important to conduct cluster analysis.Let's perform Clustering Methods on the whole dataset we created in previous analysis to segregate countries. Dataset is scaled or standardized before modeling.

Hierarchical (Ward's method): Unsupervised method to find out the unspecified number of clusters by creating dendrogram to view all groups obtained for all possible clusters, 4 in this case.



K-means : Unsupervised method to segregate the observations into K clusters.

Goal here is to minimize the total within-cluster variation as the sum of squared distances Euclidean distances between items and the corresponding centroid, for more information on k-mean please refer to [K-means Cluster Analysis · UC Business Analytics R Programming Guide (uc-r.github.io)](uc-r.github.io).

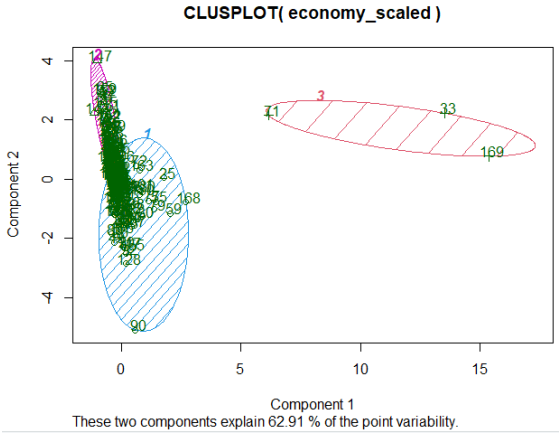


Elbow method shows that 4 clusters are optimum for this dataset, but I believe 3 clusters are good enough as total within-cluster sum of square (wss) doesn't change much from 3 to 4 clusters.

```
> aggregate(economy_scaled,by=list(clust4$cluster),FUN=mean)
  Group.1 DailyVaccinations TotalVaccinations         GDP GDPGrowthRate
1       1       -0.09362972       -0.09352399 -0.05092854    -0.1142336
2       2       -0.19474513       -0.19399379 -0.22038964     0.2091690
3       3        7.18818137        7.17049864  5.97020270     0.7281474
  GDPPerCapita EconomyAgriculture EconomyIndustry
1    0.2723320         -0.5456025       0.1608220
2   -0.6104483          1.1986922      -0.3587455
3    0.2763452         -0.1160937       0.1317560
```

From the means of three clusters we can interpret the below results.

1. Developing economies: Group1 countries that have conducted lowest dialy and total vaccinations also have lowest GDP moderate GDP growth rate, lowest GDP per capita, high agriculter and lowest Industry.
2. Emerging economies: Group2 countries that have moderately conducted daily anf total vaccinations also have moderate GDP and lowest GDP grown rate, lowest agriculter and highest Inductries.
3. Advanced economies: Group3 countries that have conducted highest dialy and total vaccinations also have higher GDP growth, GDP Growth Rate.



Conclusion: Advanced economies are ahead on vaccination.

limitations of data collection

Please note that all the derived results from the above analysis are subjected to dataset.

Dataset1: Unless specified, the vaccination data used here is compiled by Our World in Data from a variety of official and other sources such as local media. Each links to the ultimate source of its data; Where the latest available data are those reported to or compiled from local official sources by the World Health Organization, the link leads to the relevant WHO coronavirus dashboard.

Dataset2: UNdata *is a web-based data service for the global user community. It brings international statistical databases within easy reach of users through a single-entry point. Users can search and download a variety of statistical resources compiled by the United Nations (UN) statistical system and other international agencies. The numerous databases or tables collectively known as "datamarts" contain over 60 million data points and cover a wide range of statistical themes including agriculture, crime, communication, development assistance, education, energy, environment, finance, gender, health, labour market, manufacturing, national accounts, population and migration, science and technology, tourism, transport and trade.*

Terms and Conditions of use

UNdata *is a service provided by the United Nations to users of statistical data around the world. It provides easy access to data compiled and produced by UN agencies as well as other international agencies.*

Terms of Use: *All data and metadata provided on UNdata's website are available free of charge and may be copied freely, duplicated and further distributed provided that UNdata is cited as the reference.*

Disclaimers: *The data and metadata presented on UNdata are supplied as contained in the source database without any addition, subtraction, amendment, or modification by the United Nations Statistics Division.*

Accuracy and Currency of Data:

The United Nations Statistics Division strives for the highest level of accuracy and is committed to promptly correcting any errors on its part. It does not guarantee or make any express or implied representations regarding the accuracy, reliability, correctness, fitness for use for a particular purpose, or otherwise, whatsoever, of any of the databases in UNdata.

The United Nations Statistics Division periodically incorporates, without notice, revisions, updates and improvements to UNdata's content according to the sources' availability but undertakes no obligation to do so, timeously or at all.

Information about the quality or limitations of the data and metadata should be obtained from the organization responsible for the source database.

References:

Covid-19 Vaccine Tracker Updates: The Latest - The New York Times (nytimes.com)

COVID-19 vaccines: Lower case rates, hope against new variants (medicalnewstoday.com)

Efficacy, politics influence public trust in COVID-19 vaccine | Cornell Chronicle

Covid-19 vaccine tracker (ft.com)

GitHub - owid/covid-19-data: Data on COVID-19 (coronavirus) cases, deaths, hospitalizations, tests • All countries • Updated daily by Our World in Data

http://data.un.org/Host.aspx?Content=About

Country Statistics - UNData | Kaggle

Introduction to Multivariate Regression Analysis (nih.gov)