CST4070 - Comp. 3

Student:

- Name: Sarima
- Surname:Iyayi

1. Problem Definition

The following proposed research questions aim to determine whether vaccines are equally effective in both developed and developing countries or if there are any disparities in the effectiveness

1. What are the factors associated with COVID-19 vaccine effectiveness, such as vaccination coverage, vaccination rollout strategies,healthcare infrastructure,COVID-19 related factors, population demographics and socioeconomic indicators, across different countries?
2. Based on the identified factors from Research question 1, can we predict the development status of a country (i.e. developed or developing)?
3. Is there a difference in COVID-19 vaccine effectiveness between developed and developing countries using the dataset derived from research question 2 ?

As the research questions are interdependent, with each question building upon the results and data obtained from the previous question, I will define the data analysis lifecycle process for each research question as the analysis progresses.

Rq 1. What are the factors associated with COVID-19 vaccine effectiveness?

1. Data Description

The following code reads the primary data for analysis and creates a subset that includes only the relevant features related to vaccination coverage, vaccination rollout strategies, healthcare infrastructure, COVID-19 related factors, population demographics and socioeconomic indicators that are necessary for addressing the research question.

```
#Library to load and manipulate tabular data
library(data.table)
library(dplyr)

# Read the CSV file from the URL
url <- "https://github.com/owid/covid-19-data/raw/master/public/data/owid-covid-data.csv"
covid_data <- read.csv(url)
```

```
# Extract the relevant features for research question 1
features_rq1 = select(covid_data, date, population_density, median_age,
                      aged_65_older, aged_70_older, total_vaccinations,
                      people_fully_vaccinated, total_vaccinations_per_hundred,
                      hospital_beds_per_thousand, gdp_per_capita,
                      human_development_index, new_cases, new_deaths)
```

Let us examine our features_rq1 data table below

```
str(features_rq1)
```

```
## 'data.frame':    302628 obs. of  13 variables:
##  $ date                          : chr  "2020-01-03" "2020-01-04" "2020-01-05" "2020-01-06" ...
##  $ population_density            : num  54.4 54.4 54.4 54.4 54.4 ...
##  $ median_age                    : num  18.6 18.6 18.6 18.6 18.6 18.6 18.6 18.6 18.6 18.6 ...
##  $ aged_65_older                 : num  2.58 2.58 2.58 2.58 2.58 ...
##  $ aged_70_older                 : num  1.34 1.34 1.34 1.34 1.34 ...
##  $ total_vaccinations            : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ people_fully_vaccinated       : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ total_vaccinations_per_hundred: num  NA NA NA NA NA NA NA NA NA NA ...
##  $ hospital_beds_per_thousand    : num  0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 ...
##  $ gdp_per_capita                : num  1804 1804 1804 1804 1804 ...
##  $ human_development_index       : num  0.511 0.511 0.511 0.511 0.511 0.511 0.511 0.511 0.511 0.511
##  $ new_cases                     : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ new_deaths                    : num  0 0 0 0 0 0 0 0 0 0 ...
```

After inspecting our data table, we can see that it comprises of 301,127 rows and 13 columns. Although most of the attributes in the table are of quantitative data types, some may have null values, however, during our analysis, we will take steps to address and clean these null values as necessary to ensure the accuracy and reliability of our data.

2. Feature Engineering & Data Processing

The block of code below cleans the data and aggregates the features using the mean for easy comparison and benchmarking over the years

```
# Convert date column to date format
features_rq1$date = as.Date(features_rq1$date)

# extract the year and convert to integer format
features_rq1$year <- as.integer(format(features_rq1$date, "%Y"))

# Delete the date column as its irrelevant for now
features_rq1$date = NULL

# Remove rows with missing values
features_rq1 = na.omit(features_rq1)


# Aggregate mean of data at yearly level using the summarize_all()

rq1_by_year = summarize_all(group_by(features_rq1, year), mean)
```

```
str(rq1_by_year)
```

```
## tibble [4 x 13] (S3: tbl_df/tbl/data.frame)
##  $ year                          : int [1:4] 2020 2021 2022 2023
##  $ population_density            : num [1:4] 279 293 307 155
##  $ median_age                    : num [1:4] 37.4 35.1 35.6 36.8
##  $ aged_65_older                 : num [1:4] 14.6 11.8 12.4 13.9
##  $ aged_70_older                 : num [1:4] 9.47 7.68 8.08 9.17
##  $ total_vaccinations            : num [1:4] 1.26e+06 8.42e+07 3.28e+08 6.37e+08
##  $ people_fully_vaccinated       : num [1:4] 8.22e+03 3.19e+07 1.30e+08 2.44e+08
##  $ total_vaccinations_per_hundred: num [1:4] 0.643 63.905 164.876 192.645
##  $ hospital_beds_per_thousand    : num [1:4] 3.55 3.61 3.72 4.06
##  $ gdp_per_capita                : num [1:4] 39336 25812 26255 28087
##  $ human_development_index       : num [1:4] 0.885 0.808 0.815 0.831
##  $ new_cases                     : num [1:4] 159516 14709 37859 14712
##  $ new_deaths                    : num [1:4] 2824.6 253.7 112.6 99.5
```

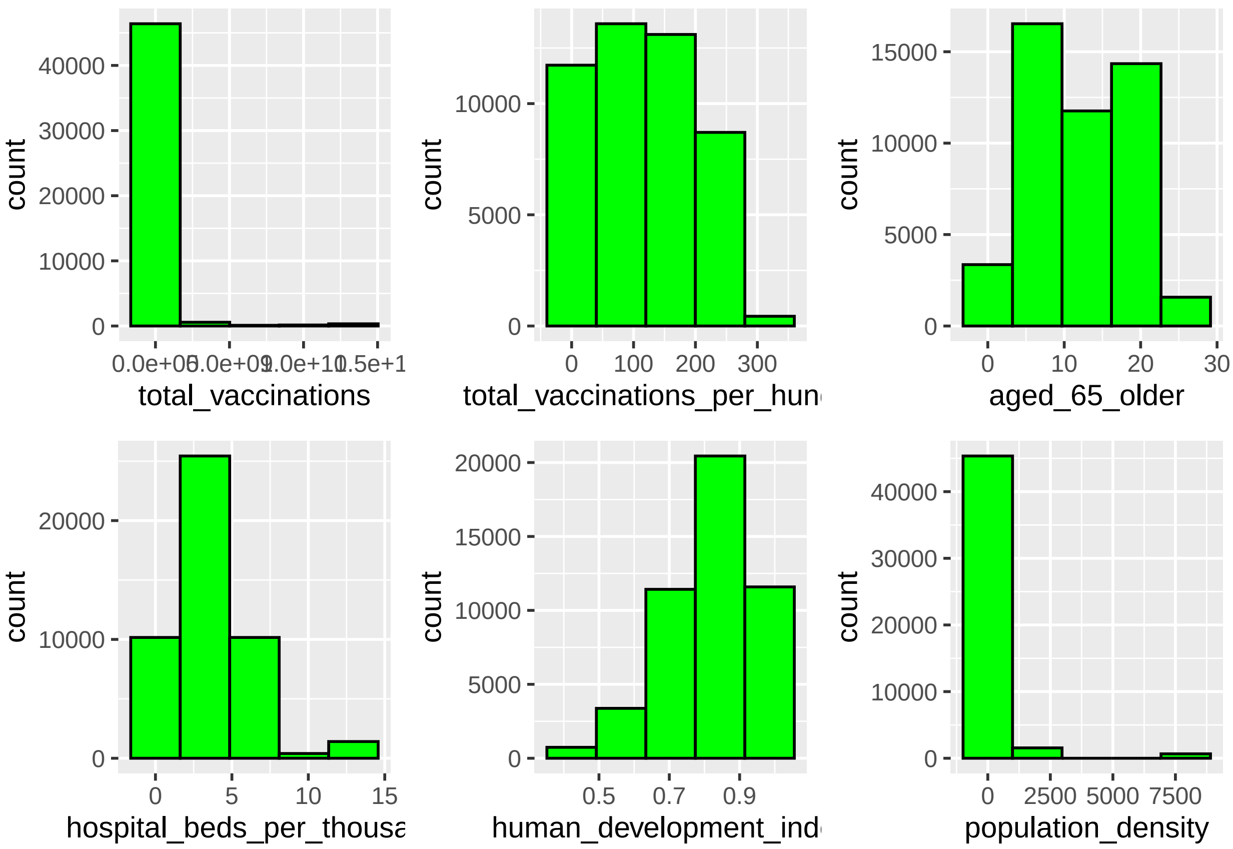
3. Modelling

To investigate research question 1, we plan to use Exploratory Data Analysis (EDA) to explore the relationship between features, using techniques such as a correlation matrix and solidifying our analysis using a regression analysis to properly understand the role of our features, which includes vaccination data, socioeconomic data, and healthcare infrastructure data, in explaining the variability of the vaccine effectiveness. From the results, we will determine the main factors associated with COVID-19 vaccine effectiveness, providing valuable insights for our next research question.

4. Results

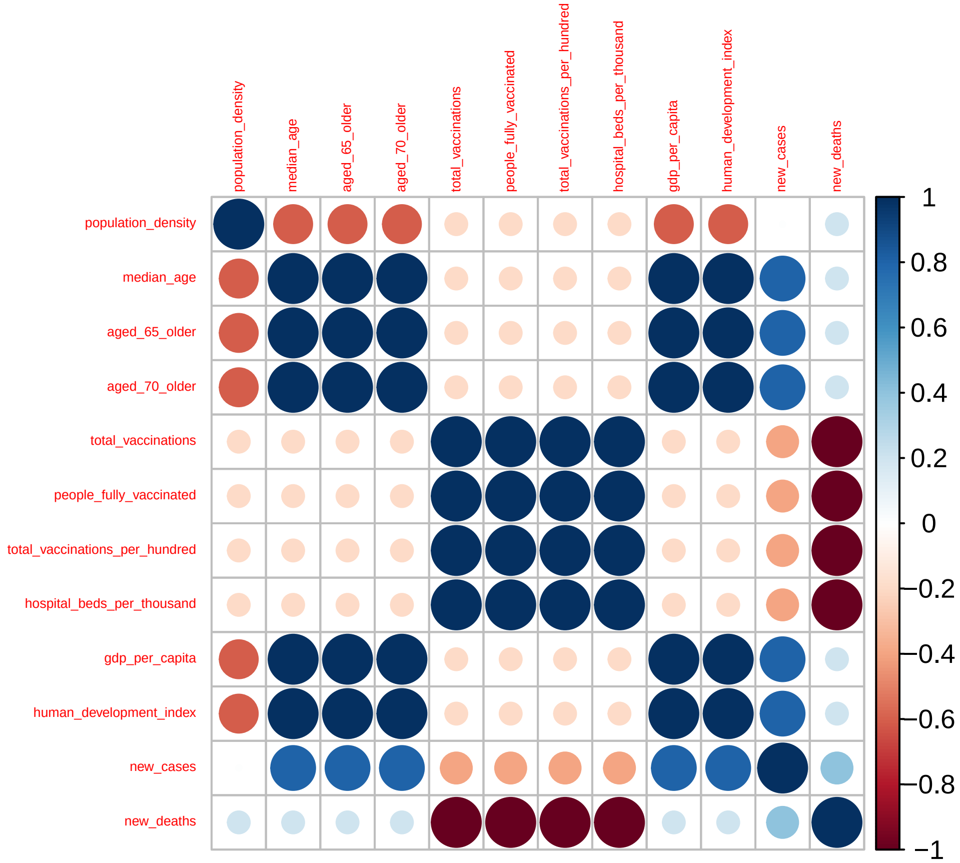
To investigate the relationships between selected features, we generate histogram charts for 6 randomly selected columns from the dataset to examine the data distribution in our dataset and determine the appropriate correlation analysis to perform using a correlation matrix.

```
# load library to plot frequency distribution
library(ggplot2)
library(gridExtra)

# Create an empty list to store the histograms
hist_list <- list()

# Loop through each index and create a histogram
for (i in c('total_vaccinations','total_vaccinations_per_hundred',
            'aged_65_older', 'hospital_beds_per_thousand',
            'human_development_index', 'population_density')) {
  x <- as.numeric(features_rq1[[i]])  # Convert column to numeric
  label <- paste( i)
  hist_list[[i]] <- ggplot(data.frame(x = x), aes(x = x)) +
    geom_histogram(bins=5, fill = 'green', col = 'black') +
    labs(x = label)
}

# Combine histograms into a grid using grid.arrange with specified nrow and ncol
grid.arrange(grobs = hist_list, nrow = 2)
```



Since most of the features in the frequency distribution do not exhibit a normal distribution, indicating that they are not parametric, we will be employing a Spearman correlation matrix for our correlation analysis.

```
#load librabry to compute correlation matrix
library(corrplot)

# Calculate Spearman correlation matrix, excluding the year column
correlation_matrix = cor(rq1_by_year[, -1],
                         method = "spearman")

# Visualize Spearman correlation matrix
corrplot(correlation_matrix, tl.cex = 0.38)
```



From the Spearman correlation matrix some significant correlations are observed between the socio-economic factors, `gdp_per_capita` and `human_development_index`, as well as the population demographic factors, `median_age`, `aged_60_older`, `aged_70_older age`, and other covid related factors, `new_cases`, `new_deaths`

To establish a more robust understanding of the factors associated with COVID-19 vaccine effectiveness, we would perform a further analysis using regression analysis to determine the strength, direction, and significance of the relationships between the factors and COVID-19 vaccine effectiveness using `'new_cases'` as our measure of effectiveness, all in the code below.

```
model = lm(data= features_rq1, formula = new_cases ~ population_density +
             median_age + aged_65_older + aged_70_older +
             hospital_beds_per_thousand + gdp_per_capita +
             human_development_index  + total_vaccinations_per_hundred +
             total_vaccinations + people_fully_vaccinated + new_deaths )

summary(model)
```

```
##
## Call:
## lm(formula = new_cases ~ population_density + median_age + aged_65_older +
##     aged_70_older + hospital_beds_per_thousand + gdp_per_capita +
##     human_development_index + total_vaccinations_per_hundred +
##     total_vaccinations + people_fully_vaccinated + new_deaths,
##     data = features_rq1)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -817941   -7047     679    7540 6636229
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   -2.737e+03  6.992e+03  -0.391 0.695483
## population_density             7.348e-01  6.474e-01   1.135 0.256386
## median_age                    -1.379e+03  2.233e+02  -6.177 6.59e-10 ***
## aged_65_older                 -2.754e+03  7.465e+02  -3.689 0.000225 ***
## aged_70_older                  5.557e+03  9.905e+02   5.610 2.03e-08 ***
## hospital_beds_per_thousand     2.515e+03  2.733e+02   9.204  < 2e-16 ***
## gdp_per_capita                 1.637e-01  5.077e-02   3.225 0.001260 **
## human_development_index        2.499e+04  1.184e+04   2.110 0.034843 *
## total_vaccinations_per_hundred  6.045e+01  7.167e+00   8.435  < 2e-16 ***
## total_vaccinations            -1.498e-04  6.520e-06 -22.975  < 2e-16 ***
## people_fully_vaccinated        5.227e-04  1.642e-05  31.836  < 2e-16 ***
## new_deaths                     6.756e+01  5.653e-01 119.518  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 111200 on 47562 degrees of freedom
## Multiple R-squared:  0.5292, Adjusted R-squared:  0.5291
## F-statistic:  4860 on 11 and 47562 DF,  p-value: < 2.2e-16
```

Findings 1: from the result of the regression analysis, an **R-squared** value of **0.52** indicates that approximately 52% of the variance in the dependent variable (in this case, “new_cases”) can be explained by the independent variables. Also based on the estimated coefficients and standard errors provided from the result,`people_fully_vaccinated` , `gdp_per_capita` , `human_development_index` , `aged_70_older` ,`hospital_beds_per_thousand`, `new_deaths`, `total_vaccinations_per_hundred`,show a statistically significant association with vaccine effectiveness.

Rq 2. Based on the identified factors from Research question 1, can we predict the development status of a country (i.e. developed or developing)?

1. Data Description

The following code creates the main dataset, `features_rq2`, required to investigate research question 2 using the factors derived from research question 1 and then we’ll examine them.

```r
# Extract the relevant features for research question 2 and rename
#location column to country for easy reference

features_rq2 = select(covid_data, country = location, new_cases,
                      people_fully_vaccinated, gdp_per_capita,
                      human_development_index, aged_70_older,
                      hospital_beds_per_thousand, new_deaths,
                      total_vaccinations_per_hundred)

str(features_rq2)
```

```
## 'data.frame':    302628 obs. of  9 variables:
##  $ country                       : chr  "Afghanistan" "Afghanistan" "Afghanistan" "Afghanistan" ...
##  $ new_cases                     : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ people_fully_vaccinated       : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ gdp_per_capita                : num  1804 1804 1804 1804 1804 ...
##  $ human_development_index       : num  0.511 0.511 0.511 0.511 0.511 0.511 0.511 0.511 0.511 0.511
##  $ aged_70_older                 : num  1.34 1.34 1.34 1.34 1.34 ...
##  $ hospital_beds_per_thousand    : num  0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 ...
##  $ new_deaths                    : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ total_vaccinations_per_hundred: num  NA NA NA NA NA NA NA NA NA NA ...
```

features_rq2 has 8 columns, 302571 rows, and contains data on countries, vaccination coverage, socio-economy, health infrastructure, vaccination rollout strategies, and COVID-19 related data. We will use this dataset to predict countries' development status, handling any Null values where neccessary.

2. Feature Engineering & Data Processing

The block of code below cleans the data, creates a new feature, converts it to factors to ensure compatibility with the random forest model.

```
# Remove missing values
features_rq2 = na.omit(features_rq2)

# Create new a feature, development_status for prediction
features_rq2 = mutate(features_rq2, development_status = ifelse(gdp_per_capita >=
                                         25000 & human_development_index >=
                                         0.8, "Developed","Developing"))

# Convert development_status to factor
features_rq2$development_status = as.factor(features_rq2$development_status)


str(features_rq2)
```

```
## 'data.frame':    47574 obs. of  10 variables:
##  $ country                       : chr  "Afghanistan" "Afghanistan" "Afghanistan" "Afghanistan" ...
##  $ new_cases                     : num  340 453 547 840 623 ...
##  $ people_fully_vaccinated       : num  55624 77560 96910 111082 113739 ...
##  $ gdp_per_capita                : num  1804 1804 1804 1804 1804 ...
##  $ human_development_index       : num  0.511 0.511 0.511 0.511 0.511 0.511 0.511 0.511 0.511 0.511
##  $ aged_70_older                 : num  1.34 1.34 1.34 1.34 1.34 ...
##  $ hospital_beds_per_thousand    : num  0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 ...
##  $ new_deaths                    : num  12 10 10 19 14 20 34 27 64 85 ...
##  $ total_vaccinations_per_hundred: num  1.23 1.33 1.39 1.44 1.44 1.46 1.52 1.53 1.56 1.61 ...
##  $ development_status            : Factor w/ 2 levels "Developed","Developing": 2 2 2 2 2 2 2 2 2 2
##  - attr(*, "na.action")= 'omit' Named int [1:255054] 1 2 3 4 5 6 7 8 9 10 ...
##   ..- attr(*, "names")= chr [1:255054] "1" "2" "3" "4" ...
```

3. Modelling

The machine learning model I have chosen for my analysis is the Random Forest algorithm.

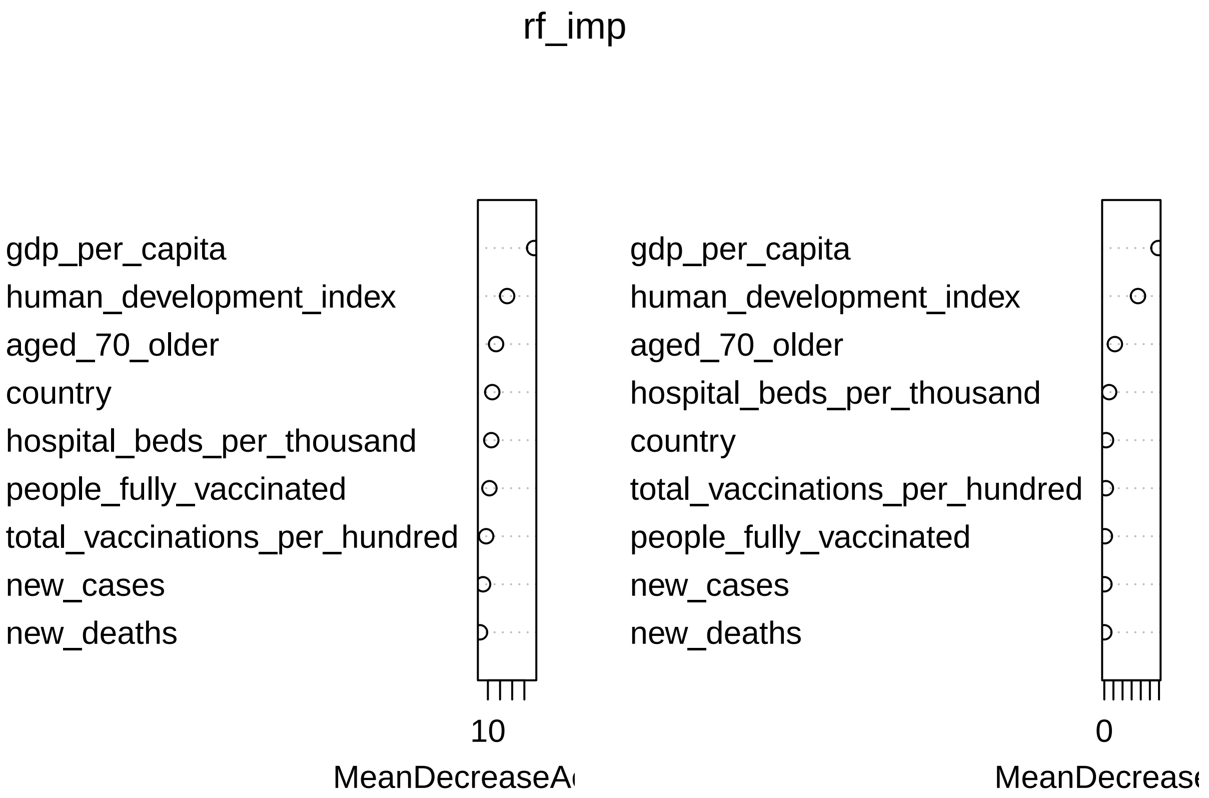
One of the main reasons for choosing Random Forest is its ability to *handle nonlinear relationships* in the data, as the relationships between predictors and the outcome variable may not be linear. Another key advantage of Random Forest is its robustness to overfitting, it *reduces the risk of capturing noise* in the training data and improves generalization performance. Random Forest also *handles missing values* effectively without requiring imputation. This is important as our dataset may still contain missing information. Random Forest can perform these, allowing for a more accurate and robust prediction model.

Furthermore, Random Forest provides a measure of *feature importance*, which would help in identifying the most relevant predictors for the classification task. This can provide insights into which vaccination effectiveness factors are most important in predicting the development status of a country, and can guide feature selection decisions.

```
# Load required libraries
library(randomForest)

#set a seed for reproducibility
set.seed(10)

#Perform random forest classification and view the
#relative importance of each feature for the prediction
rf_imp = randomForest(development_status ~ ., data =
                        features_rq2, importance = TRUE)

rf_imp$importance
```

```
##                                   Developed   Developing MeanDecreaseAccuracy MeanDecreaseGini
## country                        0.0127629399 6.938930e-03         9.556626e-03       381.421693
## new_cases                      0.0004964367 2.728677e-04         3.732045e-04        31.440555
## people_fully_vaccinated        0.0043656316 3.656578e-03         3.974702e-03       128.363018
## gdp_per_capita                 0.3895735435 2.359884e-01         3.051074e-01     11872.491603
## human_development_index        0.2541400983 1.203397e-01         1.805691e-01      7390.314362
## aged_70_older                  0.0510847417 2.145764e-02         3.477654e-02      2321.097466
## hospital_beds_per_thousand     0.0184904005 8.515320e-03         1.299875e-02      1065.200762
## new_deaths                     0.0001270296 4.018923e-05         7.926546e-05         7.833633
## total_vaccinations_per_hundred 0.0016394107 2.181001e-03         1.938754e-03       351.576877
```

```
#visualize result
varImpPlot(rf_imp)
```



In summary, important predictors for determining development status are `gdp_per_capita`, `human_development_index`, and `aged_70_older`, while `country`, `people_fully_vaccinated`, `hospital_beds_per_thousand`, `new_deaths`, `new_cases` and `total_vaccinations_per_hundred` may have lower importance, but excluding less important predictors may result in loss of valuable information and potential loss of predictive accuracy, so the complete dataset will be used for further analysis.

The script below uses train data to create a Random Forest model to make the predictions. The results would be shown in the next phase

```
# Split the data into training and testing sets
train_indices <- sample(1:nrow(features_rq2), nrow(features_rq2)*0.8) # 80% for training
train_data <- features_rq2[train_indices, ]
test_data <- features_rq2[-train_indices, ]

# Perform random forest classification
rf_model = randomForest(development_status ~ ., data = train_data)

# Predict the development status on test data and have in
#output the probability of generating a give class
rf_predictions= predict(rf_model, test_data)

rf_predictions_prob= predict(rf_model, test_data,  type = "prob")

head(rf_predictions_prob)
```

```
##     Developed Developing
## 504         0           1
## 518         0           1
## 545         0           1
## 596         0           1
## 774         0           1
## 780         0           1
```

4. Results

The following code block conducts an evaluation of the modeling results and uses test data to visualise the Confusion Matrix.

```
# Evaluate the model
confusion_matrix = table(test_data$development_status, rf_predictions)

confusion_matrix
```

```
##             rf_predictions
##              Developed Developing
##   Developed       4206          0
##   Developing         0       5309
```

In the cofusion matrix above, the model correctly predicted 4206 instances as "Developed" and 5309 instances as "Developing" and there were 0 instances misclassified as "Developed" and 0 instances misclassified as "Developing".

The code below evaluates the accuracy of the model

```
#Evaluate the accuracy of the model
accuracy = sum(diag(confusion_matrix)) / sum(confusion_matrix)

print(paste("Accuracy:", accuracy))
```

```
## [1] "Accuracy: 1"
```

Findings: The above results indicates that the model achieved perfect accuracy in this prediction task and that th.

Rq 3. Is there a difference in COVID-19 vaccine effectiveness between developed and developing countries using the dataset derived from research question 2?

1. Data Description

The script below loads the dataset required to investigate research question 3. The dataset would be gotten from the dataset of research question 2 as it contains the development status of each country. To compare COVID-19 vaccine effectiveness between developed and developing countries, we would need to analyze `new_cases` and `new_deaths`, the relevant vaccine effectiveness data and compare it between the two groups.

```
# Extract the relevant features for research question 3
#from the previous research question.
features_rq3 = select(features_rq2, country, new_cases, new_deaths, development_status)

features_rq3$year = features_rq1$year

str(features_rq3)
```

```
## ’data.frame’:    47574 obs. of  5 variables:
##  $ country           : chr  "Afghanistan" "Afghanistan" "Afghanistan" "Afghanistan" ...
##  $ new_cases         : num  340 453 547 840 623 ...
##  $ new_deaths        : num  12 10 10 19 14 20 34 27 64 85 ...
##  $ development_status: Factor w/ 2 levels "Developed","Developing": 2 2 2 2 2 2 2 2 2 2 ...
##  $ year              : int  2021 2021 2021 2021 2021 2021 2021 2021 2021 2021 ...
##  - attr(*, "na.action")= ’omit’ Named int [1:255054] 1 2 3 4 5 6 7 8 9 10 ...
##   ..- attr(*, "names")= chr [1:255054] "1" "2" "3" "4" ...
```

The dataset contains 47,573 observations (rows) and 4 variables (columns).`new_cases` and `new_deaths` are numerical values, `country` is a character variable and `development_status`, a factor variable. These features are relevant for the analysis as they provide information about the COVID-19 cases and deaths in different countries, and the development status of each country which can be used to investigate potential differences or patterns related to the COVID-19 vaccine effectiveness.

2. Feature Engineering & Data Processing

The block of code below prepares the data for analysis.

```
# Convert development_status to character format
features_rq3$development_status <- as.character(features_rq3$development_status)


#create separate data frames for developed and developing countries
#and Filter out rows with zero new_cases values
developed_countries <- features_rq3 %>%
  filter(development_status == "Developed", new_cases != 0)

developing_countries <- features_rq3 %>%
  filter(development_status == "Developing", new_cases != 0)

#Group dataset by year
rq3_by_year <- features_rq3 %>%
 group_by(development_status, year) %>%
  summarize(total_new_cases = sum(new_cases, na.rm = TRUE))#
```

```
## ‘summarise()‘ has grouped output by ’development_status’. You can override using the ‘.groups‘
## argument.
```

```
str(developed_countries)
```

```
## ’data.frame’:    20387 obs. of  5 variables:
##  $ country           : chr  "Australia" "Australia" "Australia" "Australia" ...
##  $ new_cases         : num  4 7 2 8 11 7 5 8 8 10 ...
##  $ new_deaths        : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ development_status: chr  "Developed" "Developed" "Developed" "Developed" ...
##  $ year              : int  2021 2021 2021 2021 2021 2021 2021 2021 2021 2021 ...
##  - attr(*, "na.action")= ’omit’ Named int [1:255054] 1 2 3 4 5 6 7 8 9 10 ...
##   ..- attr(*, "names")= chr [1:255054] "1" "2" "3" "4" ...
```

```
str(developing_countries)
```

```
## ’data.frame’:    22975 obs. of  5 variables:
##  $ country           : chr  "Afghanistan" "Afghanistan" "Afghanistan" "Afghanistan" ...
##  $ new_cases         : num  340 453 547 840 623 ...
##  $ new_deaths        : num  12 10 10 19 14 20 34 27 64 85 ...
##  $ development_status: chr  "Developing" "Developing" "Developing" "Developing" ...
##  $ year              : int  2021 2021 2021 2021 2021 2021 2021 2021 2021 2021 ...
##  - attr(*, "na.action")= ’omit’ Named int [1:255054] 1 2 3 4 5 6 7 8 9 10 ...
##   ..- attr(*, "names")= chr [1:255054] "1" "2" "3" "4" ...
```

3. Modelling

The analysis for this task relies on basic statistical methods to compare and summarize data, rather than employing any machine learning techniques.

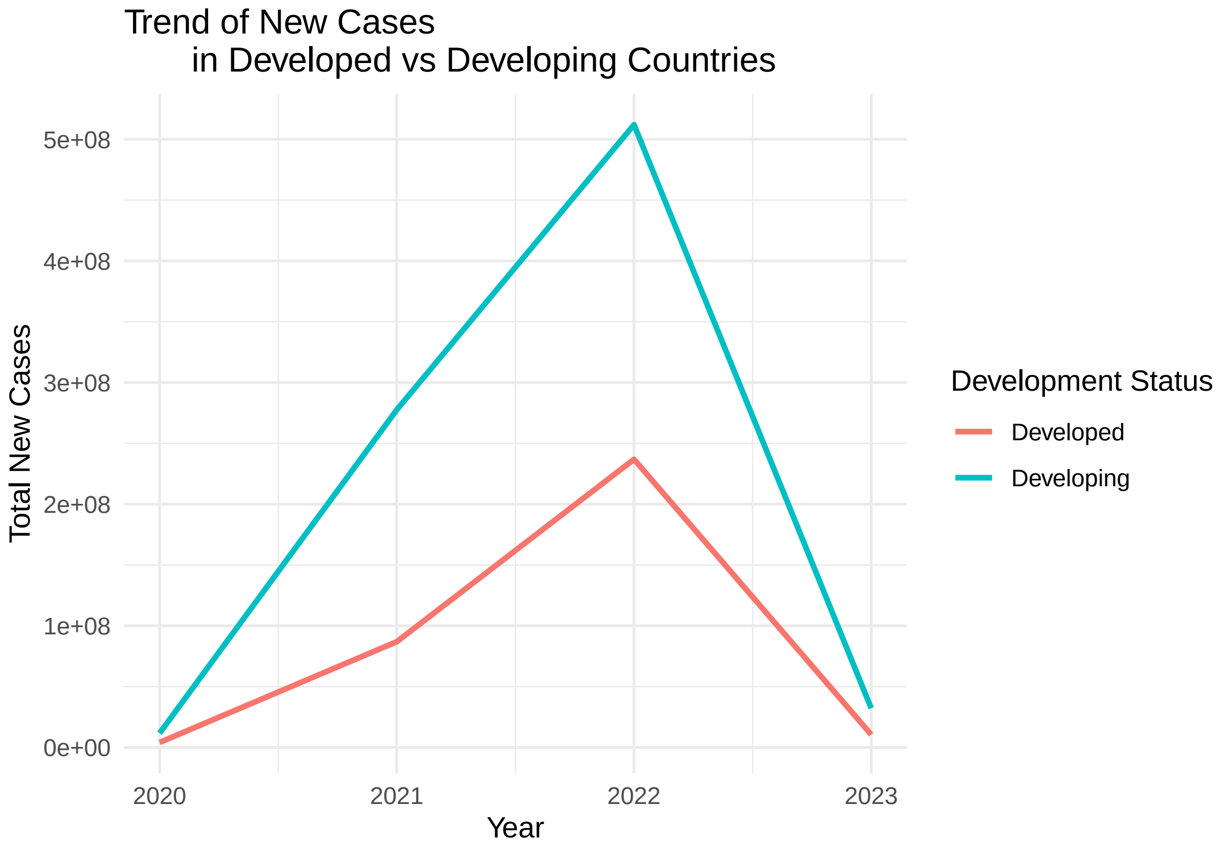
A Statistical analysis would be preformed using t-test to test the significance of differences in the number of new COVID-19 cases(which would be our indicator of COVID vaccine effectiveness) between these two groups. Before this, we would perform Data visualizations to provide insights into the data and help identify any differences or patterns that may not be apparent through statistical analyses alone

The analysis and results are visualized in the subsequent step.

4. Results

To explore the patterns and trends in vaccine effectiveness for developed and developing countries, we would plot a line chart below

```
# Create a line plot to visualize the trend of new cases over the years
line_plot <- ggplot(rq3_by_year, aes(x = year, y = total_new_cases, color =
  development_status)) +
  geom_line(size = 1) +
  labs(x = "Year", y = "Total New Cases", title = "Trend of New Cases
       in Developed vs Developing Countries") +
  theme_minimal() +
  scale_color_discrete(name = "Development Status")

# Display the line plot
print(line_plot)
```



To further analyze this, we would perform a Two Sample t-test hypothesis testing technique to provide a more rigorous evidence of whether there is a statistically significant difference in vaccine effectiveness between developed and developing countries.

```
# Perform t-test for new_cases between developed and developing countries
t_test_result <- t.test(developed_countries$new_cases, developing_countries$new_cases)

# Print the t-test result
cat("T-Test Result:\n")
```

```
## T-Test Result:
```

```
print(t_test_result)
```

```
##
##  Welch Two Sample t-test
##
## data:  developed_countries$new_cases and developing_countries$new_cases
## t = -12.764, df = 25221, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -22702.06 -16658.01
## sample estimates:
## mean of x mean of y
##  16601.62  36281.66
```

```r
# Check if the p-value is less than a chosen significance
#level (e.g., 0.05) for statistical significance
if(t_test_result$p.value < 0.05){
 cat("There is a statistically significant difference in new_cases
     between developed and developing countries.\n")
} else {
  cat("There is no statistically significant difference in new_cases
      between developed and developing countries.\n")
}
```

```
## There is a statistically significant difference in new_cases
##      between developed and developing countries.
```

Findings: Based on this t-test result, it can be inferred that there is a statistically significant difference in COVID-19 vaccine effectiveness between developed and developing countries, with the mean vaccine effectiveness being higher in developed countries compared to developing countries. Specifically, developed countries show significantly lower new cases compared to developing countries, with a mean of 16602.42 cases in developed countries and 36281.66 cases in developing countries, and a confidence interval of -22701.27 to -16657.21. This suggests that COVID vaccination effectiveness may vary between developed and developing countries, with developed countries showing a relatively lower number of new cases.

Conclusion

In this study, we aimed to investigate:

1. The factors associated with COVID-19 vaccine effectiveness across different countries
2. Predict the development status of a country based on these factors, and
3. Evaluate the vaccine effectiveness in developed and developing countries.

From the investigation the following conclusions were inferred:

1. Our findings revealed that several factors, including people fully vaccinated, GDP per capita, human development index, aged 70 and older, hospital beds per thousand, new deaths, and total vaccinations per hundred, showed statistically significant associations with vaccine effectiveness. However, the R-squared value of 0.52 indicated that only approximately 52% of the variance in new cases could be explained by the independent variables. This suggests that other factors not included in our analysis may also influence vaccine effectiveness.

2. Our results from the Random forest classification analysis showed that GDP per capita, human development index and people aged 70 and older were the most important predictors for determining the development status of a country. This suggests that these factors play a crucial role in differentiating between developed and developing countries in the context of COVID-19 vaccine effectiveness.

3. Although the analysis did not specifically examine the difference in COVID vaccination effectiveness between developed and developing countries, as originally intended, the t-test revealed a statistically significant difference in new cases between developed and developing countries, suggesting unequal vaccine effectiveness with more new cases seen in developing countries.

It's important to note that our study has some limitations:

1. The dataset used may have inherent biases and limitations in data quality, as it relies on reported data from different countries.

2. Our analysis is based on correlation, regression and statistical analyses, which cannot establish causation.Caution should be exercised in interpreting the findings.
3. The analysis focused on a limited set of variables, and other potential factors influencing COVID-19 vaccine effectiveness or country development status may not have been considered.

Despite these limitations, this study has important implications. The findings of this analysis indicate that COVID-19 vaccination is associated with a decrease in new cases. Also, The identified factors associated with COVID-19 vaccine effectiveness and development status of countries can inform policymakers and healthcare practitioners in designing and implementing effective vaccination strategies, allocating healthcare resources, and addressing health inequalities.