**Assessment 2 - Part 2**

**COVID – 19 Pandemic Data Analysis**

**Logo, company name

Description automatically generated**

**MATH 2032 – Statistics Using R**

Author: Jose Manuel Alvarez Gonzalez

Student Id: 110364322

Email Id: [alvjy005@mymail.unisa.edu.au](mailto:alvjy005@mymail.unisa.edu.au)

Date: 15/3/2022

**Introduction**

In the following project we are going to be working with two different data sets relating to the Covid-19 pandemic, the data sets are records of each country around the world. For the first data set, we have a time series report from the start of the pandemic in 2020 till the end of 2021. For this data set we will be working with the variables looking at the daily new cases, daily deaths and new vaccinations per day.  
  
The second data set, also looks at all the countries around the world, in this data set we can find statistical summaries for each country. For this data set we will be working with the variables looking at total population, GDP per capita and median age.   
  
For the propose of this project, we are going to be merging the information of both data sets and creating new variables. The propose of manipulating the data is to help us analyse each variable better and to be able to relate it with the different countries. During this project we will be using descriptive statistics to have numerical summaries for the variables. At the same time we will be using visual plots that will help to visualise the findings, we will be working with: Boxplot, Cumulative series, Density plot, Bar plot, Dot plot and Linear regression.  
  
During the merge of the two data sets, we might have to dismiss some countries as they are missing the information we require for the analysis. For the project, we will be dismissing the NA entries that are present on the variables we will be working.  
  
The analysis of this report will be divided on two parts:

- Part 1. Analyses 7 particular countries of the data set, we will be looking at the new cases and later we will compare them according to their population (per million)

- Part 2. Works with all the countries found on the data set, here we will be categorising the countries according to their GDP level and we will find how this relates to the total vaccinations. Then we will look at total deaths (per million) and try to find a relationship with the population median age of each country.

**Analysis**

**Part 1**

For this part of the project we will be working with 7 different countries: Australia, China, India, Sweden, Russia, United Kingdom, United States. In this part of the analysis we will be focusing on the daily new cases for the total period of the data set.  
  
First we are going to do a cumulative series on the new cases, we will start analysing each country from the moment they reached over a total of 1,000 cases. The idea is to be able to identify any patterns during this period. The following plot presents the information on a Log-10 scale:

Chart, line chart

Description automatically generated

Figure Cumulative series of new Covid-19 cases for 2020 and 2021

In Figure 1, first we can identify that China was the first out of the seven countries to accumulate over a total of 1,000 cases, for the remaining of the countries it happened at really similar times.  
  
Regarding the increase of cases, we can see that at the start of the pandemic, all of the countries experienced an exponential increase of the daily new cases, we can identify this finding by the step lines at the start of the graph. For this exponential increase to reduce, it took about two months for each country to control, each country settles at different total: just under 100,000 total cases for China and under 1,000,000 total cases for the USA.

Another interesting finding is that China was the only country that didn’t experienced another peak (increase in daily new cases), we can see that their slope slowly reduces, on the contrary we can see how Australia had two other peaks later on the time frame, similar to most countries.

Next, we will be manipulating the daily new cases according to one million people. For that we will need to use the population of each country.

First we will look at the descriptive statistics of this new variable on the following table:



Figure Descriptive statistics for daily new cases per one million people

The first unique finding is that the UK has a negative value, this might mean that during the data set it might had a correction. We will keep the information as its, we don’t have information from the data set according to negative entries.

From the summaries, the country that clearly sets apart from the rest is China, we can see that their distribution is right skewed and they have a high Kurtosis of 218, also the maximum and minimum range is only from 0 to 10, much smaller than any other country.

To help understand the distribution of the remaining of the countries, we will use a boxplot and a density plot. Here we can find that most countries have outliers: Median and IQR might be the best measures.   
  
In the plots we can confirm what we already found about China, the peak on their density plot and it’s median value is close to cero. This is quiet similar for Australia but here we can find many outliers, this says that they were good at controlling the spread during the total period but they also experienced different increases on new cases.  
  
India is the only country that can be compared on a population size to China, in their case, we can see that they did experience more impact on the pandemic. We can see that they have a median value of 19.53 cases per million. Still relatively small to the rest.

Sweden, has a median value of 33.2. But they have the highest range on new cases, with a maximum value of 3,197.29 and many outliers along the period. This generates more question in the way their data was collected.

Russia, USA and UK have the three highest median values of new cases, USA being the highest with a median value of 152.17 new cases per million ppl. In their density plots, we can see that the three countries also experience two peaks on the frequency of new cases during the total period.

Graphical user interface, application

Description automatically generated

Figure Boxplot for the distribution of new cases per one million people

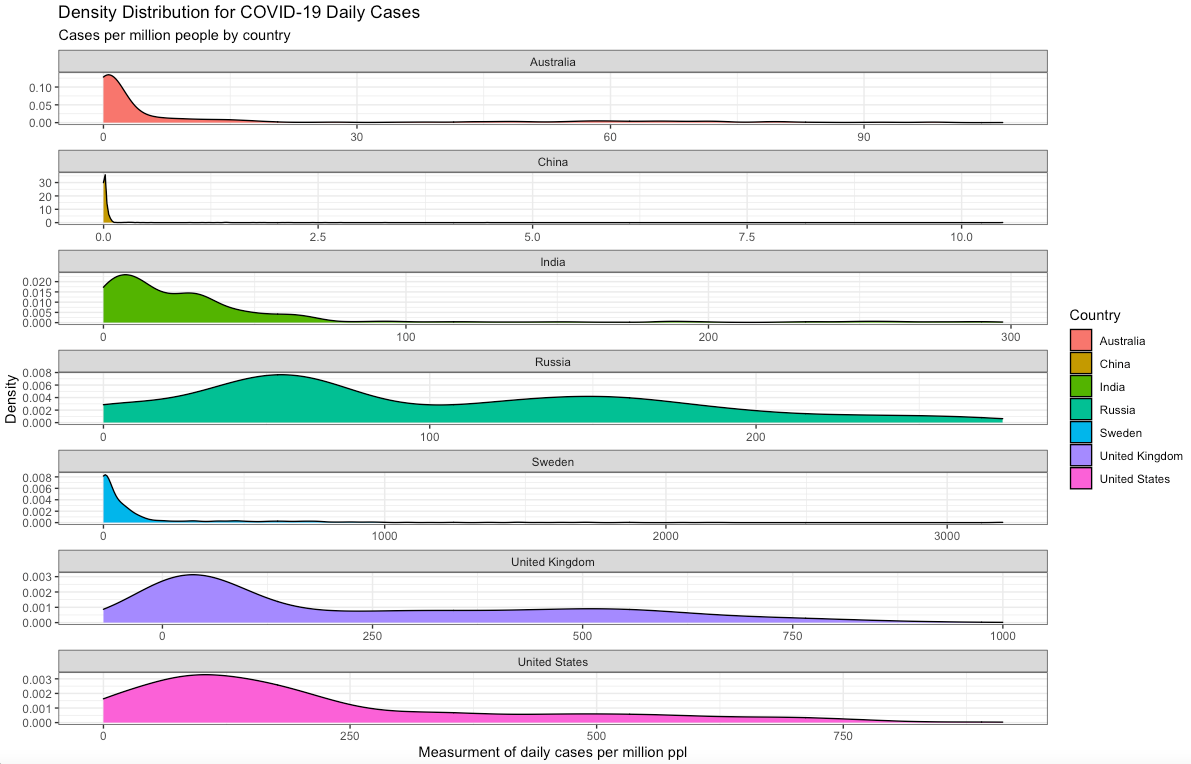


Figure Density plot for the distribution of new cases per one million people

**Part 2**

Now we will be working with all the countries from the two data sets. First we will analyse the total vaccinations according to their GDP levels.  
  
The first stage is to categorise all the countries according to their GDP per capita, for that we will have the following rules:

Rich – Top 25%

Average – Above 50%, but below the top 25%

Poor – Below 50% of the countries

Then we will create a new variable of the total number of vaccinations per one million people.

With this new data set, we can create a summary to understand each category, we will do this on two stages. The first stage will look at numerical values, focusing on the median value for the GDP per capita, population and median age, we are including median age as we will work with this variable later. Median age is also relevant to understand the demographic of each category.

Next, we will look at totals for each variable, starting with total vaccinations and total population, following with the total vaccination per one million people. For this part we will be using bar plots.



Figure Statistical summaries for each GDP category

Looking at the median values for each category we can start with the GDP, here we can find that the median value from the Poor GDP countries is 6,197.22, seven times less than the median value of 44,017.59 for Rich GDP countries. The difference on median values between Average and Rich is twice the value, from 20,422.08 to 44,017.59.  
  
Median population for each category is different, Poor is the category with the highest median value of 11,884,350. Rich countries have the lowest median population of 7,552,800. The difference here is not as significant as with GDP but we will look next at their total population in the plots.  
  
Regarding population median age, we can say that Poor countries have the median younger population from the three, with an age of 25. Rich countries have an older median age of 41.20.

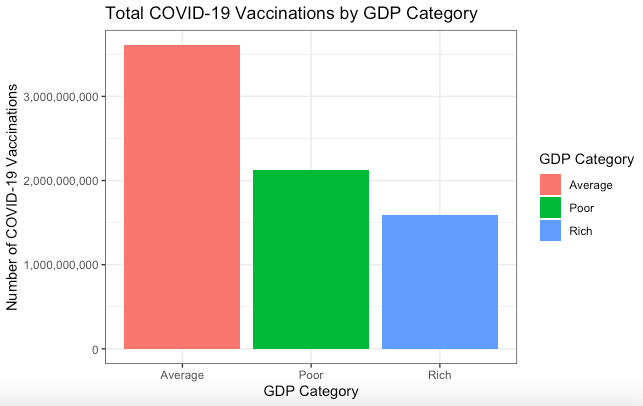


Figure Bar plot of total covid vaccinations according to GDP category

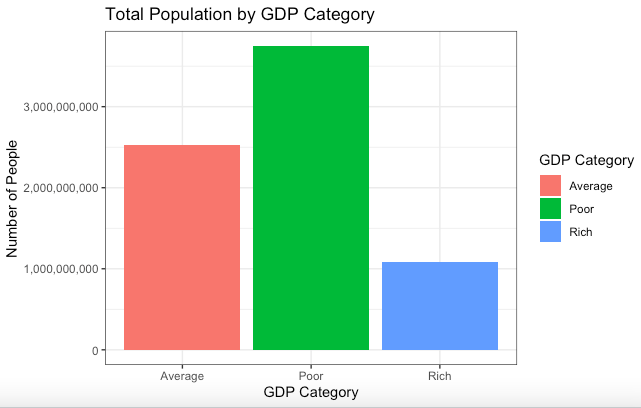


Figure Bar plot of total population according to GDP category

Here we can see that the total population of Poor countries is three times more than the one from Rich countries. Average countries have twice the total number of people than the Rich ones.

Average countries are ahead on the total number of vaccinations, not much of a difference between Poor and Rich. This is not the best way to compare the rate of vaccinations as we previously saw that in the total number of population exist a big difference between categories. For that reason we will be looking at the number of vaccinations per one million people.

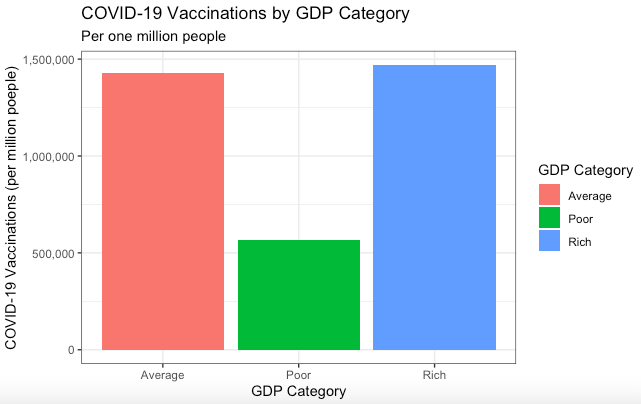


Figure Bar plot of total covid vaccinations per one million people according to GDP category

Here we can find a big shift on totals, now the Rich countries have the highest amount of vaccination with a total of 1’469,365.6 vaccines per one million people. Poor countries have almost three times less that amount with 565,841.9. On the other hand we can find really similar totals between Rich and Average countries, with Average countries having a total of 1’426,290 vaccines per one million people.



Figure Total summaries for each GDP category

For the final stage of this project we want to identify if is any relationship between the median age of each country and their total number of deaths per one million people. We start by plotting the variables for each country, we will have the total number of deaths on the Y axis, this is the variable we want to predict.

Chart, scatter chart

Description automatically generated

Figure Dot plot of each country Covid-19 total death and country median age

In the figure 9, I divided the plot on four quadrants according to the median value of each variable. I’m doing this to be able to identify an increasing relationship. We can see that it’s an increasing relationship as the Q1 and Q3 are the quadrants with more entries on the plot. To confirm this finding, we run a correlation analysis and we have a value of 0.56, the positive value says that we do have an increasing relationship and we can interpret that is a moderate relationship.

Then we create a regression model, with Total deaths as the variable we want to predict, we get the following equation:

64.705 \* Median Age – 1,015.075 = Total Deaths (per one million people),   
with an R squared of .31

Here we can find that when median age is 0, we have minus amount of deaths, this is due to the fact that no country have a median age of 0, the lowest median age is about 15. Then we can see that for each year the median age increases we also increase the total deaths by 64.705.   
  
The R squared value of .31, says that this is not a really good model to create predictions. The main reason affecting this, is the high amount of outliers we can find on the data. In Figure 10, we can see the same plot with a the regression line.

Chart, scatter chart

Description automatically generated

Figure Figure 9 Dot plot of each country Covid-19 total death and country median age with Linear Regression Line

**Conclusion**

After analysing the data, we can have different conclusions of each variable. First we start with the Covid new cases.  
  
We can see that from our chosen 7 countries: from their start of the pandemic, all of them experienced an exponential increase on the number of new cases. For all of them it took about 2 months to reduce that exponential increase, settling at different total cases.  
  
From the 7 countries, China is the only country that didn’t experienced a second increase on the number of new cases after it settled. At the same time this is the country that it got affected the least by the number of cases, based on their population size.

The USA is the country that experienced the highest median value of 152.17 new covid cases per one million people, during the total period. Sweden had the highest maximum value of 3,197.29 cases per million people in one day.

Categorising the countries according to their GDP levels, we can say that the Poor countries, have the highest total population and the youngest median age of 25. Their difference on median GDP per capita compared to the Rich countries is seven times less, with the correspondent values of 6,197.22 for Poor countries and 44,017.59 for Rich countries.

When it comes to vaccinations number per one million people, Rich and Average countries have a similar total, on the other hand Poor countries have almost a third less than the other two.

Looking at the relationship between Total Deaths per one million people and Median Age per country, we can find that is an increasing and moderate relationship with a correlation value of .56. Creating a liner regression model, confirms the increasing relationship, but the existence of high value outliers makes the model not a good prediction of total deaths, according to median age.