**Assessment 2**

**COVID – 19 Pandemic Data Analysis**

****

**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: 9/2/2022

**Introduction**

In this project we are going to work with a data set that analyses the spread of Covid in two different countries: USA and Russia. The data we are going to work, analyses the spread of the virus during a period of two years, from the first reported case for each country at the start of the pandemic in 2020, following with daily reports to the end of 2021.  
  
We are going to look at two different variables for each country and work with those findings to identify any pattern and compare them with the other country. The main two variables we are looking at are the daily new cases and the daily new deaths caused by Covid. With this information, we are going to create measurements of spread and dispersion for the totality of the period covered in the data set, is really important to recall that the measurements will be made for the totality of the findings and it won’t do a time series analysis as the spread evolved.

Once we collect all the data, we are going to see how the different variables (cases and deaths) evolved over time, we will compare the countries with each other, look for any patterns with the findings that will helps us understand better how the spread evolved. At the same time we are going to compare the two variables with each other and try to identify if is any relation between them.

**Analysis**

**Summaries**

The first step to understand the data we have ahead of us, is to find how the totality of it, is spread during the two years. For that we are going to look at two measurements of centrality for the amount of new cases and daily deaths. We are going to look at the median and mean of each distribution:

**Russia:**

New Cases New Deaths

mean 🡪 15,545 mean 🡪 448

median 🡪 12,673 median 🡪 390

**USA:**

New Cases New Deaths

mean 🡪 76,635 mean 🡪 1,221

median 🡪 54,117 median 🡪 979

We can see a relative difference between the two, the next step to decide which one we should use, is to calculate the Kurtosis and identify how well distributed is the data. The kurtosis for the countries:

**Russia: USA:**New cases 🡪 -0.58 New cases 🡪 0.60  
New deaths 🡪 -0.36 New deaths 🡪 0.81

The kurtosis for both countries is close to cero and no more than one total unit, this says that the data is well distributed and the mean will be a good value to use for our measurement of centrality.

Next we will see how the data is spread, as all our variables start with cero as a starting point, we can look at the maximum value of each one to understand our range on each variable. For that we have that the maximum value for each is the next:

**Russia: USA:**  
New cases 🡪 40,210 New cases 🡪 303,487  
New deaths 🡪 1,222 New deaths 🡪 4,442

To keep building our understanding of the spread for the data, now we will look at the skewness, this will give me an idea of where I can find most of the values. For that we have, that the skewness on each variable:

**Russia: USA:**  
New cases 🡪 0.59 New cases 🡪 1.22  
New deaths 🡪 0.70 New deaths 🡪 1.12

Looking at the skewness we can see that the occurrences of the values are relatively consistent towards their central point, but with their positive number we can see a relatively right hand side skewness, which means that is slightly more occurrences before the middle point of measurement.

**Outliers and Spread**

To continue with the measurement of spread, I decided to look for the outliers and use this point as a reference of spread for the totality of the two years period. As I mentioned previously, the data is consistently updating with each new day, each day poses a different range of how spread is the data according to that period. What I decided to do, is to look at the upper outliers values for the total period of the data. The outliers I used for that follow the formula 1.5 \* IQR + 3 Quartile.  
  
This values are going to be helpful at identifying at what moment of the period the variables went out of control, based on what we currently know are the highest maximum values reached for each variable. This will become more evident when we look at the plots for each variable and it will make it really easy to identify the particular periods when it was a high increase on values.

The outliers values to measure the spread for each country:

**Russia: USA:**  
New cases 🡪 45,031 New cases🡪 229,358  
New deaths 🡪 1,547 New deaths 🡪 3,577

From looking at these higher values measures of spread, we can already identify that Russia didn’t exceeded these values from their maximum measurements reached. These is contrary to the USA, we will look more in to these when we compare the different plots.

**Mortality Rate**

The last step in the analysis of our data is to look at the mortality rate, this is a new variable we will create by comparing the percent rate of deaths according to new cases. What we are doing here, is try to identify the total number of deaths for each new case. The way I will analyse these new data, will be like a distribution for the total period range of the data, exactly the same way as I analysed the data for the previous two variables. The reason to analyse the data this way, it will help me plotting the information during the time of the spread and we will be able to identify how the death rate changes over time.

First we will create a mortality rate for each new daily entry, then we will create the same summary measurements as we did with the previous variables. For that we have the following:

**Russia: USA:**

max 🡪 0.0510 max 🡪 0.2174

IQR 🡪 0.0178 IQR 🡪 0.0139

Median 🡪 0.0276 Median 🡪 0.0154

Mean 🡪 0.0273 Mean 🡪 0.0216

SD 🡪 0.0115 SD 🡪 0.0193

Kurt 🡪 -0.8382 Kurt 🡪 19.92

Skw 🡪 -0.1044 Skw 🡪 3.3793

\*As a note, is important to remember these values are rates or % of deaths per each new case.

Looking at the summaries we can find some really interesting findings, we will start with Russia. First we can see that they have a really normal distributed mortality rate, the centrality values of median and mean are almost the same, the skewness is almost cero, which says to us that the frequency of their rate is close to the central value. The Kurtosis is below cero, which says is not many tails and as a consequence it would not be outliers in their data.  
  
When we compare these information to the USA we have a completely different result, first we can see a big discrepancy on their centrality measurements. Following that we look at the dispersion, we start with the maximum range, we can compare this with Russia, the USA has a value four times higher than the one Russia does, this says that for at least one day the USA experienced a 21.74% mortality rate according to new infection cases. The next step to understand the data is to look at the kurtosis, this is incredible high, something we could already assume by looking at the range, we can confidently say that we will find some outliers on these data set. The next value we look at is the skewness, this value is also high but with positive value, this is actually a good result if we care about lower death rates frequency, what this skewness says to us is that we will find many more death rates occurrences below the centrality point.

Looking at the big differences found between the two countries is what made me choose this particular way of analysing the death rate and help understand how it changed during the spread period. The next step that we will follow to understand these measurements better is to look at the different plots.  
  
For the death rate we will look at the median as a measure of centrality and we will also look at the outlier maximum spread rate that we followed previously, this rate is only applicable for the USA and will have a value of 0.0456 – 4.56%

**New Cases Analysis**

Chart, line chart

Description automatically generated

Here we can see the spread of new cases during the two years for Russia and the USA, the first main point to recall for the future of the analysis, we need to subdivide the period for 2020 and 2021 as we can find really different findings on each year for both countries, particularly the USA. In this particular plot we can see that both countries suffered from different spikes at similar points in time, starting with the first quarter of 2020, the incline slope is really high for both countries. What is important to consider for this first spike, is that the severity of it is quiet mild comparing to the overall of the two year periods.  
  
The next important point to consider is on the reduction of cases during the middle of 2020, we can see that for Russia it was a smoother and quicker reduction on cases than the one for the USA, contrary from Russia, the USA suffered a second spike during the same period. The next important and consistent point for both countries is on the spike suffered at the end of 2020. For both countries this spike reached over their mean values of daily new cases, for the USA in particular, this is when the country suffered with the highest values reached over the total period, at the same time we can see that some days the values went over the outlier top ranges of new cases. We can confidently say that this has been the hardest period for the USA.

At the middle of 2021 we can see that both countries managed to control the spread and consistently maintain below their mean values of new cases, this was a good achievement for both countries, particularly for Russia as we can see a more consistent pattern and less spikes than the USA.  
  
The most important and last point to consider is the last half of 2021, here we can see that both countries suffered spike on new cases but really different results for the two. Starting with Russia we can see that during the period the daily new cases is consistently over the mean value and the daily new cases has kept above it, at the same time their spikes haven’t reached their maximum outlier range, which is important to consider that it has kept relatively under control.

During the same range for the USA we can see that the spike got to some higher values above the mean and also above their top outlier range, but at the same time we can see that the majority of their time most of the cases were under their mean value. Is quiet hard to fully see on the plot where the majority of occurrences are located, I decided to create an histogram for the USA frequency on new cases for the second half of 2021, here we can see that the majority of daily new cases are under the mean value.

Chart, histogram

Description automatically generated

**Daily Deaths Analysis**

Chart

Description automatically generated

Here we have the daily deaths during the two years of the spread, at first instance the plots look relatively similar than the daily new cases, this can says that is a relationship between new cases and daily new deaths. As we start looking with more detail at the analysis we can start finding some big differences, especially for the USA.  
  
For Russia, the first year is really similar than the one found for new cases, they suffered a spike during the first quarter of 2020 but this spike didn’t go over their mean value. They suffered a similar spike at the end of the year and here we can find that the daily deaths started raising over their mean value, but at the same time the strength of the spike is not as strong as for the totality of their period. They managed to control the spike during the first half of 2021 but the situation completely changed for the second half of 2021, here we can see a really steep incline on the amount of new deaths, it’s a really big increase on the numbers but at the same time is important to consider that is a consistent incline and it didn’t get over their top outlier range. What we can find is that for this second half of the year, they haven’t been able to reduce the amount of daily deaths, and the numbers are consistently increasing at a relatively steep level.

For the USA we can see more of a roller coaster of spikes, contrary to Russia, the USA suffered a big spike on the amount of daily deaths on the first quarter of 2020, from a relatively early time they managed to get values above their mean for daily new deaths. For the end of 2020 and consistent with daily new cases, we can see that they had the highest values during this period and their values reached above their top outlier range.

During the start of 2021 and similar to Russia they managed to control the amount of deaths and reduce the numbers of new deaths, but contrary to Russia, the USA managed to maintain the amount of new deaths relatively stable and under their mean value for the rest of the year. We can easily identify that the USA similarly to Russia, suffered an increase of new deaths during the second half of 2021, with some values above their top outlier range, but the frequencies of this occurrences were quiet low. Opposite to Russia, the USA managed to maintain these rates of occurrences below their mean value and actually kept the majority on new deaths well under the mean for the majority of the year.   
  
You can see on the histogram of the second half of 2021 for the USA, how the majority of new deaths occurrences are under their mean value. This is a good achievement for them, especially if we compare it with the start of the previous year.

Chart, histogram

Description automatically generated

**Mortality Rate Analysis**

Chart, scatter chart

Description automatically generated

In my opinion this is the most important part of the analysis, as here we can identify how the two variables affect each other. We could come up with a quick conclusion and just look at the final death rates for each country 2.73% for Russia and 1.54% for the USA, this is clearly a better result for the USA. But just looking at the final result doesn’t tell the full story, I find is really important in this case to look at the evolution on how the death rate changed for the two countries.   
  
First we can identify that for Russia 2020 and 2021 are completely the opposite to each other. On the first year we can see how they managed to keep the death rate under their mean, really good achievement even with the early spikes. For the second year we can see that for the vast majority of occurrences, the death rate is above the mean, with some ups and downs through the year, it will be interesting to further study the factors that are not allowing the death rate to reduce, maybe the higher volumes on new cases is affecting the possibility of the death rate to reduce. At the same time is important to notice that the death rates never went over a top outlier range.

For the USA we can subdivide the data for the first half of 2020 and then the remaining of the period. First we can see how the death rates were above the mean from the beginning of the pandemic, with many occurrences well above the top outlier range, this was by far the deathliest period of the pandemic on terms of death rate for the USA. Contrary to Russia, after their first spike on death rates, they have managed to control their rate through the remaining of the pandemic, we can see the rate going up and above the mean but with the vast majority of the frequencies close to the mean value.

**USA New Cases vs New Deaths**

Chart, scatter chart

Description automatically generated

Here we can find the relationship between new cases and new deaths for the USA, we can clearly identify that is an increasing relationship between the two. We could subdivide the plot in to four quadrants and we will find that the vast majority of the points are in between the first and third quadrants, this means that is an increasing relationship between the two variables. At the same time I decided to draw a linear model to represent the increasing slope and relationship between the two.

**Conclusion**

After looking at the historical data and summaries of the Covid spread for the USA and Russia we can conclude that 2020 and 2021 were different outcomes for the two countries.

First we can note that is three main spikes during the two years for both countries: first quarter of 2020, end of 2020 and second half of 2021. It affected the countries on similar ways, but the outcomes on the long term are different for each country.

We can say that for the first year Russia did a really good job at controlling the spread and the number of deaths, sadly the second year with the higher increase on cases has made it more difficult for them to reduce the rate of deaths for each new case, they haven’t been able to reduce the number of new cases and deaths below their middle points.  
  
For the USA we can say that the first year of the pandemic was their most difficult, with higher values on deaths and their respective rates. For them the big consideration is to remark how well they managed to change their situation and actually control better the spread for the second year.  
  
As a conclusion we could say that Russia was better at fighting and controlling the virus at the start of the pandemic but this control has been difficult to keep on the long term. For the USA the virus affected the most at the start and they were able to change the situation and control it better on the long term.