Big Data Engineering and Architecture

COVID-19 Pandemic Analysis in Apache Spark

Ansh Sikka

Introduction

The 2020 novel Coronavirus pandemic has gripped the world in a matter of weeks. As of the beginning of May 2020, 3.53 million people are affected with over 248 thousand dead. It is important to look at the underlying factors that contributed to such an enormous infection rate worldwide.

COVID-19 Analysis Plan

There are many factors that contributed to the high amount of cases during the pandemic. A few of the main factors include population, healthcare, and travel. Although it may not give a complete (and completely accurate) description on why/how the pandemic is growing at the pace it is, it will give us some type of insight to make future decisions. Performing descriptive and prescriptive analysis on coronavirus case/death, population, healthcare, and travel data would possibly influence national authorities worldwide to make wise decisions based on the recent data during the pandemic or in future pandemics.

Goal: *See relationships between demographics & travel rates of a country vs. the extent of the pandemic.*

We will ask a few questions about how related the number of cases and deaths are based on population, healthcare, and travel. We will additionally examine where the world is headed with the pandemic. The questions fall under 4 categories: quantity, quality, spread, and future lookahead. There are also hypotheses to show initial predications. ***However, since a lot of the data is so recent and it is being compared to more historical data, the outcomes that support/reject the hypothesis cannot be completely accurate.*** The questions are shown below:

Category	Question	Hypothesis
Quantity of People	Does the population of a country affect the number of confirmed cases?	As population increases, so does the number of cases. As a population increases in a certain area, the higher

		the chance that people are closer to each other, causing more spread.
Quantity of People	Does the population of a country affect the number of confirmed deaths?	As population increases, deaths would increase too since more people sick would strain hospitals, and death rates would be higher.
Quality of Healthcare	Does health expenditure of a country affect the number of confirmed cases?	As health expenditure increases, the number of confirmed cases wouldn't see a significant change, since there is no vaccine for preventative measures, just social distancing, which is the responsibility of the people.
Quality of Healthcare	Does health expenditure of a country affect the number of confirmed deaths?	As health expenditure increases, the number of deaths would decrease. This might be due to the fact that healthcare systems may have better capacity to take care of the ill.
Spread	Does country arrival amount affect the number of confirmed cases?	As the number of arrivals increases, so does confirmed cases. More traveling leads to more spread since interaction is worldwide.
Spread	Does country arrival amount affect the number of confirmed deaths?	Just like the cases, the number of deaths would also increase.
Lookahead	Have we hit our peak? Is the rate of new infections decreasing?	Looking at areas that are reopening give a pretty good idea that the rate of infection has decreased.

Data and Methodology

3 datasets are being used in the analysis. The bolded columns are filtered out and used for analysis. The columns that are highlighted in yellow are the ones that join the datasets together.

Global Coronavirus (COVID-19) Data (Johns Hopkins)

The dataset contains 18 columns. The ones that are bolded are the ones that we will use for our analysis. Null rows are dropped, and the dataset was split into two separate data frames: Cases and Deaths

COLUMN NAME	DESCRIPTION
CASE_TYPE	**Type of Case: Death or Case**
NUMBER_OF_CASES	**Point in time snapshot of to-date totals**
DATE	**Jan 23, 2020 - Present**
COMBINED_KEY	Full Name of Country
COUNTRY_REGION	Name of Country/Region
PROVINCE_STATE	Province/state name
ADMIN2	County Name
ISO2	2-Digit Country Code

ISO3	**3-Digit Country Code**
FIPS	5-Digit Federal Info Processing Standard
COMBINED_KEY	US only- Combo of Admin2, State_province, and Country_region
LAT	Latitude
LONG	Longitude
POPULATION_COUNT	**Number of people in country**
PEOPLE_TOTAL_TESTED	Number of People Tested (There were 0 values in this column)
PEOPLE_TOTAL_HOSPITALIZED	Number of people hospitalized (There were 0 values in this column)

Source: https://data.world/covid-19-data-resource-hub/covid-19-case-counts

International Tourism, Number of Arrivals

The data contains 64 columns. The ones that are bolded are the ones that we will use for our analysis. The only issue that might affect accuracy in analysis with this data is that it only goes up to 2018. However, we can make an assumption that the popular destinations for tourisms/arrivals haven't changed that drastically in the past 2 years. Null rows are dropped.

COLUMN NAME	DESCRIPTION
COUNTRY NAME	Name of Country
COUNTRY CODE	**3-Digit Country Code**
INDICATOR NAME	All were international arrivals
INDICATOR CODE	All were ST.INT.ARVL
YEARS COLUMNS (1964-2018)	**Used 2018 Column: Number of Arrivals**

Source: https://data.worldbank.org/indicator/st.int.arvl

Current Health Expenditure Per Capita

The data contains 64 columns. The ones that are bolded are the ones that we will use for our analysis. The only issue that might affect accuracy in analysis with this data is that it only goes up to 2016, so further representations may not be completely accurate. Null rows are dropped.

COLUMN NAME	DESCRIPTION
COUNTRY NAME	Name of Country
COUNTRY CODE	**3-Digit Country Code**
INDICATOR NAME	All were international arrivals
INDICATOR CODE	All were SH.XPD.CHEX.GD.ZS
YEARS COLUMNS (1960-2016)	**Used 2016 Column: Health Expenditure ($)**

Source: https://data.worldbank.org/indicator/SH.XPD.CHEX.PC.CD

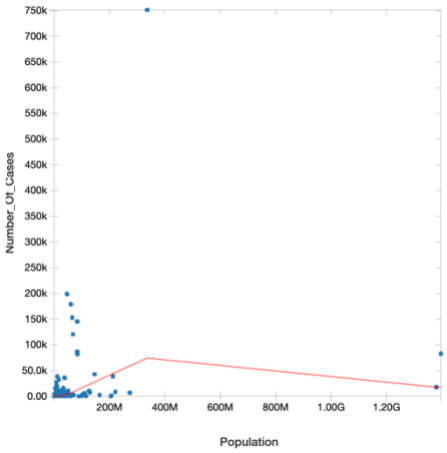
There were a number of manipulations performed on the data to create the visualizations and produce statistics.

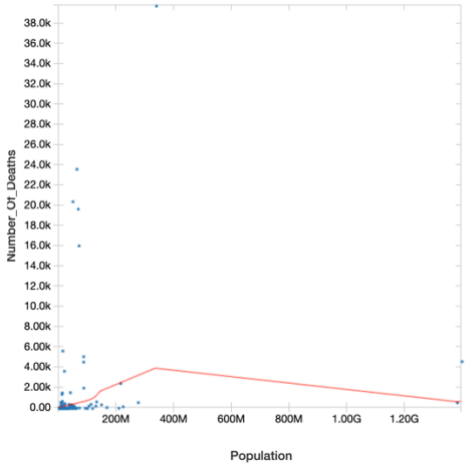
Question	Manipulations
Does the population of a country affect the number of confirmed cases?	• **Filter** dataframe for number of cases and deaths to the latest data (04/19/2020)

	• **Aggregate** number of cases (sum) and population (sum) for each country
Does the population of a country affect the number of confirmed deaths?	• **Aggregate** number of deaths (sum) and population (sum) for each country
Does health expenditure of a country affect the number of confirmed cases?	• **Join** health expenditure dataframe with COVID-19 case dataframe on country code.
Does health expenditure of a country affect the number of confirmed deaths?	• **Join** health expenditure dataframe with COVID-19 deaths dataframe on country code.
Does country arrival amount affect the number of confirmed cases?	• **Join** arrival count dataframe with COVID-19 cases dataframe on country code.
Does country arrival amount affect the number of confirmed deaths?	• **Join** arrival count dataframe with COVID-19 deaths dataframe.
Have we hit our peak? Is the rate of new infections decreasing?	• **Aggregate** number of cases (sum) by date. • Set up a **window** function to order by date • Set up a **lag** function that computes the column from 1 day earlier based on the window function on date. • Add a **new column** that computes the difference in the number of new infections from the previous day.

Summary of Findings

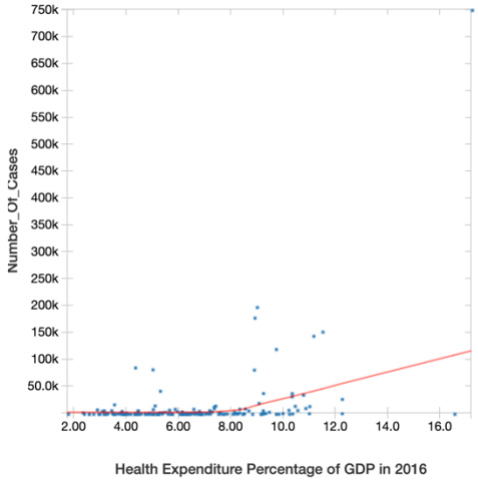
1. *Does the population of a country affect the number of confirmed cases and deaths?*





As we can see, there is a very small positive correlation between population size and the number of cases (*0.236)*. Additionally, there is even a smaller positive correlation between population size and number of deaths attributed to COVID-19 (*0.19).* Looking at the graphs, we can see a lot of countries with higher populations don't have that many cases or deaths either. However, what is that one outlier on the top of the 200M-400M population size? You guessed it, that's the USA. We can see that the USA is a breeding ground for this. From this, **we cannot accurately predict the number of cases or deaths by the country's population**. This is due to the fact that it is too soon to get accurate data and sufficient testing in high population countries. The true number of cases would also depend on different factors like availability of testing (this is where the USA leads), healthcare system, tourism, etc.

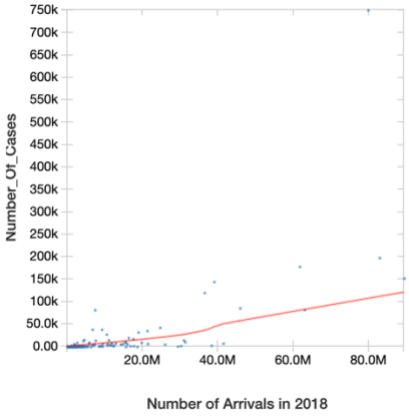
2. *Does healthcare expenditure affect the number of cases and deaths?*

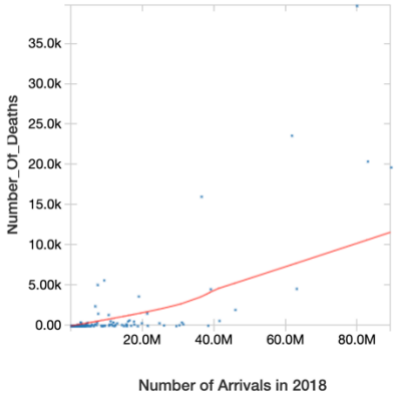


We can see the number of cases tends to increase with the health expenditure percentage. This might be due to outdated data. However, we can assume that healthcare expenditure won't

always affect the number of people infected. A possible explanation for this might be the fact that most people get treated for the virus back at home. I better way to view the effect would be looking at the death rates. This way we can see how well the healthcare system would respond to these sick patients. We can see that there is a positive correlation (0.40) for both case rates and death rates. This means that as health expenditure tends to increase, so does the rate of infections and rate of deaths. Some reasons why this may be is that people take a society's health system for granted and don't monitor their health as much. **We cannot be completely accurate about the data since the health expenditure data is from 2016.** There are additionally many other factors of a healthcare system in a country that can influence this data.

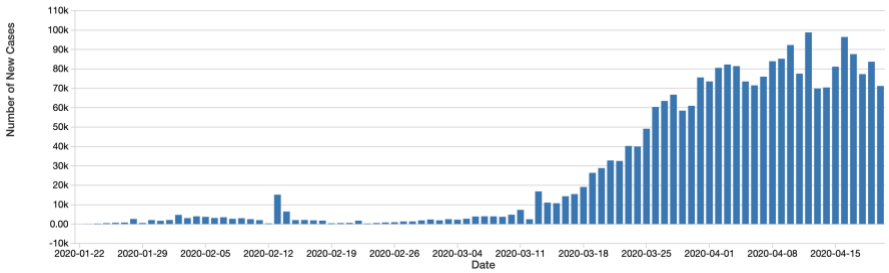
3. *Does country arrival amount affect the number of cases and deaths? (Tracking Spread)*





From the travel data, it is clear the correlation is higher than the other factors. A correlation coefficient of 0.67 and 0.75 for cases and deaths respectively based on average arrivals means that there might be some kind of pattern in travel and disease spread. We can't be completely accurate since the travel data is from 2018, but popular tourism destinations don't change as easily. So far, we can see that countries with high populations don't really affect the number of cases as much (again, there are a bunch of other factors). However, we can see that **possibly the movement of people throughout the world has an effect on the spread**. This can be useful for future pandemics because authorities can halt travel to certain popular destinations immediately to slow down the spread. This is extremely important since travel wasn’t banned right away as the epidemic in China and Italy started. It felt as if it was an unplanned ban that came too late.

4. *Have we hit our peak?*



This is a good way to see how the curve is actually responding to the pandemic. Looking currently, we can see that the new number of cases is actually starting to decrease, giving us an optimistic view. We can't confirm though, since this can just be a temporary dip.

Conclusion
As we can see, a lot of factors can influence a pandemic. The best thing to do is take preventative measures. The biggest pattern seen was from the tourism data, since a lot of disease *did* spread through travel. It was also surprising to see that things like travel and health expenditure didn't have as large of an effect than people assume. Then again, the velocity of this data for healthcare and travel isn't the best. In healthcare, many policies have changed since 2016. Popular travel destinations change more slowly, so those can be significant in analysis. Furthermore, using historical data puts some kind of control into predicting and taking measures into what governments should do to stop the spread of disease.

Takeaways and Challenges of Apache Spark

Takeaways

- The lazy evaluation model in Spark made execution faster since the queries got optimized as they were built, and runtimes were more efficient based on previous code.
- Spark provides ways to clean trivial data with ease. Functions like `df.na.drop()` made it easier to filter out null rows.
- Even though inferring the schema shouldn’t be used in production environments, it makes it easier to explore the data, especially with a lot of columns
- Along with the lazy evaluation, the `explain` function helped test out what code would work more efficiently and in what orders. It allowed for a lower-level view of the operations being performed on the data.
- Immutability is a very powerful concept as it made every output deterministic. There was no need to overwrite dataframes since I was able to use them repeatedly if needed. It is also thread safe. New dataframes were created every time a transformation was made, which made it easy to test.
- The ability for Spark to accommodate to simple SQL didn’t bring any extreme learning curves. Window and lag functions were similar.
- The statistical functions of Spark are extremely useful and intuitive since it just required column names and a function, no extra transformations.

Challenges

- Finding data for a pandemic during the pandemic → Data is updated every day and aggregating it against historical data will inevitably lead to inaccuracies.
- There were some small confusions between the cluster filesystem and Databricks filesystem.
- If one is coming from a language that allows mutability, the concept of immutability takes getting used to. There were times where it would make sense to reassign values, but it didn't work.
- The `explain` function was useful, however it became overwhelming as more transformations were applied.
- Scala has a slight learning curve with how multiple operations are performed on dataframes. For example, the fact that one can add `df.<function1>.<function2>` made it tricky to see any intermediate transformations.

As more data is produced from the current pandemic, using tools such as Apache Spark for analysis and prediction will only become more powerful. In the future, authorities and scientists would be able to use these analyses to make sufficient preparations and save more lives.