Nico Kaiser, Pascal Hurni, Pascal Neiger
nico.kaiser@bluewin.ch
pascalyveshurni@gmail.com
pneiger@bluewin.ch

Data Science Project

Covid-19 and Market Sentiment

Conceptual Design Report

16th October 2020

Abstract

This project aims to analyze the global Covid-19 data made available by the European Centre for Disease Prevention and Control (ECDC) for the ten biggest economies in the world (by GDP in 2019) and the impact of Covid-19 numbers on the leading stock market indices for said economies in the interval of the 1st of October 2019 to the 30th of September 2020. Our results show a significant correlation between rising global Covid-19 numbers (cases and deaths) and stock market performance in the first quarter of 2020, however not thereafter. Also, correlation between local Covid-19 numbers and stock market performance is generally weaker than between global numbers, most likely reflecting today's globalized economy.

Table of Contents

1 Project Objectives

A general assumption in financial market theory is that the market valuation of a company reflects the public sentiment about the future prosperity of that company, or in other words its discounted expected future earnings. It is therefore reasonable to assume that the leading stock market indices for a country's economy, comprising the stock price movements of that country's highest valued companies, are good indicators for the public sentiment about the future prosperity of that whole economy.

With this project we try to explore the impact of the rise and fall in Covid-19 cases and deaths on the public sentiment of the world's 10 strongest economies (based on their GDP in 2019, see Table 1). More generally, we try to investigate whether there are indicators that the general public believes the Coronavirus has a long-lasting effect on economic prosperity of these countries or whether this global pandemic is not expected to have a large impact on future economic performance.

Table 1: Strongest economies and their leading stock market indices

Country	Stock index	RIC
United States of America	S&P 500	^SPX
China	SSE Composite	^SSEC
Japan	Nikkei 225	^N225
Germany	DAX	^GDAXI
India	BSE Sensex	^BSESN
United Kingdom	FTSE 100	^FTSE
France	CAC 40	^FCHI
Italy	FTSE MIB	^FTMIB
Brazil	IBOVESPA	^BVSP
Canada	S&P/TSX Composite	^GSPTSE

There are certainly some caveats to such an analysis since stock market performance (and public sentiment) is impossible to predict and explain accurately due to the inherent complexity of human psychology and decisions. However, Covid-19 has certainly been a relevant piece of the puzzle when talking about public sentiment and economic outlook in 2020 and investigating the relationship between its development and that of major economic markets seems like a good point to start.

2 Methods

a) Infrastructure:

The data analysis will be done on a Lenovo Legion 5, a Lenovo ThinkPad and a Lenovo Legion 17, all running Microsoft Windows 10 Enterprise as their operating system.

All code has been tested on the Lenovo Legion 5 using a AMD Ryzen 7 4800H with an x64bit architecture. As the results were consistent, we assume that the code is simple enough to works relatively independent of the chosen workstation.

b) Tools

We will be using the Jupyter Notebook as provided by the platform Anaconda (version 2020.07; conda version 4.8.3) with Python 3 (Version 3.8.3final.0). The version of the jupyter notebook server is 6.0.3. It is running Python 3.8.3 (default, July 2 2020) [MSC v. 1916 64 AMD 64]. IPython 7.16.1.]

c) Modules and Libraries in Python

- Pandas is used for handling dataframes and data manipulation.
- Numpy and SciPyfor for calculating with data and tables
- Matplotlib for plotting and visualization of data.
- Tabulate is used to produce reader friendly tables.

d) Statistics

The data is structured as a time series. Descriptive statistics is used to describe and visualize the data. E.g. are scatterplots to quickly have an initial impression. Then inferential statistics such as linear regression is used to further assess correlation between the analyzed data. E.g. In order to assess correlation between the stock indices and the various clustering of Covid-19-Cases, we check the corresponding r-values (Pearson’s r).

3 Data

This project is based on Covid-19 data from the ECDC [1] and historical financial data from yahoo.finance [2] for the period of the 1st of October 2019 to the 30th of September 2020. The different indices can be found on yahoo.finance using the Reuters Instrument Code (RIC) listed in Table 1.

3.1 Covid-19 Data

ECDC provides daily updates of new reported cases of COVID-19 by country worldwide, as they are collected from health authorities worldwide [3]. This data set contains the following attributes for every country that has had incidents of Covid-19 starting from the 31st of December 2019 (the first reported cases in China):

- Date
- Day
- Month
- Year
- Cases (on that day)
- Deaths (on that day)
- Countries and Territories
- geoID
- Country Code
- Population in 2019
- Continent
- Cumulative Cases for 14 days per 100'000 population

Relevant for our purposes are Date, Cases, Deaths and Countries and Territories. In order to make it easier to work in combination with the stock market data, one data frame was created for each country and the files were reindexed with a date index from 2019-10-01 to 2020-09-30. NaN in Cases and Deaths were treated as 0, as the reflected the respective periods before the first cases were reported. Figure 1 below shows an example of the last 10 rows of the data frame created for the USA, reindexed with a date index.

Table 2: Tail of the Covid-19 Data Frame in the USA

	cases	deaths	country	geoID	countrycode	population	continent	cases_per100k_2_weeks
2020-09-21	39852.0	251.0	United_States_of_America	US	USA	329064917.0	America	160.574091
2020-09-22	53153.0	372.0	United_States_of_America	US	USA	329064917.0	America	169.357464
2020-09-23	38307.0	926.0	United_States_of_America	US	USA	329064917.0	America	172.756490
2020-09-24	37930.0	1102.0	United_States_of_America	US	USA	329064917.0	America	174.580750
2020-09-25	44213.0	901.0	United_States_of_America	US	USA	329064917.0	America	176.618646
2020-09-26	55013.0	964.0	United_States_of_America	US	USA	329064917.0	America	178.731299
2020-09-27	45368.0	723.0	United_States_of_America	US	USA	329064917.0	America	180.113397
2020-09-28	36248.0	259.0	United_States_of_America	US	USA	329064917.0	America	180.835747
2020-09-29	32998.0	314.0	United_States_of_America	US	USA	329064917.0	America	180.275675
2020-09-30	43017.0	928.0	United_States_of_America	US	USA	329064917.0	America	177.705969

Figures 2 and 3 show the cumulated development of cases and deaths per country, normalized by 100’000 persons of the respective populations.

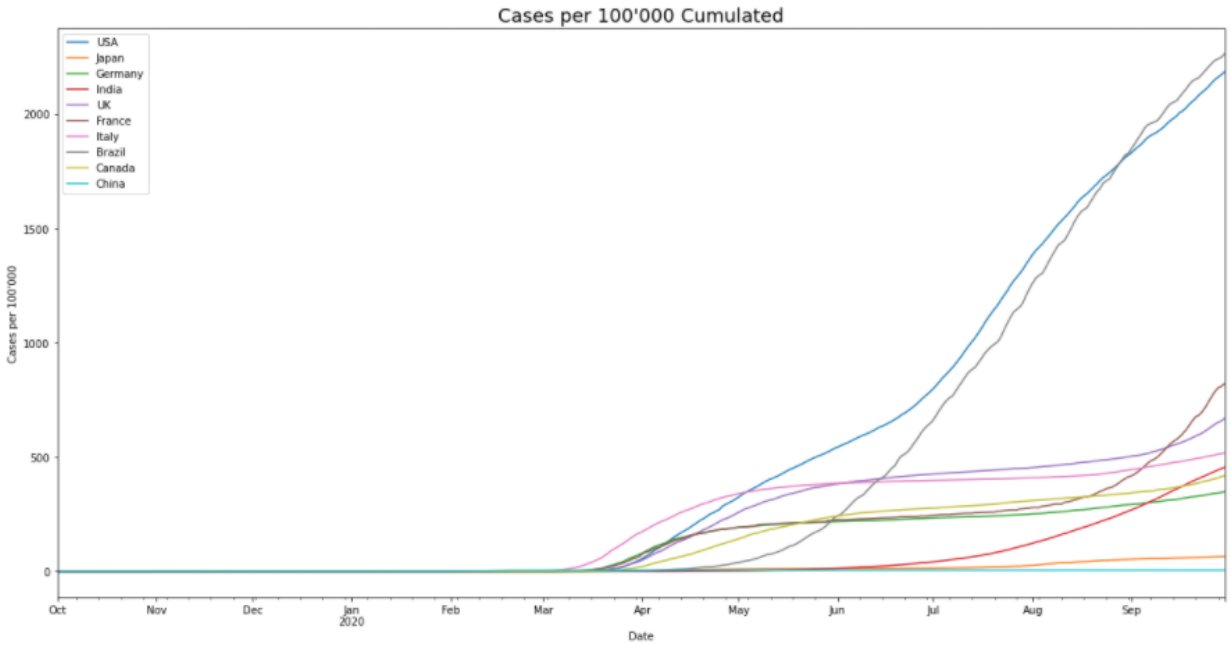

Figure 2: Cases per 100’000 Persons of the Respective Countries Population

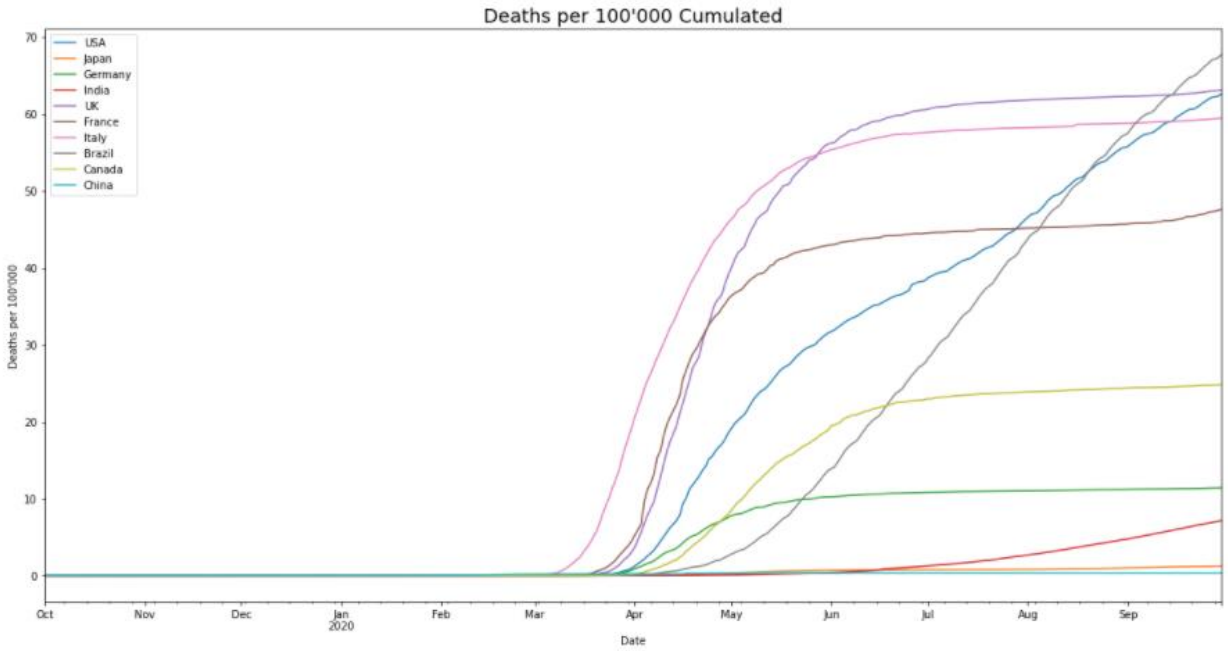

Figure 3: Deaths per 100'000 Persons of the Respective Countries Population

3.2 Historical stock market data

yahoo.finance provides public data sets of stock market prices for free. Periods can be freely defined. The files are separated for each individual index and contain the following attributes:

- Date
- Open Price
- Daily High
- Daily Low
- Close Price
- Adjusted Close Price
- Trading Volume

Only Date and the Adjusted Close Price are relevant for our purposes. The adjusted close price is preferable over the close price since it also factors in events such as dividends or stock splits, Thus they represent a better fitting representation of the actual value of the index at day's end for our purpose.

Again, one data frame was created for each country and reindexed with a date index from 2019-10-01 to 2020-09-30. In order to allow for more accurate correlation values, the time index has been shifted by -1 day (so for example, the close price for the DAX on the 1st of May 2020 has now been shifted to the 2nd of May 2020 in the data frame). This has been done in

order to correlate Covid-19 values from one day prior with current stock market performance, since Covid-19 numbers are usually communicated for one day prior and not in real time. On weekends and bank holidays no prices are reported as the is no trading. We interpolated NaN values to be able to correlate the stock prices with the reported Covid-19 data.

Figure 4 shows all leading indices normalized to 1 on the 1st of October 2019 in order to show relative movements irrespective of the total value of the indices.

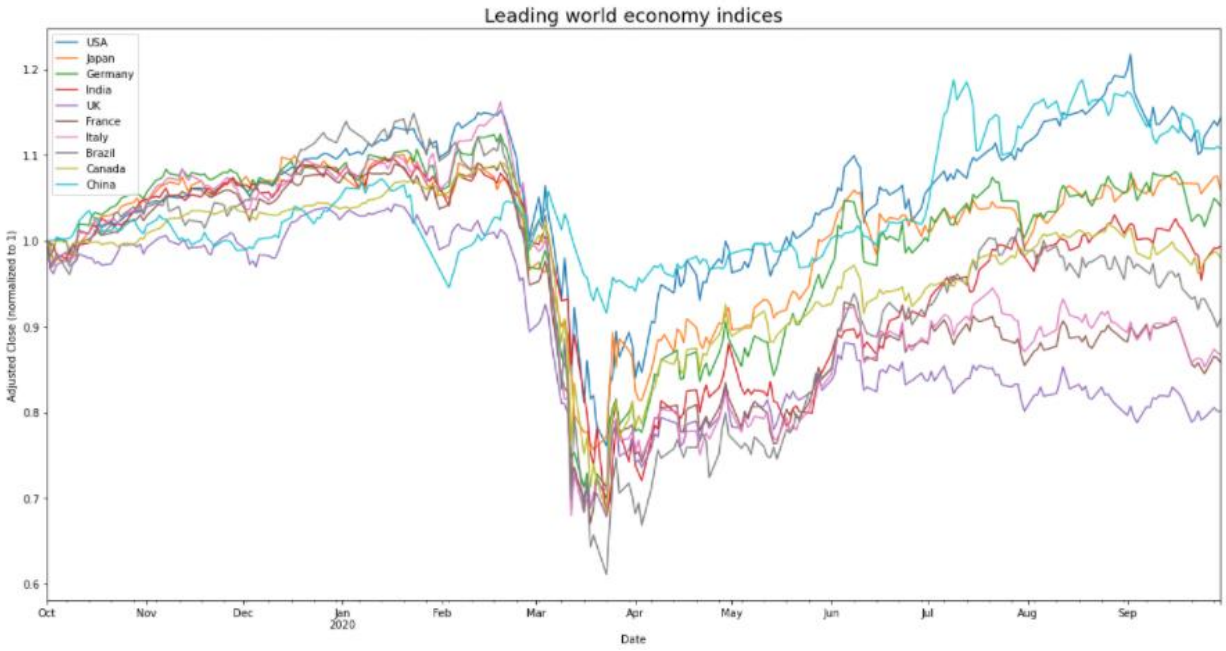

Figure 4: Development of the stock market indices (normalized per first of october 2019)

Comparing this to figures 2 and 3, it is apparent that a strong sell-off took place in all major economies (with the exception of the Chinese SSE Composite) between March and April, when Covid-19 cases and deaths steeply increased in many of the major economies. Most markets subsequently recovered in the months from May to October when Covid-19 cases and deaths decreased (in most countries).

4 Metadata

The dataset for the coronavirus data is available on the homepage of the ECDC and can be downloaded as a CSV file. The different datasets containing the historical stock index data can be downloaded in as CSV format for the desired period from yahoo.finance. Metadata for reproduction of our results has been stored in two separate readme files ("readme Covid-19 data" and "readme leading indices data"). The readme files together with the Jupyter Notebooks, CSV files and the current report are stored on the github repository.

5 Data Quality

The Covid-19 data as well as the stock market index data stem from reliable and official data sources. The data sources were selected based on how easy the data is accessible. Both data sources are always accessible and since we analyze historical data, all results are fully reproducible. To analyze data with a granularity of single days, both data sources are very well suited, since the data is already formatted on a single day basis. Further, both data sources provide data for all countries. As a third data source, a list of countries and their corresponding stock market indexes with their full names need to be passed to the analysis script. Nevertheless, some points need to be accounted for concerning the data manipulation and analysis.

5.1 Covid-19 data

While the Covid-19 data has daily data entries, the data frame doesn't have a standard date index. Luckily, the data set contains the year, month and day as separate columns, from where it is possible to create a standard date index for the upcoming time series analysis. The data frame has entries for all days, since potentially missing data, in example cases or deaths, is already substituted with zeros.

5.2 Stock Market data

The Stock Market data already has a standard date index and therefore needs no further index processing. In contrast to the Covid-19 data, the Stock Market data doesn't have data entries for all days, since the Stock Market is closed during weekends and some special holidays. To achieve completeness of our data so we can compare it to the daily Covid-19 data, missing entries were interpolated linearly.

Additionally, both data sources can be accessed via Web UI and API interface, which allows a higher automatization of the code.

6 Data Flow

The Covid-19 data contains date data (year, month and day in separate columns), geographic data (country name, country abbreviation, country code, continent) and Covid-19 related data (cases, deaths, population, cumulative cases per 100'000 for 14-days periods).

The Stock Market data contains the absolute value of a share at opening and closing of the stock market, its daily highest and lowest value, the adjusted close value, and the absolute volume of shares.

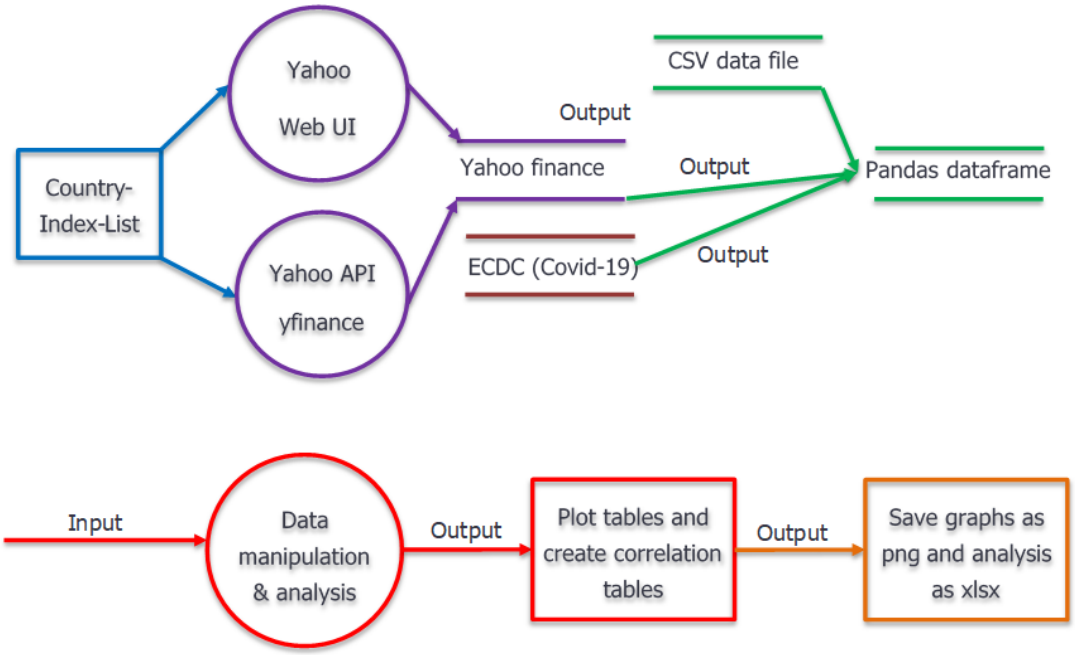

Figure 5: Data Flow model

One risk is that the chosen observation period has no data on the first or last day of the Stock Market data. Therefore, the selected observation period needs to be manually validated preliminarily. In the same way the stock market index abbreviation needs to be validated which is passed to the yahoo API.

Since we are only working with pandas data frames, there is no risk of loss of data while manipulation and analysis.

7 Data Model

In the following sections, the different data models for our data analysis project are explained.

7.1 Conceptual model

With the Country-Index-List and a predefined observation period we have all necessary information to collect the data from yahoo finance and the ECDC website. This data is the imported into python, where the manipulation and the analysis is done. The output is formatted to be used in everyday data visualization programs (i.e. excel files, image files as png).

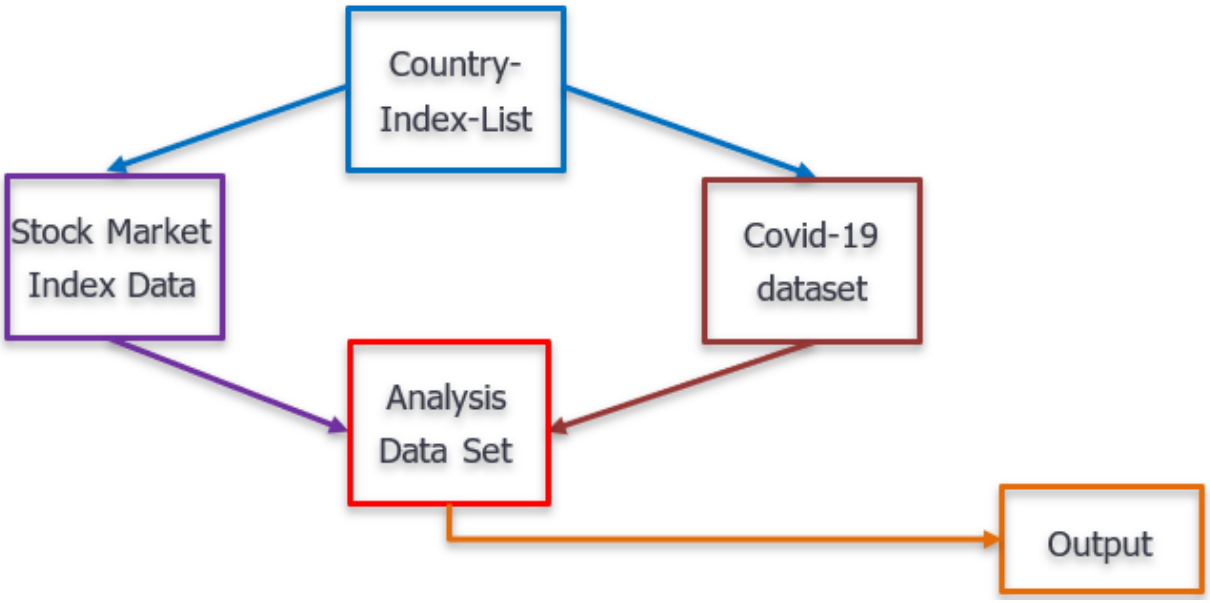

Figure 6: Conceptual Data Model

7.2 Logical model

The Covid-19 data from the European Center for Disease Prevention and Control will be analyzed separately for daily cases and deaths. To compare the Covid-19 data, we selected the adjusted close value, which is a corrected share value. To compare the data, we selected the pearson correlation for a first overall comparison, and in a second step we calculate weekly and quarterly correlations. Additionally, the skewness (second moment) of the data will be compared so that we also have test statistics with a relative measure (cp. Table 3).

Table 3: Description of the data frames

Parameter\Data	Country-Index-List	Stock Market Data	Covid-19 data set	Analysis data set	Output
Data structure	Table	Table	Table	Table	Table/Graphs
Data set organization	One data frame as csv	One data frame per country index as csv/ pandas data frame	One data frame as csv/pandas data frame	One data frame per country as pandas data frame	Tables for all correlations and graphs for effects and data distribution
Nr. of rows	Number of countries of interest	Number of days of observation period minus eventual holidays and weekends	Countries (210)*days since 31.12.2019	Number of days of observation period (here one year / 366 days)	Number of days of observation period (here one year / 366 days)
Nr. of columns	3	6	12	3	Varies
Index	None	Date	None	Date	Date
Columns	Country name, Index name, index code	Open, High, Low, Close Adjusted Close, Volume	Date data, geographic data, Covid-19 related data	Cases, Deaths, Stock Index	Deaths and Cases correlated to adjusted Close

7.3 Physical model

Since the data set size depends on the chosen observation period and the number of countries analyzed, the sizes of the data sets might differ to some extent, but all together the data should maximally take up a few megabytes.

The data is stored on the yahoo finance servers, where it is downloadable as csv, same with the Covid-19 data from ECDC. This data is stored on their servers as well as on our computers. Therefore enough redundancy is implemented to prevent loss of data.

After reading the data into our scripts, the data is represented as pandas data frames. In this format, the data is manipulated, analyzed, and visualized. The output then mainly consists of excel-files for data in tables and png-files for important graphs.

Regarding analysis time and velocity, these parameters weren't of utter importance, since finest granularity of the data is in daily steps. Therefore, the flexibility to analyze multiple countries at once was valued more.

8 Risks

The following risks have been identified

- Data loss: To reduce the risk of data loss, we download the data and store the raw data on the hardware and additionally on github.
- Data collection errors: As we use public data and do not collect own data, the quality is corresponding to the quality controls of the providing parties, i.e. yahoo finance and the European Centre for Disease Prevention and Control. Both are professional organizations widely used for similar cases. Thus, we consider the risks of poor data collection and pool data quality as negligible for our analysis.
- Data analysis mistakes: Incorrect application of data analysis / statistical methods and their interpretation shall be minimalized by support and advice from colleagues and peers. Pre-defined treatment of missing data ("NaN") will provide a consistent treatment within the research group.

9 Preliminary Studies

Table 4: Correlation between stock prices and cases respectively stock prices and deaths

Country	Correlation between cases and stock prices	Correlation between deaths and stock prices
USA	0.246	-0.176
China	-0.159	-0.119
Japan	0.132	-0.205
Germany	-0.585	-0.557
India	0.303	0.255
UK	-0.575	-0.378
France	-0.261	-0.422
Italy	-0.675	-0.63
Brazil	0.111	-0.031
Canada	-0.464	-0.379

Table 4 shows the correlations between deaths / cases and the respective leading indices since the start of Covid-19 on December 31st 2019. For some cases it seems counterintuitive: For example, the positive correlation between deaths and stock prices in the USA. One reason for this may be that public sentiment was only strongly influenced by the Coronavirus in certain periods, i.e. an initial shock due to high uncertainty with more positive outlook as knowledge grew. In order to test this hypothesis, we are going to look at the correlation for the first three quarters of 2020 in isolation (Table 5).

Table 5: Correlation between cases / deaths and leading stock market indices per quarter

Country	Cases Q1	Deaths Q1	Cases Q2	Deaths Q2	Cases Q3	Deaths Q3
USA	-0.576	-0.483	-0.188	-0.451	-0.536	0.097
China	-0.128	0.041	-0.262	-0.065	0.024	0.227
Japan	-0.664	-0.755	-0.646	-0.31	-0.164	0.546
Germany	-0.665	-0.465	-0.703	-0.734	0.174	-0.365
India	-0.75	-0.61	0.825	0.566	0.745	0.732
UK	-0.63	-0.484	-0.843	-0.72	-0.606	0.141
France	-0.679	-0.551	-0.484	-0.534	-0.386	-0.306
Italy	-0.863	-0.776	-0.677	-0.722	-0.564	-0.049
Brazil	-0.69	-0.57	0.854	0.743	0.274	0.235
Canada	-0.544	-0.461	-0.611	-0.081	-0.055	-0.433

In table 5 we can see a relatively strong negative correlation between deaths / cases and the stock market indices in Q1 for all countries except China. Q2 and Q3 show more varied correlation coefficients due to the uneven recovery of markets and the variation in the evolution of Covid-19 numbers in different countries. Results were statistically significant with $p<0.01$ for all results in Q1 and Q2. However, results were statistically insignificant for the most part in Q3 most likely due to the very high variance in Covid-19 numbers.

One interesting fact to consider is also that in general the stock market indices correlate rather strongly with each other (Table 6) with the major exception once again being the Chinese SSEC. Results were all statistically significant with $p<0.01$.

Table 6: Correlation between Stock Prices

	S&P 500	SSEC	N225	DAX	BSESN	FTSE	FCHI	FTMIB	BVSP	GSPTSE
S&P 500	1.000	0.701	0.893	0.894	0.756	0.429	0.565	0.587	0.764	0.845
SSEC	0.701	1.000	0.495	0.494	0.390	-0.106	0.049	0.080	0.340	0.369
N225	0.893	0.495	1.000	0.982	0.876	0.698	0.799	0.801	0.897	0.918
DAX	0.894	0.494	0.982	1.000	0.907	0.733	0.837	0.847	0.929	0.950
BSESN	0.756	0.390	0.876	0.907	1.000	0.817	0.901	0.910	0.964	0.936
FTSE	0.429	-0.106	0.698	0.733	0.817	1.000	0.976	0.959	0.866	0.819
FCHI	0.565	0.049	0.799	0.837	0.901	0.976	1.000	0.993	0.933	0.890
FTMIB	0.587	0.080	0.801	0.847	0.910	0.959	0.993	1.000	0.942	0.901
BVSP	0.764	0.340	0.897	0.929	0.964	0.866	0.933	0.942	1.000	0.952
GSPTSE	0.845	0.369	0.918	0.950	0.936	0.819	0.890	0.901	0.952	1.000

This might be an indicator that today's markets and corresponding market sentiment react to world news rather than to local information. To test this hypothesis, we again looked at the correlation between cases / deaths and the stock market indices for Q1-Q3 2020, however this time with the cumulated numbers for cases and deaths for all 10 economies. Table 7 shows an even stronger correlation between this globalized view at the pandemic and the leading market indices compared to the local numbers. Results again were significant with $p<0.01$ for all correlations in Q1 and Q2 and not significant for most correlations in Q3 with $p>0.05$. This last result leads to a possible conclusion that public sentiment was very negative due to the shock by the onset of Covid-19 in Q1 of 2020 and the corresponding uncertainty. However, outlook was adjusted in Q2 and Q3, albeit with slight differences between countries. The big exception is again China which might lead to the conclusion that the Chinese stock market outlook is unusually positive for the future or that it does not reflect actual economic outlook but the fact that some other factors might be in play.

Table 7: Correlation between cumulated cases / deaths and leading stock market indices per quarter of 2020.

Country	Cases Q1	Deaths Q1	Cases Q2	Deaths Q2	Cases Q3	Deaths Q3
USA	-0.651	-0.662	0.448	-0.511	0.4	0.15
China	-0.653	-0.614	0.626	-0.543	-0.074	-0.01
Japan	-0.575	-0.595	0.548	-0.578	0.33	-0.095
Germany	-0.617	-0.632	0.574	-0.574	0.075	-0.094
India	-0.782	-0.788	0.638	-0.388	0.476	0.173
UK	-0.633	-0.646	0.456	-0.558	-0.59	-0.202
France	-0.618	-0.634	0.588	-0.495	-0.41	-0.203
Italy	-0.619	-0.637	0.615	-0.555	-0.436	-0.154
Brazil	-0.704	-0.713	0.649	-0.573	-0.044	0.149
Canada	-0.681	-0.689	0.412	-0.531	0.339	0.226

The following analysis concentrates on the skewness resp. relative change rate of Covid-19 cases and deaths in comparison to the daily change of the stock market index.

For the Covid-19 data, the daily cases and deaths were divided by the population for every country. For the stock market index, the opening value was divided by the adjusted close value.

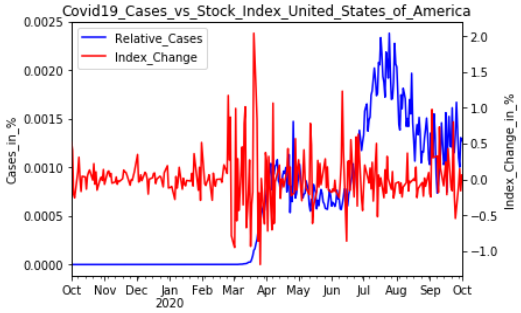

Figure 5: Stock prices and Cases, USA

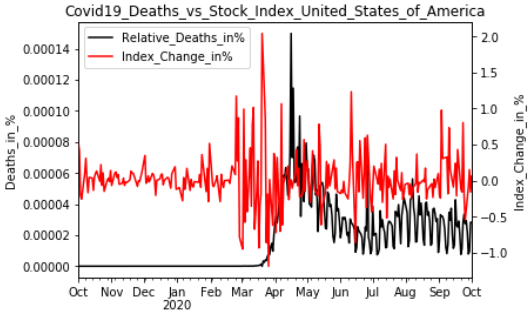

Figure 6: Stock prices and Deaths, USA

Above, the daily Covid-19 cases and deaths are compared to the stock index change of the United States of America. Below, the same for France.

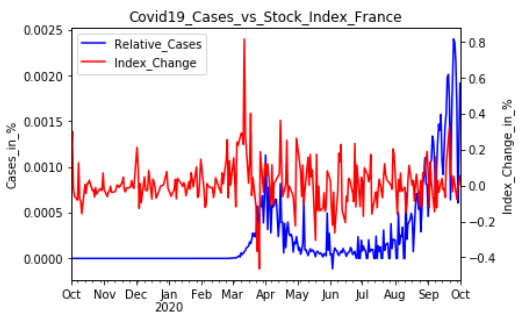

Figure 7: Cases and Stock Prices, France

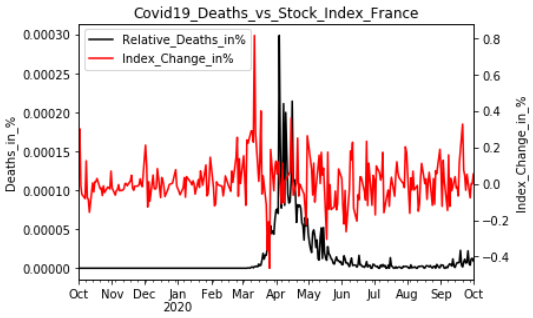

Figure 8: Deaths and Stock prices, France

Especially the comparisons with the cases show, that the stock market is not reflecting a potential increase or decrease of the Covid-19 cases. In all countries there is a stock index instability in march, exactly at the time, when Covid-19 was promoted to be a pandemic. Afterwards, the stock index begins to stabilize, but remains more fragile than before march. The overall correlations are at best marginally significant, and weekly chunk correlations show that the correlations vary greatly over time.

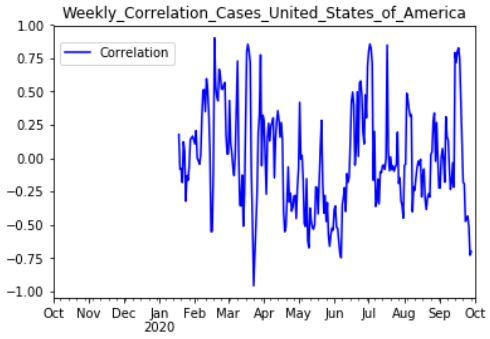

Figure 9: Weekly Correlation, USA

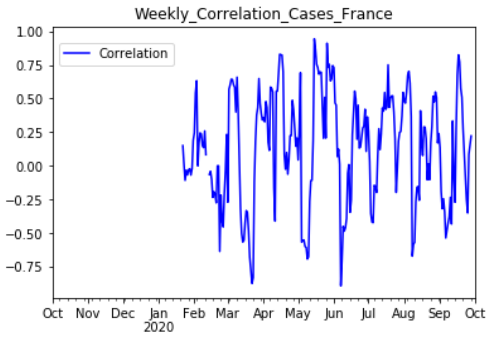

Figure 10: Weekly Correlation, France

10 Conclusions

As shown, there are strong indicators that there is a connection between market sentiment and the rise and fall of Covid-19 cases and deaths in the world's 10 biggest economies. The connection seems especially strong in Q1 of 2020 where a steep rise in global Covid-19 cases occurred and the markets reacted with a major sell-off of stocks in all major economies. Additionally, it seems that market sentiment takes on a globalized view in the major economies, reacting with higher intensity to the cumulated Covid-19 cases and deaths of the major

economies rather than to local numbers. This most likely reflects today's globalized economy, where most large companies have international business and locations that are affected by global news when it comes to a pandemic such as this. One exception to these statements is China, where markets did not react strongly to either changes in local or global Covid-19 developments, which might lead to the conclusion that other forces than public sentiment are the driving factors in Chinese stock prices. Although public sentiment seems to have somewhat recovered from the strongly negative outlook in Q1 2020 with respect to Covid-19 cases and deaths what remains is a high uncertainty in the markets, represented by the strong increase in volatility.

Appendix A

Excel-Sheet on the correlation (Pearson's r) over whole period and all variables.

References and Bibliography

Please number any information source you used in the report with corresponding links here [1]:

[1] ECDC, daily number of new reported cases of COVID-19 by country worldwide: https://www.ecdc.europa.eu/en/publications-data/download-todays-data-geographic-distribution-covid-19-cases-worldwide

[2] yahoo.finance: https://finance.yahoo.com/

[3] Notes on how data is collected by the ECDC: https://www.ecdc.europa.eu/en/covid-19/data-collection