Comparative Analysis of Clustering Techniques in Macroeconomics: Unraveling Global Economic Patterns during the 2009 Financial Crisis

Alessandro Dodon, 0001032158

Introduction

The Financial Crises remains one of the most "mysterious" events in modern history. The debate about the true causes of the crises remains unsettled, and furthermore, it was one of the biggest global economic downturns of the last century, very comparable in a way to the Great Depression. Because of those reasons, it captured the attentions of many historians, economists, mathematicians, statisticians and even public intellectuals. Plenty of qualitative insights have been offered, but I can not say the same for more quantitative forms of analysis.

In this study I present an application of Unsupervised Machine Learning and Data Analysis to study the Financial Crises of 2007-2009, exploring panel data from the World Bank, available at https://data.worldbank.org/indicator. To download the datasets, users can utilize the search function on the provided website. This project, part of the "Big Data in Social Sciences" course, extends my prior work from "Programming Lab 2," with a new emphasis on statistical analysis and the application of various models.

My main research question is: "How were countries impacted by the peak year of the Financial Crises in 2009?"

This PDF has been rendered using Quarto, and to economize on space the code is not shown. Here I outline the main results of my analysis and the crucial steps, but all of the calculations and codes can be seen in the R script.

The initial step was to download, unzip, and load datasets comprising thirteen macroeconomic/ financial indicators, quantitative, numerical and continuous in nature.

Using a combination of macroeconomic theory and a correlation plot, I picked the most significant ones for my analysis, which are:

- GDP Growth (annual %): Measures the year-over-year increase in a country's economic output, indicating economic health and productivity.
- Market Capitalization of Listed Domestic Companies (% of GDP): Reflects the total value of all publicly traded companies as a proportion of GDP, indicating the development and efficiency of the stock market and its role in economic growth.
- Inflation (Consumer Prices, annual %): Tracks the annual rate of increase in consumer prices, important for evaluating purchasing power and the effectiveness of monetary policies.

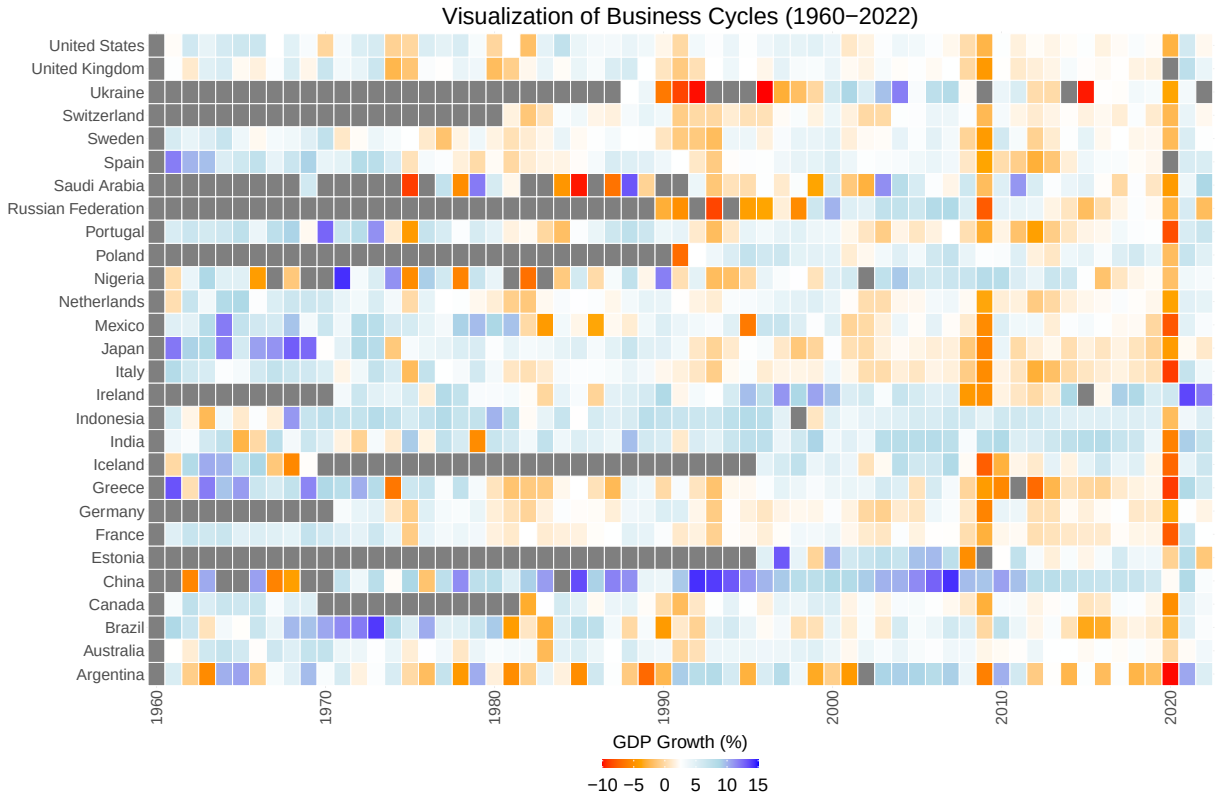
- Central Government Debt (Total % of GDP): Represents the government's debt as a percentage of GDP, providing insight into fiscal sustainability and economic impact.
- Foreign Direct Investment (Net Inflows % of GDP): Measures net inflows of investment from abroad as a percentage of GDP, indicating the level of external investment and its role in economic development.
- Unemployment (Total % of Total Labor Force, National Estimate): Indicates the percentage of the labor force that is unemployed, essential for assessing labor market conditions and economic health.

Due to the length constraints, this report primarily focuses on the results, interpretations and the thought process behind each step. However, the complete analysis can be found in the R script. This approach assures a concise yet comprehensive review of the study.

Heat map Analysis of Global Business Cycles

To gain a preliminary understanding of the global effect of the crises I plotted a heat map using the GDP growth indicator. A large list of countries was selected (chosen for their diverse characteristics and global relevance), many of which will be studied later. The interpretation is very straightforward, displaying in red the years where the countries faced negative growth and in blue the years where they faced positive

Two years indicated very severe economic downturns, 2009, peak of the Financial Crises, and 2020, marked by the COVID-19 pandemic. To simplify the complexity of the time component of the data I decided to focus exclusively on the year 2009.

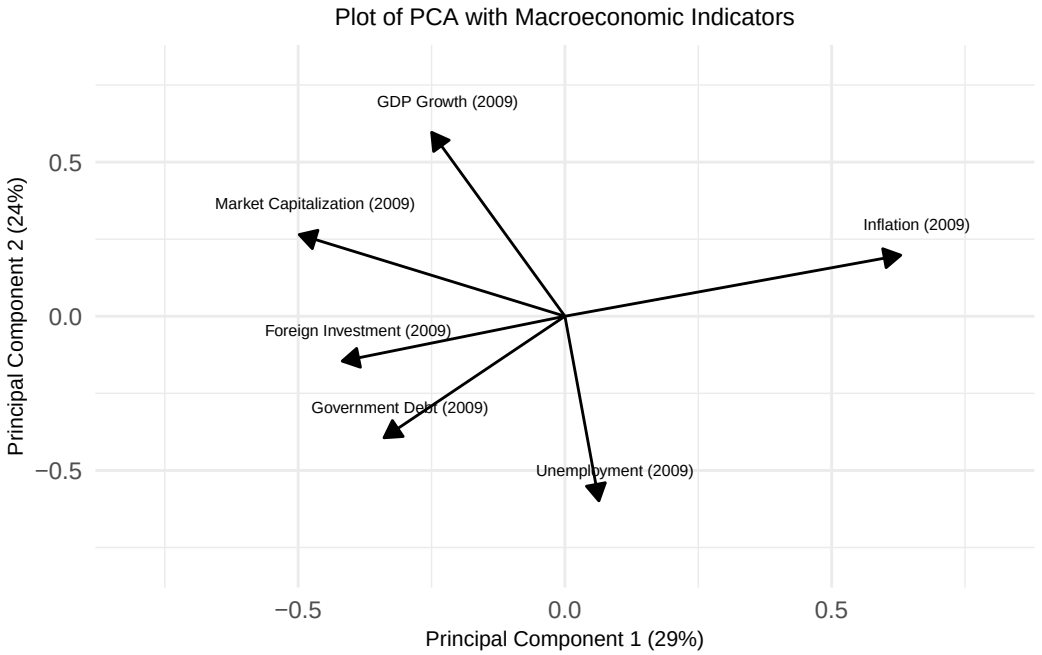


Data Pre-processing, Clustering Techniques and PCA Interpretation

As always, the data has to be pre-processed before the statistical analysis. I first removed the missing values, which unfortunately excluded many countries, then I created dataframes with the appropriate structure for the study.
As the indicators were on different scales, the data was standardized.
I now explored different clustering techniques, starting with the classing k-means algorithm. This method, applied directly to the standardized data, showed slightly higher precision (compared to the other techniques), and could operate on three or four clusters effectively based on the elbow plot. However, it was incompatible with 2D plotting.

To address this, I implemented PCA into the analysis. While using only the first two principal components, which captured over half of the variance, may not handle all of the complexities, allowed for a manageable 2D plot.

To interpret the PCA I plotted the correlations (as well as a table in the R script to confirm the results). Inflation shows a strong positive correlation with PC1 and a slight positive correlation with PC2. GDP growth has a strong positive correlation with PC2 but a negative one with PC1. Market capitalization exhibits a strong negative correlation with PC1 and a slight positive correlation with PC2. Foreign investment displayed negative correlations with both PCs, while unemployment had a strong negative correlation primarily with PC2.
This passage enables the advancement to the k-means analysis of the PCA and the successful interpretation of the results.

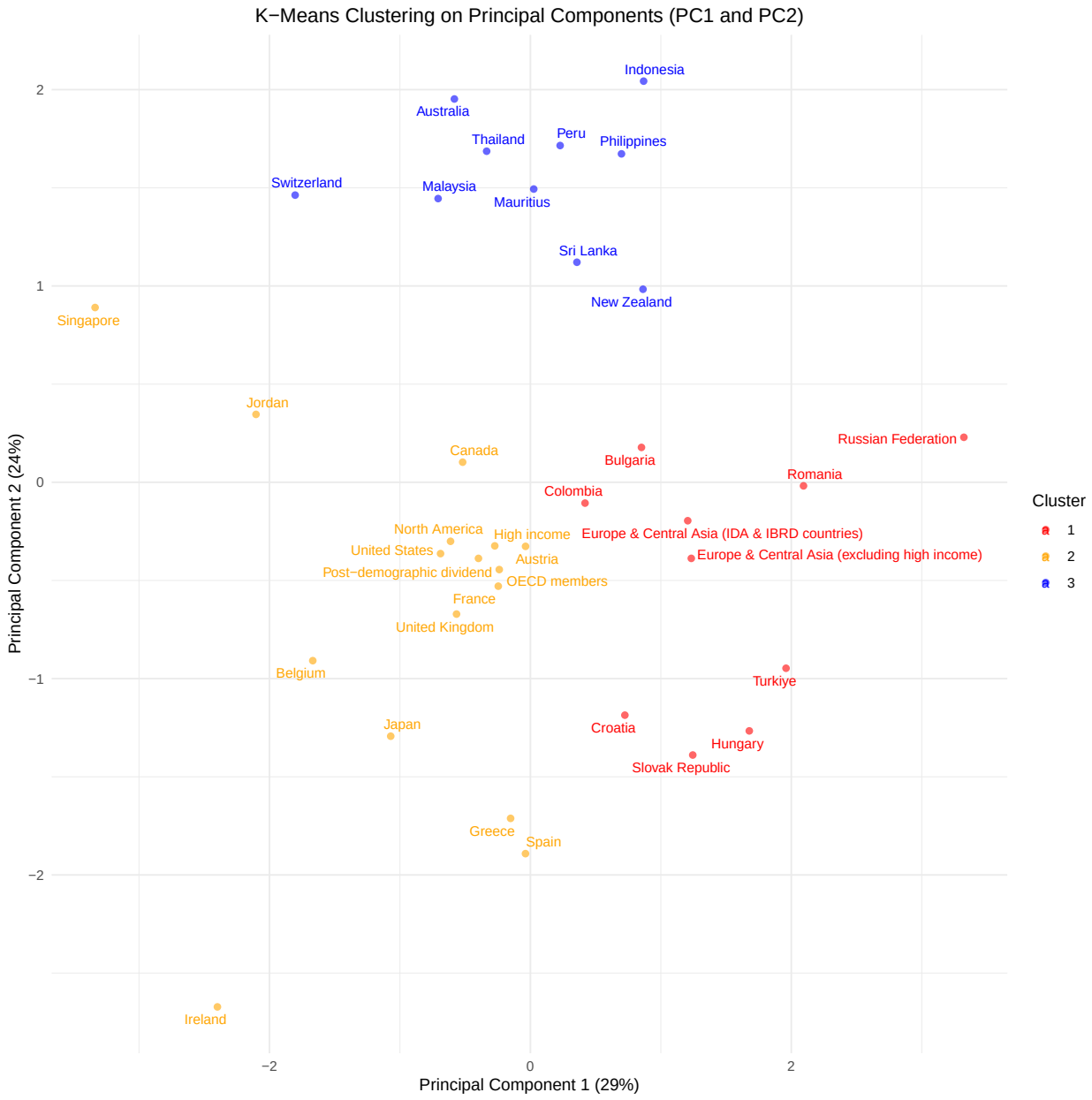


K-Means Clustering on PCA-Reduced Data

The combination of k-means clustering with PCA now clusters countries that have a comparable macroeconomic condition in 2009, showing how different economies responded to the Financial Crises.

Despite working with just over half of the variance, the k-means clustering on the PCA-reduced data shows a direct comparability to the clustering conduced over the standardized data (this passage can be found in the R script). For both, the elbow method was used to choose the number of clusters in a more evidence based way, which suggested either three or four clusters.

The interpretation of those results can be found below, after the exploration of another clustering technique.



Hierarchical Clustering Insights

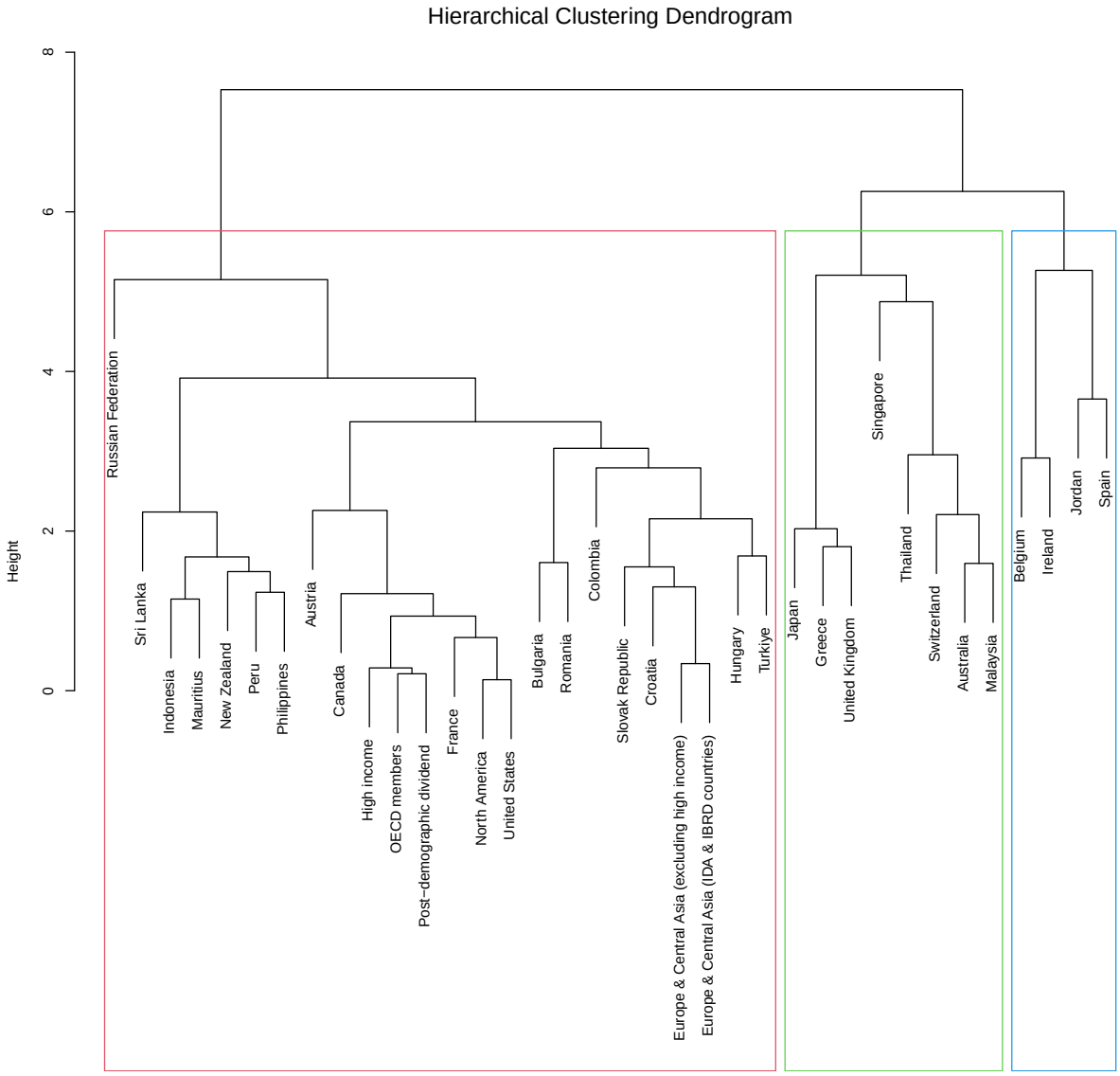
With the hierarchical clustering, using directly the standardized data, a different perspective was gained. The main advantage of this method is that it is not necessary to explicitly choose the number of clusters. Instead, the algorithm determines the grouping based on the data structure,

offering a more "complete" picture of the macroeconomic conditions.

A great deal of experimenting was done in the R script, trying to "cut" the trees at different heights, to find an optimal result. What to me seemed the best solutions were either three or two main clusters, and I choose the former option for a direct comparison with the other technique adopted.

However the clusters are notably different. The first, colored red, includes a broad range of countries, including Russia, the US, and various European nations. The second cluster groups eight countries with other distinct macroeconomic profiles. The third contains just four. These variations between the two clustering techniques may arise from the different algorithms and their respective inputs, hierarchical clustering working on standardized data and k-means on PCA-transformed data.

With that being said, the next paragraph will focus on the interpretation of the k-means clustering with PCA.



Interpreting K-Means Clustering Outcomes with Data Tables

To interpret the results of the k-means on the PCA reduced data, I have displayed two sets of tables. The first directly connects each country to its cluster number, as well as displaying the original (and non standardized) data for each indicator.
The second table provides the mean values of each macroeconomic indicator across clusters, showing some common trends for each group.

- Cluster 1 is with no doubt the one that suffered the most severe economic downturn. Countries such as Bulgaria, Colombia, Croatia, Hungary, Romania, the Russian Federation, Turkiye and the Slovak Republic show the most negative GDP mean growth and highest unemployment. Mean market capitalization is also very low, indicating that either those financial markets were not as developed in the first place (which I suspect as those countries are not so "prosperous" to begin with), or that they lost most of their value during the crises. Mean inflation is much higher and mean government debt is much lower compared to the other two clusters, which to me indicates that those countries did not employ "appropriate" monetary policies to combat the crises. The mean foreign investment remains to me more puzzling, as it is not the lowest value out of the three clusters, as I would have expected. One possibility would be that more advanced economies simply invest more in growing countries, as Economist and Nobel laureate R. famously Lucas argued.

- Cluster 2 indicates still a very severe response to the crises, as the mean GDP growth is negative and unemployment is still very high. We are now taking into consideration many European countries like Austria, Belgium, France, Spain, Greece, but also the US, UK, Canada, Japan, Singapore and others. Mean market capitalization and foreign investment being much higher (compared to the first group) suggests to me that those countries have much more developed economies and financial markets. The very high government debt indicates that as well, and possibly that those countries are adopting different monetary policies to combat the crises. Another possible explanation is that those countries simply spend more with welfare initiatives and public services. The very low inflation could be explained by high unemployment, the asset price deflation, the low GDP growth or to a demand pull inflation reduction. Macroeconomic theory is confirmed here: a country that is facing low demand, high unemployment, and failing of many business and a financial market crush is likely to have very low inflation.

- Cluster 3 seems to be relatively less effected by the crises. It's composed by countries like Australia, Indonesia, Malaysia, New Zealand, Peru, Switzerland and many others. They show a relatively high mean market capitalization (which I suspect would be similar in the previous years), a positive, even though not stellar, GDP mean growth, and a not dangerously high unemployment figure. The inflation number is higher than that of the previous cluster but I would not consider that an anomaly, a country that is growing needs to pump more money into the economy (at a considered rate) to facilitate the growing amount of economic transactions and maintain the velocity of money. Mean government debt and foreign investment could be interpreted in different ways, I would suspect that those figures have not been impacted as much as well.

Country Name	GDP Growth (2009)	Market Capitalization (2009)	Inflation (2009)	Government Debt (2009)	Foreign Investment (2009)	Unemployment (2009)	Cluster
Bulgaria	-3.27	16.63	2.75	20.16	7.49	6.82	1
Colombia	1.14	60.45	4.20	69.00	3.46	12.07	1
Croatia	-7.19	42.97	2.38	56.45	4.96	9.20	1
Europe & Central Asia (excluding high income)	-5.73	50.10	3.44	22.39	3.56	9.17	1
Europe & Central Asia (IDA & IBRD countries)	-4.67	44.74	3.47	23.31	3.48	8.98	1
Hungary	-6.60	22.92	4.21	81.11	-2.13	10.03	1
Romania	-5.52	7.51	5.59	28.23	2.66	6.86	1
Russian Federation	-7.80	62.30	11.65	8.70	2.99	8.30	1
Slovak Republic	-5.46	5.70	1.62	42.04	1.70	12.02	1
Turkiye	-4.82	35.68	6.25	48.47	1.32	12.55	1
Austria	-3.76	28.39	0.51	83.33	3.56	5.37	2
Belgium	-2.02	53.75	-0.05	95.94	16.01	7.91	2
Canada	-2.93	122.03	0.30	53.19	1.52	8.46	2
France	-2.87	72.06	0.09	77.95	0.68	9.12	2
Greece	-4.30	34.00	1.21	143.98	0.83	9.55	2
High income	-3.25	89.79	1.56	85.23	2.26	8.03	2
Ireland	-5.10	25.92	-4.48	66.62	22.83	12.61	2
Japan	-5.69	62.50	-1.35	156.98	0.23	5.07	2
Jordan	5.02	129.70	-0.74	59.22	9.83	12.90	2
North America	-2.63	105.65	-0.03	73.91	1.15	9.17	2
OECD members	-3.37	81.91	1.10	84.39	1.76	8.30	2
Post-demographic dividend	-3.39	90.68	0.84	85.45	1.78	8.04	2
Singapore	0.13	247.87	0.60	106.37	12.07	5.86	2
Spain	-3.76	96.18	-0.29	48.36	0.64	17.86	2
United Kingdom	-4.51	115.67	1.96	123.39	0.60	7.54	2
United States	-2.60	104.14	-0.36	75.84	1.11	9.25	2
Australia	1.87	135.89	1.77	23.87	3.09	5.57	3
Indonesia	4.63	39.83	4.39	30.00	0.90	6.11	3
Malaysia	-1.51	143.00	0.58	50.84	0.06	3.66	3
Mauritius	3.32	72.10	2.52	36.56	2.81	7.26	3
New Zealand	-0.11	29.25	2.12	31.94	-0.04	6.12	3
Peru	1.10	59.31	2.94	26.40	5.32	3.96	3
Philippines	1.45	49.07	4.22	52.40	1.17	3.86	3
Sri Lanka	3.54	22.69	3.46	86.06	0.96	5.85	3
Switzerland	-2.30	192.12	-0.48	21.70	8.60	4.11	3
Thailand	-0.69	62.81	-0.85	26.78	2.28	1.49	3

Cluster	Mean_GDP_Growth	Mean_Market_Capitalization	Mean_Inflation	Mean_Government_Debt	Mean_Foreign_Investment	Mean_Unemployment
1	-4.99	34.90	4.56	39.99	2.95	9.60
2	-2.81	91.27	0.05	88.76	4.80	9.06
3	1.13	80.61	2.07	38.65	2.51	4.80

Limitations and Potential Future Development

To conclude this discussion, in the name of intellectual honesty, and for potential future development, I have to highlight the main limitations of the research.

First, the pre-processing component introduced a trade-off between precision and number of countries I was able to study. As I decided to remove each row that had at least one missing value, many countries were completely excluded, such as China for example. According to the heat map analysis, and based on my knowledge it exhibited an interesting behavior during the Financial Crises, as it continued to maintain around 10% GDP growth. Similar things could be said for many other countries. Re-elaborated in statistical terms, I was not able to study the entire population, but was in a way forced to take as small sample (of less than forty countries) because of the pre-processing problems. This clearly diminishes the external validity of the study. The alternative was filling the missing values, but that would have introduced a much higher possibility of error. As it is often said in Economics "There are no solutions, only trade-offs".

Another problem was addressing the time component in panel-data, which was challenging due to autocorrelation, where each year's data (like inflation and GDP growth) depends on previous years. To manage this, I focused on 2009, a crisis peak year for many countries. While this simplified my analysis, it was not the most comprehensive approach for working with this type of data.

The experience from my "Programming Lab 2" project, which concentrated on similar topics for the US, was particularly insightful. Employing techniques like web scraping, data pre-processing, summary statistics, and creating US-centric interactive visualizations helped me better understand and interpret the time series trends.

Also the interpretation of the cluster analysis is not always straightforward, macroeconomic theory

allowed me to speculate on some trends, but I am dealing with complex macroeconomic situations of many different countries. I am also, for the sake of simplicity, not directly taking into consideration the variance of those clusters, there could be countries that are outliers and are strongly influencing the mean. In cluster 3 for example, Switzerland has a value of market capitalization of around 192.12%, much higher than any other country in that group. Theory clearly has it limits, the crises was indeed an unexpected phenomenon. Very few predicted or believed that the situation was dangerous before 2007, notably Economist and Nobel laureate Ben Bernanke was one of the very few exceptions.

Also, I am not capturing any degree of causality, as no semi-experimental or regression method was used, so I could formally say that the internal validity is very low. For potential future development they could be implemented, as well as time-series analysis, to further enrich this project. What could be particularly interesting is to find a country that never greatly suffered the crises (like China) and to compare that with a country that did greatly suffer the impact of the crises, in a Did method (also comparing the macroeconomic situation before 2007 to verify the parallel-trend assumption). Another option could be that of a logistic regression, having a dependent variable as binary outcome, for instance, "1" for countries that suffered (e.g., experienced very low or negative GDP growth) and "0" for those that did not suffer (experienced positive GDP growth) during the financial crisis. That could answer interesting policy questions like "Are countries with stricter financial regulations less likely to suffer from negative GDP growth during a financial crisis?" or many others, as there is still minimum consensus among experts in that regard.

Ultimately, I do regard the cluster analysis as a good exploratory operation, but to truly understand how each country was effected I believe more quantitative and qualitative insights are needed. Despite the obvious limitations, this study greatly deepened my understanding of the impact of one of the most significant economic downturns of modern history. My approach intentionally avoided a predefined hypothesis, allowing the data inherent patterns to surface.

Recognizing the immense value of modern quantitative techniques in Economics and Finance, I aspire, within my capabilities and sphere of influence, to contribute to a broader and more profound change. I believe indeed we are all witnessing profound paradigm shifts in those fields. It is my hope that such techniques will be increasingly and more effectively applied in the following years, providing clearer insights into complex economic phenomena.

Bibliography

- Kurlat, P. (2020). *A Course in Modern Macroeconomics*. Independently Published.
- Imai, K. (2017). *Quantitative Social Science: An Introduction*. Princeton University Press.
- Wickham, H., Çetinkaya-Rundel, M., & Grolemund, G. (2023). *R for Data Science* (2nd ed.). O'Reilly Media, Inc.
- Sowell, T. (2014). *Basic Economics, Fifth Edition: A Common Sense Guide to the Economy*. Blackstone Publishing.
- Sowell, T. (2009). *The Housing Boom and Bust*. Blackstone Publishing.
- Stock, J. H., & Watson, M. W. (2020). *Introduction to Econometrics* (4th ed., Global Edition). Pearson.