Insights on the connotation of the virus for the Spanish economy

An academic essay on the impact of covid-19 in the Spanish economic

Eduardo Sánchez López  
 Escuela Superior de Informática Universidad De Castilla La Mancha  
 Ciudad Real España  
eduardo.sanchez00@outlook.es

Maira Torres Medina  
 Escuela Superior de Informática Universidad De Castilla La Mancha  
 Ciudad Real España  
maira.torres@alu.uclm.es

0 Introduction

It is known that covid-19 pandemic has started a global recession with a scope that is hard to predict. Economy almost total cooldown neared two months in certain countries and safety measures are making more difficult to get back into business. No country was prepared for the economic fallout in part due to the low amount of studies that could be perform since the virus propagated to the rest of the world. Without having solid data to perform analysis to extract knowledge and conclusions, every country tried their own approach to minimize the effects on the pandemic.

Every sector was taken by surprise when the number of new cases and deaths kept rising. Special measures had to be taken to avoid a collapse on the sanitary system which would end up in more casualties. The investing sector was not different. Many investors liquidated an important number of assets because of the fear of economy collapse. One could ask himself many questions. Are the statistics of covid-19 spread the most important reason of the downturn for the stock market? Is a temporary crisis? Would the assets recover their value when the new cases and deaths are nearing zero? Are all the companies affected equally, or are there some groups of companies that are not affected, or even benefitted?

Naturally, the needed work to obtain real world valuable conclusions about the effect of the virus in the economy to answer these and other questions is outside the scope of an academic essay. There is a necessity for extensive data cleaning and data transformation. The trustfulness of the current data at the day of writing is on the air. Each country recollected their statistics with different methodologies [1] and countries like Spain are updating past data according to validations that could not be done on site [2].

Nonetheless, there are some techniques that could be useful to learn more about the market behaviour during a pandemic.

1 Objectives

The economy is a vast subject which involves a great number of actors and variables. Not all markets are equally important or behave important in all the regions and circumstances. As an example, the tourism sector peak in Japan is from March to mid-May [3] while in Spain the biggest peak is from July to August [4]. Because of the regional differences, it does not make sense to analyse the global markets with a general perspective.

The same reasoning can be applied when discussing the effects of covid-19 over said markets. Not all countries were affected or responded the same way, and some would recover faster than others. For this reason, the scope of the model is limited to the most important market index in Spain, IBEX35.

**The main objective** is to obtain valuable knowledge about the Spanish economy developing a predictor of the close time price of a share of all the companies in IBEX35 with Spanish covid-19 metrics as features. This predictor would be most useful in the case of a new global outbreak, which could happen in the winter of 2020 [8]. If this model deems accurate enough, it is probable that would accurately represent the patterns and behaviours of the next possible outbreak. There are multiple institutions and organizations that are making a coronavirus predictor [9] [10]. Combining those predictors with a variation of the one that is proposed in this essay could estimate the stock’s price of a company with only those coronavirus estimations. This way economic and social measures can be settled in advance to minimize the damage to the Spanish markets.

It is important to specify that the main task is not the implementation of an accurate predictor, but to correctly identify **general patterns** of the Spanish economy when exposed to a large enough virus outbreak. It is possible that the economic crisis that started in March 2020 has other agents apart from covid-19. The IBEX35 is a good enough, but not complete, representation of the current state of the Spanish economy, so it could help smaller business.

2 Scope and design considerations

This predictor would use the data from **the 20th of February to the 20th of May**, which is the first and last date that the Spanish government shared the daily new cases, deaths, hospitalizations, and critical hospitalizations by region. The coronavirus data that the model uses is the aggregation of the region’s samples into the national total. Due to the size and area of operations of the companies at IBEX35 it was deem as improbable that, if a casualisation between national covid-19 statistics and stock market prices exists, said casualisation would also happen at the regional level. If other, smaller, companies would get included in future iterations of the models, it is interesting to make this division. Small companies tend to have a smaller area of operations. If a region in which they operate is affected more severed by covid-19, the negative impact in that company would be more important than other companies operating at national or international level.

The **close time price** was chosen because the Madrid stock market closes at 17:30 and the covid-19 data was shared by the government in the morning, giving it enough time for the investors to make buy/hold/sell decision with this new information. If instead the start time price was selected, a lag of 1 day would be introduced into the model. It also makes worst predictions for Mondays, which is historically the most variable day regarding start prices due to bad news happening on the weekends [5]. This is specially true in the covid-19 situation in which a lot of the bad news for the markets where announced on the weekends, such as the start of the state of alarm [6] and the two weeks halt for non-essential economy activities [7].

**3 Data structure**

This model uses two types of data models, the IBEX35 companies daily close time and the daily new cases, deaths, hospitalized and critical hospitalized from covid-19 during the chosen timeframe. Saturday and Sunday are not included in these datasets because the stock market is closed in those days. The chosen algorithms for the model do not consider lags in the timeseries and the covid-19 statistics during the weekend effect on the stock market on Mondays is mitigated by using the close time value.

**IBEX35 datase**t has as features the corporative label of the company and as index the day for which the closed value was collected. The value is scaled from 0 to 1, being 0 the day with the lowest values and 1 the day with highest. The algorithms used in the model do not need that the data is scaled like in other algorithms based on distance, but it helps reading and understanding the model. As a remainder, the real value of a company is not important, the meaningful information is the value flow during the pandemic. In figure 1 there is a peek of the five first rows and three first columns on this dataset, and in figure 2 a graph of the 35 companies during the chosen timeframe.



Figure 1 IBEX35 dataset

Figure 2 gives already some very useful insights. Most of the companies start this period with the highest, or at least close to, closed value, and they lose the most value between the first and second week of March. This coincides with the categorization of the covid-19 as a pandemic by the WHO (11th of March) and the start of the state of alarm in Spain (16th of March).

From there, some companies recover faster than others, or even get close to maximum values before the crisis. This recovery pattern is especially interesting because it can give insights about what type of market is most affected by the disease and how fast can they recover from it.

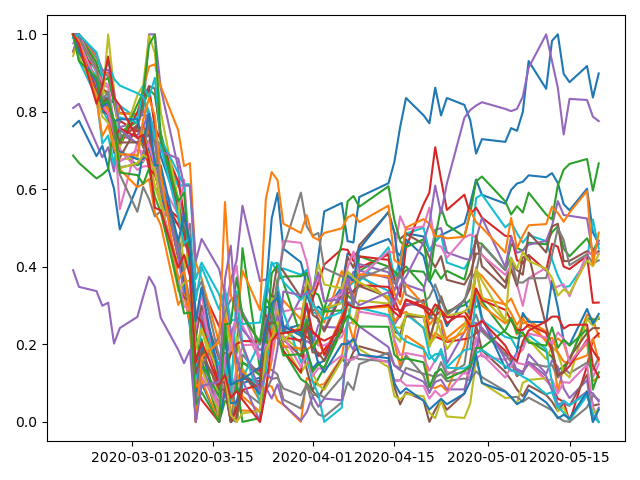
****

Figure 2 Companies closed value flow during the pandemic

Nonetheless, it is important to mention that this dataset, while given a general view of how the economy of Spain performs, it is not complete. Not all markets are represented in this subset of big companies, so it is quite probable that there are uncovered sectors that are still very important for Spanish financial growth.

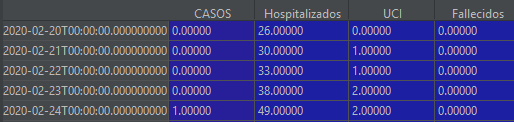


Figure 3 Covid-19 Spanish metrics.

**Coronavirus dataset** has as features the daily new cases detected, deaths, hospitalized and critical hospitalized and as index the day in which this data is recorded, weekends not included. This data is not scaled like IBEX35 dataset. Figure 3 gives a look at the first five rows and figure 4 is the evolution over time of the metrics.

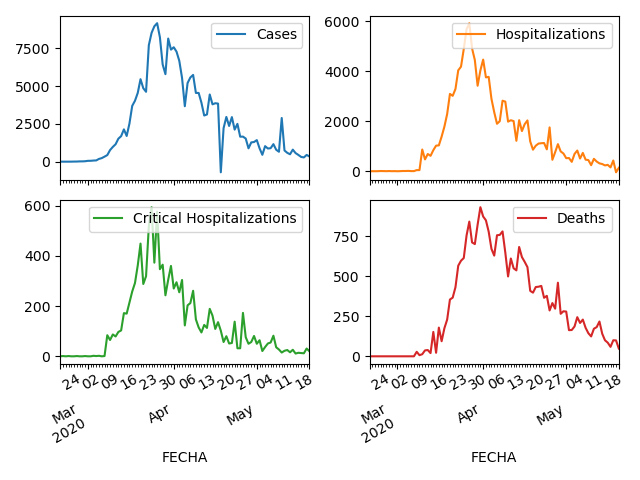


Figure 4 Daily new cases, hospitalizations, critical hospitalizations, and deaths by covid-19

Note that in the number of new cases there is a negative value. This is likely a correction done by the government to compensate with the different ways that the Spanish regions reported the data [14]. To preserve the market reaction at the time of the disclosure of the data, this erratic value is not treated in the dataset.

**4 Clustering**

As stated in the previous sections, making a single model for all the international markets would probably not yield accurate results. Tourism in Japan does not care about the same things as tourism in Spain, for example. This statement is also true when creating a model at the national level. There are some factors that are equally important for all companies, like society or social security tax rate, and when one of those factors is altered, it affects the whole economy. Part of the goal of this essay is to check whether covid-19 is one of those common factors.

For that, the first step in building the model is to divide the dataset into **clusters**. These clusters consist of companies that share a common behaviour **during the coronavirus crisis**. That is, that the close price during the timeframe from the 20th of February of 2020 to the 20th of May of 2020 have a similar pattern.

The selected technique for clustering is **hierarchical clustering**. The reason behind this election is double. The dendrogram that can be constructed with the clustering results is very helpful to understand the data behaviour and it is one of the clustering methods that allows to use the correlation as a metric. This fact is in line with the objective of the project. With this algorithm, the clusters would be formed by how the value of a company increases or decreases in comparation with its competitors. If several companies lose value at a very similar pace, the correlation between them would be high and they would be placed in the same cluster with a high probability. These are the desirable kinds of patterns for the project. The resulting dendrogram is shown in figure 5.

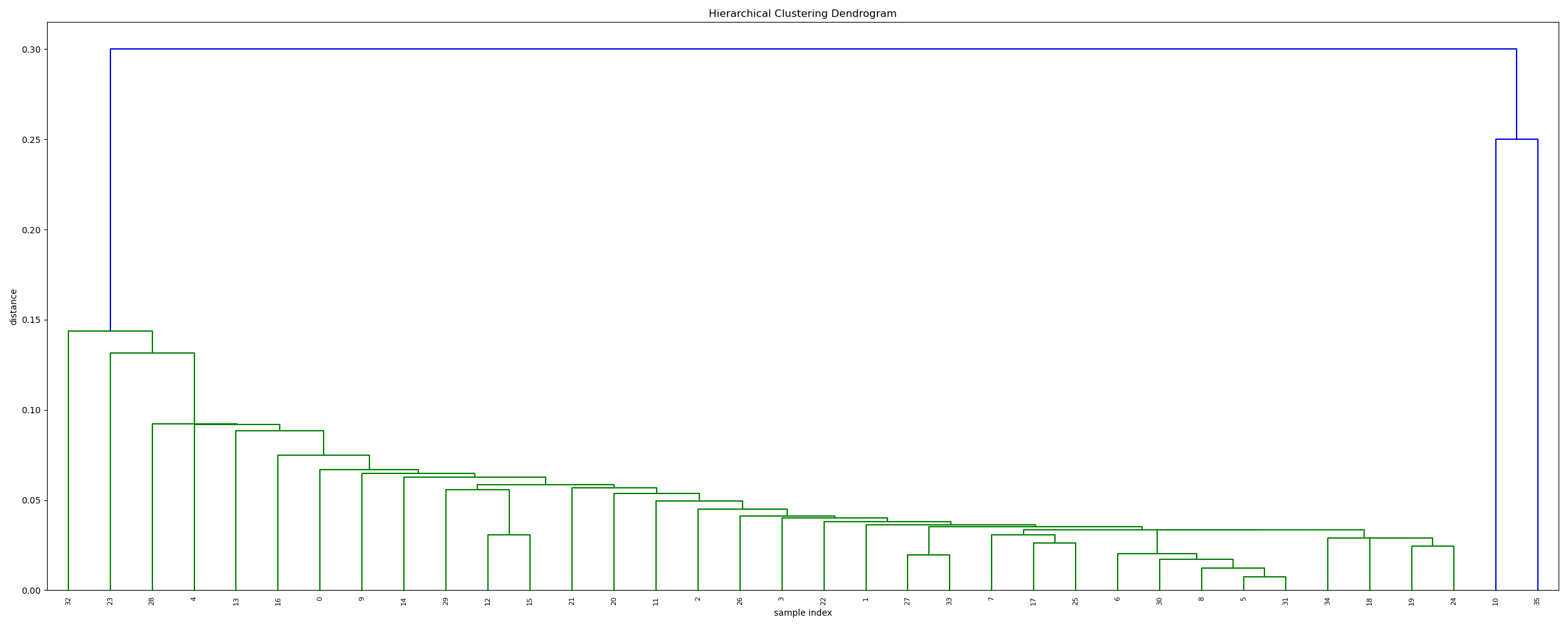


Figure 5 Dendrogram with clusters

From this dendrogram there are some conclusions that can be extracted from the data. First, there are not multiple groups of companies that behave in different ways, but one massive group and a very small group of outliers. And because of that, most of the companies go up or down in value at similar speeds. The companies at the right side on the dendrogram correspond to Cellnex Telecom and Viscofan. Both of these companies had their maximum value at the end of the timeseries, (13th and 11th of May respectively) while the other companies had their maximum at the start of the data recollection, before coronavirus was a public concern in Spain. Both companies recovered very fast and even got their local maximum during the de-escalation, while other companies were still half-way for the recuperation of the lost value. It is also quite interesting because Viscofan is **less than 5% away** from its all-time max value and is also the closest to said its has been since that peak. Cellnex is even more impressive because its all-time highest value was achieved the 8th of May. These companies should be studied separated in case there is knowledge that could be used by other entities to adapt for possible future outbreaks.

Since no clusters were clearly defined, instead of making a predictor for a cluster of companies, each company would get their own fitted predictor. This solution is not general and would probably end up overfitting the results, but it was decided that this possibility of overfitting was more acceptable than making a single predictor for 35 big companies that operate in different sectors.

**5 Feature selection**

Since there is only four candidate features to fit the model (daily new cases, deaths, hospitalized and critical hospitalized) feature selection was centered around avoiding duplicity of the features. The main concern is that the four features are very similar and maintaining all of them would introduce noise to the model.



Figure 6 Correlation heatmap

The heatmap of the correlations of every target feature (company’s values) and features used to build the model with themselves is shown in figure 4. Somethings to consider are that, as expected, the two companies with the unusual pattern have relatively high correlation with the covid-19 metrics. When the number of new cases goes up, so does the value of their stocks. The great majority of them have a negative correlation, the price goes down as the casualties are higher, while a small subset of them are indifferent with close to 0 correlation. The other important aspect is that the four non target features have correlations that could be considered as perfect. Because of this reason and to keep the model as simple as possible, only the new cases metric would be used as feature to fit the model.

6 Predictor

The chosen algorithm for the predictor is **Random Forest**. This tree-based ensemble method seems to work best with multivariate regression. In this iteration of the model only two features, the target, and the number of cases, are used to fit the model. This random forest regressor would not have a maximum depth and it would use 100 trees as assembly. Overfitting is something that should not be a worry in a regressor with two features but introducing more trees could start being a problem in this regard. The training dataset would go from the 20th of February to the 8th of May and the test dataset would go from the 11th to the 20th of May. The criterion for which the trees would agree into the best result is the mean absolute error, which is the one that is going to be calculated for each company in the testing phase. Since the target feature has been scaled from 0 to 1, the MAE gives us already an error in percentage that is easier to understand and to compare with the other metrics. All the companies with their corresponding errors are included in ANNEX 1. The mean of those errors is 0.28, which means that almost a third of the predictions were incorrect. The standard deviation is 0.13, which means that there is a high variation on the precision of the forecast. These errors are quite high, indicating that the predictor is lacking in precision for it being useful. Nonetheless, it could be adjusted for future iterations when other types of data are introduced.

7 Improvements

Since the scope for this model is limited to an academic essay and is centred more about the methodology, good practices and proposal of a solution and not about giving a complete result, there are some adjustments that can greatly improve the knowledge that can be extracted from the data. The next listing covers some ideas.

Deeper data cleaning. The data transformation stage covered the aggregation for national covid-19 stats, the conversion from total amounts to daily increases and treatment of null values in the IBEX35 dataset and the scaling of said data. There is more that can be done specially in the covid-19 dataset. As an example, as stated in previous academic essays by the same authors [11] there is a false seasonality in the increase of new cases and deaths every Monday, result of a lag due to the decrease in the workforce for the weekends. While this condition should not be as impactful in stock market analysis, it is worth exploring to check whether the “Monday’s effect” is also impacting the investments.

Data completion for market datasets. In section 3 it was discussed that this data does not give a complete figure of the financial status of Spain during the crisis. Once more data is unveiled, such as growth reports o sales report, and added to the model, the final picture would represent Spanish markets better. Stocks price on a company can be correlated to the profits of said company, but it is not a final descriptor on how well is that company doing, specially during up times [12] [13].

More focus on the creation of market groups. The proposed method for creating clusters of this subset of companies failed to give a meaningful characterization. It correctly extracted several clearly identified outliers but could not differentiate the rest of the companies from each other. While this may be the case, it is worth exploring with other techniques or with heuristics from experts on the matter in case there are unidentified differentiations that could help on the knowledge extraction. Completing the data scope as stated on the previous point should help on the exploration of different market behaviours when exposed to a pandemic.

Comparing data with other countries markets. Correlation does not mean casualization. There were multiple variables that made that the finance world regressed. Creating a model with data of other countries that had a more lax approach regarding quarantine and comparing it to Spain can identify how important were the restrictions in the economy downfall or if most of the work was done by the natural market reaction to the coronavirus stats. These insights could be very powerful to estimate the potential losses in the case of the need of a new quarantine and motivate prevention politics to save lives while avoiding market collapse. There are multiple studies about creative alternatives to avoid total lockdown, like setting two-day workweek that would avoid sanitary collapse in hospitals while maintaining a percentage of the economy active [13].

8 Conclusions

While most of the companies were very correlated to the number of cases, this variable alone did not stand as a good enough predictor for most of the corporations. This means that there are other factors that were not taken into consideration by the predictor, which do not have to be related to coronavirus, that influence on the company value.

Most of the companies have a slow recovery with a probability that it would never happen, related or not to coronavirus. With more data being recollected for the following months, it would be easier to see which companies and markets are recovering faster, and which need more time.

As final conclusion, coronavirus was the common cause for the decrease in value for 33 of the 35 IBEX35 companies. While the virus was the starting point, recovering from it does not mean that the previous financial normality is back. There are multiple factors, such as lost income due to fewer customers or insufficient stock due to the closure of the frontiers, that would not disappear the moment the number of infected are back to near zero.

ANNEXES

1 LIST OF MEAN ABSOLUTE ERRORS

|  |  |
| --- | --- |
| Company | MAE |
| ACS | 0.099413 |
| ACX | 0.289016 |
| AENA | 0.260482 |
| AMS | 0.349462 |
| ANA | 0.552032 |
| BBVA | 0.38493 |
| BKIA | 0.36397 |
| BKT | 0.303036 |
| CABK | 0.368757 |
| CIE | 0.209678 |
| CLNX | 0.262619 |
| COL | 0.497589 |
| ELE | 0.167952 |
| ENC | 0.239606 |
| ENG | 0.101905 |
| FER | 0.175693 |
| GRFS | 0.160808 |
| IAG | 0.379939 |
| IBE | 0.409952 |
| IDR | 0.545898 |
| ITX | 0.197461 |
| MAP | 0.220311 |
| MAS | 0.103902 |
| MEL | 0.315114 |
| MRL | 0.483797 |
| MTS | 0.281815 |
| NTGY | 0.250702 |
| REE | 0.103624 |
| REP | 0.103174 |
| SAB | 0.470025 |
| SAN | 0.427043 |
| SGRE | 0.119199 |
| TEF | 0.222877 |
| TL5 | 0.194428 |
| VIS | 0.443762 |

2 REFERENCES

[1] <https://www.bmj.com/content/369/bmj.m1395>

[2] <https://www.huffingtonpost.es/entry/sanidad-coronavirus-espana-fallecidos-2000-cifra_es_5ecbde8ec5b6cf747d53231f>

[3] <https://www.japan-experience.com/to-know/visiting-japan/visiting-japan-off-season>

[4] <https://mesadelturismo.org/wp-content/uploads/2019/01/ANALISIS-DE-LA-COYUNTURA-TUR%C3%8DSTICA-EN-ESPA%C3%91A-DURANTE-EL-A%C3%91O-2018.pdf>

[5] <https://www.investopedia.com/day-trading/best-time-day-week-month-trade-stocks/>

[6] <https://www.elconfidencial.com/espana/2020-03-13/pedro-sanchez-comparecencia-estado-alarma_2496668/>

[7] <https://www.elperiodico.com/es/politica/20200328/sanchez-paron-economia-total-coronavirus-7908942>

[8] <https://www.washingtonpost.com/health/2020/04/21/coronavirus-secondwave-cdcdirector/>

[9] <https://wp.bcamath.org/news/en/2020/05/15/bcam-publishes-a-report-with-predictions-about-the-evolution-of-covid-19-in-the-basque-country/>

[10] <https://www.bmj.com/content/369/bmj.m1328>

[11] <https://github.com/Edusanc95/ARIMAinCovid19>

[12]<https://www.researchgate.net/post/Why_is_the_correlation_among_stock_prices_larger_when_the_return_is_negative>

[13] <https://medium.com/@urialonw/containing-sars-cov-2-with-a-two-day-workweek-fbdea4030d30>

[14] <https://www.elnortedecastilla.es/sociedad/salud/espana-registra-nuevos-20200417123035-ntrc.html?ref=https:%2F%2Fwww.google.com%2F>

IBEX35 data was extracted from <https://es.finance.yahoo.com/>

Covid-19 data was extracted from <https://covid19.isciii.es/resources/serie_historica_acumulados.csv>