Covid 19 - Spain
Data exploration and Current Situation

Capstone
IBM
DATASCIENCE

Iván Barra

1. Introduction

1.1 Background

For the world, the current pandemic represents a challenge that involves all the habitants of the planet. To be exact, the COVID-19 virus has already killed more than 100,000 people in all over the world.

Developed nations are being hurt mainly due to lack of prevention by their rulers. Thus, thousands of elderly people die in solitude who have helped with their lifelong work to forge the quality of life in today's societies. The economic effect is another aspect to consider in this pandemic, as they are expelled from the labor market due to the bankruptcy of the company due to the paralysis of the market and business ventures.

This work analyzes the current situation in Spain in relation to the evolution and spread of the virus, an evolution that has been accelerated, creating a commotion in the population by having to attend helplessly to see how their relatives died in solitude and in a lamentable condition. Families destroyed with the loss of their loved ones and an alarmingly falling economy that in all probability will affect the most disadvantaged sectors.

The spread of this virus and the tragedy that it leaves behind is especially cruel with elder people, which in the case of Spain corresponds to 80% of the total number of deaths accumulated to date in the range of 70 to 90 years.

A recognition of the people who have made it possible for society to continue living, such as medical personnel, officials of the security forces and workers who supply food in supermarkets.

A recognition to the families who must be confined in their homes during all this quarantine.

1.2 Problem

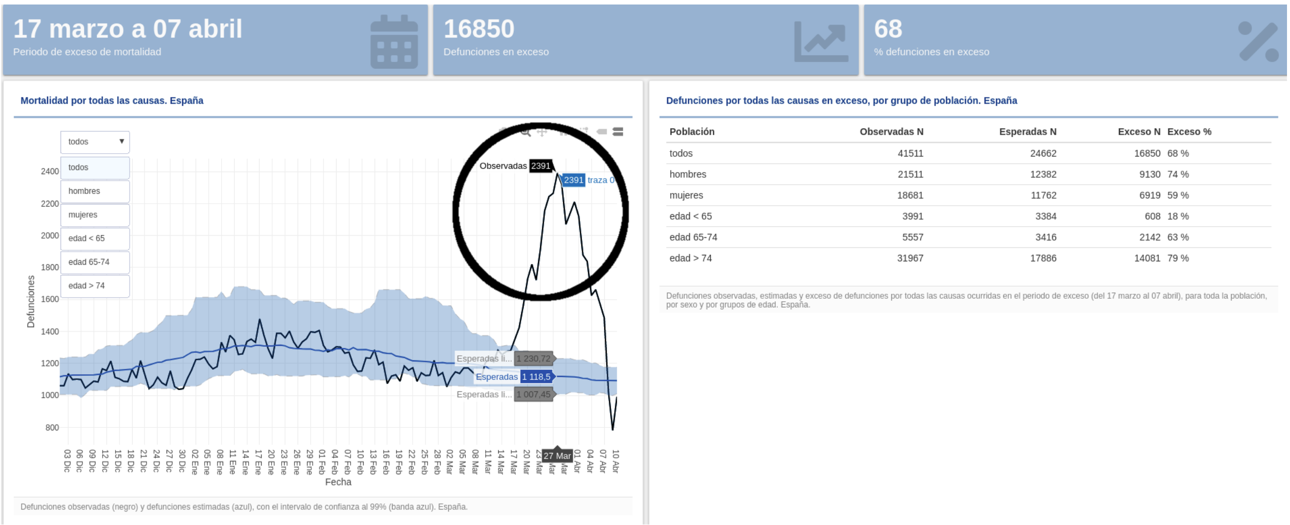
The problem that we are currently observing is the lack of verified and reliable information that allows us to make adequate decisions to resolve the pandemic.

Governments control information so that society does not express their discontent or data scientists investigate to its extent and deepen the application of prevention policies in the field of environmental health and safety.

The records of deceased persons are different from the figures that are known in practice by the institutes that observe the statistics of deaths.

https://momo.isciii.es/public/momo/dashboard/momo_dashboard.html#nacional

The following image indicates, in the circle area, the deaths observed against the deaths reported by the official authorities.



17 marzo a 07 abril
Periodo de exceso de mortalidad

16850
Defunciones en exceso

68
% defunciones en exceso

Mortalidad por todas las causas. España

Defunciones observadas (negro) y defunciones estimadas (azul), con el intervalo de confianza al 99% (banda azul). España.

Defunciones por todas las causas en exceso, por grupo de población. España

Población	Observadas N	Esperadas N	Exceso N	Exceso %
todos	41511	24662	16850	68 %
hombres	21511	12382	9130	74 %
mujeres	18681	11762	6919	59 %
edad < 65	3991	3384	608	18 %
edad 65-74	5557	3416	2142	63 %
edad > 74	31967	17886	14081	79 %

Defunciones observadas, estimadas y exceso de defunciones por todas las causas ocurridas en el periodo de exceso (del 17 marzo al 07 abril), para toda la población, por sexo y por grupos de edad. España.

One of the most significant problems has been precisely identifying and contrasting the data, considering that according to research organizations, an environment of 7.000.000 million inhabitants currently infected in the country could be calculated.

This report will be perfected over time to be of help to other people interested in studying this global pandemic, emphasizing the reality of the country.

The central problem is disinformation and a lack of transparency. This work is the beginning of an investigation that will precede the culmination of this course.

1.3 Interest

Interested professionals include data science practitioners, students, researchers, local government leaders and health workers

2. Data acquisition and cleaning

2.1 Data sources

As previously reported, the lack of verifiable information was being a problem, when identifying the sources and/or data source, we may be facing another problem, that is, too many sources making noise, transforming into what is called " Infodemic ".

The central data source is the one provided by the government health ministry https://www.mscbs.gob.es/en/profesionales/saludPublica/ccayes/alertasActual/nCov-China/situacionActual.htm and https://covid19.isciii.es/

However, we can observe more than 1000 data sources available through an international consortium that has been structured to collect the scattered data currently (all the accumulated data are of a scientific nature).

The information is updated daily. with a day of delay.
The investigation stages have been:

-Data collection
- Data preparation
- Descriptive analysis
- Predictive analytics

2.2 Data cleaning

Every attempt was made to deal with consistent data in that they came in an adaptive format. Multiple files are used and an attempt is made to optimize their consolidation.

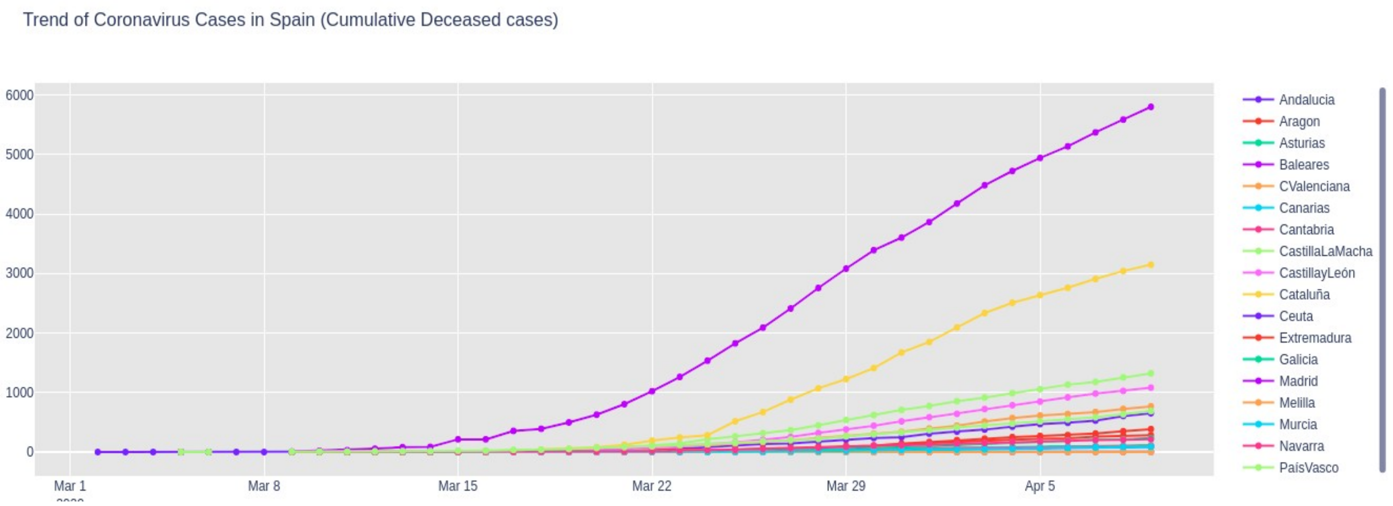
It consists of 7 to 10 .csv extension files. The files are checked before using them, and they detect files with missing or corrupt records. The data range is recent, therefore there is no loss or absence of values or corruption of the same. The most significant problem is the margin of error for creating a correct predictive model.

2.3 Feature selection

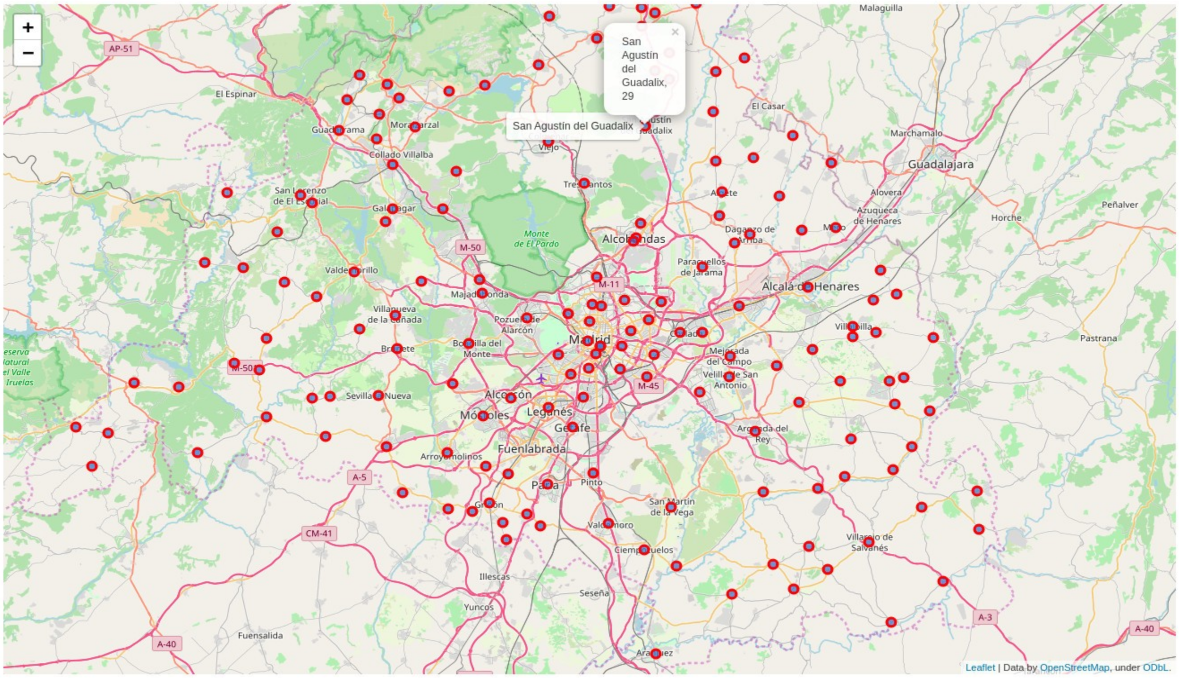
Once the data has been organized and reviewed. Being temporary in nature, the samples are not large. Let us remember that this analyzes an epidemic of recent origin in the national context. Obviously, if the goal was the analysis of two or more countries, the work would have been more complex. So seen the above, the data is small samples, the problem is still checking them. It is a problem that currently cannot be solved.

3. Exploratory Data Analysis
3.1 Calculation of target variable

With the existing data, the first objective was to carry out an analysis to find a correlation regarding, for example, the evolution of the epidemic in the different autonomous communities, it was found that the most populated cities produced the majority of deaths when just as these deaths corresponded to people who were mostly over 60 years old.

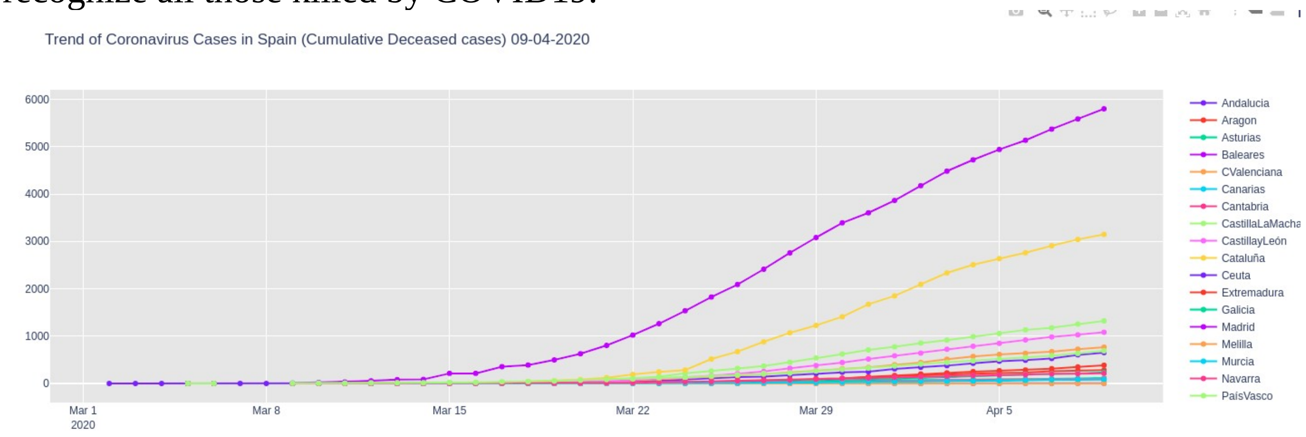


Another objective in the exploration has been to represent the city of Madrid on a plane with confirmed cases of COVID19. For this, I investigate the locations, geolocation, identification of and number of cases, among others.
In the future the algorithm will be improved for its automated update, for now the update is manual



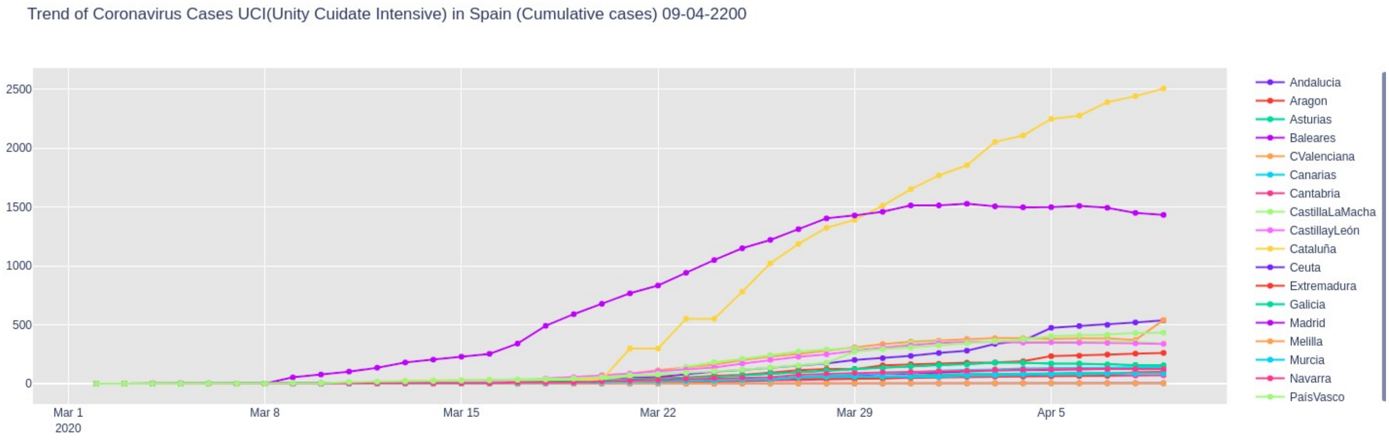
3.2 Relationship between Autonomous Communities and the deceased

The situation in the case of deceased people, the most tragic of the COVID19 pandemic. In the case of Spain, the situation has collapsed not only in hospitals also in nursing homes where residents were found dead after several days. The figures indicated by the authorities would triple as regards the classification of the deceased as it corresponds to a problem of a political nature since the authorities do not recognize all those killed by COVID19.


Trend of Coronavirus Cases in Spain (Cumulative Deceased cases) 09-04-2020

3.4 Relationship between Autonomous Communities and ICU (Intensive Care Unit)

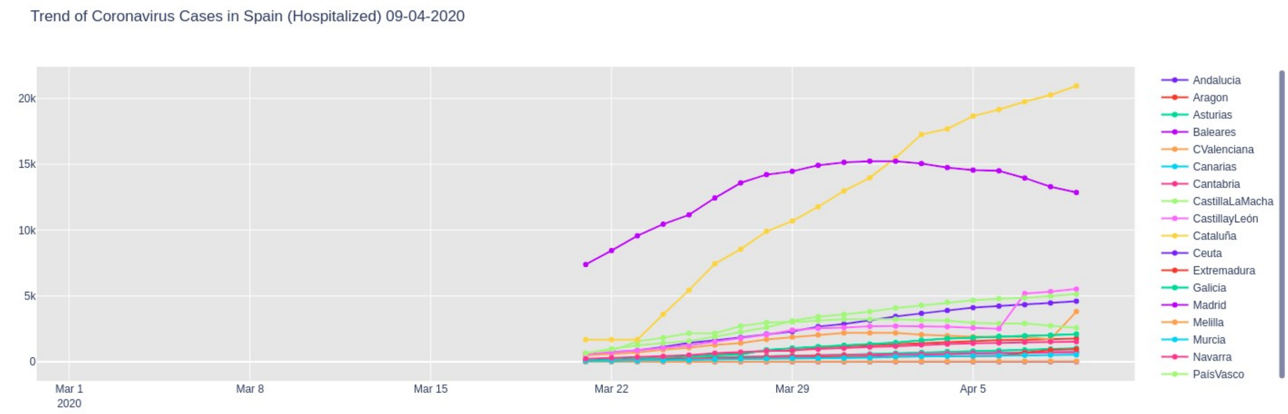
Hospital emergencies due to the virulence of COVID immediately collapse hospitals and health network. By implementing IFEMA and medicalizing hotels, the evolution of confirmed cases can be managed more efficiently. Every time that upon detecting the infection, the patient was hospitalized urgently and subsequently underwent normal hospital care. According to the graph, people spent less time in the ICU, hospital infrastructure collapsed previously.


Trend of Coronavirus Cases UCI(Unity Cuidate Intensive) in Spain (Cumulative cases) 09-04-2200

3.4 Relationship between Autonomous Communities and Hospitalized

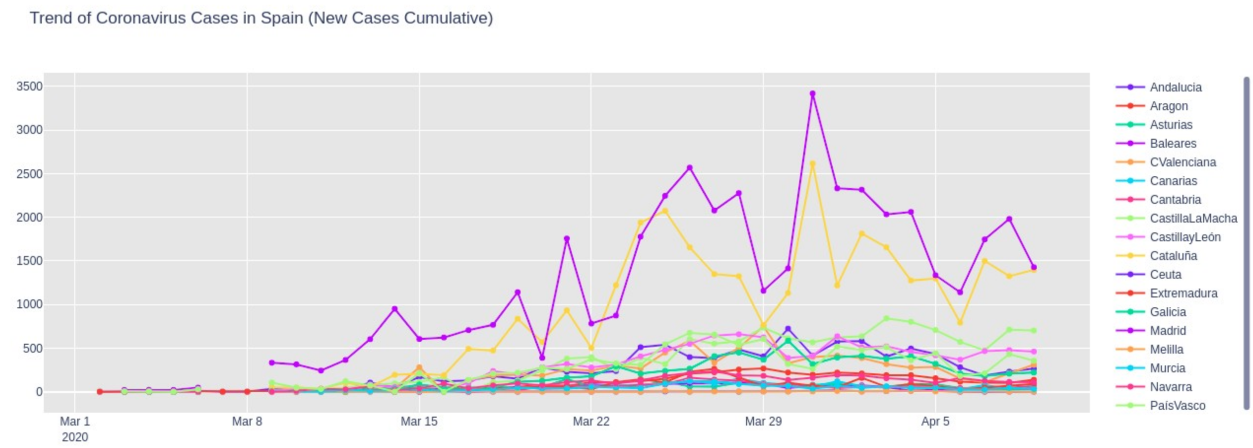
In the graph we can observe the evolution of the epidemic for people who they were detected and hospitalized. The country is considered to be 10 days behind in developing health policies for Italy.

A large field hospital (IFEMA) is installed in the community of Madrid, in which sufficient hospital beds (5500) and 500 ICU beds are equipped to cope with the collapse of hospitals that as of March 15 were saturated.


Trend of Coronavirus Cases in Spain (Hospitalized) 09-04-2020

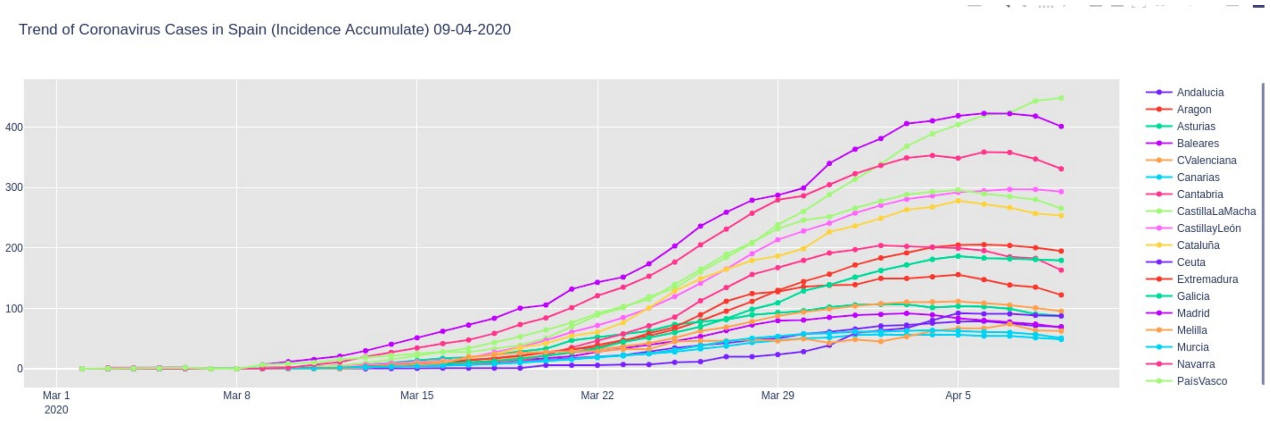
3.5 Relationship between Autonomous Communities and New Cases

If the graph is observed, the new cases are detected in the first days of the month of March, for example in the Basque Country, 4 cases are detected that would later imply closing an entire city due to the rapid contamination of those infected, in this particular case contamination occurs in religious activity (funeral).
As of today, the downward trend of new cases in the cities of Madrid and Barcelona can be observed, while in other communities the trend remains constant.


Trend of Coronavirus Cases in Spain (New Cases Cumulative)

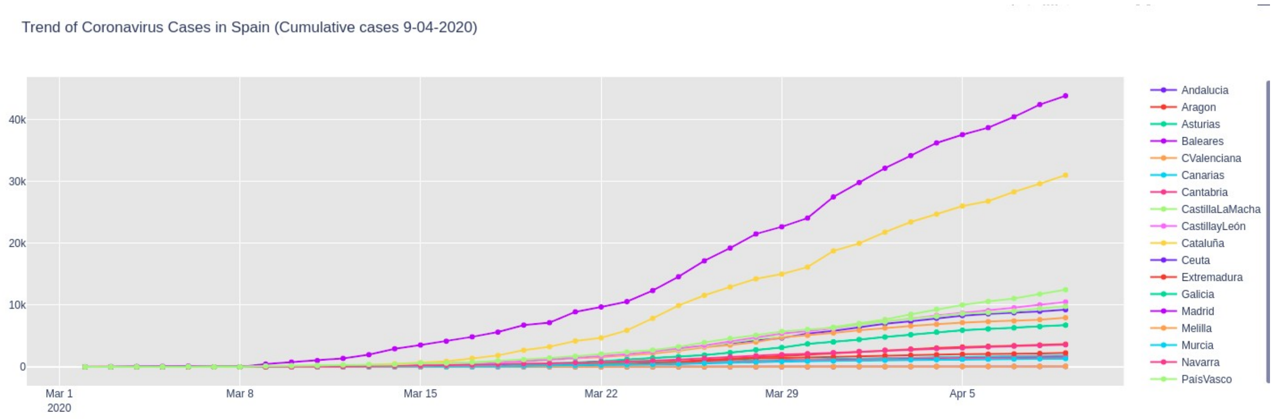
3.6 Relationship between Autonomous Communities and Incidence Accumulate

The incidence reflects the number of new "cases" in a period of time. It is a dynamic index that requires monitoring over time of the population of interest. When the disease is recurrent, it usually refers to the first appearance. The last 14 days of follow-up are considered



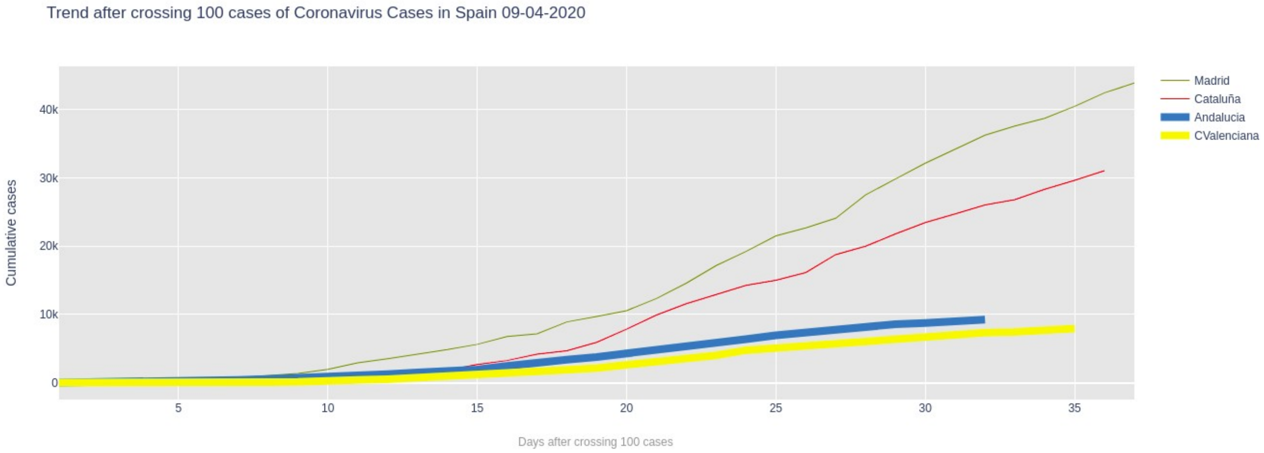
3.7 Relationship between Autonomous Communities and Cases confirmed

The communities with the highest number of confirmed cases are Madrid with 43,877 cases and Barcelona with 3,1043 confirmed cases. In the graph we can see both communities highlighting their increase in cases with respect to other communities. Consider that the visible symptoms of the virus become independant at least 15 days after being infected, in such a way that the curve increases as of March 8, making it clear that COVID was already present in the city at least 20 days in advance.



3.8 Display inference

The chart above shows the number of days after COVID-19 cases cross 100 versus the total number of cases in the country. Both Italy and South Korea have crossed the 5,600 mark in the next 13 days. The number of cases detected (trend) in India is lower compared to Italy and South Korea

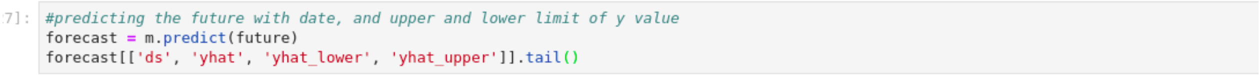


4. Predictive Modeling

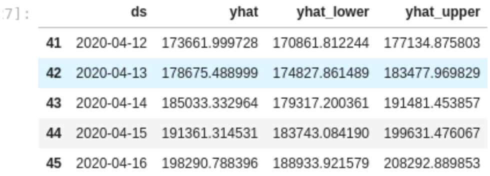
4.1 Regression models

Generating a one-week forecast before confirmed NCOVID-19 cases using Prophet, with a 95% prediction interval by creating a base model without seasonality-related parameter adjustments and additional regressors.

1)Number of cases confirmed

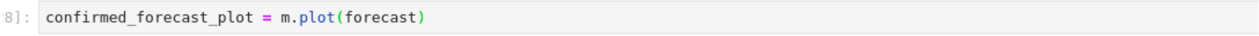


```
7]:  #predicting the future with date, and upper and lower limit of y value
     forecast = m.predict(future)
     forecast[['ds', 'yhat', 'yhat_lower', 'yhat_upper']].tail()
```

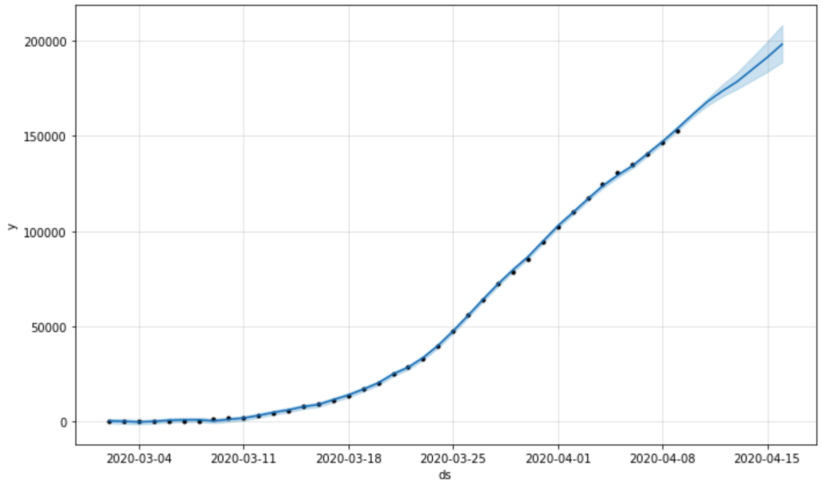


7]:

	ds	yhat	yhat_lower	yhat_upper
41	2020-04-12	173661.999728	170861.812244	177134.875803
42	2020-04-13	178675.488999	174827.861489	183477.969829
43	2020-04-14	185033.332964	179317.200361	191481.453857
44	2020-04-15	191361.314531	183743.084190	199631.476067
45	2020-04-16	198290.788396	188933.921579	208292.889853

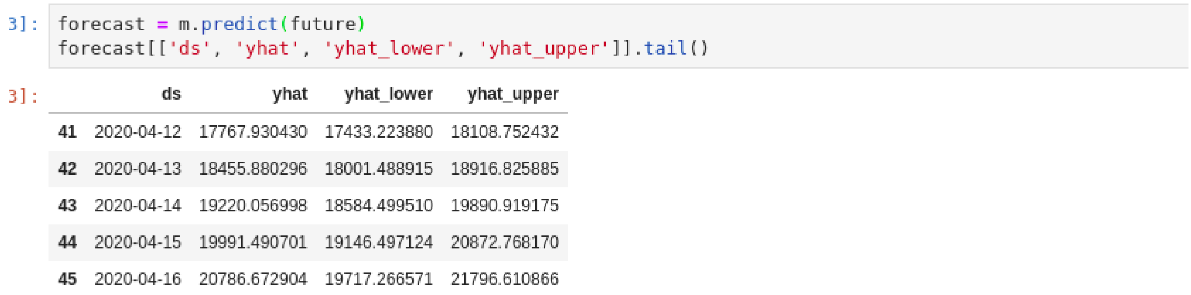


```
8]:  confirmed_forecast_plot = m.plot(forecast)
```



Observation: The projection of "confirmed cases" to April 12 according to the Prophet sample indicates 173,661 cases, at the time of writing this report. Today's official figures according to national authorities indicate, whose confirmed cases of 166,019 confirmed cases. However, as there is always a 24-hour delay in the delivery of the information, the forecast information should be confirmed on April 13 for greater security.

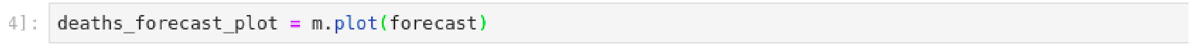
2)Number cases Deceased



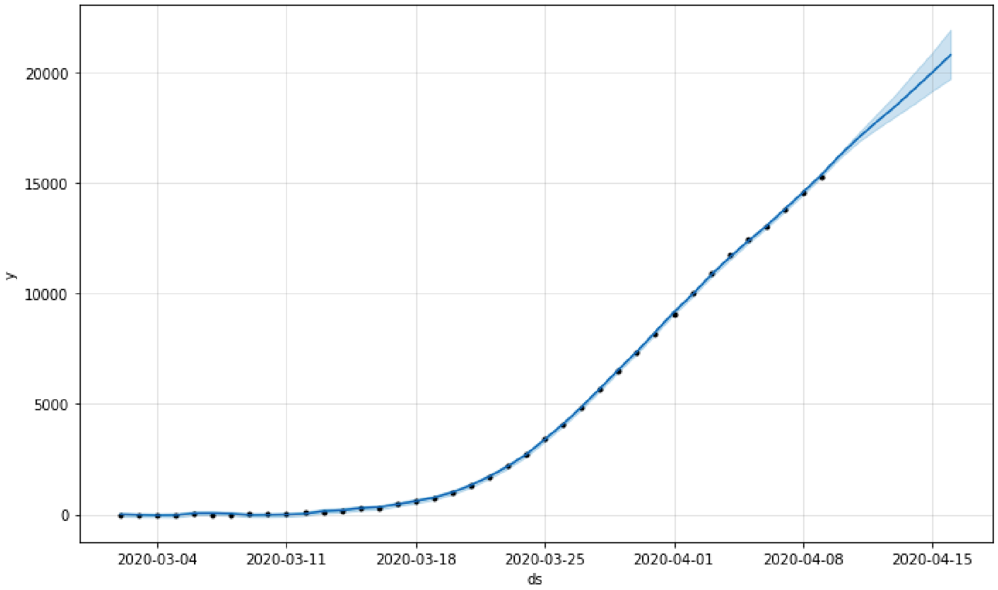
```
3]: forecast = m.predict(future)
    forecast[['ds', 'yhat', 'yhat_lower', 'yhat_upper']].tail()
```

3]:

	ds	yhat	yhat_lower	yhat_upper
41	2020-04-12	17767.930430	17433.223880	18108.752432
42	2020-04-13	18455.880296	18001.488915	18916.825885
43	2020-04-14	19220.056998	18584.499510	19890.919175
44	2020-04-15	19991.490701	19146.497124	20872.768170
45	2020-04-16	20786.672904	19717.266571	21796.610866

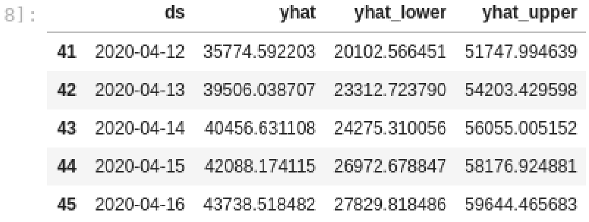


```
4]: deaths_forecast_plot = m.plot(forecast)
```



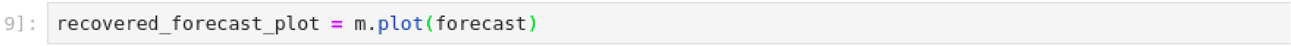
Observation: The projection of "deceased cases" to April 12 according to the Prophet sample indicates 17,767 cases, at the time of writing this report. The official figures of today according to national authorities indicate, whose deceased cases of 16,972 confirmed cases. However, as there is always a 24-hour delay in the delivery of the information, the forecast information should be confirmed on April 13 for greater security.

3)Number cases Cured/Recovered

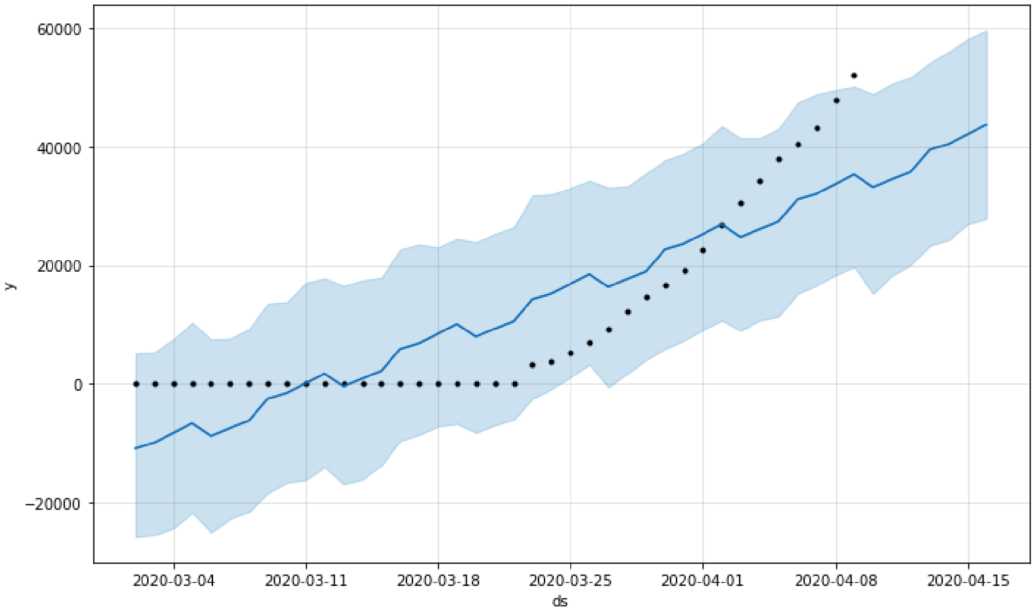


8]:

	ds	yhat	yhat_lower	yhat_upper
41	2020-04-12	35774.592203	20102.566451	51747.994639
42	2020-04-13	39506.038707	23312.723790	54203.429598
43	2020-04-14	40456.631108	24275.310056	56055.005152
44	2020-04-15	42088.174115	26972.678847	58176.924881
45	2020-04-16	43738.518482	27829.818486	59644.465683



```
9]: recovered_forecast_plot = m.plot(forecast)
```



Observacion

Observation: The projection of “recovered cases” to April 12 according to the Prophet sample indicates in the highest part 51,747 cases, at the time of writing this report. Today's official figures according to national authorities indicate the recovered cases of 62,391 cases. However, as there is always a 24-hour delay in the delivery of the information, the forecast information should be confirmed on April 13 for greater security.

5. Conclusions

With the data indicated above, many to be confirmed and others still to be discovered, there remains the feeling that there was more to look for, the data said more things than I am able to observe.

I hope that I can be of help to people who are curious and willing to cross the bridge from ignorance to knowledge.

The lack of consistency in the data has been highlighted in this crisis, despite the fact that we live in the information age. Sometimes it seems that this is the decade of the "infodemic", and if so, it is unlikely that we can use the data in the service of society.

For the good of humanity, data science becomes effective when it is able to put useful information at the service of people that could currently help save lives.