INDEX

Pg nos.

Project by:
MSc Data Science
Asmita Ingle
Nilprabha Khandagale
Lobhana Kirme
Sayali Harer
Shruti Desai

1. Project Description

Coronaviruses are a large family of viruses which may cause illness in animals or humans. In humans, several coronaviruses are known to cause respiratory infections ranging from the common cold to more severe diseases such as Middle East Respiratory Syndrome (MERS) and Severe Acute Respiratory Syndrome (SARS). The most recently discovered coronavirus causes coronavirus disease COVID-19 – WHO.

People can catch COVID-19 from others who have the virus. This has been spreading rapidly around the world and Italy is one of the most affected country. In Italy, the **first two cases** of coronavirus (covid-19) were registered at the **end of January 2020**. Then, since **22nd February**, the epidemic started to spread quickly among the Italian population. As **of 24th May 2020, Italy recorded 230 thousand cases of coronavirus (covid-19)**, representing one of the most affected countries worldwide. Furthermore, Italy counted 140.5 thousand people recovered from coronavirus or discharged from hospitals as of 24th May 2020. On **March 8, 2020 Prime Minister of Italy** announced a sweeping **coronavirus quarantine** early Sunday, restricting the movements of about a quarter of the country's population in a bid to limit contagions at the epicenter of Europe's outbreak. So our project is about the outbreak of coronavirus pandemic in the most affected country in the world **"Italy".**

During the current COVID-19 pandemic, there have been various efforts to forecast and analyze the infection cases, deaths and recoveries. Various methods and models have been adopted or developed for different contexts and purposes. However, the prediction of future is uncertain by nature. No model or data can accurately represent the complex, dynamic and heterogeneous realities of the pandemic in Italy or any other country in the world. In this case, we do not aim to make perfect predictions about the future or test how accurate it will be. Instead, to address the uncertainty of predictions in dynamic real-world scenarios, we explore the potentials of using exploratory visual analysis and predictive monitoring.

2. Problem statements

1. How does age affect the impact of covid cases?(mortality rate, number of positive cases)
2. Does covid have different effects on males and females?
3. Does any type of previous basic symptoms of diseases have any effect on covid patients?
4. Distribution of death, recovery and positive cases area wise distribution. Which region in Italy is most affected?
5. Daily deaths and recoveries.
6. Latest mortality rate and recovery rate. Also over all death and recovery rate in various regions of Italy.
7. Are patients in intensive care unit and death rate correlated?
8. Predicting the trend of deaths, positive cases and recoveries in the future using time series (date-wise)?
9. Comparison between various countries and Italy.

3. Research papers referred

https://www.maa.org/press/periodicals/loci/joma/the-sir-model-for-spread-of-disease-the-differential-equation-model

https://www.ejgm.co.uk/download/research-on-covid-19-virus-spreading-statistics-based-on-the-examples-of-the-cases-from-different-7869.

4. Data source used

This dataset is from **https://github.com/pcm-dpc/COVID-19** collected by Sito del Dipartimento della Protezione Civile - Emergenza Coronavirus: la risposta nazionale

- This dataset has following files
- **Covid_italy_region.csv** - Region level data of COVID-19 cases
- **Covid_age.csv** Distribution of cases and deaths by age classes and gender for every region.
- **Covid-disease.csv** Pre-existing chronic pathologies observed in patients deceased by COVID19. Last 4 rows are records of patients deceased and detection on how comorbidities they have.
- Git Hub -Covid 19-Master [ISTAT]- 'dpc-covid19-ita-andamento-nazionale.csv'
- http://ourworldindata.org **worldcovid.csv**

For the question of comparison between various countries we have used a different data set which contains the data of all countries together. (This data has been extracted from the world dataset for specific countries)

5. Data Description

The **covid19 region** file **contains 17 columns** and **1911 rows.**

Data Description for region file:

NAME	DESCRIPTION
Sno	name of the country (Italy)
Date	record of dates in YYY-MM-DD HH:MM:SS format, from 24th February to 24th May
Country	name of the country (Italy)
RegionCode	code of the region in Italy
RegionName	names of the regions in Italy
Latitude	latitude
Longitude	Longitude
HospitalizedPatients	Hospitalized patients with symptoms of covid on that date
IntensiveCarePatients	Intensive Care Patients
TotalHospitalizedPatients	Total hospitalised patients (both Hospitalised + Intensive Care Patients)
HomeConfinement	home confined patients
CurrentPositiveCases	Total amount of current positive cases (Hospitalised patients + Home confinement)
NewPositiveCases	New amount of current positive cases (Hospitalised patients + Home confinement)
Recovered	number of recovered cases
Deaths	deaths due to corona cases
TotalPositiveCases	Total amount of positive cases
TestsPerformed	Number of tests performed

Data Description of Covid_age.csv and Covid-disease.csv

NAME	DESCIPTION
age_classes	Unique age values
male_cases	Total values in Male Cases
male_deaths	Deaths in male cases
female_cases	Total values in Female Cases
female_deaths	deaths in Female cases
total_cases	Total positive cases
total_deaths	Total Deaths cases
region_wise_death-	Total Deaths values Region wise

Data Description for world data:

NAME	DESCRIPTION
iso_code	ISO country code
location	country name,
date	date,
total_cases	total number of confirmed/positive cases
new_cases	daily new added positive cases
total_deaths	total number of deaths
new_deaths	- daily new added death cases
total_cases_per_million	– total number of positive cases per million population,
new_cases_per_million	daily added number of positive cases per million population
total_deaths_per_million	- total number of deaths per million population
new_deaths_per_million	daily added number of deaths per million population
total_tests	total number of tests conducted till date
new_tests	daily count of tests,
total_tests_per_thousand	total number of tests per thousand people

new_tests_per_thousand	daily count of tests per thousand people
tests_units	test performed or not
population	– total number of population in the country
population_density	number of individuals per unit geographic area (population density
median_age	– middle most age
aged_65_older	– older than 65age
aged_75_older	older than 75age
gdp_per_capita	GDP (gross domestic product) per capita
extreme_poverty	– extreme dearth/poverty,
cvd_death_rate	CVD date rate per 100k population,
diabetes_prevalence	diabetic patients among confirmed cases
female_smokers	female smokers among confirmed cases
male_smokers	male smokers among confirmed cases,
handwashing_facilities	facility of washing hands,
hospital_beds_per_100k	– hospital beds per 100k population,
day	– day,
Month	month

6. Technologies used

Python (Numpy and Pandas in python)
R studio

7. Data Pre-processing required

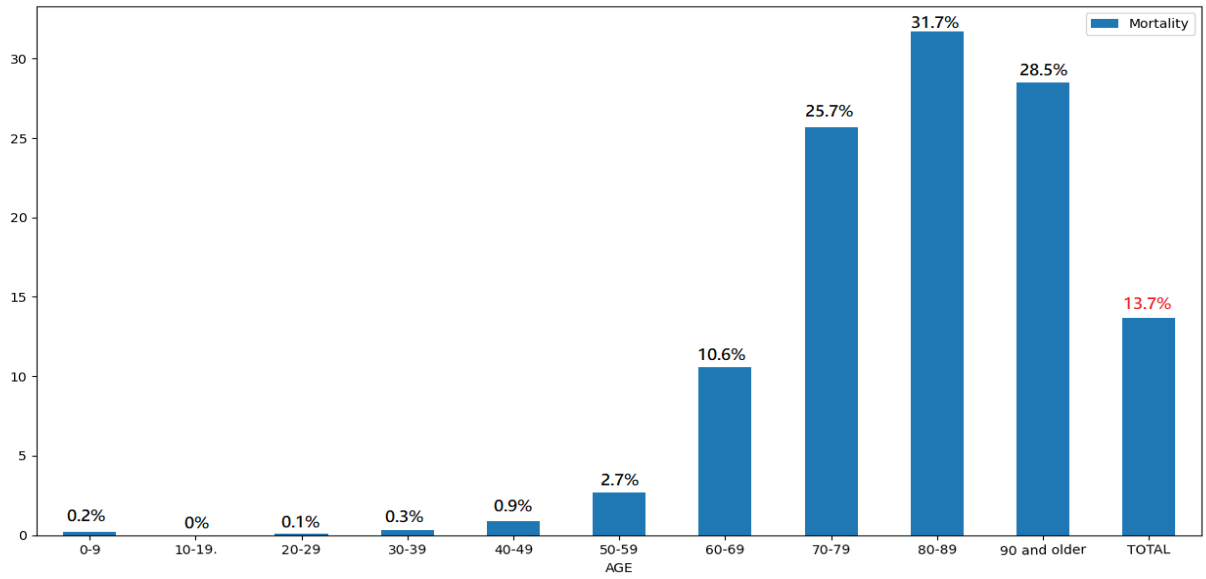
Normalization of the date column as it had date along with the time (which removed the time).Also removed the unnecessary columns like sno, region code, country name, diabetes_prevalence, female_smokers, male_smokers etc.

8. Visual Exploration and
9. Observations

a. Technologies used for visualization

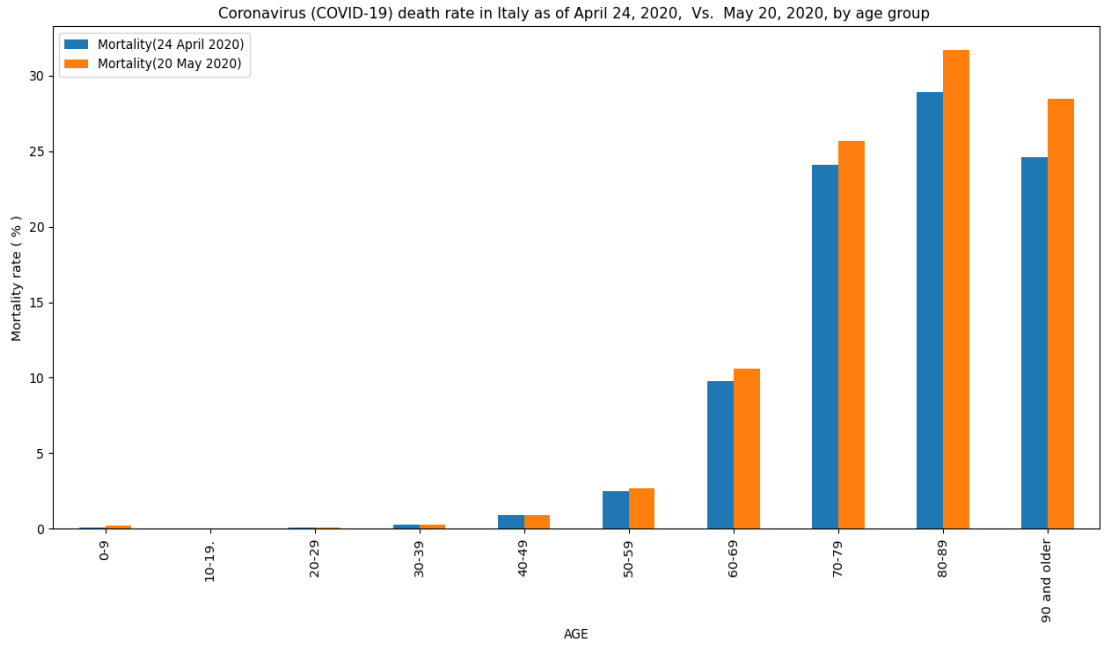
- Plotly
- seaborn
- matplotlib.pyplot
- plotly.graph_objects
- fbprophet import Prophet
- fbprophet.plot import plot_plotly, add_changepoints_to_plot
- math import sqrt
- pmdarima import auto_arima
- sklearn.metrics import mean_squared_error(for error)

1. How does age affect the impact of covid cases? (Mortality rate, number of deaths)



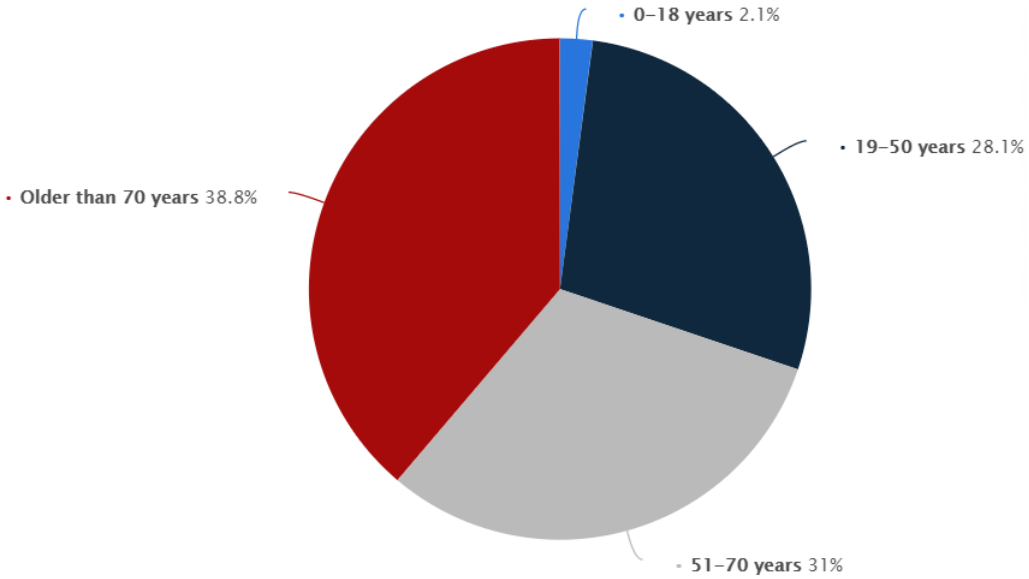
The spread of coronavirus (COVID-19) in Italy has hit mostly people over 50 years of age. Also, the virus claimed approximately 32 thousand lives since it entered the country between the end of January and the beginning of February 2020. As the chart shows, mortality rate appeared to be higher for the elderly patient. In fact, for people between 80 and 89 years of age, the mortality rate was almost 32 percent. For patients older than 90 years this figure was 28.5 percent. Overall, the mortality rate of coronavirus in Italy reached 13.7 percent, higher than that registered in other countries. Italians over the age of 70 represents more than 87% deaths.

The plot below shows the comparison of deaths according to ages as of 24th April and 20th March



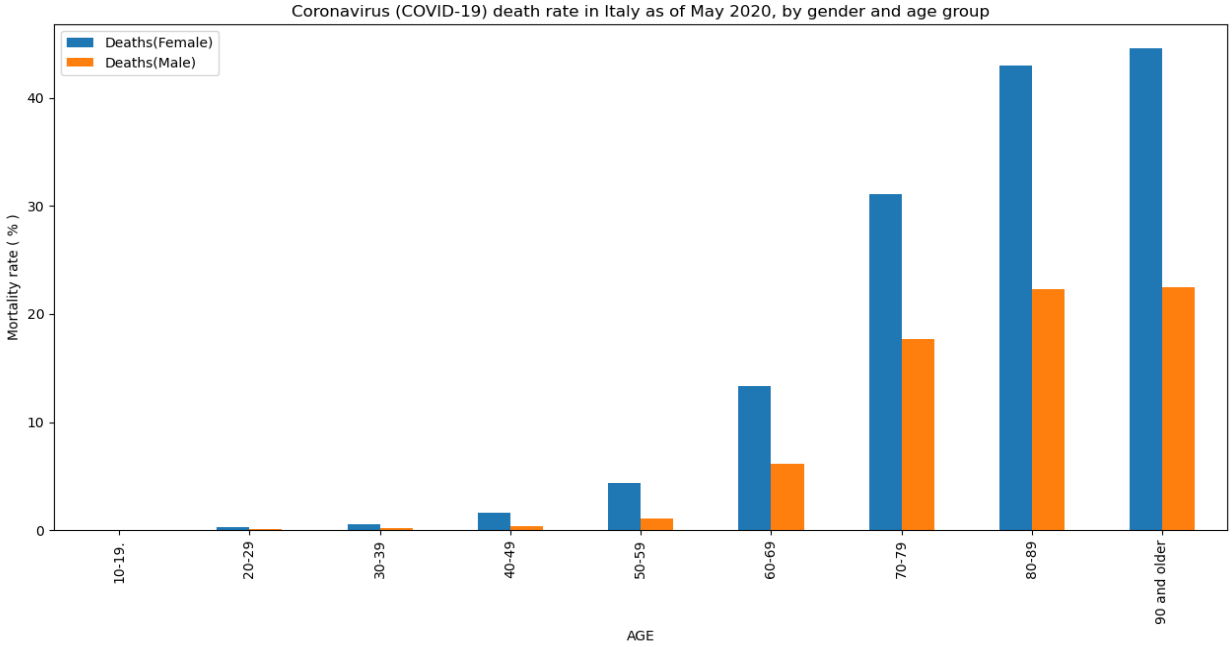
There is a drastic increase in COVID 19 mortality rate between two dates (April 24, 2020 and May 20, 2020)

Distribution of Coronavirus cases in Italy as of May 20, 2020, by age group



The spread of the Coronavirus (COVID-19) in Italy has hit mostly people over 50 years of age. In fact, as the chart shows, 70 percent of individuals infected with the virus was over 50 years old. Only one in four individuals who contracted the virus were between 19 and 50 years old.

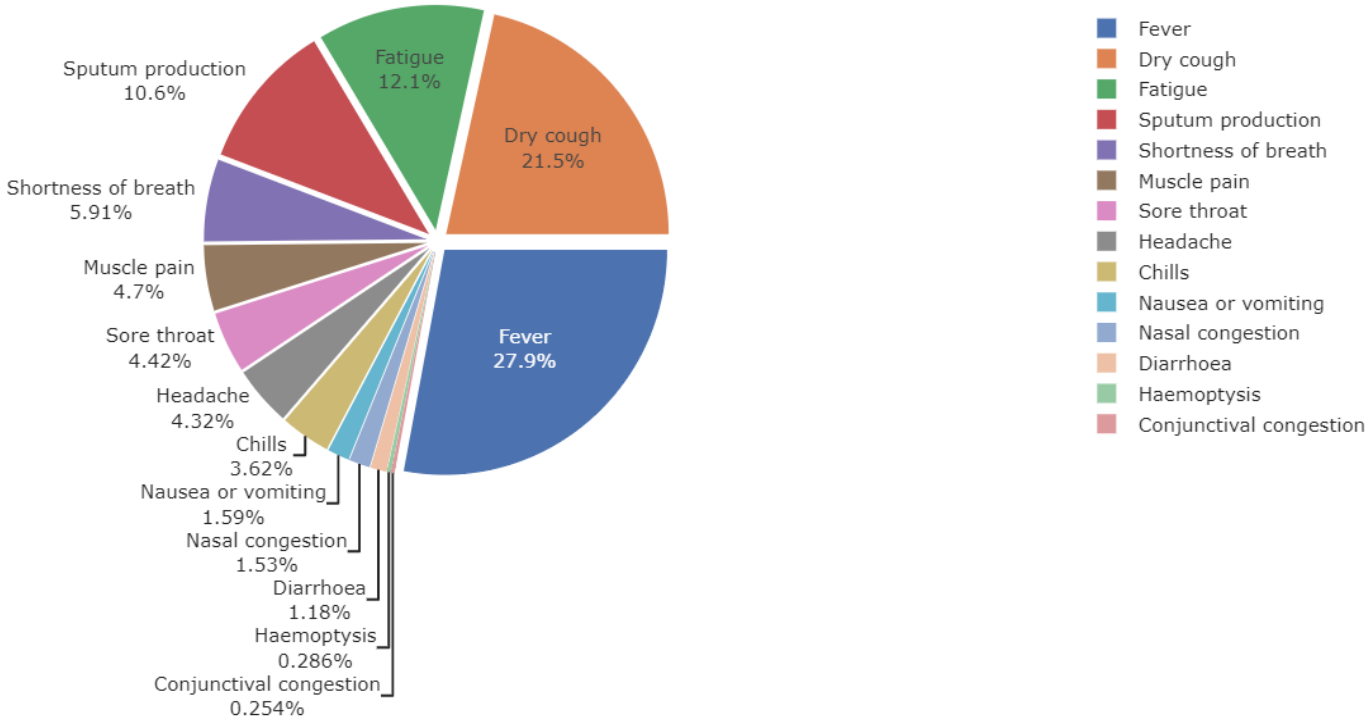
2. Does covid have different effects on males and females?



After the outbreak of the coronavirus (COVID-19) in Italy, many people died after contracting the infection. An in depth study on 30 thousand coronavirus deaths revealed that the fatality rate is much higher for men. In fact, if the mortality rate for female patients was 9.9 percent, the figure for male patients was 17.4 percent, nearly twice as high. The chart shows how this gap was recorded among all age groups. More than 70% of Italy coronavirus deaths have been among men.

At least 3,400 people in the Italy have died of devastating disease and it is announced, it had a higher deaths toll than China . But less than 1,000 of them have been women.
Men are also more likely to pick up the infection in the first place and account of 60 %

3. Does any type of previous basic symptoms of dieses have any effect on covid patients?

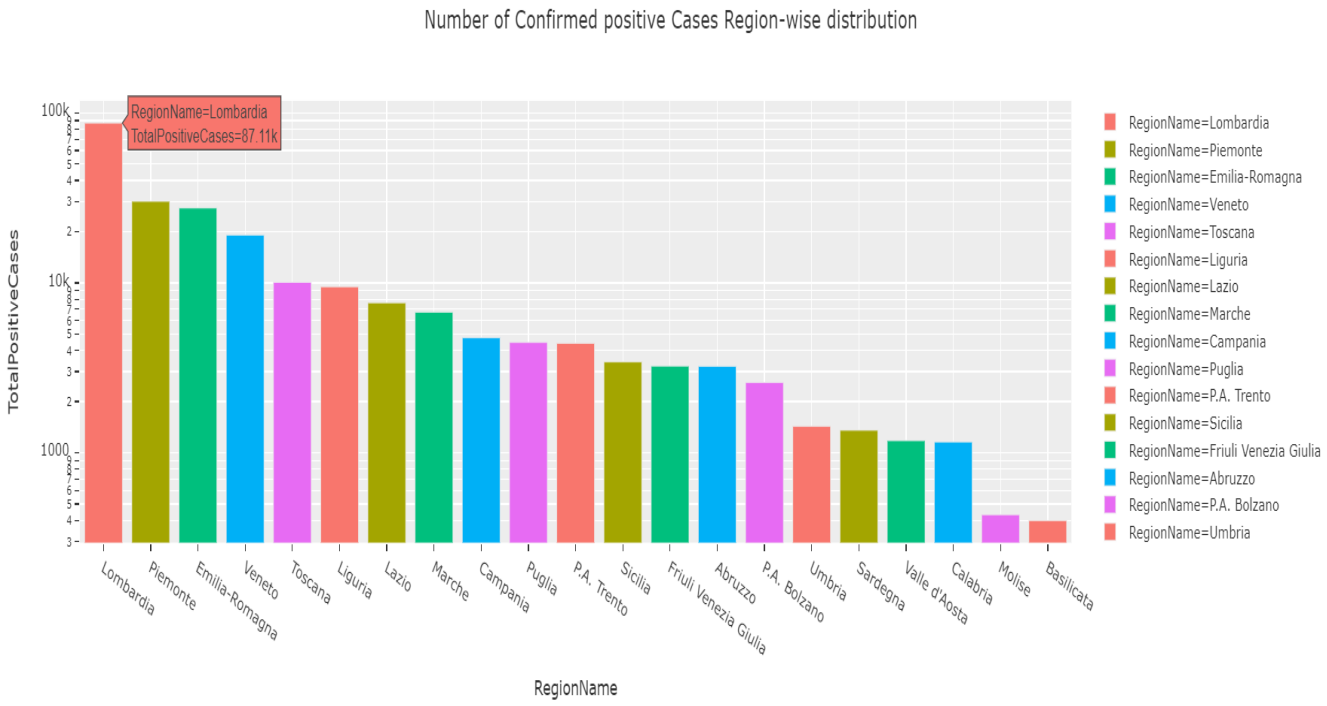


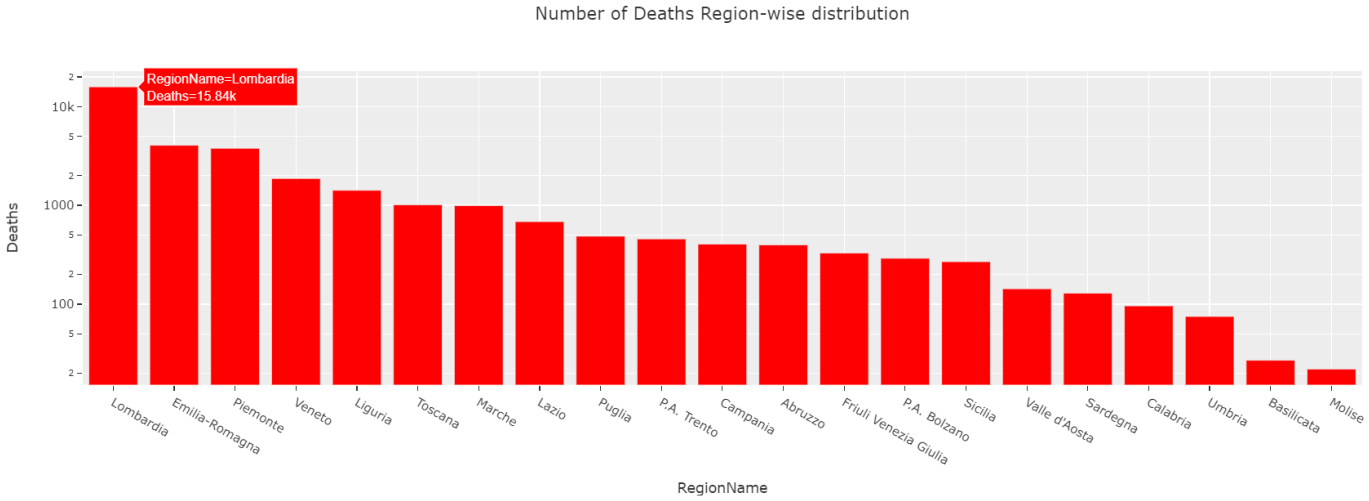
From the above graphs we observe that, the most common symptoms of COVID-19 are fever, dry cough, and tiredness. Other symptoms that are less common and may affect some patients include aches and pains, nasal congestion, headache, conjunctivitis, sore throat, diarrhea, loss of taste or smell or a rash on skin or discoloration of fingers or toes. These symptoms are usually mild and begin gradually. Some people become infected but only have very mild symptoms. Around 1 out of every 5 people who gets COVID-19 becomes seriously ill and develops difficulty breathing.

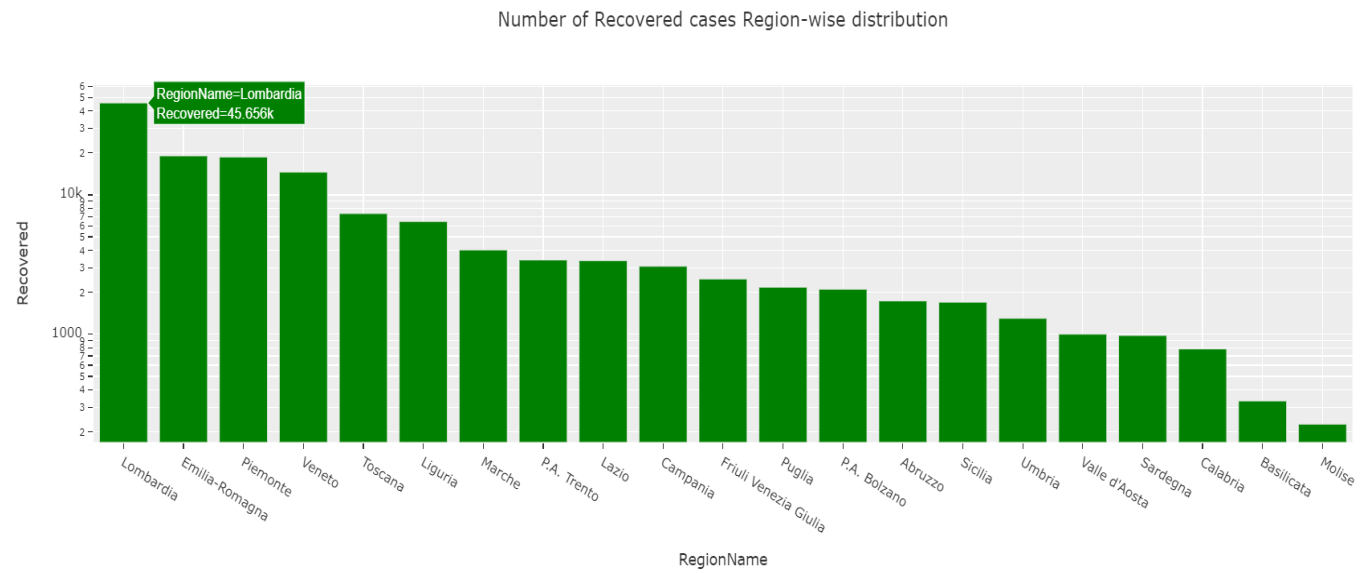
4. Distribution of death, recovery and positive cases area wise. Which region in Italy is most affected?

	RegionName	TotalHospitalizedPatients	Recovered	Deaths	TotalPositiveCases	TestsPerformed
8	Lombardia	4214	45656	15840	87110	396992.000000
13	Piemonte	1358	18694	3783	30180	188716.000000
4	Emilia-Romagna	602	19046	4055	27558	182002.000000
20	Veneto	196	14557	1869	19086	304944.000000
17	Toscana	186	7349	1013	10062	164469.000000
7	Liguria	265	6437	1419	9480	53230.000000
6	Lazio	1149	3374	684	7627	187994.000000
9	Marche	114	4028	994	6714	60949.000000
3	Campania	319	3076	405	4749	80942.000000
14	Puglia	221	2178	487	4458	71032.000000
12	P.A. Trento	31	3412	457	4404	45860.000000
16	Sicilia	100	1701	269	3423	117242.000000
5	Friuli Venezia Giulia	63	2495	329	3236	72104.000000
0	Abruzzo	152	1736	398	3226	45516.000000
11	P.A. Bolzano	35	2107	291	2593	27938.000000
18	Umbria	17	1302	75	1430	45131.000000
15	Sardegna	54	982	129	1356	43661.000000
19	Valle d'Aosta	24	1003	143	1178	11035.000000
2	Calabria	44	786	96	1157	60693.000000
10	Molise	8	227	22	432	12751.000000
1	Basilicata	14	333	27	399	25431.000000

As we see in the table above that the area mostly hit by the virus was the North particularly the region Lombardia in Italy in all positive cases, recovery and death as of 24th May,2020.





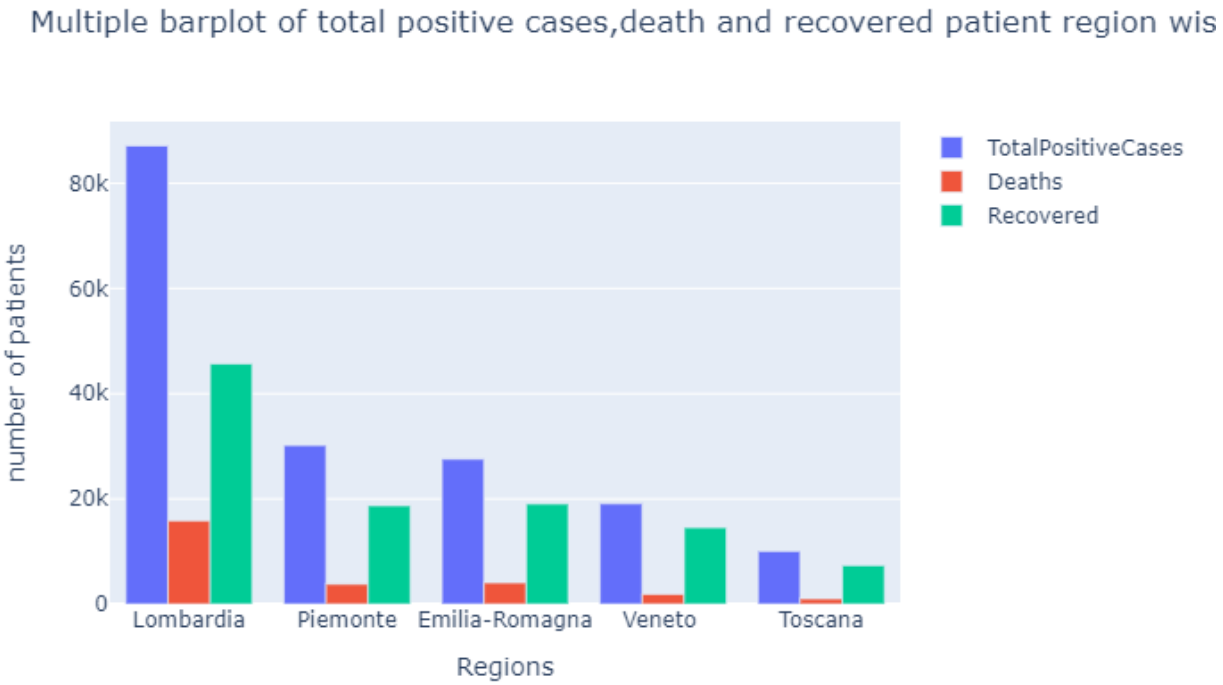


As the first graph shows **total positive cases** in different regions of Italy.
The second graph shows **deaths** in different regions of Italy.
The third graph shows **recovered cases** in different regions of Italy.

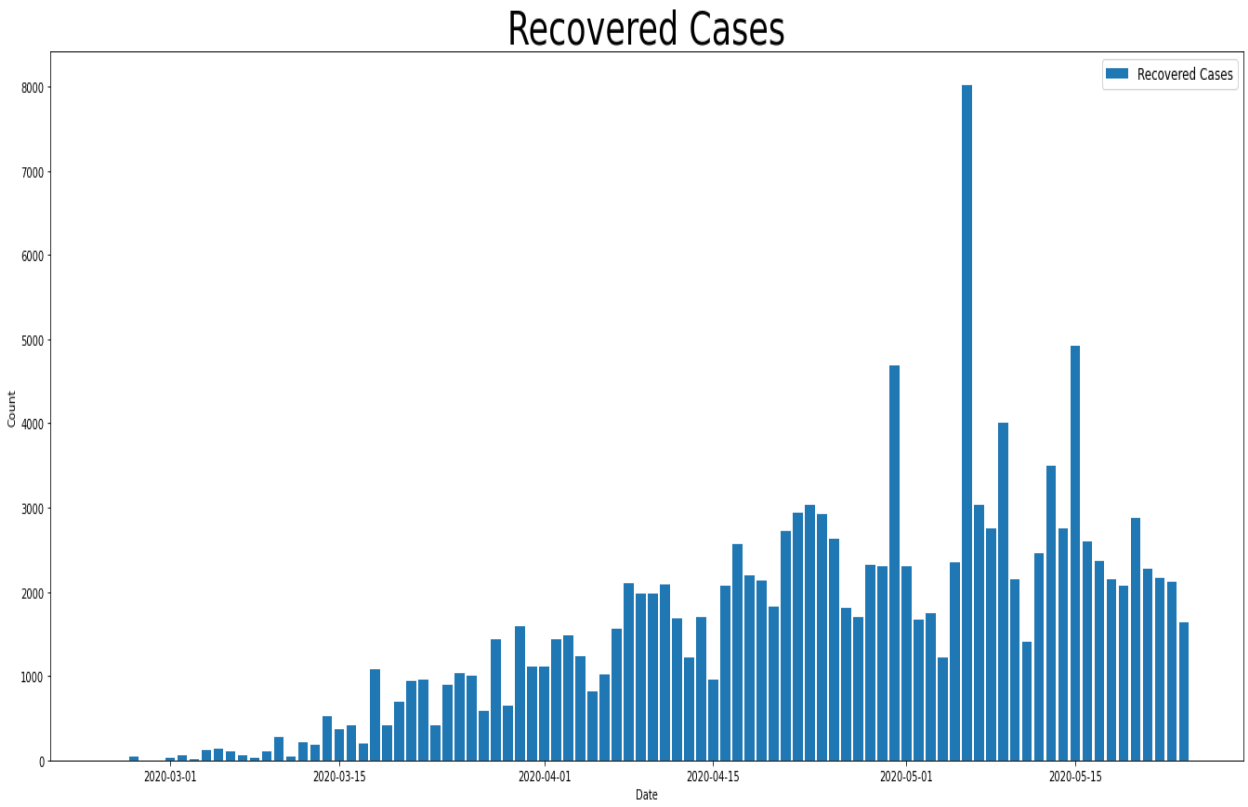
Lombardia has the maximum number **(87k+) total positive cases** of as it is the most infected in cities. As a result the next graph shows that it has the maximum number of **deaths (15840)** and **recovered cases (45656)** by coronavirus. **Piemonte** is the **second** most infected city here followed by some more countries like Emilia Romagna, Veneto, Toscana, Lazio, Liguria, Marche, Friuli V.G., Campania, Sicilia, Puglia, P.A. Trento, Calabria, Umbria, Abruzzo, Sardegna, Molisa, Basilicata, Valle d'Aosta, P.A. Bolzano etc.

But the **recovery and deaths** are more in **Emilia Romagna** as compared with **Piemonte** even though it is 2nd most affected city. All these observations are made as of 24th May 2020.

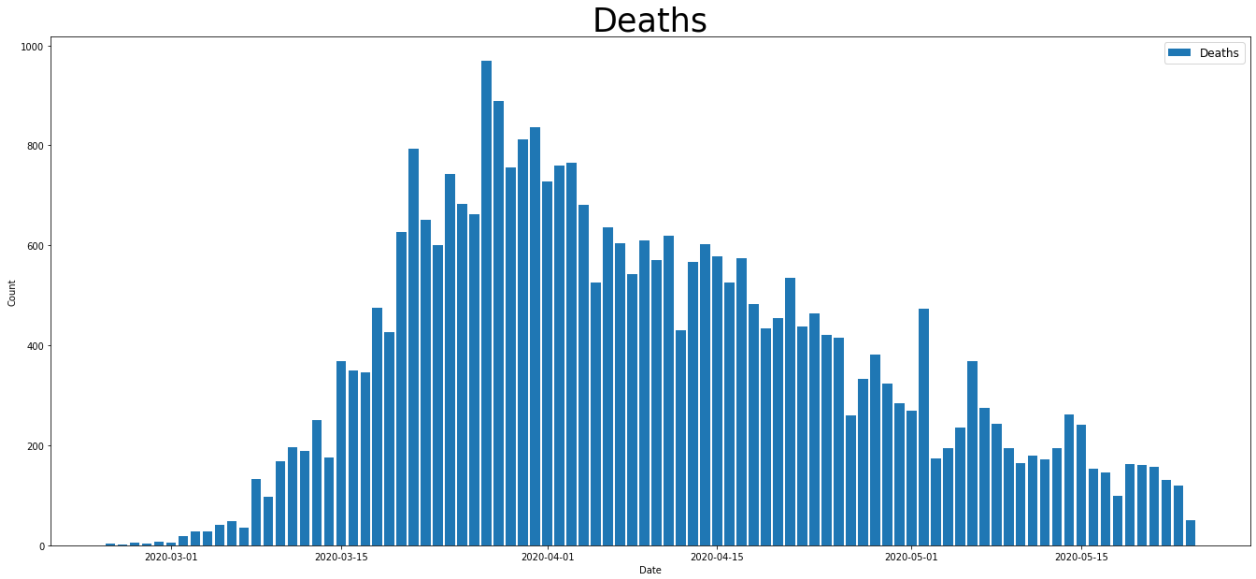
Following multiple barchart shows relative comparison between the **top five cities** that are most affected in Italy .



1. Daily deaths and recoveries.



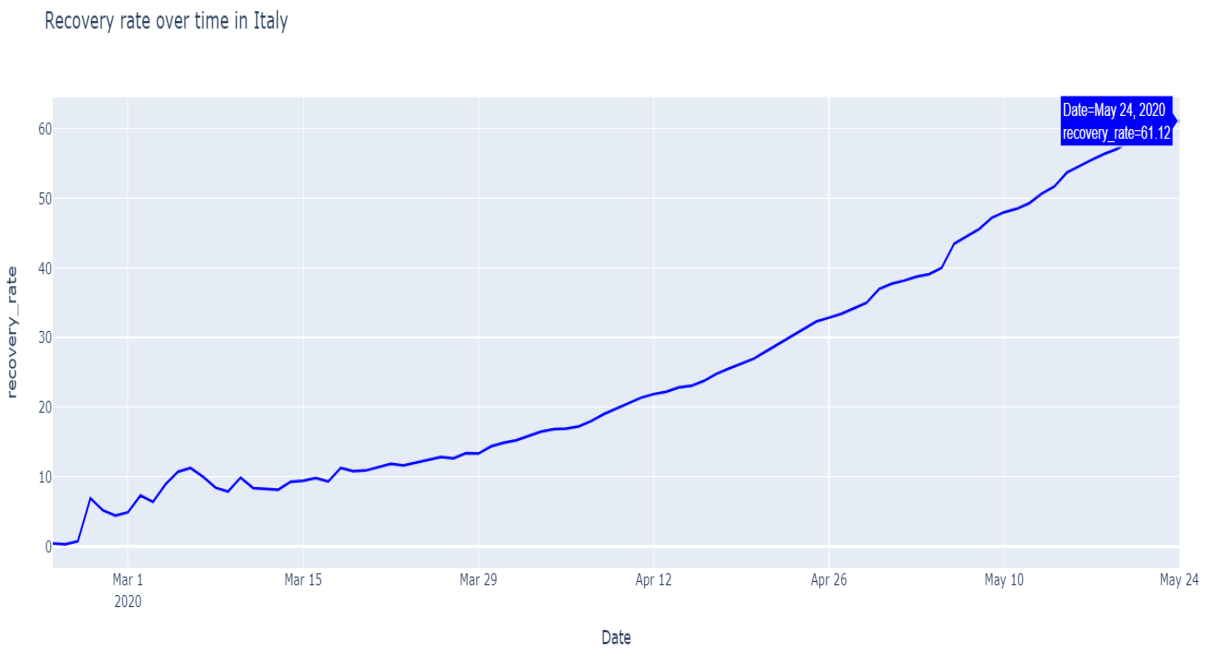
As we can see the above plot shows the daily recoveries all over Italy from 24th February to 24th May. We can observe that there is abrupt highest number recovered patients on 6th of May.

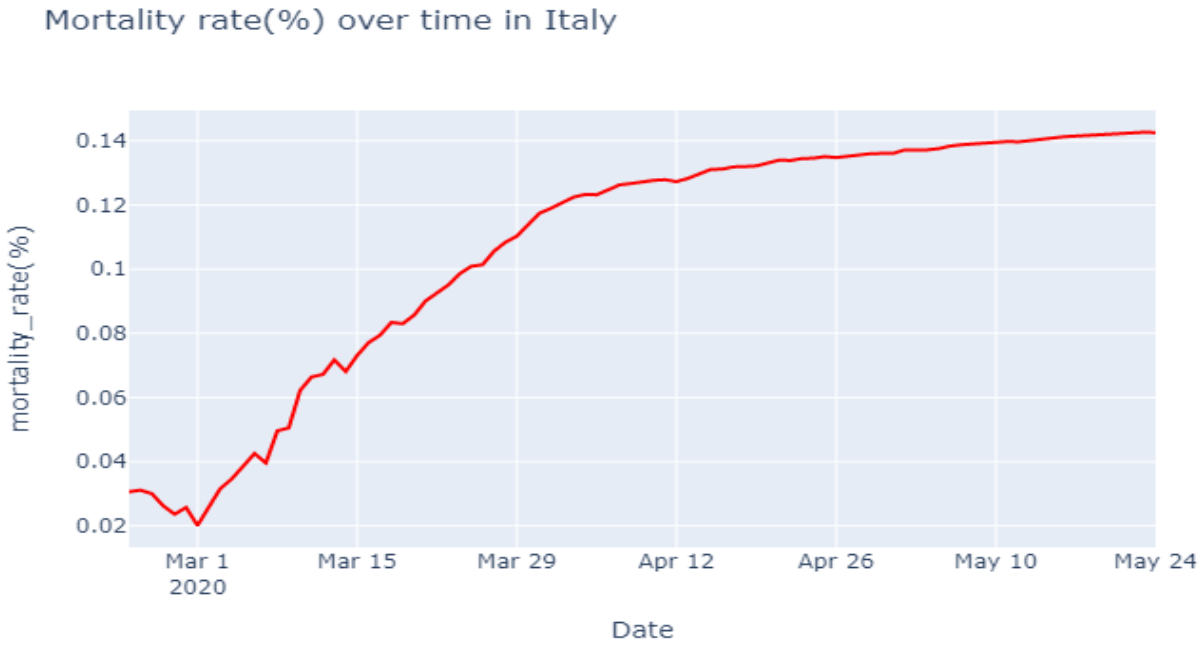


As we can see the above plot shows the daily deaths all over Italy from 24th February to 24th May. This plot show sudden rise in deaths between 15th March to 15th April, 2020 and then the daily deaths are slowly decreasing from April to May.

2. Latest mortality rate and recovery rate. Also over all deaths and recovery rate in various regions of Italy.

LATEST DETAILS AS OF 24TH MAY
The percentage of Confirmation is **10.454591764333458%**
The percentage of Death is **1.4911544997070905%**
The percentage of Death after confirmation is (i.e. mortality rate) **14.263153773199106%**
The percentage of recovery after confirmation is **61.11555830121205%**
We can say that the recovery rate has been improved tremendously over the period of time.





But as we can see along with recovery rate the death rate is also increasing. (Although not so drastic).Mortality rate is comparatively low.

Mortality rate = sum (number of deaths) /sum (number of case) = probability of dying of infected by virus (%)
Recovery rate = sum (number of recovery) /sum (number of case) = probability of recovered cases infected by virus (%)

AVERAGE RATES OVER THE TIME PERIOD OF **24TH FEBRUARY TO 24TH MAY**
Average recovery rate: **35.589716449153244%**
Average mortality rate: **13.138236885947144%**

Overall, the mortality rate of coronavirus in Italy reached **13.7 percent**, higher than that registered in other countries.

3. Are patients in intensive care unit and death rate correlated?

Italy has the highest fatalities due to coronavirus. On March 9, Italy extended the lockdown from certain areas to the whole country when it had more than 5000 confirmed coronavirus cases, out of which 733 were in intensive care and 463 were dead because of the virus. As we can see death of patients are started from mid or late Feb but from 15 March count of deaths is increasing monotonously which is terrifying for country to face. Meanwhile, for intensive care unit count of patients is at its peak between 22 March to 12 April and decreasing after 12 April.
As we can see the that patients in intensive care and deaths are moderately correlated
cor (data$IntensiveCarePatients,data$Deaths)=0.6245294 (this is the numeric value of the correlation)

14. Model building and
15. Analysis

8.Predicting the trend of deaths, positive cases and recoveries in the future using time series (date-wise)?

a.Objective

To get the future trend of attribute of concern that is forecasting

b. Algorithms used. Comparison of algorithms. Why particular algorithm is used.

Autoregressive Integrated Moving Average Model

An ARIMA model is a class of statistical models for analyzing and forecasting time series data. It explicitly caters to a suite of standard structures in time series data, and as such provides a simple yet powerful method for making skillful time series forecasts.

ARIMA is an acronym that stands **for AutoRegressive Integrated Moving Average**. It is a generalization of the simpler AutoRegressive Moving Average and adds the notion of integration.

ARIMA is a very popular statistical method for time series forecasting. ARIMA stands for **Auto-Regressive Integrated Moving Averages**. ARIMA models work on the following assumptions –

- The data series is stationary, which means that the mean and variance should not vary with time. A series can be made stationary by using log transformation or differencing the series.
- The data provided as input must be a univariate series, since arima uses the past values to predict the future values.

ARIMA has three components – AR (autoregressive term), I (differencing term) and MA (moving average term). Let us understand each of these components –

- AR term refers to the past values used for forecasting the next value. The AR term is defined by the parameter 'p' in arima. The value of 'p' is determined using the PACF plot.
- MA term is used to defines number of past forecast errors used to predict the future values. The parameter 'q' in arima represents the MA term. ACF plot is used to identify the correct 'q' value.

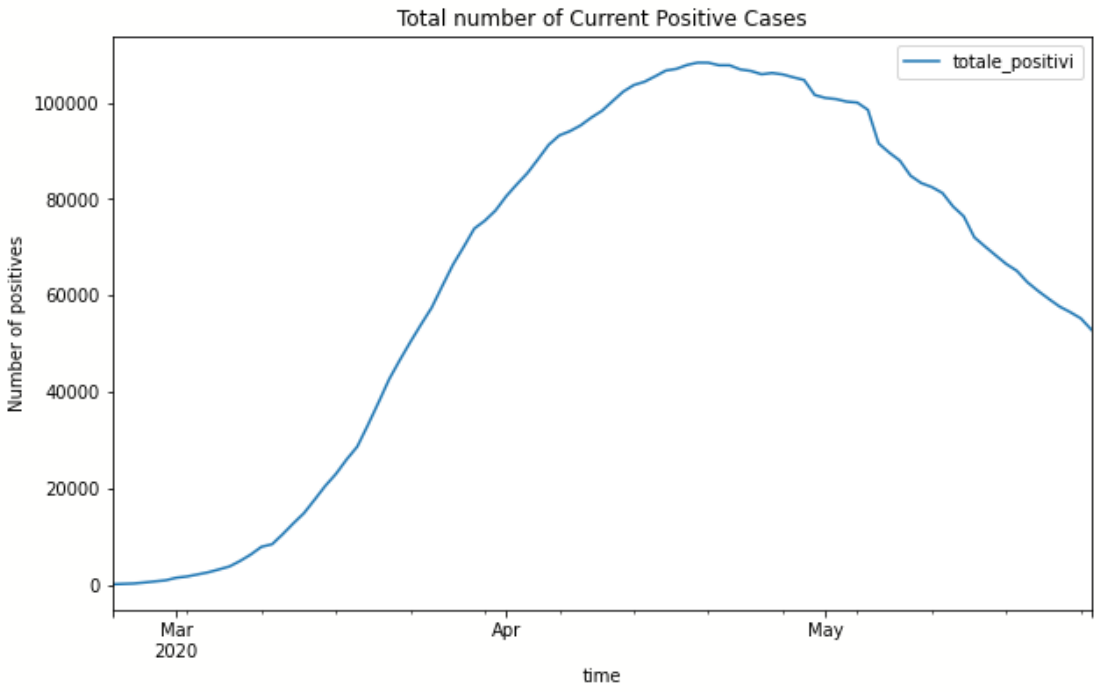
Order of differencing specifies the number of times the differencing operation is performed on series to make it stationary. Test like ADF and KPSS can be used to determine whether the series is stationary and help in identifying the d value.

How does Auto Arima select the best parameters

In the above code, we simply used the .fit() command to fit the model without having to select the combination of p, q, d. But how did the model figure out the best combination of these parameters? Auto ARIMA takes into account the AIC and BIC values generated (as you can see in the code) to determine the best combination of parameters. AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion) values are estimators to compare models. The lower these values, the better is the model.

Check out these links if you are interested in the maths behind *AIC* and *BIC*.

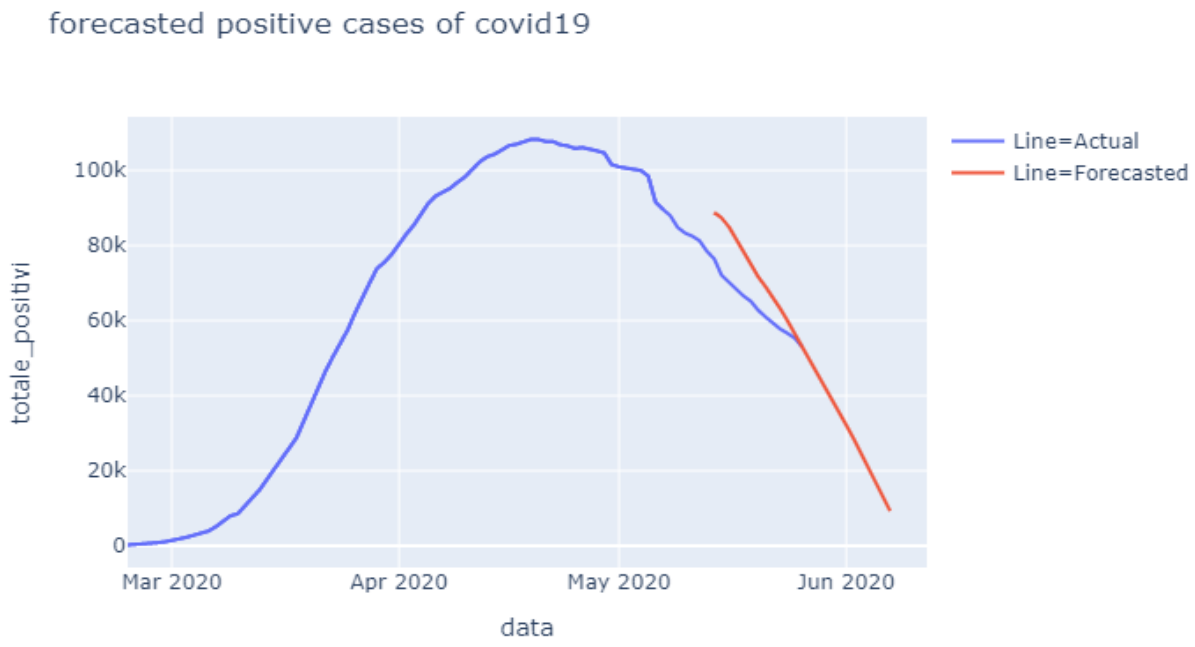
WE ARE USING AUTO ARIMA HERE



The graph tells the trend of Total number of Positive Cases per day from March to May.

FORECASTING ON THE NUMBER OF POSITIVE CASES PER DAY USING AUTO ARIMA

Line plot



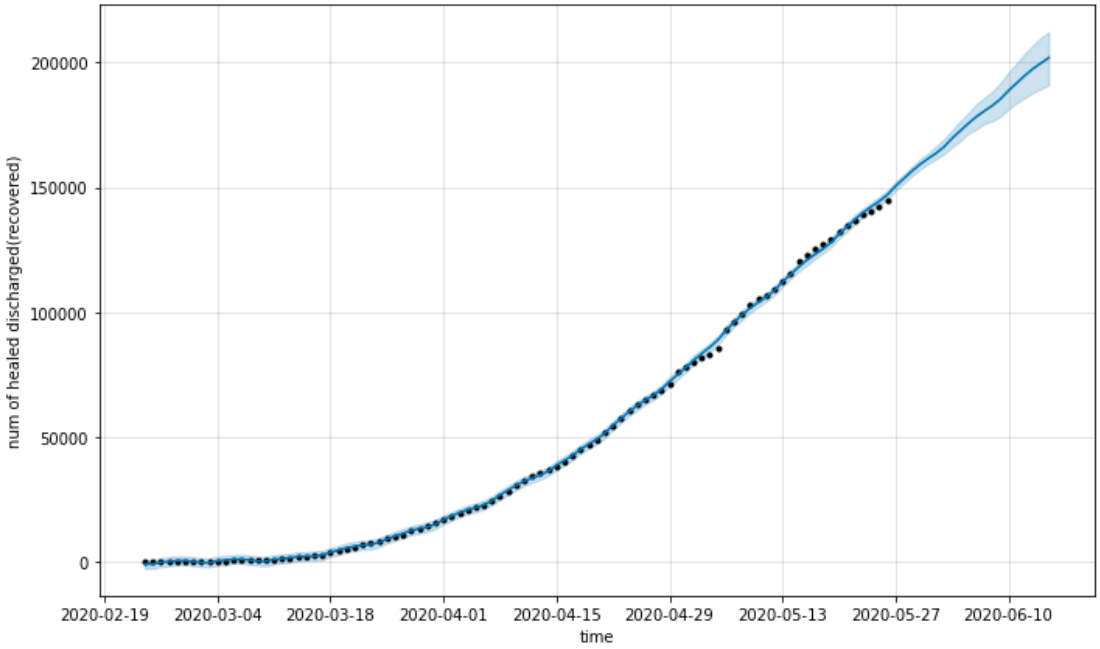
The graph gives the forecasted trend (decreasing) of the positives cases on particular day up to June 7. So we can see that the positive cases is gradually decreasing with time and was higher within the time period of April and May. The red line forecasted that up till June the cases a steeping as low as 10k, which is a good sign for Italy's progress for removing the virus.

Similarly,Forecasting using library Prophet .

The Core Data Science team at Facebook created Prophet, a forecasting library for Python and R, which they open-sourced in 2017.

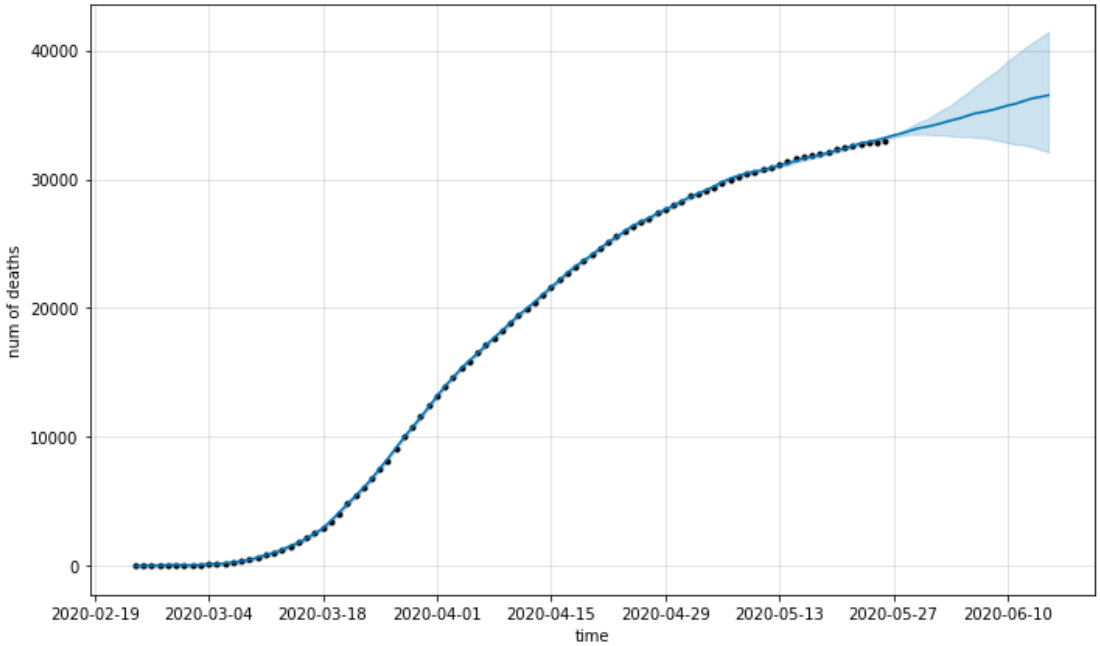
The intent behind Prophet is to "make it easier for experts and non-experts to make high-quality forecasts that keep up with demand." Prophet is able to produce reliable and robust forecasts (often performing better than other common forecasting techniques) with very little manual effort, while allowing for the application of domain knowledge via easily-interpretable parameters.

FORECASTING THE NUMBER OF RECOVERED PATIENTS USING PROPHET:



The graph shows the forecasted trend of the number of recovery cases(increasing) using Prophet. So as the patient recovery is increasing, this is a positive prospective achieved by Italy.

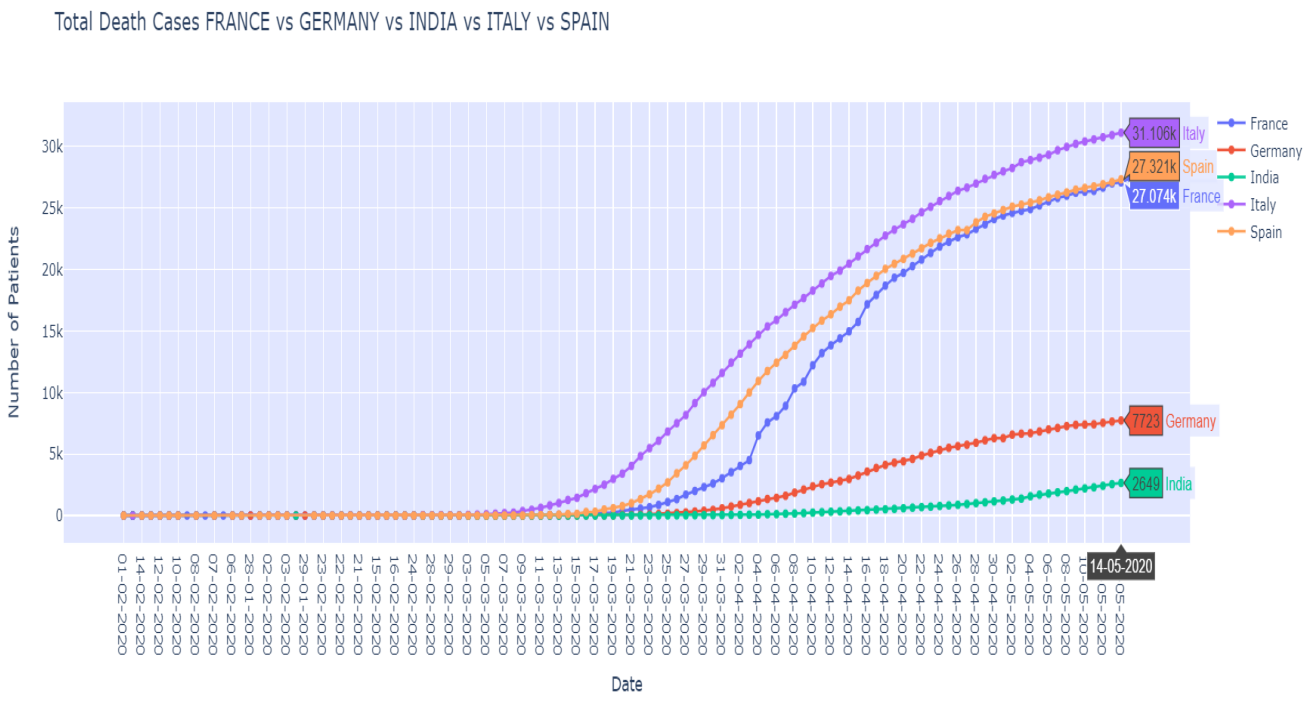
FORECASTING THE NUMBER OF PATIENT DEATHS USING PROPHET:



As we can see in the graph we have forecasted values from 24th May to a little beyond 10th June The graph shows the forecasted trend of the number of death cases(slow increase not steep/slight increase) using Prophet.

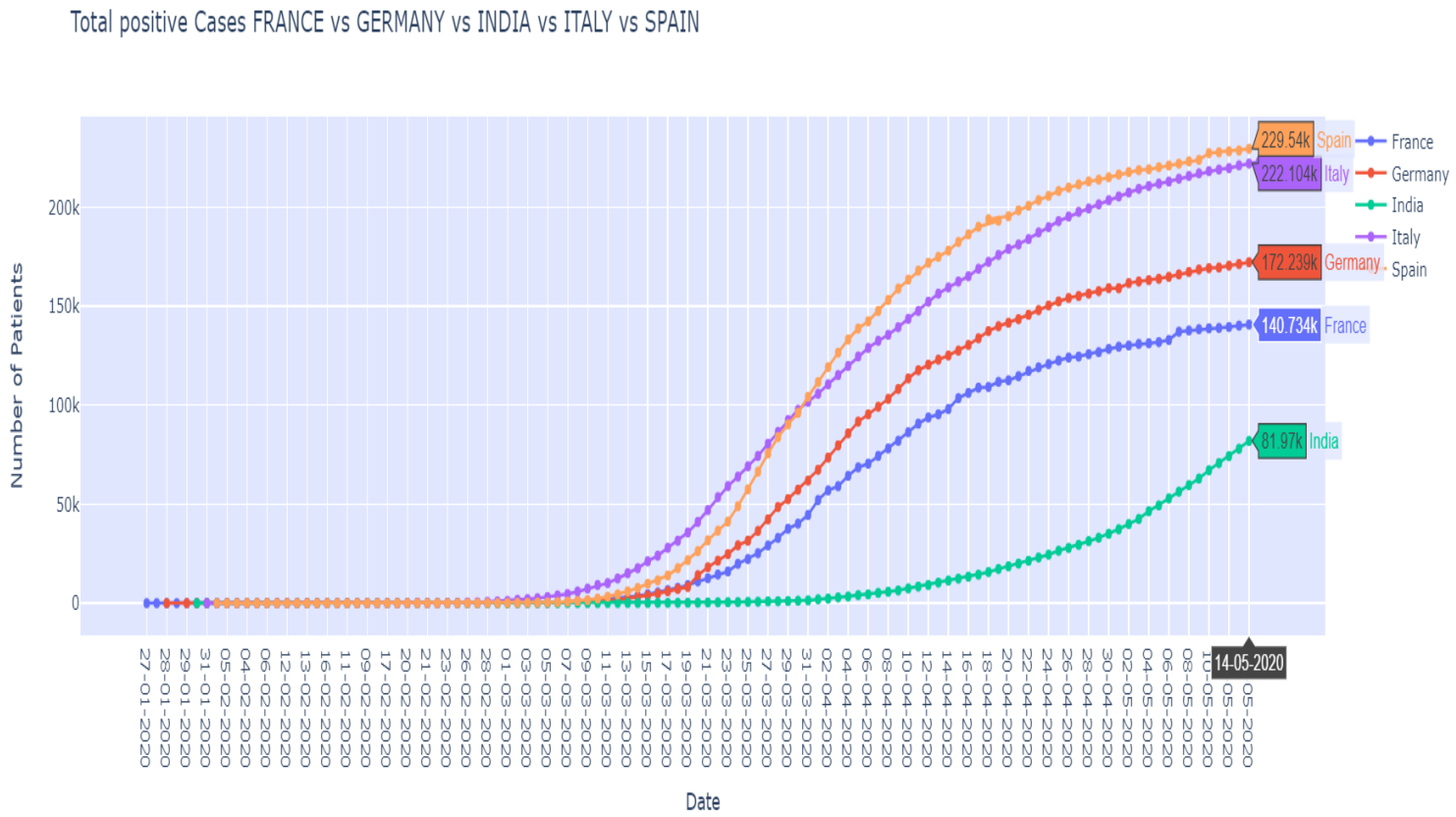
9. Comparison between various countries and Italy.

(According to its deaths and positive patients)



For Death cases - Italy is most affected country among all we are considering. After Italy, Spain and France had high mortality rate. We can see for Italy its rapidly increasing after 19 March and its monotonously increasing after 23 March for Spain.

What is more noticeable is death rate is low in Germany and very low in India compare to other countries. As for now we are considering cases only up to 14th May 2020.



Observations

For Confirmed positive cases in 5 countries - India, Italy, Spain, France and Germany

As we can see by the start of February 2020 all above mentioned countries had started detecting corona patients.

From 15 March to 20 April, all countries had rapid growth of virus that is increasing number of corona patients except India, for India, there's slow increment.

Italy was one of the most affected country but between 27th March to 31st March, it steps back as Spain had sudden rise.

As we can see from 8 May, it is getting steadier for all countries meanwhile India had monotonous increment after 2 April 2020 but very low compare to other 4 countries.

16. Future enhancement and conclusion

This project has vast scope in the future. It can be updated in near future and when requirement for the same arises. The data would be updated time to time, which is better in analysis. As the data is ongoing, it can little difficulties in the data visualization. A few parameters need a modification as they were incomplete. It is very flexible in terms of expansion. Different parts of the world are affected by corona-virus with different intensity, in which Italy was found the worst as of the starting 2 months. Various fields will be drastically change its course of action after due to this virus, so the post-corona will expect these changes:

1. More Contactless Interfaces and Interactions
2. Strengthened Digital Infrastructure
3. AI-Enabled Drug Development
4. Telemedicine
5. More Online Shopping
6. Increased Reliance on Robots
7. More Digital Events
8. Rise in Esports

COVID-19 is not unbeatable. By changing our behaviors, we can slow the spread of the disease. These mechanisms such as quarantining, avoiding crowds, and hand-washing are low-tech; they are not glamorous. But they are effective and they buy time for physicians, nurses, and scientists- who are already the heroes in this fight- to control the fallout. All of us must work together. The pandemic caught many of us by surprise earlier but we have no excuse not to act. But will learn from it, and we will be better prepared.

17. Bibliography

- https://towardsdatascience.com
- https://plotly.com
- https://www.esahq.org/esa-news/analysis-of-covid-19-data-on-numbers-in-intensive-scare-from-italy-european-society-of-anaesthesiology-esa
- https://economictimes.indiatimes.com
- https://github.com
- https://www.sciencedirect.com
- https://www.analyticsvidhya.com
- https://www.kaggle.com
- https://stackoverflow.com
- https://www.machinelearningplus.com
- https://datascience.stackexchange.com