Comparative Analysis Of Covid-19 Spread In Major Impacted Countries

Data-Science Project

16K-3620 Abdul Mannan
16K-3609 Ali Akbar

Submitted To:
Dr. Muhammad Atif Tahir

Introduction

- Covid-19 is a respiratory infection caused by severe acute respiratory syndrome
- The first patient of Covid-19 was identified in Wuhan, China in Dec 2019.
- WHO has declared it a global pandemic.
- No vaccine discovered till yet

Introduction

- Its spread is so quick that just within a few months, the total number of cases of Covid positive patients in Pakistan has increased upto 131,509 people and the death toll is around 2.5k.
- It has now spread to almost all over the world infecting a total number of 7.27M people and taking 413K lives.

Data Retrieval

Data sources are:

1. Worldometers.
2. National Institute of Health Pakistan.
3. Covid-19 Portal of Pakistan.
4. Kaggle repository.
5. John hopkins university repository

We used libraries such as Pandas, Requests and BeautifulSoup to extract the data from the "WorldoMeters" site and then parsed it before saving it onto our file system as CSV files. We also gathered the data from the above mentioned sources in an attempt to gather consistent data.

Data Preparation

- Parsing date

```
df['ObservationDate']=pd.to_datetime(df['ObservationDate'])
```

- Storing date

```
df["Last Update"]= pd.to_datetime(df["Last Update"])
```

- Checking Null values

feature	percentage_of_having_null
SNo	0.000000
ObservationDate	0.000000
Province/State	44.925916
Country/Region	0.000000
Last Update	0.000000
Confirmed	0.000000
Deaths	0.000000
Recovered	0.000000

- Feature generation

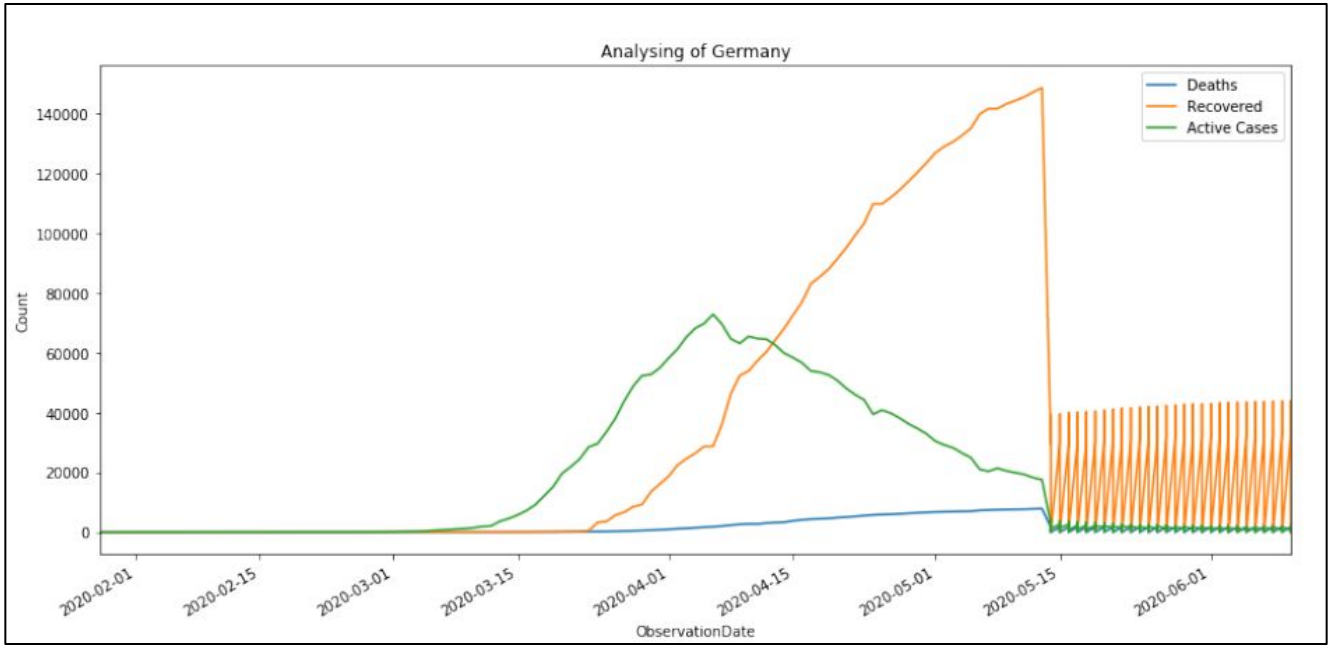
```
df['Active Cases'] = df['Confirmed'] - df['Deaths'] - df['Recovered']
```

- Creating Aggregated Variables for Analysis

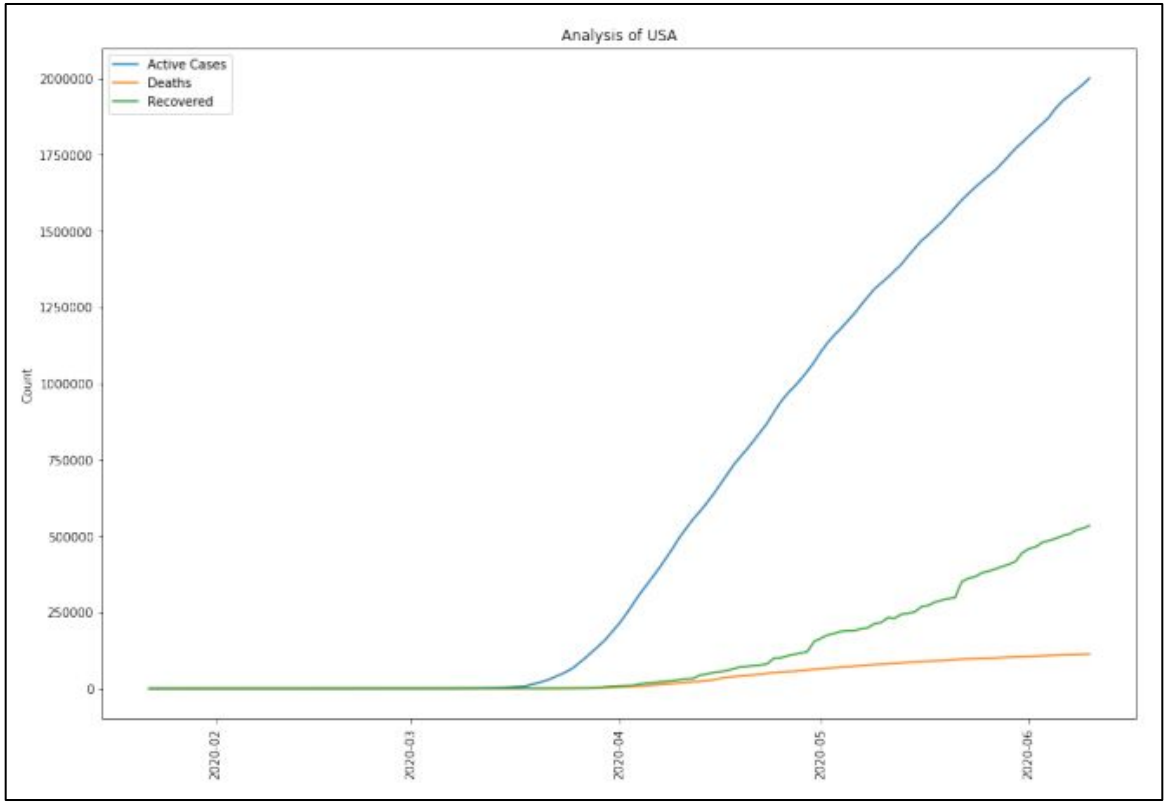
Deep-Diving inside country-wise data

```
uk_agg=pd.pivot_table(uk, index=['ObservationDate'],values=['Confirmed','Active␣
 ↪Cases','Deaths','Recovered'],aggfunc=np.sum)
#uk_agg
```

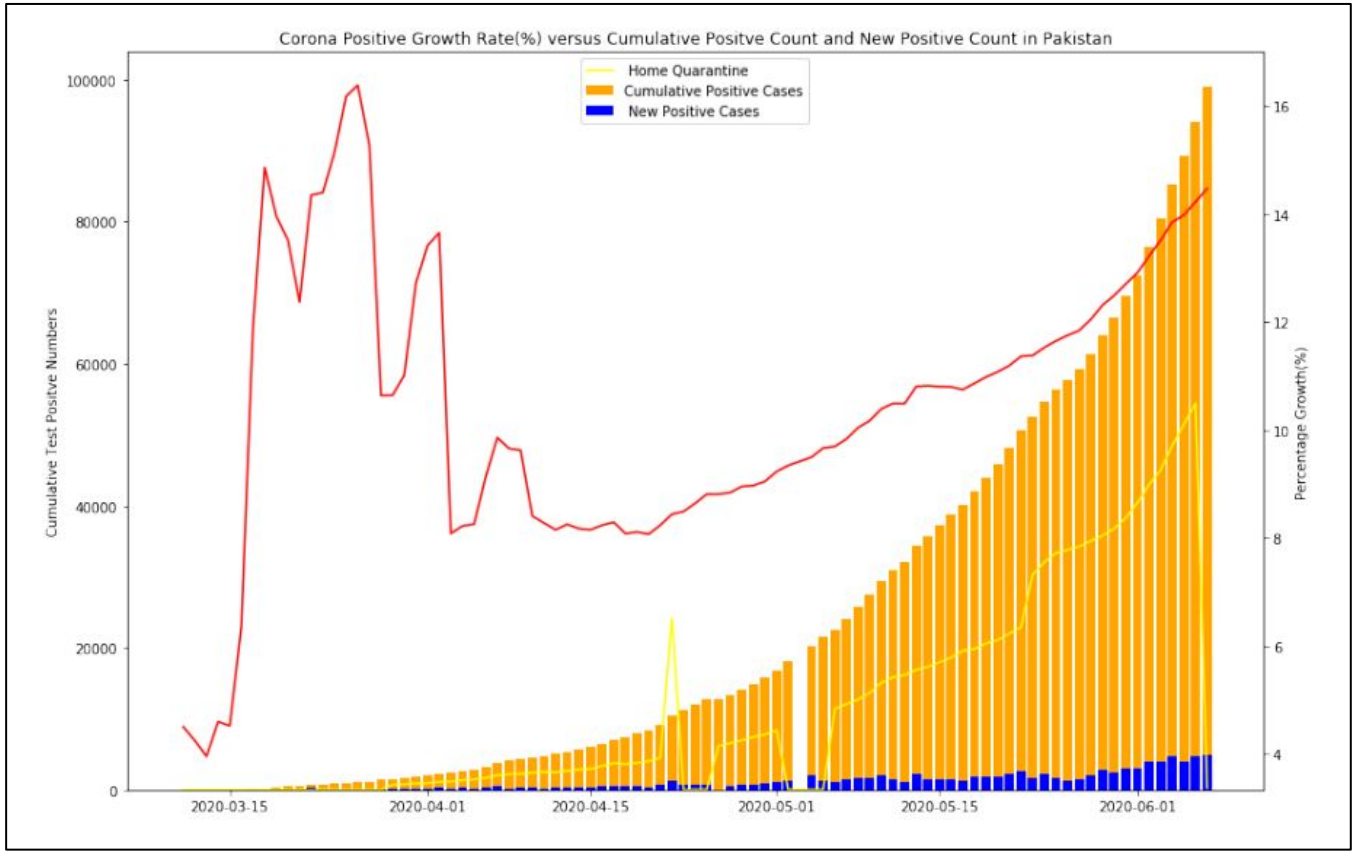
Data Exploration



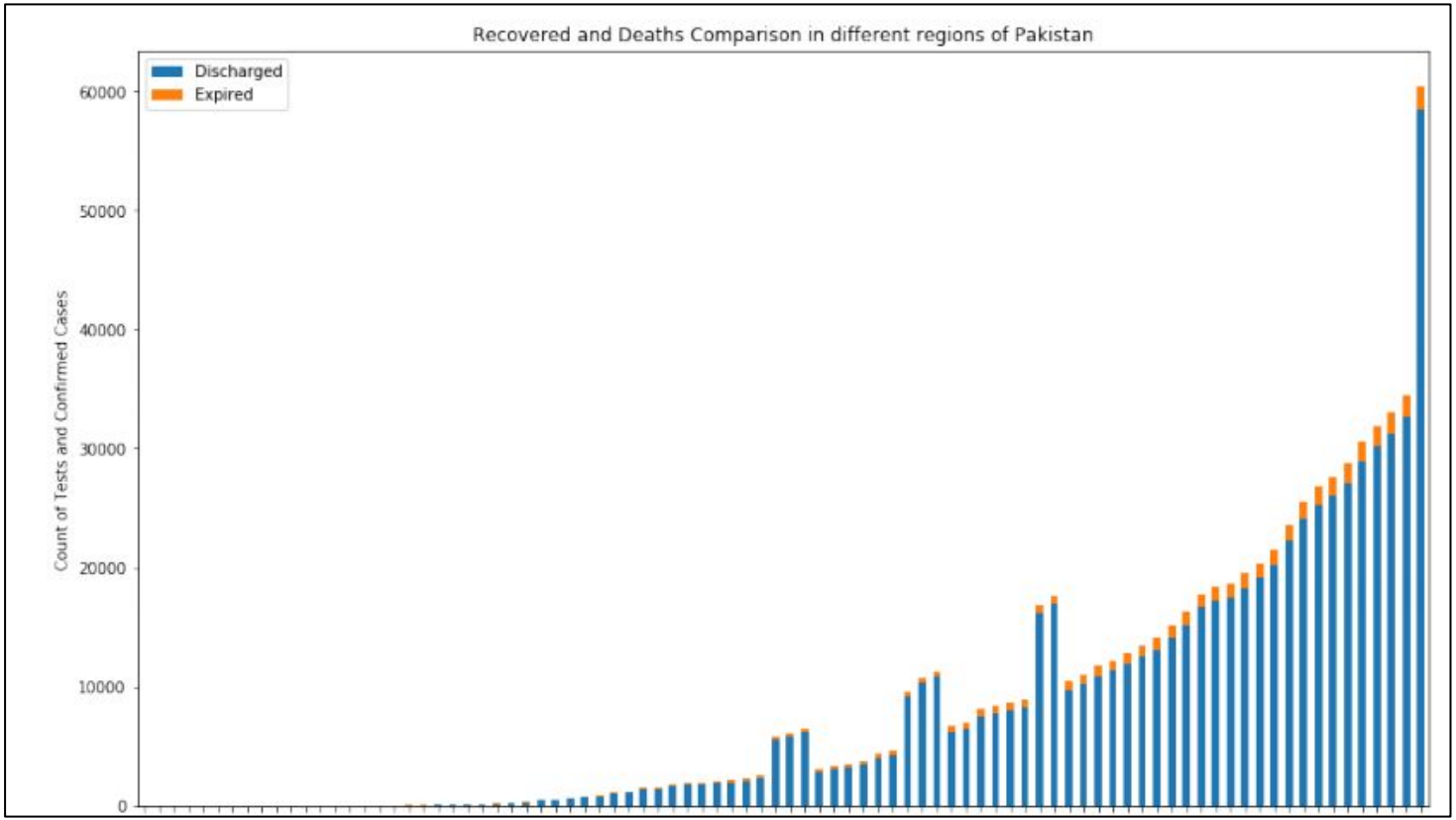
Germany has managed to keep the counts of confirmed cases and deaths lower, and higher number of recoveries. This shows that they have managed to control the spread which means their SOP's to prevent Covid spread are good.



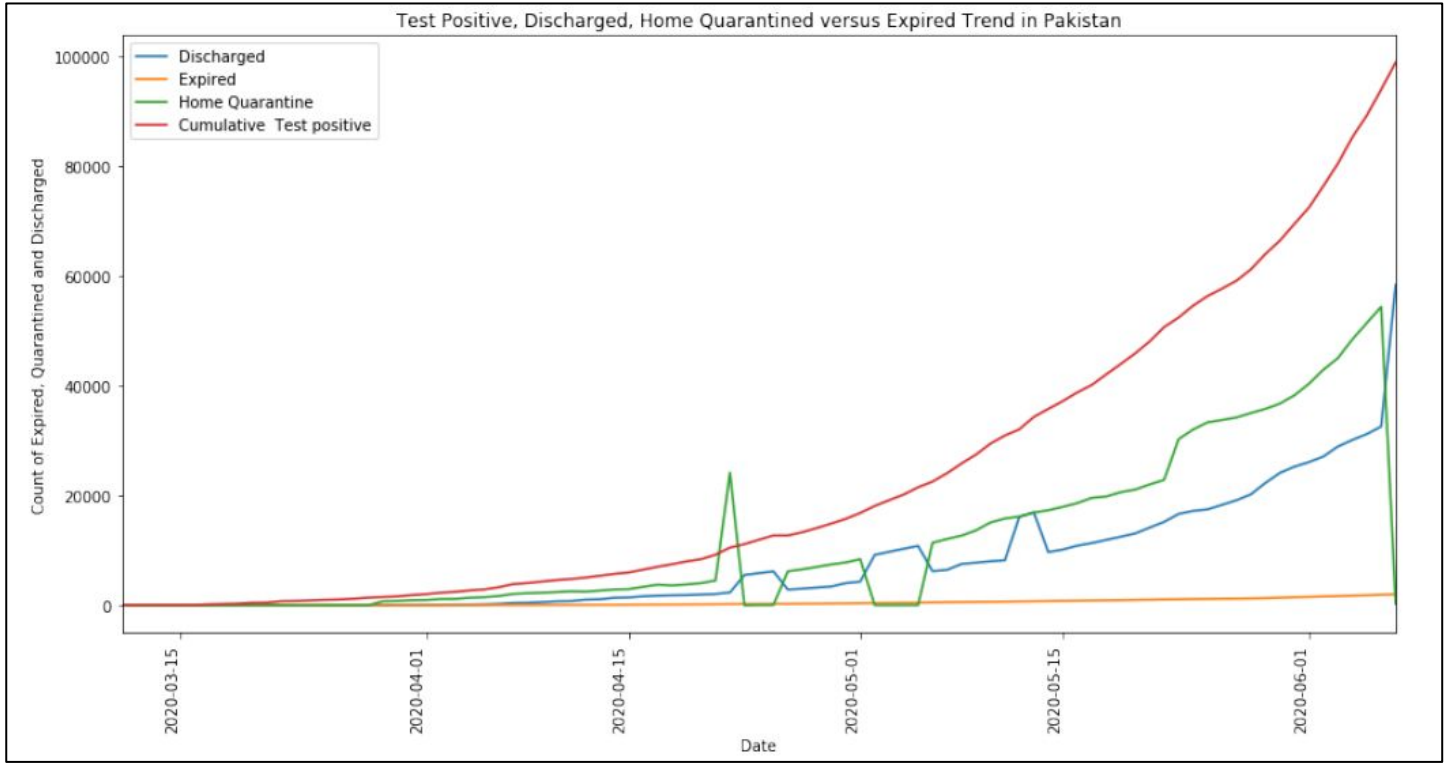
The USA is the most affected country from Corona Virus compared to other countries.



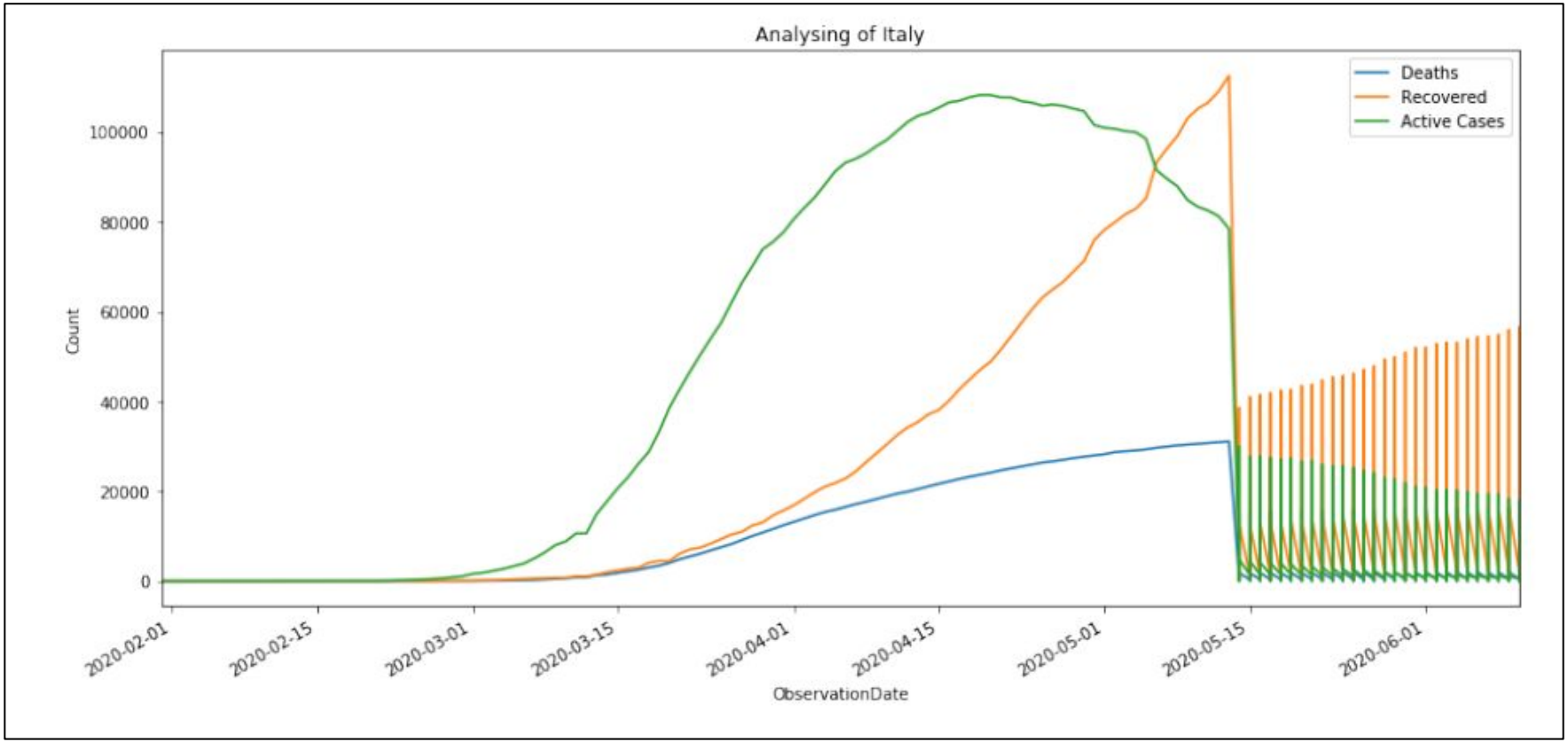
At first the number of Covid patients started to decrease after the lockdown in Pakistan, but went up again as soon as the soft lockdown was imposed, indicating the failure of lockdown strategies and the carelessness of people.



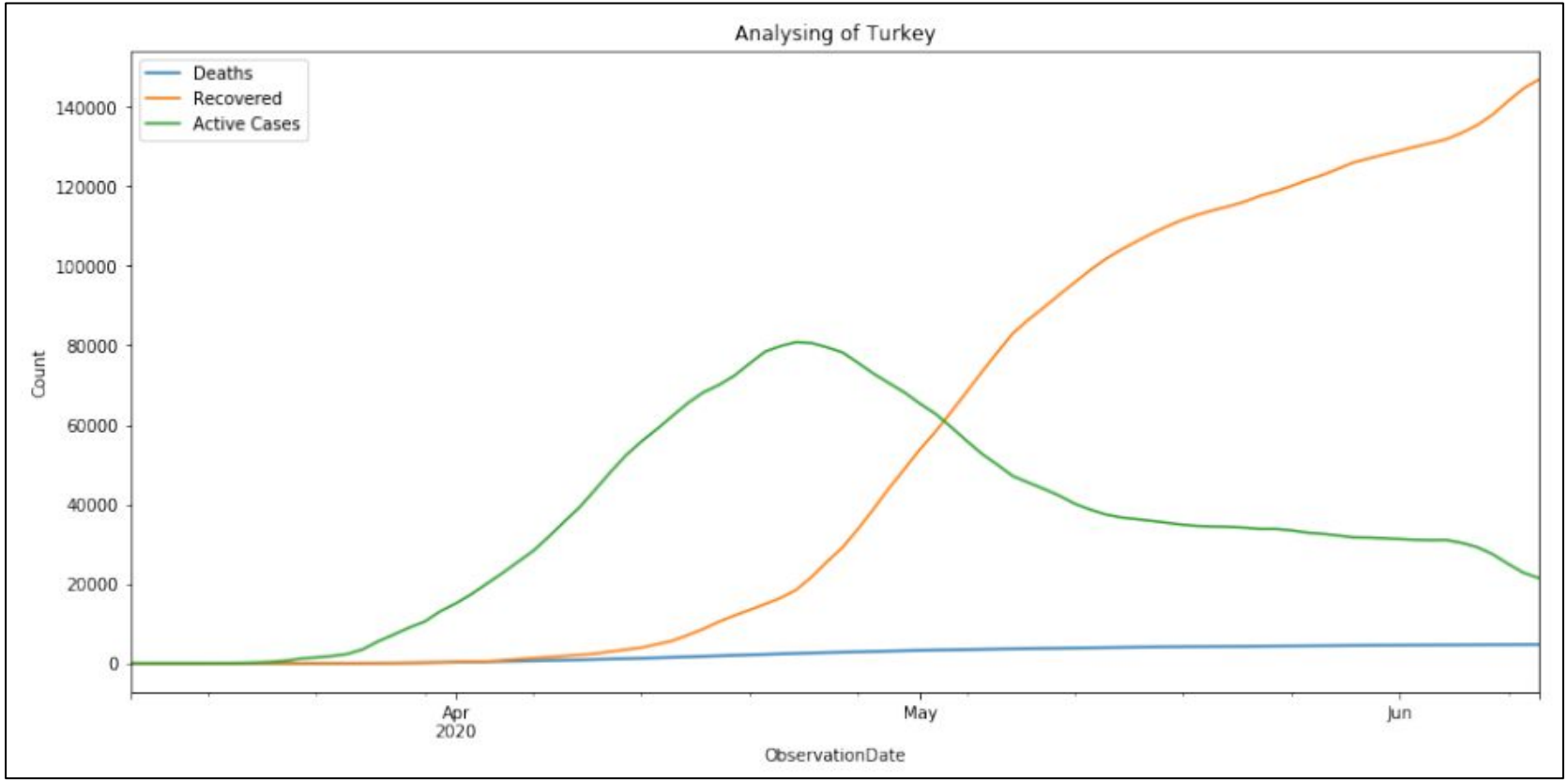
The death rate in Pakistan is much less than other countries, however considering the financial situation of Pakistan and the poor medical facilitation, the continuously increasing number of cumulative positive tests and relatively smaller rate for discharged patients points towards an alarming situation.



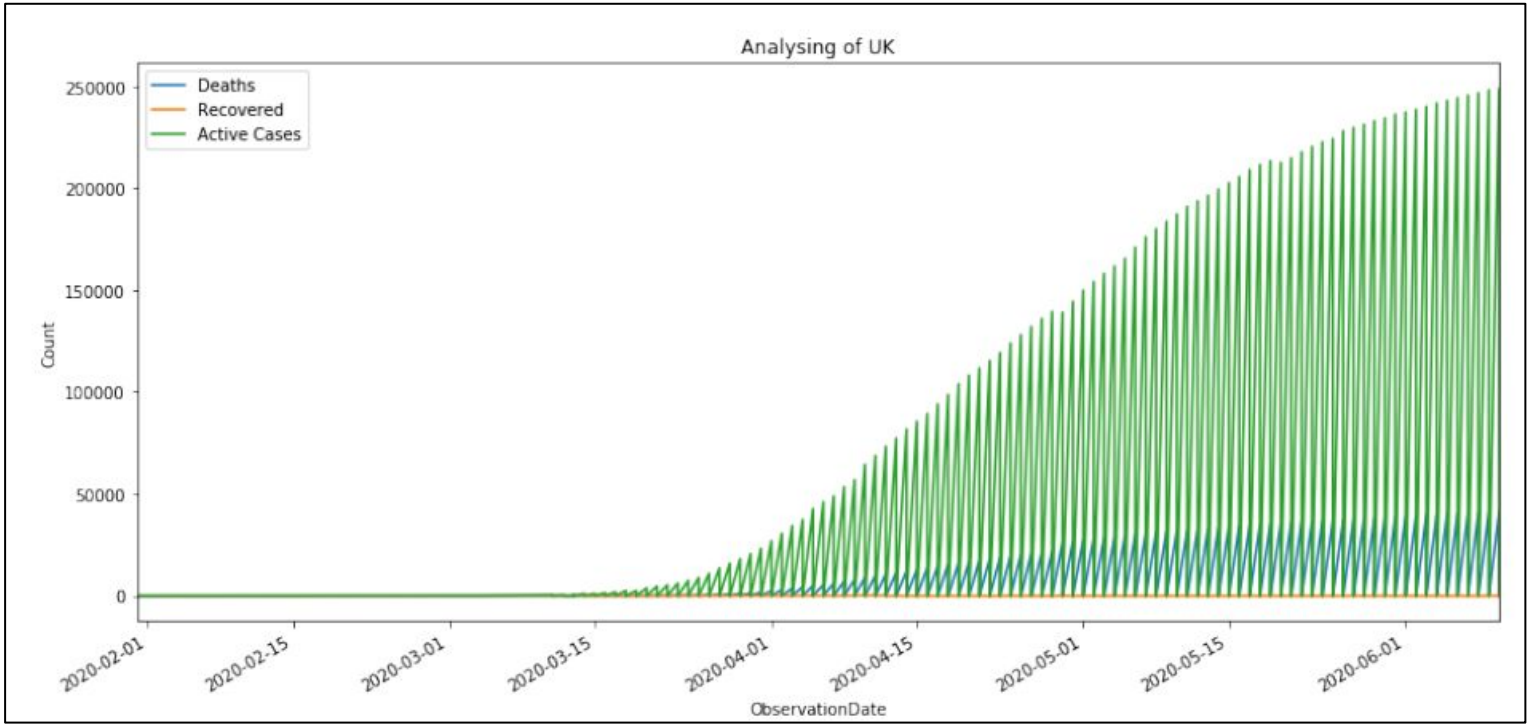
A high correlation between Cumulative Test positive and Cumulative Tests Performed tells us that as the number of tests increase, the number of positive results for the virus increase, thereby, leading to the greater number of cases in the country. This also suggests that to flatten the curve, it is highly necessary to increase the testing capacity each day substantially so that the impacted people can be quarantined immediately to minimise further spread.



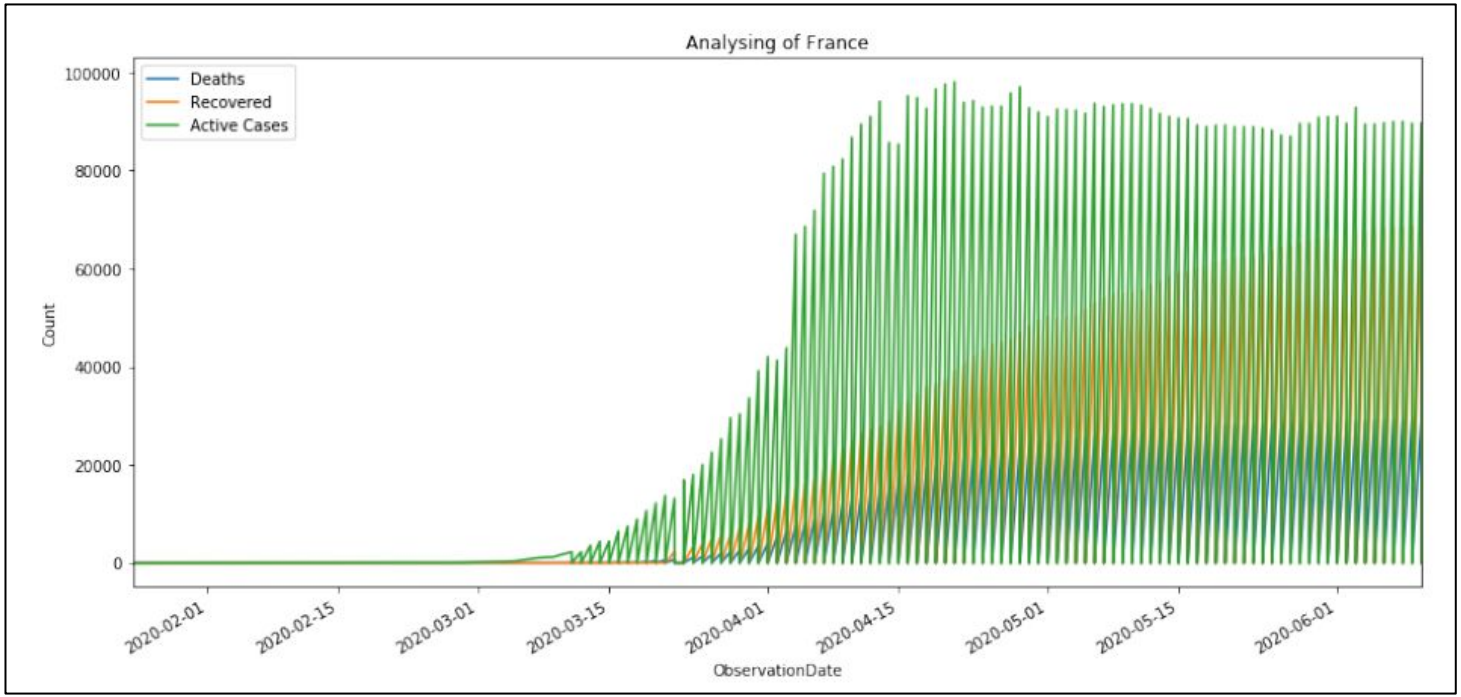
Italy managed to get control over Covid in May 2020



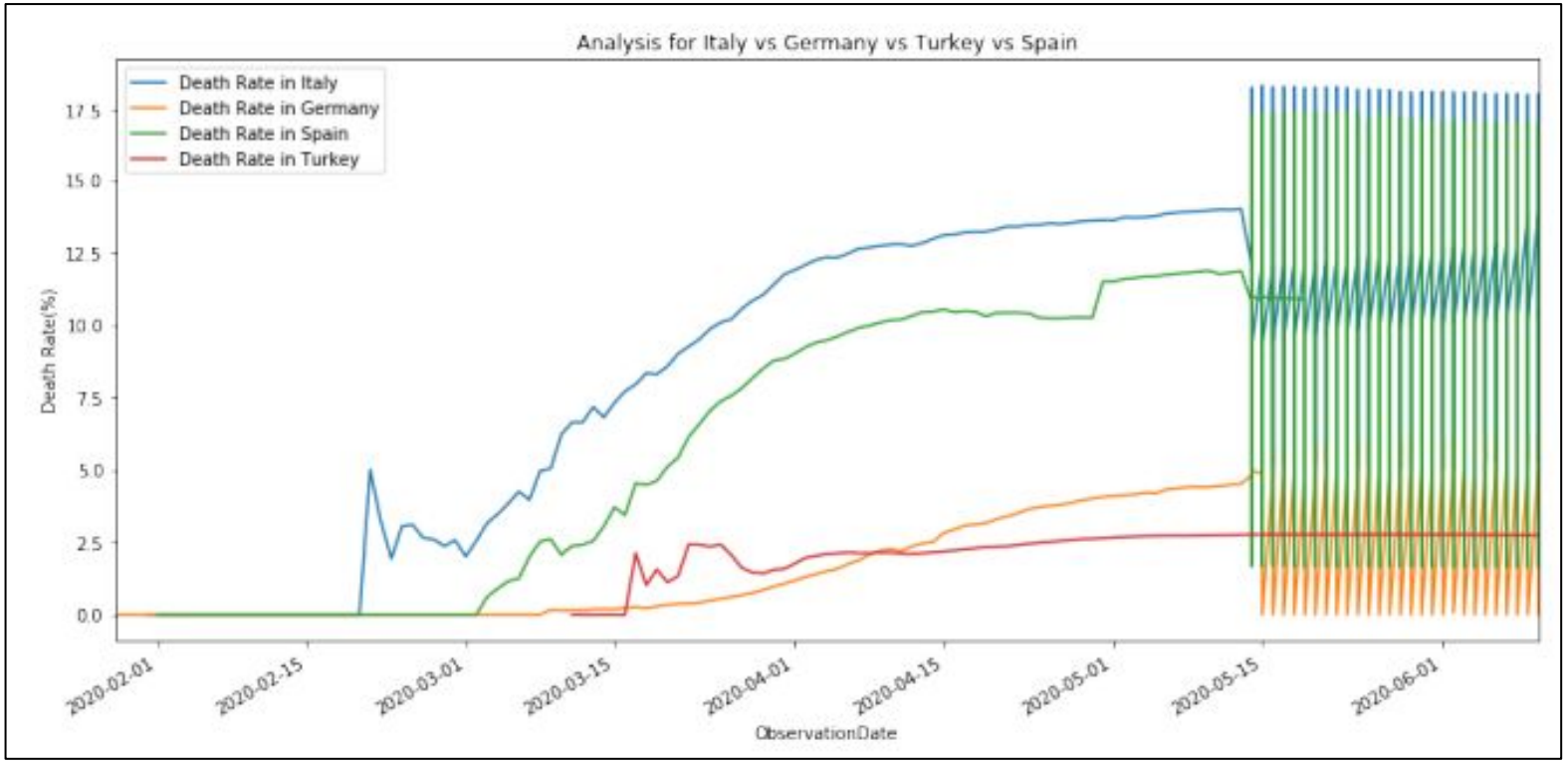
Turkey has a very high recovery rate.



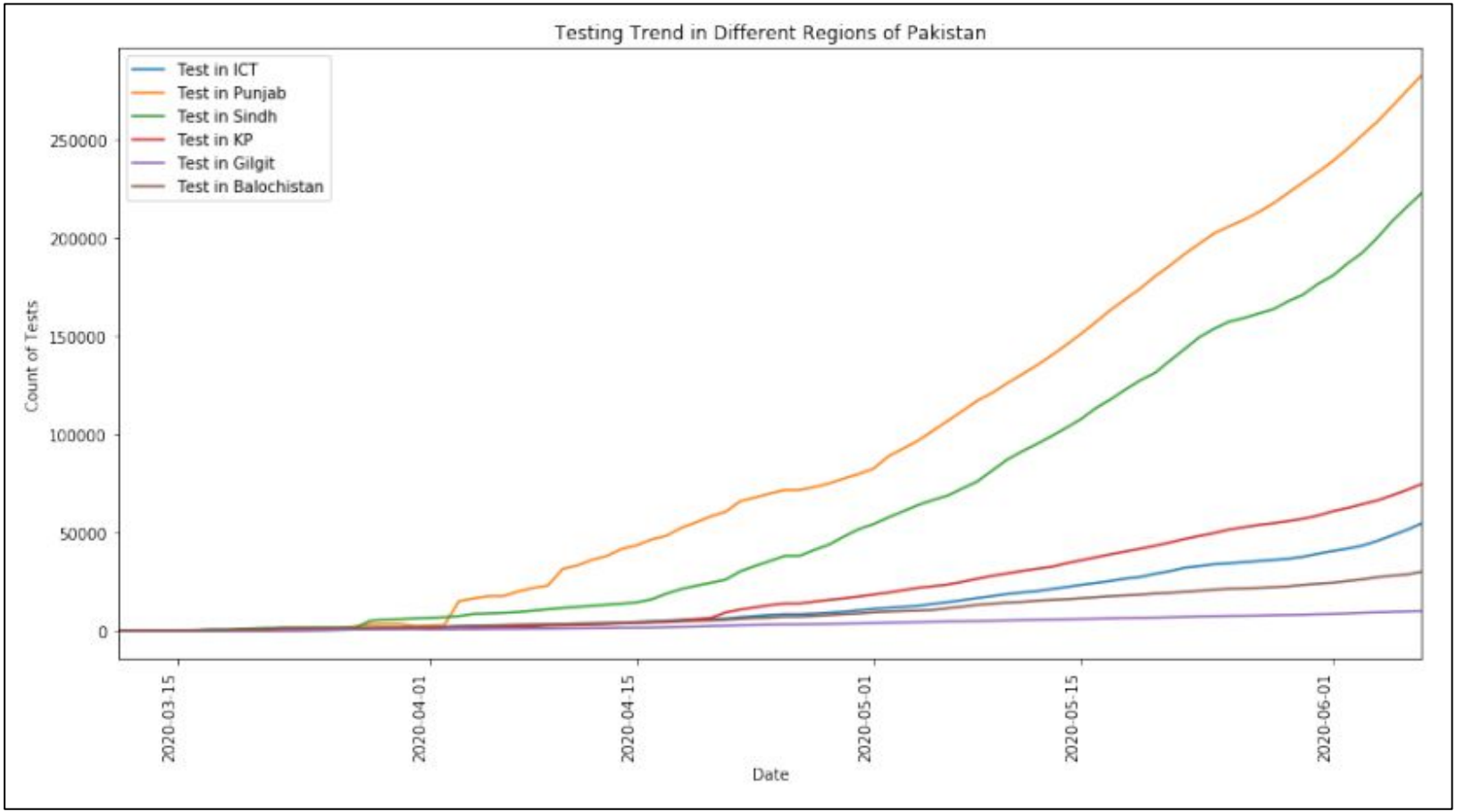
The UK has a very high spread pointing towards its bad SOP's against Covid.



France has a huge spread of Covid but manages to maintain the recovery rate high and death rate low.



The death rate in Italy is higher than that of Turkey, Germany and spain.



Punjab is performing the most testing.

Data Modeling

For modeling purposes, we have used linear regression and SVM. Linear regression is applied after generating polynomial features in order to minimize the error. SVM performed better than Linear regression.

	Dates	Polynonmial Regression Prediction	SVM Prediction
0	2020-06-11	[7466427.543321507]	7260794.727958
1	2020-06-12	[7466427.543321507]	7260794.727958
2	2020-06-13	[7466427.543321507]	7260794.727958
3	2020-06-14	[7466427.543321507]	7260794.727958
4	2020-06-15	[7466427.543321507]	7260794.727958

Presentation & Automation

Since the data for the impact of Covid-19 is varying greatly each day, it is important to inculcate the process of automation in our project pipeline. Therefore, we made sure to automate as much of the data retrieval, exploration and data modelling pipeline as possible, by using Python based notebooks, so that having the most recent data in the pipeline, the further processes can then just be executed sequentially and the outputs would vary accordingly. For the presentation part, our data is well represented in the form of report, our Jupyter notebooks added as appendices in report and this PowerPoint Presentation illustrating our results, implications and conclusions.

Conclusion

- From the data, it is clear that Covid-19 does not have a high mortality rate but its spread rate is very high. Just in few months it was able to spread to all over the world.
- The recovery rate is greater than the death rate which is a good thing but the alarming rate with which it spreads creates a problem for health facilities.

Conclusion

- The only way to control its spread is to impose a lockdown and limit the movement of people so that the virus can be contained.
- Countries like the USA, United Kingdom,and Italy have a high spreading rate of virus which is because of their carelessness that they showed in earlier days.
- Statistics show that the best possible solution to combat Covid-19 is to perform more and more testing and quarantine the affected population. This disease has no treatment and it can only be stopped by controlling its spread.