



COVID-19

Coursera capstone

Data Analysis Covid-19 pandemic

By: Pallavi kumari
August 2020

INTRODUCTION

Coronavirus(COVID-19) has become the most buzzed topic these days. COVID-19 is the disease caused by the new coronavirus that emerged in China in December 2019. The source of this virus is believed to be a 'wet market' in Wuhan which sold both dead and live animals including fishes and birds. novel coronavirus has inflicted havoc across the globe for lives and livelihoods. The impact of the pandemic on human lives is severe, but the effects on the global economy and on sustainable development's future are also a concern. The International Monetary Fund already declared that the world is into a recession. The full economic impact of the crisis is still difficult to predict but preliminary estimates are US\$2 trillion.

COVID-19 symptoms include cough, fever, shortness of breath, dry cough, headache, pneumonia. COVID-19 can be severe and some cases have caused death.

We will analyze the outbreak of coronavirus across various regions, visualize them using charts and graphs and predict the number of upcoming cases for the next 10 days using linear regression and polynomial regression models in python. The data has information from 31st December 2019 to till .



Collecting data

Collecting data from kaggle

Kaggle is a platform for predictive modelling and analytics competitions in which statisticians and data miners compete to produce the best models for predicting and describing the datasets uploaded by companies and users. This crowdsourcing approach relies on the fact that there are countless strategies that can be applied to any predictive modelling task and it is impossible to know beforehand which technique or analyst will be most effective. On 8 March 2017, Google announced that they were acquiring Kaggle. They will join the Google Cloud team and continue to be a distinct brand. In January 2018, Booz Allen and Kaggle launched Data Science Bowl, a machine learning competition to analyze cell images and identify nuclei.

Methodology

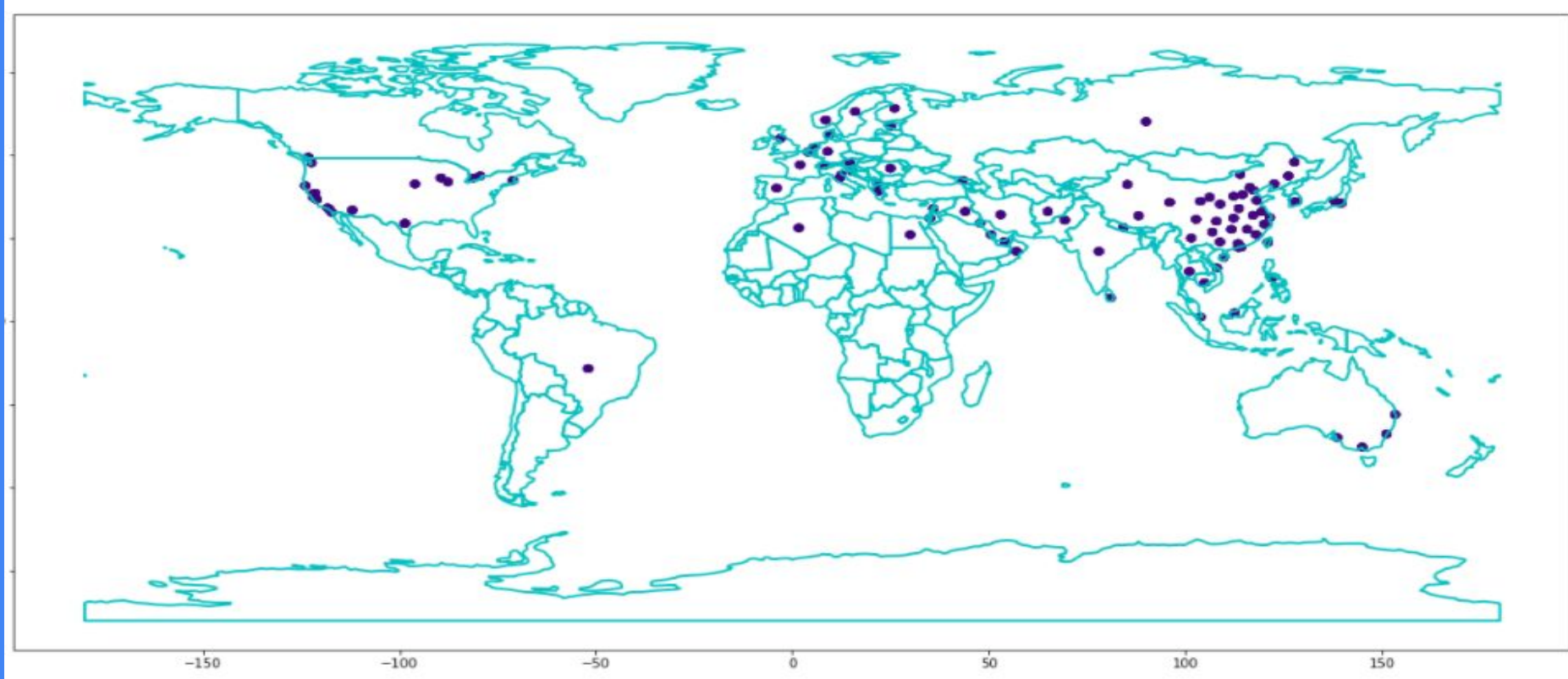
Implementing Polynomial Regression With scikit-learn

Implementing polynomial regression with scikit-learn is very similar to linear regression. There is only one extra step: you need to transform the array of inputs to include non-linear terms such as x^2 .

Degree is an integer (2 by default) that represents the degree of the polynomial regression function. In our project, we have taken the degree 3 of polynomial regression for better accuracy.

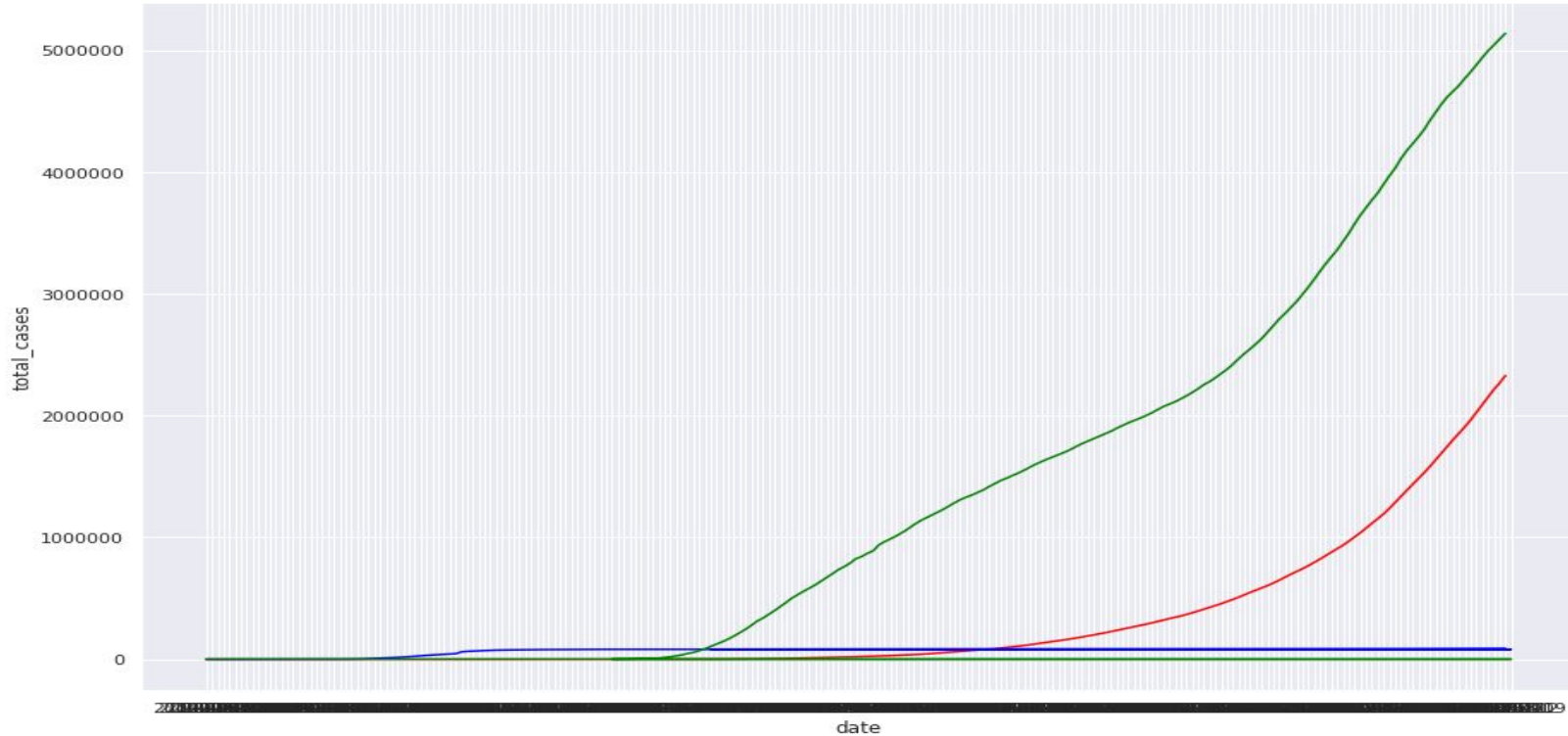
The result came out through this model is very good and has very good accuracy of 98.7 .

Countries affected by covid-19



Comparison of cases between India, China and US

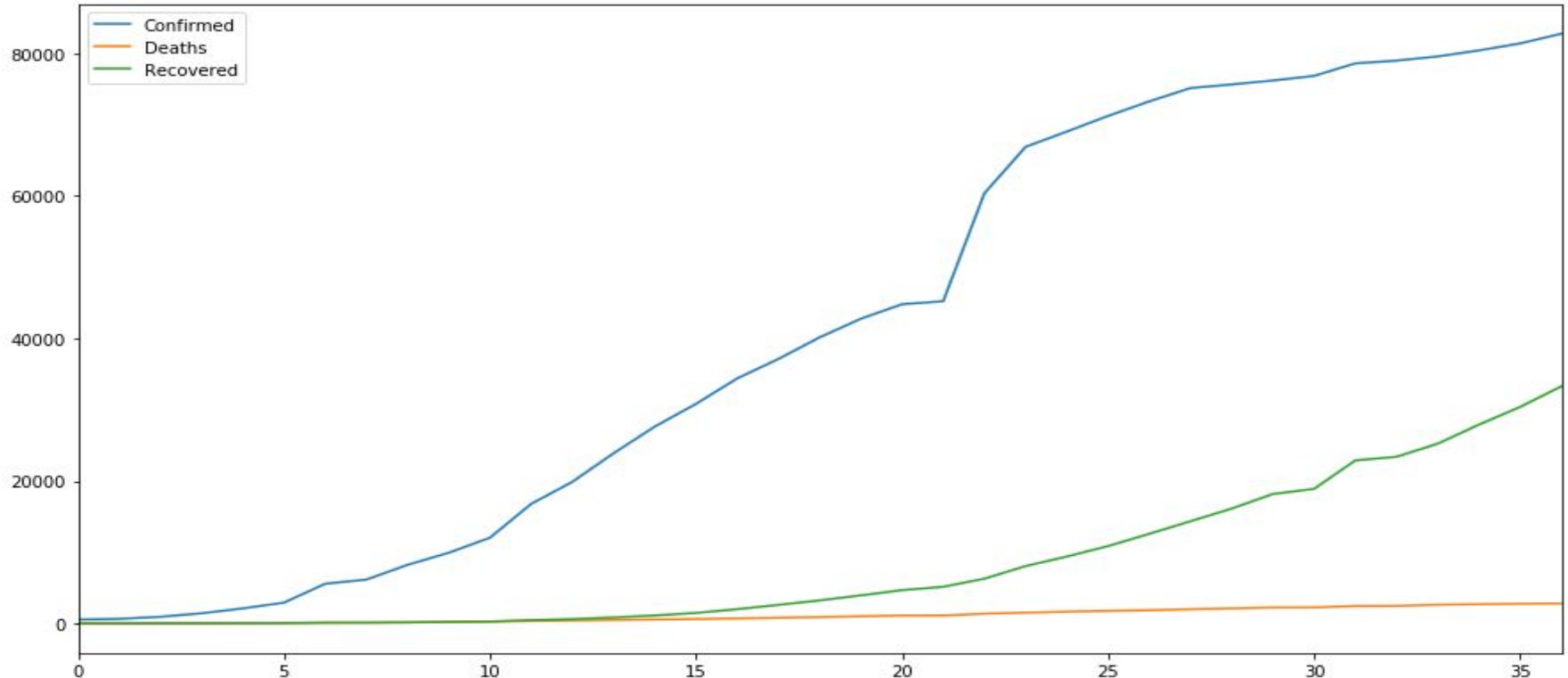
```
sb.lineplot(x='date',y='total_cases',data=india_case,color='red')  
sb.lineplot(x='date',y='total_cases',data=china_case,color='blue')  
sb.lineplot(x='date',y='total_cases',data=USA_case,color='green')  
plt.show()
```



Visualization Confirmed,Deaths,Recovered Cases

```
df2_by_date[ ['Confirmed', 'Deaths', 'Recovered'] ].plot(kind = 'line', figsize = (15, 8))
```

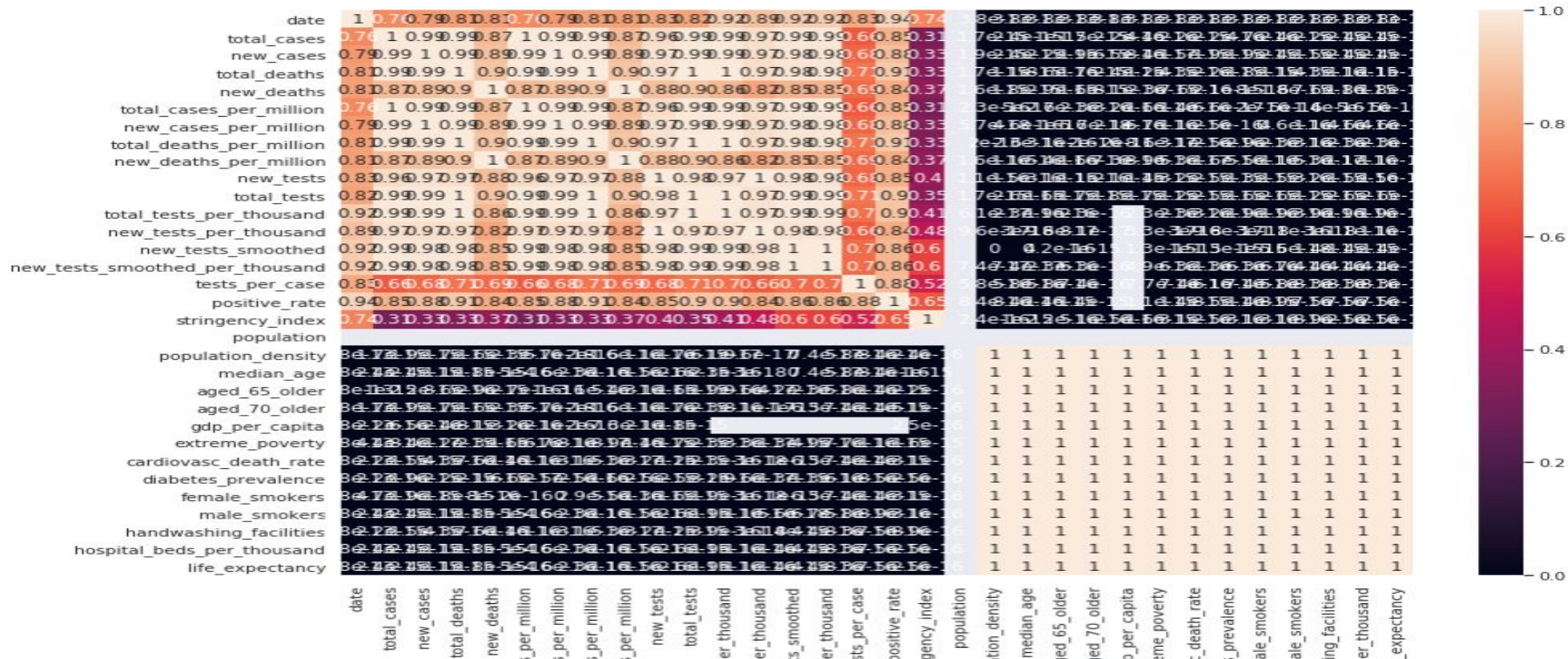
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f65ec65ee10>
```



Heatmap representation (checking correlation)

```
sb.heatmap(india_case.corr().abs(),annot=True)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7effd0131748>
```



RESULT

Accuracy achieved with using polynomial regression = 98.7%

Error comes out = 0.010

CONCLUSION

COVID-19 is increasing rapidly throughout the world. So we have made this project that is predicting the number of corona cases in upcoming days so that one can take safety measures and precautions in advance.

In this project we have applied a linear regression model first but didn't get good accuracy. So we applied polynomial regression model having 3rd degree to achieve good accuracy.

