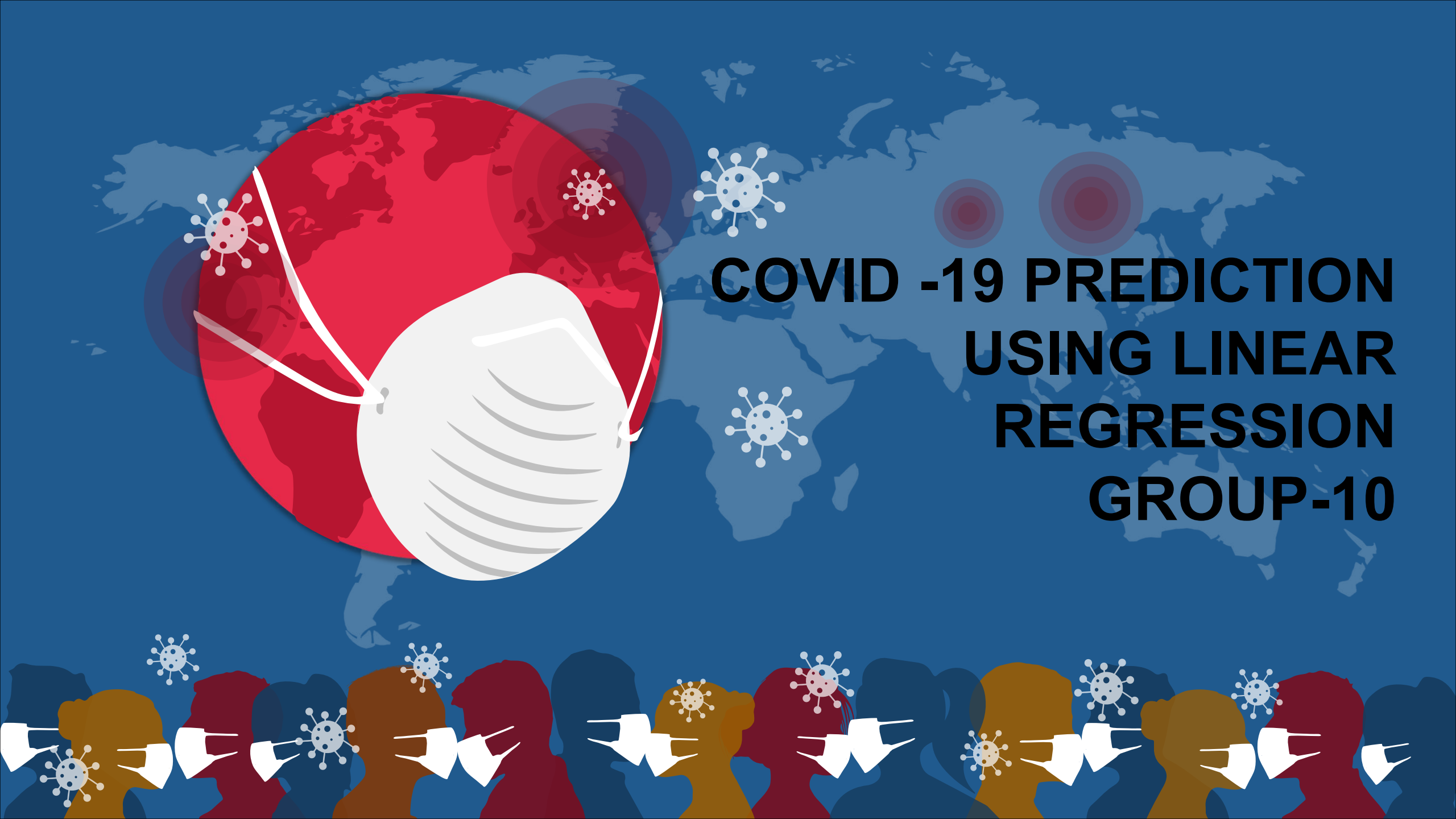


COVID -19 PREDICTION USING LINEAR REGRESSION

GROUP-10

GROUP MEMBERS

MUHAMMED SHAJAHAN-(AM.EN.U4AIE21144)

AKSHAY KRISHNAN T-(AM.EN.U4AIE21109)

NAVNEETH SURESH-(AM.EN.U4AIE21147)

NIRANJAN PRASANTH-(AM.EN.U4AIE21148)

ASWIN US-(AM.EN.U4AIE21119)

WHAT IS COVID-19 ?

01 Coronavirus disease (COVID-19) is an infectious disease caused by the SARS-CoV-2 virus.

02 Most people infected with the virus will experience mild to moderate respiratory illness and recover without requiring special treatment.

03 The best way to prevent and slow down transmission is to be well informed about the disease and how the virus spreads.

04 Protect yourself and others from infection by staying at least 1 metre apart from others, and get vaccinated when it your’s turn

INTRODUCTION

COVID-19

- As we know there are many terms related to numbers like confirmed cases, deaths and recovery.
- The trend of the Covid 19 cases is an important aspect in fighting Covid 19.
- Hence machine learning models could be applied to find future trends of Covid 19.
- This can help to make precautionary measures to prevent Covid 19.
- Hence for predicting the Covid 19 cases we will be using a Linear Regression model.
- We also visualized the dataset using bar graphs and pie charts.
- This can help better understanding the dataset.

DATASET

Dataset

There are three datasets used in the project. Confirmed cases, deaths, recovered. These datasets contain information on province,country,latitude,longitude, dates which shows the number of cases on that date.

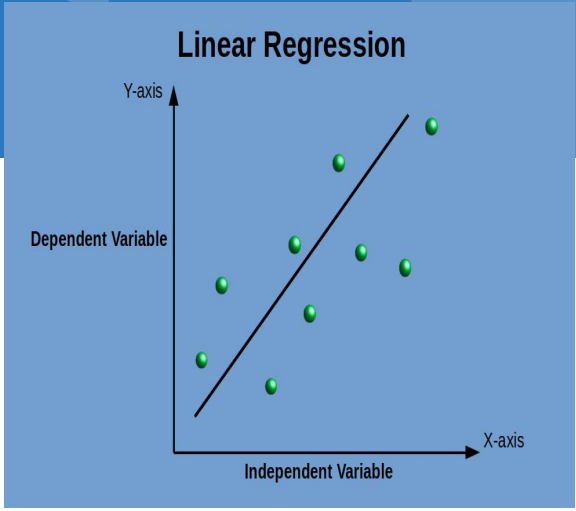
This dataset is from kaggle.

a) Confirmed cases:
https://drive.google.com/file/d/1i1ZdEmn2pEl4y8G2QMOkdxGPtt71mrRf/view?usp=sharing

b) Deaths:
https://drive.google.com/file/d/15kqh3wEeilsR0nILCl2V2kQtgWHMd2Yk/view?usp=sharing

c) Recovered:
https://drive.google.com/file/d/1P9ivCZ8Jf3OcC3fHM2XO2CFWaH3NV5M/view?usp=sharing

What is Linear Regression?



COVID-19

Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data.The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable.For example, a modeler might want to relate the weights of individuals to their heights using a linear regression model.

$Y_i = f(X_i, \beta) + e_i$

Y_i = dependent variable

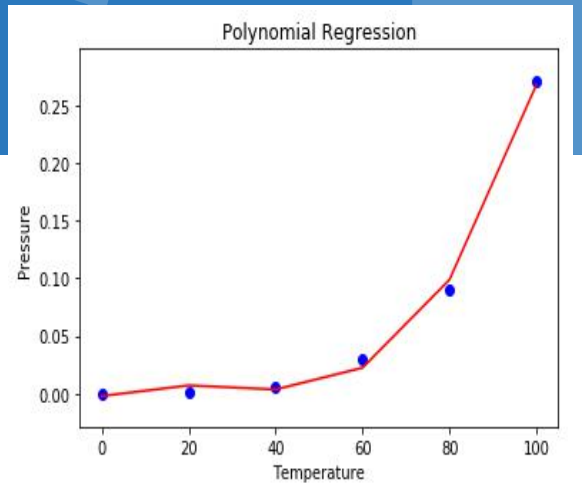
f = function

X_i = independent variable

β = unkown parameters

e_i = error terms

What is Polynomial Regression?



COVID-19

- Polynomial regression, abbreviated E(y |x), describes the fitting of a nonlinear relationship between the value of x and the conditional mean of y.
- In this project we have used two types

i) Quadratic Regression

ii) Cubic Regression

- A quadratic regression is the process of finding the equation of the parabola that best fits a set of data.
- The best way to find this equation manually is by using the least squares method. A quadratic regression is the process of finding the equation of the parabola that best fits a set of data.
- As a result, we get an equation of the form: $y=ax^2+bx+c$ where $a\neq0$.

- That is, we need to find the values of a,b and c such that the squared vertical distance between each point (x_i,y_i) and the quadratic curve $y = ax^2+bx+c$ is minimal.
- In the cubic regression model, we deal with cubic functions, that is, polynomials of degree 3.
- The cubic regression function takes the form:
  $y = a + bx + cx^2 + dx^3$,
  where a, b, c, d are real numbers, called coefficients of the cubic regression model
- As you can see, we model how the change in x affects the value of y. In other words, we assume here that x is the independent (explanatory) variable and y is the dependent (response) variable.

Mean Squared Error

COVID-19

- The mean squared error (MSE) tells you how close a regression line is to a set of points.
- It does this by taking the distances from the points to the regression line (these distances are the “errors”) and squaring them.
- The squaring is necessary to remove any negative signs.
- It also gives more weight to larger differences. It’s called the mean squared error as you’re finding the average of a set of errors. The lower the MSE, the better the forecast.
- MSE formula = $(1/n) * \Sigma(\text{actual} - \text{forecast})^2$

n = number of items,
Σ = summation notation,
Actual = original or observed y-value,
Forecast = y-value from regression.

ROOT MEAN SQUARED ERROR

COVID-19

- Root mean square error tells us the average distance between the predicted values from the model and the actual values in the dataset.
- It is a way to assess how well a regression model fits a dataset.
- The lower the RMSE, the better a given model is able to fit a dataset.
- RMSE = $\sqrt{\Sigma(P_i - O_i)^2 / n}$
- Here:

  Σ = summation notation

  P_i = is the predicted value for the i^{th} observation in the dataset

  O_i = is the observed value for the i^{th} observation in the dataset

  n = Sample size

MEAN ABSOLUTE ERROR

COVID-19

- The mean absolute error is a way to measure the accuracy of a given model.
- Mean absolute error is a loss function used for regression.
- The loss is the mean over the absolute differences between true and predicted values.
- It is calculated as:
- $MAE = (1/n) * \Sigma|y_i – x_i|$
  where:
  Σ: Sumamtion Notation
  y_i: The observed value for the i^{th} observation
  x_i: The predicted value for the i^{th} observation
  n: The total number of observations
- In general, the lower the value for the MAE the better a model is able to fit a dataset. When comparing two different models, we can compare the MAE of each model to know which one offers a better fit to a dataset.

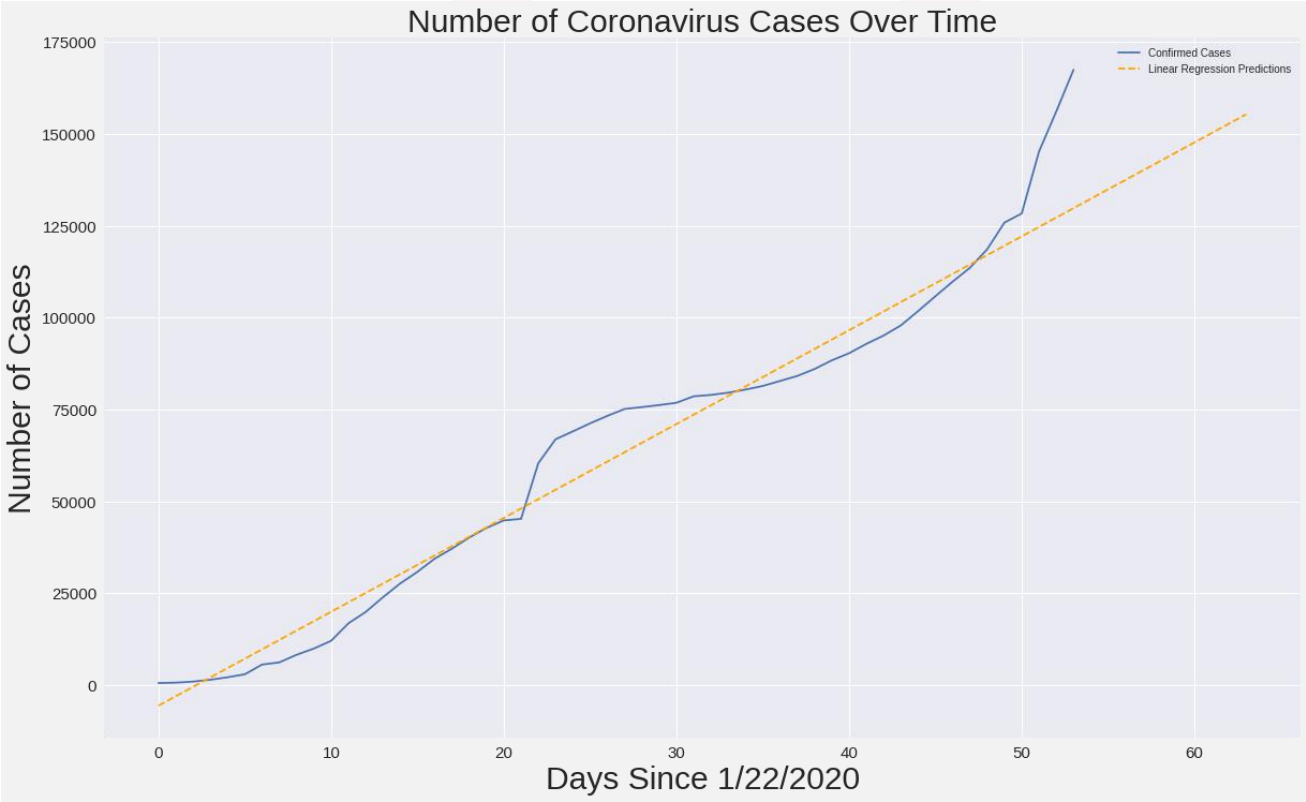
IMPLEMENTATION

LINEAR REGRESSION

- We use Sklearn library for performing linear regression
- First we see the confirmed cases
- Here we convert the required confirmed cases into a numpy array as each element of the array the sum of each column.
- We split the model into train and test using respective function in the sklearn library
- Then we fit the best line using the training data and predict the output using test data.
- We plot the graph with the matplotlib library.
- Then we find the mean absolute error, mean squared error, root mean squared error.
- We repeat the similar method for the datasets of death and recovery.
- Show plots.

PLOTS

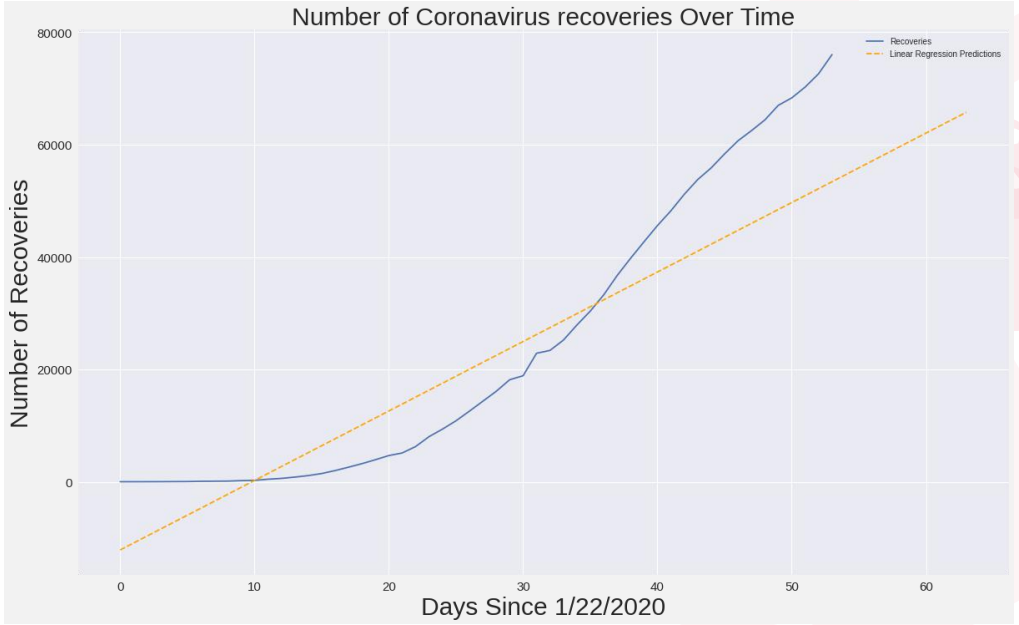
Confirmed Cases



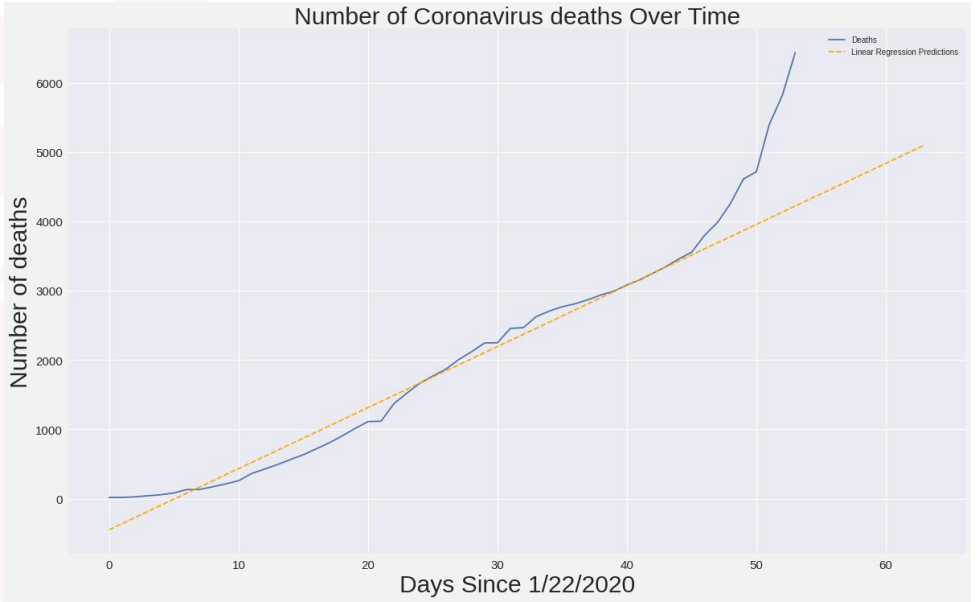
COVID-19

PLOTS

Recoveries



Deaths



COVID-19

POLYNOMIAL REGRESSION

- We use PolynomialFeatures from sklearn library.
- First we see the second degree
- Here also we split the dataset into train and test using the respective function in the sklearn library
- Polynomial Feature will generate a new feature matrix consisting of all polynomial combinations of the features with degree less than or equal to the specified degree
- We repeat the similar method for 3rd degree and 4th degree.

ERRORS

Linear Regression

```
MAE: 11965.537037037033
MSE: 307996364.0108404
RMSE:  17549.82518462336
```

Quadratic Regression

```
MAE: 10419.971004746621
MSE: 169131996.17456436
RMSE:  13005.075785037332
```

Cubic Regeression

```
MAE: 8574.477890075488
MSE: 94808210.52774853
RMSE:  9736.950781828391
```

Polynomial Regression

```
MAE: 1831.0198382906772
MSE: 6115553.428541872
RMSE:  2472.964502078805
```

CONCLUSION

- Solved a real life problem
- Efficiency can be further increased by using higher order regression and increasing the accuracy of the dataset taken
- If we introduce a new variabe “contact tracing” can also increase the efficiency of model

THANK YOU