Project Stage - IV

Develop Linear and Non-Linear (polynomial) regression models for predicting cases and deaths in US:

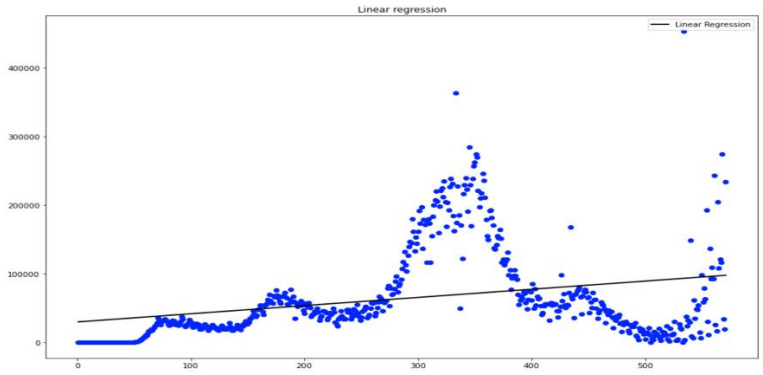
REGRESSION:

We used the regression models concept here to see how the data points of new cases and new deaths from the start day are fitted across United States and other countries whose population is similar to the United States and predict the best line that fits the data points based on the training. The types of regression models we implemented in this project are linear regression and Polynomial regression.

Simple linear regression is an approach for predicting a quantitative response using a single feature. And the linear regression works on the continuous data. It is given as follows:
$y = b + ax$
Here y is the response x is the feature
b is the intercept
a is the coefficient of x

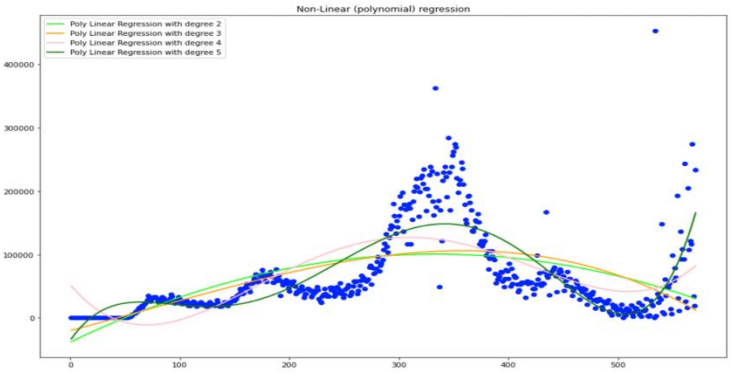
a, b together are called model coefficients. Before creating the model, we must learn the values of these model coefficients. Once after learning these coefficients are used to predict the best line of the data points.
But one of the problem with linear regression model is it will be highly biased for given data sometimes if the input data are correlated. And generally linear regression model has low variance. This low variance will help us if we have very less amount of data.
Polynomial regression model is the special case of the linear regression model where we fit a polynomial equation on the data with a curvilinear relationship between the target variables and the independent variables. It is given as follows:
$y=b+ax+ax2 +.....+axn$
where n is the degree of the polynomial.
In this project we implemented the polynomial regression
model until the degree 5. The respective equations until degree 5 are as follows:
For
$n = 2, y = b + ax + ax2$
$n = 3, y = b + ax + ax2 + ax3$
$n = 4, y = b + ax + ax2 + ax3 + ax4$
$n = 5, y = b + ax + ax2 + ax3 + ax4 + ax5$
The polynomial regression can provide good approximations between the target variable and the independent variables. But this regression model is very sensitive to the outliers. Even if there is one outlier it effects the shape of the curve.
In this project we even calculated the root mean square error for both linear regression and polynomial regression. Root mean square error is the most common metric for evaluating performance of linear regression model. It is measure of how the predicted values are different from the actual values.

In order to implement this linear regression model and polynomial regression model we imported LinearRegres- sion from sklearn. linear model , Polynomial Features from sklearn.preprocessing. Also, for calculating the root mean square error we imported the mean squared error from the sklearn.metrics.
The linear regression model for the new cases across USA from the start date of the infection is as follows:



RMSE for Linear Regression: 62714.21609478131

The polynomial regression model for the new cases across USA from the start date of the infection is as follows:



Polynomial Regression model for new cases across USA

RMSE for Polynomial Regression with degree 2: 54916.14374069963
RMSE for Polynomial Regression with degree 3: 54487.18974655347
RMSE for Polynomial Regression with degree 4: 49067.62208492077
RMSE for Polynomial Regression with degree 5: 41726.69191009244

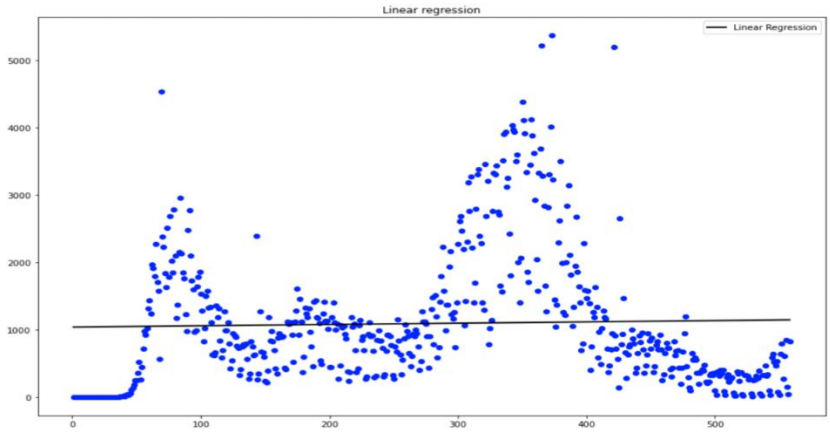
The linear regression model for the new deaths across USA from the start date of the deaths is as follows:

RMSE for Linear Regression: 992.6817817081455

The polynomial regression model for the new deaths across USA from the start date of the deaths is as follows:

RMSE for Polynomial Regression with degree 2: 869.9035082404372
RMSE for Polynomial Regression with degree 3: 849.4157029730479
RMSE for Polynomial Regression with degree 4: 849.2227607850864
RMSE for Polynomial Regression with degree 5: 698.2840913225779

The linear regression model for the new cases across Indonesia from the start date of the infection is as follows:



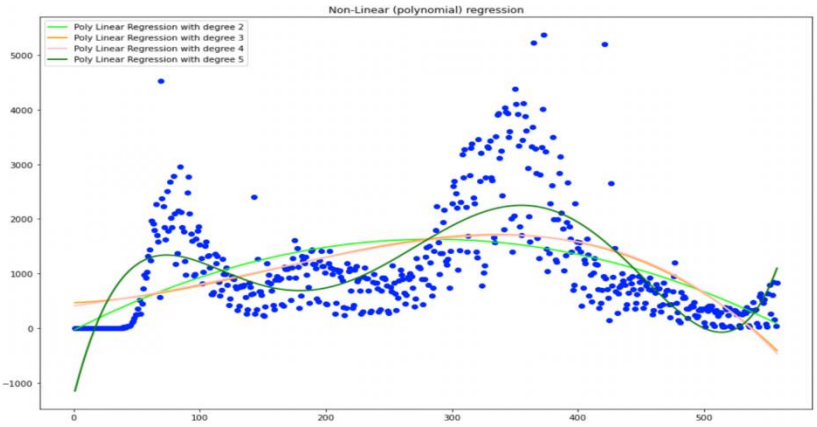
Linear Regression model for new deaths across USA

RMSE for Linear Regression: 992.6817817081455

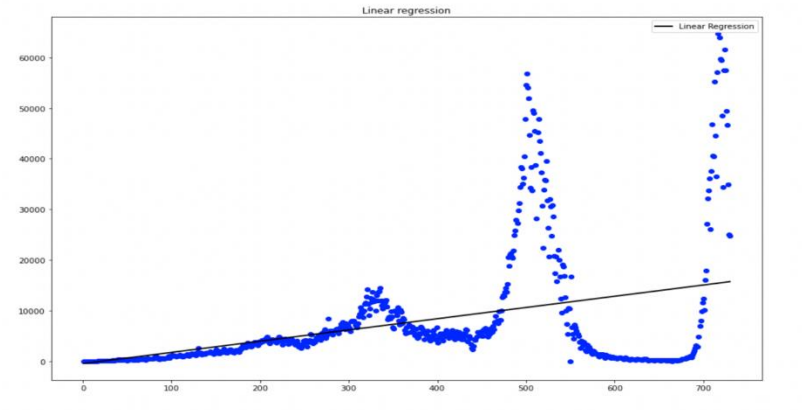
The polynomial regression model for the new deaths across USA from the start date of the deaths is as follows:

RMSE for Polynomial Regression with degree 2: 869.9035082404372
RMSE for Polynomial Regression with degree 3: 849.4157029730479
RMSE for Polynomial Regression with degree 4: 849.2227607850864
RMSE for Polynomial Regression with degree 5: 698.2840913225779

The linear regression model for the new cases across Indonesia from the start date of the infection is as follows:



Polynomial Regression model for new deaths across USA



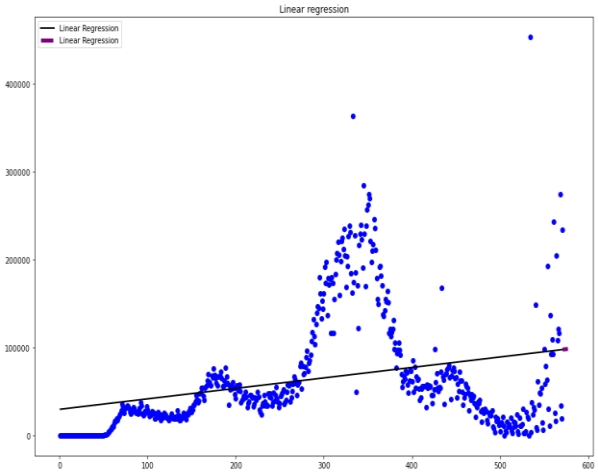
Calculate and report Root Mean Square Error (RMSE) for your models (linear and non-linear). Discuss bias versus variance tradeoff.

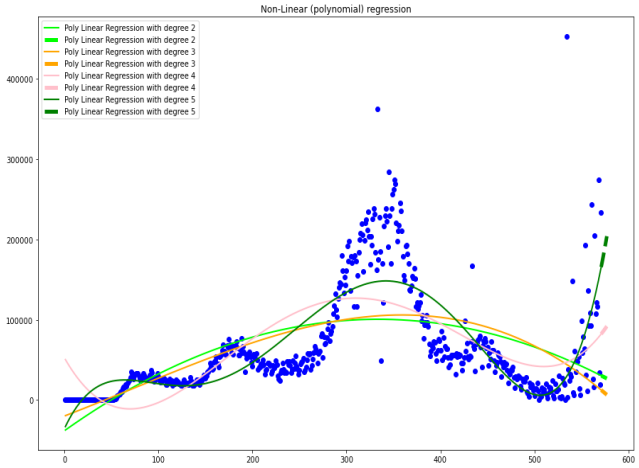
For every plot there are RMSE and trend line. While predicting forecast line is seem in plots.

The data is mostly high bias and low variance for linear regression model. While working with non-linear regression models as degree increases, we will see low bias and high variance.

Even high variance may result in overfitting problem, so we have to take model with moderate bias and moderate variance.

Plot trend line along for the data along with the forecast of 1 week ahead:



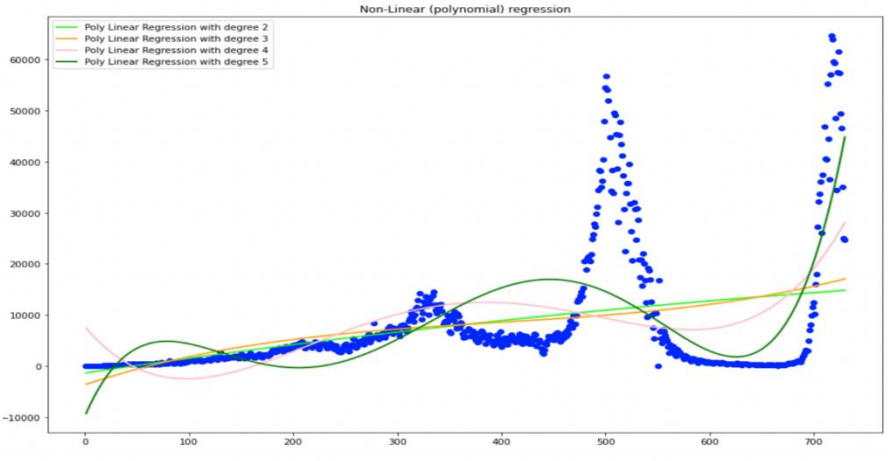


Describe the trends as compared to other countries.

Linear Regression model for new cases across Indonesia

RMSE for Linear Regression: 10930.77626509826

The polynomial regression model for the new cases across Indonesia from the start date of the infection is as follows:

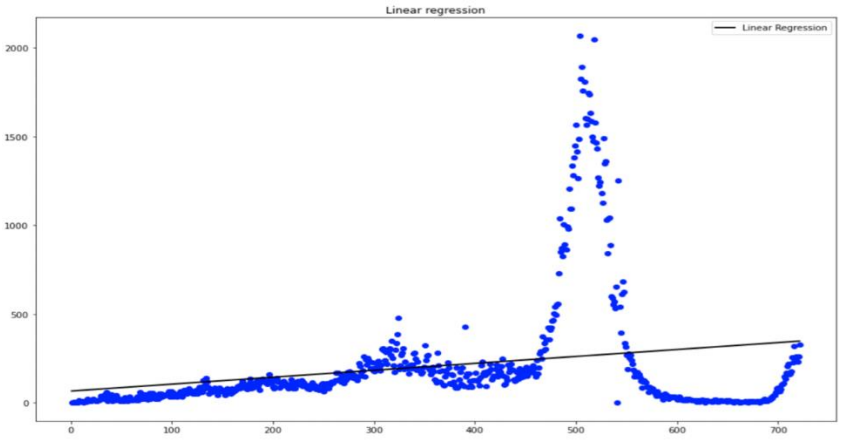


Polynomial Regression model for new cases across Indonesia

RMSE for Polynomial Regression with degree 2: 10922.654186622234
RMSE for Polynomial Regression with degree 3: 10888.86675677789
RMSE for Polynomial Regression with degree 4: 10230.262758803476

RMSE for Polynomial Regression with degree 5: 8838.388991239179

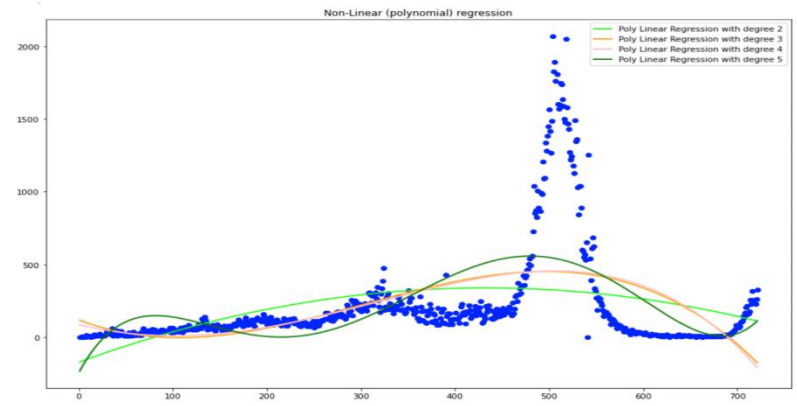
The linear regression model for the new deaths across Indonesia from the start date of the deaths is as follows:



Linear Regression model for new deaths across Indonesia

RMSE for Linear Regression: 336.1906293092541

The polynomial regression model for the new deaths across Indonesia from the start date of the deaths is as follows:

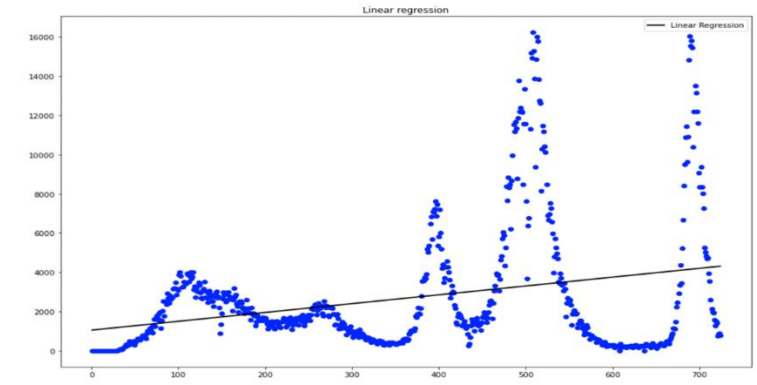


Polynomial Regression model for new deaths across Indonesia

RMSE for Polynomial Regression with degree 2: 319.07309224190163
RMSE for Polynomial Regression with degree 3: 299.8972299295997
RMSE for Polynomial Regression with degree 4: 299.7012546247724
RMSE for Polynomial Regression with degree 5: 283.32004291049594

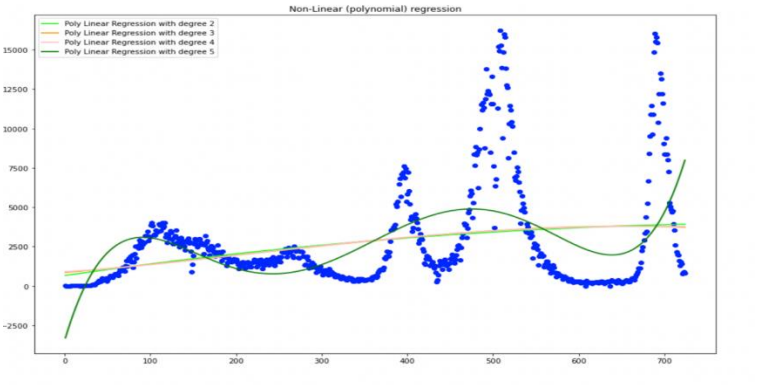
The linear regression model for the new cases across Bangladesh from the start date of the infection is as follows:

RMSE for Linear Regression: 3123.8503432568887



Linear Regression model for new cases across Bangladesh

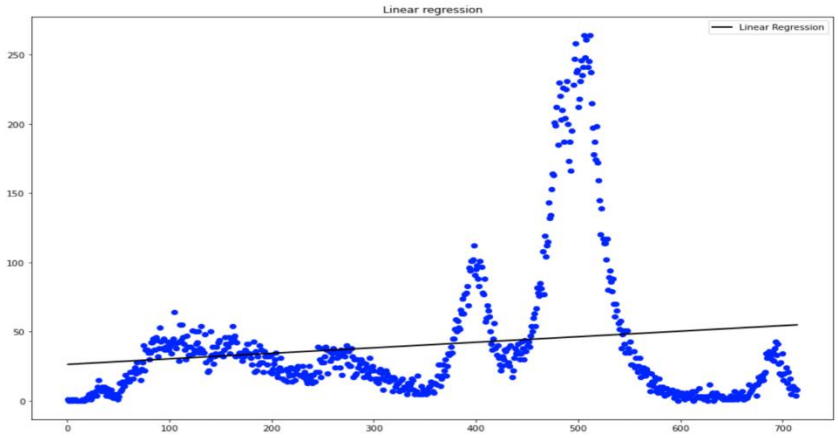
The polynomial regression model for the new cases across Bangladesh from the start date of the infection is as follows:



Polynomial Regression model for new cases across Bangladesh

RMSE for Polynomial Regression with degree 2: 3118.7339668553464
RMSE for Polynomial Regression with degree 3: 3117.8792187901226
RMSE for Polynomial Regression with degree 4: 3117.833054249292
RMSE for Polynomial Regression with degree 5: 2836.047737972691

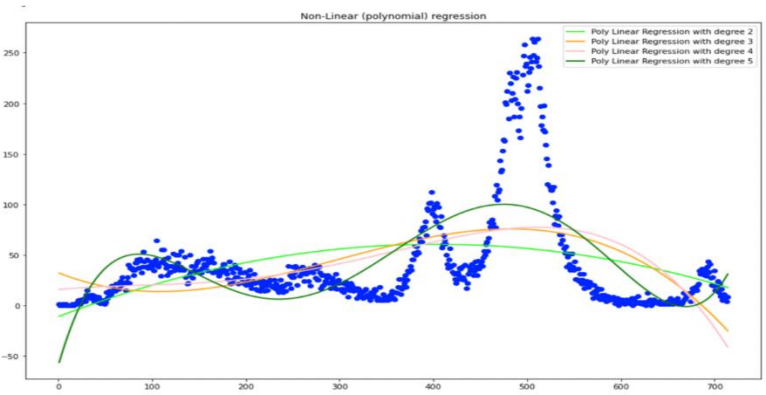
The linear regression model for the new deaths across Bangladesh from the start date of the deaths is as follows:



Linear Regression model for new deaths across Bangladesh

RMSE for Linear Regression: 51.703607927760416

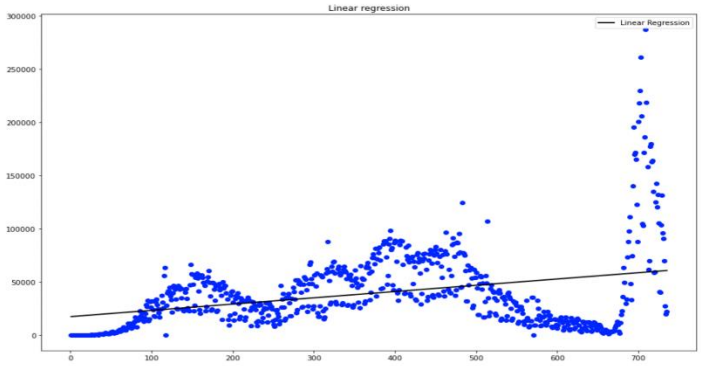
The polynomial regression model for the new deaths across Bangladesh from the start date of the deaths is as follows:



Polynomial Regression model for new deaths across Bangladesh

RMSE for Polynomial Regression with degree 2: 48.900523850989565
RMSE for Polynomial Regression with degree 3: 46.09793793526125
RMSE for Polynomial Regression with degree 4: 45.77068289447135
RMSE for Polynomial Regression with degree 5: 39.997982855183935

The linear regression model for the new cases across Brazil from the start date of the infection is as follows:



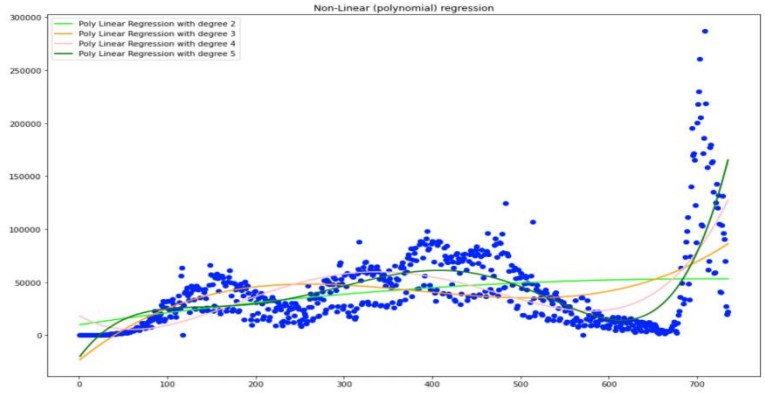
Linear Regression model for new cases across Brazil

RMSE for Linear Regression: 34966.427103926806

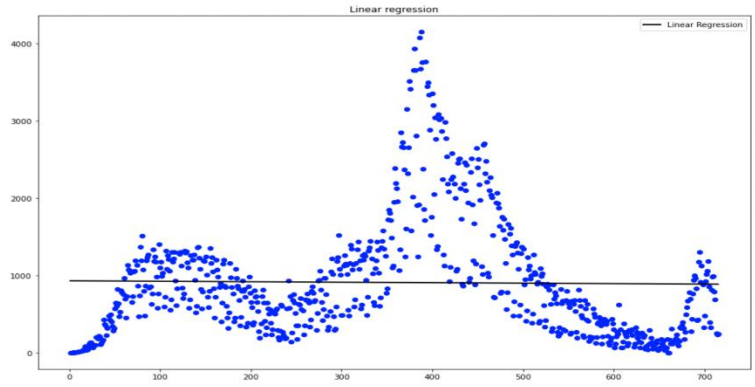
The polynomial regression model for the new cases across Brazil from the start date of the infection is as follows:

RMSE for Polynomial Regression with degree 2: 34799.078797647
RMSE for Polynomial Regression with degree 3: 32442.112007262265
RMSE for Polynomial Regression with degree 4: 29315.801679955133
RMSE for Polynomial Regression with degree 5: 26877.093540463535

The linear regression model for the new deaths across Brazil from the start date of the deaths is as follows:
RMSE for Linear Regression: 763.417967871343



Polynomial Regression model for new cases across Brazil



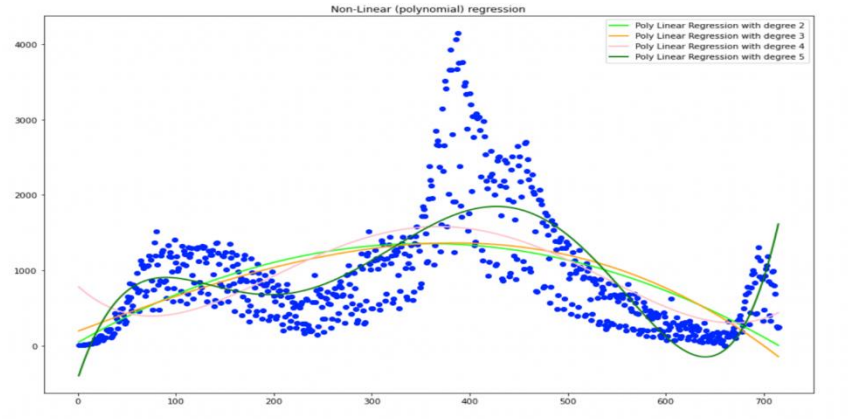
Linear Regression model for new deaths across Brazil

The polynomial regression model for the new deaths across Brazil from the start date of the deaths is as follows:

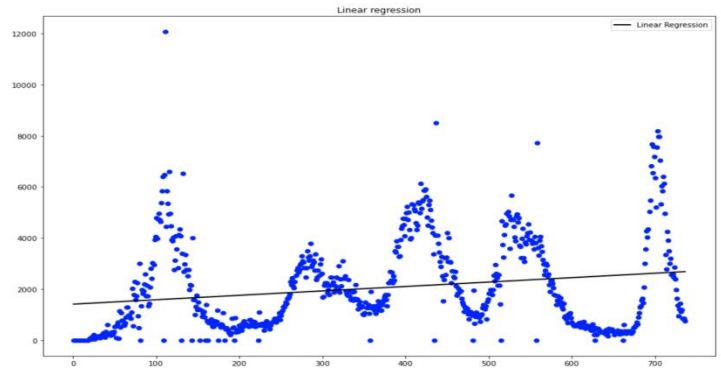
RMSE for Polynomial Regression with degree 2: 651.6114627305836
RMSE for Polynomial Regression with degree 3: 649.1296024416567
RMSE for Polynomial Regression with degree 4: 618.4815951857444
RMSE for Polynomial Regression with degree 5: 500.82959463162194

The linear regression model for the new cases across Pakistan from the start date of the infection is as follows:

RMSE for Linear Regression: 1735.550176636532

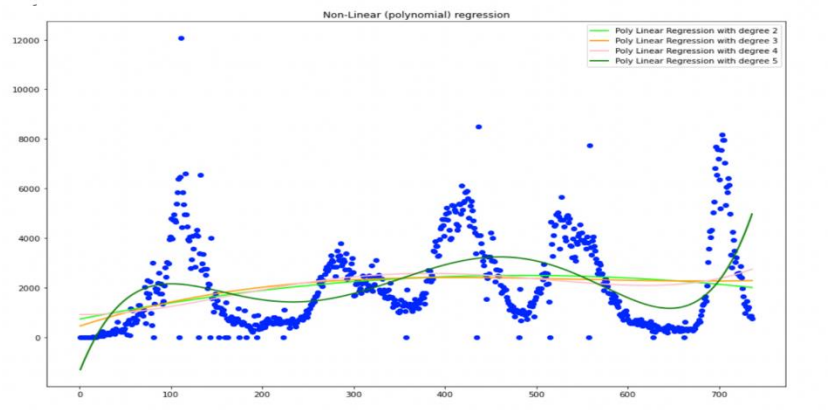


Polynomial Regression model for new deaths across Brazil



Linear Regression model for new cases across Pakistan

The polynomial regression model for the new cases across Pakistan from the start date of the infection is as follows:



Polynomial Regression model for new cases across Pakistan

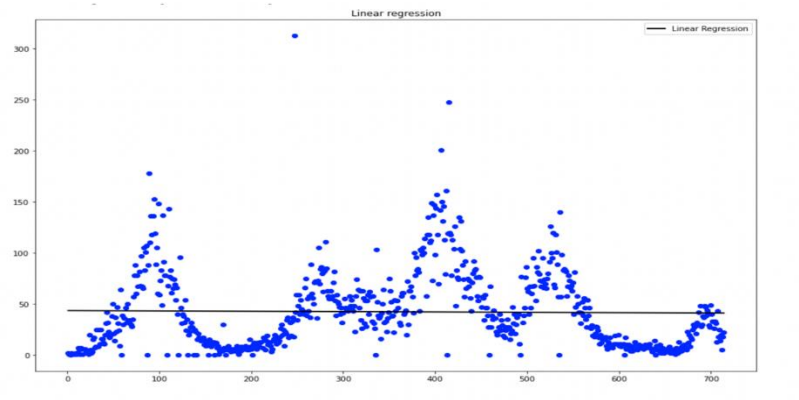
RMSE for Polynomial Regression with degree 2: 1708.3511565822494

RMSE for Polynomial Regression with degree 3: 1705.0261854157989

RMSE for Polynomial Regression with degree 4: 1697.9152986120528

RMSE for Polynomial Regression with degree 5: 1553.894096394138

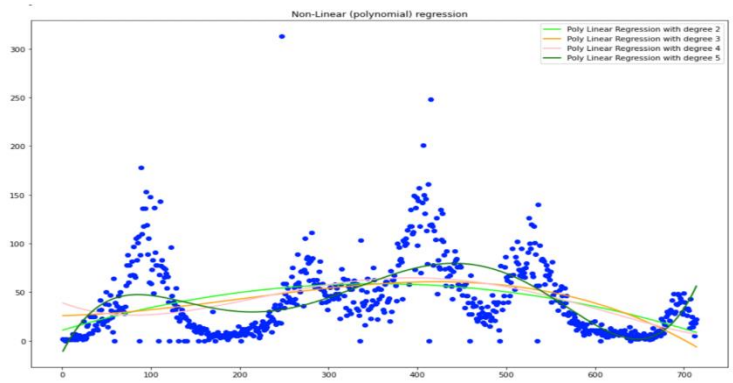
The linear regression model for the new deaths across Pakistan from the start date of the deaths is as follows:



Linear Regression model for new deaths across Pakistan

RMSE for Linear Regression: 38.02243548519602

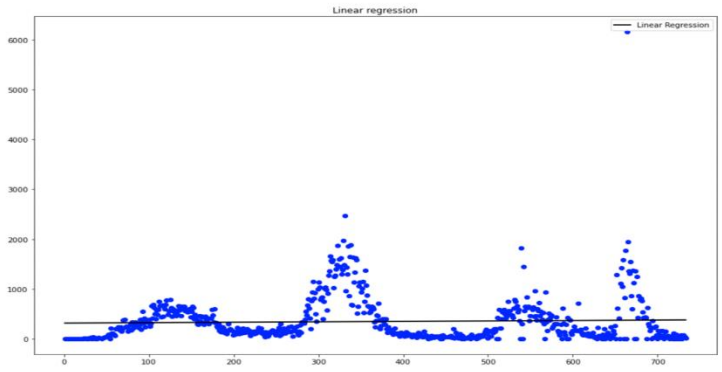
The polynomial regression model for the new deaths across Pakistan from the start date of the deaths is as follows:



Polynomial Regression model for new deaths across Pakistan

RMSE for Polynomial 35.10467409608458
RMSE for Polynomial 34.64611386919402
RMSE for Polynomial 34.36943911840733
RMSE for Polynomial 30.80674131081514

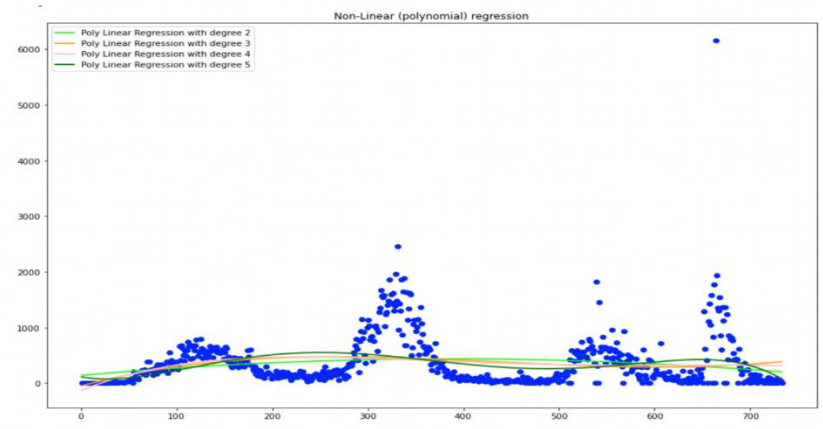
The linear regression model for the new cases across Nigeria from the start date of the infection is as follows:



Linear Regression model for new cases across Nigeria

RMSE for Linear Regression: 450.3012177692088

The polynomial regression model for the new cases across Nigeria from the start date of the infection is as follows:

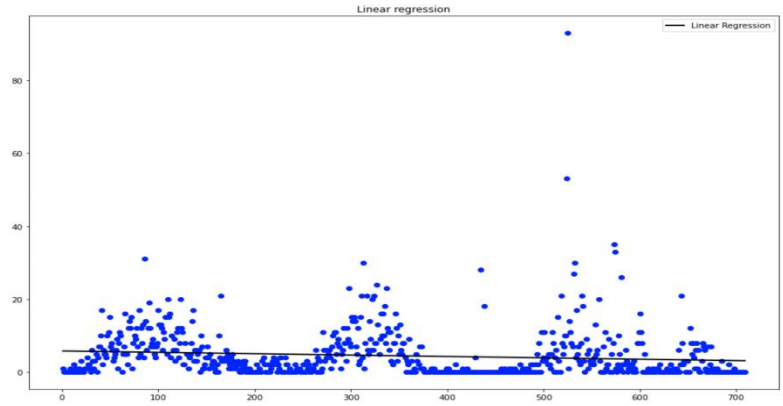


Polynomial Regression model for new cases across Nigeria

RMSE for Polynomial Regression with degree 2: 443.3029885752807

RMSE for Polynomial 437.64579007815996
RMSE for Polynomial 436.8553113620279
RMSE for Polynomial 430.87747411855594

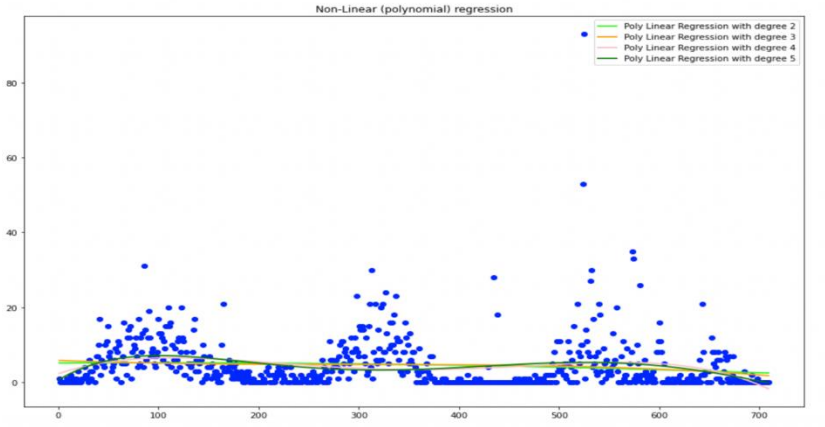
The linear regression model for the Nigeria from the start date of the deaths new deaths across is as follows:



Linear Regression model for new deaths across Nigeria

RMSE for Linear Regression: 6.785153479633278

The polynomial regression model for the new deaths across Nigeria from the start date of the deaths is as follows:



Polynomial Regression model for new deaths across Nigeria

RMSE for Polynomial Regression with degree 2: 6.7794968504038815
RMSE for Polynomial Regression with degree 3: 6.774229350584591
RMSE for Polynomial Regression with degree 4: 6.6712646996433795
RMSE for Polynomial Regression with degree 5: 6.653105962400884

From the above plots of linear regression model and Polynomial regression model we can observe that the root mean square error of linear regression models are high compared to the polynomial regression models. And we can also see that as we increase the degree of the polynomial regression model the root mean square error will be reduced.