Data Science Project Stage-3

Nikhil Bolisetty

1.ILLINOIS State Cases

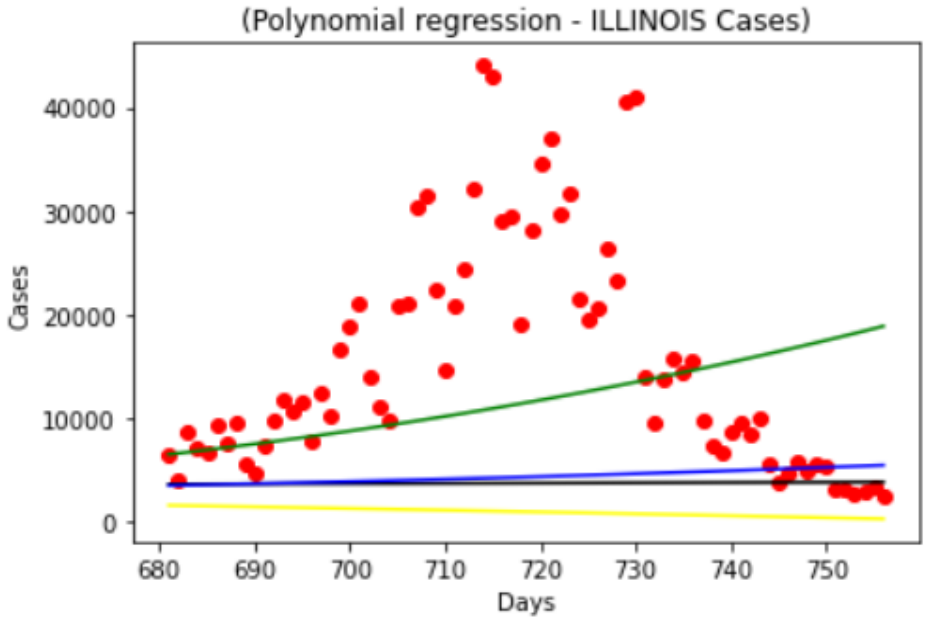
For Regression Model

```
#getiing the rmse value for linear regression model
rmse = sqrt(mean_squared_error(y_test, lr_pred_test))
rmse
```

```
16094.725929956741
```

For Polynomial Model: From the RMSE values the best fit for Illinois state cases is polynomial regression with degree 4

```
RMSE for degree 1 is 16094.725929956741
RMSE for degree 2 is 18192.552578151302
RMSE for degree 3 is 15700.947640312419
RMSE for degree 4 is 12691.911900861121
```



2.ILLINOIS State Cases

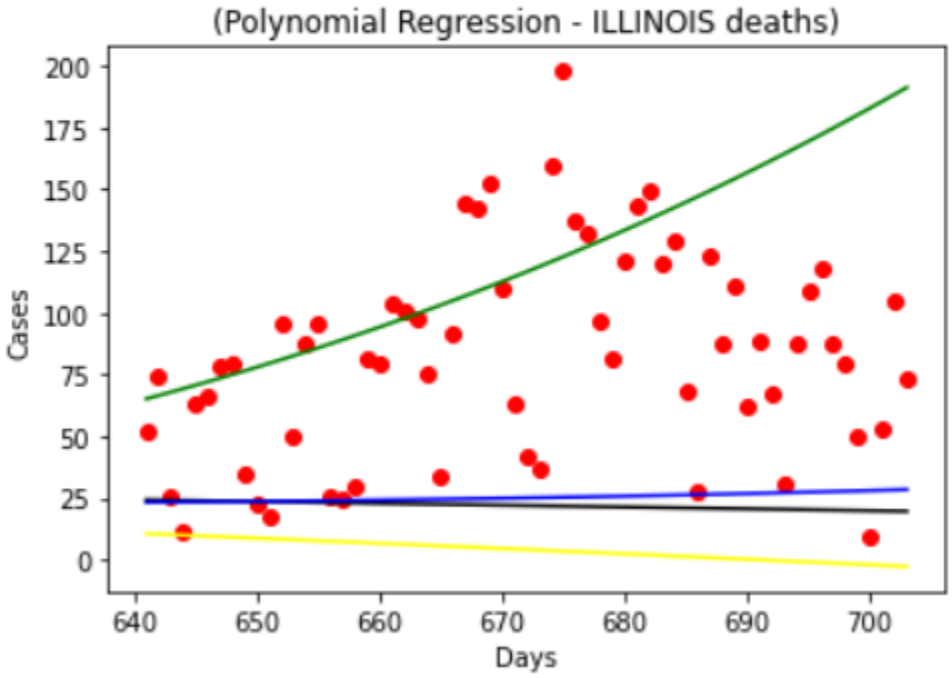
For Regression Model

```
rmse = sqrt(mean_squared_error(y_test, lr_pred_test))
rmse
```

```
73.60445192305579
```

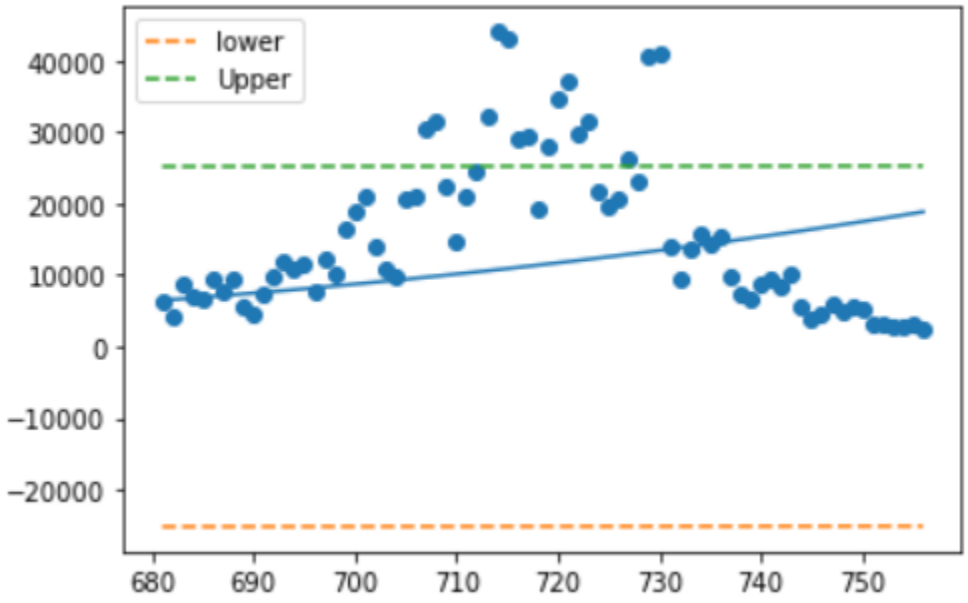
For Polynomial Model: From the RMSE values the best fit for Illinois state cases is polynomial regression with degree 4

```
RMSE for degree 1 is 73.6044519230558
RMSE for degree 2 is 89.27842159272681
RMSE for degree 3 is 70.50746305123863
RMSE for degree 4 is 62.780836819401955
```

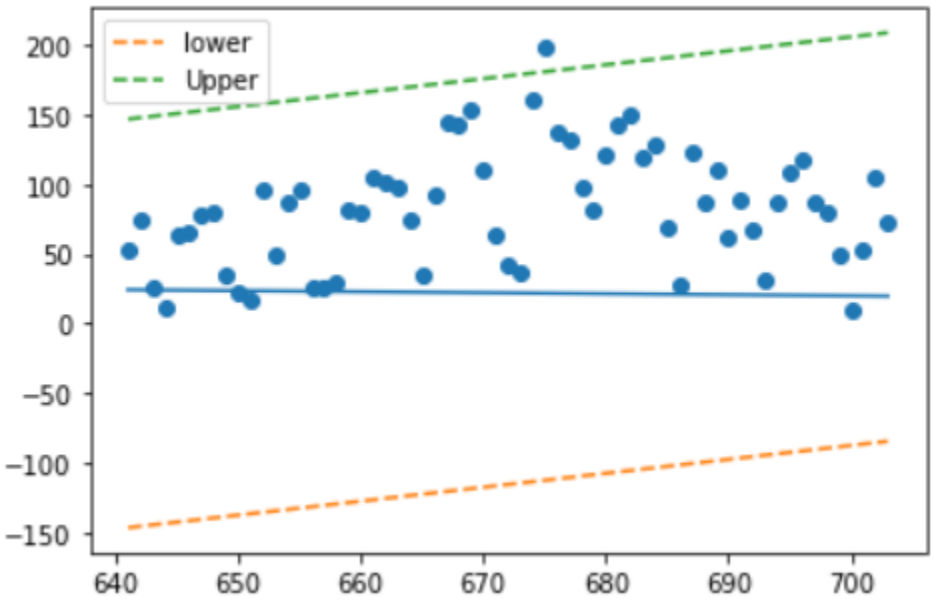


Confidence interval for ILLINOIS Cases and Deaths

1. **Cases**

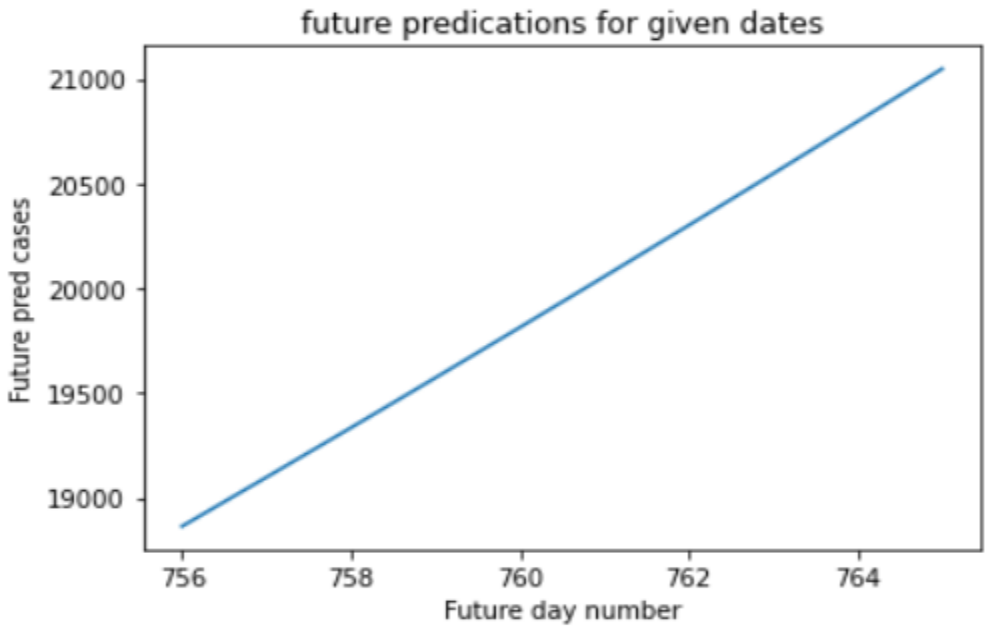


2. **Deaths**

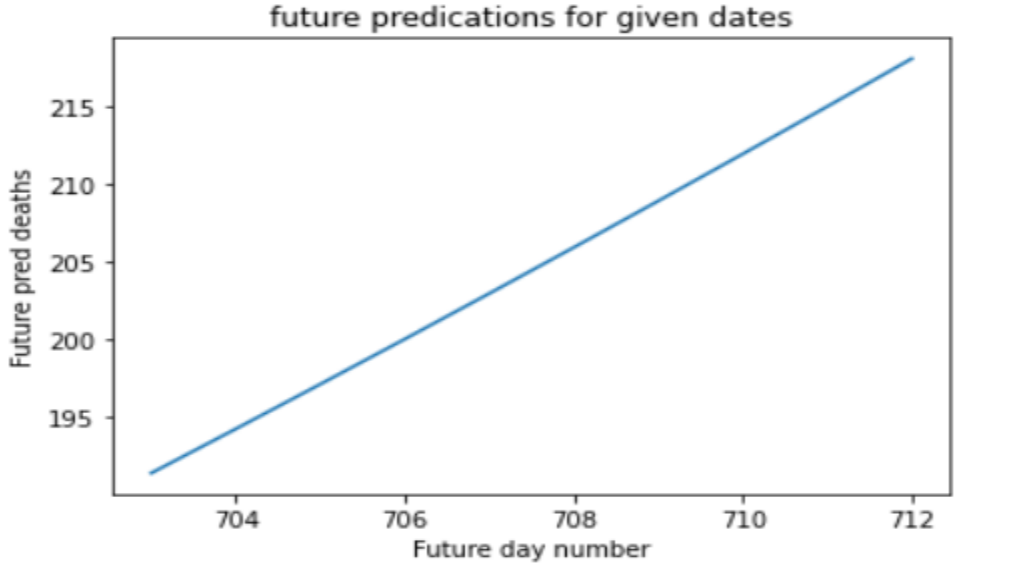


Graph for prediction path (forecast): For Future cases and deaths prediction

1. **Cases:**



2.Deaths:



TOP 5 Counties with High Infection Rate:

1. **Cass County**

Cases:

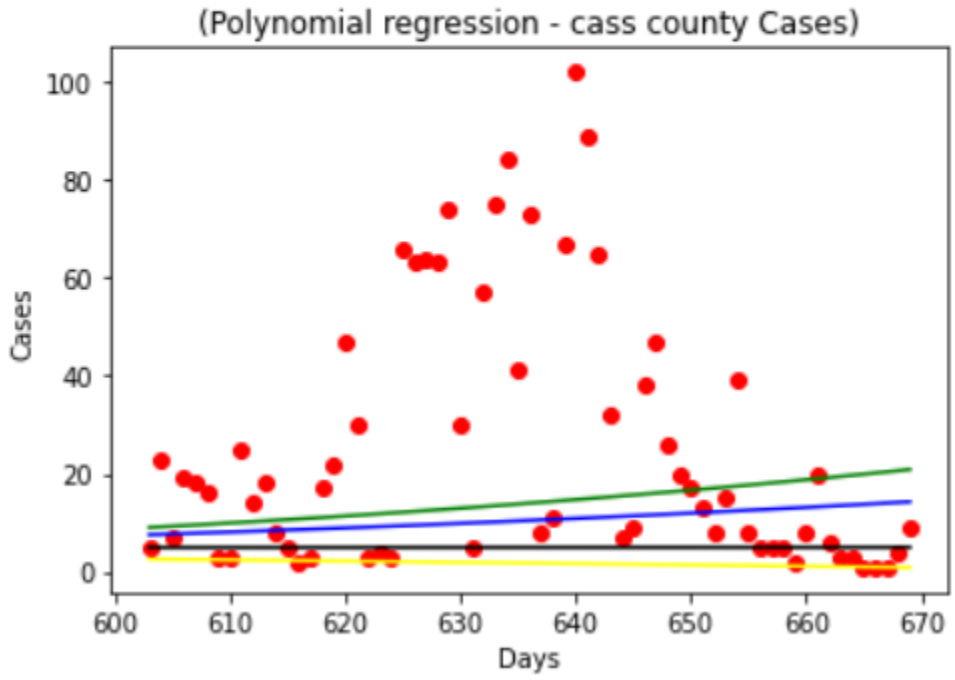
Regression Model:

```
1 rmse = sqrt(mean_squared_error(y_test, lr_pred_test))
2 rmse
```

```
33.1260400548033
```

Polynomial Model: As polynomial regression with degree 4 has less RMSE value it is the best fit.

```
RMSE for degree 1 is 33.1260400548033
RMSE for degree 2 is 35.13932505173838
RMSE for degree 3 is 30.334199040802314
RMSE for degree 4 is 29.106810590741084
```



2. Vermilion County

Cases:

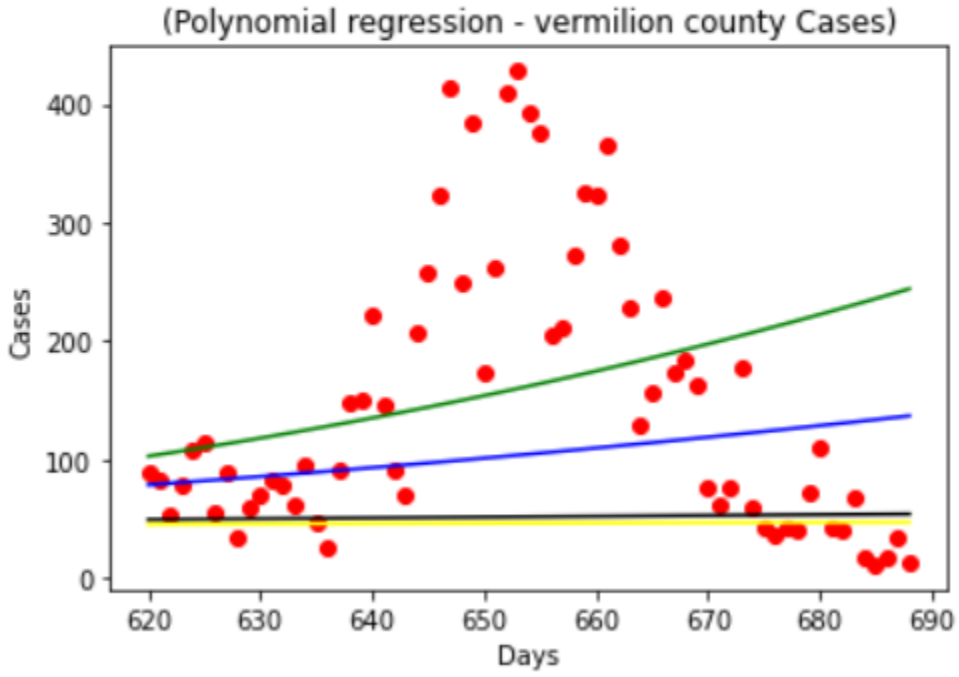
Linear Regression Model:

```
1 rmse = sqrt(mean_squared_error(y_test, lr_pred_test))
2 rmse
```

```
152.3084590952631
```

Polynomial Regression Model: As the polynomial regression with Degree 4 has least RSME value it is the best Fit**.**

```
RMSE for degree 1 is 152.30845909526312
RMSE for degree 2 is 155.7392825623117
RMSE for degree 3 is 127.59252822381627
RMSE for degree 4 is 129.96355244654322
```



3. Brown County

Cases:

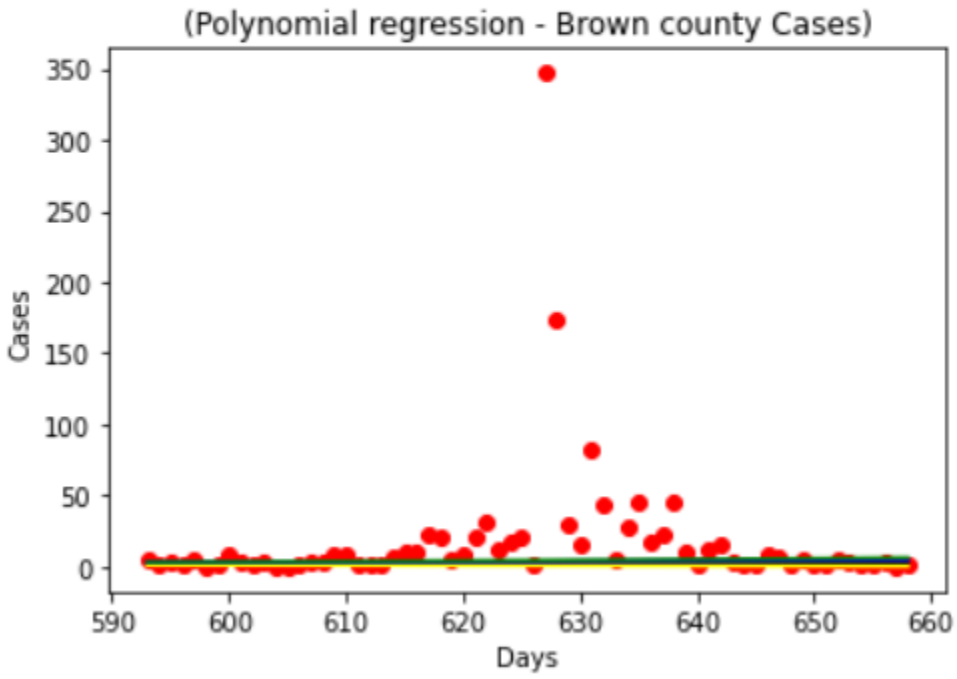
Linear Regression:

```
1 rmse = sqrt(mean_squared_error(y_test, lr_pred_test))
2 rmse
```

```
49.96582392914809
```

Polynomial Regression: As the polynomial regression with Degree 4 has least RSME value it is the best Fit**.**

```
RMSE for degree 1 is 49.96582392914809
RMSE for degree 2 is 50.56460656929865
RMSE for degree 3 is 49.6160023600729
RMSE for degree 4 is 49.39362710090286
```



4. Clay County

Cases:

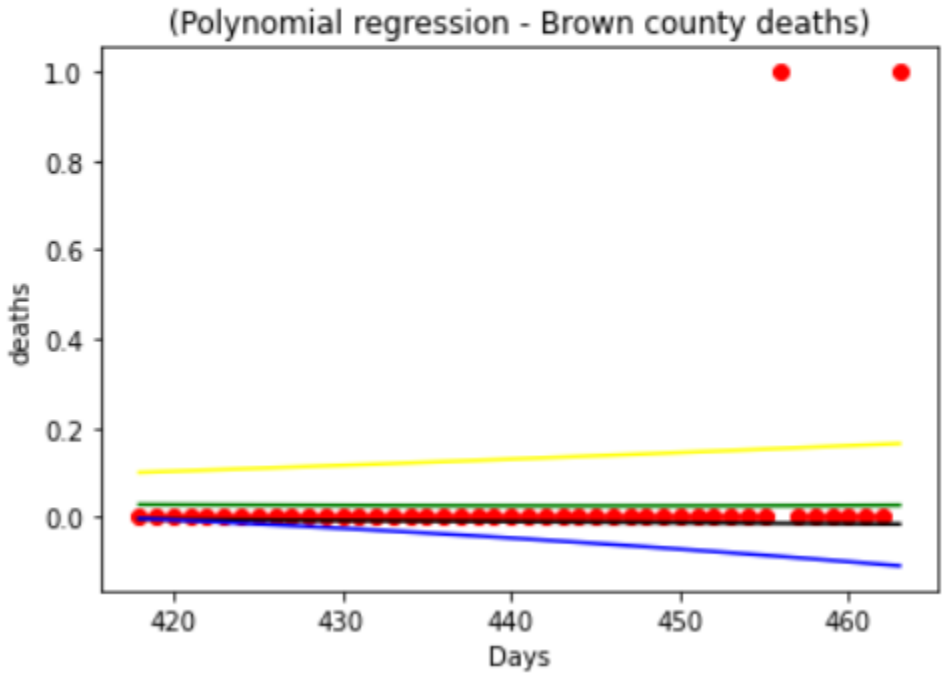
Linear Regression:

```
1 rmse = sqrt(mean_squared_error(y_test, lr_pred_test))
2 rmse
```

```
0.2119472027310547
```

Polynomial regression: As the polynomial regression with Degree 4 has least RSME value it is the best Fit**.**

```
RMSE for degree 1 is 0.2119472027310547
RMSE for degree 2 is 0.21727404629655792
RMSE for degree 3 is 0.23612752546717852
RMSE for degree 4 is 0.20467569638696714
```



5. Johnson County:

Cases:

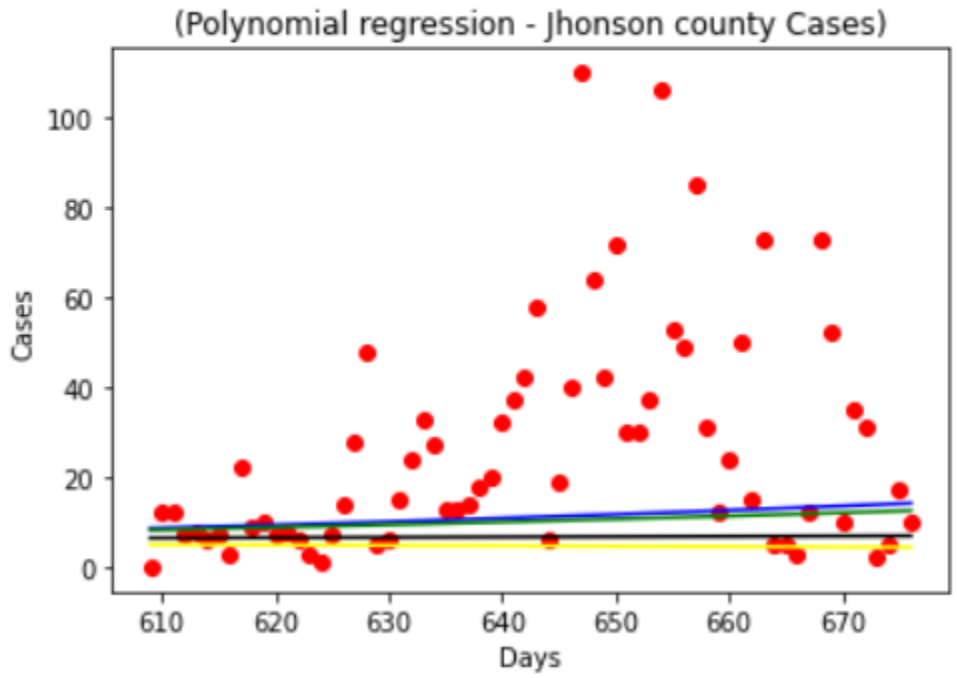
Linear regression:

```
1  rmse = sqrt(mean_squared_error(y_test, lr_pred_test))
2  rmse
```

```
: 31.597524180008534
```

Polynomial regression: As the polynomial regression with Degree 3 has least RSME value it is the best Fit**.**

```
RMSE for degree 1 is 31.597524180008534
RMSE for degree 2 is 32.94572448350795
RMSE for degree 3 is 28.676411759470238
RMSE for degree 4 is 29.25282076088395
```



HYPOTHESIS TESTING:

Can we say that households with age greater than 65 are highly effected to covid 19

In [ ]:
```
1 stats.ttest_ind(a=social_char['Cases'], b= social_char['Households with Senior'],equal_var=False)
```

As the p-value is less than 0.05, we reject the null hypothesis states that households with seniors are not much affected by covid

can we say there is a decrease in High school enrollments due to the increase in the Covid cases

In [ ]:
```
1 stats.ttest_ind(a=social_char['Cases'], b= social_char['Number of Enrollments in High School'],equal_var=False)
```

As the p-value is less than 0.05, we reject the null hypothesis states that Number of Enrollments in High School are not much affected by covid