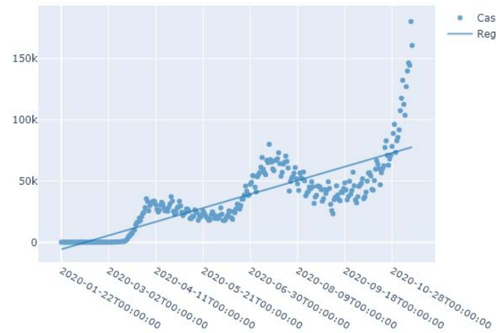
Stage 3 Report

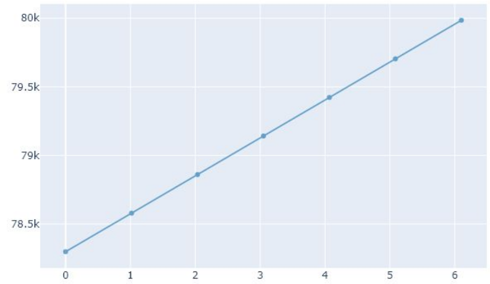
Team Task:

For the team task, we first developed Linear and Non-Linear regression models for cases and deaths in the United States. Below are models of Linear and Non-Linear regression models of cases and deaths.

Linear:

cases

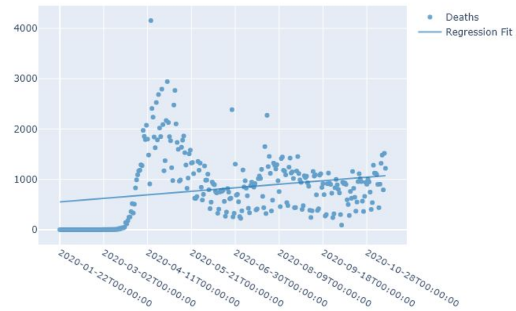


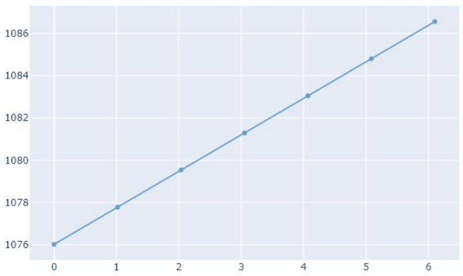


Rmse = 43191.514603438

From the graph above, We can see the trend line of New cases in the United States is gradually rising and on the right side, we can see the predictions of new cases for the next seven days which are rising greatly.

deaths



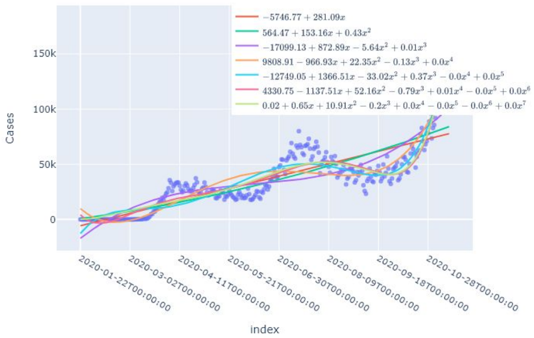


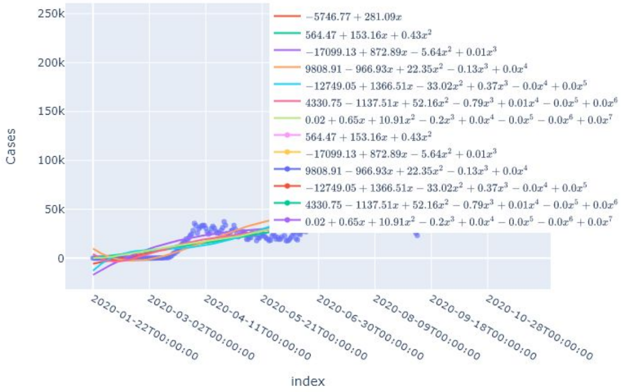
Rmse = 666.3027864314114

From the graph above, we see a linear regression of new deaths in the United States. Compared to new cases regression line, the new death regression line is rising exponentially but slowly. On the right side, we can see predictions of new deaths for the next seven days which is rising to almost 3000 deaths.

Non-Linear:

cases

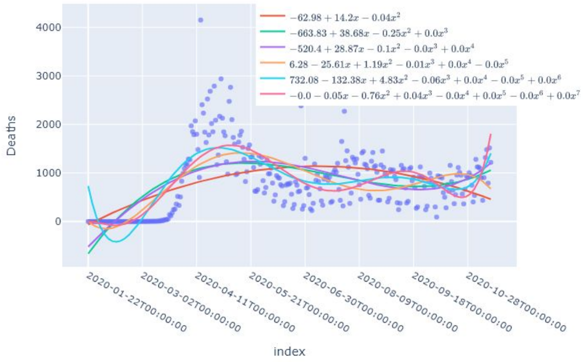


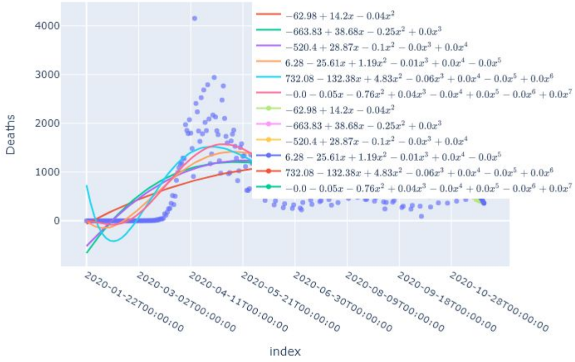


Rmse = 46570.32880050813

For Non-Linear regression of New Cases for the United States, We can see the curvature of the slope rise by a lot with predictions hitting to almost 200k new cases per day. With Non-linear regression we can see a better analysis of predictions compared to just linear regression as we can see the gradual rise of new cases.

deaths



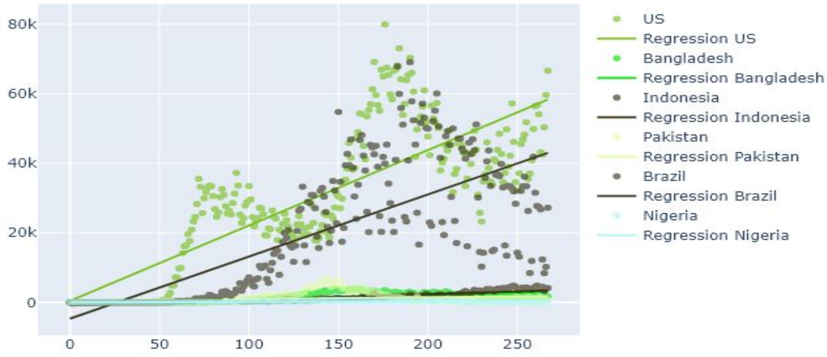


Degree 7 Rmse = 807.8588900795557

For Nonlinear regression of new deaths in the United States, the regression line is similar to the linear regression line of new deaths as it relatively stays below 1000 deaths. The prediction however gives us a better look as to where deaths will lead to as the non-linear regression line rises towards 2000-3000 deaths per day.

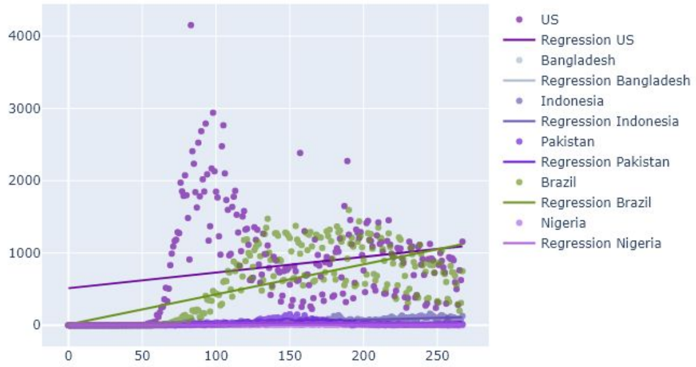
Compare trends with other countries:

cases



Compared to other countries with similar populations, the United States has the highest regression line which is rapidly increasing. The closest country to the United States regression line is Brazil which has a similar slope to The United States line. Other countries have a very small slope and low COVID-19 case numbers.

Deaths

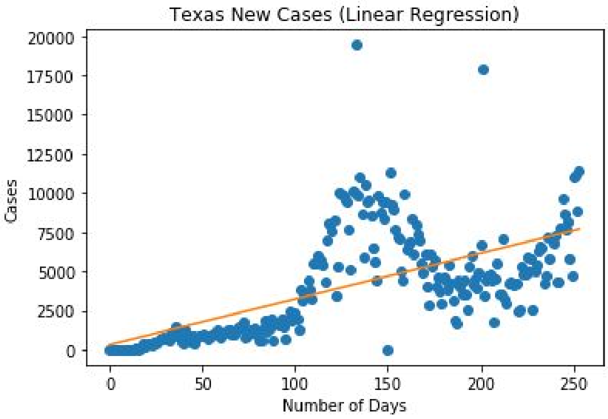


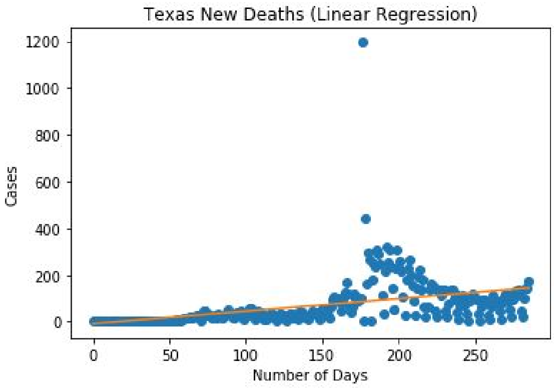
Looking at the new deaths regression line and comparing it with other countries, the United States and Brazil are close to each other with the United States regression line stabilizing and going lower but Brazil's regression line is increasing. Similar to new cases, Other countries such as Bangladesh, Indonesia, and Nigeria regression line are below 1000 deaths and a lot lower than United States and Brazil's regression line.

Member Tasks:

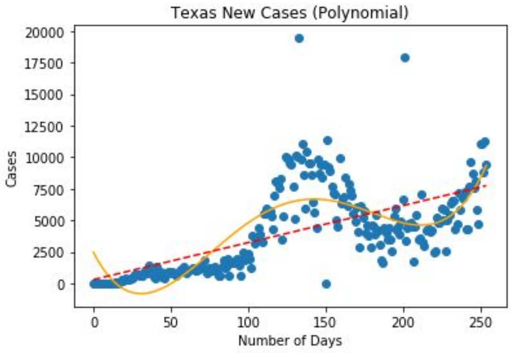
Harinder Badesha :

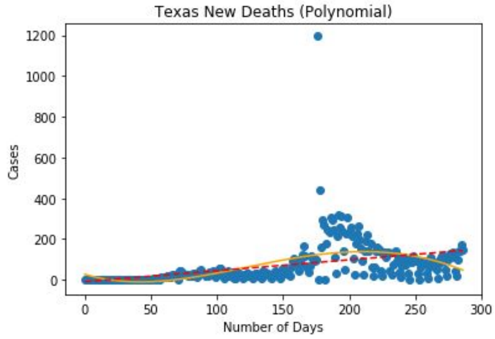
For the first member task, we utilized Linear and Non-linear regression models for a single state and compared its trends of cases and deaths.



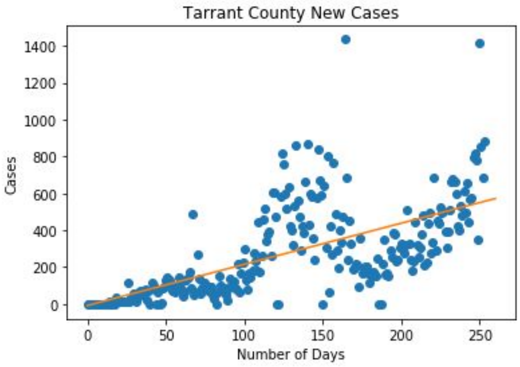


From the two models, we can see that cases have a higher slope than deaths and that the linear trend for cases is rapidly increasing with predictions going to almost going to 100,000 cases per day. With Non--Linear Regression, we get a different look at the trend.

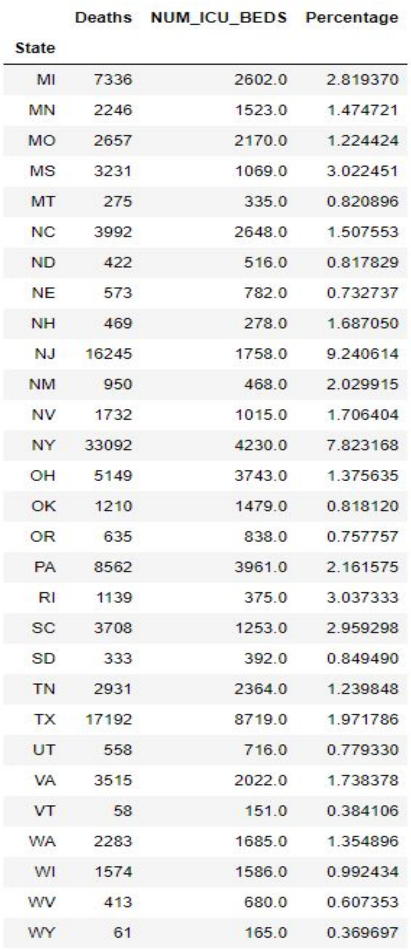




With Non-Linear regression, we can better see the trend line and that for new deaths it is relatively going down instead. With cases, the trend line seems similar to the trend line of Linear Regression. When using Linear regression on the top 5 counties in Texas, We aim to see which county is at most risk. Below is the linear regression of Tarrant County:



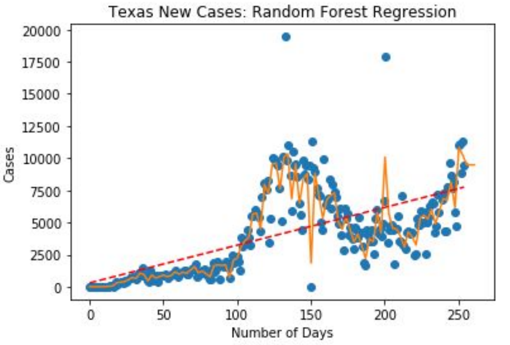
Compared to other counties, Tarrant County had the highest increasing trend line which shows that it is the county which is most at risk of spreading COVID-19.

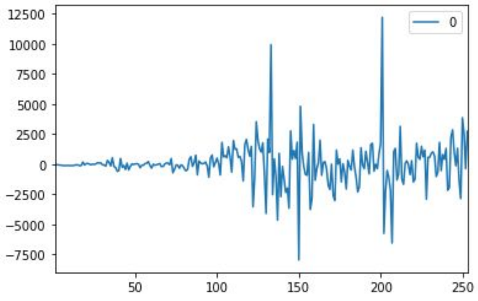


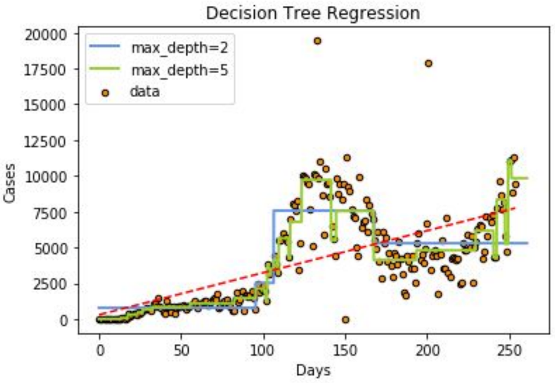
State	Deaths	NUM_ICU_BEDS	Percentage
MI	7336	2602.0	2.819370
MN	2246	1523.0	1.474721
MO	2657	2170.0	1.224424
MS	3231	1069.0	3.022451
MT	275	335.0	0.820896
NC	3992	2648.0	1.507553
ND	422	516.0	0.817829
NE	573	782.0	0.732737
NH	469	278.0	1.687050
NJ	16245	1758.0	9.240614
NM	950	468.0	2.029915
NV	1732	1015.0	1.706404
NY	33092	4230.0	7.823168
OH	5149	3743.0	1.375635
OK	1210	1479.0	0.818120
OR	635	838.0	0.757757
PA	8562	3961.0	2.161575
RI	1139	375.0	3.037333
SC	3708	1253.0	2.959298
SD	333	392.0	0.849490
TN	2931	2364.0	1.239848
TX	17192	8719.0	1.971786
UT	558	716.0	0.779330
VA	3515	2022.0	1.738378
VT	58	151.0	0.384106
WA	2283	1685.0	1.354896
WI	1574	1586.0	0.992434
WV	413	680.0	0.607353
WY	61	165.0	0.369697

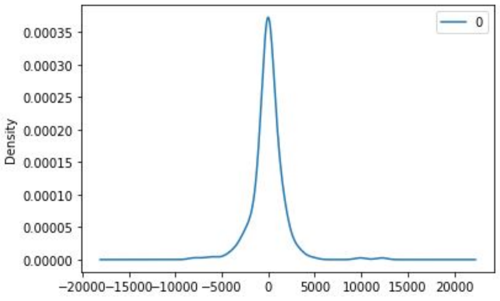
Next task was to utilize the hospital data set and calculate the point of no return of a state. To do this, I calculated the total deaths of each state and divided it by ICU Beds. In my Analysis, I discovered that New Jersey had the largest number of deaths to ICU Beds percentage and recently Wisconsin is closing in to the “point of no return” as its total deaths are nearing the maximum number of ICU Beds.

Decision Tree, Random Forest, and ARIMA:



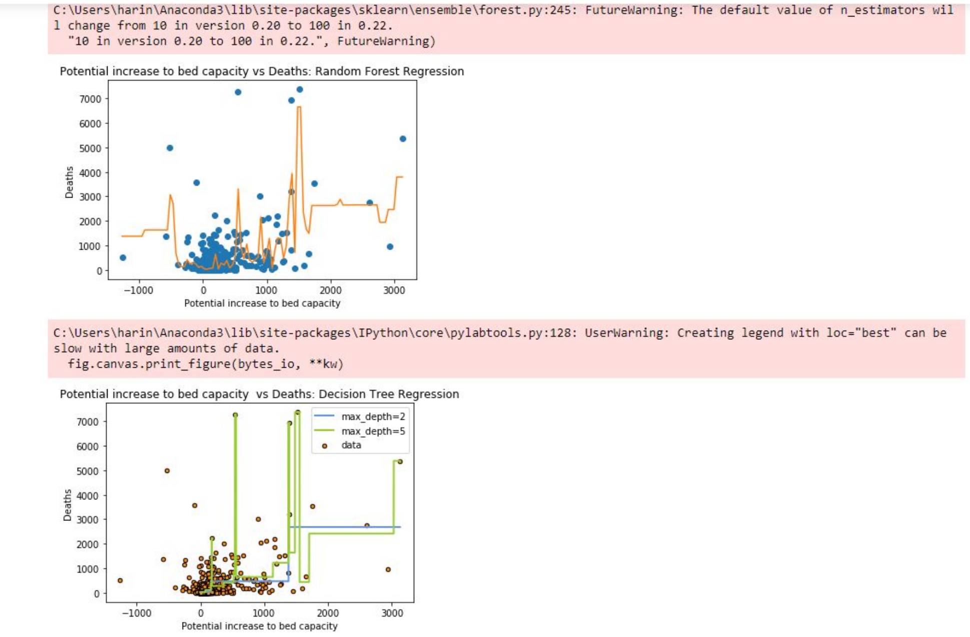






As the graph shows, Random Forest and Decision Tree regression have similar trends whereas the ARIMA plot is different.

I then used Random forest and Decision Trees for hospital bed data variables, Below are the some of the following plots:



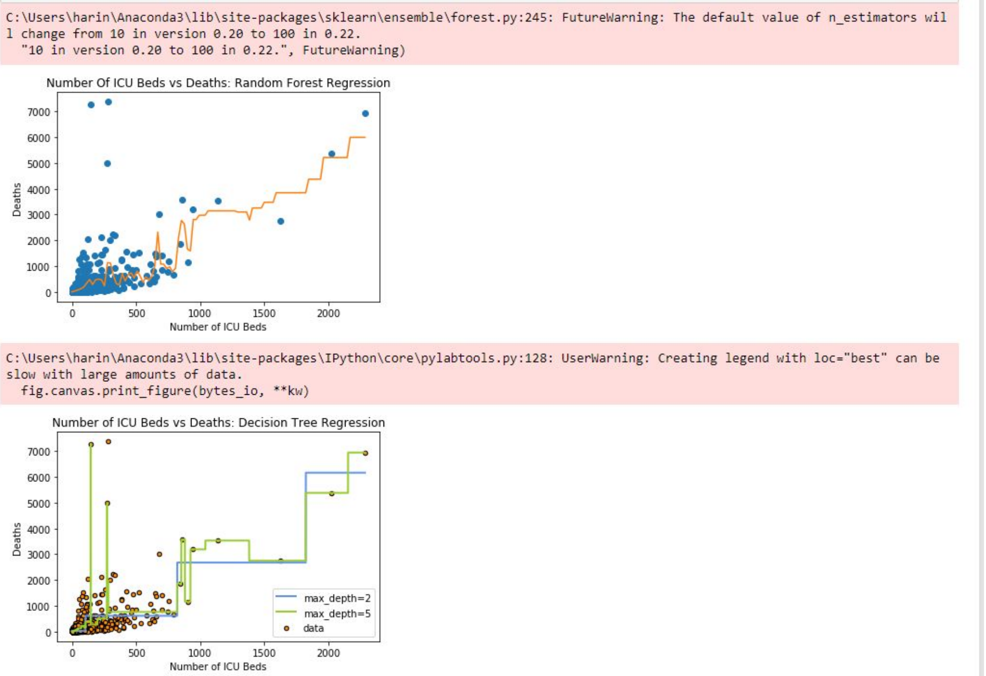
```
C:\Users\harin\Anaconda3\lib\site-packages\sklearn\ensemble\forest.py:245: FutureWarning: The default value of n_estimators wil
l change from 10 in version 0.20 to 100 in 0.22.
  "10 in version 0.20 to 100 in 0.22.", FutureWarning)
```

Potential increase to bed capacity vs Deaths: Random Forest Regression
Deaths
Potential increase to bed capacity
0
1000
2000
3000
4000
5000
6000
7000
−1000
0
1000
2000
3000

```
C:\Users\harin\Anaconda3\lib\site-packages\IPython\core\pylabtools.py:128: UserWarning: Creating legend with loc="best" can be
slow with large amounts of data.
  fig.canvas.print_figure(bytes_io, **kw)
```

Potential increase to bed capacity vs Deaths: Decision Tree Regression
max_depth=2
max_depth=5
data
Deaths
Potential increase to bed capacity
0
1000
2000
3000
4000
5000
6000
7000
−1000
0
1000
2000
3000

Most of the enrichment variables had a similar trend in the random forest regression plot. We can see that with ICU beds vs Deaths of COVID, there is a strong trend that is rapidly increasing indicating that as there are more ICU beds, The deaths will increase to almost 7,000 in total. With Bed utilization, the trend line is a little bit more different. Forest Regression RMSE seemed to be greater in bed capacity and bed utilization but it's RMSE is lesser in number of staffed beds, ventilator usage and ICU beds. ICU beds and Staffed Beds are relatively important variables because their trend has a very positive slope which predicts that the more beds that are staffed, the more deaths will occur.

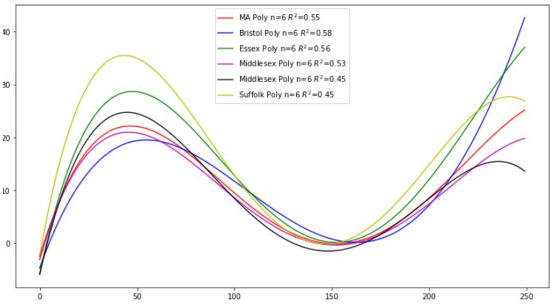


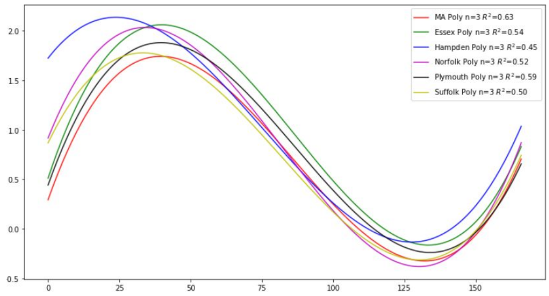
```
C:\Users\harin\Anaconda3\lib\site-packages\sklearn\ensemble\forest.py:245: FutureWarning: The default value of n_estimators wil
l change from 10 in version 0.20 to 100 in 0.22.
  "10 in version 0.20 to 100 in 0.22.", FutureWarning)
```

```
C:\Users\harin\Anaconda3\lib\site-packages\IPython\core\pylabtools.py:128: UserWarning: Creating legend with loc="best" can be
slow with large amounts of data.
  fig.canvas.print_figure(bytes_io, **kw)
```

Sanam Khalili:

In the first part, after I checked the R^2 and RMSE for polynomial and linear regression, I figured it out that polynomial n = 4 and n=3 for cases and deaths fit my data better. In cases Bristol county and in death Plymouth counties are most at risk because of their slope.





In the hospital dataset, I used ICU beds and then, I calculated the 7 days predictions for all states. I compare 7 days prediction of deaths with the number of ICU beds for all states. There is no point of no return for states.

By plotting predictions based on Random Forest, Decision Tree and ARIMA and calculating their RMSE, we can conclude that the Random Forest model fits the data better.

I selected “65 years and over”, “Number of White and Asian”, “number of hispanic latino”,”median ages”and “ male ratio per 100 female” from demographic dataset, and the result are:

By increasing the number of65 years and older the number of cases also increases.

Number of White and Asian does not have a very significant effect on the number of cases.

By increasing the number of hispanic latino number of cases is increasing..

Number of median ages affects the number of cases less than other variables.

Number of male ratios per 100 females does not affect the number of cases significantly.

Number of male ratios per 100 females does not affect the number of deaths significantly.

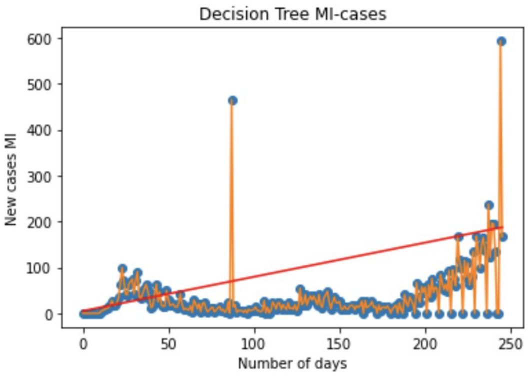
The median age does not affect the number of deaths significantly.

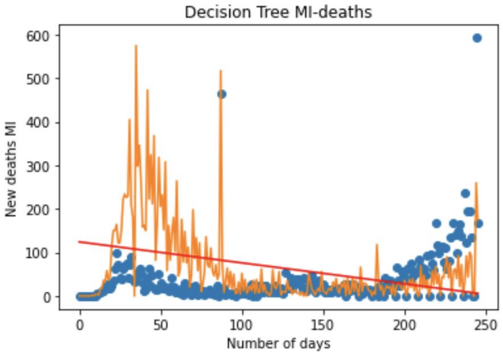
The graph shows that by increasing the number of hispanic, the number of deaths increased a little bit.

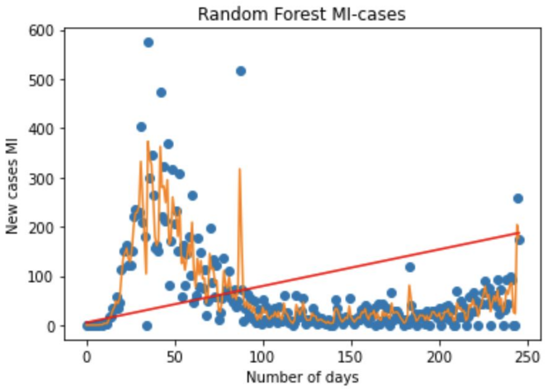
Isaac Taylor:

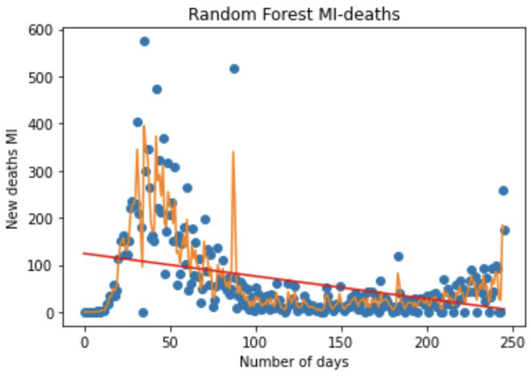
Nadia Doudou:

From the analysis of all the models, Decision Tree seems to be the model that fits the data best. For both the number of cases and number of deaths, the R^2 was 1 and the RMSE was relatively low compared to the other models. This result indicates that the model is doing good; the predictions can be said to be accurate. Random Forest Model is the second top model that works for the data with a R^2 value of 0.90. The Linear and polynomial model outperformed the ARIMA model.The reason why the R^2 value for the ARIMA model was the smallest is because ARIMA model focuses more on the moving average of the data so the R^2 metrics is not really of a big importance in this case.



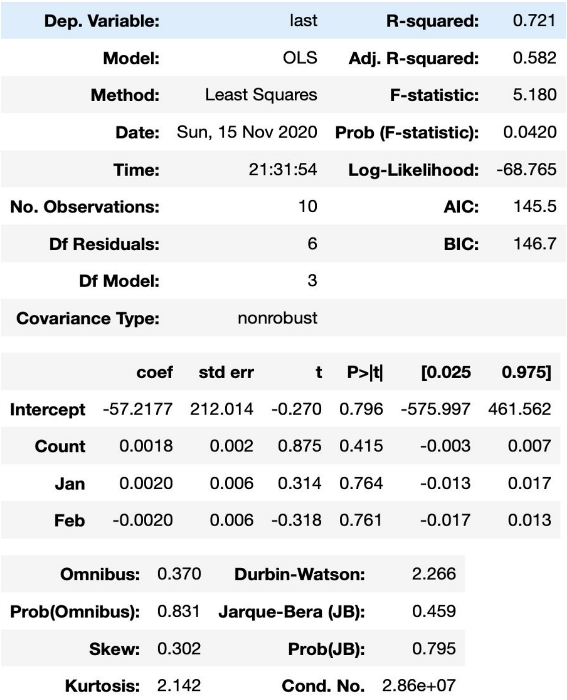






From stage II, we have already seen that there was a strong correlation between the employment level and the number of Covid cases. By modeling a multiple linear regression, the p-values were analyzed and checked if the values are less than 0.05 that indicates that there is a relationship between the variables and the number of cases and deaths and so those variables contribute to the spread of the virus. From this study, it confirmed the result from stage II as the p-values were less than 0.05 so the confidence interval doesn’t include zero.

All values in the confidence interval are plausible values for the parameter. Below is the summary of the model.



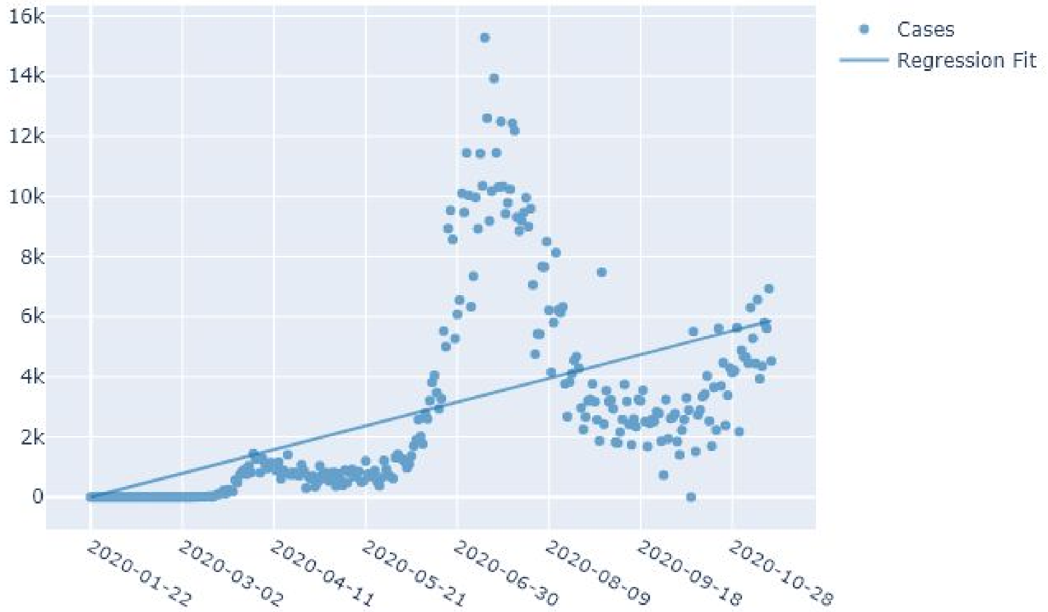
Dep. Variable:	last	**R-squared:**	0.721
Model:	OLS	**Adj. R-squared:**	0.582
Method:	Least Squares	**F-statistic:**	5.180
Date:	Sun, 15 Nov 2020	**Prob (F-statistic):**	0.0420
Time:	21:31:54	**Log-Likelihood:**	-68.765
No. Observations:	10	**AIC:**	145.5
Df Residuals:	6	**BIC:**	146.7
Df Model:	3		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	-57.2177	212.014	-0.270	0.796	-575.997	461.562
Count	0.0018	0.002	0.875	0.415	-0.003	0.007
Jan	0.0020	0.006	0.314	0.764	-0.013	0.017
Feb	-0.0020	0.006	-0.318	0.761	-0.017	0.013

Omnibus:	0.370	**Durbin-Watson:**	2.266
Prob(Omnibus):	0.831	**Jarque-Bera (JB):**	0.459
Skew:	0.302	**Prob(JB):**	0.795
Kurtosis:	2.142	**Cond. No.**	2.86e+07

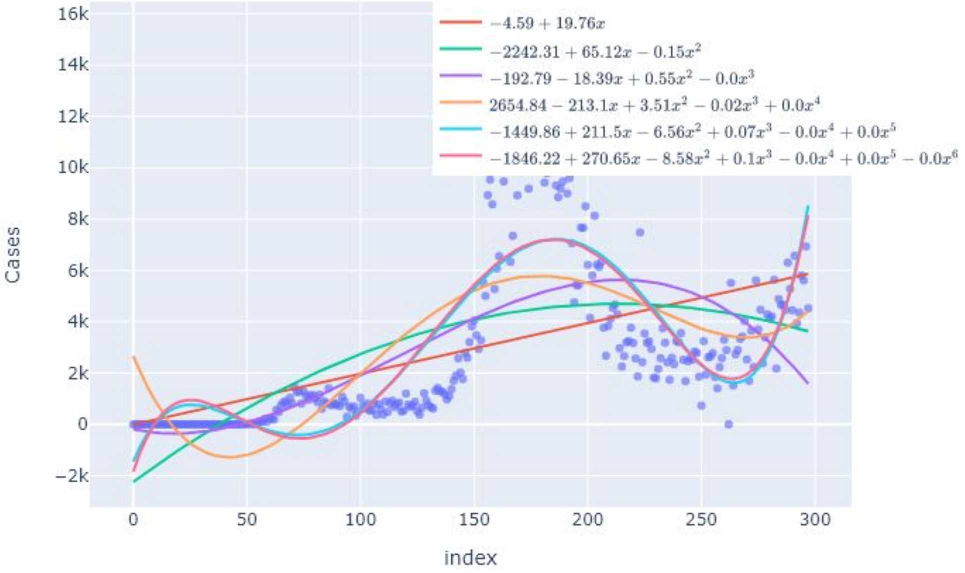
Ali Altamimi:

Fit a linear regression model using sklearn's LinearRegression package to see the trend of Florida’s cases:

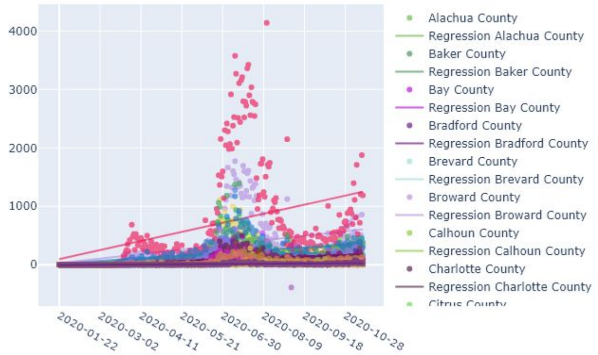


Root mean square error (RMSE) for this linear regression is 3220.

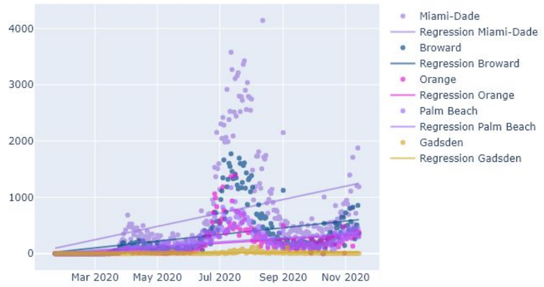
Fit a nonlinear regression model using sklearn's LinearRegression package to see the trend of Florida's cases:



Root mean square error (RMSE) for this none-linear regression when the degree is 5 is 3846. After that I observed all Florida's counties trends:

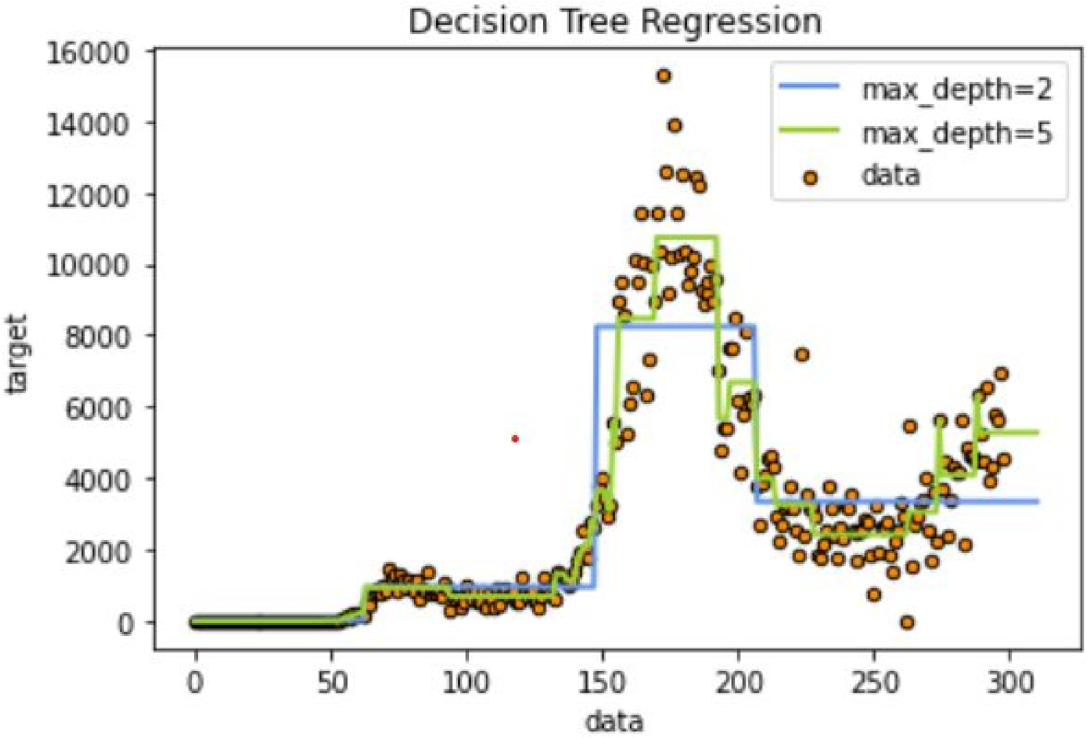


From that i concluded that the top 5 counties with cases are Miami, Broward, Orange, Palm, Gadsden



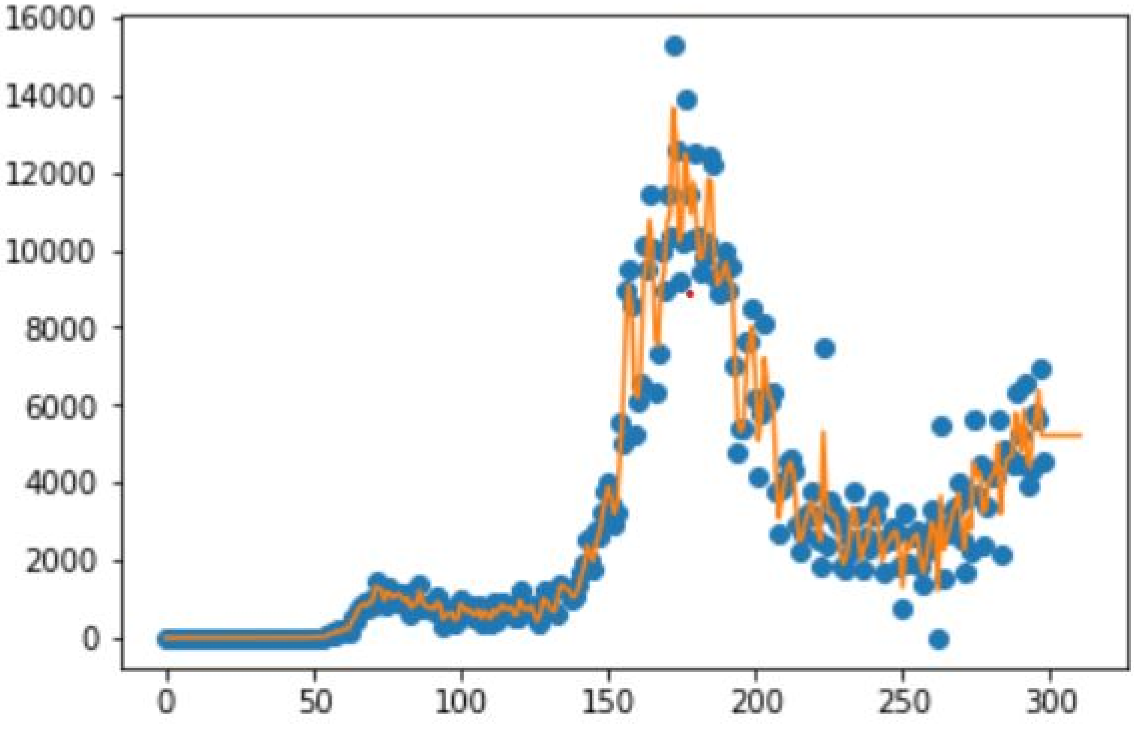
After that I utilize the hospital data to calculate the point of no return for a state. The result for florida’s hospital utilized bed is 0.33 which mean there is no point of no return.

decision tree:



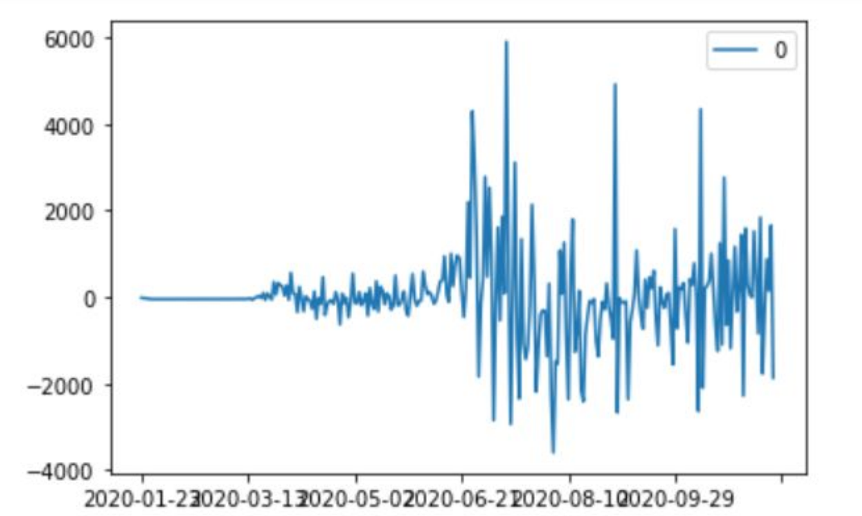
The RMSE for DT is 157.90731085373952

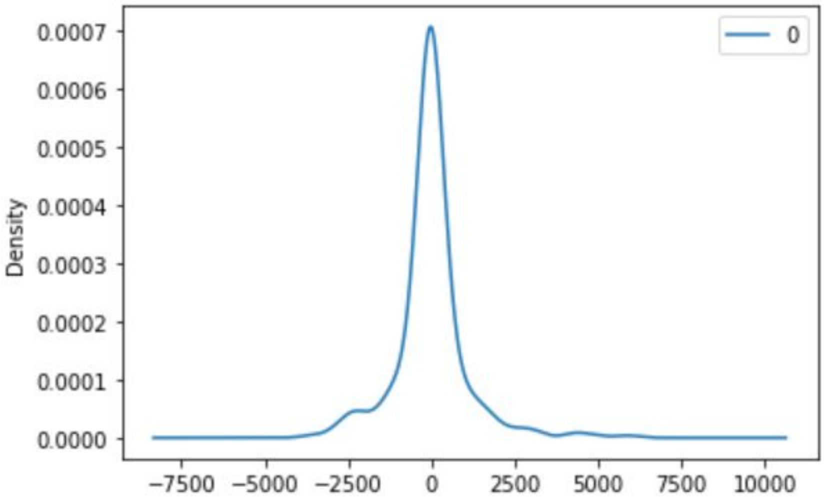
random forest:

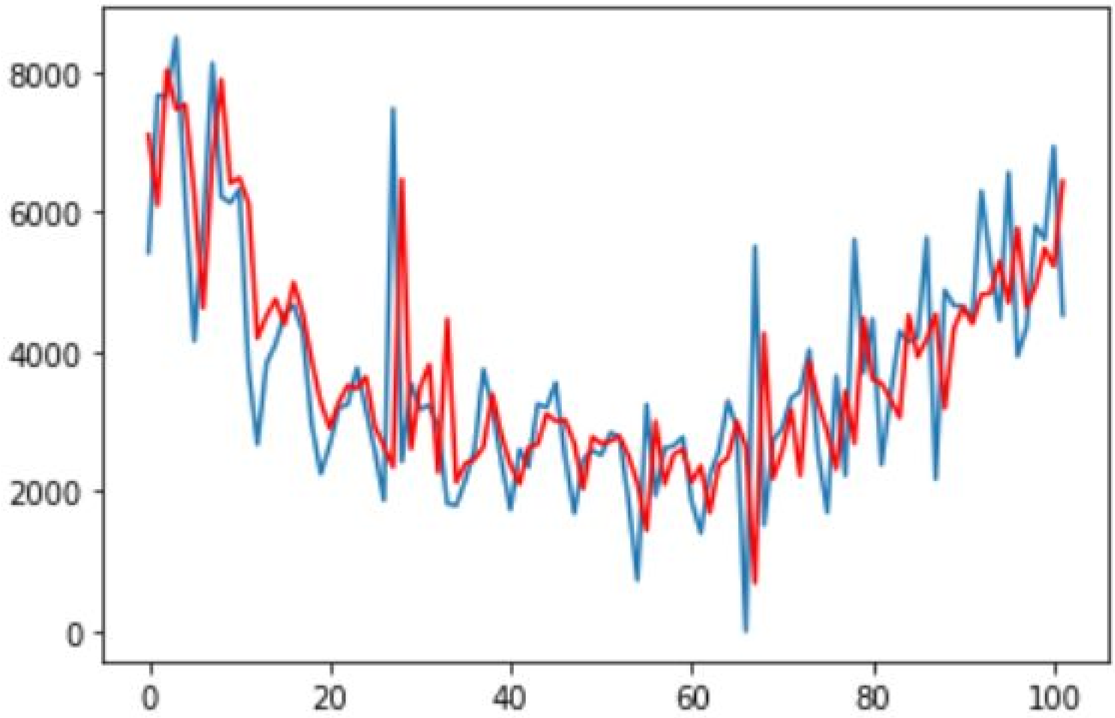


The RMSE for RF is 157.90731085373952

ARIMA



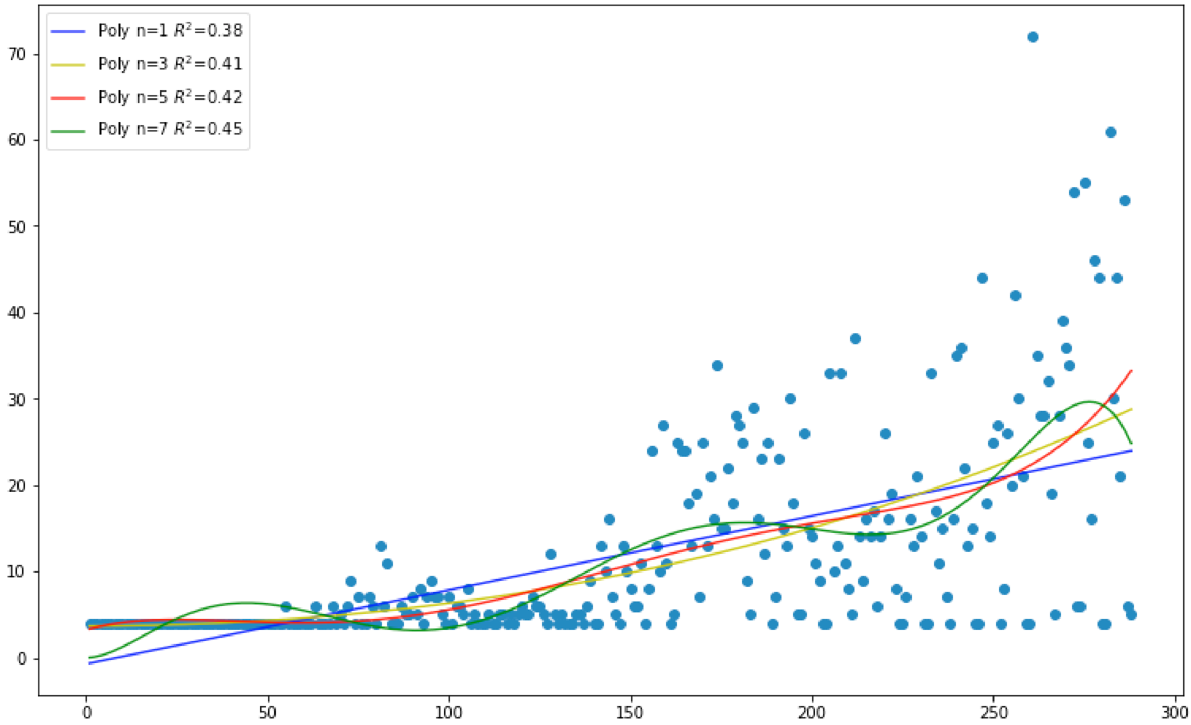




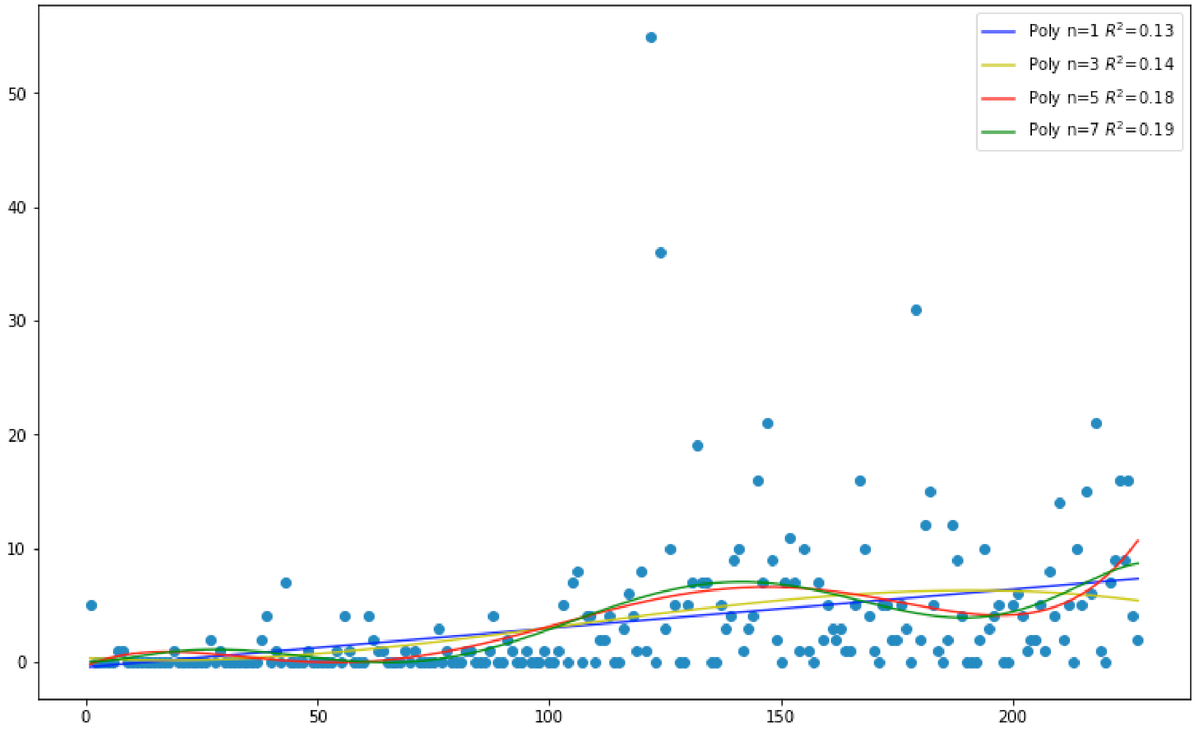
RMSE: 1334.692

Isaac Taylor:

In this stage I analyzed the state of Texas by using regression. I found that some of the polynomial regressors seemed to fit the New Cases data better than the degree one regression models, and after comparing each model's RMSE, I saw that the degree one regression model had the highest error. Despite that, it can be argued that the degree one model performed fairly well in comparison to its counterparts for predicting new cases. In some cases it may be best to choose this model because of its simplicity. The higher degree models were much more complex. I performed the same calculations for New Deaths.



The Image above shows each of the models for daily new case predictions

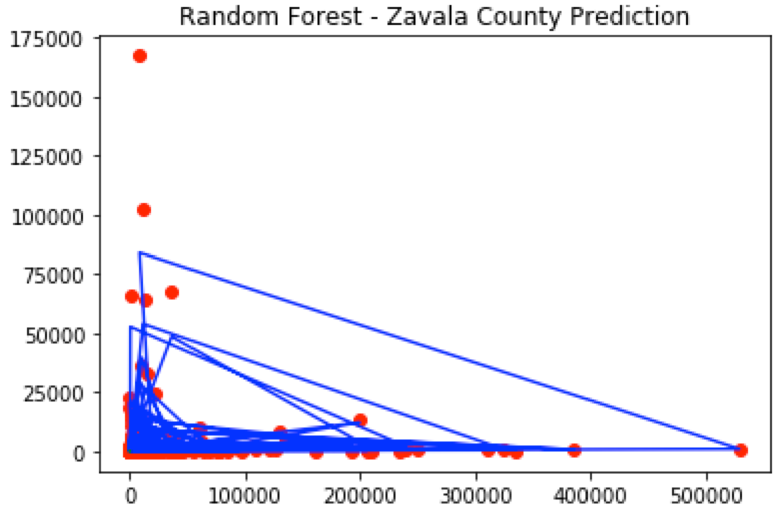


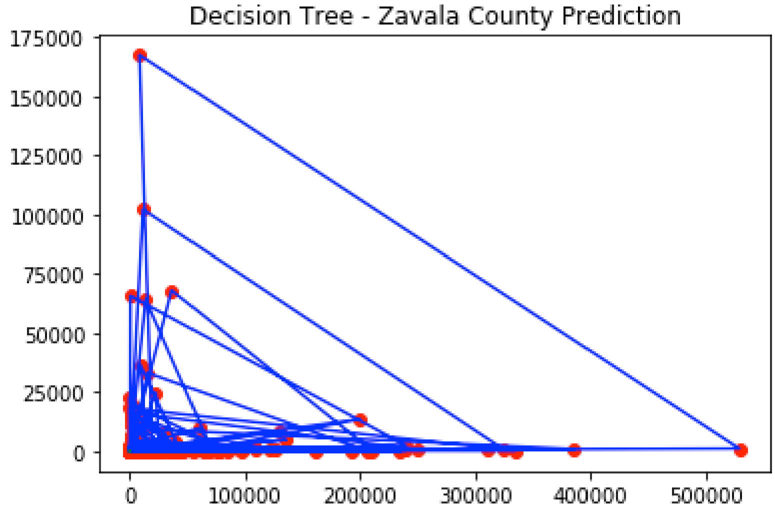
The Image above shows each of the models for daily new death predictions

Another particularly interesting part of this stage occured when I tried to predict cases with random forest and decision tree regressors for Zavala County in Texas. Both models did poorly, but the Random Forest Regressor performed the best and had the closet prediction to the number of cases in Zavala County. This task was interesting because multiple linear regression was used. I used the social dataset and predicted cases using the following:

non_family	non_family_65	single_female_no_husband	computer_internet	usa_born_total_pop

These features are regarding household type.


Random Forest - Zavala County Prediction


Decision Tree - Zavala County Prediction