Orie 3120 Final Project :

Data Analysis on Coronavirus Disease (COVID- 19) Cases in United States and New York State

May 22nd, 2021

Introduction

Since the very beginning of 2020, the world has been attacked by the new pandemic called COVID-19, and the human race is facing one of its most severe situations. We want to take a closer look into the COVID infections through a data-driven perspective. This led us to explore the COVID-19 dataset from Centers for Disease Control and Prevention (CDC) that contains the daily update of COVID cases by countries, as well as by states in the U.S., including cases such as confirmed, deaths, recovered, case fatality ratio, testing rate, etc. With the data, we hope to get a more detailed look in specific areas. For the first part, our team focused on the U.S. as a whole to analyze the daily trend. For the second part, we analyzed the data in New York State with greater detail. We hope to answer the following questions:

1. What did the trend of COVID spread look like? When did it increase and decrease?
2. Does vaccination have any effect on the infection rate?
3. What will the trend for the future next two months look like?
4. How is the population related to the COVID spread in counties of New York State?
5. Where in New York State do we see most cases? How is the population related to the death rate in New York State?

Part I. COVID Spread in the U.S. as a whole

Visualizations

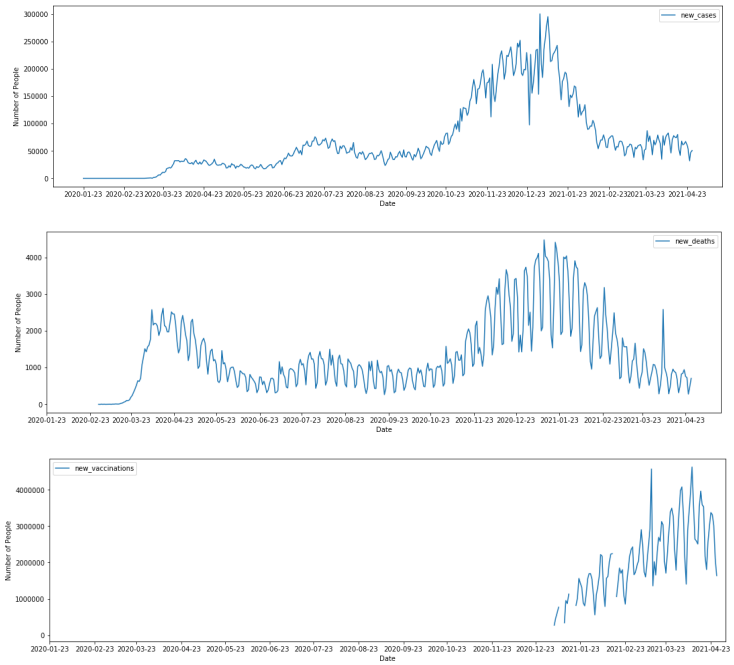


Figure 1. Date vs. Number of New Cases, Deaths, Vaccinations in the U.S. (note the scales are different for all three)

In the figure above, we graphed the number of new cases, deaths, and vaccinations in the United States against the date since Jan 23, 2020. In the first graph, we see that before October, 2020, the number of cases was increasing relatively slowly. However, there was a rapid increase

since then and the number reached a peak around Christmas. We presumed that this was due to various large gatherings at that time, particularly events related to the election. Though this presumption needs further analysis to back up. Furthermore, we see that after January 2021, the number of cases decreased at a steady rate. We presumed that this can be mainly attributed to the increasing number of vaccinations across the country, as we see in the third graph. We will analyze the correlation between vaccination and infection rate later.

In addition, we can conclude that the number of new deaths in the second graph is closely correlated with the trend of new cases in the first graph. This is easy to comprehend since more people died when more people were infected. To see this more clearly, we plotted another figure below:

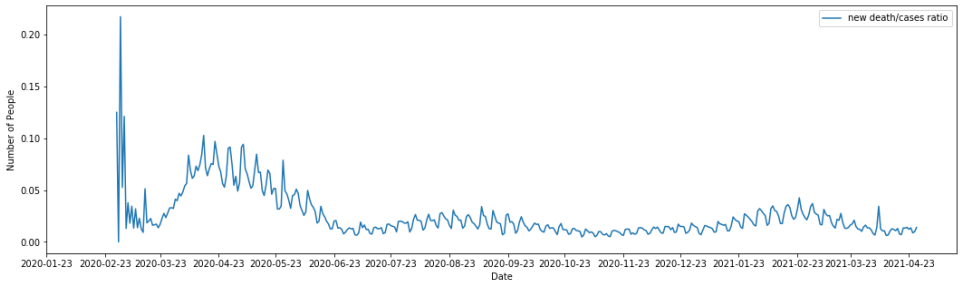


Figure 2. Date vs. the Ratio of new_deaths/new_cases

In the figure above, we see that the death ratio was as high as 20% in the beginning of the outbreak. This might be due to that we did not employ a good amount of testing in the beginning, so a large number of cases we knew were severe cases. Moreover, hospitals might also not prepare for enough resources to handle a large number of patients in a short time. The ratio was still relatively high until June 2020. At that time, the country was allocating enough resources to treat the patients, and the death ratio remained low since then.

When we were doing the analysis, we found that there were periods of ups and downs of the new cases and new deaths, so we wondered if this was an indication of seasonality. Now we will use some forecasting techniques to fit the model and predict the trend for the future.

Data Analysis: Exponential smoothing & Linear Regression

We first try out Simple Exponential Smoothing to fit the number of new deaths from July 1, 2020 to September 1, 2020, and also predict for the next two months:

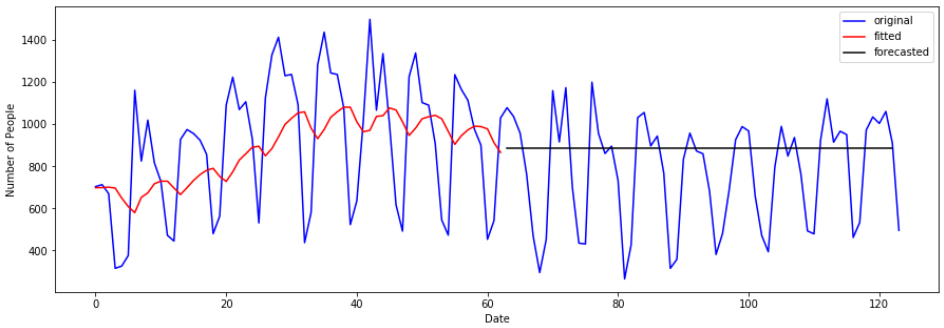


Figure 4. *Simple Exponential Smoothing* of Number of New Deaths from July 1, 2020 to September 1, 2020

The residual sum of squares is 61375991. We see that this fitting is not quite good and the prediction is also not reflective of the trend.

Next, we try out Holt-Winters Exponential Smoothing, we use an additive seasonality, and a seasonal period = 7 days (a week). The following figure is what we get.

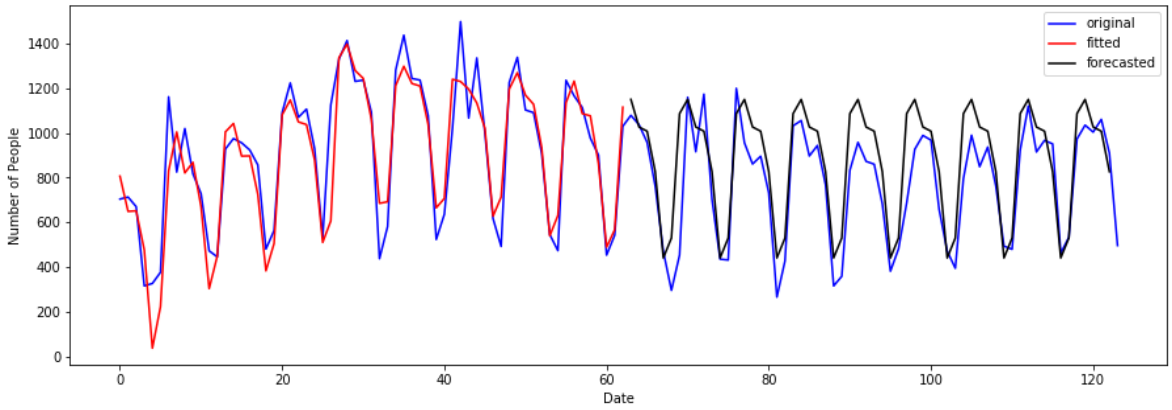


Figure 5. *Holt-Winters Exponential Smoothing* of Number of New Deaths from July 1, 2020 to September 1, 2020

The residual sum of squares now is 1124390, which is a lot better than before. The fitting is also relatively good, and the prediction is close to the actual trend. We are confident to say that we achieved a nice model for fitting. However, one feature we see for this model is that the prediction we make exhibits a seasonal pattern with constant values. That is to say, the predictions for each period are the same in Figure 5. This is not a problem for the period from July 1, 2020 to September 1, 2020 as we see in Figure 1. However, if we look at the time after January 1, 2021, we see that there is a downward trend for new deaths and new cases. To reflect this feature, we add an argument, additive trend, into the model. We get

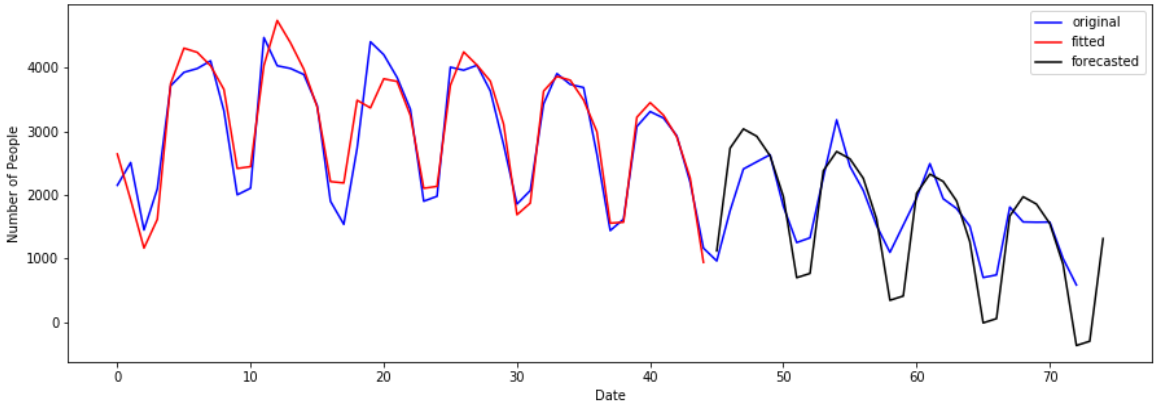


Figure 6. Holt-Winters Exponential Smoothing to *Predict* the Number of New Deaths *After Jan 1, 2021*

We see from Figure 6 that our model takes into account the downward trend of new deaths and makes fairly good predictions as well. Now, as we predict for the trend after April 23, we will use the trend argument as well.

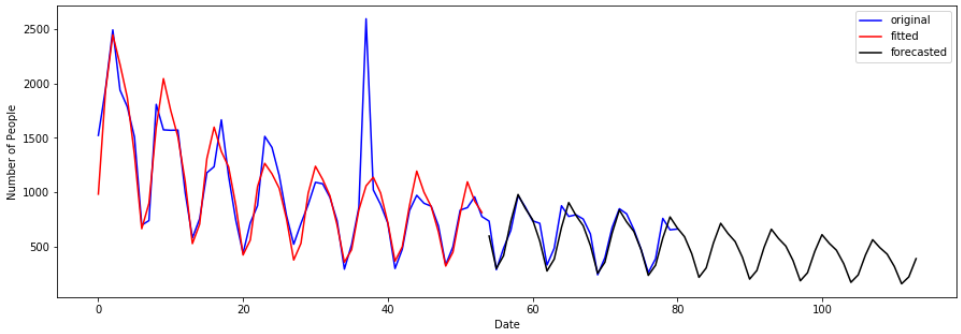


Figure 7. Holt-Winters Exponential Smoothing to *Predict* the Number of New Deaths *After April 23, 2021*

In Figure 7, we used the same modeling method for data from March 1, 2021 to April 23, 2021, to train the model and predict the number of new deaths in May and June, 2021.

We do the same thing to predict the number of new cases and new vaccinations.

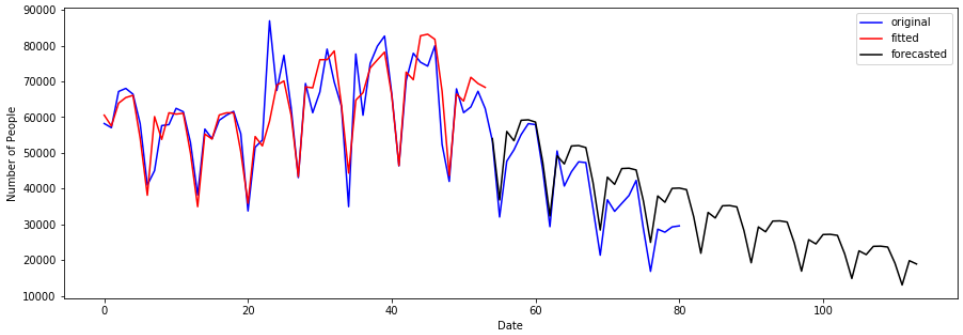


Figure 8. Holt-Winters Exponential Smoothing to *Predict* the Number of *New Cases* After April 23, 2021

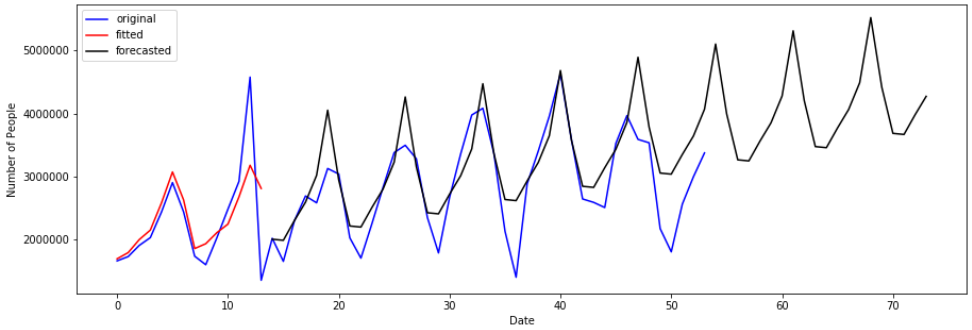


Figure 9.Holt-Winters Exponential Smoothing to *Predict* the Number of *New Vaccinations* After April 23, 2021

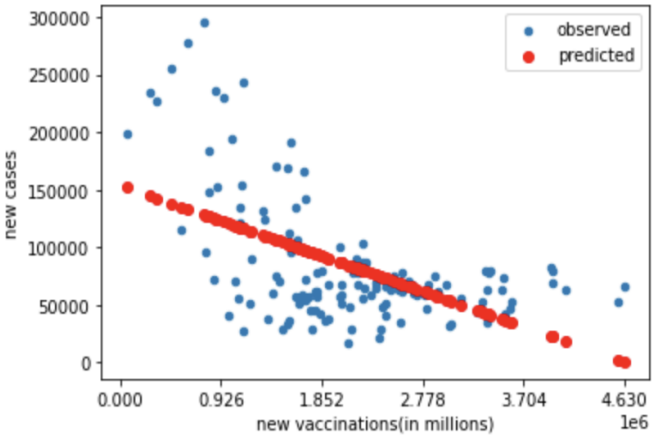
Our modeling to fit the current trend and predict the future trend is important because it helps the hospitals to prepare for the incoming patients, and the governments to implement policies accordingly.

Lastly, we decided to use linear regression to analyze the correlation between vaccination and infection rate, as we mentioned

	coef	std err	t	P>|t|	[0.025	0.975]
const	6.746e+04	3333.783	20.235	0.000	6.09e+04	7.4e+04
new_vaccinations	0.0014	0.003	0.498	0.618	-0.004	0.007

earlier. With our first attempt, it seems that the vaccination and infection rate have a positive relationship, which is not what we expected.

We think this result may be affected by the data points when the vaccine was not yet available to the public. Therefore, we improved our model by deleting these data points.



OLS Regression Results

Dep. Variable:	new_cases	**R-squared:**	0.291
Model:	OLS	**Adj. R-squared:**	0.285
Method:	Least Squares	**F-statistic:**	51.30
Date:	Thu, 20 May 2021	**Prob (F-statistic):**	6.01e-11
Time:	22:47:37	**Log-Likelihood:**	-1549.6
No. Observations:	127	**AIC:**	3103.
Df Residuals:	125	**BIC:**	3109.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	1.537e+05	1.05e+04	14.583	0.000	1.33e+05	1.75e+05
new_vaccinations	-0.0331	0.005	-7.162	0.000	-0.042	-0.024

Omnibus:	18.407	**Durbin-Watson:**	0.306
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	21.663
Skew:	0.896	**Prob(JB):**	1.98e-05
Kurtosis:	3.940	**Cond. No.**	5.59e+06

The result we got from the new model seems much more reasonable. From the result, we can see that there is a negative correlation between new vaccinations and new infection cases, and a p-value < .05. With more people getting vaccinated, the growth rate of the number of new infections will slow down.

Part II. COVID Spread in the New York Area

As students of Cornell University, located at Tompkins, New York, our team paid special attention to the COVID-19 infection within the New York state area. Ranked fourth in the state population across the country, New York has been the one of the epicenters of the COVID-19 pandemic since the outbreak. We are interested in exploring the New York area in more detail.

Visualizations

According to the latest update by Centers for Disease Control and Prevention (CDC), the total number of people with positive molecular tests in New York state by April 26, 8 pm, 2021, is 2,031,095. To better understand the COVID spread within the state, we examined the cumulative positive cases by each county.

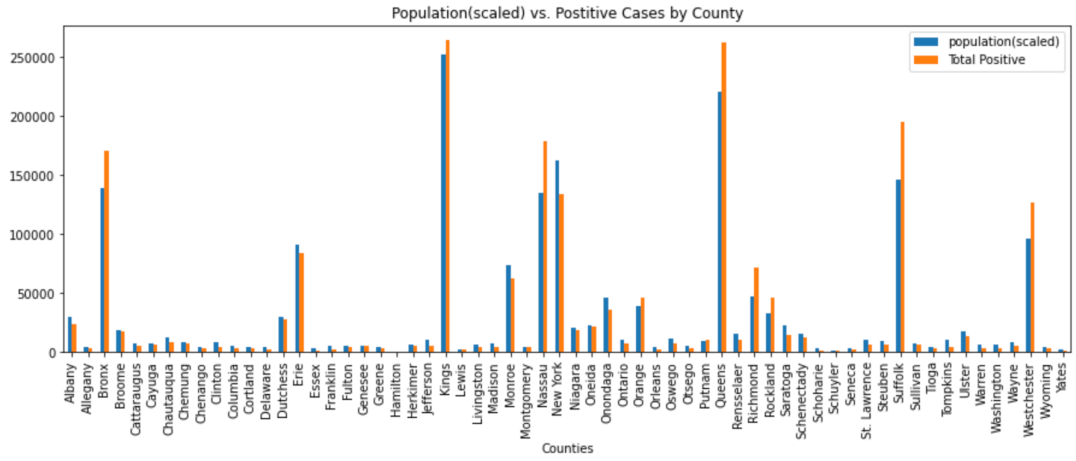


Figure 10. Number of Positive Cases vs. Population by County
(Scale has been adjusted to show the correspondence of two datasets: size factor: 1/10)

We approached the question by making a plot with each county's population (census 2020) and reported positive cases. According to the graph, a strong correlation between the population of the county and the positive cases can be discovered. Counties with larger populations, such as Kings, Queens, and Suffolk, have significantly more number of positive cases than other counties. However, as cautious researchers, we realized that no conclusion can be drawn without further investigation. To determine if the population has a statistically significant effect on the number of cases, we employ the technique of linear regression.

Data Analysis: Linear Regression

To the right, we see the linear regression model showing that population and number of positive cases have a negative relationship and with a very large p-value. It indicates that this model is doing an unsatisfactory job, and the assumption we made was not statistically correct.

	coef	std err	t	P>|t|	[0.025	0.975]
const	3.343e+04	9163.079	3.648	0.001	1.51e+04	5.18e+04
population	-0.0034	0.015	-0.226	0.822	-0.033	0.026

Dep. Variable:	cases	**R-squared:**	0.132
Model:	OLS	**Adj. R-squared:**	0.103
Method:	Least Squares	**F-statistic:**	4.498
Date:	Fri, 21 May 2021	**Prob (F-statistic):**	0.0152
Time:	11:13:54	**Log-Likelihood:**	-766.99
No. Observations:	62	**AIC:**	1540.
Df Residuals:	59	**BIC:**	1546.
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	3.426e+04	8615.574	3.976	0.000	1.7e+04	5.15e+04
population	-0.0265	0.016	-1.656	0.103	-0.058	0.006
population density (per sq mi)	7.0395	2.354	2.990	0.004	2.328	11.751

Omnibus:	54.154	**Durbin-Watson:**	2.122
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	200.076
Skew:	2.714	**Prob(JB):**	3.58e-44
Kurtosis:	9.928	**Cond. No.**	7.14e+05

Therefore, to improve our model, we introduced a new dataset that includes the information of areas of counties and obtains the population density of each, using the equation

$$population\ density\ = \frac{populaiton}{area} * 100\%.$$

By including the new factor, the model returns a reasonable result that the number of confirmed cases in each county is positively correlated with the population density, with a p-value < .05.

Visualizations

In addition to positive cases, our team looked into the number of death cases of each county in New York state.

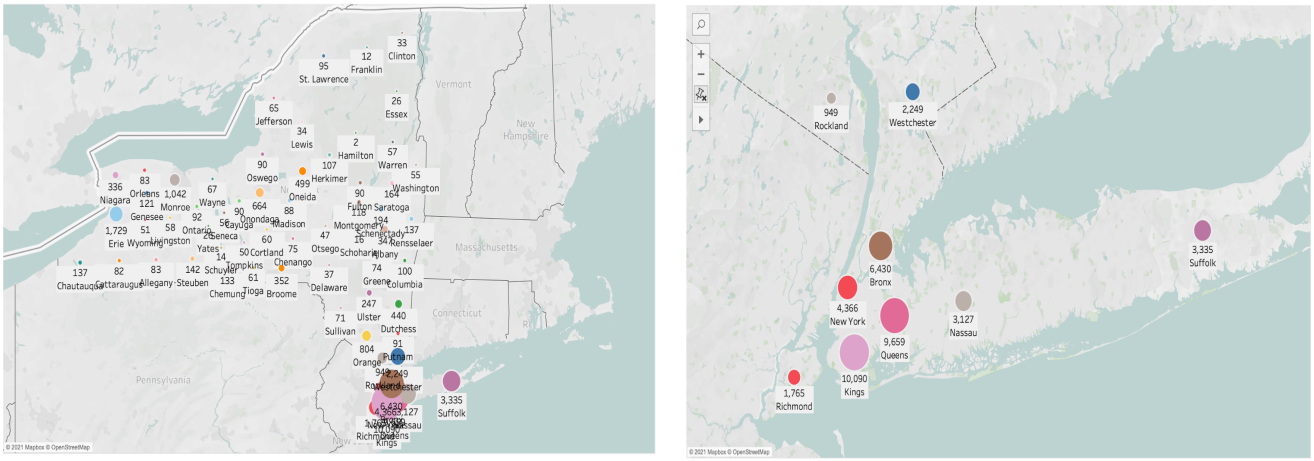


Figure 11. Number of Death cases of New York State Counties & New York City

The figures above show the numbers of death cases of New York State counties (Figure 7 is a zoomed in map of New York City, which is part of figure 6). From the figures, we can see that the five counties that have the most death cases are Kings, Queens, Bronx, New York (County), and Suffolk. Most of these counties are in New York City. This result may be correlated to the population of these five counties as they have the largest populations among all the counties of New York State. Until April 25th, 2021, Kings County has the highest number of death cases, which is 10090. Six counties have more than 5000 death cases while most of the counties have less than 1000 death cases.

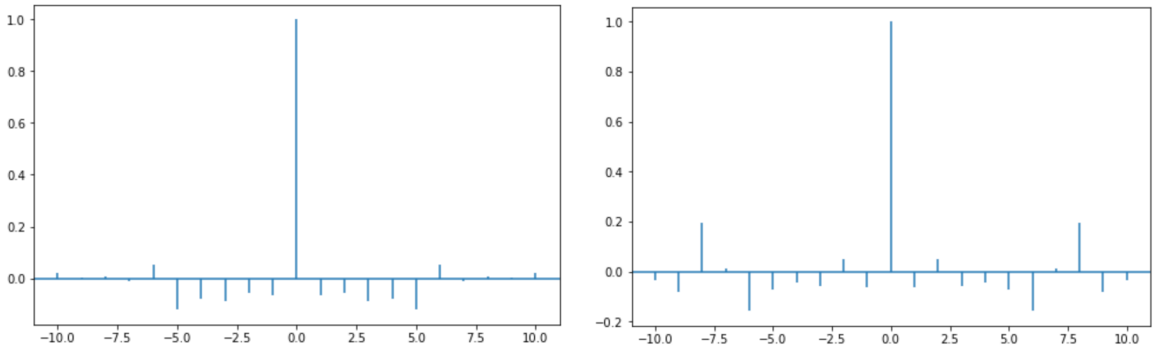
Data Analysis: Linear Regression & AIC

Dep. Variable:	4/25/21	**R-squared:**	0.909
Model:	OLS	**Adj. R-squared:**	0.904
Method:	Least Squares	**F-statistic:**	192.1
Date:	Fri, 21 May 2021	**Prob (F-statistic):**	4.38e-30
Time:	12:38:24	**Log-Likelihood:**	-485.16
No. Observations:	62	**AIC:**	978.3
Df Residuals:	58	**BIC:**	986.8
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-167.9079	103.854	-1.617	0.111	-375.795	39.979
population	-6.778e-05	0.000	-0.387	0.700	-0.000	0.000
density	0.0607	0.027	2.244	0.029	0.007	0.115
positive	0.0299	0.001	21.459	0.000	0.027	0.033

To analyze the impact of population and population density of each county's COVID death cases, we took a similar approach as before, with additional covariate of number of positive cases.

By looking at the summary of the model, we noticed that the covariate of population density now has a p-value less than 0.05. It also shows that the number of death cases is strongly correlated with the number of positive cases with coefficient 0.0299. The result makes intuitive sense, but our team went one step further to test our assumptions.



Number of Positive Cases Number of Death Cases

Figure 12. Akaike information criterion (AIC)

Assuming the acceptance level is 0.2, we can conclude that the residuals of the model are mutually independent of each other, making our previous linear model of strong significance.

Conclusion

Question 1: We generated data visualizations to see the daily trend of COVID cases and compared the situations in different time periods. From our visualizations, we found that the number of new COVID cases and death cases in the United States increased relatively faster between November, 2020 to January, 2021 compared to other time periods. The death ratio was high at the beginning of the disease outbreak but became lower after May, 2020.

Question 2: We used linear regression to see how vaccinations affect the infection rate. Our model shows that a negative correlation appears between the growth of vaccination data and new infected cases data.

Question 3: We used Exponential Smoothing, especially Holt's method, to predict the number of new deaths, new infected cases, and new vaccinations of the future two months. We first tried to use a Simple Exponential Smoothing which did not work well. We then took seasonality and the downward trend of new deaths and new cases into account and got a model that makes fairly good predictions.

Question 4: We first used visualizations to roughly see if populations of the counties seem to be related to the number of positive cases. Then we decided to use linear regression to further investigate the problem. The first model we've used gave us an incorrect result by showing that the population and number of positive cases have a negative relationship. After improving our model by adding population density, we got a reasonable result.

Question 5: For the first part of this question, we used tableau to visualize the number of death cases in each county of New York State. From the figure, we found the five counties that have the highest number of positive cases and death cases are: Kings, Queens, Bronx, New York (County), and Suffolk. For the second part of this question, we used linear regression and AIC to analyze the relationship between population, death cases, and positive cases. The result shows that the number of death cases is strongly related to the number of positive cases and they are mutually independen

Appendix

In [1]:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from datetime import datetime, timedelta
from statsmodels.tsa.holtwinters import Holt, ExponentialSmoothing, SimpleExpSmoothing
```

In [ ]:

```
dff=pd.read_csv('owid-covid-data.csv')
dff_us=dff[dff['location']=='United States']
```

In [ ]:

```
def plot_data(x):
    plt.figure(figsize=(18,5))
    plt.plot(dff_us['date'],dff_us[x],label=x)
    month_starts = [1,32,61,92,122,153,183,214,245,275,306,336,336+31,
                    336+31+31,336+31+31+28,336+31+31+28+31]
    plt.xticks(month_starts)
    plt.xlabel('Date')
    plt.ylabel('Number of People')
    plt.legend()
    plt.show()
```

In [ ]:

```
# Figure 1. Date vs. Number of New Cases, Deaths, Vaccinations in the U.S.

plot_data('new_cases')
plot_data('new_deaths')
plot_data('new_vaccinations')
```

In [ ]:

```
# Figure 2. Date vs. the Ratio of new_deaths/new_cases

dff_us['new death/cases ratio']=dff_us['new_deaths']/dff_us['new_cases']
plot_data('new death/cases ratio')
```

In [ ]:

```
# Figure 3. Number of New Deaths from July 1, 2020 to September 1, 2020

dff_us['date']=pd.to_datetime(dff_us['date'])
dff_us_summer=dff_us[(dff_us['date']>=pd.to_datetime('2020-07-01'))
                     & (dff_us['date']<=pd.to_datetime('2020-09-01'))]

plt.figure(figsize=(18,5))
plt.plot(dff_us_summer['date'],dff_us_summer['new_deaths'],label='new_deaths')
plt.xlabel('Date')
plt.ylabel('Number of People')
plt.legend()
plt.show()
```

In [ ]:

```
# Figure 4. Simple Exponential Smoothing of Number of New Deaths from
# July 1, 2020 to September 1, 2020

dff_us_summer=dff_us[(dff_us['date']>=pd.to_datetime('2020-07-01'))
                     & (dff_us['date']<=pd.to_datetime('2020-09-01'))]
dff_us_summer=dff_us_summer.reset_index()

y=dff_us_summer['new_deaths']
fit1=SimpleExpSmoothing(y).fit()
fore1=fit1.forecast(60)

fitted=fit1.fittedvalues.to_frame()
fitted=fitted.rename(columns={0: "y"})

plt.figure(figsize=(15,5))
plt.plot(df['new_deaths'], label="original",color='blue')
plt.plot(fitted['y'], label="fitted",color='red')
fore1.plot(label='forecasted',color='black')
plt.xlabel('Date')
plt.ylabel('Number of People')
plt.legend()
plt.show()
```

In [113]:

```
fit1.sse
```

Out[113]:

```
61375991.0
```

In [ ]:

```
def holt(y):
    fit=ExponentialSmoothing(y, trend = 'add', seasonal = "add",
                             seasonal_periods = 7).fit()
    fore=fit.forecast(60)

    fitted=fit.fittedvalues.to_frame()
    fitted=fitted.rename(columns={0: "y"})
    fitted=fitted.reset_index()

    plt.figure(figsize=(15,5))
    plt.plot(y, label="original",color='blue')
    plt.plot(fitted['y'], label="fitted",color='red')
    fore.plot(label='forecasted',color='black')
    plt.xlabel('Date')
    plt.ylabel('Number of People')
    plt.legend()
    plt.show()
```

In [ ]:

```
# Figure 5. Holt-Winters Exponential Smoothing of Number of New Deaths
# from July 1, 2020 to September 1, 2020

holt(y)
```

In [ ]:

```
# Figure 6. Holt-Winters Exponential Smoothing to Predict the Number of
# New Deaths After Jan 1, 2021

y3=dff_us[(dff_us['date']>=pd.to_datetime('2021-01-01'))]
y3=y3.reset_index()

holt(y3['new_deaths'])
```

In [ ]:

```
# Figure 7. Holt-Winters Exponential Smoothing to Predict the Number of
# New Deaths After April 23, 2021

y3=dff_us[(dff_us['date']>=pd.to_datetime('2021-03-01'))]
y3=y3.reset_index()

holt(y3['new_deaths'])
```

In [ ]:

```
# Figure 8. Figure 9. Holt-Winters to Predict the Number of New Cases and
# New Vaccinations After April 23, 2021

holt(y3['new_cases'])
holt(y3['new_vaccinations'])
```

In [1]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import statsmodels.api as sm
from sklearn.model_selection import train_test_split, cross_val_score
from datetime import datetime
import seaborn as sns
```

In [2]:
```
NewYork = pd.read_csv('New_York_State_Statewide_COVID-19_Testing.csv')
pop = pd.read_csv('csvData.csv')
```

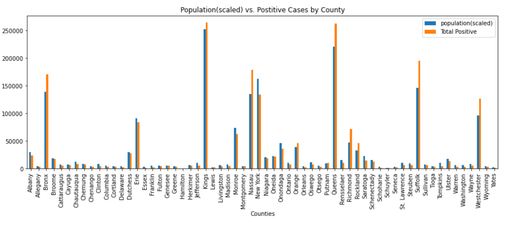
In [3]:
```
pop['CTYNAME'] = pop['CTYNAME'].str.replace( ' County',  '')
pop = pop.sort_values('CTYNAME')
LatestNoSort = NewYork[NewYork['Test Date'] == '04/22/2021']
```

In [4]:
```
plotdata1 = pd.DataFrame({"population(scaled)" : (1/10)*pop['pop2021'].array, \
                          "Total Positive":LatestNoSort['Cumulative Number of Positives'].array }, index = pop['CTYNAME'])
```

In [5]:
```
from pylab import rcParams
rcParams['figure.figsize'] = 15, 5

plotdata1.plot(kind = "bar")
plt.title("Population(scaled) vs. Postitive Cases by County")
plt.xlabel("Counties")
```

Out[5]:
```
Text(0.5, 0, 'Counties')
```



In [6]:
```
plott=pd.concat([pop['pop2021'],LatestNoSort['Cumulative Number of Positives']], ignore_index = True)
```

In [7]:
```
pd.set_option('display.max_rows', None)
pd.set_option('display.max_columns', None)
pd.set_option('display.width', None)
pd.set_option('display.max_colwidth', -1)
LatestNoSort = LatestNoSort.reset_index(drop=True)
```

```
<ipython-input-7-721f5df77400>:4: FutureWarning: Passing a negative integer is deprecated in version 1.0 and will not be supported in future version. Instead, use None to not limit the column width.
  pd.set_option('display.max_colwidth', -1)
```

In [8]:
```
percent = pd.DataFrame()
percent['county'] = LatestNoSort['County']
percent['population']= pop['pop2021']
percent['cases'] = LatestNoSort['Cumulative Number of Positives']
```

In [9]:
```
s = []
for i in range(62):
    s.append(1)
```

In [37]:
```
s = [533, 1034, 57.43, 715, 1310, 864, 1500,410.81, 898.85,1110,648, \
     502,1468,825,1227,1916,1697,533,495,658,1808,1458,1857,96.9,1290,640,662,1366,410,453,33.77,1140,\
     1213,806,662,839,817,1312,1003,246,178.28,665,102.5,199,2821,844,210,626,342,325,1404,2373,997,523,476,1161,870,\
     846,1384,500,596,376]
ss = pd.Series(s)
len(s)
```

Out[37]:
```
62
```

In [11]:
```
percent['area'] =ss
percent['population density (per sq mi)'] = percent['population']/ percent['area']
suffolk = percent[percent['county'] == "Kings"]
percent['test'] = LatestNoSort['Total Number of Tests Performed']
```

In [12]:
```
X = percent[['population','population density (per sq mi)']]
y = percent['cases']
X = sm.add_constant(X)

model = sm.OLS(y, X).fit()
resid = model.resid
model.summary()
```

Out[12]:

OLS Regression Results

Dep. Variable:	cases	**R-squared:**	0.132
Model:	OLS	**Adj. R-squared:**	0.103
Method:	Least Squares	**F-statistic:**	4.498
Date:	Sat, 22 May 2021	**Prob (F-statistic):**	0.0152
Time:	12:19:12	**Log-Likelihood:**	-766.99
No. Observations:	62	**AIC:**	1540.
Df Residuals:	59	**BIC:**	1546.
Df Model:	2		
Covariance Type:	nonrobust		

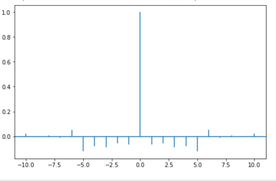
	coef	std err	t	P>|t|	[0.025	0.975]
const	3.426e+04	8615.574	3.976	0.000	1.7e+04	5.15e+04
population	-0.0265	0.016	-1.656	0.103	-0.058	0.006
population density (per sq mi)	7.0395	2.354	2.990	0.004	2.328	11.751

Omnibus:	54.154	**Durbin-Watson:**	2.122
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	200.076
Skew:	2.714	**Prob(JB):**	3.58e-44
Kurtosis:	9.928	**Cond. No.**	7.14e+05

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 7.14e+05. This might indicate that there are
strong multicollinearity or other numerical problems.

In [13]:
```
rcParams['figure.figsize'] = 8, 5
plt.acorr(resid)
```

Out[13]:
```
(array([-10,  -9,  -8,  -7,  -6,  -5,  -4,  -3,  -2,  -1,   0,   1,   2,
         3,   4,   5,   6,   7,   8,   9,  10]),
 array([ 0.01099121,  0.0035195 ,  0.00491244, -0.01109468,  0.0531842 ,
        -0.12216925, -0.07757795, -0.08627005, -0.05651572, -0.06498708,
         1.        , -0.06498708, -0.05651572, -0.08627005, -0.07757795,
        -0.12216925,  0.0531842 , -0.01109468,  0.00491244,  0.0035195 ,
         0.01099121]),
 <matplotlib.collections.LineCollection at 0x7fcd609688b0>,
 <matplotlib.lines.Line2D at 0x7fcd60968880>)
```



In [14]:
```
death = pd.read_csv('time_series_covid19_deaths_US.csv')
death = death[death['Province_State'] == "New York"]
deathNY = death[['Admin2', '4/25/21']]
deathNY = deathNY.reset_index(drop = True)
deathNY = deathNY.drop([39,57])
deathNY = deathNY.reset_index(drop = True)
deathNY ['population']  = pop['pop2021']
deathNY ['density'] = percent['population density (per sq mi)']
deathNY ['positive'] = percent['cases']
```

In [15]:
```
X = deathNY[['population','density','positive']]
X = sm.add_constant(X)
y = deathNY['4/25/21']
model2 = sm.OLS(y,X).fit()
resid = model2.resid
model2.summary()
```

Out[15]:

OLS Regression Results

Dep. Variable:	4/25/21	**R-squared:**	0.909
Model:	OLS	**Adj. R-squared:**	0.904
Method:	Least Squares	**F-statistic:**	192.1
Date:	Sat, 22 May 2021	**Prob (F-statistic):**	4.38e-30
Time:	12:19:13	**Log-Likelihood:**	-485.16
No. Observations:	62	**AIC:**	978.3
Df Residuals:	58	**BIC:**	986.8
Df Model:	3		
Covariance Type:	nonrobust		

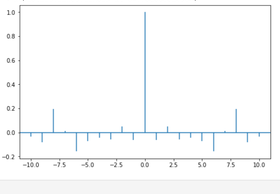
	coef	std err	t	P>|t|	[0.025	0.975]
const	-167.9079	103.854	-1.617	0.111	-375.795	39.979
population	-6.778e-05	0.000	-0.387	0.700	-0.000	0.000
density	0.0607	0.027	2.244	0.029	0.007	0.115
positive	0.0299	0.001	21.459	0.000	0.027	0.033

Omnibus:	22.421	**Durbin-Watson:**	2.119
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	176.675
Skew:	-0.370	**Prob(JB):**	4.32e-39
Kurtosis:	11.237	**Cond. No.**	8.04e+05

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 8.04e+05. This might indicate that there are
strong multicollinearity or other numerical problems.

In [16]:
```
rcParams['figure.figsize'] = 8, 5
plt.acorr(resid)
```

Out[16]:
```
(array([-10,  -9,  -8,  -7,  -6,  -5,  -4,  -3,  -2,  -1,   0,   1,   2,
         3,   4,   5,   6,   7,   8,   9,  10]),
 array([-0.03187777, -0.08236819,  0.19580811,  0.01388873, -0.15798002,
        -0.07199377, -0.04408861, -0.05762396,  0.0506597 , -0.06214473,
         1.        , -0.06214473,  0.0506597 , -0.05762396, -0.04400861,
        -0.07199377, -0.15798002,  0.01388873,  0.19580811, -0.08236819,
        -0.03187777]),
 <matplotlib.collections.LineCollection at 0x7fcd66de14c0>,
 <matplotlib.lines.Line2D at 0x7fcd66de1490>)
```



In [ ]:

In [ ]:

In [14]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import statsmodels.api as sm
from sklearn.model_selection import train_test_split, cross_val_score
from datetime import datetime
import seaborn as sns
from datetime import datetime, timedelta
from statsmodels.tsa.holtwinters import Holt, ExponentialSmoothing, SimpleExpSmoothing
from IPython.core.display import Image
```

In [2]:
```
dff=pd.read_csv('owid-covid-data.csv')
dff.head()
```

Out[2]:

	iso_code	continent	location	date	total_cases	new_cases	new_cases_smoothed	total_deaths	new_deaths	new_deaths_smoothed	...	gdp_per_capita	extreme_poverty	cardiovasc_death_rate	diabete
0	AFG	Asia	Afghanistan	2020-02-24	1.0	1.0	NaN	NaN	NaN	NaN	...	1803.987	NaN	597.029	
1	AFG	Asia	Afghanistan	2020-02-25	1.0	0.0	NaN	NaN	NaN	NaN	...	1803.987	NaN	597.029	
2	AFG	Asia	Afghanistan	2020-02-26	1.0	0.0	NaN	NaN	NaN	NaN	...	1803.987	NaN	597.029	
3	AFG	Asia	Afghanistan	2020-02-27	1.0	0.0	NaN	NaN	NaN	NaN	...	1803.987	NaN	597.029	
4	AFG	Asia	Afghanistan	2020-02-28	1.0	0.0	NaN	NaN	NaN	NaN	...	1803.987	NaN	597.029	

5 rows × 59 columns

In [18]:
```
US = dff[dff['location']=='United States']
US.head()
```

Out[18]:

	iso_code	continent	location	date	total_cases	new_cases	new_cases_smoothed	total_deaths	new_deaths	new_deaths_smoothed	...	gdp_per_capita	extreme_poverty	cardiovasc_death_rate	diabe
85198	USA	North America	United States	2020-01-22	1.0	NaN	NaN	NaN	NaN	NaN	...	54225.446	1.2	151.089	
85199	USA	North America	United States	2020-01-23	1.0	0.0	NaN	NaN	NaN	NaN	...	54225.446	1.2	151.089	
85200	USA	North America	United States	2020-01-24	2.0	1.0	NaN	NaN	NaN	NaN	...	54225.446	1.2	151.089	
85201	USA	North America	United States	2020-01-25	2.0	0.0	NaN	NaN	NaN	NaN	...	54225.446	1.2	151.089	
85202	USA	North America	United States	2020-01-26	5.0	3.0	NaN	NaN	NaN	NaN	...	54225.446	1.2	151.089	

5 rows × 59 columns

In [4]:
```
US = US.replace(np.nan, 0.0)
US.head()
```

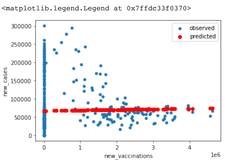
Out[4]:

	iso_code	continent	location	date	total_cases	new_cases	new_cases_smoothed	total_deaths	new_deaths	new_deaths_smoothed	...	gdp_per_capita	extreme_poverty	cardiovasc_death_rate	diabe
85198	USA	North America	United States	2020-01-22	1.0	0.0	0.0	0.0	0.0	0.0	...	54225.446	1.2	151.089	
85199	USA	North America	United States	2020-01-23	1.0	0.0	0.0	0.0	0.0	0.0	...	54225.446	1.2	151.089	
85200	USA	North America	United States	2020-01-24	2.0	1.0	0.0	0.0	0.0	0.0	...	54225.446	1.2	151.089	
85201	USA	North America	United States	2020-01-25	2.0	0.0	0.0	0.0	0.0	0.0	...	54225.446	1.2	151.089	
85202	USA	North America	United States	2020-01-26	5.0	3.0	0.0	0.0	0.0	0.0	...	54225.446	1.2	151.089	

5 rows × 59 columns

In [5]:
```
#Vaccination's impact on number of new cases: Original model
Y = US['new_cases']
X = sm.add_constant(US['new_vaccinations'])
#fit the model
model = sm.OLS(Y, X).fit()
Yhat = model.predict(X)
# plot
p = US.plot.scatter(x='new_vaccinations', y='new_cases', label='observed')
p.scatter(x=US['new_vaccinations'], y=Yhat, color='r', label='predicted')
plt.legend()
```

Out[5]:
```
<matplotlib.legend.Legend at 0x7ffdc33f0370>
```



In [6]:
```
model.summary()
```

Out[6]:

OLS Regression Results

Dep. Variable:	new_cases	**R-squared:**	0.001
Model:	OLS	**Adj. R-squared:**	-0.002
Method:	Least Squares	**F-statistic:**	0.2484
Date:	Sat, 22 May 2021	**Prob (F-statistic):**	0.618
Time:	09:17:48	**Log-Likelihood:**	-6048.3
No. Observations:	484	**AIC:**	1.210e+04
Df Residuals:	482	**BIC:**	1.211e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	6.746e+04	3333.783	20.235	0.000	6.09e+04	7.4e+04
new_vaccinations	0.0014	0.003	0.498	0.618	-0.004	0.007

Omnibus:	115.662	**Durbin-Watson:**	0.076
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	201.425
Skew:	1.446	**Prob(JB):**	1.82e-44
Kurtosis:	4.275	**Cond. No.**	1.32e+06

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.32e+06. This might indicate that there are
strong multicollinearity or other numerical problems.

In [7]:
```
#Improved model
US1 = US[US.new_vaccinations != 0.0]
US1.head()
```

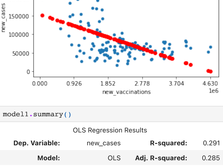
Out[7]:

	iso_code	continent	location	date	total_cases	new_cases	new_cases_smoothed	total_deaths	new_deaths	new_deaths_smoothed	...	gdp_per_capita	extreme_poverty	cardiovasc_death_rate	diabe
85532	USA	North America	United States	2020-12-21	18153724.0	199049.0	218024.857	324844.0	1920.0	2791.857	...	54225.446	1.2	151.089	
85547	USA	North America	United States	2021-01-05	21181440.0	235111.0	221632.571	364076.0	3715.0	2732.429	...	54225.446	1.2	151.089	
85548	USA	North America	United States	2021-01-06	21436884.0	255444.0	224741.143	368003.0	3927.0	2760.143	...	54225.446	1.2	151.089	
85549	USA	North America	United States	2021-01-07	21715174.0	278290.0	230830.143	371990.0	3987.0	2832.571	...	54225.446	1.2	151.089	
85550	USA	North America	United States	2021-01-08	22010389.0	295215.0	251056.857	376098.0	4108.0	3112.286	...	54225.446	1.2	151.089	

5 rows × 59 columns

In [11]:
```
Y = US1['new_cases']
X = sm.add_constant(US1['new_vaccinations'])
#fit the model
model1 = sm.OLS(Y, X).fit()
Yhat = model1.predict(X)
# plot
p = US1.plot.scatter(x='new_vaccinations', y='new_cases', label='observed')
p.scatter(x=US1['new_vaccinations'], y=Yhat, color='r', label='predicted')
tick =  [0,max(US1["new_vaccinations"])/5,
         (max(US1["new_vaccinations"])/5)*2,(max(US1["new_vaccinations"])/5)*3,
         (max(US1["new_vaccinations"])/5)*4,max(US1["new_vaccinations"])]
plt.xticks(tick)
plt.legend()
```

Out[11]:
```
<matplotlib.legend.Legend at 0x7ffda6ed47c0>
```



In [12]:
```
model1.summary()
```

Out[12]:

OLS Regression Results

Dep. Variable:	new_cases	**R-squared:**	0.291
Model:	OLS	**Adj. R-squared:**	0.285
Method:	Least Squares	**F-statistic:**	51.30
Date:	Sat, 22 May 2021	**Prob (F-statistic):**	6.01e-11
Time:	09:18:14	**Log-Likelihood:**	-1549.6
No. Observations:	127	**AIC:**	3103.
Df Residuals:	125	**BIC:**	3109.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.537e+05	1.05e+04	14.583	0.000	1.33e+05	1.75e+05
new_vaccinations	-0.0331	0.005	-7.162	0.000	-0.042	-0.024

Omnibus:	18.407	**Durbin-Watson:**	0.306
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	21.663
Skew:	0.896	**Prob(JB):**	1.98e-05
Kurtosis:	3.940	**Cond. No.**	5.59e+06

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 5.59e+06. This might indicate that there are
strong multicollinearity or other numerical problems.

In [17]:
```
#Figure 12 was made with Tableau
Image(filename = 'figure12.png')
```

Out[17]: