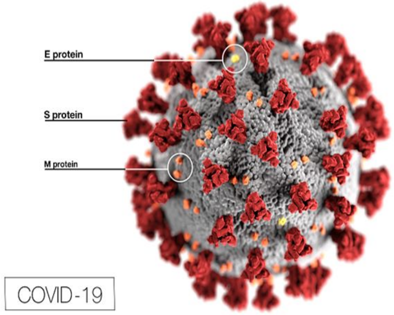
Covid-19 Data Analysis



Vandana Jain

This project predicts the number of Covid-19 cases within US Based on Dataset using :

- Support vector regression(SVR)
- Random forest (RF)

Note : Using SVR and RF as my data is not running linear

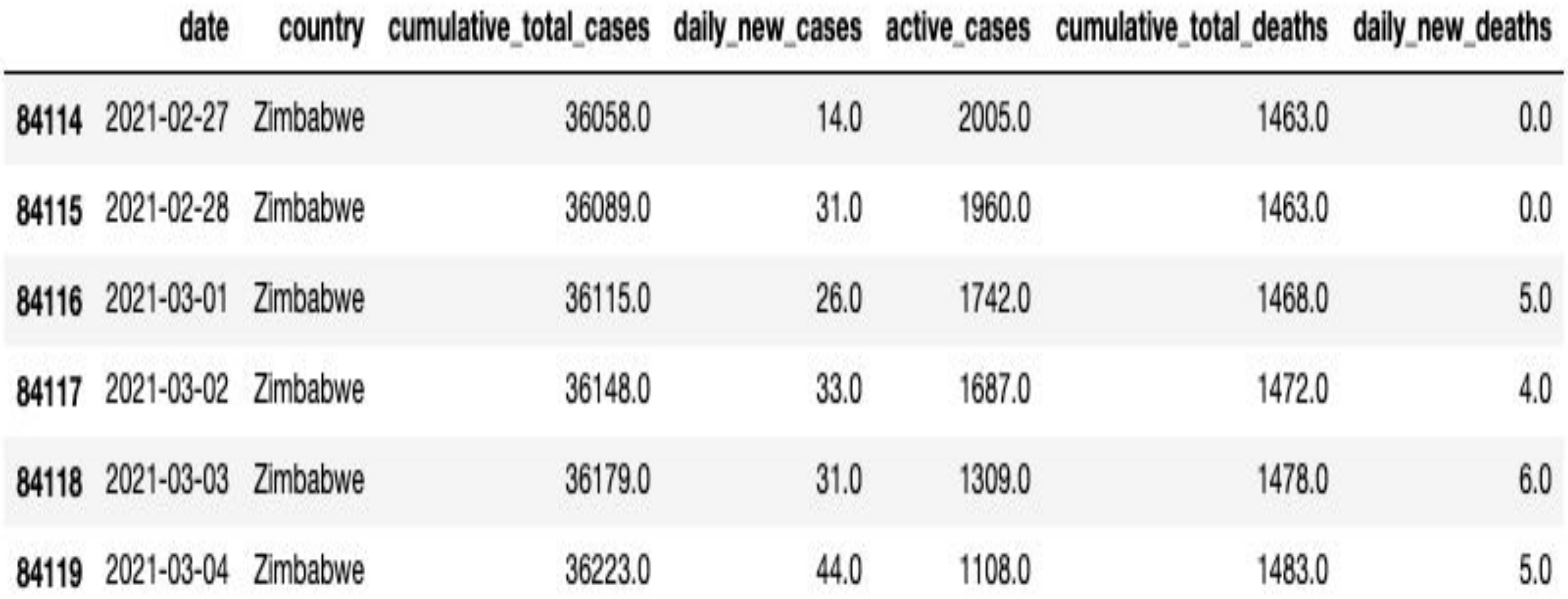
Dataset :

Data : worldometer_coronavirus_daily_data.csv

Source :Kaggle

Source Link

Dataset overview :



	date	country	cumulative_total_cases	daily_new_cases	active_cases	cumulative_total_deaths	daily_new_deaths
84114	2021-02-27	Zimbabwe	36058.0	14.0	2005.0	1463.0	0.0
84115	2021-02-28	Zimbabwe	36089.0	31.0	1960.0	1463.0	0.0
84116	2021-03-01	Zimbabwe	36115.0	26.0	1742.0	1468.0	5.0
84117	2021-03-02	Zimbabwe	36148.0	33.0	1687.0	1472.0	4.0
84118	2021-03-03	Zimbabwe	36179.0	31.0	1309.0	1478.0	6.0
84119	2021-03-04	Zimbabwe	36223.0	44.0	1108.0	1483.0	5.0

Df.shape ： (84120, 7)

Checking for Nulls:

```
df.isnull().sum()
```

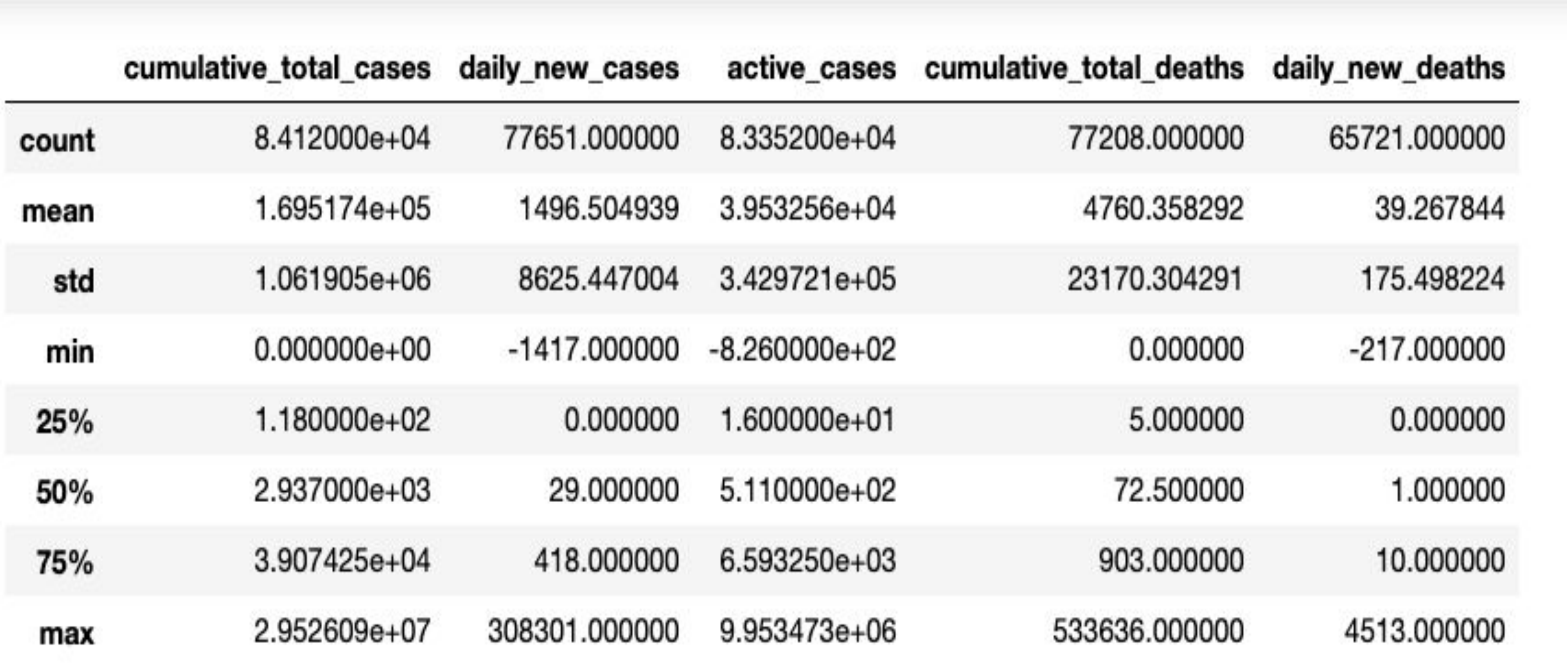
```
date                      0
country                   0
cumulative_total_cases    0
daily_new_cases        6469
active_cases            768
cumulative_total_deaths 6912
daily_new_deaths      18399
mnth_yr                   0
month                     0
year                      0
```

```
round(df.isnull().sum(axis=0).sort_values
(ascending=False)/len(df)*100,0)
```

```
daily_new_deaths         22.0
cumulative_total_deaths   8.0
daily_new_cases           8.0
active_cases              1.0
year                      0.0
month                     0.0
mnth_yr                   0.0
cumulative_total_cases    0.0
country                   0.0
date                      0.0
```

```
df.dropna()
```

Description statistics:



	cumulative_total_cases	daily_new_cases	active_cases	cumulative_total_deaths	daily_new_deaths
count	8.412000e+04	77651.000000	8.335200e+04	77208.000000	65721.000000
mean	1.695174e+05	1496.504939	3.953256e+04	4760.358292	39.267844
std	1.061905e+06	8625.447004	3.429721e+05	23170.304291	175.498224
min	0.000000e+00	-1417.000000	-8.260000e+02	0.000000	-217.000000
25%	1.180000e+02	0.000000	1.600000e+01	5.000000	0.000000
50%	2.937000e+03	29.000000	5.110000e+02	72.500000	1.000000
75%	3.907425e+04	418.000000	6.593250e+03	903.000000	10.000000
max	2.952609e+07	308301.000000	9.953473e+06	533636.000000	4513.000000

Checking total cumulative cases in US (Verified with CDC Website)

```
print("basic info")

print("total no of cases",US['cumulative_total_cases'].iloc[-1])

print("total no of deaths",US['cumulative_total_deaths'].iloc[-1])
```

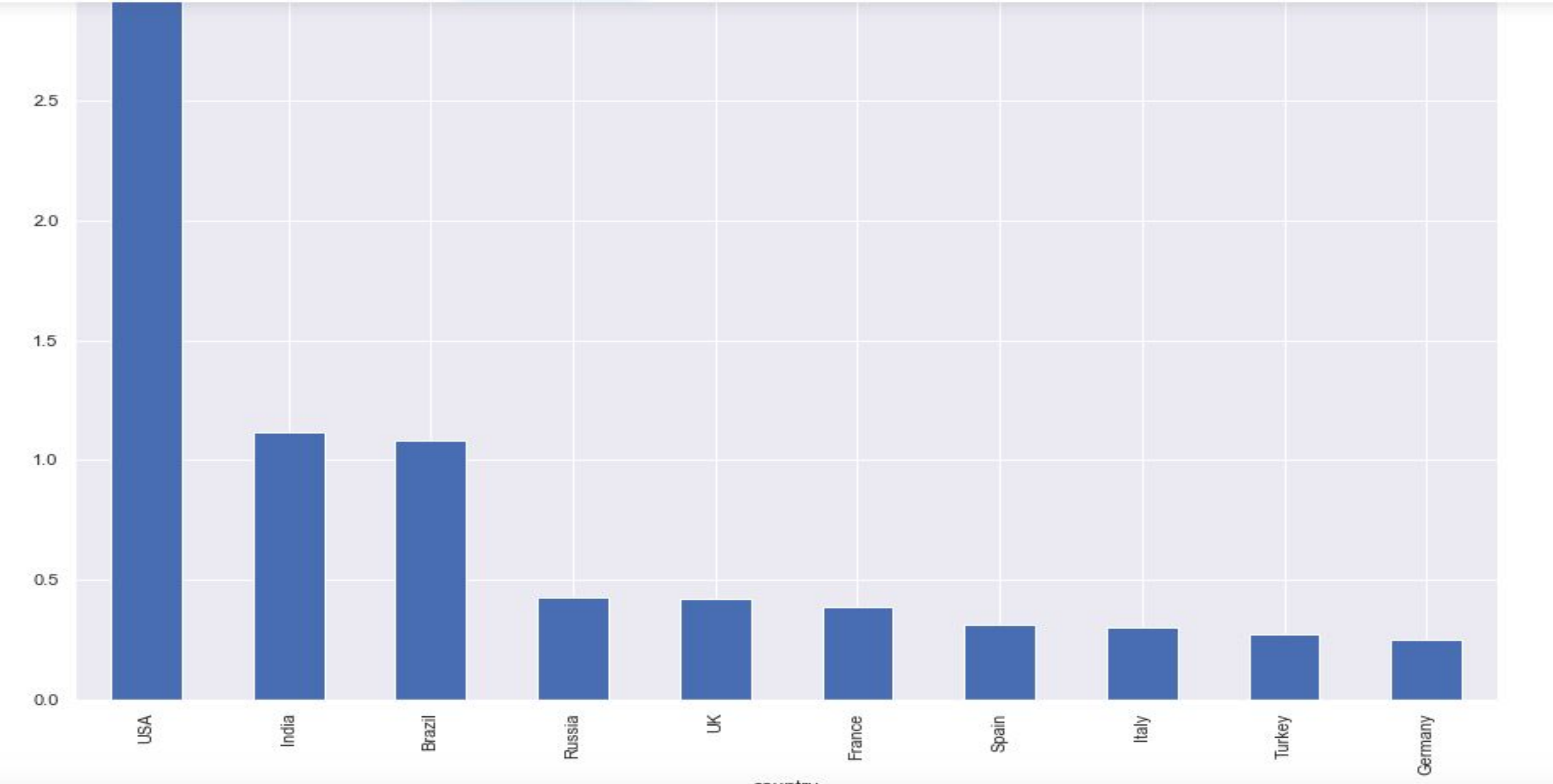
```
basic info

total no of cases 29526086.0

total no of deaths 533636.0
```

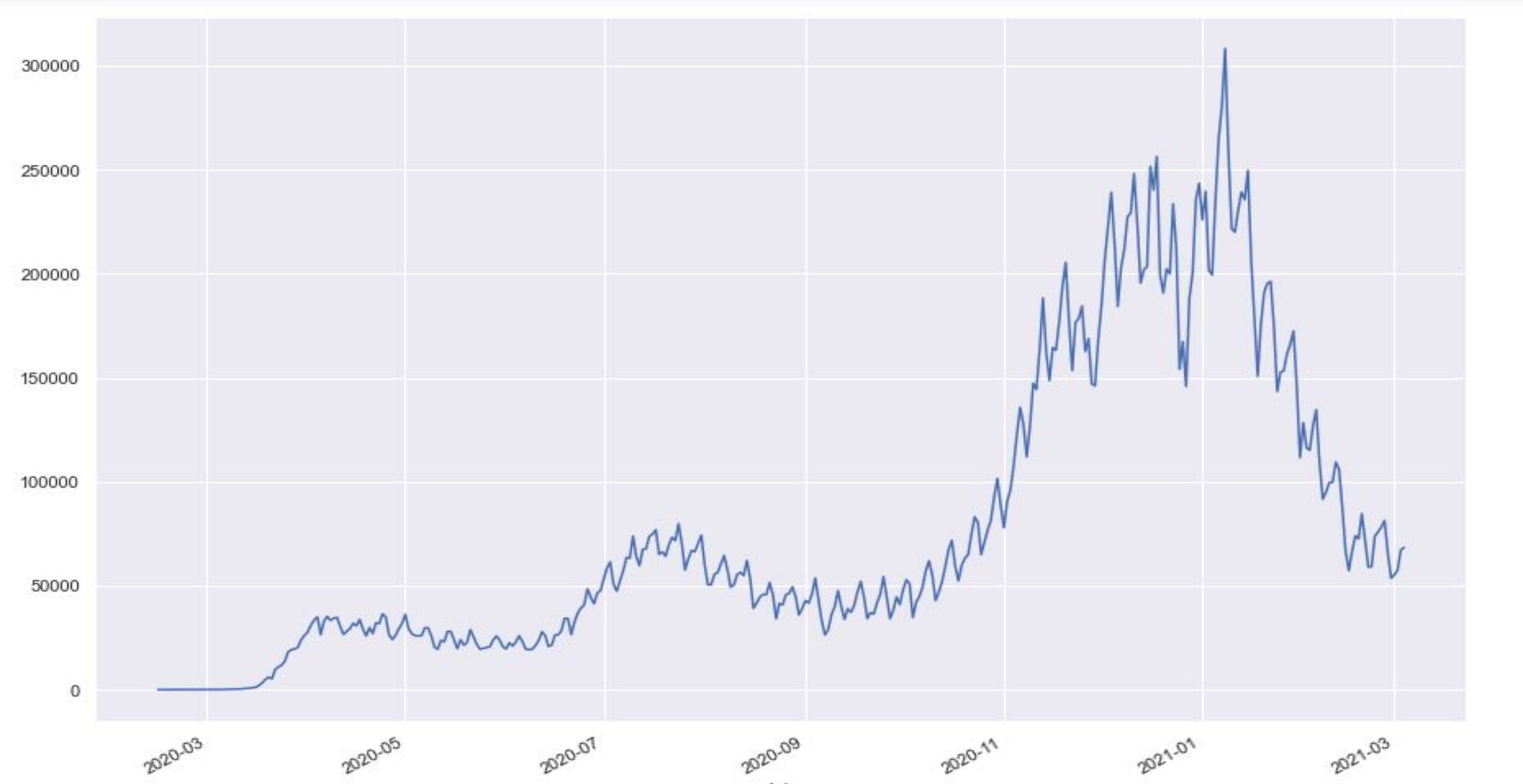
Countrywide daily cases :

```
d1=df.groupby('country')['daily_new_cases'].sum().sort_values(ascending=False).head(10)
```



Covid daily trend in US:

dd=df[**df.country=="USA"**].groupby('date')['daily_new_cases'].sum().sort_values()



Predictive Modelling:

```
x=US['date']

y=US['daily_new_cases']

x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.3)
```

Train Data=70% **Test Data= 30%**

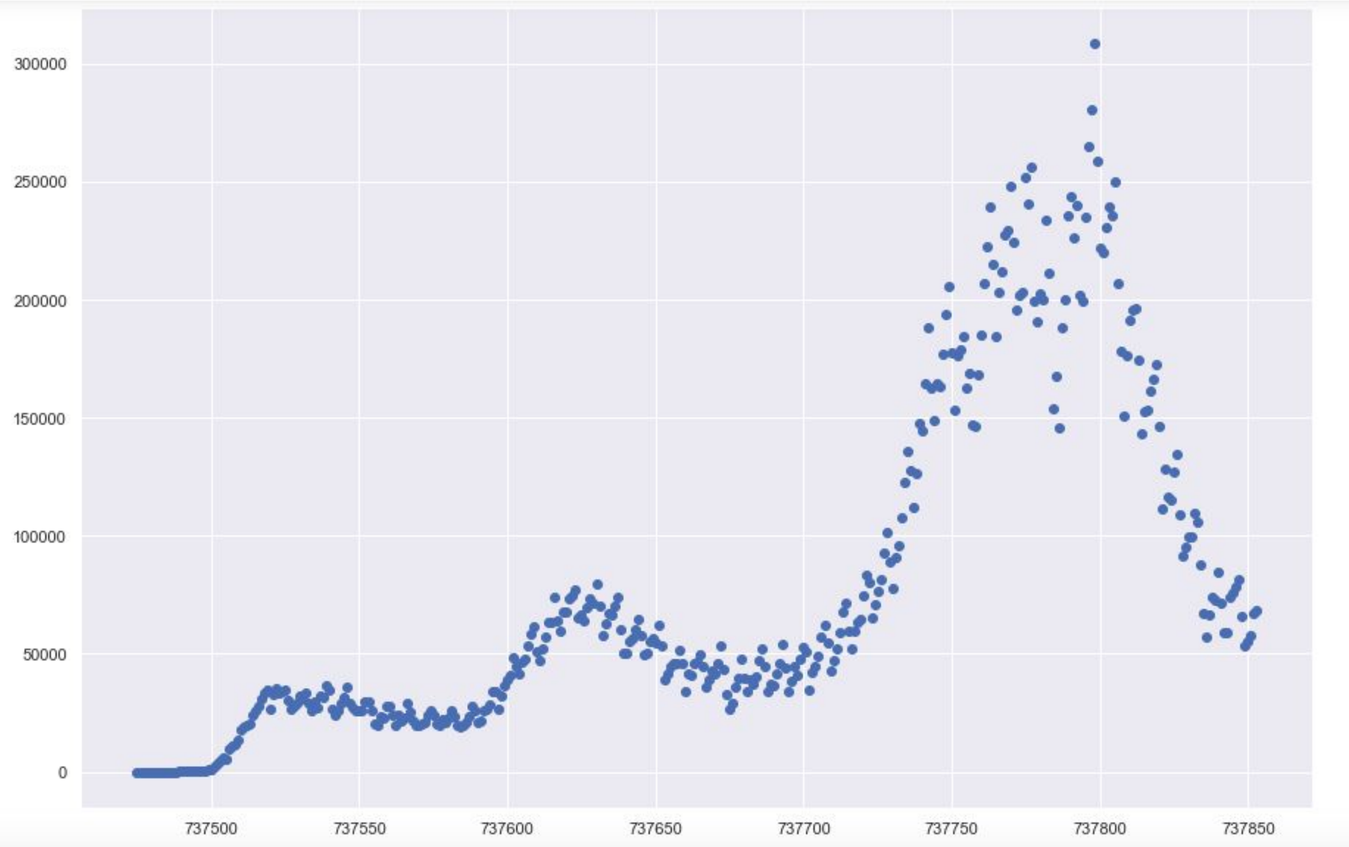
```
lr=LinearRegression()

lr.fit(np.array(x_train).reshape(-1,1),np.array(y_train).reshape(-1,1))
```

Continued :

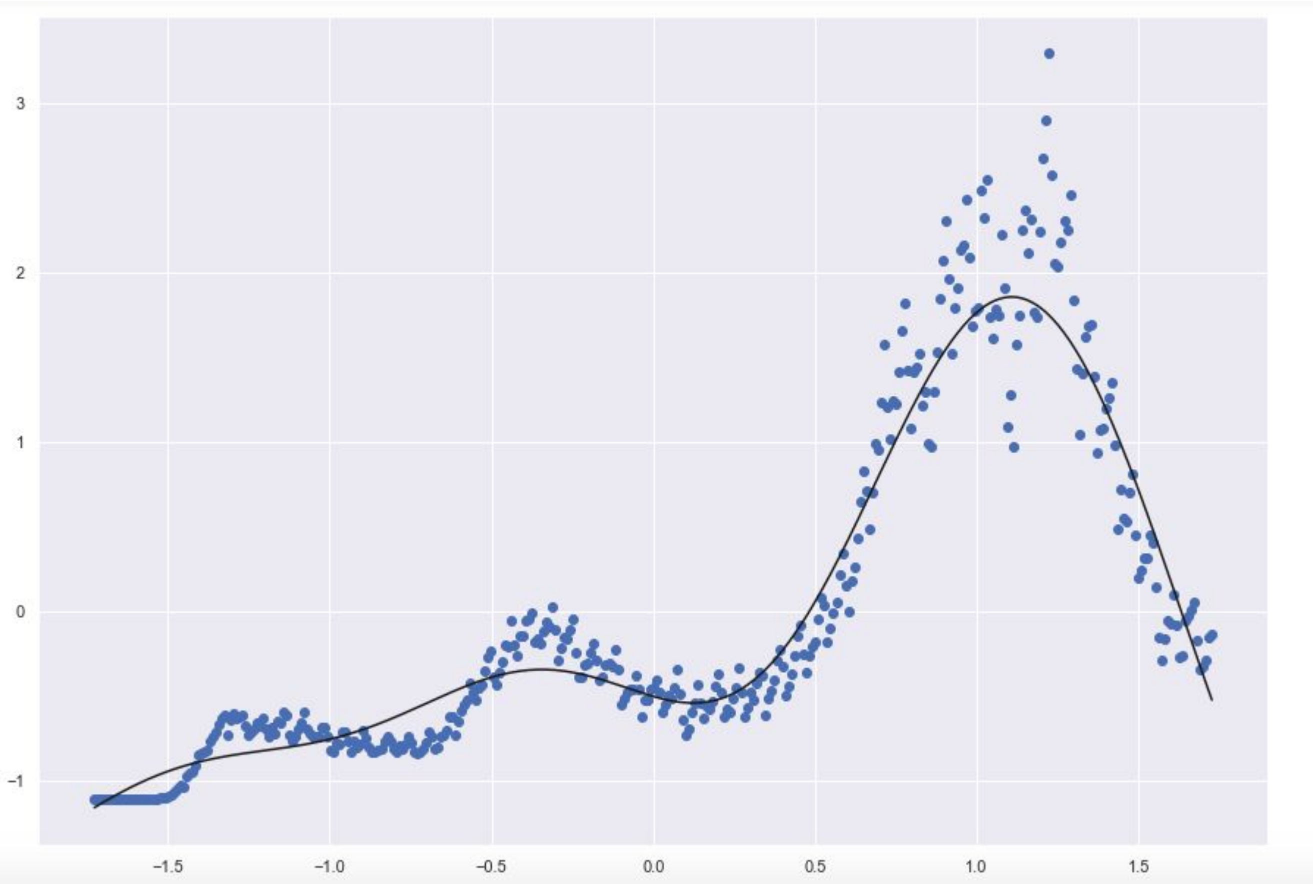
```
plt.scatter(x,y)

plt.show()
```



SVM (Using kernel :rbf(radial bias function)):

```
from sklearn.preprocessing i
from sklearn.svm import SVR
sc_x=StandardScaler()
sc_y=StandardScaler()
sx=sc_x.fit_transform(x)
sy=sc_y.fit_transform(y)
from sklearn.svm import SVR
reg=SVR(kernel='rbf')
reg.fit(sx,sy)
reg.score(sx,sy)
0.930073567204027
```



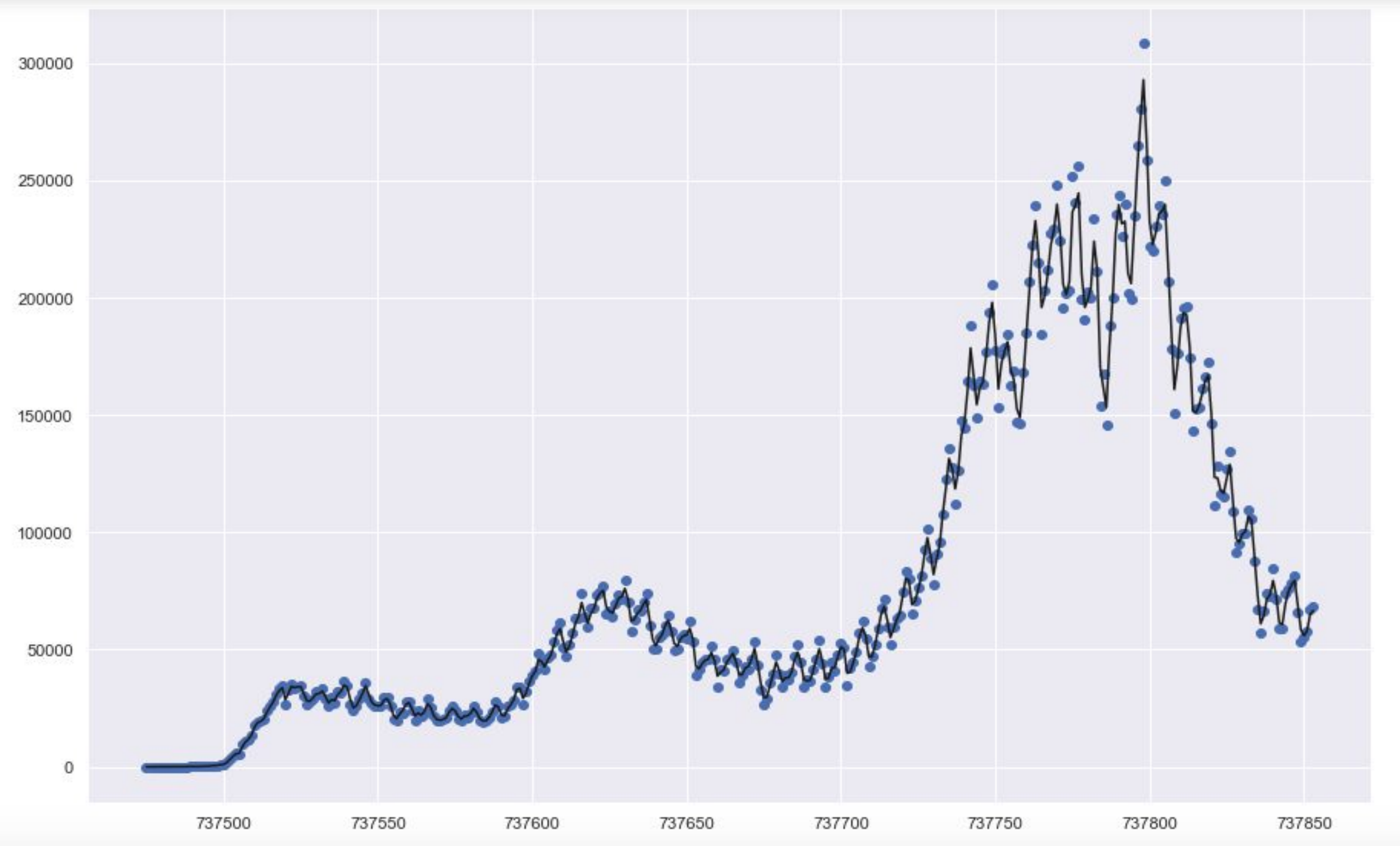
Random Forest:

regf=RandomForest

regf.fit(x,y)

regf.score(x,y)

0.997528468985273



Comparison:

Model :	Score
SVR	0.930073567204027
RF	**0.997528468985273**

Winner : RF

Thankyou