EAST WEST UNIVERSITY
Department of Computer Science and Engineering
B.Sc. in Computer Science and Engineering Program

Project Report

Project Title: ***Linear Regression on COVID-19*: *Global Forecasting (Week-1)***

Course Code: CSE475

Course Title: Machine Learning

Semester: Spring2020

Submitted To:

Dr. Md. Golam Rabiul Alam

Lecture,

Department of Computer Science and Engineering

Submitted By:

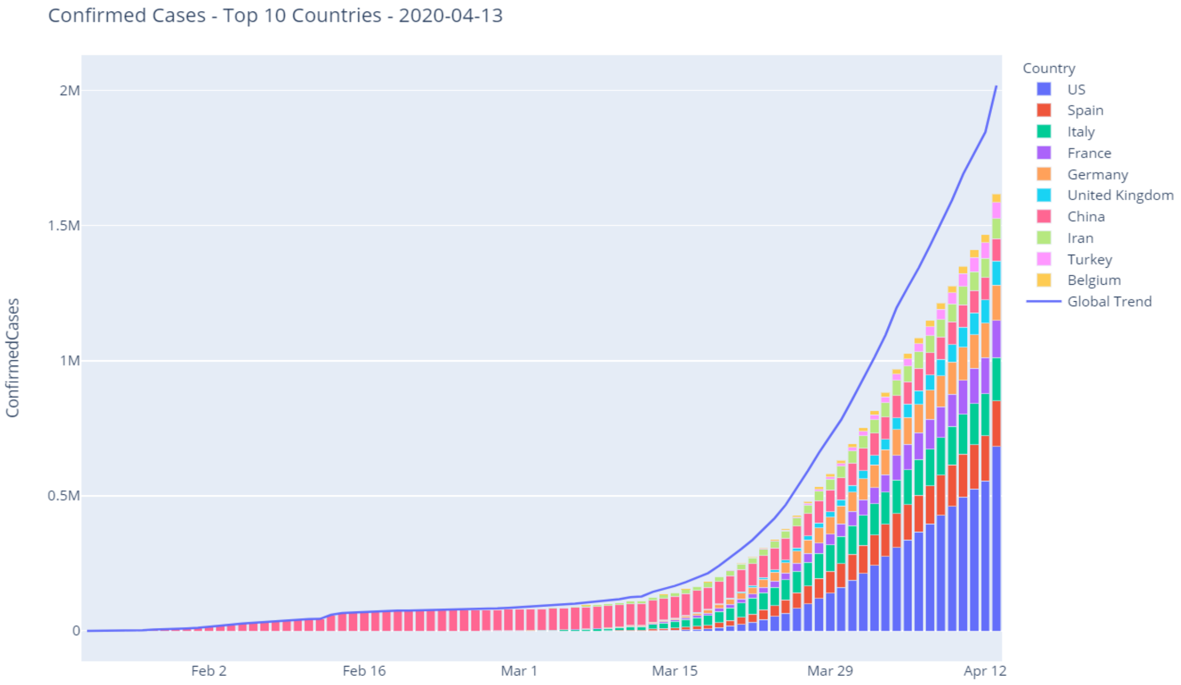
Md. Shanewaz Akib

ID: 2016-3-60-050

Submission Deadline: 19 May 2020

Introduction

COVID-19 is a strain of coronavirus that first broke out in Wuhan, China in December 2019 and has since become a global pandemic. As of 19 May 2020, more than 4.8 million cases have been reported across 188 countries and territories, resulting in more than 318,000 deaths. More than 1.78 million people have recovered.



Background

The White House Office of Science and Technology Policy (OSTP) pulled together a coalition research groups and companies to prepare the COVID-19 Open Research Dataset (CORD-19) to attempt to address key open scientific questions on COVID-19. Those questions are drawn from National Academies of Sciences, Engineering, and Medicine's (NASEM) and the World Health Organization (WHO).

Challenge

Kaggle is launching two companion COVID-19 forecasting challenges to help answer a subset of the NASEM/WHO questions. While the challenge involves forecasting confirmed cases and fatalities between March 25 and April 22 by region, the primary goal **isn't to produce accurate forecasts**. It's to identify factors that appear to impact the transmission rate of COVID-19.

As the data becomes available, we will update the leaderboard with live results based on date made available from the Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE).

We have received support and guidance from health and policy organizations in launching these challenges. We're hopeful can make valuable contributions to developing a better understanding of factors that impact the transmission of COVID-19.

Evaluation

Submissions are evaluated using the column-wise root mean squared logarithmic error.

The RMSLE for a single column calculated as

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\log(p_i+1)-\log(a_i+1))^2},$$

Where,

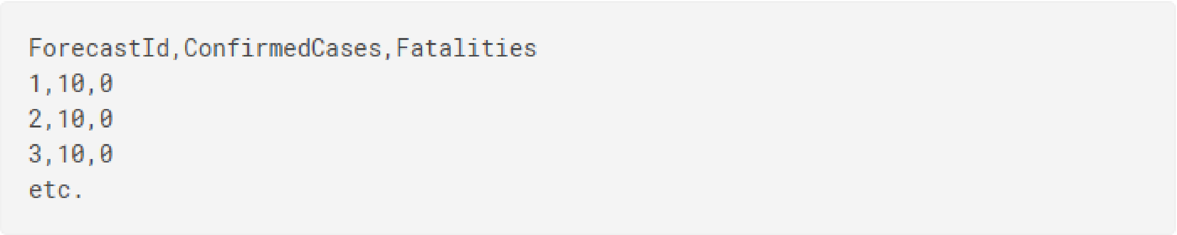
n is the total number of observations
p_i is your prediction
a_i is the actual value
$\log(x)$ is the natural logarithm of x

The final score is the mean of the RMSLE over all columns.

Submission File

We understand this is a serious situation, and in no way want to trivialize the human impact this crisis is causing by predicting fatalities. Our goal is to provide better methods for estimates that can assist medical and governmental institutions to prepare and adjust as pandemics unfold.

For each ForecastId in the test set, you'll predict the cumulative COVID-19 cases and fatalities to date. The file should contain a header and have the following format:



```
ForecastId,ConfirmedCases,Fatalities
1,10,0
2,10,0
3,10,0
etc.
```

You will get the ForecastId for the corresponding date and location from the test.csv file.

Date Description

In this challenge, you will be predicting the *cumulative* number of confirmed COVID19 cases in various locations across the world, as well as the number of resulting fatalities, for *future* dates.

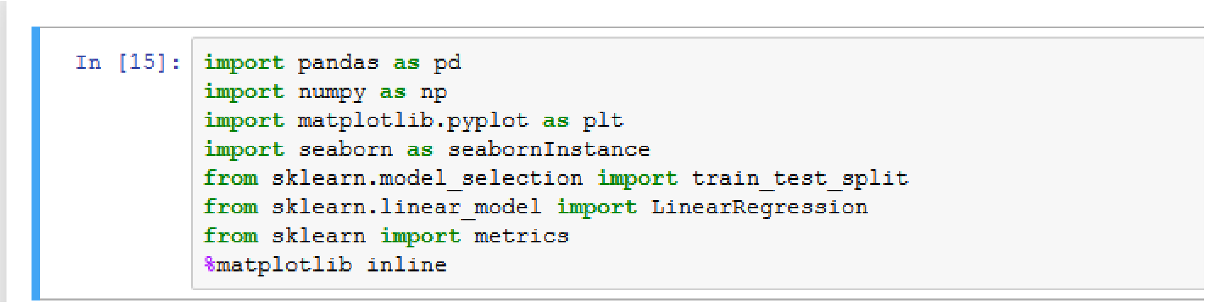
We understand this is a serious situation, and in no way want to trivialize the human impact this crisis is causing by predicting fatalities. Our goal is to provide better methods for estimates that can assist medical and governmental institutions to prepare and adjust as pandemics unfold.

Files:

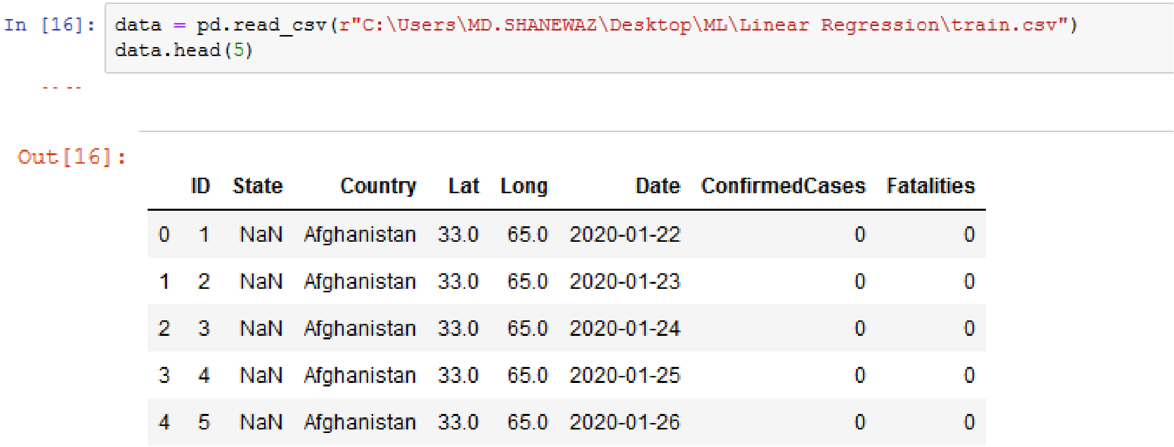
- train.csv - the training data up to Mar 18, 2020.
- test.csv - the dates to predict; there is a week of overlap with the training data for the initial Public leaderboard. Once submissions are paused, the Public leaderboard will update based on last 28 days of predicted data.
- submission.csv - a sample submission in the correct format; again, predictions should be *cumulative*

This evaluation data for this competition comes from **John Hopkins CSSE**, which is uninvolved in the competition.

Screenshot of Implementation



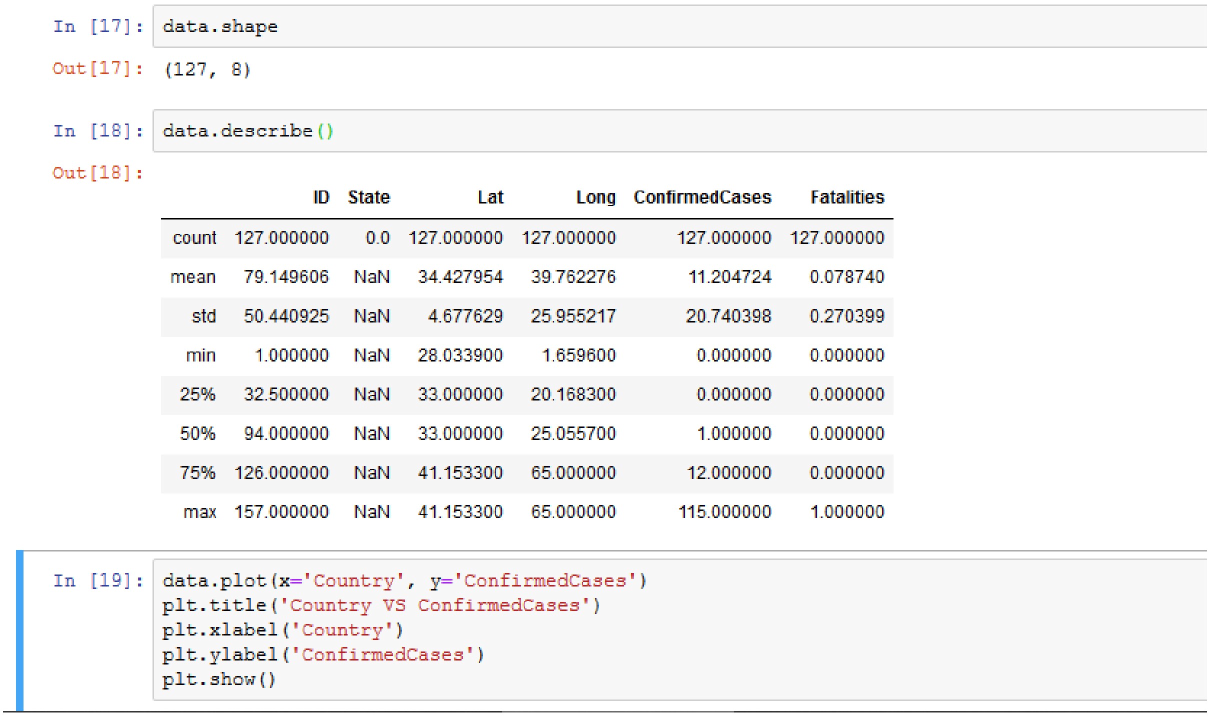
```
In [15]: import pandas as pd
         import numpy as np
         import matplotlib.pyplot as plt
         import seaborn as seabornInstance
         from sklearn.model_selection import train_test_split
         from sklearn.linear_model import LinearRegression
         from sklearn import metrics
         %matplotlib inline
```



```
In [16]: data = pd.read_csv(r"C:\Users\MD.SHANEWAZ\Desktop\ML\Linear Regression\train.csv")
         data.head(5)
```

Out[16]:

	ID	State	Country	Lat	Long	Date	ConfirmedCases	Fatalities
0	1	NaN	Afghanistan	33.0	65.0	2020-01-22	0	0
1	2	NaN	Afghanistan	33.0	65.0	2020-01-23	0	0
2	3	NaN	Afghanistan	33.0	65.0	2020-01-24	0	0
3	4	NaN	Afghanistan	33.0	65.0	2020-01-25	0	0
4	5	NaN	Afghanistan	33.0	65.0	2020-01-26	0	0



```
In [17]: data.shape
```

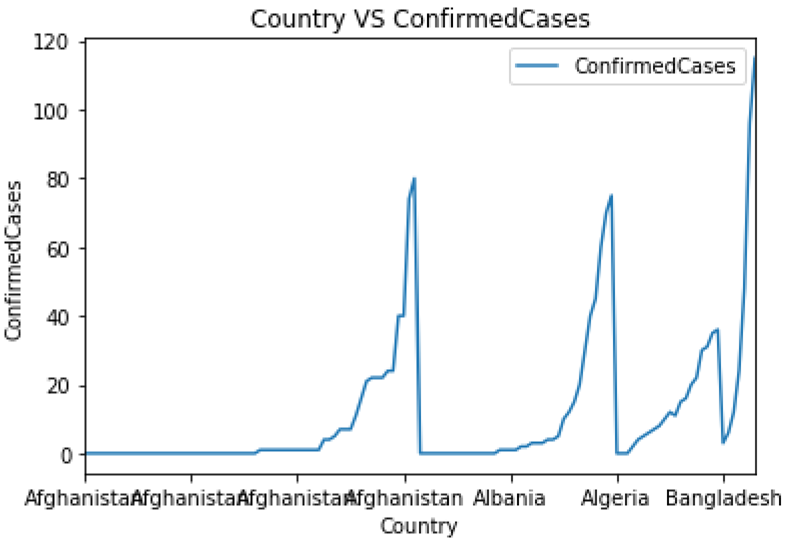
```
Out[17]: (127, 8)
```

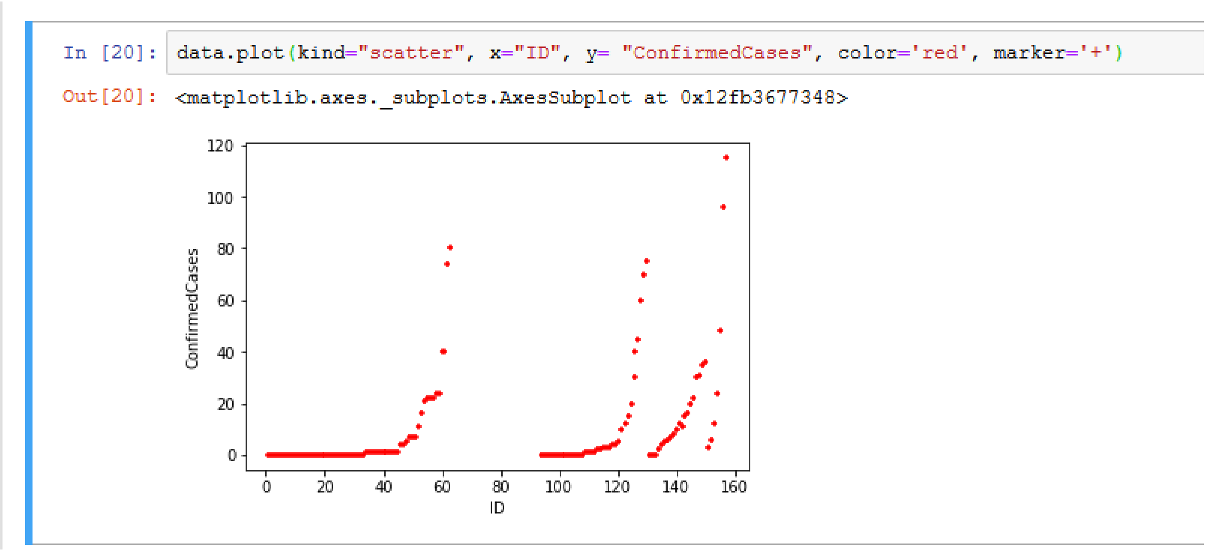
```
In [18]: data.describe()
```

Out[18]:

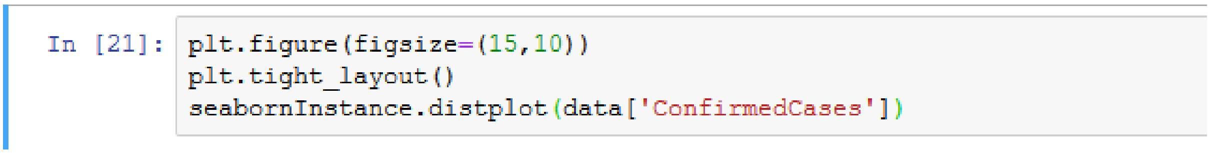
	ID	State	Lat	Long	ConfirmedCases	Fatalities
count	127.000000	0.0	127.000000	127.000000	127.000000	127.000000
mean	79.149606	NaN	34.427954	39.762276	11.204724	0.078740
std	50.440925	NaN	4.677629	25.955217	20.740398	0.270399
min	1.000000	NaN	28.033900	1.659600	0.000000	0.000000
25%	32.500000	NaN	33.000000	20.168300	0.000000	0.000000
50%	94.000000	NaN	33.000000	25.055700	1.000000	0.000000
75%	126.000000	NaN	41.153300	65.000000	12.000000	0.000000
max	157.000000	NaN	41.153300	65.000000	115.000000	1.000000

```
In [19]: data.plot(x='Country', y='ConfirmedCases')
         plt.title('Country VS ConfirmedCases')
         plt.xlabel('Country')
         plt.ylabel('ConfirmedCases')
         plt.show()
```

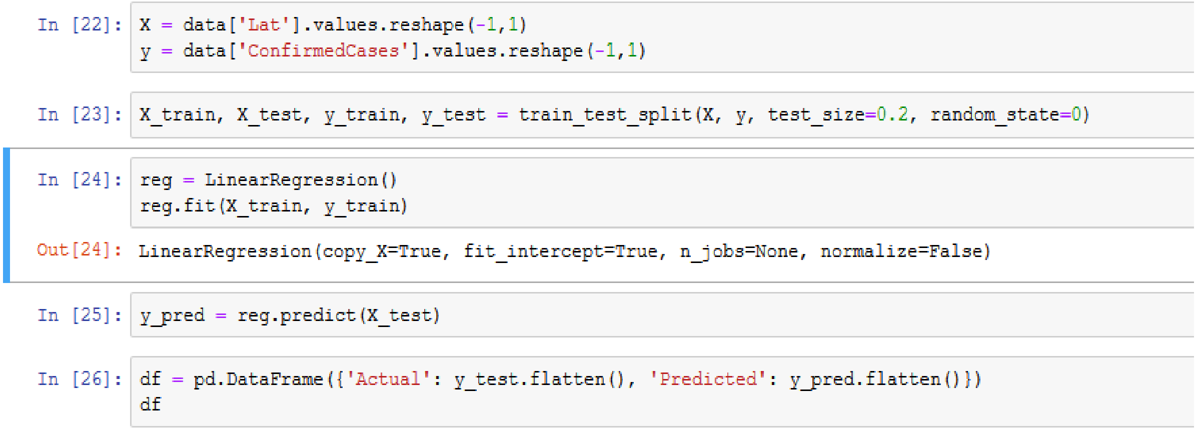




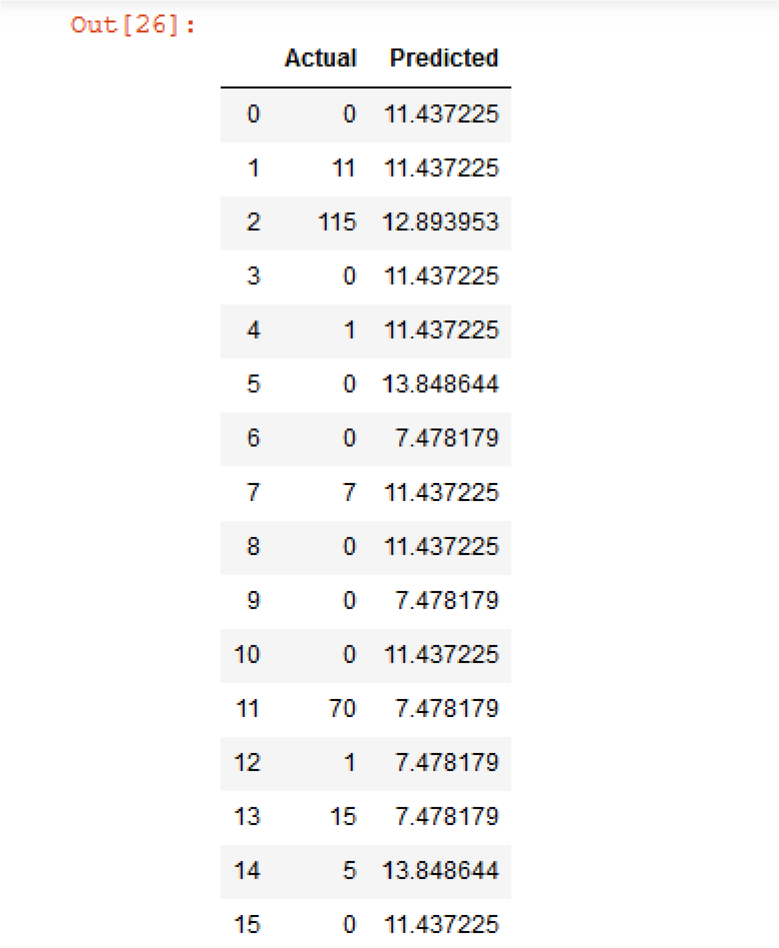
```
In [20]: data.plot(kind="scatter", x="ID", y= "ConfirmedCases", color='red', marker='+')
Out[20]: <matplotlib.axes._subplots.AxesSubplot at 0x12fb3677348>
```



```
In [21]: plt.figure(figsize=(15,10))
         plt.tight_layout()
         seabornInstance.distplot(data['ConfirmedCases'])
```

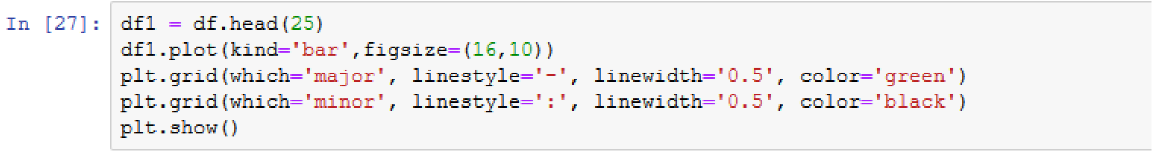


```
In [22]: X = data['Lat'].values.reshape(-1,1)
         y = data['ConfirmedCases'].values.reshape(-1,1)

In [23]: X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=0)

In [24]: reg = LinearRegression()
         reg.fit(X_train, y_train)

Out[24]: LinearRegression(copy_X=True, fit_intercept=True, n_jobs=None, normalize=False)

In [25]: y_pred = reg.predict(X_test)

In [26]: df = pd.DataFrame({'Actual': y_test.flatten(), 'Predicted': y_pred.flatten()})
         df
```



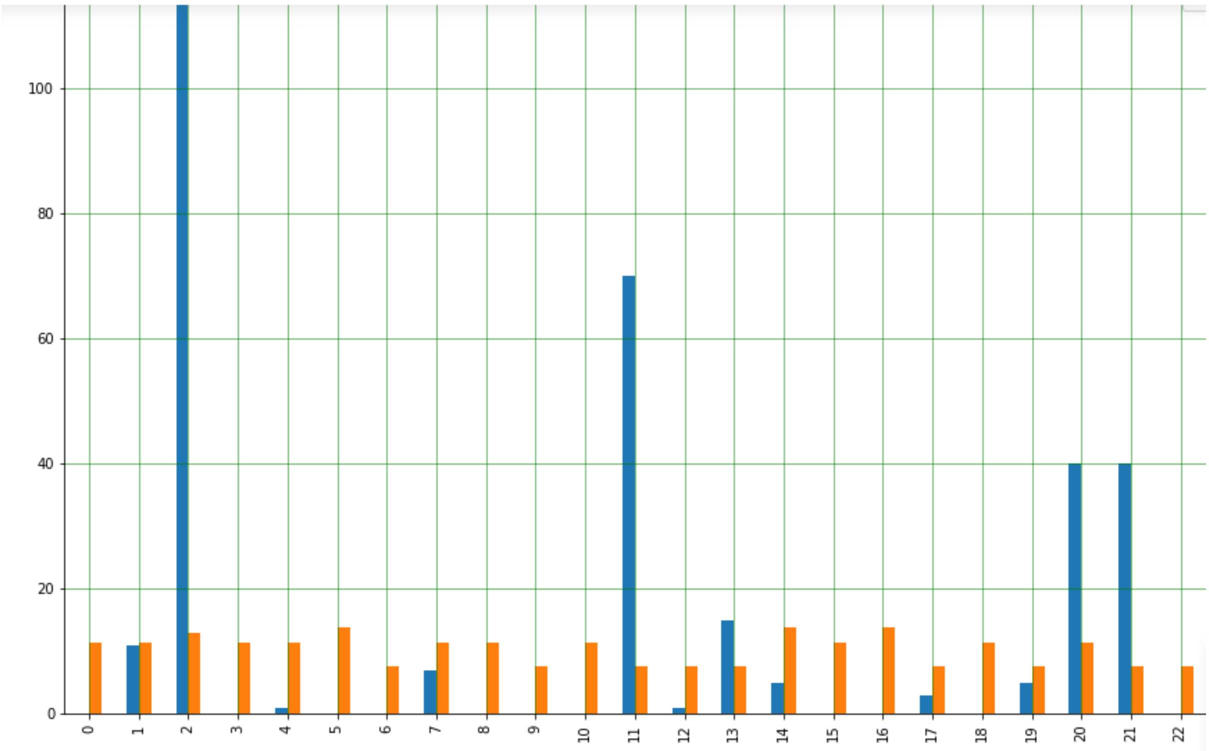
Out[26]:

	Actual	Predicted
0	0	11.437225
1	11	11.437225
2	115	12.893953
3	0	11.437225
4	1	11.437225
5	0	13.848644
6	0	7.478179
7	7	11.437225
8	0	11.437225
9	0	7.478179
10	0	11.437225
11	70	7.478179
12	1	7.478179
13	15	7.478179
14	5	13.848644
15	0	11.437225

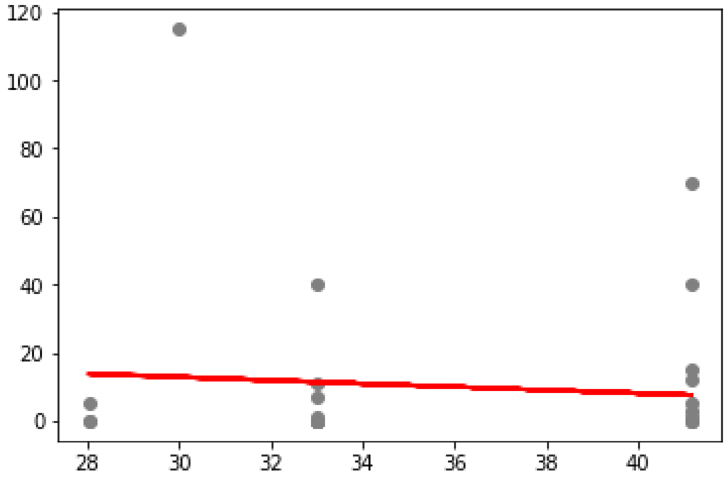


In [27]:

```
df1 = df.head(25)
df1.plot(kind='bar',figsize=(16,10))
plt.grid(which='major', linestyle='-', linewidth='0.5', color='green')
plt.grid(which='minor', linestyle=':', linewidth='0.5', color='black')
plt.show()
```



```
In [28]:  plt.scatter(X_test, y_test,  color='gray')
          plt.plot(X_test, y_pred, color='red', linewidth=2)
          plt.show()
```



Conclusions

The global pandemic of the severe acute respiratory syndrome Coronavirus 2 (SARS-CoV-2) has become the primary national security issue of many nations. Advancement of accurate prediction models for the outbreak is essential to provide insights into the spread and consequences of this infectious disease. Due to the high level of uncertainty and lack of crucial data, standard

epidemiological models have shown low accuracy for long-term prediction. This report presents a comparative analysis of ML and soft computing models to predict the COVID-19 outbreak. The results of two ML models (MLP and ANFIS) reported a high generalization ability for long-term prediction. With respect to the results reported in this paper and due to the highly complex nature of the COVID-19 outbreak and differences from nation-to-nation, this study suggests ML as an effective tool to model the time series of outbreak. For the advancement of higher performance models for long-term prediction, future research should be devoted to comparative studies on various ML models for individual countries. Due to the fundamental differences between the outbreak in various countries, advancement of global models with generalization ability would not be feasible. As observed and reported in many studies, it is unlikely that an individual outbreak will be replicated elsewhere

References:

- *https://www.kaggle.com/mdshanewazakib/global-forecasting-week-1-covid-19-rf-regression/data*
- ***https://www.kaggle.com/c/covid19-global-forecasting-week-1***
- ***https://www.researchgate.net/publication/340782507_COVID-19_Outbreak_Prediction_with_Machine_Learning***
- *Remuzzi, A.; Remuzzi, G. COVID-19 and Italy: what next? Lancet* ***2020***.
- *https://covid19-projections.com*
- *https://www.datarevenue.com/en-blog/machine-learning-covid-19*