קורס באנליזה של ביג דאטה

שאלת המחקר שבחרנו לעבודה זו היא האם קיים מתאם בין מספר המתים מווירוס הקורונה בארצות הברית לבין מדד הפחד המכונה VIX[1]. השערת המחקר הייתה שעל אף הסיוע הנרחב שקיבלו השווקים על ידי נשיא ארצות הברית, דונלד טרמפ, יש קשר בין מספר המתים לעליית המדד.

לצורך התרגיל השתמשנו בשני מקורות: מהראשון לקחנו את נתוני החולים והמתים מנגיף הקורונה בעולם, מהשני שאבנו את נתוני מדד הפחד[2] (אנחנו השתמשנו בעמודת 'VIX high') .

20 השורות הראשונות של הדאטה אחרי filter:

date	countriesAndTerritories	deaths
2020-04-25	United_States_of_...	1054
2020-04-24	United_States_of_...	3179
2020-04-23	United_States_of_...	1721
2020-04-22	United_States_of_...	2524
2020-04-21	United_States_of_...	1857
2020-04-20	United_States_of_...	1772
2020-04-19	United_States_of_...	1856
2020-04-18	United_States_of_...	3770
2020-04-17	United_States_of_...	2299
2020-04-16	United_States_of_...	4928
2020-04-15	United_States_of_...	2408
2020-04-14	United_States_of_...	1541
2020-04-13	United_States_of_...	1500
2020-04-12	United_States_of_...	1831
2020-04-11	United_States_of_...	2087
2020-04-10	United_States_of_...	1873
2020-04-09	United_States_of_...	1922
2020-04-08	United_States_of_...	1906
2020-04-07	United_States_of_...	1342
2020-04-06	United_States_of_...	1146

only showing top 20 rows

20 השורות הראשונות של הדאטה VIX אחרי filter[3]:

date	VIX High
2019-12-31	15.39
2020-01-02	13.72
2020-01-03	16.20
2020-01-06	16.39
2020-01-07	14.46
2020-01-08	15.24
2020-01-09	13.24
2020-01-10	12.87
2020-01-13	13.09
2020-01-14	13.82
2020-01-15	12.83
2020-01-16	12.42
2020-01-17	12.48
2020-01-21	13.33
2020-01-22	13.01
2020-01-23	14.15
2020-01-24	15.98
2020-01-27	19.02
2020-01-28	18.03
2020-01-29	16.65

only showing top 20 rows

[1] https://he.wikipedia.org/wiki/VIX

[2] http://www.cboe.com/products/vix-index-volatility/vix-options-and-futures/vix-index/vix-historical-data

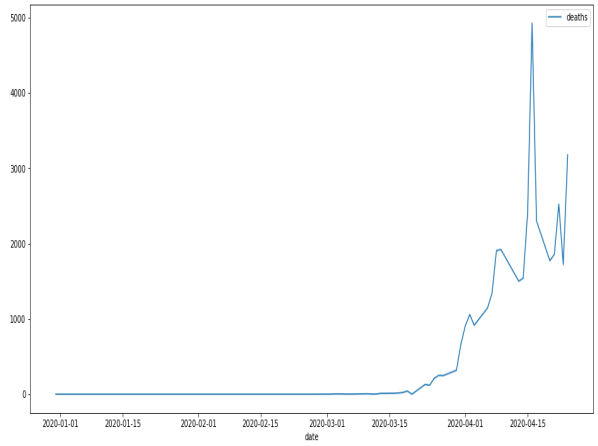
[3] בשתי קבצי הcsv אנחנו מקבלים את הנתונים ברזולוציה יומית ותאריכי המדידה הם מה31.12.19 ועד ה25.4.20

צירוף שתי הטבלאות:

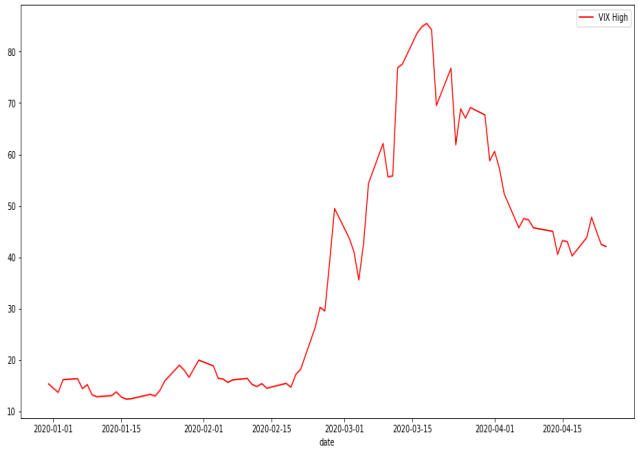
date	countriesAndTerritories	deaths	VIX High
2020-04-24	United_States_of_...	3179	42.08
2020-04-23	United_States_of_...	1721	42.47
2020-04-22	United_States_of_...	2524	45.07
2020-04-21	United_States_of_...	1857	47.77
2020-04-20	United_States_of_...	1772	43.83
2020-04-17	United_States_of_...	2299	40.26
2020-04-16	United_States_of_...	4928	43.02
2020-04-15	United_States_of_...	2408	43.23
2020-04-14	United_States_of_...	1541	40.57
2020-04-13	United_States_of_...	1500	45.04
2020-04-09	United_States_of_...	1922	45.73
2020-04-08	United_States_of_...	1906	47.28
2020-04-07	United_States_of_...	1342	47.51
2020-04-06	United_States_of_...	1146	45.73
2020-04-03	United_States_of_...	915	52.29
2020-04-02	United_States_of_...	1059	57.24
2020-04-01	United_States_of_...	909	60.59
2020-03-31	United_States_of_...	661	58.75
2020-03-30	United_States_of_...	318	67.69
2020-03-27	United_States_of_...	246	69.10

only showing top 20 rows

גרף המציג את מקרי המוות על פי תאריך:



גרף המציג את מדד הVIX על פי תאריך:



פילוח ערכי מדד VIX לפי חודשים:

מקסימום

month(date)	max(VIX High)
12	15.39
1	19.99
3	85.47
4	60.59
2	49.48

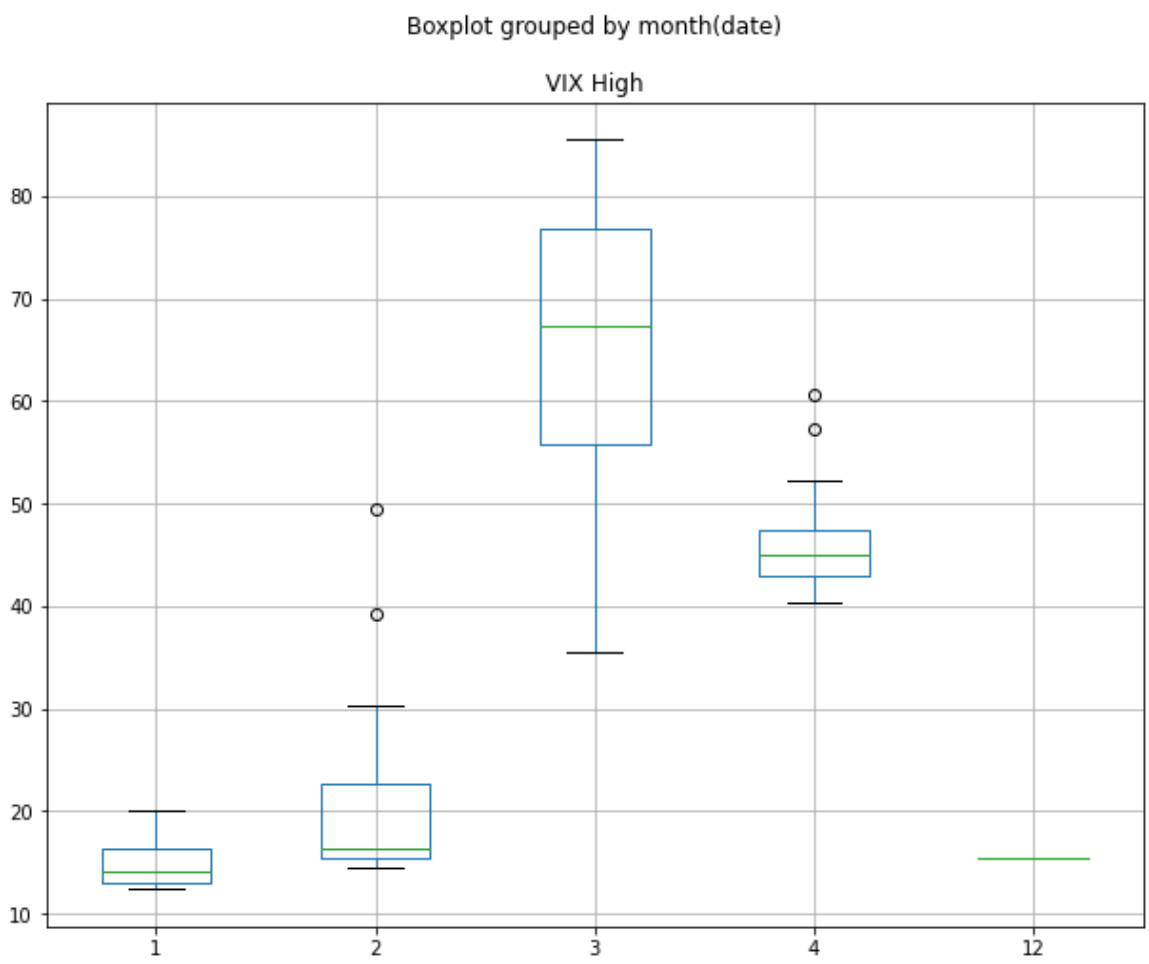
ממוצע

month(date)	avg(VIX High)
12	15.390000343322754
1	15.014761833917527
3	64.6977275501598
4	46.453529582304114
2	21.086841934605648

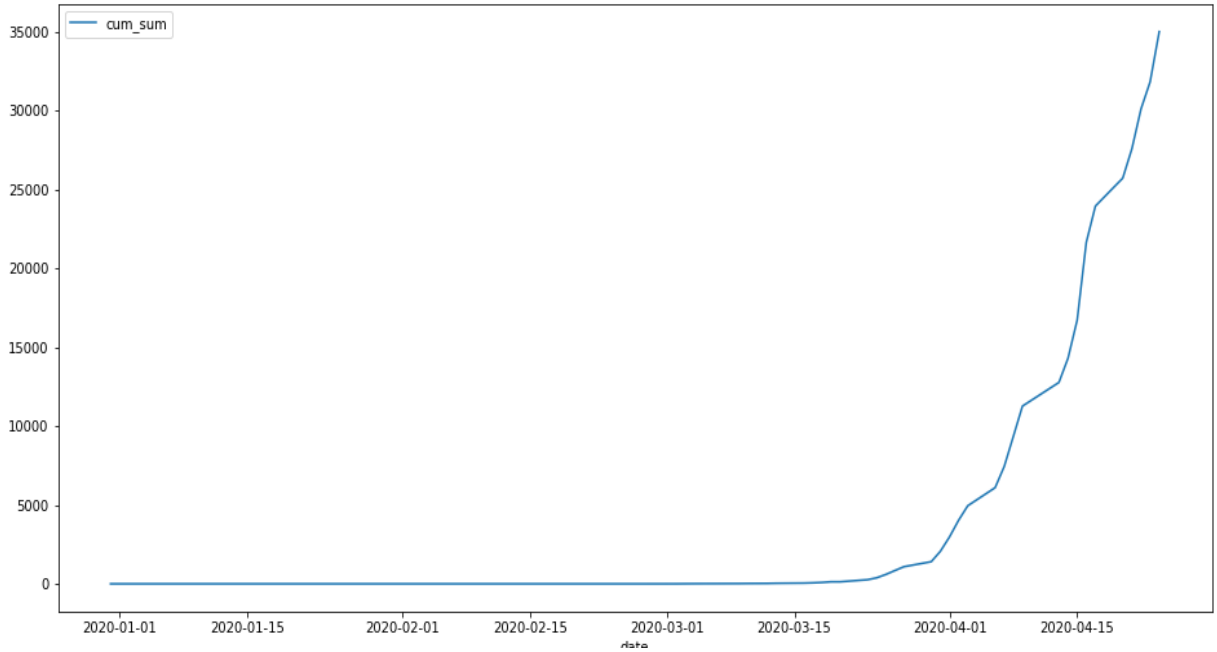
מינימום

month(date)	min(VIX High)
12	15.39
1	12.42
3	35.58
4	40.26
2	14.54

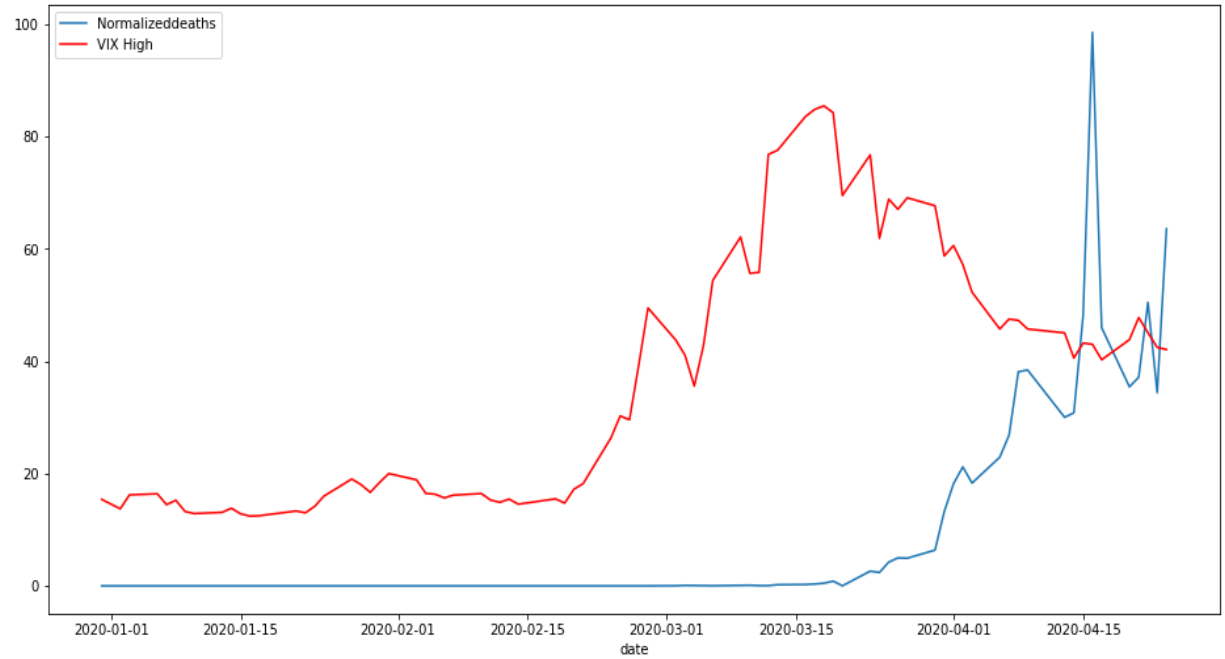
גרף "קופסה" המציג את ערכי המדד VIX:



גרף המציג את חישוב הסכומים החלקיים של כמות המתים בכל יום בארצות הברית:



קשר בין עמודת VIX high לעמודת death[4]:



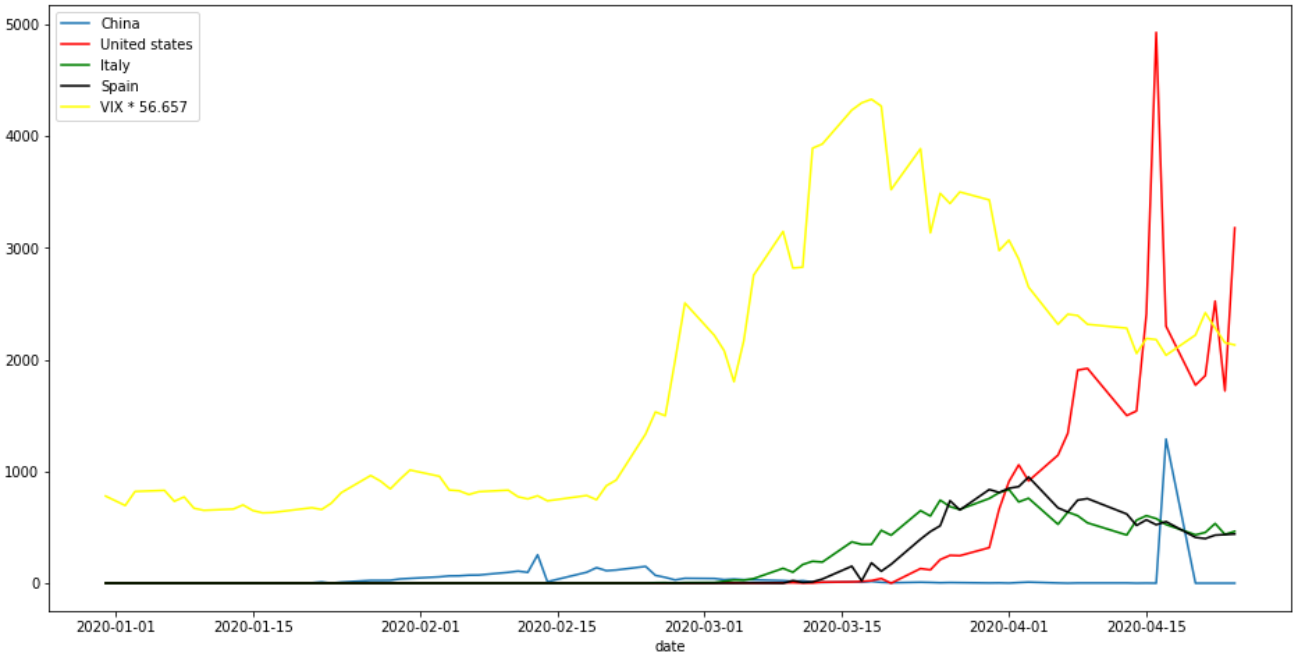
[4] כאשר עמדת deaths הוקטנה ביחס מחושב - לפי הערכים כאן

טבלה שמשלבת את מקרי המוות של סין, איטליה וארצות הברית עם מדד הפחד VIX:

	date	United states	China	Italy	Spain	VIX High
0	2019-12-31	0	0	0	0	15.390000
1	2020-01-02	0	0	0	0	13.720000
2	2020-01-03	0	0	0	0	16.200001
3	2020-01-06	0	0	0	0	16.389999
4	2020-01-07	0	0	0	0	14.460000
...	...	...	...	...	...	...
75	2020-04-20	1772	0	433	410	43.830002
76	2020-04-21	1857	0	454	399	47.770000
77	2020-04-22	2524	0	534	430	45.070000
78	2020-04-23	1721	0	437	435	42.470001
79	2020-04-24	3179	0	464	440	42.080002

80 rows × 6 columns

גרף המתאר את מדד VIX[5] ביחד לכמות המתים בסין ארצות הברית ואיטליה – ניתן לראות את הטבלה המשולבת כאן:

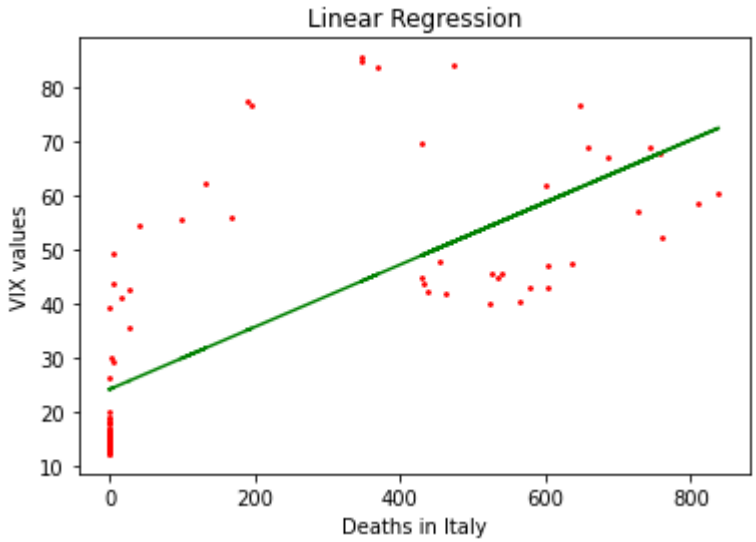


[5] כאשר הכפלנו את היחס בין מקס כמות המתים בערך המקסימלי של VIX – לפי הערכים כאן

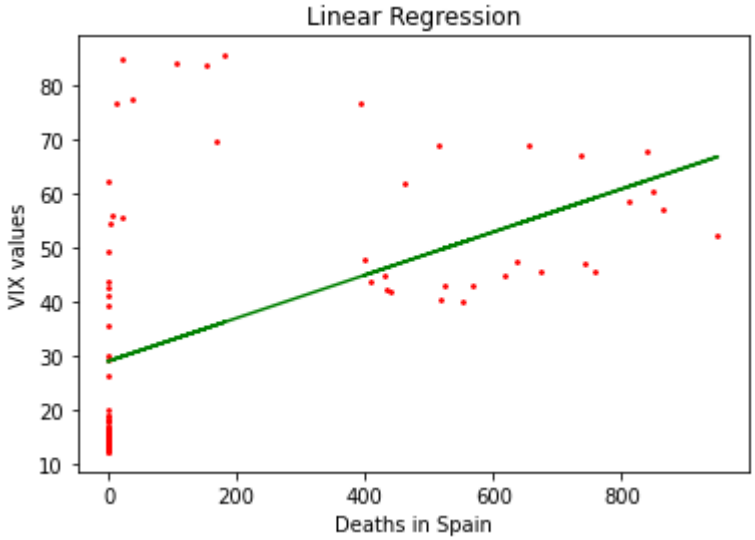
רגרסיה ליניארית:

בתהליך זה נרצה למצוא קשר בין מספר המתים באיטליה וספרד- המשתנים הבלתי תלוי, לבין מדד הפחד - המשתנה התלוי.[6] [7] [8]

רגרסיה ליניארית עבור איטליה:



רגרסיה ליניארית עבור ספרד:



[6] https://becominghuman.ai/jump-start-with-linear-regression-using-pyspark-mllib-f7f1578a684a

[7] https://medium.com/kharpann/perform-linear-regression-on-big-data-using-python-spark-and-mllib-b1204769547e

[8] https://he.wikipedia.org/wiki/%D7%A8%D7%92%D7%A8%D7%A1%D7%99%D7%94_%D7%9C%D7%99%D7%A0%D7%99%D7%90%D7%A8%D7%99%D7%AA

נתאר את השלבים שעברנו על מנת לייצר את הרגרסיה:

1. יצרנו טבלה שעמודותיה ממושקלות על פי היחסים [0.7, 0.3][9] והן: features, label . כאשר features הוא מערך בעל 2 איברים אשר מכיל את מספר המתים מהמדינות איטליה וספרד (משתנים בלתי תלויים) ו label הוא ערך המדד VIX (משתנה תלוי). קישור להצגה
2. אימון המודל וביצוע תחזיות. קישור להצגה. קישור להצגה
3. ניתוח התחזית:

שורה ראשונה בתמונה מטה מייצגת את R^2 – מקדם מתאם מרובה (מייצג כמה מתאים האלגוריתם שנבחר לנתונים שלנו, ככל שקרוב יותר ל1 יותר מתאים).

שורה שנייה בתמונה מטה מייצגת את הטעות הריבועית הממוצעת (mean square error) – זה ההבדל בין האומד לבין מה שנאמד.[10]

שורה שלישית היא שורש הטעות הממוצעת (rmse)

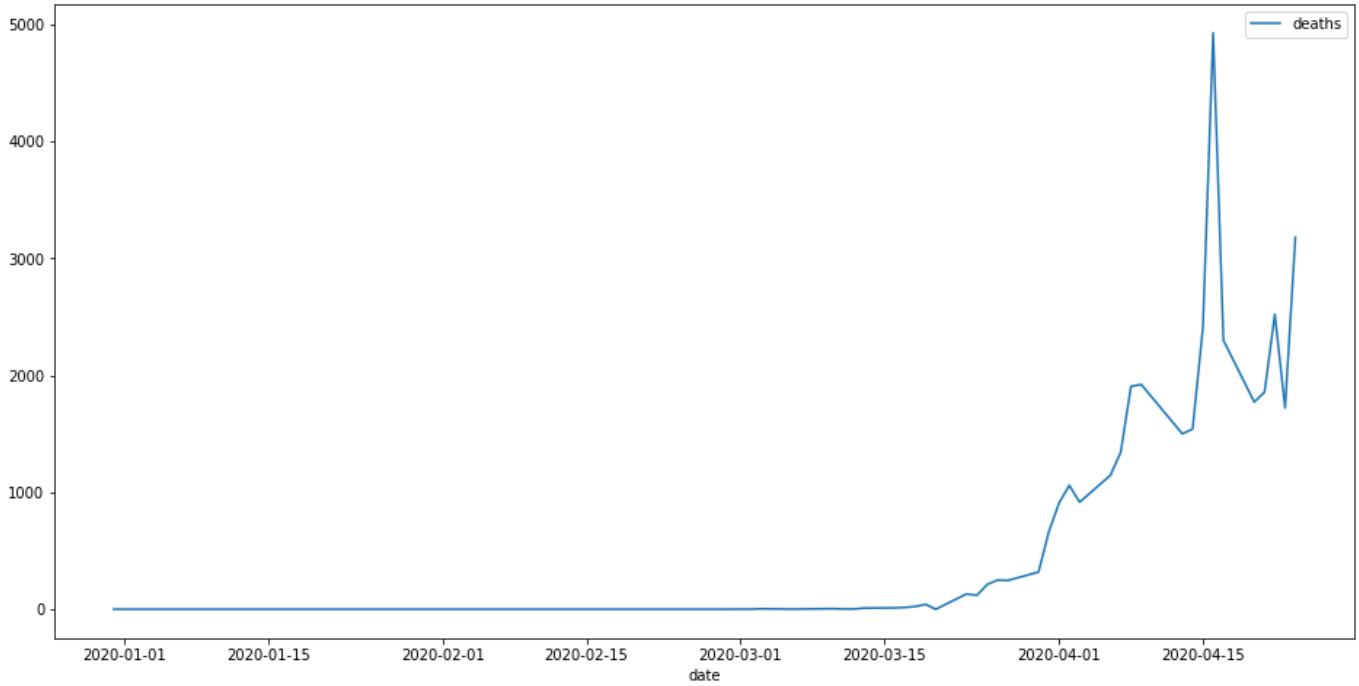
```
0.800176763529667
104.61325049439789
10.2280619129138
```

[9] ראינו לפני הדוגמאות באינטרנט

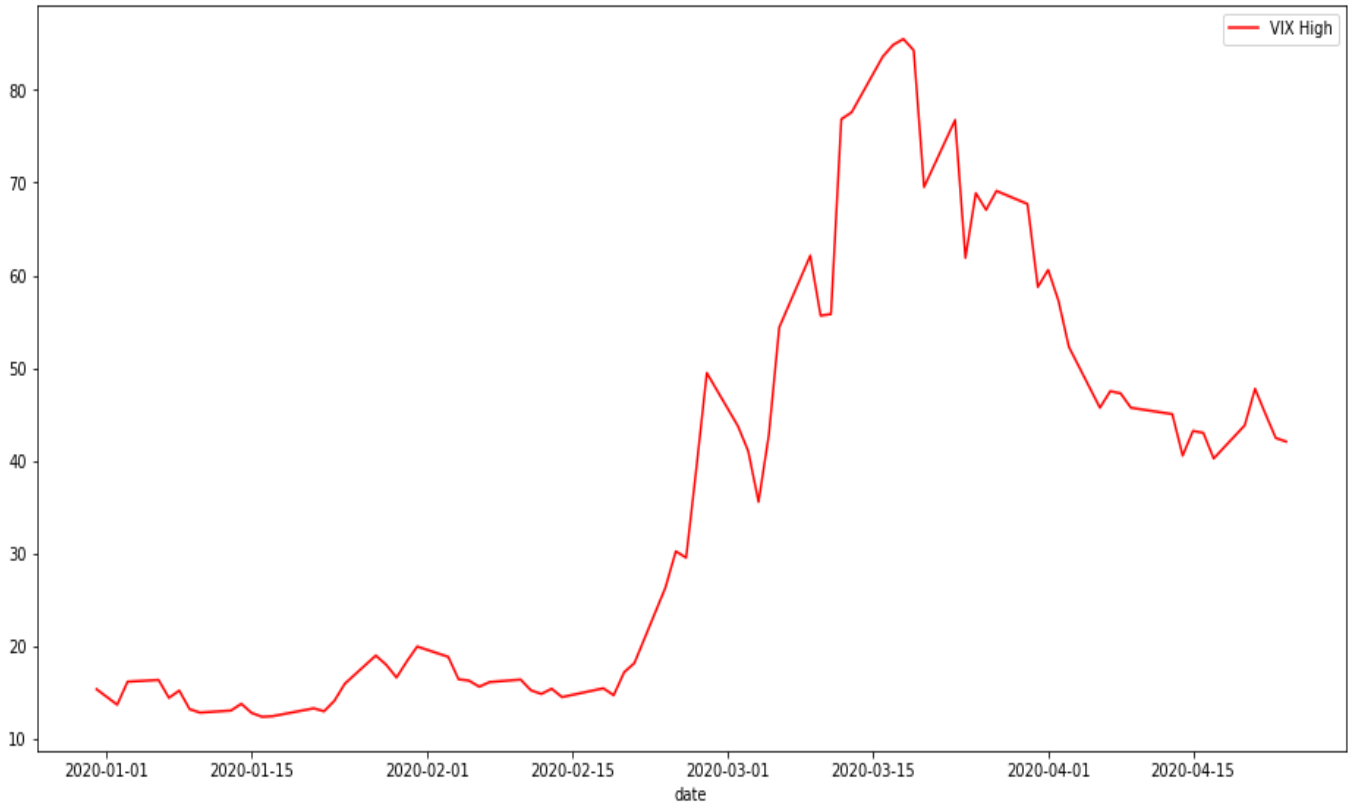
[10] https://he.wikipedia.org/wiki/%D7%98%D7%A2%D7%95%D7%AA_%D7%A8%D7%99%D7%91%D7%95%D7%A2%D7%99%D7%AA_%D7%9E%D7%9E%D7%95%D7%A6%D7%A2%D7%AA

נספחים נוספים:

גרף מקרי מוות לפי תאריך בארצות הברית – מוגדל



גרף VIX High לפי תאריך – מוגדל



תיאור הטבלה המשולבת הסופית:

	0	1	2	3	4
summary	count	mean	stddev	min	max
United states	80	437.375	904.8556409796307	0	4928
China	80	44.3625	147.82484022002512	0	1290
Italy	80	218.075	282.30281681157	0	839
Spain	80	193.6375	294.81391873046834	0	950
VIX High	80	36.805125057697296	22.742882981526396	12.42	85.47

טבלה משולבת של סין, איטליה ארצות הברית ומדד VIX לפני ואחרי נרמול יחסי:

```
+----------+-------------+-----+-----+-----+--------+-----------------+
|      date|United states|China|Italy|Spain|VIX High|     VIX * 56.657|
+----------+-------------+-----+-----+-----+--------+-----------------+
|2019-12-31|            0|    0|    0|    0|   15.39|779.6112473917007|
|2020-01-02|            0|    0|    0|    0|   13.72|695.0140535268783|
|2020-01-03|            0|    0|    0|    0|    16.2|820.6434386482238|
|2020-01-06|            0|    0|    0|    0|   16.39|830.2681990814208|
|2020-01-07|            0|    0|    0|    0|   14.46|732.5002219324111|
|2020-01-08|            0|    0|    0|    0|   15.24|772.0126684055327|
|2020-01-09|            0|    0|    0|    0|   13.24|670.6986684055328|
|2020-01-10|            0|    0|    0|    0|   12.87|651.9555842027663|
|2020-01-13|            0|    0|    0|    0|   13.09|663.1001377296448|
|2020-01-14|            0|    0|    0|    0|   13.82|700.0797245407103|
|2020-01-15|            0|    1|    0|    0|   12.83|649.9293061351775|
|2020-01-16|            0|    0|    0|    0|   12.42|629.1599438648224|
|2020-01-17|            0|    0|    0|    0|   12.48|632.1993368110657|
|2020-01-21|            0|    3|    0|    0|   13.33|675.2578061351776|
|2020-01-22|            0|   11|    0|    0|   13.01|659.0475815944671|
|2020-01-23|            0|    0|    0|    0|   14.15| 716.796530675888|
|2020-01-24|            0|    9|    0|    0|   15.98|809.4988368110656|
|2020-01-27|            0|   25|    0|    0|   19.02|963.4961631889342|
|2020-01-28|            0|   25|    0|    0|   18.03|913.3457447834014|
|2020-01-29|            0|   26|    0|    0|   16.65| 843.439030675888|
+----------+-------------+-----+-----+-----+--------+-----------------+
only showing top 20 rows
```

טבלה (חלקית) ממושקלת של features and label:

features	label
[28.0,0.0]	35.58
[27.0,1.0]	42.84
[133.0,0.0]	62.12
[98.0,23.0]	55.66
[196.0,12.0]	76.83
[189.0,37.0]	77.57
[370.0,152.0]	83.56
[347.0,21.0]	84.83
[347.0,182.0]	85.47
[743.0,514.0]	68.86
[685.0,738.0]	67.06
[839.0,849.0]	60.59
[727.0,864.0]	57.24
[527.0,674.0]	45.73
[636.0,637.0]	47.51
[604.0,743.0]	47.28
[540.0,757.0]	45.73
[431.0,619.0]	45.04
[564.0,517.0]	40.57
[604.0,567.0]	43.23
[578.0,523.0]	43.02
[433.0,410.0]	43.83
[534.0,430.0]	45.07

פלט התחזיות של המודל:

```
+-------------+-----+------------------+
|features     |label|prediction        |
+-------------+-----+------------------+
|(2,[],[])    |16.39|22.629243586508487|
|(2,[],[])    |13.09|22.629243586508487|
|(2,[],[])    |12.83|22.629243586508487|
|(2,[],[])    |13.01|22.629243586508487|
|(2,[],[])    |18.39|22.629243586508487|
|(2,[],[])    |19.99|22.629243586508487|
|(2,[],[])    |16.46|22.629243586508487|
|(2,[],[])    |15.66|22.629243586508487|
|(2,[],[])    |14.88|22.629243586508487|
|(2,[],[])    |14.54|22.629243586508487|
|(2,[],[])    |15.49|22.629243586508487|
|(2,[],[])    |18.21|22.629243586508487|
|(2,[],[])    |26.35|22.629243586508487|
|[5.0,0.0]    |49.48|23.486974345668678|
|[41.0,2.0]   |54.39|29.431332356792012|
|[167.0,7.0]  |55.82|50.46788885055373 |
|[473.0,107.0]|84.26|91.39583856965518 |
|[429.0,169.0]|69.51|76.6774007893141  |
|[649.0,394.0]|76.74|88.39591552398247 |
|[601.0,462.0]|61.88|72.2973827718231  |
+-------------+-----+------------------+
only showing top 20 rows
```