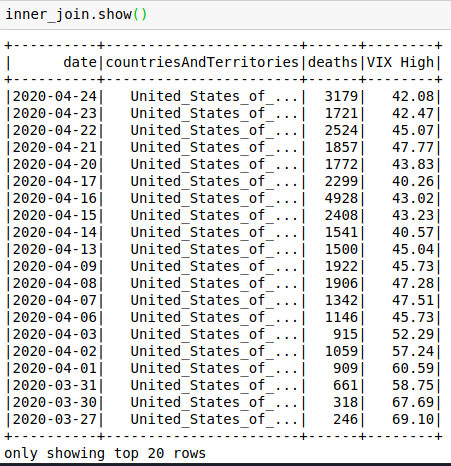
**קורס באנליזה של ביג דאטה**

שאלת המחקר שבחרנו לעבודה זו היא האם קיים מתאם בין מספר המתים מווירוס הקורונה בארצות הברית לבין מדד הפחד המכונה VIX[[1]](#footnote-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]](#footnote-2) (אנחנו השתמשנו בעמודת 'VIX high' (.

|  |  |
| --- | --- |
| 20 השורות הראשונות של הדאטה אחרי filter: | 20 השורות הראשונות של הדאטה VIX אחרי filter[[3]](#footnote-3): |
|  |  |

## צירוף שתי הטבלאות:

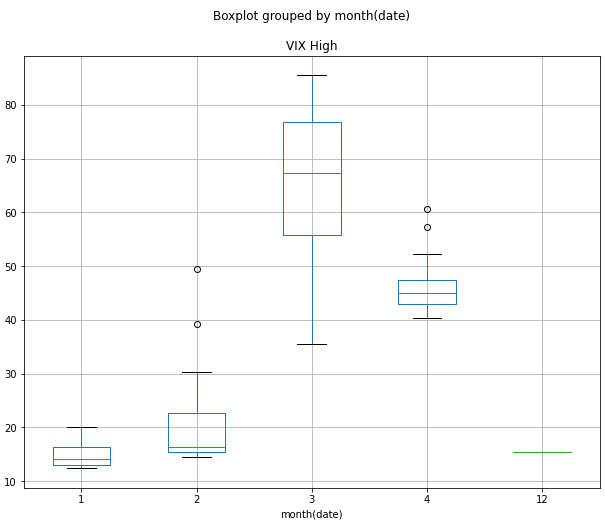


|  |  |
| --- | --- |
| גרף המציג את מקרי המוות על פי תאריך: | גרף המציג את מדד הVIX על פי תאריך: |
|  |  |

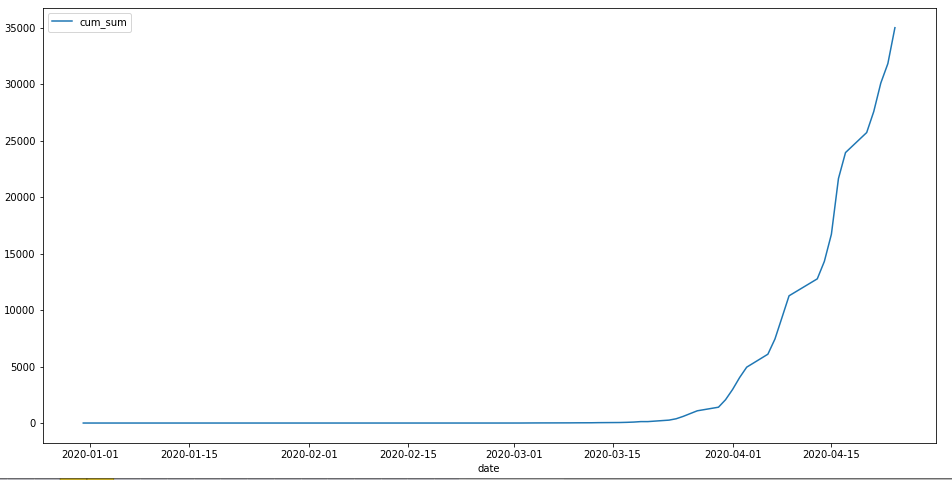
## פילוח ערכי מדד VIX לפי חודשים:

|  |  |  |
| --- | --- | --- |
| **מינימום** | **ממוצע** | **מקסימום** |
|  |  |  |

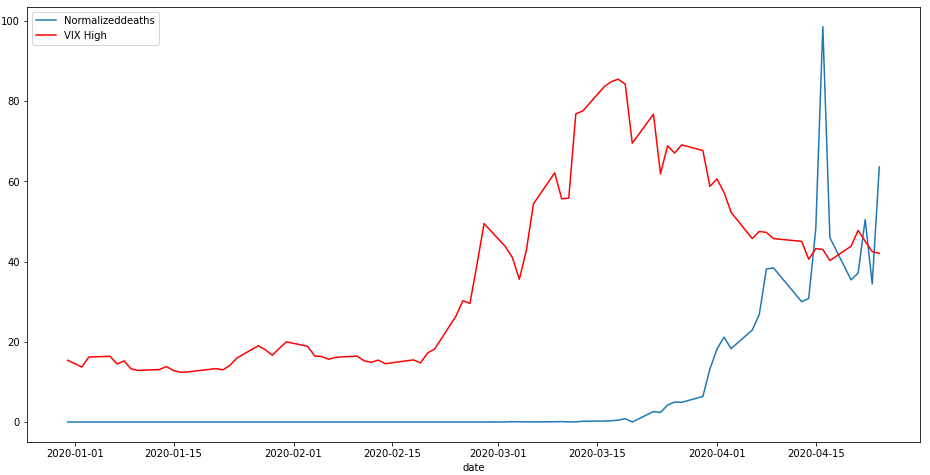
# גרף "קופסה" המציג את ערכי המדד VIX:



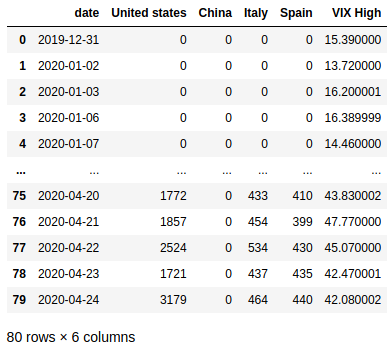
## גרף המציג את חישוב הסכומים החלקיים של כמות המתים בכל יום בארצות הברית:



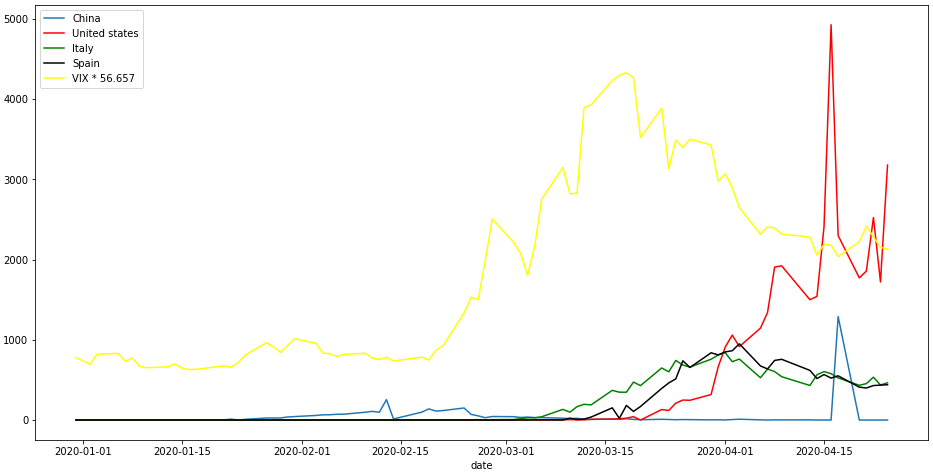
## קשר בין עמודת VIX high לעמודת death[[4]](#footnote-4):



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



## גרף המתאר את מדד [[5]](#footnote-5)VIX ביחד לכמות המתים בסין ארצות הברית ואיטליה – ניתן לראות את הטבלה המשולבת [כאן](#טבלה_מלאה):



## רגרסיה ליניארית:

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

|  |
| --- |
| רגרסיה ליניארית עבור איטליה: |
|  |
| רגרסיה ליניארית עבור ספרד: |
|  |

נתאר את השלבים שעברנו על מנת לייצר את הרגרסיה:

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

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

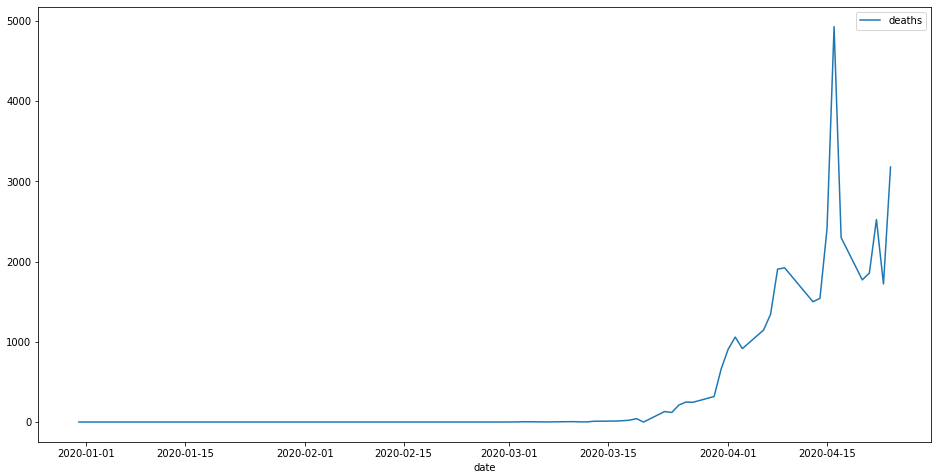
**שורה שנייה** בתמונה מטה מייצגת את הטעות הריבועית הממוצעת (mean square error) – זה ההבדל בין האומד לבין מה שנאמד.[[10]](#footnote-10)

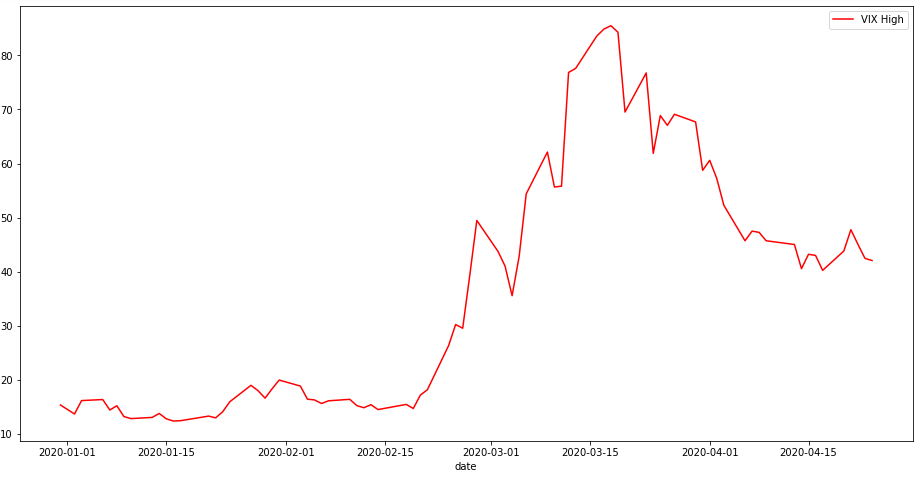
**שורה שלישית** היא שורש הטעות הממוצעת (rmse)



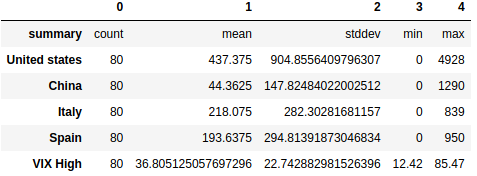
# נספחים נוספים:

## גרף מקרי מוות לפי תאריך בארצות הברית – מוגדל

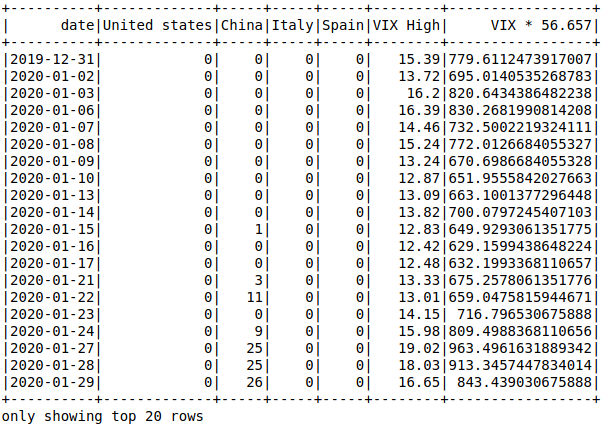


גרף VIX High לפי תאריך – מוגדל

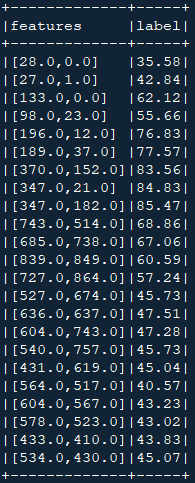
## תיאור הטבלה המשולבת הסופית:



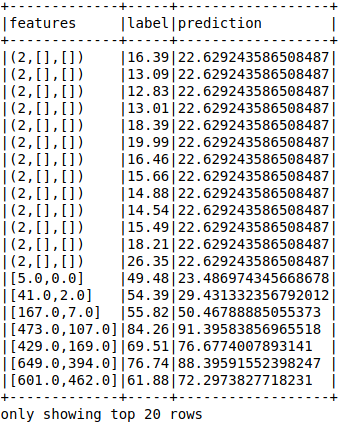
## טבלה משולבת של סין, איטליה ארצות הברית ומדד VIX לפני ואחרי נרמול יחסי:



## טבלה (חלקית) ממושקלת של features and label:



פלט התחזיות של המודל:



1. <https://he.wikipedia.org/wiki/VIX> [↑](#footnote-ref-1)
2. <http://www.cboe.com/products/vix-index-volatility/vix-options-and-futures/vix-index/vix-historical-data> [↑](#footnote-ref-2)
3. בשתי קבצי הcsv אנחנו מקבלים את הנתונים ברזולוציה יומית ותאריכי המדידה הם מה31.12.19 ועד ה25.4.20 [↑](#footnote-ref-3)
4. כאשר עמדת deaths הוקטנה ביחס מחושב - לפי הערכים [כאן](#טבלה_מתארת) [↑](#footnote-ref-4)
5. כאשר הכפלנו את היחס בין מקס כמות המתים בערך המקסימלי של VIX – לפי הערכים [כאן](#טבלה_מתארת) [↑](#footnote-ref-5)
6. <https://becominghuman.ai/jump-start-with-linear-regression-using-pyspark-mllib-f7f1578a684a> [↑](#footnote-ref-6)
7. <https://medium.com/kharpann/perform-linear-regression-on-big-data-using-python-spark-and-mllib-b1204769547e> [↑](#footnote-ref-7)
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> [↑](#footnote-ref-8)
9. ראינו לפני הדוגמאות באינטרנט [↑](#footnote-ref-9)
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> [↑](#footnote-ref-10)