TIME SERIES MODELING OF COVID-19

רקע

בפרוייקט זה אנו מנסים למדל ולחזות את התפרצות הקורונה באזורים נבחרים

ליאור רזניק, אלכס בוייב

מבוא

סדרות עתיות(או time-series) הן רצף של תצפיות הנקלטות לאורך פרקי זמן שווים, בין אם מדובר בשעות, ימים, חודשים, שנים ואף עשרות שנים.

בסדרות עתיות נעזרים בתחומים רבים כאשר המשותף לכל התחומים הוא הניסיון ללמוד איך הסדרה מתנהגת ובעזרת הבנה זו לנסות ולחזות את העתיד דבר שלכשעצמו נורא חשוב.

דוגמאות לבעיות שניתן לפתור בתחום: ערך פתיחה או סגירה של מניה בבורסה, רווחים של חברה ברבעון, כמות השיחות היומיות למוקד המשטרתי וכו'.

סדרת עתית היא סוג של תהליך סטוכסטי כלומר, תהליך שכל ההתפתחות שלו תלויה בגורמים מקריים, במילים אחרות: מדובר בתהליך לא-דטרמיניסטי, קיימים מספר מצבים שאליהם הסדרה יכולה להתפתח.

עובדה זו גורמת לכך שחיזוי של סדרה עתית יכול להיות קשה ובעייתי מאוד.

בנוסף, חלק נרחב מהעיבוד המקדים והטיפול במידע שאנחנו מכירים מבעיות אחרות יהיה שונה כאן (למשל: החלוקה לסטים של אימון, פיתוח ובדיקה יהיה שונה).

בחלק הראשון של עבודה זו אנו נסקור כלים סטטיסטים בסיסים,את הבסיס התאורתי העומד מאוחרי סדרות עתיות ומודלים לחיזוי ומידול של סדרות עתיות.בחלק השני נתאר את תהליך העבודה, סביבות העבודה והמודלים שבחרנו לבנות ובחלק האחרון נדבר על הערכת המודלים.

הערה: מידלנו עבור ישראל ועבור כל העולם, בכדי לחסוך במקום נסקור כאן רק את העבודה על ישראל (בעולם השתמשנו באותם הכלים).

חלק ראשון: רקע תאורתי , סקירה כללית

אנליזה של סדרה עתית

כאשר אנו באים למדל סדרת עתית(לפחות במידול הקלאסי), השלב הראשון הוא לבצע אנליזה עליה, להכיר אותה יותר לעמוק, לתאר אותה מבחינה סטטיסטית, להבין איך היא מתנהגת, האם היא סציונרית, האם היא עונתית, האם היא מורכבת אך ורק מרעש לבן או שאם יש בה רעש לבן כלשהו, האם ישנה קורלציה בין הסדרה לבין עצמה. כל השאלות הללו ורבות אחרות הן שאלות שאם לא נשאל אותן לא נוכל להכין את המידע נכון ולא נוכל לבנות מודל אשר באמת מתאים לבעיה.

הסיבה לכך שהאנליזה של סדרה עתית שונה מאנליזות מידע אחרות שאנחנו מכירים היא כאמור העובדה שמדובר בתהליך לא-דטרמיניסטי.

בחלק זה נדבר על כל הדברים הללו (לצד דברים נוספים מתורת הסטטיסטיקה) ומשמעותם למידול נכון של סדרה עתית.

סטטיסטיקה תיאורית בסיסית

כמו בכל אנליזה של מידע, ישנם שלבים אשר ניתן לעשות בכדי להבין יותר טוב את המידע: סיכום חמשת המספרים, טווח בין רבעוני (IQR), ממוצע מדגמי, סטיית תקן מדגמית, ייצוג גרפי (Scatterplot, boxplot , histogram) וכו'. בעזרת כל הכלים הללו אנו יכולים לקבל מושג כללי על המידע אשר ברשותנו, את ההתפלגות שלו.

אז בעצם מה כל המושגים הללו אומרים? ובכן,

1. <u>חמשת המספרים(Five-Number Summary)</u>: כשמם, מדובר בחמשת מספרים אשר מביאים לנו תיאור בסיסי של הנתונים אשר ברשותנו, דבר אשר יכול לתת לנו מושג כללי לגבי:
   - התפלגות ופיזור הנתונים.
   - אנומליות הקיימות במידע.
   - הטווח של המידע.

   חמשת המספרים הללו כוללים את:

   - הערך המינימלי.
   - $Q1$ - הרבעון הנמוך או למעשה, 25 האחוזים השכחים של המידע.
   - $Q2$ - או למעשה החציון (Median) כלומר, 50 האחוזים.
   - $Q3$ - הרבעון העליון כלומר, 75 האחוזים.
   - הערך המקסימאלי.

הנגזרת הראשונה מחמשת המספרים הוא הטווח הבין רבעוני (IQR) והוא מחושב בצורה הבאה:

$IQR = Q3 - Q1$ הדברים הראשונים שאנו יכולים לצפות בעזרתו הם: פיזור והתפשטות הנתונים או במילים אחרות עד כמה הנתונים מפוזרים. בנוסף, אנו יכולים לזהות אנומליות במידע.

ברמת העקרון, ככל שהטווח הבין רבעוני גדול יותר כך הפיזור של הנתונים רחב יותר. ככל שהפיזור רחב יותר כך ההבדל בין התצפיות או הערכים האינדיבודלים גדול יותר.

מצד שני, ככל שהטווח הבין רבעוני קטן יותר כך הפיזור של הנתונים קטן יותר, כלומר הם קרובים יותר אחד לשני ולחציון וכך המידע עקבי יותר.

טוב, אז הבנו מה הוא פיזור ואיך ניתן לראות אם הערכים עקביים או לא אבל איך אנחנו יכולים לזהות חריגות? ובכן, חריגה נחשבת ככזאת אם היא נופלת מתחת ל: $Q1-1.5IQR$ או מעל ל: $Q3+1.5IQR$.

2. <u>Boxplot</u>: תרשים גרפי של חמשת המספרים.
3. <u>ממוצע מדגמי וסטיית תקן מדגמית</u>: לרוב, לא יהיה לנו את כל המידע על האוכלוסיה , יהיה לנו מדגם מייצג כלשהו שעליו נצטרך לבצע חישובים. שתי מדדים שאנו יכולים לחשב על המדגם הם: ממוצע (mean) וסטטית תקן (standard deviation) או למעשה את הפיזור של הנתונים שלנו ביחס לממוצע.
4. <u>היסטוגרמה(Histogram)</u>: מדובר בייצוג גרפי נוסף שבעזרתו אנחנו יכולים למדוד את הפיזור של התצפיות שלנו או למעשה את השכיחות של כל אחת מהתצפיות שלנו. בנוסף, אנו גם יכולים להטיל את פונקציית הצפיפות על ההיסטוגרמה.
5. <u>Scatterplot:</u> בעזרת סוג גרף זה ניתן לבדוק אם ישנה קורלציה בין גורמים מסויימים כמו למשל: כמות החולים המאומתים לבין מזג האוויר.
   כמובן שניתן לבצע חישוב מתמטי ולדעת בוודאוות אם קיים קשר ומה מקדמי הקשר, אך ניתן לקבל אינטואיציה על קשר כזה על ידי חיפוש תבניות בגרף מסוג זה, ברמת העקרון בעזרת גרף זה אנו יכולים לראות את טיב הקשר, את חוזקו והאם הוא שלילי או חיובי.
   בעזרת גרף זה אנו יכולים להחליט אם כדי להוסיף עוד פיצרים למודל שלנו.
   חשוב שנשים לב כי קיום קורלציה איננו מחייב קיום של סיבתיות אלא רק קיום של קשר.

סטטיסטיקה היסקית בסיסית

כמובן שאי אפשר למדל דבר מבלי סטטיסטיקה היסקית, בסטטיסטיקה זו בניגוד לסטטיסטיקה התיאורית ניתן לומר שאנו יותר "מפוכחים", שאנחנו מבינים שאין בידנו את כלל המידע על האוכלוסיה וכי אנחנו מסתכלים על מדגם מסוים בלבד.

פעמים רבות אנחנו נרצה לבצע אנליזה בעזרת רגרסיה לינארית בכדי לנסות ולהבין יותר טוב את הקשר בין המשתנים הבלתי תלוים למשתנה המטרה שלנו.

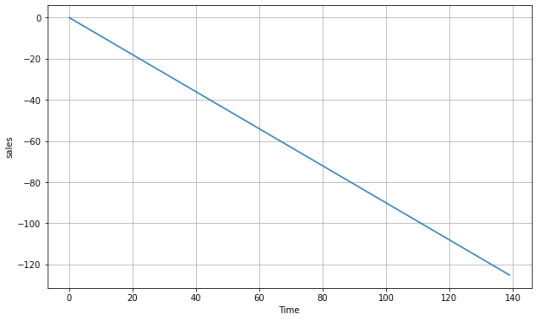
חשוב לשים לב כי, כאשר אנו באים לבצע אנלזיה בעזרת רגרסיה לינארית ההנחות הבאות(על הטעויות השאריות) הן סבירות(אם כי לא תמיד מתקיימות באמת):

- ההתפלגות היא התפלגות נורמלית והתוחלת היא 0 (ניתן לבדוק את ההנחה הנ"ל בעזרת היסטוגרמה או qq-plot).
- הן הומוסקדסטיות , כלומר בעלות שונות זהה או בשפה פשוטה פיזור אחיד סביב קו הרגרסיה.
- חוסר תלות בין תצפיות שונות.

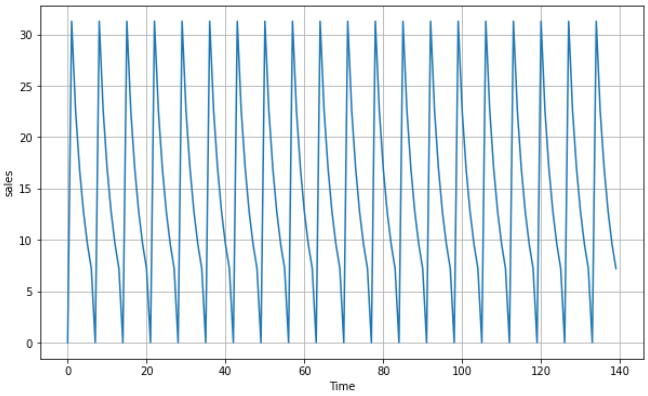
כמובן, שכאשר אנו באים למדל סדרה עתית אנחנו נשתמש ברגרסיה כלשהי (שכן, בסופו של דבר אנחנו באים לחזות משתנה רציף), אך לא בטוח שנרצה להשתמש ברגרסיה הלינארית הקלאסית (לפחות לא ללא הכנות מתאימות), שכן: סדרות עתיות לרוב אינן עומדות בהנחות אלו.

כאשר אנו באים למדל סדרה עתית(לפחות בעזרת כלים סטטיסטים קלאסיים) עלינו לבדוק דברים נוספים, כמו למשל:

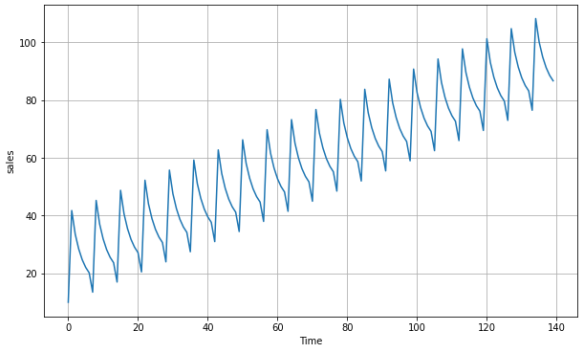
1. Trend: כלומר, אם יש כיוון או מגמה מסויימת שאליו המידע זורם.



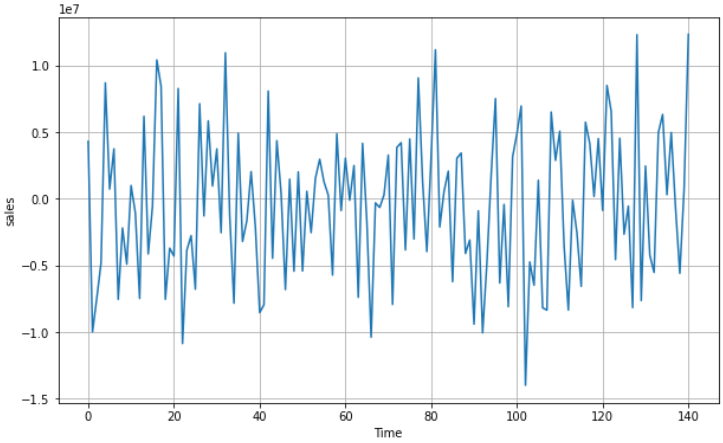
2. Seasonality: עונתיות, האם יש תבניות החוזרות על עצמן לאורך אינטרוולים שווים: למשל:
   2.1. אם אנחנו מדברים על מזגאוויר ברמה חודשית: אז חום בחודשי הקיץ.
   2.2. אם אנחנו מדברים על מכירות ברשת במהלך השבוע, ניתן לראות דפוסים חוזרים: למשל: עלייה במכירות בתחילת שבוע(ביחס לסופש שהחנות הייתה סגורה), ירידה באמצע השבוע והפסקת מכירות בסופש.



3. שילוב של השתיים ביחד: אנו יכולים לראות גרף במגמת עלייה ועם פיקים החוזרים על עצמם לאורך אינטרוולים שווים.
למשל:
חנות אשר המכירות בה מתנהלות כפי שמצויין מעלה אך בראייה של מספר חודשים ניתן לראות כי החנות ריווחית.



4. White Noise: רעש לבן, קיימות סדרות שהן כל כולן רעש לבן, כלומר לא ניתן למצוא אצלהם דפוסי התנהגות ולכן, אין כל כך מה לעשות איתם.



כמובן שסדרה יכולה להיות ללא טרנד ולא עונתית ועדיין כזאת שאפשר לחזות.

5. Auto Correlated: סדרה יכולה להיות בעלת אוטו-קורלציה(או מספר כאלו), כלומר: ערך הסדרה בנקודת זמן מסויימת תלוי בערך או ערכי הסדרה בנקודות זמן קודמות. למעשה, סדרה יכולה להיות בעלת כמה אוטוקורלציות ובינהן, שינוים הלא ניתנים לחיזוי.
במילים פשוטות: Auto Correlation היא ממד להשפעת נקודות בסט על נקודות מאוחרות יותר.

רוב הסדרות בעולם האמיתי מורכבות משילוב של כל הדברים יחדיו.

חיזוי של time series יכול להיות טריקי, בסופו של דבר את התבניות אפשר ללמוד אבל את הרעש והקפיצות קשה עד בלתי אפשרי לחזות.

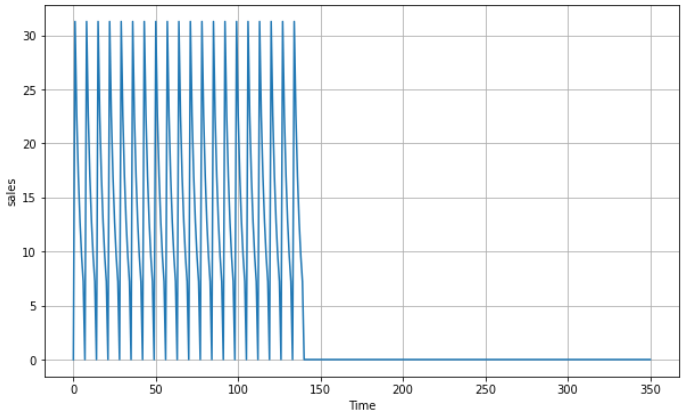
בנוסף, יכול לקרות אירוע שלא צפינו שישנה את כל ההתנהגות של הסדרה כמו למשל: במקרה של מניות השקה של טכנולוגיה חדשה או כישלון של טכנולוגיה צפויה.

בקורונה: יציאה מהסגר או התפרצות מחודשת עקב חג.

וכאן הנקודה השישית נכנסת לסיפור:

6. <u>סוטצינאריות:</u> סדרות כאלו הן סדרות שאין בהן שינוי בממוצע או בשונות, כלומר: אינן תלויות במרכיב הזמן. במקרה שהוצג מעלה, כאשר קורה אירוע שלא צפינו ומשנה את כל ההתנהגות, המודל אינו-סוטצינארי ודבר זה גורם לבעיות רבות בחיזוי.
בכדי לטפל במודלים כאלו ניתן לבצע מספר פעולות:
- בעיקר במודלים ממשפחת ARMA, להפוך את הסדרה לסוטצינארית או לקרב אותה לכזאת על ידי אינטגרטיביות או על ידי שיטות אחרות כמו log scale.
- במקרה של כל המודלים (עד כמה שזה יכול להשמע מוזר) לאמן את המודל על חלק נורא קטן, לאחר השינוי, לאחר האירוע המכונן.

דוגמא לסדרה לא סוטצינארית:



בשורה התחתונה, ניתן להסתכל על סדרה עיתית כסכום או מכפלת סך מרכיביה (מרכיבים אשר נמצאים מעלה).
חיזוי:
גם כאן בכדי לבדוק את איכות המודל עלינו לחלק אותו ל-2(3) סטים שונים:train,(dev),test אך מכיוון שאנחנו מדברים על בעיות חיזוי זמן, ישנם כמה דגשים שחשוב לשים לב אליהם(במיוחד כאשר אנו באים לעבוד בשיטות הקלאסיות יותר):
- עלינו לחלק את הסטים כך שלא תהיה פגיעה ברצף הזמן.
- אם קיימת עונתיות במידע, עלינו לדאוג כי כל סט יכיל עונות שלמות ואת כל העונות הקיימות במידע. כלומר: אם יש לנו עונתיות של ימים, לא נרצה שהסט יכיל חלקי ימים בשביל לא ליצור הטייה לכיוון שעות (למשל: בחיזוי של טמפרוטרה, לא נרצה שיופיעו לנו יותר שעות לילה מיום).

סקירת מודלים וכלים

קיימות שלושה קטגרויות של כלים ומודלים לחיזוי סדרות עתיות אך לפני שנכיר אותן , תחילה עלינו להבין מה הוא מרחב ההיפוטזה שלנו (Hypothesis Space), מה הוא "מרחב המודלים האפשריים".

מרחב ההיפוטזה(Hypothesis Space)

בסופו של דבר בבואנו לבצע חיזוי עתידי אנחנו באים לחזות משתנה רציף ולכן, המרחב שלנו הוא מרחב הרגרסיה.

אך להגיד שהמרחב הוא מרחב הרגרסיה זה לא מספיק, איננו יכולים למדל את כל המרחב הזה , מידול של מרחב כל כך גדול (ובעיקר כאשר אין הרבה מידע זמין) עלול לגרום לשונות או לפיזור רחב בין תוצאות המודלים השונים (יש סיבה לכך שמודלי למידה עמוקה עובדים טוב יותר עם כמה שיותר מידע).

במילים אחרות: התאמת יתר (Overfitting).

מצד שני, אם נצמצם יותר מדי את המרחב שלנו אנחנו נקבל מרחב שבו כמות המודלים האפשריים מוגבלת למודלים פשוטים מדי עבור הבעיה ולכן, אנו עלולים למצוא את עצמנו במצב של הטייה (Bias) גבוהה מדי או למעשה, Underfitting.

אם כך, מרחב ההיפוטזה שלנו הוא כמעט לחלוטין מרחב הרגרסיה הלינארית (במודלי הלמידה העמוקה אנחנו נראה כי אומנם השכבה האחרונה תהיה לינארית אך שכבות ביניים יהיו לא לינאריות וזאת בכדי לנסות ולמצוא קשרים לא לינארים).

סוגי מודלים

את שיטות החיזוי ניתן לחלק לקטגוריות או משפחות בשתי צורות עקריות:

- מידול של סדרה יחידהבעלת משתנה יחיד (Univariate) או סדרות מרובות משתנים (Multi Variate) או במילים אחרות האם החיזוי מתבצע עבור משתנה מטרה יחיד או לא.
- ישנן כמה משפחות של כלים שניתן להשתמש בהן: כלים סטטיסטים קלאסיים, אלגוריתמי למידת מכונה קלאסיים (ML) ואלגוריתמי למידה עמוקה.
  אנו נרחיב בנושא הכלים הקלאסיים, אך כסקירה קצרה:
  בתחום למידת המכונה הקלאסית אנו יכולים להשתמש בכלי רגרסיה כמו: SVR, Polynomial regression , עצי החלטה , catboost וכו'.

כלים סטטיסטים קלאסיים

קיימים כלים סטטיסטים קלאסיים (שאף לא נחשבים חלק מסט מודלי הלמידה) אשר נחשבים ככלים אשר יכולים לעשות את עבודת החיזוי נאמנה ואף בצורה איכותית יותר ממודלי הלמידה בקטגוריה זו ניתן למנות כלים רבים אנחנו נתמקד בעיקרים שבהם.

משפחת ARIMA

משפחת ARIMA היא משפחה של מודלים אשר מנסה למדל את ה-lag (צעד) הבא ברצף בעזרת שילוב של אחד או יותר מהגורמים הבאים:

Autoregression(AR)

החלק הראשון של שיטת ARIMA (ולמעשה, שיטה בפני עצמה) היא שיטת ה- Autoregression(אוטורגרסי) בשיטה זו החיזוי תלוי אך ורק בתצפיות העבר(בתוספת טעות והטייה).

למעשה מדובר באלגוריתם מסוג Univariate.

המשוואה של שיטה זו:

$$Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + .. + \beta_p Y_{t-p} + \epsilon_1$$

הפרמטר היחיד שניתן לכוונן בשיטה הוא p ומשמעותו הוא כמות הלגים בעבר שיש להתייחס אליהם כאשר באים לבצע חיזוי.

אחת מהדרכים שנעזרים בהן בכדי לדעת איזה p הכי מתאים לבעיה היא על ידי חישוב partial autocorrelation או בקצרה PACF. למעשה מדובר בפונקציית קורלציה עצמית וישירה אשר בודקת את השפעת lag-ים מהעבר על lag-ים מאוחרים יותר או למעשה את הקשר שבין הסדרה לבין עצמה וזאת תוך כדי הורדת השפעות ביניים (למשל: אם המחיר של היום תלוי במחיר של אתמול ושל שלשום והמחיר של אתמול תלוי בשל שלשום, בפונקצייה זו לא נתייחס לאתמול, אלא נתייחס אך ורק לשלשום).

ניתן לספור את כמות ה-lag-ים מתחילת הסדרה ועד הפעם הראשונה שהם נכנסים לתוך הרווח סמך (confidence interval (CI)), ה-lag-ים הללו הם המשפיעים ביותר על הסדרה ולכן, הכמות שלהם יכולה להיות ערך טוב עבור p.

<u>מספר דגשים</u> AR לכשעצמו טוב לחיזוי של סדרות חסרות טרנד ועונתיות.

Moving Average(MA)

החלק השלישי של שיטת ARIMA וגם שיטה בפני עצמה היא שיטת הממוצע הנע של טעויות השארית (residual errors), גם שיטה זו היא שיטה לינארית אשר מנסה לחזות את הlag או הצעד הבא ברצף על ידי פונקציה לינארית, ההבדל כאן הוא הניסיון למדל בעזרת ממוצע נע על השאריות.

גם שיטה זו בתור שיטה עצמאית איננה טובה בחיזוי (לרוב) אך היא כן טובה במשימה של החלקה והעלמת רעשים.

למעשה, בעזרתה ניתן להעלים רעשים, לבצע חיזוי על סדרה נקייה מרעשים ולאחר מכן להחזיר את הערכים אשר הוחסרו ולבצע שוב סינון מחודש של רעשים.

<u>נשים לב כי:</u> שימוש בחלון ממורכז עבור חישוב ממוצע נע יביא לנו דיוק יותר טוב משימוש בחלון מטייל, הבעיה שאין לנו אפשרות להשתמש בחלון ממרוכז עבור ערכים בהווה שכן, אין לנו את ערכי העתיד.

ולכן, עבור ערכי עבר אנחנו יכולים להשתמש בממוצע נע עם חלון ממורכז (לצורך העניין ערך מלפני שנה פחות כמה ימים לערך עבור שנה אחרי) ועבור ערכי הווה אנחנו יכולים להשתמש בערך נע בעל חלון מטייל.

ולכן למעשה במודל MA לאחר דפרנציציה התחזית שלנו תהיה מורכבת מחלון ממוצע נע מטייל על הסדרה שעברה דפרנציציה בתוספת חלון ממוצע נע ממורכז על מידע מן העבר.

המשוואה של שיטה זו:

$$Y_t = \alpha + \epsilon_t + \phi_1\epsilon_{t-1} + \phi_2\epsilon_{t-2} + .. + \phi_q\epsilon_{t-q}$$

הפרמטר של השיטה הוא q אותו אנו יכולים למצוא בעזרת פונקציית Autocorrelation- , פונקציה זו מאפשרת לנו לראות את ההשפעה של הסדרה על עצמה כולל השפעות לא ישירות.

פרמטר זה למעשה אומר לנו, על כמה lag-ים צריך להסתכל אחורה במידול.

גם שיטה זו בשימוש עצמי טובה אך ורק עבור סדרות ללא טרנד ועונתיות וכמו כן, סדרות שהן univariate.

Autoregression Moving Average (ARMA)

שיטה המשלבת את שתי השיטות מעלה, למעשה מדובר במודל לינארי אשר מנסה לחזות או למדל את הצעד הבא בעזרת פונקציה לינארית של תחזיות ושאריות עבר.

כמובן ש:

- עבור שיטה זו צריך לספק גם את p וגם את q.
- שיטה זו בשימוש עצמי טובה אך ורק עבור סדרות ללא טרנד ועונתיות וכמו כן, סדרון שהן univariate.

Autoregression Integrated Moving Average (ARIMA)

מודל ARMA ומרכיביו סובלים משתי בעיות עקריות: חוסר יכולת לעבוד עם סדרה בעלת טרנד , סדרה לא סוטצינאריות (כלומר בעלת שונות וממוצע משתנים) וחוסר יכולת לעבוד (או לפחות בצורה מעולה) עם סדרה בעלת עונתיות.

I ב-Arima מאפשר לפתור את הבעיה של הטרנד/ חוסר סוטצינאריות על ידי ביצוע דפרנציציה

(differencing) או למעשה החסרת לאגים מהעבר מלאגים נוכחיים.

צריך להזהר כאשר אנו עושים דפרציציה שכן, אנו יכולים להביא את הסדרה למצב של החסרת-יתר. סדרה כזאת, יכולה להיות סטצוינארית אך החסרת-יתר יכולה להשפיע על שאר הפרמטרים של המודל ולמעשה, אנו עלולים להפסיד מידע.

למען האמת, זאת היא לא השיטה היחידה להשיג סטוציונריות קיימות שיטות רבות אחרות כמו למשל log scale ולעיתים בכדי לא לאבד מידע, נעדיף להשתמש בהן.

דיברנו על איך לגרום לסדרה להיות סטוציונרית אך לא דיברנו על איך לגלות אם היא כזאת או לא. ובכן, קיימות מספר שיטות אך אנחנו נשתמש במבחן Augmented Dickey-Fuller (ADF), במבחן זה השארת ה-0 שלנו (null hypothesis) היא כי הסדרה איננה סטוצינורית.

למעשה, מבחן ADF בודק האם השינוי במשתנה המטרה (Y) יכול להיות מוסבר על ידי ערך lag שלו (על ידי תצפית קודמת שלו), בתוספת של טרנד לינארי.

בהנחה וקיים טרנד לינארי אבל המשתנה לא יכול להיות מוסבר בעזרת ערך עבר שלו אז אנו מתמודדים עם סדרה לא סטוציונארית.

מבחינתנו, אנו סך הכל צריכים לתרגם את תוצאות המבחן . במידה ואנחנו מקבלים test statistic הקטן מהערכים הקריטים ו- p-value ברמת מובהקות של 95 אחוז (כלומר, קטן מ-0.05) אנו יכולים לדחות את השארת ה-0 ולאמר בבטחה כי הסדרה סטוציונארית.

למעלה, נתנו ניתוח דיי פשטני עבור פונקציות ה-Autocorrelation וה-Partial Autocorrelation האמת היא שניתוח יותר מעמיק שלהם יכול לתת לנו מספר כיוונים.

תחילה חשוב להבין שכאשר משתמשים במודל ARIMA לא תמיד משתמשים גם בMA וגם בAR . לעיתים , שימוש בשניהם יחדיו יכול להוות overkill ולכן, לפעמים נבדוק את האפשרות לעבוד רק עם אחד מהם, וזאת לפי כמות המידע שהוא מביא לנו.

אנחנו ננסה להשתמש רק ב-AR כאשר:

- פונקציית ה-Autocorrelation מראה כי הקורלציה בירידה אל עבר האפס.
- פונקציית ה-Autocorrelation של הסדרה במצבה הסוטוצינארי חיובית ב-lag הראשון.
- פונקציית ה-Partial Autocorrelation מגיעה במהירות לאפס.

לעומת זאת, אנו ננסה להשתמש רק ב-MA כאשר:

- ה-lag הראשון הוא שלילי בפונקציית ה-Autocorrelation , כלומר- ישנה קורלציה שלילית בלאג הראשון.
- פונקציית Autocorrelation אשר יורדת בצורה חדה לאחר מספר מועט של lag-ים.
- פונקציית ה-Partial Autocorrelation יורדת בהדרגתיות.

אם נראה כי התוצאות של אחד החלקים בנפרד לא טובות מספיק או לחילופין אף אחד מהתנאים הללו לא מתקיים, ננסה לשלב את שתי החלקים לפי השיטה הפשטנית אשר נדונה בעמוד הקודם.

אופצייה נוספת היא אופצייה של grid search באופצייה זו ננסה לחפש את כל השילובים האפשריים (עד גבול עליון מסויים) של p q ו-d.

הרחבות למשפחת ARIMA

Seasonal Autoregressive Integrated Moving-Average (SARIMA)

מודל זה מוסיף התייחסות למרכיב העונתיות, למעשה במודל זה אנו מנסים לחזות את הנקודת זמן הבאה ברצף כקומבינציה לינארית של המרכיבים שדנו עליהם במודל ARIMA ובנוסף אליהם אותם המשתנים רק עבור מרכיב העונתיות.

למעשה, המודל הזה מאפשר לנו למדל את הסדרה שלנו ברמה העונתית.

שיטה זו טובה אך ורק עבור סדרות שהן univariate.

(Seasonal) Autoregressive Integrated Moving-Average with Exogenous Regressors (S/ARIMAX)

אנו יכולים לחשוב על מודל זה כמודל אשר משתמש בערכים נוספים/ תצפיות נוספות אשר נמדדו באותו הזמן (בסדרות מקבילות).

המודל מקבל את התצפיות הללו כפי שהן ולא מבצע עליהן את כל התהליך כפי שהוא מבצע על הסדרה שאנו מנסים לחזות.

שיטה זו טובה עבור סדרות שהן univariate בעלות טרנד, עונתיות וסדרות מקבילות/ תצפיות אשר הסדרה נמצאת איתם בקורלציה.

Vector Autoregression(VAR)

במודל מסוג זה אנחנו מנסים לחזות את הצעד הבא בכל סדרת זמן. למעשה, מדובר בהכללה של מודל AR הרגיל בכדי שיוכל לעבוד על multivariate.

קיימות הרחבות דומות עבור שאר מודלי המשפחה.

משפחת Exponential Smoothing (es)

Simple Exponential Smoothing (SES)

שיטת SES היא שיטה אשר ממדלת או מנסה לחזות את נקודת הזמן הבאה על ידי ידי פונקציה לינארית ממושקלת מעריכית של תצפיות עבר.

השיטה מתאימה לסדרה חסרת טרנד ועונתיות , כמו כן, ל-univariate .

Holt Winter's Exponential Smoothing (HWES)

שיטה זו דומה לשיטה הקודמת רק שהיא לוקחת בחשבון טרנדים ועונתיות.

משפחת ARCH

הבעיה עם משפחת ARIMA היא שאינה עובדת בצורה טובה עם תנודתיות, עם סדרות שהשונות ו/או הממוצע שלהם משתנה לאורך זמן.

פתרון לבעיה זו נמצא בהרחבות של ARIMA בדמות משפחת ARCH.

מודל Autoregressive Conditional Heteroskedasticity(ARCH)

שיטה זו מאפשרת לנו למדל את השינוי בפיזור של הסדרה (שונות) כפונקציה של שגיאות השארית.
למעשה, בעזרת שיטה זו ניתן לחזות את השינוי בשונות לאורך זמן.

מספר דגשים:

- גם כאן אנו צריכים לספק את מספר ה-lag-ים שעל המודל להתייחס אליהם.
- הנחת העבודה של מודל זה היא כי מדובר בסדרה סוצטיונרית, ללא טרנד או עונתיות.
- המודל לכשעצמו לא מועיל יתר על המידה, עלינו להשתמש בו לאחר מידול קודם עם מודל ARMA לדוגמא.
- קיימות הרחבות והכללות של המודל כמו לדוגמא GARCH.

חלק שני: מחשבות על העבר

סביבות עבודה

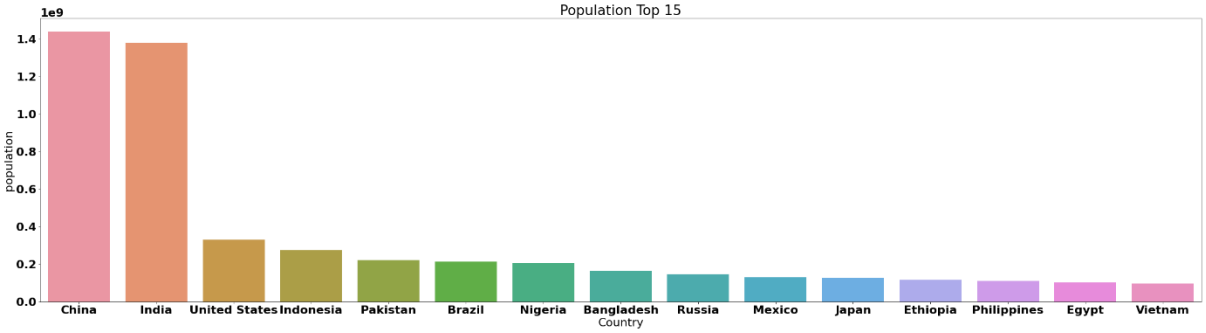
בחרנו לעבוד עם שפת פייתון, פייתון היא שפת תכנות לשימוש כללי ולשימוש בתחום הסטטיסטי ולמידת המכונה, לפייתון יש מגוון רחב של כלים וחבילות המתאימות לתחום הנ"ל.

EDA בסיסי

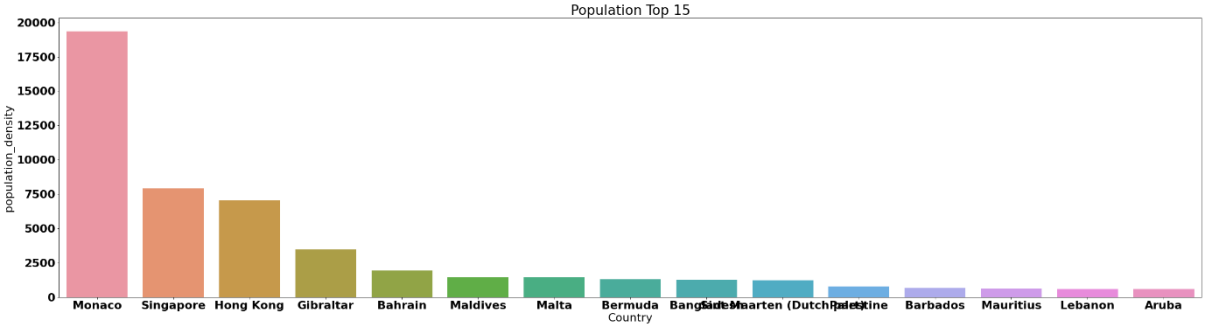
ביצענו EDA בסיסי אשר כלל:

עבור העולם כולו:

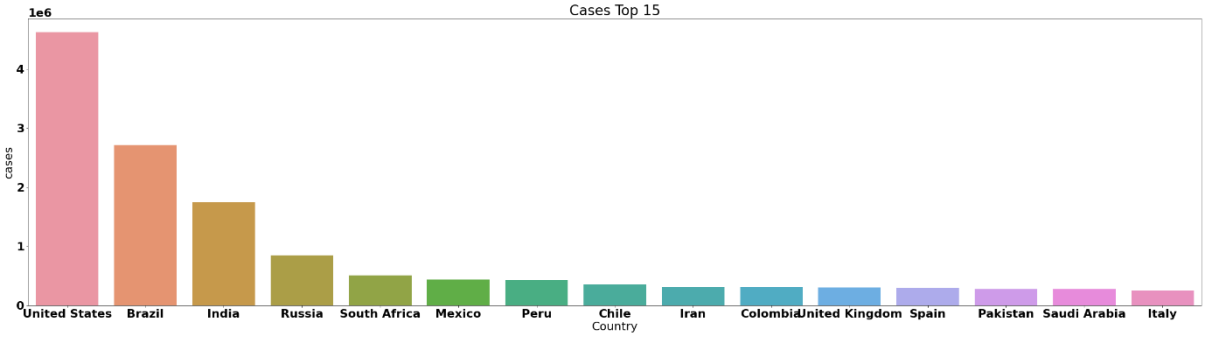
- 15 המדינות בעלות הכי הרבה אוכלוסיה:



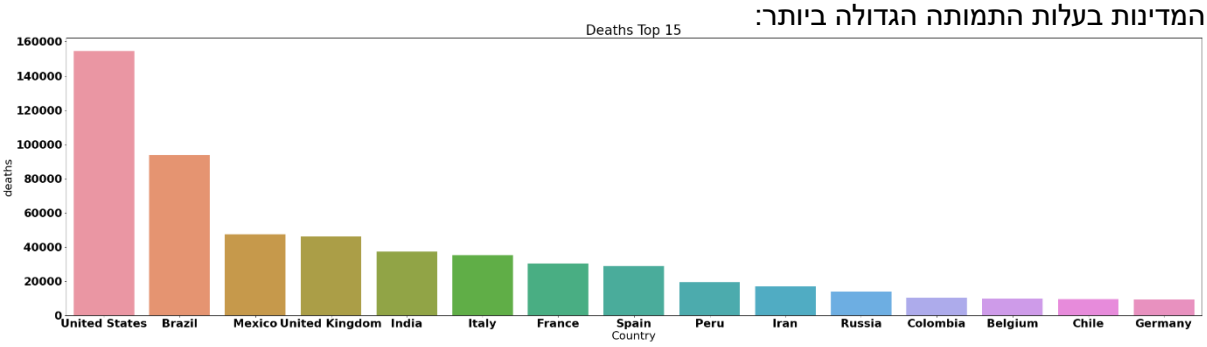
- המדינות הצפופות ביותר בעולם:



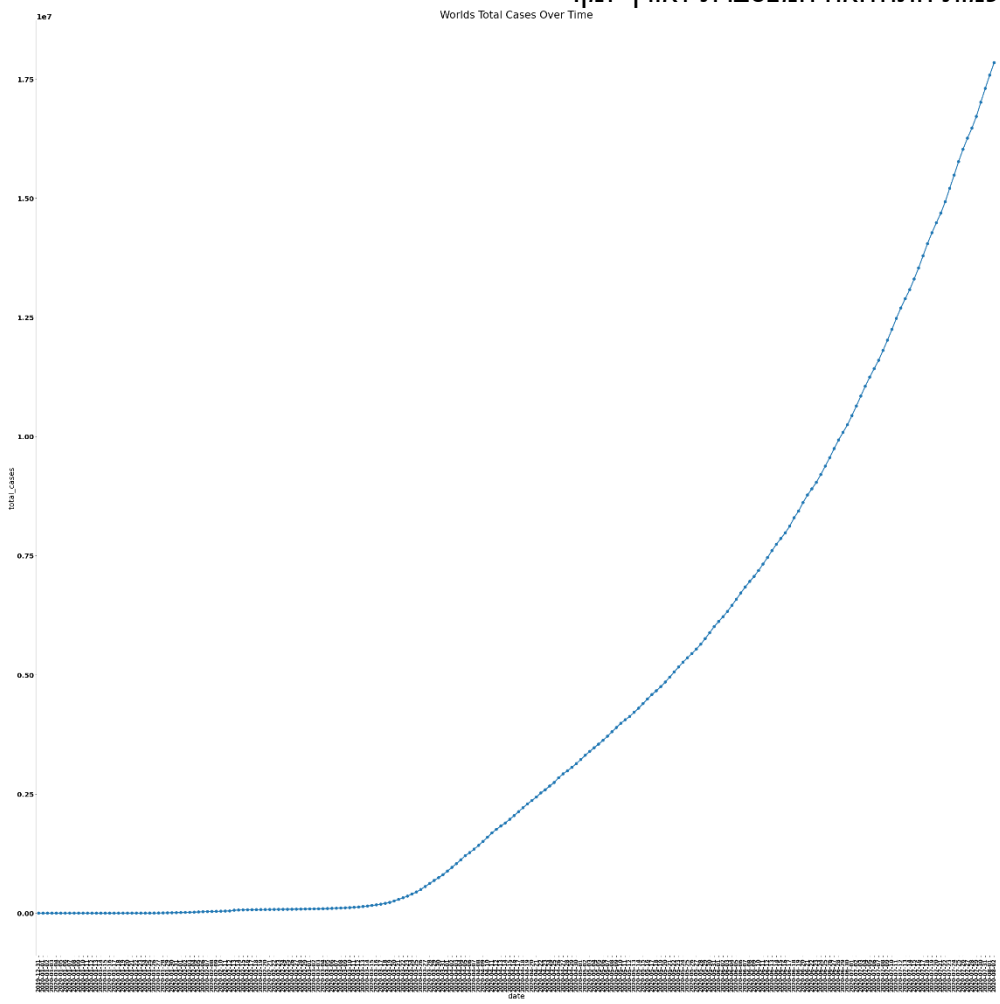
- המדינות בעלות מקרי ההדבקה הרבים ביותר:



- המדינות בעלות התמותה הגדולה ביותר:



- כמות התחלואה המצטברת לאורך זמן:



- קורלצייה בין כל משתני הסט:

total_cases_per_million	total_deaths_per_million	population_density	median_age	hospital_beds_per_thousand	life_expectancy	total_tests_per_thousand	handwashing_facilities	
total_cases_per_million	1.000000	0.821825	0.090319	0.238200	0.203663	0.302779	0.854329	0.154884
total_deaths_per_million	0.821825	1.000000	0.062610	0.243126	0.178588	0.250525	0.752286	0.137977
population_density	0.090319	0.062610	1.000000	0.201351	0.073791	0.284515	-0.015966	0.094684
median_age	0.238200	0.243126	0.201351	1.000000	0.762060	0.849199	0.406706	0.796859
hospital_beds_per_thousand	0.203663	0.178588	0.073791	0.762060	1.000000	0.556276	0.350091	0.534031
life_expectancy	0.302779	0.250525	0.284515	0.849199	0.556276	1.000000	0.441807	0.835157
total_tests_per_thousand	0.854329	0.752286	-0.015966	0.406706	0.350091	0.441807	1.000000	0.235523
handwashing_facilities	0.154884	0.137977	0.094684	0.796859	0.534031	0.835157	0.235523	1.000000

כמה הבחנות:

- ✓ ניתן לראות קורלצייה חיובית חזקה בין התחלואה (מנורמל למיליון איש) לבין התמותה(מנורמל למיליון איש).
- ✓ ניתן לראות קורלצייה חיובית חזקה בין התחלואה והתמותה (מנורמל למיליון איש) לבין כמות הבדיקות (מנורמל לאלף).
- ✓ ניתן לראות קורלצייה חיבות חזקה בין כמות מקומות לשטיפת ידיים לבין תוחלת חיים, חציון הגיל.

עד כמה שזה יכול להפתיע:

- ✓ לא ניתן לראות קורלציה חזקה בין תוחלת החיים במדינה לבין המחלה.
- ✓ לא ניתן לראות קורלציה חזקה בין גיל חציון לבין התחלואה והמוות.
- ✓ לא ניתן לראות קורלציה חזקה בין התמותה לבין כמות המיתות במדינה.

חשוב לשים לב: קורלצייה לאו דווקא מצביעה על סיבתיות.

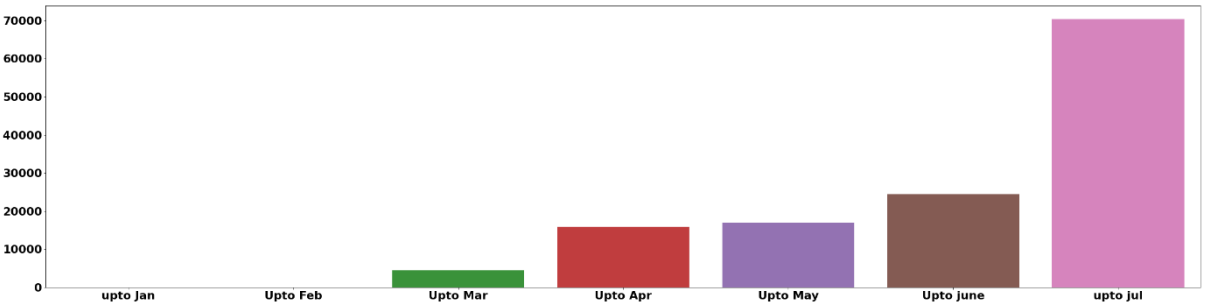
עבור ישראל (ומדינות נוספות, אך כאן נתייחס רק לישראל):

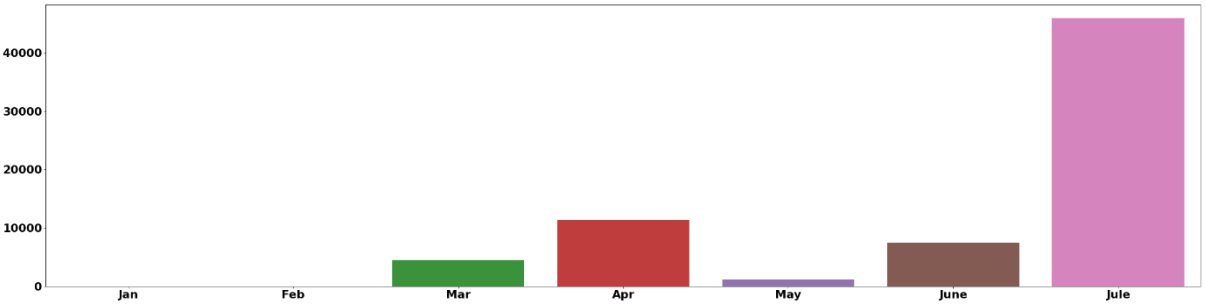
- סכימת חמשת המספרים:

	total_cases	new_cases	total_deaths	new_deaths
count	213.000000	213.000000	213.000000	213.000000
mean	14224.798122	339.356808	160.319249	2.474178
std	17002.747580	546.716877	159.111897	3.516208
min	0.000000	0.000000	0.000000	0.000000
25%	1.000000	0.000000	0.000000	0.000000
50%	12982.000000	79.000000	151.000000	1.000000
75%	18268.000000	412.000000	299.000000	4.000000
max	72283.000000	2502.000000	527.000000	15.000000

ניתן לראות כי אין מדובר בהתפלגות נורמלית.

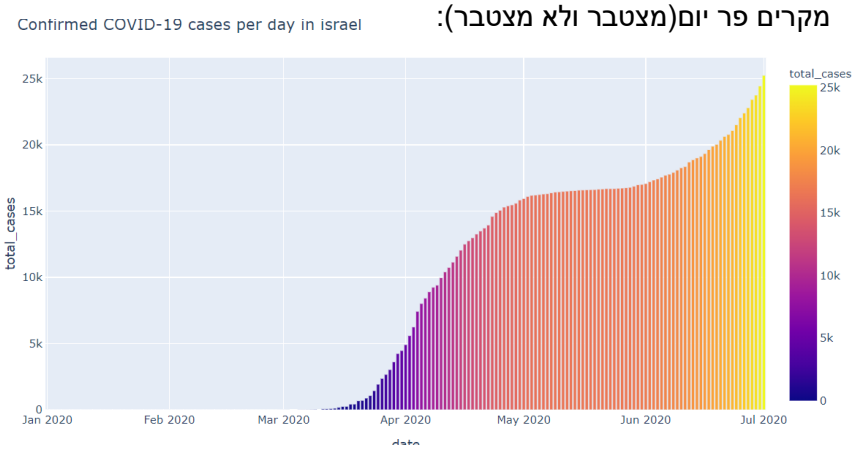
- התקדמות חודשית של המחלה:

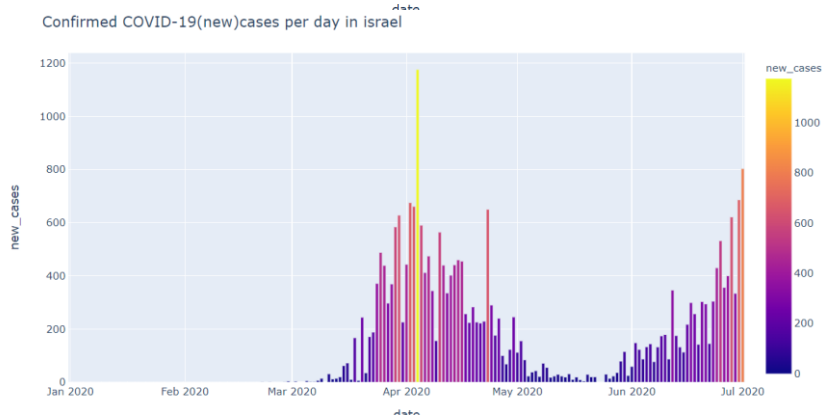




ניתן לראות כי עד מרץ, כמעט ולא היו מקרים בארץ.

- מקרים פר יום(מצטבר ולא מצטבר):





את הקפיצות הרציניות ניתן להסביר על ידי שינוי בכמות הבדיקות היומי, שינוי במדיניות.

מודלים שהשתמשנו בהם

כאמור לבעיות מסוג זה ניתן לגשת בשלוש גישות שונות:

1. הגישה הסטטיסטית הקלאסית.
2. גישת למידת המכונה הקלאסית.
3. גישת הלמידה העמוקה.

החלטנו לבחור במודלים הבאים:

1. עבור הגישה הסטטיסטית בחרנו במודל ARIMA.
מצד אחד, מודל זה נחשב למודל איכותי אשר חוזה את העתיד בהצלחה ולעיתים רבות אף בהצלחה רבה על פני מודלי למידת המכונה (הקלאסים והעמוקים).
מצד שני, בחירה זו בעייתית שכן, ל-ARIMA יש הנחה חזקה של חוסר תלות בזמן וגם כמות הנדבקים היומית תלויה בכמות הבדיקות המשתנה.
<u>ניתן לפתור את הבעיה הראשונה</u> על ידי שימוש במודל אחר כמו למשל HWES או GRACH. <u>את הבעיה השנייה</u> ניתן לפתור על ידי ההרחבה של ARIMA ל-ARIMAX.
2. עבור גישת הלמידה הקלאסית בחרנו ב: רגרסיה לינארית פולינומיאלית(Linear Polynomial Regression).
<u>אפשרות נוספת</u> היא שימוש ב-SVR עם קרנל פולינומיאלי.
3. עבור גישת הלמידה העמוקה בחרנו ב: רשתות LSTM בשל הזיכרון שלהם והיכולת לקלוט תלויות גלובאליות וברשתות CNN בשל הקלילות שלהן והיכולת לקלוט תלויות לוקאליות.
<u>פתרון נוסף:</u> שימוש ב – TIME VECTOR.

הציפייה לחיזוי טוב לא הייתה גבוהה, בעיקר בגישה של הלמידה העמוקה, מהסיבה הפשוטה שאין לנו מספיק מידע. כידוע, מודלי למידה עמוקה הם מודלים בעלי מרחב היפוטזי גדול , זהו יתרון אך תכונה זו גם דורשת מידע רב לאימון המודל בכדי שהוא יוכל להכליל בצורה טובה.

נשים לב כי : בשל מחסור במידע, ביצענו חלוקה אך ורק ל-train/test.

חלק שלישי:תוצאות ומחשבות על העתיד

תוצאות נבחרות

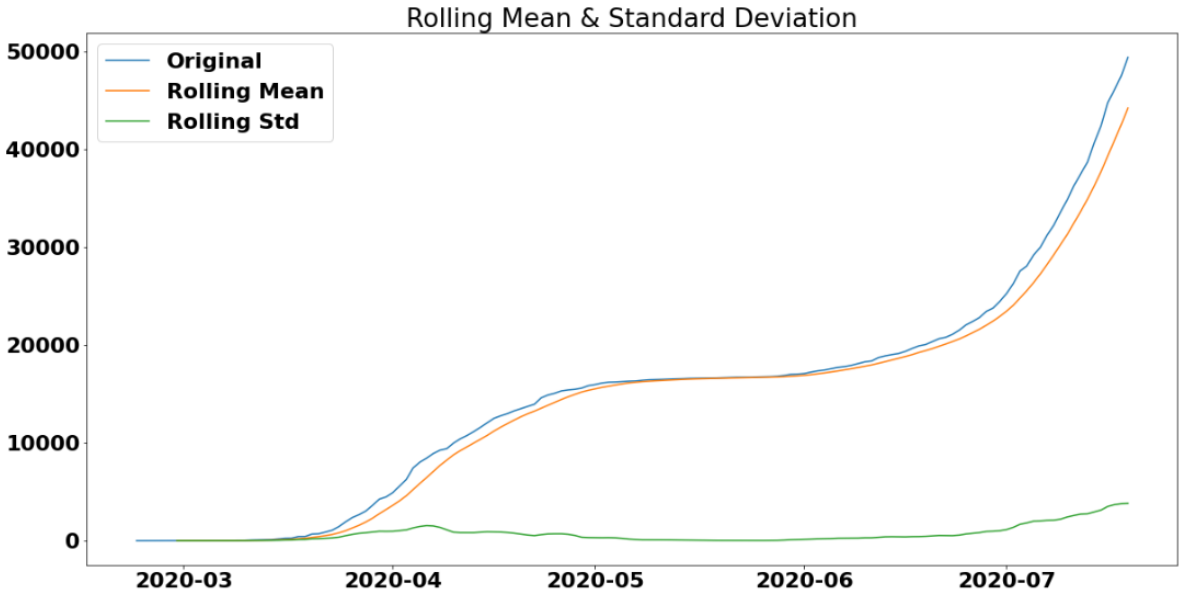
ARIMA

מידול

כאמור בשביל למדל את ARIMA אנו צריכים לספק לו את מספר ה-lagים שצריך לקחת בחשבון בכל אחד מהחלקים.

תחילה, עלינו להבין האם הסדרה סטציונרית, עשינו זאת על ידי מבחן ADF (כפי שהוסבר למעלה) ווויזואלית.

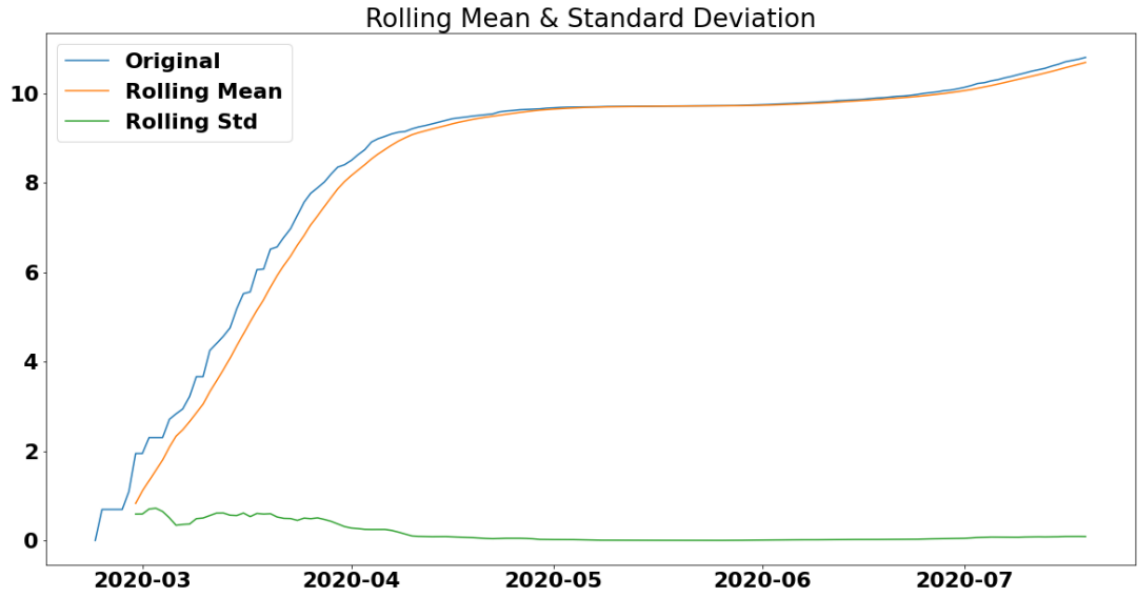
```
Test Statistic                  1.271369
p-value                         0.996444
Lags Used                      14.000000
Number of Observations Used   133.000000
Critical Value 1%              -3.480500
Critical Value 5%              -2.883528
Critical Value 10%             -2.578496
dtype: float64
```



מבחינה וויזואלית ניתן לראות כי הסדרה לא סטוציונארית. בנוסף, לפי המבחן ערך הבדיקה שלנו גדול משמעותית מכל הערכים הקריטיים וגם ה-pvalue שלנו גדול מ-0.05 . כלומר, אנו לא יכולים לסתור את הנחת האפס.

בכדי לא לפגוע בסדרה , החלטנו לבצע log scale, המבחן לאחר ה-log scale:

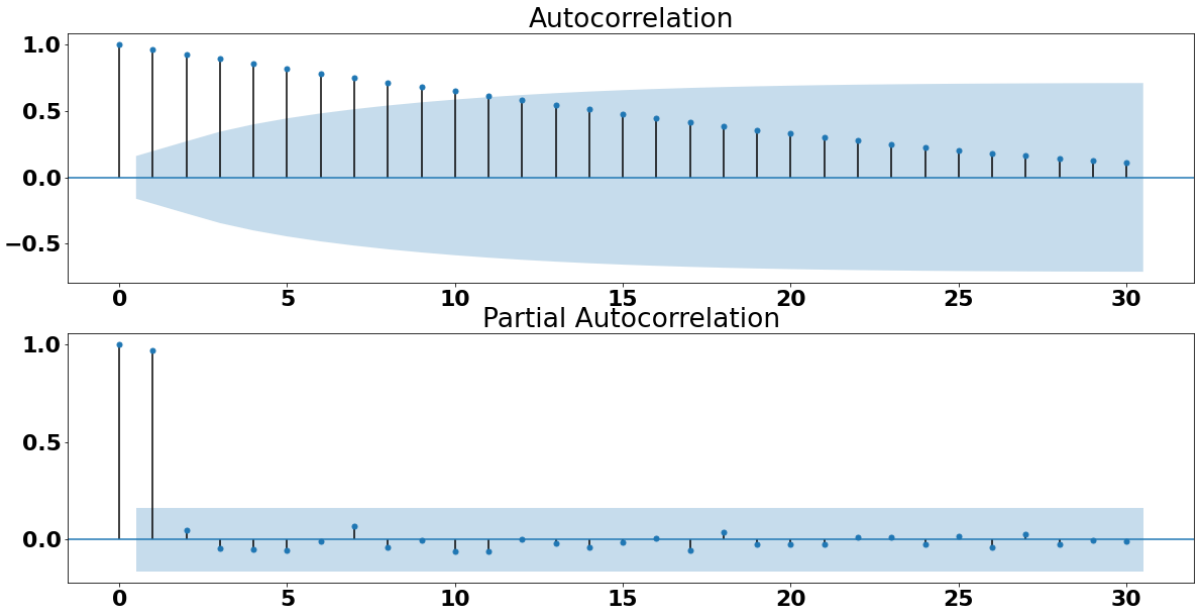
```
Test Statistic                 -4.044924
p-value                         0.001193
Lags Used                      14.000000
Number of Observations Used   133.000000
Critical Value 1%              -3.480500
Critical Value 5%              -2.883528
Critical Value 10%             -2.578496
dtype: float64
```



ניתן לראות כי אנו יכולים לדחות את השארת ה-0.

השלב הבא הוא להחליט מה יהיו הערכים של q ו-p כלומר לכמה lag-ים מודלי ה-AR וה-MA צריכים להתייחס . נשים לב כי: d שלנו הוא 0 שכן, בסקאלה הלוגריתמית הסדרה שלנו סטוציונארית.

את ערכי ה-q וה-p מצאנו בעזרת פונקציות האוטוקורלציה והאוטוקרלציה החלקית.

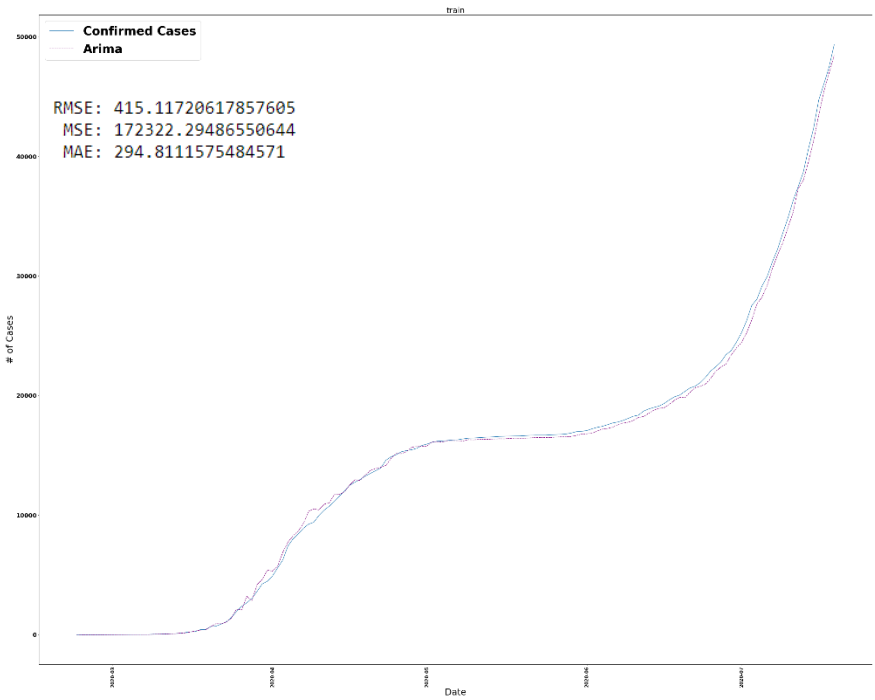


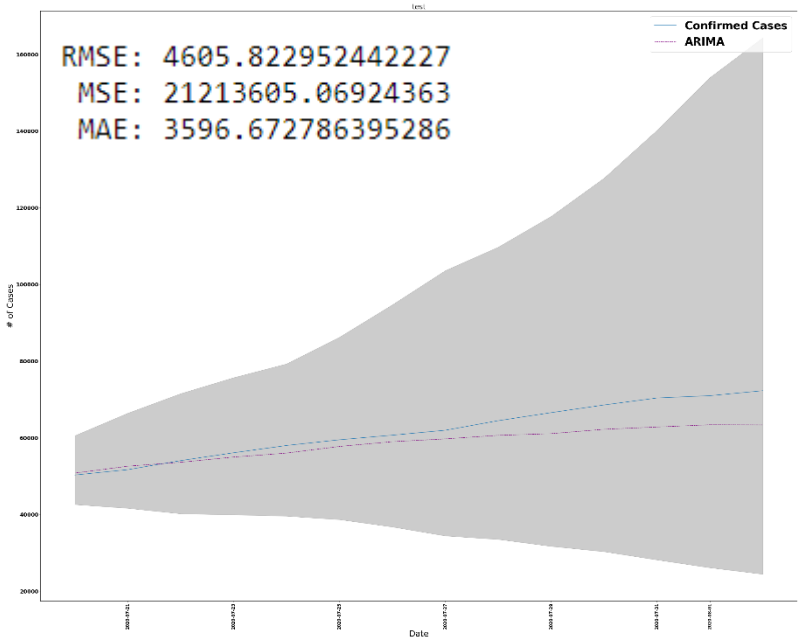
ניתן לראות כי:

- פונקציית ה-Autocorrelation מראה כי הקורלציה בירידה אל עבר האפס.
- פונקציית ה-Autocorrelation של הסדרה במצבה הסוטוצינארי חיובית ב-lag הראשון.
- פונקציית ה-Partial Autocorrelation מגיעה במהירות לאפס.

ולכן, לא השתמשנו ב-MA (Q=0) ועבור AR השתמשנו ב-11 (שכן, הם הכי משמעותיים).

תוצאות

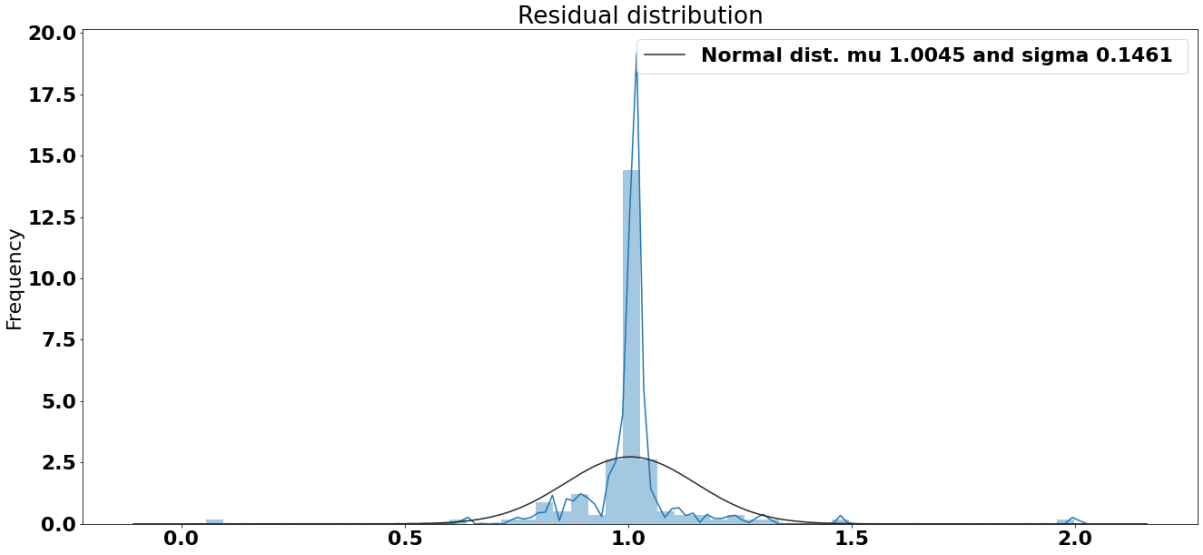




מסקנות:

- ניתן לראות כי ישנו overfitting .
- ניתן לראות כי שגיאת ה-RMSE גדולה יחסית בשתי הסטים (415 אנשים בסט האימון ו- 4605 בסט הבדיקה).
- ניתן לראות כי מגמת החיזוי נכונה , אך המודל שלנו חוזה תמיד כמות קטנה יותר.
- בדיקת נורמליות:

בבדיקה זו אנו לוקחים את התפלגות השגיאות (אם לדייק, השאריות) ובודקים עד כמה ההתפלגות קרובה להתפלגות נורמלית, הצפי או התקווה שלנו היא שהיא תהיה קרובה להתפלגות זו.



ניתן לראות כי ההתפלגות לא ממש נורמלית , לא בצורה ולא מבחינת הממוצע והפיזור. בנוסף, ניתן ניתן לראות כי יש לנו הטייה.

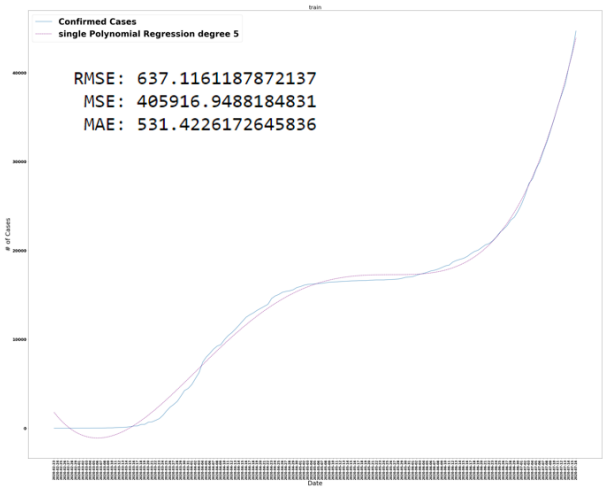
הערה: התייחסנו רק ל-RMSE (שכן, הוא מביא לנו את השגיאה בסקלה של החיזוי).

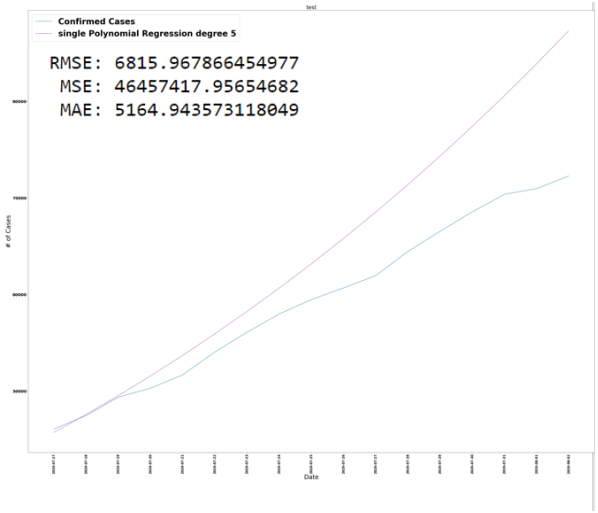
Polynomial linear regression

מידול

ניסינו למדל מספר מודלים כאשר בכל פעם אנו בונים מטריצת פיצרים עד מעלה אחרת, ניסינו מעלות בין 1-5.

המודל אשר הביא את התוצאה הכי טובה הוא מודל ממעלה 5:





מסקנות:

- יש לנו overfitting.
- יש לנו שגיאה גדולה , 637 אנשים בממוצע בסט האימון ו-6815 בסט הבדיקה.
- יש לנו איבוד מגמה בכל מיני שלבים.
- מודל ARIMA עבד טוב יותר.

גישת הלמידה העמוקה

מידול

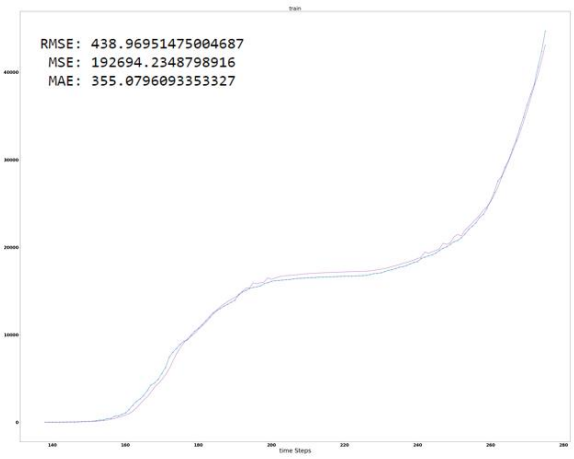
בגישה זו הכנו את המידע על ידי יצירת חלונות בגדלים 5,7,3 וניסיון לחזות את הצעד הבא עבור החלון.

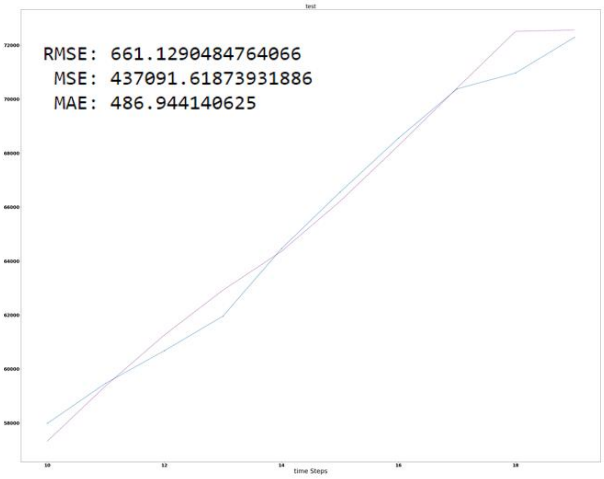
כלומר כל דוגמית x שלנו מורכבת מחלון בגודל מסויים ודוגמית y היא היום הבא שאנחנו מנסים לחזות.

ניסינו להשתמש במודל LSTM ובמודל CLSTM .

תוצאות

התוצאה הכי טובה הייתה עבור מודל LSTM עם חלון 7:





מסקנות:

1. ניתן לראות כי ישנו overfitting אם כי קטן יותר מבמודלים האחרים .
2. ניתן לראות כי שגיאת ה-RMSE גדולה יחסית בשתי הסטים (438 אנשים בסט האימון ו- 661 בסט הבדיקה).
3. ניתן לראות כי מגמת החיזוי אינה נכונה תמיד.

לסיכום, מבחינת המטריקה זהו המודל הטוב ביותר, מצד שני, מבחינת חיזוי מגמה היו לנו טובים יותר.

שיפורים אפשריים

מעבר לשיפורים אשר הוצאו בחלק של בחירת המודלים, ניתן לבצע את השיפורים הבאים:

1. נשים לב כי התייחסנו לסדרה כתלות בעצמה בלבד, לא התבוננו על גורמים חיצוניים,
יכול להיות שישנו קשר בין מקרי התחלואה לאחד או יותר מהגורמים הבאים: מזג-אוויר, סגר, צפיפות האכולוסיה, האם מדובר בעולם ראשון או לא וכו'.
וכמובן, שקיים קשר בין כמות הבדיקות לבין התחלואה.
ניתן לנסות להשתמש בהרחבה Exogenous Regressors בשביל זה.
2. ביצוע של cross-validation.
3. בכדי לקרב את המודל לזמן שאנחנו מנסים לחזות, לדאוג כל פעם לאמן אותו מחדש על סט הבדיקה, במצב זה אנחנו נמנע (או לפחות ננסה) את הכישלון של המודל שלנו בחיזוי של נקודות זמן אחרי אירוע מכונן. שכן, בסופו של דבר סט הבדיקה הוא הסט הקרוב ביותר לנקודת זמן הנוכחית.
4. בהמשך ל-3 , אימון המודל על תחזיות חדשות.
5. בהמשך ל-3 ו-4, עד כה ראינו שיטה הנקראת fixed partitioning כלומר חלוקה קשיחה של המידע, שיטה נוספת היא roll-forward partitioning שבה אנו מאמנים את המודל על חלק קטן מהמידע, ואז מנסים לחזות כל הזמן נקודות מהמידע הנמצא בסט dev , לאחר מכן מאמנים את המודל על המידע שחזה וכך הלאה.
6. שימוש ב- grid search לביצוע של hyper-params tuning (ביצענו את זה בחלק מהמודלים אך לא בכולם).
7. לנסות למדל בעזרת sigmoid , אם נניח לרגע שאין הדבקה חוזרת אז בסופו של דבר המגפה תפסק אם כולם ידבקו או לחילופין יחוסנו (חיסון העדר או חיסון במעבדה) ולכן, המגפה לא יכולה להתפשט לאורך זמן בצורה מעריכית, באיזשהו שלב היא תצטרך להאט ולהפסיק לתפס בצורה הזאת (לעבור את נקודת ההשתקפות) עד שהעקומה תתיישר לחלוטין.
ולכן, בהנחה כי קיימת חסינות נגד הדבקה חוזרת ו/או ימצא חיסון שניתן להתחסן בו פעם אחת ולסיים אם זה, מידול זה יכול להתאים.