נכתב על ידי [ניל צור](https://www.linkedin.com/in/neal-tsur/)

**תרחיש מקצועי**

**רקע**

**סקטור:** GovTech

**תיאור:** המחלקה האסטרטגית של עיריית ירושלים שכרה את שירותכם לצורך התמודדות מקומית עם התפרצות מחודשת של קורונה. המטרה שלהם היא ליצור יכולת זיהוי, חיזוי וניטור במצב של התפרצויות חדשות עקב זן עמיד לחיסון הנוכחי. לצורך כך, ברצונם לחזק ככל שניתן את השימוש של כל מרכז קהילתי בנתוני תחלואה.

במהלך שיחת העבודה עימם הועלו הכיוונים הבאים ככיוונים אשר עשויים לייצר ערך חברתי וכלכלי:

* לזהות עבור כל מרכז קהילתי נבחר אילו שכונות *מצומדות*(coupled) אליו - כלומר, בהינתן התפרצות בשכונה אחת, אילו שכונות נמצאות בסכנה להתפרצות גם אצלהן. המטרה אינה רק לזהות קורלציות אלא גם סיבתיות, כלומר שרשרת הדבקות ברמה השכונתית ו/או האיזור הסטטיסטי(שכונה מורכבת ממספר איזורים סטטיסטיים). הדבר יאפשר לבצע פעולות מנע והסברה נקודתיות בהמשך השרשרת.
  + האם הפתרון הוא בנאלי? - שכונות מדביקות שכונות לידן או שאינן בהכרח שכנות. הלקוח מתעניין בפתרונות שאינם מובנים מאליהם.
  + חיזוי תחלואה בשכונה אחת על בסיס התחלואה בשכונה אחרת
* חיזוי עבור כל שכונה את כמות התחלואה X ימים קדימה
  + עד כמה ימים קדימה התחזית אמינה
  + יש רצון שתבוצע גם אגרגציה של חיזוי תחלואה ברמה העירונית
  + לייצר מדד וודאות לכל תחזית(ויזואלי ו/או מספרי)
  + לזהות נקודות כשל למודל

**דגשים**

* לא כל דבר שנכתב פה אכן נדרש לביצוע ולא כל חבילה או מושג אכן תשמשו בפועל. נסו לחשוב מה אתם יכולים לפתור בזמן שיש לכם, אבל כזה שיביא ערך. קחו בחשבון שכל דבר יקח פי 4 זמן ממה שאתם חושבים שיקח.
* **הפתרון חייב להיות ויזואלי ונגיש**. **אחרת לא יעשה שימוש בכלי הזה**. המשתמשים הם מנהלים קהילתיים ופונקציונרים בעירייה. במידה והפיילוט יצליח, חזון החברה הוא לשכפלו לעיריות נוספות ברחבי העולם. מודל עסקי SaaS.
* שימו לב גם ליצור את הכלי, אך לחשוב איך תשכנעו אחרים שהוא עובד.
* על הפתרון להיות כזה הניתן לעדכון אחד ליום.
* הפתרון לא ידרוש התקנות מיוחדות במחשבי הקצה.
* נקו את הנתונים וודאו אותם! לפני שתתחילו ליישם את הפתרון שלכם.
* מומלץ לבצע EDA(Exploratory Data Analysis) בתחילת העבודה.
* מוזמנים להוסיף כל נתון שאתם מרגישים שחסר מעבר לנתונים שמוזכרים מטה(shape files, חבילות פייתון, נתוני למ"ס וכולי)
* רצוי לחשוב על מטריקה שדרכה ניתן יהיה לבחון את הפתרון שלכם. למרות זאת, חשוב להבין ולזכור איפה נמצא **הערך** של הפתרון. המטריקה עלולה לא לייצג כהלכה את הפתרון.
* עשוי גם לעניין אתכם לעשות שימוש [בtransfer entropy](https://en.wikipedia.org/wiki/Transfer_entropy)

**נתונים**

הקבצים נמצאים [בלינק הזה](https://drive.google.com/drive/folders/1CYfZAjeKFoxEZxhK6nxVah6yLt4u4fE1?usp=sharing)

* קובץ CSV נתוני תחלואה(חולים, מתים) פייר אזור סטטיסטי ופייר יום ממשרד הבריאות עבור כל האיזורים הסטטיסטיים במדינת ישראל
* קובץ PDF המתאר את קובץ הCSV של נתוני התחלואה באיזורים הסטטיסטיים
* קובץ xlsx עם המספר והשם(עברית ואנגלית) של כל איזור סטטיסטי בירושלים

**חבילות פייתון מומלצות**

חלק מהחבילות הללו בסיסיות לכל מדען נתונים. חלקן הוספתי כי עשויות רלוונטיות ספציפית לכם. במידה ואתם עובדים ב- R, Java או Matlab, תעשו חיפוש לפתרונות דומים, אולם הסטנדרט הוא פייתון.

* [Scikit-learn](https://sklearn.org/) - **(A.k.A sklearn)** חבילת Machine learning הבסיסית ביותר שמכילה עשרות אלגוריתמים ודאטה preprocessing
* [Pandas](https://pandas.pydata.org/) - מקל את העבודה עם טבלאות נתונים ומניפולציה על הדאטה
* [Numpy](https://numpy.org/) - לעיתים נעשה בה שימוש באופן ישיר לזירוז החישוב. נמצאת מתחת למנוע של pandas.
* [Scipy](https://docs.scipy.org/doc/scipy/reference/) - חבילת מדע ומתמטיקה סטנדרטית
* [Matplotlib](https://matplotlib.org/stable/contents.html) / [seaborn](https://seaborn.pydata.org/) - חבילות ויזואליזציה
* [scikit-optimize](https://scikit-optimize.github.io/stable/) - מאפשרת אופטימיזציה מבוססת סטטיסטיקה ביאסיאנית למודלים של sklearn.
* [pyvis](https://pyvis.readthedocs.io/en/latest/) - החלק משמעותי בחבילה הוא היכולת ליצור רשת(גרף) אינטרקטיבית בפורמט HTML - כזה שניתן לפתוח בכל דפדפן בלי שום צורך להתקין תוסף.
* [networkX](https://networkx.org/) - החבילה הבסיסית לניתוח ותצוגה של רשתות.
* [pysidny](https://pysindy.readthedocs.io/en/latest/) - חבילת רגרסיה המיועדת לטיפול במערכות דינמיות לא לינאריות(דומה קצת לsymbolic regression)