

מטלה מסכמת - קורס כלכלה בעולם הביג דאטה

מרצה: ד"ר רועי ששון

עוזר הוראה: אופיר בצר

מגישים: מתן המי (305610792), מאור קריטנברג (307930982)

תאריך: 17.8.21

דו"ח מחקר - עבור חיזוי נתוני התעבורה על בסיס מסד הנתונים של Apple

**חלק ראשון - הכנת מסד הנתונים**

ראשית, ביצענו את השינויים הנדרשים עבור מסד נתוני התעבורה של אפל. העברנו את מסד הנתונים לפורמט לונג, הסרנו ערכים חסרים וניקינו עמודות לא רלוונטיות עבור ביצוע התחזית. בשל העובדה כי להתפרצות הקורנה היתה השפעה מרחיקת לכת על מדיניות הממשלות ברחבי העולם וגררה צעדים מצמצמי תנועה כמו: סגירת מקומות בילוי ובתי ספר, סגרים טוטאליים על ערים ומדינות שונות וכו' המשתנים המסבירים הנוספים שבחרנו להוסיף למסד הנתונים התמקדו בנתוני תחלואה שונים הנאספים ברמה העולמית ונתונים המרכזים את צעדי המדיניות של המדינות השונות כתגובה להתפשטות מגפת הקורונה.

**את נתוני התחלואה** לקחנו ממסד נתונים של ארגון "Our World In Data" - ארגון המנגיש בצורה רחבה מסדי נתונים שונים לציבור הרחב. לארגון יש מסד נתונים המרכז נתוני תחלואה של מגפת הקורונה המתעדכן על בסיס יומי. עבור ביצוע התחזית השתמשנו בנתוני תחלואה (מספר חולים חדשים, מספר מתים למיליון וכו') לצד נתוני התחסנות ברמת המדינות השונות. בנוסף, כללנו במסד הנתונים של המטלה נתונים כמו צפיפות האוכלוסין ורמת התוצר של המדינה, מתוך מחשבה כי יש להם השפעה על התפשטות המגיפה, היקף השימוש ברכבים והם יכולים לתרום ליכולת החיזוי של המודל.

**את נתוני המדיניות** לקחנו ממסד נתונים של אוניברסיטת אוקספורד. מדובר בעבודה משותפת של חוקרים מרחבי העולם המעדכנים את צעדי המדיניות של המדינות השונות ברמה היומית. מסד הנתונים מבצע מעקב אחר מדיניות מגבילות שונות כמו סגירת בתי ספר, סגירת מקומות בילוי, מדיניות ביצוע בדיקות ועוד. עבור כל צעד מדיניות מבצעים תהליך של תקנון המצביע על מידת החומרה של צעד זה (למשל: האם כל בתי הספר נסגרו, או שרק גני הילדים). בנוסף, מציינים האם צעד המדיניות בוצע ברמה הארצית וברמה מוגבלת יותר (כמו למשל הערים "האדומות" בישראל שבהם היו הגבלות מקיפות יותר).

אנחנו התבססו על האינדקסים שמסד הנתונים של אוקספורד מייצר. אינדקסים אלו סוכמים את צעדי המדיניות ומסדרים אותם בסקאלה של 0-100, ציון גבוה מציין כי המדינה נקטה בהגבלות רבות ברמה המדינית. לדעתנו, השימוש באינדקס רציף יוצר דרך מדידה אחידה עבור צעדי המדיניות השונים בכל מדינה וניתן להבין טוב יותר מאילו מדינות צעדי המדיניות היו חמורים יותר ובאילו הצעדים היו פחות חמורים. בנוסף, האינדקסים מתגברים על פערי מדיניות בתוך המדינה, יש משקל באינדקס במידה וצעד המדיניות הוצע ברמה הארצית או המחוזית וכו'.

בנוסף, הוספנו למסד הנתונים את מועדי החגים בכל המדינות הרלוונטיות. לשם כך התבססנו על נתונים מהאתרים "Nager.date" ו- "windcalander.com". אתרים אלו מרכזים את ימי החגים עבור המדינות שעליהן אנחנו נדרשים לבצע את הפרדיקציה. הוספה של ימי החגים מגיעה מתוך מחשבה שבימים אלו יש שינויים בהרגלי התנועה וזה עלול להשפיע על התחזית שאנחנו נבצע. השתמשנו **רק**  בימי חג הנחגגים ברמה הארצית ומצוינים בצורה רשמית על ידי המדינה.

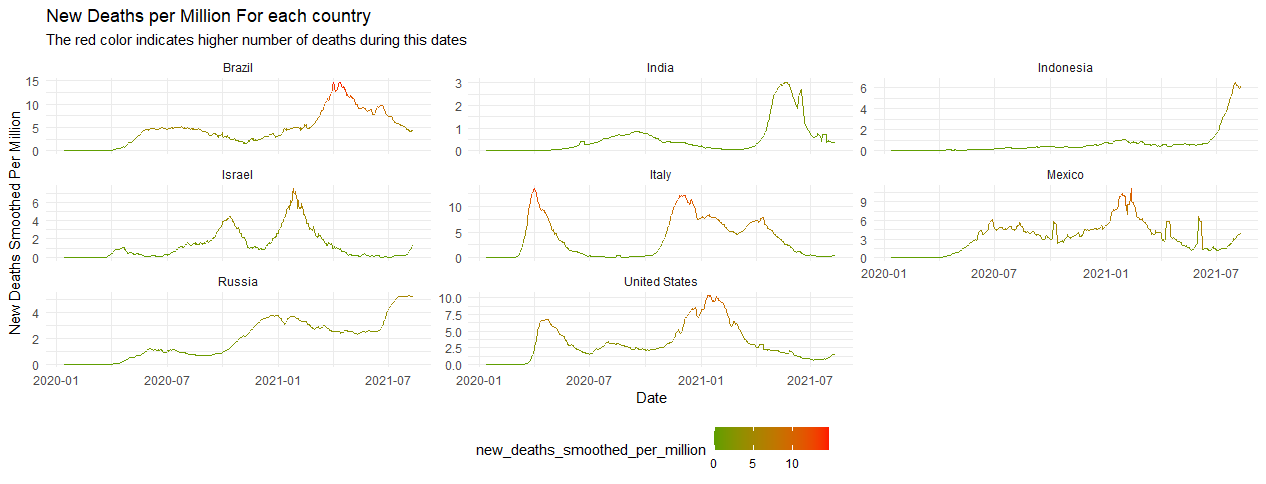
לאחר איחוד מסדי הנתונים נכנסו לשלב של ניקוי ערכים חסרים וסידור המשתנים המסבירים. בשל העובדה שהחיסונים היו זמינים רק בתחילת שנת 2021, הפכנו את הערכים החסרים ל - 0 עד למועד תחילת מתן הזריקות הראשונות. בנוסף בחלק מהשמשתנים המסבירים היו ערכים חסרים שנובעים בהיעדר דיווח. בחרנו להשוות אותם לערך המדווח האחרון מכיוון שמדובר בצעדים שאינם משתנים בצורה דרסטית ברמה היומית.

לסיכום, לאחר ביצוע עבודת איחוד וניקוי מסדי הנתונים. נותרנו עם קובץ הבנוי מ-26 משתנים וכולל 9212 תצפיות. **להלן תיאור מסד הנתונים:**

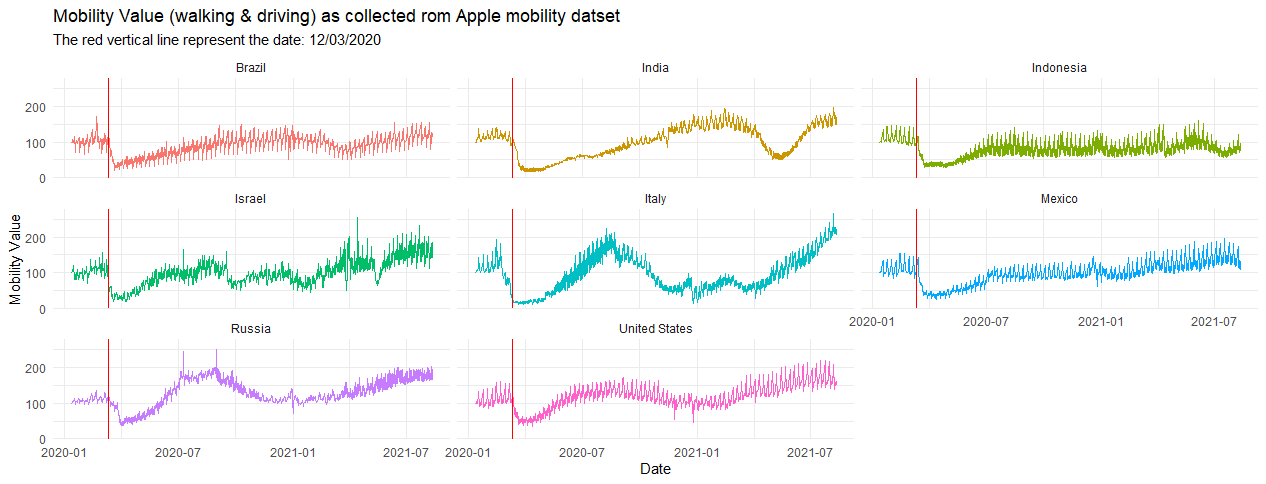


**חלק שני - תיאור ויזואלי של מסד הנתונים**

להתפשטות מגפת הקורונה הייתה השפעה מהותית על צעדי המדיניות שנקטו הממשלות ברחבי העולם. מה שכמובן הוביל לשינויים בתנועת הולכי הרגל ותנועת הרכבים במדינות. הגרף הבא מתאר את מספר הנפטרים ממחלת הקורונה לכל מיליון איש בכל התקופה:

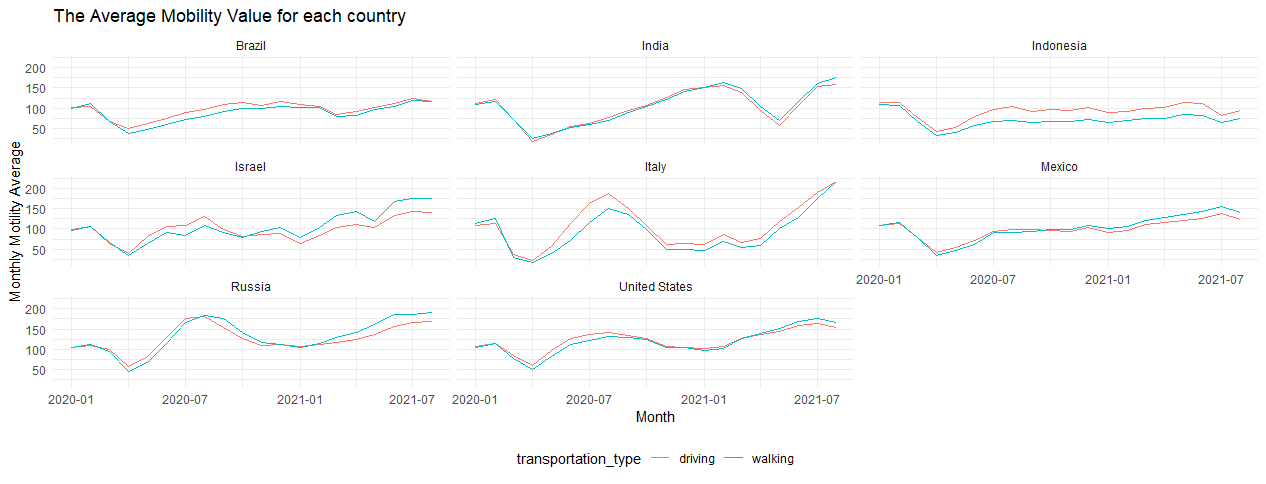


העלייה במספר החולים והנפטרים ממגיפת הקורונה גררה ירידה משמעותית בתנועה בתקופה הנבחנת כתוצאה ממדיניות ממשלתית (סגרים, סגירת מקומות בילוי וכו') ושינויי התנהגות חברתיים (רצון לשמור על ריחוק חברתי, פחד מהידבקות וכו'). להלן גרף המתאר את השינוי (הקו האנכי האדום מסמן את התאריך 12/03/2020 סביבו החלו דיווחים משמעותיים על התפשטות הקורונה מחוץ לתחומי סין):



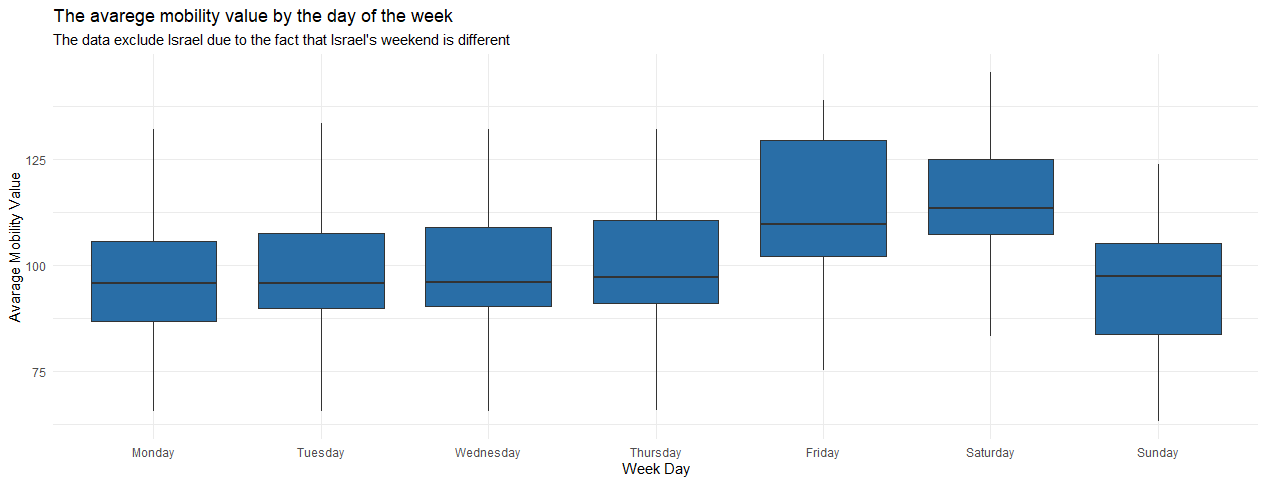
ניתן לראות ירידות משמעותיות בהיקפי התנועה, בזמנים בהם חלו התפרצויות נוספות של המחלה (לדוגמה גרף התמותה בהודו מראה תמותה גבוהה ברבעון השני של 2021 ובגרף התנועות רואים ירידה באותה התקופה)

בגרף הבא בחרנו לבצע ממוצע חודשי של ערכי התנועה, על מנת לצמצם את תנודתיות הגרפים הנובעת מהמדידה היומית. בנוסף, ביצענו הפרדה הן ברמת המדינות והן ברמת סוגי התעבורה (הליכה ונהיגה):

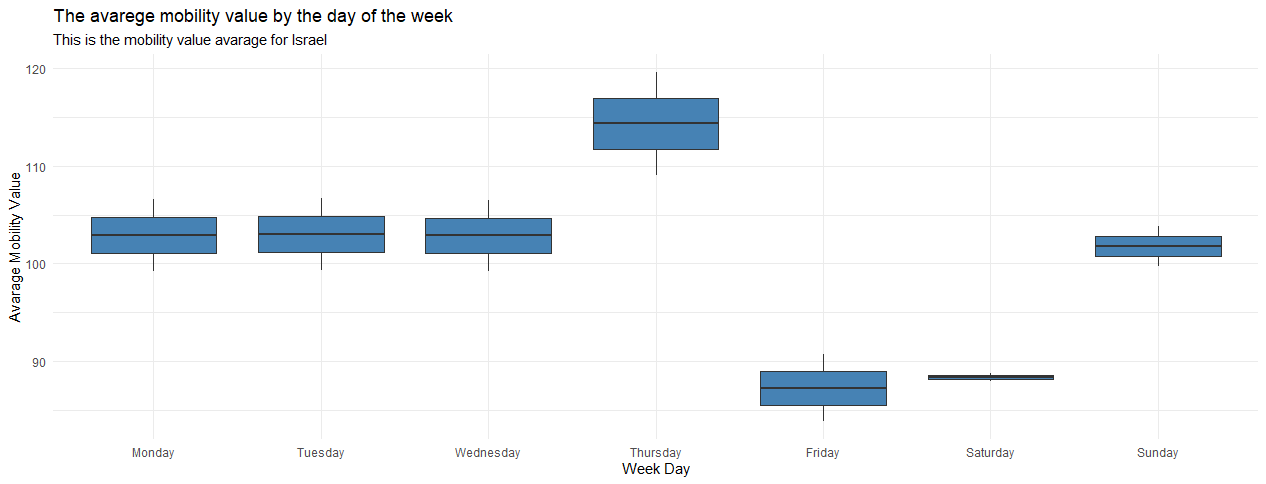


**למעט הודו ומקסיקו**, בתקופה שבה התנועה חוזרת לערכים גבוהים יותר בעקבות הסרת המגבלות הממשלתיות, יש נטיה לשימוש ברכב לעומת הליכה ברגל. ניתן ליחס זאת הן מתוך רצון לנקוט במשנה זהירות בכל מה שקשור להסתובבות בקרב אנשים והן כתוצאה מהגבלות המוטלות על הסתובבות רגלית תחבורה ציבורית וכיוצא בזה.

בחנו גם את דפוסי התנועה ביחס לימי השבוע. ניתן לראות כי בעולם יש עליה בהיקף התנועה לקראת סוף השבוע (בישראל מדובר ביום חמישי ובשאר המדינות מדובר ביום שישי) וירידה בהיקף התנועה ביום החופש (ראשון). ישראל מתאפיינת בשיעור תנועה נמוך במהלך הסופ"ש שקשור להגבלות על פעילות המסחר, אי הפעלת תחבורה ציבורית ומטעמים אישיים של שמירת שבת:

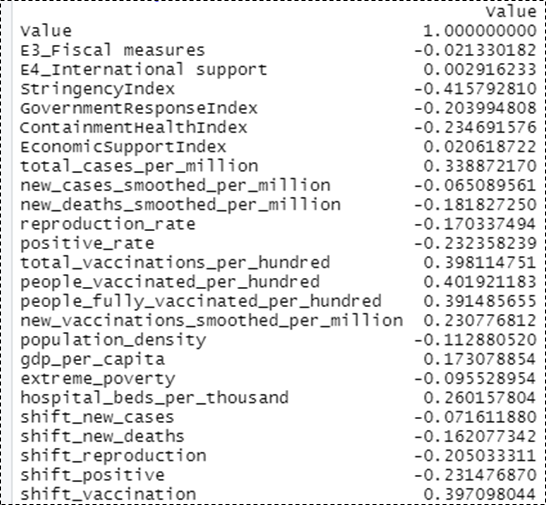


**גרף המתאר את שיעורי התנועה בישראל:**



**חלק שלישי - מודל חיזוי לתאריכים 1-10/08/2021**

על מנת לתפוס את ההבדלים בין הליכה ולנהיגה, הוספנו למסד הנתונים משתנה בינארי המקבל את הערך 1 כאשר מדובר בנתונים עבור הולכי רגל ו-0 במידה ומדובר בנהיגה. בנוסף, הוספנו משתנים הבנויים על הסטה של עשרה ימים קדימה עבור נתוני התחלואה (מס' החולים הפעילים, מס' המתים, שיעור החיוביים ומקדם ההדבקה) מתוך חשיבה כי ההשפעה של נתונים אלו על התנועה מתעכבת כתוצאה מזמן התגובה של הממשלה, האופי האקספוננציאלי של התפשטות המחלה ועוד. בחנו את הקשר בין המשתנים המסבירים ל - Value (המשתנה החזוי) באמצעות מטריצת קורלציות. **להלן התוצאות**:



לטובת אימון המודל השתמשנו בשיטת cross validation. את מסד הנתונים חילקנו לקובץ אימון המכיל 85% מסך התצפיות. את קובץ אימון זה חילקנו ל- 10 תקיות שעליהן בחנו את שלושת המודלים. החלטנו לאמוד את נתוני התעבורה באמצעות מודל ריגרסיה ליניארית המתאימה ביותר להערכתינו לאמידת משתנה ה-Value בשל העובדה שהוא רציף. בהתאם לתוצאות בחרנו לבחון 3 מודלים אשר כוללים מספר משתנים שונים. המודל הראשון כולל את המשתנים בעלי הקורלציה החזקה ביותר מבין כל המשתנים הכלולים במסד הנתונים. המודל השני כולל את כל המשתנים הקיימים בניקוי משתנים בעלי קורלציה חלשה במיוחד למשתנה value ומודל שלישי עם כלל המשתנים הזמינים. תוצאות ההרצה, כולל הבחינה אל מול ה- Test data set הראו r^2 הגבוה ביותר (0.66) עבור המודל השלישי ולכן בחרנו בו.

**חלק רביעי - קורלציה אל מול סיבתיות**

כאשר אנחנו מדברים על סיבתיות הכוונה היא להראות כי שינוי בגורם א' גורר בצורה מובהקת לשינוי בגורם ב'. קשר קורלטיבי בין משתנים שונים מזהה את אופן השינוי של משתנים מסוימים ביחס למשתנים אחרים אך לא בהכרח מראה כי משתנה אחד משפיע על משתנה השני. את הקשר הסיבתי בין משתנים שונים קשה להראות מחוץ לתנאי מעבדה סטריליים, מכיוון שבעולם האמיתי מדובר במערכת דינמית שקשה לבודד את ההשפעה הבודדת של משתנה יחיד.

בשל כך, לדעתינו המודל שלנו לא תופס קשר סיבתי בין נתוני התעבורה למשתנים המסבירים שבחרנו לכלול במודל. אין אפשרות להחזיק את כל המשתנים קבועים ולשנות משתנה מסוים על מנת לבחון את השפעתו על ערכי התנועה. לפיכך המודל שלנו בוחן את הקורלציה בין כלל המשתנים המסבירים לבין נפח התנועה. בחיזוי עבור עתידי, כאשר הנתונים אינם זמינים לנו ישנו רובד נוסף של אי וודאות שנוסף למודל. אנו נדרשים לבצע חיזוי גם עבור ערכי המשתנים המסבירים עצמם, ולכן טווח הטעות גדל ויכול הניבוי של המודל קטנה.

לפיכך, אנו מעריכים כי איכות הפרדיקציה שלנו תהיה פחות טובה ומדויקת ביחס למודל בחלק השלישי בשל המידע שאיננו זמין. בשל העובדה כי אנו עדים להתפרצות של ואריאנטים חדשים של נגיף הקורונה לצד ירידה ביעילות החיסון אנחנו מעריכים כי בתקופה שבה אנו מעוניינים לבצע תחזית ממשלות ברחבי העולם ינקטו בצעדי מדיניות במטרה לבלום את ההתפרצות המחודשת. לכן, היינו שמחים לקבל מידע על מדיניות הממשלות הרלוונטיות לתחזית בתקופת התאריכים הזו. כמובן שנתוני תחלואה עתידיים היו תורמים לשיפור התחזית שלנו.

**חלק חמישי - מודל חיזוי עבור 18/8 עד ה-1/9**

בחלק החמישי של המטלה התבקשנו לבצע חיזוי עבור ערכי התנועה של הולכי הרגל והנהגים במדינות השונות עבור תאריכים עתידיים. האתגר המרכזי בחלק זה הוא ההתמודדות עם העובדה כי הערכים של המשתנים המסבירים אינם ידועים לנו ולכן עלינו לחשוב אילו משתנים מסבירים עלינו לכלול במודל החיזוי עבור סעיף זה ואיך להעריך את הערך העתידי שלהם. לשם כך, השתמשנו בממוצע מתגלגל, אשר לוקח את 30 הערכים האחרונים של המשתנים המסבירים השונים ומפיק מהם ממוצע אשר מדמה את הערך של המשתנה בתאריך העתידי. בסיום התהליך נותרנו עם מסד נתונים הכולל נתונים עבור התאריכים הרלוונטיים.

לטובת ביצוע התחזית, השתמשנו במודל דומה למודל הנבחר בסעיף 3 של המטלה: שימוש ברגרסיה ליניארית שלוקחת בחשבון את כלל המשתנים המסבירים במסד הנתונים. השמטנו משתנים שאמידתם הייתה מורכבת, ווידאנו שאין פגיעה בr^2 של המודל. לדוגמא השמטנו את משתני החיסונים המצטברים שלא ניתן לאמוד אותם באמצעות ממוצע, אך השארנו את משתנה המתחסנים החדשים. ביצענו תהליך דומה של אימון ובדיקת המודל וקיבלנו ערך של R^2 דומה. **תוצאות הפרדיקציה עבור התאריכים הנדרשים מפורצות בקובץ ה-CSV המצורף למטלה.**