

**SafeRoads**

**Data Understanding Report**

**רום קדוש – 207691841**

**סנדרה זייגרמכר -206987521**

**תוכן עניינים**

1. **איסוף נתונים ............................................................................. 3-12**
2. **תיאור הנתונים ........................................................................... 12-13**
3. **חקירת נתונים ............................................................................ 13-22**
4. **איכות הנתונים ........................................................................... 23**
   1. **איסוף נתונים**
   2. **מקורות הנתונים**

הנתונים אשר נתבסס עליהם בפרויקט שלנו נאספו ממאגרים ציבוריים הכוללים את Data.gov.il .

**Data.gov**

הנתונים מכילים מידע כמו מיקומי תאונות (יישוב, רחובות), נתוני זמן (שנה, חודש, שעה), תנאי דרך (מהירות מותרת, סימון ותמרור, תאורה), סוג התאונה וחומרתה, ומאפייני הכביש. כמו כן, הנתונים כוללים מידע על האזור המנהלי בו אירעה התאונה ומאפיינים סביבתיים כמו מזג האוויר ופני השטח. הנתונים מגיעים בקובץ CSV אשר מכיל מידע מפורט על תאונות עירוניות ובין-עירוניות בשנים 2022-2023.

**נתונים שנרכשו:**

בפרויקט לא נעשה שימוש בנתונים שנרכשו ממקורות חיצוניים, כגון נתוני דמוגרפיה. אם יתעורר צורך בהבנת פרופיל הנהגים או התאונות לפי מאפיינים דמוגרפיים (כמו גיל, מגדר), ניתן יהיה לשקול שימוש בנתונים משלימים ממקורות כמו סקרים או מאגרים פרטיים.

**נתונים נוספים:**

במידה ויתברר שהמידע הקיים אינו מספיק או שיעלה צורך בנתונים משלימים שאין ברשותנו, נוכל להוסיף נתונים נוספים על ידי שיתוף פעולה עם גורמים חיצוניים כמו המשטרה, מד"א והרשויות המקומיות. בנוסף נוכל להשתמש בנתונים מהלשכה המרכזית לסטטיסטיקה (הלמ"ס).

* 1. **בדיקה ראשונית של הנתונים**

פירוט של קובץ הנתונים מאתר data gov.

|  |  |  |  |
| --- | --- | --- | --- |
| מס׳ | מאפיין | תיאור מאפיין | סינון ראשוני לעיבוד נתונים בהמשך |
| 1 | PK\_TEUNA\_FIKT | מספר תאונה פיקטיבי ייחודי |  |
| 2 | SUG\_TIK | סוג התיק, אופן הטיפול בתאונה |  |
| 3 | THUM\_GEOGRAFI | האזור הגיאוגרפי בו התרחשה התאונה, 1 – ישראל ללא יהודה ושומרון, 2 -יהודה והשומרון. |  |
| 4 | SUG\_DEREH | סוג דרך – 1 - עירוני בצומת, 2 - עירוני לא בצומת, 3 - לא עירוני בצומת, 4 - לא עירוני לא בצומת. | כללי ברזל - בלוק - Block - מרכז הגנת סייבר לאזרחים |
| 5 | SEMEL\_YISHUV | סמל ישוב – מספרים שמסמנים את היישובים בארץ ישראל. מגיע עם קידוד מתאים במילון עזר. | כללי ברזל - בלוק - Block - מרכז הגנת סייבר לאזרחים |
| 6 | REHOV1 | רחוב 1 – הראשי, מילון רחובות לפי מספר מקבלים את הרחוב בו התרחשה התאונה. | כללי ברזל - בלוק - Block - מרכז הגנת סייבר לאזרחים |
| 7 | REHOV2 | רחוב 2 – המשני במידה והתאונה קרתה בצומת, מילון רחובות לפי מספר מקבלים את הרחוב בו התרחשה התאונה. | כללי ברזל - בלוק - Block - מרכז הגנת סייבר לאזרחים |
| 8 | MISPAR\_BAIT | מספר בית. |  |
| 9 | ZOMET\_IRONI | צמתים עירוניים (המילון עזר לא נמצא). |  |
| 10 | KVISH1 | דרך לא עירונית – 1. |  |
| 11 | KVISH2 | דרך לא עירונית – 2. |  |
| 12 | KM | ק״מ. |  |
| 13 | ZOMET\_LO\_IRONI | צמתים לא עירוניים (המילון עזר לא נמצא). |  |
| 14 | YEHIDA | יחידה משטרתית, 11 – מרחב חוף (חיפה), 12 – מרחב גליל, 14 – מרחב עמקים, 20 – מרחב תל אביב, 33 – מרחב אילת, 34 – מרחב הנגב, 36 – מרחב שמשון (עד 1999), 37 – מרחב שמשון (החל מ2004), 38 – מרחב לכיש, 41 – מרחב שומרון, 43 – מרחב יהודה, 51 – מרחב השרון, 52 – מרחב השפלה, 61 – מחוז ירושלים. |  |
| 15 | SHNET\_TEUNA | שנת תאונה – 2022 /2023. | כללי ברזל - בלוק - Block - מרכז הגנת סייבר לאזרחים |
| 16 | HODESH\_TEUNA | חודש תאונה – לפי מס׳ חודשים לועזי 1 -12. | כללי ברזל - בלוק - Block - מרכז הגנת סייבר לאזרחים |
| 17 | YOM\_BE\_HODES | יום בחודש (המילון עזר לא נמצא). |  |
| 18 | SHAA | שעה (המילון עזר לא נמצא). |  |
| 19 | SUG\_YOM | סוג יום, 1 – חג, 2 – ערב חג, 3 – חול המועד, 4 – יום אחר. |  |
| 20 | YOM\_LAYLA | יום/לילה – יום 1, לילה 5 | כללי ברזל - בלוק - Block - מרכז הגנת סייבר לאזרחים |
| 21 | YOM\_BASHAVUA | יום בשבוע – כמס׳ הימים בשבוע 1 -7 | כללי ברזל - בלוק - Block - מרכז הגנת סייבר לאזרחים |
| 22 | RAMZOR | מרומזר/לא מרומזר – לא זה 0, כן זה 1, לא ידוע זה 9. | כללי ברזל - בלוק - Block - מרכז הגנת סייבר לאזרחים |
| 23 | HUMRAT\_TEUNA | חומרת התאונה – לפי דרגות, 1 – זה קטלנית, 2 – קשה, 3- קלה. | כללי ברזל - בלוק - Block - מרכז הגנת סייבר לאזרחים |
| 24 | SUG\_TEUNA | סוג תאונה – 1 – פגיעה בהולך רגל, 2 – התנגשות חזית אל צד, 3 - התנגשות חזית באחור, 4 – התנגשות צד בצד, 5 – התנגשות חזית אל חזית, 6 – התנגשות עם רכב שנעצר ללא חנייה, 7 – התנגשות עם רכב חונה, 8 – התנגשות עם עצם דומם, 9 – ירידה מהכביש/ עלייה למדרכה, 10 – התהפכות, 11- החלקה, 12 – פגיעה בנוסע בתוך כלי רכב, 13 – נפילה ברכב נע, 14 – שריפה, 15 – אחר, 17 – התנגשות אחור אל חזית, 18 – התנגשות אחור אל צד, 19 – התנגשות עם בעל חיים, 20 – פגיעה ממטען של רכב. | כללי ברזל - בלוק - Block - מרכז הגנת סייבר לאזרחים |
| 25 | ZURAT\_DEREH | צורת הדרך, 1 – כניסה למחלף, 2 – ביציאה ממחלף, 3 – מ.חנייה/ת דלק, 4 – שיפוע תלול, 5 – עקום חד, 6 – על גשר מנהרה, 7 – מפגש מסילת ברזל, 8 – כביש ישר/ צומת, 9 – אחר, 10 – תחנת אוטובוס, 11 – נתיב תחבורה ציבורית. |  |
| 26 | HAD\_MASLUL | דרך חד מסלולית (חלופה לדרך רב מסלולית), 1 – חד סטרי, 2 – דו סיטרי + קו הפרדה רצוף, 3 – דו סיטרי אין קו הפרדה רצוף, 4 -אחר, 9 – לא ידוע מספר מסלולים. |  |
| 27 | RAV\_MASLUL | דרך רב מסלולית (חלופה לדרך חד מסלולית), 1 – מיפרדה מסומנת בצבע, 2 – מיפרדה עם גדר בטיחות, 3 – מיפרדה בנויה ללא גדר בטיחות, 4 – מיפרדה לא בנויה, 5 – אחר. |  |
| 28 | MEHIRUT\_MUTERET | מהירות מותרת (בקמ״ש) – 0 – לא ידוע, 1 – עד 50, 2 – 60, 3 – 70, 4 – 80, 5 – 90, 6 – 100, 7 – 110. | כללי ברזל - בלוק - Block - מרכז הגנת סייבר לאזרחים |
| 29 | TKINUT | תקינות הדרך – 0 – לא ידוע, 1 – אין ליקוי, 2 – שוליים גרועים, 3- כביש משובש, 4 – שוליים גורעים בכביש משובש. | כללי ברזל - בלוק - Block - מרכז הגנת סייבר לאזרחים |
| 30 | ROHAV | רוחב הכביש, 0 – לא ידוע, 1 – עד חמש מטר, 2 – חמש עד שבע מטר, 3 – שבע עד עשר וחצי, 4 – עשר וחצי עד ארבע עשרה, 5 – יותר מארבע עשרה. |  |
| 31 | SIMUN\_TIMRUR | סימון/תימרור – 1 – סימון לקוי/ חסר, 2 – תמרור לקוי / חסר, 3 – אין ליקוי, 4 – לא נדרש תמרור, 5 – לא ידוע. | כללי ברזל - בלוק - Block - מרכז הגנת סייבר לאזרחים |
| 32 | TEURA | תאורה – 1 – אור יום רגיל, 2 – ראות מוגבלת עקב מזג אוויר (עשן, ערפל), 3 – לילה פעלה תאורה, 4 – קיימת תאורה בלתי תקינה/ לא פועלת, 5- לילה לא קיימת תאורה, 6 – לילה לא ידוע, 7 – לילה תאורה תקינה עם ראות מוגבלת, 8 – לילה תאורה לא תקינה עם ראות מוגבלת, 9 – לילה ללא תאורה עם ראות מוגבלת, 10 – דמדומים, 11 – יום לא ידוע. | כללי ברזל - בלוק - Block - מרכז הגנת סייבר לאזרחים |
| 33 | BAKARA | בקרה בצומת – רק בעירוני בצומת או לא עירוני בצומת. 1 – אין בקרה, 2 – רמזור תקין, 3 – רמזור מהבהב צהוב, 4 – רמזור לא תקין, 5 – תמרור עצור, 6- תמרור זכות קדימה, 7 – אחר. | כללי ברזל - בלוק - Block - מרכז הגנת סייבר לאזרחים |
| 34 | MEZEG\_AVIR | מזג אויר – 1 – בהיר, 2 – גשום, 3 – שרבי, 4 – ערפילי, 5 – אחר, 9 – לא ידוע. | כללי ברזל - בלוק - Block - מרכז הגנת סייבר לאזרחים |
| 35 | PNE\_KVISH | מצב פני הכביש – 1 – יבש, 2 – רטוב ממים, 3 – מרוח בחומר דלק, 4 – מכוסה בבוץ, 5 – חול או חצץ על הכביש, 6 – אחר, 9 – לא ידוע. | כללי ברזל - בלוק - Block - מרכז הגנת סייבר לאזרחים |
| 36 | SUG\_EZEM | סוג העצם (רק בסוג תאונה 8 – התנגשות עם עצם דומם), 1 – עץ, 2 – עמוד חשמל/ תאורה /טלפון, 3 – תמרור ושלט, 4 – גשר סימניו ומגניו, 5 – מבנה, 6 – גדר בטיחות לרכב, 7 – חבית, 8 – אחר, 9 – לא ידוע. |  |
| 37 | MERHAK\_EZEM | מרחק העצם (רק בסוג תאונה 8 – התנגשות עם עצם דומם), 1 – עד מטר, 2 – אחד עד שלוש מטר, 3 – על הכביש, 4 – על שטח הפרדה, 9 – לא ידוע. |  |
| 38 | LO\_HAZA | לא חצה ( רק בסוג תאונה 1, חלופה לאופן חצייה, תאונה מסוג 1 זה פגיעה בהולך רגל), 1 – הלך בכיוון התנועה, 2 – הלך נגד, 3 – שיחק על הכביש, 4 – עמד על הכביש, 5 – היה על אי הפרדה, 6 – היה על שוליים/ מדרכה, 7 – אחר, 9 – לא ידוע. |  |
| 39 | OFEN\_HAZIYA | אופן חצייה (רק בסוג תאונה 1, חלופה לאופן חצייה, תאונה מסוג 1 זה פגיעה בהולך רגל), 1 -התפרץ אל הכביש, 2 – חצה שהוא מוסתר, 3 – חצה רגיל, 4 – אחר, 8 – לא חצה, 9 – לא ידוע אם חצה. |  |
| 40 | MEKOM\_HAZIYA | מקום חצייה (רק בסוג תאונה 1, תאונה מסוג 1 זה פגיעה בהולך רגל), ), 1 – לא במעבר חצייה ליד צומת, 2 - לא במעבר חצייה לא ליד צומת, 3 – במעבר חצייה בלי רמזור, 4 – במעבר חצייה עם רמזור, 9 – לא ידוע מקום חצייה. |  |
| 41 | KIVUN\_HAZIYA | כיוון חצייה (רק בסוג תאונה 1, תאונה מסוג 1 זה פגיעה בהולך רגל), 1 – מימין לשמאל, 2 – משמאל לימין, 9 – לא ידוע. |  |
| 42 | MAHOZ | מחוז (רק בסוג דרך 1,2 – 1 זה עירוני בצומת, 2 זה עירוני לא בצומת) המילון עזר לא נמצא. |  |
| 43 | NAFA | נפה (רק בסוג דרך 1,2 – 1 זה עירוני בצומת, 2 זה עירוני לא בצומת) המילון עזר לא נמצא. |  |
| 44 | EZOR\_TIVI | אזור טבעי (רק בסוג דרך 1,2 – 1 זה עירוני בצומת, 2 זה עירוני לא בצומת) המילון עזר לא נמצא. |  |
| 45 | MAAMAD\_MINIZIPALI | מעמד מוניציפלי (רק בסוג דרך 1,2 – 1 זה עירוני בצומת, 2 זה עירוני לא בצומת) המילון עזר לא נמצא. |  |
| 46 | ZURAT\_ISHUV | צורת יישוב (רק בסוג דרך 1,2 – 1 זה עירוני בצומת, 2 זה עירוני לא בצומת) המילון עזר לא נמצא. |  |
| 47 | IGUN\_MEKUBAZ | איכות עיגון, 1 – עיגון מדויק, 2 – מרכז יישוב, 3 – מרכז דרך, 4 – מרכז קילומטר, 9 – לא עוגן. |  |
| 48 | X\_KORDINAT | קורדינטה איקס. | כללי ברזל - בלוק - Block - מרכז הגנת סייבר לאזרחים |
| 49 | Y\_KORDINAT | קורדינטה וואי. | כללי ברזל - בלוק - Block - מרכז הגנת סייבר לאזרחים |

הנתונים שנאספו כוללים מידע רחב ומפורט על תאונות דרכים משנים 2022 ו- 2023, בחרנו להתמקד בשנים האלו כיוון שיש המון רשומות ותיעוד על תאונות דרכים וגם הן עדכניות. כך נוכל לדייק את התחזיות באופן המקסימלי. בנוסף, הנתונים בdataset שלנו מכילים מספר רב של תכונות המאפשרות לבצע ניתוחים סטטיסטיים והסקת מסקנות. מאפיינים כמו שעת התאונה, חומרת התאונה, סוג הדרך, תנאי מזג האוויר ומיקום גיאוגרפי מספקים בסיס מצוין לבניית מודלים חיזויים ולזיהוי מגמות בתאונות דרכים. עם בסיס נתונים הכולל מידע רחב על שנתיים מלאות, ניתן להפיק תובנות משמעותיות וליישם מודלים לחיזוי תאונות עתידיות. בנתונים המקוריים נכללו 49 עמודות (מאפיינים), אך לא כולן רלוונטיות לניתוח וחיזוי תאונות דרכים. לאחר סינון והסרת העמודות הלא רלוונטיות, נותרו 18 עמודות שמתארות את מאפייני התאונה בצורה ממוקדת ואפקטיבית. מספר זה נחשב מתאים ליישום מודלים חיזויים, תוך שמירה על איזון הנתונים ומיצוי מלא של המודל.

* על פי הניתוח הראשוני וההבנה שלנו על הנתונים בdataset, החלטנו להוסיף עמודה חדשה אשר בה סימנו בוי ירוק את המאפיינים שלפי דעתנו ישמשו אותנו בהמשך כבר בעיבוד הנתונים נציין כי זה סינון ראשוני וייתכן ונשנה את הסיווג של המאפיין.
* במהלך הטיפול בנתונים, זוהו עמודות מסוימות המכילות ערכים חסרים, כפי שנראה במידע הגולמי. ערכים חסרים בדאטה עשויים להופיע עקב טעויות באיסוף נתונים או העדר מידע באירועים מסוימים, כמו מיקום גיאוגרפי (X, Y), מספר בית, או רחובות חסרים. הטיפול בערכים החסרים לא בוצע בדאטה שלנו, יש פשוט תאים ריקים בנתונים עצמם, במידת הצורך, בהמשך הפרויקט נבצע טיפול בערכים חסרים על מנת לעבד את הנתונים באופן המקסימלי.

1. **תיאור הנתונים**
   1. כמות הנתונים:

בשנת 2022 ישנם מספר תצפיות כולל של [10,404].

בשנת 2023 ישנם מספר תצפיות כולל של [8,832].

מספר המאפיינים (עמודות): הדאטה מכיל 49 מאפיינים שמתארים את התאונות ואת נסיבותיהן.

* 1. סוגי הערכים:

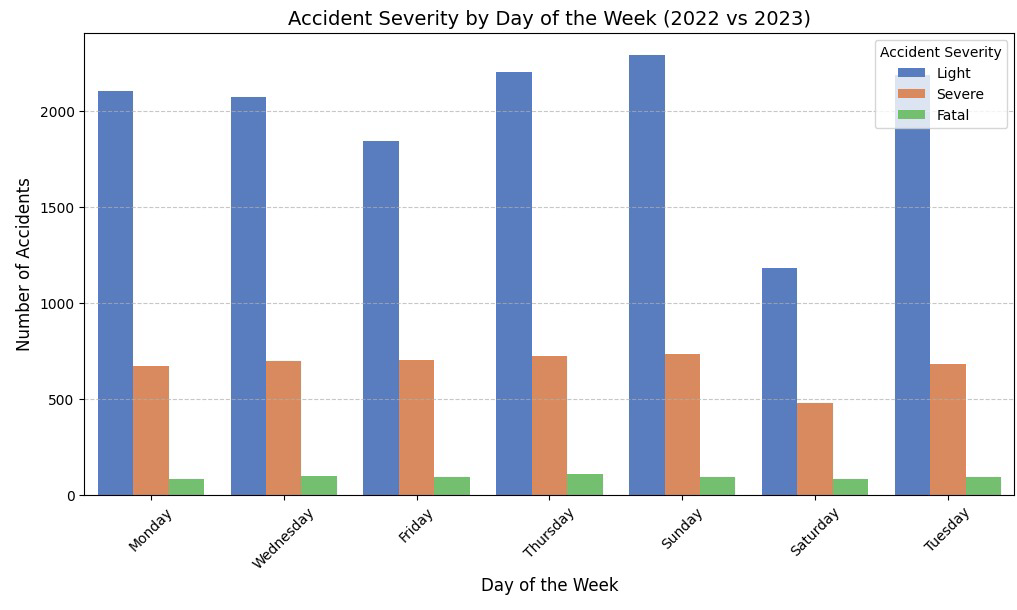
כל הנתונים בדאטה שלנו הם מסוג מספרי (Numeric), כולל ערכים שמייצגים מספרים רציפים כמו קואורדינטות או ערכים שלמים המייצגים קטגוריות מקודדות, כמו סוג יום, חומרת תאונה, ומאפייני דרך. הקטגוריות המקודדות מובילות למילון עזר שבו על פי הנתון המספרי נקבל נתון טקסטואלי מתאים בערכו.

* 1. סכמות קידוד:

בנתונים שלנו נעשה שימוש בסכמות קידוד מספריות עבור מאפיינים קטגוריאליים, כאשר ערכים מספריים מייצגים קטגוריות שונות. לדוגמה, עמודות כמו SUG\_YOM (סוג יום) ו-SUG\_TEUNA (סוג תאונה) מקודדות באמצעות מספרים שלמים כדי לייצג מצבים או קטגוריות שונות, כגון יום חול, סוף שבוע, או סוגי תאונות שונים.

סכמות קידוד אלו מאפשרות עבודה נוחה עם המידע במודלים סטטיסטיים ולמידת מכונה. לנוחות ולהבנה טובה יותר, צירפנו למעלה טבלה המסבירה את המשמעות של כל קוד בעמודות השונות, כפי שנמצא בנתונים.

1. **חקירת נתונים**
   1. במהלך חקירת הנתונים, השתמשנו בגרפים ובכלים שונים בשפת פייתון באמצעות Google Colab כדי להמחיש ולהבין את הקשרים בין המשתנים השונים. העבודה כללה יצירת גרפים שמטרתם הייתה לזהות אם קיימת השפעה הדדית בין המשתנים, ולבדוק את מידת השפעתם של משתנים אחד על השני. באמצעות הגרפים, ניתחנו את המידע בצורה ויזואלית, מה שסייע לנו להבין את הדינמיקה בין הנתונים ולזהות מגמות, קשרים ויחסים פוטנציאליים בין המשתנים השונים.

**ניתוח גרף ראשון: חומרת תאונות לפי יום בשבוע (השוואה בין 2022 ל-2023)**

מטרת הגרף היא לבחון את הקשר בין יום בשבוע לבין כמות התאונות בכל רמות החומרה (קלות, קשות וקטלניות) לאורך השנים 2022 ו-2023. בנוסף, הגרף נועד לזהות תבניות והתנהגויות הקשורות לימים מסוימים שבהם מתרחשות יותר תאונות, וכיצד חומרתן משתנה בין טווח השנים.

בחרנו להציג את הגרף בצורה זו כיוון שהוא מאפשר השוואה ברורה בין שתי השנים עבור כל יום בשבוע. הוא מראה גם את כמות התאונות הכללית וגם את ההתפלגות לפי חומרת התאונות.

מסקנות מהגרף:

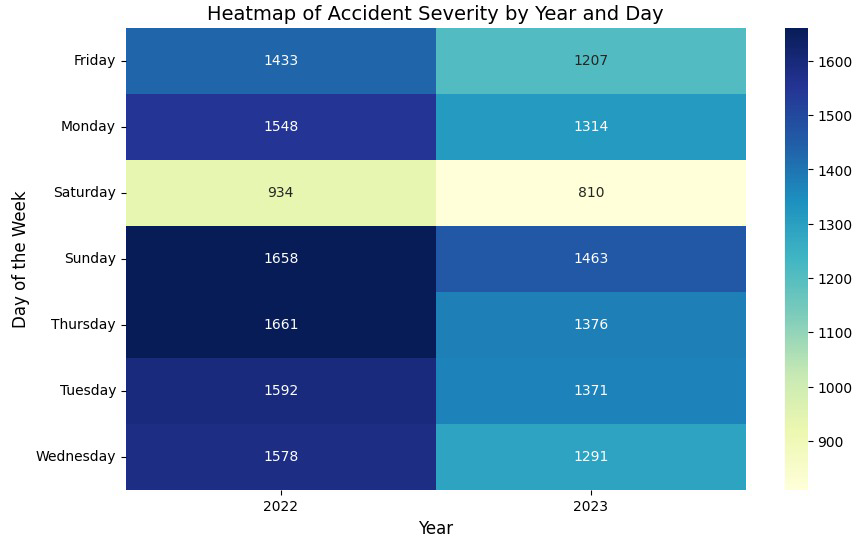
* ימים כמו ראשון ושלישי בולטים בכמות תאונות גדולה, מה שעשוי להעיד על דפוסי תנועה מסוימים, כמו עומסים או התנהגויות נהיגה.
* רוב התאונות הן קלות, עם שיעור קטן של תאונות קשות וקטלניות. עם זאת, ייתכן שתאונות קשות בולטות יותר בימים עם פחות תנועה.
* שינוי בין שנים - ייתכנו שינויים בכמות התאונות או בהתפלגות החומרה בין 2022 ל-2023, מה שעשוי להצביע על גורמים כמו שינויים במדיניות התנועה, תנאי מזג אוויר, או התנהגות נהגים.

כיצד הגרף תורם לפרויקט:

* קבלת החלטות מבוססת נתונים: הגרף מספק בסיס להשוואה בין שנים, ומאפשר בחינה האם צעדים מסוימים שננקטו (כגון הגברת האכיפה) השפיעו על הפחתת התאונות.
* תחזיות עתידיות: המידע יכול לתרום לבניית מודלים לחיזוי תאונות לפי יום בשבוע, מה שיאפשר היערכות טובה יותר.

מסקנות וצעדים עתידיים:

* כדאי לבחון את הימים עם כמות תאונות גבוהה יותר באופן מפורט ולזהות האם יש מאפיינים ייחודיים (שעות מסוימות, מיקומים ספציפיים).
* שילוב נתוני מזג אוויר או עומסים עשוי לתת פרספקטיבה רחבה יותר על הסיבות לעלייה בכמות התאונות בימים מסוימים.

**ניתוח גרף שני: מפת חום של חומרת תאונות לפי יום ושנה**

מטרת מפת החום היא להוות כלי ויזואלי עוצמתי להצגת קשרים ודפוסים במערכת נתונים מרובת משתנים. במקרה זה, מפת החום נועדה להמחיש את הקשר בין ימי השבוע, השנים (2022 ו-2023), וכמות התאונות בכל יום. הצבעים במפה מייצגים את כמות התאונות, כאשר צבעים כהים יותר מייצגים ערכים גבוהים וצבעים בהירים מייצגים ערכים נמוכים יותר.

בחרנו להציג את הגרף בצורה זו כיוון שמפת חום מאפשרת זיהוי חזותי מהיר, במקרה שלנו של ימים מסוימים בשבוע שבהם כמות התאונות גבוהה יותר, והשוואה פשוטה בין השנים, הצבעים מקלים על הבנת הריכוזיות וההתפלגות של התאונות לאורך השבוע.

מסקנות מהגרף:

* ימי חמישי וראשון נראים כבולטים במיוחד בכמות התאונות לשנת 2022. לעומת זאת, ב-2023 ניתן לראות ירידה מסוימת בימים אלו.
* ירידה או עלייה בכמות תאונות: לדוגמה, יום שישי רשם ירידה בכמות התאונות בין 2022 ל-2023, מה שמרמז על אפשרות של שינוי בהתנהגות הנהגים או עומסים שונים ביום זה.
* זיהוי ימים פחות מסוכנים: יום שבת הוא יום שבו כמות התאונות היא הנמוכה ביותר בשני השנים. תובנה זו יכולה להועיל בקביעת מדיניות או פעולות אכיפה מוגברת ביום זה.

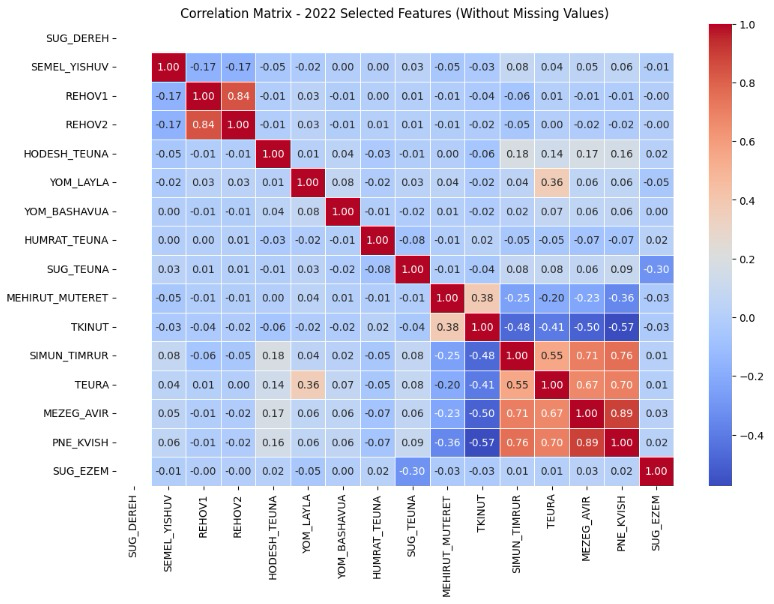
ֿ

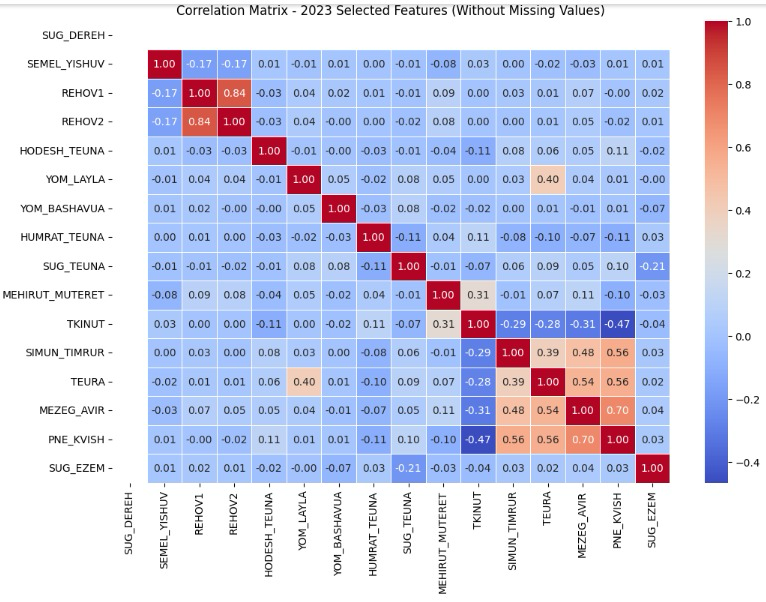
כיצד הגרף תורם לפרויקט:

* תובנות מעשיות: ניתן להשתמש בתוצאות על מנת להמליץ על פעולות מניעה ממוקדות, כמו הגברת אכיפה בימים מועדים לפורענות.
* השוואה רחבה: מאפשרת זיהוי תבניות כלליות במערכת נתונים בצורה פשוטה וויזואלית.
* הנחיות להמשך ניתוח: תובנות ממפת החום יכולות להנחות לניתוח מעמיק נוסף, כמו השפעת תנאי מזג אוויר או שעות עומס בימים שנמצאו כבעייתיים.

מסקנות וצעדים לעתיד:

* ניתן לבצע ניתוח נוסף של הקשרים בין שעות עומס ותנאי מזג אוויר לימים שבהם כמות התאונות גבוהה.
* כחול כהה מצביע על ימי שיא בתאונות, שבהם נדרשת תשומת לב מוגברת ואולי גם התערבות.
* צהוב בהיר מצביע על ימים רגועים יותר, שבהם ניתן ללמוד על גורמים שמפחיתים תאונות, כגון תנועה דלילה יותר (לדוגמה, מנוחה בשבת).

**ניתוח גרף שלישי : מטריצת מתאם - השוואת משתנים נבחרים בין השנים 2022 ל-2023**



מטריצת המתאם (Correlation Matrix) היא כלי סטטיסטי חשוב בניתוח נתונים, שמטרתה למדוד את הקשר הלינארי בין משתנים שונים. המתאם נע בטווח שבין -1 ל-1, כאשר:

* מתאם חיובי גבוה (קרוב ל-1) מציין קשר ישיר בין המשתנים, כלומר כשמשתנה אחד עולה גם השני נוטה לעלות.
* מתאם שלילי גבוה (קרוב ל--1) מציין קשר הפוך בין המשתנים, כלומר כשמשתנה אחד עולה השני נוטה לרדת.
* מתאם קרוב ל-0 מציין שאין קשר לינארי משמעותי בין המשתנים.

מטרת השימוש במטריצת מתאם:

1. זיהוי קשרים חזקים בין משתנים - מאפשר לזהות משתנים שיכולים להיות תלויים אחד בשני או להשפיע זה על זה.
2. סינון משתנים לא רלוונטיים - בעבודה עם מודלים של Machine Learning, משתנים עם מתאם גבוה ביניהם עשויים להיות מיותרים (Multicollinearity).
3. בחירת משתנים חשובים - עוזר לבחור משתנים שיכולים להסביר טוב יותר את התופעה הנחקרת.

משמעות הצבעים במטריצת המתאם:

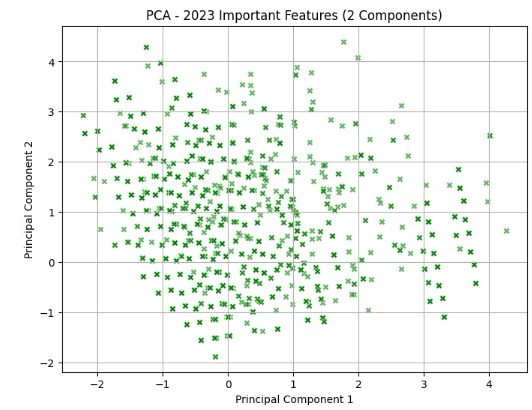
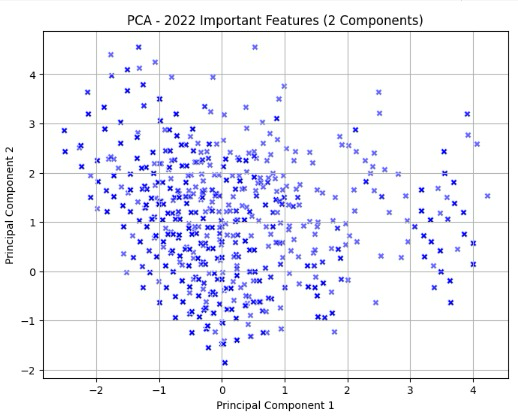
* אדום כהה: מתאם חיובי חזק מאוד (קרוב ל-1).
* כחול כהה: מתאם שלילי חזק מאוד (קרוב ל--1).
* צבעים בהירים: מתאם חלש או לא קיים (קרוב ל-0).

ניתוח המטריצה :

* הבדלים בין השנים: במטריצה של 2022 ושל 2023, ניתן לראות קשרים שונים בין משתנים, מה שמספק תובנות לגבי שינויים דינמיים בתאונות.
* משתנים עם קשר חזק: ניתן לראות, למשל, קשרים חיוביים חזקים בין REHOV1 ל-REHOV2 בשתי השנים. הדבר מצביע על כך שהמשתנים הללו עשויים להיות קשורים גאוגרפית או מבניים.
* משתנים עם קשר שלילי: הקשר השלילי בין מהירות מותרת (MEHIRUT\_MUTERET) לבין משתנים אחרים, כמו חומרת התאונה (HUMRAT\_TEUNA), עשוי לרמוז על השפעת המהירות על חומרת התאונה.

מסקנות וצעדים לעתיד:

* באמצעות המטריצה נוכל לזהות משתנים אשר משפיעים משמעותית על תאונות הדרכים.
* בבוא העת נוכל לסנן משתנים שאינם מוסיפים מידע חדש.
* נצליח להבין קשרים בין משתנים אשר בעתיד יוכלו לשפר את ניתוחי המודל שלנו.

** ניתוח גרף רביעי: PCA לשנים 2022 ו-2023**

גרפי ה-PCA לשנים 2022 ו-2023 ממחישים את הפיזור של תאונות הדרכים בהתבסס על מאפיינים עיקריים שנבחרו מתוך הדאטה. מיפוי הרכיבים איפשר לנו לצמצם את המידע הרב לשני צירים עיקריים (רכיב עיקרי 1 ורכיב עיקרי 2) כך שניתן לנתח את הנתונים בצורה ויזואלית ברורה.

מטרת גרף ה-PCA:

* הבנת הקשרים המורכבים בין המשתנים השונים ותיאורם בצמצום ממדים.
* זיהוי תבניות דומיננטיות או התנהגויות יוצאות דופן בנתונים.
* השוואה בין השנים - באמצעות מיפוי בשני גרפים נפרדים ניתן לזהות שינויים בהתנהגות התאונות.

מה ניתן ללמוד מהגרפים:

* פיזור נתונים: הפיזור המרוכז במרכז מצביע על דפוסי תאונות דומים בין רוב התאונות, בעוד הנקודות הרחוקות ממרכז הפיזור עשויות להצביע על תאונות יוצאות דופן עם מאפיינים חריגים.
* השוואה בין שנים: ניתן לראות אם פיזור הנתונים השתנה בין השנים ולזהות מגמות. לדוגמה, אם ב-2023 קיימת התפשטות רחבה יותר של הנקודות, הדבר עשוי להעיד על שינוי במאפייני התאונות.
* חיזוי: אם קיימות קבוצות ברורות של נקודות (Cluster), ניתן יהיה להשתמש במידע זה כדי לפתח מודלים שינתחו את הקבוצות השונות.

המשמעות של ה-PCA בפרויקט:

* חידוד משתנים: גרף ה-PCA מאפשר להעריך האם הבחירה במאפיינים שנבחרו באמת מייצגת את הנתונים בצורה מיטבית.
* בסיס לניתוחים מתקדמים: התוצאות יכולות לשמש לבניית מודלים חיזויים, כמו קלסיפיקציה של סוג תאונה.
* ניטור שינויים בין שנים: הבדלים במבנה הפיזור יכולים להצביע על שינוי מגמות, מדיניות או נסיבות סביבתיות בין השנים.

מה רואים בגרפים בפועל?

שנת 2022:

רוב הנקודות קרובות יותר למרכז הפיזור, מה שמעיד על כך שרוב התאונות בשנת 2022 היו קרובות לממוצע מבחינת התכונות שנבחרו (חומרת התאונה, סוג הדרך, מזג האוויר, ועוד).

עם זאת, ישנם ערכי X חזקים ורחוקים מהמרכז, המייצגים תאונות קיצוניות או חריגות מבחינת המאפיינים שלהן.

שנת 2023:

בגרף ה-PCA לשנת 2023, הנקודות נראות מפוזרות באופן דומה, אך יש נטייה של חלק מהנקודות להיות מרוכזות בצורה צפופה יותר באזורים מסוימים.

גם כאן ניכרים ערכי X חזקים בקצוות, שמרמזים על מקרים יוצאי דופן, ייתכן שתאונות בתנאים חריגים או שילובים פחות שכיחים של מאפיינים.

**גרף חמישי: השפעת המהירות המותרת על חומרת התאונה (2022 ו-2023)**

גרף התיבה (Boxplot) מציג את ההשפעה של מהירות מותרת על חומרת התאונה. זהו גרף סטטיסטי המאפשר לנו להבין את התפלגות הנתונים עבור כל קטגוריה של מהירות, באמצעות תיאור גרפי של הערכים המרכזיים, הפריסה של הנתונים, וחריגים.

אלמנטים בולטים בגרף:

* ציר ה-X: מייצג את קטגוריות המהירויות המותרות (למשל עד 50 קמ"ש, 60 קמ"ש).
* ציר ה-Y: מייצג את רמות חומרת התאונה (קטלנית, קשה, קלה).
* תיבות (Box): מתארות את אחוזוני הנתונים (25%-75%).
* קו בתוך התיבה: מייצג את החציון (Median), שהוא הערך המרכזי.
* קווי תוחלת (Whiskers): מראים את טווח הנתונים ללא חריגים.
* נקודות בודדות: מייצגות חריגים - מקרים שבהם הנתונים חורגים מהטווח הצפוי.

מגמת ההתפלגות:

* כל קטגוריות המהירות מציגות תיבה רחבה יחסית, דבר המצביע על התפלגות רחבה של חומרת תאונות עבור כל קטגוריה.
* רוב התאונות הן בדרגת חומרה "קלה" (Light), אך ישנן גם תאונות בדרגת "קשה" ו"קטלנית" בכל מהירויות.

קטגוריות ספציפיות:

* עבור קטגוריות "Up to 50 km/h" ו-"Unknown", ההתפלגות מציגה דפוס דומה מאוד, דבר שעשוי להעיד על דמיון בין המקרים שבהם המהירות אינה ידועה לבין מהירויות נמוכות.
* בקטגוריות מהירויות גבוהות יותר (90 קמ"ש ו-110 קמ"ש), ניתן לראות שעדיין יש אחוז גבוה של תאונות בדרגה קלה, אך ייתכן שיש חריגים נוספים בדרגת קשה.

ערכים חריגים:

* נקודות חריגות המייצגות תאונות קטלניות (Fatal) מופיעות בכל הקטגוריות, דבר שמעיד על כך שחומרת התאונה אינה תלויה אך ורק במהירות.

מסקנות וצעדים לעתיד:

* השפעה מוגבלת של מהירות מותרת על חומרת התאונה - למרות שמהירות גבוהה נחשבת למסוכנת יותר, רוב התאונות הקלות מתרחשות גם במהירויות נמוכות. ייתכן שהגורם המשמעותי הוא תנאי דרך ולא המהירות בלבד.
* צורך בניתוח נוסף - ייתכן שמהירות לבד אינה מספיקה להסיק מסקנות משמעותיות, וצריך לשלב פרמטרים נוספים כמו תנאי מזג אוויר או סוג הכביש.
* העדר תלות ברורה במהירות - הגרף מדגיש את היעדר קשר חד משמעי בין המהירות המותרת לבין חומרת התאונה, מה שאומר כי יש גורמים נוספים המשפיעים על תוצאות התאונה.

יתרונות הגרף

* ויזואליזציה של התפלגות: הגרף מאפשר לראות את ההתפלגות המלאה של חומרת התאונות בכל קטגוריית מהירות.
* הצגת חריגים: נקודות חריגות מבליטות מקרים יוצאי דופן שדורשים תשומת לב מיוחדת.
* השוואה ברורה בין קטגוריות: ניתן לראות את הדמיון והשוני בין קטגוריות המהירויות השונות.

נציין כי הגרפים משקפים ממצאים ראשוניים, הם הביאו לנו אינדיקציה ראשונה לגבי הנתונים והקשרים בין המשתנים. בנוסף הם סיפקו לנו רעיונות המשך לשלב עיבוד הנתונים.

1. **איכות הנתונים**

מאגר הנתונים שלנו מכיל בין היתר חוסר עקביות - ערכים חסרים, שגיאות שונות וערכים לא ידועים, כל אלו מקשים על תהליך הניתוח. בעוד שמדובר באחד הכלים החשובים ביותר להבנת תופעות מורכבות כמו תאונות דרכים וזיהוי דפוסים, איכות הנתונים היא קריטית. כאשר נתונים גולמיים אינם עוברים תהליך ניקוי ובקרה, הדבר עלול להוביל למסקנות שגויות, מודלי חיזוי לא מדויקים ואף לפספוס של תובנות קריטיות. לכן, אנו בוחנות באופן יסודי את הנתונים הזמינים, מזהים בעיות פוטנציאליות ומתמודדים עימן לפני תחילת עבודת הקמת המודלים והניתוחים המתקדמים. בעיות נפוצות שעלולות לצוץ כוללות נתונים חסרים, שגיאות הקלדה, חוסר עקביות בקידוד, מטא-נתונים לא מדויקים ושגיאות מדידה הן דברים שקיימים כמעט בכל מאגר נתונים, לכן אנחנו סוקרות אותו באופן מלא.

**נתונים חסרים:**

* קואורדינטות גיאוגרפיות (X, Y): מיקום התאונה חסר או לא נרשם במדויק.
* שמות רחובות: רחוב אחד או יותר אינם מצוינים בקובץ.
* סוג התאונה: ערכים ריקים בסיווג סוג התאונה – למשל תאונה חזיתית או פגיעת הולך רגל.
* נתונים כרונולוגיים: שדות כמו שעה, חודש או שנה עלולים להכיל ערכים חסרים.

איך נתמודד:

* איתור: שימוש בכלים כמו Python או Excel לזיהוי ערכים חסרים פקודות isnull()  או סינון ערכים ריקים.
* השלמה: ניתן להחליף נתונים חסרים בערכים ממוצעים (בשדות מספריים כמו מהירות מותרת).
* חיזוק המודל: נתעלם מנתונים חסרים בשדות לא קריטיים או לשלב טכניקות חיזוי להשלמת ערכים חסרים.

הנתונים שאנו משתמשות בהם בפרויקט מבוססים על נתוני אמת ממקורות ממשלתיים, דבר המבטיח רמת אמינות גבוהה ואיכות מידע יציבה יחסית. מכיוון שהנתונים נאספים ומנוהלים על ידי גופים מוסמכים, רמת הדיוק והעיבוד הבסיסי שלהם מונעת היווצרות של רוב הבעיות הנפוצות הקיימות במאגרי מידע גולמיים. השימוש ב"מילון" המצורף, שמפרט ומפרש באופן מלא כל קוד בערכי הנתונים, מספק בהירות ומקל על ניתוחם. המילון מאפשר הבנה טקסטואלית של הקודים הסטטיסטיים ומבטיח שהמשמעות מאחורי כל נתון תהיה ברורה ומדויקת.