רמת חומרת פציעה בתאונות
דרכים

פרויקט גמר

Data Preparation Report

שם הסטודנטיות :

מאי יוסף 318608072

אלמוג יבדייב 305325417

שם המנחה : יואב זיו

תוכן עניינים

בחירת נתונים

מטרת הפרויקט היא לפתח מערכת חיזוי מבוססת נתונים שתאפשר להעריך את חומרת הפציעות בתאונות דרכים. המערכת תתבסס על נתונים היסטוריים ותשמש כלי תומך החלטה עבור שירותי חירום ורשויות התחבורה. בנוסף, המערכת תאפשר לזהות מגמות וסיכונים גאוגרפיים.

הנתונים שלנו נלקחו מתוך מאגר ממשלתי רשמי (Data.gov) והינם דיווחים של תאונות דרכים במדינת מרילנד. הקובץ כולל כ־ 189,333 שורות ו־39 עמודות כאשר חלק מהעמודות הן מספריות, אך הרוב קטגוריאליות. עם ריבוי ערכים ייחודיים, תתי-קטגוריות, טקסט חופשי, ולעיתים גם שגיאות כתיב, משמעות לא ברורה או ערכים לא אחידים מבחינת קידוד.

קיימות עמודות שבהן מופיעים עשרות, מאות ואף אלפים ערכים ייחודיים, בחלק מהמקרים בשילוב אותיות גדולות/קטנות, סימני פיסוק, פסיקים או ערכים שמייצגים יותר מקטגוריה אחת למשל: Not Suspect of Alcohol Use, Not Suspect of Drug Use.
בהתבסס על שלב איסוף והבנת הנתונים שביצענו בשלב הקודם בדוח data understanding, בשלב זה ביצענו בחינה מחדש של מבנה הנתונים במטרה למקד את העבודה ולבחור בעמודות שתומכות ישירות במטרת הפרויקט: חיזוי חומרת הפציעה בתאונות דרכים.

<u>בחירת שורות:</u>

הסרנו את כל הרשומות (שורות) שבהן לא הופיע ערך בעמודת המטרה Injury Severity, מאחר ולא ניתן לבצע למידה מונחית על שורות ללא ערך מטרה.

<u>בחירת עמודות:</u>

זיהינו מספר עמודות שאינן תורמות לתהליך הלמידה והן:

- עמודות מזהות: עמודות Report Number, Local Case Number, Person ID ו־Vehicle ID הן עמודות מזהות בלבד ואינן מספקות תרומה לחיזוי או ללמידת מכונה. מאחר והן משמשות אך ורק לצורכי תיעוד, הן הוסרו.
- עמודות עם מידע כפול או חופף: העמודה Location הוסרה מאחר והיא מספקת מידע שכבר קיים בעמודות Latitude־וLongitude.
- עמודות עם אחוז גבוה מאוד של ערכים חסרים: Off-Road Description, Municipality, Related Non-Motorist, Non-Motorist Substance Abuse ו־Circumstance - עמודות אלו הכילו יותר מ־80% ערכים חסרים, ולכן החלטנו להסירן כדי לא לפגוע באיכות המודל.
- עמודות בעלות ערך אחיד או כמעט אחיד: בעמודת Driverless Vehicle, כל הערכים הם "No" או "UnKnown" לכן אינה יכולה לתרום והוסרה. עמודת Parked vehicle- עמודה בינארית שבה כמעט כל הערכים הם "No" (184,777 מתוך כ־189,000), ולכן בשל חוסר האיזון הקיצוני הוסרה גם היא.
- עמודת Vehicle Model הוסרה מאחר והחלטנו כי אינה רלוונטית לחיזוי שלנו בגלל משמעותה (מודל הרכב) וגם כי היא מכילה המון ערכים ייחודיים כ-7000, מה שמקשה מאוד על קידוד אפקטיבי ועלול להכניס רעש מיותר למודל.
- שימוש במיקום גאוגרפי ואיחוד לאזורים: במקור, הנתונים כללו מיקום מדויק של התאונה באמצעות העמודות: Latitude, Longitude, Road Name־ו.Cross-Street Name.
מידע זה מדויק אך בעייתי מבחינת מודל חיזוי, שכן הוא כולל אלפי ערכים שונים ,דבר שמקשה על הכללה, מוסיף רעש לנתונים, ויוצר קושי למודל לזהות דפוסים מרחביים.
כדי לשמר מידע גאוגרפי תוך הפחתת רזולוציה והכללה מרחבית, נעשה שימוש בספריית h3 של חברת Uber, הממפה את כדור הארץ ל־משושים גאוגרפיים בגודל קבוע, לפי רמות רזולוציה.
יישמנו את הפונקציה ()h3.latlng_to_cell על העמודות Latitude־וLongitude.
התווספה עמודה חדשה לנתונים: h3_index ולכל שורת תאונה הוקצה ערך חדש בעמודה שמייצג את המשושה שאליו שייכת המיקום.
נעשה שימוש ברזולוציה 6, המייצגת חלוקה לאזורים בגודל בינוני - איזון בין פירוט לבין הכללה. הרזולוציה שנבחרה הייתה מספיק קטנה כדי לשמר מידע גאוגרפי שימושי, כגון שיוך לרמת שכונה או אזור מסחרי, אך לא קטנה מדי כך שלא תיווצר בעיית פיצול יתר (Overfitting) או רעש מנתונים מפורטים מדי. במקביל, הרזולוציה הייתה מספיק גדולה כדי לקבץ נקודות קרובות גאוגרפית לאותו משושה, ובכך לפשט את המידע מבלי לאבד דפוסים מרחביים חשובים. הבחירה ברזולוציה זו מאפשרת למודל ללמוד תבניות אזוריות בצורה יעילה, תוך שמירה על איזון בין דיוק לפרקטיות בלמידת מכונה.
לאחר השיוך, העמודות המקוריות של מיקום (Latitude, Longitude, Road Name, Cross-Street Name) הוסרו מהמערך.
- עמודת Crash Date/Time הכילה את תאריך ושעת התאונה בפורמט טקסטואלי. ביצענו המרה לפורמט תאריך־שעה תקני (datetime), ומתוכה חילצנו שלושה מאפיינים חדשים, עמודות חדשות שהתווספו לנתונים:
Crash Hour: שעת התאונה- לצורך ניתוח הבדלים בין תאונות ביום לעומת לילה.
Crash Day: יום בשבוע שבו התרחשה התאונה- מאפשר לזהות דפוסים לפי ימות השבוע.
Crash Month :חודש בשנה- לבחינת עונתיות או הבדלים בין תקופות השנה.
לאחר חילוץ המאפיינים החדשים, עמודת Crash Date/Time המקורית הוסרה, מאחר שלא נדרש עוד מידע נוסף ממנה.

לאחר בחינה של העמודות שנותרו, זיהינו כי למרות הסרת עמודות בעייתיות רבות, עדיין קיימות עמודות מרכזיות עם שיעור גבוה של ערכים חסרים או ריבוי ערכים ייחודיים באופן שעלול להשפיע לרעה על ביצועי המודל. מצב זה הצריך גישה זהירה לקבלת החלטות לגבי שימור או הסרה של משתנים מסוימים.

לכן, במקום להסתמך רק על כמות החסרים או ריבוי הקטגוריות, בחרנו ליישם גישה מבוססת ניסוי: הרצנו מספר מודלים ומספר גרסאות של המודל תוך שינוי אופן הטיפול בערכים החסרים ובמבנה העמודות. בין השיטות שנוסו: מילוי ערכים חסרים בעמודות מספריות באמצעות חציון או ממוצע, מילוי ערכים חסרים בעמודות קטגוריאליות באמצעות הערך השכיח (mode) או איחוד הערכים החסרים תחת "Other", בנוסף ניסינו גם שילוב- חלק מהעמודות מילאנו לפי שכיח וחלק לפי Other.

הרצת המודלים עם ובלי עמודות מסוימות, תוך שימוש במדדים כמו Feature Importance של מודלים מבוססי עצים. בניתוח חשיבות המשתנים (feature importance), גילינו כי חלק מהעמודות שנראו בתחילה כבעייתיות, כן תורמות במידה מסוימת להסבר המודל, גם אם לא מדובר בתרומה גבוהה. כאשר הסרנו עמודות עם חשיבות נמוכה, לא נצפה שיפור משמעותי בביצועי המודל (ולעיתים אף חלה ירידה קלה ביציבות התוצאה). לאור זאת, קיבלנו החלטה לשמר את כל העמודות שנותרו לאחר סינון ראשוני, תוך השלמה מושכלת של ערכים חסרים לפי סוג המשתנה וזאת על מנת לשמר מידע פוטנציאלי חשוב ולהימנע מהפסד של תכונות שעשויות להשפיע במקרים מסוימים.

נשארנו עם 25 עמודות (21 עמודות מקוריות ו4 עמודות חדשות) אשר עשויות להיות קשורות ישירות לחומרת הפציעה.

<u>עמודות שהשארנו:</u>

לאחר ניתוח הנתונים בשלב ההכנה, נשמרו לצורך המודל רק העמודות המקוריות שנמצאו רלוונטיות לתחזית חומרת הפציעה, והוסרו עמודות בעייתיות או כאלו שלא צפוי שיתרמו למידול. מתוך 39 העמודות המקוריות, נשמרו 21 עמודות, ובהן: Agency Name, ACRS Report Type, Route Type, Collision Type, Weather, Surface Condition, Light, Traffic Control, Driver Substance Abuse, Driver At Fault, Injury Severity, Driver Distracted By, Drivers License State, Vehicle Damage Extent, Vehicle First Impact Location, Vehicle Body Type, Vehicle Movement, Vehicle Going Dir, Speed Limit, Vehicle Year ו־Vehicle Make. אלו עמודות שמייצגות את תנאי הדרך, פרטי הרכב והנהג, ופרטים כלליים על סוג התאונה.

בנוסף, נוספו 4 עמודות חדשות שנבנו כחלק מהנדסת תכונות: h3_index - מזהה אזור גאוגרפי כללי שחושב על בסיס קואורדינטות המיקום באמצעות ספריית H3, וכן שלוש תכונות זמן שנחצבו מתוך עמודת התאריך המקורית: Crash Hour, Crash Day, ו־Crash Month.

<u>השוואה למסמך הקודם Data Understanding:</u>

בשלב הקודם, במסמך הבנת הנתונים (Data Understanding), זיהינו מספר עמודות שיש להסירן מהנתונים בשל חוסר תרומתן לחיזוי או בעיות איכות. בין העמודות שסומנו להסרה היו עמודות עם שיעור גבוה מאוד של ערכים חסרים (Off-Road Description, Municipality, Related Non-Motorist, Non-Motorist Substance Abuse, Circumstance), עמודות מזהות (Report Number, Local Case Number, Vehicle ID, Person ID), עמודת Location שסיפקה מידע שכבר קיים בקואורדינטות, וכן Driverless Vehicle ו־Drivers License State שסומנו כעמודות ללא תרומה צפויה למודל. בשלב ההכנה (Data Preparation), הסרנו בפועל את כל העמודות שצוינו, למעט Drivers License State, שנשארה בסט הנתונים בשלב זה, על מנת לבחון האם ייתכן קשר בין מדינת הנפקת הרישיון לבין חומרת הפציעה- למשל עקב הבדלים בחוקי נהיגה, אכיפה או ניסיון נהיגה טיפוסי במדינות שונות. בנוסף לכך, הסרנו עמודות נוספות שהתגלו כבעייתיות במהלך העבודה: Parked Vehicle, שבה כמעט כל הערכים היו "No", ו־Vehicle Model שהכילה אלפי ערכים ייחודיים שהקשו על קידוד אפקטיבי. עמודת Crash Date/Time לא הוסרה מיד, אלא עברה המרה לשלוש תכונות חדשות – שעת התאונה, יום בשבוע וחודש ולאחר מכן הוסרה. גם עמודות המיקום (Latitude, Longitude, Road Name, Cross-Street Name) הומרו לעמודה אחת חדשה בשם h3_index, המייצגת אזור גאוגרפי כללי באמצעות חלוקה למשושים ברזולוציה 6 בעזרת ספריית H3. בכך, מעבר להסרת עמודות לא רלוונטיות, נוספו לנתונים תכונות חדשות ומשמעותיות שתומכות ביעדי המודל, זיהוי דפוסים גאוגרפיים וזמניים של חומרת פציעות.

ניקוי נתונים

בשלב זה ביצענו סדרת פעולות שמטרתן לשפר את איכות הנתונים ולהכין אותם בצורה מיטבית ללמידת מכונה.

<u>טיפול בנתונים חסרים:</u>

להלן טבלה המציגה את מספר הערכים החסרים בכל עמודה בנתונים המקוריים לפני שינויים (מוצגים בטבלה רק העמודות שבהן יש ערכים חסרים ולא כל העמודות):

	Missing Values	Percentage Missing
Route Type	18517	9.780123
Road Name	19623	10.364279
Cross-Street Name	26698	14.101081
Off-Road Description	171854	90.768118
Municipality	170207	89.898222
Related Non-Motorist	183246	96.785030
Collision Type	585	0.308979
Weather	13356	7.054238
Surface Condition	21730	11.477133
Light	1445	0.763206
Traffic Control	26970	14.244743
Driver Substance Abuse	31320	16.542283
Non-Motorist Substance Abuse	184392	97.390312
Injury Severity	1056	0.557747
Circumstance	153413	81.028136
Driver Distracted By	1151	0.607924
Drivers License State	11907	6.288920
Vehicle Damage Extent	316	0.166902
Vehicle First Impact Location	156	0.082395
Vehicle Body Type	2830	1.494721
Vehicle Movement	948	0.500705
Vehicle Going Dir	5518	2.914442
Parked Vehicle	1534	0.810213
Vehicle Make	473	0.249824
Vehicle Model	515	0.272008

ביצענו את הפעולות הבאות:

1. <u>טיפול בערכים חסרי משמעות:</u> בשלב זה המרנו ערכים שאינם נחשבים מידע תקף או שימושי לערכים חסרים (NaN), כדי לאחד את ההתייחסות לערכים בעייתיים. כללנו בתהליך זה מגוון ערכים חוזרים ושכיחים כמו "", " ", "ZZKNOWN", "UNKN", "UNK", "n/a", "NA", "Unknown" ועוד, אשר הופיעו בצורות שונות בכל מערך הנתונים. המרה זו מאפשרת לנו להתייחס אליהם כאל ערכים חסרים באופן אחיד.
2. <u>מחיקת עמודות עם ערכים חסרים רבים:</u> כפי שפורט בסעיף הקודם, הסרנו עמודות שבהן למעלה מ־80% מהתצפיות הכילו ערכים חסרים:- Off-Road Description, Municipality, Related Non-Motorist, Non-Motorist Substance Abuse, Circumstance.
3. <u>התמודדות עם עמודות המכילות ערכים חסרים:</u>

- בעמודת המטרה Injury Severity הסרנו את כל השורות שבהן הערך היה חסר (1056), מאחר שלא ניתן לבצע למידת מכונה על תצפיות ללא ערך מטרה. ונותרנו עם 188,277 רשומות.
- טיפול בערכים חסרים בעמודות מספריות: במקום להסיר שורות, השלמנו את הערכים החסרים באמצעות חציון (median) - שיטה שאינה רגישה לערכים קיצוניים, ומשמרת את מרכז הפיזור של הנתונים.
  בעמודה המספרית Vehicle Year (שנת ייצור הרכב): בעמודה זו נמצאו ערכים לא תקינים, ולכן הגדרנו טווח סביר בין השנים 1980 ל־2025. ערכים מחוץ לטווח הוסרו והוגדרו כ־NaN. לאחר הסינון, נותרו כ־4,460 ערכים חסרים.
- עמודות קטגוריאליות: כל שאר העמודות עם ערכים חסרים בטבלה הן קטגוריאליות (טקסטואליות). כדי להימנע מאיבוד של תצפיות שלמות, בחרנו לא להסיר שורות עם ערכים חסרים, העדפנו לטפל באמצעות מילוי מותאם- "Other" ולא לוותר על תצפיות שלמות שיכולות לתרום ללמידה של המודל.

*היה ניסיון למחוק שורות מהנתונים שבהם יש מעל 40% ערכים חסרים בכל רשומה. אך ראינו כי דבר זה אינו משפיע על המודל ועל תוצאותיו ולכן לא עשינו זאת.

במהלך שלב זה, בחנו לעומק את איכות הנתונים שבחרנו לניתוח, וזיהינו מספר בעיות חוזרות ונפוצות. בין הבעיות שזיהינו היו ערכים חסרים, שגיאות הקלדה, אי-עקביות בקידוד הערכים, ולעיתים אף ערכים לא ברורים שנדרשה בדיקה ידנית להבנת משמעותם. להלן נפרט עמודה אחר עמודה (רק את העמודות שבהן זיהינו בעיות אלה), את סוגי הבעיות שזוהו ואת הפעולות שנקטנו כדי לנקות ולהכשיר את הנתונים באופן מיטבי להמשך תהליך הלמידה החישובית.

<u>Injury Severity:</u>

עמודה זו מהווה את משתנה המטרה במערך הנתונים, והיא מתארת את רמת חומרת הפציעה שנגרמה בתאונה. בבחינה הראשונית של העמודה נמצאה בעיה משמעותית של כפילויות כתוצאה מאי-אחידות ברישום הקטגוריות. ערכים זהים הופיעו בצורות שונות- חלקם באותיות גדולות בלבד (למשל NO APPARENT INJURY) וחלקם בשילוב אותיות קטנות וגדולות (No Apparent Injury).
אי-אחידות זו גרמה לריבוי קטגוריות מלאכותי, לפיצול נתונים שמייצגים את אותה משמעות סמנטית, ולהטיה אפשרית בתוצאות הניתוחים והמודלים.

לצורך הניקוי, בוצעה המרה של כל הערכים לאותיות קטנות, כתוצאה מכך מספר הקטגוריות ירד מ־10 וריאציות שונות לחמש קטגוריות סטנדרטיות בלבד:
no apparent injury, possible injury, suspected minor injury, suspected serious injury, ו־fatal injury.
פעולה זו תרמה ליצירת עמודת מטרה תקנית, אחידה וללא כפילויות, ובכך אפשרה להמשיך לביצוע ניתוחים מדויקים ולהכשיר את הנתונים לשלב המידול.

Agency Name:

העמודה Agency Name, המתארת את שם סוכנות האכיפה שדיווחה על התאונה, סבלה מבעיה של חוסר עקביות בקידוד. אותן סוכנויות הופיעו בצורות שונות- חלק מהערכים הופיעו באותיות גדולות וכללו את שם העיר בלבד (GAITHERSBURG), וחלק תיארו את הסוכנות במלואה (Gaithersburg Police Depar) מצבים אלו יוצרים כפילויות מלאכותיות, מקשים על חישוב מדויק של נתונים לפי סוכנות, ועלולים להטעות את המודל. לשם כך, בוצעה סקירה ידנית של כלל הערכים הייחודיים בעמודה וזוהו קבוצות של ערכים שקולים סמנטית. לאחר מכן נבנה מילון מיפוי אשר איחד את כל הצורות השונות לערך תקני אחד לכל סוכנות. פעולה זו תרמה ליצירת עמודה אחידה, שמאפשרת ניתוח מדויק וללא רעש של תאונות לפי סוכנות.

Crash Date/Time:

העמודה Crash Date/Time, שמכילה את מועד התאונה (תאריך ושעה), הכילה ערכים תקניים אך הופיעה בחלק מהמקרים עם סיומת AM/PM ובחלקם ללא. כל הערכים הומרו בהצלחה לפורמט datetime של pandas, ללא ערכים שגויים או חסרים. על בסיס עמודה זו גזרנו שלוש תכונות חדשות: שעת התאונה (Crash Hour), יום השבוע (Crash Day) וחודש התרחשותה (Crash Month). לבסוף, העמודה המקורית הוסרה לצורך שמירה על מבנה נתונים נקי ולמניעת כפילויות. תהליך זה אפשר להפיק מידע מובנה ורלוונטי על התפלגות תאונות לאורך זמן.

Route Type:

עמודת Route Type מציינת את סוג הדרך שבה התרחשה התאונה (למשל כביש עירוני, מדינתי, מחוזי). בבחינה הראשונית נמצאה אי-עקביות ברישום, לדוגמה: County לעומת County Route, או Maryland (State) לעומת Maryland (State) Route. הבדל ניסוחי זה, למרות שמשמעותו זהה, גרם לריבוי קטגוריות מיותרות. הפתרון כלל מיפוי של ערכים חופפים והמרה לגרסה אחידה באמצעות מילון שהוכן ידנית. התוצאה היא עמודה אחידה וברורה, שמייצגת נכונה את סוג הדרך, ללא בלבול או כפילויות מיותרות.

Collision Type:

העמודה Collision Type, המתארת את סוג ההתנגשות בתאונה (כגון פגיעה חזיתית, אחורית, צדית וכדומה), הכילה ניסוחים מגוונים לאותו סוג תאונה. לדוגמה: FRONT TO REAR מול SAME DIR REAR END, או HEAD ON מול FRONT TO FRONT. מצב זה יצר פיצול מלאכותי בנתונים, שהקשה על זיהוי תבניות והוריד מאיכות הניתוח. לפיכך, כלל הערכים הומרו לאותיות גדולות (uppercase) ולאחר מכן בוצע מיפוי של ערכים סמנטית זהים לאותו ניסוח תקני. כתוצאה מכך הופחת רעש, צומצם מספר הקטגוריות, והעמודה הפכה ליציבה ומוכנה לניתוח לפי סוגי תאונות.

Weather:

עמודת Weather, אשר מתארת את תנאי מזג האוויר בעת התאונה, כללה ערכים שנכתבו בצורות שונות אך בעלות משמעות זהה, למשל: RAINING לעומת Rain. בוצעה המרה של כלל הערכים לאותיות קטנות (lowercase) והוסרו רווחים מיותרים. לאחר מכן, בוצע מיפוי של ערכים סמנטיים שקולים (כגון: raining הוחלף ל־rain).

Surface Condition:

עמודת Surface Condition, אשר מתארת את מצב פני השטח בזמן התאונה (למשל יבש, רטוב, קפוא), הציגה חוסר אחידות בכתיב הקטגוריות- גם מבחינת אותיות גדולות/קטנות לדוגמה Dry לעומת DRY, וגם מבחינת ניסוחים חופפים שונים כמו: ice/frost לעומת ICE. בנוסף, חלק מהערכים כללו תיאורים מורכבים או משולבים שהקשו על אחידות הקטגוריות, כמו water (standing, moving). לצורך סטנדרטיזציה, כל הערכים הומרו לאותיות קטנות, רווחים מיותרים הוסרו, ובוצע מיפוי ידני של ערכים שקולים סמנטית: ice/frost הפך ל־ice, ו־water (standing, moving) הוחלף ל־water(standing/moving).
ניקוי זה צמצם את מספר הקטגוריות מ 19 ל 10 בלבד, יצר אחידות סמנטית, ושיפר את מוכנות העמודה לשימוש במודל.

Light:

עמודת Light, המתארת את תנאי התאורה בעת התאונה (כגון אור יום, חושך עם תאורה, בין ערביים), הכילה ערכים דומים בניסוחים שונים. לדוגמה, Daylight לעומת DAYLIGHT, או dark lights on לעומת dark - lighted. חוסר אחידות זה גרם לריבוי קטגוריות מיותרות. כלל הערכים הומרו לאותיות קטנות (lowercase), רווחים מיותרים בתחילת וסוף הערך הוסרו, ובוצע מיפוי של ערכים סמנטית זהים לניסוח תקני אחד. לדוגמה, dark lights on הומר ל־dark - lighted, ו־dark no lights ל־dark - not lighted. התוצאה היא עמודה נקייה ואחידה, שמאפשרת ניתוח ברור של השפעת תנאי התאורה על חומרת התאונה.

Traffic Control:

העמודה Traffic Control מתארת את אמצעי הבקרה שהיו קיימים בזירת התאונה- כגון תמרורים, רמזורים, שלטים, שוטר תנועה ועוד. בבחינה הראשונית נמצאו בעיות של ריבוי ניסוחים לקטגוריות דומות: No Controls מול NO CONTROLS, Warning Sign מול Other Warning Sign, ושילובים נוספים. התוצאה הייתה פיצול מיותר בין ערכים שמתארים תופעות דומות, והקושי בניתוח לפי סוג בקרה. לצורך פתרון, כלל הערכים הומרו לאותיות קטנות אחידות, והוסרו רווחים מיותרים. לאחר מכן, בוצע מיפוי ערכים בעלי משמעות דומה, לדוגמה: traffic control signal הומר ל־traffic signal, ו־ school zone sign device ל־school zone sign. העמודה כעת אחידה ומפוקחת, מוכנה לניתוח השפעת אמצעי הבקרה על חומרת התאונה.

Driver Substance Abuse:

עמודת Driver Substance Abuse מציינת האם הנהג היה תחת השפעת אלכוהול, סמים או תרופות בעת התאונה, ואם יש חשד לכך או שהנושא לא ידוע. בבדיקה ראשונית נמצאו ערכים טקסטואליים ארוכים, לא אחידים, ולעיתים שילוב של כמה מונחים באותו ערך (למשל: "Suspect of Alcohol Use, Unknown"). כמו כן, בעמודה זו הופיעו מספר רב של ערכים חסרים (נשקלה האפשרות למחוק את עמודה זאת, אך לאחר הרצת המודל עם העמודה ובלעדיו, החלטנו להשאיר את העמודה על מנת לא לאבד מידע שיכול להיות קריטי). לצורך ניקוי, כל הערכים הומרו לאותיות קטנות והוסרו רווחים מיותרים. בנוסף, בוצע מיפוי ערכים שמופיע בהם ניסוח לא ברור או חופף לניסוחים מדויקים ואחידים יותר.
לדוגמה, הערך "unknown, suspect of drug use" הוחלף ל־"alcohol unknown, drug suspected" כדי לשמור על מבנה אחיד של "alcohol X, drug Y" בכל השורות. מטרת הפעולה הייתה להקטין את מספר הקטגוריות ולשפר את היכולת של המודל לזהות דפוסים הקשורים לשימוש באלכוהול או סמים.

Driver Distracted By:

עמודה זו מתארת את הסיבה האפשרית להסחת דעתו של הנהג בזמן התאונה, כגון שימוש בטלפון נייד, שיחה עם נוסעים ועוד. בבדיקה ראשונית נמצא כי הערכים בעמודה אינם אחידים חלקם נכתבו באותיות גדולות או קטנות לצורך ניקוי, כל הערכים הומרו לאותיות קטנות והוסרו רווחים מיותרים.

Drivers License State:

עמודה זו מתארת את מדינת הרישוי של הנהג המעורב בתאונה. בבחינה הראשונית זוהו 77 ערכים ייחודיים, שכללו קודים של מדינות בארצות הברית, טריטוריות שונות ואף מדינות זרות. בנוסף, הכילה ערכים חסרים או הכילו את הקוד XX, שאינו מייצג מדינה תקפה.
לצורך ניקוי וסטנדרטיזציה, בוצע מיפוי של קודי המדינות לשמותיהם המלאים, לדוגמה: MD הומר ל־Maryland, ו־DC ל־District of Columbia. הקוד XX סווג כערך חסר (NaN), מתוך הנחה שמדובר במידע שגוי או שאינו זמין.
כתוצאה מהתהליך, העמודה כעת אחידה, תקנית וברורה, עם שמות מדינות מלאים במקום קודים מקוצרים - ומוכנה לשלב הקידוד והניתוח.

Vehicle Damage Extent:

עמודה זו מתארת את רמת הנזק שנגרם לרכב, כפי שנרשמה בדוח המשטרתי. בבחינה הראשונית זוהו 11 ערכים ייחודיים, כאשר ההבדלים ביניהם נבעו אך ורק מכתיב שונה באותיות גדולות וקטנות - לדוגמה: Disabling לעומת DISABLING, או Superficial מול SUPERFICIAL. לצורך ניקוי, כלל הערכים הומרו לאותיות קטנות אחידות, והוסרו רווחים מיותרים.
כתוצאה מכך, מספר הקטגוריות ירד מ־11 ל־7 בלבד, והעמודה כעת אחידה, ברורה, ומוכנה להמרה לקידוד קטגוריאלי ולניתוח במודל.

Vehicle First Impact Location:

עמודה זו מתארת את מיקום הפגיעה הראשונית ברכב, בהתאם למיקום יחסי על פני שעון (למשל: TWELVE OCLOCK מייצג פגיעה חזיתית).
בבחינה ראשונית זוהו 32 ערכים ייחודיים, כאשר רבים מהם תיארו את אותו מיקום בדיוק אך נרשמו בצורות שונות, לדוגמה: Twelve O Clock לעומת TWELVE OCLOCK, או Six O Clock לעומת SIX OCLOCK.
לצורך אחידות, בוצע מיפוי ידני של הערכים השקולים, כך שכל ערכי השעון נכתבים כעת בצורה אחידה באותיות גדולות (לדוגמה: One O Clock הומר ל־ONE OCLOCK).
כתוצאה מכך, מספר הקטגוריות ירד ל־18 בלבד, והעמודה כעת אחידה, ברורה ומוכנה לקידוד קטגוריאלי, תוך שמירה על מידע שימושי לתיאור מיקום הפגיעה ברכב.

Vehicle Body Type

עמודת Vehicle Body Type מתארת את סוג הרכב המעורב בתאונה- (למשל רכב פרטי, רכב שטח, רכב חירום, משאית, אוטובוס, ועוד). העמודה הכילה 56 קטגוריות שונות, כולל כפילויות שנבעו מהבדלים בין אותיות קטנות וגדולות, ורווחים מיותרים. לדוגמה: Passenger Car לעומת PASSENGER CAR, או UTILITY VEHICLE (SPORT) לעומת Sport Utility Vehicle. לצורך ניקוי הנתונים, הערכים הומרו לאותיות קטנות, רווחים מיותרים הוסרו, ובוצע מיפוי ידני לקטגוריות שקולות – לדוגמה: (sport) utility vehicle הומר ל־sport utility vehicle. בעקבות תהליך זה, מספר הקטגוריות ירד מ־56 ל־44 בלבד, והעמודה הפכה לאחידה, ומוכנה לניתוח השפעת סוג הרכב על חומרת הפציעה.

Vehicle Movement

עמודת Vehicle Movement מתארת את אופן תנועת הרכב בזמן התאונה, כגון תנועה במהירות קבועה, האצה, בלימה, פנייה או חנייה. בבדיקה נמצאו כפילויות שנבעו מהבדלים באותיות גדולות וקטנות (למשל: Moving Constant Speed לעומת moving constant speed), וכן וריאציות ניסוח שונות המתארות את אותה פעולה (כגון: Turning Left לעומת Making Left Turn, או Parked לעומת Parking). לצורך ניקוי הנתונים, כל הערכים הומרו לאותיות קטנות ורווחים הוסרו. בנוסף, בוצע מיפוי ידני לאיחוד ניסוחים שקולים לערכים אחידים. בעקבות כך, מספר הקטגוריות ירד מ־35 ל־21 בלבד. תהליך זה שיפר את עקביות העמודה, הפחית רעש מיותר, וייעל את ניתוח דפוסי התנועה לצורכי חיזוי חומרת התאונה.

Vehicle Year

עמודת Vehicle Year מייצגת את שנת הייצור של הרכב. מדובר במאפיין חשוב, שעשוי להעיד על טכנולוגיות בטיחות, מצב תחזוקתי וסיכון יחסי. בבחינה ראשונית נמצאו בעמודה זו ערכים בלתי סבירים: חלק מהשורות הכילו תווים לא תקניים (כגון 0, 999, 3), חלק כללו שנים מוקדמות מאוד (1900), ואחרות כללו שנים עתידיות מדי (2099). כדי לנקות את העמודה, בוצעה המרה בטוחה של כל הערכים לפורמט מספרי (to_numeric), כאשר כל ערך שלא ניתן להמיר סווג כ־NaN. לאחר מכן הוגדר טווח סביר לשנות ייצור בין 1980 ל־2025 וכל ערך מחוץ לטווח זה הומר גם הוא לערך חסר. כתוצאה מכך, נוספו ערכים חסרים לעמודה, אך איכות הנתונים עלתה באופן משמעותי, כאשר כל ערך שנותר מייצג כעת שנת ייצור אמינה ורלוונטית למידול.

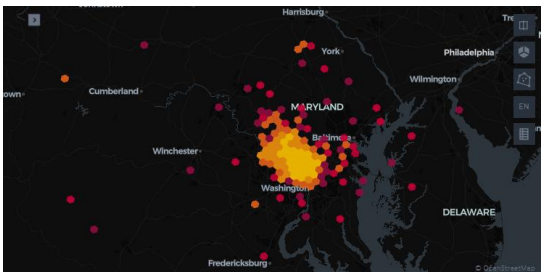
Vehicle Make

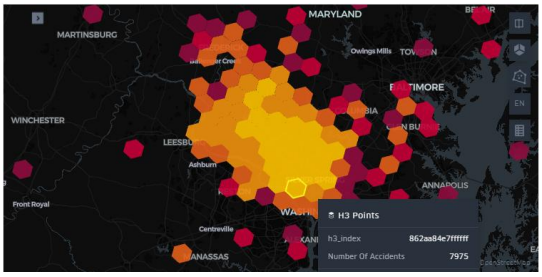
עמודת Vehicle Make, המתארת את יצרן הרכב, נמצאה כבעייתית במיוחד: התגלו בה מעל ל־1,900 ערכים ייחודיים, בשל קיצורים, שגיאות כתיב, תווים חסרים, שמות מותגים חלקיים או בלתי מזוהים. מצב זה יוצר רעש משמעותי, מקשה על מידול תקין ועלול להטעות את המודל.
לצורך ניקוי העמודה בוצעו מספר שלבים: ראשית, כל הערכים הומרו לאותיות גדולות ורווחים מיותרים הוסרו. לאחר מכן בוצע תיקון ידני לערכים שכיחים עם קיצורים או שגיאות נפוצות, למשל: TOYT ו־TOYO הומרו ל־TOYOTA ועוד.
בשלב הבא, כל הערכים בעמודה הושוו לרשימת מותגים תקניים שהוגדרה ידנית על ידינו, רשימה הכוללת את שמות היצרנים המרכזיים הקיימים. ההשוואה התבצעה באמצעות fuzzy matching תוך שימוש במדד הדמיון token_sort_ratio . ערכים שקיבלו ציון התאמה של 80% ומעלה הומרו למותג התקני הקרוב ביותר, בעוד ערכים שאינם עומדים בסף זה סווגו לקטגוריה כללית בשם OTHER. התוצאה היא עמודה נקייה, אחידה וממוקדת, המוכנה לשימוש בשלב המידול תוך צמצום רעש.

```
✅ מספר ערכים ייחודיים לאחר ניקוי: 55
📄 מספר רשומות: 188277
❌ מספר ערכים שהפכו ל-OTHER: 13130
```

Road Name, Cross-Street Name: (שימוש ב2 העמודות Latitude, Longitude):

העמודות Road Name ו־Cross-Street Name, שתיארו את שמות הרחובות בהם התרחשה התאונה (רחוב ראשי ורחוב חציה בהתאמה), נמצאו כעמודות בעלות גיוון קיצוני: אלפי ערכים ייחודיים, כולל שגיאות כתיב, קיצורים, רמות פירוט שונות, וריבוי גרסאות לאותו רחוב. תהליך סטנדרטיזציה לשמות רחובות היה מורכב, עתיר עבודה ולא בהכרח אפקטיבי מבחינה תחזיתית. לכן, הוחלט על גישה שונה: מיקום התאונה ייוצג באמצעות קואורדינטות (Latitude ו־Longitude) בשילוב ספריית H3, מערכת אינדוקס מרחבית של חברת Uber. באמצעות הפונקציה h3.latlng_to_cell והגדרת רזולוציה 6, המרות גאוגרפיות בוצעו לכל התצפיות, ונוצרה עמודה חדשה בשם h3_index, אשר מייצגת את האזור הגאוגרפי של התאונה בצורה אחידה. לאחר מכן, הוסרו העמודות Road Name, Cross-Street Name, Latitude, ו־Longitude. פעולה זו הפחיתה רעש, פתרה את בעיית חוסר האחידות, ושיפרה את ההכנה הגאוגרפית של הנתונים לצרכי מידול.





יצירת נתונים חדשים

במהלך שלב הכנת הנתונים, נגזרו מספר תכונות חדשות (עמודות) מתוך השדות הקיימים, במטרה לשפר את ייצוג המידע, לצמצם רעש ובעיות ייחודיות מיותרות, ולהנגיש למודלים מידע סמוי בצורה מובנית וברורה. לא נוספו רשומות חדשות (שורות) לסט הנתונים המקורי בשלב זה, אך בשלב המידול נעשה שימוש באלגוריתם SMOTE לצורך יצירת דוגמאות סינתטיות לסט האימון בלבד, פירוט נוסף בהמשך.

הפקת תכונות חדשות

פירוק תאריך התאונה (Crash Date/Time):
כיוון שרוב האלגוריתמים שחשבנו להשתמש בהם אינם תומכים בטיפוס תאריך (datetime), בוצע פיצול של שדה זה לשלוש תכונות נפרדות:
Crash_Month - החודש שבו התרחשה התאונה
Crash_DayOfWeek - היום בשבוע בו התרחשה התאונה
Crash_Hour - השעה ביום שבה התרחשה התאונה
פירוק זה אפשר למודלים ללמוד דפוסים עונתיים, יומיים ושעתיים, ולזהות זמני סיכון אופייניים.

התכונה h3_index:
h3_index נוצרה מתוך עמודות קואורדינטות גולמיות (Latitude, Longitude) באמצעות ספריית H3, וממפה את מיקום התאונה לאזור מרחבי מוגדר.
השימוש באזורים במקום בקואורדינטות ישירות נועד להפחית את רמת הייחודיות הגבוהה שהייתה בנתוני המיקום המקוריים, ולאפשר למודל לזהות דפוסים מרחביים מבלי להיות רגיש לרמות דיוק מיותרות.
תכונה זו החליפה את העמודות:
Road Name (4383 ערכים ייחודיים), Cross-Street Name (כ־7250 ערכים ייחודיים), Latitude ו־Longitude.
הנתונים נאספו במדינת מרילנד בארה"ב, בה שמות רחובות יכולים לייצג מקטעים עצומים, כך ששמות הרחובות עצמם אינם מספקים אינדיקציה מדויקת למיקום התאונה, אלא יוצרים רעש וממדיות מיותרת.
לעומת זאת, h3_index מזהה אזור גאוגרפי לפי ספריית H3 ברזולוציה 6, המאפשר קיבוץ מרחבי אפקטיבי יותר. תכונה זו החליפה את כלל עמודות המיקום המקוריות, והציגה ייצוג מרחבי יעיל ונקי מרעש. בנוסף, היא מספקת מיקוד אזורי ברור ומסייעת למודל להבין טוב יותר את השפעת המיקום הגאוגרפי על חומרת הפציעה.

יצירת רשומות חדשות

אף שלא נוצרו רשומות חדשות במהלך שלב ההכנה הראשוני, בשלב בניית המודלים נעשה שימוש באלגוריתם SMOTE לצורך איזון מחלקות. האלגוריתם יצר דוגמאות סינתטיות חדשות ממחלקות מיעוט, אך ורק עבור סט האימון (Training Set). לא נוסף מידע חדש לקובץ הנתונים המקורי. התהליך בוצע כחלק בלתי נפרד מהכנת הדאטה למידול בלבד.

התאמות לדרישות האלגוריתם

כחלק מתהליך הפקת התכונות והכנת הנתונים למידול, נשקלו דרישות טיפוסיות של אלגוריתמים שונים ללמידת מכונה. אף ששלב המידול המלא טרם הושלם, נערכו ניסויים מקדימים שנועדו להכוון את תהליך ההכנה, ובפרט לאפשר קבלת החלטות מושכלות בנוגע לבחירת מאפיינים, טיפוסים וסוגי קידוד.

בהתבסס על ניסיונות ראשוניים והיכרות עם האלגוריתמים השכיחים שבכוונתנו לבחון (LightGBM, Random Forest, XGBoost,Logistic Regression, CatBoost), התקבלה ההבנה הבאה:

- מרבית האלגוריתמים דורשים קלט מספרי בלבד, ולכן בוצעה המרה של משתנים קטגוריאליים באמצעות שיטות שונות, כגון One-Hot Encoding, Label Encoding. בנוסף, נעשה שימוש בהגדרת משתנים כ־category אשר הומרו לערכים מספריים זמניים במידת הצורך (למשל לצורך התאמה ל־SMOTE) בהתאם למודל שייבחר.
- מאחר שרוב האלגוריתמים אינם תומכים ישירות בטיפוס datetime, העמודה פורקה למרכיבים מספריים (שעה, יום, חודש) כדי להפוך את המידע לזמין ללמידה.
- נרמול של המשתנים לא בוצע עבור מרבית המודלים, מאחר והם מבוססי עצים ואינם רגישים להבדלי סקאלה. בנוסף, רוב המשתנים בדאטה הם קטגוריאליים, למעט שתי עמודות מספריות. חריג לכך הוא מודל הרגרסיה הלוגיסטית ,שבו כן נדרש נרמול שבוצע באמצעות- StandardScaler, מאחר והוא רגיש לסקאלת המשתנים
- לא בוצעה הפקה של תכונות חדשות באמצעות איגוד או ממוצע, אך כן נגזרו תכונות חדשות מתוך עמודות מקוריות כמו תאריך ומיקום- כמפורט לעיל.

כל שלב ההכנה התבצע מתוך כוונה לשמור על גמישות, לאפשר התאמה למספר אלגוריתמים אפשריים, ולייעל את שלב המידול שיבוא בהמשך.

אינטגרציית נתונים

בפרויקט זה נעשה שימוש במקור נתונים אחד: קובץ CSV שהופק ממאגר ממשלתי של מדינת מרילנד (Data.gov), המכיל דיווחים על תאונות דרכים. לא נדרשה אינטגרציה של טבלאות ממקורות חיצוניים, ולא בוצע מיזוג בין מקורות נתונים שונים. עם זאת, בוצעה אינטגרציה פנימית בין עמודות מתוך אותו קובץ. לדוגמה, יצירת מזהה מרחבי (h3_index) על בסיס קואורדינטות גולמיות (Latitude ו־Longitude), במטרה לשפר את הייצוג הגאוגרפי של המידע.

עיצוב מחדש של נתונים

לקראת שלב המידול, הנתונים הותאמו לפורמט הנדרש על ידי האלגוריתמים השונים, תוך שמירה על מבנה עקבי ונקי. בשלב זה נשקל שימוש במספר אלגוריתמים ללמידת מכונה, ביניהם : LightGBM, XGBoost, CatBoost, Random Forest ו-Logistic Regression.
מרבית המודלים דורשים קלט מספרי ולעיתים גם קידוד אחיד או הסרת משתנים שאינם תורמים. בהתאם לכך, בוצעו הפעולות הבאות:

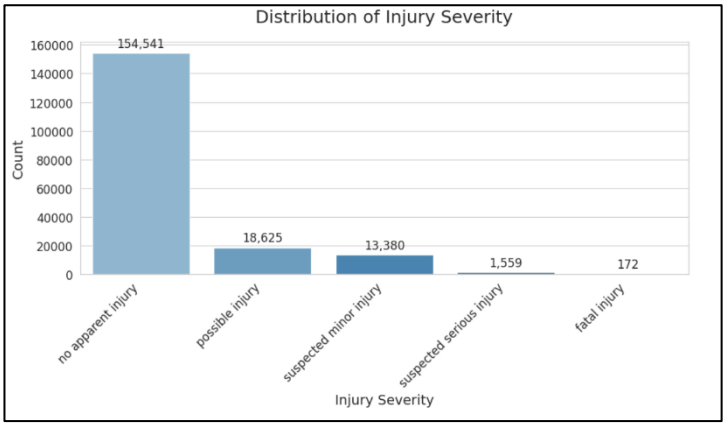
<u>פעולות עיצוב שבוצעו בפועל:</u>

- טיפול במשתנים קטגוריאליים: כל העמודות הקטגוריאליות הומרו לערכים מספריים על ידי: One-Hot Encoding, Label Encoding, או המרתן לטיפוס category בשילוב cat.codes.
- קידוד עמודת המטרה: עמודת Injury Severity הומרה לערכים מספריים באמצעות LabelEncoder.
- איזון מחלקות: בשל חוסר איזון בין קבוצות המטרה (למשל, פציעות קטלניות שהן נדירות יחסית), נעשה שימוש באלגוריתם SMOTE לצורך יצירת דוגמאות סינתטיות לאחר חלוקת הנתונים ורק על קבוצת האימון.
- המרת תאריך: עמודת התאריך פורקה למאפיינים חדשים (Crash_Hour, Crash_DayOfWeek, Crash_Month) אשר צורפו לפיצ'רים והומרו לערכים מספריים.

ניתוח נתונים ראשוני (EDA):

לאחר ביצוע כל השלבים, המערך הסופי לאחר ניקוי וטרנספורמציות כולל 188,277 שורות ו-25 עמודות. תחילה בוצע ניתוח חקרני של עמודות הנתונים המרכזיות, תוך בחינת ההתפלגות הפנימית של כל עמודה באופן עצמאי.

<u>גרף התפלגות משתנה המטרה</u>
משתנה המטרה במערך הנתונים הוא Injury Severity, אשר מתאר את רמת חומרת הפציעה כתוצאה מהתאונה.
הגרף והטבלה שלפנינו מציגים את התפלגות הנתונים לפי קטגוריות שונות של פציעה:




```
Injury Severity
no apparent injury          154541
possible injury              18625
suspected minor injury       13380
suspected serious injury      1559
fatal injury                   172
Name: count, dtype: int64
```

כפי שניתן לראות, הרוב המכריע של התצפיות משתייכות לקטגוריית "No Apparent Injury" עם 154,541 מקרים מתוך כ-189,277.
לעומת זאת, קטגוריות של פציעות חמורות או קטלניות נדירות הרבה יותר (1,559 ו- 172 מקרים בהתאמה), מה שממחיש את חוסר האיזון הבולט בין הקטגוריות.
תופעה זו עשויה להשפיע על ביצועי האלגוריתמים, מאחר קיימת נטייה למודלים להעדיף את הקטגוריה השכיחה ביותר.
בהמשך בוצעה התמודדות עם חוסר איזון זה באמצעות שימוש באלגוריתם SMOTE.

סטטיסטיקות תיאוריות של משתנים מספריים

להלן טבלת סיכום סטטיסטי של שני המאפיינים המספריים במערך הנתונים: Speed Limit ו-Vehicle Year.
עבור כל משתנה מוצגים: מספר התצפיות (count), ממוצע (mean), סטיית תקן (std), ערך מינימלי (min), רבעון ראשון (25%), חציון (50%), רבעון שלישי (75%) והערך המקסימלי (max).

סטטיסטיקות תיאוריות עבור משתנים מספריים

	count	mean	std	min	25%	50%	75%	max
Speed Limit	188277.00	32.42	11.13	0.00	25.00	35.00	40.00	75.00
Vehicle Year	188277.00	2011.01	6.35	1980.00	2007.00	2012.00	2016.00	2025.00

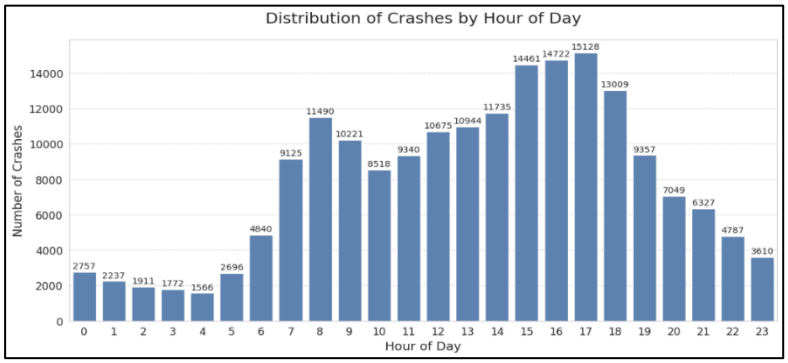
- Speed Limit: ערכי המהירות המותרת בכביש נע, בין 0 ל־75 קמ"ש, כאשר הממוצע הוא 32.42 קמ"ש וסטיית התקן 11.13.
- Vehicle Year: שנות הייצור של הרכבים נעות בין 1980 ל־2025 (טווח שהגדרנו בתהליך הניקוי), עם ממוצע של שנת 2011.01 וסטיית תקן של 6.35 שנים.

נתונים אלו מאפשרים להבין את הפיזור והמאפיינים הבסיסיים של הנתונים, ומהווים בסיס לניתוחים נוספים בשלב המידול.

גרפים לניתוח מאפייני זמן

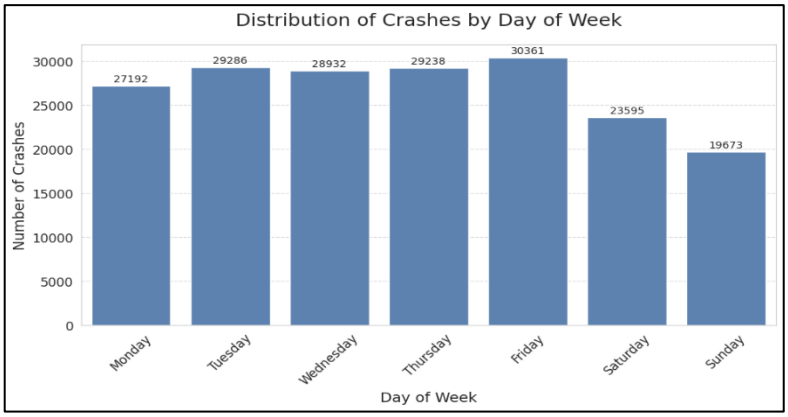
מאפייני הזמן במערך הנתונים נגזרו מתוך שדה התאריך המקורי וכוללים את שעת התאונה (Crash Hour), יום התאונה (Crash Day) וחודש התאונה (Crash Month).
מאחר ומדובר בנתונים מחזוריים וקטגוריאליים סדורים, לא בוצעה עבורם סטטיסטיקה תיאורית רגילה (כגון ממוצע, חציון, סטיית תקן), במקום זאת, הוצגו טבלאות שכיחויות וגרפים שמטרתם להמחיש את התפלגות התאונות על פני צירי הזמן.
מוצגת התפלגות התאונות לפי:

- שעה ביממה



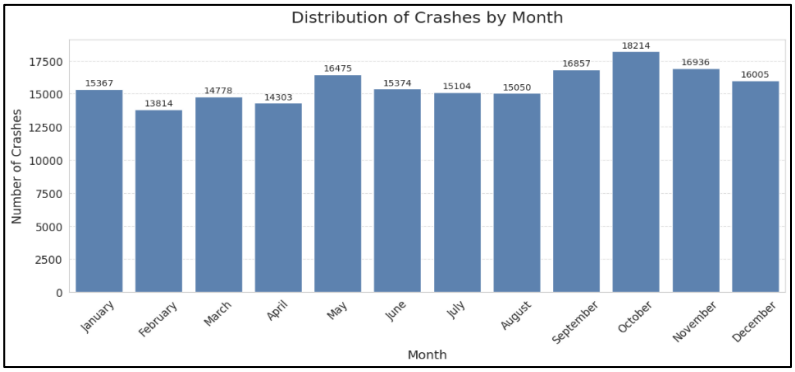
התפלגות התאונות לפי שעה ביום מראה כי שעות השיא מבחינת כמות התאונות הן בין השעה 15:00 ל־18:00, עם שיא בולט סביב השעה 17:00. ייתכן וניתן לייחס תבנית זו לשעות העומס בכבישים, ונכון להביאו בחשבון בעת בניית תוכניות לשיפור בטיחות בדרכים והגברת אכיפה בשעות רגישות אלו. לעומת זאת, בשעות הלילה המאוחרות (0:00–5:00) נרשמות רמות נמוכות יותר של תאונות, מה שמעיד על תנועה דלילה בכבישים בשעות אלו.

- יום בשבוע



הגרף מציג את התפלגות התאונות לפי ימי השבוע (יום שני מיוצג ב-0 ויום ראשון ב- 6 בהתאם לברירת המחדל המקובלת). מהתפלגות הנתונים עולה כי מספר התאונות גבוה יותר בימי חול, במיוחד בימים שני עד שישי, ונמוך יחסית בסופי השבוע (שבת וראשון). מגמה זו עשויה להעיד על נפח תנועה גבוה יותר בימי עבודה, בהשוואה לימי סוף השבוע שבהם הכבישים פחות עמוסים.

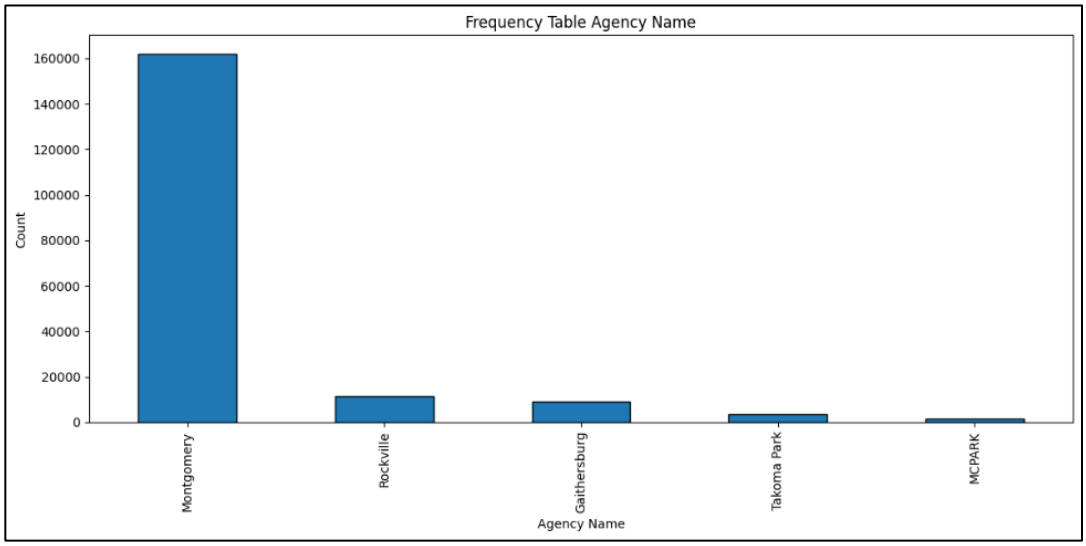
- חודש בשנה



הגרף מציג את התפלגות התאונות לאורך חודשי השנה. ניתן לראות שהפערים במספר התאונות בין החודשים אינם גדולים במיוחד, כאשר החודשים בהם התרחשו הכי הרבה תאונות הם אוקטובר (18,214), ספטמבר (16,857) ונובמבר (16,936). לעומתם, חודש פברואר רשם את מספר התאונות הנמוך ביותר (13,814), ככל הנראה בשל מספר ימי הפעילות הקצר בו.

<u>גרף התפלגות התאונות לפי שם הסוכנות המדווחת (Agency Name)</u>

לצורך ניתוח חקרני של הנתונים, נבחנו ערכי העמודה Agency Name, המייצגת את שם הסוכנות המדווחת בכל רשומת תאונה. העמודה היא קטגוריאלית וכוללת חמש קטגוריות בלבד.

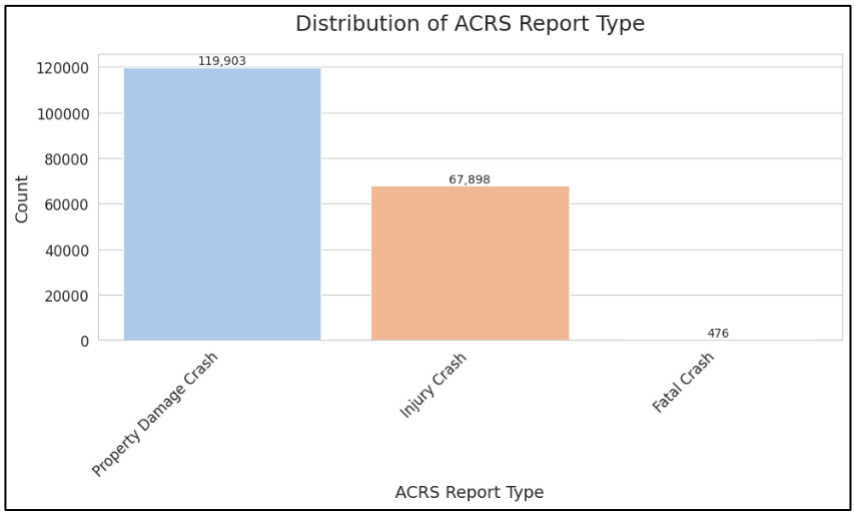


	Agency Name	Count	Percentage
0	Montgomery	162731	86.43
1	Rockville	11548	6.13
2	Gaithersburg	9223	4.90
3	Takoma Park	3428	1.82
4	MCPARK	1347	0.72

הניתוח מצביע על פיזור לא אחיד: כ־86% מהרשומות בדאטה דווחו על ידי Montgomery, בעוד יתר הסוכנויות אחראיות על שיעור קטן מהנתונים (למשל MCPARK – פחות מ־1%). ייתכן שהפער נובע מהבדלים בגודל האזור הגאוגרפי שבאחריות כל סוכנות, או בהבדלים במדיניות הדיווח.

גרף התפלגות התאונות לפי סוג הדיווח (ACRS Report Type)

העמודה ACRS Report Type מסווגת את התאונות לשלוש קטגוריות:
Property Damage Crash- תאונה עם נזק לרכוש בלבד
Injury Crash- תאונה שבה נגרמה פציעה
Fatal Crash- תאונה קטלנית



מהגרף עולה כי מרבית הדיווחים (119,903) מתייחסים לנזק רכוש בלבד, בעוד ש- 67,898 מהדיווחים כוללים פציעות. מספר התאונות הקטלניות נמוך ביותר ועומד על 476 בלבד מכלל המקרים, דבר הממחיש את נדירותן ואת חוסר האיזון הקיים בין סוגי הדיווח.
בהמשך תיבחן תרומתה של עמודה זו למודל החיזוי באמצעות ניתוח הקשר בינה לבין חומרת הפציעה.
.

טבלת התפלגות התאונות לפי סוג הדרך (Route Type)

עמודת Route Type מייצגת את סוג הדרך בה התרחשה התאונה (למשל: דרך מדינתית, דרך מחוזית, דרך עירונית, דרך פרטית, מסלול אופניים וכו'). להלן טבלת השכיחויות:

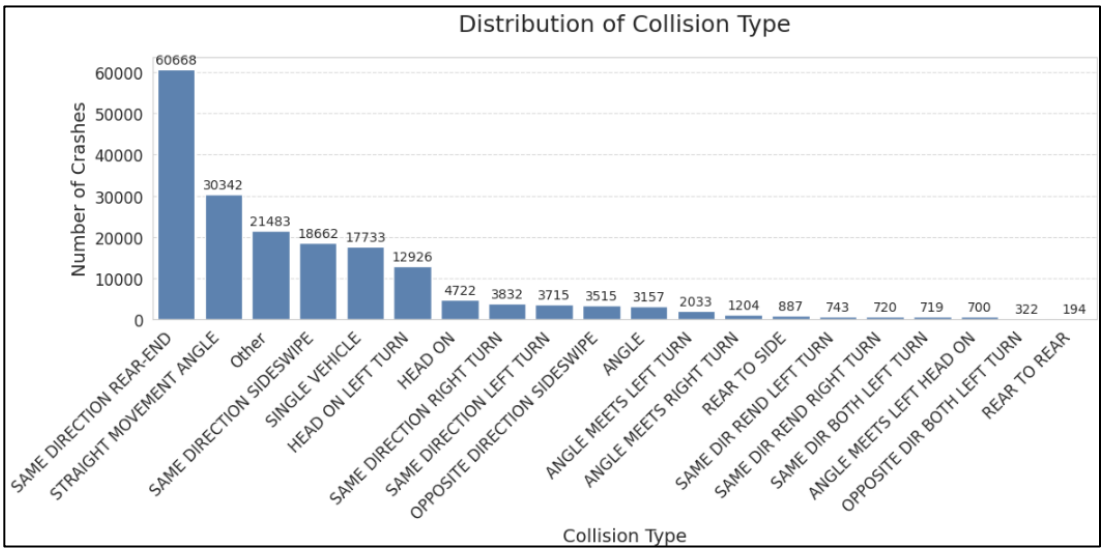
טבלת שכיחויות לעמודה Route Type:

	Route Type	Count	Percentage
0	Maryland (State) Route	82690	43.92
1	County Route	61362	32.59
2	Other	18262	9.70
3	Municipality Route	10555	5.61
4	US (State)	8278	4.40
5	Interstate (State)	3363	1.79
6	Other Public Roadway	1261	0.67
7	Ramp	878	0.47
8	Government Route	696	0.37
9	Local Route	309	0.16
10	Bicycle Route	197	0.10
11	Spur	148	0.08
12	Private Route	131	0.07
13	Crossover	100	0.05
14	Service Road	47	0.02

מהטבלה עולה שהתפלגות סוגי הדרכים אינה אחידה. מרבית התאונות התרחשו על שני סוגי דרכים עיקריים: " Maryland (State) Route" עם כ- 43.9% מהמקרים, ו־"County Route" עם כ- 32.6%. כ־9.7% מהתאונות סווגו כ־"Other", קטגוריה שאיחדה ערכים חסרים. לעומתם, קטגוריות אחרות כמו "Municipality Route" או "US (State)" מופיעות בשכיחויות נמוכות יחסית.
ואילו קטגוריות נדירות במיוחד (כמו Service Road, Crossover, Private Route) מהוות פחות מ־0.1% מכלל הנתונים.

גרף התפלגות התאונות לפי סוגי ההתנגשות (Collision Type)

עמודת Collision Type מסווגת את סוג ההתנגשות שהתרחשה בתאונה, לפי כיווני תנועת הרכבים המעורבים.

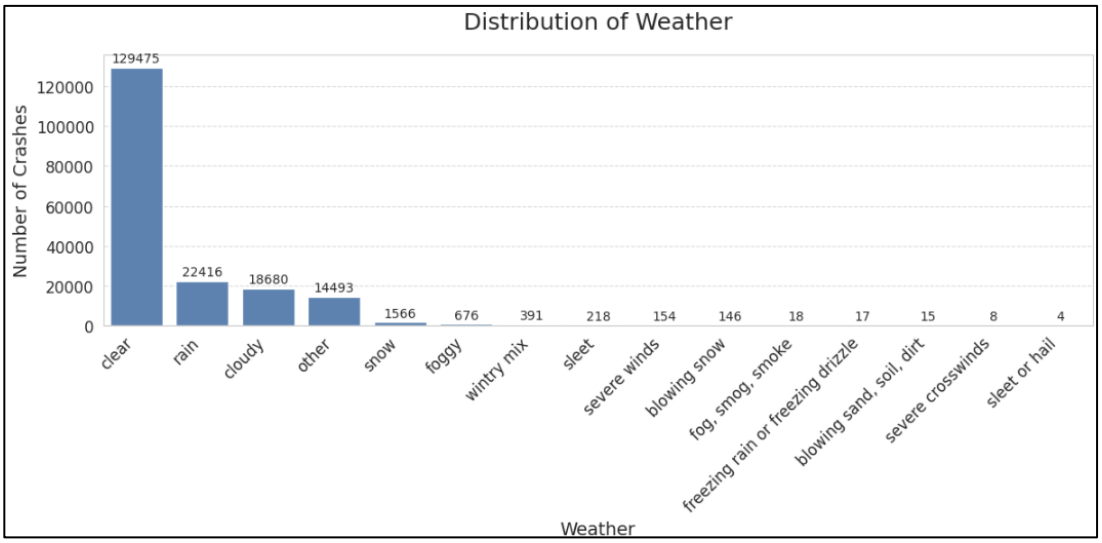


מהגרף עולה כי סוג ההתנגשות הנפוץ ביותר הוא פגיעה מאחור בתנועה באותו כיוון (SAME DIRECTION REAR-END), עם למעלה מ 60 אלף מקרים. סוגים נוספים בעלי שכיחות גבוהה הם התנגשות בזווית בתנועה ישרה (STRAIGHT MOVEMENT ANGLE). וקטגוריית Other, שכוללת מקרים שלא סווגו לאחת הקטגוריות המוגדרות וגם ערכים חסרים ששייכנו לקבוצה זו. לעומת זאת, סוגי התנגשות כגון REAR TO REAR ו־ OPPOSITE DIRECTION BOTH LEFT TURN נדירים יחסית ומופיעים בפחות מ-500 מקרים.

הבנת התפלגות סוגי ההתנגשויות תורמת לזיהוי דפוסים עיקריים, ומספקת בסיס להמשך ניתוח חומרת הפגיעות.

גרף התפלגות התאונות לפי תנאי מזג האוויר (Weather)

עמודת Weather מתארת את תנאי מזג האוויר בזמן התרחשות התאונה, כפי שדווחו על ידי גורמי האכיפה או המעורבים באירוע.



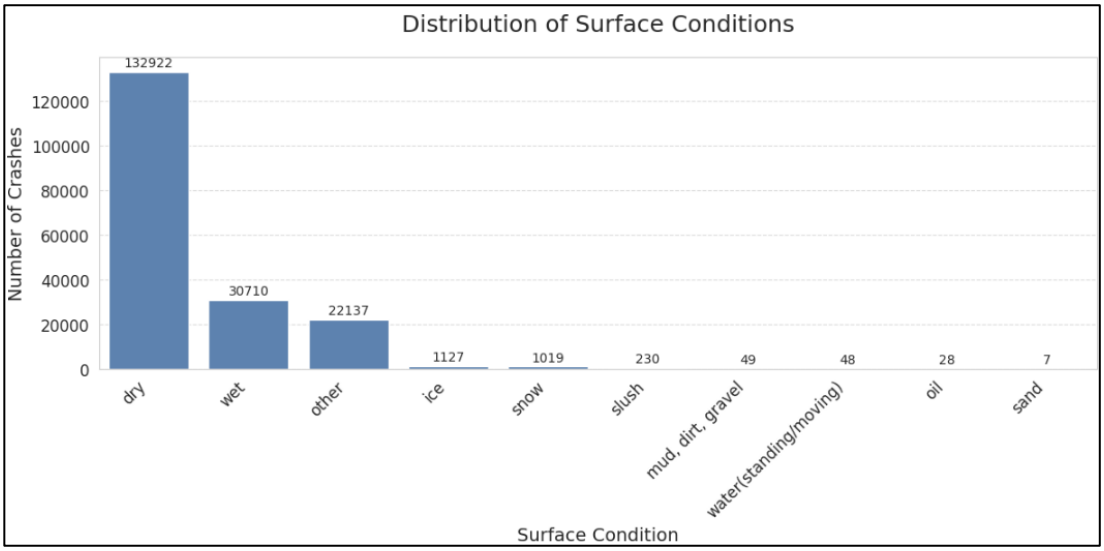
מהגרף עולה כי מרבית התאונות התרחשו בתנאים של שמיים בהירים (clear), עם יותר מ־120,000 מקרים. קטגוריות נוספות בעלות שכיחות משמעותית הן גשם (rain) וענגות (cloudy), עם מעל 22,000 ומעל 18,000 תאונות בהתאמה.

לעומת זאת, תנאי מזג אוויר קיצוניים יותר, כמו שלג (snow), ערפל (foggy), גשמים קפואים (sleet) או רוחות חזקות (severe winds) מופיעים בשכיחות נמוכה בהרבה. ומקרים נדירים במיוחד כמו: טפטוף קפוא (freezing drizzle) ורוחות צד עזות (severe crosswinds),כמעט ואינם נוכחים במערך, מופעים בפחות מ-20 מקרים.

נתונים אלו מדגישים כי ריבוי התאונות נגרם דווקא בתנאים שנחשבים נוחים או ניטרליים לנהיגה ולאו דווקא בתנאים מסוכנים. ייתכן בשל תחושת ביטחון-יתר של הנהגים במזג אוויר נוח מה שעלול להוביל להפחתת ערנות וזהירות.

גרף התפלגות התאונות לפי מצב פני הכביש (Surface Condition)

עמודת Surface Condition מתארת את מצב פני הכביש בעת התאונה, ומכילה מידע על תנאי האחיזה של הכביש, כגון: כביש יבש, רטוב, מכוסה קרח, שלג, וכו'.



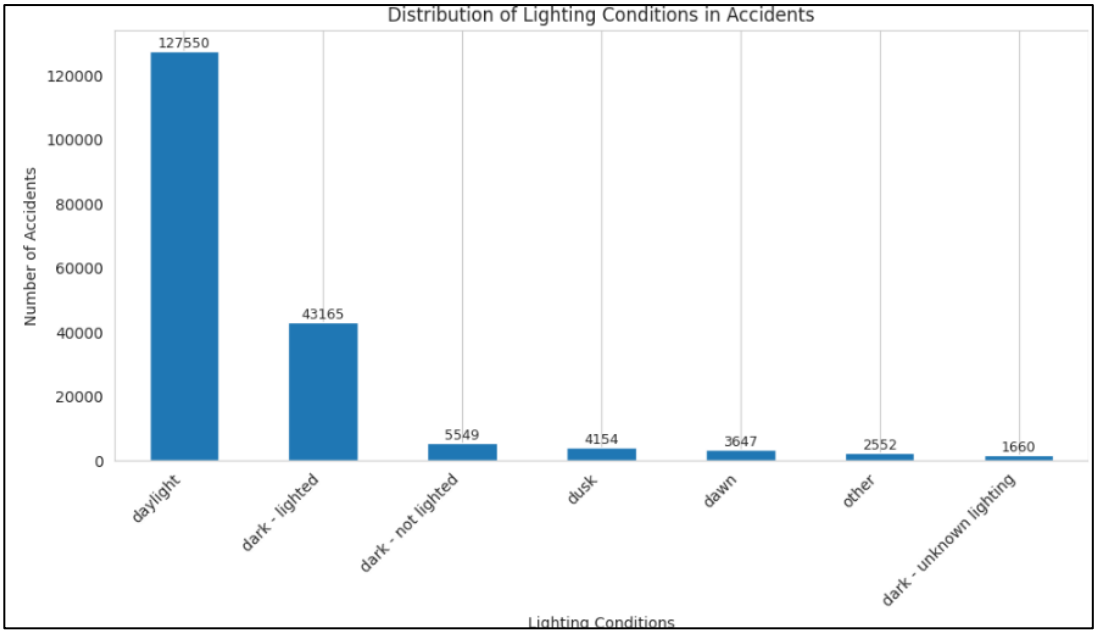
מהגרף עולה כי מרבית התאונות התרחשו כאשר הכביש היה יבש (dry), עם למעלה מ־130,000 מקרים. כלומר, רוב התאונות התרחשו בתנאי דרך "תקינים". קטגוריית wet מופיעה במקום השני עם כ־30,710 תאונות. אחריה, תנאים קיצוניים יותר כמו ice, snow ו־slush, מופיעים בשכיחויות נמוכות משמעותית.

תנאים חריגים נוספים, כגון "oil", "mud, dirt, gravel" או "water(standing/moving)" כמעט ואינם נוכחים במערך, עם עשרות תצפיות בלבד.

ממצאים אלו מחזקים את ההשערה כי רוב התאונות אינן נובעות מתנאי דרך קיצוניים, אלא ככל הנראה מגורמים התנהגותיים, כגון: נהיגה במהירות לא מותאמת, חוסר תשומת לב או הסחות דעת, גם כאשר תנאי הדרך מאפשרים אחיזה טובה.

גרף התפלגות התאונות לפי תנאי התאורה (Light)

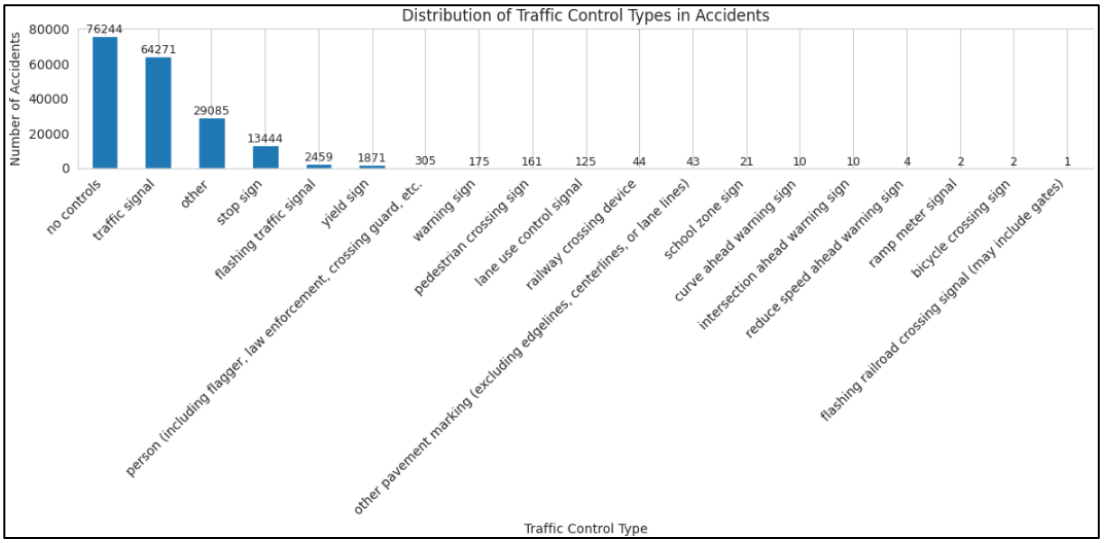
עמודה זו מתארת את תנאי התאורה בעת התאונה, לדוגמה: אור יום, חשכה, תאורה מלאכותית, דמדומים וכו'.



מהגרף ניתן לראות כי מרבית התאונות התרחשו בתנאי תאורה של אור יום (daylight), בפער משמעותי משאר הקטגוריות(פי 3 מהבאה אחריה). הקטגוריה הנפוצה אחריה היא חושך - מואר (dark - lighted). לעומתן, יתר תנאי התאורה כגון חושך - לא מואר, דמדומים (dusk) ושחר (dawn) מופיעים בשכיחות נמוכה בהרבה. ממצא זה מצביע על כך שתאונות מתרחשות בשכיחות גבוהה דווקא בשעות האור יום, ייתכן בשל נפח תנועה גבוה יותר בשעות היום.

גרף התפלגות התאונות לפי אמצעי בקרת התנועה (Traffic Control)

עמודה זו מתארת את סוג אמצעי בקרת התנועה והניהול שהיו זמינים בזירת התאונה, כגון תמרורים, רמזורים, מעגלי תנועה ואמצעים אחרים.

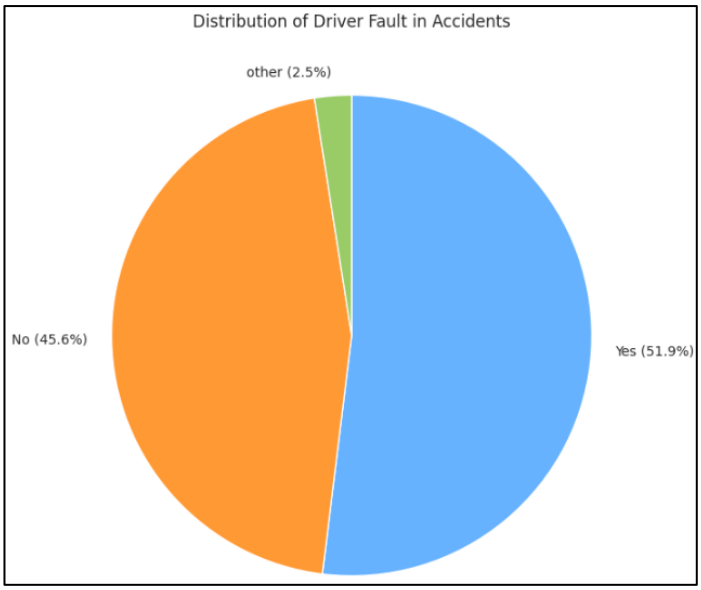


מהגרף עולה כי ברוב מקרי התאונות (כ־76,244) לא היה כלל אמצעי בקרה (no controls) באזור התאונה. לאחר מכן מופיעים ברשימה רמזורים (traffic signal) עם כ־64,271 מקרים, ותמרורי עצור (stop sign) עם כ־13,444 מקרים. אמצעים נוספים כמו, תמרורי מתן זכות קדימה, תמרורי אזהרה, שלטים להולכי רגל ועוד.. מופיעים בשכיחויות נמוכות בהרבה, לעיתים עשרות מקרים בלבד.

נתון זה עשוי להעיד על כך שהיעדר אמצעי בקרה או הכוונה תורם לעלייה בשכיחות התאונות, ייתכן בשל חוסר ודאות בקרב הנהגים או קושי בקבלת החלטות במצבים לא מוסדרים.

גרף התפלגות התאונות לפי אחריות נהג (Driver At Fault)

העמודה Driver At Fault מציינת האם נהג הרכב היה האחראי לתאונה. נתון זה מאפשר להבין את חלקו של נהג הרכב בהתרחשות התאונה.



Driver At Fault	
Yes	97809
No	85786
other	4682

מהגרף עולה כי בכ־51.9% מהמקרים הנהג נמצא כאחראי לתאונה, לעומת כ־45.6% מהמקרים שבהם הנהג לא נמצא אשם. ממצא זה מצביע על כך שיותר ממחצית מהתאונות נובעות מגורמים הקשורים להתנהגות הנהג, כגון: חוסר זהירות, עבירות תנועה, או הסחות דעת. (other מייצג ערכים חסרים).

יחד עם זאת, יש לשים לב שהפער בין הקבוצות אינו קיצוני. המשמעות היא שגם לתנאים חיצוניים (כמו מזג אוויר, תשתיות, תאורה, תנאי דרך, או טעויות של נהגים אחרים) יש השפעה לא מבוטלת.

טבלת התפלגות התאונות לפי מקור הסחת דעת הנהג (Driver Distracted By)

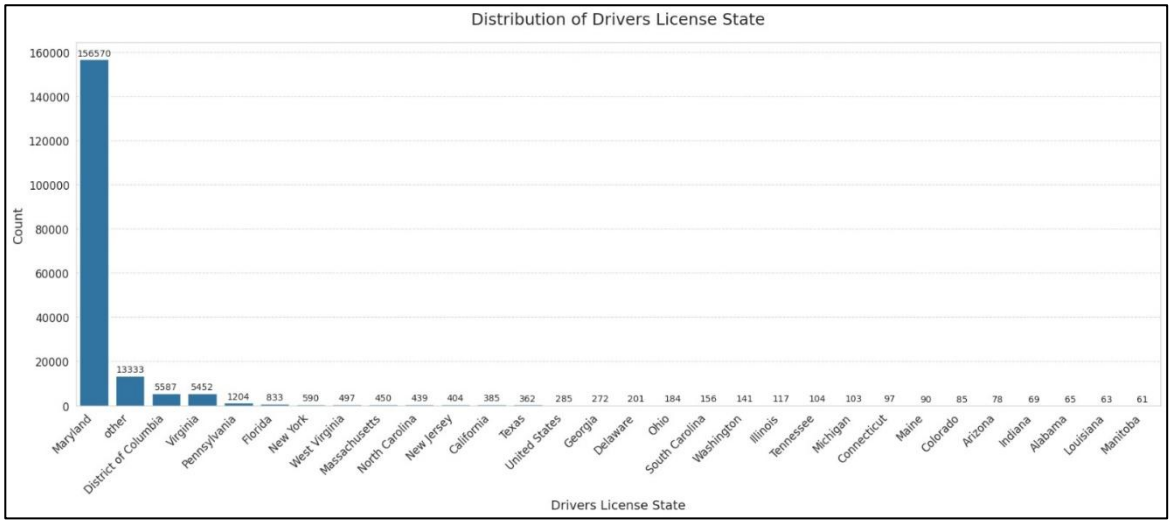
עמודה זו מתארת את הגורם שגרם להסחת דעת הנהג בעת התאונה, כגון שימוש בטלפון נייד, הפרעות מנוסעים, אכילה ושתייה או הסחות דעת אחרות. ניתוח עמודת הסחת דעת מספק תובנות חשובות על דפוסי התנהגות של נהגים.

Driver Distracted By	
not distracted	118108
other	36973
looked but did not see	21546
inattentive or lost in thought	4320
other distraction	3245
distracted by outside person object or event	970
other action (looking away from task, etc.)	541
by other occupants	418
other cellular phone related	372
other electronic device (navigational palm pilot)	325
talking or listening to cellular phone	282
no driver present	280
by moving object in vehicle	212
eating or drinking	196
adjusting audio and or climate controls	134
using other device controls integral to vehicle	92
texting from a cellular phone	67
using device object brought into vehicle	63
dialing cellular phone	48
talking/listening	35
smoking related	26
manually operating (dialing, playing game, etc.)	24
Name: count, dtype: int64	

מהנתונים עולה כי ברוב מקרי התאונות (כ-118,108 מקרים) הנהגים לא היו מוסחים כלל בזמן התאונה. מבין הנהגים שהיו מוסחים, הגורמים העיקריים להסחת דעת היו:
הסתכלות מבלי לזהות (looked but did not see),
חוסר ריכוז או היסחפות במחשבות (inattentive or lost in thought),
והסחות חיצוניות כמו אנשים או עצמים בסביבה (other distraction, distracted by outside person, object or event).
גורמים כגון שימוש בטלפון נייד או במכשירים אלקטרוניים הופיעו בשכיחות נמוכה יחסית.
הממצאים מצביעים על כך שלא רק גורמים טכנולוגיים אלא גם חוסר ערנות רגעי מהווים סיכון משמעותי להתרחשות תאונה.

גרף התפלגות התאונות לפי מדינת רישוי הנהג (Drivers License State)

עמודה זו מתארת את המדינה או הטריטוריה שהנפיקה את רישיון הנהיגה של הנהג המעורב בתאונה. מידע זה יכול להצביע על מגמות גאוגרפיות ולזהות אם יש דפוסים בין תושבי מדינות שונות לבין מעורבות בתאונות.



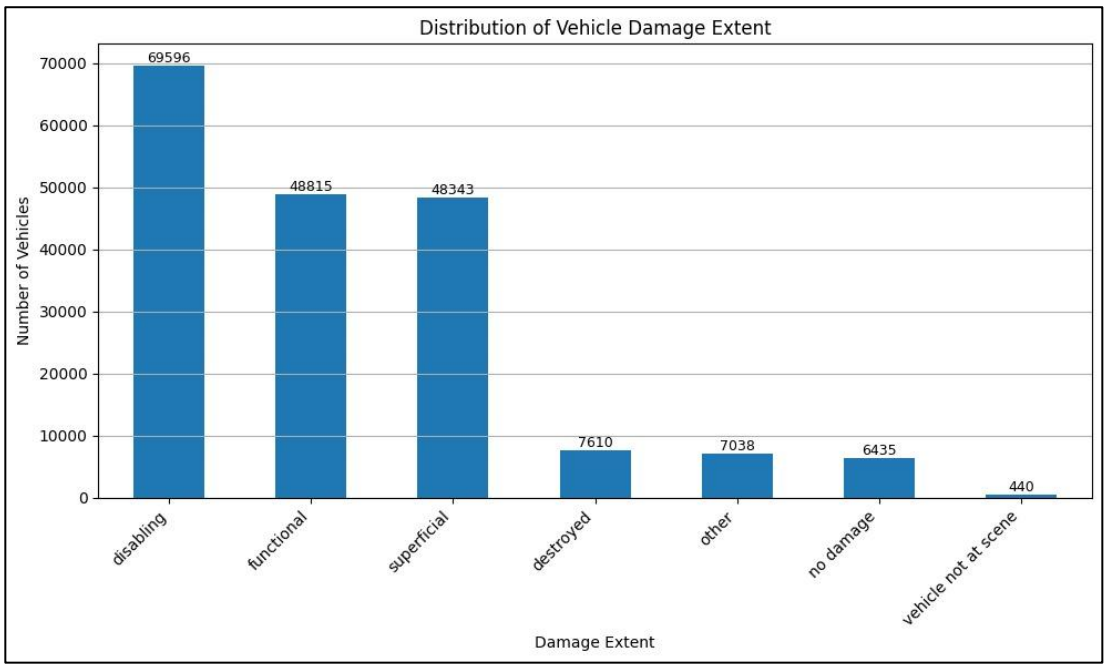
Drivers License State	
Maryland	156570
other	13333
District of Columbia	5587
Virginia	5452
Pennsylvania	1204
Florida	833
New York	590
West Virginia	497
Massachusetts	450
North Carolina	439
New Jersey	404
California	385
Texas	362
United States	285
Georgia	272
Delaware	201
Ohio	184
South Carolina	156
Washington	141
Illinois	117
Tennessee	104
Michigan	103
Connecticut	97
Maine	90
Colorado	85
Arizona	78
Indiana	69
Alabama	65
Louisiana	63
Manitoba	61

עמודה זו מציינת את המדינה או הטריטוריה שהנפיקה את רישיון הנהיגה לנהג המעורב בתאונה. ניתוח התפלגות הנתונים מצביע על כך שרוב מוחלט של הנהגים מחזיקים ברישיון ממדינת מרילנד (Maryland), עם פער משמעותי לעומת מדינות נוספות כמו District of Columbia ו־Virginia. שאר המדינות מופיעות בשכיחויות נמוכות מאוד, כאשר רבות מהן כוללות פחות מ־100 מקרים, וחלקן אף פחות מ־10. נתון זה מתיישב עם העובדה שהנתונים נאספו בעיקר ממדינת מרילנד.

על מנת לצמצם רעש ולשפר את מבנה הנתונים, אוחדו קטגוריות נדירות (פחות מ־50 מופעים) תחת הקטגוריה "other" המכילה גם ערכים חסרים שהחלטנו לשים תחתיה.
בבדיקה שבוצעה על מודל לחיזוי פציעה לעומת אין פציעה נמצא כי האיחוד לא השפיע על מדדי הביצוע המרכזיים (precision, recall, accuracy), אך תרם לשיפור קל בזיהוי מקרי פציעה: ירידה במספר המקרים שלא זוהו FN (יש פציעה והחיזוי פספס אותה: מ־715 ל־709, לצד עלייה זניחה בטעות מסוג FP (אין פציעה שזוהתה כפציעה: מ־5874 ל־5879).

גרף התפלגות התאונות לפי היקף הנזק לרכב (Vehicle Damage Extent)

עמודה זו מתארת את דרגת הנזק שנגרמה לרכב בעקבות התאונה, החל מנזק שטחי ועד לנזק כבד/טוטאלוס או היעדר נזק. ניתוח העמודה מספק מידע חשוב על עוצמת התאונה, ויכול לשמש כמשתנה מסביר לחומרת הפציעה שנגרמה מהתאונה.



מהגרף עולה כי ברוב התאונות נגרם לרכב נזק בדרגת השבתה (disabling) עם כ-69,596 מקרים, כלומר נזק המשבית את פעולת הרכב.

קטגוריות נפוצות נוספות הן: Functional- רכב שניזוק אך עדין כשיר לנסיעה (48,815 מקרים), superficial- נזק חיצוני שטחי בלבד (48,343 מקרים).

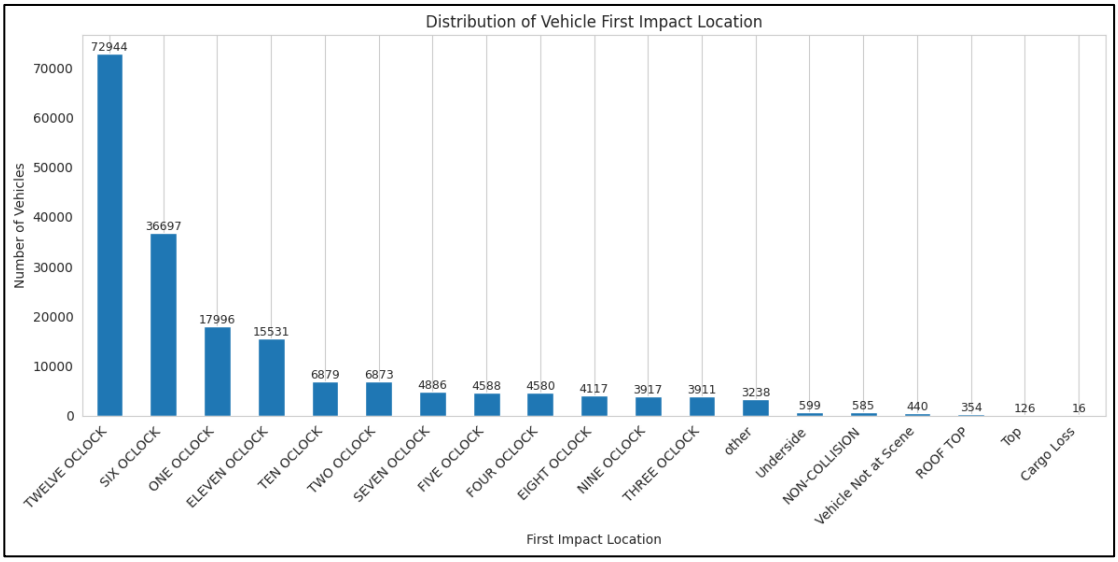
לעומתן נזק חמור יותר כמו destroyed (הריסה מוחלטת של הרכב) מהווה חלק קטן מהתאונות (7,610 מקרים), וכך גם מקרי no damage- תאונות ללא נזק נראה (6,435 מקרים).

ממצא זה מצביע על כך שלמרבית התאונות נגרם נזק פיזי לרכב, גם אם לא מדובר בנזק הרסני, דבר שעשוי להעיד על עוצמת ההתנגשות ברוב המקרים.

גרף התפלגות התאונות לפי מיקום הפגיעה הראשון ברכב (Vehicle First Impact Location)

עמודה זו מתארת את מיקום הפגיעה הראשונית ברכב באמצעות סימון שעות על פני שעון אנלוגי, כאשר:
O'Clock 12 מייצג פגיעה חזיתית, O'Clock 6 פגיעה אחורית, 3 ו-O'Clock 9 מייצגים פגיעות צידיות, ושעות אחרות מייצגות פגיעות בזוויות שונות מסביב לרכב.

שיטת קידוד זו מאפשרת הבנה ברורה של דפוסי הפגיעה ברכב בהתאם לזווית ההתנגשות.



מהגרף עולה כי הפגיעה הנפוצה ביותר היא פגיעה חזיתית (TWELVE O'CLOCK), עם כ-72,944 מקרים. לאחר מכן, נפוצה הפגיעה האחורית (SIX O'CLOCK), עם כ-36,697 מקרים. פגיעות צידיות, כמו בשעות ONE O'CLOCK ו-ELEVEN O'CLOCK, מופיעות בשכיחות נמוכה יותר, ואילו פגיעות בזוויות אחרות הן נדירות יחסית. נתונים אלו עשויים להעיד על כך שרוב ההתנגשויות מתרחשות במהלך נסיעה קדימה.

גרף התפלגות התאונות לפי סוג המבנה של הרכב (Vehicle Body Type)

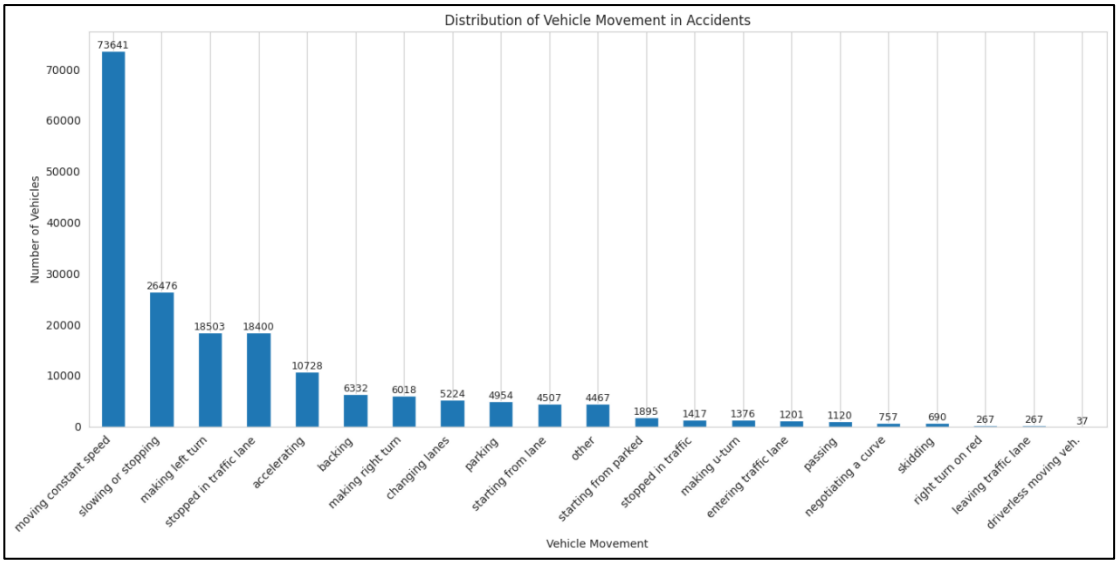
עמודה זו מתארת את סוג המבנה של הרכב שהיה מעורב בתאונה, כגון: רכב פרטי, רכב שטח (SUV), משאית קלה, אופנוע וכדומה. ניתוח מאפיין זה מסייע לזהות דפוסים שונים של מעורבות בתאונות בהתאם לסוג הרכב והשימוש בו.

Vehicle Body Type	
passenger car	129579
sport utility vehicle	18623
pickup truck	6787
other	5510
van	4958
transit bus	4078
school bus	3378
police vehicle/non emergency	2119
other light trucks (10,000lbs (4,536kg) or less)	1909
cargo van/light truck 2 axles (over 10,000lbs (4,536 kg))	1858
police vehicle/emergency	1497
medium/heavy trucks 3 axles (over 10,000lbs (4,536kg))	1469
motorcycle	889
pickup	817
station wagon	807
truck tractor	553
ambulance/emergency	466
fire vehicle/emergency	437
other bus	435
van - passenger (<9 seats)	379
fire vehicle/non emergency	296
other trucks	218
recreational vehicle	197
ambulance/non emergency	185
snowmobile	127
single-unit truck	124
all terrain vehicle (atv)	119
motorcycle - 2 wheeled	100
van - cargo	99
moped	87
autocycle	42
low speed vehicle	37
cross country bus	34
all-terrain vehicle/all-terrain cycle (atv/atc)	23
farm vehicle	21
moped or motorized bicycle	20

מנתונים עולה כי מרבית הרכבים המעורבים בתאונות הם מסוג רכב פרטי (Passenger car), המהווים את החלק העיקרי מכלל התאונות. אחריהם מופיעים רכבי שטח (SUV) וטנדרים (Pickup truck) בפער משמעותי.
יתר סוגי המרכב, כגון רכבי הסעות (Van, Transit bus), רכבי חירום, אופנועים ומשאיות כבדות, מופיעים בשכיחויות נמוכות יחסית.
ממצאים אלו תומכים בהנחה כי רכבים פרטיים, בהיותם הנפוצים ביותר על הכביש, הם גם המעורבים ביותר בתאונות. ניתן לראות כי קיימות מספר קטגוריות נדירות. לצורך בדיקה האם איחודן יתרום לביצועי המודל, בוצעה הרצת מודל עם שלושה ספים שונים לאיחוד קטגוריות נדירות תחת הקטגוריה 'other'. מהלך זה בוצע במטרה לפשט את העמודה ולצמצם רעש מנתונים נדירים. נמצא כי כאשר בחרנו בסף 20 ראינו שיפור קל בתוצאות המודל: ירידה במספר מקרי הפציעה שלא זוהו (FN מ־709 ל־708) וירידה במספר המקרים שנחזו בטעות כפציעה (FP מ־5879 ל־5870). למרות שהשיפור היה מזערי, הוחלט ליישם את האיחוד עם סף 20.

גרף התפלגות התאונות לפי תנועת הרכב (Vehicle Movement)

עמודה זו מתארת את מצב התנועה של הרכב ברגע התאונה, למשל: נסיעה במהירות קבועה, האטה, פנייה, מעבר נתיב או עצירה. מידע זה חיוני להבנת דפוסי התנועה שנפוצים בעת תאונה, ולזיהוי מצבים שעלולים להגביר את הסיכון למעורבות בתאונות.

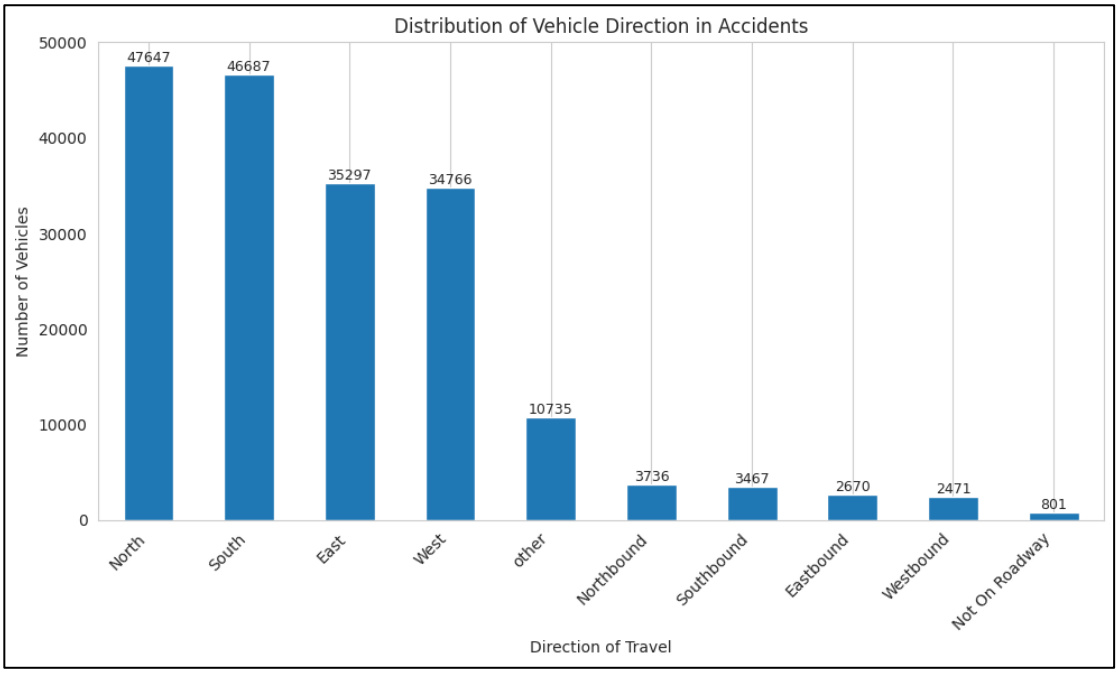


מהגרף עולה כי המצב הנפוץ ביותר בעת התאונה היה נסיעה במהירות קבועה (moving constant speed), עם כ־73,641 מקרים. אחריו בולטים המצבים של: האטה או עצירה (slowing or stopping) כ־26,476 מקרים, פנייה שמאלה (making left turn) כ־18,503 מקרים ועצירה בנתיב תנועה (stopped in traffic lane) כ־18,400 מקרים. מצבים נוספים כמו האצה, נסיעה לאחור, שינוי נתיב או פנייה ימינה מופיעים בתדירות נמוכה יותר.

ממצאים אלו עשויים להעיד כי גם במהלך נסיעה רגילה לכאורה, כמו נסיעה במהירות קבועה, קיימת רמת סיכון גבוהה לתאונות, ייתכן בשל חוסר ערנות, הפתעות בכביש או טעויות אנוש בלתי צפויות.

גרף התפלגות התאונות לפי כיוון נסיעת הרכב (Vehicle Going Dir)

עמודה זו מתארת את כיוון התנועה של הרכב בעת התאונה (כגון צפון, דרום, מזרח, מערב או כיוונים משולבים). מידע זה יכול לשמש להבנת כיוונים מועדים לפורענות בתשתית הכבישים.

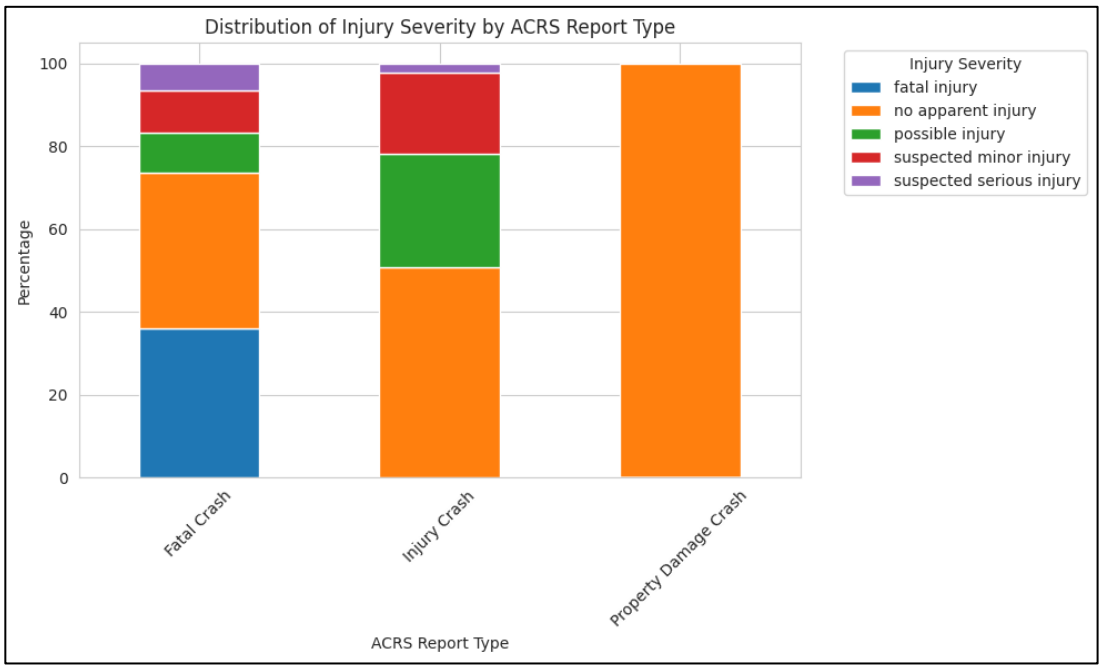


מהגרף עולה כי מרבית התאונות התרחשו כאשר הרכב נע בכיוונים ראשיים: צפון (North) עם כ־47,647 מקרים, ו־דרום (South) עם כ־46,687 מקרים. כיוונים נוספים כמו מזרח (East) ו־מערב (West) מציגים גם הם שכיחויות גבוהות יחסית (כ־35,297 ו־34,766 מקרים בהתאמה), אם כי מעט נמוכות יותר. לעומת זאת, כיווני נסיעה מפורטים יותר כמו Northbound, Southbound, וכן Eastbound ו־Westbound, מופיעים בשכיחויות נמוכות באופן משמעותי. מונחים אלה מתייחסים לרוב לנסיעה בכיוון מוגדר בתוך נתיב תנועה (bound), להבדיל מכיוונים כלליים כמו North או South. גם קטגוריית Not On Roadway- תנועה מחוץ לכביש מדווחת במספר מקרים בודדים בלבד (801 מקרים).
ממצאים אלו עשויים להעיד כי תאונות מתרחשות בעיקר בכיוונים המרכזיים של מערכת הדרכים, כאשר ייתכן שהשכיחות הגבוהה מושפעת גם מתנועת כלי רכב רבה יותר בכיוונים אלה.

ניתוח קשרים בין המאפיינים לעמודת המטרה (Injury Severity):

ניתוח קשר בין סוג הדיווח (ACRS Report Type) לבין חומרת הפציעה (Injury Severity):

Injury Severity ACRS Report Type	fatal injury	no apparent injury	possible injury	suspected minor injury	suspected serious injury
Fatal Crash	35.924370	37.815126	9.453782	10.294118	6.512605
Injury Crash	0.001473	50.749654	27.364576	19.633863	2.250434
Property Damage Crash	0.000000	100.000000	0.000000	0.000000	0.000000



לצורך בדיקת הקשר בין סוג הדיווח בתאונה (ACRS Report Type) לבין חומרת הפציעה (Injury Severity), חושבה טבלת שכיחויות יחסית (באחוזים) והוצג גרף עמודות מדורג. מהתצוגה עולה:

- תאונות שדווחו כ- "Property Damage Crash" כוללות אך ורק מקרים ללא פציעה.
- תאונות שדווחו כ- "Injury Crash" כוללות בעיקר מקרים ללא פציעה (כ-50%), לצד שיעור משמעותי של פציעות אפשריות (27%), פציעות קלות (כ-20%), ו־2% פציעות חמורות. מקרים קטלניים נדירים (פחות מ-0.5%).
- לעומתן, תאונות המדווחות כ- "Fatal Crash" כוללות שיעור גבוה במיוחד של פציעות קטלניות (כ־36%) וללא פציעות (כ-38%), לצד פציעות קלות וחמורות.

```
Chi-Square Statistic: 139170.3770
Degrees of Freedom: 8
P-Value: 0.0000
Cramér's V: 0.6079
```

לצורך בחינת מובהקות הקשר, בוצע מבחן חי-בריבוע שהניב: ערך סטטיסטי של 139,170 וערך p קטן מ-0.0001, המעידים על קיום קשר מובהק סטטיסטית בין סוג הדיווח לבין חומרת הפציעה. בנוסף, חושב מדד Cramér's V שהניב ערך של 0.608, המעיד על קשר חזק בין המשתנים.

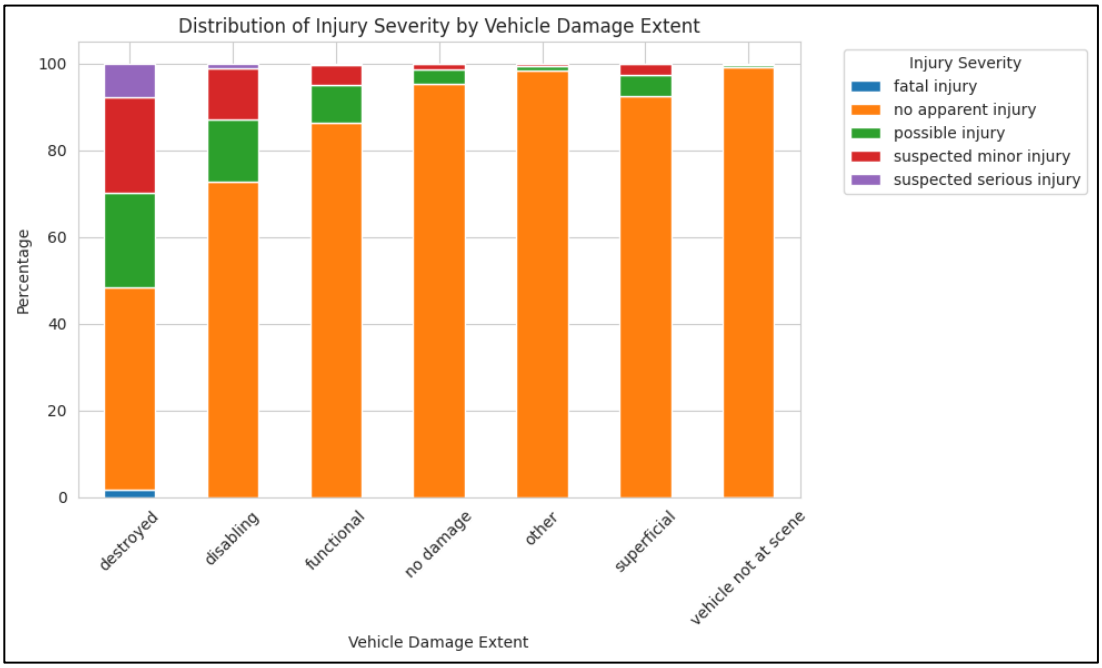
על סמך תוצאות אלו, קיימת תלות מובהקת וחזקה בין סוג הדיווח לבין חומרת הפציעה, בהתאם להיגיון המצופה. כלומר, סוג הדיווח מהווה אינדיקציה משמעותית לעוצמת התאונה ולחומרת הפגיעה.

ניתוח קשר בין היקף הנזק לרכב (Vehicle Damage Extent) לבין חומרת הפציעה (Injury Severity):

לצורך ניתוח הקשר בין מידת הנזק לרכב לבין חומרת הפציעה, הוצגו טבלת שכיחויות יחסית וגרף עמודות מדורג. מהנתונים עולה כי בכולן, גם בדרגות הנזק החמורות ביותר, הקטגוריה השכיחה ביותר היא "ללא פציעה נראית לעין", דבר המדגיש את חוסר האיזון הכללי בנתונים.

- כאשר הרכב הושמד לחלוטין (destroyed), עדיין כ־47% מהמקרים הסתיימו ללא פציעה, יחד עם זאת שיעורי הפציעות הקלות, החמורות והקטלניות גבוהים יחסית למידות הנזק האחרות. למשל, 7.6% פציעות חמורות ו־1.6% פציעות קטלניות.
- ככל שמידת הנזק לרכב פוחתת ,כך גם שיעור הפציעות החמורות פוחת לדוגמה, ברכבים עם נזק פונקציונלי (functional) נצפו רק 0.23% פציעות חמורות ו־0.008% פציעות קטלניות.
- בקטגוריות של נזק שטחי או ללא נזק כלל, מעל 90% מהמקרים הסתיימו ללא פציעה, עם שיעורים זניחים של פציעות חמורות או קטלניות.

Injury Severity Vehicle Damage Extent	fatal injury	no apparent injury	possible injury	suspected minor injury	suspected serious injury
destroyed	1.655716	46.819974	21.865966	22.049934	7.608410
disabling	0.058911	72.844704	14.351400	11.589747	1.155239
functional	0.008194	86.266516	8.749360	4.744443	0.231486
no damage	0.000000	95.260295	3.403263	1.274281	0.062160
other	0.000000	98.309179	1.108269	0.497300	0.085251
superficial	0.002069	92.433237	4.968661	2.486399	0.109633
vehicle not at scene	0.000000	99.090909	0.681818	0.227273	0.000000




```
Chi-Square Statistic: 21626.1220
Degrees of Freedom: 24
P-Value: 0.0000
Cramér's V: 0.1695
```

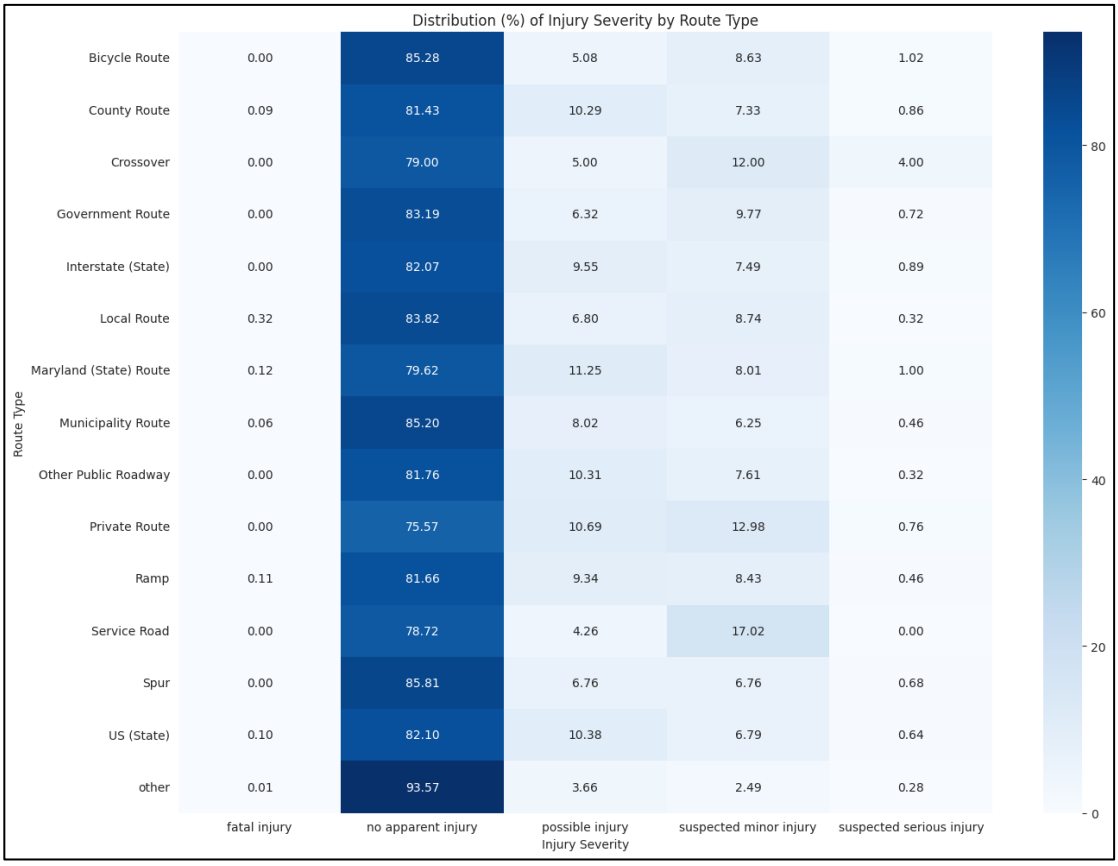
לצורך בחינת מובהקות הקשר, בוצע מבחן חי-בריבוע שהניב ערך סטטיסטי של 21,626 וערך p קטן מ־0.0001, המעידים על קיום קשר מובהק סטטיסטית בין היקף הנזק לרכב לבין חומרת הפציעה. בנוסף, חושב מדד Cramér's V שהניב ערך של 0.1695, המעיד על קשר חלש בין המשתנים.

מכאן נובע שלמרות הקשר המובהק, ייתכן שהוא מושפע באופן חזק מחוסר האיזון בנתוני הפציעות.

ניתוח קשר בין סוג הדרך (Route Type) לבין חומרת הפציעה (Injury Severity):

לצורך בחינת הקשר בין סוג הדרך לבין חומרת הפציעה, נותחה התפלגות אחוזית של חומרת הפציעה לפי קטגוריות הדרך באמצעות heatmap. ממצאים עיקריים:

- ברוב סוגי הדרכים, שיעור התאונות שמסתיימות ללא פציעה נראית לעין (No apparent injury) הוא הגבוה ביותר. למשל, ב- Bicycle Route שיעור זה עומד על 85.3%, וב- Spur על 85.8%.
- עם זאת, בדרכים כמו Service Road ו- Private Route שיעור הפציעות הקלות (Suspected Minor Injury) גבוה מהממוצע: 17.0% ו־13.0% בהתאמה.
- פציעות קטלניות (Fatal Injury) מופיעות בשיעור נמוך מאוד בכל הקטגוריות, כאשר Local Route מציג את השיעור הגבוה ביותר (0.32%).




```
Chi-Square Statistic: 296.57734385665674
P-Value: 1.3038388817973847e-53
Cramér's V: 0.019844508875074027
```

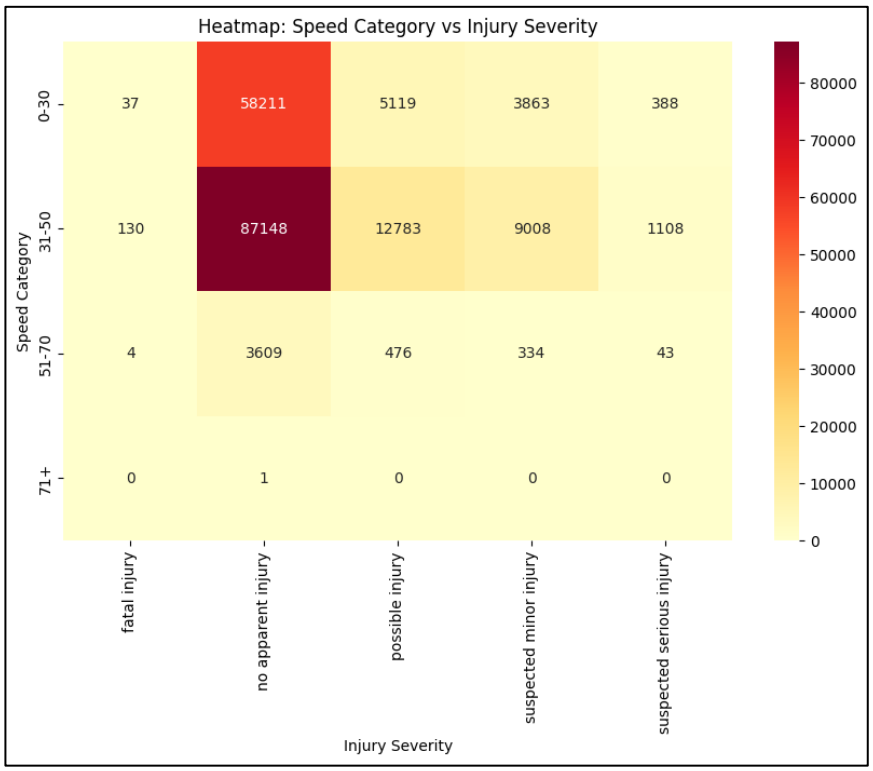
לצורך בדיקת מובהקות הקשר, בוצע מבחן חי-בריבוע שהניב ערך סטטיסטי של 296.57 וערך p הקטן מ־0.0001, המעידים על קיום קשר מובהק סטטיסטית בין סוג הדרך לבין חומרת הפציעה. עם זאת, מדד Cramér's V הנמוך (0.0198) מעיד על כך שעוצמת הקשר בין המשתנים חלשה מאוד. ייתכן שסוג הדרך אינו משתנה מסביר משמעותי בפני עצמו, אלא עשוי לתרום למודל רק כאשר משולב עם משתנים נוספים.

ניתוח קשר בין הגבלת המהירות (Speed Limit) לבין חומרת הפציעה (Injury Severity):

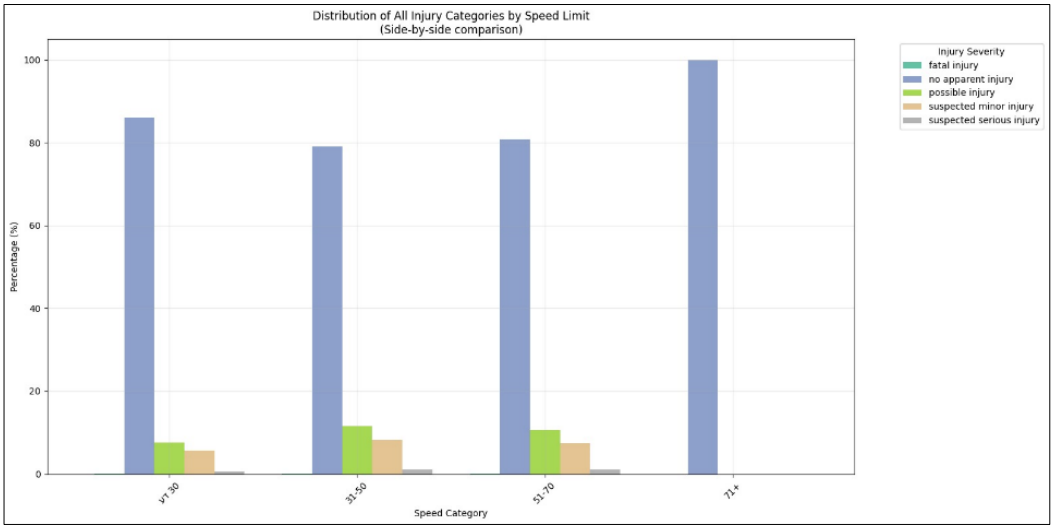
Heatmap המציג את הקשר בין מהירות מותרת לבין חומרת הפציעה מראה כי מרבית התאונות התרחשו בטווחי המהירות 0-30 ו-31-50 קמ"ש. עם זאת, בקטגוריית 31-50 קמ"ש נרשם גם המספר האבסולוטי הגבוה ביותר של פציעות חמורות וקטלניות (1,238).

בקטגוריית 51-70 קמ"ש, אף שמספר התאונות הכולל נמוך בהרבה, שיעור הפציעות החמורות והקטלניות מתוך כלל התאונות (כ-1.05%) דומה לזה שבקטגוריית 31-50 וגבוה כמעט פי שניים לעומת קטגוריית 0-30.

ממצא זה עשוי להצביע על כך שמרבית התאונות מתרחשות באזורים עירוניים שבהם המהירות המותרת נעה בין 0 ל־50 קמ"ש, כגון סמוך למוסדות חינוך, אזורי מגורים ואזורים עם הולכי רגל רבים. כמו כן, ייתכן שתחומי מהירות בינוניים מציגים הרכב פציעות חמור יותר באופן יחסי ,גם אם לא מתרחשות בהם בהכרח יותר תאונות. עם זאת, המגמה לא לינארית: בקטגוריות מהירות גבוהות יותר (71+) כמעט ואין תצפיות, ולכן אין להסיק מסקנות מבוססות.



בנוסף לכך, גרף ההתפלגות היחסית מציג את האחוזים של כל רמות הפציעה בתוך כל קטגוריית מהירות. מהגרף עולה כי בקטגוריית 0-30 קמ"ש כ־85% מהתאונות אינן כוללות פציעה נראית לעין, בעוד שבקטגוריות 31-50 ו־51-70 אחוז זה יורד מעט. במקביל, אחוז הפציעות הקלות (possible/minor) עולה, ונצפית גם עלייה קטנה בשיעור הפציעות החמורות בהשוואה לקטגוריית 0-30. בקטגוריית 71+ קמ"ש מופיעה תאונה אחת בלבד, ולכן אין אפשרות להסיק מסקנות לגביה. ממצאים אלו עשויים להצביע על כך שככל שהמהירות עולה, הרכב הפציעות משתנה גם אם מספר התאונות יורד, אך הקשר עדיין אינו חד משמעי.



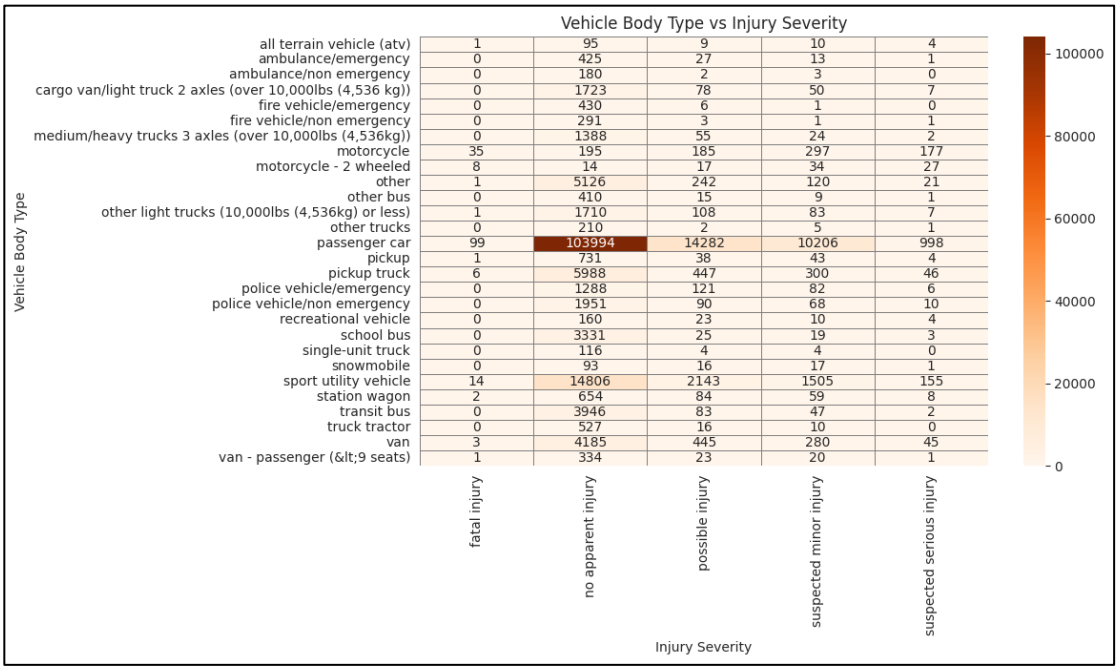
לצורך בדיקת מובהקות הקשר בין המשתנים, בוצע מבחן חי-בריבוע (Chi-Square Test) על טבלת שכיחויות בין קטגוריות המהירות ורמות הפציעה. תוצאות המבחן:

```
Chi-Square Statistic: 1390.550
P-Value: 1.51551e-290
Cramér's V: 0.050430
```

למרות שערך ה־p מעיד על קשר מובהק סטטיסטית ($p < 0.0001$), ערך Cramér's V הנמוך מעיד על עוצמת קשר חלשה. ממצאים אלה מחזקים את ההבנה כי למהירות מותרת עשויה להיות השפעה מסוימת על חומרת הפציעה, אך ייתכן שהשפעה זו מתווכת או מושפעת ממשתנים נוספים.

<u>ניתוח קשר בין סוג מבנה הרכב (Vehicle Body Type) לחומרת הפציעה (Injury Severity):</u>

Heatmap המציג את הקשר בין סוג הרכב המעורב בתאונה לבין חומרת הפציעה מראה כי
רכב פרטי (Passenger car) הוא סוג הרכב השכיח ביותר במאגר (103,994 תאונות ללא פציעה), והוא גם מעורב במספר האבסולוטי הגבוה ביותר של פציעות קלות, חמורות וקטלניות.
עם זאת, ייתכן שהדבר נובע מהשכיחות הגבוהה של רכב זה בכבישים, ולא מהיותו מסוכן יותר.
רכב שטח (Sport Utility Vehicle) מציג שכיחות בינונית אך שיעור ניכר של פציעות חמורות יחסית (155 מתוך 18,623 מקרים).
אופנועים בולטים בכך שלמרות שכיחותם הנמוכה יחסית בנתונים (כ-889 מקרים בהם אופנועים מעורבים בתאונות), שיעור הפציעות הקטלניות והחמורות גבוה משמעותית ביחס למספר התאונות הכולל, לדוגמה, 35 תאונות קטלניות (המהוות כ-4% מכלל התאונות עם אופנועים). נתון זה מחזק את הסברה כי רוכבי דו־גלגלי חשופים יותר לפגיעות חמורות.
רכבי הסעות ואוטובוסים (כגון: School Bus, Transit Bus) מציגים לרוב שיעור נמוך של פציעות חמורות, ייתכן בשל מבנה מגונן או אופי נהיגה זהיר.




```
Chi-Square Statistic: 11727.181
P-Value: 0.000000e+00
Cramér's V: 0.124907
```

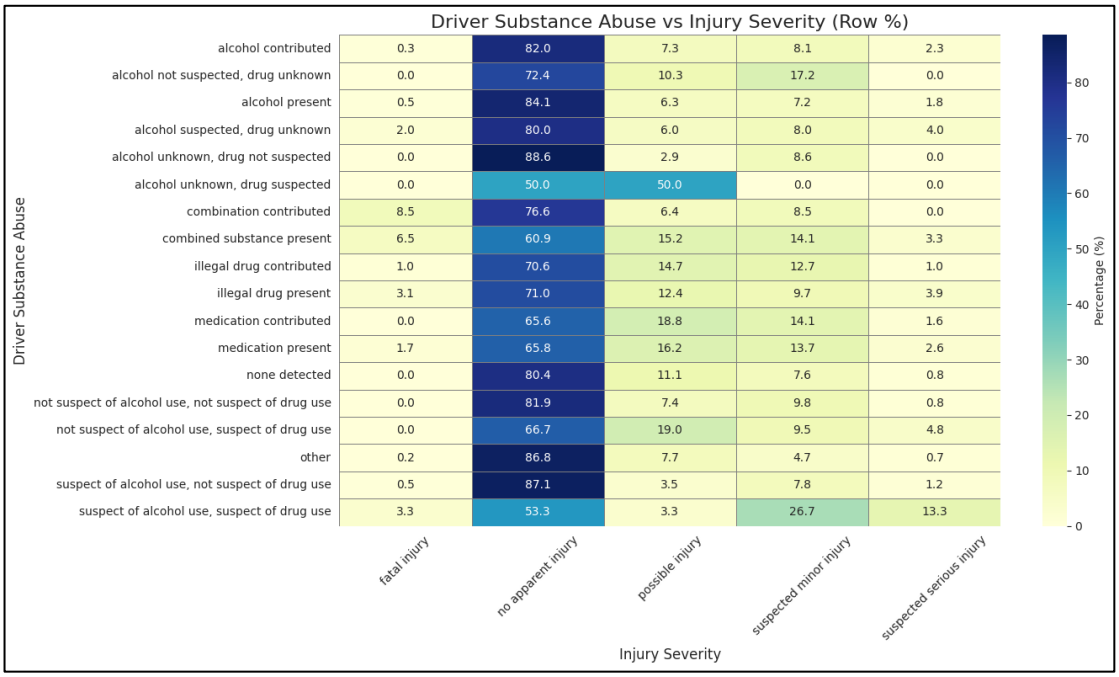
לצורך בדיקת מובהקות הקשר, בוצע מבחן חי-בריבוע שהניב ערך סטטיסטי של 11,727 וערך p הקטן מ־0.0001, המעידים על קיום קשר מובהק סטטיסטית בין סוג הרכב לבין חומרת הפציעה. בנוסף, מדד Cramér's V (0.12) מעיד על כך שעוצמת הקשר בין המשתנים חלשה עד בינונית.

מסקנה:
קיים קשר בין סוג הרכב המעורב בתאונה לבין חומרת הפציעה: רכבים דו-גלגליים נוטים להיות מעורבים בשיעור גבוה יותר של פציעות חמורות וקטלניות, בעוד שרכבים גדולים (כמו אוטובוסים) מציגים נטייה נמוכה יותר לפציעות חמורות . עם זאת, יש להתחשב גם בהבדלים בשכיחות סוגי הרכבים בכביש, אשר עשויים להשפיע על הממצאים.

ניתוח קשר בין שימוש חומרים מסוכנים (Driver Substance Abuse) לחומרת הפציעה (Injury Severity):

על מנת לבחון את הקשר בין משתנה "Driver Substance Abuse" לבין משתנה המטרה "Injury Severity", בוצע ניתוח באמצעות heatmap המציג את התפלגות אחוזי הפציעות בכל קטגוריה של שימוש בחומרים, מתוך סך המקרים באותה הקטגוריה (כל שורה מסתכמת ל־100%).
הצגת הנתונים באחוזים מאפשרת השוואה בין קטגוריות שונות גם כאשר שכיחותן בדאטה שונה משמעותית.

הניתוח מראה כי במרבית המקרים שבהם לא זוהה שימוש בחומרים (none detected), שיעור התאונות ללא פציעה נראית לעין גבוה במיוחד (80.4%), בעוד ששיעור הפציעות הקטלניות והחמורות נמוך.
לעומת זאת, כאשר קיימת נוכחות של חומרים מסוכנים – במיוחד בשילוב חומרים (combined substance present, combination contributed) או שימוש בסמים לא חוקיים – נצפים שיעורים גבוהים יותר של פציעות חמורות וקטלניות.
לדוגמה, בקבוצת combination contributed, כ־8.5% מהתאונות הן קטלניות, שיעור הגבוה פי כמה בהשוואה לקטגוריות אחרות.
גם בקבוצות כמו illegal drug present או suspect of alcohol use, suspect of drug use, שיעור הפציעות הקטלניות (3.1%–3.3%) והחמורות (13.3% ו־3.9% בהתאמה) משמעותי יותר, מה שמרמז על קשר בין שימוש בחומרים לבין חומרת הפציעה.




```
Statistical Test Results:
Chi-Square Statistic: 2891.927
P-Value: 0.000000e+00
Cramér's V: 0.061968
```

לצורך בדיקת מובהקות הקשר, בוצע מבחן חי־בריבוע שהניב ערך סטטיסטי של 2,891 וערך p הקטן מ־0.0001, המעידים על קשר מובהק סטטיסטית בין המשתנים. עם זאת, מדד Cramér's V (0.06) מצביע על כך שעוצמת הקשר חלשה.

מסקנה:
קיים קשר בין שימוש או חשד לשימוש בחומרים מסוכנים מצד הנהג לבין חומרת הפציעה: כאשר יש שימוש או חשד לשימוש בחומרים, במיוחד חומרים משולבים או סמים לא חוקיים שיעור הפציעות החמורות והקטלניות גבוה יותר. עם זאת, עוצמת הקשר הכללית חלשה, ולכן מומלץ לשלב משתנה זה כחלק ממערך משתנים רחב יותר במודל, אך לא להסתמך עליו לבדו.