האקדמית עמק יזרעאל
ע"ש מקס שטרן
היחידה ללימודי חוץ והמשך

DATA UNDERSTANDING REPORT

פרויקט גמר

מערכת לחיזוי רמת חומרת הפציעה
בתאונות דרכים וזיהוי אזורים מסוכנים

מגישות

אלמוג יבדייב 305325417
מאי יוסף 318608072

מוגש ליואב זיו

איסוף נתונים

1.1 מקורות נתונים:

הנתונים בפרויקט שלנו נלקחו ממקור ציבורי וזמין: אתר Data.gov, שהוא מאגר הנתונים הרשמי של ממשלת ארצות הברית שמספק גישה למגוון רחב של נתונים ציבוריים. המידע נלקח מתוך מאגר הנתונים "Crash Reporting - Drivers Data", שמספק מידע מפורט על תאונות דרכים שדווחו על ידי הרשויות בארצות הברית. נתונים אלו כוללים מידע על מאפייני התאונה, תנאי הדרך, כלי הרכב המעורבים, מידת הפציעה של המעורבים ועוד גורמים רלוונטיים.

סוגי נתונים:

נתונים קיימים: מדובר במאגר נתונים קיים שנוצר לצרכים ממשלתיים, הכולל כ-189,333 רשומות עם 39 עמודות שמספקות מידע מגוון על תאונות דרכים.

נכון לעכשיו, איננו משתמשות בנתונים שנרכשו או בנתונים נוספים, שכן הנתונים הקיימים מספיקים כדי לענות על מטרת המחקר שלנו: חיזוי רמת הפציעה בעקבות תאונות דרכים. במידה ויהיה צורך בהמשך הפרויקט, נבחן את האפשרות להעשיר את המידע עם מקורות נוספים, כמו נתונים דמוגרפיים של האזורים בהם התרחשו התאונות.

1.2 בדיקת נתונים ראשונית:

במסגרת בדיקה ראשונית של מאגר הנתונים, נבחנו התכונות השונות על מנת להעריך את מידת התאמתן למטרות הפרויקט ולזהות אתגרים אפשריים בעיבוד הנתונים.

<u>תכונות מבטיחות לחיזוי רמת הפציעה:</u>

- **Injury Severity (רמת הפציעה):** עמודת המטרה שמודל החיזוי יתמקד בה.
- **Weather (מזג אוויר):** תנאי מזג האוויר עשויים להשפיע על חומרת התאונה.
- **Surface Condition (מצב הכביש):** מצב פני הדרך הוא גורם משפיע על שליטה ברכב.
- **Light (תנאי תאורה):** האם התאונה התרחשה ביום או בלילה, גורם חשוב.
- **Speed Limit (מהירות מותרת):** למהירות המותרת יש קשר ישיר לחומרת הפציעות.
- **Collision Type (סוג התנגשות):** סוג ההתנגשות (חזיתית, צדית) משפיע על רמת הפציעה.
- **Vehicle Damage Extent (רמת נזק לרכב):** אינדיקציה ישירה לחומרת הפציעות.
- **Driver Substance Abuse (שימוש בסמים/אלכוהול):** גורם קריטי במיוחד בתאונות חמורות.
- **Latitude ו-Longitude (קו רוחב וקו אורך):** מאפשרים ניתוח גיאוגרפי לאזורים מסוכנים.
- **Crash Date/Time (תאריך ושעת התאונה):** עשוי לחשוף מגמות כמו הבדל בין יום ולילה.
- **Vehicle Body Type(סוג הרכב):**משפי על רמת ההגנה ועל חומרת הפציעה בתאונה.

<u>תכונות שאינן רלוונטיות או מכילות ערכים חסרים רבים:</u>

- **Off-Road Description, Municipality, Related Non-Motorist, Non-Motorist Substance Abuse, Circumstance**: מכילות מעל 80% ערכים חסרים ואינן תורמות למודל.
- **Report Number, Local Case Number, Vehicle ID, Person ID**: מזהים טכניים שאינם רלוונטיים לניתוח רמת הפציעה.
- **Drivers License State**: לא נראה שיש קשר ישיר בינו לבין חומרת הפציעות.
- **Location**: נתון כפול לנתוני קו רוחב ואורך.
- **Driverless Vehicle** - אין מידע משמעותי בעמודה זו, מרבית התאונות לא כוללות כלי רכב ללא נהג. לכן לא צפוי להשפיע על המודל.

האם יש מספיק נתונים?

המאגר מכיל 189,333 רשומות עם 39 עמודות, דבר המהווה בסיס טוב לניתוח סטטיסטי ולבניית מודל חיזוי. עם זאת, יש להתמודד עם הערכים החסרים בעמודות כמו, Weather ו-Surface Condition. ובנוסף , בעמודת המטרה Injury Severity יש חוסר איזון בין הקטגוריות שיצריך טיפול.

האם יש יותר מדי עמודות למודל?

הכמות הכוללת של העמודות אינה גדולה מדי (39), אך ישנם נתונים שאינם רלוונטיים או עם כמות גדולה של ערכים חסרים, ולכן ניתן לצמצם את העמודות כדי לשפר את ביצועי המודל.

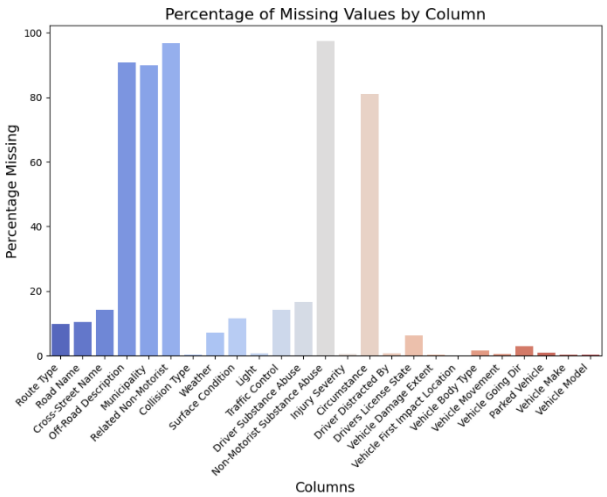
האם נעשה שימוש במקורות נתונים נוספים?

לא. הנתונים נאספו ממקור אחד בלבד (Data.gov). בשלב זה אין צורך בשילוב נתונים נוספים, אך ניתן לשקול בעתיד להוסיף נתונים דמוגרפיים (כמו גיל או מגדר) אם נרצה להרחיב את הניתוח.

ערכים חסרים:

במהלך הניתוח הראשוני, זוהו תכונות עם אחוזים שונים של ערכים חסרים. הגרף והטבלה המצורפים מציגים את אחוז הערכים החסרים בעמודות עם החוסרים.

	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



<u>ממצאים מרכזיים:</u>

1. **עמודות עם אחוזים גבוהים של ערכים חסרים:**
   - **Off-Road Description** (90.77%) ו-**Municipality** (89.90%) הן תכונות עם אחוזים גבוהים במיוחד של ערכים חסרים, מה שמעיד על חוסר מידע משמעותי בעמודות אלו ועל פוטנציאל נמוך לשימוש יעיל בהן במסגרת הניתוח.
   - **Related Non-Motorist** (96.79%) ו-**Non-Motorist Substance Abuse** (97.39%) מכילות כמעט לחלוטין נתונים חסרים, מה שמעלה ספקות לגבי הרלוונטיות שלהן לניתוח או השפעתן על תוצאות המודל.
2. **עמודות עם אחוזים נמוכים של ערכים חסרים:**
   - עמודות קריטיות כמו **Injury Severity** (0.56%) ו-**Vehicle Damage Extent** (0.17%) מכילות ערכים חסרים באחוזים נמוכים, ולכן השפעת החסרים על הניתוח הכולל היא מינימלית.
3. **עמודות ביניים:**
   - עמודות כמו **Road Name** (10.36%) ו-**Traffic Control** (14.24%) מכילות אחוזים בינוניים של ערכים חסרים, יש לשקול את מידת התרומה שלהן למודל החיזוי ואת העלות מול התועלת בטיפול בערכים חסרים בעמודות אלו.

<u>טיפול בערכים חסרים:</u>

1. **שמירה על עמודות קריטיות:** עמודות כמו **Injury Severity** שבהן אחוז הערכים החסרים נמוך (0.56%) יושלמו באמצעות שיטות כגון מילוי בערכים נפוצים, ממוצע או חציון, כדי למנוע איבוד מידע חשוב לניתוח.
2. **הסרת עמודות עם ערכים חסרים רבים:** עמודות עם אחוזים גבוהים מאוד של ערכים חסרים, עשויות להימחק אם אינן רלוונטיות למטרת הניתוח.
3. עמודות עם חסרים בינוניים ייבחנו להשלמה או להסרה בהתאם לחשיבותן בניתוח.

תיאור הנתונים:

2.1. כמות הנתונים

במאגר הנתונים שברשותנו:

- מספר הרשומות (שורות): 189,333.
- מספר התכונות (עמודות): 39.

2.2 סוגי ערכים:

הנתונים במאגר מכילים שלושה סוגים עיקריים של ערכים: מספריים, קטגוריאליים ובוליאניים. להלן סיווג מלא של כל המשתנים בהתאם לסוגי הערכים שלהם:

משתנים מספריים (Numeric)

משתנים אלו מכילים ערכים מספריים רציפים או שלמים, אשר משמשים למדידה או לכימות:

- Speed Limit – מגבלת המהירות באזור התאונה (int64).
- Vehicle Year – שנת הייצור של הרכב (int64).
- Latitude – קו רוחב של מיקום התאונה (float64).
- Longitude – קו אורך של מיקום התאונה (float64).

משתנים קטגוריאליים (Categorical)

עמודות עם מספר מוגבל של ערכים אפשריים (קטגוריות מוגדרות מראש):

- Agency Name: שם הסוכנות.
- ACRS Report Type: סוג דוח ACRS(מערכת דיווח סוג תאונות אוטומטית).
- Injury Severity – רמת הפציעה בתאונה.
- Weather – תנאי מזג האוויר בזמן התאונה
- Surface Condition – מצב הכביש בזמן התאונה .
- Light – תנאי התאורה .
- Driver Substance Abuse – שימוש בסמים/אלכוהול ע"י הנהג.
- Vehicle Damage Extent – היקף הנזק שנגרם לרכב.
- Route Type – סוג נתיב/מסלול
- Collision Type – סוג ההתנגשות
- Vehicle Movement – תנועת הרכב בעת התאונה.
- Vehicle Body Type – סוג הרכב.
- Vehicle First Impact Location – מקום הפגיעה הראשון ברכב
- Driver Distracted By – מה גרם להסחת דעת של הנהג
- Drivers License State – מדינת רישיון הנהיגה של הנהג.
- Municipality – שם הרשות המקומית/ העירייה בה התרחשה התאונה.
- Related Non-Motorist-הולך רגל/משתמש לא ממונע.
- Non-Motorist Substance Abuse-שימוש בסמים/אלכוהול ע"י לא נהגים
- Circumstance- נסיבות התאונה.
- Vehicle Going Dir- כיוון נסיעת הרכב.
- Driver At Fault – האם הנהג היה אשם בתאונה (Yes/No/Unknown).
- Vehicle Make: יצרן הרכב.

משתנים רבים סווגו כקטגוריאליים מכיוון שלאחר תהליך ניקוי, אחידות, ותהליך דמיז (Dummy Variables) סביר להניח שהם יתאימו לניתוח כמשתנים קטגוריאליים. חלק מהמשתנים כוללים:

- **שגיאות כתיב או סיווג לא אחיד :**עמודות כמו Vehicle Make כוללות ערכים כמו "TOYOTA", "TOYT", "TOYTA" לאחר תיקון שגיאות אלה, ניתן לסווג אותן כקטגוריאליות.
- **ערכים מרובי קטגוריות :**עמודות שמכילות ערכים מופרדים בפסיקים או בתווים אחרים. תהליך פיצול לערכים נפרדים יהפוך אותן לקטגוריאליות. למשל Related Non-Motorist, עם ערך כמו "Bicyclist, Pedestrian"

תהליך זה מאפשר לסדר את הנתונים בצורה שתתאים לניתוח כקטגוריות מוגדרות היטב.

משתנים בוליאניים (Boolean)

עמודות עם שני ערכים אפשריים:

- Driverless Vehicle – האם הרכב היה ללא נהג (Unknown/No).
- Parked Vehicle – האם הרכב היה חונה בזמן התאונה (Yes/No).

משתנה תאריכי

- Crash Date/Time- תאריך/שעת התאונה

משתנים טקסטואליים או מזהים (Text/Identifiers)

עמודות עם תוכן חופשי, שאינו מוגבל למספר מוגדר של ערכים או לזיהוי רשומות:

- Report Number – מזהה הדוח.
- Local Case Number – מזהה מספר תיק.
- Road Name – שם הכביש בו התרחשה התאונה.
- Person ID – מזהה ייחודי של הנהג או המעורב בתאונה.
- Vehicle ID – מזהה ייחודי של הרכב המעורב בתאונה.
- Off-Road Description – תיאור נוסף של מיקום התאונה.
- Location- מיקום התאונה לפי קו אורך ורוחב
- Traffic Control – סוג השליטה בתנועה .
- Cross-Street Name – שם הרחוב החוצה.
- Vehicle Model: דגם הרכב.

2.3 סכמות קידוד:

Drivers License State (מדינת רישיון הנהיגה):

עמודה זו מציינת את המדינה שהנפיקה את רישיון הנהיגה לפי קוד סטנדרטי בן שתי אותיות.

- מדינות בארה"ב לדוגמא:
  - MD = Maryland
  - DC = District of Columbia
  - VA = Virginia
  - PA = Pennsylvania
- מדינות זרות לדוגמא:
  - XX = רישיון זר או לא מזוהה
  - MX-ROO = Quintana Roo, Mexico
  - MX-MEX = Mexico City, Mexico

קידוד זה מאפשר ניתוח דמוגרפי והשוואת נהגים ממדינות שונות.

Vehicle First Impact Location (מיקום הפגיעה הראשון ברכב):

עמודה זו מציינת את נקודת הפגיעה הראשונה ברכב במהלך התאונה, תוך שימוש בכיוונים בשעון אנלוגי וערכים נוספים.

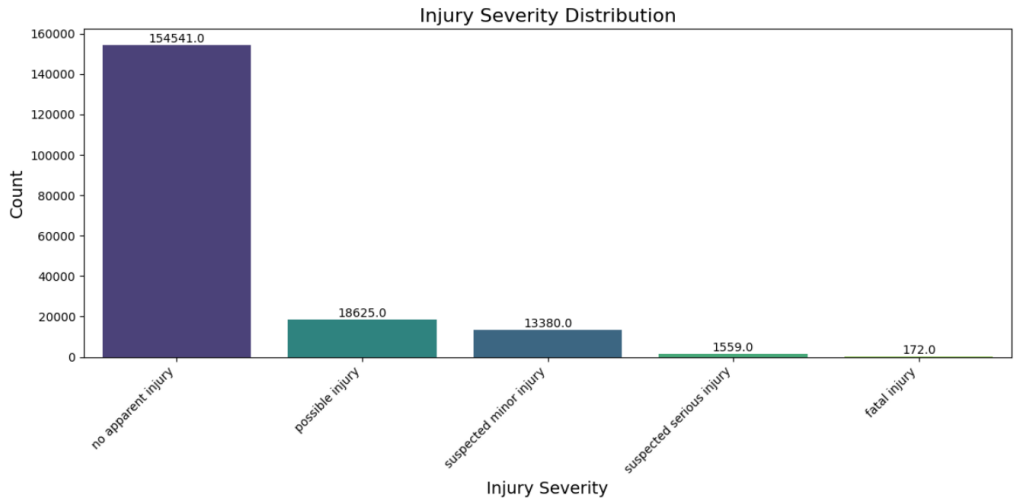
תיאור הערכים בעמודה לדוגמה:

- **TWELVE OCLOCK (12:00):** פגיעה קדמית (68,026 מופעים).
- **SIX OCLOCK (6:00):** פגיעה אחורית (34,275 מופעים).
- **ELEVEN OCLOCK / (1:00) ONE OCLOCK (11:00):** פגיעות זוויתיות קדמיות.
- **UNKNOWN:** מיקום לא ידוע.
- **NON-COLLISION:** תאונה ללא פגיעה ישירה.
- ערכים נוספים כוללים פגיעות בגג (ROOF TOP), תחתית הרכב (UNDERSIDE) ואיבוד מטען (CARGO LOSS).

הקידוד מאפשר ניתוח מגמות פגיעות, השוואת סוגי תאונות ותכנון שיפורי בטיחות.

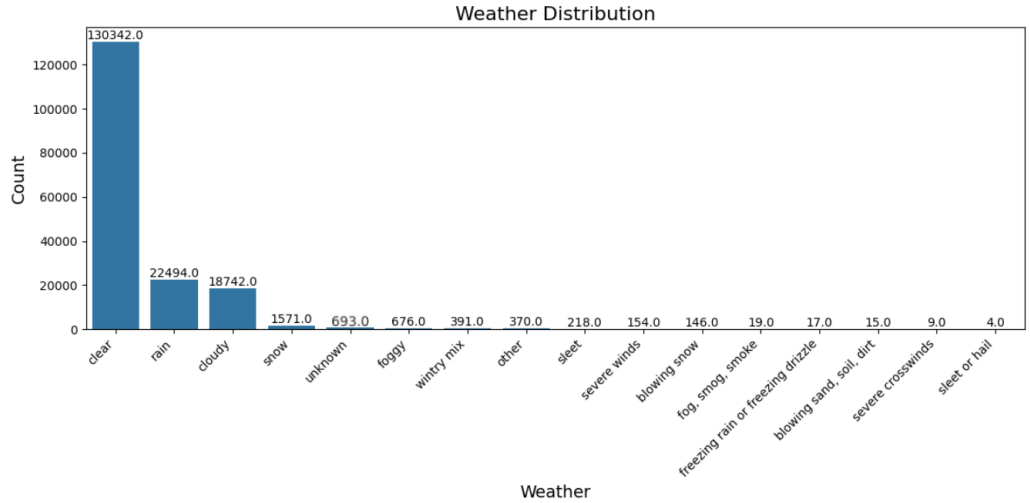
חקר נתונים:

<u>גרף 1:</u> עמודת המטרה "injury Severity ": פיזור חומרת פציעות בתאונות דרכים



הגרף מציג את פיזור חומרת הפציעות בתאונות דרכים (עמודת המטרה-"injury Severity"), ומראה חוסר איזון משמעותי בנתונים. הרוב המוחלט של התאונות (154,541) מסווגות כ "No Apparent Injury" (ללא פציעה נראית), בעוד קטגוריות חמורות יותר כמו "Fatal Injury" ו "Suspected Serious Injury" מכילות מעט מאוד נתונים, עם 172 ו-1,559 מופעים בלבד, בהתאמה. קטגוריות מתונות יותר, כמו "Possible Injury" (18,625) ו "Suspected Minor Injury" (13,380), נמצאות בתווך. חוסר האיזון הקיצוני מצביע על צורך בטכניקות כמו Smote או Weighted Loss כדי לשפר את יכולת המודל לחזות פציעות חמורות, שהן נדירות יחסית בנתונים.

<u>גרף 2:</u> פיזור כמות תאונות לפי תנאי מזג האוויר ("Weather"):



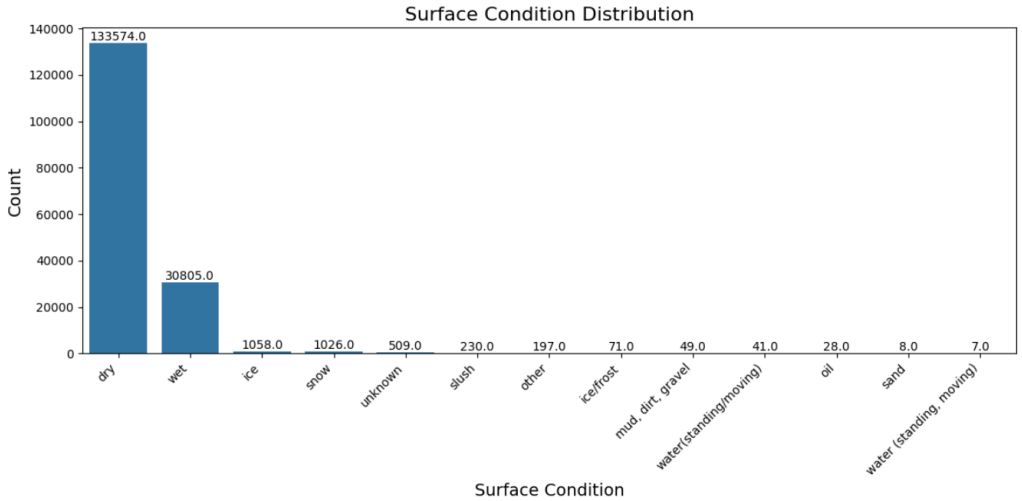
הגרף מציג את פיזור כמות התאונות לפי תנאי מזג האוויר ומראה כי רוב התאונות התרחשו במזג אוויר בהיר (Clear) עם כ-130,342 מופעים, בעוד מזג אוויר גשום (Rain) ומעונן (Cloudy) הם התנאים הבולטים הבאים עם 22,494 ו-18,742 מופעים, בהתאמה. תנאי מזג אוויר קשים כמו שלג (Snow) או ערפל (Foggy) נדירים יחסית, עם פחות מ-2,000 מקרים כל אחד. כמו כן, נתוני מזג אוויר לא ידועים (Unknown) מופיעים ב-693 מקרים ועשויים להצביע על חוסר דיוק בתיעוד. ממצאים אלו מראים שתנאי מזג אוויר קשים פחות שכיחים, אך ייתכן שיש להם השפעה משמעותית על חומרת התאונות.

המשך לגרף הקודם: ניתוח אחוז מקרי הפציעות לפי תנאי מזג האוויר

Weather	Percentage Not "No Apparent Injury"	Row Counts
blowing sand, soil, dirt	33.333333	15
blowing snow	19.178082	146
clear	18.321800	130342
cloudy	19.544339	18742
fog, smog, smoke	15.789474	19
foggy	19.082840	676
freezing rain or freezing drizzle	11.764706	17
other	18.918919	370
rain	19.049524	22494
severe crosswinds	33.333333	9
severe winds	17.532468	154
sleet	16.513761	218
sleet or hail	25.000000	4
snow	16.231700	1571
unknown	3.896104	693
wintry mix	18.414322	391

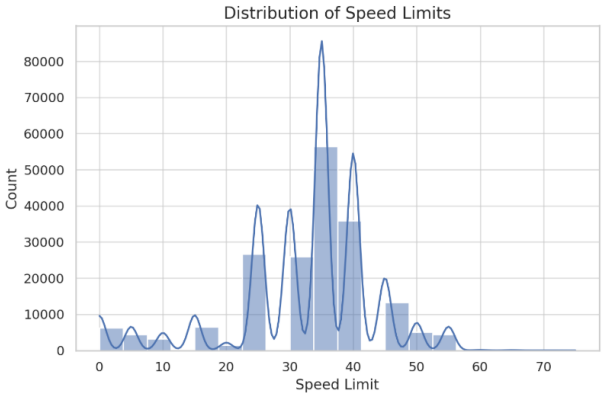
הטבלה מראה את האחוז היחסי של מקרי תאונות שאינן "No Apparent Injury" בכל תנאי מזג אוויר. תנאים כמו: **"Blowing sand, soil, dirt"**ו **"Severe crosswinds"** מציגים אחוזים גבוהים של פציעות יחסית למקרים אחרים, אך מתרחשים במספר קטן של תאונות. לעומתם, תנאים נפוצים כמו **Clear**ו **Rain** מציגים אחוזים נמוכים יחסית אך משפיעים משמעותית בשל כמות המקרים הגבוהה שלהם. גישה זו מאפשרת לבחון את ההשפעה היחסית של כל תנאי מזג אוויר בצורה מבודדת, תוך התמקדות בזיהוי תבניות ייחודיות.

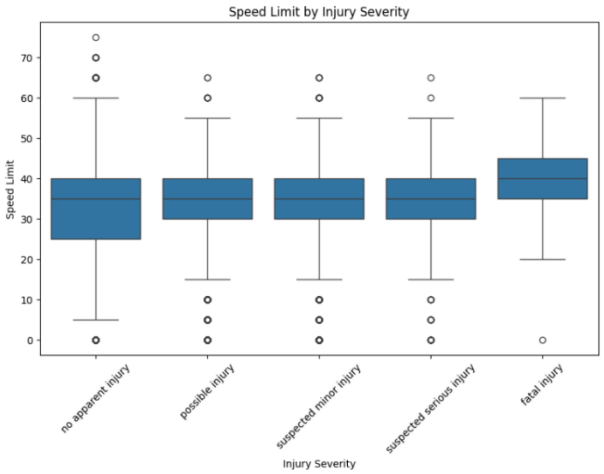
<u>גרף 3</u>: פיזור תאונות דרכים לפי "surface condition":



הגרף מציג את פיזור התאונות לפי מצב פני השטח. ניתן לראות כי רוב התאונות התרחשו במשטחים יבשים (**Dry**) עם כ-133,574 מקרים, בעוד משטחים רטובים (**Wet**) הם התנאי שטח השני בשכיחותו עם כ-30,805 מקרים. תנאים נדירים כמו קרח (**Ice**) ושלג (**Snow**) כוללים כמות מקרים נמוכה משמעותית, פחות מ-1,100 כל אחד. התוצאות מדגישות את השכיחות הגבוהה של תאונות בתנאים יבשים ורטובים, אך ייתכן שתנאים נדירים יותר הם בעלי השפעה משמעותית על חומרת התאונות.

גרף 4: הקשר בין מגבלת המהירות לרמת הפציעה בתאונות דרכים:



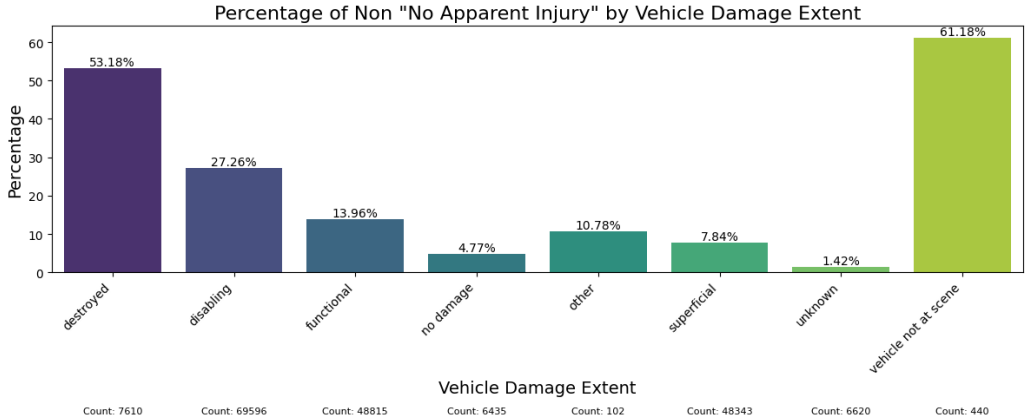


הגרף מציג את הקשר בין מגבלת המהירות באזורים שבהם התרחשו תאונות לבין רמות הפציעה השונות. ניתן לראות כי בתאונות קטלניות (Fatal Injury), מגבלת המהירות הממוצעת נוטה להיות מעט גבוהה יותר, כפי שמשתקף בחציון (Median) המעט יותר גבוה שלה בהשוואה לשאר רמות הפציעה. עם זאת, רוב התאונות, ללא קשר לרמת הפציעה, מתרחשות באזורים עם מגבלת מהירות דומה של כ-35-40 קמ"ש.

הגרף המציג את התפלגות מגבלות המהירות מחזק את הממצא הזה בכך שהוא מראה שמרבית התאונות מתרחשות בטווח המהירויות הנפוץ ביותר (35-40 קמ"ש), ללא קשר לרמת הפציעה, תוצאה זו מצביעה על כך שהמגבלה המהירות היא רק מרכיב אחד מתוך מגוון משתנים שיכולים להשפיע על רמת הפציעה.

הדמיון בטווחי מגבלת המהירות בין רמות הפציעה השונות מצביע על כך שמגבלת המהירות לבדה אינה מסבירה באופן ישיר את חומרת הפציעה. ייתכן כי גורמים נוספים, כמו למשל: תנאי הדרך, סוג ההתנגשות ,סוג הרכב ועוד, הם בעלי השפעה משמעותית יותר על רמת הפציעה, ולכן יש לבחון אותם בניתוחים נוספים.

גרף 5: מציג את הקשר בין רמת הנזק לרכב מתאונה לכמות הנפצעים בתאונה:



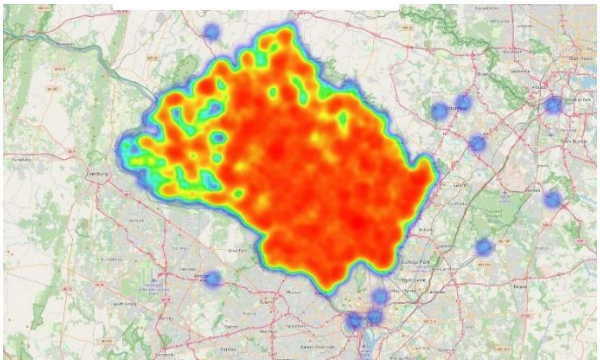
גרף זה מציג את אחוז המקרים שבהם דווח על פציעות (לא כולל את: "ללא פציעה נראית לעין") ביחס לרמת הנזק שנגרם לרכב בתאונה. ניתן לראות כי ככל שרמת הנזק לרכב חמורה יותר, כך עולה הסבירות לפציעות. לדוגמה, עבור רכבים שנהרסו לחלוטין ("destroyed"), אחוז המקרים עם פציעות עומד על 53.18%, בעוד שעבור רכבים ללא נזק נראה לעין (" no damage"), האחוז נמוך משמעותית ועומד על 4.77%. קטגוריות ביניים כמו "functional" ו-"superficial" מציגות אחוזים מתונים יותר של פציעות, העומדים על 13.96% ו-7.84% בהתאמה.

קטגוריה חריגה שמצביעה על אחוז פציעות גבוה במיוחד היא "vehicle not at scene", עם שיעור של 61.18%. קטגוריה זו עשויה לשקף תאונות שבהן הרכב המעורב לא היה נוכח בזירת התאונה או לא תועד כראוי. הסבר אפשרי לכך עשוי להיות מקרים של תאונות חמורות יותר, כגון תאונות פגע וברח, או חוסר מידע מלא על מאפייני התאונה והרכב.

נתונים אלו מדגישים את החשיבות של רמת הנזק לכלי הרכב כמדד מרכזי שיכול לסייע בחיזוי רמת הפציעה, תוך התייחסות למקרים חריגים שדורשים ניתוח נוסף.

גרף 6: מפת חום לפיזור תאונות עם פציעות לפי אזורים גאוגרפים גיאוגרפיים:

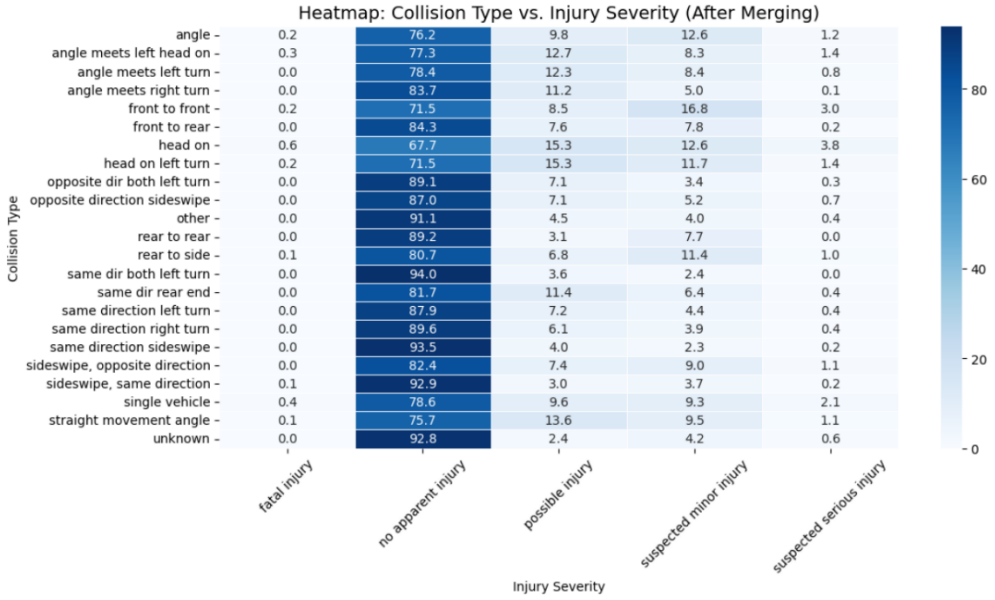
מפת החום מרחוק



מפת החום מקרוב

המפות המצורפות הן מפת חום (Heatmap) המציגות את פיזור התאונות שבהן דווח על פציעות (לא כולל "ללא פציעה נראית לעין"), על בסיס מיקומים גאוגרפיים לפי קווי אורך ורוחב. צבעי מפת החום משקפים את ריכוז התאונות באזורים שונים: צבע אדום מציין ריכוז גבוה של תאונות עם פציעות, בעוד שצבע כחול מייצג ריכוזים נמוכים יותר. ניתן לראות בבירור מוקדים גאוגרפיים שבהם מתרחשות תאונות רבות המובילות לפציעות, למשל אם זה אזורים עירוניים, כבישים מרכזיים או צמתים. ניתוח זה מדגיש את חשיבות המיקום הגאוגרפי בזיהוי מוקדי סיכון, דבר שעשוי לסייע בקבלת החלטות לשיפור הבטיחות באזורים אלו ולהפחתת התאונות באזורים בעלי ריכוז גבוה של תאונות.

גרף 7: הקשר בין סוג התאונה (Collision Type) לבין רמת הפציעה (Injury Severity)



ה-Heatmap מציג את ההתפלגות היחסית של **רמות הפציעה (Injury Severity)** עבור כל **סוג תאונה (Collision Type)**. כל תא בטבלה מראה את האחוז של אותה רמת פציעה מתוך סך כל התאונות באותו סוג תאונה.

ה-Heatmap מדגימה את הקשר בין סוגי התאונות לבין רמת הפציעה, ומראה כי רוב התאונות מסתיימות ב"אין פציעה נראית לעין", במיוחד בסוגי תאונות כמו "Sideswipe באותו כיוון" (93.5%) ו"אחר" (91.1%). סוגי תאונות בסיכון גבוה, כמו "חזית לחזית" ו"זווית פוגשת פנייה שמאלה", מציגים אחוז גבוה יותר של פציעות חמורות ומקרי מוות. פציעות קטלניות הן נדירות באופן כללי, אך הן נפוצות יותר בהתנגשויות חזיתיות. תובנות אלו מדגישות את הצורך באמצעי בטיחות ממוקדים, כמו שילוט משופר ואכיפה מחמירה יותר, עבור סוגי תאונות בסיכון גבוה, במטרה להפחית את ההשלכות החמורות.

סיכום שלב ניתוח הנתונים (Data Exploration)

במהלך שלב זה, בוצעו ניתוחים ויזואליים של הנתונים באמצעות גרפים וכלים סטטיסטיים, במטרה להבין את מאפייני הנתונים ולבחון את הקשרים האפשריים בין המשתנים לעמודת המטרה. להלן התשובות לשאלות המרכזיות שעלו בשלב זה:

1. **השערות ראשוניות עיקריות לגבי הנתונים:**
   - תאונות עם תנאי מזג אוויר קשים (כמו גשם, שלג או ערפל) יובילו לשיעור גבוה יותר של פציעות חמורות ביחס למזג אוויר בהיר.
   - ככל שמגבלת המהירות גבוהה יותר, כך עולה הסבירות לפציעות חמורות או קטלניות.
   - רמת הנזק לרכב (Vehicle Damage Extent) קשורה ישירות לחומרת הפציעה – נזק חמור יותר צפוי להוביל לפציעות חמורות יותר.
   - סוגי תאונות "חזית לחזית" או "זווית פוגשת פנייה" יובילו לשיעור גבוה יותר של פציעות חמורות ומוות ביחס לתאונות צדדיות או קלות יותר.
2. **מאפיינים הנראים מבטיחים לניתוח נוסף:**
   - **Injury Severity:** עמודת המטרה, המהווה את ליבת הניתוח והחיזוי.
   - **Vehicle Damage Extent:** מדד חשוב לקשר בין רמת הנזק לרכב לרמת הפציעה.
   - **Speed Limit:** משתנה מרכזי שיכול לשקף את רמת הסיכון באזורים שונים.
   - **Surface Condition & Weather:** תנאי מזג האוויר והכביש עשויים להיות בעלי השפעה על חומרת הפציעות.
   - **Collision Type:** סוג ההתנגשות קריטי להבנת חומרת הפציעה.
   - **Driver Substance Abuse:** משתנה שעשוי להציג קשר ישיר בין שימוש בחומרים מסוכנים לבין תאונות חמורות.
   - **Longitude & Latitude:** מיקום גאוגרפי לזיהוי מוקדי תאונות עם פציעות חמורות.
3. **ממצאים חדשים שהתגלו במהלך החקירה:**
   - חוסר איזון משמעותי בעמודת המטרה: רוב המקרים מסווגים כ-"No Apparent Injury," מה שמחייב שימוש בטכניקות כמו SMOTE לאיזון הנתונים.
   - תנאי דרך קשים כמו קרח או שלג נדירים יחסית אך עשויים להשפיע באופן משמעותי על חומרת הפציעות.
   - מגבלת מהירות אינה משפיעה באופן ברור: מרבית התאונות מתרחשות בטווח מהירויות בינוני (35-40 קמ"ש), ללא קשר ישיר לחומרת הפציעה.
4. **עדכון ההשערות הראשוניות בעקבות החקירה:**
   - ההנחה הראשונית הייתה שתנאי מזג אוויר קשים תמיד מובילים לתאונות חמורות, כלומר תאונות עם פציעות. אך הנתונים מראים שרוב התאונות מתרחשות במזג אוויר בהיר ומשטחים יבשים.
   - תנאי דרך קשים לא גורמים ליותר תאונות ,כפי שחשבנו אך כאשר הן קורות - יש להן סיכוי גבוה יותר לגרום לפציעות חמורות.
   - השערה נוספת הייתה שמגבלת מהירות גבוהה מובילה לפציעות חמורות יותר, אך נמצא שטווח מהירויות בינוני הוא השכיח ביותר. יש צורך לבחון גורמים נוספים, כמו סוג ההתנגשות והתנהגות הנהגים.
5. **השפעת החקירה על מטרות הפרויקט:**
   - המטרה המרכזית – חיזוי רמת הפציעה בתאונות דרכים – נותרה בעינה.
   - עם זאת, נוספו דגשים משניים:
     - טיפול בחוסר האיזון בנתונים באמצעות טכניקות כמו SMOTE או Weighted Loss.
     - ניתוח משתנים מרובי-ממדים כמו סוג התאונה ותנאי הדרך להשגת חיזוי מדויק יותר.
     - בדיקת נתונים חיצוניים נוספים, כגון נתונים דמוגרפיים או צפיפות תנועה, כדי לשפר את מודל החיזוי

איכות הנתונים:

נתונים חסרים (Missing Data)

במהלך ניתוח הנתונים נמצא כי מספר עמודות מכילות אחוזים גבוהים של ערכים חסרים, מה שעשוי להשפיע על איכות המודל ודיוק התחזיות. בנוסף, בטבלת הנתונים שלנו ערכים חסרים מקודדים לא רק כריקים, אלא גם בצורות שונות כגון Uknown, N/A (חלק מהעמודות יכול להיות בעל משמעות וחלק יכול להיות ערך חסר) ועוד, מה שמצריך אחידות בטיפול בהם.

דוגמאות לעמודות בהן יש N/A: Weather , Light, Traffic Control וכולי.

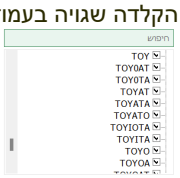
- ישנן עמודות שבהם חלק מהערכים הינם: -, 0,? ומספרים רנדומליים (כמו 99) שאינם קשורים לתוכן העמודה- ערכים אלו יכולים לייצג ערכים חסרים או יכולים להיות טעות הקלדה. לדוגמה בעמודת Vehicle Make ובעמודת Vehicle Model.

שגיאות נתונים (Data Errors)

נמצאו מספר אי-סדירויות בנתונים, כולל שגיאות הקלדה וחוסר אחידות בערכים, מה שעלול לגרום לבעיות בניתוח. דוגמאות כוללות:

- בעמודת Vehicle Body Type אחד הערכים מופיע ככה: "Van - Passenger (<9 Seats)" .

- הקלדה שגויה בעמודת Vehicle Make, לדוגמה הרכב Toyota מופיע כמה פעמים עם טעויות כתיב:



אישור ביטול דוגמה לכמה מהם.

שגיאות מדידה (Measurement Errors)

שגיאות מדידה מתרחשות כאשר הערכים שהוזנו נכונים מבחינה טכנית אך מבוססים על שיטת מדידה שגויה או חוסר דיוק במקור הנתונים. דוגמאות כוללות:

- מהירות מותרת– (Speed Limit) : ערכים נעים בין 0 ל-75, לדוגמה ערך 0 למהירות מותרת עלול להיות שגיאת הקלדה או להצביע על חניה.

- Vehicle Year - ערכים מחוץ לטווח ההגיוני למשל, שנת ייצור 9999,5,0, 2911, 15 וכולי. (ערכים כמו 2911 יכולים להיות טעויות הקלדה, וערכים כמו 9999 או 0 יכול להיות ערך חסר).

חוסר עקביות בקידוד(Coding Inconsistencies)

בעיות עקביות בקידוד נובעות מהיעדר אחידות בערכים הקטגוריאליים. דוגמאות:

1. עמודות בהן יש ערכים עם אותה משמעות אך יש אי אחדות מבחינת אותיות גדולות/קטנות. דוגמאות:

- בעמודת המטרה שלנו Injury Severity לכל ערך יש ערך כפול לדוגמה- "NO APPARENT INJURY " ו-" No Apparent Injury".
- בעמודת Weather ערך אחד הוא Clear ואחר- CLEAR.
- בעמודת Vehicle First Impact Location ערך אחד הוא "Eight O Clock" ואחר- "EIGHT OCLOCK" (גם הערך כתוב קצת אחרת).

2. חלק מהעמודות מכילות ערכים לא אחידים עבור אותו משתנה, קיימות וריאציות שונות, מה שעלול להוביל לקשיים בניתוח הנתונים. דוגמאות:

- בעמודת Route Type יש ערך: Maryland (State) ויש ערך אחר- "Maryland (State) Route"
- בעמודת Surface Condition יש ערך: "ICE" ויש ערך "Ice/Frost".
- בעמודת Light יש ערך: "Dark- Lighted" ו- "DARK LIGHTS ON".
- בעמודת Non-Motorist Substance Abuse יש ערך: "UNKNOWN" ו- "Unknown, Unknown".

3. תאריך ושעה (Crash Date/Time) - התאריכים מופיעים בפורמטים שונים:

לדוגמה: בחלק מהשדות הפורמט הוא: "08/17/2018 02:25:00 PM" כלומר עם AM/PM, וחלק בפורמט: "09/11/2015 13:29:00".

מטא-נתונים שגויים (Bad Metadata)

- Location- נתון כפול ל Latitude-וLongitude.

נקודה נוספת:

כמו כן יש לנו עמודות שמכילות בשדה אחד 2 ערכים שונים שמופרדים על ידי ",". לאחר מכן יתכן שבחלק מהמודלים נצטרך לעשות dummies . דוגמאות: בעמודת Related Non-Motorist אחד הערכים הינו: BICYCLIST, "PEDESTRIAN", בעמודת Driver Substance Abuse אחד הערכים הוא: Not Suspect of Alcohol Use, Not "Suspect of Drug Use".