# פרויקט מסכם בינה עסקית - **40205**

תום כספי 206347072

עומר בן יעקב 205689060

מתן דוברובנסקי 318624426

## שאלה עסקית + KPIs

מטרת הפרויקט הינה לנתח ולהבין את נתוני התעבורה של קליפורניה כמדגם מייצג, בכדי לזהות בעיות, מודלים וטכניקות של סיכונים ואכיפת חוק בכבישים. הפרויקט מתמקד בחשיפת מדדים ומצביעים על התנהלות התעבורה ופעילות האכיפה לצורך הבנת השאלות המרכזיות של המחקר. באמצעות הניתוחים והמדדים, אנו נספק מידע מועיל ורלוונטי המשמש לשיפור התעבורה, האכיפה והבטיחות בכבישים, ולסיוע בקבלת החלטות מבוססות על נתונים אמיתיים.

1. ראש הטופס

**שאלת מחקר 1** :האם יש קשר בין מגדר הנהגים לסוג העבירה שבוצעה? במידה ויש, האם קשר זה סטטיסטית חשוב ורלוונטי?

מדדי KPI עבור שאלה מחקר 1:

1.אחוז הנשים (driver gender = "Female") שביצעו עבירות מסוימות (violation) מתוך כל מי שביצע את העבירה הספציפית. אם יימצא שלמעלה מ55% נגיד שקיים קשר

2.אחוז העצירות שהובילו למעצר (Arrest Rate) לפי מין הנהג (driver\_gender). ניתן לחשב את האחוז על פי כמות העצירות שהובילו למעצר ולחלק בכלל העצירות לפי מין הנהג.

3.יחס הימצאות של חומר סמים אסורים (Contraband Hit Rate) בין גברים לנשים. ניתן לחשב את היחס בין כמות העצירות בהן נמצא חומר סמים אסורים (contraband\_found = True) לגברים לבין היחס לנשים.

**שאלת מחקר 2**: האם קיים קשר בין גיל הנהגים (driver raw age) להימצאות של חומרים אסורים ברכב לאחר שבוצעה בדיקה

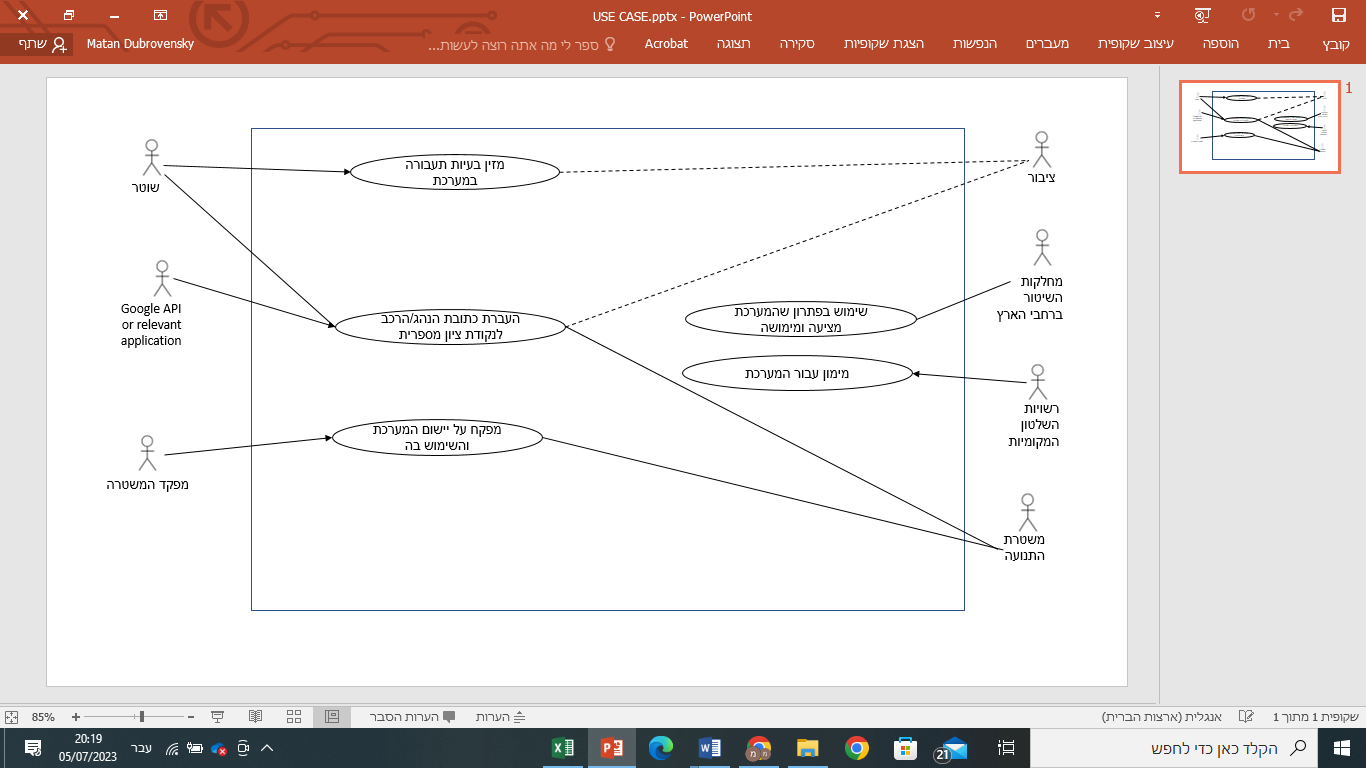
מדדי KPI עבור שאלה מחקר 2:

1. אחוז החיפושים שבוצעו במהלך העצירה (search conducted = True) מתוך כל העצירות, לפי קבוצות גילאים.
2. אחוז המציאות של חומר סמים אסורים (contraband found = True) בין העצירות בהן בוצע חיפוש, לפי קבוצות גילאים.
3. יחס המציאות של חומר סמים אסורים (contraband found = True) אם יימצא ש 65% מהימצאותן של חומרים אסורים הן בקרב צעירים אז יהיה ניתן לומר שקיים קשר.

## הגדרת Data Warehouse

פשטות: סכמת Star מציעה עיצוב פשוט ואינטואיטיבי. היא כוללת טבלת עובדות מרכזית, שמכילה את המדדים המדויקים לדוגמה contraband found is arrestedd, המוקפת טבלאות ממד שמספקות אפיונים תיאוריים לדוגמה:  
(police department, driver gender, driver age, driver race violation).

ביצועי חיפוש מיטביים: סכמת כוכב מתאימה היטב לניתוח ולקיום השאילתות. המבנה הדינמי שלה מפחית את מספר השימוש ב join בשאילתות, בעזרת סכמת הכוכב, המשתמש יוכל לפנות לטבלה ייחודית ומשפר את ביצועי החיפוש. ניתן לבצע פונקציות אגרגטיביות וסינון בקלות בטבלאות ממדים, בעוד שטבלת העובדות מכילה את מדדי המפתח, מה שמאפשר אחזור נתונים מהיר יותר.



## תהליך הETL

* Extract - שלב חילוץ המידע כולל את תהליך ייבוא הנתונים מהמקורות השונים שקיימים לנו, המקור העיקרי הוא דאטה סט מאתר Kaggle. בנוסף, תהליך זה כולל תקשורת עם מקורות הנתונים ואת השליפה של הנתונים עצמם. המקורות יכולים להיות מסדי נתונים, קבצים מקומיים ועוד.  
  בשלב זה, הנתונים מועברים ממקורם המקורי לשלב הבא בתהליך.
* Transform - עיבוד ותיקון: שלב העיבוד והתיקון מתרחש לאחר שהנתונים מיובאים מהאתר עצמו. בשלב זה, אנו דואגים כי הנתונים הגולמיים יעברו מס' שלבים כגון:
  + ניקוי הנתונים מפגמים כגון ערכים חסרים, ערכים לא תקינים, ערכים מותאמים אישית וכו'.
  + פילוח הנתונים לפי מילון שדות ופרמטרים מוגדרים מראש.
  + שינוי פורמט הנתונים לתקן מסוים או פורמט שמתאים למטרות העיבוד והאנליזה המבוקשות.
  + הוספת חישובים, תקופות זמן, שינויי מבנה ועוד.
  + לסיכום שלב העיבוד והתיקון מתחיל עם הנתונים הגולמיים שנאספו בשלב החילוץ ומסתיים עם יצירת נתונים מועילים ומקוננים בצורה מתואמת ומוכנה לטעינה.
* Load - טעינה: שלב הטעינה מתרחש לאחר השלמת שלב העיבוד והתיקון. בשלב זה, הנתונים מועברים לExcel שבו הדאטה נשמר ומנוהל. יתרון הטעינה הוא שניתן לחבר מספר מקורות של נתונים באותו מסד נתונים ולטעון אותם למסדי נתונים שונים לפי הצורך.   
  בשלב זה, ניתן גם לייצר אינדקסים, מפתחות עיקריים ותנאים עבור הנתונים כדי לשפר את ביצועי השאילתות והפעולות האנליטיות על הנתונים.

## ניתוח הData Warehouse

## חישוב מונה מספר הפעמים שסוג העבירה הינה מהירות מופרזת וגם נערך חיפוש ברכב:

SELECT COUNT(\*) FILTER (WHERE violation = 'speeding' AND search\_conducted = True) OVER () AS count\_speeding\_searches

FROM BIData;

## חישוב ממוצע גיל עבור כל סוג עבירה:

SELECT violation, AVG(driver\_age\_raw) OVER (PARTITION BY violation) AS average\_age

FROM BIData;

## זיהוי העבירה הנפוצה לפי מוצא:

SELECT DISTINCT ON (driver\_race) driver\_race, violation, violation\_count

FROM (

SELECT driver\_race, violation,

COUNT(\*) OVER (PARTITION BY driver\_race, violation) AS violation\_count,

ROW\_NUMBER() OVER (PARTITION BY driver\_race ORDER BY COUNT(\*) DESC) AS rn

FROM BIData

GROUP BY driver\_race, violation

) AS subquery

WHERE rn = 1;

## חישוב חמשת העבירות הנפוצות בכל מדינה בסדר יורד:

SELECT country, violation, violation\_count

FROM (

SELECT country, violation, COUNT(\*) AS violation\_count, ROW\_NUMBER() OVER (PARTITION BY country ORDER BY COUNT(\*) DESC) AS rn

FROM BIData

GROUP BY country, violation

) AS subquery

WHERE rn <= 5;

## חישוב אחוז העבירות שגרמו למעצר לפי מדינה :

SELECT country, violation, arrest\_percentage

FROM (

SELECT country, violation,

COUNT(\*) FILTER (WHERE stop\_outcome = 'arrested') \* 100.0 / COUNT(\*) OVER (PARTITION BY country) AS arrest\_percentage,

ROW\_NUMBER() OVER (PARTITION BY country ORDER BY COUNT(\*) DESC) AS rn

FROM BIData

GROUP BY country, violation

) AS subquery

WHERE rn = 1;

## אחוז העבירות שתוצאתן הייתה אזהרה מילולית- גברים למול נשים לפי סוג עבירה:

SELECT violation,

100.0 \* COUNT(\*) FILTER (WHERE driver\_gender = 'F' AND stop\_outcome = 'verbal warning')

/ SUM(COUNT(\*) FILTER (WHERE driver\_gender = 'F') OVER (PARTITION BY violation))

OVER (PARTITION BY violation) AS women\_verbal\_warning\_percentage,

100.0 \* COUNT(\*) FILTER (WHERE driver\_gender = 'M' AND stop\_outcome = 'verbal warning')

/ SUM(COUNT(\*) FILTER (WHERE driver\_gender = 'M') OVER (PARTITION BY violation))

OVER (PARTITION BY violation) AS men\_verbal\_warning\_percentage

FROM BIData

WINDOW w AS (PARTITION BY violation);

## אחוז העבירות שנמצא בהם חומר אסור לפי סוג עבירה:

SELECT

violation,

100.0 \* COUNT(\*) FILTER (WHERE contraband\_found = TRUE)

/ SUM(COUNT(\*)) OVER (PARTITION BY violation) AS contraband\_percentage

FROM BIData

WINDOW w AS (PARTITION BY violation);

## מציג את המיקומים בהם קרו הכי הרבה מקרים ואת חמשת המקרים הנפוצים ממוינים בסדר יורד:

SELECT

location raw,

violation,

incident count

FROM (

SELECT

location\_raw, violation,

COUNT(\*) AS incident\_count,

ROW\_NUMBER() OVER (PARTITION BY location\_raw ORDER BY COUNT(\*) DESC) AS rn

FROM BIData

WINDOW w AS (PARTITION BY location\_raw)

) AS subquery

WHERE rn <= 5

ORDER BY incident\_count DESC;

## מסקנות

## בהיבטי חישוב חמשת העבירות הנפוצות בכל מדינה בסדר יורד ביחס לכל הערים שנמצאות בדאטה סט עלה כי סן דייגו בקירוב עם הרבה מקרי הפרות (incident\_count = 198), המדינות עם מספר האירועים הגבוה ביותר הן קליפורניה, טקסס ופנסילבניה.

## עבור חישוב ממוצע הגיל עבור כל סוג של הפרה (עבירה)התקבל כי נהיגה תחת השפעה (DUI) (31.1) אי שימוש בחגורת בטיחות (28) הפרת מהירות (28) ועוד דברים נוספים (36) הם אלה הגורמים העיקריים לסוגי העבירות.

## עבור חישוב חמשת העבירות הנפוצות בכל מדינה בסדר יורד קיבלנו כי עבירת מהירות הינה העבירה הנפוצה ביותר.

## יש לציין כי נתונים אלו ממחישים את הצורך בהעלאת המודעות בצורה טובה יותר לסכנות הנהיגה תחת השפעות שונות (אלכוהול, סמים ועוד). בלי קשר, חשוב גם לחנך את הציבור על חשיבות חגורות בטיחות (Equipment) והקפדה על מהירות הנסיעה המותרת. המטרה הגדולה של כולנו הינה להפחית את מספר תאונות הדרכים ולהציל חיים.