ענווה אברהם 208886333;
רעות בן חמו 209299478;

ויזואליזציה של מידע- דו"ח מסכם לפרויקט

30.06.2023

חלק 1 – מבוא

הנושא שאנו בחרנו לעשות עליו את הפרויקט הוא תאונות דרכים בארה"ב, אנחנו נרצה לבחון בעזרת דאטה סט על תאונות בארה"ב שנאסף בין השנים 2016-2021 את המאפיינים השונים שגרמו לאותה תאונה ונרצה לבחון האם ישנם פרמטרים שמשפיעים יותר מאחרים ונוכל להגיד שהם אחד מהגורמים שגרמו לאותה התאונה.

מקור הדאטה סט שלנו הוא מאתר Kaggle. (יש לציין כי מאז הורדת הדאטה לוקאלית אצלנו הדאטה באתר התעדכן עד מרץ 2023)

בדאטה סט שלנו יש 2.85 מיליון רשומות ו47 עמודות פיצ'רים שונות שנאספו מכ-49 מדינות שונות. כל רשומה מתארת תאונה שקרתה ומולאו כלל העמודות עבור אותה תאונה, על העמודות השונות נפרט מייד.

הפיצ'רים בקובץ מתארים מאפיינים שונים של התאונה ושל הסביבה בה התרחשה התאונה כגון נתוני מזג אוויר, עיקולים בכביש, זמן התאונה, מיקום התאונה וכולי.

לכל תאונה בדאטה מזהה ייחודי משלה בעמודה ID והוא משמש כמפתח ראשי בנתונים שלנו.

הדאטה סט שלנו הוא מסוג flat table משום שכל תאונה היא ישות ונמצאת בשורה נפרדת, וכן כל עמודה היא תכונה שונה.

המידע נאסף במספר דרכים: משרדי תחבורה במדינות השונות בארה"ב, רשויות החוק, מצלמות תנועה וחיישני תנועה בכבישים וכו'..

שאלת המחקר המרכזית עליה נרצה לענות היא "מה הם הגורמים המשפיעים על הסיכוי לתאונה?" נרצה לדעת האם אותם הגורמים מגדילים או מקטינים את הסיכוי לתאונה משום שנרצה למנוע את התאונות כמה שניתן ולכן נרצה למצוא את הגורמים המשפיעים ביותר.

שאלות משנה עליהן נרצה לענות:

a) "האם שינויי מזג אוויר כאלו או אחרים משפיעים באופן ישיר על סיכוי גבוה יותר לתאונות, גם כאשר הדבר איננו ברור ממבט ראשוני?"

b) "האם תמרורים, רמזורים, עיקולים, שטחי הפרדה וכולי.. הם גורמים שמעלים את הסיכוי או מורידים את הסיכוי לתאונה במקום מסוים? האם ישנה קורלציה מסוימת בין הגורמים הללו ?"

c) "באיזו מידה גורמים משניים למזג אוויר כמו טמפרטורה ורוח משפיעים על קיום תאונה?"

d) "באילו מדינות (בארה"ב) יש הכי הרבה תאונות?"

ענווה אברהם 208886333;
רעות בן חמו 209299478;

חלק 2 – נתונים ומטלות

1. נתונים :

יש לנו מספר data types והם: categorical (Nominal, Ordinal, Binary), quantitative, Date. מאחר ויש לנו 47 Attributes אין לנו צורך בכולם ולכן התמקדנו במס' attributes מרכזיים אשר יסייעו לנו לענות על שאלת המחקר שלנו. בנספחים [1] ניתן למצוא טבלה המפרטת את כלל המשתנים שלנו וכן פירוט מלא עליהם. נחלק אותם כעת לקטגוריות:

1. **Nominal**: ID, City, State, Wind Direction, weather condition, amenity, Sunrise sunset, Civil twilight, Nautical twilight, Astronomical twilight
2. **Ordinal**: Severity
3. **Quantitative**: Temperature (F), wind chill, humidity, pressure, visibility, wind speed, precipitation
4. **Binary**: Bump, Crossing, Give way, Junction, No exit, Railway, Roundabout,

2. מטלות:

מטלות המשתמש הן להבין אילו גורמים עלולים להגביר את הסיכוי או להוריד את הסיכוי לתאונה, בסמוך לשאלות המחקר אותן אנו בודקות.

באמצעות הוויזואליזציה ניתן לסדר את הנתונים האלה בצורה ברורה שתקל עלינו לזהות מגמות ולנתח את הדאטה בצורה המיטבית.

בדאטה קיימות 2.85 מיליון רשומות כאשר כל רשומה היא תאונה . למשתמש יהיה בלתי אפשרי לעבור על כל הרשומות ולהסיק מסקנות. לכן אנו נרצה לבנות את הוויזואליזציה הטובה ביותר כך שתוכל לסייע למשתמש לענות על השאלות הללו.

משום שאין לנו דיי מקום כדי להכניס את כלל ההסברים המפורטים ניתן לראות תחת נספחים [2] את הטבלאות המלאות. נראה כעת טבלה מצומצמת מאוד עבור השאלות שלנו:

Tasks	Target	Action
זיהוי attributes הקשורים למז"א וחקירת השפעתם על תאונה	Attributes: Distribution	Analyze: Discover
		Search: Lookup
		Query: Identify
ניתוח הקשרים בין הattributes השונים	Attributes: Correlation	Analyze: Consume – Discover & Present
		Search: explore
		Query: Compare & Summarize
חקירת השפעתם של מצבי מזג אוויר שונים על חומרת התאונה	Attributes: Similarity	Analyze: Consume - Discover
		Search: Locate
		Query: Compare
הבנה באילו מדינות (בארה"ב) יש הכי הרבה תאונות	All Data: Features	Analyze: present
		Search: explore
		Query: Summarize

ענווה אברהם 208886333;
רעות בן חמו 209299478;

חלק 3 - וויזואליזציה

1. קישור לוויזואליזציה , קישור לגיט
2. **Preprocessing**:

הדאטה שלנו מכיל 2.85 מיליון רשומות ולא הייתה לנו את היכולת לעבוד עם כל הרשומות הללו.
משום כך, בחרנו לבצע דגימה מתוך הדאטה סט ולעבוד על כמות רשומות קטנה יותר. בשביל זה,
ביצענו דגימות של כל 40 רשומות מתוך הדאטה סט וקיבלנו בסוף דאטה של 193,210 רשומות.
תחילה ייבאנו את הקובץ מתוך Kaggle:

```
! mkdir ~/.kaggle
! cp kaggle.json ~/.kaggle/
! chmod 600 ~/.kaggle/kaggle.json
! kaggle datasets download -d sobhanmoosavi/us-accidents
```

לאחר מכן ביצענו חילוץ של הקובץ ואז יצירת הקובץ החדש:

```
[ ] !unzip us-accidents.zip

[ ] step = 40

    # Custom function to skip rows based on the step size
    def skip_rows_func(index):
        return index % step != 0

    # Read the CSV file and skip rows based on the step size
    df = pd.read_csv('/content/US_Accidents_March23.csv', skiprows=skip_rows_func)

    df.to_csv('data_new2.csv', index=False)
```

לבסוף כיווצנו את הקובץ לZIP על מנת שנוכל להעלות אותו לGIT.

3. **הסבר על הקוד –**

בקוד ביצענו ויזואליזציות וניתוח על סט הנתונים. השתמשנו במספר חבילות Python:

- streamlit: משמש לבניית הדש בורד וממשק המשתמש.
- Zipfile: משמש לחילוץ קובץ CSV מארכיון ZIP.
- matplotlib.pyplot: משמש ליצירת Heatmap של מטריצת הקורלציה.
- numpy: משמש לפעולות וחישובים.
- Pandas: משמש למניפולציה וניתוח נתונים.
- plotly.express: משמש ליצירת הדמיות אינטראקטיביות, כגון Bar Plot .
- plotly.graph_objects: משמש ליצירת הדמיות מותאמות אישית, כגון Box Plot.
- seaborn: משמש ליצירת Heatmap של מטריצת הקורלציה.

הקוד כולל את הוויזואליזציות הבאות:

Box Plot: מנתח את הקשר בין טמפרטורה לחומרת התאונה.

מטריצת קורלציה: מחשבת ומציגה את המתאם בין הattributes השונים באמצעות Heatmap.

Bar Plot: מראה כיצד תנאי מזג אוויר שונים משפיעים על חומרת התאונה.

מפת ארה"ב: מציג את מספר התאונות לכל מדינה בארה"ב.

ענווה אברהם 208886333;
רעות בן חמו 209299478;

תצוגת נתונים: הוספנו בסוף הדשבורד הצגה של מערך הנתונים על מנת להציג למשתמש כיצד נראה סט הנתונים.

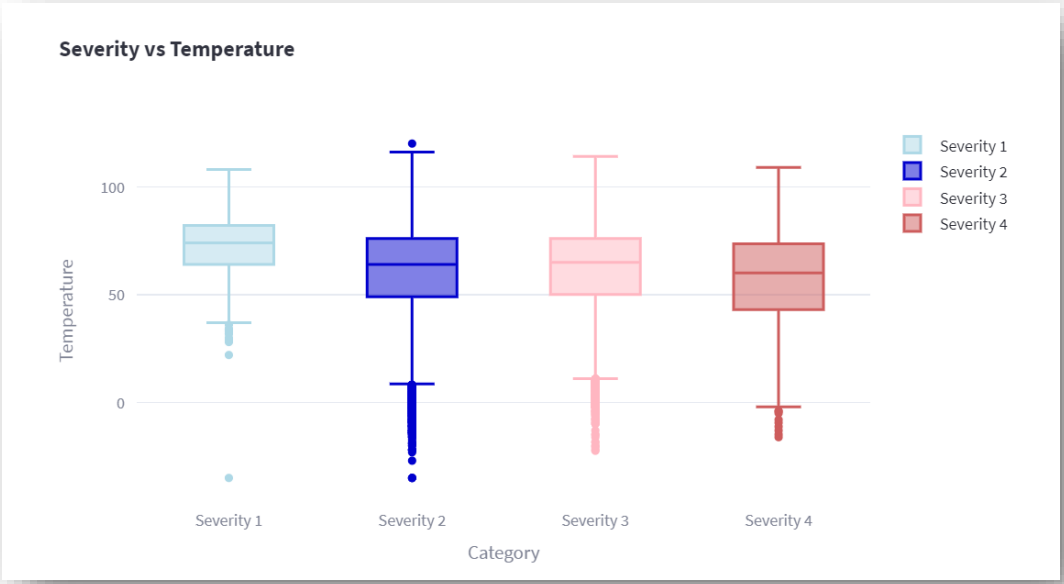
וויזואליזציות אלו מספקות תובנות לגבי השפעת הטמפרטורה, תנאי הדרך, תנאי מזג האוויר והמיקום הגיאוגרפי על חומרת התאונה.

4. **הסבר על התרשימים**:

i. **Box Plot** – How does temperature (Fahrenheit) affect accidents?

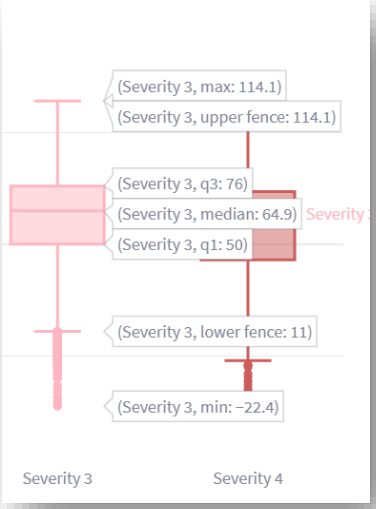
מטרת התרשים היא לענות על השאלה האם טמפרטורה משפיעה בצורה משמעותית על דרגת חומרת תאונה גבוהה יותר\נמוכה יותר. כלומר, האם כאשר יש טמפרטורות קיצוניות יש תאונות חמורות יותר, או אילו מסקנות ניתן להסיק מהבנה של האם טמפרטורה משפיעה בצורה ניכרת על חומרת התאונה?

למשל, ניתן לראות שעבור חומרת תאונה 4 הטמפרטורות נמוכות יותר בממוצע מאשר חומרת 1. נוכל למשל להסיק שטמפרטורות נמוכות גוררת גשם\רוח\שלג\קרח ואלו גורמים שיכולים להשפיע מאוד על תאונות. קרח עלול לגרום להחלקה, גשם ורוח עלולים לגרום לחוסר יציבות ונראות לקויה.

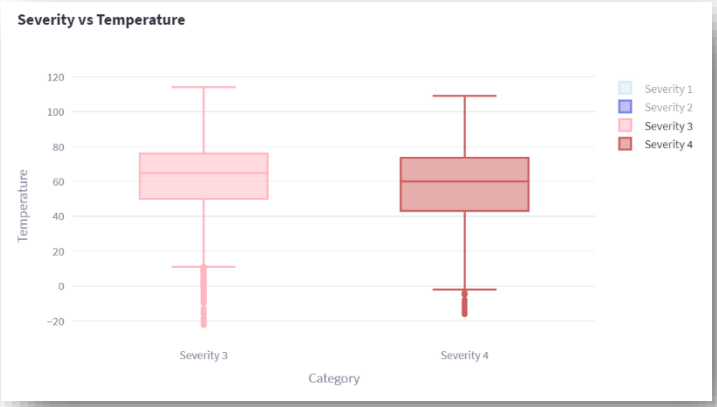


ענווה אברהם 208886333;
רעות בן חמו 209299478;

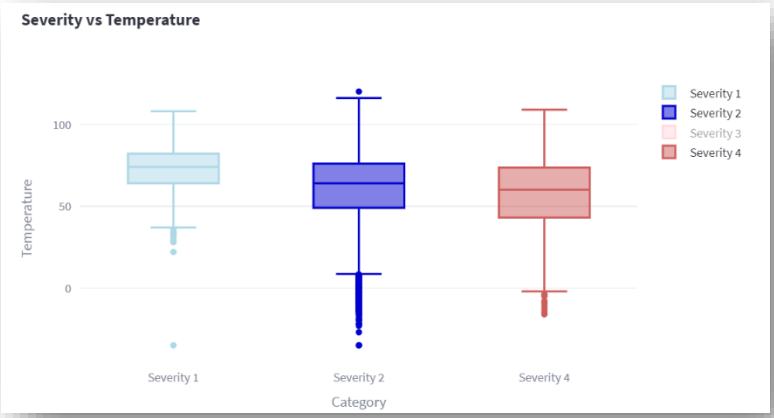
כאשר עומדים עם העכבר על אחת הקטגוריות מופיע ה-hover:



נוכל לסנן ולראות רק דרגות מסוימות ע"י סימונן :



ונוכל גם למשל להוריד את אחת הדרגות ע"י לחיצה על דרגת חומרתה בצד ימין במקרא. למשל לחצנו על 3 וקיבלנו:



ענווה אברהם 208886333;
רעות בן חמו 209299478;

Marks & Channels

Marks: נקודות, קווים.

Channels: מיקום ביחס לציר ה-Y (vertical position). המיקום על ציר ה-Y משקף את ערכי הטמפרטורה (בפרנהייט) עבור דרגת החומרה.

צבע – Hue בהתאם לדרגות החומרה. הצבע (Hue) משקף את הקטגוריות השונות – דרגות החומרה השונות.

Effectiveness and Expressiveness

אקספרסיביות: עבור הערכים הקטגוריאליים השתמשנו בצבעים שונים ע"י Hue אשר כל צבע מבטא קטגוריה אחרת בדרגות החומרה השונות. בנוסף מיינו את הערכים מקטן לגדול לצורך התמצאות נוחה יותר בגרף וכן הצבעים מבחול בהיר לאדום כהה המבטאים גם כן את העלייה בדרגת החומרה.

עבור הערכים הנומריים השתמשנו במיקום האנכי ע"ג ציר ה-Y אשר מבטא סטטיסטיקות שונות אודות הטמפרטורה עבור אותה קטגוריה. לדוגמה, הפס בתוך ה-box מבטא את החציון וגודל הקופסא עצמה מבטא את 50% האמצעיים של הדאטה. כלומר המיקום ביחס לציר שופך אור על הערכים עבור אותה קטגוריה. משום כך, הגרף עונה על הקריטריונים של אקספרסיביות.

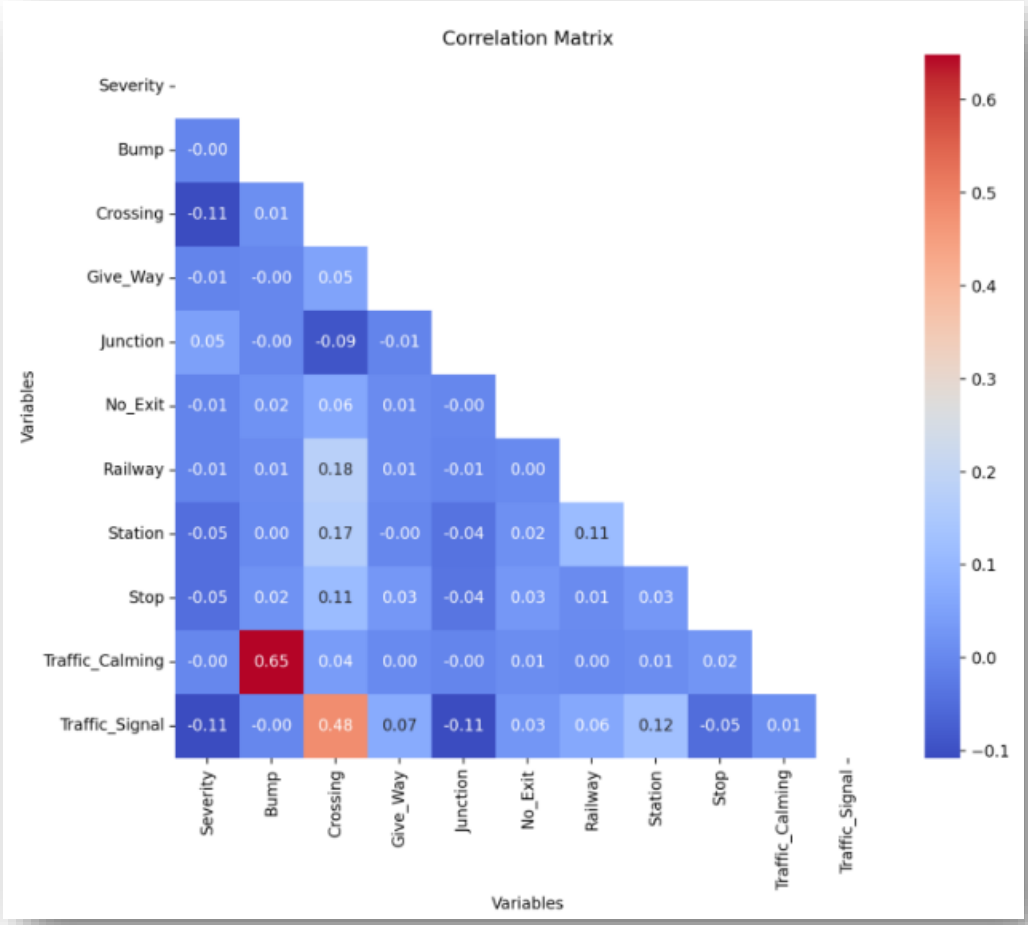
אפקטיביות: נבחן לפי 4 מדדים :

1) Accuracy – ניתן להבחין בבירור בהבדלים בין הboxes השונים משום שכל Box בצבע שונה – מבטא קטגוריה מבין דרגות החומרה. בנוסף, ניתן להבדיל בין הערכים הסטטיסטיים השונים אשר נמצאים לכל אורך ציר ה-Y מאחר ולכל box ערכים משלו.
2) Discriminability- עבור ערך זה יש לנו 4 רמות, כאשר כל רמה מייצגת קטגוריה (מיוצג ע"י הצבע, Hue).
3) Separability- הchannels אינם תלויים אחד בשני משום שהbox plot השונים נמצאים פחות או יותר באותו הטווח ולכן ציר ה-Y לא משתנה גם כאשר מסננים חלק מהצבעים.
4) Pop up- אפשר להבחין בקלות בין הקטגוריות השונות בגלל הצבעים השונים.

ענווה אברהם 208886333;
רעות בן חמו 209299478;

ii. **Correlation matrix –** How does one road condition affect another?
מטרת התרשים היא להבין האם ישנם קשרים ברורים בין תנאי דרך שונים.
להשקפתנו, תנאי דרך הם גורם משמעותי בתאונות הדרכים ואנו נרצה להבין האם ישנם תנאיי דרך אשר להם קורלציה חזקה. נרצה לבחון מה הם אותם תנאיי הדרך וכיצד הם מתקיימים בפועל, ולבחון מדוע השילוב הזה הוא בעל קורלציה חזקה יותר מאחר.

כאשר המשתמש יסתכל על גרף זה הוא יכול להסיק מגוון מסקנות בהתאם למה שמעניין אותו. למשל, מעניין לראות שעבור מעבר חצייה, יש קשר מאוד חזק עם רמזור אבל קשר משמעותית יותר חלש עם שלט עצור. המשתמש יוכל לראות את הנתונים הללו ולהבין באילו צמדים להתרכז.



הגרף איננו אטרקטיבי מאחר ובחרנו לשים את הערכים בריבועים עצמם לא ראינו צורך בהוספה מיותרת של ערכים נוספים כאשר "נעמדים" על קובייה מסוימת.

Marks & Channels

Marks: שטח (ריבועים).

Channels: הצבע (סטורציה) אשר מתאר את טווח הצבעים (ערכים) האפשרי – מכחול כהה (קשר חלש מאוד) לאדום כהה (קשר חזק מאוד).

ענווה אברהם 208886333;
רעות בן חמו 209299478;

מיקום אשר מבטא את הצמדים – לכל צמד יש ריבוע ייעודי עבור הקשר שלו. המיקום הינו לפי מיקום הערכים השונים על הצירים ומתוך כך נגזר מיקום ההצלבה ביניהם.

Effectiveness and Expressiveness

<u>אקספרסיביות:</u> השתמשנו בצבע (סטורציה) מאחר ומדובר כאן על טווח ערכים בין 1- ל1 שמבטאים את חוזק הקשר בין שני משתנים והצבע מבטא את המיקום היחסי בין אותו הטווח. בחרנו בצבעים הבסיסיים כחול ואדום משום שבעינינו הם מבטאים טוב את היחסיות.
בחרנו להשתמש רק בחצי מטריצה משום שלטעמינו אין צורך בכל המטריצה מאחר והיא סימטרית ולכן אין טעם בוויזואליזציה מיותרת.
האלכסון הראשי לא קיים מאחר והוא לא רלוונטי משום שהערכים בו הם תמיד 1 (בין משתנה לעצמו) ולא מעניין אותנו הקשר בין משתנה לעצמו.
מיקום הריבועים מתאים להצטלבות בין כל שני attributes.
כל ריבוע במקומו קיבל צבע המתאר את חוזק הקשר בין אותם שני attributes שהצטלבו.
משום כך, הגרף עונה על הקריטריונים של אקספרסיביות.

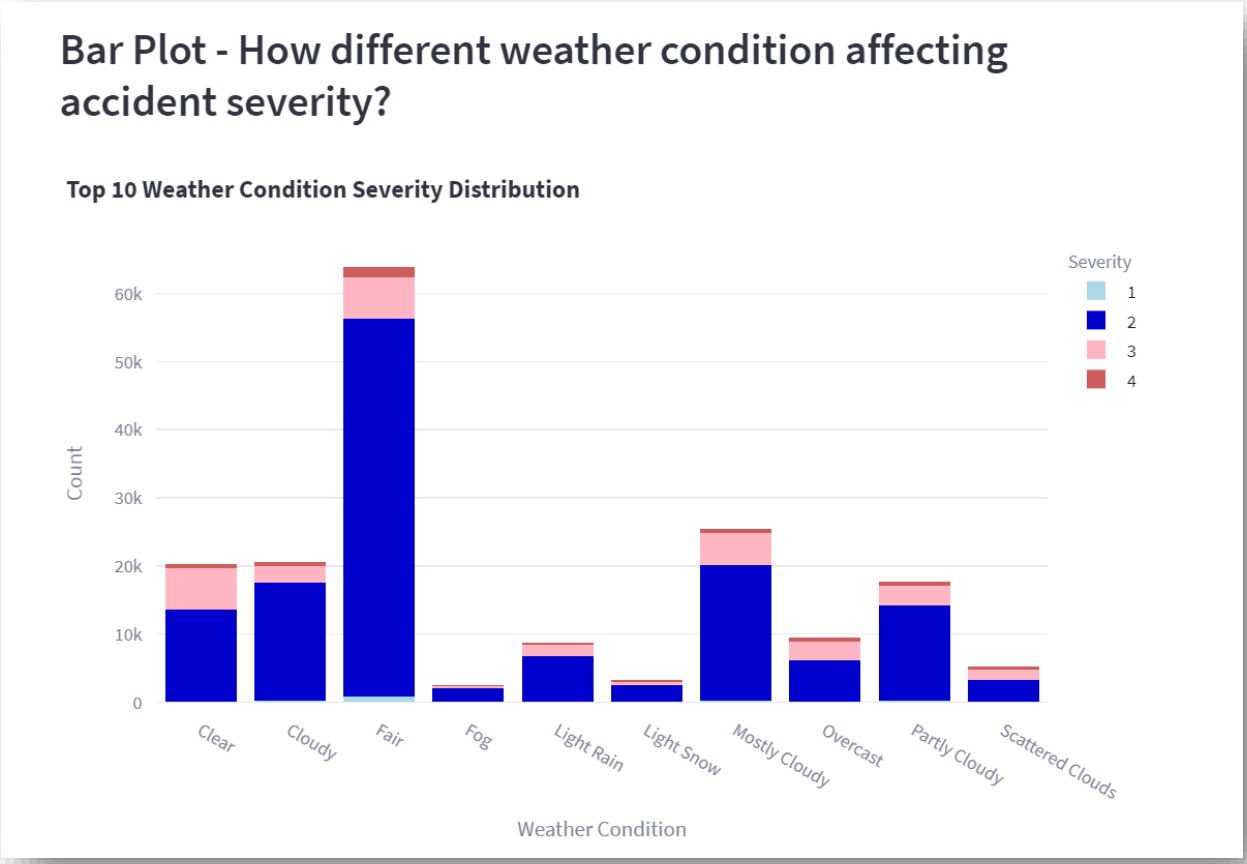
<u>אפקטיביות:</u> נבחן לפי 4 מדדים :

1) Accuracy – ניתן להבחין בבירור בהבדלים בין הריבועים השונים משום שבין כל ריבוע וריבוע יש קו שחור המסמן את הגבולות של כל ריבוע. בנוסף, מגוון הערכים השונים בריבועים בטווח בין 1- ל1 מסייע גם כן משום שניתן להבדיל טוב בין הצבעים השונים. גם בחירת הצבעים המנוגדים (כחול ואדום) מסייעת לכך.
2) Discriminability- יש כאן רמות רבות משום שיכולים להיות מגוון גווני צבע שונים בטווח המתואר.
3) Separability - הchannels תלויים אחד בשני משום שאם נשנה את הסדר של הattributes בצירים אז הריבועים של ההצלבות שלו יזוזו בהתאמה – וכן הצבעים המקושרים לכל ריבוע. מאחר וכל ריבוע מקושר לצבע נוכל להגיד כי קיים קשר ביניהם.
4) Pop up- אפשר להבחין בקלות בין הריבועים השונים בגלל הצבעים וגבולות הריבועים.

ענווה אברהם 208886333;
רעות בן חמו 209299478;

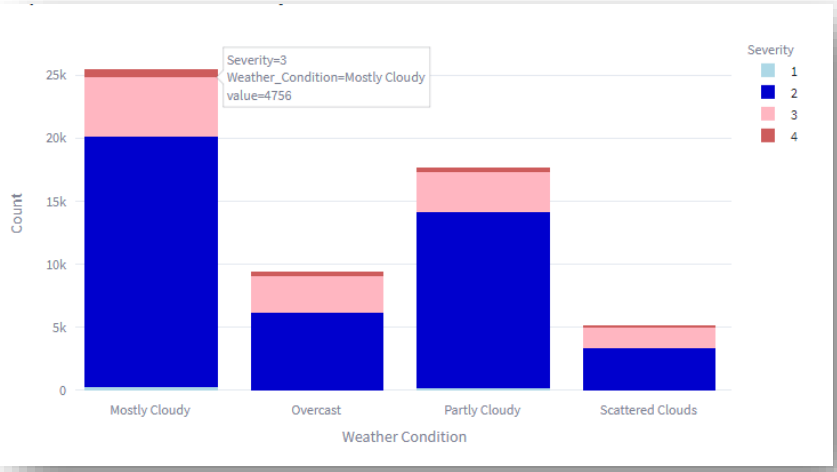
iii. **Bar plot –** How do different weather conditions affect accident severity?

מטרת התרשים היא להבין כיצד מצבי מזג האוויר השונים משפיעים על חומרת התאונה.
התרשים מחולק לפי קטגוריות של מזג האוויר (ציר ה-X) כאשר בציר ה-Y יש את כמות התאונות שקרו עבור אותו מזג האוויר. בנוסף לכך בתוך כל bar יש חלוקה לצבעים לפי כמות התאונות עבור כל חומרת תאונה.
ניתן להסיק מסקנות שונות מהתרשים ובין היתר להבין לאיזה תנאי מזג האוויר יש השפעה הכי גדולה על חומרת התאונה. ניתן לחבר גרף זה עם גרף הbox plot למשל שהראינו מקודם כאשר בחרנו ערך כמו טמפרטורה שיכול לתמוך בחלק מהמסקנות שניתן להסיק מגרף זה. ואף ניתן להרחיב לattributes נוספים שמעניינים את המשתמש.
למשל, ניתן לראות שרוב התאונות קרו במזג אוויר רגיל, אך ישנה כמות גבוהה של תאונות שקרו במזג אוויר מעונן.

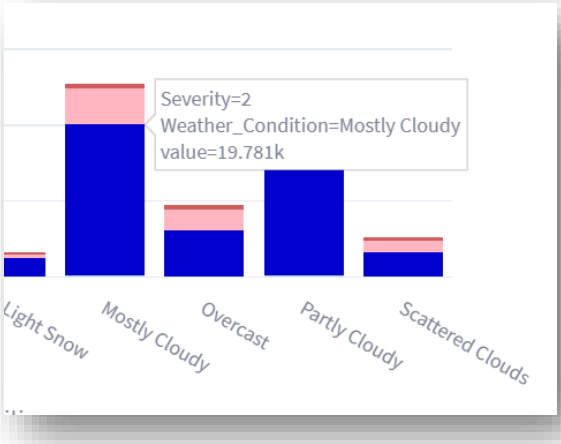


ענווה אברהם 208886333;
רעות בן חמו 209299478;

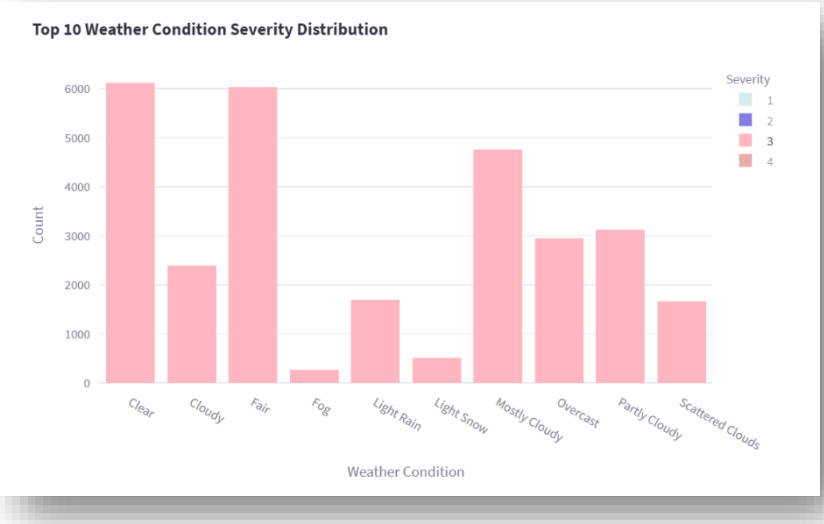
אם נרצה למשל להתמקד רק בחלק ממצבי מזג האוויר נוכל לסמנם ולקבל למשל:



בנוסף, נוכל לקבל מידע מהhover כאשר נעמוד עם העכבר על הערכים:



נוכל גם להסיר את הבחירה בחלק מהדרגות חומרה ולהסתכל רק על דרגת חומרה ספציפית. למשל נסתכל על 3:



ענווה אברהם 208886333;
רעות בן חמו 209299478;

Marks & Channels

Marks: קווים.

Channels: צבע (Hue) המבטא את הקטגוריות השונות בחומרת התאונה.
גובה הbar מבטא את כמות התאונות הכוללת עבור אותו מזג אוויר.

Effectiveness and Expressiveness

אקספרסיביות: השתמשנו בצבע (Hue) על מנת להבדיל בין חומרות התאונה השונות, כל חומרה קיבלה צבע משלה וניתן להבדיל בין הצבעים בקלות. בנוסף הצבעים מסודרים בהתאם לדרגות החומרה מכחול בהיר לאדום כהה כאשר כחול בהיר מסמן דרגה נמוכה ואדום כהה מסמן דרגה גבוהה.
השתמשנו במיקום על ציר ה-Y אשר מבטא את הערך המספרי (סכום) של כמות התאונות עבור הקטגוריות שנמצאות בציר ה-X כאשר נק' הייחוס היא 0.
משום כך, הגרף עונה על הקריטריונים של אקספרסיביות.

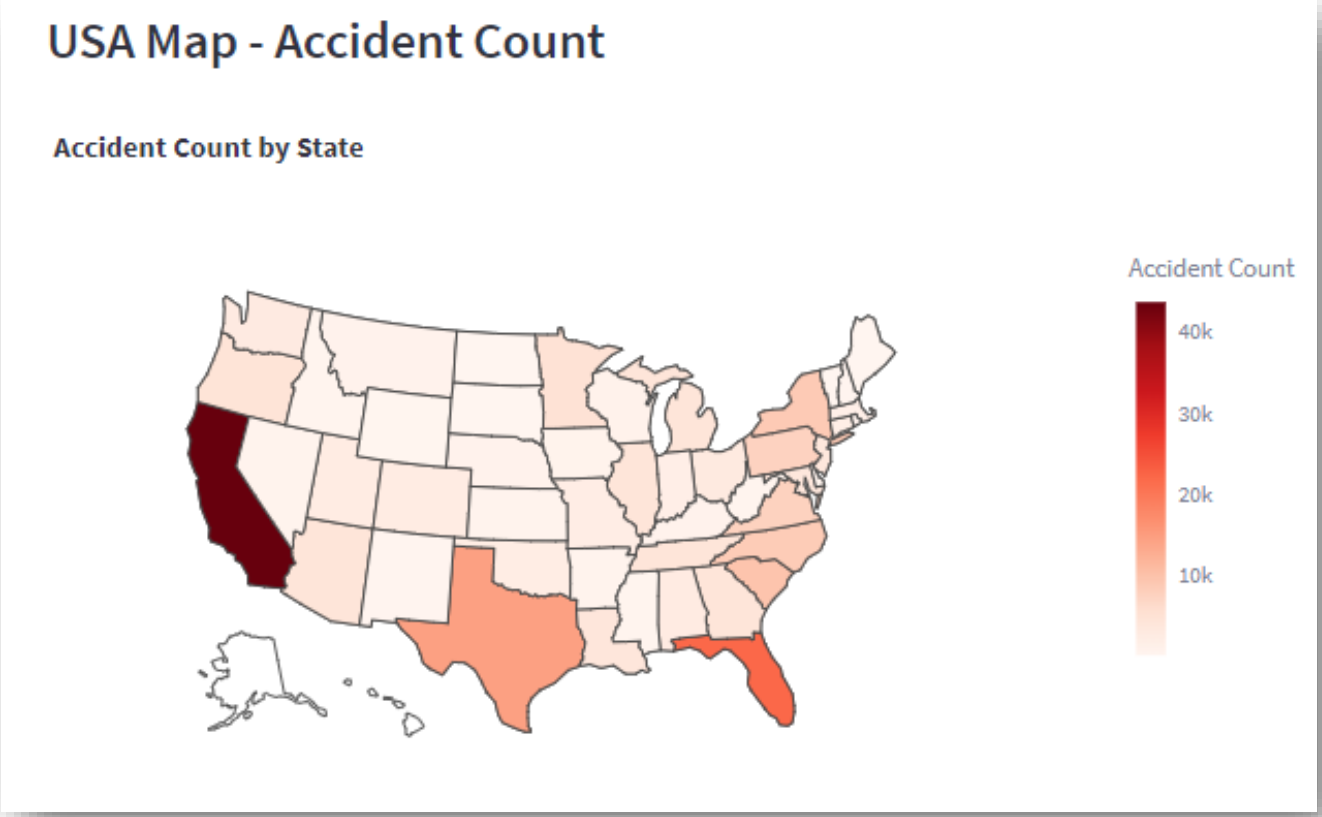
אפקטיביות: נבחן לפי 4 מדדים :

1) Accuracy – יש כאן מספר הבחנות שניתן לעשות. ראשית, ניתן להבחין בקלות בצבעים השונים המתאים את דרגות החומרה השונות של התאונות. הצבעים מנוגדים ומאפשרים לנו להבדיל ביניהם.
בנוסף, ניתן להבדיל בין הbars השונים משום שיש ביניהם מספיק מרווחים המאפשרים הבדלה ברורה.
כמו כן, לפי הערכים בציר ה-Y וגובה הbar אפשר להבחין בהבדלי הגובה בין הbars השונים.
2) Discriminability- נסתכל בשני אופנים: עבור הערכים בציר ה-Y זה תלוי לפי המצב בו נמצאים (כלומר אם בתרשים הראשי או כחלק מאינטראקטיביות) ולכן נוכל להגיד שיש שם רמות רבות.
עבור הצבעים ישנן 4 רמות – אחת לכל דרגת חומרה שונה והצבע שלה.
3) Separability- הchannels תלויים אחד בשני משום שאם נסיר למשל חלק מדרגות החומרה (כלומר חלק מהצבעים) אז הערכים בציר ה-Y ישתנו.
4) Pop up- אפשר להבחין בקלות בין הbars השונים בגלל הרווחים ביניהם, וכן ניתן להבחין בקלות בין הצבעים השונים בכל bar משום שהם מספיק מנוגדים.
עם זאת, מאחר ויש דרגות חומרה להן כמות התאונות קטנה (דרגה 1) כאשר מסתכלים על גרף המלא קיים קושי לזהותם, לכן על מנת להסתכל עליהם באופן ברור מומלץ להסיר את שאר הדרגות מהוויזואליזציה. האינטראקטיביות מסייעת לשמירה על האפקטיביות.
בנוסף, על ידי הסתכלות על גובה הbar אפשר להבחין בקלות בהבדלים בין הbars השונים ולהסיק מסקנות בסיסיות יחסית מהר.

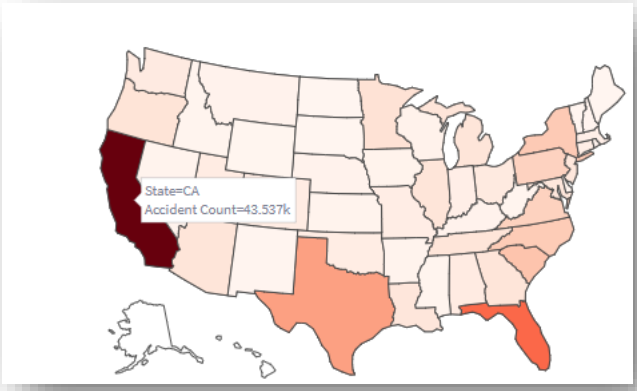
ענווה אברהם 208886333;
רעות בן חמו 209299478;

iv. **Map \ geographic plot –** do specific states (in the US) have more accidents?

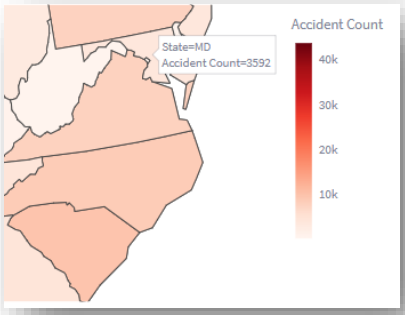
מטרת התרשים היא להבין באילו מדינות בארה"ב יש את כמות התאונות הגדולה ביותר.
התרשים הוא תרשים מפה – כאשר המפה כולה היא מפת ארה"ב והיא מחולקת לפי המדינות השונות ביבשת. עבור כל מדינה אספנו את כמות התאונות שקרו בשטחה ורצינו להציג זאת וויזואלית כדי להבחין האם קיים הבדל מהותי בין המדינות השונות.



אם נעמוד עם העכבר על שטח מדינה מסוימת נוכל לקבל מידע ספציפי עליה (עם hover) :



ענווה אברהם 208886333;
רעות בן חמו 209299478;



בנוסף אפשר עם Zoom-in להתקרב למדינות קטנות יותר ולראות את הגבולות בצורה יותר נוחה:

Marks & Channels

Marks: מיקום (Area).

Channels: צבע (סטורציה) אשר מבטא את כמות התאונות בסדר עולה – כלומר מבהיר לכהה.

מיקום גיאוגרפי של המדינות השונות אשר מבטא את מיקומן האמיתי ביבשת.

Effectiveness and Expressiveness

אקספרסיביות: הצבע (סטורציה) מבטא את כמות התאונות באותה מדינה, כאשר טווח הצבעים נע בין בהיר לכהה כך שמדינה עם מספר תאונות מועט צבועה בצבע בהיר ומדינה עם מספר תאונות רב צבועה בצבע כהה.
המיקום הגיאוגרפי במפה מבטא את המיקום האמיתי של המדינה ביחס ליבשת.
לפי הצבע של כל מדינה במפה נוכל לדעת כמה תאונות יש בכל מדינה וכן לראות מיהן המדינות בעלות כמות תאונות גדולה.
לדוגמה, קליפורניה (CA) צבועה בצבע אדום כהה במפה, דבר המרמז על הרבה תאונות ביחס לשאר המדינות.

אפקטיביות: נבחן לפי 4 מדדים :

1) Accuracy – ניתן להבחין בקלות בין המדינות השונות בגלל אופן נראות המפה – גבולות המדינות בצבע שחור וכן כל מדינה צבועה בצבע. באם יש מדינות סמוכות בעלות צבע דומה, המשתמש יוכל להשתמש בhover על מנת לקבל מידע נוסף.
2) Discriminability- יש כאן רמות רבות משום שיכולים להיות מגוון גווני צבע שונים בטווח המתואר, זאת משום שמדובר בכמות תאונות וזה משתנה רציף בעל כמות מרובה של ערכים.
3) Separability- אין תלות בין הchannels השונים.
4) Pop up- ניתן להבחין בקלות במדינות בעלות כמות גדולה מאוד של תאונות (צבע אדום כהה) לבין מדינות עם כמות קטנה מאוד של תאונות (קרוב ללבן). נוכל להסתכל גם לפי אזורים ביבשת ולהבחין האם ישנן קבוצות מדינות סמוכות להן יש כמות תאונות דומה – אולי זה מבטא משהו.

ענווה אברהם 208886333;
רעות בן חמו 209299478;

סיכום

ניתן להגיד כי כלל הוויזואליזציות שלנו עונות על שאלות המחקר שהצגנו בחלק 1.

עם זאת, סט הנתונים שלנו רחב ובעל attributes רבים וניתן להמשיך לחקור אותו, לבצע וויזואליזציות נוספות או אף לשפר את הוויזואליזציות שהוצגו בפרויקט זה.
כמו כן, נציין שנית כי הוויזואליזציות שלנו בנויות על חלק מתוך סט הנתונים ועל מנת לקבל תמונת מצב אמיתית יש לבצע את הוויזואליזציות על סט הנתונים המלא. בנוסף, בעזרת סט הנתונים המלא יהיה ניתן לבחון שינויים לאורך זמן.

כפי שציינו בחלק 1, סט הנתונים התעדכן מאז תחילת עבודתנו על הפרויקט, וככל הנראה גם ימשיך להתעדכן, ולכן יהיה ניתן לדייק את הוויזואליזציות לפי הצורך.

במסגרת הזמן שניתנה לחלק זה בפרויקט מיקדנו את שאלות המחקר בצורה ספציפית ובהתאם לכך הצגנו וויזואליזציות מתאימות.

ענווה אברהם 208886333;
רעות בן חמו 209299478;

נספחים

1. טבלת attributes

Attribute name	Data type	Attribute description
ID	Categorical - Nominal	This is a unique identifier of the accident record.
Severity	Categorical - Ordinal	Shows the severity of the accident, a number between 1 and 4
City	Categorical - Nominal	Shows the city in address field.
Country	Categorical - Nominal	Shows the country in address field.
State	Categorical - Nominal	Shows the state in address field.
Temperature(F)	Quantitative	Shows the temperature (in Fahrenheit).
Wind_Chill(F)	Quantitative	Shows the wind chill (in Fahrenheit).
Humidity(%)	Quantitative	Shows the humidity (in percentage).
Pressure(in)	Quantitative	Shows the air pressure (in inches).
Visibility(mi)	Quantitative	Shows visibility (in miles).
Wind_Direction	Categorical - Nominal	Shows wind direction.
Wind_Speed(mph)	Quantitative	Shows wind speed (in miles per hour).
Precipitation(in)	Quantitative	Shows precipitation amount in inches, if there is any.
Weather_Condition	Categorical - Nominal	Shows the weather condition (rain, snow, thunderstorm, fog, etc.)
Amenity	Categorical - Nominal	A POI annotation which indicates presence of amenity in a nearby location.
Bump	Binary	A POI annotation which indicates presence of speed bump or hump in a nearby location.
Crossing	Binary	A POI annotation which indicates presence of crossing in a nearby location.
Give_Way	Binary	A POI annotation which indicates presence of give_way in a nearby location.
Junction	Binary	A POI annotation which indicates presence of junction in a nearby location.

No_Exit	Binary	A POI annotation which indicates presence of no_exit in a nearby location.
Railway	Binary	A POI annotation which indicates presence of railway in a nearby location.\
Roundabout	Binary	A POI annotation which indicates presence of roundabout in a nearby location.
Station	Binary	A POI annotation which indicates presence of station in a nearby location.
Stop	Binary	A POI annotation which indicates presence of stop in a nearby location.
Traffic_Calming	Binary	A POI annotation which indicates presence of traffic_calming in a nearby location.
Traffic_Signal	Binary	A POI annotation which indicates presence of traffic_signal in a nearby location.
Turning_Loop	Binary	A POI annotation which indicates presence of turning_loop in a nearby location.
Sunrise_Sunset	Categorical - Nominal	Shows the period of day (i.e., day or night) based on sunrise/sunset.
Civil_Twilight	Categorical - Nominal	Shows the period of day (i.e., day or night) based on civil twilight.
Nautical_Twilight	Categorical - Nominal	Shows the period of day (i.e., day or night) based on nautical twilight.
Astronomical_Twilight	Categorical - Nominal	Shows the period of day (i.e., day or night) based on astronomical twilight.

ענווה אברהם 208886333;
רעות בן חמו 209299478;

2. טבלאות מלאות עבור target , action :

a. באיזו מידה גורמים משניים למזג אוויר כמו טמפרטורה משפיעים על קיום תאונה?

Target	Action	
Attributes: Distribution המשתמש ירצה להבין כיצד הattribute משפיע על חומרת התאונה.	המשתמש יחקור אילו attributes משפיעים על חומרת תאונה גבוהה\נמוכה	Analyze: Discover
	המשתמש יודע על אילו attributes מדובר אבל לא יודע בדיוק מי הם ואת מי לחפש	Search: Lookup
	המשתמש יזהה כיצד הattributes משפיעים על חומרת התאונה	Query: Identify

b. האם תמרורים, רמזורים, עיקולים, שטחי הפרדה וכולי.. הם גורמים שמעלים את הסיכוי או מורידים את הסיכוי לתאונה במקום מסוים? האם ישנה קורלציה מסוימת בין הגורמים הללו ?

Target	Action	
Attributes: Correlation. המשתמש ירצה להבין מה הקורלציות השונות בין הattributes.	המשתמש ירצה לחקור ולהסיק מסקנות מתוך הקשרים בין הattributes ולהציג אותם	Analyze: Consume – Discover & Present
	המשתמש לא יודע אילו קורלציות לחפש ואיפה להסתכל	Search: Explore
	המשתמש יסתכל על כלל הקשרים ויבין מי הם הקשרים החזקים ומי החלשים	Query: Compare & summarize

c. האם שינויי מזג אוויר כאלו או אחרים משפיעים באופן ישיר על סיכוי גבוה יותר לתאונות, גם כאשר הדבר איננו ברור ממבט ראשוני? – ניתוח מצבי מזג אוויר שונים לפי חומרת תאונה, וזיהוי מצבים בעייתיים

Target	Action	
Attributes: Similarity המשתמש ירצה לראות מי אלו מצבי מזג האוויר שמשפיעים הכי הרבה על תאונות בדרגת חומרה גבוהה.	המשתמש יחקור על מצבי מזג האוויר השונים ואת אופן השפעתם על דרגות חומרת התאונה	Analyze: Consume - Discover
	המשתמש יזהה מצבי מזג אוויר בעייתיים אך מאחר והוא לא יודע בדיוק איפה הם נמצאים הוא צריך לחפש אותם	Search: Locate
	המשתמש ישווה בין התוצאות שיקבל וינסה להבין מי אלו הבעייתיים ביותר- הם אלו שמעניינים אותנו.	Query: Compare

ענווה אברהם 208886333;
רעות בן חמו 209299478;

d. באילו מדינות (בארה"ב) יש הכי הרבה תאונות?

Target	Action	
All Data: Features ההסתכלות כאן תלוית מטרה. האם המשתמש מחפש את המדינה בעלת הכי הרבה תאונות? האם המשתמש מחפש את כל המדינות מעל סף מסוים של תאונות? המטרה היא לקבל סיכום כללי עבור כל מדינה על מנת לבצע את הניתוחים ולהסיק מסקנות.	המשתמש ירצה להבין מי הן המדינות הבעייתיות, כלומר באילו מדינות יש הרבה תאונות	Analyze : Present
	המשתמש לא יודע באיזה אזור ביבשת לחפש ואיזו מדינה לחפש	Search: Explore
	כלל הנתונים נסכמים לכל מדינה וההסתכלות היא כללית.	Query: Summarize

Tasks	Target	Action
זיהוי attributes הקשורים למז"א וחקירת השפעתם על תאונה	Attributes: Distribution	Analyze: Discover
		Search: Lookup
		Query: Identify
ניתוח הקשרים בין הattributes השונים	Attributes: Correlation	Analyze: Consume – Discover & Present
		Search: explore
		Query: Compare & Summarize
חקירת השפעתם של מצבי מזג אוויר שונים על חומרת התאונה	Attributes: Similarity	Analyze: Consume - Discover
		Search: Locate
		Query: Compare
הבנה באילו מדינות (בארה"ב) יש הכי הרבה תאונות	All Data: Features	Analyze: present
		Search: explore
		Query: Summarize