**פרויקט סיום קורס בוויזואליזציה**

**בנושא הריגות אזרחים על ידי שוטרים בארה"ב בין השנים 2015 ל2022**

מגישים: איתי ויינר ושלו אביב

ת"ז: 323014068, 323128298

שם הקורס: ויזואליזציה של מידע  
תאריך הגשה: 27.6.22

**מבוא:**

שאלת המחקר שלנו היא: איך כמות המתים על ידי יריה משוטר השתנתה לאורך השנים האחרונות.

שאלות המשנה שלנו:

1. איך גזע משפיע על כמות המתים על ידי יריה משוטר לאורך זמן.
2. מה הם ההבדלים בכמות המתים על ידי יריה משוטר במדינות השונות לאורך זמן.

המאגר מכיל נתונים על מקרים המתעדים הריגות אזרחים בארה"ב בין השנים 2015 עד 2022 על ידי יריה או יריה וטייזר. הנתונים לקוחים מrepository בגיטהאב של Washington Post שפורסם במאמר שלהם בנושא, כמו שאפשר לראות בקישורים הבאים.

<https://www.washingtonpost.com/graphics/investigations/police-shootings-database/>

<https://github.com/washingtonpost/data-police-shootings>

היקף הנתונים לפני Preprocessing היה טבלה של 7504 רשומות ו17 פיצ'רים.  
לאחר הpreprocessing, נשארנו עם טבלה של 6969 רשומות ו23 פיצ'רים (הוספנו עמודות נוספות לצורך נוחות).

**נתונים:**

**הדאטה סט מכיל רשומות עבור כל נהרג על ידי שוטר במילוי תפקיד בארה"ב בין התאריכים ה2 לינואר 2015 ועד לתאריך ה19 ליוני 2022 (הדאטה ממשיך להתעדכן בגיט, אך הדאטה שבו אנו משתמשים הוא דאטה שהורדנו מראש וכבר לאחר preprocessing ולכן הוא מכיל רק עד ה19 ביוני 2022). בין היתר מפורט שם הנפטר, המיקום בו נורה על ידי שוטר, מין הנפטר, גיל הנפטר ועוד.**

**הדאטהסט שלנו הוא מסוג Flat Table.**

|  |  |  |
| --- | --- | --- |
| **Feature** | **Data Type** | **Description** |
| **id** | **Categorical** | **Unique identifier for victim** |
| **name** | **Categorical** | **Victim’s name** |
| **date** | **Quantitative** | **Date of shooting in YYYY-MM-DD format** |
| **manner\_of\_death** | **Categorical** | **Manner of victim’s death** |
| **armed** | **Categorical** | **Describes whether the victim was armed (and if so with what), unarmed or unknown whether he had a weapon or not** |
| **age** | **Ordinal** | **Victim’s age when died** |
| **gender** | **Categorical** | **Victim’s gender** |
| **race** | **Categorical** | **Victim’s race in one letter representation or unknown** |
| **city** | **Categorical** | **The city, the victim got shot in** |
| **state** | **Categorical** | **The state, the victim got shot in** |
| **signs\_of\_mental\_illness** | **Categorical** | **Boolean, describes whether the victim has a history of mental illness** |
| **threat\_level** | **Categorical** | **Describes whether the victim was a threat to the officers in duty** |
| **flee** | **Categorical** | **Describes whether the victim fled the scene and if so whether it was by foot or by car** |
| **body\_camera** | **Categorical** | **Boolean, describes whether the on-scene officer had a body camera on him** |
| **longitude** | **Quantitative** | **Used with latitude to express location as , express location of death** |
| **latitude** | **Quantitative** | **Used with latitude to express location as , express location of death** |
| **is\_geocoding\_exact** | **Categorical** | **Boolean, describes whether the location expressed by (lat, lon) is exact or expresses the center of the city or the county** |
| **race\_full** | **Categorical** | **Derived Column. Created from `race`, describes the race in full name** |
| **state\_full** | **Categorical** | **Derived Column. Created from `state`, describes the state in full name** |
| **year** | **Ordinal** | **Derived Column. The year component of `date`.** |
| **month** | **Ordinal** | **Derived Column. The month component of `date`.** |
| **day** | **Ordinal** | **Derived Column. The day component of `date`.** |
| **gender\_binary** | **Categorical** | **Derived Column. Binary. Gender in binary format, M=1, F=0.** |

**מטלות משתמש:**

**הוויזואליזציה עוסקת במקרים אשר התרחשו בארצות הברית שבהם המשטרה הרגה בירייה אנשים כחלק מניסיון מעצר.  
הוויזואליזציה עונה על שאלות מתחום ביקורת על המשטרה וניתוח האגרסיביות שלה.**

**השאלות עליהן ניתן לענות הן:**

|  |  |  |
| --- | --- | --- |
| **השאלה** | **Action** | **Target** |
| **האם רמת האגרסיביות של השוטרים בארצות הברית קטנה עם הזמן?** | **Compare** | **כמות האנשים שמתו כל שנה.** |
| **באיזה מדינות המשטרה מתנהגת באופן אלים במיוחד?** | **Locate** | **המדינות עם כמות המתים הגבוהות ביותר.** |
| **האם אפשר לראות שינוי בהתנהגות המשטרה עבור כל מין?** | **Compare** | **ההבדל בין כמות האנשים מאותו המין שמתו בטווח תאריכים כלשהו (כל שנה/חודש וכדומה)** |
| **עבור איזה גזעים היו יותר מקרים?** | **Locate** | **הגזעים עם אחוז המתים הגבוהה ביותר.** |
| **האם יש הבדל בין המדינות עבור גזע ספציפי?** | **Discover** | **המדינות שבהן מתו אנשים מכל גזע.** |
| **האם רמת האגרסיביות של השוטרים ירדה עם הזמן עבור גזעים שונים?** | **Compare** | **כמות האנשים שמתו מכל גזע כל שנה.** |
| **באיזה מדינות מתו יותר מ100 אנשים?** | **Browse** | **המדינות שבהן מתו יותר מ-100 אנשים.** |

**רעיונות ועיצובים אלטרנטיביים:**

1. אלטרנטיבות:

**עבור וויזואליזצית ה-bar**s **(1):**

1. **הצגת הנתונים הרלוונטיים עבור כל מין בו זמנית בפרופורציות שוות בסגנון של 2 בארים אחד ליד השני  
   marks-** lines**.  
   channels-** **צירים אנכיים ואופקיים, אורך וצבע לכל מין.**
2. **במקום בר להציג בר תלת ממדי שבעומק שלו אפשר לראות את השינוי עבור כל גזע.  
   marks-** lines**.  
   channels- צירים אנכיים ואופקיים, צבע לכל גזע, אורך ונפח.**

**עבור וויזואליזצית ה-pie (2):**

1. **הצגת גרף Sunburst במקום הפאי כך שבאמצע יהיו לנו מדינות, ושבשכבה השנייה לכל מדינה יהיה מפורט האחוזים מכל גזע שמתו שם  
   marks-** areas**.  
   channels- צבע ושטח.**
2. **הצגת גרף bar אשר מראה את כמות המתים לכל גזע.  
   marks-**lines**.  
   channels- צירים אנכיים ואופקיים, ואורך.**

**עבור וויזואליזצית ה-map (3):**

1. **הצגת heat-map במקום המפה אשר מראה חלקים יותר כהים עבור מדינות עם יותר מקרים.  
   marks-** areas**.  
   channels- שטח וצבע.**
2. **הצגת גרף bar אשר מראה את כמות המקרים לכל מדינה.  
   marks-**lines**.  
   channels- צירים אנכיים ואופקיים, ואורך.**
3. יתרונות חסרונות:
   1. Chart, bar chart

      Description automatically generated **יתרונות: קל למימוש, מראה השוואה מיידית קבועה.  
      חסרונות: קיים הבדל גדול בפרופורציות המקרים לכל מין (סך הכל 6645 מקרים לזכר ו-324 מקרים לנקבה) לכן הצגתם אחד ליד השני תהיה מאוד לא אינפורמטיבית.**
   2. Chart, bar chart

      Description automatically generated **יתרונות: מציג את כל הנתונים בו זמנית באופן יחסית מובן.  
      חסרונות: קשה למימוש, יכול להיות שיהיו בארים שיסתירו אחרים.**

**למה בחרנו את הפתרון שלנו:**

**בחרנו את הפתרון שלנו כיוון שהוא פשוט למימוש (לעומת 1.2 שהוא קשה למימוש) הוא פשוט להבנה. יש לו את האפשרות להציג את המינים בנפרד, מה שעונה על השאלה שרוצה לבדוק הבדל עבור כל מין בנפרד.  
גרף זה נחשב למאוד יעיל, הוא נמצא על צירים (Magnitude) ובעל פרופורציות שוות (Magnitude) מה שגורם לו להיות פשוט להבנה ואפקטיבי.**

1. Chart, sunburst chart

   Description automatically generated **יתרונות: נותן דרך נוחה להצגת התפלגות לכל ערך.  
   חסרונות: לא מתאים לנתונים בעלי הרבה קטגוריות וקטגוריות משניות, במקרה שלנו 51 מדינות ולכל אחת 7 קטגוריות לגזעים, יוצא 357 ערכים לעיגול החיצוני ולכן זה לא יהיה אינפורמטיבי.**
2. Chart

   Description automatically generated **יתרונות: פשוט למימוש, קל להבנה ומציג הבדלים טוב.  
   חסרונות: לא טוב בלהציג ולהשוות יחסים מתוך אוכלוסייה, פחות מתמקד בתמונה הגדולה ויותר מציג כל ערך בנפרד.**

**למה בחרנו את הפתרון שלנו:**

**בחרנו את הפתרון שלנו כיוון שהוא פשוט למימוש, הוא פשוט להבנה, הוא נותן לנו לראות תמונת עולם תוך כדי השוואה בין הקבוצות השונות ויש לנו כמות קטנה של קטגוריות ולכן לכל ערך בו יש מקום בולט וברור.  
גרף זה נחשב למאוד יעיל, הוא מכיל חלקים המתאימים לגודל הערך (Identity) ויש צבע לכל קטגוריה (Identity) מה שעוזר לו להיות ברור וקל והבנה.**

1. Chart, treemap chart

   Description automatically generated **יתרונות: טוב להשוואה בין ערכים, מציאת ערכי קיצון ופשוט למימוש.  
   חסרונות: חוץ ממציאת ערכי הקיצון הוא לא תורם לנו לשאלות המחקר.**
2. Chart

   Description automatically generated **יתרונות: פשוט למימוש, קל להבנה, מאפשר מציאת ערכי מינימום ומקסימום.  
   חסרונות: נותן מידע מצומצם, פחות מתמקד בתמונה הגדולה ויותר מציג כל ערך בנפרד, ממש הצגה ופחות השווה בין ערכים.**

**למה בחרנו את הפתרון שלנו:**

**בחרנו את הפתרון שלנו כיוון שהוא הוא פשוט להבנה, הוא נותן לנו לראות תמונת עולם תוך כדי השוואה בין הקבוצות השונות, הוא נותן לנו גישה להרבה מידע ומציג את זה על בסיס מפה אשר מתחבר למשמעות הנתונים.  
גרף זה הוא מאוד אינפורמטיבי, הוא נותן לנו גם תמונת מצב גדולה על כל המדינה עם צביעה הדרגתית (Magnitude) לפי כמות המקרים בכל מקום וגם נותן גישה למידע יותר מפורט וספציפי על כל מקרה וצובע באדום וכחול כל מין (Identity) בשביל להוסיף עוד פאן של השוואה.**

**הסבר על היישום:**

1. **התחלנו את העיבוד המקדים בבדיקת הערכים החסרים, דרך הReadMe של הrepository בדקנו לאילו ערכים יש משמעות ולאילו אין, לאחר מכן יצרנו מילונים שיעזרו במילוי ערכים יותר אינפורמטיביים כדון מילון הממיר את השם המקוצר של המדינה לשם המלא שלה ושממיר לנו את הגזע מאות אחת למילה המלאה (W->White וכו'). לאחר מכן, ביצענו הורדה ותיקון של ערכי הnull.**

Table

Description automatically generated

**החלטנו שרשומות חסרות name או age לא תורמות לנו ולכן מחקנו את רשומות אלו.**

**לאחר מכן, תיקנו את ערכי הnull של העמודות armed, gender, race, flee בהתאם להגדרות שהיו על הדאטה בReadMe. לאחר מכן מחקנו את הרשומות בעלות gender=Unknown (כי היה כתוב שUnknown=null, היה בהתחלה גם וגם; והחלטנו שgender ריק לא תורם לנו).  
לבסוף, השתמשנו בחבילה geopy.geocoders.Nominatim בשביל להשיג נקודות אורך רוחב במקום הnull, עשינו זאת לפי העיר והמדינה שהוגדרו בתור המקום בו הקורבן נורה.**

**ובנוסף, שינינו את is\_geocoding\_exact שאומר אם הקואורדינטות מדויקות עד מרחק של כ100 מטר מהנקודה האמיתית לfalse כיוון שאנחנו לא יודעים את המיקום המדויק שבו האירוע התרחש.**

**לסיום הpreprocessing יצרנו עמודות עזר מעמודות קיימות שהן: year, month, day שהוצאנו אותן מdate; הוספנו עמודות לשם המלא לכל race ולכל state, וגם עמודה עם ערכים בינאריים למין הקורבן.**

* **לשם הנוחות, את העיבוד המקדים ביצענו בקולאב ומשם שמרנו את קובץ הcsv לאחר העריכות בשביל שימוש עתידי.**

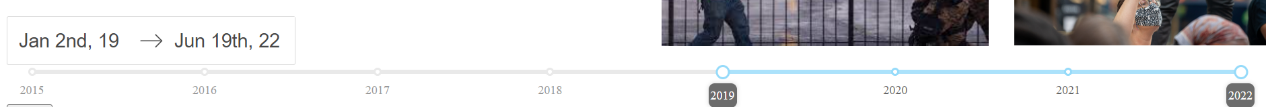
1. **תמונות של הויזואליזציה:**

Graphical user interface, chart, application, Excel

Description automatically generated

**אחד הפילטרים העיקריים שלנו הם הפילטרים של התאריכים והשנים אשר ניתנים לשינוי בעזרת הסליידר לפי שנים או בחירת טווח תאריכים ספציפי.**

**הפילטרים:**

**Graphical user interface, application, table

Description automatically generated**

**הוויזואליזציה:**

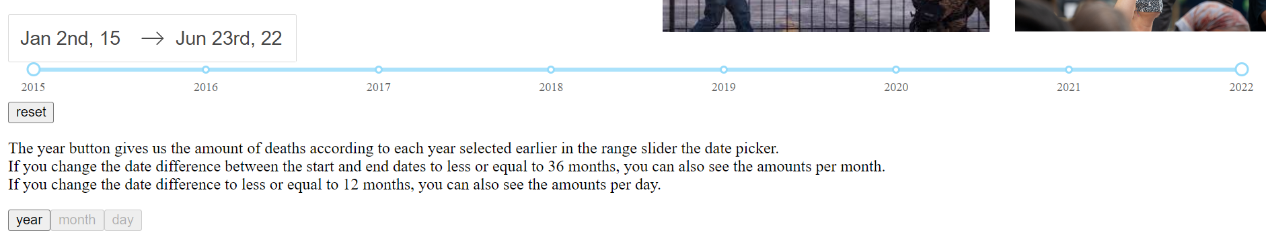
Chart

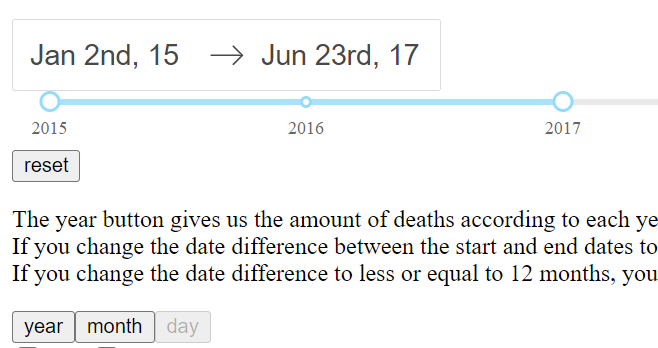
Description automatically generated

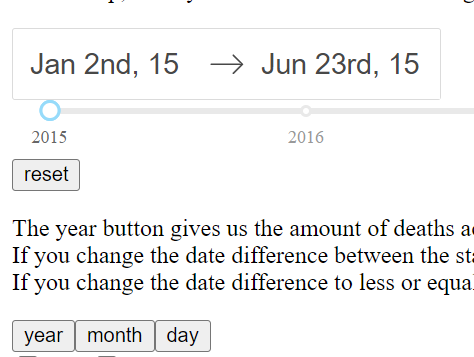
**קיימים לנו כפתורים אשר משנים את ה**-bar **פלוט במצבים שבהם הוא פחות אינפורמטיבי לאחד יותר אינפורמטיבי, כאשר יש נתונים בטווח של פחות מ3 שנים ניתן ללחוץ על כפתור של חודשים אשר הופך את הגרף להיות על בסיס חודשים במקום שנים, וכאשר אנו נמצאים בטווח של פחות משנה ניתן ללחוץ על כפתור של יום אשר הופך את הגרף המוצג להיות על בסיס ימים.**

**הפילטרים:**

**המצב ההתחלטי:**

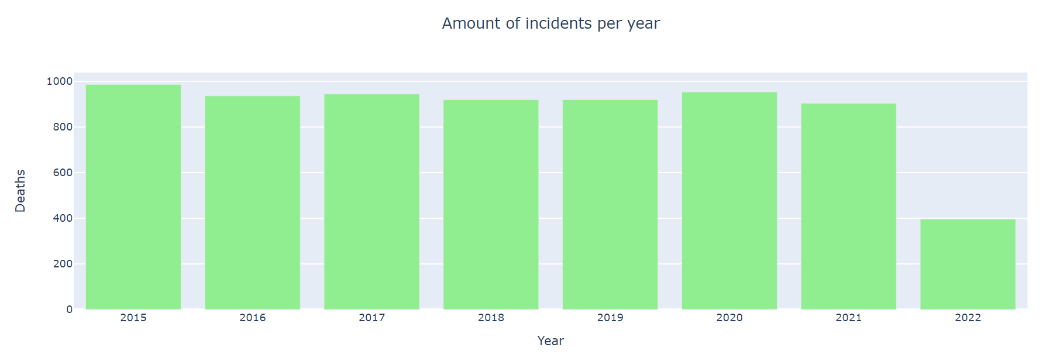


**הגבלה לפי חודשים:  
  
הגבלה לפי ימים:**

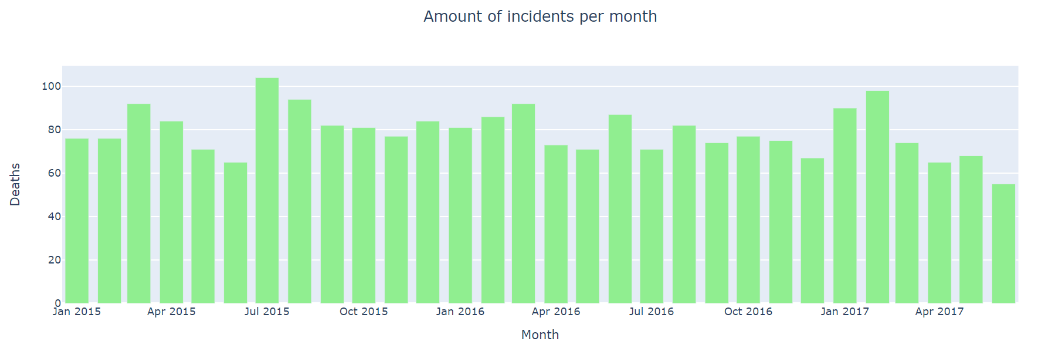
****

**גרף ה-**bars **לפי הכפתורים השונים הנ"ל:**

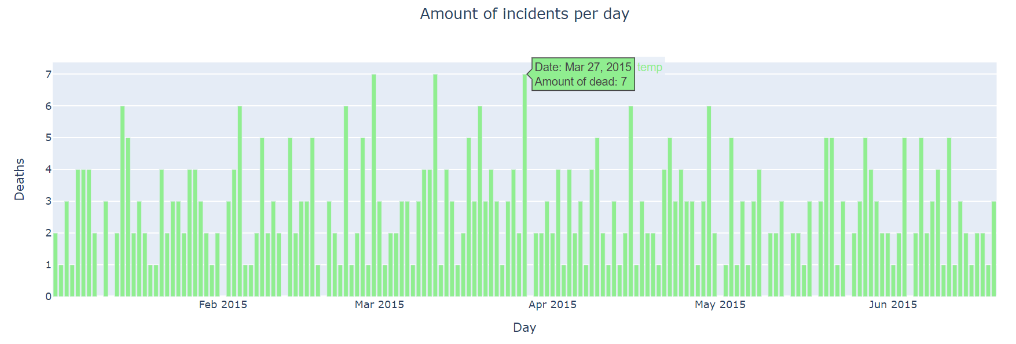
**שנים:**

****

**חודשים:**

****

**ימים:**

****

**כפתורי** check-box **אשר בוחרים את המינים המוצגים.**

**הפילטרים:**

**המצב הדיפולטי:**

****

**בחירת נשים בלבד:**

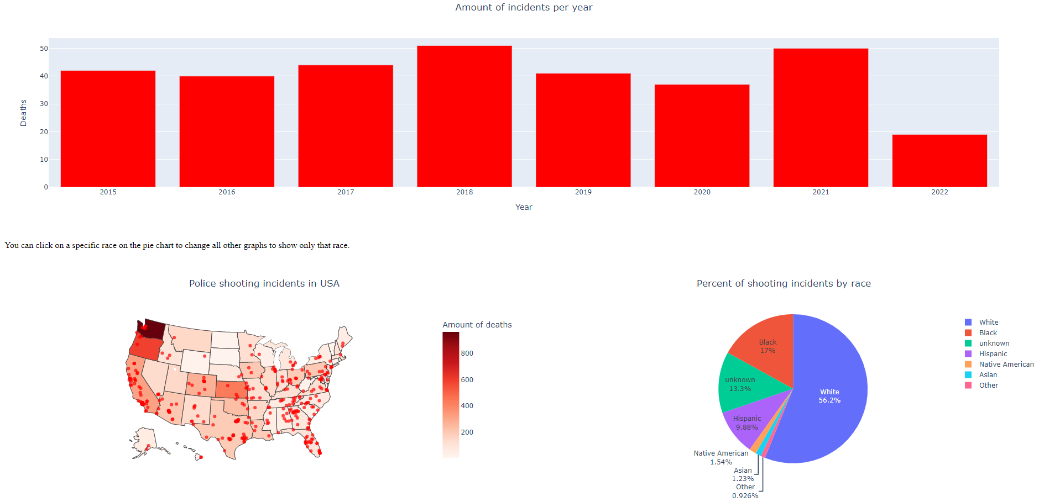
****

**בחירת גברים בלבד:**

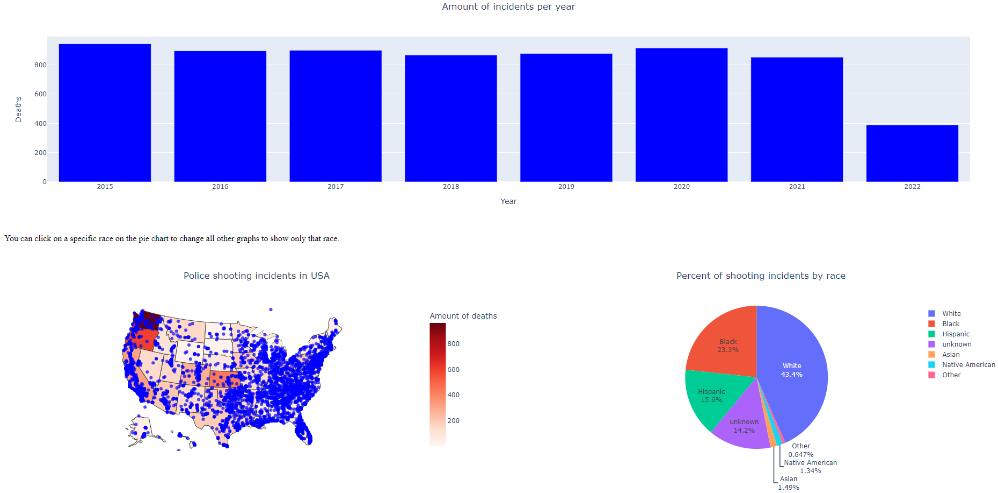
****

**הוויזואליזציות:**

**עבור נשים:**

****

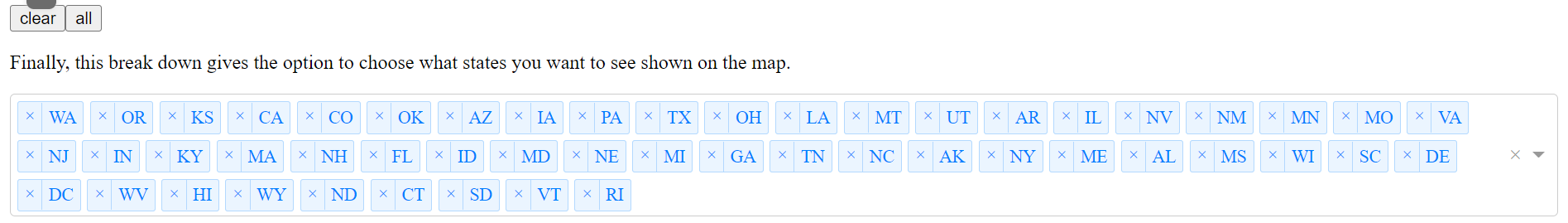
**עבור גברים:**

****

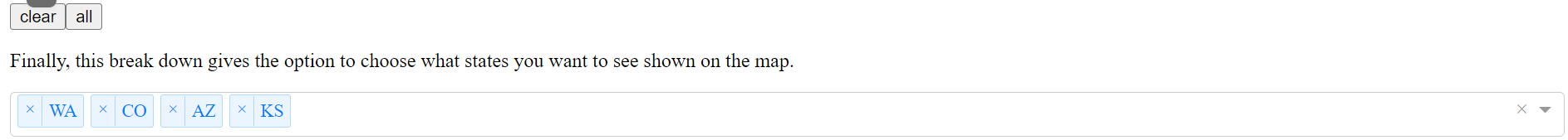
**הפילטר במרכזי הבא הוא הפילטר של המדינות שבו אפשר לבחור איזה מדינות יוצגו בגרף ואיזה לא, בנוסף אליו יש גם 2 כפתורים אשר עוזרים להתעסק איתו והוספו כדי להקל על המשתמש והם clear אשר מנקה את המדינות שנבחרו ו-all אשר מציג את כל המדינות הרלוונטיות.**

**הפילטר:**

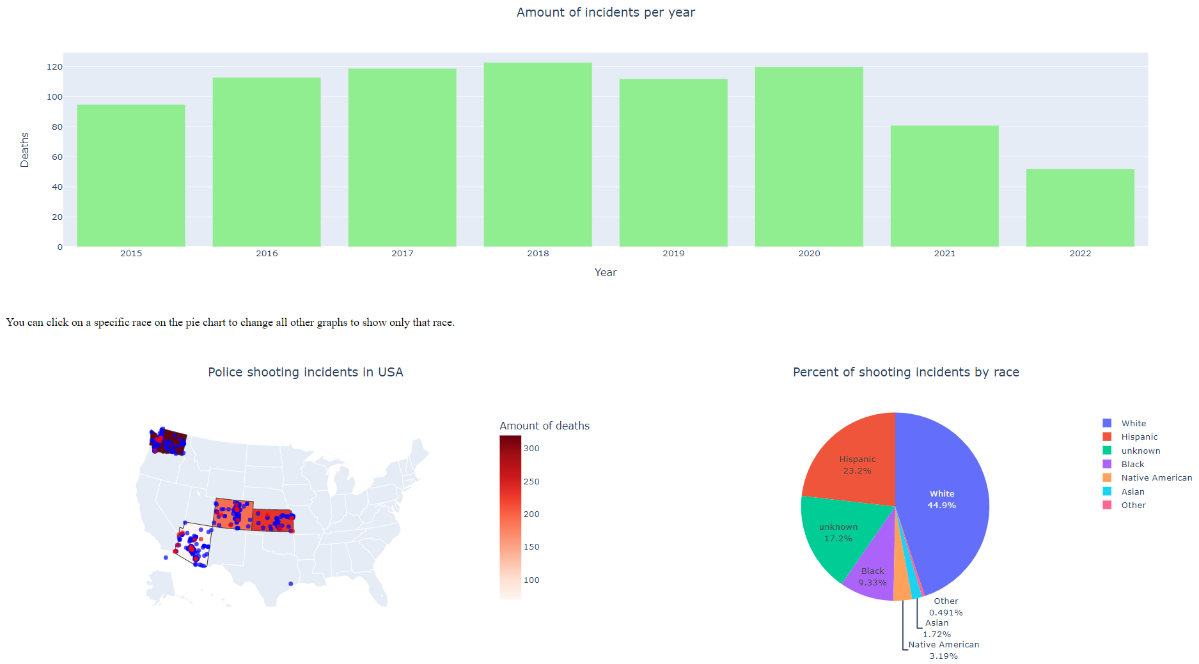
**במצב דיפולטי:**

****

**כאשר בוחרים מדינות:**

****

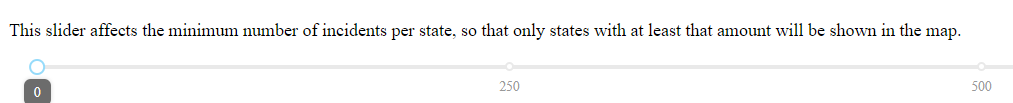
**הוויזואליזציות עבור בחירת מדינות:**

****

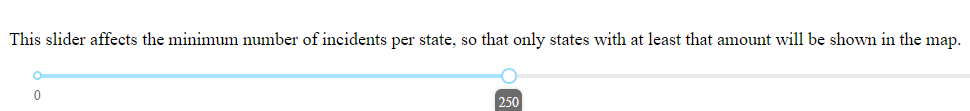
**עוד פילטר אשר משפיע על המדינות הוא slider אשר מסנן מדינות שיש להן פחות מקרי מוות מהערך שבחרו בו.**

**הפילטר:**

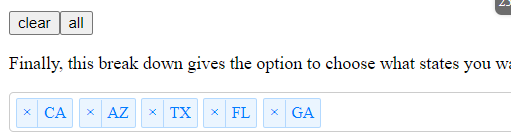
**במצב דיפולטי:**

****

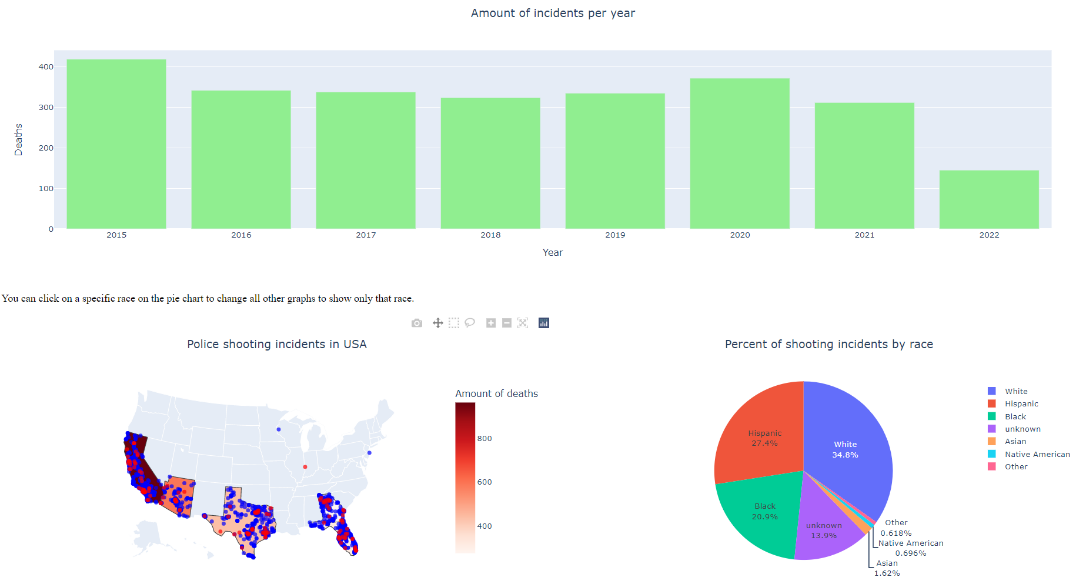
**כאשר בוחרים ערך:**

****

**ההשפעה על הפילטר של המדינות:**

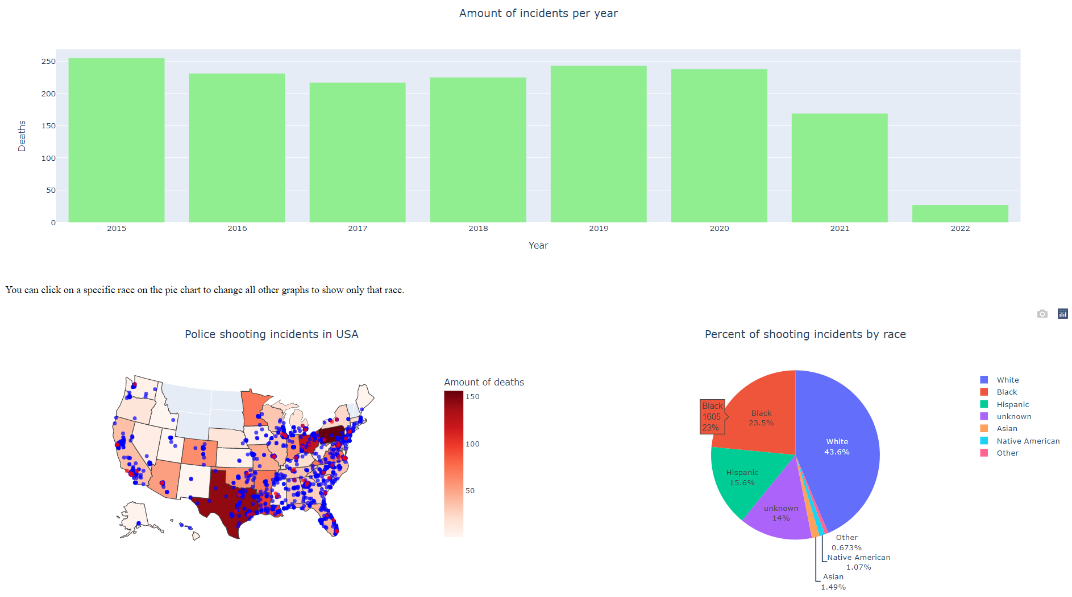
****

**ההשפעה על הוויזואליזציות:**

****

**הפילטר האחרון שקיים לנו הוא לחיצה על גזע בגרף פאי אשר משנה את שאר הגרפים להיות רלוונטיים לאותו בגזע.**

**הפילטר+ השפעה:**

****

**בנוסף לפילטרים קיים לנו טקסט מפורט כאשר שמים את העכבר על כל אחד מהגרפים.**

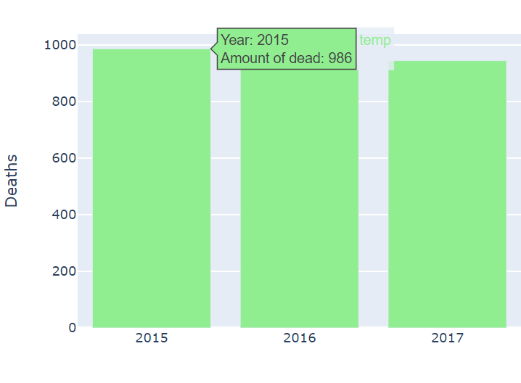
**דוגמאות:**

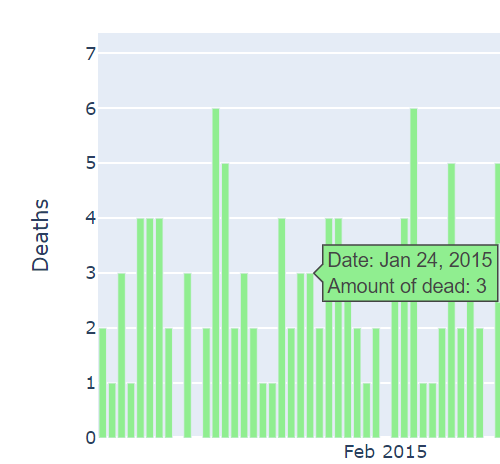
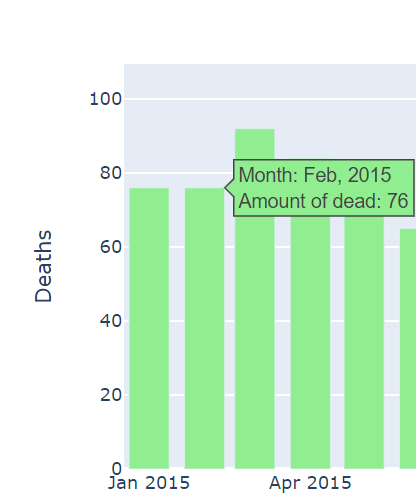
**Chart

Description automatically generatedA picture containing chart

Description automatically generated**

**Chart, pie chart

Description automatically generated**

****

1. **בקוד השתמשנו בחבילות plotly.graph\_objects, dash, pandas, datetime.**

**התחלנו ביצירת העמוד html עצמו, בשביל לסדר את הדף html עצמו ולמקם את כל האיברים הנמצאים בו.  
לאחר מכן, עברנו לפונקציות שעובדות עם callbacks האחראיות על יצירת הגרפים ועדכונם לפי הפילטרים והפיצ'רים השונים בדף. הפונקציה הראשונה והראשית שיש לנו בקוד שמקבלת את הטריגר מהפילטרים המרכזיים שיש לנו שהם: שינוי בתאריך, שינוי במדינות הנבחרות או במינים אשר רוצים להציג, בנוסף בחירת גזע או שינוי בסוג הbar שרוצים שיוצג קורים דרך פונקציה זו. בסופו של דבר אלה הדברים העיקריים שמגדירים לנו את הגרפים וכך אנחנו דואגים שכל הגרפים ישתנו בהתאם.  
כל אחד מהcallbacks האחרים שיש לנו אחראים על שינוי בקטגוריה שונה, הcallbacks השונים הם:**

* **change\_country- callback זה נקרא בעזרת פילטרים שאמורים לשנות את המדינות המוצגות כגון כפתורי הclear וall שנמצאים מעל הbreakdown, הסליידר שמגביל מדינות לפי מקומות שמתו בהם יותר מהערך הנבחר, או בחירת race מגרף הפאי כדי להציג רק מדינות שרלוונטיות לגזע הנוכחי. בקולבק זה אנחנו מורידים מהאופציות המקסימליות מדינות שלא עומדות בתנאי ומעדכנים את האופציות שנבחרו בהתאם.**
* **change\_date- בcallback זה, אנחנו מתייחסים לשינויים בrange slider של השנים, ומעדכנים את התאריכים האחרים בהתאם.**
* **reset\_- כפתור הreset מאפס לנו את הסינונים חזרה למקור, כפתור זה קורא לכל פונקציות הcallback ללא main (שchange\_date יקרא לו), ובכך מאפס את הערכים חזרה לדיפולטיים. בcallback זה אנו מאפסים את כל הפילטרים אשר לא משתנים דרך הcallbacks האחרים שהם gender והvalue של הסליידר המציין את כמות ההריגות.**

פעולות מיוחדות של הmain- **כדי לגרום לכך שההבדלים ישפיעו על כל הגרפים. בתחילת הפונקציה יצרנו לנו דאטה זמני אשר מתאים לפילטרים השונים. בנוסף, בשביל פעולות הפילטרים הקטגורייאליות (כפתורי הyear, month, day ובחירת המינים), יצרנו גרפים קצת שונים עבור כל תנאי ופעולה זו קוראת אוטומטית בקוד.**