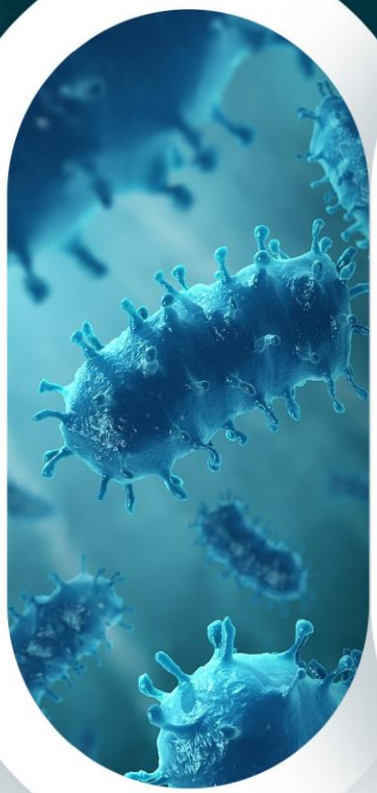
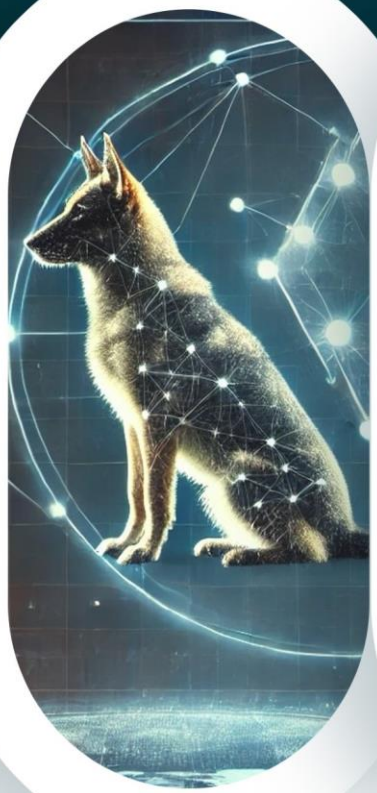
The Academic College of Yezreel Valley
Department of Information Systems

המכללה האקדמית עמק יזרעאל
חוג למערכות מידע

Data preparation report
פרויקט גמר מסמך הכנת הנתונים

מערכת לחיזוי אזורים ומועדים בסיכון לחשיפה
ולהתפרצות מחלת הכלבת והצגת מפת סיכונים.

סטודנטיות מגישות:
אל קוזלי אימאן, ת.ז 212175582
רם דניאל, ת.ז 208220509


קורס פרויקט גמר
שנת לימודים: תשפ"ה
תאריך הגשה: 5.4.2025
שם המרצה: מר זכאי אבי
שם המנחה: גב' גוטפריד ג'ניה

תוכן עניינים

1. בחירת הנתונים

כל הבחירות שביצענו נעשו בקפידה, כאשר כל עמודה שהשארנו נבחרה על סמך המשמעות שלה לתהליך ולמטרה שלנו. אנחנו מאמינים כי הנתונים שנשארו הם אלו שיכולים לתרום בצורה הטובה ביותר להפקת תובנות מדויקת ומשמעותית. לגבי הנתונים שהורדנו, עשינו זאת לאחר חשיבה מושכלת ומדוקדקת, תוך התמקדות במה שיכול להועיל ולשפר את המודל. החלטנו להוריד עמודות שהיו חסרות ערך או שיצרו חפיפות מיותרות, ובכך לייעל את התהליך ולהתמקד במה שבאמת חשוב. ייתכן שבהמשך, בתהליך בחירת האלגוריתם, נחליט לשנות או לעדכן חלק מהעמודות.

נציג את העמודות אחת אחת, נסביר את המשמעות שלהן ואת המטרה שהן משרתות בתהליך. כך נוכל להבהיר את תרומתן למטרות שלנו ולוודא שהן אכן רלוונטיות ומועילות להפקת התובנות הרצויות.

העמודות של הטבלה המקורית:

OBJECTID	Animal_Lab_ID	Event	Date	LinkToTest	OpenLink	LinkToMre	LinkMoreOpen	Year	SpeciesNameEng	RegionEng
RegionHeb	SpeciesNameHeb	Species	SettlementHeb	SettlementEng	LocationNotSettlementEng	LocationNotSettlementHeb	GlobalID	CreationDate	Creator	
EditDate	Editor	x	y							

<u>עמודות שהסרנו:</u>

Animal_Lab_ID- מכילה מזהה פנימי של המעבדה עבור כל חיה, והוא משמש ככלי לניהול רשומות בתוך המעבדה. המזהה הזה בדרך כלל לא מספק מידע חשוב או משמעותי לניתוח נתונים או חיזוי מגמות בתהליך שלנו, ולכן החלטנו להסיר אותו. הוא לא תורם ישירות למטרות כריית הנתונים שלנו, ומאחר ולא ישפיע על תוצאות המודל, הורדנו אותו כדי להימנע מעומס מיותר ולהתמקד בנתונים הרלוונטיים יותר.

העמודות LinkToTest, OpenLink, LinkToMre, ו-LinkMoreOpen הוסרו כיוון שהן מכילות לינקים (קישורים) לדוחות מעבדה/דוחות אירוע עבור כל רשומה כל אירוע כלבת עם פירוט על האירוע. שתיים מהעמודות עם הקישורים אינן זמינות וחסומות לפתיחה, אך שתי עמודות אחרות עם הקישורים כן פעילות וניתנות לפתיחה וקריאה. הלינקים הפעילים שימושיים במילוי נתונים חסרים והשלמת פערים או הבנת כל אירוע לפרטיו בצורה מעמיקה, אך עבור השלב הבא של מידול הנתונים והרצת האלגוריתמים עמודות אלו שמכילות קישורים לדוחות פחות מתאימות ורלוונטיות. לכן, כדי לשמור על מידול עם נתונים מדויקים וממוקדים, החלטנו להסיר עמודות אלו ולהתמקד בעמודות שמספקות מידע ישיר יותר.

העמודות RegionHeb, LocationNotSettelmentHeb ו- SpeciesNameHeb , SettlementHeb הוסרו מכיוון שהן מכילות את אותם הנתונים שקיימים כבר בעמודות באנגלית SettlementEng, LocationNotSettelmentEng , RegionEng ו SpeciesNameEng -העדפנו לעבוד עם עמודות באנגלית כדי לשמור על אחידות הנתונים ולהקל על תהליך המידול והקידוד עבור העיבוד והניתוח והסיווג שכן עבודה עם נתונים בעברית יכולה לסבך את התהליך, והיה כפל מידע עם עמודות שמכילות את אותם נתונים בעברית ובאנגלית. בנוסף ווידאנו שהמידע המופיע בעמודות האנגלית תואם למידע שהיה בעמודות בעברית כך שאין איבוד מידע או שגיאות.

GlobalID- מספקת מזהה ייחודי גלובלי לכל רשומה, שמיועד לאפיין את הנתונים בצורה ייחודית במערכות נתונים שונות. עם זאת, במודל שלנו אנחנו מתמקדים רק במקרים בישראל, ולכן המזהה הגלובלי אינו קריטי לצורך החיזוי והניתוחים, ולכן הוחלט להוריד אותה.

CreationDate- מציינת את תאריך יצירת הרשומה. הורדנו אותה כי היא לא תורמת ישירות לניתוח הנתונים או לחיזוי הכלבת, אלא רק נותנת מידע על מועד יצירת הרשומות.

Creator- מציינת את שם האדם שיצר את הרשומה. הורדנו אותה כי המידע על היוצר לא משפיע על הניתוחים שלנו ואין בו ערך ישיר בחיזוי הכלבת ואותו ערך לא משמעותי חזר על עצמו ברוב הרשומות.

EditDate- מציינת את תאריך עריכת הרשומה. הורדנו אותה כי היא לא מספקת מידע מספיק רלוונטי לחיזוי או לניתוחים שאנחנו מבצעים, אלא רק מעידה על מועד העדכון של הרשומות.

Editor- מציינת את שם האדם שערך את הרשומה. הורדנו אותה כי המידע על מי שערך את הרשומה אינו תורם לחיזוי הכלבת ואינו משפיע על איכות או דיוק המודל שלנו, ואותו ערך לא משמעותי חזר על עצמו ברוב הרשומות.

<u>עמודות שהשארנו והסבר:</u>

OBJECTID- מכילה מספר רשומה ייחודי לכל שורה בנתונים, ולכן השארנו אותה. היא חשובה לשמירה על ייחודיות הנתונים ומאפשרת לנו לעקוב אחרי כל רשומה בצורה ברורה. גם אם היא לא מספקת מידע ישיר או משמעותי בתהליך האנליזה, היא יכולה לשמש ככלי לניהול נתונים וזיהוי רשומות בהמשך התהליך שכן היא ממספרת לנו את האירועים מתחילת הטבלה ב2006 ועד האירוע הכי עדכני ב2025.

Event- מכילה את מספר האירוע בתוך השנה, כלומר מזהה ייחודי לכל אירוע שהתרחש באותה שנה. השארנו את העמודה הזו משום שהיא יכולה לספק אינדיקציה על כמות האירועים שהתרחשו בכל שנה. כרגע אנחנו לא בטוחים אם היא תהיה מועילה בשלב ניתוח הנתונים והאלגוריתם, אך בחרנו להשאיר אותה כדי שנוכל לבחון את השפעתה בהמשך. ייתכן שבעתיד נחליט להסיר אותה או שלא נשתמש בה אם נראה שאין לה תרומה משמעותית לחיזוי.

Date- מכילה את התאריך שבו התרחש כל אירוע. השארנו את העמודה הזו מכיוון שהתאריך מהווה מידע חיוני שיכול לעזור לנו להבין את ההתפלגות הזמנית של האירועים. תאריך עשוי להוות גורם חשוב בהבנת מגמות לאורך זמן, ולכן הוא נשאר כדי שנוכל לזהות דפוסים עונתיים או תקופתיים. כמו כן, הוספנו את עמודת החודש (Month) כדי לשפר את הדיוק במידת הצורך, ולוודא שהחיזוי יתבצע על פי תקופות ולא רק לפי ימים ספציפיים.

Year- מכילה את שנת האירוע. השארנו את העמודה הזו כי היא מספקת מידע חשוב לגבי ההתפלגות השנתית של האירועים. השנתון יכול לשמש לנו לזיהוי מגמות עונתיות או שנתיות ולהבנה של השפעת הזמן על האירועים. בנוסף, עמודת השנה חשובה במיוחד כשמשלבים אותה עם עמודת החודש, והיא יכולה לשפר את דיוק החיזוי במודלים עתידיים.

SpeciesNameEng- מכילה את שם סוג החיה הנגועה באנגלית. השארנו את העמודה הזו כי היא מספקת מידע חיוני על סוג החיה המעורבת בכל אירוע. הבנת סוגי החיות הנגועות יכולה להיות משמעותית בניתוח מגמות של התפשטות מחלות, זיהוי אזורים רגישים יותר ועוד. לכן, העמודה הזו נותרה, כי היא יכולה לתרום לתהליך החיזוי ולהפקת תובנות לגבי ההתפשטות או השפעת סוגי החיות.

RegionEng- מכילה את שם האזור באנגלית. בטבלה המקורית של אירועי כלבת ישנם 6 אזורים שונים מדווחים ברשומות. השארנו את העמודה הזו כי היא מספקת מידע חשוב על המיקום הגיאוגרפי של האירוע, וזה יכול להיות משמעותי להבנת ההתפשטות של מחלות או זיהוי אזורים עם סיכון גבוה יותר. השימוש בשמות אזורים באנגלית גם מקל על עיבוד הנתונים, במיוחד כשמדובר בעבודה עם מקורות בינלאומיים או אלגוריתמים שדורשים עקביות בשפה.

Species- מציינת את סוג מחלת הכלבת. כרגע אנחנו לא בטוחים אם היא תהיה משמעותית לניתוח הנתונים, אבל השארנו אותה כי ייתכן שהיא תסייע להבין את הדפוסים הקשורים למחלה. בעתיד, נוכל לבדוק את תרומתה למודל ובמידת הצורך להחליט אם וכיצד להשתמש בה.

SettlementEng- חשובה לחיזוי הכלבת כי היא מספקת מידע על המיקום הגיאוגרפי של האירועים במקרה הזה, שמות היישובים באנגלית. המיקום הוא פרמטר קריטי במודלים במיוחד כשמדובר במחלות זיהומיות כמו כלבת. ידע על היישוב בו התרחשו מקרי הכלבת יכול לעזור לזהות דפוסים גיאוגרפיים, כמו איזורים עם סיכון גבוה יותר, ולהתאים את החיזוי לפי מאפיינים מקומיים של כל יישוב. השם באנגלית חשוב לשימוש במערכות שמטפלות בנתונים בינלאומיים או דורשות התאמה של המידע עם מאגרי נתונים גלובליים. ת הנתונים השלמנו מהעמודה המקבילה בעברית, ווידאנו שהמידע תואם בכל השורות. השארנו את העמודה כדי להבטיח שהמידע יהיה אחיד ומלא.

LocationNotSettlementEng- היא עמודה שציינה מיקומים שאינם יישובים, כמו אזורים גיאוגרפיים או מקומות שאין להם שם יישוב מוגדר. את העמודה LocationNotSettlementEng הורדנו ואיחדנו אותה עם SettlementEng כי שמנו לב שבמקרים שבהם חסר נתון ביישוב באנגלית, נתון זה הופיע בעמודת LocationNotSettlementEng והעמודה הזו כמעט ריקה יש בה רק מקרים בודדים.
לכן החלטנו לאחד את העמודות כדי ליצור עמודה אחת שמסמלת את המיקום הכללי, בין אם מדובר ביישוב ובין אם לא, ובכך הפכנו את הנתונים ליותר מסודרים ואחידים. זה גם חוסך מקום בקובץ ומפשט את המידע, ומאפשר גישה למיקום באופן ברור יותר.

העמודות **x** ו **y-** מציינות את המיקום הגיאוגרפי המדויק של התרחשות האירועים, באמצעות קואורדינטות גיאוגרפיות קו רוחב וקו אורך. החלטנו להשאיר את העמודות הללו כי ייתכן שהן יעזרו לדייק את המידע על כל אירוע. המידע הגיאוגרפי יכול לעזור בזיהוי אזורים עם סיכון גבוה יותר לכלבת, ולשפר את הדיוק של המודל.

<u>להלן העמודות שהשארנו אחרי מחיקה ואיחוד נתונים:</u>

OBJECTID	Event	Date	Month	Year	SpeciesNameEng	RegionEng	Species	SettlementEng	x	y

עמודות אלו נבחרו מאחר והן משקפות פרטי מיקום, זמן וסוג האירוע הנחוצים לחיזוי.
ההחלטות לגבי הנתונים שנבחרו או הוסרו נעשו תוך שיקול דעת לגבי הרלוונטיות שלהם למודל, איכותם והצורך לשמור על עקביות ודיוק. כל הנתונים שנשמרו מציעים אינדיקציות חזקות לתהליך החיזוי ומספקים תמונה ברורה וממוקדת של נתוני כלבת בישראל.
בסעיף זה פירטנו לגבי בחירת העמודות בטבלה, ולגבי בחירת השורות בטבלה נרחיב בסעיף הבא.

2. ניקוי הנתונים

בשלב הניקוי, עברנו על הנתונים שנבחרו והנחנו דגש על תיקון בעיות שנמצאו בנתונים כדי להבטיח שהם יהיו מדויקים, עקביים וניתנים לשימוש בהמשך.

במהלך תהליך הניקוי, זיהינו מספר בעיות בנתונים ויישמנו פתרונות שונים:

1. נתונים חסרים

בחלק מהנתונים, השלמנו ידנית את המידע לפי הנתונים האחרים שהיו בקובץ. נתונים חסרים במיקומים גיאוגרפיים ופרטי החיות הושלמו גם על ידי קישורים למידע נוסף שהיה זמין באירועים (לינקים שמפרטים אודות האירועים ודוחות מעבדה). במקרים בודדים וחריגים בהם לא ניתן היה להשלים את הנתונים, במיוחד כאשר היו שורות ריקות, הוסרו השורות.

2. שגיאות נתונים

בעמודת species זן הכלבת זיהינו חוסר עקביות: בהתחלה היו מתעדים את זן הכלבת לדוגמא VII לרוב בעקביות עד אפריל 2024 לאחר מכן עברו לעדכן את זן החיה או שנשאר ערך ריק ולא תיעדו את זן הכלבת בעמודה זו עד סוף הטבלה העדכנית נכון לסוף מרץ 2025. את השורות עם הערכים החסרים האלו החלטנו למלא ב ערך" NULL כדי לשמור על עקביות ונתונים מדויקים.

בנוסף היו רשומות בודדות עם שגיאות בנתונים כגון חוסר התאמה בין הערכים בעברית לאנגלית או חוסר התאמה בין המיקום/היישוב של האירוע לבין אזור המדווח ותיקנו את הנתונים באופן מציאותי.

3. חוסר אחידות בקידוד

מקרה של חוסר אחידות בקידוד לדוגמא שמות היישובים, שמות אלו היו כתובים באנגלית והיו מקרים בהם נתונים היו חסרים או לא עקביים. השלמנו את השמות החסרים על פי העמודה בעברית שהורדה. מה שלא מצאנו חיפשנו את שם היישוב באינטרנט ותיקנו בהתאם את שם הישוב באנגלית, כך שהשמות יהיו אחידים בכל הקובץ. במקרים שבהם היו שמות חוזרים מאירועים קודמים, העתקנו את השם הקיים כדי לשמור על עקביות. בנוסף השתמשנו בקוד פייתון כדי לעבור על הנתונים ולוודא שאין בעיות של אחידות. כמו כן, הסרנו סימונים כמו מקף וגרשיים והשארנו רק גרש יחיד כדי לשמור על אחידות ותקינות.

כדי לנקות את הנתונים, השתמשנו בטכניקות הבאות:

הסרה של שורות לא תקינות :

שורות עם ערכים רבים חסרים או שגויים ולא הגיוניים שלא ניתן היה לתקן, הוסרו.

השלמה של נתונים חסרים :

במיקומים גיאוגרפיים, נתוני חיות, ושמות ישובים, השתמשנו בהשלמות ידניות וממקורות חיצוניים ומהעמודות שהכילו לינקים עם דוחות מפורטים לכל מקרה שסיפקו לנו מידע מפורט אודות כל אירוע.

תיקון שגיאות ידניות :

תיקנו בעיות עקביות בעמודות כגון עמודת זן הכלבת ושמות ישובים ושמות אזורים.

שימוש בקידוד עקבי :

אחרי שעברנו על השמות, ווידאנו אחידות בעזרת קוד פייתון ופונקציות באקסל, הסרנו סימנים לא נחוצים והספקנו אחידות בנתונים.

מקרים שלא ניתן לשחזר:

במהלך תהליך ניקוי הנתונים של הדאטה בייס המקורי של אירועי כלבת בישראל, זיהינו מקרים בודדים שבהם לא היה ניתן לשחזר את המידע בצורה תקינה.

מדובר ברשומות שבהן נמצאו שגיאות שנבעו מעדכון ידני שגוי שבוצע על ידי גורמים שהזינו את המידע לדאטהבייס המקורי (עובדים במשרד החקלאות).

ברוב המקרים שבהם נמצאו ערכים חסרים, השלמנו את הנתונים בצורה מושכלת או השארנו ערכים חסרים (למשל, None) מבלי למחוק את הרשומה, כל עוד שאר הערכים ברשומה היו קיימים ותקינים.

עם זאת, כאשר זוהו רשומות שבהן היו חסרים מספר רב של ערכים חשובים או כאשר נמצאו שגיאות חמורות ברוב שדות הרשומה (למשל, חוסר במיקום, סוג חיה וזן הכלבת גם יחד), ולא היה אפשר להשלים את המידע בצורה אמינה, החלטנו להסיר את אותן רשומות מהקובץ.

מדובר במספר קטן מאוד של מקרים, שבוצעו בזהירות ותוך שמירה על שלמות ואיכות מאגר הנתונים לצורך ניתוח מהימן בהמשך.

בכל מקרה שבו לא היה ניתן לשחזר או לתקן את הנתונים בצורה ברורה, נמנעו מהוספת הערכים השגויים ובמקום זאת הוסרו מהקובץ כדי לשמור על איכות הנתונים ולהימנע מהכנסת רעש או נתונים לא נכונים לתוך המודל. חשוב לציין שלא היו הרבה מקרים כאלה אלה כמה רשומות בודדות.

דוגמא צילום מסך של הטבלה עם הנתונים אחרי בחירה וניקוי:

OBJECTID	Event	Date	Month	Year	SpeciesNameEng	RegionEng	Species	SettlementEng	x	y
735	25	20/03/2025	3	2025	Jackal	Galil Golan	Na	Goren	35.237415	33.057974
734	24	20/03/2025	3	2025	Dog	Amakim	Na	iksal	35.321526	32.681132
733	23	20/03/2025	3	2025	Cattle	Shfela Vahar	Na	Shadmot Mehola	35.530511	32.347125
732	22	20/03/2025	3	2025	Jackal	Galil Golan	Na	Netu'a	35.323554	33.067101
731	21	18/03/2025	3	2025	Dog	Hasharon	Na	Ma'anit	35.029613	32.457490

3. יצירת נתונים חדשים

במטרה להעשיר את בסיס הנתונים הקיים של אירועי הכלבת ולשפר את יכולת הניתוח והחיזוי בפרויקט, החלטנו להוסיף עמודות (תכונות) חדשות אשר עשויות להיות בעלות השפעה על התפרצות מחלת הכלבת בישראל לאורך השנים.

לשם כך יצרנו, באמצעות קוד פייתון בסביבת Google Colab, שתי טבלאות נתונים חדשות, שכל אחת מהן מותאמת מבחינת מבנה לטבלת אירועי הכלבת המקורית:

- טבלת מזג אוויר:

יצרנו טבלה חדשה הכוללת נתוני מזג אוויר עבור כל חודש בין ינואר 2006 ועד מרץ 2025, עבור שלושה אזורים גאוגרפיים כלליים בישראל: צפון (North), מרכז (Center), ודרום (South).

שם הקובץ/טבלה: weather_summary_3_regions_israel_2006_2025

הנתונים בטבלה כוללים שלוש עמודות חדשות:

Avg Temperature (°C) – טמפרטורה ממוצעת חודשית

Monthly Precipitation (mm) – כמות משקעים חודשית במילימטרים

Rainy Days – מספר ימי גשם בכל חודש

הנתונים התקבלו באמצעות ספריית Meteostat ובוצעו התאמות על מנת למלא חודשים חסרים על סמך תחנה סמוכה לפי קואורדינטות גאוגרפיות.

מיפוי לאזורים כלליים עבור נתוני מזג האוויר בהתאמה לאזורים הגיאוגרפיים בטבלה הראשית:
מאחר שבטבלת הכלבת קיימים שישה אזורים גאוגרפיים (למשל: Galil Golan, Negev, Hasharon),
יצרנו עמודה חדשה בשם General Region שממפה כל אזור לאחת משלוש קבוצות עיקריות – North, Center, South – על מנת לאפשר התאמה לנתוני מזג האוויר. הנתונים והשמות באנגלית לצורך המשך עבודה נוחה באנגלית בקידוד בשלב מידול הנתונים ובהתאמה לשאר הנתונים בטבלאות.

לינק לקוד עם הסברים עבור יצירת הטבלה עם הנתונים החדשים ב google colab:
meteostat פרויקט גמר קוד נתונים מזג אוויר Colab - .ipynb

דוגמא צילום מסך של הטבלה עם הנתונים החדשים:

Region	Year	Month	Avg Temperature (Â°C)	Monthly Precipitation (mm)	Rainy Days
Center	2006	1	13.17	175.6	18
Center	2006	2	14.71	49.9	8

❖ טבלת היסטוריית מלחמות בישראל:

טבלה נוספת שנבנתה כוללת מידע על מלחמות ומבצעים צבאיים בין השנים 2006 ועד מרץ 2025.
שם הקובץ/טבלה: israel_war_history_2006_to_2025_Binary_with_war_names
הטבלה מחולקת לפי שנים וחודשים בפורמט מספרי, לכל חודש מוגדר האם התקיימה מלחמה (War in Israel – ערך בינארי Yes/No), ובמקרה שכן – מצוין שם המלחמה (War Name).
השמות נקבעו על בסיס תאריכי התחלה וסיום מדויקים כפי שמופיעים במקורות רשמיים (שמות כגון Second Lebanon War, Operation Cast Lead, Iron Swords War ועוד). הנתונים והשמות באנגלית לצורך המשך עבודה נוחה באנגלית בקידוד בשלב מידול הנתונים ובהתאמה לשאר הנתונים בטבלאות.

לינק לקוד עם הסברים עבור יצירת הטבלה עם הנתונים החדשים ב google colab:
פרויקט גמר קוד נתוני היסטוריית מלחמות ישראל Colab - .ipynb

דוגמא צילום מסך של הטבלה עם הנתונים החדשים:

Year	Month	War in Israel	War Name
2006	1	No	None
2006	2	No	None

---

4. שילוב הנתונים

לאחר יצירת שתי טבלאות הנתונים החדשות – טבלת מזג האוויר וטבלת היסטוריית המלחמות – ביצענו את שילובן עם טבלת אירועי הכלבת, שהיא בסיס הנתונים המרכזי שלנו בפרויקט.
התהליך בוצע באמצעות קוד פייתון בסביבת Google Colab, תוך שימוש בפונקציית merge של ספריית Pandas. במהלך תהליך הוספת הנתונים החדשים ומיזוג הנתונים עם הטבלה הראשית הנקייה, השתמשנו ונעזרנו בנוסף גם בתכנת power BI כולל power query ותכנת אקסל.

שילוב עם טבלת מזג האוויר:

ראשית, מיזגנו את טבלת הכלבת עם טבלת מזג האוויר לפי שלושה שדות מפתח משותפים:
Year, Month, ו-General Region.

המטרה הייתה לצרף לכל שורה בטבלת הכלבת את תנאי מזג האוויר ששררו באותו חודש ובאותו אזור כללי. השילוב בוצע באמצעות left join כך שכל שורת מידע על אירוע כלבת תישמר, גם אם חסר לגביה מידע מטבלת מזג האוויר.

שילוב עם טבלת המלחמות:

לאחר מכן, ביצענו מיזוג נוסף עם טבלת היסטוריית המלחמות, על פי שדות הזמן Year ו-Month בלבד (שכן טבלת המלחמות אינה מחולקת לפי אזורים).

כך צירפנו לכל אירוע כלבת את המידע האם הייתה מלחמה באותו חודש (War in Israel) ואת שם המלחמה (War Name), אם רלוונטי. גם כאן נעשה שימוש ב־**left join** כדי להבטיח שכל האירועים יישמרו, גם אם לא הייתה מלחמה באותו חודש.

המטרה העיקרית של שילוב זה היא ליצור טבלה אחת מרכזית המאחדת את כלל הנתונים הדרושים למחקר, תוך שמירה על קשר ישיר בין כל מקרה כלבת ובין התנאים הסביבתיים, האקלימיים והביטחוניים שהתקיימו בזמן ובמיקום שבו התרחש.

לינק לקוד עם הסברים עבור יצירת הטבלה הסופית בעזרת מיזוג הנתונים מהדאטהבייס עם הנתונים החדשים בgoogle colab:

פרויקט גמר כלבת קוד מיזוג נתוניים.ipynb - Colab

שם הקובץ/ טבלה הסופי: Rabies__Weather__War_Combined_1.4.25

5.עיצוב ופורמט הנתונים

לפני המעבר לשלב הבא של Data Modeling, שבו נריץ אלגוריתמים לצורך ניתוח וחיזוי, עיצבנו את הטבלה, וביצענו התאמה סופית למבנה הטבלה המרכזית שכללה את כל הנתונים החדשים ששולבו עד כה. שינינו ועדכנו את שמות העמודות לשמות קריאים באנגלית, באופן אחיד ומובן, כך שכל עמודה מייצגת בצורה מדויקת את המשמעות שלה, כלומר לכל העמודות יש שמות ברורים ואינטואיטיביים בהתאמה לנתונים.

לדוגמא:

Index Event ID – מזהה ייחודי של אירוע הכלבת

Event Per Year – מספר סידורי של האירוע באותה שנה

Animal Species, Rabies Species – סוג בעל החיים וסוג הכלבת

Region, Settlement, x, y – מיקום גאוגרפי, מיקום/ישוב, אזור/מחוז – לפי 6 סוגים שונים, וערכי נ"ץ.

Region_Weather – אזור כללי לצורך חיבור עם מזג האוויר - מכיל 3 אזורים צפון/מרכז/דרום מה שהיה בטבלת מזג אוויר עמודת General Region, במקום 6 אזורים שכפי שיש בעמודה הנוספת region מטבלת אירועי הכלבת.

Avg Temperature (°C), Monthly Precipitation (mm), Rainy Days – נתוני אקלים חודשיים כגון טמפרטורה ממוצעת לכל חודש בהתאם לשנה, כמות משקעים ממוצעת (גשם) בכל חודש בהתאם לשנה ובהתאמה לכך עמודת כמות ימי גשם בכל חודש.

War in Israel, War Name – נתוני מלחמות לפי חודש, עמודה בינארית כן/לא ועמודה עם שם המלחמה.

צילום מסך מקובץ אקסל של כל העמודות עם השמות העדכניים בדאטהבייס המעודכן והסופי:

Index Event ID	Event Per Year	Date	Year	Month	Animal Species	Rabies Species	Region	Settlement	x	y

Region_Weather	Avg Temperature (Ã‚Â°C)	Monthly Precipitation (mm)	Rainy Days	War in Israel	War Name

<u>בדיקת סוגי נתונים:</u>

ווידאנו שהעמודות המספריות מוגדרות כ־int או float בהתאם, ושעמודות של טקסט מוגדרות כ־string. כמו כן, ערכים חסרים (למשל כאשר לא הייתה מלחמה באותו חודש) מולאו בערכי ברירת מחדל כמו No או None. השתמשנו בכלים של אקסל, קוד פייתון בסביבת גוגל קולאב, ותוכנת power query שבתוך תכנת power BI עבור כל שלבי העריכה, התאמה, ניקוי נתונים, בדיקת הנתונים, התאמת עמודות ושורות, שיפור המבנה והפורמט של הנתונים והדאטהבייס.

<u>מודלים מתאימים לשימוש בטבלה המאוחדת:</u>

בהתאם לאופי הנתונים בפרויקט (אירועים בזמן, תכונות קטגוריות וכמותיות, שילוב מידע סביבתי וביטחוני), אנו שוקלות להשתמש במספר מודלים פשוטים לצד מודלים מתקדמים:

1. מודל Logistic Regression - מודל סיווג בסיסי שמעריך הסתברות לאירוע בינארי לפי תכונות קלט. מתאים לחיזוי האם התרחש אירוע כלבת בהתבסס על טמפרטורה, אזור, עונה או מלחמה.
2. מודל Decision Tree Classifier - מודל שמייצר עץ החלטות ברור להבנת התנאים המובילים לתוצאה. מאפשר לזהות בקלות השפעות של משתנים כמו אזור, גשם או מלחמה על מקרי כלבת.
3. מודל Random Forest - אוסף של עצי החלטה שמייצר תחזית משוקללת ומדויקת יותר. מפחית השפעה של רעש בדאטה, ומספק חיזוי יציב לאורך זמן.
4. מודל XGBoost - אלגוריתם Boosting עוצמתי ומהיר שמתאים לדאטה מורכב עם תכונות שונות. מתאים במיוחד לשימוש בעמודות כמו מזג אוויר, אזור, עונת השנה וסוג החיה.
5. מודל CatBoost - Boosting המיועד לעבודה עם תכונות קטגוריות באופן ישיר, ללא קידוד. מתאים לעבודה עם עמודות כמו Animal Species, Region, War Name מבלי צורך בהמרה ל־one-hot.
6. מודל Prophet (Time Series) - אלגוריתם סדרות זמן לחיזוי עונתי ומתמשך של אירועים לאורך זמן. מאפשר לזהות מגמות בעונתיות של מקרי כלבת לפי חודשים, ולהבין השפעות מזג אוויר עונתיות.
7. מודל LightGBM - אלגוריתם Boosting מתקדם שמתמקד בביצועים מהירים גם על דאטה גדול. מתאים מאוד למודלים הכוללים הרבה תכונות, ומומלץ לשימוש כאשר יש ערכים חסרים או קטגוריים.

לסיכום, הטבלה הסופית עוצבה באופן שמותאם לעבודה עם מגוון אלגוריתמים של למידת מכונה, כולל מודלים מבוססי סיווג, רגרסיה וסדרות זמן. עיצוב הנתונים בוצע תוך הקפדה על פורמט עקבי, סוגי משתנים נכונים ושמות עמודות ברורים, במטרה להבטיח המשך עבודה חלקה, מדויקת ויעילה בשלב בניית המודלים.

6. ניתוח נתונים ראשוני

ניתוח נתונים ראשוני (EDA - Exploratory Data Analysis) :

גרף עמודות עבור התפלגות מקרי כלבת לפי סוג בעל החיים

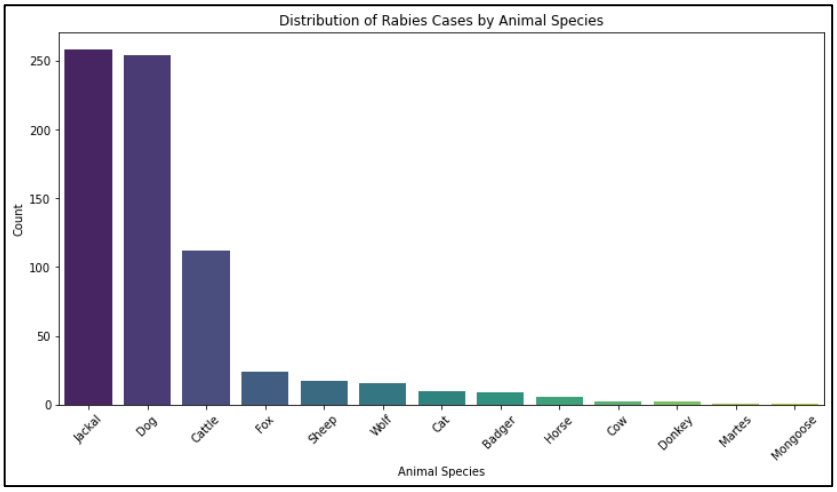
הגרף מציג את התפלגות מקרי הכלבת לפי סוג בעלי החיים שנפגעו.

בעלי החיים העיקריים המעורבים במקרי הכלבת הם תנים (Jackal) וכלבים (Dog), המובילים במספר המקרים ובפער משמעותי משאר בעלי החיים. לאחר מכן, מופיעים בקר וצאן (Cattle ו־Sheep), כאשר גם אצלם נצפים מקרים, אך בשכיחות נמוכה יותר.

בעלי חיים נוספים כמו שועלים (Fox), זאבים (Wolf), חתולים (Cat), גיריות (Badger), חמורים (Donkey) וסוסים (Horse) נמצאים במעורבות מועטה מאוד. קיימים מקרים בודדים בלבד של כלבת בקרב בעלי חיים אחרים כגון מרטס (Martes) ו־נמייה (Mongoose).

מהגרף עולה כי בעלי חיים משוטטים או פראיים למחצה (תנים, כלבים) מהווים את עיקר הסיכון להעברת מחלת הכלבת, מה שמדגיש את הצורך במיקוד פעולות המניעה והחיסון באוכלוסיות אלו.

Animal Species	
Jackal	258
Dog	254
Cattle	112
Fox	24
Sheep	17
Wolf	16
Cat	10
Badger	9
Horse	6
Cow	2
Donkey	2
Martes	1
Mongoose	1



גרף עמודות עבור התפלגות מקרי כלבת לפי אזור גיאוגרפי

הגרף מציג את מספר מקרי הכלבת בכל אחד מ6 האזורים השונים בישראל לפי טבלת אירועי הכלבת.

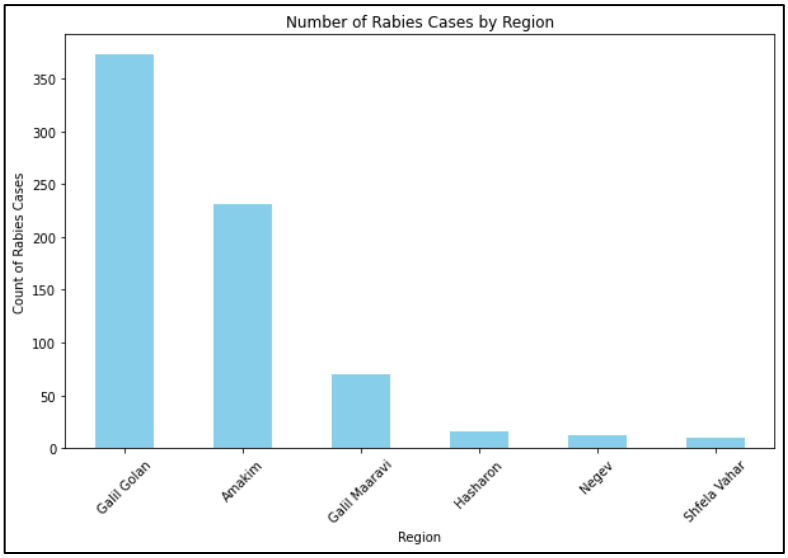
האזור עם המספר הגבוה ביותר של מקרי כלבת הוא גליל גולן (Galil Golan) עם 373 מקרים, מה שמצביע על סיכון גבוה במיוחד בהשוואה לשאר האזורים.

לאחר מכן, בולט אזור העמקים (Amakim) עם 231 מקרים, גם הוא עם שיעור גבוה של מקרי כלבת.

גליל מערבי (Galil Maaravi) מציג מספר מקרים נמוך יותר (70 מקרים), ואחריו אזורי השרון (Hasharon), הנגב (Negev) ושפלה והר (Shfela Vahar), עם מספר מקרי כלבת בודדים בלבד (פחות מ-20 מקרים בכל אחד). תוצאה זו עשויה להעיד כי אזורי הצפון (גליל גולן והעמקים) חשופים יותר להתפרצות מחלת הכלבת, ייתכן בשל גורמים סביבתיים כגון : אוכלוסייה גבוהה של בעלי חיים משוטטים או חיות בר, סמיכות לאזורים חקלאיים פתוחים, תנאים גיאוגרפיים ואקלימיים ייחודיים.

בהשוואה לאזורי המרכז והדרום, בהם יש פחות חשיפה לבעלי חיים נגועים והפיקוח המחמיר יותר, הסיכון באזורים אלו נמוך משמעותית.

Region	
Galil Golan	373
Amakim	231
Galil Maaravi	70
Hasharon	16
Negev	12
Shfela Vahar	10



<u>גרף קו של מספר מקרי הכלבת לאורך השנים:</u>

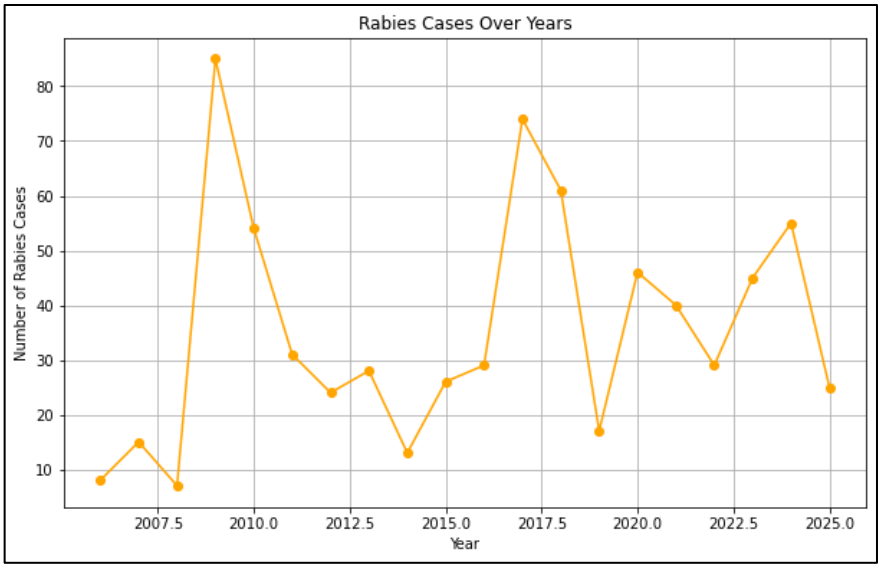
הגרף מציג את המגמה והתפלגות של מקרי הכלבת בישראל בשנים 2006–2025.

בהתבוננות בנתונים, ניתן לראות שבשנים 2009 ו-2017 ו2024 היה מספר גבוה במיוחד של מקרי כלבת. נכון לסוף הטבלה במרץ 2025 יש עדיין עליה בכמות המקרי כלבת בישראל, אך כיון שהנתונים הם לא משנה מלאה ב2025 רואים ירידה בגרף בשנה זו. מגמות אלו עשוית להעיד על רמת השיפור בעבודה במניעת התפרצות המחלה, כמו פעילויות חיסון לבע״ח, צעדים רגולטוריים או שינויים במזג האוויר.

העליה במספר מקרי הכלבת מסוף שנת 2023 ובמהלך 2024 ו2025 נובעת ככל הנראה ממלחמת חרבות ברזל. לאחר כל שיא במספר מקרי הכלבת נצפתה ירידה חלקית במספר המקרים, אך לאורך התקופה קיימת תנודתיות בולטת, כאשר מספר מקרי הכלבת ממשיך לעלות ולרדת בין השנים השונות.

למרות מאמצים של חיסון, ניטור ומניעה, הנתונים מצביעים על כך שמקרי הכלבת בישראל לא ירדו באופן עקבי ומתמשך, ומצריכים המשך טיפול אפקטיבי ופעולות ממוקדות להפחתת התפרצות המחלה.

Year	
2006	8
2007	15
2008	7
2009	85
2010	54
2011	31
2012	24
2013	28
2014	13
2015	26
2016	29
2017	74
2018	61
2019	17
2020	46
2021	40
2022	29
2023	45
2024	55
2025	25

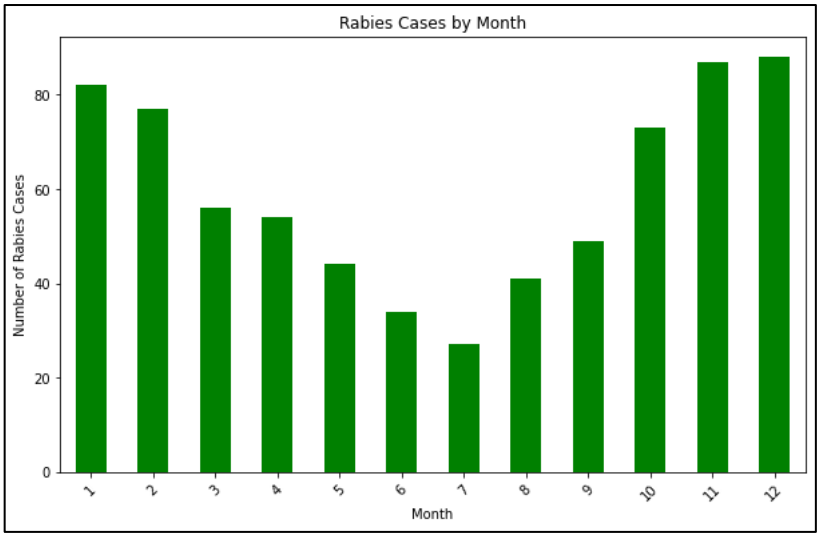


גרף עמודות עבור מגמות חודשיות במספר מקרי הכלבת

הגרף מציג את מספר מקרי הכלבת הממוצע בכל חודש לאורך השנים, ומאפשר לנו לזהות אם ישנם חודשים שבהם מספר המקרים גבוה יותר, כלומר אם קיימת עונתיות.

ניתן להבחין במגמה ברורה: בחודשי החורף (נובמבר–מרץ) יש עלייה במספר מקרי הכלבת, עם שיאים בחודשים דצמבר וינואר. לעומת זאת, בחודשי הקיץ (יוני–אוגוסט) נרשמת ירידה במספר המקרים.

לפי גרף זה ניתן להצביע על עונתיות או על תנאים סביבתיים שמעלים את הסיכון להידבקות בכלבת בתקופה זו, כגון ירידה ביעילות החיסון של חיות הבר והגורים שלהם בחודשי הקיץ החמים, מה שגורם לתחלואה גבוה יותר של בעלי חיים צעירים ובוגרים בחודשי החורף.



גרף ממוצע טמפרטורה מול מספר מקרי כלבת

הגרף מציג את הקשר בין ממוצע הטמפרטורה החודשית לבין מספר מקרי הכלבת בכל אזור בישראל.

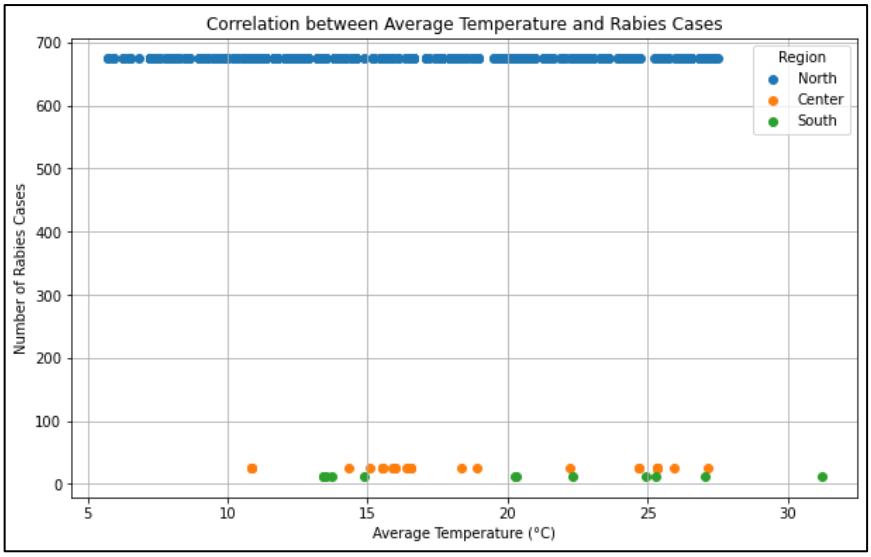
מהגרף ניתן לראות כי:

בצפון הארץ (North): מספר מקרי הכלבת גבוה מאוד באופן קבוע, ללא תלות מובהקת בטמפרטורה. רוב המקרים מתרחשים בטווח ממוצע של 10–20 מעלות צלזיוס.

באזור המרכז (Center) ובאזור הדרום (South): מספר מקרי הכלבת נמוך משמעותית. נראה כי באזורים אלו, כאשר הטמפרטורות עולות (לכיוון 25–30 מעלות), מספר המקרים אינו משתנה באופן מהותי.

באופן כללי, אזורים עם טמפרטורות נמוכות יותר (כמו הצפון) מציגים מספר גבוה יותר של מקרי כלבת, לעומת אזורים חמים יותר.

תוצאה זו יכולה להעיד על כך שהצפון מהווה אזור עם סיכון גבוה יותר להידבקות בכלבת, כנראה בשל גורמים נוספים ולא רק טמפרטורה. ובנוסף תוצאה זו עשויה להצביע על כך שמזג אוויר קריר יחסית תורם להתפרצות מחלת הכלבת, אך ייתכן שמעורבים גם משתנים נוספים כמו סוגי בעלי חיים, צפיפות אוכלוסיית חיות בר, ואופי גאוגרפי.



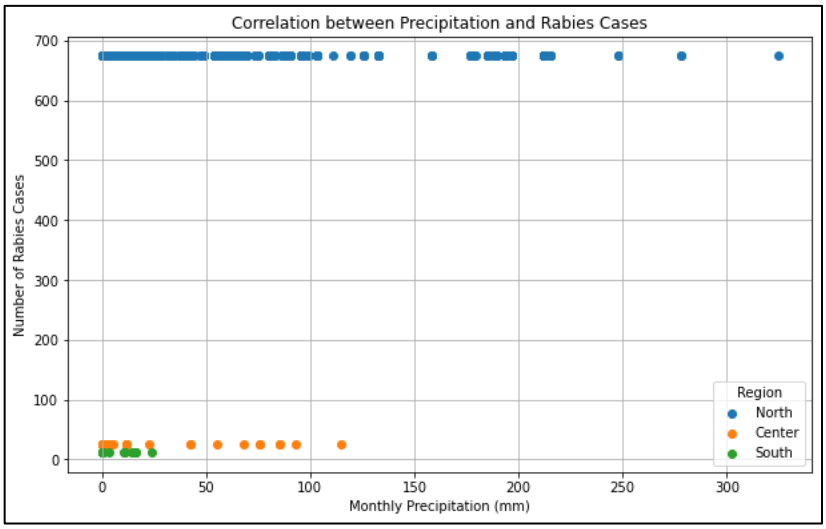
<u>גרף כמות המשקעים מול מספר מקרי כלבת</u>

באזור הצפון, מספר מקרי הכלבת גבוה מאוד לאורך התקופה, ונראה כי הוא מתרחש בעיקר כאשר כמות המשקעים נמוכה עד בינונית. ככל שכמות המשקעים גבוהה יותר, נצפית ירידה קלה במספר האירועים.

באזור המרכז, מספר מקרי הכלבת נמוך יחסית, ונראה כי הוא מתרחש גם כן בעיקר בכמויות משקעים נמוכות.

באזור הדרום, מספר מקרי הכלבת נמוך מאוד, עם ריכוז מקרים בעיקר באזורים ובחודשים בהם הכמות החודשית של המשקעים נמוכה.

מהגרף ניתן להסיק כי יש נטייה להתרחשות מקרי כלבת בעיקר בחודשים ואזורים בהם כמות המשקעים נמוכה, אם כי הקשר אינו ליניארי מוחלט ודורש בדיקה סטטיסטית נוספת לחיזוק המסקנה.



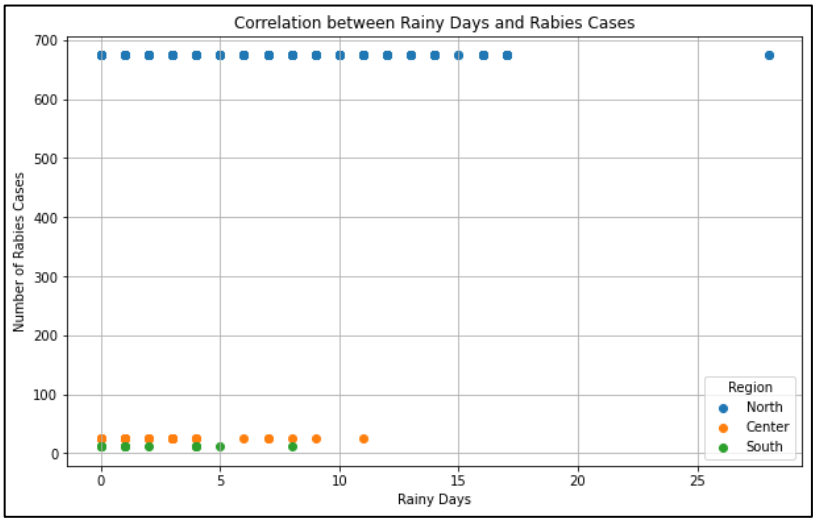
<u>גרף עבור מגמת הקשר בין ימי גשם למקרי כלבת</u>

הגרף מציג את הקשר בין מספר ימי הגשם החודשיים לבין מספר מקרי הכלבת בשלושת אזורי הארץ (צפון, מרכז ודרום).

ניתן לראות כי באזור הצפון (נקודות כחולות), מספר מקרי הכלבת נותר גבוה ועקבי לאורך כל טווח ימי הגשם, גם כאשר יש מעט מאוד ימי גשם בחודש.

במרכז ובדרום (נקודות כתומות וירוקות), לעומת זאת, מספר מקרי הכלבת נמוך יותר ומתרכז בעיקר בחודשים עם כמות ימי גשם מועטה.

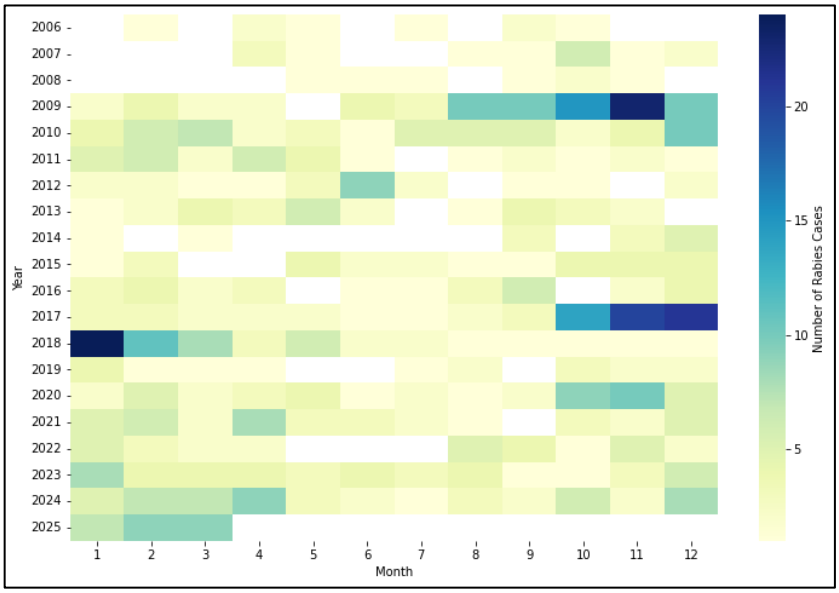
הממצאים מצביעים על כך שכמות ימי הגשם אינה משנה משמעותית את רמת התפרצות מחלת הכלבת, בעיקר בצפון. ייתכן כי גורמים אחרים הם בעלי השפעה דומיננטית יותר מאשר פרמטר ימי הגשם.



<u>גרף מפת חום עבור פיזור מקרי הכלבת לפי חודש ושנה</u>

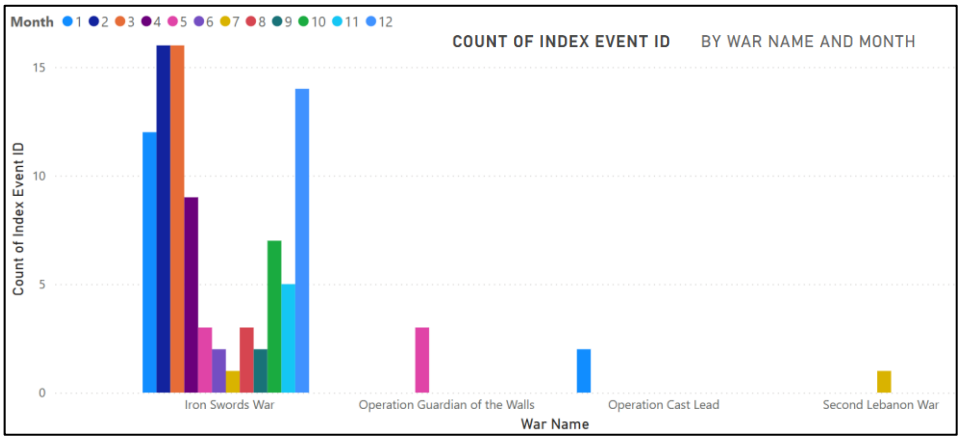
הגרף מציג את התפלגות מקרי הכלבת לאורך חודשים ושנים בעזרת מפת חום (Heatmap), שבה צבעים כהים יותר מייצגים מספר גבוה יותר של מקרים.

ניתן לזהות כי קיימות שנים עם ריכוז גבוה של מקרי כלבת, בעיקר בסוף השנה (חודשים נובמבר–דצמבר), למשל בשנים 2009, 2017, ו־2024. בחלק מהשנים (כגון 2010, 2019 ו־2022) נראית פעילות מתונה יותר של מקרי כלבת, המתפרסת על פני חודשים שונים. הנתונים מצביעים על כך שבחלק מהשנים קיימת עונתיות מסוימת עם ריבוי מקרי כלבת בחודשי החורף (סוף השנה), אך יש גם שנים שבהן ההתפלגות פחות עקבית.



גרף התפלגות מקרי כלבת לפי מלחמה וחודש:

הגרף מציג את מספר מקרי הכלבת במהלך תקופות של מלחמות שונות, תוך חלוקה לפי חודשי השנה. ניתן לראות כי רוב מקרי הכלבת התרחשו במהלך מלחמת חרבות ברזל (Iron Swords War), במיוחד בחודשי החורף (חודשים 1, 2, ו-12). לעומת זאת, בתקופות מלחמות אחרות כמו שומר החומות (Operation Guardian of the Walls), עופרת יצוקה (Operation Cast Lead) ומלחמת לבנון השנייה (Second Lebanon War), התרחשו פחות מקרי כלבת. התוצאה מחזקת את ההשערה כי בתקופות לחימה, ובפרט במלחמה האחרונה, הייתה מגמת עלייה במספר מקרי הכלבת בישראל, ייתכן בשל שינויי סביבה, תנועת בעלי חיים או ירידה בפעולות הפיקוח והחיסון באותם זמנים, בנוסף כמובן שהניתן להבין שמלחמות/מבצעים קודמים היו קצרים בהרבה לעומת מלחמת חרבות ברזל שנמשכת חודשים רבים ולכן מספר מקרי הכלבת גבוה יותר, בנוסף כנראה שבעבר היה פחות דיווח ותיעוד של מקרי הכלבת באופן מסודר לעומת השנים האחרונות עם השיפור הטכנולוגי וזמינות המידע.



סטטיסטיקות מזג אוויר: משקעים חודשיים, ימי גשם וטמפרטורה ממוצעת

משקעים: כמות המשקעים החודשית הממוצעת היא 48.44 מ״מ, עם סטיית תקן של 64.3 מ״מ, המעידה על פיזור רחב ושונות גבוהה בין חודשי השנה. ימי גשם: מספר ימי הגשם החודשי הממוצע עומד על 5.28 ימים, עם סטיית תקן של 4.99 ימים, מה שמרמז על הבדל ניכר בין עונות גשומות ליבשות.

טמפרטורה: הטמפרטורה החודשית הממוצעת היא C16.16°, עם סטיית תקן של C5.75°, מה שמעיד על שונות עונתית ברורה אך מתונה יחסית לאורך השנה.

מדד	משקעים חודשיים (מ"מ)	ימי גשם	טמפרטורה ממוצעת (מעלות צלזיוס)
סה"כ	712	712	712
ממוצע	48.44	5.28	16.16
סטיית תקן	64.3	4.99	5.75
מינימום	0	0	5.71
רבעון 25%	0.45	1	11.77
רבעון 50%	16.9	4	15.14
רבעון 75%	69.6	9	20.88
מקסימום	324.7	28	31.21

במהלך העבודה על מסמך זה כחלק מפרויקט הגמר, ביצענו ניקוי, יצירה, שילוב ועיצוב נתונים ממספר מקורות תוך הקפדה על דיוק והתאמה לצורכי הפרויקט. ערכנו ניתוח נתונים ראשוני (EDA) לזיהוי דפוסים ומגמות עיקריות. כעת יש בסיס נתונים איכותי ומוכן להמשך בניית המודלים והחיזוי.

בברכה,

דניאל ואימאן.