The Academic College of Yezreel Valley
Department of Information Systems

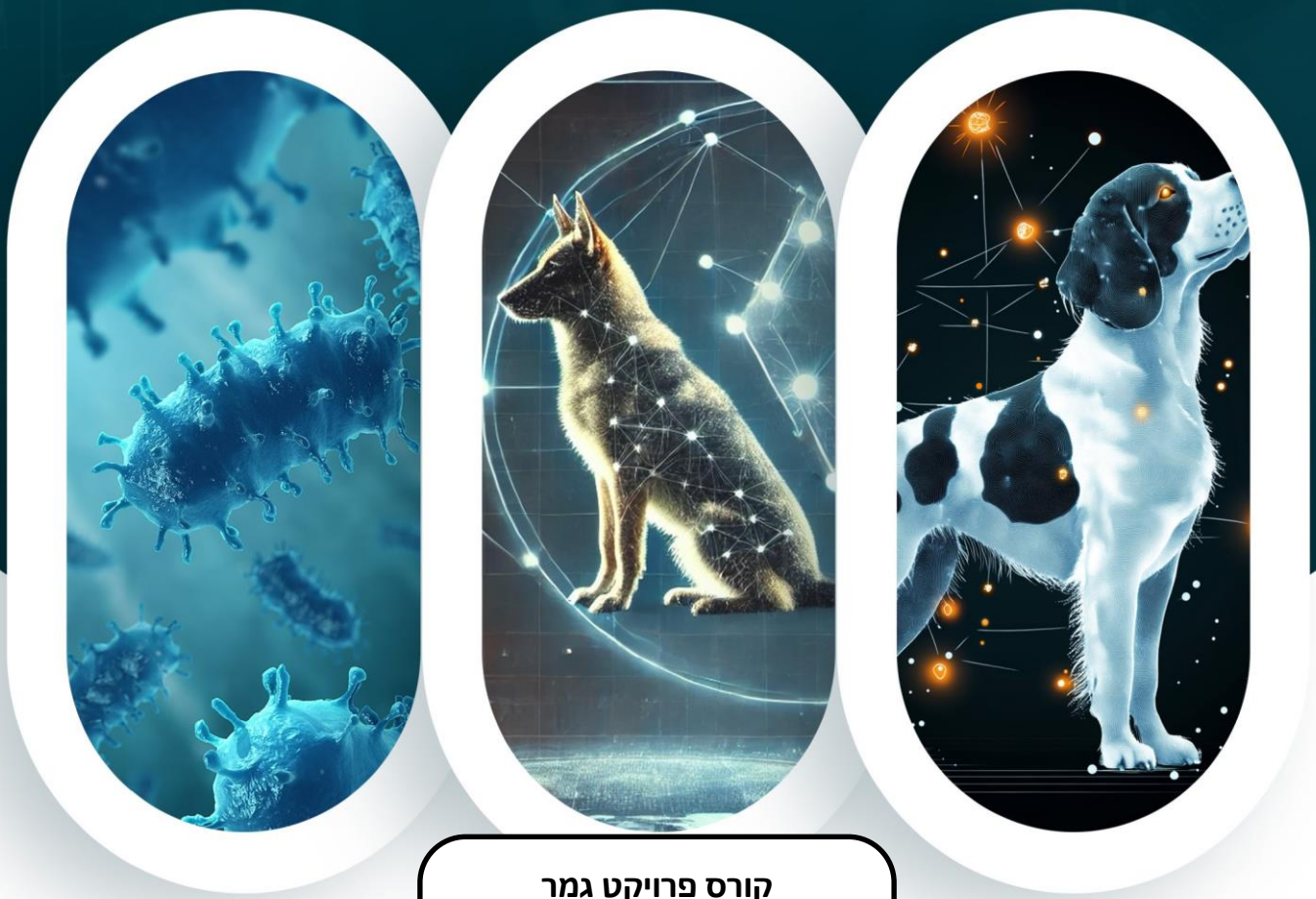
המכללה האקדמית עמק יזרעאל
חוג למערכות מידע

Data understanding report

פרויקט גמר מסמך הבנת 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: 31.1.2025
שם המרצה: מר זכאי אבי
שם המנחה: גב' גוטפריד ג'ניה

תוכן עניינים

1. איסוף נתונים

1.1 מקורות נתונים

- המקור נתונים הראשון:

https://data.gov.il/dataset/rabies-occurrence/resource/85cd1c27-7824-45ff-9982-22da0285fafb

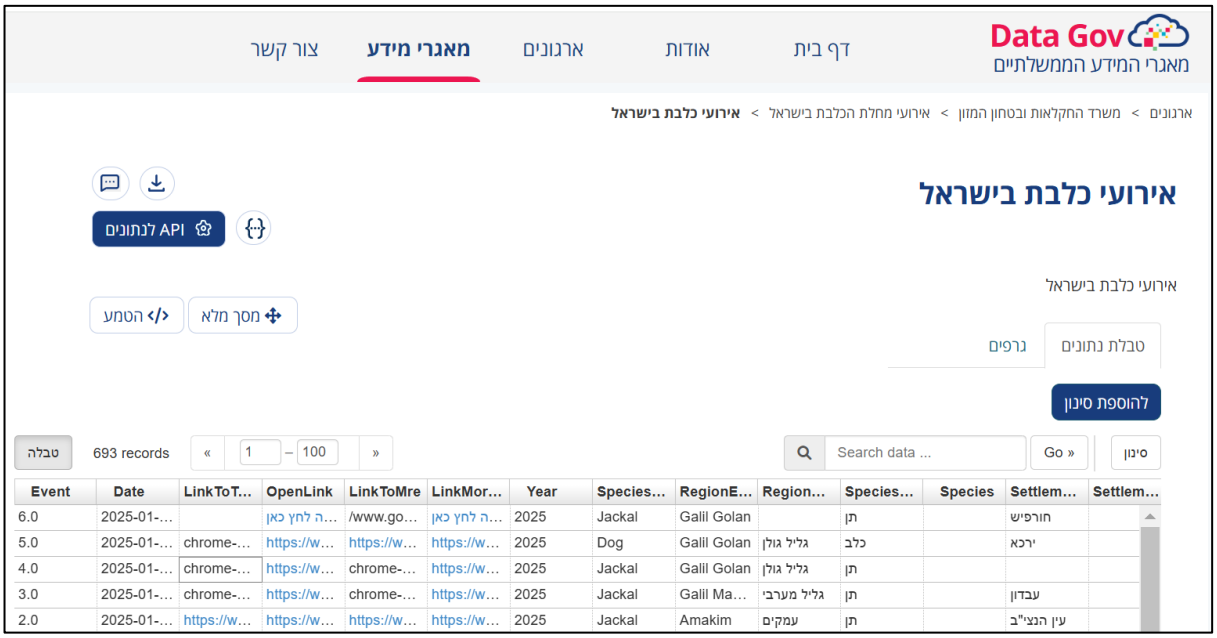
האתר בלינק הזה הוא מאגר נתונים ממשלתי ב *data.gov.il*-המכיל מידע מפורט על מקרי כלבת בישראל, כולל נתוני זיהוי של האירועים, סוגי בעלי חיים נגועים, תאריכים, מיקומים גיאוגרפיים, וקישורים לדוחות PDF המפרטים כל אירוע, נתוני GIS: נתוני רוחב ואורך גיאוגרפיים מאפשרים מיפוי מדויק. המאגר מתעדכן באופן שוטף ומאפשר ניתוח מגמות אפידמיולוגיות.

- המקור נתונים השני (שהוא משלים של המקור הראשון):

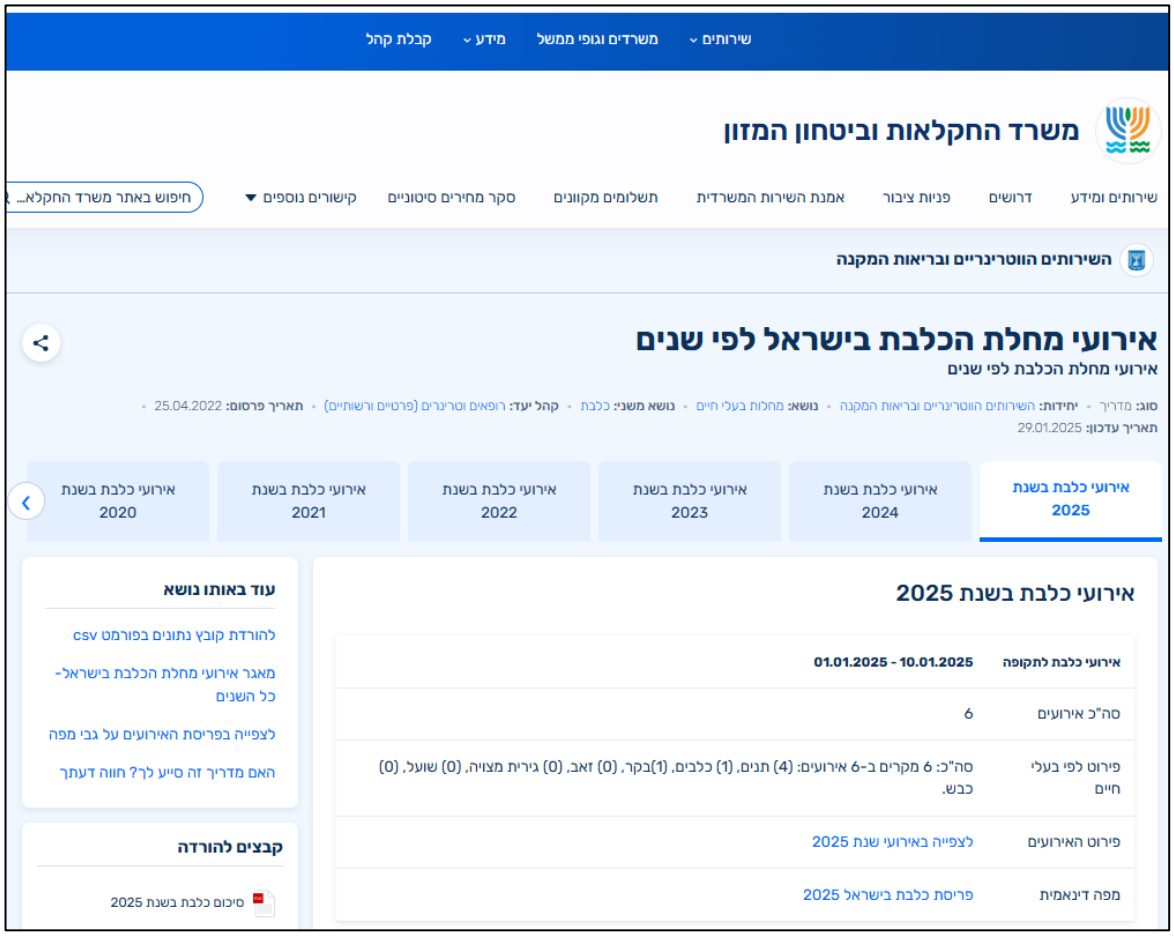
https://www.gov.il/he/pages/rabies-occurrence-years

האתר בלינק מספק מידע ממשלתי רשמי על מקרי כלבת בישראל לאורך השנים, כולל נתונים סטטיסטיים, מפות, ודוחות מפורטים על התפרצויות ומיקומים גיאוגרפיים של אירועים מתועדים. המקור הראשון, הכולל חיבור API וקישור לקובץ Excel, מכיל רשומות שבהן חסרים לפעמים פרטים כמו יישובים או מיקומים גיאוגרפיים. לעומת זאת, המקור השני מספק מידע מעודכן ומפורט יותר, מה שמאפשר להצליב ולהשלים את הנתונים החסרים, במיוחד עבור שדות הקשורים למיקום. שילוב המידע משני המקורות יכול לשפר את שלמות ואמינות הנתונים לניתוח מעמיק יותר.

- **צילום מסך מקור הראשון:**



- **צילום מסך מקור שני:**



מקורות הנתונים כוללים:

- **נתונים קיימים:** המידע נאסף מדיווחים קיימים ורשמיים, נתונים מעודכנים ומבוקרים שנאספים ומדווחים על ידי משרד החקלאות וביטחון המזון, על ידי השירותים הווטרינריים ובריאות המקנה, ועל ידי רשויות מקומיות. הנתונים כוללים תיעוד של מקרים מאומתים ופרטים רלוונטיים הקשורים אליהם.
- **נתונים ממשלתיים פתוחים:** המאגר המרכזי ב-**data.gov.il** מספק קובץ **Excel** ו-**API** עם נתונים עדכניים על מקרים חדשים, לצד מידע היסטורי על התפרצויות כלבת בישראל. הנתונים זמינים לציבור הרחב ואין צורך ברכישתם, שכן מדובר במידע ממשלתי פתוח ונגיש לכל. המאגר מתעדכן באופן שוטף וכולל קישורים לדוחות רשמיים מפורטים.
- **מקורות נוספים ושילוב נתונים עתידי:** המאגר מכיל נתונים היסטוריים ומעודכנים לאורך תקופה של שני עשורים, מה שמאפשר ניתוח מגמות ארוכות טווח. במקרה הצורך, ננסה להצליב את הנתונים עם מקורות חיצוניים נוספים, כגון נתוני מזג אוויר (כמו כמות משקעים וטמפרטורה) המתאימים לתאריכי האירועים. שילוב זה עשוי לסייע בזיהוי קשרים בין תנאי האקלים להתפרצות מקרי כלבת, כלומר לנתח האם קיימת השפעה, ולזהות דפוסים ומגמות נוספות.

1.2.בדיקה ראשונית של הנתונים

בשלב זה ניתחנו את הנתונים על מנת לזהות מאפיינים מרכזיים, לבחון אילו שדות רלוונטיים להמשך הניתוח, ולזהות אתגרים פוטנציאליים במיזוג נתונים ממקורות שונים.

מאפיינים מבטיחים:

העמודות המרכזיות שנבחרו כחשובות לניתוח כוללות את **Date, SettlementHeb, RegionHeb, SpeciesNameHeb**, ו-**Year**, שכן הן מאפשרות ניתוח מגמות אזוריות, השוואה בין סוגי בעלי חיים נגועים, וזיהוי דפוסים לאורך זמן. בנוסף, ייתכן שגם העמודות **x** ו-**y**, המכילות את המיקום הגיאוגרפי המדויק של האירוע, יהיו שימושיות למיפוי ולניתוח גיאוגרפי מעמיק יותר.

מאפיינים לא רלוונטיים:

חלק מהעמודות, כמו **GlobalID** ו-**Creator**, אינן תורמות הרבה לניתוח ולכן ייתכן שניתן להסירן. יש לבחון האם שדות נוספים, כמו מזהי רשומות פנימיים, נדרשים או שניתן לצמצמם.
בנוסף, קיימות עמודות המכילות קישורים לדוחות ומקורות מידע נוספים, אך חלק מהקישורים אינם פעילים או חסומים לגישה, מה שהופך אותם לפחות רלוונטיים עבור המשך העבודה. אנו נבחן האם ניתן להחליפם במידע זמין אחר או להסירם במידת הצורך.

האם יש מספיק נתונים לניתוח?

מערך הנתונים כולל 693 רשומות, נכון ל-29.1.2025. נתונים אלו מספקים בשלב זה תשתית מספקת לפרויקט, ומהווים בסיס לניתוח סטטיסטי, ניתוח מגמות, הפקת תובנות וביצוע חיזויים. עם זאת, ככל שנתקדם בשלבי הפרויקט מעיבוד הנתונים והלאה, ייתכן שיהיה צורך להרחיב את מאגר הנתונים על מנת לשפר את דיוק המודל ולחדד את יכולת ניבוי המגמות כדי לשפר ולהרחיב את הפרויקט. הרחבת מאגר הנתונים עשויה לכלול שילוב נתונים נוספים ממאגרים חיצוניים, בהתאם לצרכים שיעלו בהמשך הפרויקט.

בעיות אפשריות במיזוג נתונים:

בעת שילוב הנתונים עם מקורות נוספים, כמו נתוני מזג אוויר, נצטרך לוודא שיש התאמה חד-חד ערכית בין המפתחות של מאגרי הנתונים לצורך המיזוג, וכן התאמה בפורמטים השונים, כי עשויות להתעורר בעיות התאמה בין פורמטים שונים של נתונים, לדוגמא בתאריכים אך בעיקר בקידוד שמות מקומות ושמות הזנים של בעלי החיים הנמצאים במאגרים בשפות שונות (מבחינת שמות בעברית מול אנגלית), ולכן נצטרך לטפל בכך, לדאוג שתהיה אחידות בקידוד לבצע המרות נתונים במידת הצורך, ולבחון מנגנונים אוטומטיים לזיהוי והתאמת ערכים חופפים.

התמודדות עם ערכים חסרים:

בניתוח הנתונים זיהינו כי קיימים ערכים חסרים בעמודות מסוימות, בעיקר **LocationNotSettlementEng, LocationNotSettlementHeb**, ו-**Species**. ערכים חסרים

אלו עלולים להשפיע על איכות הניתוחים הסטטיסטיים והמסקנות המתקבלות, במיוחד כאשר הם קשורים למידע גיאוגרפי חיוני או לסיווג וירוס הכלבת.

בבדיקת המקור ממנו התקבלו הנתונים, לא מצאנו תיעוד רשמי לשיטה שבה טופלו הערכים החסרים. לא נראה כי הוחלפו בערכים ברירת מחדל או הושלמו באמצעות אמידת נתונים. הרשומות החסרות פשוט נותרו ריקות, ללא סימון מיוחד (למשל **NA, NULL**, או ערך סמן אחר).

בנוסף לערכים החסרים, זוהתה בעיית חוסר עקביות בעמודת **Species**:

- ברוב הרשומות, הערך בעמודה זו מציין את זן נגיף הכלבת שנמצא בבדיקה, למשל **VII A**.
- עם זאת, בחלק מהרשומות הערך מציין זן של בעל החיים, למשל "**זהוב**" עבור תן זהוב, במקום זן נגיף הכלבת.
- חוסר העקביות עלול לפגוע בדיוק הניתוחים ולהוביל לבלבול בין סוגי הנתונים.

<u>דרכי טיפול מוצעות:</u>

- בדיקה אם ניתן להשליך מהערכים הקיימים – ייתכן שבמקרים מסוימים ניתן להסיק את המידע החסר מתוך עמודות אחרות או על סמך רשומות דומות.
- הצלבת מידע ממקורות נוספים – אם המידע חסר באופן משמעותי, ניתן לנסות להשלימו באמצעות נתונים חיצוניים, כגון מידע גיאוגרפי רשמי.
- שימוש במנגנון סימון לערכים חסרים – לצורך שקיפות ודיוק, ניתן לסמן ערכים חסרים במפורש (**NA** או **Unknown**) כדי למנוע עיבוד שגוי של הנתונים.
- סטנדרטיזציה של קידוד המידע בעמודת **Species** – נוודא שכל הערכים יתייחסו לזני נגיף הכלבת בלבד, ונעביר מידע על זן בעל החיים לעמודות המתאימות (**SpeciesNameHeb, SpeciesNameEng**).
- תיקון ערכים לא עקביים – נבצע בדיקות לאיתור רשומות שבהן נרשם בטעות זן של בעל חיים במקום זן הנגיף, ונעדכן בהתאם לנתונים הקיימים במערכת.

בהמשך, בהתאם להתקדמות הפרויקט, נבחן אילו שיטות מתאימות ביותר לטיפול בערכים החסרים ולשמירה על עקביות הנתונים.

<u>2. תיאור הנתונים</u>

<u>2.1. כמות הנתונים</u>

- **מספר רשומות**: מערך הנתונים מכיל 693 רשומות (מעודכן לפי 29.1.2025).
- **מספר עמודות:** הקובץ מכיל 25 עמודות בגרסת ה-**Excel** הגולמית, בעוד שבנתונים בטבלה בגרסת האתר **API** קיימות 26 עמודות, כלומר קיימת עמודה נוספת של **id**, שאינה מופיעה בקובץ ה-**Excel**, עמודה זו **id** משמשת כמזהה ייחודי (**Identifier**) לכל אירוע כלבת, וממספרת את האירועים/רשומות באופן רציף מ-1 ועד 693.

ניקח בחשבון הבדל זה במספר העמודות, בעת שילוב ועיבוד הנתונים ממקורות שונים בפרויקט.

2.2 סוגי הערכים

חלק מהעמודות במאגר הנתונים נדרשות להמרה או להתאמות בעת עבודה עם קובץ **CSV** או **Excel**. לדוגמה, עמודת **Date** מכילה ערכים בפורמט תאריך ולכן יש להגדירה כנתוני **DATE** ולא כטקסט (**text**) כפי שהיא מופיעה תחילה בהורדת הנתונים. באופן דומה, עמודת **Year** צריכה להיות מוגדרת כנתון מספרי (**numeric**) ולא כטקסט. כלומר בעת הורדת הנתונים, חלק מהעמודות עלולות להופיע בפורמט שגוי, ולכן נדרש להגדיר אותן כראוי בהתאם לפלטפורמה/תוכנה שבה משתמשים (למשל **Power BI / Python**) כדי להבטיח ניתוח תקין ויצירת גרפים מדויקים.

סוגי הנתונים העיקריים:

- **נתוני תאריכים**: עמודת **Date** מכילה תאריכים בפורמט אחיד **(dd-mm-yyyy)**.
- **נתוני טקסט**: **SpeciesNameEng, RegionHeb** ו-**SettlementHeb** הן עמודות לדוגמא שמכילות ערכים של שמות זני בעלי חיים ושמות של מיקומים בישראל: אזורים ויישובים, הסיומת **Eng** או **Heb** בשם העמודה מציינת האם הערך מופיע בעברית או באנגלית.
- **נתוני מיקום**: כפי שציינו המידע כולל שמות אזורים ויישובים (בעברית ובאנגלית) , אבל ישנם גם נתוני מיקום גיאוגרפיים מדויקים (קואורדינטות). העמודות **x** ו-**y** מייצגות נתוני קואורדינטות גיאוגרפיות (קו רוחב וקו אורך), ולכן נדרש להגדירן כנתונים מספריים לצורך ניתוח מיקום מדויק.

על מנת להבטיח עקביות בעיבוד הנתונים, נוודא כי סוגי הנתונים מוגדרים נכון בהתאם למערכת שבה נעשה שימוש.

מילון נתונים:

רשימת העמודות וסוגי הנתונים/ערכים במאגר (כפי שמופיעים במקור לפני המרה):

*סוגי הנתונים בטבלה זו משקפים את הפורמט המקורי כפי שמופיע במאגר, לפני ביצוע התאמות והמרות לצורך עיבוד הנתונים.

שם עמודה	סוג נתון
id	numeric
OBJECTID	numeric
Animal_Lab_ID	text
Event	text
Date	text
LinkToTest	text

OpenLink	text
LinkToMre	text
LinkMoreOpen	text
Year	text
SpeciesNameEng	text
RegionEng	text
RegionHeb	text
SpeciesNameHeb	text
Species	text
SettlementHeb	text
SettlementEng	text
LocationNotSettlementEng	text
LocationNotSettlementHeb	text
GlobalID	text
CreationDate	text
Creator	text
EditDate	text
Editor	text
x	numeric
y	numeric

2.3 סכמות קידוד

מערך הנתונים מכיל מספר עמודות שבהן נעשה שימוש בקידוד לייצוג משתנים קטגוריים:

סוגי בעלי חיים:

העמודות **SpeciesNameEng** ו- **SpeciesNameHeb** מציינות שמות בעלי החיים באנגלית ובעברית. בנוסף, סוג בעלי החיים מקודד גם בעמודת **Species**, אך במקרים רבים ערך זה אינו מתייחס לסוג החיה אלא לזן נגיף הכלבת שנמצא בבדיקה, לדוגמה **VIIA**.

אזור גיאוגרפי:

שם האזור מופיע בשתי עמודות: **RegionHeb** (בעברית) ו- **RegionEng** (באנגלית).
שמות היישובים מופיעים בעמודות **SettlementHeb** ו- **SettlementEng**, בהתאם לשפה.

נתוני קישור:

שדות כמו **LinkToTest** מספקים קישורים ישירים לדוחות מפורטים בפורמט **PDF** אונליין.
בנוסף, ייתכנו הבדלים וחוסר עקביות בקידוד שמות היישובים ושמות הזנים, ולכן ייתכן שיהיה צורך לבצע המרות כדי להבטיח אחידות בעיבוד הנתונים.

---

3. חקר הנתונים

במהלך חקר הנתונים, גיבשנו מספר השערות ראשוניות שהתבססו על מאפייני המידע הקיים.
ההשערות הראשוניות כוללות:

❖ **השערות:**

- **הקשר בין מיקום גיאוגרפי להתפרצות**: יש אזורים גיאוגרפיים מסוימים שבהם התפרצות הכלבת נפוצה יותר. כגון הגליל, העמקים והגולן, מציגים שכיחות גבוהה יותר של מקרים.
- **עונתיות**: ייתכן שיש עונתיות בהתרחשות המקרים, מספר המקרים עולה בעונות מסוימות. לפי הנתונים שיש לנו אנחנו רואים עלייה במיוחד בחורף ובסתיו.
- **סוגי בעלי חיים**: יש הבדל בין סוגי בעלי חיים בהקשר להתפרצות המחלה, מרבית מקרי הכלבת מתרכזים בסוגי בעלי חיים מסוימים, כמו תנים וכלבים.

❖ **מאפיינים הנראים מבטיחים להמשך הניתוח:**

מיקום גיאוגרפי (RegionHeb, RegionEng):

תורמים לזיהוי מוקדי התפרצות וזה יכול לכלול ערים, אזורים כפריים, או אזורים עם צפיפות אוכלוסין גבוהה של בעלי חיים.

תאריכים (Date):

יכולים לעזור לזהות דפוסים עונתיים או מגמות לאורך זמן, ניתן לבדוק אם יש עלייה במקרים בעונות מסוימות.

סוגי בעלי חיים (SpeciesNameEng, SpeciesNameHeb):

בעלי חיים מעורבים בהתפרצות יכולים להצביע על מקורות שונים של המחלה. לדוגמה, איזה סוג או מין הם המקור העיקרי.

מזהה ייחודי לכל בעל חיים שנבדק (Animal_Lab_ID):

יכול לשמש למעקב אחרי מקרים ספציפיים ולבצע ניתוחים על מקרים חוזרים.

יישוב שבו התרחש האירוע (SettlementHeb ,SettlementEng):

מאפשר ניתוח ברמה מקומית. ניתן לבדוק אם יש יישובים עם שיעור גבוה יותר של מקרים.

מיקום שאינו יישוב- שטחים פתוחים (LocationNotSettlementEng, LocationNotSettlementHeb):

יכול לעזור להבין את הקשרים בין מיקומים לא מאוכלסים לבין התפרצות המחלה.

❖ **<u>האם הגילויים שחשפתם שינו את ההשערות הראשוניות שלכם?</u>**

הגילויים שחשפנו במהלך הניתוח שינו את ההשערות הראשוניות שלנו באופן משמעותי.
במהלך הניתוח, גילינו מאפיינים נוספים שהעשירו את ההבנה שלנו לגבי התפרצות מחלת הכלבת.
אחד הגילויים המרכזיים היה השימוש בקואורדינטות גיאוגרפיות, אשר חשף מוקדים גיאוגרפיים ספציפיים שבהם התפרצות המחלה הייתה גבוהה יותר. גילוי זה הדגיש את החשיבות של ניתוח גיאוגרפי במעקב אחר התפרצויות מחלה, והראה לנו כיצד ניתן לנצל את המידע הגיאוגרפי כדי להבין טוב יותר את התפשטות המחלה.
בנוסף, נמצא כי ישנה עלייה מובהקת במספר המקרים באזורים מסוימים בצפון הארץ, במיוחד בסמוך לריכוזי חיות בר. גילוי זה שינה את ההשערות שלנו לגבי הקשרים בין סוגי בעלי חיים לבין התפרצות המחלה. בעוד שההשערה המקורית הייתה שמקרי הכלבת מתרכזים בעיקר בבעלי חיים מסוימים, גילוי זה הראה שגם מינים פחות צפויים מעורבים כגון בקר וגירית.
בהתאם לכך, הותאמו מטרות הניתוח כך שיתמקדו בזיהוי מוקדים גיאוגרפיים עם סיכון גבוה להתפרצות בעתיד. זה כולל את הצורך לשקול שילוב של נתונים סביבתיים, כמו תנאי מזג אוויר, כדי להבין את הקשר בין הסביבה לשכיחות המקרים.
לסיכום, גילויים אלו לא רק שיפרו את ההבנה שלנו לגבי התפשטות מחלת הכלבת, אלא גם הובילו אותנו לשקול אסטרטגיות חדשות למניעת התפרצות, כמו חיזוק המעקב והטיפול באזורים עם ריכוז גבוה של חיות משק וחיות בר.

❖ **<u>האם חקר הנתונים שינה את מטרות מדעי הנתונים שלכם?</u>**

במהלך תהליך הניתוח, גילינו תובנות חדשות שהובילו אותנו להבין את המורכבות של התפרצות מחלת הכלבת. גילויים כמו מוקדים גיאוגרפיים ספציפיים שבהם יש ריכוז גבוה של מקרים, גם עלייה מובהקת באזורים עם ריכוזי חיות בר.

1. **מיקוד במוקדים גיאוגרפיים:** המידע החדש הוביל אותנו לשים דגש על זיהוי מוקדים גיאוגרפיים עם סיכון גבוה להתפרצות בעתיד. במקום להתמקד רק בניתוח כללי של מקרי הכלבת, אנו מתכוונים לפתח מודלים שיכולים לחזות התפרצויות באזורים ספציפיים.
2. **שילוב נתונים סביבתיים:** גילויים אלו גם הובילו אותנו לשקול לשלב נתונים סביבתיים, כמו תנאי מזג האוויר, כדי להבין את הקשרים בין הסביבה לשכיחות המקרים. מטרתנו היא לפתח מודלים שיכולים לחזות את השפעת התנאים הסביבתיים על התפרצות המחלה.

3. **פיתוח אסטרטגיות מניעה**: אנו שואפים לפתח אסטרטגיות מניעה ממוקדות יותר, שיכולות לכלול חיזוק המעקב והטיפול באזורים עם ריכוז גבוה של חיות בר. המטרה היא לא רק להבין את התופעה, אלא גם לפעול למניעת התפרצות מחלת הכלבת.

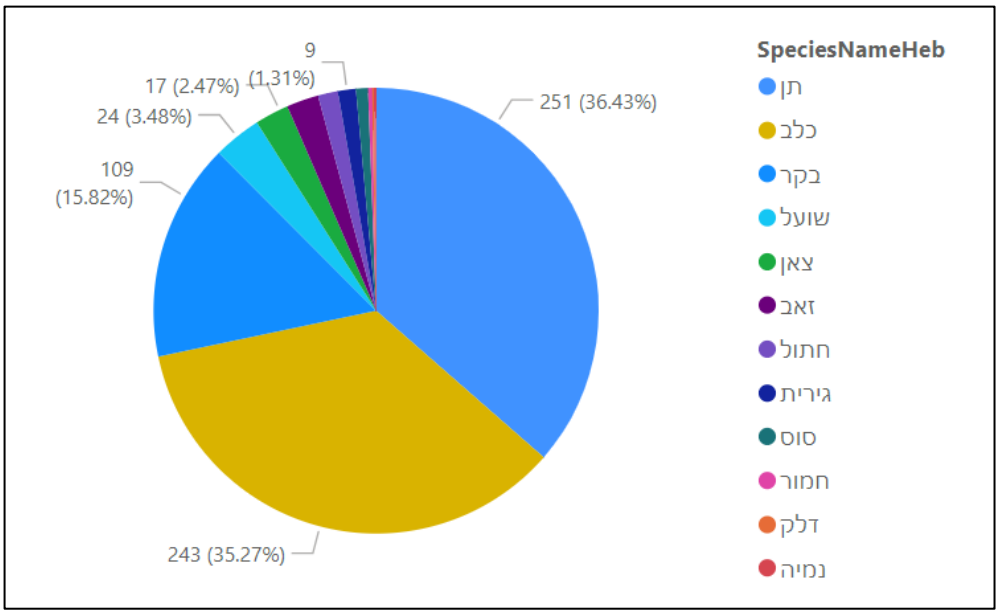
חקר הנתונים לא שינה באופן מהותי את מטרות מדעי הנתונים שלנו, אך כן שיפר את ההבנה שלנו לגבי התפרצות מחלת הכלבת, אלא גם הוביל אותנו למקד את המטרות שלנו במדעי הנתונים, כך שיתמקדו בזיהוי מוקדים גיאוגרפיים, שילוב נתונים סביבתיים ופיתוח אסטרטגיות מניעה אפקטיביות.

❖ **<u>הצגת גרפים כולל הסברים:</u>**

- **מקרי כלבת לפי מיני בעלי חיים**
  - **תרשים עוגה**

תרשים העוגה מציג את מספר מקרי הכלבת בישראל לפי מיני בעלי חיים.

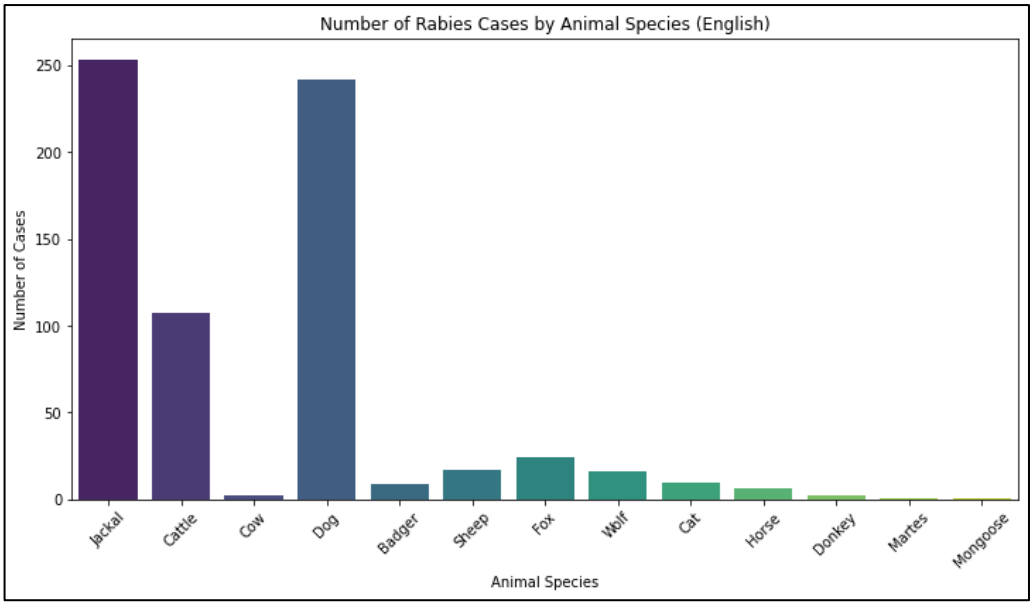
1. תנים הם המין הנפוץ ביותר, עם 251 מקרים (36.43% מהסך הכולל).
2. כלבים הם המין השני הנפוץ ביותר, עם 243 מקרים (35.27% מהסך הכולל).
3. בקר הם המין השלישי הנפוץ ביותר, עם 109 מקרים (15.82% מהסך הכולל).



  - **תרשים עמודות**

בדומה לגרף תרשים עוגה אך בגרף עמודות אשר מציג את מספר מקרי הכלבת לפי מיני בעלי חיים.

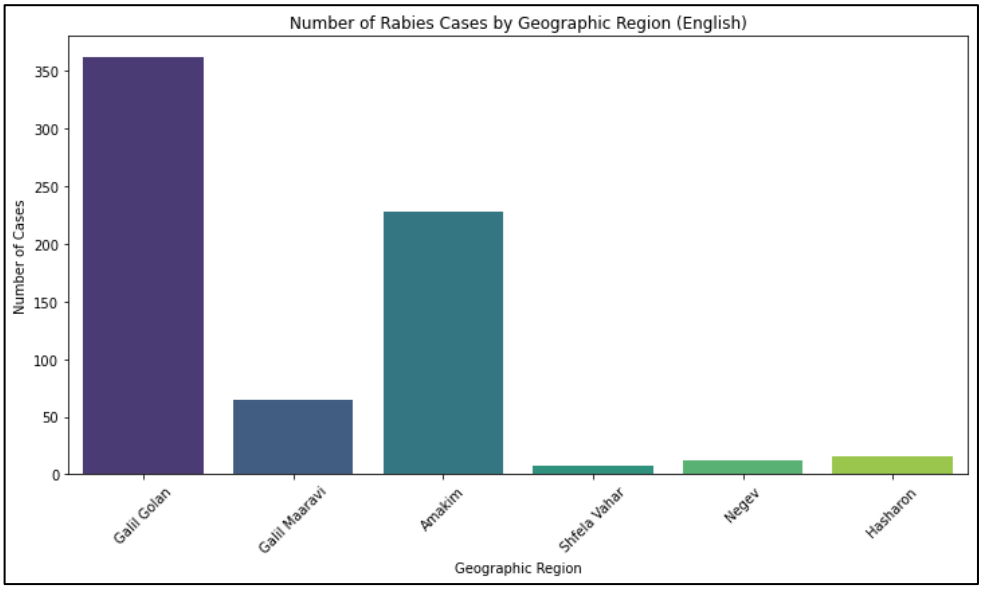
תנים הם המין עם מספר מקרי הכלבת הגבוה ביותר, ואחריהם כלבים, ואז בקר, יחסית ל3 מינים אלו, מספר מקרי הכלבת נמוך יותר עבור שאר מיני בעלי החיים.



❖ **מספר מקרי כלבת לפי אזור גיאוגרפי:**

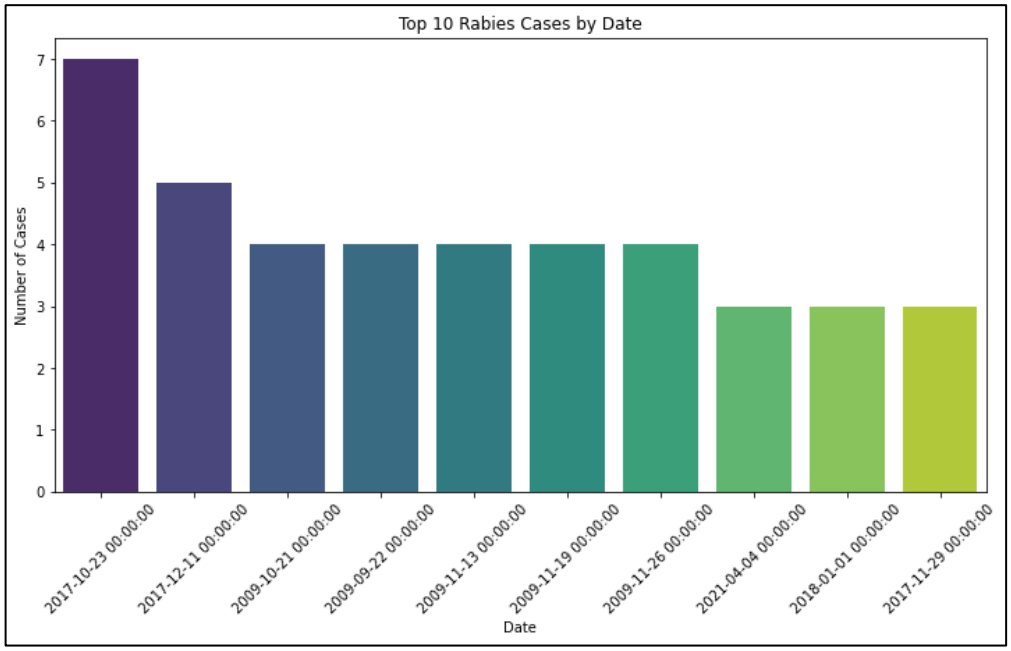
הנתונים מציגים את הפיזור הגיאוגרפי של מקרי הכלבת בישראל ומדגישים הבדלים משמעותיים בין האזורים השונים.

- גליל גולן: היה האזור עם מספר מקרי הכלבת הגבוה ביותר, מעל 350 מקרים.
- עמקים: נרשמו מעל 200 מקרים באזור זה.
- גליל מערבי: באזור זה נרשמו כ-60 מקרים.
- **אזורים אחרים**, **שפלה והר**, **נגב**, ו**השרון**, מציגים מספרים נמוכים בהרבה, דבר שיכול להעיד על פיקוח נמוך יותר, אמצעים טובים יותר למניעה או ריכוז נמוך יותר של בעלי חיים פגיעים וחשיפה למחלה.



❖ 10 התאריכים עם מספר מקרי הכלבת הגבוה ביותר:

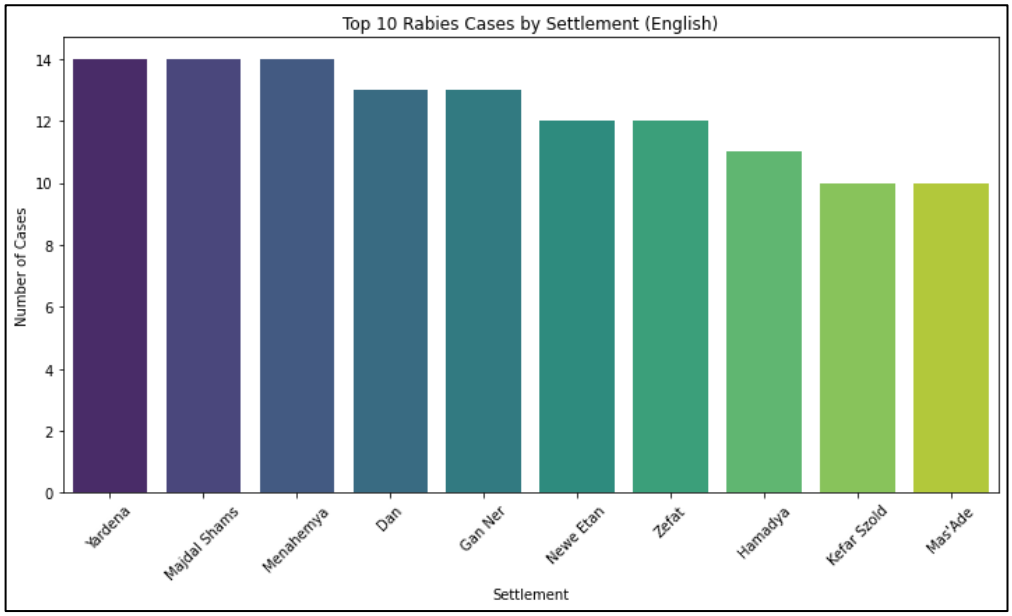
תרשים העמודות מציג את 10 התאריכים שבהם נרשם מספר מקרי הכלבת הגבוה ביותר. ייתכן כי מדובר בתקופות שבהן הייתה התפרצות כלבת מוגברת, מה שעשוי להצביע על שינויים עונתיים, אירועים מיוחדים, או גורמים סביבתיים שתורמים להדבקה. נתונים אלו יכולים לשמש בסיס למעקב וניתוח מעמיק יותר כדי לזהות מגמות או דפוסים.



❖ 10 היישובים המובילים במספר מקרי הכלבת

תרשים זה מציג את עשרת היישובים עם מספר מקרי הכלבת הגבוה ביותר.

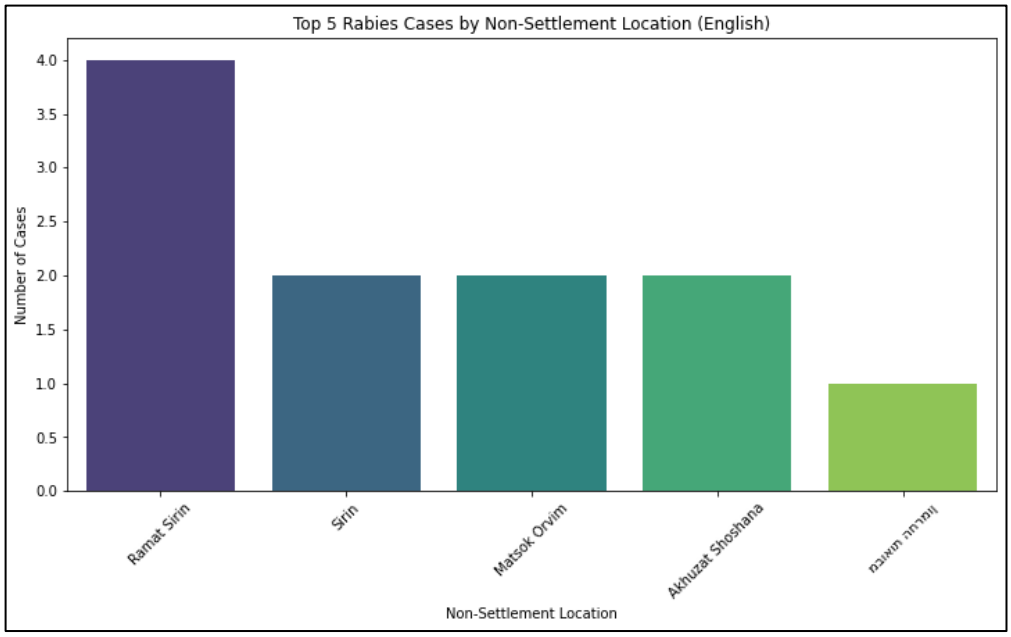
הנתונים מדגישים יישובים כמו **ירדנה** ו**מג'דל שמס** כמובילים במספר מקרי הכלבת. ייתכן שבאזורים אלו יש ריכוז גבוה של בעלי חיים או גישה מוגבלת לשירותי בריאות וטרינריים, או פיקוח מוגבר של הרשויות.



❖ 5 האתרים הלא-יישוביים המובילים במספר מקרי הכלבת

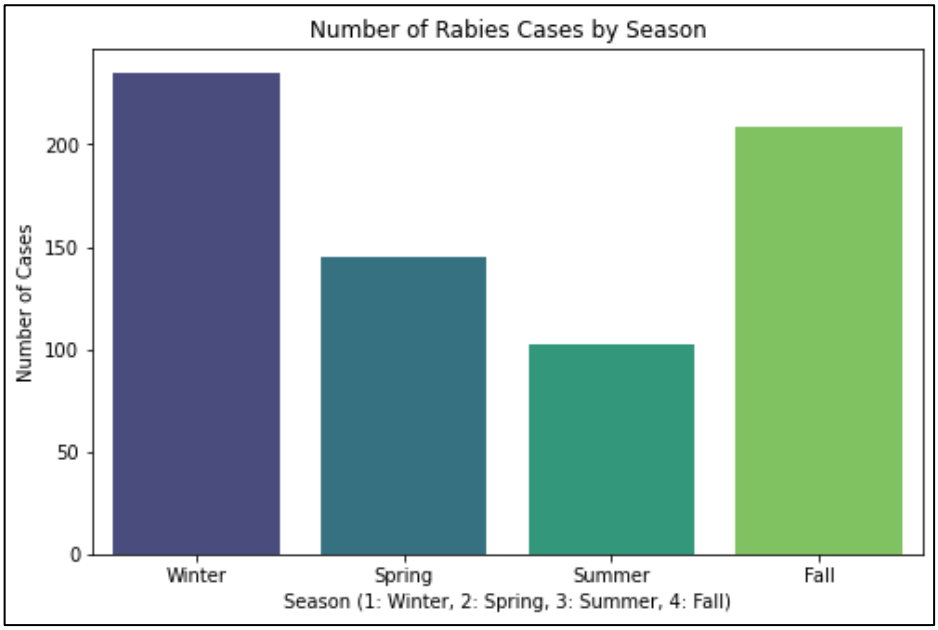
תרשים זה מציג את חמשת המיקומים הלא-יישוביים – שטח פתוח שבהם נרשם מספר מקרי הכלבת הגבוה ביותר.

האזורים מחוץ ליישובים, כמו **רמת סירין** ו**מצוק עורבים** ,עשויים להיות נקודות חמות בשל פעילות של חיות בר רבות החיות בלהקות או עדרים, או עקב יכולת נמוכה של הרשויות בפיקוח מניעת הדבקה .נתונים אלו מדגישים את הצורך בנקיטת פעולות מניעה ממוקדות גם ביישובים וגם בשטחים פתוחים.



❖ מקרי כלבת לפי עונה

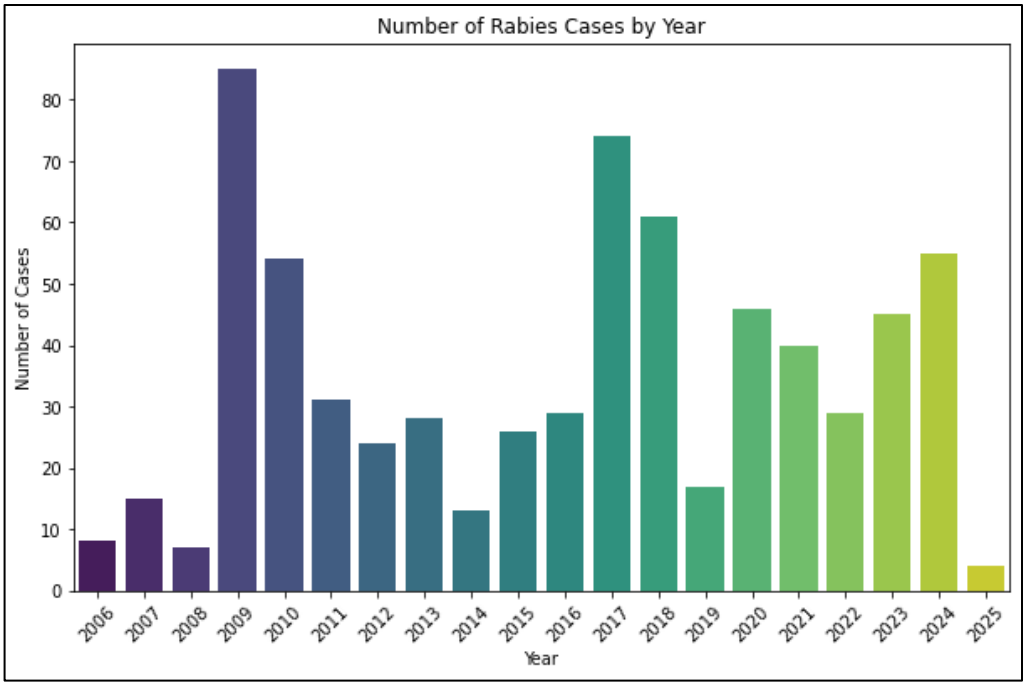
לפי תוצאות הגרף ניתן לראות שמספר מקרי הכלבת הגבוה ביותר התרחש בעונת החורף, במקום השני בעונת הסתיו, במקום השלישי בעונת האביב ואופן מפתיע במקום אחרון בעונת הקיץ.



❖ **מקרי כלבת לפי שנה**

לפי גרף זה ניתן להסיק שהשנים עם מספר מקרי הכלבת הגבוה ביותר לפי סדר יורד:

- **2009**
- **2017**
- **2010**
- **2024**



4.איכות הנתונים

במציאות לעיתים קרובות במאגרי מידע, הנתונים אינם מושלמים. למעשה, רוב הנתונים מכילים שגיאות קידוד, ערכים חסרים או סוגים אחרים של חוסר עקביות שעלולים להקשות על ניתוח הנתונים. כדי להימנע מבעיות פוטנציאליות בהמשך הפרויקט, נערוך ניתוח איכות מעמיק של הנתונים הזמינים לפני שלב עיבוד הנתונים והמודלים.

נתייחס לבעיות כגון:

❖ **נתונים חסרים:**

נתונים חסרים כוללים ערכים שהם ריקים או מקודדים כתשובות חסרות. בחלק מהעמודות, כמו LocationNotSettlementHeb, LocationNotSettlementEng, ו- Species קיימים ערכים ריקים או שבחלק מהרשומות חסרים ערכים. לא מצאנו ברשומות סימנים מוסכמים משלימים כמו NA או 999. הערכים החסרים עשויים להוביל לאי-דיוק בנתונים ולהשפיע על תוצאות הניתוח.

- ניתן להשלים את הערכים החסרים ממקורות מידע חיצוניים במידה והם זמינים, בהתאם למקורות הנתונים שציינו בתחילת המסמך. בנוסף, ניתן להשתמש בכמה שיטות לניהול נתונים חסרים, כמו השלמת הערכים החסרים לפי ממוצע או ערך שכיח, או להשמיט שורות במקרים שבהם לא ניתן להשלים את הערכים החסרים בצורה מהימנה.

❖ **שגיאות נתונים:**

שגיאות נתונים הן בדרך כלל טעויות הקלדה שנעשו בעת הזנת הנתונים. את המאגר נתונים מעדכנים באופן ידני ואנושי מה שגורם לטעויות שלרוב לא מתוקנות מיידית אלא רק בהמשך אותה שנה – לדוגמא עבור סוף שנת 2024 וחודש ינואר של שנת 2025 האחראי לעדכון המאגר עוד לא תיקן את שגיאותיו. לדוגמא בשדות כמו RegionHeb ו- RegionEng עשויה להיות אי-עקביות בין איות השמות בשפות עברית ואנגלית, דבר שעלול לגרום לחוסר התאמה בין הנתונים. דוגמא נוספת - בעמודת Species התרחשה שגיאת הזנת נתונים, בחלק מהרשומות מצוין זן בעל החיים בעברית/אנגלית ובחלק מהרשומות מצוין הנגיף של מחלת הכלבת. זו בעיה של שגיאות נתונים וגם בעיית מטא דאטה.

- יש לבצע בדיקות של שלמות הנתונים ונכונותם ולתקן בהתאם לשאר הפרטים באותה רשומה ובהתאם למקוות המידע המשלימים ככל הניתן. ניתן לעשות תהליך סטנדרטיזציה של שמות האזורים כך שיהיו אחידים בשני השדות, וכן לוודא אחידות בקידוד המונח, או להעדיף שפה מסוימת לעומת שפה אחרת בהתאם לדרישות.

❖ **שגיאות מדידה:**

שגיאות מדידה כוללות נתונים שהוזנו נכון אך מבוססים על שיטת מדידה שגויה.

לא ראינו בנתונים שגיאות מדידה לפי שיטה שגויה באופן וודאי. אולי ברישום נתוני הקואורדינטות (x, y) יש שגיאה בשיטת המדידה מה שגורם לקיום ערכים חריגים שאינם תואמים מיקומים גיאוגרפיים תקינים, דבר שיכול להשפיע על איכות המידע.

- יש לבצע בדיקה של שיטות המדידה וניתוח סטטיסטי לזיהוי ערכים חריגים ולבצע תיקון של נתונים אלו על ידי הצלבה עם מקורות חיצוניים, כמו מפות גיאוגרפיות, לצורך ווידוא דיוקם.

❖ **אי התאמות וחוסר עקביות בקידוד**:

אי-התאמות בקידוד כוללות לרוב שימוש ביחידות מדידה לא סטנדרטיות או חוסר עקביות בערכים. (כמו השימוש גם ב "נ" וגם ב"נקבה" לציון מגדר כדוגמא כללית לא עבור הנתונים שלנו) שימוש במונחים לא סטנדרטיים או חוסר עקביות בין יחידות מדידה עלול לגרום לבעיות בהשוואת נתונים, כפי שציינו מקודם ישנן מספר עמודות שמופיע בהן חוסר עקביות בערכים כגון עמודת Species בסוג הערכים ואיות סוג הזן, בנוסף לעמודות הכפולות מבחינת שפה

(עברית/אנגלית) , עמודות המציינות את שם היישוב/אזור/בע"ח ולפעמים האיות עבור חלק מהשמות בוצע באופן שונה, בחוסר עקביות.

- ניתן לבצע בדיקת נוספת עבור יחידות מדידה לא סטנדרטיות למרות שבבדיקה ראשונית לא מצאנו שיש שימוש כזה במאגר נתונים שלנו, ניתן ליצור מערכת סטנדרטית לקידוד הנתונים (כגון, אחידות בשמות אזורים או בעלי חיים) מה שיסייע בשמירה על עקביות המידע. ניתן להחיל מערכת אוטומטית שתתקן את אי העקביות בהזנה.

❖ **מטה-דאטה (נתונים) שגויים**:

מטא-נתונים שגויים כוללים אי-התאמה בין המשמעות הגלויה של שדה לבין המשמעות המוצהרת בשם השדה או בהגדרתו. אי התאמה בין משמעות השדה הנראה לעין לבין המשמעות המתוארת בשמו או בהגדרתו יכולה להוביל לבלבול ופגיעה במידע. חלק מהשדות והעמודות שחשודות בשגיאת מטה-דאטה לא טובה, זה Species כפי שהסברנו מקודם, בנוסף למספר עמודות המכילות קישורים אשר בפועל חלקם חסומים או לא עובדים, ורק חלקם פועלים כפי שמצופה מהם. בנוסף גם הבנת עמודת GlobalID והערכים שהיא מכילה זה לא מובן מאליו, אין התאמה מיידית בין הגדרת השם לבין הערכים בפועל, דבר מבלבל ומטעה.

- יש לערוך סקירה מעמיקה של המטה-דאטה כדי לוודא שכל שדה מתאר את המידע בצורה מדויקת. ניתן לבצע תקנון והגדרת שמות שדות בצורה ברורה ומובנת, כך שיהיה תיאום מוחלט בין המשמעות של השדה והשמות שהוקצו לו.

לסיכום, על ידי טיפול נכון בבעיות בנושא איכות ושלמות הנתונים כגון נתונים חסרים, תיקון שגיאות קידוד ויישום שיטות לניהול שגיאות מדידה, ויישום ההמלצות המצורפות כל אלו יסייעו לשפר את איכות הנתונים, להבטיח עקביות ודיוק, ולשפר את עיבוד הנתונים אמינות הניתוחים והמודלים.