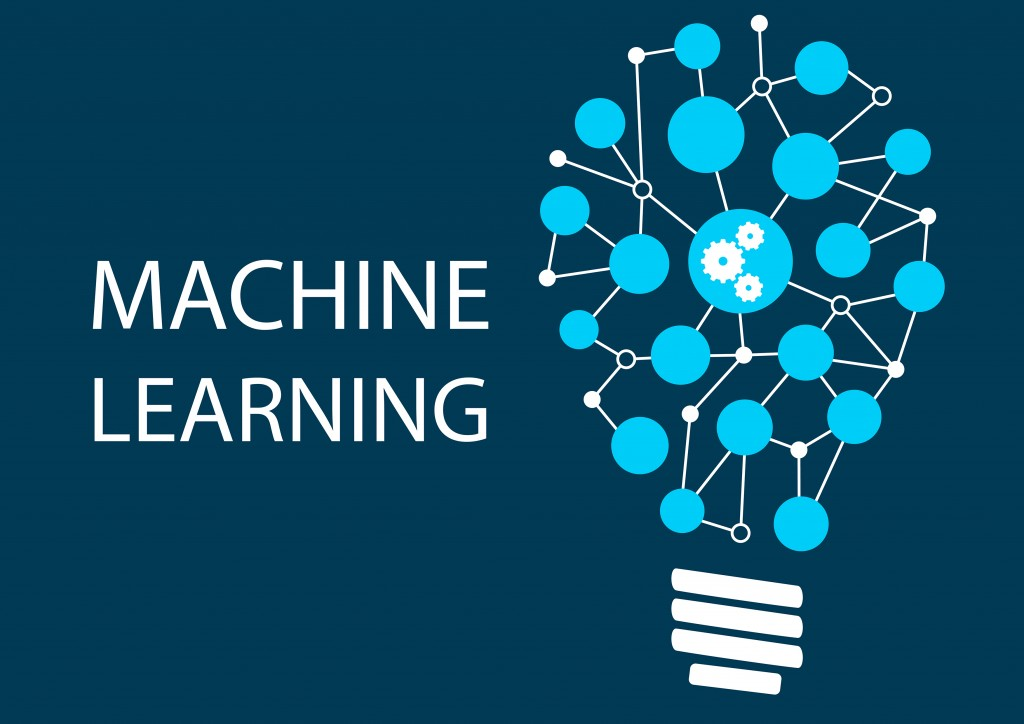


**פרויקט בלימוד מכונה**

**חלק א'**



מגישים:

204097737

305719221

203625355

**מרצה הקורס:** בעז לרנר

**מתרגל הקורס:** בן חדד

תוכן עניינים

[הגדרת הבעיה 2](#_Toc7278230)

[1.תיאור כללי של עולם התוכן הנחקר: 2](#_Toc7278231)

[2.שאלת המחקר: 3](#_Toc7278232)

[הבנת הנתונים: 3](#_Toc7278233)

[1.תיעוד מקורות הנתונים ומשמעותם 3](#_Toc7278234)

[2.הסתברויות אפריוריות וקשרים בין מאפיינים 4](#_Toc7278235)

[3.איכות הנתונים 9](#_Toc7278236)

[הכנת הנתונים 10](#_Toc7278237)

[Feature Selection 10](#_Toc7278238)

[טיפול פרטני במאפיניים 11](#_Toc7278239)

[הכנת הנתונים לאימון ובניית מערכת לומדת 12](#_Toc7278240)

[נספחים 14](#_Toc7278241)

[נספח 1 :טבלאות שכיחות והסתברויות בייסניות למציאת הסתברות אפריורית 14](#_Toc7278242)

# הגדרת הבעיה

## 1.תיאור כללי של עולם התוכן הנחקר:

בפרויקט זה בוצע חקר על תחום אימוץ כלבים וחתולים על פי נתונים שהתקבלו.

בפרויקט אנו מתמודדים עם בעיה מחקרית מעולם לימוד המכונה מסוג Classification. בתוך עולם בעיות זה אנו נמצאים בתת תחום של Supervised Learning בו ידועים לנו ערכי המשתנה המוסבר אשר הוגדרו בהוראות העבודה.

בבדיקה שבוצעה נמצא מחקר אשר עסק בבעיה דומה אך השתמש בכלים אחרים . המחקרמאונברסיטת קליפורניה בשם Prediction of Adoption. (ב[נספח מאמרים](#_4.2))

במחקר זה נבדקו 4813 חתולים ו3303 כלבים ונבחנו משתנים שונים והשפעתם על הסיכוי לאימוץ של הכלבים והחתולים שנבדקו. במחקר זה בוצעה רגרסיה לוגיסטית מרובה על מספר רב של משתנים. בפרויקט אותו נבצע, בניגוד למחקר זה, אנו לא מעוניינים רק בניבוי בינארי האם החיה אומצה או לא, אלא בניבוי מורכב יותר המחלק ל-3 מחלקות על פי מהירות אימוץ החיה.

## 2.שאלת המחקר:

בפרויקט נרצה להגיע לדיוק מרבי בניבוי האם יבוצע אימוץ של כלבים וחתולים ובקצב האימוץ בהם יאומצו. נרצה לבדוק האם אומצו ביום בו נרשמו לאימוץ, בשלושת החודשים העוקבים לרישום, או שלא נלקחו לאימוץ במהלך שלושת החודשים הראשונים לאחר הרשמתם. מענה על שאלת מחקר זו יעזור לארגונים המבצעים את התהליך להשקיע משאבים בקידום הכלבים והחתולים שאינם בסבירות גבוהה לאימוץ ולחסוך משאבים על קידום כלבים וחתולים בעלי סבירות גבוהה לאימוץ מוקדם.

# הבנת הנתונים:

## 1.תיעוד מקורות הנתונים ומשמעותם

הנתונים בפרויקט נלקחו מאתר petFinder.my, אתר המשמש כפלטפורמה המובילה של מלזיה לרווחת בעלי חיים מאז 2008 עם מסד נתונים של יותר מ150,000 בעלי חיים.

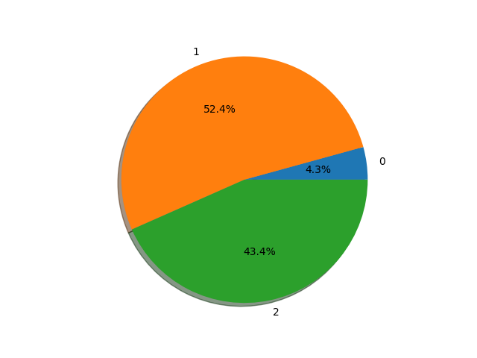
באתר זה, מפורסמים בעלי החיים הזקוקים לאימוץ באמצעות פרופילים מקוונים , כאשר לכל בעל חיים פירוט בנוגע לסוג בעל החיים ומידע נוסף אודות בעל החיים עצמו כגון גיל, צבע ומצב בריאותי.

סביר כי חלק מהנתונים עבור כל בעל חיים הוקלד בצורה ידנית על ידי אדם שאינו מומחה, על פי כן סביר כי בנתונים טעויות הנובעות מטעויות הקלדה וכן טעויות שנגרמו על ידי אפיון בעל החיים על ידי בן אדם שאינו מומחה.

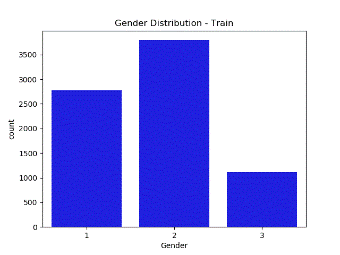
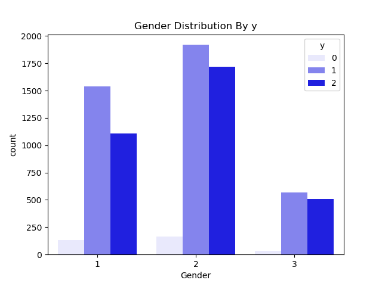
|  |  |  |  |
| --- | --- | --- | --- |
| שם המשתנה | תיאור | טיפוס | מרחב ערכים |
| Type | סוג בעל החיים | קטגוריאלי | 1=כלב  2=חתול |
| Age | גיל בעל החיים בחודשים | בדיד | 0-255 |
| Breed1 | גזע עיקרי | קטגוריאלי | 0-307 |
| Breed2 | גזע משני | קטגוריאלי | 0-307  0=לא קיים |
| Gender | מין בעל החיים | קטגוריאלי | 1=זכר  2=נקבה  3=מעורב |
| Color1 | צבע מס' 1 | קטגוריאלי | 1-7 |
| Color2 | צבע מס' 2 | קטגוריאלי | 0-7  0=לא קיים |
| Color3 | צבע מס' 3 | קטגוריאלי | 0-7  0=לא קיים |
| MaturitySize | גודל בבגרות | קטגוריאלי | 1-4  1=קטן  4=גדול מאוד |
| FurLength | אורך הפרווה | קטגוריאלי | 1-3  1=קצר  3=ארוך |
| Vaccinated | האם בעל החיים חוסן | קטגוריאלי | 1=כן  2=לא  3=לא ידוע |
| Dewormed | האם בעל החיים עבר טיפול נגד תולעים | קטגוריאלי | 1=כן  2=לא  3=לא ידוע |
| Sterilized | האם עבר סירוס/עיקור | קטגוריאלי | 1=כן  2=לא  3=לא ידוע |
| Health | מצב בריאותי | קטגוריאלי | 1=בריא  2=פציעה מינורית  3=פציעה משמעותית |
| Quantity | מספר החיות המוצגות בפרופיל | בדיד אורדינאלי | 1-20 |
| Fee | מחיר האימוץ | רציף | 0-1000  0=חינם |
| State | מדינה במלזיה | קטגוריאלי | 41324-41415 |
| VideoAmt | כמות סרטוני וידאו של בעל החיים שהועלו לפרופיל | בדיד אורדינאלי | 0-8 |
| PhotoAmt | כמות התמונות של בעל החיים שהועלו לפרופיל | בדיד אורדינאלי | 0-30 |
| Y | משתנה המטרה על מהירות האימוץ | קטגוריאלי | 0-2 |

## 2.הסתברויות אפריוריות וקשרים בין מאפיינים

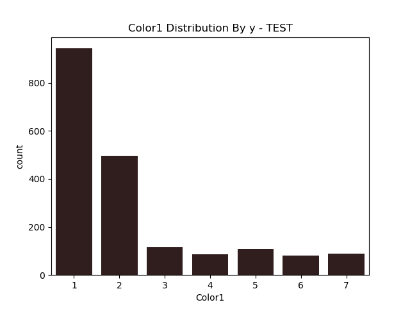
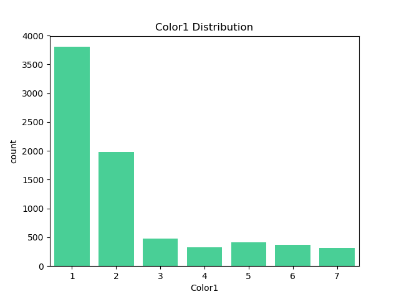
בחרנו להשתמש בהסתברויות הבייסניות כהסתברויות האפריוריות של משתנה המטרה.ניתן לראות ב[(נספח)](#_נספח_1_:טבלאות) את טבלאות השכיחות ומציאת ההסתברויות האפריוריות המותנות לחלק מן המשתנים.

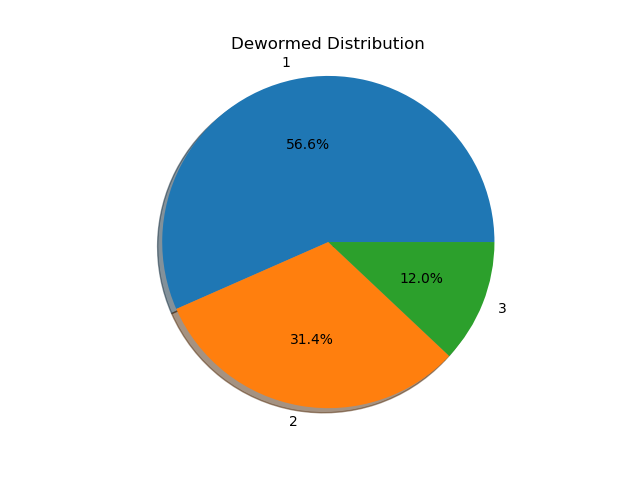
****

בסט האימון ניתן לראות כי התפלגות המשתנה המוסבר אינה אחידה מה שמרמז על סט לא מאוזן ויש לטפל בכך בהמשך.ניתן לראות כי ההסתברות האפריורית לבעל חיים להיות מאומץ באותו היום ( (Y=0הינה נמוכה מאוד. לעומת זאת החלוקה בין 1 ו-2 פחות או יותר מאוזנת.

****

ניתן לראות שבשינוי המינים אין הבדל מאוד מובהק במהירויות האימוץ. ב"לא ידוע" ניתן לראות כי ההסתברות לy=0 קטנה מאוד. זה מסתדר עם ההגיון. ייתכן וחלק מהחיות שמגיעות לאימוץ בתחילה מוגדרות כמין לא ידוע ותשמר בבסיס הנתונים כך.אך כעבור מס' ימים מתגלה המין.לכן יתכן וזה מסביר את כך שחיות עם מין לא ידוע אינן מאומצות ביומן הראשון.

****

****עבור המשתנה צבע ניתן לראות כי חלק גדול מהמופעים הם צבע שחור. הדבר הגיוני מבחינת המציאות (שכיח ב2 סוגי החיות) ניתן לראות שזה גם מתקיים בסט הבחינה הסופי.עבור הצבעים הנוספים – פירוט ב[נספח צבעים](#_1.17_colors)

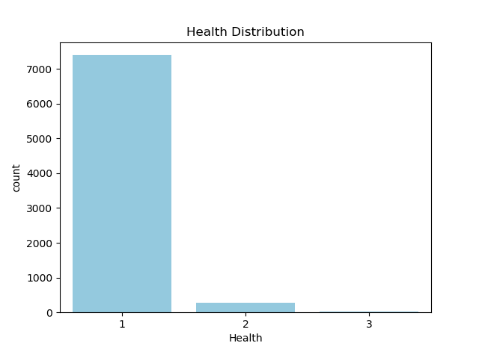
****

עבור טיפול נגד תולעים נראה כי רוב בעלי החיים אכן עברו טיפול. גם כאן ניתן לראות כי יש נתח יחסית גדול של חיות אשר לא ידוע אם עברו טיפול או לא.

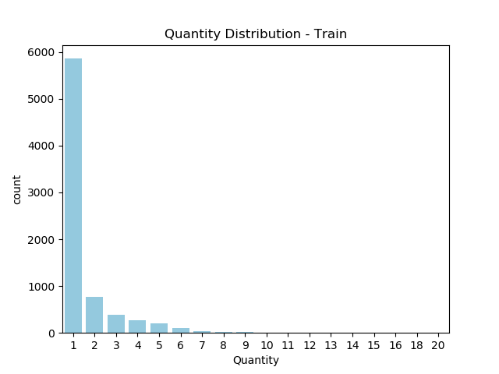
אין אינדיקציה חזקה להשפעה על משתנה המטרה. ([נספח](#_1.11))

****

במשתנה זה ניתן להבחין כי אין איזון באופן גורף. כ- 5\4 מהתצפיות משתייכות לזן שלא קיים או מעורב (0 או 307). אנו משערים שמשתנה זה מוסיף רעש גדול לפונקציית המטרה, על אחת כמה וכמה שגם אם קיים זן (לא 0 או 370) הוא מוגדר באופן "זן שני".

****

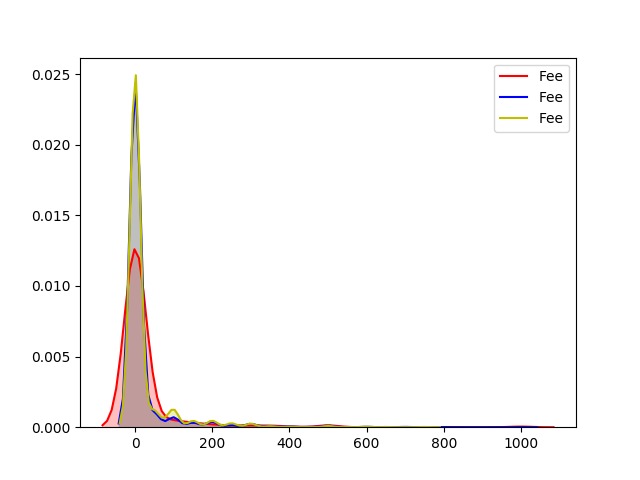
ניתן לראות כי עבור התפלגות המצב הבריאותי קיים חוסר איזון מוחלט. יש לזכור גם שמשתנה זה הוא מאוד סובייקטיבי שכן בעוד שאחד יכול להגדיר חיה כבריאה האחר יכול לזהות אותה דווקא כפצועה\חולה. כלומר אין כאן קו הפרדה ברור. הדבר יכול לרמוז על רעש.

****

ניתן לראות כי התפלגות מספר החיות המוצגות בפרופיל היא בעלת זנב ימני ארוך. עובדה זו מתקיימת גם בגלל שהכמות חסומה מלמטה ובד"כ חיה אחת המוצגת בפרופיל עשויה להספיק. מעבר לכך נראה כי יותר מ10 חיות המוצגות בפרופיל היא תופעה נדירה. בעתיד נבצע ניתוח של איחוד קטגוריות .

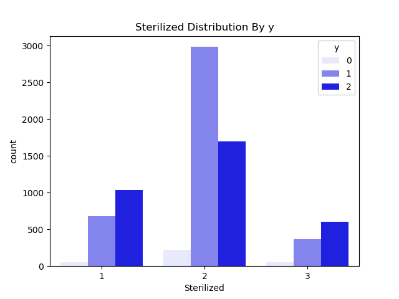
****

אדום – y=0, כחול y=1, צהוב, y=2

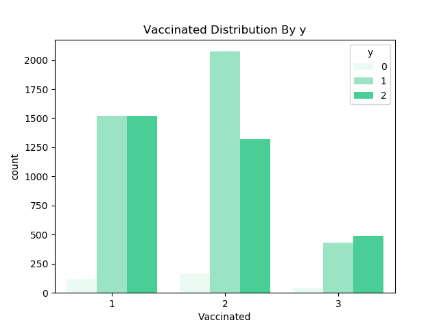
*ניתן לראות כי ההתפלגויות המותנות בכל הקטגוריות לא משתנות מהותית. מרכז הפעמון קרוב ל-0 וזהו גם הערך השכיח בכל אחת מן ההתפלגויות. ניתן לראות כי קיים זנב ימני ארוך בגלל הסכומים הגבוהים מאוד המופיעים במעט מאוד מקרים. בעתיד יישקל ביצוע התמרה.*

****

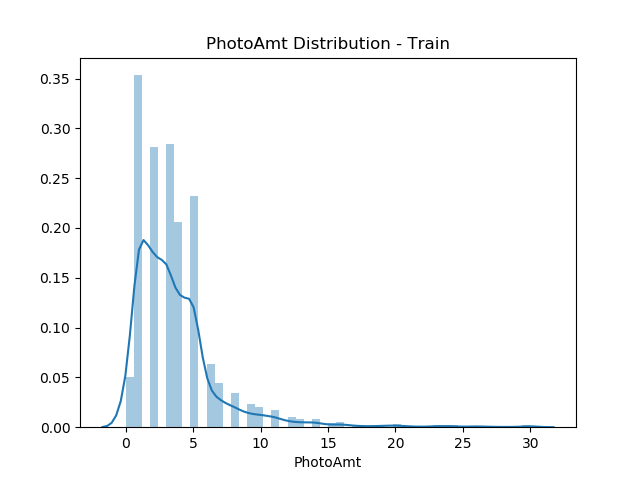
כיוון שמשתנה זה קטגוריאלי ומקבל מס' רב של  ערכים בחרנו כאן להציג את ה10 השכיחים ביותר. ניתן לראות שיותר משליש מהמופעים מוגדרים כ"מעורבים" בזנם הראשון. הדבר מתיישב עם ההגיון אך מתעצם כשמדובר במדגם ממלזיה. כיוון שרוב חיות הבית מוחזקות לא לשם מותרות.

**משתנים עם חשד להשפעה על פונקציית המטרה:**

****

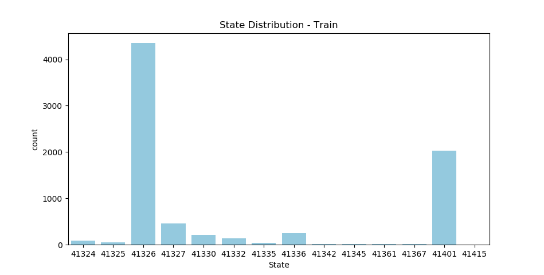
****עבור בעל חיים שעבר סירוס או עיקור ניתן לראות כי ההתפלגויות המותנות במשתנה המטרה משתנות. לדוגמא עבור חיות שעברו סירוס\עיקור אשר רובן לא נמסרו תוך 3 חודשים. בתחילה זה נראה סותר את ההגיון כיוון שניתן לחשוב כי אנשים ינטו לחיות מעוקרות. אך יש לזכור שחלק מהאנשים דווקא כן ירצו להרבות את חיית המחמד שלהם. לכן זהו קשר מעט מפתיע.

****

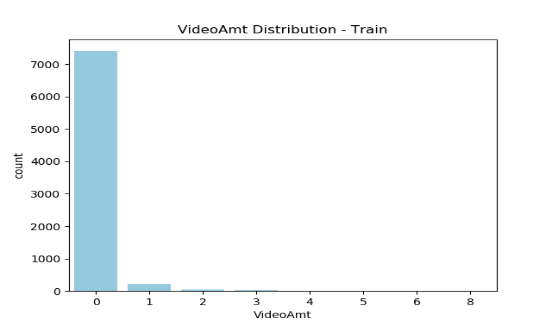
 עבור המשתנה האם בעל החיים חוסן ניתן לראות שיש קשר למשתנה המסובר. התפלגות ערכי הY שונה בכל אחת מן הקבוצות – ולכן אנו מניחים כי השתייכות לקטגוריה מסויימת אכן משפיעה על מהירות האימוץ. את ההסתברויות המתונות האפריוריות ניתן לראות ב([נספח](#_1.9))

****

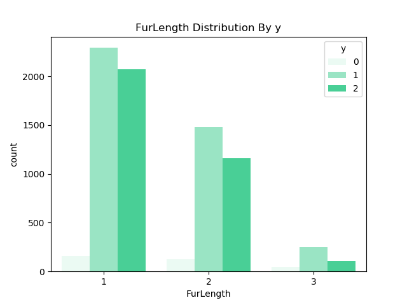
כמות התמונות הוא משתנה חשוב בעינינו. כמו בכל רכישה\אימוץ באינטרנט חשוב לראות וויזואלית את המוצר לפני. באיור ניתן לראות היסטוגרמה כללית עבור כמות התמונות. ניתן לראות כי ההתפלגות בצורת פעמון עם זנב ימני.

****

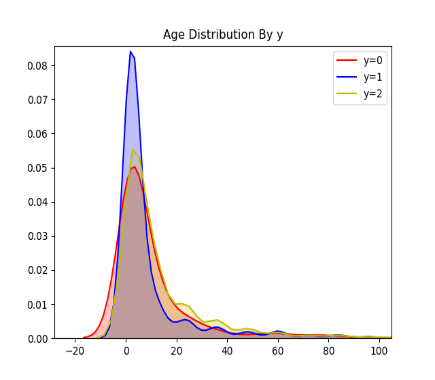
ניתן לראות שחלק גדול מן התצפיות מגיע ממחוז Selangor וממחוז Kuala Lumpur. הדבר מתיישב עם ההגיון. Selangor הוא המחוז עם הכי הרבה אוכלוסיה בפער. Kuala Lumpur הוא המחוז עם התל"ג הגבוה ביותר במדינה. בעתיד נשקול להכניס משתנים אלו ולבחון השפעה. (ניתן לראות בנספח את המפה ואת הקוד הממחישים את הכמויות)

****

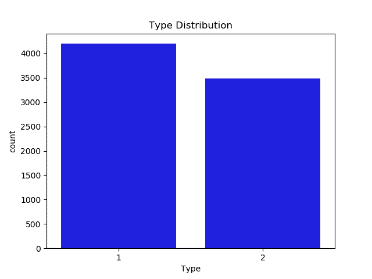
יותר מ95% מהתפציות הן ללא וידאו. זהו משתנה לא מאוזן כלל. בעתיד יתכן ויישקל פתרון דיסקיטיזציה בינארית או איחודים אחרים.

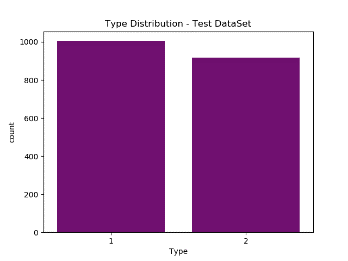
****

ניתן לראות שלאורך הפרווה אכן השפעה על מהירות האימוץ. כך לדוגמא לפי ה([נספח](#_1.8)) ניתן לראות שיותר מ11% מתוך החיות בעלות פרווה ארוכה אומצו ביום הראשון. עם זאת – יש לשים לב כי הסט לא מאוזן. ישנם הרבה יותר מופעים עם פרווה קצרה ממופעים עם פרווה ארוכה.

****

על פי ההסטוגרמות המותנות ניתן לראות כי קיים הבדל קטן בצפיפיות של ערכי הגיל. מדובר בהתפלגות פעמון עם זנב ימני - זאת גם מכיוון שערכי המשתנה חסומים מלמטה (ע"י 0).ניתן לראות כי קיימות מעט תצפיות עם ערכי גיל גבוהים.בעתיד נשקול לבצע התמרות שונות למשתנה זה כדי לקבל השפעה גבוהה יותר.אנו משערים כי קיימת השפעה על סמך נסיון אישי בלבד.

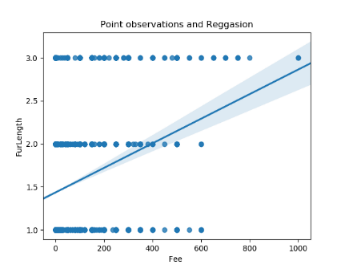
****

סוג בעל החיים הוא גורם משמעותי באימוץ. ניתן לראות כי ישנם יותר כלבים המגיעים לאימוץ. בנוסף – עפ"י ההסתברויות האפריוריות המותנות – ניתן להסיק עפ"י סט האימון כי לחתולים יש נטייה להיות מאומצים מהר יותר. ([נספח](#_1.1_סוג_החיה))

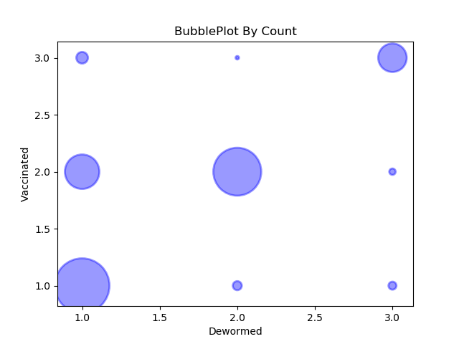
****

עבור משתנה זה ניתן לראטות קשר למוסבר באמצעות הסתברויות אפריוריות מותנות. פרטים נוספים ב[נספח](#_1.7).

**קשרים מעניינים בין משתנים:**

קשר בין מחיר ואורך פרווה:

ניתן לראות כי על פי הנתונים קיים קשר חיובי בין מחיר ואורך הפרווה. ייתכן וזו הטייה מדגמית בגלל התצפית החריגה(1000), אך ניתן לראות שביחס לכמות החיות עם שיער ארוך ישנן לא מעט חיות שמאומצות במחיר גבוה.

קשר בין האם חוסן לבין האם עבר טיפול בתולעים:

קשר זה הוא קשר חיובי ומתבקש. הוא נובע לרוב מהעובדה ההגיונית שקיים סיכוי גבוה שכלב שמגיע לעבור חיסון יעבור באותה העת גם טיפול תולעים. גודל הבועה פרופורציוני למס' המופעים, ולכן ניתן לראות כי: עבור חיות שלא חוסנו – רובן לא עברו טיפול. עבור חיות שכן חוסנו – רובן עברו טיפול. עבור חיות שלא ידוע אם עברו חיסון – עבור רובן לא ידוע אם טופלו.

## 3.איכות הנתונים

סט הנתונים שהתקבל, בצורתו הגולמית, מכיל ערכים חסרים ב-3 משתנים מסבירים, וכן מידע לא ידוע במספר משתנים נוספים. למשל משתנה "מחוסן" כאשר הערך שלו שווה ל-3 לא ניתן לדעת האם בעל החיים מחוסן או לא. להלן שיעור הערכים ה**חסרים** בכל אחת מן העמודות;

|  |  |  |
| --- | --- | --- |
| **משתנה** | **שכיחות** | **פרופורציית המידע החסר** |
| Breed2 | 32 | 0.00416 |
| Color2 | 15 | 0.00195 |
| Color3 | 39 | 0.00507 |

עבור משתנים אשר ערכם "**אינו ידוע**";

|  |  |  |
| --- | --- | --- |
| **משתנה** | **שכיחות** | **פרופורציית המידע החסר** |
| MaturitySize | 15 | 0.00195 |
| FurLength | 400 | 0.05205 |
| Vaccinated | 968 | 0.12597 |
| Dewormed | 924 | 0.12024 |
| Sterilized | 1016 | 0.13222 |
| Health | 16 | 0.00208 |

**טיפול בערכים חסרים:**

Breed2- במקרה של רשומות אלו, הוחלט לתת להם את הערך 0 כיוון שההנחה היא שטעויות אלו נגרמו מטעויות הקלדה וכוונתן הייתה שלא קיים גזע שני. מבדיקה ראינו כי הרבה ממופעים אלו מקבלים את הערך 307 (מעורב) בBreed1. כך שממילא אין משמעות לגזע השני – מה שמחזק את טענתינו.

Color2-ישנן 15 רשומות עם ערך חסר, ברשומות אלו ערכם של Color3 הוא 0, כלומר לבעל החיים אין צבע נוסף,לכן נבחר להחליף את ערכי המשתנה ברשומות אלו ב0, כלומר 'לא קיים'.

Color3- ערך משתנה זה מעיד על צבע נוסף, גם במקרה זה נבחר להחליף את הרשומות בהן הערך לא קיים, בערך 0. ניתן לראות כי רבים ממופעים אלו גם מקבלים את הערך 0 בColor2 מה שמעלה את הסבירות שאכן מדובר בטעויות אנוש.

**טיפול בנתונים עם ערך 'לא ידוע':**

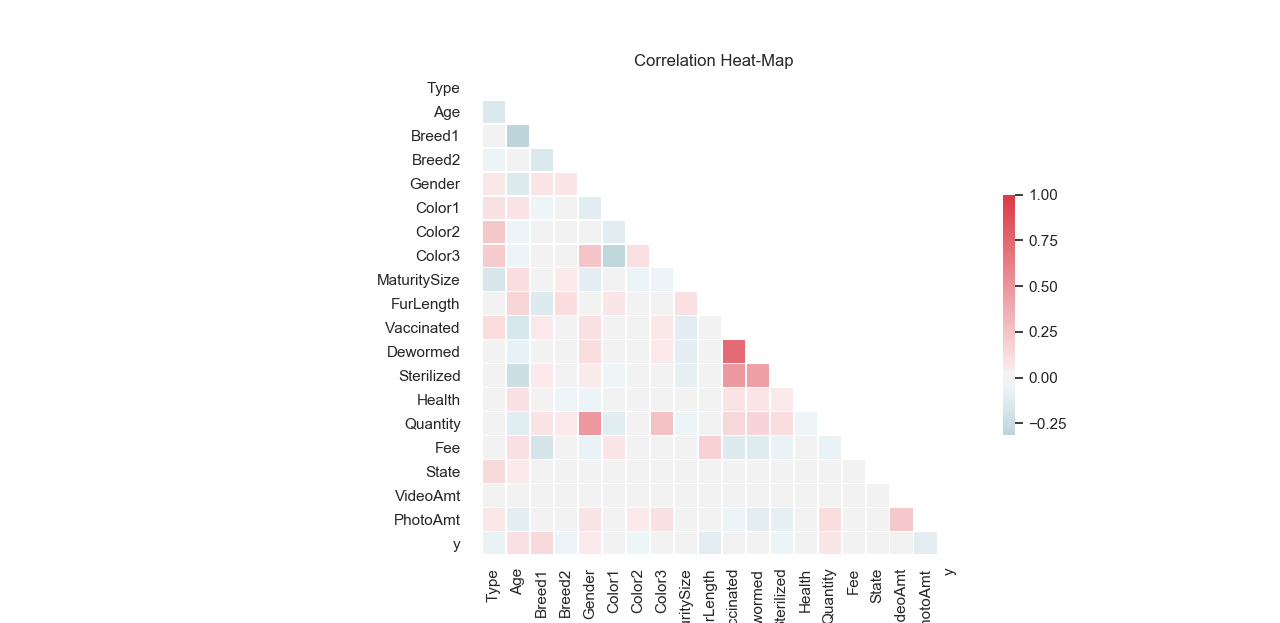
לאחר בחינת הנתונים שלהם קיים ערך לא ידוע, תוך התחשבות בפרופורציה ובהתפלגות שאר הערכים עבור המשתנים, הוחלט לתת לכל משתנה עם ערך שכיח מובהק את הערך השכיח ביותר. הפרמטר היחיד המתאים לדרישה זו הוא המשתנה ' MaturitySize' אשר מעל 70% מהרשומות מקבלות את הערך 2 ולכן הוחלט להחליף את הערך 'לא ידוע' בערך 2. בשאר המקרים הוחלט להתייחס לקטגוריה "אינו ידוע" כערך בעל חשיבות ולכן יושארו משתנים אלו כפי שהתקבלו במאגר הנתונים.

**בנוסף**, לאחר מעבר על הנתונים נמצא כי ישנם 4 מופעים של משתנה Breed1 אשר מקבלים את הערך 0, כאשר למשתנה Breed2 ערך שונה מ-0.

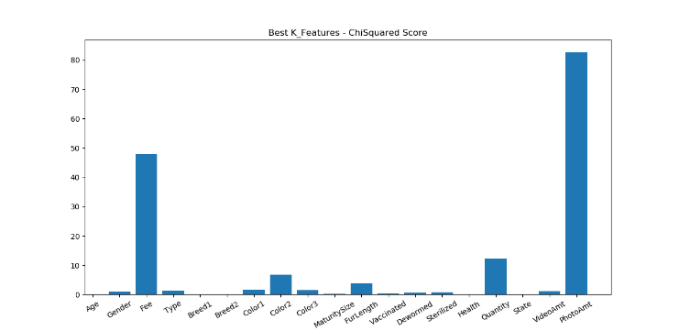
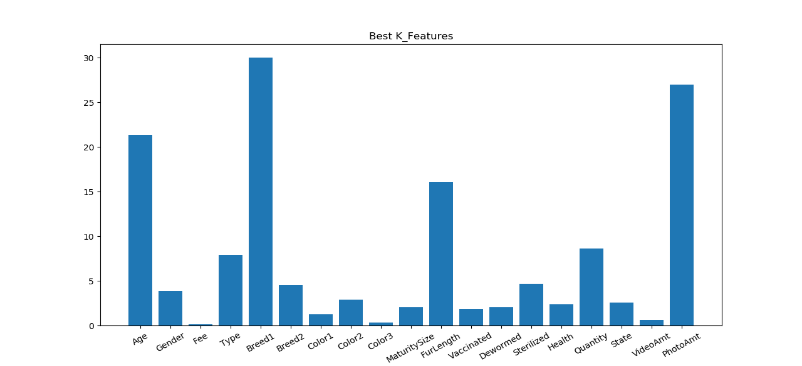
בסתירה להיגיון, לא יתכן כי בעל חיים יהיה חסר גזע ראשי ובעל גזע משני. לכן הוחלט להחליף את הערכים המופיעים בBreed2 בכל מופע של Breed1 המקבל ערך 0 ולתת לBreed2 ערך 0. חשוב לציין ששינוים אלו יבוצעו גם על סט נתוני הבחינה.

# הכנת הנתונים

## Feature Selection

לצורך קבלת מושג כללי אודות טיב המשתנים שהתקבלו הוחלט בתחילה לבצע אלגוריתם בתכנה הנותן ניקוד לכל משתנה(פעם לפי התפלגות F ופעם לפי כי בריבוע).

לאחר מכן חושבה מטריצת קורלציות העשוייה לנבא קשרים ליניאריים בין מאפניים. ניתן לראות עפ"י מטריצת הקורלציות כי אין משתנה שמשמפיע באופן מאוד מובהק על משתנה המטרה.מטריצת קורלציות ב[נספח](#_3.1).

מנתונים ומבחנים אלו , נבחר להסיר בשלב זה את משתנה Color3 בלבד. בהמשך, כתלות במשתנים נוספים שיגזרו מהמשתנים הקיימים תישקל הסרת משתנים קיימים בינהם Breed2 , VideoAmt ועוד (יפורט בהמשך).

ניתן בנוסף לראות כי קיימת קורלציה גבוהה בין Dewormed ומשתנה Vaccinated כפי שהוצג בסעיפים הקודמים. בשלב זה על אף המתאם הגבוה (0.73) הוחלט שלא להסיר את אחד מהם כיוון שכרגע אנו רואים חשיבות ל2 המשתנים ביחד ולא ניתן להחליט באופן חד משמעי שאחד מהם משפיע על המוסבר בצורה משמעותית יותר מהאחר.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Vaccinated | Dewormed | Sterilized | Health | Quantity | Fee | State | VideoAmt | PhotoAmt | y |
| -0.03 | 0.01 | -0.05 | 0.03 | 0.07 | 0.01 | 0.03 | -0.02 | -0.09 | 1.00 |

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Type | Age | Breed1 | Breed2 | Gender | Color1 | Color2 | Color3 | MaturitySize | FurLength | |
| y | -0.07 | 0.10 | 0.13 | -0.05 | 0.05 | -0.02 | -0.04 | -0.01 | 0.03 | | -0.09 |

## טיפול פרטני במאפיניים

****

עקב סט לא מאוזן, עבור המשתנה נבנה משתנה חדש:

****

*עבור משתנים המינים נבנה משתנה חדש  עפ"י החוקיות הבאה:*

*If Breed1 Not 307 And Breed2 Is 0 🡪 IsPure = 1*

*Else 🡪 IsPure = 0*

*ממשתנה זה התקבלו יותר תצפיות על חיות שIsPure=0 . ניתן לראות ב[נספח](#_2.5) ו[בנספח](#_2.6).*

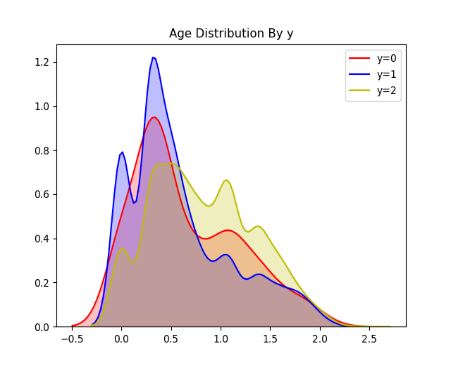
בוצעה בדיקה האם ישנה השפעה לגזעיות החיה על מהירות אימוצו. התגלה כי אכן ישנה השפעה וכאשר החיה גזעית יש עלייה של כ- 0.012 בהסתברות לאימוץ ביום הראשון, עלייה של כ- 0.058 בהסתברות לאימוץ במהלך שלושת החודשים הראשונים וירידה של כ-0.07 בהסתברות לכך שלא יאומצו ב3 החודשים הראשונים. כלומר ניתן לראות שהיות החיה גזעית משפיע לטובה על מהירות האימוץ.

****

*עבור משתנה קטגוריאלי זה ניתן להפיק עוד מידע. הוחלט להוסיף לכל רשומה – בצמוד למחוז 3 משתנים נוספים:*

**

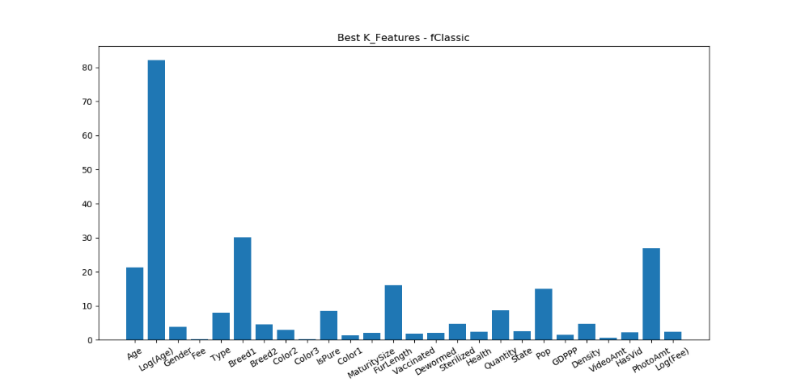
*כלומר לייצג בנוסף את כמות האוכ' במחוז, את הצפיפות במחוז ואת תל"ג המחוז.*

****

*כיון שמשתנים אלה מקבלים ערכים גבוהים ביחס לכך שהערך השכיח בהם הוא 0 הוחלט להוסיף 2 משתנים חדשים, שהם* ***התמרה לוגריתמית*** *של המשתנים הללו. משתנים אלו יבחנו אל מול המוסבר ואז יוחלט האם יש בהם צורך.*

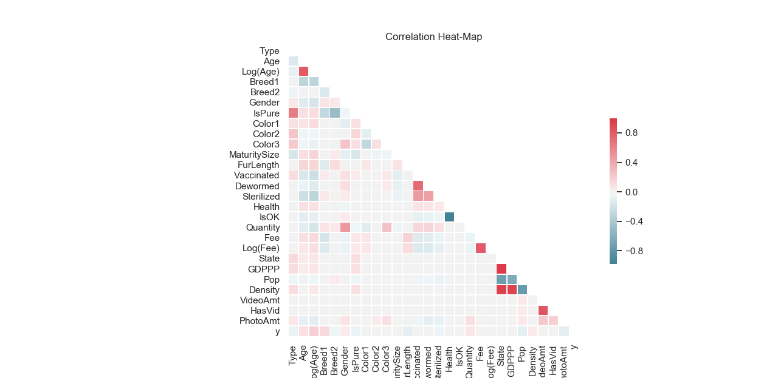
****

*מפאת סט לא מאוזן במשתנה Health , הוחלט לבצע טרנפורמציה למשתנה בצורה הבאה:*

*לאחר ביצוע הטרנפורמציות והוספת המשתנים בוצע מחדש תהליך של בחינת מתאם ומבחני בחירת משתנים כפי שניתן לראות בגרפים הבאים:*

Correlation Heat-Map with all variables

K Features selection with new variables



## 

## הכנת הנתונים לאימון ובניית מערכת לומדת

על מנת להכין את הנתונים לאימון ובחירת מערכת לומדת, יש לחלק את הנתונים לסטים של אימון אימות ובחינה. חלוקת הנתונים לסטים בלתי תלויים, תורמת למציאת המודל הנכון ביותר תוך התחשבות בנושא התאמת היתר, כאשר לכל סט ישנן מטרות שונות. מטרת סט האימון היא לאמן מודלים עם פרמטרים מוגדרים, מטרת סט האימות היא בחינה של המודלים השונים עם פרמטרים מסוימים והשוואת ציוני האימות שלהם .בכך ניתן לאפשר בחירה במודל המקבל את הציון הגבוה ביותר. מטרת סט הבחינה היא מתן ציון סופי למודל שנבחר. במסגרת סט האימון, נשתמש בשיטת ה Cross Validation, לשם כך נעשה שימוש באלגוריתם K-Folds. אלגוריתם זה מחלק את הנתונים לK סטים. בכל איטרציה סט אחד נבחר להיות סט האימות סט אחר להיות סט בחירה ושאר הסטים מהווים את סט האימון, כלומר מקבלים K חלוקות שונות.

נבחר להשתמש בK=4 , כאשר בכל חלוקה נחלק את הסט שמיועד לאימות ובחינה לשני חלקים, כאשר פעם אחת החצי הראשון לבחינה והשני לאימות ובפעם השניה להפך. כך למעשה נקבל שמונה חלוקות שונות שבכל אחת מהן 12.5% לבחינה, 12.5% לאימון ו75% ישמשו לסט הנתונים.

# נספחים

## נספח 1 :טבלאות שכיחות והסתברויות בייסניות למציאת הסתברות אפריורית

### 1.1 סוג החיה

|  |  |  |  |
| --- | --- | --- | --- |
| Type/Y | 1 | 2 | 3 |
| 0 | 140 | 188 | 328 |
| 1 | 2133 | 1890 | 4023 |
| 2 | 1926 | 1407 | 3333 |
| **סה"כ** | **4199** | **3485** | **7684** |

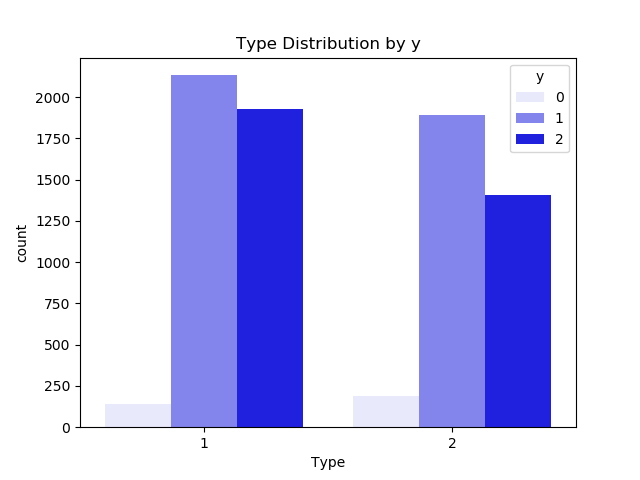
### 1.2

|  |  |  |  |
| --- | --- | --- | --- |
| Type/Y |  |  |  |
| 0 | 0.033341 | 0.053945 | 0.042686 |
| 1 | 0.507978 | 0.542324 | 0.523555 |
| 2 | 0.458681 | 0.40373 | 0.433758 |
| **סה"כ** | **1** | **1** | **1** |

**מהירות האימוץ כתלות בסוג בעל החיים-**

ניתן לראות כי ההסתברות כי חתול יאומץ באותו היום שנרשם היא כמעט כפולה מההסתברות שכלב יאומץ באותו היום. כמו כן ניתן לראות כי ההסתברות שחתול יאומץ בין יום לשלושה חודשים משלב ההרשמה גבוה במעט מההסתברות המקבילה עבור כלב. כמובן שההסתברות המשלימה, בעל החיים לא אומץ כלל, גבוה יותר עבור כלב מאשר חתול. ניתן לראות כי ההסתברות של בעל חיים לא מוגדר נמצאת על הסקלה בין 2 סוגי בעלי החיים.

### 1.3



### 1.4

|  |  |  |  |
| --- | --- | --- | --- |
| Gender/Y |  | 2 | 3 |
| 0 | 130 | 162 | 36 |
| 1 | 1537 | 1919 | 567 |
| 2 | 1110 | 1717 | 506 |
| **סה"כ** | **2777** | **3789** | **1109** |

1.5

|  |  |  |  |
| --- | --- | --- | --- |
| Gender/Y |  |  |  |
| 0 | 0.046813 | 0.042755 | 0.032462 |
| 1 | 0.553475 | 0.506466 | 0.511271 |
| 2 | 0.399712 | 0.453154 | 0.456267 |
| **סה"כ** | **1** | **1** | **1** |

**מהירות האימוץ כתלות במגדר-**

ניתן לראות כי בנוגע לאימוץ באותו יום הרישום, אין הבדלים מהותיים בין זכר לנקבה. עם זאת, ניתן לראות כי כאשר המגדר לא ידוע, ההסתברות לאימוץ באותו היום קטן בצורה משמעותית.

כמו כן ניתן לראות כי ההסתברות שבעל חיים לא יאומץ כלל גדולה יותר כאשר בעל החיים הוא נקבה או לא מוגדר, כלומר, הסיכוי עבור בעל חיים זכרי לאימוץ גדול יותר.

### 1.6

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| MaturitySize/Y | 1 | 2 | 3 | 4 |
| 0 | 109 | 179 | 38 | 2 |
| 1 | 908 | 2783 | 322 | 10 |
| 2 | 692 | 2365 | 273 | 3 |
| **סה"כ** | **1709** | **5327** | **633** | **15** |

### 1.7

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| MaturitySize/Y |  |  |  |  |
| 0 | 0.06378 | 0.033602 | 0.060032 | 0.133333 |
| 1 | 0.531305 | 0.522433 | 0.508689 | 0.666667 |
| 2 | 0.404915 | 0.443965 | 0.43128 | 0.2 |
| **סה"כ** | **1** | **1** | **1** | **1** |

קשר בין גודל החיב בבגרות לבין מיהרות האימוץ מתבקש. העדפות האנשים משתנות כתלות בגודל – אילוצי דירה, מקום וכו'.

### 1.7

|  |  |  |  |
| --- | --- | --- | --- |
| FurLength/Y | 1 | 2 | 3 |
| 0 | 155 | 128 | 45 |
| 1 | 2293 | 1480 | 250 |
| 2 | 2070 | 1158 | 105 |
| **סה"כ** | **4518** | **2766** | **400** |

|  |  |  |  |
| --- | --- | --- | --- |
| FurLength/Y |  |  |  |
| 0 | 0.034307 | 0.046276 | 0.1125 |
| 1 | 0.507525 | 0.535069 | 0.625 |
| 2 | 0.458167 | 0.418655 | 0.2625 |
| **סה"כ** | **1** | **1** | **1** |

### 1.8

### 1.9

|  |  |  |  |
| --- | --- | --- | --- |
| Vaccinated/Y | 1 | 2 | 3 |
| 0 | 119 | 164 | 45 |
| 1 | 1519 | 2073 | 431 |
| 2 | 1522 | 1319 | 492 |
| **סה"כ** | **3160** | **3556** | **968** |

### 1.10

|  |  |  |  |
| --- | --- | --- | --- |
| Vaccinated/Y |  |  |  |
| 0 | 0.037658 | 0.046119 | 0.046488 |
| 1 | 0.480696 | 0.582958 | 0.445248 |
| 2 | 0.481646 | 0.370922 | 0.508264 |
| **סה"כ** | **1** | **1** | **1** |

### 1.11

|  |  |  |  |
| --- | --- | --- | --- |
| Dewormed/Y | 1 | 2 | 3 |
| 0 | 167 | 118 | 43 |
| 1 | 2285 | 1333 | 405 |
| 2 | 1897 | 960 | 476 |
| **סה"כ** | **4349** | **2411** | **924** |

### 1.12

|  |  |  |  |
| --- | --- | --- | --- |
| Dewormed/Y |  |  |  |
| 0 | 0.0384 | 0.048942 | 0.046537 |
| 1 | 0.525408 | 0.552883 | 0.438312 |
| 2 | 0.436192 | 0.398175 | 0.515152 |
| סך הכל | 1 | 1 | 1 |

### 1.13

|  |  |  |  |
| --- | --- | --- | --- |
| Sterilized/Y | 1 | 2 | 3 |
| 0 | 54 | 223 | 51 |
| 1 | 677 | 2982 | 364 |
| 2 | 1034 | 1698 | 601 |
| **סה"כ** | **1765** | **4903** | **1016** |

### 1.14

|  |  |  |  |
| --- | --- | --- | --- |
| Sterilized/Y |  |  |  |
| 0 | 0.030595 | 0.045482 | 0.050197 |
| 1 | 0.383569 | 0.608199 | 0.358268 |
| 2 | 0.585836 | 0.346319 | 0.591535 |
| **סה"כ** | 1 | 1 | 1 |

### 1.15

|  |  |  |  |
| --- | --- | --- | --- |
| Health/Y | 1 | 2 | 3 |
| 0 | 313 | 14 | 1 |
| 1 | 3900 | 118 | 5 |
| 2 | 3186 | 137 | 10 |
| **סה"כ** | **7399** | **269** | **16** |

### 1.16

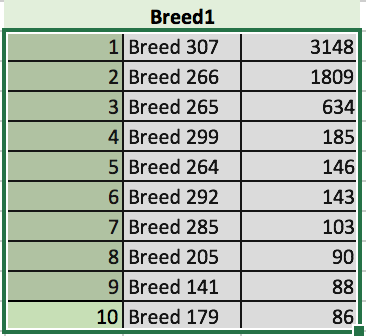
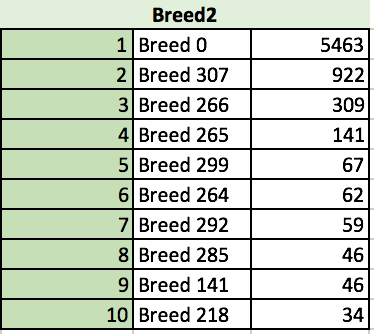
עבור המשתנה מצב בריאותי ניתן לראות התפלגויות מותנות שונות. לדוגמא בעבור חיה עם פציעה חמורה נראה שיש נטייה לקצוות. כלומר – או לאמץ מיד או לא לאמץ בכלל.

|  |  |  |  |
| --- | --- | --- | --- |
| Health/Y |  |  |  |
| 0 | 0.042303 | 0.052045 | 0.0625 |
| 1 | 0.527098 | 0.438662 | 0.3125 |
| 2 | 0.430599 | 0.509294 | 0.625 |
| **סה"כ** | 1 | 1 | 1 |

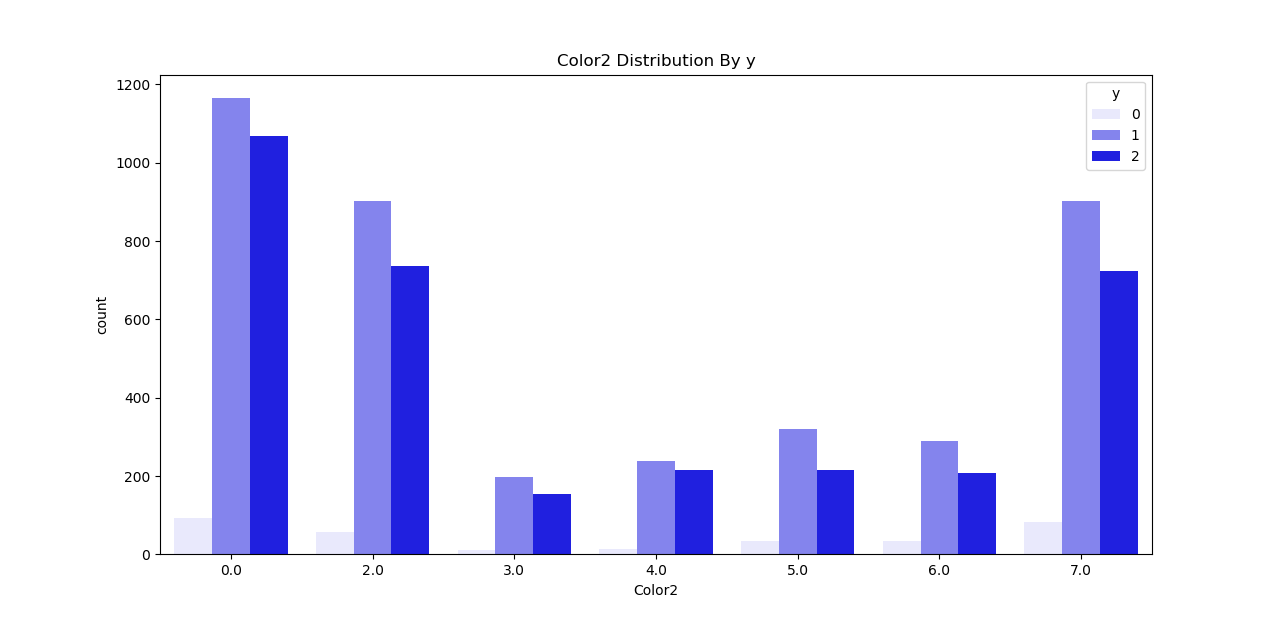
מאידך חשוב לזכור כי יש מס' מאוד קטן של תצפיות במצבים 3 או 2. לכן קשה להגיד שזה מדגם מייצג.

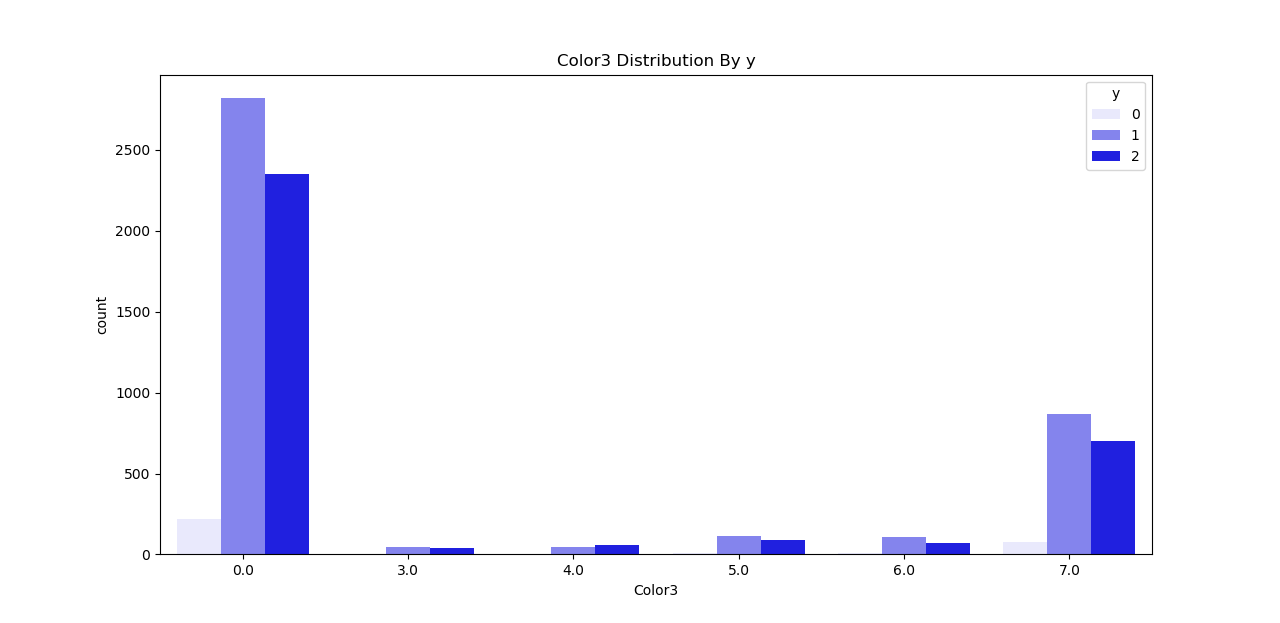
### 1.17

התפלגויות עשרת הזנים המובילים בBreed1 וב-Breed2.



### 1.17 colors

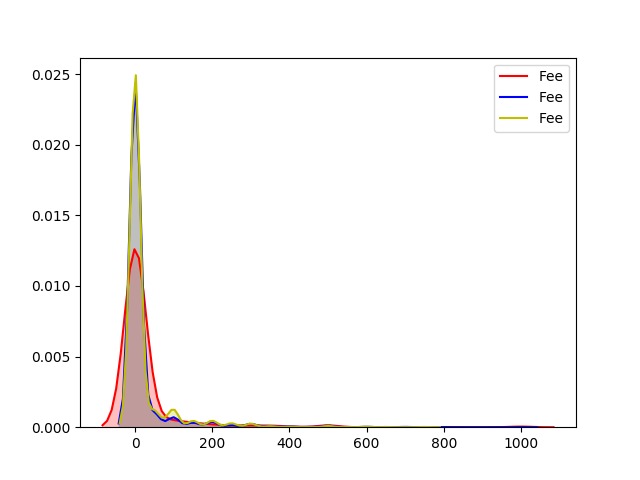




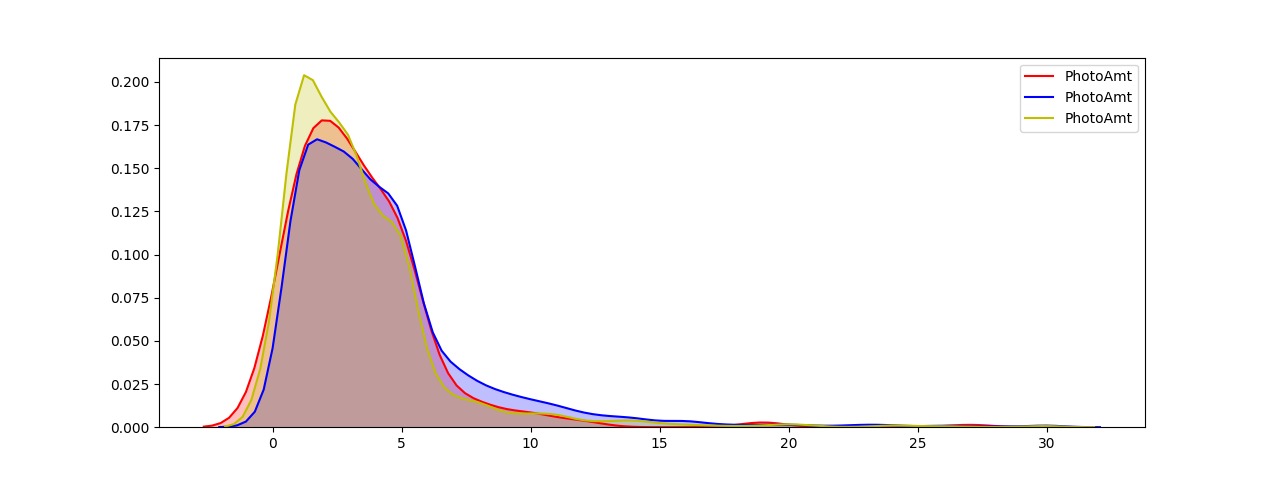
ניתן לראות כי עבור התפלגות 2 הצבעים אין שינוי מיוחד במערב בין הצבעים. בנוסף Color3 לא מאוזן כלל. הוא יוצר הטייה והמון מידע חסר.

### 2.1

אדום – y=0, כחול y=1, צהוב, y=2

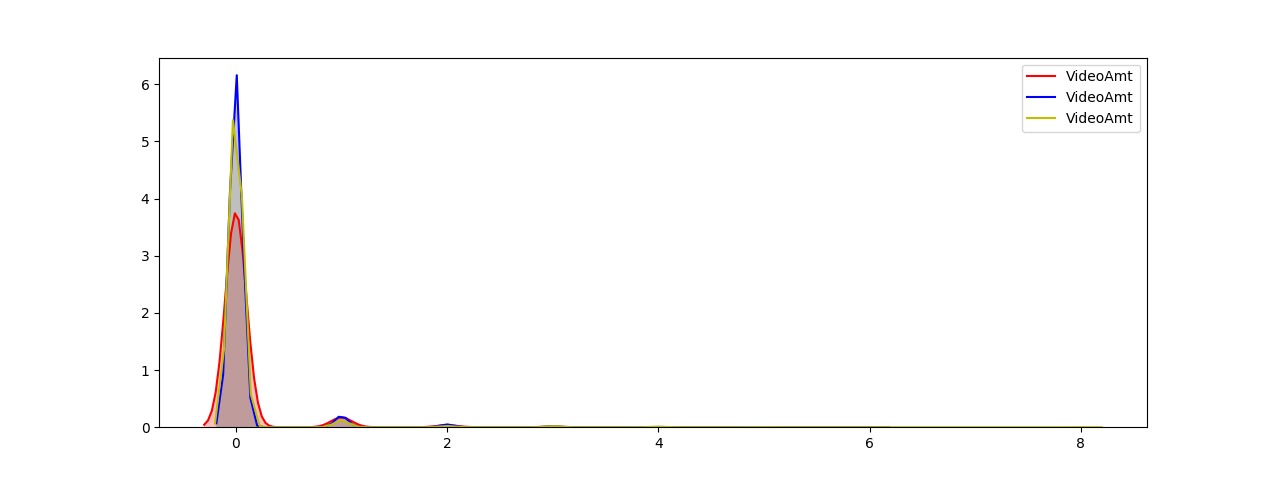
2.2

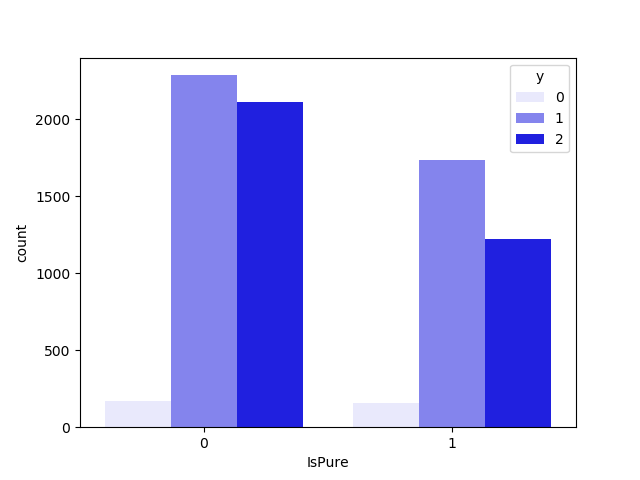
אדום – y=0, כחול y=1, צהוב, y=2



### 2.3

אדום – y=0, כחול y=1, צהוב, y=2





### 2.5

### 2.6

Table P(Y/IsPure)

|  |  |  |
| --- | --- | --- |
|  | התפלגות המידע | |
|  | 0 | 1 |
| 0 | 0.527439024 | 0.472560976 |
| 1 | 0.568232662 | 0.431767338 |
| 2 | 0.633963396 | 0.366036604 |

### 3.1

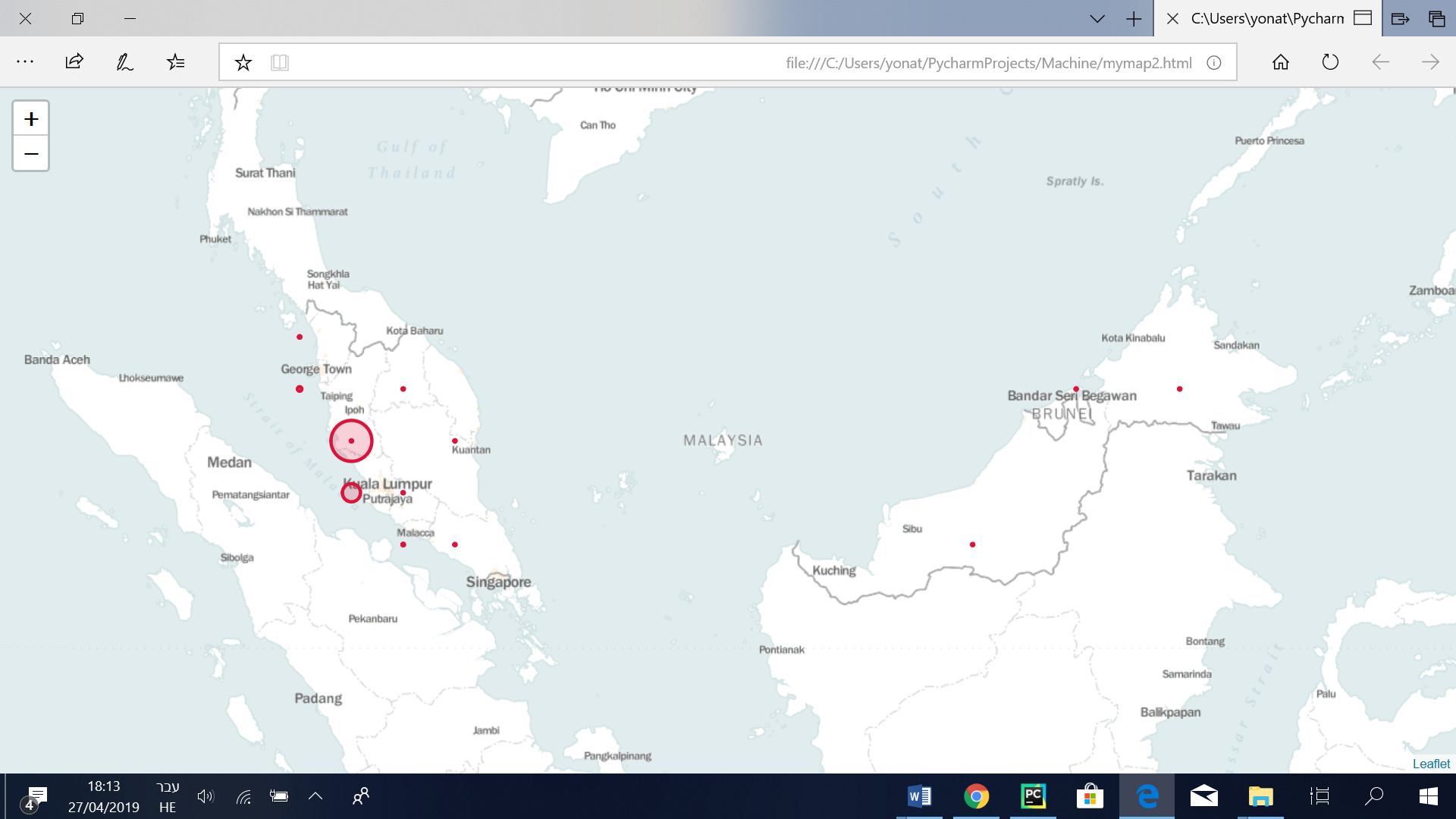
Corralation Matrix

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Type | Age | Breed1 | Breed2 | Gender | Color1 | Color2 | Color3 | Maturity Size | Fur Length | Vaccinated | Dewormed | Sterilized | Health | Quantity | Fee | State | VideoAmt | PhotoAmt | y |
| Type | 1.00 | -0.14 | 0.03 | -0.05 | 0.06 | 0.10 | 0.24 | 0.21 | -0.16 | 0.01 | 0.12 | 0.04 | 0.01 | 0.00 | 0.04 | -0.03 | 0.13 | -0.02 | 0.06 | -0.07 |
| Age | -0.14 | 1.00 | -0.31 | -0.04 | -0.12 | 0.08 | -0.05 | -0.05 | 0.12 | 0.16 | -0.16 | -0.07 | -0.22 | 0.09 | -0.12 | 0.09 | 0.04 | -0.01 | -0.09 | 0.10 |
| Breed1 | 0.03 | -0.31 | 1.00 | -0.15 | 0.08 | -0.04 | -0.02 | -0.01 | 0.00 | -0.13 | 0.05 | 0.03 | 0.05 | -0.03 | 0.09 | -0.19 | -0.03 | 0.01 | 0.03 | 0.13 |
| Breed2 | -0.05 | -0.04 | -0.15 | 1.00 | 0.06 | -0.01 | 0.01 | 0.03 | 0.04 | 0.11 | 0.00 | -0.01 | -0.01 | -0.04 | 0.05 | 0.02 | -0.03 | 0.02 | 0.04 | -0.05 |
| Gender | 0.06 | -0.12 | 0.08 | 0.06 | 1.00 | -0.11 | 0.03 | 0.25 | -0.09 | -0.03 | 0.10 | 0.11 | 0.05 | -0.05 | 0.49 | -0.06 | 0.00 | 0.00 | 0.09 | 0.05 |
| Color1 | 0.10 | 0.08 | -0.04 | -0.01 | -0.11 | 1.00 | -0.11 | -0.28 | -0.01 | 0.07 | -0.02 | -0.03 | -0.04 | 0.03 | -0.11 | 0.07 | 0.01 | -0.01 | -0.04 | -0.02 |
| Color2 | 0.24 | -0.05 | -0.02 | 0.01 | 0.03 | -0.11 | 1.00 | 0.09 | -0.06 | -0.03 | 0.03 | 0.01 | 0.01 | -0.01 | 0.02 | -0.01 | 0.02 | 0.01 | 0.05 | -0.04 |
| Color3 | 0.21 | -0.05 | -0.01 | 0.03 | 0.25 | -0.28 | 0.09 | 1.00 | -0.05 | 0.01 | 0.06 | 0.05 | 0.04 | -0.02 | 0.27 | -0.03 | 0.01 | 0.02 | 0.10 | -0.01 |
| Maturity Size | -0.16 | 0.12 | 0.00 | 0.04 | -0.09 | -0.01 | -0.06 | -0.05 | 1.00 | 0.09 | -0.11 | -0.09 | -0.08 | 0.00 | -0.05 | 0.04 | -0.04 | 0.01 | 0.02 | 0.03 |
| Fur Length | 0.01 | 0.16 | -0.13 | 0.11 | -0.03 | 0.07 | -0.03 | 0.01 | 0.09 | 1.00 | 0.00 | 0.01 | 0.03 | 0.02 | -0.03 | 0.17 | -0.02 | -0.02 | -0.03 | -0.09 |
| Vaccinated | 0.12 | -0.16 | 0.05 | 0.00 | 0.10 | -0.02 | 0.03 | 0.06 | -0.11 | 0.00 | 1.00 | 0.73 | 0.48 | 0.08 | 0.14 | -0.13 | 0.03 | -0.04 | -0.05 | -0.03 |
| Dewormed | 0.04 | -0.07 | 0.03 | -0.01 | 0.11 | -0.03 | 0.01 | 0.05 | -0.09 | 0.01 | 0.73 | 1.00 | 0.44 | 0.08 | 0.16 | -0.12 | 0.01 | -0.03 | -0.09 | 0.01 |
| Sterilized | 0.01 | -0.22 | 0.05 | -0.01 | 0.05 | -0.04 | 0.01 | 0.04 | -0.08 | 0.03 | 0.48 | 0.44 | 1.00 | 0.05 | 0.11 | -0.07 | 0.02 | -0.02 | -0.08 | -0.05 |
| Health | 0.00 | 0.09 | -0.03 | -0.04 | -0.05 | 0.03 | -0.01 | -0.02 | 0.00 | 0.02 | 0.08 | 0.08 | 0.05 | 1.00 | -0.04 | -0.02 | 0.04 | 0.01 | -0.02 | 0.03 |
| Quantity | 0.04 | -0.12 | 0.09 | 0.05 | 0.49 | -0.11 | 0.02 | 0.27 | -0.05 | -0.03 | 0.14 | 0.16 | 0.11 | -0.04 | 1.00 | -0.06 | 0.01 | 0.01 | 0.11 | 0.07 |
| Fee | -0.03 | 0.09 | -0.19 | 0.02 | -0.06 | 0.07 | -0.01 | -0.03 | 0.04 | 0.17 | -0.13 | -0.12 | -0.07 | -0.02 | -0.06 | 1.00 | -0.01 | 0.00 | 0.01 | 0.01 |
| State | 0.13 | 0.04 | -0.03 | -0.03 | 0.00 | 0.01 | 0.02 | 0.01 | -0.04 | -0.02 | 0.03 | 0.01 | 0.02 | 0.04 | 0.01 | -0.01 | 1.00 | -0.03 | -0.01 | 0.03 |
| VideoAmt | -0.02 | -0.01 | 0.01 | 0.02 | 0.00 | -0.01 | 0.01 | 0.02 | 0.01 | -0.02 | -0.04 | -0.03 | -0.02 | 0.01 | 0.01 | 0.00 | -0.03 | 1.00 | 0.22 | -0.02 |
| PhotoAmt | 0.06 | -0.09 | 0.03 | 0.04 | 0.09 | -0.04 | 0.05 | 0.10 | 0.02 | -0.03 | -0.05 | -0.09 | -0.08 | -0.02 | 0.11 | 0.01 | -0.01 | 0.22 | 1.00 | -0.09 |
| y | -0.07 | 0.10 | 0.13 | -0.05 | 0.05 | -0.02 | -0.04 | -0.01 | 0.03 | -0.09 | -0.03 | 0.01 | -0.05 | 0.03 | 0.07 | 0.01 | 0.03 | -0.02 | -0.09 | 1.00 |

### 3.2 מפת שכיחויות וקוד הרצה

גודל הנקודות האדומות פרופורציוני לכמות התצפיות מאותו מחוז

**Bubble Map with Python**



##Drawing a Map and save it as an HTML Document  
m = folium.Map(location=[20, 0], tiles=**"Mapbox Bright"**, zoom\_start=2)  
**for** i **in** range(0, len(DataBubbles)):

folium.Circle(  
 location=[int(DataBubbles.iloc[i][**'Lat'**]), int(DataBubbles.iloc[i][**'Long'**])],  
 popup=DataBubbles.iloc[i][**'Name'**],  
 radius=int(DataBubbles.iloc[i][**'Count'**])\*10 ,  
 color=**'crimson'**,  
 fill=**True**,  
 fill\_color=**'crimson'** ).add\_to(m)  
m.save(**'mymap2.html'**)

### 4.1

**import** pandas **as** pd  
**import** seaborn **as** sns  
**import** matplotlib.pyplot **as** plt  
**import** numpy **as** np  
**import** numpy **as** np  
**from** scipy.stats **import** gaussian\_kde  
**import** folium  
  
  
  
Data = pd.read\_csv(**"C:\\Users\yonat\Desktop\תעשייה וניהול\שנה ג'\סמס ב\לימוד מכונה\פרויקט\Data.csv"**)  
EditedData = pd.read\_csv(**"C:\\Users\yonat\Desktop\תעשייה וניהול\שנה ג'\סמס ב\לימוד מכונה\פרויקט\EditedData.csv"**)  
DataTest = pd.read\_csv(**"C:\\Users\yonat\Desktop\תעשייה וניהול\שנה ג'\סמס ב\לימוד מכונה\פרויקט\Test.csv"**)  
DataBubbles = pd.read\_csv(**"C:\\Users\yonat\Desktop\תעשייה וניהול\שנה ג'\סמס ב\לימוד מכונה\פרויקט\Bubbles.csv"**)  
  
*##Print For Check*print(DataBubbles.head())  
  
  
*#ax = sns.countplot(x="y", data=Data , color="Blue")  
#plt.title('y Counting')  
  
#ax = sns.countplot(x="IsPure", data=Data,hue='y' , color="Blue")  
#plt.title('IsPure and y')  
  
#ax = sns.countplot(x="Vaccinated", data=Data,hue='y' , color="Blue")  
#plt.title('Vaccinated and y')  
  
#ax = sns.countplot(x="Type", data=DataTest , color="Purple")  
#plt.title('Type Distribution - Test DataSet')  
  
#ax = sns.countplot(x="Gender", data=Data, color="Blue")  
#plt.title('Gender Distribution - Train')  
  
#ax = sns.countplot(x="Gender", data=Data,hue='y' ,color="Blue")  
#plt.title('Gender Distribution By y')  
  
#ax = sns.countplot(x="NumberOfColors" ,data =Data ,hue='y',color=(0.2, 0.1, 0.1))  
#plt.title('Color1 Distribution - TEST')  
  
#ax = sns.countplot(x="NumberOfColors" ,data =Data ,hue='y',color=(0.2, 0.1, 0.1))  
#plt.title('Color1 Distribution - TEST')  
  
##-------Pie Chart----------##  
#labels = Data['y'].astype('category').cat.categories.tolist()  
#counts = Data['y'].value\_counts()  
#sizes = [counts[var\_cat] for var\_cat in labels]  
#fig1, ax1 = plt.subplots()  
#ax1.pie(sizes, labels=labels, autopct='%1.1f%%', shadow=True) #autopct is show the % on plot  
#ax1.axis('equal')  
  
  
#ax = sns.distplot( Data["PhotoAmt"] )  
#plt.title('PhotoAmt Distribution - Train')  
  
#ax = sns.countplot(x="Color3" ,data =Data,color=(0.2, 0.1, 0.1))  
#plt.title('Color1 Distribution - TEST')  
#Coorelation  
  
#corrr=np.corrcoef(EditedData[:,0],EditedData[:,1],EditedData[:,2],EditedData[:,3,],  
#EditedData[:,4])  
# marriage ~balance  
  
#xy = np.vstack([EditedData["Color2"],EditedData["Gender"]])  
#z = gaussian\_kde(xy)(xy)  
#fig, ax = plt.subplots()  
#plt.hist2d(EditedData["Dewormed"], EditedData["Vaccinated"], (50, 50), cmap=plt.cm.jet)  
#plt.colorbar()  
#plt.xlabel("Color3")  
#plt.ylabel("Gender")  
#plt.title("Point observations and Reggasion")  
#plt.show()  
  
  
#plt.scatter(DataBubbles['Dewormed'], DataBubbles['Vaccinated'], s=DataBubbles['Count'], c="blue", alpha=0.4, linewidth=2)  
#plt.xlabel("Dewormed")  
#plt.ylabel("Vaccinated")  
#plt.title("BubblePlot By Count")  
#plt.show()*

K-selection :

**import** pandas **as** pd  
**import** seaborn **as** sns  
**import** matplotlib.pyplot **as** plt  
**import** numpy **as** np  
  
**from** sklearn.feature\_selection **import** SelectKBest, f\_classif , chi2  
  
Data = pd.read\_csv(**"C:\\Users\yonat\Desktop\תעשייה וניהול\שנה ג'\סמס ב\לימוד מכונה\פרויקט\EditedData.csv"**)  
Data.fillna(**'Breed1'**)  
Data.fillna(**'Breed2'**)  
  
*#Full*predictors = [**"Age"**,**'Log(Age)'** ,**"Gender"** , **"Fee"** , **"Type"**,**'Breed1'**,**'Breed2'**,**'Color2'**,**'Color3'**,**'IsPure'**,**'Color1'**,**'MaturitySize'**,**'FurLength'**,**'Vaccinated'**,**'Dewormed'**,**'Sterilized'**,**'Health'**,**'Quantity'**,**'State'**,**'Pop'**,**'GDPPP'**,**'Density'**,**'VideoAmt'**,**'HasVid'**,**'PhotoAmt'**,**'Log(Fee)'**]  
  
*#Semi  
#predictors = ["Age", "Gender" , "Fee" , "Type",'Breed1','Breed2','Color1','Color2','Color3','MaturitySize','FurLength','Vaccinated','Dewormed','Sterilized','Health','Quantity','State','VideoAmt','PhotoAmt']  
  
# Perform feature selection*selector = SelectKBest(f\_classif, k=7)  
selector.fit(Data[predictors], Data[**"y"**])  
  
*# Get the raw p-values for each feature, and transform from p-values into scores*scores = -np.log10(selector.pvalues\_)  
  
*# Plot the scores. See how "Pclass", "Sex", "Title", and "Fare" are the best?*plt.bar(range(len(predictors)), scores)  
plt.xticks(range(len(predictors)), predictors, rotation=30)  
plt.title(**'Best K\_Features - fClassic'**)  
plt.show()

Correlation Matrix:

**import** pandas **as** pd  
**import** seaborn **as** sns  
**import** matplotlib.pyplot **as** plt  
**import** numpy **as** np  
**import** numpy **as** np  
**from** matplotlib **import** pyplot **as** plt  
**from** matplotlib **import** cm **as** cm  
  
Data = pd.read\_csv(**"C:\\Users\yonat\Desktop\תעשייה וניהול\שנה ג'\סמס ב\לימוד מכונה\פרויקט\Data.csv"**)  
EditedData = pd.read\_csv(**"C:\\Users\yonat\Desktop\תעשייה וניהול\שנה ג'\סמס ב\לימוד מכונה\פרויקט\EditedData.csv"**)  
DataTest = pd.read\_csv(**"C:\\Users\yonat\Desktop\תעשייה וניהול\שנה ג'\סמס ב\לימוד מכונה\פרויקט\Test.csv"**)  
  
  
  
**from** string **import** ascii\_letters  
**import** numpy **as** np  
**import** pandas **as** pd  
**import** seaborn **as** sns  
**import** matplotlib.pyplot **as** plt  
  
  
  
  
sns.set(style=**"white"**)  
  
  
*# Compute the correlation matrix*corr = EditedData.corr()  
  
*# Generate a mask for the upper triangle*mask = np.zeros\_like(corr, dtype=np.bool)  
mask[np.triu\_indices\_from(mask)] = **True***# Set up the matplotlib figure*f, ax = plt.subplots(figsize=(11, 9))  
  
*# Generate a custom diverging colormap*cmap = sns.diverging\_palette(220, 10, as\_cmap=**True**)  
  
*# Draw the heatmap with the mask and correct aspect ratio*sns.heatmap(corr, mask=mask, cmap=cmap, vmax=1, center=0,  
 square=**True**, linewidths=.5, cbar\_kws={**"shrink"**: .5})  
plt.title(**'Correlation Heat-Map'**)  
*#corr.to\_csv("C:\\Users\yonat\Desktop\תעשייה וניהול\שנה ג'\סמס ב\לימוד מכונה\פרויקט\CorrelationMatrix.csv")*print(corr)  
plt.show()

### 4.2 נספח מאמרים

Prediction of Adoption Versus Euthanasia Among Dogs and Cats in a California Animal  
 Shelter, שנכתב על ידי Merry Lepper, Philip H. Kass, and Lynette A. Hart