לימוד מכונה 364-1-1811

**פרויקט - חלק א'**

תוכן עניינים

[תוכן עניינים 1](#_Toc8145549)

[מבוא 2](#_Toc8145550)

[הגדרת הבעיה 2](#_Toc8145551)

[תיאור כללי של עולם התוכן הנחקר 2](#_Toc8145552)

[הגדרת שאלת המחקר 2](#_Toc8145553)

[הבנת הנתונים 2](#_Toc8145554)

[תיעוד מקורות הנתונים ומשמעותם 2](#_Toc8145555)

[הסתברויות אפריוריות וקשרים בין מאפיינים 4](#_Toc8145556)

[משתנים קטגוריאלים 4](#_Toc8145557)

[משתנים רציפים 4](#_Toc8145558)

[קשרים בין משתנים 5](#_Toc8145559)

[איכות הנתונים 6](#_Toc8145560)

[הכנת הנתונים 7](#_Toc8145561)

[בחירת מאפיינים 7](#_Toc8145562)

[טיפול פרטני במאפיינים 8](#_Toc8145563)

[המצב החדש 9](#_Toc8145564)

[הכנת הנתונים לאימון ובחינת מערכת לומדת 10](#_Toc8145565)

[נספחים 11](#_Toc8145566)

[טבלאות שכיחות 11](#_Toc8145567)

[קטגוריאלים – גרפי פורפורציה 11](#_Toc8145568)

[משתנים רציפים – היסטוגרמות 15](#_Toc8145569)

[קשרים בין משתנים – גרפים וטבלאות 16](#_Toc8145570)

[קשרים עם פיצ'רים חדשים לאחר טרנספורמציה: 18](#_Toc8145571)

[קוד R 20](#_Toc8145572)

[קובץ איתחול 20](#_Toc8145573)

[קובץ ראשי 20](#_Toc8145574)

# מבוא

בעבודה זו קיבלנו סט נתונים טבלאי ומספר טבלאות עזר, נתונים אלו שייכים לאתר אימוץ חיות במדינת מלזיה. זהו חלק ראשון של הפרויקט בוא אנו לומדים את הנתונים והתצפיות, ביצאנו טרנספורמציות ושינויים על המשתנים המסבירים (מפורט בדוח). למידת הנתונים והניתוחים הסטטיסטים התבצעו בשפת R דרך ממשק R – STUDIO וכמו כן גם כל המניפולציות על הנתונים. הלמידה של סט נתונים חדש היותה אתגר מעניין ולמדנו רבות איך לבחון מידע חדש ואיזה ערך ניתן להפיק ממנו, ואיך ניתן אף לשפר אותו להמשך עבודה עתידי.

<https://www.petfinder.my/-> האתר ממנו נאספו הנתונים.

# הגדרת הבעיה

## תיאור כללי של עולם התוכן הנחקר

הבעיה המחקרית איתה אנו מתמודדים הינה מה הפרמטרים המשפיעים על אימוץ חיות מחמד (כלבים וחתולים) באתר אימוץ החיות petfinder. אנשים ברובם שונים זה מזה באופי ובטעם, ולכן קיים שוני בין הפרמטרים אשר משפיעים על אדם מסויים לאמץ חיית מחמד. ממחקרים שנעשו בנושא, גילינו כי אחד האלמטים המשמעותיים ביותר המשפעים על אימוץ הוא תמונה של חיית המחמד (<https://heartsspeak.org/how-photos-are-important-to-pet-adoption-a-study/>). 9 מתוך 10 מאמצים פוטנציאלים השתמשו בתמונות במרשתת על מנת להשוות בין חיות מחמד לפני קבלת החלטה. בנוסף, 65% העידו כי צפו התמונות ברשת לפני שהתחילו בתהליך האימוץ. מעבר לכך, מחקרים נוספים הראו כי לעונת השנה יש קשר לאימוץ/נטישת חיות מחמד, וכי אימוץ חיות מחמד המגיעות מהסגר יומר תחושת גאווה ושליחות אצל המאמצים.

(<https://chewonthis.maddiesfund.org/2017/10/spp-research/>)

## הגדרת שאלת המחקר

מה אנו מצפים לפתור בעזרת הכלים והשיטות של מערכות לומדות?

בעזרת הכלים של מערכות לומדות, אנו מצפים ללמוד ולהבין אלו משתנים ופרמטרים, הנאספים על כלבים וחתולים, יש השפעה על תהליך האימוץ של החיה וגם מה העוצמה/ההשפעה של כל משתנה בפני עצמו. כלומר, אני בודקים אילו משתנים ופרמטרים שניתן לאסוף, על בעל חיים המיועד לאימוץ, יעזרו לחזות באופן המירבי את אימוץ בעל החיים.

# הבנת הנתונים

תיעוד מקורות הנתונים ומשמעותם

מקור הנתונים שלנו הוא אתר אימוץ חיות במדינת מלזיה. האתר מספק שירותי אימוץ של מגוון חיות אך מתמקד בשני סוגים עיקריים: כלבים וחתולים.

לא מפורט כיצד נאספו הנתונים אך ניתן לשער בסבירות גבוהה כי רובם הוזנו ידנית למערכת מידע של אתר האינטרנט דרך ממשק "הוספת חיה לאימוץ", מודול שכנראה קיים במערכת המידע, ומאפשר הוספת פרטים שונים המספקים תיאור של החיה המיועדת לאימוץ על ידי המוסר. אנו מאמינים שהנתונים נאספו בדרך זו כיוון שכאשר נכנסים לפרופיל של "חיה לאימוץ" מתקבלים פרטים על החיה שמורכבים מטבלאות שונות שקיבלנו מבסיס הנתונים. לכן סביר להניח כי קיימת מערכת מידע המאגדת את המידע דרך טבלאות שונות ולא הזנה ידנית של המידע לטבלאות אקסל. לאחר בדיקה מעמיקה מצאנו כי כל משתמש רשום באתר יכול ליצור פרופיל של חיה לאימוץ.

לסיכום, ייתכן כי קיימת בעיה באיכות הנתונים מהבחינה שהתגיות נעשות על ידי אנשים, ולא מומחים. בנוסף, כיוון שכל משתמש רשאי ליצור פרופיל אימוץ, דבר שיכול להוביל לנתונים חסרים וחסר אחידות.

### פירוט המשתנים השונים כפי שהתקבלו

Type – סוג בעל החיים (1 = כלב, 2 = חתול)

Age – גיל בעל החיים

Breed1 – גזע עיקרי (ערכים בטבלת BreedLabels)

Breed2 – גזע משני (ערכים בטבלת BreedLabels, 0 = לא קיים(

Gender – מין בעל החיים (1 = זכר, 2 = נקבה, 3 = מעורב, כאשר הפרופיל מייצג קבוצה של בעלי חיים)

Color1 – צבע מס' 1 (ערכים בטבלת ColorLabels)

Color2 – צבע מס' 2 (ערכים בטבלת ColorLabels, 0 = לא קיים)

Color3 – צבע מס' 3 (ערכים בטבלת ColorLabels, 0 = לא קיים)

MaturitySize -גודל בבגרות (1 = קטן, 2 = בינוני, 3 = גדול, 4 = מאוד גדול, 0 = לא צוין)

FurLength – אורך הפרווה (1 = קצר, 2 = בינוני, 3 = ארוך, 4 = לא צוין)

Vaccinated – האם בעל החיים חוסן (1 = כן, 2 = לא, 3 = לא ידוע)

Dewormed – האם בעל החיים עבר טיפול נגד תולעים (1 = כן, 2 = לא, 3 = לא ידוע)

Sterilized – האם עבר סירוס/עיקור (1 = כן, 2 = לא, 3 = לא ידוע)

Health – מצב בריאותי (1 = בריא, 2 = פציעה מינורית, 3 = פציעה משמעותית, 4 = לא צוין)

Quantity – מספר החיות המוצגות בפרופיל

Fee – מחיר האימוץ (0 = חינם)

State – מדינה במלזיה (ערכים בטבלת StateLabels)

VideoAmt – כמות סרטוני וידאו של בעל החיים שהועלו לפרופיל

PhotoAmt – כמות התמונות של בעל החיים שהועלו לפרופיל

### משתנה מטרה

y – "מהירות האימוץ"

0 = בעל החיים נלקח לאימוץ באותו יום שנרשם

1 = בין יום לשלושה חודשים

2 = לא נלקח לאימוץ במהלך השלושת החודשים הראשונים

## הסתברויות אפריוריות וקשרים בין מאפיינים

### משתנים קטגוריאלים

על מנת לקבל סדר גודל של הנתונים וטיבם בחלטנו לעשות טבלת שכיחויות של המשתנים הקטגוריאליים. (נספח טבלאות שכיחות)

משתנה מטרה – ניתן לראות כי רק כ-4% מהחיות מאומצות כבר ביום החשיפה הראשוני וכי כמחצית מאומצים בטווח של עד שלושה חודשים והשאר כלל לא. נתון זה מדגיש כי כמות גדולה מאוד של חיות בית מחכות מעל שלושה חודשים לאימוץ או לא מאומצות כלל.

במשתנים הבאים: Gender, Color, FurLength, Vaccinated אין משהו יוצא דופן שבולט לעין.

במשתנה Dewormed ניתן לראות כי רוב החיות לא עברו טיפול – מה שיכול להפיע על החלטת האימוץ

כמו כן, גם במשתנה Sterilized ניתן לראות כי אחוז גבוה מהחיות אינן מסורסות אך לא ברור מה מידת השפעה של נתון זה. (עניין תרבותי)

נתון ה – Health בולט בכך שכ- 96% החיות הבית מוגדרות במצב בריאותי תקין. לדעתנו נתון זה חשוב כיוון שרוב האנשים לא כנראה יאמצו חיות בית הנמצאות בסכנת חיים.

בנוגע למשתנה State. מעניין לראות שרוב בעלי החיים באתר מגיעים משתי מדינות. כ-56% מ Selangor וכ- 26% ממדינת Kuala Lumpur (מעל 80%).

לכן אנו יכולים להניח כי שתי מדינות אלו בעלי אוכלוסיות גדולות יחסית לשאר המדינות.

מכיוון שישנם מספר רב של מדינות בעלות מעט תצפיות, בהמשך נחליט כיצד להפוך את המשתנה ליותר רלוונטי.

### משתנים רציפים

במשתנים הרציפים בחרנו לעשות היסטוגרמות על מנת לקבל ויזואליזציה של התפלגות הנתונים. כלל ההיסטוגרמות מוצגות בנספחים.

**Age –** משתנה זה בעל ערכים גבוהים מאד ואינו תואם לחיי אדם בשנים. לכן אנו משערים כי הנתונים מוצגים בחודשים, כי אחרת למספרים אין הגיון. לאחר יצירת ההיסטוגרמה נראה כי הנתונים בעלי התפלגות מעריכית שלילית. בנוסף, קיימים מספר ערכים גבוהים חריגים אשר תופסים חלק מאד קטן מכמות התצפיות.

**Quantity –** במשתנה זה קיים רוב גדול של ערכי 1. (כ-76%). ניתן להסיק כי רוב החיות המיועדות לאימוץ באתר הינן פרטים המגיעים ביחידים. גם כאן נראה כי הנתונים בעלי התפלגות מעריכית שלילית. ייתכן כי מתוך מחשבה עסקית גם משתנה זה יהפוך למשתנה בינארי המפריד בין מקרים בהם נמסרת חיה לבדה לאימוץ, אל מול מקרים בהם נמסרים מס' חיות יחדיו. זאת מכיוון שישנן קטגוריות מאוד דלות וישנו היגיון עסקי בכך שיהיה הבדל בין "אימוץ יחידים" ל"אימוץ רבים" ונרצה לתפוס הבדל זה.

**Fee** **–** משתנה המעיד על עלות עבור האימוץ. קיימת בעייה בהערכת משתנה זה. כ-84.1% מהתצפיות בעלי ערך 0, וכ- 94.2% מתחת לערך 100. במצב הנוכחי לא נראה כי משתנה זה יכול לעזור לנו.

**Video Amount –** במשתנה זה מצאנו כי 96% מבעלי החיים באתר **ללא** אף סרטון וידאו. עקב כך, רצינו לראות כמה מבעלי החיים באמת מאומצים תוך שלושה חודשים. כלומר בעלי סרטון וידאו וערך מטרה Y 1 או 0. קיבלנו כי רק 2.5% החיות המחמד אומצו ולכן אנו חושבים כי אנשים לא בוחרים האם לאמץ חיה או לא על סמך הקיום של סרטון וידאו.

**Photo Amount –** נתון חשוב מאד. מצאנו כי לכ- 97% מחיות המחמד באתר יש לפחות תמונה אחת. דבר זה מעיד לדעתנו כי לתמונות החיה יש תפקיד מהותי בקבלת החלטה על אימוץ.

ניתן לראות כי מרבית המשתנים אינם מאוזנים. עם זאת, לעתים הדבר כך דווקא כיוון שמייצג את המציאות. למשל, רק כ-4% מהחיות מאומצות באותו יום שהועלו. יש בכך הגיון כיוון שכנראה נדיר שחיה מאומצת כל כך מהר. באופן כללי, מכיוון שמדובר במסד נתונים מספיק גדול, סביר להניח שהוא מייצג את המציאות במדינת מלזיה.

## קשרים בין משתנים

ראשית באופן צפוי ניתן לראות כי הערכים במשתנים סירוס, חיסון ועבר תלוע (dewormed) בעלי חפיפה משמעותית של כ80% בקטגוריות. כלומר הם יסבירו את אותה השונות ויש להחליט לגבי ייעול בשימוש בהם.

באופן מעניין ניתן לראות תלות מסוימת בין מין החיה לבין סיכוייה להיות מאומצת. באופן כללי ישנן יותר ממין נקבה לאימוץ, אך למין זכר יש סיכוי של 5% יותר להיות מאומץ (בין אם באותו יום או תוך שלושה חודשים), תלות שנמצאה מובהקת דרך מבחן חי בריבוע.

בבדיקה לגבי התלות בין משתנה המטרה ובריאות החיה, נמצא כי ישנם הבדלים. מתוך הסתכלות על טבלת השכיחויות, ניתן לראות שכלבים וחתולים שאינם בריאים בעלי סיכוי נמוך יותר לאימוץ. חשוב לציין כי ישנם מס' מצומצם של תצפיות לא בריאות. לכן, אמנם מבחן חי בריבוע יצא מובהק, אך חלק מהנחות המבחן לא התקיימו במלואן.

ישנם עוד קשרים רבים, אך אלו הקשרים שבעיננו מוסרים הכי הרבה מידע לפני שנדבר על טרנספורמציות שונות שערכנו על משתנים ובכך יצרנו תופעות מעניינות נוספות, עליהן נרחיב בהמשך ונציג נתונים בנספחים.

כפי שצוין בהקדמה, אנו חושדים שמס' התמונות יהיה גורם אשר משפיע על משתנה המטרה, זאת בעקבות ספרות בתחום. בנוסף, לדעתנו קיום של תשלום ומצב הכלב (האם מסורס\מעוקר ומחוסן) גם הם יהוו משתנים משמעותיים, גם כן מתוך ההיגיון העסקי. זאת מכיוון שבמידה והחיה אינה מטופלת עד הסוף הדבר עלול לגרור עלויות לא מעטות עבור המאמץ ולהוות שיקול משמעותי, גם בקשרים שמצאנו היה ניתן לראות כי ישנה תלות בין משתנים אלה לבין המשתנה התלוי, ועל כן יעזרו בניבויו. בנוסף, לדעתנו בריאות הינו משתנה נוסף משמעותי, אך במקרה הנוכחי מכיוון שכמעט כל התצפיות בריאות והסיכוי האפריורי לחיה שאינה בריאה כה קטן, לא בטוח שההבדל ימלא את מלוא הפוטנציאל שיש לו מבחינת הגיון עסקי. ישנם כמובן משתנים הדורשים טיפול לפני שיתרמו, כמו לדוגמה משתנים עם קטגוריות רבות נדירות כמו מדינה, צבע, גזע משניים ומספר סרטונים. ייתכן מאוד כי חלק מהמשתנים הללו יתרמו ערך משמעותי ליכולת הניבוי רק לאחר טרנספורמציה כזו או אחרת עליה נפרט בסעיף המשך.

## איכות הנתונים

ישנם מעט מאוד נתונים חסרים, וכולם נמצאים במשתנים משניים כמו למשל color2,3 וbreed2. מכיוון שלמשתנים אלה מלכתחילה ערך של 0 עבור המשמעות "לא קיים", הוחלט להמיר את כולם ל0. הסיבה לכך הינה שאין לנו סיבה להאמין שבהכרח היה אמור להיות במקומם ערך, ומכיוון שאנו לא רוצים להכניס הטייה לנתונים נניח כי אכן לא אמור להיות ערך במקומם. מעבר לכך לא הייתה התמודדות עם ערכים חסרים. חשוב לציין כי סיבה נוספת לכך שלא הייתה התמודדות עם נתונים חסרים היא שבסופו של דבר החלטנו למחוק או לשנות משתנים אלה באופן שמבטל את הנתונים החסרים. החלטה זו תורחב לעומק בהמשך.

המשתנים היחידים בעלי ערכים קיצוניים באופן לא הגיוני היו משתנים גיל וכמות. לדוגמה, נתקלנו בכלב בן 255, גם אם המספר מייצג חודשים, מדובר בכלב בן מעל ל21, גיל לא סביר עבור כלבים, בטח לא עבור כאלה הנמסרים לאימוץ. בנוסף במאפיין כמות היה קבוצה של 20 חיות למסירה ביחד, נתון שגם הוא אינו סביר. התלבטנו לגבי אופן הטיפול בערכים קיצוניים, כיוון שלא רצינו להסירם – זאת כיוון שהמודל שנאמן עשוי להיות צריך להתמודד עם ערכים כאלו בעתיד. על כן, החלטנו להעביר את משתנה גיל טרנספורמציה למקטעים (דיסקרטיזציה) ולהכיל את כלל הערכים הקיצוניים במקטע האחרון אשר מייצג כלבים "מבוגרים" לאימוץ. את המשתנה כמות הפכנו לבינארי. נפרט לגבי הטרנספורמציות בהמשך העבודה ונפרט כיצד החלטנו לייצר את המקטעים השונים.

# הכנת הנתונים

## בחירת מאפיינים

בחלק זה, לאחר שלמדנו וניתחנו את המשתנים השונים, ביצאנו מספר פעולות על מנת לפשט את המודל ובכך את הבעיה. לשם כך ביצענו:

* השמטת מאפיינים "רועשים" או חסרי חשיבות.
* השמטת מאפיינים בעלי איכות נמוכה מדי (שגיאות, ערכים חסרים וכו').
* השמטת תצפיות בעלות חוסר רב.
* התמקדות בפלחי אוכלוסייה נבחרים.

על מנת להוריד את ממד הבעיה, חיפשנו משתנים שלא נראה בהסתכלות ראשונית שיש להם ערך, הן מבחינת סטטיסטיקה תיאורית והן מבחינת הגיון עסקי. הוחלט להוריד לחלוטין את המשתנים color 2, color 3 וbreed2. הסיבה לכך היא שמשתנים אלה מקבלים באופן כמעט מוחלט את הערך 0 (כלומר ‘אין’). בנוסף, ניתן להניח כי הצבעים המשניים אינם מהווים שיקול עיקרי בהחלטת אימוץ, דבר שעולה גם מבדיקות שערכנו, כגון טבלאות שכיחות משותפות. המשתנה breed2 מהווה גזע משני במקרה של חיה מעורבת. מכיוון שראינו כי גם המשתנה breed1 וגם breed2 מסתכמים לגזע אחד מרכזי ומגוון קטגוריות אחרות נדירות, ושהקטגוריות המשמעותיות מציגות מעורבות או גזעיות של החיה, הורדנו מאפיינים אלה ושינינו אותם למשתנה חדש, עליו נרחיב בהמשך.

בנוסף, ערכנו טרנספורמציה וביטלנו את המשתנים "חיסון", "תלוע" ו"סירוס" כיוון שראינו שיש ביניהם חפיפה משמעותית מאוד. על המשתנה החדש נרחיב בהמשך.

מעבר לכך לא היו חוסרים או נתונים בעלי איכות נמוכה שדרשו התייחסות מיוחדת.

## טיפול פרטני במאפיינים

ראשית, ערכנו דיסקרטיזציה לשני משתנים: גיל וכמות תמונות. את תהליך הדיסקרטיזציה וקביעת הסיפים יצרנו כך שנשמור על כמויות כמה שיותר שוות בתוך כל קטגוריה. על מנת לתת מקום לטווחים השונים, חילקנו כל משתנה לכמות שונה של קטגוריות, כאשר age חולק ל7 קטגוריות וPhotoAmt חולק ל3 קטגוריות. סיבה נוספת לחלוקה זו היא שראינו שנותנת פיצול משמעותי יותר בחלוקת משתנה המטרה, קצת כמו ניסיון למזער את האנטרופיה, אך בלי לחשב את המדד, אלא בהסתכלות על טבלאות שכיחות משותפות.

מעבר לתהליך הדיסקרטיזציה שינינו מספר מאפיינים:

המאפיין Health הפך למשתנה בינארי שנקרא IsHealthy, אשר מקבל ערך 1 כאשר מדובר בחיה בריאה וערך 0 כאשר מדובר בחיה לא בריאה. אמנם כך אנו מאבדים את הדקות שבין פציעה\פגיעה קלה לקשה, אך מכיוון שהיו בקטגוריות אלו כמויות מזעריות של תצפיות, הוחלט שיהוו קטגוריה משמעותית וברורה יותר יחדיו.

המאפיין Fee הפך גם הוא למשתנה בינארי HasFee, אשר מציין האם יש תשלום על האימוץ. הסיבה לכך היא שהיו רמות שונות של תשלומים ללא קוהרנטיות ולכל כמות היו תצפיות בודדות בלבד. הן מכיוון שרצינו כמות משמעותית בכל קטגוריה והן כי בעיננו לפי ההיגיון העסקי עצם קיום התשלום הוא שמשמעותי, ולא גובהו הספציפי, החלטנו לאבד את הכמות הספציפית לטובת קטגוריות ברורות יותר בעלות יכולת הכללה גבוהה.

המאפיין VideoAmt עבר תהליך דומה של הפיכה לבינארי, ושמו החדש HasVideos. הסיבה לכך היא קודם כל שכמעט כל התצפיות הינן ללא סרטונים, כך שהיו שיקולי שמירה על כמויות מינימאליות בכל קטגוריה, אך בנוסף לא ראינו מההסתכלות הראשונית שכמויות הסרטונים משנות עבור האימוץ. לעומת זאת השאלה "האם יש סרטון" מניבה פיצול טוב וחד בהרבה עבור המשתנה.

המשתנים breed 1 ו2 הפכו למשתנה אחד בינארי גם כן שנקרא IsPure אשר בוחן האם מדובר בכלב מעורב או גזעי. החלטה זו נעשתה לאחר חקר מקדים לעומק של המשתנים והבנה שבמילא עיקר מהותם להפריד בין כלבים מעורבים וגזעיים. לאחר הטרנספורמציה ניתן לראות שהמשתנה אכן מביא לפיצול טוב ומובהק של משתנה המטרה ופותר את בעיית הקטגוריות הנדירות.

המשתנה Quantity גם הוא הכיל קטגוריות קיצוניות עד לא הגיוניות ובנוסף היה בעל סקאלה מוזרה והרבה ערכים נדירים. בעיננו, לפי ההיגיון העסקי רוב המאמצים מחפשים לאמץ חיה בודדת, וייתכן שכאלה שמוכנים לאמץ מס' חיות לא מחפשים מס' ספציפי. כך זה לפחות נראה מההסתכלות הראשונית והצגה ויזואלית של הנתונים. בסופו של דבר החלטנו לאבד את הייחודיות והכמות של המשתנה לטובת מאפיין בינארי פשוט יותר שרק מציין האם מדובר באימוץ של חיה אחת או יותר ונקרא SingleAdoption.

לבסוף, היינו צריכים להחליט מה לעשות עם שלושת המשתנים "מסורס", "תלוע" ו"מחוסן". מכיוון שלערכים אלה חפיפה כה גדולה לא רצינו להשאיר את שלושתם, אך גם לא רצינו להוריד 2 מהם ברנדומאליות. לכן, החלטנו לייצר משתנה חדש המורכב מההיגיון העסקי של שלושתם. המשתנה החדש נקרא Treated והוא מקבל ערך 1 עבור מקרים שישנו 1 בכל שלושת המשתנים, 2 במקרה שבחלק החיה טופלה וחלק לא, ו3 עבור מקרים שהחיה לא עברה\לא ידוע שעברה אף טיפול. כך בעצם אידנו את שלושת המאפיינים למשתנה אחד שההיגיון מאחוריו לבחון האם החיה עברה את כלל הטיפולים, או שצריכה לעבור חלק על ידי המאמץ או את כולם.

חשוב לציין כי עבור כל המשתנים הרצנו בדיקות מובהקות וויזואליזציה מתאימה לפני ואחרי הטרנספורמציה על מנת לוודא שקיבלנו תוצאה יעילה יותר, לפחות לפי המדדים הנוכחיים, וברורה יותר מהבחינה העסקית בהתמודדות עם הבעיה. ההצגות המהותיות והנתונים מוצגים בנספחים.

## המצב החדש

להלן פירוט המשתנים לאחר כל השינויים:

**Type** – סוג בעל החיים (1 = כלב, 2 = חתול)

**Age** – גיל בעל החיים מחולק לקטגוריות לפי טווחי גיל (לא כלל הערך המקסימאלי)

1 = 0 – 1

2 = 1 – 2

3 = 2 – 3

4 = 3 – 5

5 = 5 – 11

6 = 11 – 24

7 = 24 +

**Gender** – מין בעל החיים (1 = זכר, 2 = נקבה, 3 = מעורב, כאשר הפרופיל מייצג קבוצה של בעלי חיים)

**Color1** – צבע עיקרי (ערכים בטבלת ColorLabels)

**MaturitySize** -גודל בבגרות (1 = קטן, 2 = בינוני, 3 = גדול, 4 = מאוד גדול, 0 = לא צוין)

**FurLength** – אורך הפרווה (1 = קצר, 2 = בינוני, 3 = ארוך, 4 = לא צוין)

**State** – מדינה במלזיה (ערכים בטבלת StateLabels)

**PhotoAmt** – כמות התמונות של בעל החיים שהועלו לפרופיל

**HasFee** – משתנה בינארי (0 = לא, 1 = כן)

**HasVideos** – משתנה בינארי. (0 = לא, 1 = כן)

**SingleAdoption** – משתנה בינארי. (0 = לא, 1 = כן)

**Treated** – משתנה בינארי. (0 = עבר טיפול מלא, 1 = עבר טיפול חלקי,

3 = לא עבר טיפול כלל)

**IsHealthy** – משתנה בינארי. (1 =בריא, 0 = לא בריא)

**IsPure** – משתנה בנארי. במקום משתני ה-Breed (1 = גזעי, 0 = מעורב)

## הכנת הנתונים לאימון ובחינת מערכת לומדת

על מנת לייצר את חלוקת סט הנתונים, החלטנו לחלק לפי כלל אצבע של 60% לסט אימון ו20% לאימות ובחינה, זאת על מנת לספק כמות מספקת של נתונים לאימון איכותי של המודל ולאחריו כמות שווה ובלתי תלויה עבור אימות ובחינה. הדבר החשוב ביותר עבורנו היה לייצר שלושה סטים בלתי תלויים, ולכן הדגש היה לערוך רנדומיזציה של האינדקסים לפני הפיצול ובנוסף דאגנו שלא יהיה חפיפות כלל בין הסטים. סיבה נוספת שרצינו לתת נפח גדול יותר לסט האימון הוא הכמות המצומצמת של קטגוריה 0 במשתנה המטרה, והיה לנו חשוב שהמודל יתאמן על כמות מספקת של תצפיות אשר קיבלו ערך זה על מנת להבטיח את טיב ניבויו.

# נספחים

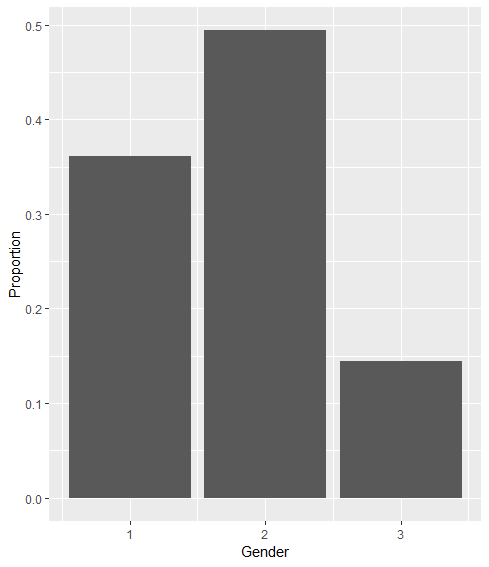
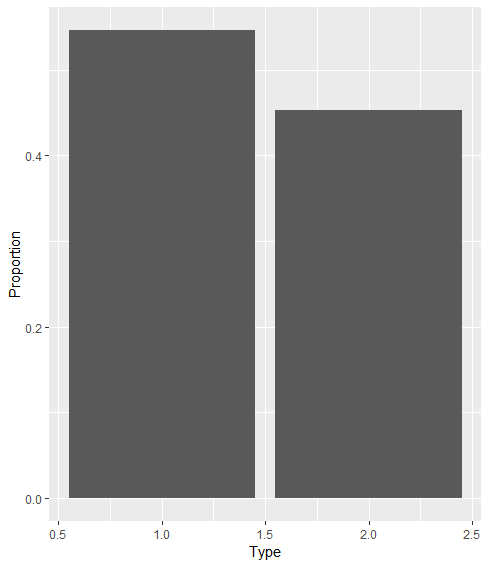
## טבלאות שכיחות

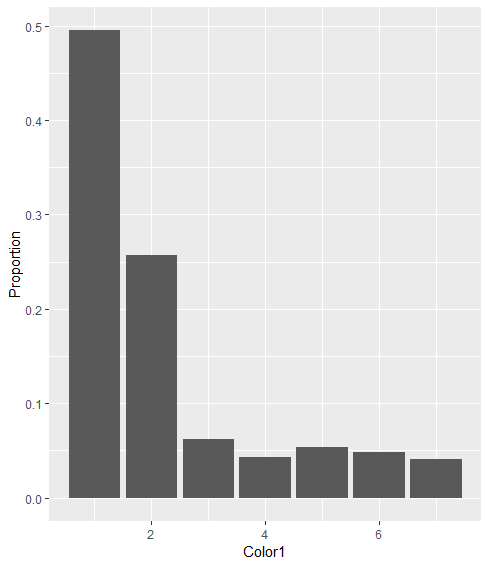
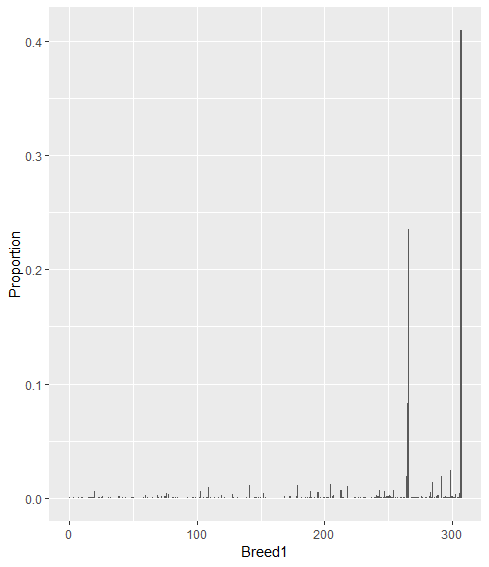


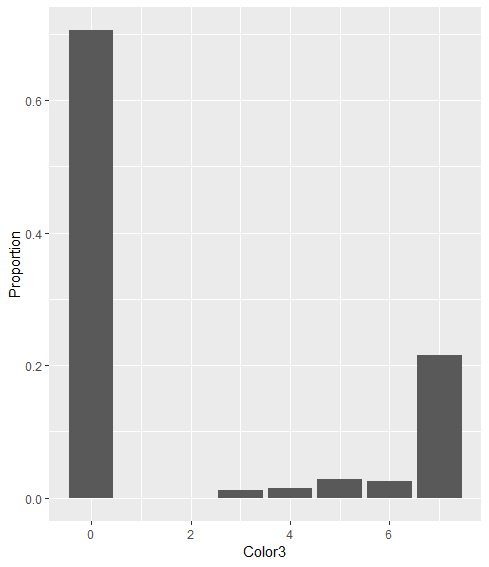
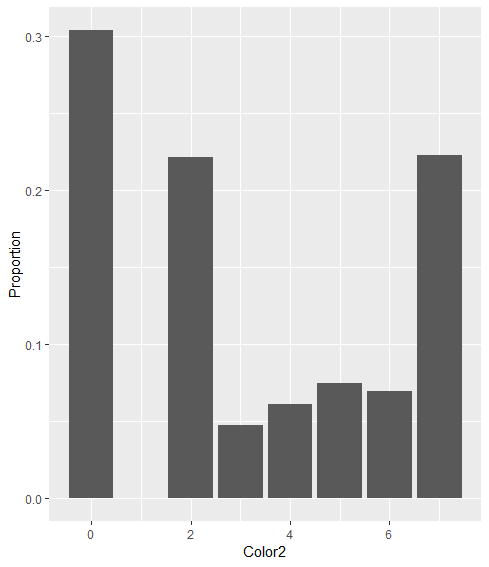


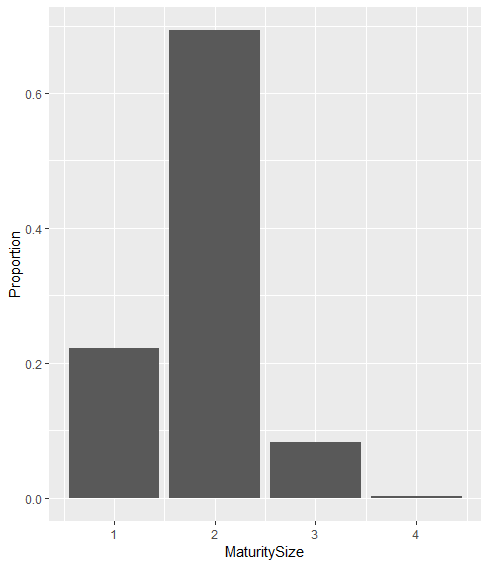
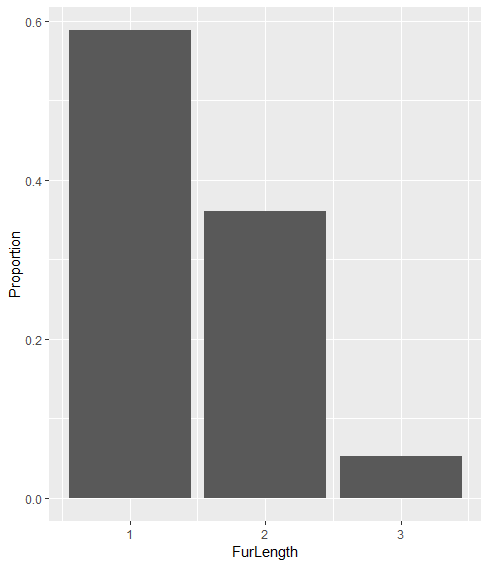


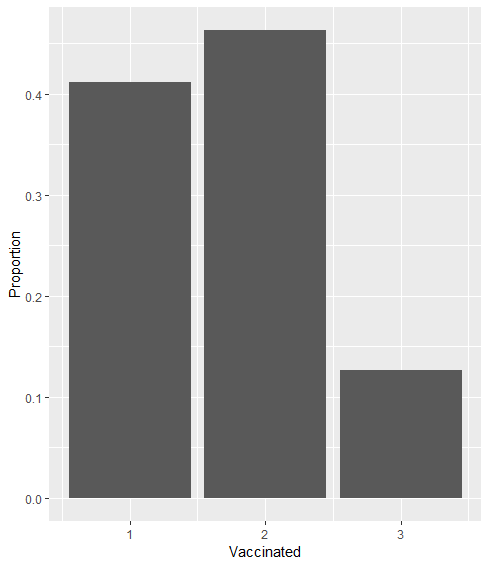
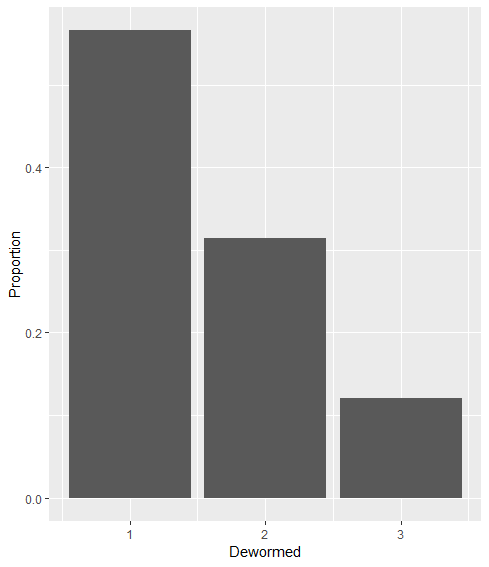
## קטגוריאלים – גרפי פורפורציה

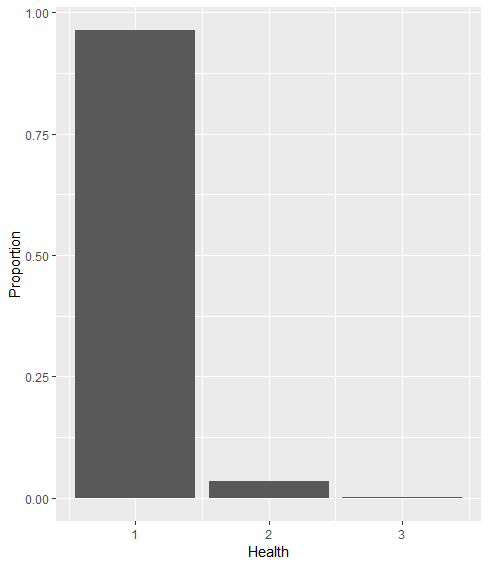
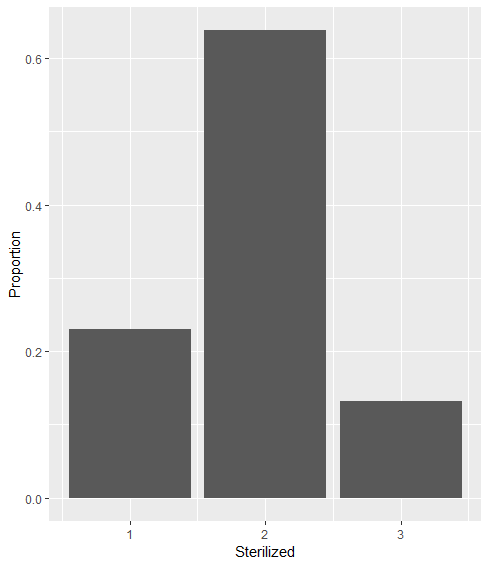


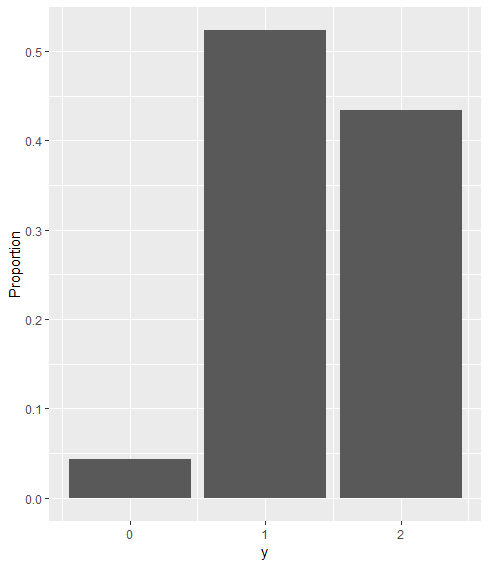
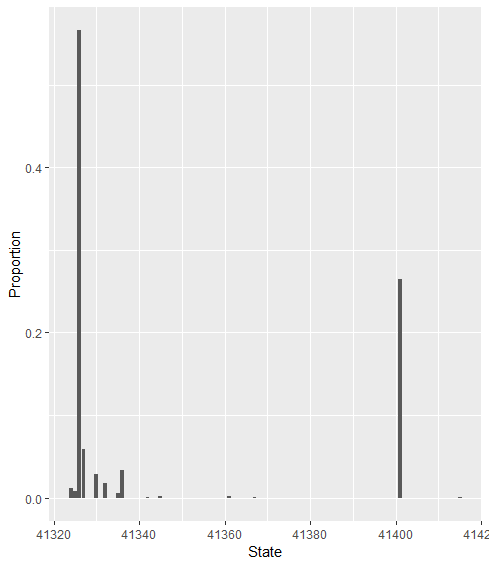




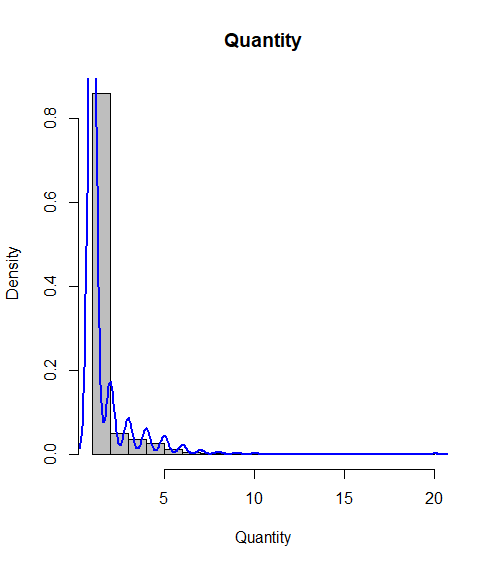
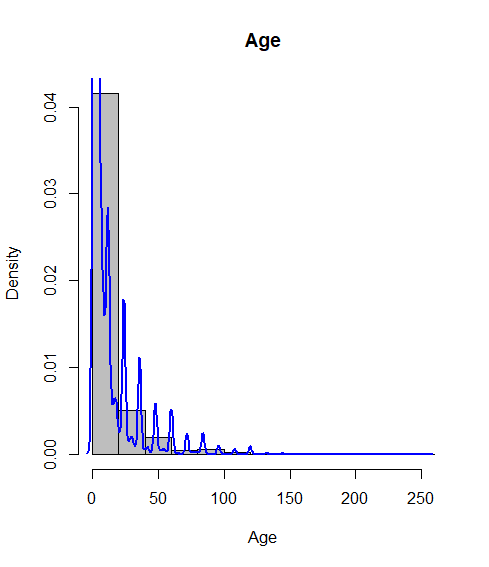
 

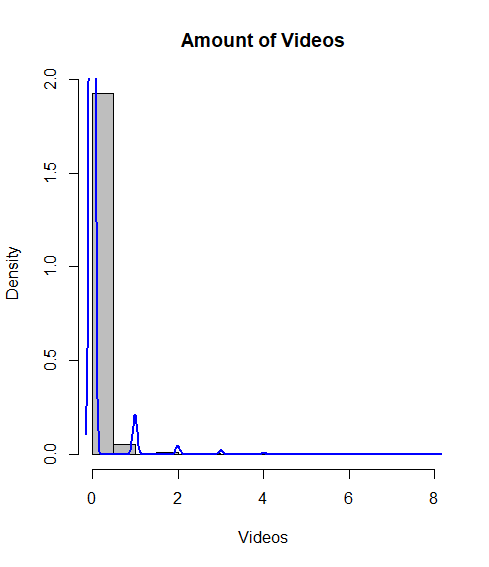
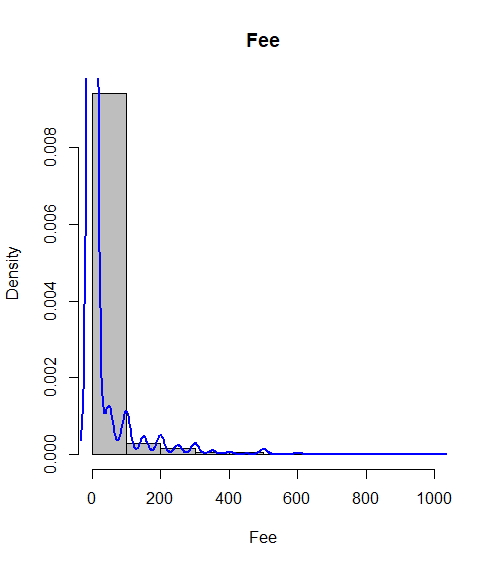
 

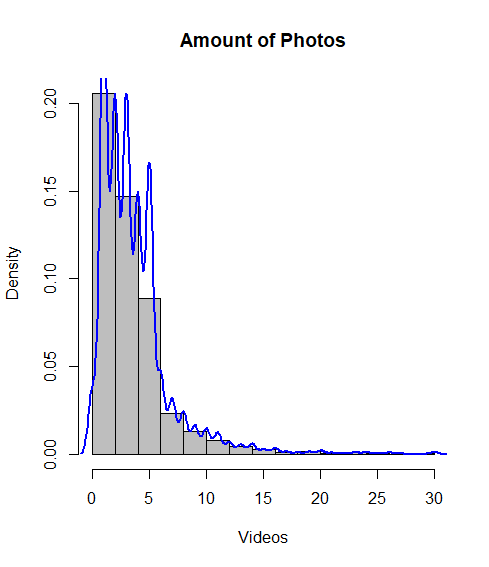




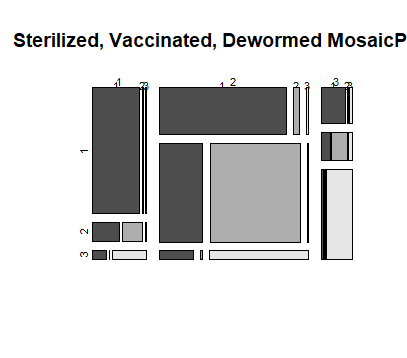
## משתנים רציפים – היסטוגרמות







## קשרים בין משתנים – גרפים וטבלאות



table(cleandata$Gender,cleandata$y) %>% prop.table

0 1 2

1 0.01691827 0.20002603 0.14445601

2 0.02108277 0.24973972 0.22345133

3 0.00468506 0.07378969 0.06585112

> chisq.test(cleandata$y,cleandata$HasFee)

Pearson's Chi-squared test

data: cleandata$y and cleandata$HasFee

X-squared = 14.895, df = 2, p-value = 0.0005828

|  |
| --- |
| table(cleandata$y,cleandata$Health) %>% prop.table    1 2 3  0 0.0407339927 0.0018219677 0.0001301406  1 0.5075481520 0.0153565851 0.0006507028  2 0.4146277980 0.0178292556 0.0013014055  > chisq.test(cleandata$Health,cleandata$y)  Pearson's Chi-squared test  data: cleandata$Health and cleandata$y  X-squared = 11.024, df = 4, p-value = 0.0263  Warning message:  In chisq.test(cleandata$Health, cleandata$y) :  Chi-squared approximation may be incorrect  A screenshot of a cell phone  Description automatically generatedA screenshot of a cell phone  Description automatically generated |
|  |
| |  | | --- | |  | |

A screenshot of a cell phone

Description automatically generated

## קשרים עם פיצ'רים חדשים לאחר טרנספורמציה:

table(cleandata$y,discretize(cleandata$PhotoAmt, breaks = 3)) %>% prop.table

[0,2) [2,4) [4,30]

0 0.01067153 0.01548673 0.01652785

1 0.10541385 0.16957314 0.24856845

2 0.12636648 0.15460698 0.15278501

> chisq.test(discretize(cleandata$PhotoAmt, breaks = 3),cleandata$y)

Pearson's Chi-squared test

data: discretize(cleandata$PhotoAmt, breaks = 3) and cleandata$y

X-squared = 133.45, df = 4, p-value < 2.2e-16

table(cleandata$y,cleandata$HasFee) %>% prop.table

0 1

0 0.037220198 0.005465903

1 0.447032795 0.076522644

2 0.357235815 0.076522644

> chisq.test(cleandata$y,cleandata$HasFee)

Pearson's Chi-squared test

data: cleandata$y and cleandata$HasFee

X-squared = 14.895, df = 2, p-value = 0.0005828

table(cleandata$PhotoAmt,cleandata$y) %>% prop.table

0 1 2

0 0.002602811 0.005075482 0.022774597

1 0.008068714 0.100338365 0.103591879

2 0.015096304 0.168662155 0.153826132

3 0.016918272 0.249479438 0.153565851

> chisq.test(cleandata$PhotoAmt,cleandata$y)

Pearson's Chi-squared test

data: cleandata$PhotoAmt and cleandata$y

X-squared = 211.7, df = 6, p-value < 2.2e-16

table(cleandata$Age,cleandata$y) %>% prop.table

0 1 2

1 0.001041124 0.005596044 0.003253514

2 0.005465903 0.094742322 0.035007808

3 0.012363352 0.137818844 0.063899011

4 0.006507028 0.111660593 0.084200937

5 0.004945341 0.063248308 0.081988548

6 0.006376887 0.042946382 0.075871942

7 0.005986465 0.067542946 0.089536700

> chisq.test(cleandata$PhotoAmt,cleandata$y)

Pearson's Chi-squared test

data: cleandata$PhotoAmt and cleandata$y

X-squared = 211.7, df = 6, p-value < 2.2e-16

table(cleandata$IsPure,cleandata$y) %>% prop.table

0 1 2

0 0.008849558 0.189354503 0.211478397

1 0.033836543 0.334200937 0.222280062

> chisq.test(cleandata$IsPure,cleandata$y)

Pearson's Chi-squared test

data: cleandata$IsPure and cleandata$y

X-squared = 177.45, df = 2, p-value < 2.2e-16

table(cleandata$IsHealthy,cleandata$y) %>% prop.table

0 1 2

0 0.001952108 0.016007288 0.019130661

1 0.040733993 0.507548152 0.414627798

> chisq.test(cleandata$IsPure,cleandata$y)

Pearson's Chi-squared test

data: cleandata$IsPure and cleandata$y

X-squared = 177.45, df = 2, p-value < 2.2e-16

table(cleandata$Treated,cleandata$y) %>% prop.table

0 1 2

1 0.005856325 0.070145757 0.106194690

2 0.017438834 0.242581989 0.165408641

3 0.019390942 0.210827694 0.162155128

> chisq.test(cleandata$Treated,cleandata$y)

Pearson's Chi-squared test

data: cleandata$Treated and cleandata$y

X-squared = 162.98, df = 4, p-value < 2.2e-16

0 1 2

0 0.04034357 0.50169183 0.42074440

1 0.00234253 0.02186361 0.01301406

> chisq.test(cleandata$HasVideos,cleandata$y)

Pearson's Chi-squared test

data: cleandata$HasVideos and cleandata$y

X-squared = 10.012, df = 2, p-value = 0.006697

## קוד R

### קובץ איתחול

library(magrittr)

library(data.table)

library(ggplot2)

library(tidyverse)

library(MASS)

library(arules)

library(sqldf)

"Yoav's Computer Directory"

dataset <- fread("C:\\Users\\BIGVU\\Documents\\ML Project\\Data.csv")

"Yaron's Computer Directory"

dataset <- fread("C:\\Users\\Yaron\\Desktop\\working proccess\\Academy\\5\\ML\\projectML\\Data.csv")

### קובץ ראשי

---

output: html\_document

editor\_options:

chunk\_output\_type: console

---

First, let's look at the different variables, compute new relevant ones and tidy up.

The proportion of each group in the dataset, in a table and then barchart. This part will

focus on each varialbe individually, and specifically categorial data.

```{r}

dataset$Type %>% table %>% prop.table()

ggplot(dataset, aes(Type)) + geom\_bar(aes(y = (..count..)/sum(..count..))) + ylab('Proportion')

dataset$Gender %>% table %>% prop.table()

ggplot(dataset, aes(Gender)) + geom\_bar(aes(y = (..count..)/sum(..count..))) + ylab('Proportion')

dataset$Breed1 %>% table %>% prop.table()

ggplot(dataset, aes(Breed1)) + geom\_bar(aes(y = (..count..)/sum(..count..))) + ylab('Proportion')

dataset$Color1 %>% table %>% prop.table()

ggplot(dataset, aes(Color1)) + geom\_bar(aes(y = (..count..)/sum(..count..))) + ylab('Proportion')

ggplot(dataset, aes(Color2)) + geom\_bar(aes(y = (..count..)/sum(..count..))) + ylab('Proportion')

ggplot(dataset, aes(Color3)) + geom\_bar(aes(y = (..count..)/sum(..count..))) + ylab('Proportion')

dataset$Breed2 %>% table %>% prop.table()

dataset$MaturitySize %>% table %>% prop.table()

ggplot(dataset, aes(MaturitySize)) + geom\_bar(aes(y = (..count..)/sum(..count..))) + ylab('Proportion')

dataset$FurLength %>% table %>% prop.table()

ggplot(dataset, aes(FurLength)) + geom\_bar(aes(y = (..count..)/sum(..count..))) + ylab('Proportion')

dataset$Vaccinated %>% table %>% prop.table()

ggplot(dataset, aes(Vaccinated)) + geom\_bar(aes(y = (..count..)/sum(..count..))) + ylab('Proportion')

dataset$Dewormed %>% table %>% prop.table()

ggplot(dataset, aes(Dewormed)) + geom\_bar(aes(y = (..count..)/sum(..count..))) + ylab('Proportion')

dataset$Sterilized %>% table %>% prop.table()

ggplot(dataset, aes(Sterilized)) + geom\_bar(aes(y = (..count..)/sum(..count..))) + ylab('Proportion')

dataset$Health %>% table %>% prop.table()

ggplot(dataset, aes(Health)) + geom\_bar(aes(y = (..count..)/sum(..count..))) + ylab('Proportion')

dataset$State %>% table %>% prop.table()

ggplot(dataset, aes(State)) + geom\_bar(aes(y = (..count..)/sum(..count..))) + ylab('Proportion')

dataset$y %>% table %>% prop.table()

ggplot(dataset, aes(y)) + geom\_bar(aes(y = (..count..)/sum(..count..))) + ylab('Proportion') # Almost no cases of 0. We need to take this into consideration when we split into Train and Validation

```

Now the numeric variables. I'm not sure it's really good to show these with Histograms, as they are integers, but whatever. It gives the general idea.

```{r}

hist(dataset$Age,prob=T, main='Age',xlab = 'Age',col="grey")

lines(density(dataset$Age),col="blue",lwd=2) # Some of the outliers here are disturbing. We will remove them and create this again later

hist(dataset$Quantity,prob=TRUE, main='Quantity',xlab = 'Quantity',col="grey")

lines(density(dataset$Quantity),col="blue",lwd=2) #Same here. let's see exactly how many we have above 8

dataset[Quantity > 8, .N,] # only 49, out of 7684. We'll need to decide what to do with these profiles.

dataset[Quantity == 1, .N/7684,] # i think we shuld focus on the pets that not in groups and do the prediction on them.

dataset[Quantity > 1 & dataset$y > 1, .N/7684,] #test to see how much goupe pets DONT get adopted

dataset[Quantity == 1 & dataset$y > 1, .N/7684,] #test for the amount of NOT adopted pets that come in group of one.

hist(dataset$Fee,prob=TRUE, main='Fee',xlab = 'Fee',col="grey")

lines(density(dataset$Fee),col="blue",lwd=2) # This is also problematic. Most are zeroes and then disperesed. We'll probably be best to create a binary variable indicating a fee existed

dataset[Fee == 0 , .N/7684,] # 84.1% of pets in the dataset Don't have adoption fees.

dataset[Fee <= 100 , .N/7684,]

hist(dataset$VideoAmt,prob=T, main='Amount of Videos',xlab = 'Videos',col="grey")

lines(density(dataset$VideoAmt),col="blue",lwd=2) #Also here, most profiles have 0 videos. Might be better as a binary variable indicating the existent of videos

dataset[VideoAmt == 0 , .N/7684,] # 96.3% of pets in the dataset have 0 videos in their profiles.

dataset[VideoAmt != 0 & dataset$y <=1 , .N/7684,]

hist(dataset$PhotoAmt,prob=TRUE, main='Amount of Photos',xlab = 'Videos',col="grey")

lines(density(dataset$PhotoAmt),col="blue",lwd=2) #This is more interesting. Seems like you usually have 1 picture, but it is rather common to have more.

dataset[PhotoAmt > 1 & dataset$y <2, .N/7684,]

dataset[PhotoAmt == 1 , .N/7684,]

dataset[PhotoAmt == 2 , .N/7684,]

dataset[PhotoAmt == 3 , .N/7684,]

dataset[PhotoAmt == 4 , .N/7684,]

dataset[PhotoAmt == 5 , .N/7684,]

dataset[PhotoAmt != 0 , .N/7684,]

```

Time for some boxplots. Let's see how valuable are our numeric variables

```{r}

boxplot(dataset$Age~dataset$y, main = 'Age~Target variable Boxplot') # Not extremely helpful. The "not adopted" category is slightly older, but not enough.

boxplot(dataset$Quantity~dataset$y, main = 'Quantity~Target variable Boxplot') #This is a bad variable, so obviously it is also bad here. In general it seems people are less inclined to adopt pets in groups, which makes sense.

boxplot(dataset$Fee~dataset$y) # Really not helpful. Again, we will turn this into a binary variable and see if that makes a difference.

boxplot(dataset$VideoAmt~dataset$y) #Same as Fee.

boxplot(dataset$PhotoAmt~dataset$y, main = 'PhotoAmt~Target variable Boxplot') # It seems more photos are slightly helpful for increasing the odds for adoption, but only slightly.

```

Let's make sure we only have NA on secondary variables

```{r}

cleandata <- dataset

cleandata[is.na(cleandata),] <- 0 #SWAP ALL NA TO 0

# cleandata <- cleandata[complete.cases(cleandata), ] #clearrows with NA

cleandata [is.na(Type),.N] # I ran this line of code 20 times, for each variable manually, because I couldn't figure out how to do it more efficiantly.

```

Indeed that is the case, so we can replace every NA with 0

```{r}

# Again, couldn't figure out a better way. I go through each column and change only the NA with 0.

for (i in names(cleandata)) {

cleandata[is.na(get(i)), (i):=0]

}

```

Creating the new binary variables from Variables that are currently extremely odd shaped and unhelpful. Also removing old ones.

```{r}

cleandata[,HasFee := ifelse(Fee>0,1,0), ] #84% of pets don't have adoption fee. In order for the those who have fees those who have fees to become a meaningful group we are turning this Binary

cleandata[,HasVideos := ifelse(VideoAmt>0,1,0), ] #96% of pets don't have videos on their profiles. This is why it is more interesting to ask if having videos at all make any difference, hence we turn this Binary as well.

cleandata[,SingleAdoption := ifelse(Quantity==1,1,0), ]

# cleandata[,isPure := ifelse(Breed2==0,0,1), ]

cleandata <- cleandata [,-'Fee',]

cleandata <- cleandata [,-'VideoAmt',]

cleandata <- cleandata [,-'Quantity',]

cleandata <- cleandata [,-'Breed2',]

cleandata <- cleandata [,-'Color2',]

cleandata <- cleandata [,-'Color3',]

# marge 3 colums on health condition into new veriable: 1= Treated, 2= semi treated, 3= not treated

cleandata[,Treated:= sqldf('

SELECT (CASE

WHEN Vaccinated = 1 AND Dewormed = 1 AND Sterilized = 1 THEN 1

WHEN Vaccinated = 1 OR Dewormed = 1 OR Sterilized = 1 THEN 2

ELSE 3

END

) AS Treated

FROM cleandata

')]

# tranpose breed lable to binari values: 0 = mixed, 1 = pedigree

cleandata[,IsPure := ifelse(dataset$Breed1 ==307,0,1), ]

# transpose health lable to binary values: 1 = healthy, 0 = not

cleandata[,IsHealthy := ifelse(Health ==1,1,0), ]

cleandata <- cleandata [,-'Health',]

cleandata <- cleandata [,-'Breed1',]

cleandata <- cleandata [,-'Vaccinated',]

cleandata <- cleandata [,-'Dewormed',]

cleandata <- cleandata [,-'Sterilized',]

cleandata[,State:= sqldf('

SELECT (CASE

WHEN State = 41326 THEN 1

WHEN State = 41401 THEN 2

ELSE 0

END

) AS State

FROM cleandata

')]

```

Connections/correlation between variables

```{r}

cleandata %>% names

mosaicplot(table(cleandata$Sterilized,cleandata$Vaccinated,cleandata$Dewormed), color = T, main = 'Sterilized, Vaccinated, Dewormed MosaicPlot')

table(cleandata$Gender,cleandata$y) %>% prop.table

chisq.test(cleandata$y,cleandata$HasFee)

table(cleandata$MaturitySize,cleandata$y) %>% prop.table

chisq.test(cleandata$MaturitySize,cleandata$FurLength)

table(cleandata$Gender,cleandata$y) %>% prop.table

chisq.test(cleandata$y,cleandata$HasFee)

table(cleandata$y,cleandata$HasFee) %>% prop.table

chisq.test(cleandata$Gender,cleandata$y)

table(cleandata$y,cleandata$Health) %>% prop.table

chisq.test(cleandata$Health,cleandata$y)

table(cleandata$y,discretize(cleandata$PhotoAmt, breaks = 3)) %>% prop.table

chisq.test(discretize(cleandata$PhotoAmt, breaks = 3),cleandata$y)

table(cleandata$HasVideos,discretize(cleandata$PhotoAmt, breaks = 3)) %>% prop.table

chisq.test(discretize(cleandata$PhotoAmt, breaks = 3),cleandata$y)

```

discretizing continous variables

```{r}

discretize(cleandata$Age, breaks = 7)

# tanspose age valuse to categorial bin according to discretize function

cleandata[,Age:= sqldf('

SELECT (CASE

WHEN Age < 1 THEN 1

WHEN Age < 2 AND Age >=1 THEN 2

WHEN Age < 3 AND Age >=2 THEN 3

WHEN Age < 5 AND Age >=3 THEN 4

WHEN Age < 11 AND Age >=5 THEN 5

WHEN Age < 24 AND Age >=11 THEN 6

ELSE 7

END

) AS Age

FROM cleandata

')]

discretize(cleandata$PhotoAmt, breaks = 3)

# transpose PhotiAmt valuse to categorial bin according to discretize function

cleandata[,PhotoAmt:= sqldf('

SELECT (CASE

WHEN PhotoAmt = 0 THEN 0

WHEN PhotoAmt < 2 AND Age >0 THEN 1

WHEN PhotoAmt < 4 AND Age >=2 THEN 2

ELSE 3

END

) AS PhotoAmt

FROM cleandata

')]

```

Let's see how the new features are working

```{r}

table(cleandata$PhotoAmt,cleandata$y) %>% prop.table

chisq.test(cleandata$PhotoAmt,cleandata$y)

table(cleandata$Age,cleandata$y) %>% prop.table

chisq.test(cleandata$PhotoAmt,cleandata$y)

table(cleandata$IsPure,cleandata$y) %>% prop.table

chisq.test(cleandata$IsPure,cleandata$y)

table(cleandata$IsHealthy,cleandata$y) %>% prop.table

chisq.test(cleandata$IsPure,cleandata$y)

table(cleandata$Treated,cleandata$y) %>% prop.table

chisq.test(cleandata$Treated,cleandata$y)

table(cleandata$HasVideos,cleandata$y) %>% prop.table

chisq.test(cleandata$HasVideos,cleandata$y)

```

Splitting the data-set

```{r}

# Compute sample sizes.

sampleSizeTraining <- floor(0.6 \* nrow(cleandata))

sampleSizeValidation <- floor(0.2 \* nrow(cleandata))

sampleSizeTest <- floor(0.2 \* nrow(cleandata))

# Create the randomly-sampled indices for the dataframe. Use setdiff() to

# avoid overlapping subsets of indices.

indicesTraining <- sort(sample(seq\_len(nrow(cleandata)), size=sampleSizeTraining))

indicesNotTraining <- setdiff(seq\_len(nrow(cleandata)), indicesTraining)

indicesValidation <- sort(sample(indicesNotTraining, size=sampleSizeValidation))

indicesTest <- setdiff(indicesNotTraining, indicesValidation)

# Finally, output the three dataframes for training, validation and test.

cleandataTraining <- cleandata[indicesTraining, ]

cleandataValidation <- cleandata[indicesValidation, ]

cleandataTest <- cleandata[indicesTest, ]

table(cleandataTraining$y) %>% prop.table()

```