לימוד מכונה 364-1-1811

**פרויקט - חלק א'**

זהו החלק הראשון מבין שני חלקים.

**הנחיות הגשה:** **דו"ח התרגיל הראשון וקוד ה-R שכתבתם יוגשו לתיבת ההגשה במודל עד לתאריך ה-5.5.2019 בשעה 23:55. מספיקה הגשה של אחד מבני הזוג.**

משקל התרגיל: 15% מהציון הסופי בקורס.

מטרת התרגיל: בתרגיל זה נתרגל טעינת, עיבוד, הצגת וניתוח נתונים. פעולות אלה גם ישמשו אותנו בהבנת והכנת הנתונים לקראת התרגיל השני בו נשתמש בנתונים לאימון ובחינה של מערכות לומדות והשוואה ביניהן.

צוותי הגשה: הגשת התרגיל הינה בשלשות.

דגשים לדו"ח: אורך הדו"ח לא יעלה על 10 עמודים (לא כולל הקודים שנכתבו ועמודים נלווים כמו שער ותוכן עניינים), בגודל כתב 12, פונט Arial ורווח של שורה וחצי. חריגה ממספר עמודים זה תגרור הורדת נק'. יש לשמור על תמציתיות ולהתמקד בתובנות המרכזיות שלכם בכל סעיף.

שפת תכנות: R (המותקנת במעבדות 040, 041, 042).

מאגר הנתונים לתרגילים: הפרויקט יתבצע על מאגר נתונים של אתר לאימוץ חיות מחמד (כלבים וחתולים).

<https://www.petfinder.my/>

הסבר אודות בסיס הנתונים ב-**Moodel**:

Data – נתונים עבור אימון המערכת. כולל משתנים המסבירים ומשתנה המטרה. **זהו בסיס הנתונים שישמש אתכם לפרויקט.**

Test – משמש לטובת הגשת החיזוי הסופי. כולל משתנים מסבירים בלבד. בחלק זה אינכם צריכים להגיש עדיין חיזויים, אבל הקובץ פורסם כדי שתוכלו לראות איך נראות הרשומות עבורן יש לבצע חיזוי (זה חשוב בעיקר לטובת ההחלטה איך להתמודד עם נתונים חסרים וכד').

הסבר למשתנים בבסיס הנתונים:

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

Age – גיל בעל החיים (לעשות קטגוריאלי לפי מקטעים)

Breed1 – גזע עיקרי (ערכים בטבלת BreedLabels) (בינארי גזעי/לא גזעי – 307 = מעורב)

Breed2 – גזע משני (ערכים בטבלת BreedLabels, 0 = לא קיים) (הפךלמשתנה בינאי העיד של גזעיות או לא)

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) מעיפים

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

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) => איחוד של שלושתם!

Health – מצב בריאותי (1 = בריא, 2 = פציעה מינורית, 3 = פציעה משמעותית, 4 = לא צוין) – (בינארי 0 בריא , 1- לא בריא)

Quantity – מספר החיות המוצגות בפרופיל ( בינארי)

Fee – מחיר האימוץ (0 = חינם) ( בינארי)

State – מדינה במלזיה (ערכים בטבלת StateLabels) (בינארי 3 ערכים – מדינה גדולה , בינונית וכל השאר)

VideoAmt – כמות סרטוני וידאו של בעל החיים שהועלו לפרופיל – (בנארי – יש/אין)

PhotoAmt – כמות התמונות של בעל החיים שהועלו לפרופיל (

y – "מהירות האימוץ" (0 = בעל החיים נלקח לאימוץ באותו יום שנרשם, 1 = בין יום לשלושה חודשים, 2 = לא נלקח לאימוץ במהלך השלושת החודשים הראשונים).

הקדמה:

* **ההוראות בתרגיל זה הינן כלליות (כלומר לא נכתבו ספציפית עבור בסיס הנתונים הנתון) ומטרתן להנחות אתכם כיצד יש לנתח ולעבוד עם כל בסיס נתונים שהוא. ייתכן כי שאלה מנחה כלשהי המופיעה בהוראות לא רלוונטית או ניתנת למענה עבור בסיס הנתונים הנתון. במקרה כזה, הסבירו מדוע לא ניתן/רלוונטי לענות עליה.**
* מבנה התרגיל חופף חלקית למבנה מערכת כריית נתונים כפי המופיע במדריך CRISP-DM (ראה באתר הקורס). היעזרו במדריך זה כדי להעשיר את עבודתכם.
* בתרגיל יושם דגש על שימוש בתוכנה לצורך מענה על שאלות הקשורות בנתונים ובניתוחם. לא פחות חשובות הן התובנות משימוש זה לגבי עולם התוכן של הבעיה הנחקרת, כשהשאלה המרכזית הינה: מה בעצם למדנו מתרגיל זה? באפשרותכם לשלב בדו"ח העבודה screenshots של פלטים, סקריפטים של R, טבלאות, גרפים וכו', לנתחם ולהשליך מהם על עולם התוכן הנחקר.
* עבור כל פעולה שבוצעה בתוכנה יש גם לתעד: מדוע ביצענו אותה? מה התקבל בפלט? מה למדנו מהפלט? מהן ההשלכות על השלבים הבאים?
  + דוגמה: ע"ס ידע אישי, חשדנו שישנו קשר חזק בין שני משתנים🡨בחרנו להציג מתאם זה כדי לבחון את הקשר ביניהם🡨הגרף מראה מתאם גבוה🡨ניתן ללמוד כי שני המשתנים מתארים את אותה התופעה🡨נמחק את אחד מהמשתנים, כך שבשלב הבא נוכל להשיג מודל פשוט יותר.
* יש להגיש את כל קבצי R עליהם עבדתם.
* שאלותבנוגע לתרגיל, יש לפרסם בפורום הייעודי שייפתח במודל.

**מבנה העבודה:**

**הגדרת הבעיה (8 נק')**

1. **תיאור כללי של עולם התוכן הנחקר**

* מהי הבעיה המחקרית? מה עשו מחקרים קודמים שעסקו בנושא כדי להתמודד עם הבעיה?

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/>)

1. **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.

משתנה המטרה בעבודה שלנו הוא "מהירות האימוץ", המוגדר על ידי שלושה ערכים:

0 = בעל החיים נלקח לאימוץ באותו יום שנרשם

1 = בין יום לשלושה חודשים

2 = לא נלקח לאימוץ במהלך השלושת החודשים הראשונים

**הבנת הנתונים (53 נק')**

1. **תיעוד מקורות הנתונים ומשמעותם** (אם יש צורך, היעזרו במידע מהאינטרנט)

* מהו מקור הנתונים במאגר אתו אתם עובדים וכיצד הנתונים נוצרו (למשל: מדידות חיישנים, נתונים סטטיסטיים, ידנית, נתוני מומחה, וכו')?
* מהם המאפיינים המופיעים בסט הנתונים? מה משמעותם? מהו משתנה המטרה ואיזה ערכים הוא מקבל?(לעשות העתק הדבק מההוראות?? לא כזה הבנתי מה מצפים מאיתנו...)

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. לסיכום, ייתכן כי קיימת בעיה באיכות הנתונים מהבחינה שהתגיות נעשות על ידי אנשים, ולא מומחים. מצד שני לפחות ניתן לצפות לעקביות שכן כל המידע ממקור אחד והוזן בדרך אחידה.

1. **הסתברויות אפריוריות וקשרים בין מאפיינים**

* מהן ההסתברויות האפריוריות של משתנה המטרה ושל שאר המאפיינים? על מה מלמדות הסתברויות אלה? עבור משתנים רציפים ניתן להשתמש בהיסטוגרמה.

## 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.

בנוסף, קיימת מדינה בטבלת מדינות בשם Perlis שאינה מופיעה כלל בטבלת הנתונים ולא באתר. מכיוון שישנם מספר רב של מדינות בעלות מעט תצפיות, בהמשך נחליט כיצד להפוך את המשתנה ליותר רלוונטי.

## משתנים רציפים

במשתנים הרצים 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. (יעבור לסעיף הבא)

**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.

* האם סט הנתונים מאוזן? האם לדעתכם הוא מייצג את המציאות?

כפי שניתן לראות, 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 באתר המלזי.

* חפשו והציגו קשרים "מעניינים" בין מאפיינים - צפויים ולא-צפויים. הסבירו המשמעות של קשרים אלה.

## 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.

* מהם המאפיינים בהם ניתן "לחשוד" כבעלי השפעה על משתנה המטרה (ע"ס ידע אישי/מוקדם, מחקרים, ניסיון בתחום וכו'), ולפיכך כנראה יבחרו ע"י המערכות הלומדות שנבחן בתרגיל השני?

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.

1. **איכות הנתונים**

* האם ישנם נתונים חסרים? אם כן, מה ניתן לומר על עליהם?

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.

* האם ישנם נתונים שאינם הגיוניים? מה אתם מציעים לעשות עם נתונים אלה?

אם יש לכם תובנות נוספות לגבי הנתונים, זה המקום להציגן.

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 (32 נק')**

1. **על פי הצורך, בצעו ונמקו בחירת מאפיינים שביצעתם**

* 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.

1. **על פי הצורך, תנו טיפול פרטני במאפיינים**

* דיסקרטיזציה של משתנים רציפים, למשל בצורה שמייצגת את ההתפלגות או בצורה שמייצגת דרישות של עולם התוכן.
  + - דוגמה: משתנה רציף "גיל" ניתן להפוך למשתנה בדיד ע"י הגדרת סיפים כגון 6, 18, 21, 65 וכו'.
* גזירת מאפיינים חדשים (פונקציות של משתנים קיימים).

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.

1. **הכנת הנתונים לאימון ובחינת מערכת לומדת**

* לאחר העיבוד שביצעתם קודם, חלקו את סט הנתונים המלא לסטי אימון, אימות ובחינה. כתבו מה השיקולים שלכם באופן החלוקה שבחרתם? מה מבטיחה החלוקה שבחרתם לגבי אימון ובחינת מערכות לומדות?

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.

**איכות הדו"ח ורמת שימוש בתכנת R (7 נק')** – נק' אלו ינתנו בהתאם להערכה כללית בנוגע לאיכות הדו"ח ובהתאם לרמת השימוש בתכנת ה-R (האם נעשה שימוש מקיף/חלקי/כלל לא).

בהצלחה !

# נספחים

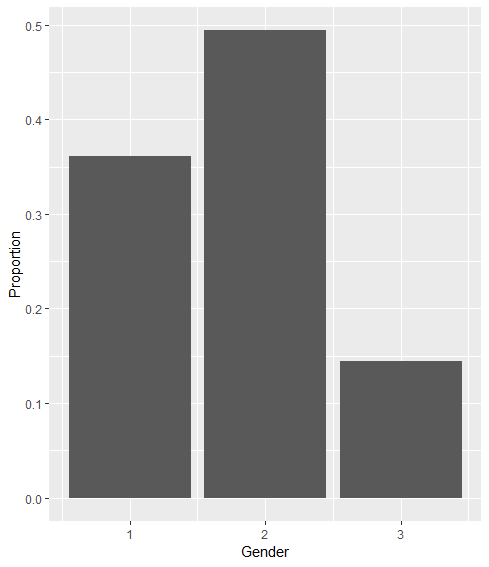
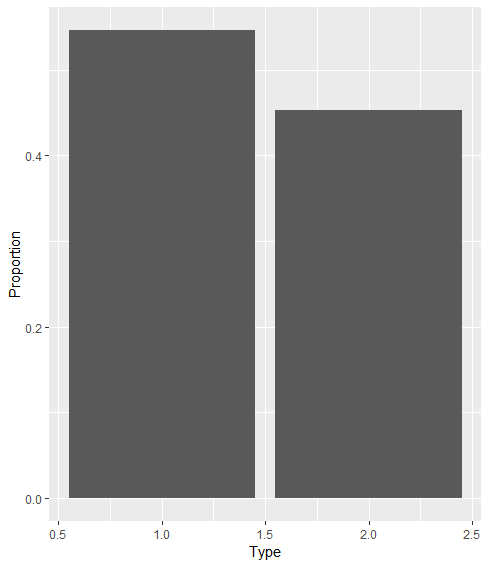
## טבלאות שכיחות

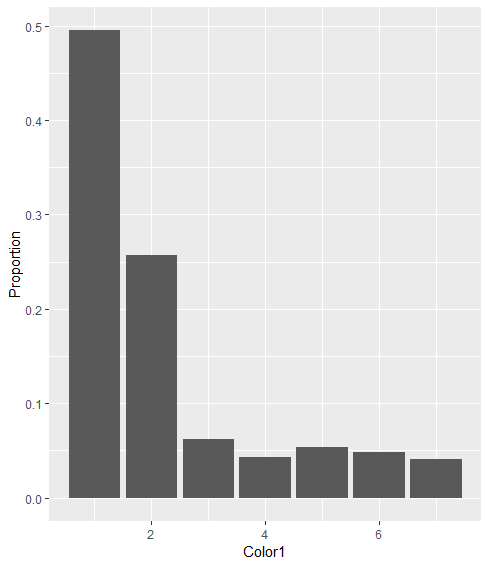
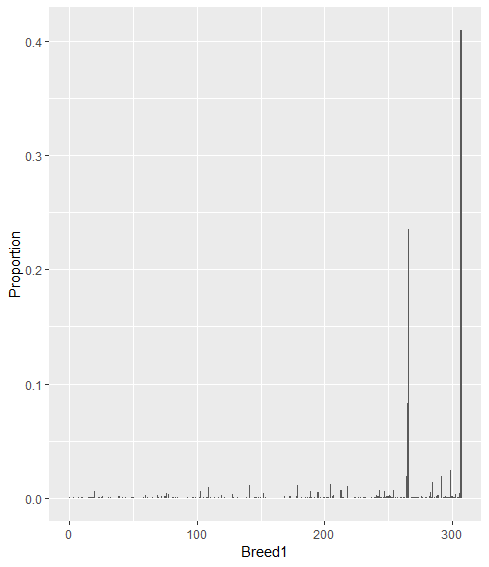


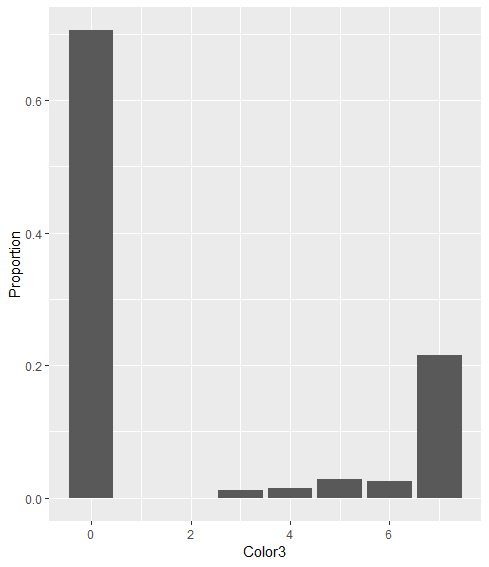
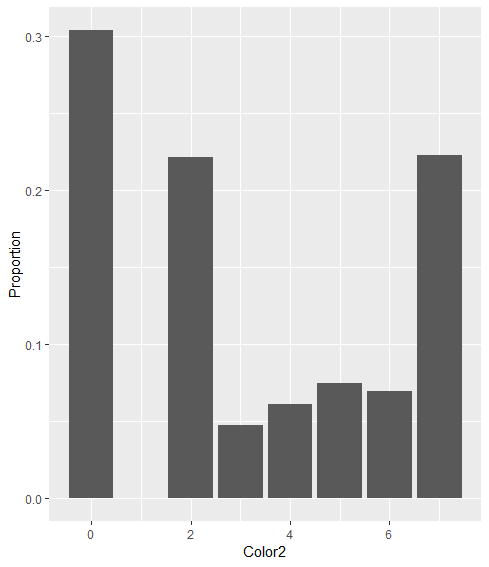


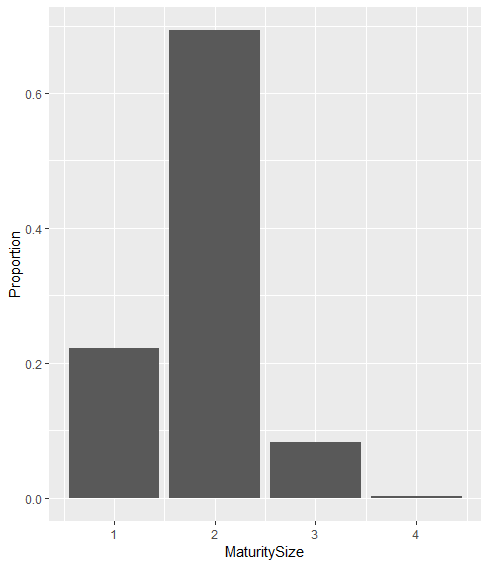
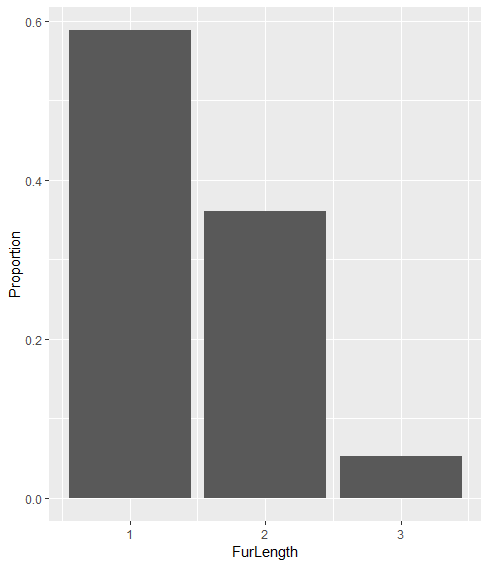


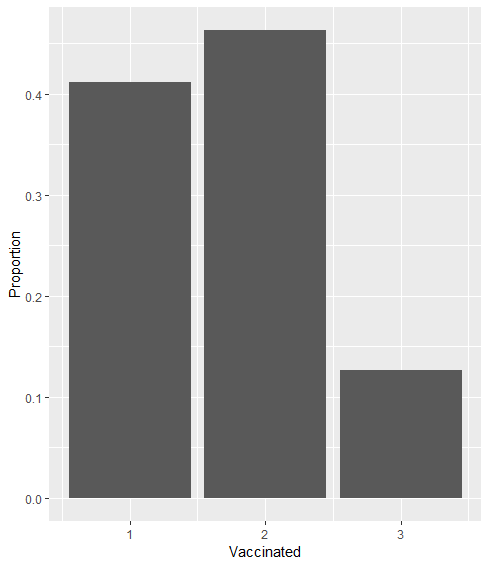
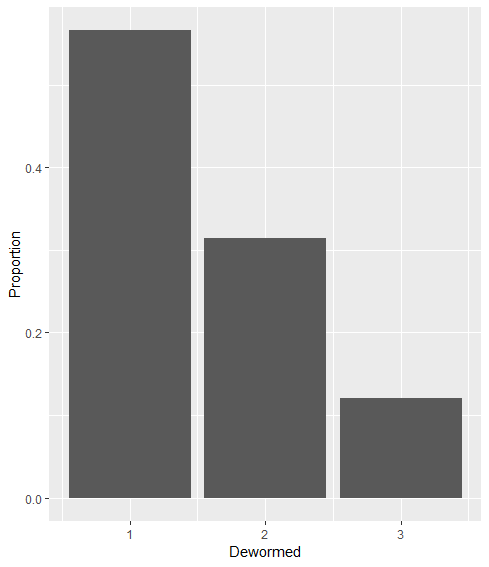
## קטגוריאלים – גרפי פורפורציה

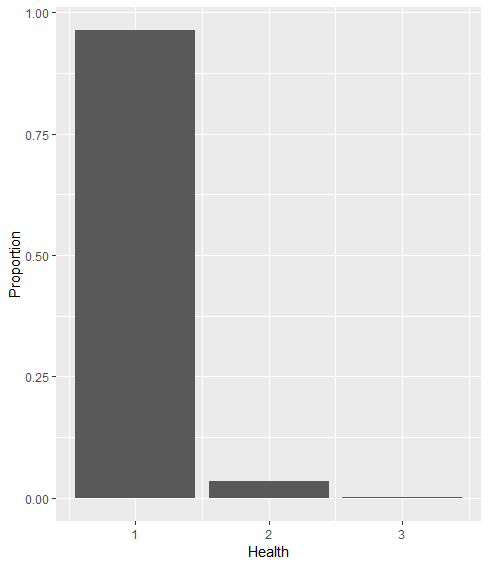
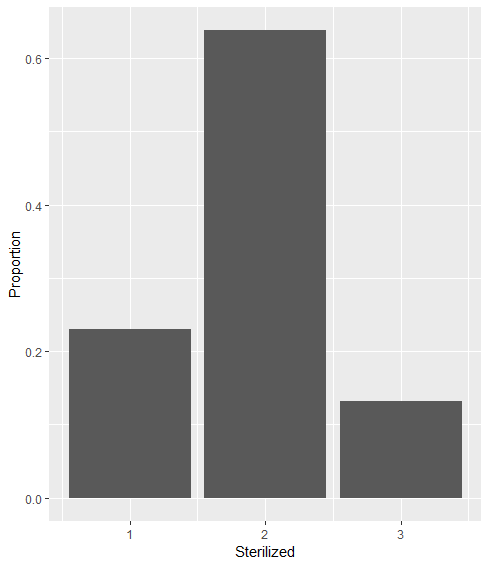


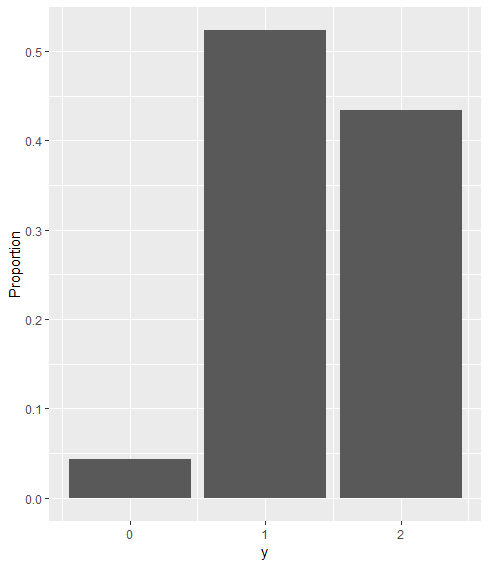
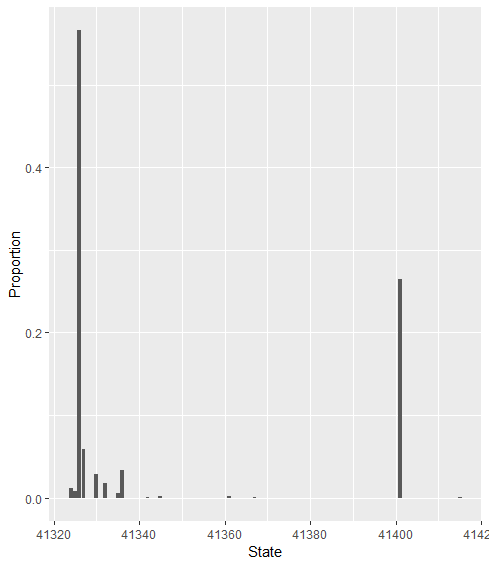




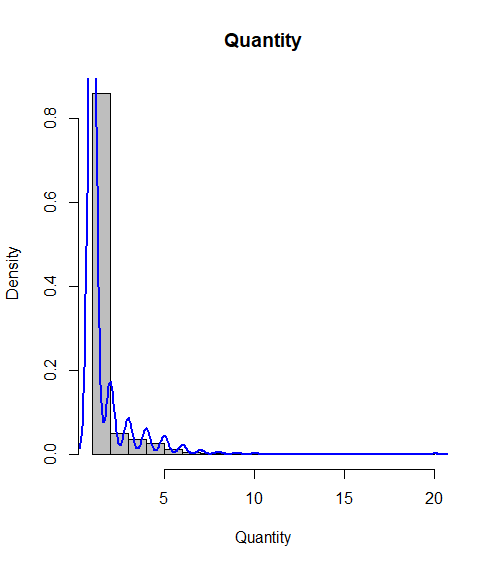
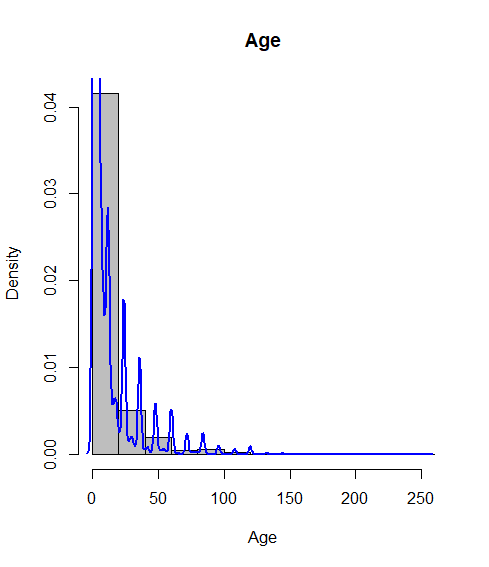
 

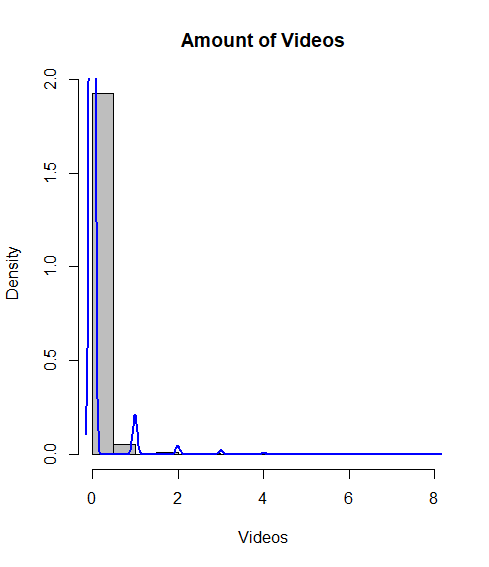
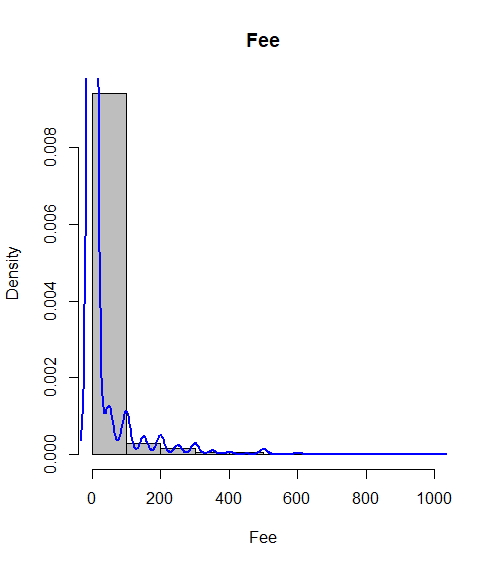
 

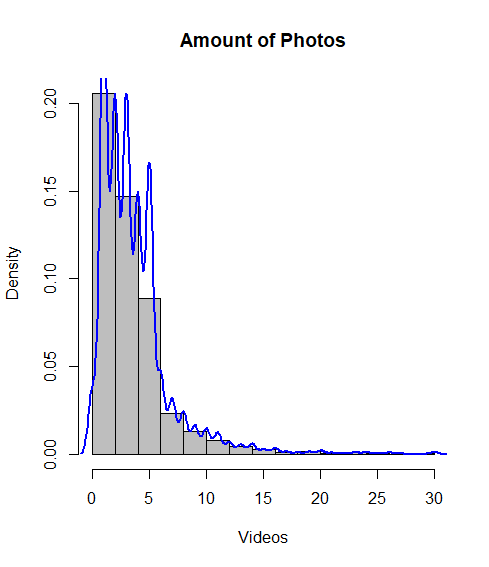




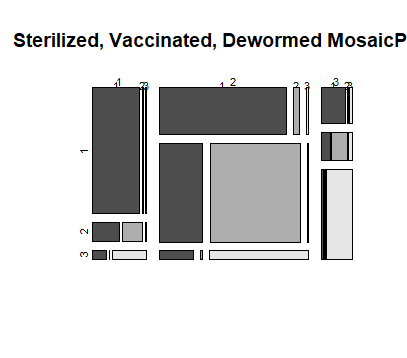
## משתנים רציפים – היסטוגרמות







## 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