המכללה האקדמית תל-חי למדע, סביבה וחברה
המכללה האקדמית תל-חי | ד.נ. גליל עליון 12210 | טלפון: 04-8181818

פרויקט גמר בקורס:

BIG DATA

מרצה:

ד"ר איתי שרון

מגישים:

עדן בהלול 203820139
שרון חן 301491643

תש"ף סמסטר אביב

רקע:

פרויקט זה עוסק בחקר שוק של בעלי כלבים, בנסיון לאתר אילו שירותים הם היו מעוניינים להשתמש.
השירות העיקרי שבחרנו להתייחס אליו במסגרת הפרויקט ולייחס לו את קבוצת המטרה, הוא מידת העניין של המשתמשים באפליקציה לדעת אילו מקומות ציבוריים נגישים לכלבים- כלומר Dog Friendly.
קבוצת המטרה הוגדרה כבינארית: 1 : מעוניין להשתמש 0 : לא מעוניין להשתמש.

כמו כן, בדקנו צרכי שוק של שירותים נוספים כגון: "שמירה על כלבים", "הולכת כלבים", "עבודה כשומר/מוליך כלבים", "הכרויות לבעלי לכלבים" וכדומה.

איסוף הנתונים התבצע באמצעות שאלון אנונימי שהופץ בקרב בעלי כלבים (משפחה, חברים, ובעיקר קבוצות יעודיות בפייסבוק). הושגו 204 תצפיות, המכילות מידע על בעלי הכלבים, וכמו כן על הכלבים עצמם.

מתוך נתונים אלו, נרצה לזהות את פרופיל משתמש פוטנציאלי, בעל ההסתברות הגבוהה ביותר לשימוש באפליקציה- מציאת מקומות נגישים לכלבים. יתרה מזאת, עם המסקנות שנסיק, נוכל לבצע פרסום ייעודי בדפי אינטרנט שבהן קהל היעד מתאים לפרופיל שבנינו.

שלב א' : ניקוי הנתונים

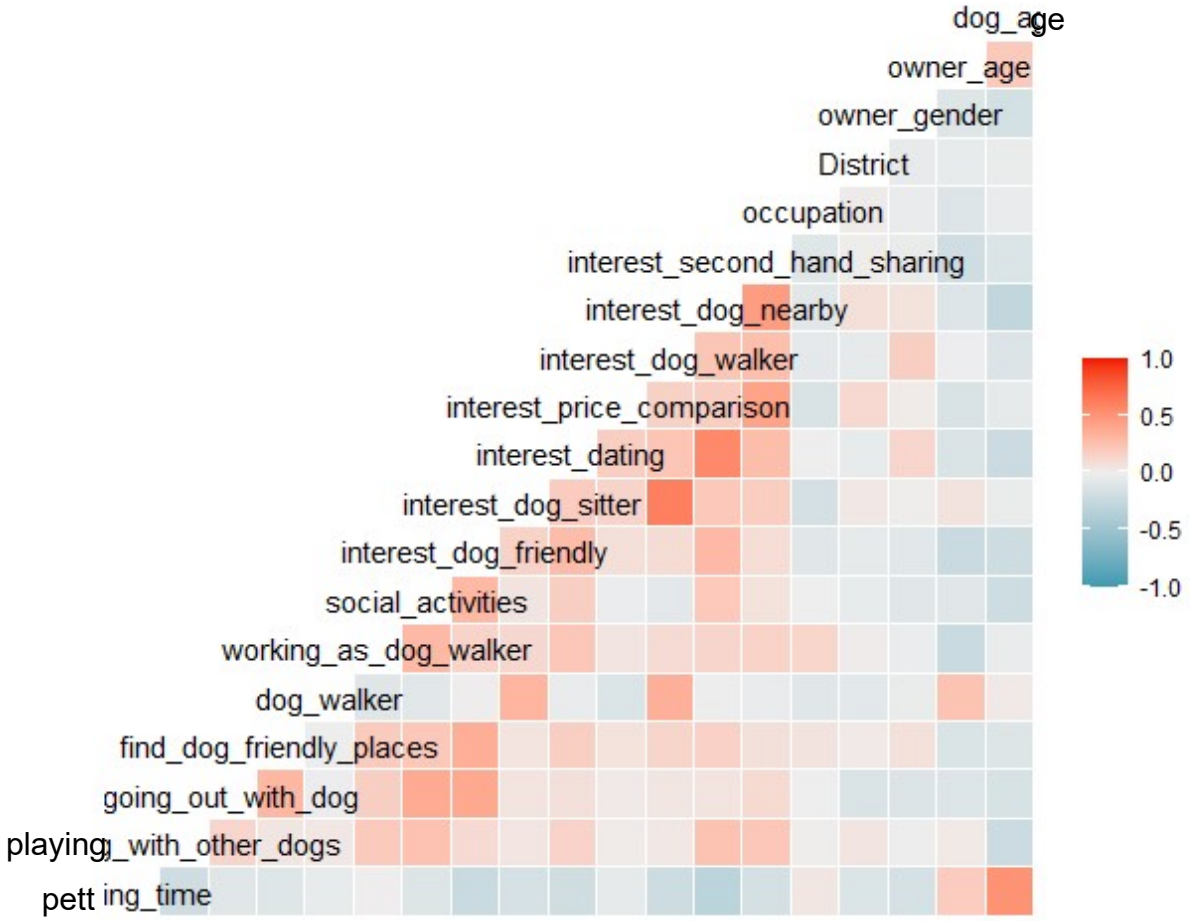
המשתנים בבסיס הנתונים שלנו התחלקו לשלושה סוגים: מנה (Ratio), סדר (Ordinal) ושמי (Nominal) מכיוון שלא ניתן לבצע ניתוח סטטיסטי וחישוב מתמטי עם מחרוזות, התבקשנו להשתמש ב- LEVELS

חריגות: לאחר שביצענו בדיקה לאיתור outliers נמצא כי קיימת חריגה בשורה 118 בגיל הכלב, השתמשנו בערך החציון כדי להחליף את המידע השגוי.

שלב ב' : היכרות עם הנתונים

1. זיהוי קורלציות בין משתנים:

בגרף שלהלן (באדיבות ספריית GGPLOT) ניתן לראות את הקורלציות באמצעות מטריצה צבעונית המייצגת את הקשרים בין כל המשתנים.



ע"פ הגרף (ובדיקה נוספת בקוד) ניתן לראות הקורלציות החזקות במדגם שאמדנו בין זוגות המשתנים הבאים:

- petting_time ~ dog_age = 0.5143641
- interest_dating ~ interest_dog_nearby = 0.5625078
- interest_dog_sitter ~ interest_dog_walker = 0.6148942

מסקנות:

נראה שהמקרה הראשון הוא טריוויאלי, כיוון ששני הפרמטרים מושפעים מהזמן שחולף באותה מידה.
מובן שהשאלון עוסק רק בכלבים מאומצים, ולכן גיל הכלב (המיוצג בשנים של בן-אדם) יגדל באותו אופן שבו גדל הזמן שהוא מאומץ. (לא קיים בנתונים מקרה של כלב לא מאומץ)

המקרה השני מצביע על כך שאדם המעוניין למצוא כלבים (או בעלי כלבים) בסביבתו, יש סיכוי גבוה שיהיה מעוניין גם בהיכרויות ודייטים עם אותם בעלי כלבים. זה כמובן הגיוני מכיוון שבשני המקרים מדובר בהיכרות כלשהי. בנוסף ניתן להסיק מכך שאנשים יעדיפו היכרויות ודייטים עם אנשים שקרובים אליהם מבחינה גיאוגרפית, כלומר בסביבתם.

המקרה השלישי והחזק ביותר הוא בין "שמירה על כלבים" ל-"הולכת כלבים".
ראשית, הגיוני לחשוב שאדם שלא מסוגל לבצע אחת ממשימות אלו באופן סדיר (וזקוק לשירות), ישתמש ככל הנראה בשני השירותים, ולא רק באחד מהם. יש דמיון רב בין שני התפקידים והקשר נראה הגיוני.
(במידה מסוימת אפשר לומר ש"הולכת כלבים" מהווה משימה אחת מתוך "שמירה על כלבים", וכמו כן בעת שמירה על הכלבים, לעיתים יש להוליך אותם)

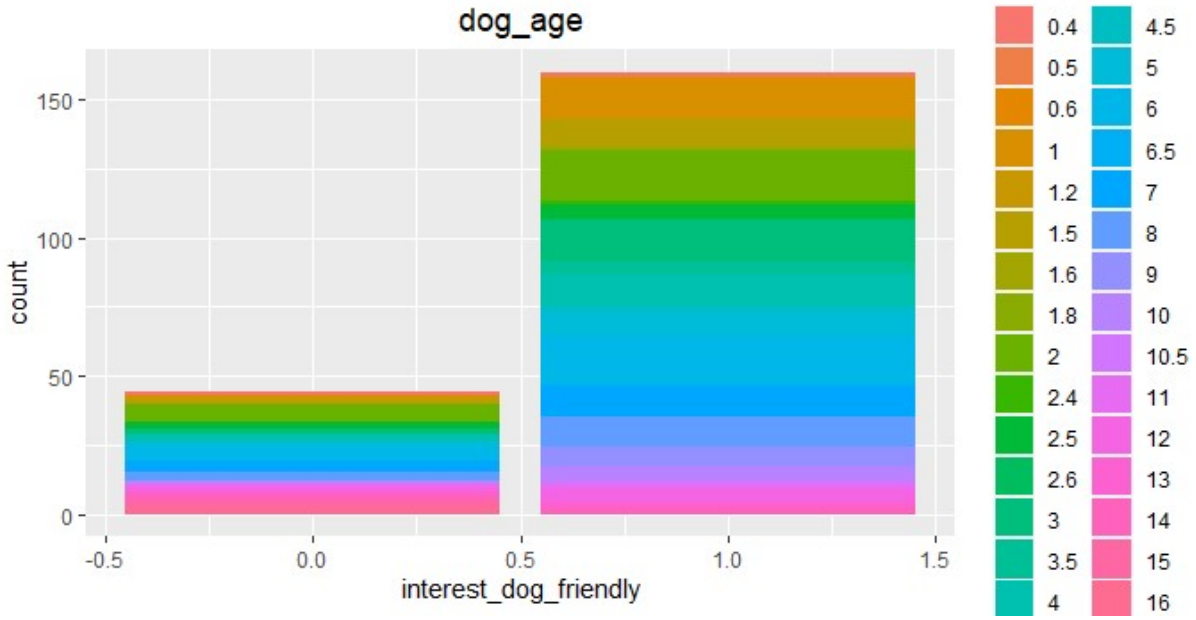
ע"פ הגרף, נראה כי לא קיימת אנטי-קורלציה משמעותית בנתונים אלו.

למרות כל המסקנות, ולאחר בניית הגרפים בעבור מקרים אלו (מופיעים רק בקובץ הקוד), נראה שלא ניתן להסיק הרבה מידע על פרופיל משתמש פוטנציאלי מהמקרים הנ"ל.
בנוסף, בניית מודל רגרסיה לינארית לא מניבה תוצאות אינפורמטיביות, ככל הנראה מכיוון שמדובר במשתנים קטגוריים ושמיים (יש צורך ברגרסיה לוגיסטית - מחוץ לחומר הנלמד בקורס)

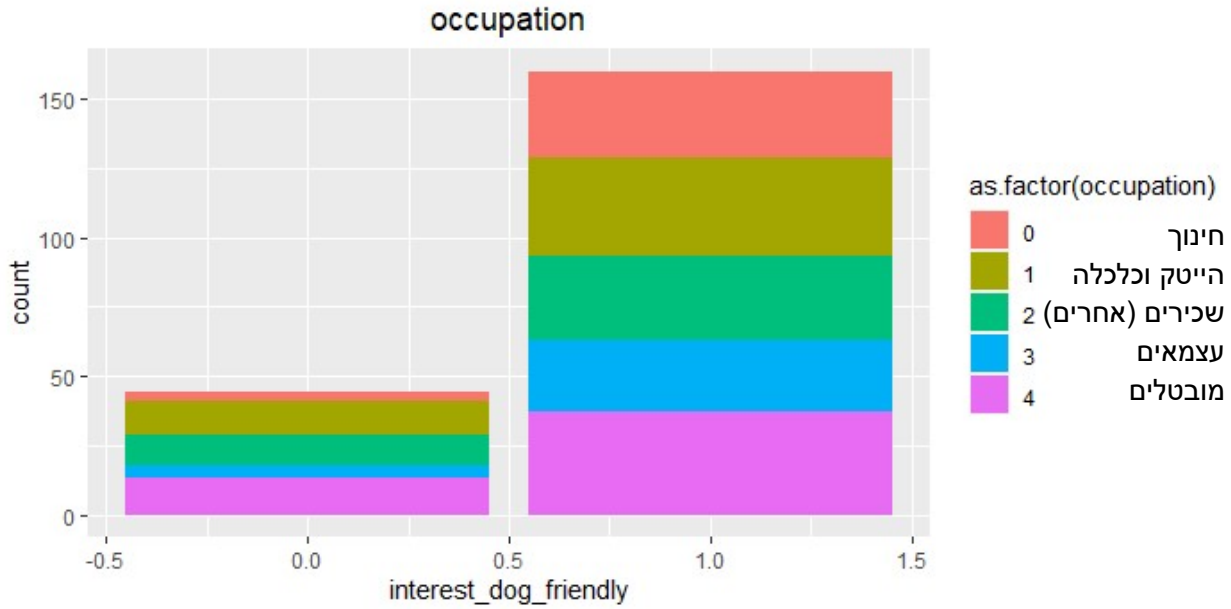
על כן, בחרנו להתמקד בפרמטר interest_dog_friendly המייצג קבוצת מטרה כללית המעוניינת להשתמש בשירותי האפליקציה לצורך מציאת מקומות ציבוריים מונגשים וידידותיים לכלבים.

2. התפלגות הנתונים לפי קבוצת המטרה:

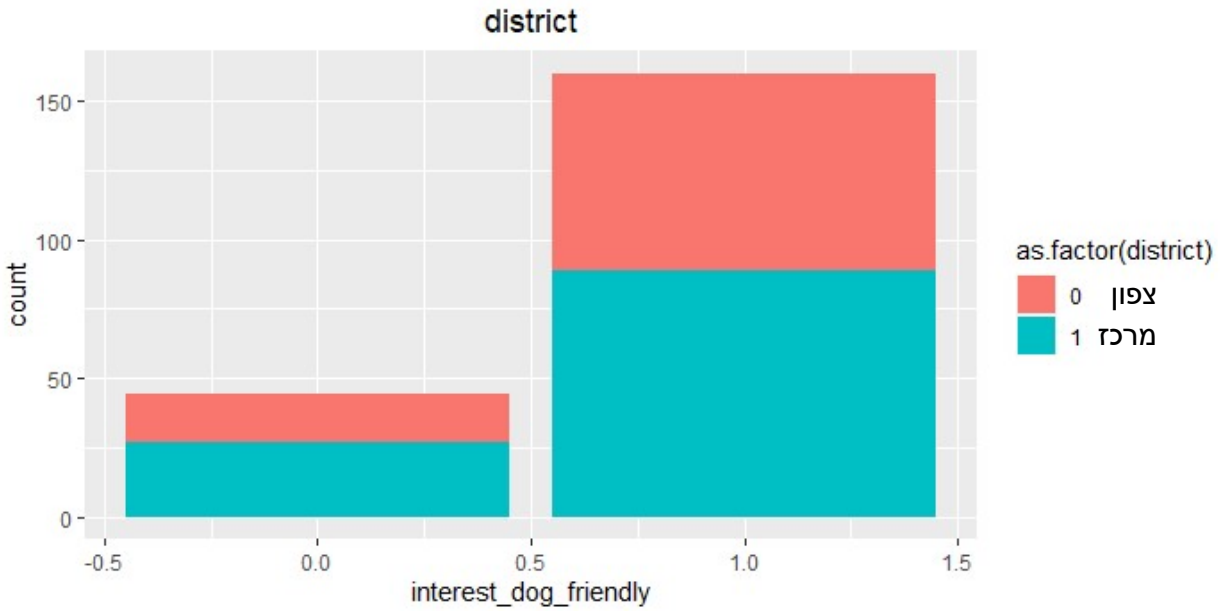
מתוך משיבי השאלון שציינו כי הם מעוניינים להשתמש באפליקציה, ניתן לראות שטווח הגילאים העיקרי לכלבים נע בין הגילאים שנתיים לשנתיים וחצי, וכמו כן בין 6 ל-6 וחצי שנים.



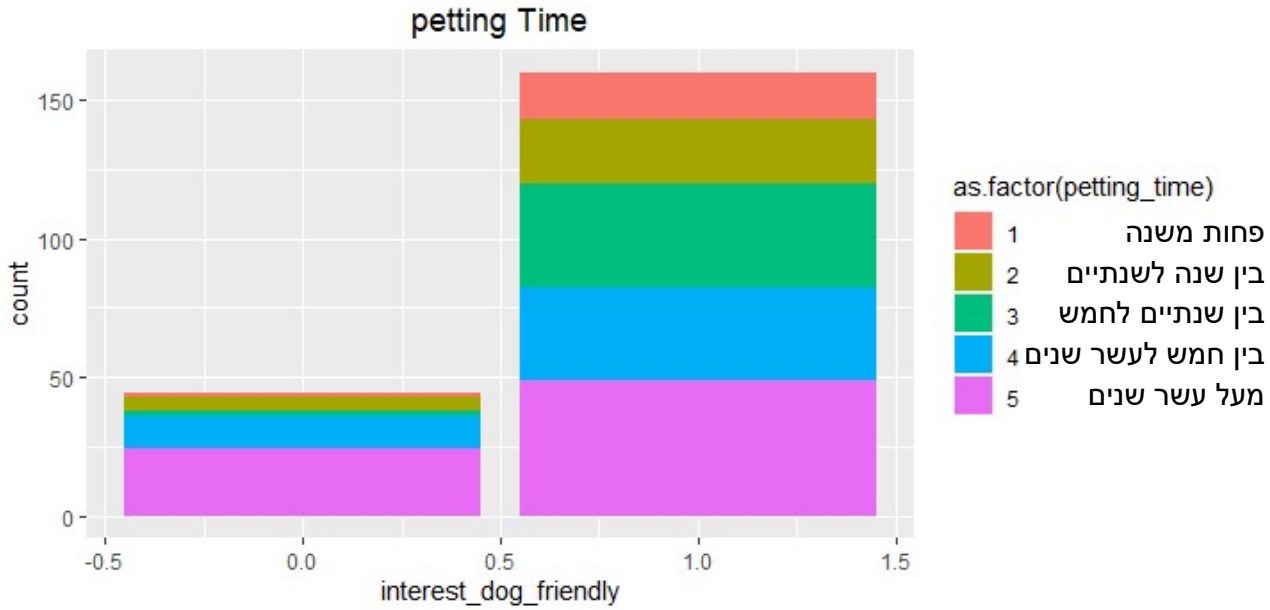
ראינו לנכון לבצע איחוד בין מקצועות בעלי אופי ריאלי תחת קבוצה אחת, מכיוון שכאשר הם נמצאים בנפרד (כלכלה והייטק) ישנה כמות קטנה מידי של נתונים. על פי הגרף ניתן לראות שקיימת **התפלגות אחידה** באופן יחסי בתחומי העיסוק של משיבי השאלון שציינו כי הם מעוניינים להשתמש באפליקציה.



באופן דומה, החלטנו לחלק את אזור המגורים לשתי קבוצות עיקריות: צפון ומרכז. מכיוון שרוב הנתונים היו מהאזורים האלו. ניתן לראות על פי הגרף שיש נטייה להשתמש במוצר לבעלי כלבים מאזור המרכז לעומת אזור הצפון. כלומר, ניתן להסיק שאיזור המגורים של המשתמש הפוטנציאלי יהיה במרכז הארץ.

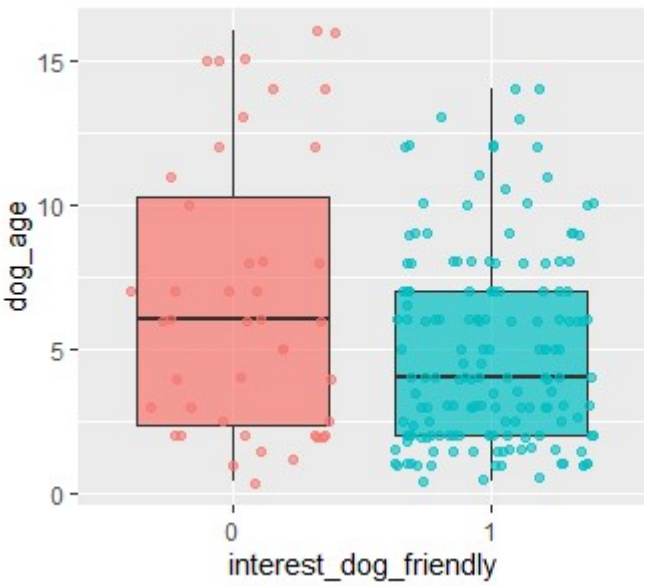


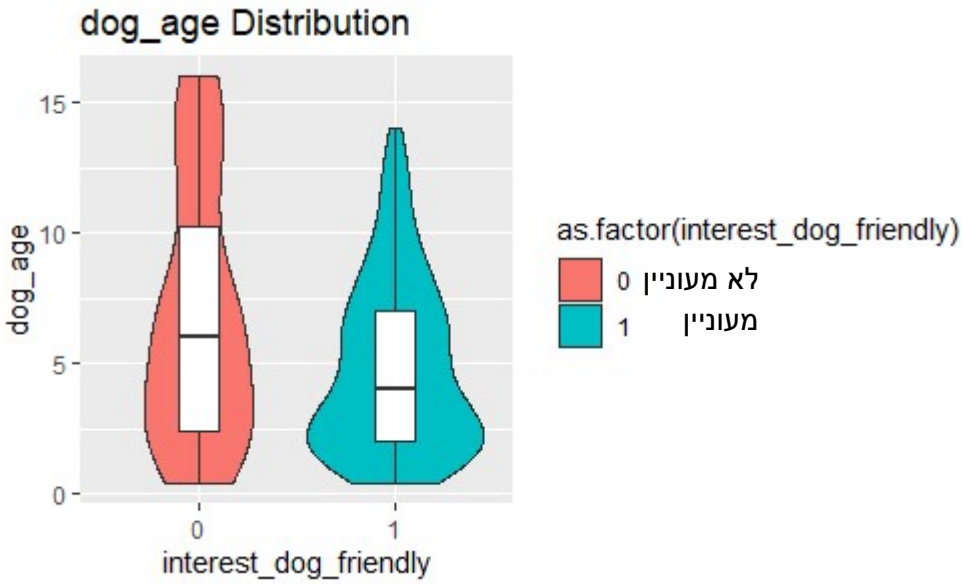
אם נתבונן בכל קבוצה (צבע) בנפרד, ונבחן את יחס החלוקה לשתי הקטגוריות "מעוניין" או "לא מעוניין", ניתן לראות מגמת עלייה באחוז ה-"לא מעוניינים" בכל קבוצה:
בקבוצה 1 (בתחילת תקופת האימוץ) אחוז קטן מאוד מבעלי הכלבים לא מעוניינים להשתמש באפליקציה.
בקבוצה 2 ישנו אחוז גדול יותר, וכן הלאה עד לקבוצה 5 (כאשר הכלב ותיק) שבה כ-33% מבעלי הכלבים לא מעוניינים להשתמש באפליקציה.
מכאן נסיק שכאשר תקופת האימוץ עולה, מידת הרצון לשימוש באפליקציה פוחתת.



3. פיזור והתפלגות הנתונים:

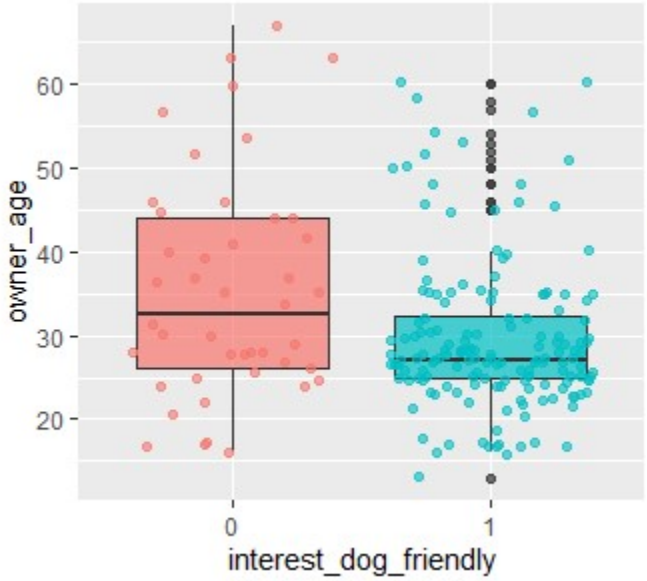
מידת הרצון להשתמש באפליקציה על פי **גיל הכלב**:

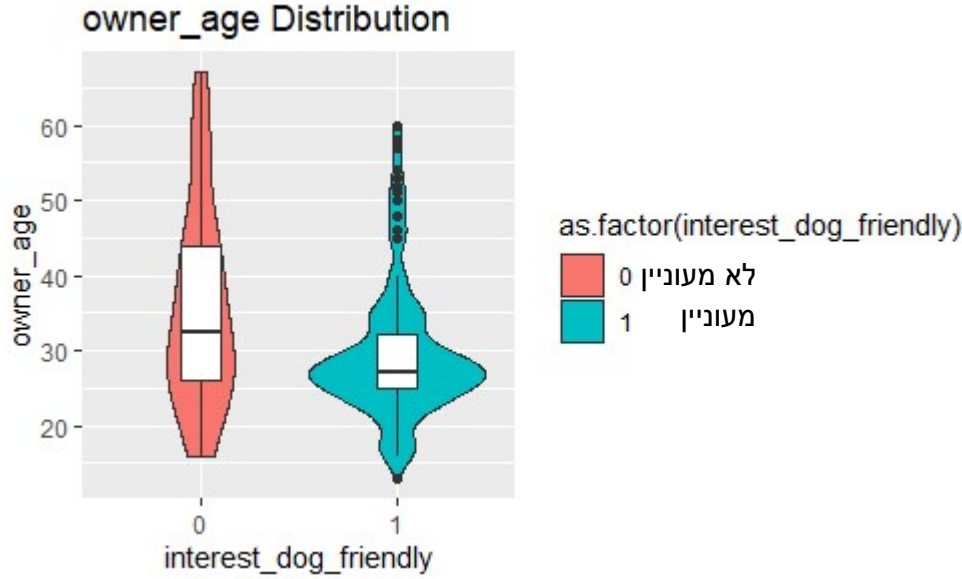




מתוך משיבי השאלון שציינו כי הם מעוניינים להשתמש באפליקציה (בכחול), ניתן לראות שככל שגיל הכלב עולה, ישנה מגמת ירידה במידת הרצון להשתמש באפליקציה.
כלומר, עבור כלבים צעירים (עד כשנתיים) מידת הרצון היא גבוהה, ובעבור כלבים מבוגרים (כ-6 עד 6 וחצי שנים) מידת הרצון נמוכה. (בהתאמה לגרף הקודם - כאמור ישנה קורלציה בין גיל הכלב לבין זמן האימוץ)

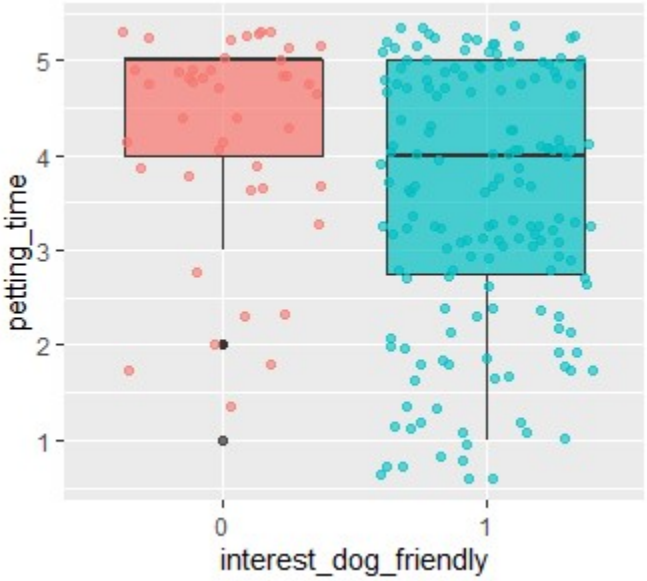
מידת הרצון להשתמש באפליקציה על פי **גיל בעל הכלב**:

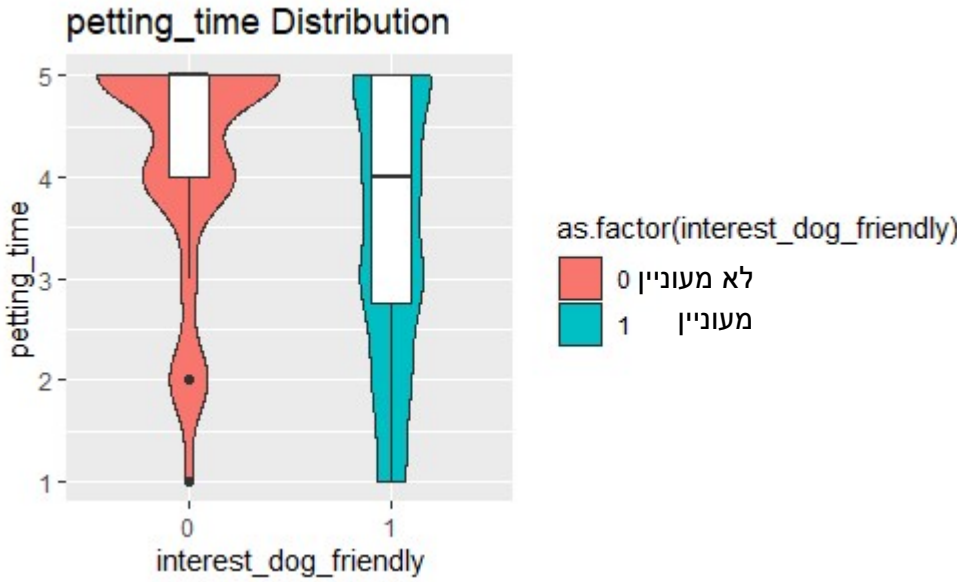




מתוך משיבי השאלון שציינו כי הם מעוניינים להשתמש באפליקציה (בכחול), ניתן לראות שגיל בעל הכלב נמצא בעיקר בטווח שבין 25-30 שנים (דור המילניום). מכאן ניתן להסיק, שבעבור פרופיל המשתמש הפוטנציאלי - יהיה טווח גילאים דומה.

מידת הרצון להשתמש באפליקציה על פי **זמן האימוץ** (בשנים).

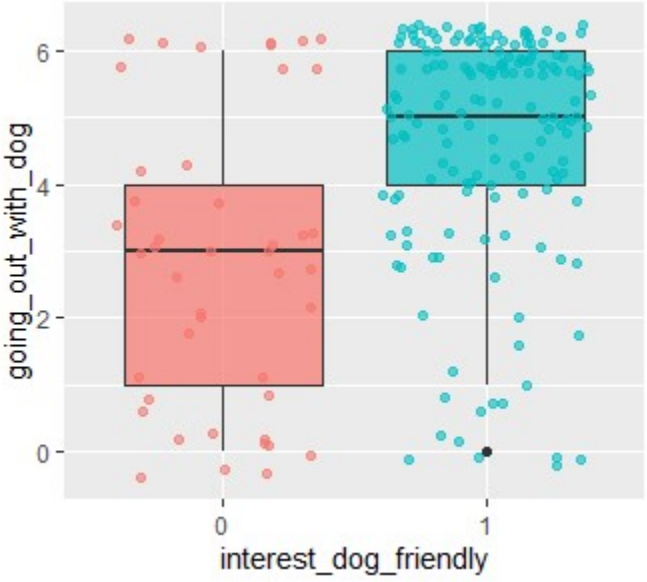


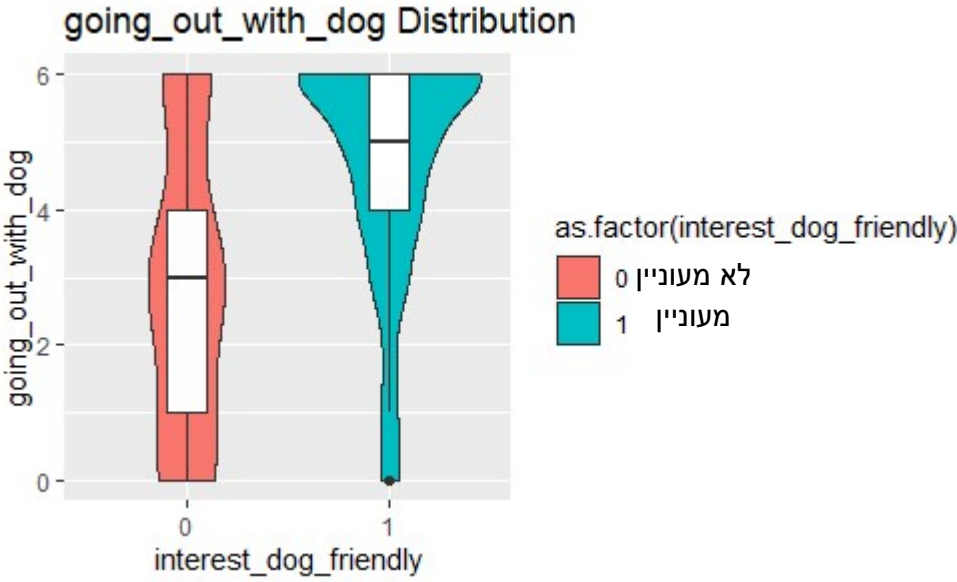


פחות משנה	1
בין שנה לשנתיים	2
בין שנתיים לחמש	3
בין חמש לעשר שנים	4
מעל עשר שנים	5

תזכורת: ציר ה-Y של גרפים אלו מייצג את זמן האימוץ ע"פ הטבלה הבאה:
מתוך משיבי השאלון שציינו כי הם מעוניינים להשתמש באפליקציה (בכחול),
ניתן לראות שתי קבוצות עיקריות המתוייגות כ- 3 ו -5
(וזאת בהתאמה לגרף הקודם של זמן האימוץ).
בנוסף נראה שמתוך משיבי השאלון שלא מעוניינים להשתמש באפליקציה,
ברוב המקרים, זמן האימוץ הוא ארוך (5 עד 10 שנים או מעל 10 שנים)

מידת הרצון להשתמש באפליקציה על פי מידת הרצון לצאת ל**מקומות בילוי**





מתוך משיבי השאלון שציינו כי הם מעוניינים להשתמש באפליקציה (בכחול), נראה כי רובם הם "בליינים"
(כאלו שנוהגים לצאת לבלות לעיתים קרובות) והיו מעוניינים לשלב יציאה עם הכלב למקומות הבילוי.
מכאן ניתן להסיק שפרופיל משתמש פוטנציאלי, יהיה בעל מידת עניין גבוהה לבילוי עם הכלב.

שלב ג': הכנת הנתונים לקראת הרצת אלגוריתמים

1. נרמול הנתונים:

ראשית, בוצעה המרה של משתני סדר ומשתנים שמיים לטיפוס מסוג factor. (ע"י המתודה (as.factor()
בנוסף, הותאמו טווחי הנתונים המספריים בעזרת שימוש במתודה scale()

2. סיווג משתנה המטרה:

0 : לא מעוניין להשתמש באפליקציה
1 : מעוניין להשתמש באפליקציה

3. חלוקת הנתונים:

אימון: 70 אחוז מהבסיס נתונים – 148 תצפיות
מבחן: 30 אחוז מבסיס הנתונים – 56 תצפיות

שלב ד': הרצת אלגוריתמים random forest ו-decision tree

1. תוצאות ריצה - RANDOM FOREST:

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0  8  4
         1  5 39

               Accuracy : 0.8393
                 95% CI : (0.7167, 0.9238)
    No Information Rate : 0.7679
    P-Value [Acc > NIR] : 0.1321

                  Kappa : 0.5368

 Mcnemar's Test P-Value : 1.0000

            Sensitivity : 0.9070
            Specificity : 0.6154
         Pos Pred Value : 0.8864
         Neg Pred Value : 0.6667
             Prevalence : 0.7679
         Detection Rate : 0.6964
   Detection Prevalence : 0.7857
      Balanced Accuracy : 0.7612

       'Positive' Class : 1
```

2. מסקנות - RANDOM FOREST:

- דיוק המודל עומד על כ- 84%

- היו 5 מקרים בהם המודל ניבא שבעל הכלב לא ישתמש באפליקציה, ובקבוצת המבחן התוצאה הייתה שהצרכן יהיה מעוניין להשתמש במוצר. (false negative)

- היו 4 מקרים בהם המודל ניבא שבעל הכלב כן ישתמש באפליקציה ובקבוצת המבחן, התוצאה הייתה שהצרכן יהיה לא מעוניין להשתמש במוצר. (false positive)

- מרווח הביטחון של דיוק המודל נע בין הטווחים 0.717 ל 0.923 – מרווח יחסית רחב.

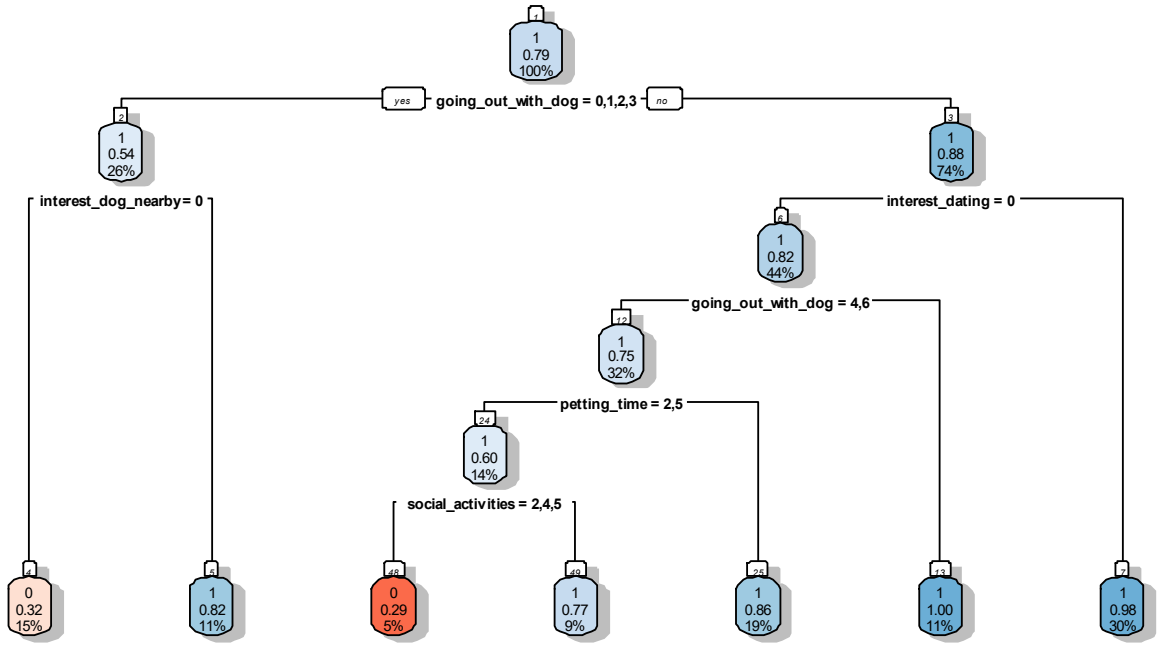
- ניתן לראות שהמחלקה הדומיננטית (1: מעוניין) תופסת 0.7679 מסיווג המודל: $\frac{38+5}{56}$
היחס שהתקבל קרוב מאוד לאחוז הדיוק של המודל, ייתכן ומצביע על כך שיש לשפר את המודל.

- קיים הבדל בין אחוז הניבוי הנכון של מחלקה 1 : sensitivity- 0.907 = $\frac{38}{38+5}$
ובין אחוז הניבוי הנכון של מחלקה 0 : specificity– 0.6923 = $\frac{9}{9+4}$
הבדל זה נובע בגלל הגודל השונה של ה-LEVELS במשתנה המטרה. לפיכך מחלקה 1 הנה מאוד דומיננטית במודל, יחד עם זאת, אחוז הניבוי המוצלח של המודל לגבי מחלקה 0 הוא כ-70 אחוז.

3. תוצאות ריצה - DECISION TREE:

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 10  7
         1  3 36

               Accuracy : 0.8214
                 95% CI : (0.696, 0.9109)
    No Information Rate : 0.7679
    P-Value [Acc > NIR] : 0.2176

                  Kappa : 0.5477

 Mcnemar's Test P-Value : 0.3428

            Sensitivity : 0.8372
            Specificity : 0.7692
         Pos Pred Value : 0.9231
         Neg Pred Value : 0.5882
             Prevalence : 0.7679
         Detection Rate : 0.6429
   Detection Prevalence : 0.6964
      Balanced Accuracy : 0.8032

       'Positive' Class : 1
```

על פי העץ ניתן לראות את התנאים הנדרשים לחיזוי פרופיל המשתמש. כלומר ניתן לראות המצבים שבהם בעל הכלב ישתמש באפליקציה.

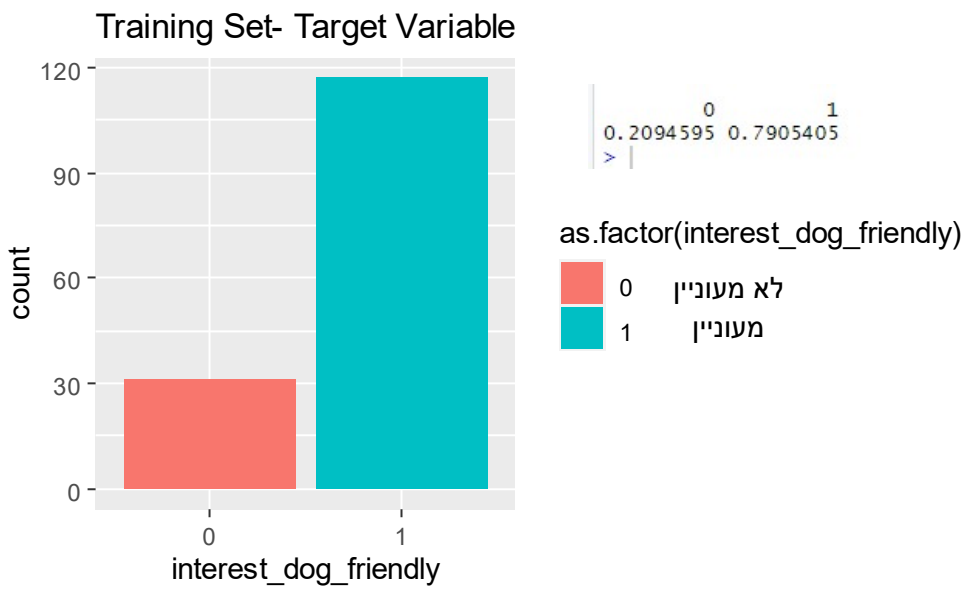


4. מסקנות - DECISION TREE:

- דיוק המודל עומד על 82 אחוזים.
- היו 7 מקרים בהם המודל ניבא שבעל הכלב ישתמש במוצר וטעה
- היו 3 מקרים שבהם המודל ניבא שבעל הכלב לא ישתמש במוצר וטעה.
- מרווח הביטחון של דיוק המודל נע בין 0.696-0.9100
- המחלקה הדומיננטית תופסת 77 אחוז מסיווג המודל.
- הפער בין sensitivity לבין specificity אינינו גדול, למרות השוני בגודל ה-LEVELS של קבוצת המטרה (השפיע על הפער ב-RANDOM FOREST)

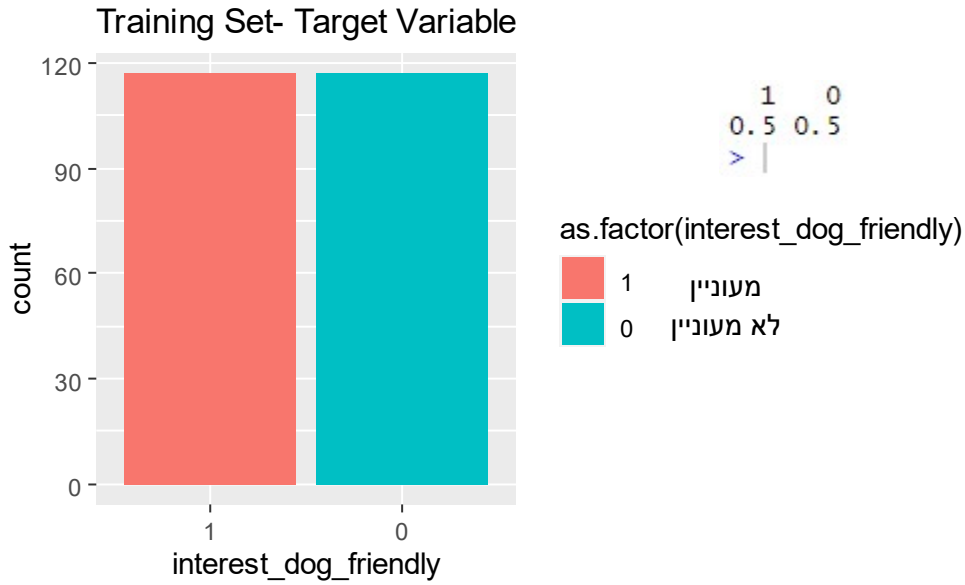
שלב ו': שיפור המודל

ננסה לשפר את המודל על ידי חלוקה שווה של קבוצות משתנה המטרה.
מצב זה נקרא: "Handling Class Imbalance Problem". לעיתים מצב זה משפיע על דיוק המודל.



קיימת ספריה ב-R בשם ROSE: Randomly Over Samples Examples- ובה קיימת מתודה המאפשרת להגדיל את המדגם- "ליצור" תצפיות חדשות. מה שקורה מאחורי הקלעים הוא שהפונקציה בוחרת באופן אקראי תצפיות קיימות ומשכפלת אותם. כמו כן הממוצע, החציון, סטיית התקן ושאר הרכיבים המופיעים ב"summary" לא משתנים באופן משמעותי.

1. ניסיון ראשון - הגדלת המדגם:




```
   1   0
 0.5 0.5
>
```

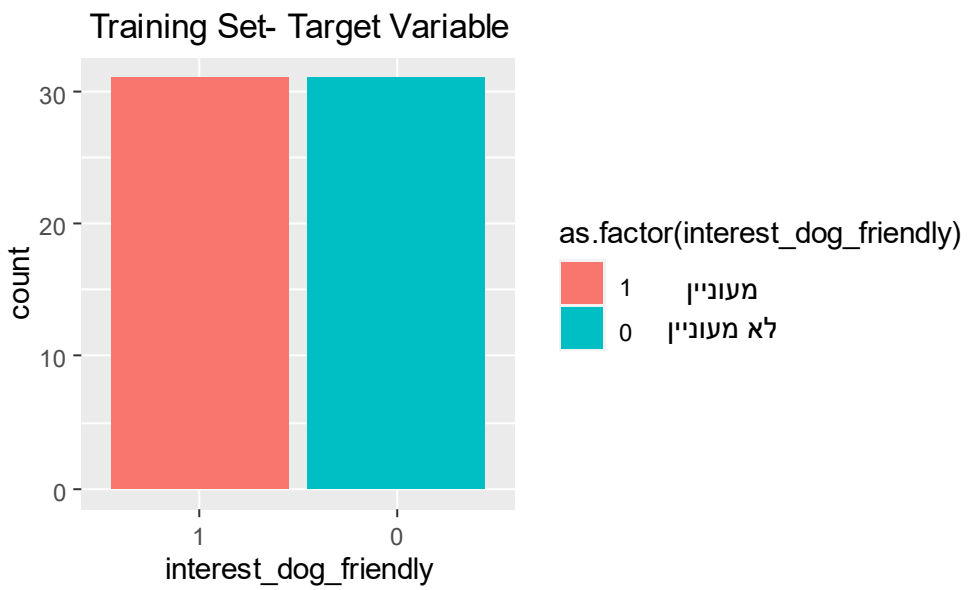
הרצת Decision Tree

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 11 15
         1  2 28

               Accuracy : 0.6964
                 95% CI : (0.559, 0.8122)
    No Information Rate : 0.7679
    P-Value [Acc > NIR] : 0.919400

                  Kappa : 0.3687

 Mcnemar's Test P-Value : 0.003609

            Sensitivity : 0.6512
            Specificity : 0.8462
         Pos Pred Value : 0.9333
         Neg Pred Value : 0.4231
             Prevalence : 0.7679
         Detection Rate : 0.5000
   Detection Prevalence : 0.5357
      Balanced Accuracy : 0.7487

       'Positive' Class : 1
```

הרצת Random Forest

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0  9  6
         1  4 37

               Accuracy : 0.8214
                 95% CI : (0.696, 0.9109)
    No Information Rate : 0.7679
    P-Value [Acc > NIR] : 0.2176

                  Kappa : 0.5246

 Mcnemar's Test P-Value : 0.7518

            Sensitivity : 0.8605
            Specificity : 0.6923
         Pos Pred Value : 0.9024
         Neg Pred Value : 0.6000
             Prevalence : 0.7679
         Detection Rate : 0.6607
   Detection Prevalence : 0.7321
      Balanced Accuracy : 0.7764

       'Positive' Class : 1
```

2. ניסיון שני - הקטנת המדגם:



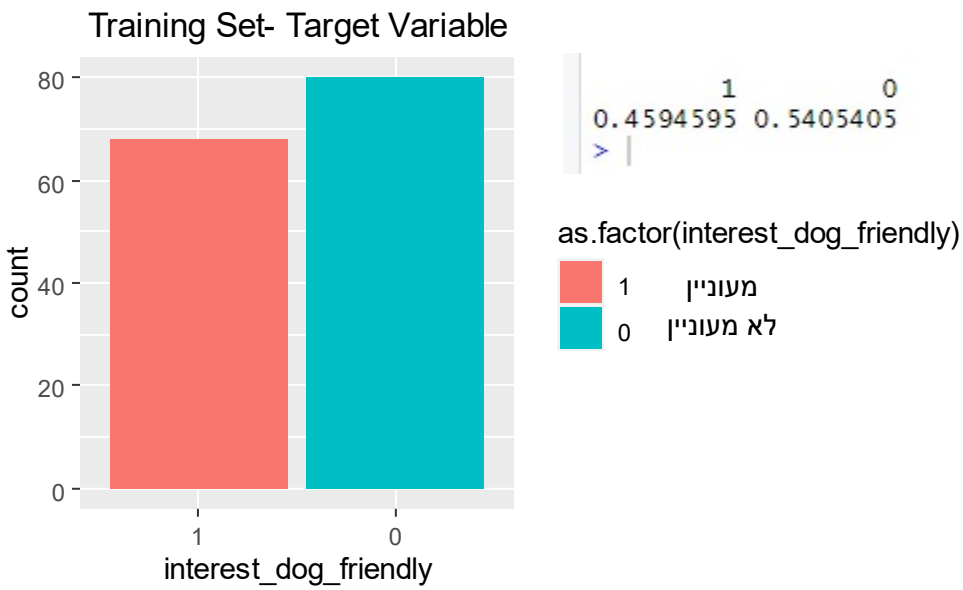
הרצת Decision Tree

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 10 23
         1  3 20

               Accuracy : 0.5357
                 95% CI : (0.3974, 0.6701)
    No Information Rate : 0.7679
    P-Value [Acc > NIR] : 0.9999625

                  Kappa : 0.1525

 Mcnemar's Test P-Value : 0.0001944

            Sensitivity : 0.4651
            Specificity : 0.7692
         Pos Pred Value : 0.8696
         Neg Pred Value : 0.3030
             Prevalence : 0.7679
         Detection Rate : 0.3571
   Detection Prevalence : 0.4107
      Balanced Accuracy : 0.6172

       'Positive' Class : 1
```

הרצת Random Forest

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 11 14
         1  2 29

               Accuracy : 0.7143
                 95% CI : (0.5779, 0.827)
    No Information Rate : 0.7679
    P-Value [Acc > NIR] : 0.86508

                  Kappa : 0.3938

 Mcnemar's Test P-Value : 0.00596

            Sensitivity : 0.6744
            Specificity : 0.8462
         Pos Pred Value : 0.9355
         Neg Pred Value : 0.4400
             Prevalence : 0.7679
         Detection Rate : 0.5179
   Detection Prevalence : 0.5536
      Balanced Accuracy : 0.7603

       'Positive' Class : 1
```

3. ניסיון שלישי - הקטנה והגדלת המדגם בו זמנית (צמצום פערים):




```
        1         0
0.4594595 0.5405405
>
```

הרצת Decision Tree

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0  7 18
         1  6 25

               Accuracy : 0.5714
                 95% CI : (0.4322, 0.7029)
    No Information Rate : 0.7679
    P-Value [Acc > NIR] : 0.99967

                  Kappa : 0.0907

 Mcnemar's Test P-Value : 0.02474

            Sensitivity : 0.5814
            Specificity : 0.5385
         Pos Pred Value : 0.8065
         Neg Pred Value : 0.2800
             Prevalence : 0.7679
         Detection Rate : 0.4464
   Detection Prevalence : 0.5536
      Balanced Accuracy : 0.5599

       'Positive' Class : 1
```

הרצת Random Forest

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0  9  5
         1  4 38

               Accuracy : 0.8393
                 95% CI : (0.7167, 0.9238)
    No Information Rate : 0.7679
    P-Value [Acc > NIR] : 0.1321

                  Kappa : 0.561

 Mcnemar's Test P-Value : 1.0000

            Sensitivity : 0.8837
            Specificity : 0.6923
         Pos Pred Value : 0.9048
         Neg Pred Value : 0.6429
             Prevalence : 0.7679
         Detection Rate : 0.6786
   Detection Prevalence : 0.7500
      Balanced Accuracy : 0.7880

       'Positive' Class : 1
```

4. מסקנות:

- שלושת המתודות לא שיפרו את המודל.
- שלושת המתודות הזיקו לדיוק המודל, מלבד המתודה השלישית המשלבת הגדלה והקטנת המדגם- לא שינתה את דיוק המודל ב-RANDOM FOREST
- לאור האמור, ועל מנת להסיק מסקנות סופיות לגבי פרופיל משתמש פוטנציאלי נחזור לעבוד עם המודל המקורי.

סיכום:

בעבודה זו ביצענו חקר שוק על מנת למצוא את פרופיל הלקוח אשר יהיה מעוניין להשתמש באפליקציה. השתמשנו בשני אלגוריתמים עיקריים: "עץ החלטה" ו-" יער רנדומלי" .

להלן פרופיל משתמש פוטנציאלי על סמך הגרפים:

- גיל הכלב: שנתיים עד שנתיים וחצי, או 6 עד 6 וחצי שנים.
- גיל בעל הכלב: 25-30 שנים
- תקופת האימוץ: קצרה, פחות משנה.
- איזור מגורים: מגוון (עם עדיפות קלה למרכז הארץ)
- תחום העיסוק: מגוון (התפלגות אחידה) עם נטייה קלה ל"הייטק וכלכלה" או "מובטל"

על סמך עץ ההחלטה, כ-30 אחוז ממשתמשי האפליקציה מעוניינים מאוד לשלב יציאה עם הכלב למקומות בילוי, וכמו כן היו מעוניינים בפלטפורמת היכרויות עם בעלי כלבים נוספים בסביבתם.

מקורות חיצוניים:

Shir Grinblat - https://www.kaggle.com/shirgrinblat/mri-alzheimer-my-project

Dr Bharatendra Rai – https://www.youtube.com/watch?v=Ho2Klvzjegg&feature=youtu.be