שער פה

תוכן עניינים?

הבנת הנתונים

1. מקור הנתונים הוא מאתר <https://www.petfinder.my/>, אשר מאפשר לאנשים ממלזיה לפרסם את חיית המחמד שלהם לאימוץ או מכירה. הזנת הנתונים נעשית באופן ידני על ידי המפרסם. להלן המאפיינים המופיעים בסט הנתונים:

* Type – סוג בעל החיים (1 = כלב, 2 = חתול)
* Age – גיל בעל החיים
* Breed1 – גזע עיקרי לפי אנדקס
* Breed2 – גזע משני לפי אנדקס
* Gender – מין בעל החיים (1 = זכר, 2 = נקבה, 3 = מעורב, כאשר הפרופיל מייצג קבוצה של בעלי חיים)
* Color1 – צבע מס' 1 לפי אנדקס
* Color2 – צבע מס' 2 לפי אנדקס
* Color3 – צבע מס' 3 לפי אנדקס
* 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 לפי אנדקס
* VideoAmt – כמות סרטוני וידאו של בעל החיים שהועלו לפרופיל
* PhotoAmt – כמות התמונות של בעל החיים שהועלו לפרופיל

משתנה המטרה הינו Y, אשר מתייחס לאימוץ חיית המחמד. למשתנה 3 ערכים. 0 = אומץ באותו יום שנרשם באתר, 1 = אומץ בין יום לשלושה חודשים לאחר הרשמה, 2 = לא אומץ במהלך שלושת החודשים הראשונים להרשמתו.

מטרת הפרויקט הינה לסווג חיות מחמד בין קטגוריות כאלה, או במילים אחרות לנבא אם יאומצו ותוך כמה זמן בקטגוריות.

1. על מנת לבחון הסתברויות אפריוריות נשתמש בbar plots עבור משתנים קטגוריאליים והיסטוגרמות עבור משתנים מספריים. נבחן את ההסתברויות האפריוריות לפי סט הנתונים. עבור כל משתנה, נציג את הקוד שבעזרתו הגענו למסקנה ונוסיף ממצאים גרפיים.

משתנה מטרה y

dataset$y %>% table

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

A screenshot of a cell phone

Description automatically generated

ניתן לראות כי קטגוריה 1 ו2 יחסית מאוזנות, אך קטגוריה 0 כלל לא ומהווה רק כ4%. מצד שני, הגיוני כי חוסר איזון זה מייצג את המציאות, שכן סביר להניח שנדיר שחיית מחמד תאומץ באותו יום שהפרופיל שלה הועלה לאתר.

**ככה לעשות לכל משתנה? רק זה ייקח איזה 3 עמודים. מה אומר?**