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

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

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.

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.

## קשרים בין משתנים

* האם סט הנתונים מאוזן? האם לדעתכם הוא מייצג את המציאות?
* חפשו והציגו קשרים "מעניינים" בין מאפיינים - צפויים ולא-צפויים. הסבירו המשמעות של קשרים אלה.
* 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?

1. **איכות הנתונים**

* האם ישנם נתונים חסרים? אם כן, מה ניתן לומר על עליהם?
* האם ישנם נתונים שאינם הגיוניים (למשל: גיל>250)? מה אתם מציעים לעשות עם נתונים אלה?

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

**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.

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).

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?

**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

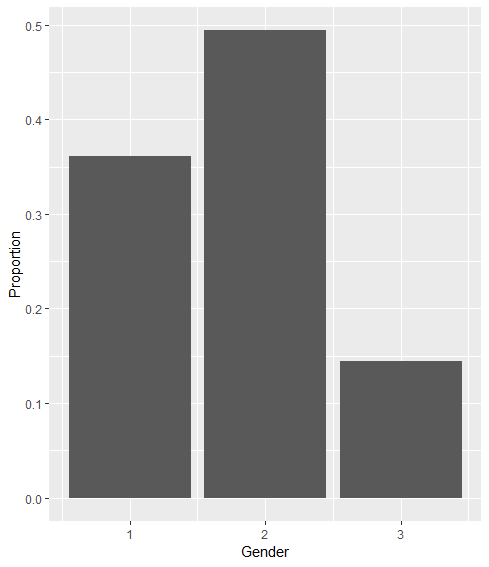
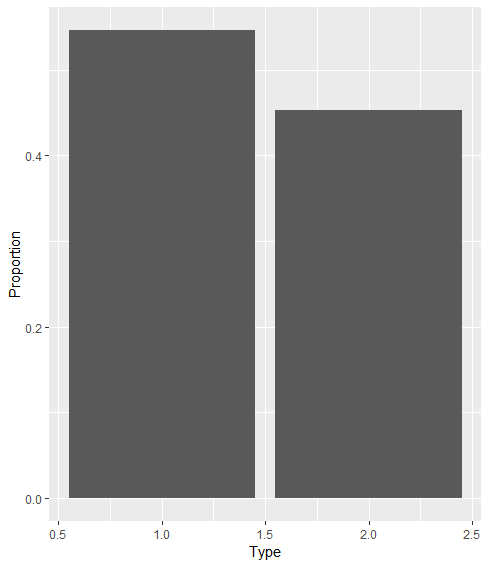
## טבלאות שכיחות

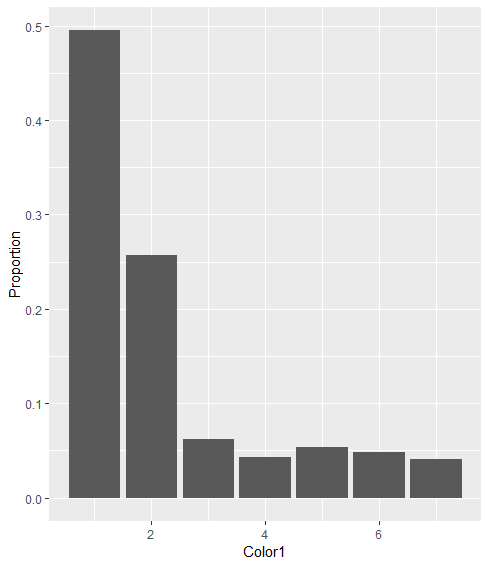
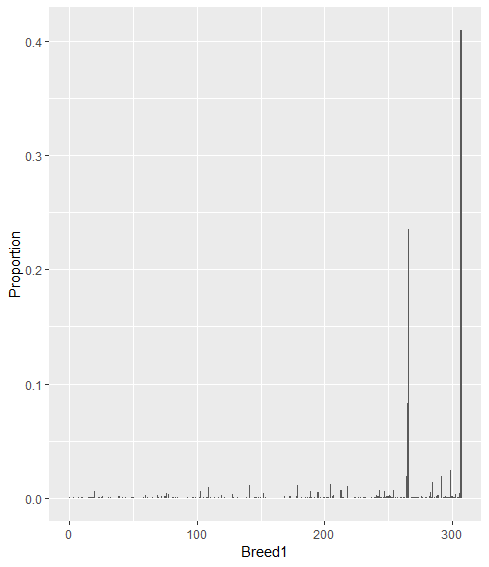


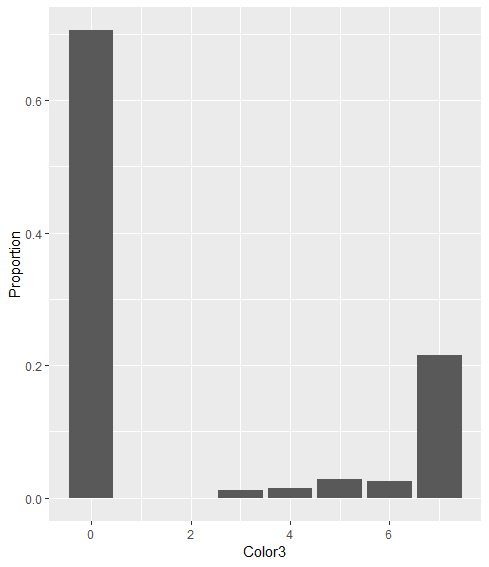
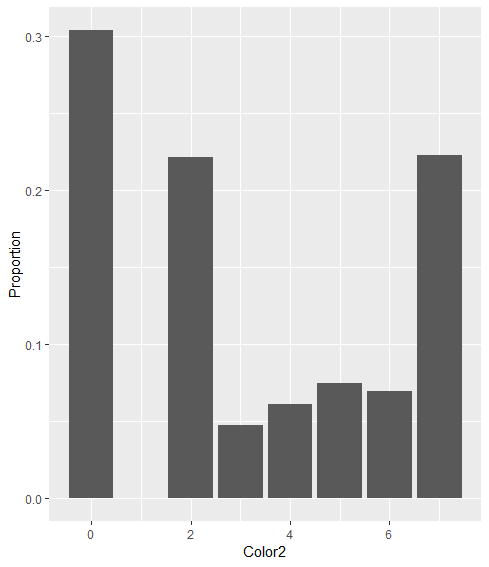


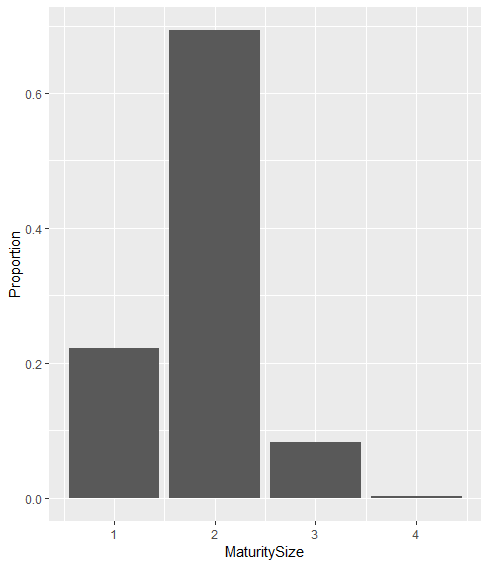
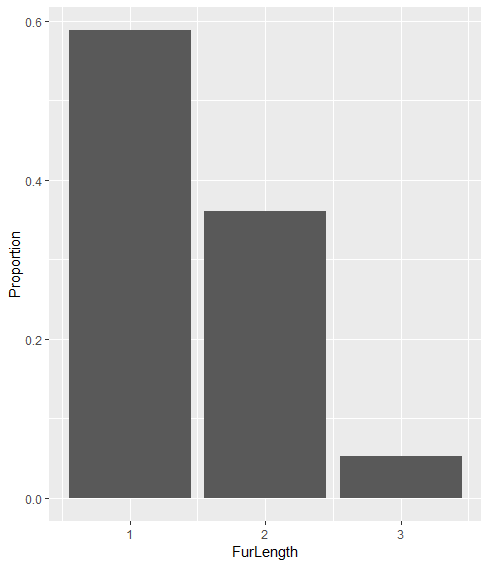


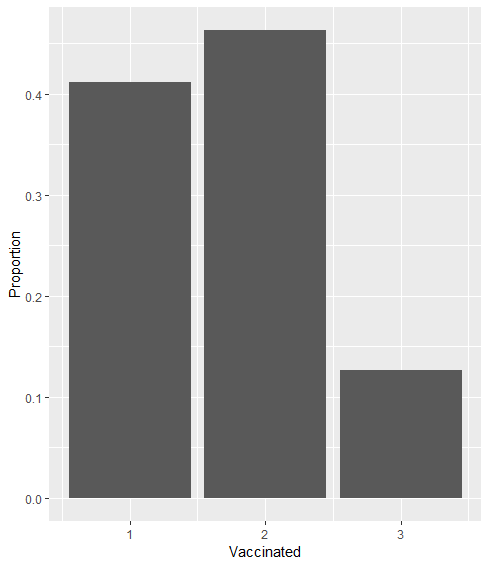
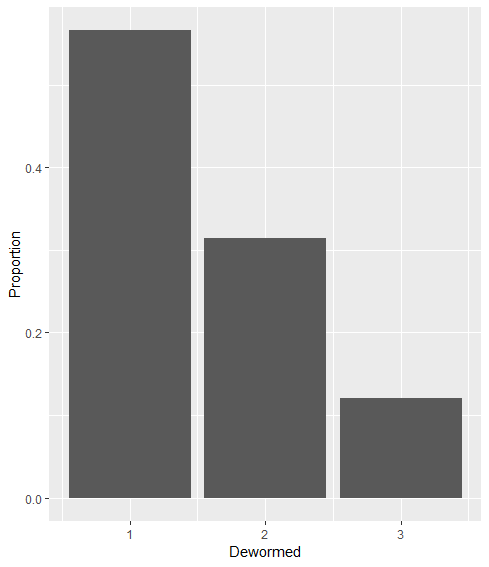
## קטגוריאלים – גרפי פורפורציה

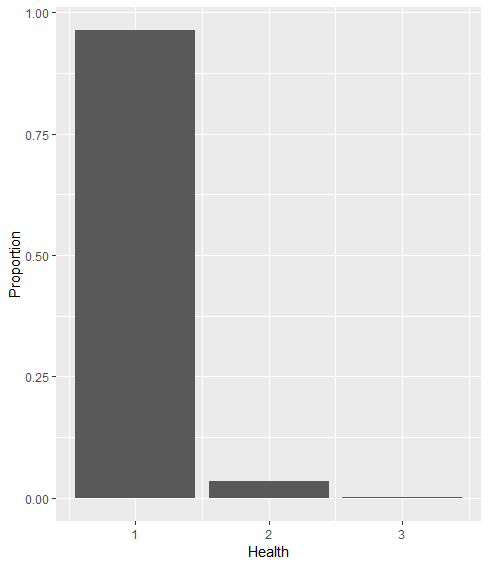
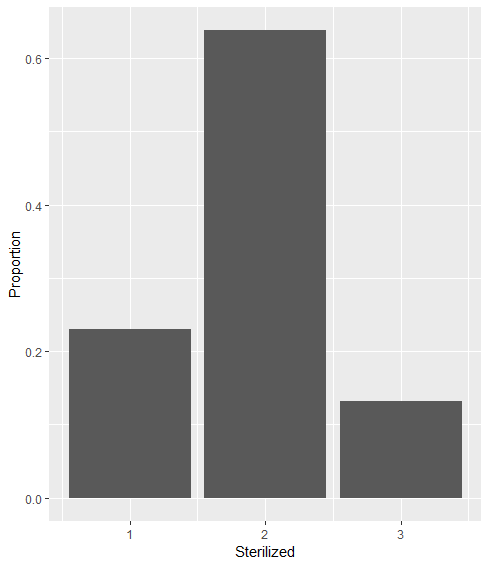


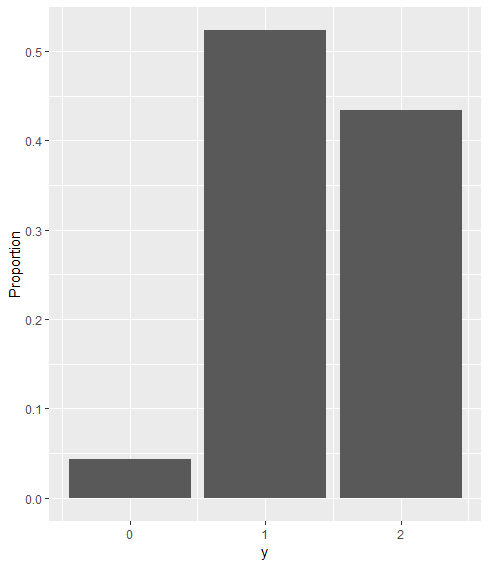
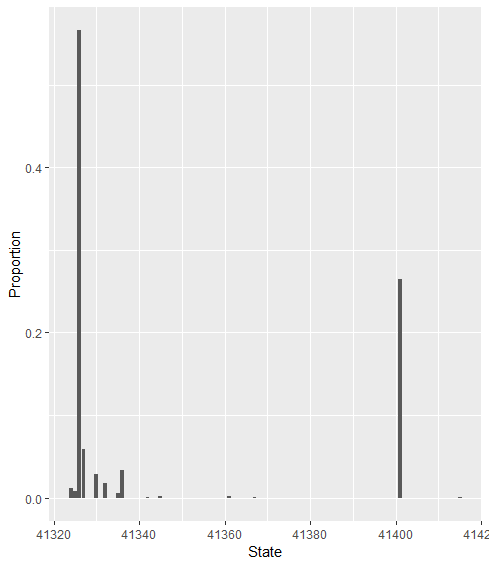




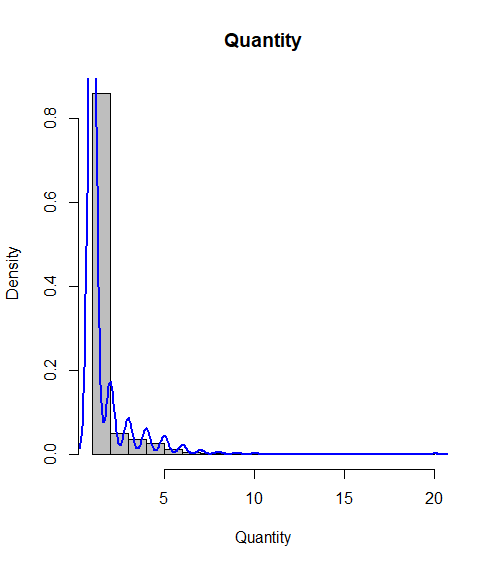
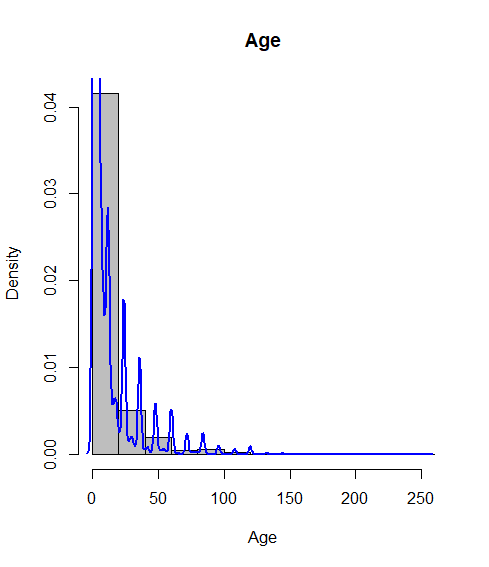
 

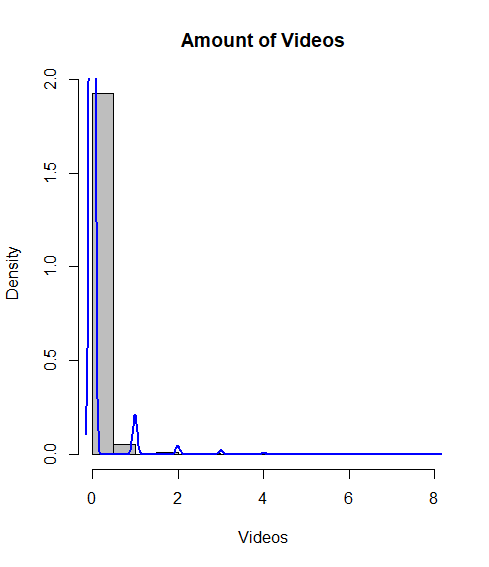
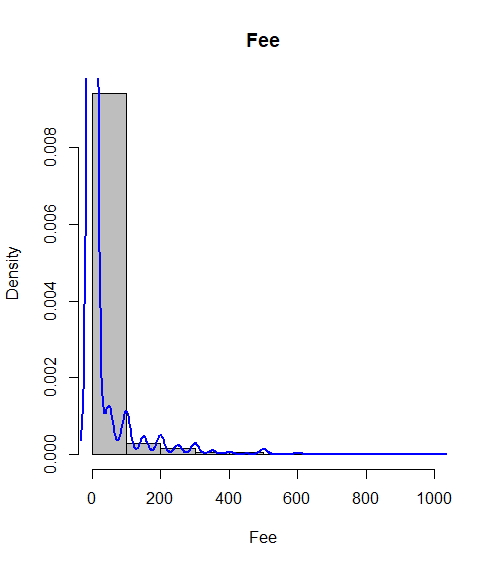
 

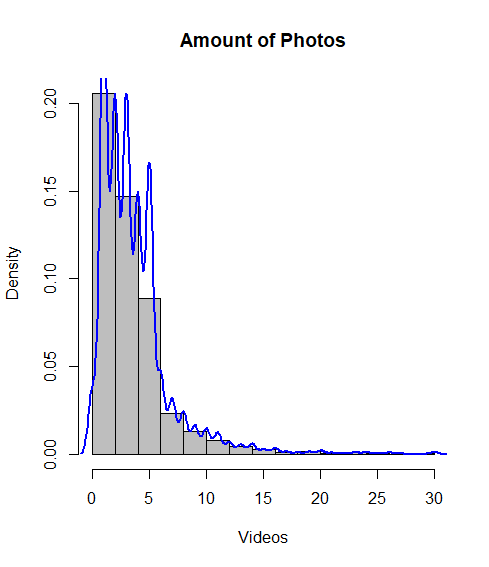




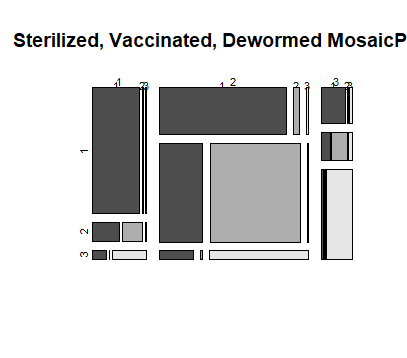
## משתנים רציפים – היסטוגרמות







## קשרים בין משתנים – גרפים וטבלאות



table(cleandata$y,cleandata$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