1 הגדרת הבעיה

1.1 תיאור כללי של עולם התוכן הנחקר

הבעיה המחקרית בה עוסק הפרויקט היא מציאת לקוחות פוטנציאליים להצטרפות לתוכנית פיקדון. בעזרת נתוני שיחות העבר ונתונים לגבי הצטרפות\ אי הצטרפות לתוכנית חיסכון נרצה לבנות מודל שיכלול בתוכו את הגורמים המשפיעים על החלטת הלקוח בכדי להפוך את קמפיין תוכנית החיסכון לאפקטיבי יותר ופחות מטריד ללקוחות לא מעוניינים.

מחקרים מראים כי מציאת לקוחות חדשים לחברה הינה משימה מורכבת שדורשת השקעה רבה, וגוזלת זמן רב ומשאבים שונים. הגדלת מספר הלקוחות משמעותה הגדלת מקור ההכנסה לחברה ותפיסת נתח שוק רחב יותר. ישנן מספר דרכים בכדי לגייס לקוחות חדשים (פרסומות, אתרי אינטרנט וכו'), אך נמצא כי שימוש בנתונים על לקוחות קיימים וניתוחם היא אחת הדרכים היעילות לגייס. כלומר, ניתוח של הלקוחות הקיימים תוך ניסיון להבין מהו המכנה המשותף ביניהם, או מהם הגורמים שהביאו לכדי הימצאותם בין לקוחות החברה, הוא שלב חשוב ביותר בגיוס מוצלח של לקוחות. (Thorleucther, Poel, & Prinzie, 2012).

Neha More (2017) ניסה לענות על הבעיה המחקרית ולמצוא מערכת לימוד מכונה שתדע לחזות באילו לקוחות כדאי להתמקד בקמפיין שיווק תוכנית הפיקדון. לרשותו היו נתוני שיחות שכללו מאפיינים על הלקוח ואת ההחלטה לגבי הצטרפות\אי הצטרפות לתוכנית חיסכון, ובעזרתם בנה מודל של רגרסיה לוגיסטית אשר שימש לביצוע התחזיות לגבי לקוחות פוטנציאליים. הרגרסיה הלוגיסטית הצליחה לחזות בדיוק של 84.15% תוך שימוש בשיטת KS.

גם Alex Shelaev אשר בחן מספר מודלים לביצוע התחזית מצא כי מודל הרגרסיה הלוגיסטית סיפק את הערך score F1 הגבוה ביותר (F1 הוא ממוצע הרמוני של precision ו-recall). לעומת זאת, עבור לקוחות חדשים מצא כי התקבל F1 נמוך ביותר, כלומר ניתן להסיק שהמודל אינו מתאים לתחזיות הצלחת הקמפיין עבור לקוחות חדשים. ולמרות שסט הנתונים יחסית קטן, ערך F1 score מעיד כי המודל יהיה טוב לתחזית עבור לקוחות קיימים.

במחקר שפורסם ב-Journal of Marketing Research , טענו כי מודלים לחיזוי הפיקדונות כללו בדרך כלל 2 משתנים אשר נשלטים על ידי הבנק (1)שיעור הדיבידנד (2)הוצאות פרסום, ושלושה משתנים אשר אינם נשלטים על ידי הבנק (1)הכנסה לכל משק בית (2)מלאי משותף (3)שיעורי רווח.

שאלת המחקר תעסוק במציאת הגורמים המשפיעים על ההחלטה של הלקוח להצטרף לתוכנית פיקדון. לאחר מציאת הגורמים המשפיעים נרצה לבנות מודל מערכת לומדת אשר יוכל לנבות האם לקוח עתיד להצטרף לתוכנית החיסכון או לא על פי מאפייניו השונים. את התחזית נבצע בעזרת שימוש בכלי מערכות לומדות בכדי לבנות מודל סיווג יעיל ומדויק ככל האפשר. הסיווג יתבסס על שימוש בהיסטוריית נתוני שיחות טלפון ונתונים דמוגרפיים של הלקוח.

2 הבנת הנתונים

2.1 תיעוד מקורות הנתונים ומשמעותם

המידע נאסף במהלך השנים 2008-2013 בבנק בפורטוגל. במחקר ניתחו 150 מאפיינים הקשורים ללקוח הבנק, תוצרת ומאפיינים סוציאל-אקונומיים.
סט הנתונים שברשותנו מצומצם מעט מסך הנתונים שנאספו ומכיל 3406 רשומות עם 16 משתנים מסבירים ואחד מוסבר. אנחנו משערים כי תיעוד השיחות (יום, חודש) נשמר במערכת השיחות, ואת הנתונים הדמוגרפיים שלפו מנתוני הלקוחות בבנק.

<u>משמעות המאפיינים:</u>

Name	R name	Description	Type	Values
age	X1	גיל	Numeric	21-78
job	X2	סוג עבודה	Categorical	אדמיניסטרציה, לא ידוע, ניהול, עובד צווארון כחול, טכנאי
marital	X3	מצב משפחתי	Categorical	נשוי, גרוש (גרוש או אלמן), רווק
education	X4	השכלת הלקוח	categorical	תואר ראשון, שני, שלישי, לא ידוע
default	X5	מצב כרטיס האשראי	Categorical	בפיגור, לא בפיגור
balance	X6	יתרה שנתית ממוצעת בבנק (יורו)	Numeric	[-1865, 38126]
housing	X7	האם הלקוח לוקח משכנתא	binary	כן\לא
loan	X8	האם יש ללקוח הלוואות	binary	כן\לא
contact	X9	אופן יצירת הקשר	Categorical	טלפון, סלולארי, לא ידוע
day	X10	יום בשבוע בו הייתה ההתקשרות האחרונה עם הלקוח	Categorical	שני, שלישי, רביעי, חמישי, שישי

month	X11	החודש בשנה בו הייתה ההתקשרות האחרונה עם הלקוח	Categorical	ינואר - דצמבר
campaign	X12	מספר הפעמים שיצרו קשר עם הלקוח בקמפיין הנוכחי	Numeric	1-58
pday	X13	מספר הימים שעברו מאז הפעם האחרונה שנוצר קשר עם הלקוח בקמפיין הקודם	Numeric	ערך 1- אומר שלא נוצר קשר עם הלקוח 1-854
previous	X14	מספר הפעמים שיצרו קשר עם הלקוח לפני הקמפיין הנוכחי	Numeric	0-32
poutcome	X15	תוצאת הקמפיין הקודם	Categorical	הצלחה, כישלון, אחר, לא ידוע
gender	X16	מין הלקוח	Categorical	זכר, נקבה, לא ידוע
y	Y	האם הלקוח הסכים לתכנית?	binary	0/1

2.2 הסתברויות אפריוריות וקשרים בין מאפיינים

- ## הסתברויות אפריוריות

בחלק זה נחשב הסתברויות אפריוריות על סמך הנתונים הקיימים. את חישוב ההסתברויות נחשב לפי יחס של ערך משתנה מסוים חלקי סך כל הרשומות (בהורדת נתונים חסרים\לא ידועים).

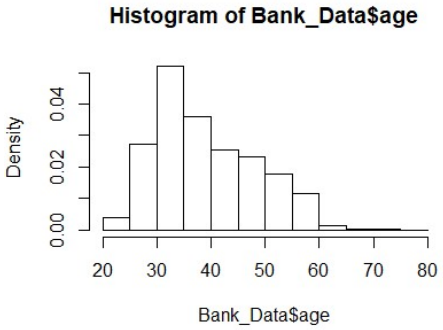
<u>משתנה המטרה Y</u>

עבור משתנה המטרה נבדוק מה ההסתברות האפריורית לכך שלקוח יסכים להצטרפות לתוכנית (yes) ואת ההסתברות האפריורית לכך שלקוח לא יסכים להצטרף לתוכנית (no). את ההסתברויות נחשב בעזרת השכיחויות של כל אחת מהאופציות להחלטה על סמך נתוני השיחות. <u>(נספח 1)</u>

$$P(yes) = 0.34997 \qquad P(no) = 0.65003$$

<u>מאפיין גיל age</u>

משתנה הגיל הוא משתנה רציף ולכן נוכל להציג את ההסתברויות האפריוריות בעזרת היסטוגרמה.

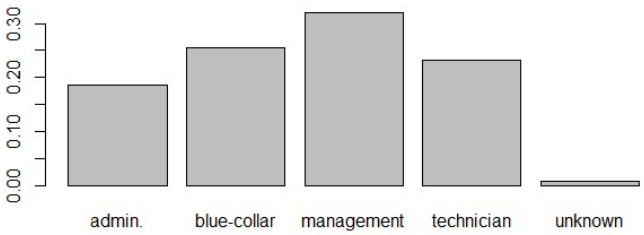


נבצע חלוקה לדליים בכדי לחשב הסתברויות אפריוריות,

$$P(25-) = 0.0205$$
$$P(26-30) = 0.1368$$
$$P(32-40) = 0.4404$$
$$P(41-55) = 0.3338$$
$$P(56+) = 0.0684$$

<u>מאפיין סוג העבודה job</u>

משתנה קטגוריאלי בעל 5 ערכים שונים, את הפיזור שלו נוכל להציג בעזרת תרשים bar.

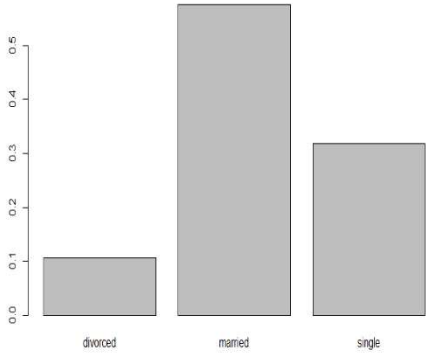


$$P(unknown) = 0.0085 \quad P(blue\ worker) = 0.2547$$
$$P(admin) = 0.1861 \quad P(managment) = 0.3194$$
$$P(technician) = 0.2313$$

ניתן לראות כי ישנה הסתברות קטנה מאוד של ערך "לא ידוע" עבור סוג העבודה (פחות מ1%), ולכן נתייחס רק לנתונים הקיימים בשביל למצוא את ההסתברויות האפריוריות המתוקנות.

$$P(blue\ worker) = 0.2567\ ;\ P(admin) = 0.1877$$
$$P(managment) = 0.3221\ ;\ P(technician) = 0.2333$$

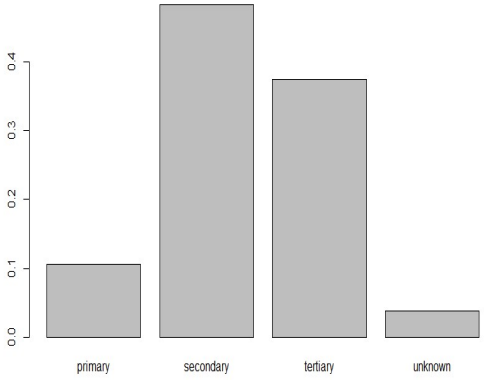
מאפיין מצב משפחתי marital



זהו משתנה קטגוריאלי המתאר את המצב המשפחתי של הנבדק. ניתן לזהות 3 סוגים של מצב משפחתי בdata שלנו – נשוי\רווק\גרוש.

$$P(single) = 0.3183$$
$$P(divorced) = 0.1059$$
$$P(mariied) = 0.5757$$

מאפיין השכלת הלקוח education



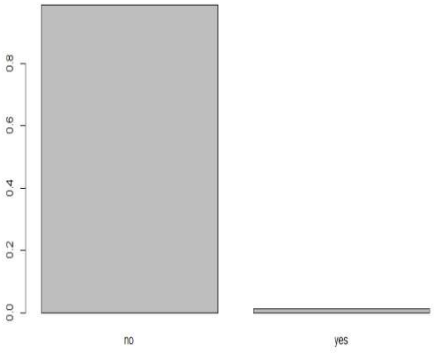
משתנה קטגוריאלי, אשר 3.76% מהנתונים אינם ידועים. חישוב ההסתברויות האפריוריות התבצע תוך חוסר התחשבות (ניקוי) של הנתונים הלא ידועים.

(ההיסטוגרמה כן כוללת הסתברות של ערכים חסרים)

$$P(primary) = 0.3893$$
$$P(secondary) = 0.5012$$
$$P(tertiary) = 0.1096$$

לפי דעתנו המדגם במקרה זה לא בדיוק מייצג את המציאות, ומתאר מדגם מהאוכלוסייה שהוא "משכיל" מאוד, אחוז בעלי התארים המתקדמים בנתונים הוא קרוב ל90%. עם זאת, יכול להיות שהיה ניסון מצד הבנק להתמקד באוכלוסייה "משכילה" יותר.

מאפיין מצב כרטיס אשראי default



משתנה מצב כרטיס אשראי מתאר האם הלקוח בפיגור או לא בפיגור. עבור סט נתונים זה אין נתונים חסרים, ונקבל את ההסתברויות האפריוריות הבאות –

$$P(yes) = P(delay) = 0.0135$$
$$P(no) = 0.9865$$

ניתן לראות שהרוב המוחלט של הלקוחות אינו בפיגור בכרטיס האשראי.

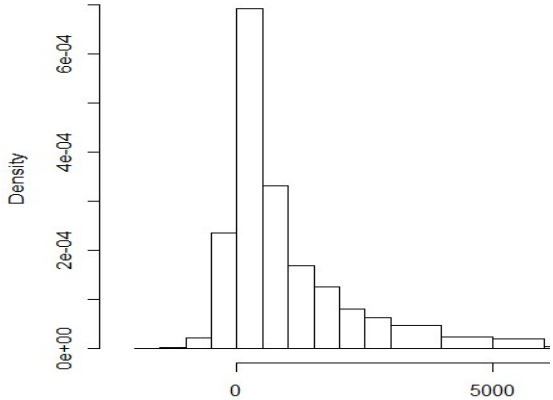
מאפיין יתרה שנתית ממוצעת בבנק balance

מצאנו את ערכי המינימום והמקסימום עבור נתונים אלו [-1865, 38126], הממוצע הוא 1520, החציון הוא 561.

לפי ההיסטוגרמה ראשונית (נספח 2), ראינו שההסתברות הולכת וקטנה עבור יתרה שנתית מעל 15,000 (כמעט אפס), באופן ברור יותר ניתן לראות כי -

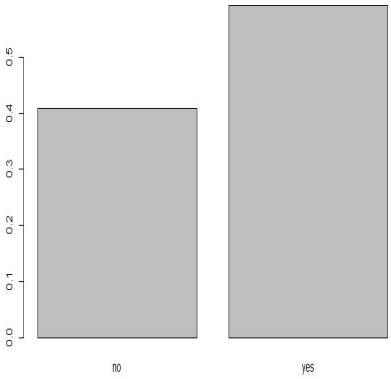
$$P(balance > 15000) = 1 - P(balance < 15000) = 1 - \frac{3380}{3406} = 7.633 * 10^{-3}$$

ולכן בחרנו להציג את ההיסטוגרמה באופן ספציפי יותר עבור הערכים ה"שולטים" בנתונים.



ההיסטוגרמה לפי דעתנו תואמת את המציאות, מכיוון שאכן הגיוני שיתרה שנתית גבוהה בבנק היא לא נפוצה, ואכן רוב היתרה מתפזרת סביב הממוצע, רק נדירים בעלי יתרה ממוצעת גבוהה מהרגיל (15000 ומעלה, עד אזור ה40,000).

$$P(balance < 0) = 0.1295\,;\;\; P(0 < balance < 500) = 0.3426$$
$$P(500 < balance < 1000) = 0.1644\,;\; P(1000 < balance < 2000) = 0.1468$$
$$P(2000 < balance < 4000) = 0.1183\,;\; P(balance > 4000) = 0.0984$$



<u>מאפיין משכנתא housing</u>

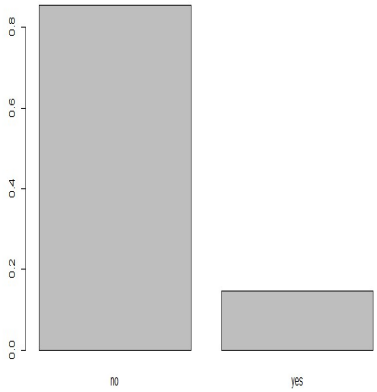
משתנה בינארי yes\no. נוכל למצוא הסתברות אפריוריות לפי היחס בין מספר התצפיות לערך המתאים חלקי מספר התצפיות הכולל.

$$P(yes) = 0.5916$$
$$P(no) = 0.4084$$

כלומר בנתונים שלנו 60% מהלקוחות לוקחים משכנתא וכ40% לא לוקחים משכנתא. נראה אכן סביר בהתייחס למציאות.

<u>מאפיין הלוואות loan</u>

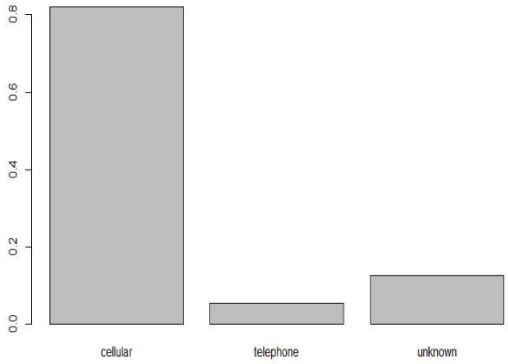
האם לקוח מסוים לוקח הלוואה, משתנה בינארי שגם לו נוכל להציג הסתברויות אפריוריות. גם כאן היחס נראה הגיוני למדי ביחס למציאות.



$$P(yes) = P(taking\ loan) = 0.1459$$
$$P(no) = P(not\ taking\ loan) = 0.8541$$

<u>מאפיין אופן יצירת הקשר contact</u>

משתנה קטגוריאלי – טלפון \סלולר\ לא ידוע. אנחנו מניחים ששתי האופציות היחידות ליצירת קשר הן טלפון או סלולר, ולכן לא נתייחס לנתונים הלא ידועים לחישוב הסתברויות אפריוריות. (הנתונים הלא ידועים מהווים כ12.5% מסך הנתונים) $\frac{426}{3406} = 0.125$



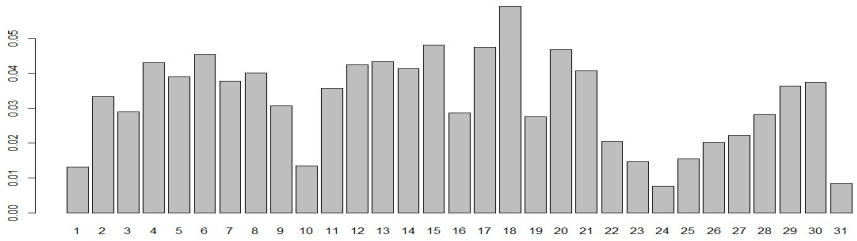
$$P(cellular) = 0.9372$$
$$P(telephone) = 0.0628$$

גם לאחר הורדת הנתונים החסרים, היתרון המשמעותי של השימוש בסלולר נשמר ולכן הנתונים הלא ידועים לא יכלו "להטות" את ההסתברות האפריורית.

ניתן לראות שרוב השיחות המתועדת בנתונים שלנו בוצעו לטלפון הנייד, נתון שאכן מתיישב עם המציאות בעידן הסלולרי של היום.

<u>מאפיין יום בחודש של התקשרות אחרונה day</u>

משתנה בדיד עם מספר רב של ערכים, נתאר בעזרת היסטוגרמה את ההסתברויות

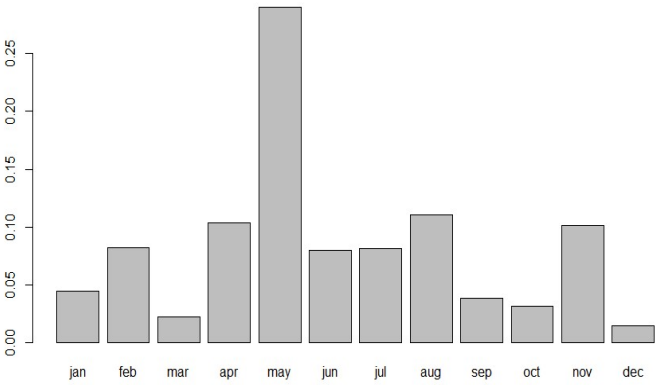


בשביל לראות הסתברויות אפריוריות נצטרך למצוא מה ההסתברות לכל אחד מהימים, אך כרגע בשלב הצגת ההסתברויות האפריוריות, נאחד את הימים לקבוצות – תחילת חודש (1-10), אמצע חודש (11-20) וסוף חודש (21-31).

$$P(beginning\ of\ the\ month) = \frac{1110}{3406} = 0.3259$$
$$P(middle\ of\ the\ month) = \frac{1436}{3406} = 0.4216$$
$$P(end\ of\ the\ month) = \frac{860}{3406} = 0.2525$$

<u>מאפיין חודש בשנה של התקשרות אחרונה month</u>

זהו משתנה קטגוריאלי של 12 ערכים (ישנם 12 חודשים במהלך השנה). נתאר את ההסתברויות בעזרת *barplot*.



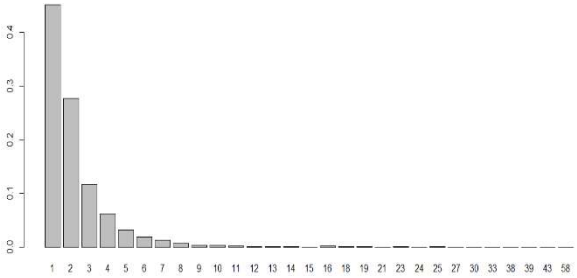
$$P(jan) = 0.0443\ ;\ P(feb) = 0.0822\ ; P(mar) = 0.0223$$
$$P(apr) = 0.01036; P(may) = 0.29;$$
$$P(june) = 0.0796\ ;\ P(july) = 0.0813\ ;\ P(aug) = 0.1107$$
$$P(sep) = 0.0385\ ;\ P(oct) = 0.0317\ ;$$
$$P(nov) = 0.1001\ ;\ P(dec) = 0.01438$$

נשקול בהמשך לאחד קטגוריות לפי עונות\רבעונים.

<u>מאפיין מספר פעמים שנוצר קשר בקמפיין נוכחי campaign</u>

משתנה בדיד המתאר ערך מספרי (שלם). ניתן לראות לפי ההיסטוגרמה שההסתברות לערכים גדולים מ9 קטנה במיוחד, ההסתברויות הגדולות ביותר יהיו לערכים נמוכים וככל שהערך עולה ההסתברות קטנה.

לכן נבחר לייצג הסתברויות אפריוריות בעזרת דליים (חלוקה לטווחים).



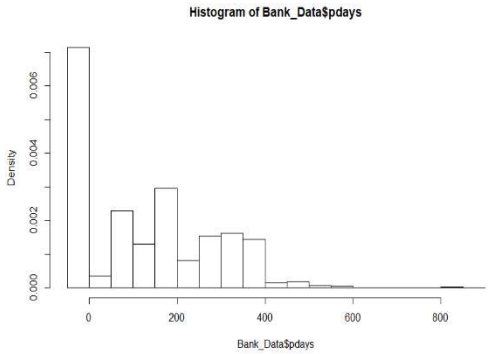
***X* – מספר הפעמים שנוצר קצר בקמפיין נוכחי.**

$$P(1 < x < 2) = 0.729\ ;\ P(3 < x < 5) = 0.211$$
$$P(6 < x < 8) = 0.03876\ ; P(x > 9) = 0.0212$$

התפלגות המאפיין הזה אכן נראית לנו הגיונית. הגיוני שמספר השיחות הנפוץ ביותר יהיה נמוך, ורק לחלק מהלקוחות נצליח להתקשר מספר רב יותר של פעמים.

<u>מאפיין מספר ימים שעברו מאז התקשרות בקמפיין קודם pday</u>

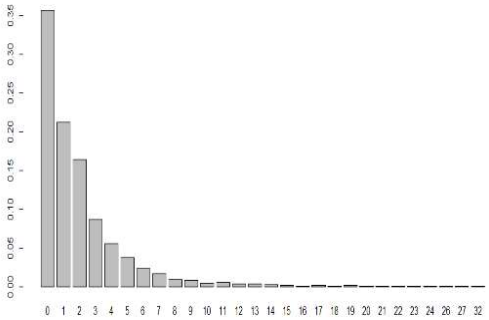
גם כאן מדובר במשתנה בדיד. ניתן לראות לפי ההיסטוגרמה שעם רוב הלקוחות לא יצרנו קשר בעבר (עבורם ערך המשתנה הנוכחי יהיה 1-). גם כאן נרצה לחלק את המשתנה לדליים בכדי לחשב הסתברויות אפריוריות.



$$P(no\ contact\ in\ the\ past) = 0.3564$$
$$P(0 < pdays < 200) = 0.3444$$
$$P(201 < pdays < 400) = 0.2716\ ; P(pdays > 401) = 0.0276$$

<u>מאפיין מספר פעמים שנוצר קשר עם הלקוח לפני קמפיין נוכחי previous</u>

מדובר במשתנה בדיד, לכן נציג את ההיסטוגרמה של המשתנה בכדי להבין את התנהגות ההסתברות.



ניתן לראות שיש צורך בחלוקה לדליים בכדי לחשב הסתברויות אפריוריות.
נבצע את החלוקה לפי קיבוץ בין ערכים סמוכים שנראים שמשפיעים באותו האופן.

$$P(previous = 0) = 0.3564\ ;\ P(1 < previous < 2) = 0.3764$$
$$P(3 < previous < 5) = 0.178\ ;P(previous > 6) = 0.0872$$

גם כאן, ההסתברויות מתארות נתונים המתיישבים עם ההיגיון. הגיוני שעם רוב הלקוחות לא הצלחנו ליצור קשר לפני הקמפיין הנוכחי, והסתברות קטנה יותר למספרים גדולים יותר של פעמים.

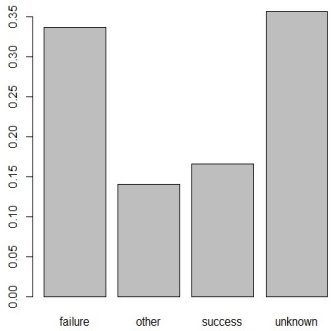
<u>מאפיין תוצאת קמפיין קודם poudcome</u>

משתנה קטגוריאלי עם 4 ערכים אפשריים. ניתן לראות שבמקרה זה ישנם יחסית הרבה ערכים חסרים (כ35%).

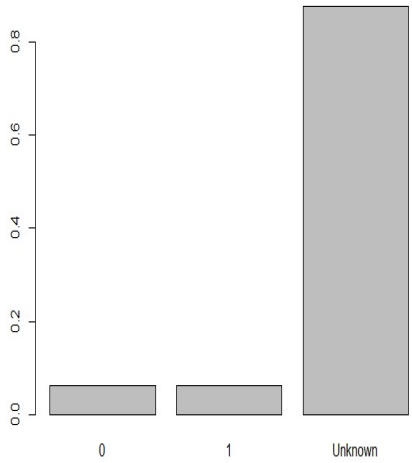
$$P(unknown) = 0.3567$$

נציג את ההסתברויות האפריוריות, ללא התייחסות לנתונים החסרים. (כלומר נחלק את מספר התצפיות של ערך מסוים במספר התצפיות הכולל ללא ערכים חסרים)

$$P(failure) = 0.5235\ ;\ P(sucesses) = 0.2578\ ;\ P(other) = \ 0.2186$$



<u>מאפיין מין הלקוח gender</u>



משתנה קטגוריאלי עם 3 ערכים – גבר\אישה\לא ידוע. ניתן לראות בהיסטוגרמה כי הרוב המוחלט לערך זה הוא "לא ידוע".

$$P(male) = \ 0.0616\ ;\ P(female) = 0.06312\ ;\ P(unknown) = 0.8752$$

בהתעלם מהנתונים החסרים, נוכל לחשב הסתברויות אפריוריות לנשים וגברים (רק באופן יחסי לתצפיות של נשים וגברים), ונקבל נתונים מאוזנים בין המינים, דבר שהגיוני לצפות לקבלו.

$$P(male) = \frac{0.0616}{0.0616 + 0.06312} = \ 0.4941$$

$$P(female) = \frac{0.06312}{0.0616 + 0.06312} = 0.5059$$

- ## האם סט הנתונים מאוזן והאם מייצג את המציאות

לפי ההסתברויות האפריוריות שבחנו בחלק הקודם, ניתן לראות כי סט הנתונים הנתון לא בהכרח מאוזן. נבחן תחילה את **משתנה המטרה y** (האם הלקוח צפוי להצטרף לתוכנית פיקדון)

$$P(yes) = 0.34997 \qquad P(no) = 0.65003$$

ניתן לראות כי השכיחות של כל אחת מהאופציות שונה. כאשר משתנה המטרה אינו מאוזן, אנחנו בעצם מאמנים את המערכת שלנו על סטים שאינם זהים ויכול להיווצר מצב של הטיה לכיוון הערך הנפוץ יותר. עם זאת, אנו מאמינים שבמקרה הספציפי הזה, טעות מסוג ראשון (כלומר לחזות שלקוח יצטרף אך שבפעול הוא אינו מעוניין) "זולה" יותר

מאשר טעות מסוג שני (לחזות שלקוח אינו יצטרף למרות שבמידה והיו פונים אליו הוא אכן יהיה מעונין להצטרף). לכן בשלב זה, לא נעשה מאמץ בכדי לאזן את סט הנתונים.
בנוסף, אפשר לראות שיש חוסר איזון גם עבור חלק מהמשתנים המסבירים.

מאפיין הגיל אינו מאוזן, אותו הדבר לגבי מאפיין **יתרה שנתית**. ניתן לזהות עבור שני מאפיינים אלו היסטוגרמה שאינה בעלת גבהים זהים. לגבי מאפיין הגיל ניתן לראות כי הנתונים לא לגמרי מתיישבים עם המציאות בפורטוגל. אבל לא כל האוכלוסייה בפורטוגל הינה האוכלוסייה של קהל הלקוחות בבנקים שאליהם פונים במטרה להצטרפות בפיקדונות, לכן לא ציפינו לקבל התאמה מלאה בין ההסתברויות האפריוריות של סט הנתונים לבין הנתונים על האוכלוסייה בפורטוגל.

	שם הקבוצה בR	סט נתונים	מציאות בפורטוגל
מתחת ל25 (15-25)	**A**	2.05%	12.01%
25-30	**B**	13.68%	6.164%
30-40	**C**	44.04%	15.6%
40-55	**D**	33.38%	26.17%
55 ומעלה	**E**	6.84%	40%

מאפיין **סוג העבודה** יחסית מאוזן, באופן כללי ההסתברויות האפריוריות מתחלקות באופן כמעט זהה בין 4 האופציות השונות (עובדי צווארון כחול – 25.67%, עובדי אדמיניסטרציה – 18.77%, עובדי ניהול – 32.221%, טכנאים – 23.333%). במקרה זה נעריך כי המאפיין מייצג את המציאות, כי ככל הנראה עובדי ניהול בעלי פוטנציאל גבוה יותר להצטרפות לפיקדון ולכן רוב הפניות מופנות אליהם. באופן כללי, ניתן לראות כי ישנה פנייה לתחומים נרחבים של סוגי עבודה, והפניות יחסית זהות בין התחומים השונים.

עבור מאפיין **מצב משפחתי** ניתן לזהות חוסר איזון, ישנם 31.83% רווקים, 10.59% גרושים ו57.57% נשואים. עם זאת, כנראה שאחוזים אלו מתיישבים עם המציאות.

מבחינת מאפיין **ההשכלה**, ניתן לראות כי הרוב הגדול בעלי השכלה גבוהה, 38.93% תואר ראשון, 50.12% תואר שני ורק 10.96% תואר שלישי. מבחינת התאמה למציאות, ניתן לראות כי באופן כללי הפניות מופנות לאוכלוסייה יותר משכילה (כ60% בעלי תואר שני\שלישי), לדעתנו התפלגות זו אינה בהכרח מתארת את המציאות, אך כאמור יכול להיות שהיה מאמץ לפנות לאוכלוסייה "משכילה" יותר ולהציע לה את תכנית הפיקדון.

עבור מאפיין **פיגור בכרטיס אשראי** ניתן לזהות חוסר איזון מוחלט. מספר הלקוחות שבפיגור הם כ 1.35%, ולעומת זאת, הלקוחות שאינם בפיגור מהווים 98.65%. כלומר הרוב המוחלט של הלקוחות אינם בפיגור, ולכן ככל הנראה המאפיין הזה אינו מספק מידע מעניין מכיוון שהוא זהה כמעט בכל הרשומות, וכמעט חסר ערך לפלג לפיו.

עבור מאפיין **לקיחת המשכנתא**, ניתן לראות שהנתונים יחסית מאוזנים. ההסתברות ללקיחת משכנתא היא 59.16%, ולעומת זאת ההסתברות לאי לקיחת משכנתא היא 40.84%. בהחלט יחס שיכול להתיישב עם נתוני המציאות.

מאפיין **לקיחת הלוואה**, ניתן לראות כי 14.59% מהלקוחות לקחו הלוואה והשאר, 85.41% אינם לוקחים הלוואה. יחס זה יכול לתאר את המציאות, או שאינו בהכרח רחוק מכך.

עבור מאפיין **יצירת קשר** ניתן לזהות חוסר איזון. הרוב המוחלט של ההתקשרויות בוצעו על ידי טלפון סלולרי 93.72%, אך בהחלט יחס שמתאר נכונה את המציאות בעידן הסלולרי של היום.

עבור מאפיין **יום בחודש** ניתן לראות שישנם ימים מסוימים שבעלי הסתברות גבוהה יותר וימים מסוימים בעלי הסתברות נמוכה יותר. במידה ומקבצים את הערכים לדליים וניתן לזהות איזון בין תחילת חודש, אמצע חודש וסוף חודש.

מאפיין **חודש בשנה** אינו מאוזן, ניתן לראות שבאופן קיצוני רוב השיחות בהתקשרות האחרונה בוצעו בחודש מאי – 29%.

מספר פעמים שנוצר קשר בקמפיין נוכחי **ומספר ימים שעברו** מאז התקשרות אחרונה אינם מאוזנים. ניתן לזהות ירידה קיצונית בהיסטוגרמה, כך שההסתברות הולכת וקטנה ככל שמספר הימים גדל\מספר הפעמים גדל .

עבור מאפיין **תוצאת הקמפיין הקודם**, ניתן לראות כי ישנו מספר גדול של נתונים חסרים 35.67% ולכן קשה לקבוע האם מאפיין זה מאוזן או לא. בהתעלמות מהנתונים החסרים, ניתן לראות כי רוב התוצאות הנצפו היו כישלון – 52.35%.

עבור מאפיין **מין הלקוח** ניתן לזהות מספר גדול במיוחד של נתונים חסרים (87.52%). בהתעלם מהנתונים החסרים (הסתכלות רק על כ400 רשומות), ניתן לראות כי קיים איזון בין מספר הגברים והנשים. מבחינת התאמה למציאות, נבחן את ההתפלגות אל מול נתוני המין בפורטוגל.

	סט הנתונים	מציאות פורטוגל
גברים	49.51%	49%
נשים	50.59%	51%

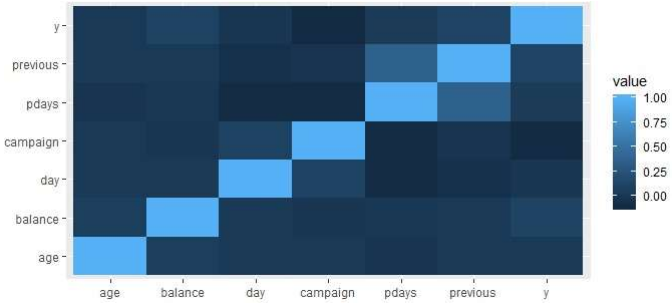
ניתן לראות בבירור שסט הנתונים, בהקשר של מאפיין המין, תואם להתפלגות בפורטוגל.

- ## קשרים מעניינים בין המאפיינים

בכדי להבין לעומק את הנתונים, נרצה לבחון קשר בין מאפיינים ובין מאפיינים למשתנה המטרה Y.
ישנו מספר לא מועט של משתנים קטגוריאליים שאינם רציפים, ולכן קשה למצוא מקדמי מתאם עבור משתנים אלו, גם בינם לבין עצמם וגם בינם לבין משתנים נומריים. נתחיל בניתוח מטריצת מקדמי המתאם של המשתנים הנומריים בסט הנתונים.

<u>מטריצת קורלציות בין משתנים נומריים</u>

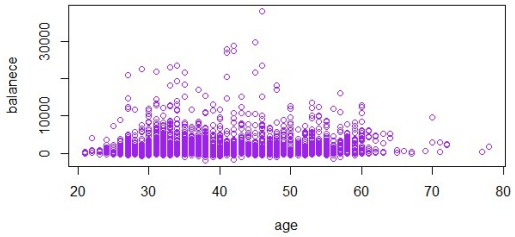
	age	balance	day	campaign	pdays	previous	y
age	1.00	0.07	0.02	0.01	-0.03	0.02	0.01
balance	0.07	1.00	0.02	-0.01	0.00	0.01	0.10
day	0.02	0.02	1.00	0.09	-0.11	-0.06	-0.01
campaign	0.01	-0.01	0.09	1.00	-0.11	-0.03	-0.12
pdays	-0.03	0.00	-0.11	-0.11	1.00	0.36	0.04
previous	0.02	0.01	-0.06	-0.03	0.36	1.00	0.11
y	0.01	0.10	-0.01	-0.12	0.04	0.11	1.00



ניתן לראות כי המתאמים לרוב נמוכים. המתאם הנצפה הגבוה ביותר הוא 0.36 בין מאפיין previous לבין pdays.

<u>קשר בין גיל לבין יתרה ממוצעת</u>

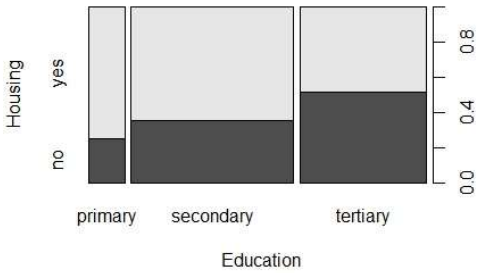
בחרנו לבחון קשר זה מכיוון שאנחנו משערים שככל הנראה ניתן לזהות דפוס מסוים בין הגיל לבין היתרה הממוצעת בבנק.



לפי גרף הפיזור, ניתן לראות כי היתרה השנתית הגבוה ביותר נצפית בגילאים צעירים, כלומר בדרך כלל בגילאים 30-50. עבור גילאים נמוכים\גבוהים יותר ניתן לצפות ביתרה שנתית נמוכה יחסית. אמנם, מקדם המתאם בין המשתנים יצא 0.069, ולא נראה כי קיימת מגמה חזקה במיוחד.

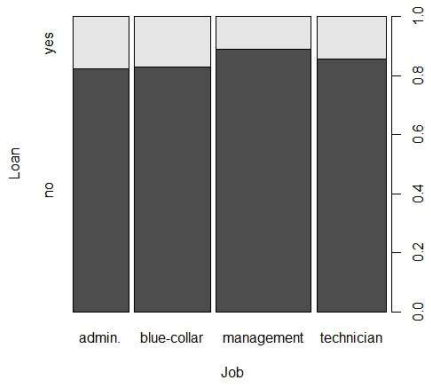
<u>קשר בין השכלה ללקיחת משכנתא</u>

ניתן לראות כי עבור השכלה נמוכה יותר, ההסתברות לקחת משכנתא גדולה יותר, ולהיפך, עבור רמת השכלה גבוהה יותר, ההסתברות לקחת משכנתא נמוכה יותר.



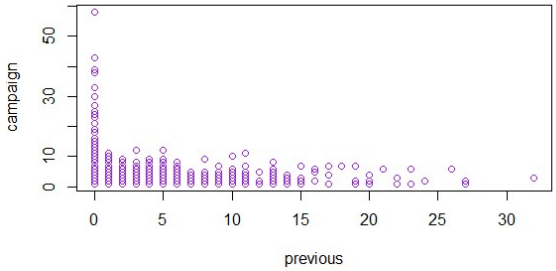
קשר בין סוג העבודה ללקיחת הלוואה

ניתן לראות שעבור המשתנה הקטגוריאלי סוג עבודה, לא ניתן לזהות שוני משמעותי בין לקיחת הלוואה או אי לקיחת הלוואה.



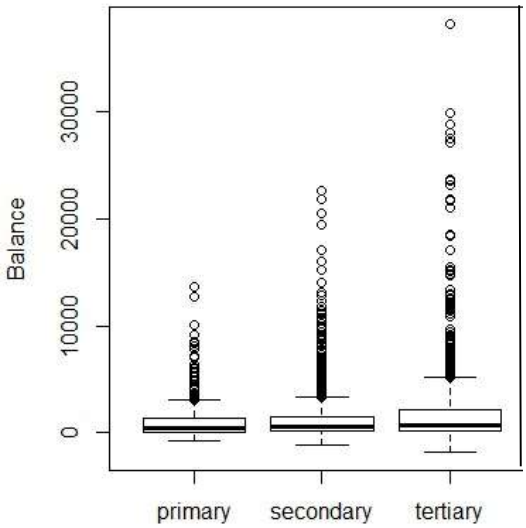
קשר בין מספר פעמים שיצרו קשר עם הלקוח בקמפיין הקודם לבין מספר הפעמים שיצרו קשר עם הלקוח בקמפיין נוכחי

בחרנו לבחון קשר זה בכדי לראות אם ישנה השפעה של מה שקרה בקמפיין הקודם על מה שמזהים בקמפיין הנוכחי. מקדם המתאם של המשתנים יוצא קשר שלילי וחלש (0.0326-). עם זאת, ניתן לראות שעבור לקוחות שכלל לא נוצר איתם קשר בקמפיין הקודם, ניתן לראות שמספר הפעמים שנוצר קשר בקמפיין הנוכחי הוא הגבוה ביותר. כלומר, ככל הנראה מנסים ליצור הכי הרבה קשר עם לקוחות שלא יצרנו איתם קשר בעבר. מעבר לאבחנה זו, ניתן לראות שאין השפעה ולכן המתאם יוצא חלש.



קשר בין רמת ההשכלה ליתרת בנק שנתית ממוצעת

ניתן לראות כי אפשר לזהות מתאם בין רמת ההשכלה לבין היתרה הממוצעת בבנק. ככל שרמת ההשכלה גבוהה יותר ניתן להניח כי העובדים עובדים בעבודות שבהן המשכורות גבוהות יותר, ועל כן ניתן לזהות יתרת בנק שנתית גבוהה יותר. בנוסף, ניתן לראות כי קיימים ערכים קיצוניים יותר עבור הקטגוריה תואר שלישי, כלומר עבור ערך זה נקבל את טווח הערכים הגדול ביותר לעומת הקטגוריות השונות.

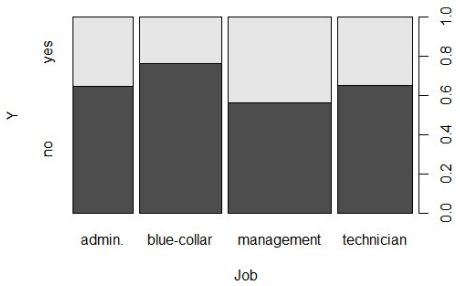


- ## מאפיינים חשודים כמשפיעים על משתנה המטרה

ישנם מספר מאפיינים שניתן לחשוד בהם בתור משפיעים על משתנה המטרה Y. בחלק זה נתייחס להשערות וידע על סמך מחקרים קיימים בנושא. כחלק מחקר סט הנתונים שקיבלנו, בחנו את הקשרים בין המאפיינים השונים למשתנה המטרה, בכדי לבחון את ההשערות ולבדוק אילו מאפיינים הם בעלי השפעה משמעותית על משתנה המטרה.

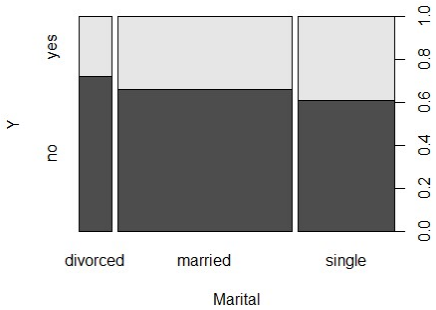
לפי מחקרים קודמים וידע מוקדם, אנחנו משערים כי המאפיינים המשפיעים ביותר על הצטרפות לתוכנית פיקדון יהיו סוג העבודה, מצב משפחתי, השכלה, יתרה שנתית ממוצעת בבנק, הלוואות ותוצאת הקמפיין הקודם. בדקנו את ההשערות בין משתנים אלו למשתנה המוסבר, וכן בנוסף ביצענו מבחן חי בריבוע לחוסר תלות בין מאפיינים אלו לבין המשתנה המוסבר, וראינו כי השערת אי תלות נדחתה בכל הקשרים הבאים ולכן נאמר שהם אכן משפיעים על המוסבר.

<u>קשר בין משתנה סוג העבודה למשתנה המוסבר</u>



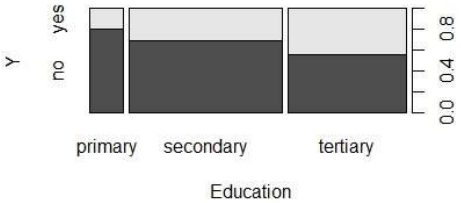
חשדנו במשתנה זה בתור משפיע על משתנה המטרה מכיוון שככל הנראה יש קשר בין סוג עבודה להכנסה, וככל שההכנסה של לקוח גדולה יותר כך יש יותר סיכוי שירצה להצטרף לתוכנית פיקדון.
ניתן לראות כי קיים הבדל בין סוג העבודה לשאלה האם לקוח יסכים להצטרף לתוכנית הפיקדון. עבור תפקידי ניהול ניתן לזהות הסכמה מקסימאלית מבין סוגי העבודה, לעומת עובדי צווארון כחול אצלם ניתן לראות את האחוזים הנמוכים ביותר להצטרפות.

<u>קשר בין משתנה מצב משפחתי למשתנה המוסבר</u>



חשדנו במאפיין זה מכיוון שחשבנו שזוגות נשואים, שלהם יש מחשבה על עתידם ועתיד ילדיהם, יצאו להשקיע יותר בתוכניות פיקדון, ולכן יהיה הבדל בינם לבין גרושים\רווקים.
עבור מאפיין מצב משפחתי ניתן לראות כי קיים שוני במשתנה המוסבר Y כתלות במצב המשפחתי. דווקא נראה שהרווקים הם בעלי אחוזים גבוהים ביותר להצטרפות לתוכנית פיקדון.

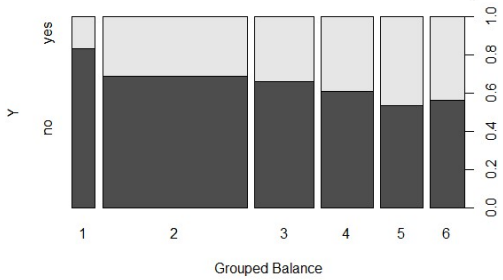
<u>קשר בין משתנה רמת ההשכלה למשתנה המוסבר</u>



חשדנו במאפיין זה בגלל ההבנה שיש קשר בין רמת השכלה לבין הכנסה. כפי שראינו קודם, עבור רמת השכלה גבוהה יותר ניתן לצפות ליתרה שנתית גבוהה יותר, ועל כן, כפי שכתבנו קודם, נצפה שלקוחות עם הכנסה גבוהה יותר ככל הנראה יהיו בעלי מוטיבציה גבוהה יותר להצטרפות לתוכנית פיקדון.
ניתן לראות כי ככל שרמת ההשכלה של הלקוח עולה, כך ההסתברות שהוא יסכים לתכנית הפיקדון תהיה גבוהה יותר. כלומר, קיים מתאם חיובי בין רמת ההשכלה לבין ההסכמה לתכנית הפיקדון.
מוצג תוצאת מבחן חי בריבוע שבוחן את התלות בין משתנים אלו, וראינו כי הp-value קטן מאוד, כלומר נדחה את השערת האפס ונאמר כי המשתנים אכן תלויים, כפי שזיהינו מהתרשים.

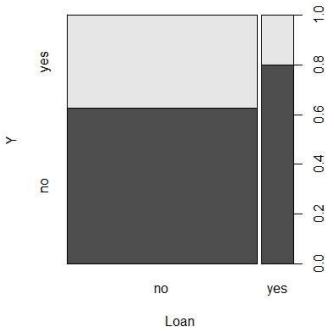
```
Pearson's Chi-squared test

data:  Bank_Data$education and Bank_Data$y
X-squared = 100.57, df = 3, p-value < 2.2e-16
```

<u>קשר בין משתנה יתרה שנתית ממוצעת למשתנה המוסבר</u>



כפי שצפינו, ניתן לראות שיש מתאם חיובי בין גודל ההכנסה להסתברות להצטרפות לתוכנית פיקדון.
ככל הנראה, לקוחות אשר יתרת הבנק השנתית שלהם גדולה יותר, יהיה להם יותר "כסף פנוי" אותו יוכלו להשקיע בתוכנית פיקדון, ועל כן נראה שההסתברות להצטרפות לתוכנית כזו גדלה ככל שהיתרה השנתית גדלה.

<u>קשר בין משתנה לקיחת הלוואה\לקיחת משכנתא למשתנה המוסבר</u>

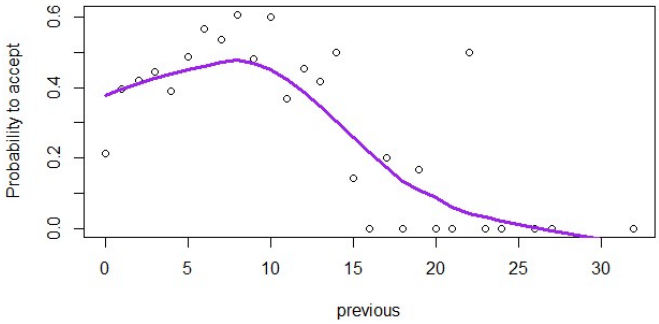


ניתן לראות כי יש קשר בין לקיחת הלוואה לבין המשתנה המוסבר. מבין אלו שלוקחים הלוואה, ניתן לראות כי 80% אינם מסכימים לתוכנית פיקדון, ולעומת זאת, מתוך אלו שלא לוקחים הלוואה, רק 60% אינם מסכימים לתוכנית הפיקדון.
אכן צפינו קשר זה מכיוון שהגיוני שאלו שלוקחים הלוואה ככל הנראה בעלי יתרה נמוכה יותר ולכן לא יתעניינו בתוכנית פיקדון שבה יש להפקיד חסכונות. לעומת זאת, לקוחות שלא לוקחים הלוואה ככל הנראה בעלי הון שניתן להשקיע בפיקדון ולכן פחות לקוחות יסרבו לתוכנית הפיקדון.
גם פה מוצג מבחן חי בריבוע, וניתן לראות כי דוחים את השערת האפס, כלומר קיימת תלות.

```
Pearson's Chi-squared test with Yates' continuity correction

data:  Bank_Data$loan and Bank_Data_yes_no$y
X-squared = 55.846, df = 1, p-value = 7.838e-14
```

גם עבור המשתנה לקיחת משכנתא ניתן לזהות את אותו סוג הקשר. מבין הלקוחות שלוקחים משכנתא, כ80% לא יסכימו לתוכנית הפיקדון, לעומת 50% שלא יסכימו מבין אלו שלא לקחו משכנתא. גם עבור משתנים אלו קיבלנו במבחן החי בריבוע כי דוחים את השערת האפס. (נספח 3)

קשר בין מספר פעמים שיצרו קשר עם לקוח לפני הקמפיין הנוכחי למשתנה מוסבר



ניתן לראות שבהתחלה, ככל שמספר הפעמים שיצרו קשר עם הלקוח בעבר עולה – כך יגדל הסיכוי שיסכים להצטרפות לקמפיין הנוכחי. עם זאת, אחרי מספר מסוימים של התקשרויות (כ8 פעמים), הסיכוי הולך ויורד. זה אכן הגיוני, בהתחלה יש משמעות להתקשרות, אך לאחר מכן ככל הנראה זה נתפס בעיני הלקוח כ"הטרדה" ולכן ישפיע לרעה על רצונו להצטרף לתוכנית הפיקדון.

קשר בין תוצאת קמפיין קודם למשתנה המוסבר

ניתן לראות הבדל מובהק בין תוצאות הקמפיין הקודם השונות. לפי הנתונים, ניתן לראות כי אם הלקוח הצטרף לתוכנית הפיקדון בקמפיין הקודם, יש לו הסתברות גבוהה להצטרף גם לקמפיין הנוכחי (כ80%), לעומת לקוחות שלא הצטרפו בקמפיין הקודם (רק כ30% יצטרפו בקמפיין הנוכחי).

2.3 איכות הנתונים

- ## נתונים חסרים

עבור סט הנתונים הנתון ישנם מספר מאפיינים בעלי נתונים חסרים.

מאפיין	הסתברות לנתונים חסרים	כמות נתונים חסרים	משמעות
job	0.85%	29	ניתן לראות כי מדובר בכמות מינורית לעומת גודל סט הנתונים הנתון, ולכן, ככל הנראה, לא יצור בעיה בהמשך. נשקול בהמשך השלמת נתונים אלו.
education	3.76%	128	עבור משתנה ההשכלה ניתן לראות כי כמות הנתונים החסרים הינה מאוד נמוכה, ולכן נשקול להשלים אותם בהמשך.
contact	12.5%	426	עבור מאפיין זה, הערך הנפוץ ביותר הינו באמצעות סלולר (90%), וגם אינו בעל השפעה גבוהה על משתנה המטרה. לכן, נשקול להסיר משתנה זה.
poutcome	35.67%	1215	עבור משתנה זה ניתן לראות כי קיימת כמות גדולה של נתונים חסרים. עם זאת, מצאנו כי למאפיין יש מתאם עם משתנה המטרה, ולכן מבחינה זו לא נרצה להסיר אותו.
gender	87.52%	2981	עבור משתנה המין מדובר על כמות גדולה מאוד של נתונים חסרים (הרוב המוחלט חסר). בנוסף, ניתן לראות כי מאפיין זה בעל מתאם נמוך עם משתנה המטרה. נשקול הסרה של מאפיין זה מסט הנתונים.

את המשתנים עבודה, השכלה ותוצאת קמפיין קודם נשלים בעזרת אלגוריתם MICE. האלגוריתם מבצע השלמה לנתונים חסרים. השיטה מתבססת על מפרט מותנה לחלוטין, כאשר כל משתנה חסר מושלם על ידי מודל נפרד. האלגוריתם יכול להתמודד עם השלמת נתונים קטגוריאליים רציפים, בינאריים, לא מסודרים ועוד. (נספח 16)

- ## נתונים שאינם הגיוניים

הרצנו פקודת summary על סט הנתונים, בכדי לראות את הטווחים של כל אחד מהמאפיינים. לאחר בחינת טווחי הנתונים מצאנו כי כל הנתונים נראים הגיוניים. (נספח 11)

3 הכנת הנתונים

3.1 בחירת מאפיינים

הסרת מאפיינים מסט הנתונים

מאפיין	תובנות	דרך פעולה
default	עבור מאפיין זה, ראינו כי 0.9865 מהערכים הינם "no", כלומר הרוב המוחלט של הלקוחות אינם מפגרים בתשלום האשראי.	משתנה זה אינו מוסיף מידע, מכיוון שזהה כמעט עבור כל הלקוחות הרשומים בסט הנתונים, ולכן נבחר להשמיט משתנה זה.
contact	עבור מאפיין זה, יש מספר גדול של משתנים חסרים, וכן ראינו כי מעבר לכך, הרוב המוחלט בוצע באמצעות סלולר ועל כן למאפיין זה אין השפעה רבה על משתנה המטרה.	בחרנו להסיר מאפיין זה בגלל חוסר השפעה על המשתנה המוסבר.
gender	ראינו כי עבור משתנה זה הרוב המוחלט של הנתונים חסרים (קרוב ל90%). מעבר לכך, אין שוני בין השפעת המין השונה על משתנה המטרה.	בחרנו להסיר משתנה זה מסט הנתונים.

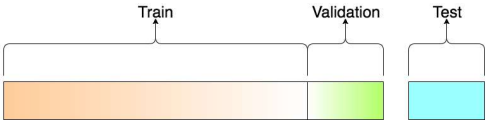
3.2 טיפול פרטני במאפיינים

חלוקה לקטגוריות

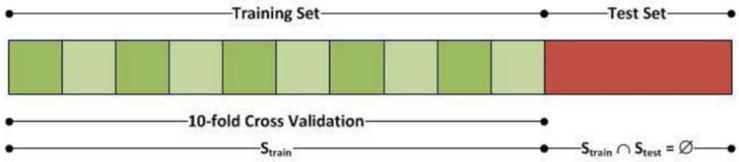
מאפיין	הסבר	חלוקה לקטגוריות
age	משתנה זה מתאר את גיל הלקוח אליו בוצעה ההתקשרות. עבור מאפיין זה ראינו כי יש טווח ערכים גדול במיוחד, ולכן נרצה לחלק את הגילאים לקבוצות גיל בכדי לפשט את המאפיין.	חלוקה ל5 קבוצות גיל: 25 ומטה, 26-30, 31-40, 41-55, 56 ומעלה.
balance	משתנה זה מתאר את יתרת הבנק הממוצעת השנתית של הלקוח. ניתן לראות סקלה נרחבת של ערכים, לכן גם כאן איחדנו לקטגוריות בכדי לפשט את ההסתכלות על המאפיין.	חלוקה ל6 קבוצות: יתרה שלילית (מתחת ל0), 0-500, 501-1000, 1001-2000, 2001-4000, 4001 ומעלה.
campaign	עבור משתנה זה זיהינו כי ישנו קשר יורד לבין משתנה המטרה. הקשר חזק עבור מספר פעמים נמוך. לעומת זאת, החל מערך מסוים של פעמים ניתן לראות כי המגמה קבועה ואין הבדל בהשפעה על משתנה המטרה.	חלוקה ל4 קבוצות: 0-2, 3-4, 5-6, 7 ומעלה. (נספח 12)

3.3 הכנת נתונים לאימון ובחינת מערכת לומדת

את סט הנתונים נדרשנו לחלק לשלושה חלקים – סט אימון (training), סט אימות (validation), ובחינה (test). את החלוקה ביצענו בשלבים – קודם כל ביצענו חלוקה לסט אימון training וסט בחינה test. בחרנו להקצות 20% מסט הנתונים לסט בחינה(681 נתונים), אחרי סקירה לגבי החלוקה הנהוגה לסדר גודל נתונים כמו שלנו. בכדי לאמן את המודל כפי שצריך, רצינו להקצות את מירב הנתונים לסט האימון (80%) בכדי לאפשר למודל להיחשף לרשומות שונות ומגוונות.



לאחר מכן, יש לחלק את סט האימון לסט אימון וסט אימות. את החלוקה נבצע בעזרת שיטת cross validation, הנקראת k-fold, אשר מבצעת חלוקה של סט האימון כולו לk חלקים, ובכל איטרציה מקצים k-1 מהקבוצות לסט האימון ואת הקבוצה הנותרת לסט אימות. שיטה זו הינה השיטה הפופולרית ביותר מבין שיטות cross validation ומעניקה בחינת כל אחת מהתצפיות גם בתור אימון וגם בתור אימות, ובוחרת את סט הנתונים שמפיק את נתוני הvalidation הטובים ביותר. בחרנו להשתמש בk=10, מכיוון שזהו הערך המומלץ ביותר לשיטה זו. כלומר, לפי ערך זה, מתקבל כי מתוך 80% המהווים 2725 נתונים, 10% המהווים 273 נתונים יהיו סט האימות, וכל השאר – 2452, יהיו סט האימון.



4 ביבליוגרפיה

1 https://medium.com/@abbdar/first-steps-in-machine-learning-predicting-subscription-for-bank-deposits-866516b90e4 First Steps in Machine Learning — Predicting Subscription for Bank Deposits\Alex Shelaev

2 http://www.rpubs.com/more11neha/banking Marketing campaign to sell term deposits: Predicting Target Customers\Neha More

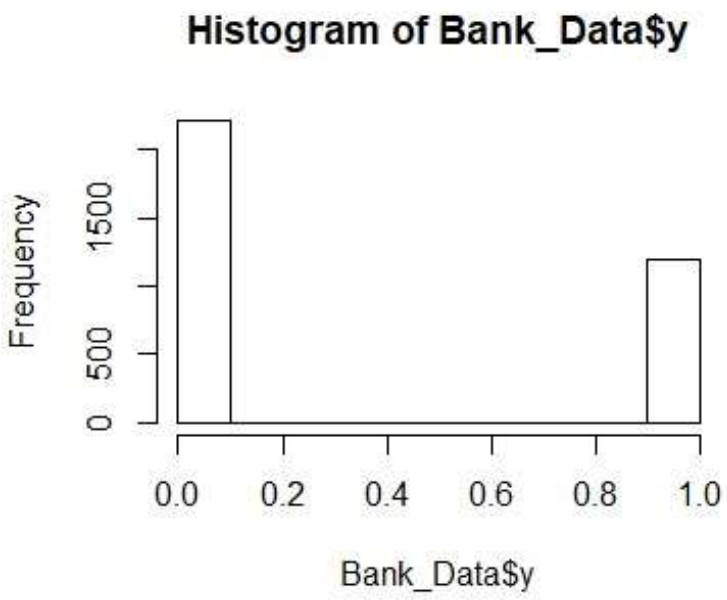
3 https://www.jstor.org/stable/pdf/3149513.pdf?refreqid=excelsior%3A4481b2dde37504691e6e83c2acee5001 Predicting Bank Deposits and Loans\G. David Hughes

4 Moro, S., Cortez, P., & Rita, P. (2014). A data-driven approach to predict the .success of bank telemarketing. Decision Support Systems, 62, 22-31

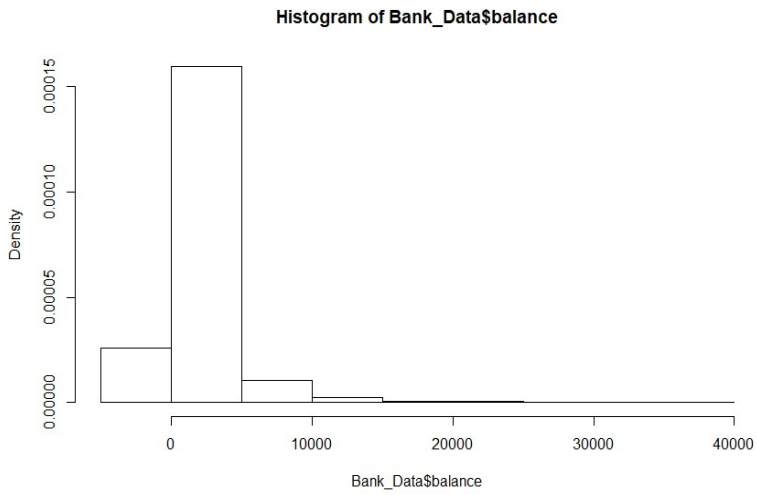
5 נספחים

5.1 היסטוגרמות - התפלגויות אפריוריות

נספח 1 – היסטוגרמה של משתנה המטרה

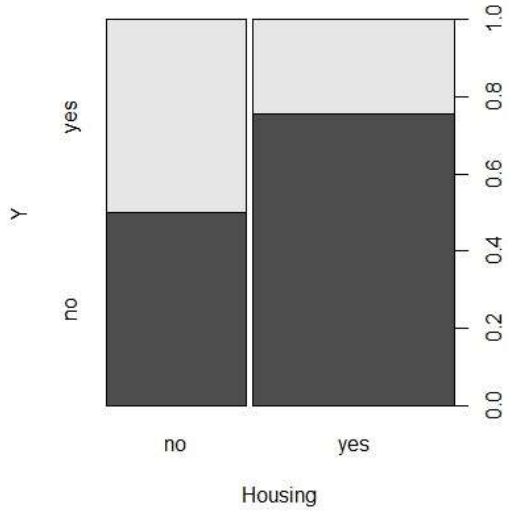


נספח 2 - היסטוגרמה של מאפיין יתרה שנתית



5.2 קשרים בין מאפיינים למשתנה מוסבר

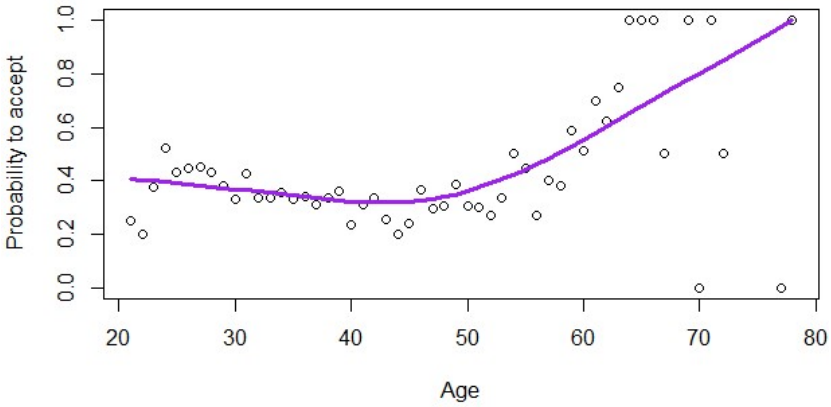
נספח 3 - קשר בין לקיחת משכנתא לבין משתנה מוסבר

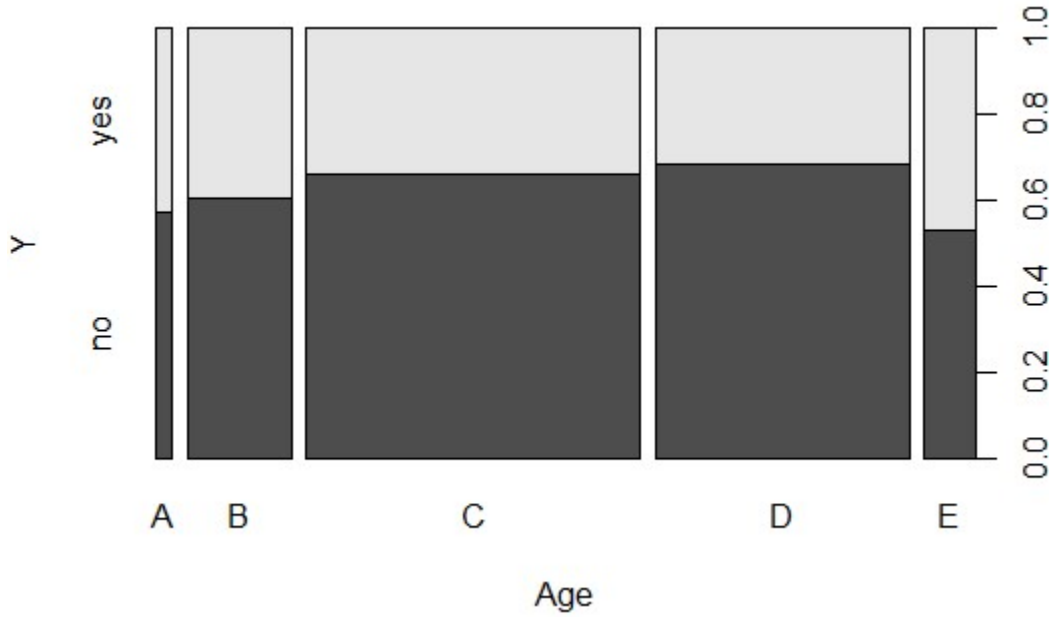



```
        Pearson's Chi-squared test with Yates' continuity correction

data:  Bank_Data$housing and Bank_Data_yes_no$y
X-squared = 237.12, df = 1, p-value < 2.2e-16
```

נספח 4- קשר בין משתנה מוסבר לגיל

עבור גיל כמשתנה רציף ניתן לחשוב כי ישנו מתאם חיובי עולה בין גיל לבין ההסתברות להצטרפות לתוכנית פיקדון. אך לפי ההסתברויות האפריוריות ראינו כי זה לא מייצג, כי מספר הלקוחות אינו זהה עבור הגילאים השונים.

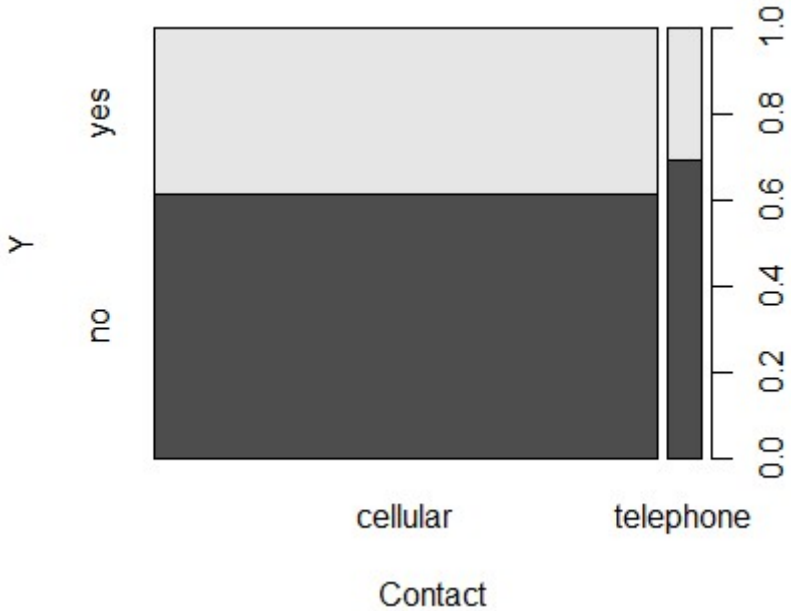




נרצה לבחון לפי קבוצות גיל, במקום עבור גיל ספציפי. כעת, ניתן לראות כי אין השפעה מובהקת של הגיל על המשתנה המוסבר.

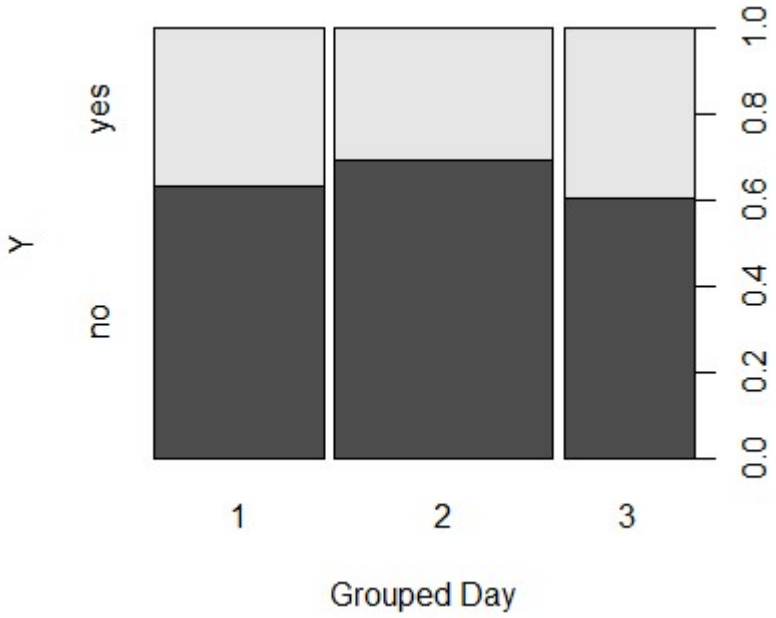
A – מתחת ל25
B – 25-30
C – 30-40
D – 40-55
E – מעל 55

<u>נספח 5 - קשר בין contact לבין משתנה מוסבר</u>



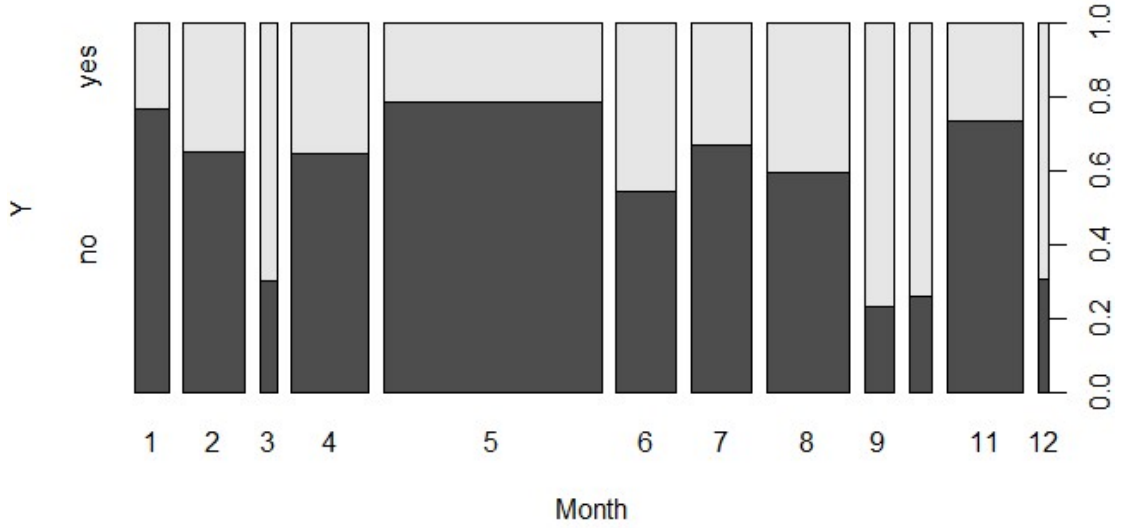
<u>נספח 6 - קשר בין יום בחודש למשתנה מוסבר</u>

ניתן לראות כי לא קיים הבדל משמעותי בין הימים השונים בהם בוצעה ההתקשרות האחרונה עם הלקוח.

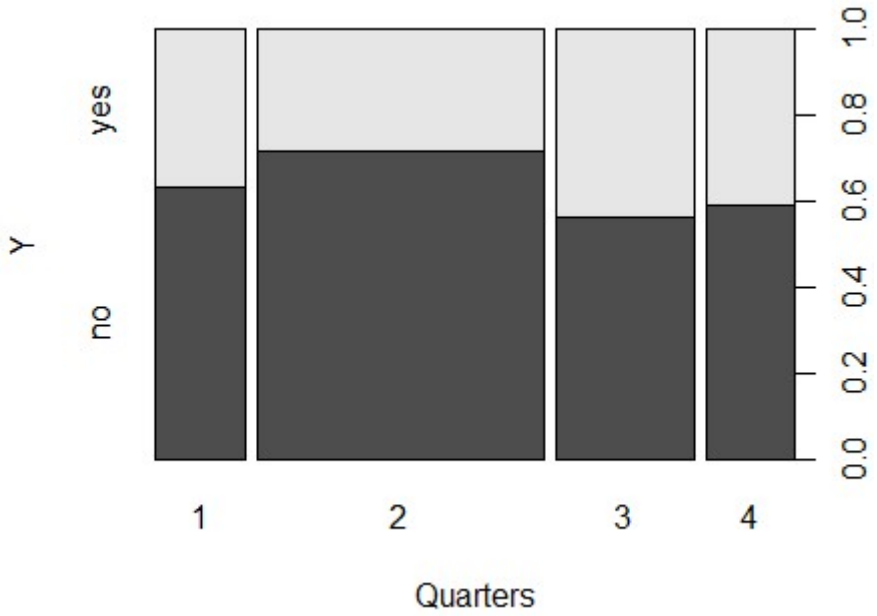


נספח 7 - קשר בין חודש למשתנה המוסבר

ניתן לראות כי אין הרבה הבדל בין הרבעון בו בוצעה ההתקשרות לבין ההחלטה שלו להצטרפות לתוכנית פיקדון.

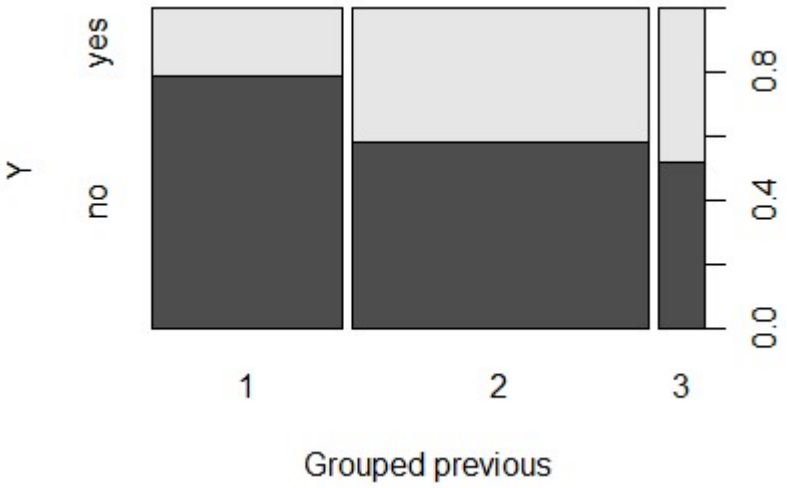


ועבור חלוקה לרבעונים נקבל:

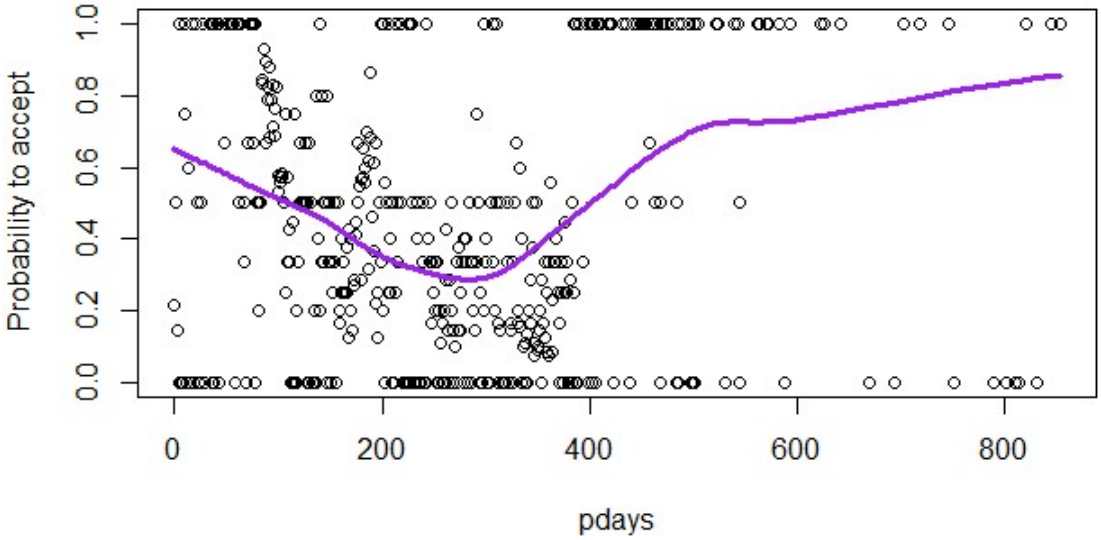


נספח 8 - קשר בין מספר פעמים שנוצר קשר עם לקוח לפני הקמפיין נוכחי למשתנה המוסבר

עבור חלוקה לקבוצות: 1- אפס פעמים, 2- פעם אחת עד חמש פעמים, 3- שש פעמים ומעלה.

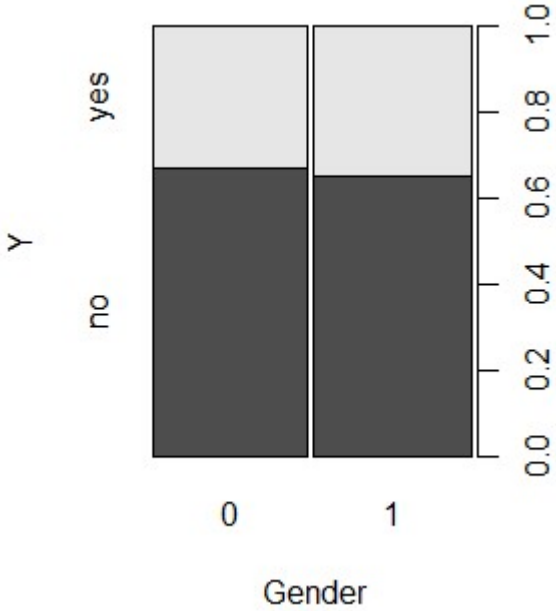


<u>נספח 9 - קשר בין מספר ימים שעברו מהפעם האחרונה שנוצר קשר למשתנה המוסבר</u>



<u>נספח 10 - קשר בין מין הלקוח למשתנה המוסבר</u>

ניתן לראות כי על פי הנתונים הקיימים, אין השפעה של מין הלקוח על הצטרפות לתוכנית פיקדון בקמפיין הנוכחי.



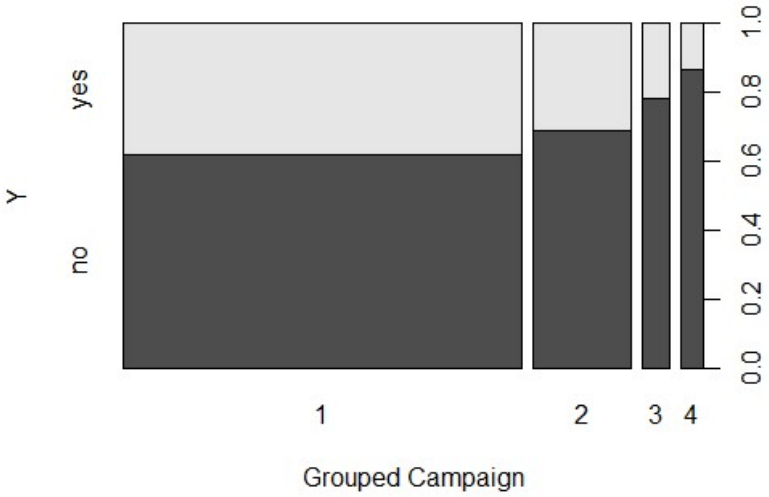
5.3 איכות נתונים

נספח 11- ניתוח summary

```
      age            job           marital        education     default      balance
 Min.   :21.00   admin.      : 634   divorced: 361   primary  : 359   no :3360   Min.   :-1865
 1st Qu.:32.00   blue-collar: 867   married :1961   secondary:1643   yes:  46   1st Qu.:  138
 Median :37.00   management :1088   single  :1084   tertiary :1276              Median :  561
 Mean   :39.65   technician : 788                   unknown  : 128              Mean   : 1520
 3rd Qu.:46.00   unknown    :  29                                               3rd Qu.: 1693
 Max.   :78.00                                                                  Max.   :38126

 housing      loan           contact        day            month       campaign        pdays
 no :1391   no :2909   cellular :2793   Min.   : 1.00   may    :988   Min.   : 1.000   Min.   : -1.0
 yes:2015   yes: 497   telephone: 187   1st Qu.: 8.00   aug    :377   1st Qu.: 1.000   1st Qu.: -1.0
                       unknown  : 426   Median :15.00   apr    :353   Median : 2.000   Median :104.0
                                        Mean   :14.96   nov    :344   Mean   : 2.358   Mean   :142.8
                                        3rd Qu.:21.00   feb    :280   3rd Qu.: 3.000   3rd Qu.:255.8
                                        Max.   :31.00   jul    :277   Max.   :58.000   Max.   :854.0
                                                        (other):787

    previous        poutcome        Gender          y
 Min.   : 0.000   failure:1147   0      : 210   Min.   :0.00
 1st Qu.: 0.000   other  : 479   1      : 215   1st Qu.:0.00
 Median : 1.000   success: 565   Unknown:2981   Median :0.00
 Mean   : 2.039   unknown:1215                   Mean   :0.35
 3rd Qu.: 3.000                                  3rd Qu.:1.00
 Max.   :32.000                                  Max.   :1.00
```

נספח 12 – חלוקה לקטגוריות מאפיין campaign

חלוקה לקטגוריות: 1: 0-2, 2: 3-4, 3: 5-6, 4: 7 ומעלה.



5.4 תיעוד קוד

נספח 13 – הסתברויות אפריוריות

```
Y_apriori_Yes <- length(which(Bank_Data$y==1))/length(Bank_Data$y)

Y_apriori_NO <- 1-Y_apriori_Yes

hist(Bank_Data$y)


hist(Bank_Data$age, freq=FALSE)

apriori_age_under_25 <- as.numeric(sqldf("select count (age) from Bank_Data where age <=25"
))/length(Bank_Data$age)

apriori_age_26_30 <- as.numeric(sqldf("select count (age) from Bank_Data where age between 26 and 30"
))/length(Bank_Data$age)
```

```
apriori_age_31_40 <- as.numeric(sqldf("select count (age) from Bank_Data where age between 31 and 40"
))/length(Bank_Data$age)

apriori_age_41_55 <- as.numeric(sqldf("select count (age) from Bank_Data where age between 41 and 55"
))/length(Bank_Data$age)

apriori_age_above_56 <- as.numeric(sqldf("select count (age) from Bank_Data where age >=56"
))/length(Bank_Data$age)



barplot(prop.table(table(Bank_Data$job)))

apriori_job_admin <- length(which(Bank_Data$job=="admin."))/(length(Bank_Data$job)-
length(which(Bank_Data$job=="unknown")))

apriori_job_blue_collar <- length(which(Bank_Data$job=="blue-collar"))/(length(Bank_Data$job)-
length(which(Bank_Data$job=="unknown")))

apriori_job_management <- length(which(Bank_Data$job=="management"))/(length(Bank_Data$job)-
length(which(Bank_Data$job=="unknown")))

apriori_job_technician <- length(which(Bank_Data$job=="technician"))/(length(Bank_Data$job)-
length(which(Bank_Data$job=="unknown")))

apriori_job_unknown <- length(which(Bank_Data$job=="unknown"))/(length(Bank_Data$job))



barplot(prop.table(table(Bank_Data$marital)))

apriori_marital_divorced <- length(which(Bank_Data$marital=="divorced"))/(length(Bank_Data$marital))

apriori_marital_married <- length(which(Bank_Data$marital=="married"))/(length(Bank_Data$marital))

apriori_marital_single <- length(which(Bank_Data$marital=="single"))/(length(Bank_Data$marital))



barplot(prop.table(table(Bank_Data$education)))

apriori_education_primary <- length(which(Bank_Data$education=="primary"))/(length(Bank_Data$education)-
length(which(Bank_Data$education=="unknown")))

apriori_education_secondary <- length(which(Bank_Data$education=="secondary"))/(length(Bank_Data$education)-
length(which(Bank_Data$education=="unknown")))

apriori_education_tertiary <- length(which(Bank_Data$education=="tertiary"))/(length(Bank_Data$education)-
length(which(Bank_Data$education=="unknown")))

apriori_education_unknown <- length(which(Bank_Data$education=="unknown"))/(length(Bank_Data$education))



barplot(prop.table(table(Bank_Data$default)))

apriori_default_yes <- length(which(Bank_Data$default=="yes"))/(length(Bank_Data$default))
```

```
apriori_default_no <- length(which(Bank_Data$default=="no"))/(length(Bank_Data$default))


hist(Bank_Data$balance, freq = FALSE)

balance_new <- sqldf("select balance from Bank_Data where balance<15000")

hist(balance_new$balance,breaks = c(-2000,-1500,-1000,-
500,0,500,1000,1500,2000,2500,3000,4000,5000,6000,15000))

apriori_balance_under_zero <- as.numeric(sqldf("select count (balance) from Bank_Data where balance <=0"
))/length(Bank_Data$balance)

apriori_balance_0_500 <- as.numeric(sqldf("select count (balance) from Bank_Data where balance between 1 and
500" ))/length(Bank_Data$balance)

apriori_balance_501_1000 <- as.numeric(sqldf("select count (balance) from Bank_Data where balance between 501
and 1000" ))/length(Bank_Data$balance)

apriori_balance_1001_2000 <- as.numeric(sqldf("select count (balance) from Bank_Data where balance between
1001 and 2000" ))/length(Bank_Data$balance)

apriori_balance_2001_4000 <- as.numeric(sqldf("select count (balance) from Bank_Data where balance between
2001 and 4000" ))/length(Bank_Data$balance)

apriori_balance_above_4001 <- as.numeric(sqldf("select count (balance) from Bank_Data where balance >=4001"
))/length(Bank_Data$balance)



barplot(prop.table(table(Bank_Data$housing)))

job_apriori_housing_yes <- length(which(Bank_Data$housing=="yes"))/(length(Bank_Data$housing))

job_apriori_housing_no <- length(which(Bank_Data$housing=="no"))/(length(Bank_Data$housing))



barplot(prop.table(table(Bank_Data$loan)))

job_apriori_loan_yes <- length(which(Bank_Data$loan=="yes"))/(length(Bank_Data$loan))

job_apriori_loan_no <- length(which(Bank_Data$loan=="no"))/(length(Bank_Data$loan))



barplot(prop.table(table(Bank_Data$contact)))

job_apriori_contact_cellular <- length(which(Bank_Data$contact=="cellular"))/(length(Bank_Data$contact)-
length(which(Bank_Data$contact=="unknown")))

job_apriori_contact_telephone <- length(which(Bank_Data$contact=="telephone"))/(length(Bank_Data$contact)-
length(which(Bank_Data$contact=="unknown")))

job_apriori_contact_unknown <- length(which(Bank_Data$contact=="unknown"))/(length(Bank_Data$contact))
```

```
barplot(prop.table(table(Bank_Data$day)))

apriori_day_start <- as.numeric(sqldf("select count (day) from Bank_Data where day <=10"
))/length(Bank_Data$day)

apriori_day_middle <- as.numeric(sqldf("select count (day) from Bank_Data where day between 11 and 20"
))/length(Bank_Data$day)

apriori_day_end <-  as.numeric(sqldf("select count (day) from Bank_Data where day >=21"
))/length(Bank_Data$day)



barplot(prop.table(table(Bank_Data$month))[c("jan","feb","mar","apr","may","jun","jul","aug","sep","oct","nov","d
ec")])

apriori_month_1 <- length(which(Bank_Data$month=="jan"))/(length(Bank_Data$month))

apriori_month_2 <- length(which(Bank_Data$month=="feb"))/(length(Bank_Data$month))

apriori_month_3 <- length(which(Bank_Data$month=="mar"))/(length(Bank_Data$month))

apriori_month_4 <- length(which(Bank_Data$month=="apr"))/(length(Bank_Data$month))

apriori_month_5 <- length(which(Bank_Data$month=="may"))/(length(Bank_Data$month))

apriori_month_6 <- length(which(Bank_Data$month=="jun"))/(length(Bank_Data$month))

apriori_month_7 <- length(which(Bank_Data$month=="jul"))/(length(Bank_Data$month))

apriori_month_8 <- length(which(Bank_Data$month=="aug"))/(length(Bank_Data$month))

apriori_month_9 <- length(which(Bank_Data$month=="sep"))/(length(Bank_Data$month))

apriori_month_10 <- length(which(Bank_Data$month=="oct"))/(length(Bank_Data$month))

apriori_month_11 <- length(which(Bank_Data$month=="nov"))/(length(Bank_Data$month))

apriori_month_12 <- length(which(Bank_Data$month=="dec"))/(length(Bank_Data$month))



barplot(prop.table(table(Bank_Data$campaign)))

apriori_campaign_1_2 <- as.numeric(sqldf("select count (campaign) from Bank_Data where campaign <=2"
))/length(Bank_Data$campaign)

apriori_campaign_3_5 <- as.numeric(sqldf("select count (campaign) from Bank_Data where campaign between 3
and 5" ))/length(Bank_Data$campaign)

apriori_campaign_6_8 <- as.numeric(sqldf("select count (campaign) from Bank_Data where campaign between 6
and 8" ))/length(Bank_Data$campaign)

apriori_campaign_above_9 <- as.numeric(sqldf("select count (campaign) from Bank_Data where campaign >=9"
))/length(Bank_Data$campaign)



hist(Bank_Data$pdays, freq = FALSE)
```

```
apriori_pdays_minus_one <- as.numeric(sqldf("select count (pdays) from Bank_Data where pdays <=-1" ))/length(Bank_Data$pdays)

apriori_pdays_0_200 <- as.numeric(sqldf("select count (pdays) from Bank_Data where pdays between 0 and 200" ))/length(Bank_Data$pdays)

apriori_pdays_201_400 <- as.numeric(sqldf("select count (pdays) from Bank_Data where pdays between 201 and 400" ))/length(Bank_Data$pdays)

apriori_pdays_above_401 <- as.numeric(sqldf("select count (pdays) from Bank_Data where pdays >=401" ))/length(Bank_Data$pdays)
```

```
barplot(prop.table(table(Bank_Data$previous)))

apriori_previous_0 <- as.numeric(sqldf("select count (previous) from Bank_Data where previous <=0" ))/length(Bank_Data$previous)

apriori_previous_1_2 <- as.numeric(sqldf("select count (previous) from Bank_Data where previous between 1 and 2" ))/length(Bank_Data$previous)

apriori_previous_3_5 <- as.numeric(sqldf("select count (previous) from Bank_Data where previous between 3 and 5" ))/length(Bank_Data$previous)

apriori_previous_above_6 <- as.numeric(sqldf("select count (previous) from Bank_Data where previous >=6" ))/length(Bank_Data$previous)
```

```
barplot(prop.table(table(Bank_Data$poutcome)))

apriori_poutcome_failure <- length(which(Bank_Data$poutcome=="failure"))/(length(Bank_Data$poutcome)-length(which(Bank_Data$poutcome=="unknown")))

apriori_poutcome_other <- length(which(Bank_Data$poutcome=="other"))/(length(Bank_Data$poutcome)-length(which(Bank_Data$poutcome=="unknown")))

apriori_poutcome_success <- length(which(Bank_Data$poutcome=="success"))/(length(Bank_Data$poutcome)-length(which(Bank_Data$poutcome=="unknown")))

apriori_poutcome_unknown <- length(which(Bank_Data$poutcome=="unknown"))/(length(Bank_Data$poutcome))
```

```
barplot(prop.table(table(Bank_Data$Gender)))

apriori_gender_male <- length(which(Bank_Data$Gender=="0"))/(length(Bank_Data$Gender)-length(which(Bank_Data$Gender=="Unknown")))

apriori_gender_female <- length(which(Bank_Data$Gender=="1"))/(length(Bank_Data$Gender)-length(which(Bank_Data$Gender=="Unknown")))

apriori_gender_unknown <- length(which(Bank_Data$Gender=="Unknown"))/(length(Bank_Data$Gender))
```

נספח 14 – קשרים בין משתנים

```
#יצירת מוסבר עם "כן" ו"לא" במקום 0 ו-1

Bank_Data_yes_no <- Bank_Data

"Bank_Data_yes_no$y[Bank_Data_yes_no$y == "1"] <- "yes

"Bank_Data_yes_no$y[Bank_Data_yes_no$y == "0"] <- "no

Bank_Data_yes_no$y <- as.factor(Bank_Data_yes_no$y)


# קשר בין השכלה למשכנתא

education_no_unknown <- sqldf("select education from Bank_Data where education !='unknown'")

housung_no_education_unknown <- sqldf("select housing from Bank_Data where education !='unknown'")

plot(education_no_unknown$education,housung_no_education_unknown$housing,xlab="Education",ylab="Housing")


# קשר בין עבודה ללקיחת הלואה

job_no_unknown <- sqldf("select job from Bank_Data where job !='unknown'")

loan_no_job_unknown <- sqldf("select loan from Bank_Data where job !='unknown'")

plot(job_no_unknown$job,loan_no_job_unknown$loan, xlab="Job", ylab="Loan")


# קשר בין השכלה ליתרת חשבון

balance_no_education_unknown <- sqldf("select balance from Bank_Data where education !='unknown'")

boxplot(balance_no_education_unknown$balance~education_no_unknown$education,ylab="Balance")


# היטמאפ בין משתנים מספריים והמוסבר

Numeric_Bank_Data <-
cbind(Bank_Data$age,Bank_Data$balance,Bank_Data$day,Bank_Data$campaign,Bank_Data$pdays,
Bank_Data$previous,Bank_Data$y)

colnames(Numeric_Bank_Data) <- c("age","balance","day","campaign","pdays","previous","y")

cormat <- round(cor(Numeric_Bank_Data),2)

melted_cormat <- melt(cormat)

print(cormat)

()ggplot(data = melted_cormat, aes(x=Var1, y=Var2, fill=value)) + geom_tile


#קשר בין השכלה למוסבר

y_no_education_unknown <- sqldf("select y from Bank_Data_yes_no where education !='unknown'")
```

```
plot(education_no_unknown$education, y_no_education_unknown$y,xlab="Education", ylab="Y" )
chisq.test(Bank_Data$education,Bank_Data$y)

#קשר בין לקיחת הלוואה למוסבר
plot(Bank_Data$loan,Bank_Data_yes_no$y, xlab="Loan",ylab="Y")
chisq.test(Bank_Data$loan,Bank_Data_yes_no$y)

#קשר בין משכנתא למוסבר
plot(Bank_Data$housing,Bank_Data_yes_no$y, xlab="Housing",ylab="Y")
chisq.test(Bank_Data$housing,Bank_Data_yes_no$y)

#קשר בין אופן יצירת קשר למוסבר
contact_no_unknown <- sqldf("select contact from Bank_Data where contact !='unknown'")
y_no_contact_unknown <- sqldf("select y from Bank_Data_yes_no where contact !='unknown'")
plot(contact_no_unknown$contact,y_no_contact_unknown$y, xlab="Contact",ylab="Y")

#קשר בין גיל למוסבר
"Bank_Data_yes_no$age[Bank_Data_yes_no$age>=56] <- "E
"Bank_Data_yes_no$age[Bank_Data_yes_no$age<=25] <- "A
"Bank_Data_yes_no$age[Bank_Data_yes_no$age<=30] <- "B
"Bank_Data_yes_no$age[Bank_Data_yes_no$age<=40] <- "C
"Bank_Data_yes_no$age[Bank_Data_yes_no$age<=55] <- "D
Bank_Data_yes_no$age <- as.factor(Bank_Data_yes_no$age)
plot(Bank_Data_yes_no$age,Bank_Data_yes_no$y, xlab="Age",ylab="Y")
# כמשתנה רציף
ageY <- sqldf("select age, count(age) as count, sum(y) as sumY, y from Bank_Data group by age")
newAgeY <- data.frame(groupedAge=ageY$age, prob=(ageY$sumY/ageY$count))
plot(x=newAgeY$groupedAge,y=newAgeY$prob, xlab = "Age", ylab= "Probability to accept")
lines(lowess(x=newAgeY$groupedAge,y=newAgeY$prob),lwd=3,col="purple")

#קשר בין עבודה למוסבר
y_no_job_unknown <- sqldf("select y from Bank_Data_yes_no where job !='unknown'")
plot(job_no_unknown$job,y_no_job_unknown$y,xlab="Job", ylab="Y")
```

#קשר בין מצב משפחתי למוסבר

```
plot(Bank_Data$marital,Bank_Data_yes_no$y, xlab="Marital", ylab="Y")
```

#קשר בין יתרת חשבון למוסבר

```
Bank_Data_yes_no$grouped_balance <- findInterval(Bank_Data_yes_no$balance,c(-2000,0,500,1000,2000,4000,40000))

Bank_Data_yes_no$grouped_balance <- as.factor(Bank_Data_yes_no$grouped_balance)

plot(Bank_Data_yes_no$grouped_balance,Bank_Data_yes_no$y, xlab="Grouped Balance",ylab="Y")
```

#קשר בין יום למוסבר

```
Bank_Data_yes_no$grouped_day <- findInterval(Bank_Data_yes_no$day,c(0,11,21,32))

Bank_Data_yes_no$grouped_day <- as.factor(Bank_Data_yes_no$grouped_day)

plot(Bank_Data_yes_no$grouped_day,Bank_Data_yes_no$y, xlab="Grouped Day",ylab="Y")
```

#קשר בין חודש למוסבר

```
Bank_Data_yes_no$month_num[Bank_Data_yes_no$month=="jan"] <- 1

Bank_Data_yes_no$month_num[Bank_Data_yes_no$month=="feb"] <- 2

Bank_Data_yes_no$month_num[Bank_Data_yes_no$month=="mar"] <- 3

Bank_Data_yes_no$month_num[Bank_Data_yes_no$month=="apr"] <- 4

Bank_Data_yes_no$month_num[Bank_Data_yes_no$month=="may"] <- 5

Bank_Data_yes_no$month_num[Bank_Data_yes_no$month=="jun"] <- 6

Bank_Data_yes_no$month_num[Bank_Data_yes_no$month=="jul"] <- 7

Bank_Data_yes_no$month_num[Bank_Data_yes_no$month=="aug"] <- 8

Bank_Data_yes_no$month_num[Bank_Data_yes_no$month=="sep"] <- 9

Bank_Data_yes_no$month_num[Bank_Data_yes_no$month=="oct"] <- 10

Bank_Data_yes_no$month_num[Bank_Data_yes_no$month=="nov"] <- 11

Bank_Data_yes_no$month_num[Bank_Data_yes_no$month=="dec"] <- 12

Bank_Data_yes_no$month_num <- as.factor(Bank_Data_yes_no$month_num)

plot(Bank_Data_yes_no$month_num,Bank_Data_yes_no$y,xlab = "Month", ylab = "Y")

Bank_Data_yes_no$month_num <- findInterval(Bank_Data_yes_no$month_num,c(0,4,7,10,13))

Bank_Data_yes_no$month_num <- as.factor(Bank_Data_yes_no$month_num)

plot(Bank_Data_yes_no$month_num,Bank_Data_yes_no$y,xlab="Quarters",ylab="Y")
```

#קשר בין קמפיין למוסבר

#כמשתנה רציף

```
campaignY <- sqldf("select campaign, count(campaign) as count, sum(y) as sumY, y from Bank_Data group by campaign")

newcampaignY <- data.frame(groupedcampaign=campaignY$campaign, prob=(campaignY$sumY/campaignY$count))

plot(x=newcampaignY$groupedcampaign,y=newcampaignY$prob, xlab = "campaign", ylab= "Probability to accept")

lines(lowess(x=newcampaignY$groupedcampaign,y=newcampaignY$prob),lwd=3,col="purple")
```

#קשר בין מספר הימים שעברו מהפעם האחרונה שנוצר קשר למוסבר

#כמשתנה רציף

```
pdaysY <- sqldf("select pdays, count(pdays) as count, sum(y) as sumY, y from Bank_Data group by pdays")

newpdaysY <- data.frame(groupedpdays=pdaysY$pdays, prob=(pdaysY$sumY/pdaysY$count))

plot(x=newpdaysY$groupedpdays,y=newpdaysY$prob, xlab = "pdays", ylab= "Probability to accept")

lines(lowess(x=newpdaysY$groupedpdays,y=newpdaysY$prob),lwd=3,col="purple")
```

#קשר בין מספר הפעמים שיצרו קשר עם הלקוח לפני הקמפיין הנוכחי

#כמשתנה רציף

```
previousY <- sqldf("select previous, count(previous) as count, sum(y) as sumY, y from Bank_Data group by previous")

newpreviousY <- data.frame(groupedprevious=previousY$previous, prob=(previousY$sumY/previousY$count))

plot(x=newpreviousY$groupedprevious,y=newpreviousY$prob, xlab = "previous", ylab= "Probability to accept")

lines(lowess(x=newpreviousY$groupedprevious,y=newpreviousY$prob),lwd=3,col="purple")
```

#קשר בין תוצאות הקמפיין הקודם למוסבר

```
poutcome_no_unknown <- sqldf("select poutcome from Bank_Data where poutcome !='unknown'")

y_no_poutcome_unknown <- sqldf("select y from Bank_Data_yes_no where poutcome !='unknown'")

plot(poutcome_no_unknown$poutcome,y_no_poutcome_unknown$y, xlab="poutcome",ylab="Y")
```

#קשר בין מין למוסבר

```
gender_no_unknown <- sqldf("select gender from Bank_Data where gender !='Unknown'")

y_no_gender_unknown <- sqldf("select y from Bank_Data_yes_no where gender !='Unknown'")

plot(gender_no_unknown$Gender,y_no_gender_unknown$y, xlab="Gender", ylab="Y")
```

```
Bank_Data_yes_no$Gender <- as.factor(Bank_Data_yes_no$Gender)
```

```
#סיכום הנתונים
print(summary(Bank_Data))
```

```
#דיסקרטיזציה של קמפיין והקשר עם המוסבר
Bank_Data_yes_no$grouped_campaign <- findInterval(Bank_Data_yes_no$campaign,c(0,3,5,7,60))
Bank_Data_yes_no$grouped_campaign <- as.factor(Bank_Data_yes_no$grouped_campaign)
plot(Bank_Data_yes_no$grouped_campaign, Bank_Data_yes_no$y, xlab="Grouped Campaign", ylab="Y")
```

```
#דיסקרטיזציה של פריבויס והקשר עם המוסבר
Bank_Data_yes_no$grouped_previous <- findInterval(Bank_Data_yes_no$previous,c(0,1,6,40))
Bank_Data_yes_no$grouped_previous <- as.factor(Bank_Data_yes_no$grouped_previous)
plot(Bank_Data_yes_no$grouped_previous, Bank_Data_yes_no$y, xlab="Grouped previous", ylab="Y")
```

נספח 15 – הכנת נתונים

```
#הכנת הנתונים לקראת השלמת החסרים
drops <- c("default","contact","Gender")
Data_After_Cleaning_Features <- Bank_Data[ , !(names(Bank_Data) %in% drops)]
```

```
Data_After_Cleaning_Features[Data_After_Cleaning_Features=="unknown"] <- NA


#ביצוע דיסקריטיזציה לנתונים הרלוונטיים

Discrete_Data_Without_Missing_Values <- Data_Without_Missing_Values

Discrete_Data_Without_Missing_Values$balance <- findInterval(Discrete_Data_Without_Missing_Values$balance,c(-2000,0,500,1000,2000,4000,40000))

Discrete_Data_Without_Missing_Values$age <- findInterval(Discrete_Data_Without_Missing_Values$age,c(0,25,31,41,56,100))

Discrete_Data_Without_Missing_Values$campaign <- findInterval(Discrete_Data_Without_Missing_Values$campaign,c(0,3,5,7,60))
```

נספח 16 – אלגוריתם mice

```
temp_Data_Without_Missing_Values <- mice(data=Data_After_Cleaning_Features, m=5, method="pmm", maxit=50, seed=50)

Data_Without_Missing_Values <- complete(temp_Data_Without_Missing_Values)
```