לימוד מכונה 364-1-1811

פרויקט לימוד מכונה

חלק א

מרצה הקורס: פרופ' בעז לרנר

בודק תרגילים ואחראי מעבדות: רועי וולף

מגיש: איתן האלי 314090838

בסיס נתונים: Paris Housing Data

תאריך הגשה: 08/12/2022

תוכן עניינים

חלק א'

1.הגדרת הבעיה

1.1 תיאור כללי של עולם התוכן הנחקר

נדל"ן בפריז הוא שוק הצומח במהירות רבה בפרט נדל"ן יוקרה אשר צומח מהר יותר מכל ענף אחר בנדל"ן של צרפת. לצרפתים יש טעם לדברים היותר טובים בחיים וכתוצאה מכך שוק הנדל"ן בדגש על בתי יוקרה הפך להיות מספר שתיים בעולם בגודלו. בנוסף צרפתים הם אנשים ציוניים, לכן אנשי צרפת יעדיפו לקנות בית בפריז מאשר בערי בירה של מדינות אחרות שעשויות להיות זולות יותר עם החזר כספי מהיר יותר ביחס להחזר בפריז.

הנדל"ן בפריז במגמת גדילה בשנים האחרונות ולא קיים אינדיקציה שזה ישתנה בשנים הקרובות, לכן הרבה משקיעים חדשים קונים בתים בפריז ואף מוכנים להשקיע הרבה מאוד כסף.

במחקר שנעשה בנושא הנדל"ן של פריז נמצא כי קיים עליה של כ-7% בכל רבעון בין 2021 לבין 2022 ועליה של 17.5% על בתים קדומים בשנים אלו.

המטרה לבחון את המדדים הרלוונטיים כדי לסווג את סוג הבית בפריז (יוקרה/ רגיל).

1.2 הגדרת שאלת המחקר

במחקר שלנו, אנו מצפים לחזות מתוך הנתונים אשר בידינו את הגורמים המשפיעים המשמעותיים על סיווג סוג הבית בפריז, ומכך לבנות מודל שמסוגל לסווג את סוג הבית בפריז בצורה יעילה ומהירה ביותר. מכך נוכל לתת סיווג נכון ככל האפשר לכל קונה/מוכר בית בפריז. שאלת המחקר עוסקת בכך, האם ניתן על פי המדדים שקיבלנו לסווג את סוג הבית בעזרת כל התכונות הנתונות עבור לקוח המגיע למערכת.

2. הבנת הנתונים

2.1 תיעוד מקורות הנתונים ומשמעותם

מקור הנתונים שלנו בפרויקט הוא Paris Housing Data . את הנתונים בו אספו בעזרת אמצעים שונים :

<u>הנחות</u>:

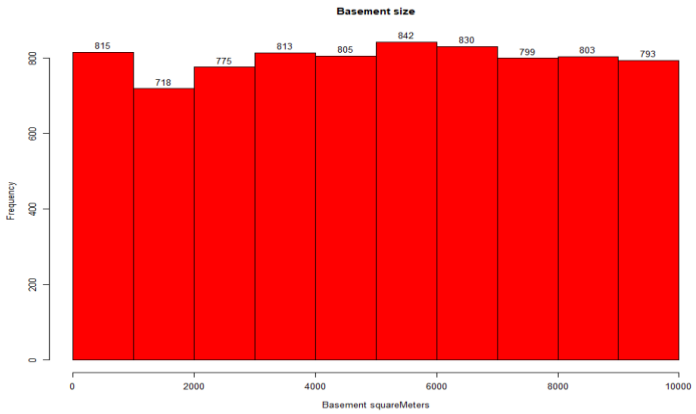
- עלות של דירה ומיקוד חייב לעמוד בתנאי אי-שליליות.
- שנת הבניה של נכנס לא יכולה להיות גדול מ2022.

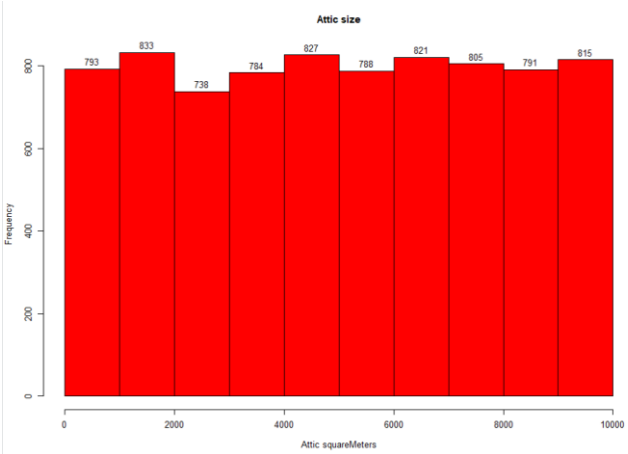
הבנת השתנה	שם המשתנה	יחידת מידה	טווח ערכים
סיווג השתנה -יוקרה/רגיל משתנה מטרה	category	0-basic, 1-luxury	0,1
מחיר הבית	Price	Money	0-10,004,278
גודל הבית במטר מרובע	squareMeters	m^2	89-99985
מספר חדרים בבית	numberOfRooms	Category	1-100
האם קיים גינה?	hasYard	0-no yard, 1-yard	0,1
האם קיים בריכה?	hasPool	0-no pool, 1-pool	0,1
מספר קומת בבית	floors	Category	1-100
מיקוד (סיווג אזור)	cityCode	Int	0-99953
רמת בלעדיות של השכונה	cityPartRange	Category	1-10
מספר בעלי הדירה עד כה	numPrevOwners	Category	1-10
שנת בנייה	made	Year	1900-2022
האם הבניה היא חדשה ?	isNewBuilt	0-old style 1-new style	0,1
האם קיים הגנה מפני סערה ?	hasStormProtector	0-no protector 1-has protector	0,1
גודל המרתף במטר מרובע	basement	m^2	1-10,000
גודל עליית גג במטר מרובע	attic	m^2	1-10,000
גודל המוסך	garage	m^2	100-1,000
קיים מחסן	hasStorageRoom	0-no storage 1-has storage	0,1
מספר חדרי אורחים	hasGuestRoom	Category	1-10

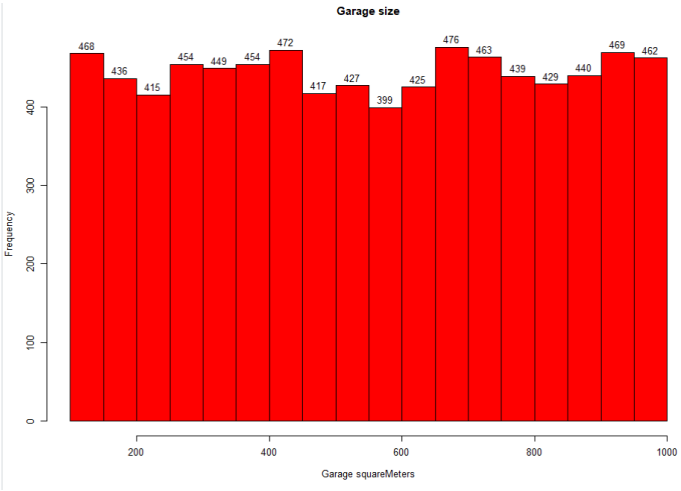
שם המשתנה	איך נמדד	הסבר על המשתנה
category	נתוני מומחה	בחלק זה אני אעשה הסכות לפי המשתנים ונגדיר איזה משתנים רלוונטיים להגדרת קטגורית הבית שהגדרנו כמשתנה המטרה
Price	נתוני מומחה	ככל שמחיר הנכס עולה יש סיכוי גבוה יותר לקבל קטגוריה של יוקרה
squareMeters	ידני	גודל הדירה עשוי להשפיע על הקטגוריה שלה ולכן יש משמעות לגודל הנכס
numberOfRooms	ספירה ידני	בתים עם מספר גדול של חדרים בדר"כ עולים יותר לכן הוא משפיע על המחיר שמשפיע על קטגורית הנכס
hasYard	בדיקה וויזואלית	גינה מגדילה את השטח של הנכס ואת איכות חיי המתגוררים בה
hasPool	בדיקה וויזואלית	פריז היא עיר מרכזית עם אוכלוסייה גדולה ולכן בריכה בנכס זה פריבילגיה
floors	ספירה ידני	מספר רב של קומות מעידה על נכס עם פוטנציאל למספר רב של חדרים
cityCode	נתוני מומחה	מיקוד בצרפת ניתן לפי האזור מגורים שעושי להעיד על אזור טוב או פחות טוב
cityPartRange	נתונים סטטיסטיים	רמת בלעדיות של אזור עשויה להעיד על קטגורית הנכס של האזור (למשל קיסריה)
numPrevOwners	נתוני הבית	מספר הבעלים עוזר לנו להסיק מסקנות חיוביות ושליליות על הנכס
made - year	נתוני מומחים	בצרפת שנת הבניה חשובה מפני שמחירי הנדלן משתנים עם השנים ולכן נוכל להסיק בכמה אחוזים מחיר הדירה עלה/ירד מאז שהוא נבנה
isNewBuilt	נתוני מומחים	בצפרת ובכל אירופה סגנון יצירה של מוצר עשוי להעלות את הביקוש. למשל יין, יין מעולם הישן מוצר בעזרת מנהגים(ללא טכנולוגיה),לעומת יין מהעולם החדש ולכן מקבל מחיר שונה
hasStormProtector	בדיקת מומחה	מצביע על נכס עם ביטחון גבוהה
basement	מדידות אקדח לייזר	מצביע על שטח נוסף לנכס שמאפשר לפתח תחביבים ואפילו עסק כמו שיעורי יוגה או פילטיס
attic	מדידות אקדח לייזר	שטח מעולה לחדר משחקים לילדים שעשוי למשוך משפחות לשכונע
garage	מדידות אקדח לייזר	מצביע על שטח נוסף לנכס שמאפשר לפתח תחביבים כמו נגרות וגם חניה לרכב שעשוי לחסוך זמן בחיפוש חניה
hasStorageRoom	בדיקה וויזואלית	מצביע על שטח נוסף לנכס אשר עשוי לעלות את מחירה
hasGuestRoom	ספירה ידני	מספר רב של חדרי אורחים מצביע על אירוח רב של אנשים, בדר"כ פריווילגיה של עשירים

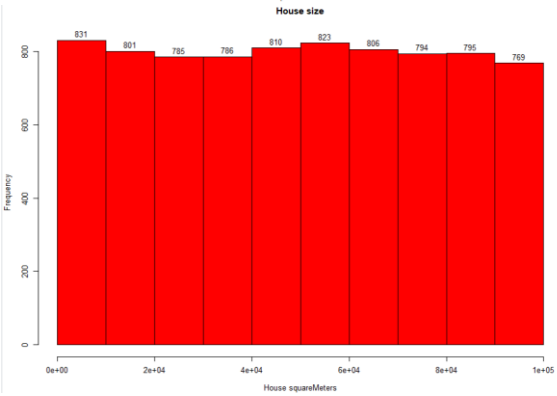
2.2 הסתברויות אפריוריות וקשרים בין מאפיינים

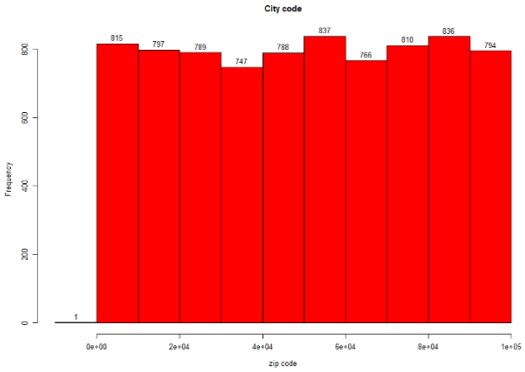
<u>משתנים רציפים - היסטוגרמה:</u>

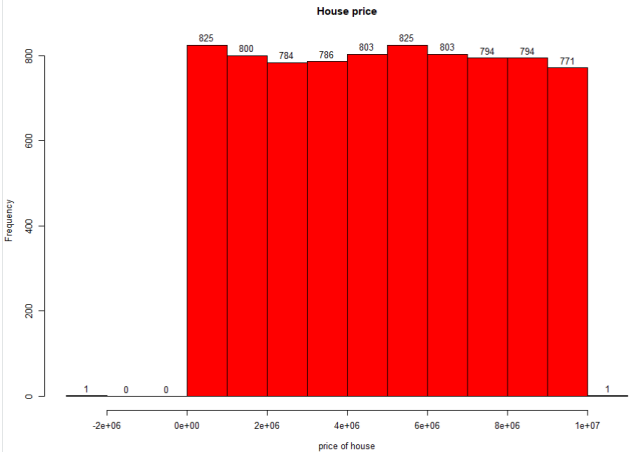












משתנים קטגוריאליים – הסתברויות אפריוריות:

numberOfRooms	
Value	**Probability**
0-20	0.198375
21-40	0.20175
41-60	0.20775
61-80	0.19625
81-100	0.19575
Incorrect values	0.000125

numberOfFloors	
Value	**Probability**
0-20	0.1995
21-40	0.205875
41-60	0.1975
61-80	0.196875
81-100	0.198125
Incorrect values	0.002125

haspool	
Value	**Probability**
0 – no pool	0.4985
1 – has pool	0.499
Incorrect values	0.0025

hasyard	
Value	**Probability**
0 – no yard	0.490125
1 – has yard	0.50775
Incorrect values	0.02125

cityPartRange	
Value	**Probability**
1-2	0.198625
3-4	0.199375
5-6	0.1965
7-8	0.206125
9-10	0.197375
Incorrect values	0.002

numPrevOwenrs	
Value	**Probability**
1-2	0.1935
3-4	0.20225
5-6	0.206375
7-8	0.193375
9-10	0.202125
Incorrect values	0.002375

made	
Value	**Probability**
1900-1925	0.000125
1926-1950	0
1951-1975	0
1976-2000	0.344125
2001-2022	0.652125
Incorrect values	0.003625

hasGuestRoom	
Value	**Probability**
0-2	0.184625
3-4	0.182125
5-6	0.180625
7-8	0.179375
9-10	0.182625
Incorrect values	0.090625

hasStormProtection	
Value	**Probability**
0 – no storm protection	0.5
1 – has storm protection	0.497875
Incorrect values	0.002125

isNewBuilt	
Value	**Probability**
0 – not new built	0.50175
1 – new built	0.495625
Incorrect values	0.002625

hasStorageRoom	
Value	**Probability**
0 – no storage room	0.49825
1 – has storage room	0.50125
Incorrect values	0.0005

category	
Value	**Probability**
Basic	0.863875
Luxury	0.136125
Incorrect values	0

בבחינת ההסתברויות האפריוריות של משתנה המטרה ושאר המשתנים, משתנה המטרה מציג תמונת מצב של 13.61% מהתצפיות בסט הוגדרו כבתי יוקרה, נתון זה לא מתיישב עם הסטטיסטיקה בפריז (לא הצלחתי למצוא כמה אחוז מכלל הבתים בפריז הם בתי יוקרה אך לא נראה לי הגיוני ש-13 מתוך 100 בתים הם בתי יוקרה, לא הצלחתי למצוא נתונים באינטרנט לגבי נושא בתים בפריז שהוא לא נדלן).

בנוסף לא נראה לי הגיוני שלא הייתה בנייה של בתים בפריז בין השנים 1926-1975 ובנייה של 0.0125% מכלל הבתים של פריז בין השנים 1900-1925 (אני מסיק זאת מפני שצרפת נכנעו לגרמניה הנאצית כדי שלא יחריבו את ערי צרפת בזמן פלישתם, בפרט את פריז בשנת 1940 אך עדיין היו הפגזות רבות עד אז ולכן היה צורך בבנייה בפריז. אם זאת, קיים עד היום הרבה בניה מלפני מלחמת העולם השנייה), לכן ניתן להסיק כי התצפיות איינן מייצגות את אוכלוסיית הבתים בפריז.

נתבונן בהיסטוגרמות של המשתנים הרציפים, ניתן לראות שהמשתנים מחולקים בצורה מאוזנת כלומר יש הסתברות שווה להיות בכל טווח אבל ניתן לראות נתונים שלילים ב- city code ו- house price.

כאשר נבחן את ההסתברויות האפריוריות של המשתנים הבדידים, חלקם הוכנסו לחמישה טווחים בכדי לבחון את הנתונים בצורה מופשטת, ניתן לראות שהנתונים מאוזנים כלומר יש הסתברויות כמעט שוות להיות בכל טווח.
בנוסף ניתן לראות בעייתיות בהסתברויות של made כפי שנאמר לעיל, וקיים בעייתיות בעמודה של has guest room מפני שקיים הסתברות של כ-10% להגריל תצפית שחסר נתון במשתנה.

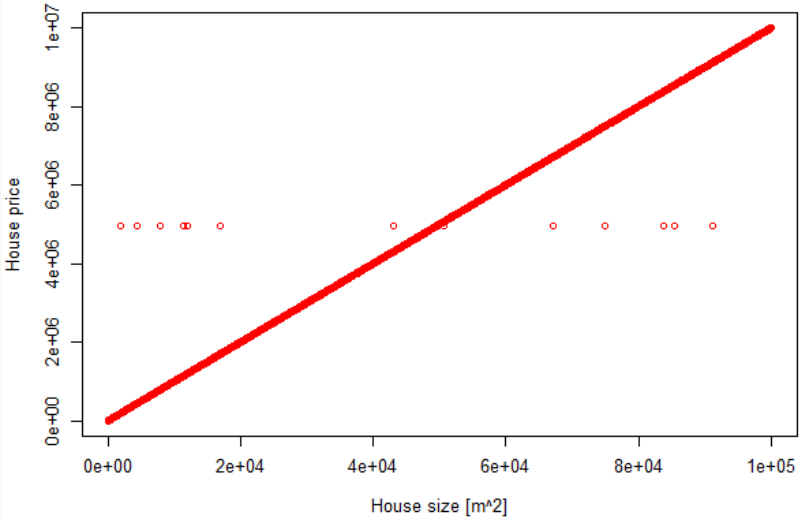
לא נראה לי הגיוני שיש הסתברות כמעט שווה להגריל תצפית עם מספר קומות בין 81-100 לבין תצפית עם מספר קומות בין 0-20, הבעייתיות הזאת חוזרת על עצמה עם כל המשתנים המסבירים למעט made(בו יש בעייתיות אחרת), לכן ניתן לחזק את הטענה שהתצפיות איינן מייצגות את אוכלוסיית הבתים בפריז.
(לא הצלחת למצוא יותר מידי נתונים על הבתים בפריז ולכן אני מביע את דעתי מתוך הגיון וידע שצברתי עם השנים).

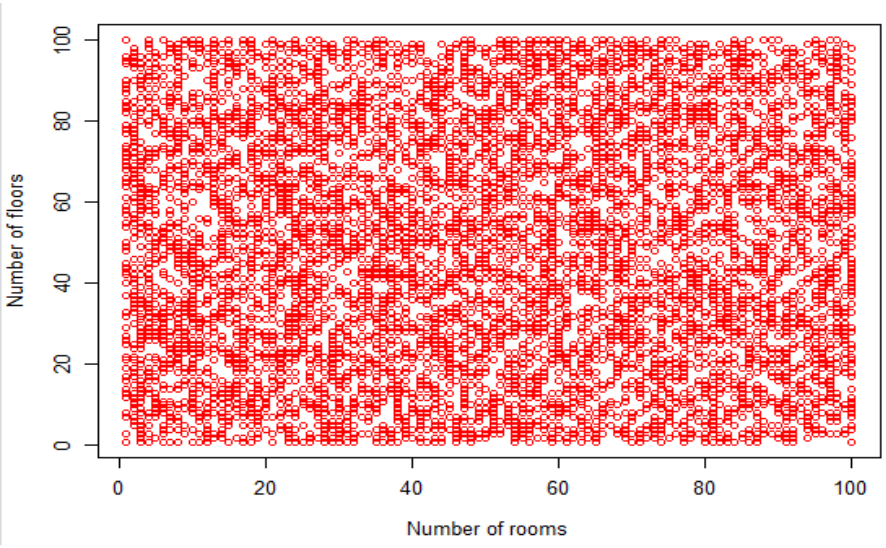
ביצעתי תיקון (לא סופי) כדי לבדוק קשרים מעניינים בין המשתנים, השינויים שנעשו הינם הפיכת שדה שגוי ל-NA במספר תצפיות, לאחר מכן החלפת NA בממוצע של אותו עמודה, אך עדיין רציתי לשנות קצת את הנתונים ולכן בעמודה של isNewBuilt ו-hasStormProtecto בכל מקום שהיה NA הוכנס הערך 0 מפני שאחוז ה-NA היה סביבות ה-0.2%. בנוסף שיניתי את המשתנה המוסבר כך שהוא יהיה בינארי כלומר יקבל ערך 1 על בית יוקרה ואפס על בית רגיל על מנת לבדוק קשר בינו לבין מספר משתנים אחרים.

<u>קשרים בין משתנים</u>:

<u>קשר בין מחיר לגודל הבית</u>:

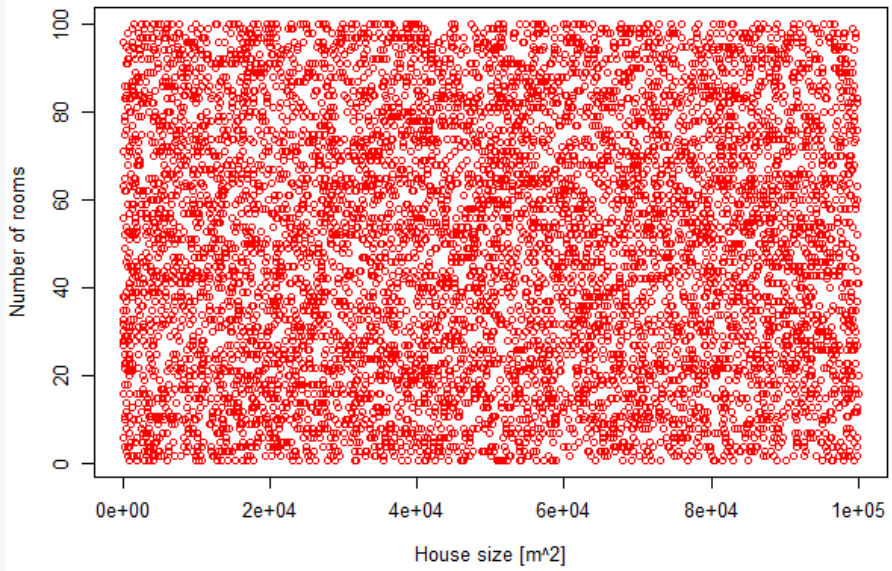
ידוע בשוק הנדלן, ככל שגודל הבית יותר גדול ככה המחיר עולה ולכן אנו נצפה לקבל קשר לינארי בין מחיר הבית לבין גודלו.
כאשר נבחן את הגרף ניתן לראות שקיים קשר לינארי עם שיפוע חיובי בין מחיר לגודל הבית למאט מספר תצפיות חריגות.





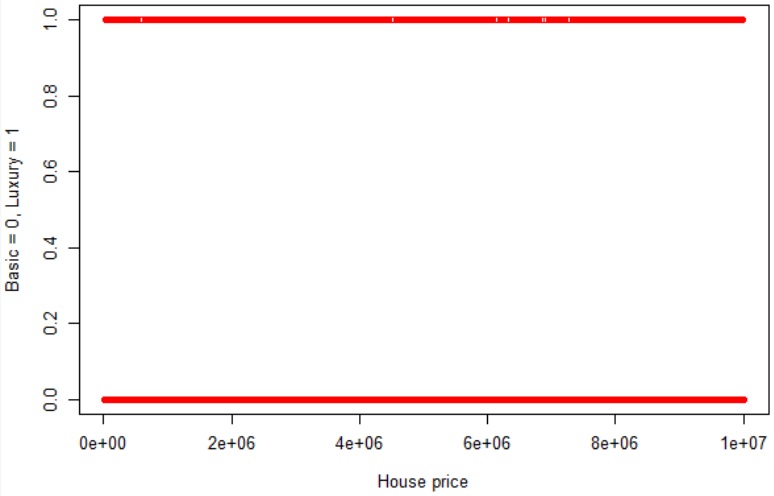
קשר בין מספר חדרים למספר קומות:

להנחתי ככל שמספר החדרים גדל נצפה למספר קומות גדול יותר או שטח גדול יותר.
ניתן לראות שאין פה קשר בין מספר חדרים למספר קומות, ממצא זה הגיוני אבל רק בתנאי שנמצא קשר בין מספר חדרים לגודל הבית.



קשר בין מספר חדרים לגודל הבית :

כאשר בדקנו את הקשר בין מספר חדרים לבין מספר קומות הבחנו בכך שאין קשר ולכן נצפנה לקבל קשר כלשהו בין מספר חדרים לבין גודל הבית.
נתבונן בפיזור, ניתן לראות שאין קשר, ממצע זה לא מתיישב עם ההנחה שלי אך זה מחזק את הטענה שהנתונים אינם מדמים את המציעות וכל הנראה הוגרלו באקראיות.

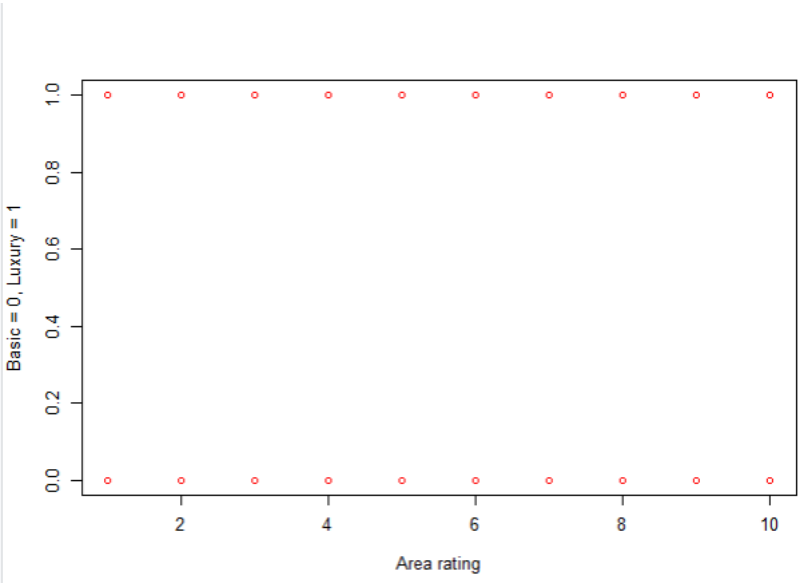


קשר בין מחיר הבית לבין קטגוריה:

לפי שוק הנדלן בתי יוקרה הם בתים מפוארים עם הרבה תוספות שאין לבתים בסיסים ולכן אני מניח, ככל שמחיר הבית עולה ככה הסיכוי שהוא יהיה בית יוקרה.
נתבונן בפיזור, לא ניתן להבחין בקשר כלשהו אפילו שציפינו לראות מדרגה משמאל לימין, בשלב זה אני מתחיל לחשוש שהקשר בין המשתנים הוא לא קשר לינארי אלא קשר הסתברותי.

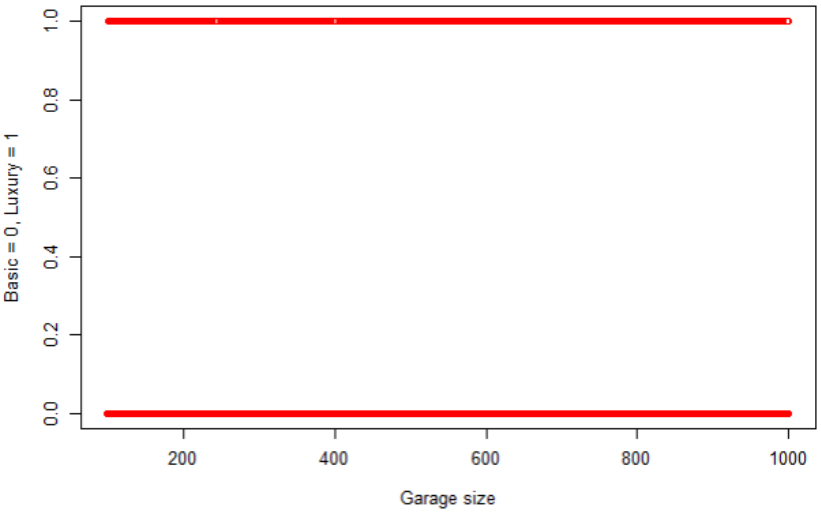
קשר בין דירוג אזורי לבין קטגורית הבית:

לפי הקשר בין מחיר הבית לבין הקטגוריה עלה חשש, כי קיים קשר הסתברותי בין המשתנים לבין הקטגוריה אך עדיין אין מספיק ממצאים כדי להגיד בוודאות. גם בגרף זה לא ניתן להבחין בקשר אפילו שהייתי מצפה לראות מספר גדול יותר של בתי יוקרה בשכונה עם דירוג גבוהה יותר. ככל הנראה אין קשר בין דירוג אזורי לבין קטגורית הבית.



קשר בין גודל מוסך לבין קטגוריה:

מוסך פרטי בערים מרכזים הוא לרוב פריבילגיה, לכן נצפה שבפריז מוסך פרטי הוא מקיים איזשהו קשר לבתי יוקרה בפרט גודל המוסך. נתבונן בממצעים, ניתן לראות שאין מדרגה משמאל לימין שמצביע על כך שאין קשר לינארי בין גודל המוסך לבין קטגורית הבית.



קשר הסתברותי של Category בהינתן HasYard:
P(Luxury|HasYard=1) = 0.2316593, ניתן לראות שקיים קשר הסתברותי יחסית חלש.
P(Luxury|HasYard=0) = 0.0372354, אבל ניתן לראות שיש הסתברות נמוכה מאוד ליפול בקטגוריה של יוקרה אם אין גינה.
P(Basic|HasYard=0) = 0.9627646, ניתן לראות שיש קשר הסתברותי חזק מאוד בין קבלת קטגוריה של בית בסיסי בהינתן שאין גינה.

קשר הסתברותי של Category בהינתן HasPool:
P(Luxury|HasPool=1) = 0.2369609, ניתן לראות שקיים קשר הסתברותי יחסית חלש.
P(Luxury|HasPool=0) = 0.03535607, כאן ניתן לראות שיש הסתברות נמוכה מאוד ליפול בקטגוריה של יוקה בהינתן שאין בריכה.
P(Basic|HasPool=0) = 0.9646439, ופה ניתן לראות שיש הסתברות גבוהה ליפול לקטגורית בית בסיסי בהינתן שאין בריכה.

<u>קשר הסתברותי של Category בהינתן CityPartRange</u>:
אני מחפש קשר כלשהו ולכן אבדוק טווחים בין 1 ל-5 ובין 6 ל-10.
P(Luxury|1 <= CityRange <= 5) = 0.1341677
P(Luxury|6 <= CityRange <= 10) = 0.1383805, לא נראה שיש קשר הסתברותי של קטגוריה בהינתן דירוג אזור אך זה לא מספיק כדי לפסול את השדה.
אבדוק אם המשתנים משפיעים על קטגורית הבית בעזרת רגרסיה לוגיסטי בסעיף 3.

2.3 איכות הנתונים

- קיים מספר יחסית קטן של נתונים חסרים אך בכל זאת עדיף לא לאבד מידע, לכן עבור משתנים רציפים נחשב את הממוצע של העמודה ונזין לנתונים החסרים. עבור משתנים קטגוריאלים נחשב את הממוצע ונעגל קלפי מטה ונזין לנתונים החסרים. עבור המשתנים הבינאריים נחשב ממוצע ,נבדוק האם הוא מעל 0.5, אם כן נזין ערך 1 לנתונים החסרים אחרת 0. הסיבה שאני חושב שזה לא יפגע באיכות הנתונים זה מפני שאחוז הנתונים החסרים הוא קטן מאוד כלומר <%5, יש כ-16 נתונים חסרים בעמודות שאכן חסר כלומר 16/8000. מעבר לכך היה קשה למצוא קשר בין הנתונים אבל אני חושב שהצלחתי להבין את הקשרים (קשר הסתברותי בהינתן ערך המותנה).

<u>**נתונים בעייתיים**</u>:

cityCode – יש שדה שערכו 3-, אני מניח שזו טעות הקלדה אך אינני רוצה לאבד נתונים ולכן אכניס מספר תקין שלא קיים בעמודה.

Made – יש שדה שערכו 2122 שבעייתי מפני שזה שנה שאינו בטווח מחייה שלנו, אני אניח שזה טעות והיה אמור להיות 2022.

isNewBuild – קיים שדות שערכם הוא מחרוזת של "כן" ו-"לא" אבל זה עמודה של ערכים בינארי, אתקן זאת בכך שאכניס 1 איפה שרשום "כן" ו-0 איפה שרשום "לא".

hasStormProtaction – קיים את אותו הבעיה כמו ב-isNewBuild ואפתור בצורה זהה.

Price – קיים ערך שלישי באחד השדות, זה לא הגיוני ולכן אחליף את הערך הזה עם הממוצע של העמודה.

HasGuestRoom – הינו משתנה בעייתי מפני שקיים מספר בתים עם מספר חדרי אירוח שגדול יותר מאשר מספר חדרי שינה בבית, לכן אציע להחליף את השדה הזה לשדה בינארי ככה שאם קיים חדרי אירוח הוא יקבל ערך 1 אחרת 0.

Attic,basement,garge – קיים 561 תצפיות בעייתיות מפני שחלכם גדולים יותר משטח הבית, לא ניתן לדעת אם זה טעות בהזנת גודל הבית או טעות בהזנת המשתנים שצוינו ולכן נמחק אותם. הבעייתיות פה זה ש-561 הוא סביבות ה-7% אחוזים מהנתונים אבל בגלל שאני מניח שהנתונים הוזנו בצורה רנדומלית ולא מראים אינדיקציה לנתונים מהמציאות אוותר עליהם.

3.הכנת הנתונים

3.1 על פי הצורך, בצעו ונמקו בחירת מאפיינים שביצעתם

נבצע תיקון משתנים ולאחר מכן רגרסיה לוגיסטית לפני שנתחיל השמטת מאפיינים.

- קיים 561 תצפיות אשר בעלי רעש מפני שהגודל המרתף, עליית גג ומחסן גדולים משטח הבית, לא ניתן לדעת אם היה טעות בהזנת גודל הבית או טעות הזנת גודל מרתף/עליית גג/מרתף ולכן בהתחלה בחרתי להסיר את התצפיות עם התחשבות בכך ש561 גדול מה-5% מכלל התצפיות. לאחר בדיקה עם רגרסיה לוגיסטית נמצא ששדות אלו לא משפעים עיקרים על קטגורית הבית ולכן אחליט לא למחוק כדי לא לאבד נתונים.
- בחרתי לא להשמיט תצפיות עם ערכים חסרים כדי לא לאבד מידע, לכן הזנתי ערכים לפי התיאור ב"נתונים בעייתיים".
- לאחר ביצוע רגרסיה לוגיסטית נמצא כי השדות הרלוונטיות לקביעת קטגורית הבית (השדות שמובהקים כלומר ה-P-val שלהם מראה שהם מובהק) הינם HasYard,HasPool,made,isNewBuild, לכן אותם נשאיר במודל שלנו. מתוך חשש להתאמת יתר אוסיף את השדות שלא מובהקות אך ה-P-val שלהם הכי קטן מתוך התוצאה של הרגרסיה הלוגיסטית, כלומר hasGuestRoom ו-cityCode. (ניתן לראות בנספחים את התוצאות של רגרסיה לוגיסטית)

3.2

- נבצע דיסקרטיזציה על המשתנה הרציף cityCode מפני שקיים הסתברות יחסית שווה ליפול בכל טווח אם נחלק לטווחים שווים(ניתן לראות זאת מתוך ההיסטוגרמה של סעיף 2.2), כך ש:

```
1 - (0<=cityCode<=20,000)
2 - (20,000<cityCode<=40,000)
3 - (40,000<cityCode<=60,000)
4 - (60,000<cityCode<=80,000)
5 - (80,000<cityCode<=100,000)
```

מעבר לכך שינונו את hasGuestRoom למשתנה בינארי כך שאם יש חדרי אירוח נזין 1 אחרת 0, מפני שאנחנו רוצים לפתור את הבעיה של ערכים חרגים (יש יותר חדרי אירוח מאשר חדרי שינה).

ניתן לראות ירידה של AIC לאחר השינויים אלו ולכן ככל הנראה קיבלנו החלטה נכונה.

- בחרתי להוסיף משתנה Error_size_measure אשר מציין אם היה טעות במדדיה של גודל מרתף, עלית גג ומוסך כך ש: אם היה מדידה אחת שגויה יקבל 1, אם היה שני מדידות שגויות יקבל 2 ואם היה שלושה מדידות שגויות יקבל 3, אחרת 0.

<u>ניתן לראות את השינויים של הרגרסיה הלוגיסטית:</u>(כל התצפיות)

1) מודל מלא ללא המשתנה החדש. (AIC = 3909.9)

```
Call:
glm(formula = as.numeric(final_dataTable$category) ~ ., family = binomial,
    data = final_dataTable)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7054  -0.4972  -0.1445  -0.0417   3.8875

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)         1.297e+01  8.903e+00   1.457   0.1451
Unnamed..0          5.166e-05  1.795e-05   2.878   0.0040 **
squareMeters       -9.422e-06  3.233e-05  -0.291   0.7707
numberOfRooms      -6.326e-04  1.452e-03  -0.436   0.6630
hasYard             2.663e+00  1.048e-01  25.410   <2e-16 ***
hasPool             2.699e+00  1.056e-01  25.560   <2e-16 ***
floors             -1.430e-03  1.447e-03  -0.989   0.3229
cityCode            3.776e-02  2.915e-02   1.295   0.1952
cityPartRange       1.485e-03  1.450e-02   0.102   0.9184
numPrevOwners       1.112e-03  1.466e-02   0.076   0.9395
made               -1.038e-02  4.448e-03  -2.335   0.0196 *
isNewBuilt          2.709e+00  1.038e-01  26.095   <2e-16 ***
hasStormProtector   7.823e-02  8.326e-02   0.939   0.3475
basement           -6.128e-07  1.452e-05  -0.042   0.9663
attic               5.813e-06  1.445e-05   0.402   0.6874
garage             -2.737e-06  1.598e-04  -0.017   0.9863
hasStorageRoom      4.719e-02  8.339e-02   0.566   0.5714
hasGuestRoom1       2.525e-01  1.516e-01   1.665   0.0959 .
price               8.688e-08  3.234e-07   0.269   0.7882
Error_size_measure  1.737e-02  1.031e-01   0.168   0.8663
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6365.6  on 7998  degrees of freedom
Residual deviance: 3869.9  on 7979  degrees of freedom
  (1 observation deleted due to missingness)
AIC: 3909.9

Number of Fisher Scoring iterations: 6
```

2) מודל עם פרמטרים מובהקים ללא המשתנה החדש.

(AIC = 3896)

```
Call:
glm(formula = as.numeric(final_dataTable$category) ~ hasYard +
    hasPool + made + isNewBuilt, family = binomial, data = final_dataTable)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5671  -0.5048  -0.1435  -0.0392   3.8647

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) 12.567190   8.879001   1.415   0.1570
hasYard      2.651691   0.104331  25.416   <2e-16 ***
hasPool      2.690832   0.105159  25.588   <2e-16 ***
made        -0.009913   0.004432  -2.237   0.0253 *
isNewBuilt   2.698732   0.103326  26.119   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6365.9  on 7999  degrees of freedom
Residual deviance: 3886.0  on 7995  degrees of freedom
AIC: 3896

Number of Fisher Scoring iterations: 6
```

3) מודל עם פרמטרים מובהקים וגם לא מובהקים אך עם P-val נמוכים ביותר, ללא המשתנה החדש.

(ניתן לראות מודל איכותי יותר לפי AIC=3895.4)

```
Call:
glm(formula = as.numeric(final_dataTable$category) ~ hasYard +
    hasPool + made + isNewBuilt + hasGuestRoom + cityCode, family = binomial,
    data = final_dataTable)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6116  -0.5064  -0.1440  -0.0401   3.9139

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)   12.470685   8.882866   1.404   0.1603
hasYard        2.653142   0.104403  25.413   <2e-16 ***
hasPool        2.689633   0.105213  25.564   <2e-16 ***
made          -0.010039   0.004434  -2.264   0.0236 *
isNewBuilt     2.703627   0.103474  26.129   <2e-16 ***
hasGuestRoom1  0.253718   0.150794   1.683   0.0925 .
cityCode       0.037211   0.029130   1.277   0.2015
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6365.9  on 7999  degrees of freedom
Residual deviance: 3881.4  on 7993  degrees of freedom
AIC: 3895.4

Number of Fisher Scoring iterations: 6
```

4) מודל כמו ב-3 אך עם הוספה של הפרמטר החדש.

ניתן לראות עליה של AIC לכן לא משפר את המודל שלנו. נבחר לא להכניס את המשתנה החדש מפני שהמודל פחות טוב לאחר ההכנסה.

(AIC = 3897.2)

```
Call:
glm(formula = as.numeric(final_dataTable$category) ~ hasYard +
    hasPool + made + isNewBuilt + hasGuestRoom + cityCode + Error_size_measure,
    family = binomial, data = final_dataTable)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6100  -0.5062  -0.1441  -0.0402   3.9154

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)         12.504939   8.883777   1.408   0.1592
hasYard              2.654237   0.104452  25.411   <2e-16 ***
hasPool              2.690391   0.105242  25.564   <2e-16 ***
made                -0.010059   0.004434  -2.268   0.0233 *
isNewBuilt           2.703816   0.103479  26.129   <2e-16 ***
hasGuestRoom1        0.253663   0.150790   1.682   0.0925 .
cityCode             0.037318   0.029130   1.281   0.2002
Error_size_measure   0.038571   0.093750   0.411   0.6808
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6365.9  on 7999  degrees of freedom
Residual deviance: 3881.2  on 7992  degrees of freedom
AIC: 3897.2

Number of Fisher Scoring iterations: 6
```

נספחים

3.1: מודל עם החסרה של 561 תצפיות.

```
Call:
glm(formula = as.numeric(clean_DataTable1$category) ~ ., family = binomial,
    data = clean_DataTable1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6434  -0.5068  -0.1487  -0.0429   3.8798

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)        1.237e+01  9.136e+00   1.354   0.1756
squareMeters       8.161e-06  3.941e-05   0.207   0.8359
numberOfRooms     -1.208e-03  1.494e-03  -0.808   0.4188
hasYard            2.594e+00  1.071e-01  24.219   <2e-16 ***
hasPool            2.659e+00  1.085e-01  24.516   <2e-16 ***
floors            -1.635e-03  1.483e-03  -1.103   0.2702
cityCode           1.482e-06  1.470e-06   1.009   0.3131
cityPartRange     -8.819e-04  1.488e-02  -0.059   0.9528
numPrevOwners      3.354e-03  1.505e-02   0.223   0.8237
made              -9.838e-03  4.562e-03  -2.156   0.0311 *
isNewBuilt         2.648e+00  1.061e-01  24.953   <2e-16 ***
hasStormProtector  6.064e-02  8.555e-02   0.709   0.4784
basement          -2.692e-06  1.492e-05  -0.180   0.8568
attic              2.845e-06  1.486e-05   0.191   0.8482
garage            -2.466e-05  1.649e-04  -0.150   0.8811
hasStorageRoom     6.823e-02  8.567e-02   0.796   0.4258
hasGuestRoom       2.391e-01  1.553e-01   1.540   0.1236
price             -9.341e-08  3.944e-07  -0.237   0.8128
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5935.6  on 7438  degrees of freedom
Residual deviance: 3654.8  on 7421  degrees of freedom
AIC: 3690.8
```

רגרסיה לוגיסטית מודל שלם,
AIC = 3690.8

```
Call:
glm(formula = as.numeric(clean_DataTable1$category) ~ hasYard +
    hasPool + made + isNewBuilt, family = binomial, data = clean_DataTable1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5530  -0.5091  -0.1480  -0.0414   3.8350

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) 12.800527   9.111311   1.405   0.1600
hasYard      2.590625   0.106875  24.240   <2e-16 ***
hasPool      2.658369   0.108239  24.560   <2e-16 ***
made        -0.009972   0.004548  -2.193   0.0283 *
isNewBuilt   2.644915   0.105820  24.995   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5935.6  on 7438  degrees of freedom
Residual deviance: 3662.0  on 7434  degrees of freedom
AIC: 3672

Number of Fisher Scoring iterations: 6
```

רגרסיה לוגיסטית מודל עם שדות מובהקים עם חשד להתאמת יתר.
AIC = 3672

```
Call:
glm(formula = as.numeric(clean_DataTable1$category) ~ hasYard +
    hasPool + made + isNewBuilt + hasGuestRoom + cityCode, family = binomial,
    data = clean_DataTable1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5956  -0.5099  -0.1486  -0.0421   3.8798

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.259e+01  9.115e+00   1.382   0.1671
hasYard       2.591e+00  1.069e-01  24.229   <2e-16 ***
hasPool       2.657e+00  1.083e-01  24.539   <2e-16 ***
made         -1.002e-02  4.549e-03  -2.202   0.0277 *
isNewBuilt    2.648e+00  1.059e-01  25.001   <2e-16 ***
hasGuestRoom  2.360e-01  1.550e-01   1.523   0.1279
cityCode      1.496e-06  1.467e-06   1.020   0.3078
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5935.6  on 7438  degrees of freedom
Residual deviance: 3658.5  on 7432  degrees of freedom
AIC: 3672.5

Number of Fisher Scoring iterations: 6
```

רגרסיה לוגיסטית על מודל מותאם עם הוספת משתנים שאינם מובהקים כדי לא לקבל התאמת יתר.
AIC = 3672.5

:3.2

```
Call:
glm(formula = as.numeric(clean_DataTable1$category) ~ hasYard +
    hasPool + made + isNewBuilt + hasGuestRoom + cityCode, family = binomial,
    data = clean_DataTable1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5965  -0.5101  -0.1484  -0.0423   3.8798

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)   12.590242   9.115341   1.381   0.1672
hasYard        2.591959   0.106946  24.236   <2e-16 ***
hasPool        2.657278   0.108295  24.537   <2e-16 ***
made          -0.010033   0.004549  -2.205   0.0274 *
isNewBuilt     2.649401   0.105962  25.003   <2e-16 ***
hasGuestRoom1  0.235262   0.154967   1.518   0.1290
cityCode       0.037576   0.029969   1.254   0.2099
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5935.6  on 7438  degrees of freedom
Residual deviance: 3658.0  on 7432  degrees of freedom
AIC: 3672

Number of Fisher Scoring iterations: 6
```

רגרסיה לוגיסטית על מודל מותאם עם הוספת משתנים שאינם מובהקים כדי לא לקבל התאמת יתר ודיסקרטיזציה על משתנים.
AIC = 3672

```
Call:
glm(formula = as.numeric(clean_DataTable1$category) ~ hasYard +
    hasPool + made + isNewBuilt + hasGuestRoom + cityCode + price_per_sqrtMeter,
    family = binomial, data = clean_DataTable1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6098  -0.5097  -0.1485  -0.0423   3.8797

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)         12.24638    9.12795   1.342   0.1797
hasYard              2.58957    0.10699  24.204   <2e-16 ***
hasPool              2.65472    0.10834  24.503   <2e-16 ***
made                -0.01006    0.00455  -2.211   0.0271 *
isNewBuilt           2.64929    0.10597  25.000   <2e-16 ***
hasGuestRoom1        0.23414    0.15491   1.511   0.1307
cityCode             0.03804    0.02998   1.269   0.2046
price_per_sqrtMeter  0.19600    0.26903   0.729   0.4663
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5935.6  on 7438  degrees of freedom
Residual deviance: 3657.5  on 7431  degrees of freedom
AIC: 3673.5

Number of Fisher Scoring iterations: 6
```

רגרסיה לוגיסטית על מודל מותאם עם הוספת משתנים שאינם מובהקים וגם הוספת משתנה קטגוריה חדשה של עלות למטר מרובע, המודל הסופי.
AIC = 3673.5