פרויקט לימוד מכונה

חלק א

ירדן עיני 204371082

פלג אליהו 318356995

הגדרת הבעיה

1. תיאור כללי של עולם התוכן הנחקר

הבעיה המחקרית אותה אנו רוצים לבחון היא בעיית סיווג נכסים לנכסי יוקרה נכסים "רגילים".

מחקרים קודמים ניסו למצוא מודל שבעזרתו ניתן יהיה להעריך נכס על פי קשר בין מאפיינים שונים. במחקר של (U B H R A J Y O T & Sahoo, n.d.) יש התייחסות לכך שישנן מספר טכניקות ומודלים שבעזרתם ניתן לחקור את השפעות הסביבה על ערך הנכס. במאמר יש התייחסות לאופן בו ערך הנכס מושפע מפרמטרים שונים, כדוגמת היקף השוק והשפעתו, השפעות הזיהום באזור הנכס המוערך על ידי חיישנים וסקרי דעת קהל. בנוסף, מתקיים דיון בנוגע לחוסר היציבות שעשוי להיווצר במודל עקב קיומם של קשרים בין המשתנים השונים (מספר חדרים, מספר חדרי אמבטיה, גודל הבית וגודל מגרש). במאמר הוצגו מודלים שונים המתארים טכניקות אקונומטריות מרחביות מתחום הכלכלה שעוזרות גם כן להעריך את שווי הנכס. מאמר נוסף (Wyatt, n.d.) מציין את השימוש במערכת מידע גאוגרפית עבור הערכת שווי הנכס. במאמר מתייחסים למושג "value map" שהוא ייצוג קרטוגרפי או מרחבי של נתונים סטטיסטיים אשר עשויים לסייע להעריך את ערך הנכס. בעזרת נתונים אלו ניתן לבצע תכנון, מיסוי, ניהול שווי נכסים ועוד (ביביליוגרפיה

2. הגדרת שאלת המחקר

בפרויקט זה, בעזרת הכלים והשיטות השונות של מערכות לומדות, נרצה לבנות מודל אשר יאפשר לסווג נכסים כנכסי יוקרה ולאפשר לסווג נכס כנכס יוקרה או לא, על סמך נתונים שונים הנאספים אודותיו. שאלת המחקר שלנו היא האם נכס מסוים הוא נכס יוקרה או לא, בהתחשב בפרמטרים שונים.

הבנת הנתונים

1. תיעוד מקורות הנתונים ומשמעותם

מקור הנתונים אותו אנו נבחן הוא מקור נתונים אודות הנכסים בעיר פריז שבצרפת. מקור הנתונים מכיל נתונים אודות הנכס, כמו מספר החדרים, האם קיימת גינה או בריכה, מספר הקומות, שטח הנכס וכו'. מהתבוננות בסט הנתונים ובערכי המאפיינים, אנו מבינים כי כל אחת מהרשומות מייצגת נכס (בניין), אשר מכיל עשוי להכיל מספר דירות שונות.

- squareMeters – מטר רבוע

משתנה זה מייצג את שטח הנכס ביחידות של מטרים רבועים. משתנה זה נמדד בעזרת חיישן מרחק עבור כל דירה הקיימת בנכס, ולאחר מכן סכימה עבור כלל הדירות הקיימות בנכס.

- numberOfRooms – מספר החדרים

משתנה זה מייצג את מספר החדרים הקיימים בנכס. משתנה זה הוא חסר יחידות והוא נמדד באמצעות ספירה של מספר החדרים הקיימים בנכס. טווח הנתונים הקיים עבור משתנה זה בסט הנתונים נע בטווח של בין 1-100 חדרים.

- hasYard – האם יש גינה

משתנה זה מתאר האם קיימת גינה חיצונית לנכס. משתנה זה הינו בינארי, כלומר מקבל את הערך 1 אם קיימת גינה בנכס, ו-0 אחרת.

- hasPool – האם יש בריכה

משתנה זה מתאר האם קיימת בריכה בנכס. משתנה זה הינו בינארי, כלומר מקבל את הערך 1 אם קיימת בריכה, ו-0 אחרת.

- Floors – מספר הקומות

משתנה זה מייצג את מספר הקומות הקיימות בנכס. משתנה זה הוא משתנה בדיד, והוא יכול לקבל ערכים חיוביים בלבד. טווח הנתונים הקיים עבור משתנה זה בסט הנתונים נע בטווח של בין 1-100 קומות.

- cityCode – מיקוד

משתנה זה מייצג את המיקוד בו ממוקם הנכנס. זהו קוד מספרי אשר מייצג את המיקום הגיאוגרפי של הנכס. משתנה זה הוא משתנה שמי, כלומר אין משמעות לערך המספרי שלו.

- cityPartRange – מדד אקסקלוסיביות השכונה

משתנה זה מייצג את רמת האקסלוסיביות של הדרה בטווח של בין 1-10. אנו משערים כי משתנה זה נאסף ע"י הערכת מומחים.

- numPrevOwners – מספר הבעלים הקודמים

משתנה זה מייצג את מספר הבעלים שהחזיקו בנכס בעבר. אנו משערים כי משתנה זה נאסף בעזרת תיעוד עבר של מחזיקי הנכסים, למשל רישום בטאבו. משתנה זה הינו משתנה בדיד בטווח של בין 1-10.

- Made – שנת בניית הנכס
משתנה זה מייצג את השנה בה נבנה הנכס. אנו משערים כי משתנה זה נאסף על ידי רישום מקרקעין.

- isNewBuilt – האם בנייה חדשה
משתנה בינארי המקבל 1 אם הנכס הוא בבנייה חדשה ו-0 אחרת. אנו משערים כי משתנה זה נאסף ע"י הערכת מומחה לגבי אופי הבנייה, החומרים ושימוש בטכניקות בנייה מתקדמות.

- hasStormProtector – האם יש מגן סערות
משתנה בינארי המקבל ערך 1 במידה והנכס מכיל מגן מסערות ו-0 אחרת. אנו משערים כי משתנה זה נאסף בהתאם לנתוני הבנייה של הנכס.

- Basement – גודל המרתף
משתנה המייצג את שטח המרתף הקיים בנכס ביחידות של מטר מרובע. אנו משערים כי משתנה זה נאסף בעזרת מדידת השטח באמצעות חיישני מרחק וחישוב השטח מתוכם, וערכיו בטווח של בין 1-10000.

- Attic – גודל עליית הגג
משתנה המייצג את שטח עליית הגג הקיימת בנכס ביחידות של מטר מרובע. אנו משערים כי משתנה זה נאסף בעזרת מדידת השטח באמצעות חיישני מרחק וחישוב השטח מתוכם, וערכיו בטווח של בין 1-10000.

- Garage – גודל המחסן
משתנה המייצג את שטח המחסן הקיים בנכס ביחידות של מטר מרובע. אנו משערים כי משתנה זה נאסף בעזרת מדידת השטח באמצעות חיישני מרחק וחישוב השטח מתוכם, וערכיו בטווח של בין 100-1000.

- hasStorageRoom – האם יש חדר אחסון
משתנה בינארי המקבל ערך 1 האם בנכס קיים חדר אחסון ו-0 אחרת. אנו משערים כי משתנה זה נאסף בעזרת נתוני המתאר של הבנייה.

- hasGuestRoom – מספר חדרי האורחים משתנה המייצג את מספר חדרי האורחים הקיימים בנכס, וערכיו בסט הנתונים נעים בטווח של בין 0-10 חדרים.

- Price – מחיר הכנס

משתנה המייצג את מחיר הנכס ביחידות של יורו. אנו משערים כי משתנה זה נאסף מאתרי נדל"ן אודות הנכס.

- Category - קטגוריית הנכס

משתנה המתאר האם מדובר בנכס יוקרה "luxury" או בנכס בסיסי "basic". אנו משערים כי משתנה זה נקבע ע"י הערכת מומחים.

2. הסתברויות אפריוריות

על מנת לבחון את ההסתברויות למשתנים בבסיס הנתונים שלנו מהעולם האמיתי, חיפשנו באתר נדל"ן צרפתי (myFrenchHouse.com) את כמות הנכסים אשר עונים על כל אחד מהמאפיינים בקרב לכלל הנכסים הקיימים באתר. מבדיקה זו עולה כי השכיחות היחסית של נכס יוקרה בצרפת היא 5% מתוך כלל הנכסים, כלומר נכסי היוקרה מהווים פלח קטן יחסית בשוק הנדל"ן הצרפתי. בבסיס הנתונים שקיבלנו, נכסי היוקרה מהווים כ-12% מכלל הנכסים הקיימים, כלומר פי 2.4 משכיחותם במציאות. כמו כן, השכיחות היחסית של נכס אשר מכיל בריכה לפי נתוני אתר זה היא כ-8% מכלל הנכסים, בעוד שבבסיס הנתונים שלנו הנכסים אשר מכילים בריכה מהווים 49.9% מכלל הנכסים, כלומר פי מעל ל-6 משכיחותם באתר. בנוסף, נראה כי ההסתברות האפריורית לנכס שמכיל גינה היא כ-6%, ובבסיס הנתונים שלנו הסתברות זו עומדת על כ-50%, כלומר פי מעל ל-8 משכיחותם באתר. מההסתברויות של משתני הבריכה והגינה בסט הנתונים ניתן להסיק כי נכסים רבים בפריז מכילים אזור חיצוני לרווחת הדיירים.

מתוך סט הנתונים שקיבלנו ומתוך מאגר הנכסים בפריז הקיים באתר הנדל"ן Green-Acres.fr, הפקנו שתי היסטוגרמות אשר מציגות כמות הנכסים הקיימים עבור כל טווח מחירים (1.1. היסטוגרמות מחירים). מהתרשימים ניתן לראות כי מחיר הנכס בסט הנתונים שקיבלנו מתפלג בקירוב יחסית אחיד, בעוד שמחיר הנכסים בפריז מתוך אתר Green-Acres דועך בקירוב יחסית מעריכי, או לפי התפלגות power-law, ונראה כי סט הנתונים שלנו לא מייצג את השכיחות היחסית המתאימה עבור כל טווח מחירים של נכסים. בנוסף, מערכי המחיר בסט הנתונים עולה המסקנה כי היצע הדירות בפריז מתאים בצורה יחסית שווה לתושבים בעלי רמות השקעה שונות בנכס.

מסט הנתונים שלנו, 49.5% מהנכסים נבנו בבנייה חדשה, והדבר יכול להעיד על סטנדרט הבנייה והתחדשות הבנייה בפריז. כ-49.7% מהנכסים בנתונים מכילים מגן סערות, וזה עשוי להעיד על כך שקיימת מודעות גבוהה בפריז להמצאות הסכנה הטמונה בסערות ולכן לכמות גדולה מהנכסים יש מגן סערות. בכ-50.1% מהנכסים בסט הנתונים קיים מחסן.

עבור שאר המשתנים הצגנו את ההסתברויות האפריוריות בעזרת תרשימי היסטוגרמה של השכיחות היחסית של כל טווח ערכים בסט הנתונים (תרשימי התפלגויות המשתנים). עבור משתנים אלו ניתן לראות כי הם מתפלגים בצורה יחסית אחידה.

מתוך נתונים אלה, אנו משערים כי סט הנתונים אינו מאוזן וכי הוא אינו מייצג כראוי את המציאות.

קשרים בין המאפיינים

הקשר בין שטח הנכס במ"ר לבין המחיר:

בחרנו לבחון קשר זה מאחר ואנו מניחים שככל ששטח הנכס גדול יותר, כך המחיר שלו צפוי לגדול. נכס בעל שטח גדול לרוב יאפשר ליותר אנשים להתגורר בו, ולמתגוררים בו יהיה מרחב מחייה ונוחות גבוהים יותר, ועל כן אנו מניחים שהמחיר יגדל בהתאמה. מתרשים הפיזור שהפקנו, ניתן לראות שמתקיים קשר לינארי חיובי בין שני המשתנים. על מנת לבחון סטטיסטית את מובהקות הקשר בין שני המשתנים, ביצענו מבחן קורלציה ביניהם וקיבלנו ערך קורלציה גבוה מאוד, ערך p-value נמוך ורווח סמך אשר אינו מכיל את הערך 0, כלומר קיבלנו תוצאה שמצביעה על כך שהקשר בין שני המשתנים מובהק. אכן בהתאם לציפיותינו, ככל ששטח הנכס גדול יותר, כך גם עולה מחירו. ניתן לראות כי קיימת תצפית חריגה ובהמשך הפרויקט נטפל בה (הקשר בין שטח הנכס במ"ר לבין המחיר).

הקשר בין מדד אקסקלוסיביות השכונה לבין המחיר:

אנו מצפים שהקשר בין מדד אקסקלוסיביות השכונה לבין מחיר הנכס יהיה יחסית ישיר, ושככל שמדד האקסקלוסיביות יהיה גבוה יותר, כך הביקוש יהיה גדול יותר ובהתאם מחיר הנכס יעלה. מתרשים הקופסה שהפקנו, ניתן לראות כי המחיר מתפלג באופן יחסית אחיד בין מדדי האקסקלוסיביות השונים, ונראה כי לא קיים קשר משמעותי בין שני המאפיינים הללו, וכי חציון ואחוזוני המחיר דומים עבור כל אחד מערכי המדד. על מנת לבחון סטטיסטית את המסקנה, הרצנו מבחן ANOVA וקיבלנו ערך p-value גבוה מ-5%, אשר מצביע כי אין הבדל בין ההתפלגויות השונות של המחיר בהתאם לכל ערך של המדד. בנוסף, ניתן לראות כי קיימת קטגוריה בעלת ערכים חסרים ובהמשך העבודה נטפל בחוסרים אלו. (הקשר בין מדד האקסקלוסיביות לבין המחיר).

הקשר בין הימצאות בריכה בנכס לבין הימצאות גינה בנכס:

בחרנו לבחון את הקשר בין שני המשתנים הללו מאחר ואנו משערים כי נכסים בעלי בריכה בסבירות גבוהה יכילו גם גינה אשר הבריכה נמצאת בה. את הקשר בין שני המשתנים ייצגנו בעזרת תרשים עמודות מועמס.

ניתן לראות כי הרכבי המשתנים מאוד דומים, ונראה כי לא קיימת השפעה משמעותית של משתנה אחד על השני. על מנת לבחון סטטיסטית את השוואת ההרכבים, הרצנו מבחן חי בריבוע וקיבלנו ערך p-value גדול מ-0.05, כלומר בר"מ של 5% נקבע כי לא קיים קשר בין שני המשתנים (הקשר בין הימצאות בריכה בנכס לבין הימצאות גינה בנכס).

הקשר בין הימצאות בריכה בנכס לבין מדד אקסקלוסיביות השכונה:

בחרנו לבחון את הקשר בין שני המשתנים מאחר ואנו משערים כי קיומן של בריכות בנכסים עשוי להצביע על מדד האקסקלוסיביות של השכונה. על מנת לבחון את הקשר בין המשתנים הפקנו תרשים קופסה ובו ניתן לראות כי התפלגות מדד האקסקלוסיביות עבור נכסים בעלי בריכה ומחוסרי בריכה נראית דומה יחסית. על מנת לבחון זאת סטטיסטית, הרצנו מבחן ANOVA להשוואת התפלגויות מדד האקסקלוסיביות ביחס למשתנה הימצאות הבריכה בנכס, וקיבלנו ערך p-value גבוה מ-5%, אשר מצביע כי אין הבדל בין ההתפלגויות השונות של המדד ביחס להימצאותה של בריכה בנכס (

הקשר בין הימצאות בריכה בנכס לבין מדד אקסקלוסיביות השכונה).

הקשרים בין המאפיינים לבין משתנה המטרה

מדד האקסקלוסיביות של השכונה וקטגוריית הנכס – משתנה זה עשוי להיות בעל השפעה משמעותית על משתנה המטרה, קטגוריית הנכס (יוקרתי/ בסיסי). דירוג האקסקלוסיביות של השכונה מושפע לרוב מרמת ההכנסה של התושבים, המדד הסוציואקונומי של השכונה, רמת הפשיעה, רמת ההשכלה של התושבים ועוד. בשכונות בעלות מדד סוציואקונומי גבוה, רמת פשיעה נמוכה ורמות הכנסה גבוהות, סביר להניח שהנכסים יהיו בעלי רמת יוקרתיות גבוהה יותר, אשר עשויה לנבוע מקיומם של מאפיינים אלו אשר הופכים את הנכס למבוקש ויוקרתי. לעומת זאת, ניתן לראות כי בסט הנתונים שלנו, דירוג האקסקלוסיביות של השכונה מתפלג באופן יחסית דומה בין שתי קטגוריות הנכסים השונות (יוקרתי/ בסיסי), וממבחן ה-ANOVA אכן נראה כי אין הבדל בין התפלגויות המדד ביחס לקטגוריית הנכס (הקשר בין מדד האקסקלוסיביות לבין קטגוריית הנכס

האם בנייה חדשה וקטגוריית הנכס – משתנה זה עשוי להצביע גם כן על יוקרתיות הנכס. נכסים שהוקמו בשנים האחרונות, ובנייתם השתמשה בחומרי גלם איכותיים וטכניקות בנייה מתקדמות, צפויים להיות יוקרתיים יותר ביחס לאחרים. בנייה חדשה לרוב מאופיינת בתשתיות חדשות ובבנייה עמידה יותר, דבר שעשוי ליצור ביקוש גדול ולהוביל לרמת יוקרה גבוהה לנכס. בסט הנתונים שלנו ניתן לראות כי הרכב משתנה זה די שונה בין שתי קטגוריות הנכסים השונות (יוקרתי/ בסיסי). (הקשר בין האם בנייה חדשה לבין קטגוריית הנכס) על מנת לבחון סטטיסטית את הקשר בין בנייה חדשה בנכס לבין קטגוריית היוקרה שלו, ביצענו מבחן

חי בריבוע להשוואת הרכבים וקיבלנו p-value קטן מ-5%, ולכן בר"מ זו ניתן להסיק כי הקשר בין המשתנים מובהק.

האם קיימת בריכה וקטגוריית הנכס – הימצאותה של בריכה בנכס עשויה להפוך אותו ליוקרתי ומבוקש, מאחר ובריכה נחשבת לעיתים כסמל סטאטוס, ומהווה "אטרקציה" נוספת לדיירי הנכס, דבר שלרוב הופך אותו למבוקש ויוקרתי יותר. על מנת לבחון את הקשר, ביצענו מבחן חי בריבוע להשוואת הרכבים וקיבלנו p-value קטן מ-5%, ולכן בר"מ זו ניתן להסיק כי הקשר בין המשתנים מובהק (הקשר בין האם יש בריכה לבין קטגוריית הנכס).

מחיר וקטגוריית הנכס – משתנה נוסף שעשוי להצביע על רמת יוקרה גבוהה לנכס הוא משתנה המחיר. לרוב, מחירים גבוהים נובעים מתוך קיומו של ביקוש נרחב לנכס, וביקוש נרחב מושפע מרמת היוקרתיות. אנו מצפים שמשתנה המחיר יהיה אחד המאפיינים הבולטים לסיווג נכס כיוקרתי, ולרוב טווח המחירים של נכסים יוקרתיים שונה משמעותית מטווח של נכסים שאינם יוקרתיים. לעומת זאת, ניתן לראות כי בסט הנתונים שלנו, המחיר מתפלג באופן יחסית דומה בין שתי קטגוריות הנכסים השוניות (יוקרתי/ בסיסי) (הקשר בין מחיר לבין קטגוריית הנכס), ובמבחן ANOVA שהרצנו להשוואת התפלגויות המחיר ביחס לקטגוריית הנכס, קיבלנו p-value גדול מ-5% שמצביע כי בר"מ זו ההבדל בין התפלגויות, המחיר ביחס לקטגוריה לא מובהק.

בנוסף, על מנת לראות את הקשרים והקורלציות בין המשתנים השונים בבסיס הנתונים, הפקנו תרשים מטריצת קורלציה. מהתרשים ניתן לראות כי בין המחיר למטר המרובע קיימת קורלציה חזקה, וכי בין קטגוריית הנכס למשתנים המציינים האם יש גינה, בריכה והאם בנייה חדשה קיימת קורלציה בינונית. בין שאר המשתנים לא נראה כי קיימת קורלציה.

3. איכות הנתונים

בסט הנתונים קיימים מספר נתונים חסרים. חוסרים אלו נמצאים בכלל המשתנים, מלבד במשתנה שטח הנכס במ"ר ובמשתנה מספר חדרי האורחים בנכס. על מנת להתמודד עם הנתונים החסרים, נבדוק האם ברשומה מסוימת מספר החוסרים גדול ומשמעותי, ואז נבחר להסיר אותה מסט הנתונים, או שמא מדובר בחוסר של מספר ערכים קטן יחסית ברשומה, ואז נבחר לטפל בה בדרכים אחרות.

עבור רשומות בעלות מספר חוסרים קטן, נבחר להשלים אותן באחת ממספר דרכים על מנת להימנע מאיבוד נתונים: הראשונה, להכניס את הערך הממוצע של אותו משתנה, והשנייה להזין ערך 0 עבור משתנים בינאריים (0/1), מתוך נקודת הנחה שהחוסר בערך השדה עשוי להיות מוסבר מערכו האמיתי (=NMAR Not Missing at Random), למשל נכסים שלא דווח אם יש בהם בריכה, סביר שהדבר נובע מאחר ובאמת לא קיימת בהם בריכה, מאחר ואם אכן הייתה בריכה, לא היו מוותרים על דיווח זה. בסעיף הבא מפורטות הדרכים בהם התמודדנו עם החוסרים. בסעיף הבא נתאר את כלל השינויים שבוצעו בסט הנתונים.

בנוסף, בסט הנתונים שלנו קיימים מספר נתונים שאינם הגיוניים: קיים נכסים ששנת הבנייה שלו גדולה מהשנה הנוכחית, נכס בעלי מחיר שלילי, ונכנס בעל קוד מיקוד שלילי. הטיפול במקרים אלה יתואר בסעיף הכנת הנתונים. כמו כן, חלק מהמשתנים הבינאריים מכילים גם ערכי 0 ו-1 וגם ערכי yes/no. על מנת לשמור על אחידות, נמיר את ערכי הyes- לערך הבינארי 1, ואת ערכי ה-no לערך הבינארי 0.

הכנת הנתונים

את כל השינויים שיתוארו כעת, ביצענו בסט הנתונים דרך האקסל.

עבור משתנים בינאריים בהם היו ערכי 0/1 ובנוסף ערכי yes/no (האם בנייה חדשה והאם מכיל מגן סערות), החלפנו את "yes" בערך הבינארי המתאים לו, 1, ואת "no" ב-0.

באחת הרשומות היה מחיר שלילי עבור הנכס. מכיוון שבתחילת העבודה מצאנו כי קיים קשר לינארי בין משתנה המטר המרובע ומשתנה המחיר, ביצענו מבחן רגרסיה לינארית על מנת לקבל את משוואת הרגרסיה המתאימה לקשר לינארי זה (רגרסיה לינארית - מחיר ומטר מרובע). לאחר קבלת הפרמטרים של המודל (החותך והשיפוע), השלמנו את שדה המחיר בעזרת הצבת הערך של השטח במ"ר במשוואת הרגרסיה.

רשומות בעלות ערך חסר יחיד עבור משתנים בינאריים (האם יש בריכה) בחרנו להשלים בעזרת הזנת הערך 0, מתוך הנחה שסביר יותר כי נתון זה לא דווח בעקבות אי קיום הבריכה בנכס (NMAR).

ברשומה בה היה ערך מיקוד שלישי שאינו הגיוני, בחרנו להשלים בעזרת הכנסת מספר רנדומלי שהוגרל בטווח הערכים של המשתנה בסט הנתונים. בחרנו להשלים בעזרת מספר רנדומלי מאחר ולערך המיקוד עצמו לא קיים משמעות כערך, ולכן בחרנו לא למצע אותו.

אחת הרשומות הייתה בעלת שנת בנייה גדולה מהשנה הנוכחית, על מנת לטפל ברשומה זו, הזנו במקומה את השנה הגדולה ביותר שהייתה קיימת בסט הנתונים על מנת לשמור על היגיון ככל שניתן ברשומה.

רשומות בעלות מספר חוסרים רבים בחרנו להסיר, על מנת להימנע מהטיה מוגזמת של הנתונים.

רשומות אשר הייתה חסרה בהן שנת הבנייה, השלמנו בעזרת השדה "האם בנייה חדשה". עבור כל ערך של משתנה זה (0/1), חישבנו את שנת הבנייה הממוצעת (שנת הבנייה הממוצעת לבנייה חדשה ולבנייה לא

חדשה). לכל שנת בנייה חסרה, בדקנו האם הנכס נבנה בבנייה חדשה ואם כן אז הזנו את השנה הממוצעת לבנייה חדשה, ולהיפך.

בשאר הרשומות בהן היו חוסרים בודדים, בחרנו להשלים את החוסרים ולא למחוק את הרשומות על מנת לא לאבד את הנתונים. על מנת להשלים את הנתונים החסרים, מיצענו את ערכי העמודה, והכנסנו את הממוצע אל תוך הנתונים החסרים. בחרנו להשלים על ידי הערך הממוצע בשדות בהם לא מצאנו קשרים משמעותיים בין המשתנים, מאחר והממוצע ייצור פחות הטיה בנתוני המשתנה.

דיסקריטיזציה של משתנים רציפים

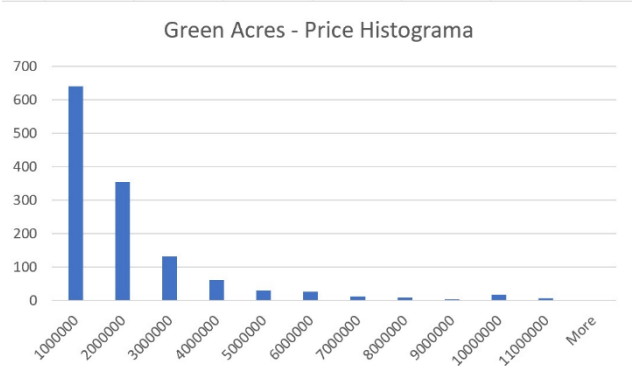
שקלנו לבצע דיסקריטיזציה של המשתנים הרציפים בסט הנתונים שקיבלנו, ועל מנת לבחון את הקבוצות שיש לאחד, בנינו תרשימים אשר מציגים את ההתפלגויות השכיחות היחסית שלהם (נספח 1.2), וגם את ההתפלגויות שלהם ביחס למשתנה המטרה. לאחר בחינה של התרשימים, ראינו כי כלל המשתנים הרציפים מתפלגים בקירוב להתפלגות אחידה, ובנוסף ההתפלגות של כל אחד מערכי המשתנים סביב המשתנה המוסבר יחסית זהה, ולכן לא ראינו לנכון לבצע דיסקריטיזציה.

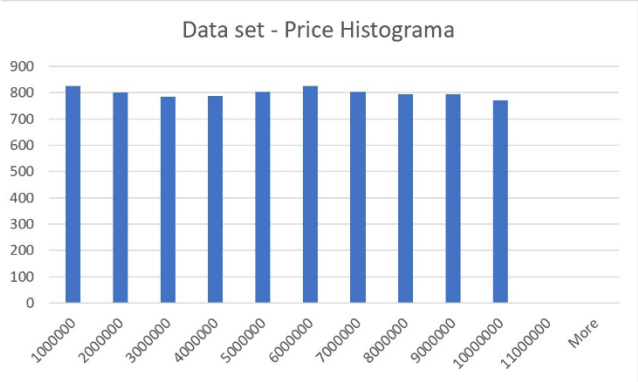
נספחים

1. הסתברויות אפריוריות

1.1. היסטוגרמות מחירים

מימין – היסטוגרמה מתוך בסיס הנתונים שלנו, משמאל – היסטוגרמה מתוך נתוני אתר Green-Acres:



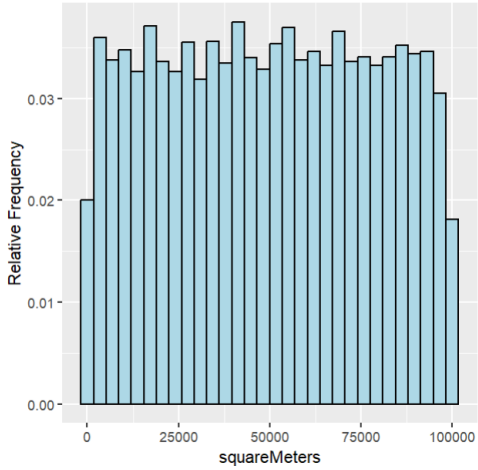


חזור

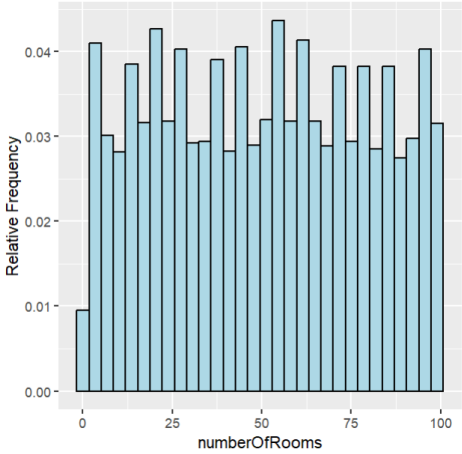
1.2. תרשימי התפלגויות המשתנים

חזור

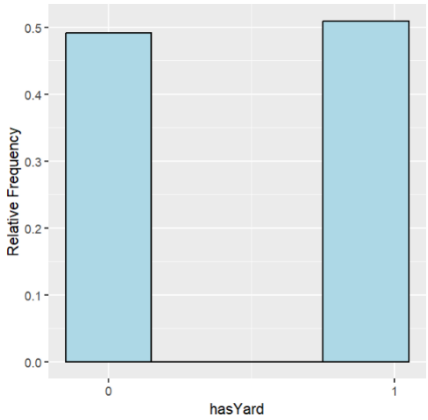
מטר רבוע



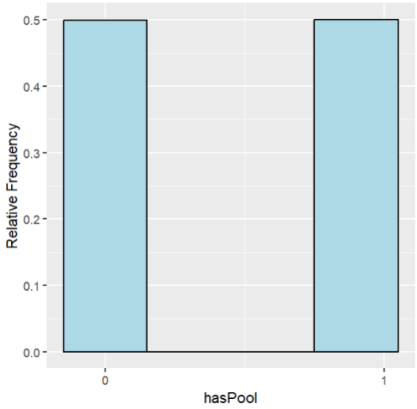
מספר החדרים



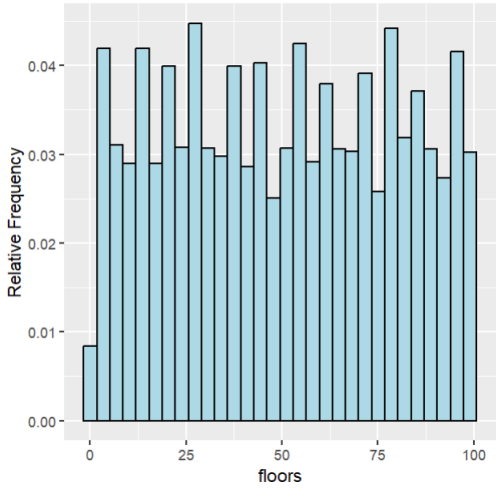
האם יש גינה



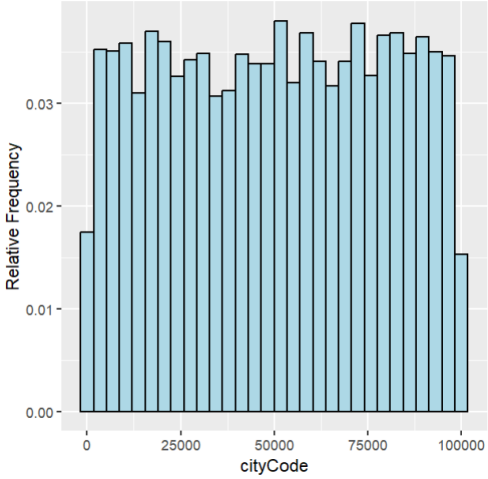
האם יש בריכה



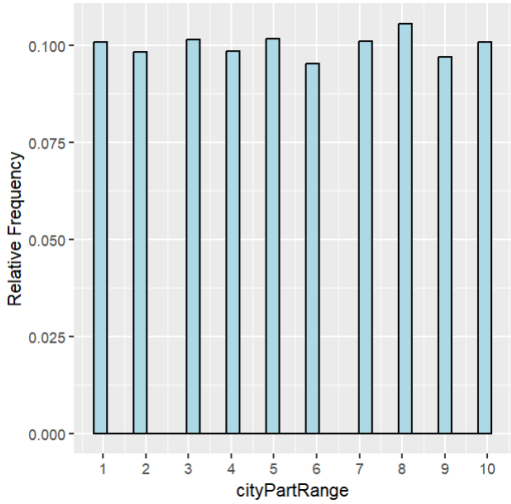
מספר הקומות



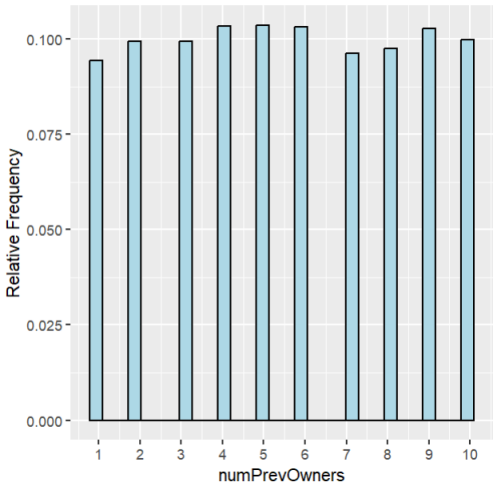
מיקוד



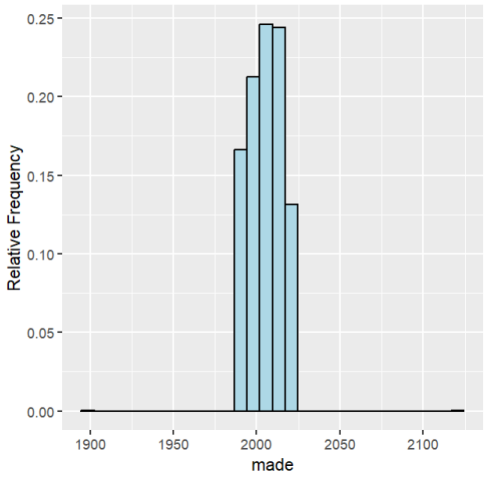
מדד אקסקלוסיביות השכונה



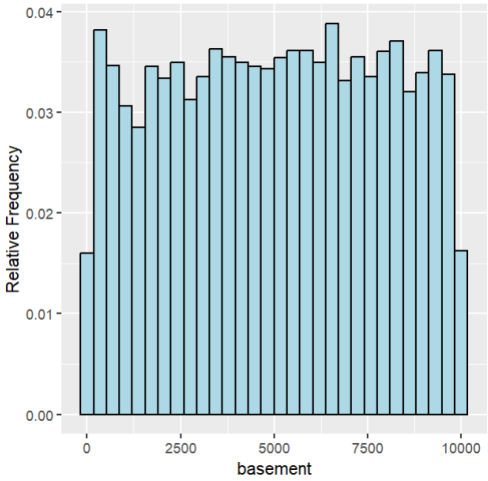
מספר הבעלים הקודמים



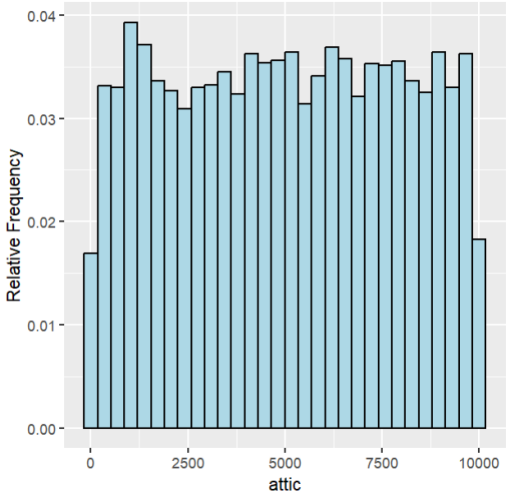
שנת בניית הנכס



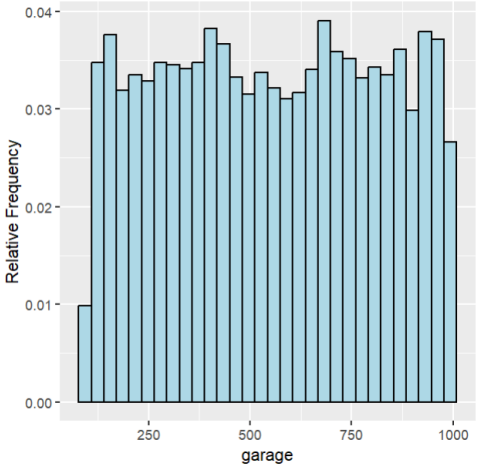
גודל המרתף



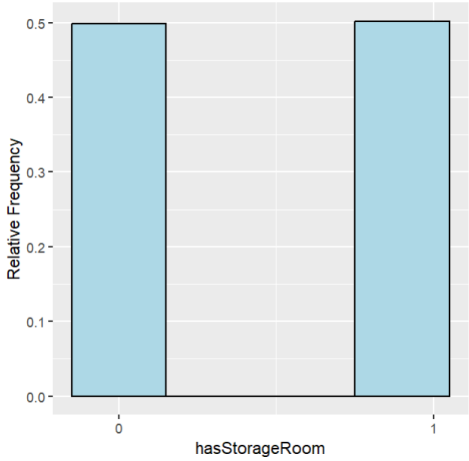
גודל עליית הגג



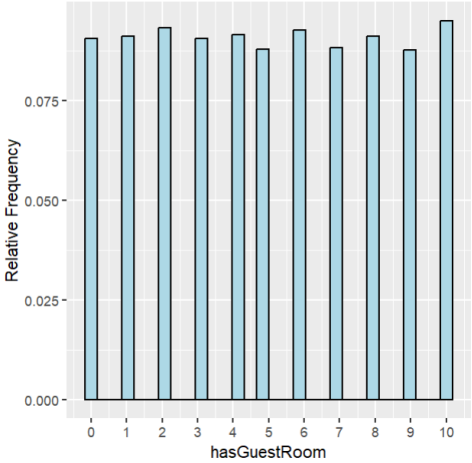
גודל המחסן



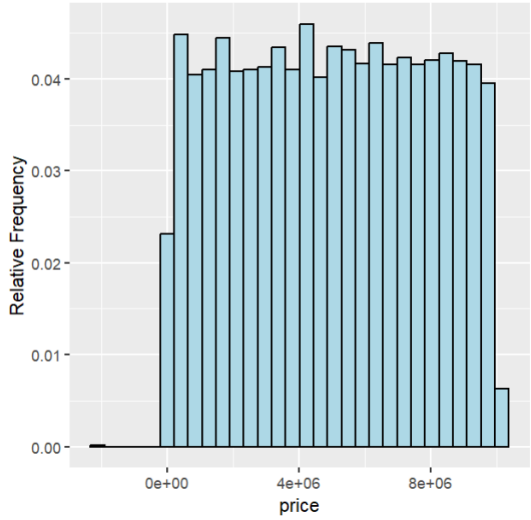
האם יש חדר אחסון



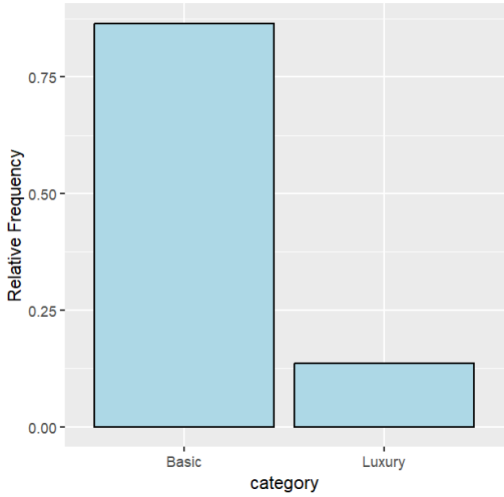
מספר חדרי האורחים



מחיר הנכס



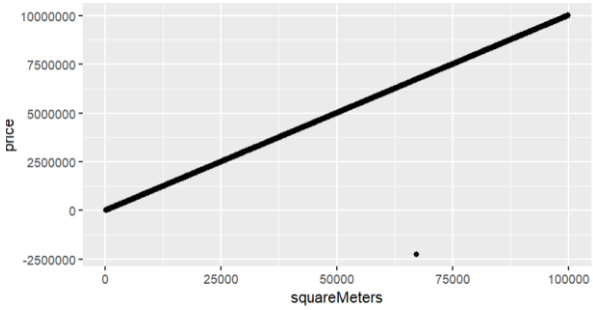
קטגוריית הנכס



חזור

2. קשרים בין משתנים

הקשר בין שטח הנכס במ"ר לבין המחיר



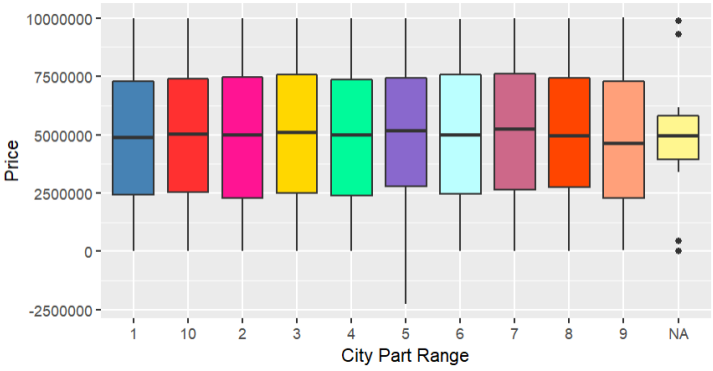

```
> correl<-cor.test(table$squareMeters, table$price, method=c("pearson"))
> correl

        Pearson's product-moment correlation

data:  table$squareMeters and table$price
t = 2560.5, df = 7985, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9993643 0.9994177
sample estimates:
      cor
0.9993916
```

חזור

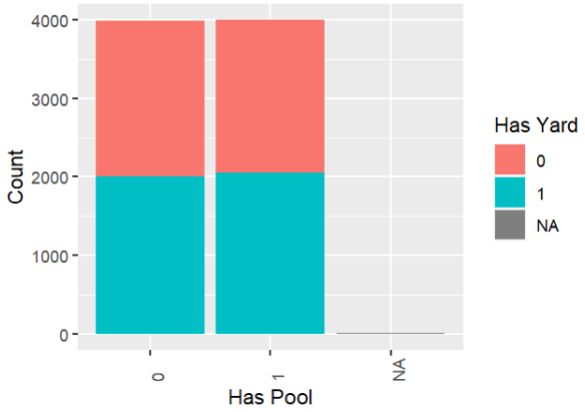
הקשר בין מדד אקסקלוסיביות השכונה לבין המחיר




```
> av1 <- aov(table$price~table$cityPartRange)
> summary(av1)
                     Df    Sum Sq   Mean Sq F value Pr(>F)
table$cityPartRange   1 1.707e+10 1.707e+10   0.002  0.964
Residuals          7969 6.630e+16 8.319e+12
29 observations deleted due to missingness
```

חזור

הקשר בין הימצאות בריכה בנכס לבין הימצאות גינה בנכס



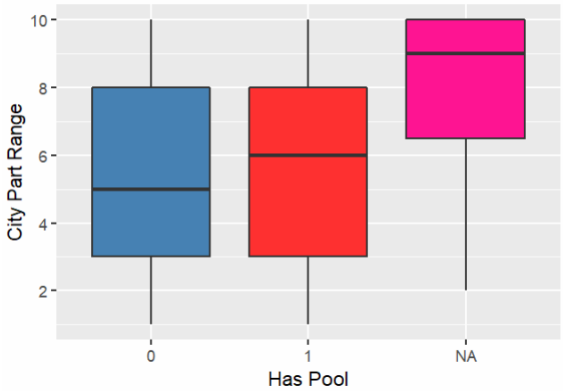

```
> chiTest<-chisq.test(table$hasPool, table$hasYard, correct=FALSE)
> chiTest

        Pearson's Chi-squared test

data:  table$hasPool and table$hasYard
X-squared = 1.0831, df = 1, p-value = 0.298
```

חזור

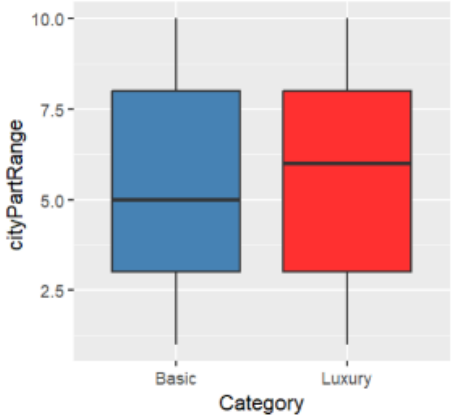
הקשר בין הימצאות בריכה בנכס לבין מדד אקסקלוסיביות השכונה




```
> av2 <- aov(table$cityPartRange~table$hasPool)
> summary(av2)
               Df Sum Sq Mean Sq F value Pr(>F)
table$hasPool   1      5   5.420   0.656  0.418
Residuals    7978  65919   8.263
20 observations deleted due to missingness
```

חזור

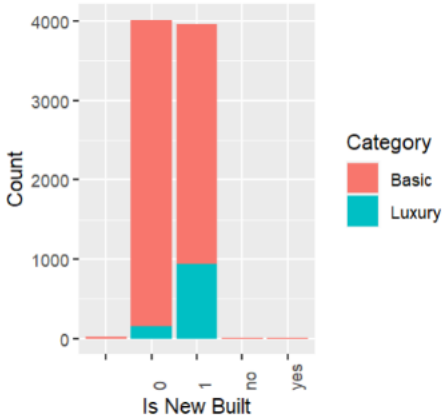
הקשר בין מדד אקסקלוסיביות השכונה לבין קטגוריית הנכס




```
> av4 <- aov(table$category~table$cityPartRange)
> summary(av4)
                      Df Sum Sq Mean Sq F value Pr(>F)
table$cityPartRange    1    0.0 0.00334   0.028  0.866
Residuals           7981  939.7 0.11774
```

חזור

הקשר בין האם בנייה חדשה לבין קטגוריית הנכס



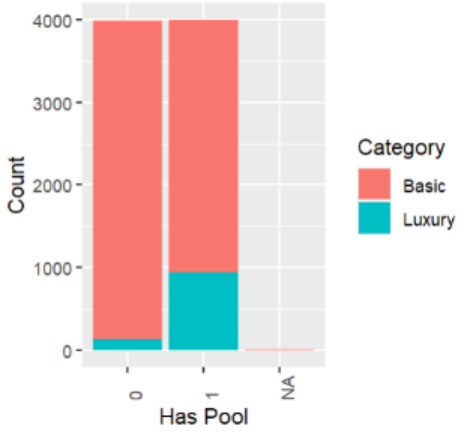

```
> chiTest2<-chisq.test(table$isNewBuilt, table$category, correct=FALSE)
> chiTest2

        Pearson's Chi-squared test

data:  table$isNewBuilt and table$category
X-squared = 675.44, df = 1, p-value < 2.2e-16
```

חזור

הקשר בין האם יש בריכה לבין קטגוריית הנכס



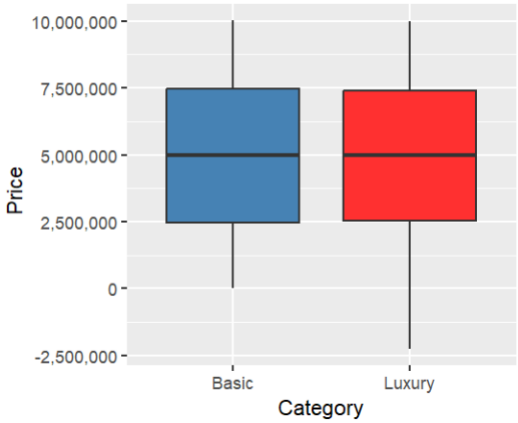

```
> chiTest1<-chisq.test(table$hasPool, table$category, correct=FALSE)
> chiTest1

        Pearson's Chi-squared test

data:  table$hasPool and table$category
X-squared = 688.06, df = 1, p-value < 2.2e-16
```

חזור

הקשר בין מחיר הנכס לבין קטגוריית הנכס

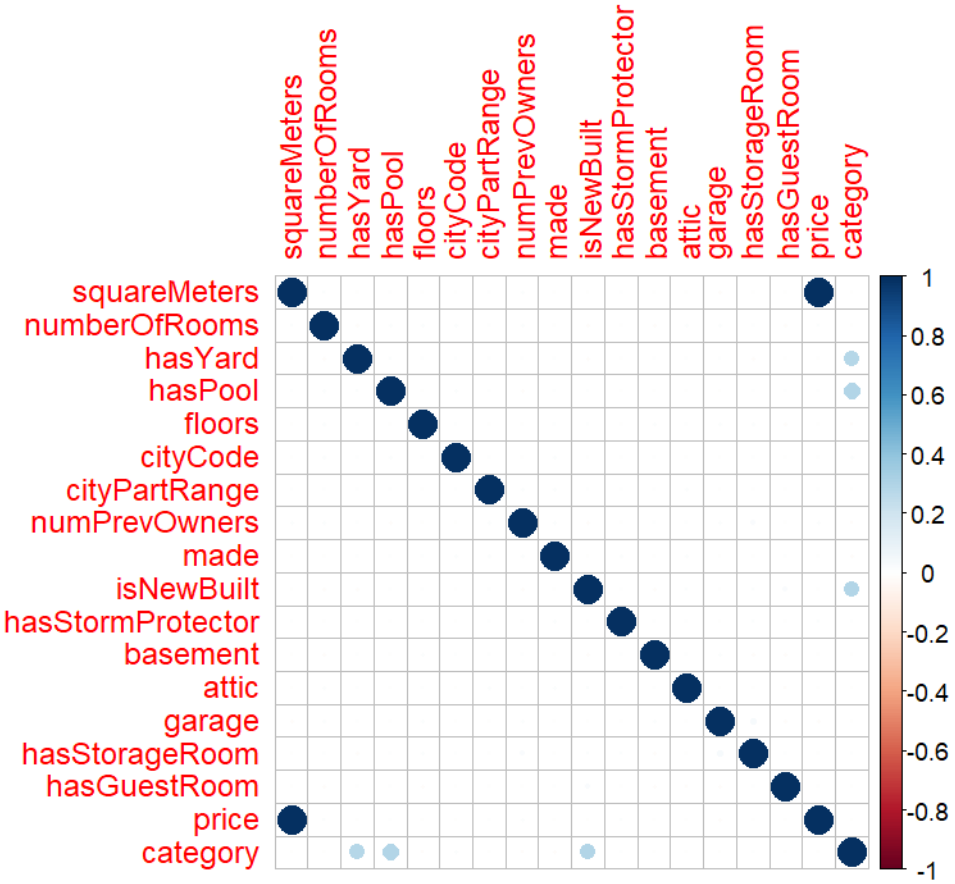



```
> av3 <- aov(table$category~table$price)
> summary(av3)
              Df Sum Sq Mean Sq F value Pr(>F)
table$price    1      0 0.03497   0.297  0.586
Residuals   7968    938 0.11772
13 observations deleted due to missingness
```

חזור

מטריצת קורלציה בין המשתנים

חזור



חזור

2. הכנת הנתונים

רגרסיה לינארית – מחיר ומטר מרובע

```
> fit <- lm(table$price ~ table$squareMeters)
> summary(fit)

Call:
lm(formula = table$price ~ table$squareMeters)

Residuals:
     Min       1Q   Median       3Q      Max 
-8982749    -1338      895     3350    13066 

Coefficients:
                    Estimate Std. Error  t value Pr(>|t|)    
(Intercept)        6.459e+03  2.246e+03    2.875  0.00405 ** 
table$squareMeters 9.998e+01  3.912e-02 2555.512  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 100700 on 7968 degrees of freedom
  (13 observations deleted due to missingness)
Multiple R-squared:  0.9988,	Adjusted R-squared:  0.9988 
F-statistic: 6.531e+06 on 1 and 7968 DF,  p-value: < 2.2e-16
```

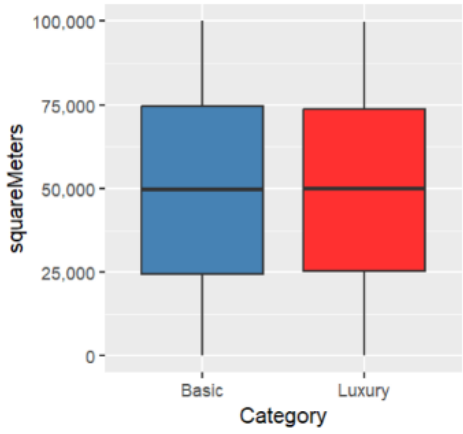
$$price = 6459 + 99.98 * squareMeters = 6459 + 99.98 * 67215 = 6726615$$

חזור

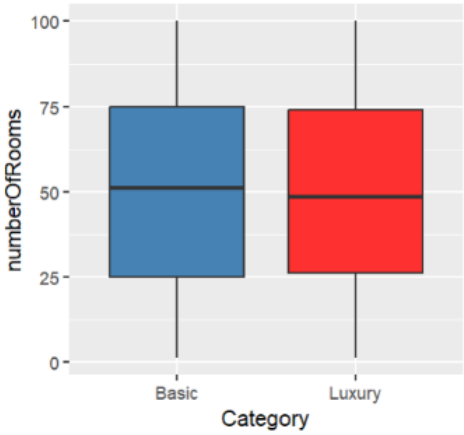
התפלגויות המשתנים הרציפים ביחס למשתנה המטרה

חזור

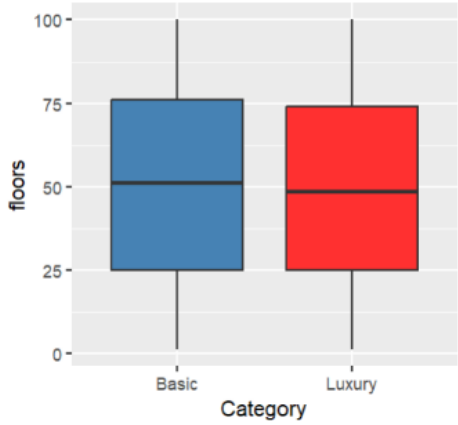
מטר רבוע וקטגוריית הנכס



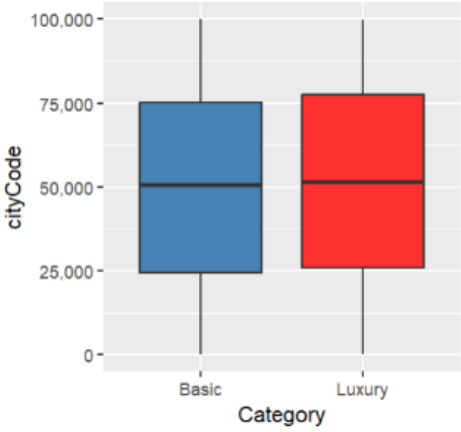
מספר החדרים וקטגוריית הנכס



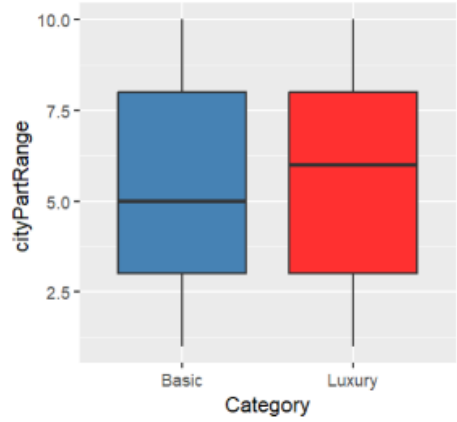
מספר הקומות וקטגוריית הנכס



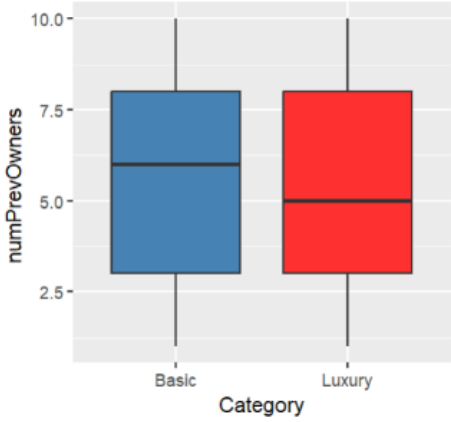
מיקוד וקטגוריית הנכס



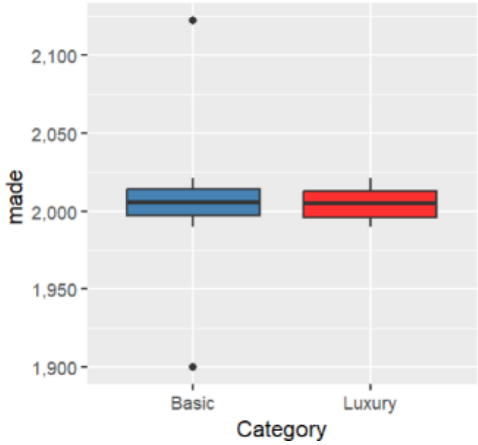
מדד אקסקלוסיביות השכונה וקטגוריית הנכס



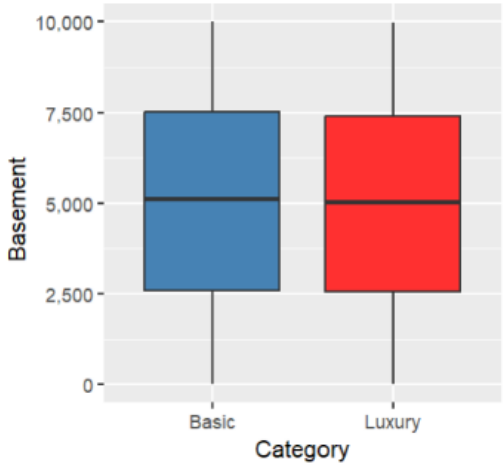
מספר הבעלים הקודמים וקטגוריית הנכס



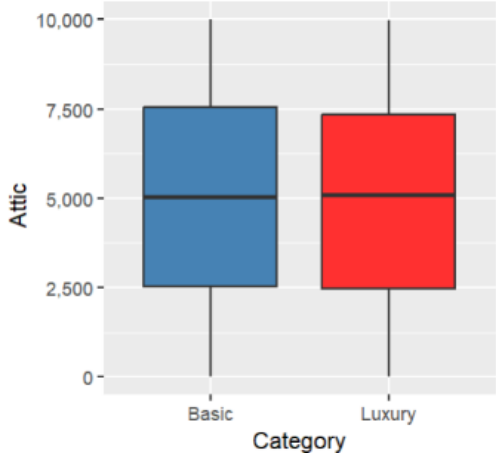
שנת בניית הנכס וקטגוריית הנכס



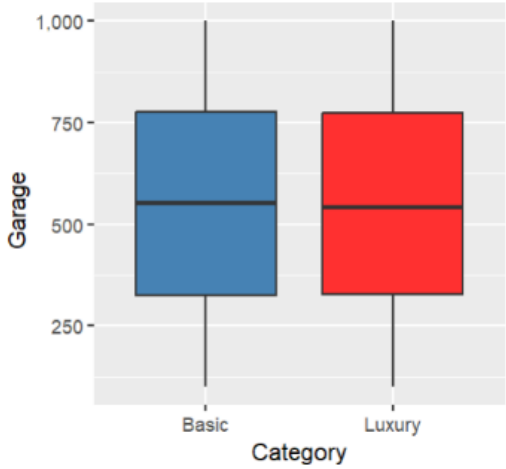
גודל המרתף וקטגוריית הנכס



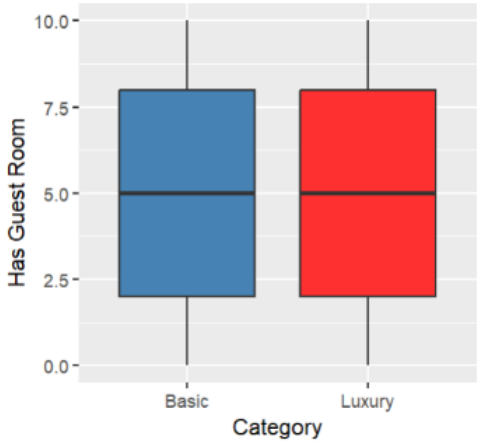
גודל עליית הגג וקטגוריית הנכס



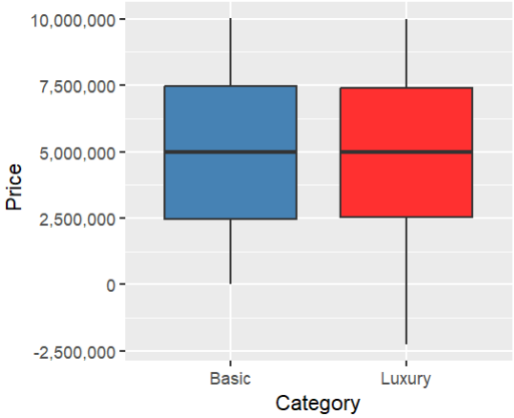
גודל המחסן וקטגוריית הנכס



מספר חדרי האורחים וקטגוריית הנכס



מחיר הנכס וקטגוריית הנכס



חזור

ביביליוגרפיה

- U B H R A J Y O T, S. S., & Sahoo, I. (n.d.). *Maler and Vincent Handbook of EnvEco, vol Related papers*.
- Wyatt, P. (n.d.). *Geographical information system Using a geographical information system for property valuation*.

חזור