אזורים ירוקים ומחירים בשכונה

מטרה:

בשנתיים האחרונות, בצל מגפת הקורונה דובר רבות על חשיבות הרכב שימושי הקרקע בשכונה ככלל וחשיבותו של אזור טבע עירוני, "אזור ירוק" בשכונה בפרט. בעבודה זו אשאף לבחון האם דרישה זו, לאזור ירוק בשכונה התקיימה גם בעבר. במהלך העבודה אשתמש בכלים אשר למדנו בקורס מבוא למדע הנתונים הגאו-מרחבי ואבחן האם אחוז השטחים בשכונה המוקצים לשטחים ירוקים משפיע על מחיר הנכסים בה ככלל ועל מגמת עליית המחירים בה בפרט. זאת מתוך נקודת ההנחה כי ביקוש יבוא לידי ביטוי בעליית מחירים. לטובת העבודה אעשה בשימוש עסקאות הנדל"ן אשר בוצעו בתל אביב בין השנים 2012-2017.

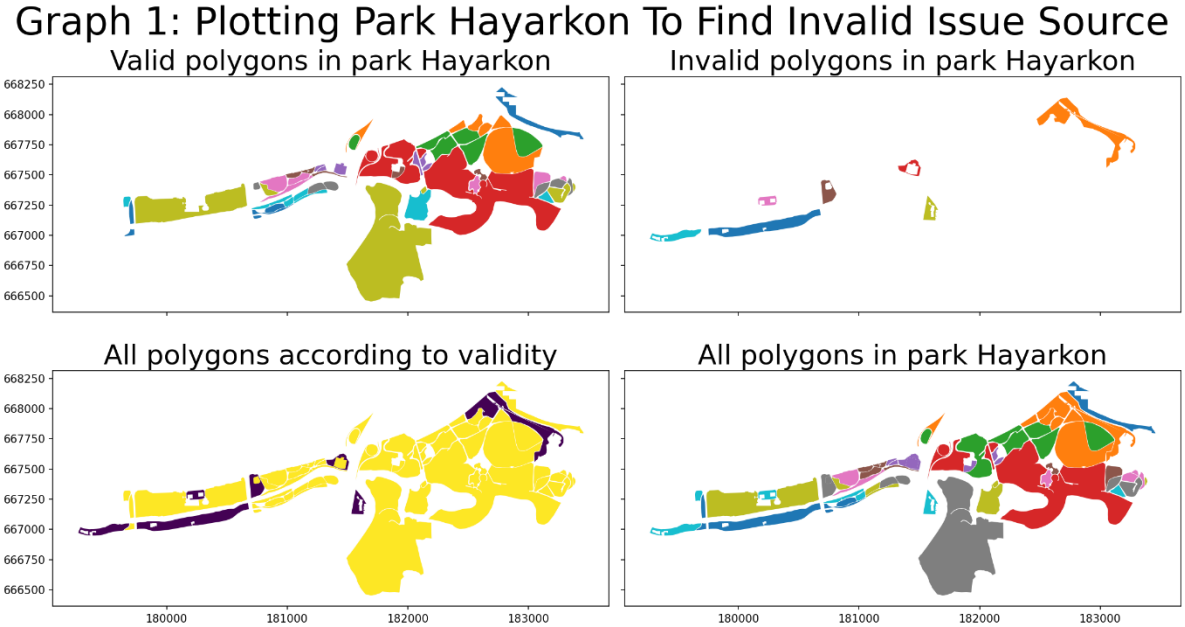
שיטות:

ייבוא הנתונים - במהלך העבודה אעשה שימוש בשלוש מקורות נתונים עיקרים אותם ייבאתי ועליהם להיות באותה תיקיה כמו המחברת. הקבצים הועלו ביחד עם העבודה בתוך תיקיות יש לעלות את כל הקבצים הנמצאים בתיקיות NBHD וGreen ולא רק את ה shapfile על הקבצים להיות מחוץ לתיקיות בהן הם נשלחו.

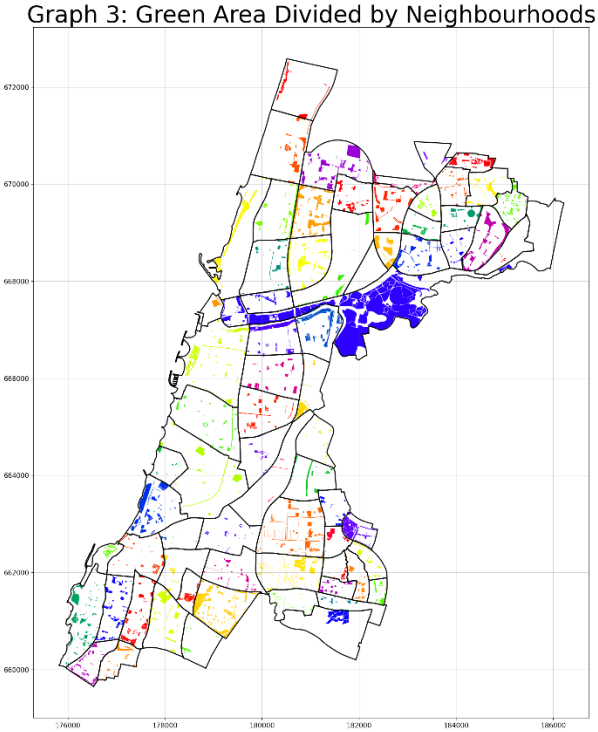
1. מחירים - קובץ csv הכולל את כל עסקאות הנדל"ן בתל-אביב בשנים 2012-2017 , כולל מחיר, מיקום הנכס, מחיר העסקה, קומה, שטח ועוד.
2. שכונות - קובץ shapefile אשר הורד מאתר עירית תל אביב הכולל פוליגונים עם מיקום במרחב של כל השכונות בתל אביב, מספר מזהה, שם השכונה ועוד.
3. שטחים ירוקים - קובץ shapefile אשר הורד מאתר הGIS של עיריית תל אביב הכולל את השטחים הירוקים המצאים בתל אביב כפוליגונים עם מיקום במרחב, שמם, מספר מזהה וסוג השטח הירוק.

עיבוד ראשוני של הנתונים:

שטחים ירוקים – בקובץ השטחים הירוקים בחנתי האם כלל הרשומות הנמצאות בטבלה זו יוצרות צורה גיאומטרית "תקפות", כאלו הניתנות לעיבוד גיאוגרפי לטובת המשך העבודה. בחינה של נתון זה הובילה למסקנה כי ישנן צורות גיאומטריות לא תקפות. בבחינה של רשומות אלו נראה כי רבות מהן נמצאות בפארק הירקון. על כן, שורטטו השטחים התקפים והלא תקפים הנמצאים בפארק הירקון על מנת להגיע לשורש הבעיה (ראה גרף 1). נראה כי, השטחים הלא תקפים הם אלו אשר יש בהם חורים, על כן נוסף "באפר" בגודל 0 אשר תיקן את בעיה זו.



Graph 3: Green Area Divided by Neighbourhoods

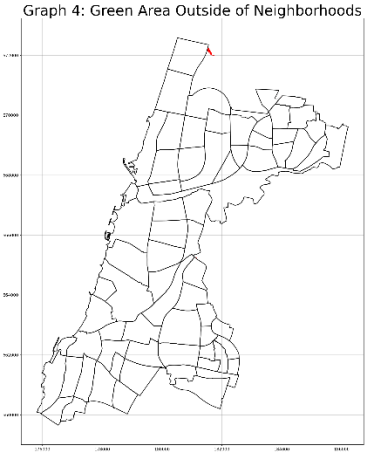


מחירים- בקובץ המחירים בוצעה המרה של הנתונים מפנדס לגיו-פנדס, זאת תוך יצירת עמודה הכוללת בתוכה את הנתונים הגיאוגרפים והקפדה על היטל זהה לשל שתי השכבות הנוספות. בנוסף, נוספה עמודה של מחירי דירות מנורמלים לשטח, עמודת התאריך הוחלפה בעמודה הכוללת את שנת המכירה בלבד. לבסוף, לטובת מיקוד וביצוע ניתוחים סטטיסטים הורדו עמודות מיותרות כדוגמת שם העיר (ראה גרף 2 בנספחים).

נתונים סטטיסטיים לכל שכונה:

לאחר העיבוד הראשוני, נבחנו הנתונים הסטטיסטיים הדרושים לטובת העבודה לכל שכונה ובהם: אחוז השטחים הירוקים בשכונה, המחיר הממוצע לדירה בשכונה לאורך כל התקופה הנבחנת והשינוי הדרסטי ביותר בין השנים כמו גם הקשר בין נתוני המכירות לבין אחוזי השטחים הירוקים.

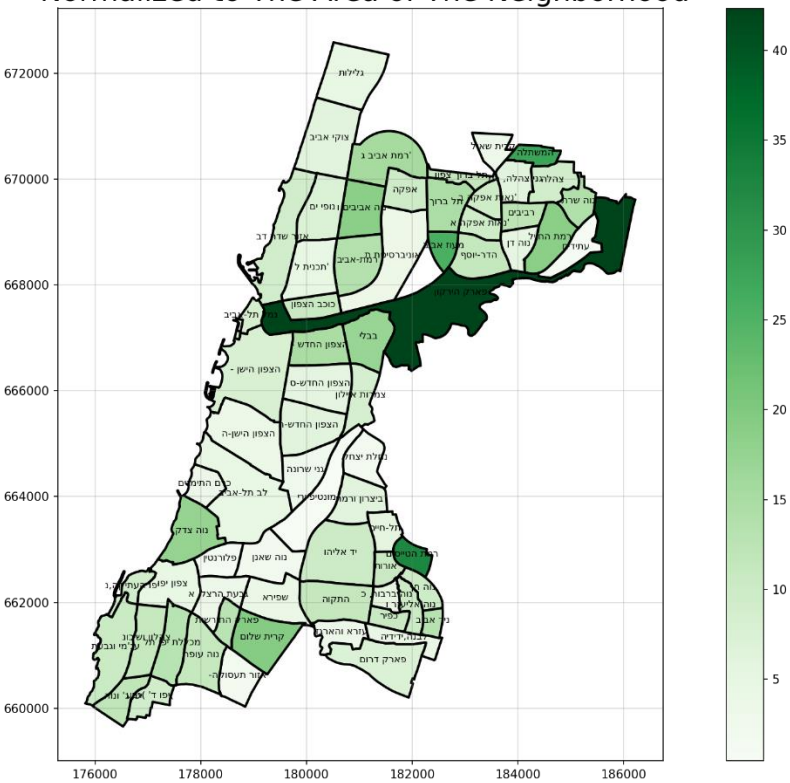
Graph 4: Green Area Outside of Neighborhoods



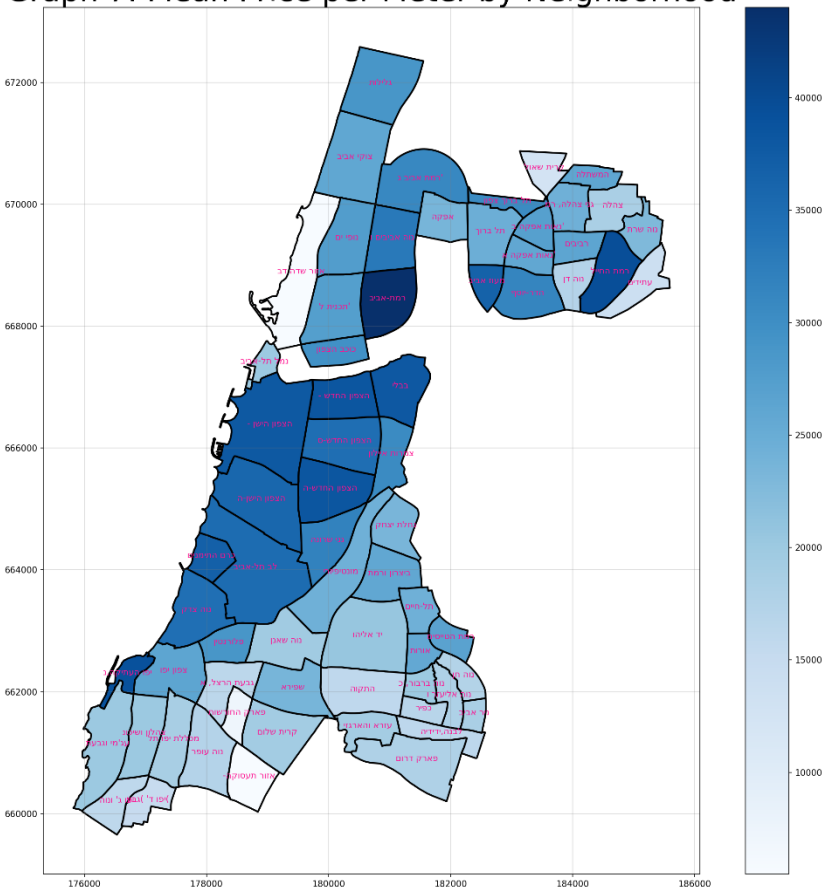
אחוז השטחים הירוקים בשכונה – לטובת בחינת מידע זה, ראשית יש לבחון באיזה שכונה נמצא כל שטח ירוק. לטובת כך נוצרה טבלה חדשה המכילה, עבור כל שטח ירוק את פרטי השכונה אליה הוא משויך. במידה ושטח ירוק נפרס על פני יותר משכונה אחת הוא ייחצה וכל חלק שלו יכיל את הנתונים הגיאוגרפים ופרטי השכונה הרלוונטיים. בגרף מספר 3, השטח הירוק צבוע לפי מספר השכונה וניתן לראות כי חלוקה זו בוצעה בהצלחה והשטחים ירוקים אשר נמצאים על קווי התפר בין שכונות פוצלו . כמו כן, נראה כי השטחים הירוקים אשר נכללו בטבלה המקורית אך לא בטבלה החדשה הם אלו אשר כוללים פוליגונים ריקים או לחלופין ממוקמים מחוץ לשטחי תל אביב (ראה שטח אדום בגרף 4).

לאחר מכן, נוספה עמודה חדשה אשר בה חושב השטח של כל שטח ירוק לאחר הניתוח הגיאוגרפי ויוצא המידע המעובד כקובץ json. בהמשך, בוצע קיבוץ של הטבלה הנוכחית לפי מספר השכונה ונוצרה טבלה חדשה המכילה עבור כל שכונה את סכום השטחים הירוקים אשר היא מכילה בתוכה. הטבלה המכילה את סכום השטחים הירוקים, אוחדה עם טבלת השכונות המקורית לטובת הצמדת הנתונים הגיאוגרפים של כל שכונה לסכום השטחים הירוקים אשר חושב. מן הנתונים הגיאוגרפים של השכונה חושב שטח השכונה ולבסוף אחוז השטחים הירוקים בתוכה (סכום השטחים הירוקים חלקי שטח השכונה כפול 100). אחוז השטחים הירוקים בכל שכונה הוצג בצורה ויזואלית ורציפה על גבי מפת של העיר תל אביב (ראה גרף 5).

Graph 5: Total Green Area in Each Neighborhood Normalized to The Area of The Neighborhood



Graph 7: Mean Price per Meter by Neighborhood

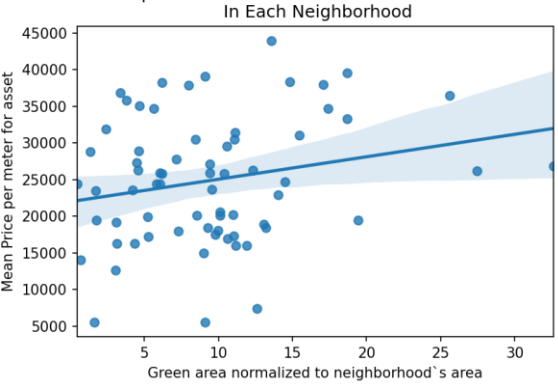


לאחר סיום העבודה על טבלה זו, ולפני המעבר לעבודה על נתוני המכירות בשכונה בוצע ניקוי של הנתונים המוצגים בטבלה. זאת על ידי הורדת עמודות מיותרות אשר כוללות בתוכן מידע אשר לא ישמש בהמשך העבודה. בכך צומצמו מספר העמודות בטבלה מ16 ל4 בלבד (ראה גרף מספר 6 בנספחים).

<u>בחינת מחירי הדירות אל מול השטחים הירוקים:</u>

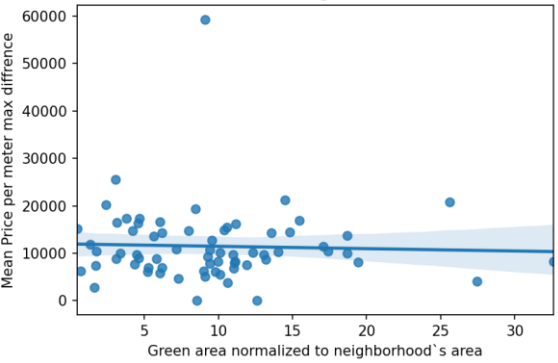
<u>המחיר הממוצע לדירה בשכונה לאורך כל התקופה</u> - לטובת בחינת המחיר הממוצע לדירה בשכונה יוצרה ראשית טבלה חדשה בעזרת איחוד מרחבי ובה עבור כל מכירה מצורפים הנתונים היבשים של השכונה בה בוצעה המכירה. טבלה זו סוכמה לידי טבלה חדשה אשר מכילה עבור כל שכונה, את המחיר הממוצע למטר לדירה בה. לבסוף, אוחדו נתונים אלו עם הטבלה אשר נוצרה בסוף החלק הקודם, אחוז השטחים הירוקים בשכונה על בסיס מספר השכונה. בכך נוצרה טבלה חדשה המכילה את הנתונים הגיאוגרפים של השכונה, המחיר הממוצע לדירה בה ואחוז השטחים הירוקים בה. המחיר הממוצע למטר בכל שכונה הוצג בצורה ויזואלית ורציפה על גבי מפת של העיר תל אביב (ראה גרף 7). לטובת בחינת הקשר בין הנתונים, אחוז השטחים הירוקים ומחיר ממוצע למטר הוצבו כל השכונות על פני גרף הכולל בתוכו קו רגרסיה לינארית (ראה גרף 8) וחושב מתאם פירסון בין השניים ל 0.22.

Graph 8: Mean Price vs Amount of Green Areas In Each Neighborhood



<u>השינוי הדרסטי ביותר בין השנים</u> – חישוב השפעת אחוז השטחים הירוקים בכל שכונה על שינוי המחירים בשכונה נעשה על ידי בחינת ההפרש בין המחיר השנתי הממוצע הגבוהה ביותר והנמוך ביותר עבור כל שכונה אל מול אחוזי השטחים הירוקים בשכונה זו. על כן, נעשה שימוש בטבלת המכירות עם נתוני השכונות מתחילת הסעיף הקודם וזו קובצה לכדי טבלה חדשה אשר מכילה את המחיר הממוצע למטר בכל שכונה בכל שנה. טבלה זו, קובצה לטבלה חדשה אשר בה עבור כל שכונה בוצע חישוב ההפרשים בין המחיר השנתי המקסימלי למינימלי. נתונים אלו, אוחדו גם כן עם הטבלה המכילה את אחוז השטחים הירוקים בשכונה על בסיס מספר השכונה. בכך, נוצרה טבלה חדשה המכילה את הנתונים הגיאוגרפים של השכונה, השינוי הדרסטי ביותר במחירים בה ואחוז השטחים הירוקים בה. לטובת בחינת הקשר בין הנתונים, אחוז השטחים הירוקים והשינוי הדרסטי ביותר במחירים הוצבו כל השכונות על פני גרף הכולל בתוכו קו רגרסיה לינארית (ראה ערך גרף 9) וחושב מתאם פירסון בין השניים ל 0.04-.

Graph 9: Changes in price over the years vs amount of green areas in each neighborhood

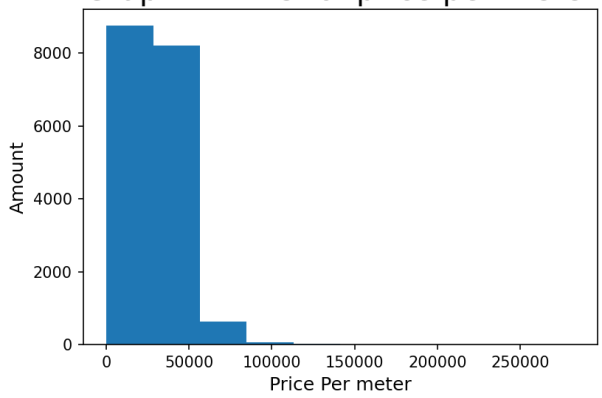


למידת מכונה

בחלקה האחרון של העבודה, ביצעתי למידת מכונה על מנת לבחון האם אחוז השטחים הירוקים בשכונה מסייעים לנו לחזות את מחירי הדירה, זאת על סמך נתונים נוספים אשר ידועים לנו על כל נכס: שנת המכירה, גודל הדירה, שנת בנייה ומספר הקומות בבניין. לטובת כך, ביצעתי שלוש למידות מכונה שונות וניסיתי לבחון האם ניתן ללמוד מאלו על הקשר בין אחוז השטחים הירוקים בשכונה למחירי הדירות בה. שלושת למידות המכונה אותם ביצעתי הן – למידת מכונה עם נתוני השטחים הירוקים אך ללא מספר הקומה בה נמצאת הדירה, למידת מכונה עם נתוני השטחים הירוקים ועם מספר הקומה בה נמצאת הדירה ולמידת מכונה עם מספר הקומה בה נמצאת הדירה אך ללא השטחים הירוקים.

לטובת מציאת מספר הקומה בה נמצאת הדירה היה צורך להפוך את מספר הקומה בה הדירה נמצאת מנתון אשר מוצג בעברית לנתון מספרי. לטובת כך, בודדתי את מספר הדירה מנתוני המכירות השכונה ובדקתי מה הם שמות הקומות המופיעות בתדירות הגבוהה ביותר. זאת, מכיוון שלבצע תרגום של השמות היה לא ריאלי במסגרת העבודה הנתונה. כצפוי, קומות -1 עד 10 היו אלו בהן התבצעו המכירות הרבות ביותר. על כן, בעזרת מילון והפונקציה "מפה" המרתי את קומות אלו לקומות בעלות ערך מספרי, בעוד שנתוני הקומה בשאר הדירה הפכו ל"ללא ערך". את נתונים אלו איחדתי חזרה לתוך הטבלה המקורית. הנתונים המעובדים נשמרו לקובץ csv חדש.

Graph 11: Hist of price per meter



בשלב הבא של הכנת הנתונים, יצרתי טבלה חדשה הכוללת אך ורק את הסדרות בהן יש לי צורך לטובת למידות המכונה כולל נתוני מחירי הדירות. לאחר מכן, בחנתי באמצעות גרף 10, את הטבלה החדשה ואת הנתונים החסרים בה המיוצגים בצהוב. מתוך נתונים אלו, השמטתי את הנתונים החסרים, מכיוון שאלו לא מלמדים אותנו דבר, לא ניתן לבצע למידת מכונה בעזרתם וקיימים כמות פריטים רבה, מעל 17 אלף גם בילדיהם. לבסוף, בחנתי את התפלגות נתוני מחירי הדירות (ראה גרף 11) והסרתי נתונים אשר המחירים בהם חריגים באופן משמעותי.

לאחר מכן, עברתי לשלב למידת המכונה עצמו, לטובת כל למידות המכונה ביצעתי הליך דומה אלא עם הנתונים הרלוונטיים ללמידת המכונה הספציפית. ראשית, הפרדתי בין נתוני מחירי הדירות לבין שאר הסדרות הרלוונטיות על ידי פיצולם לשני מאגרי נתונים נפרדים, איקס וואי. בהמשך, פיצלתי גם את מאגרי נתונים אלו, כאשר בכל אחד מהם כשבעים אחוזים מן הנתונים ישמשו לצורך למידת המכונה עצמה והיתר, לטובת בחינת הביצועים לאחר מכן. בשלב הבא, אימנתי את המודל על הנתונים אשר הקצתי לטובת כך. לבסוף, הפעלתי את המודל על נתוני המבחן ובדקתי את אמינות המודל. את אמינות המודל בחנתי במספר דרכים, בחנתי בצורה גרפית את הקורלציה בין נתוני האמת לתוצאות המודל (ראה ערך גרפים 12, 15 18 ובגרסתם הממוקדת גרפים 13,16 ו19), את ממוצע השגיאה המוחלטת, את ציון המודל ולבסוף את פיזור כמות השגיאה על פני גרף עמודות (ראה ערך גרפים 14,17 ו20).

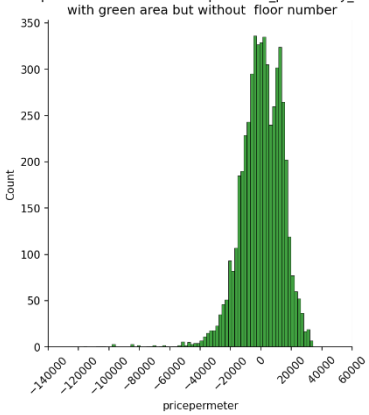
Graph 10: Null values by column

תוצאות ומסקנות

בבחינת תוצאות הנתונים, ניתן להבחין כי לא ניתן להצביע על קשר חזק בין אחוז השטחים הירוקים בשכונה למחירים בה. ראשית כבר בבחינת מחירי הדירות אל מול אחוזי השטחים הירוקים, נראה כי גרף מספר 8 ותוצאות מתאם פירסון (0.22) מצביעים על קשר חלש. הדבר נכון בבחינת מחיר הדירות הממוצע בשכונה לאורך כל התקופה אך מתחדד עוד יותר בבחינת השנוי במחירי הדירות. בבחינת קשר זה, צפינו בהעידר קשר באופן מובהק בגרף 9 ואף בנטייה קלה לקשר שלישי עם מתאם פירסון של 0.04-.

Graph 12: Prediction of price per meter with green area but without floor number

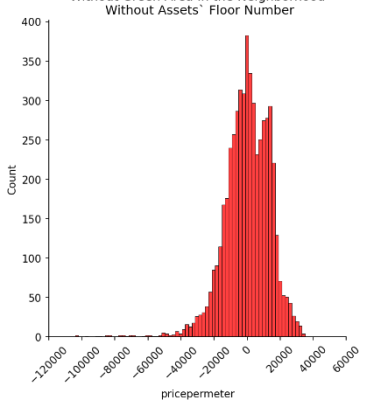
Graph 14: Diffrence between predicted_price and y_test with green area but without floor number



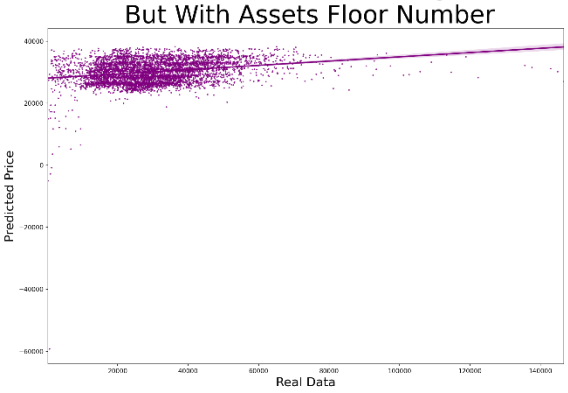
גם בבחינת תרומתו של נתון זה ללמידת המכונה ניתן להבחין כי יכולתו לסייע בחיזוי מחירי הדירות הינו מוגבלת עד אפסית. זאת, כאשר בלמידת מכונה אשר בוצעה בסיועו ממוצע השגיאה המוחלטת הינו 10638.33, וציון למידת המכונה הינו 0.075, נראה כי גרף הפיזור אינו אחיד (ראה גרף מספר 12) אך ניתן להבחין בעקומת פעמון מסוימת בגרף העמודות המצביע על השגיאה (ראה גרף מספר 14). זאת בעוד, למידת מכונה אשר בוצעה ללא מספר הקומה ולא השטחים הירוקים בשכונה הפיקה ממוצע שגיאות מוחלטות נמוך יותר, 10545.21, אך ציון למידת המכונה שלה הינו 0.067, כלומר פחות טוב, עם זאת, גרף הפיזור נראה יותר אחיד (ראה גרף מספר 18) וגרף השגיאה מפוזר יותר (ראה גרף מספר 20). לעומת זאת, למידת מכונה אשר הסתמכה על מספר הקומה בה הדירה נמצאת ללא השטחים הירוקים הפיקה ציון אף פחות טוב 0.066, אך עם ממוצע שגיאה מוחלטת קטן יותר 10595.98, גרף פיזור הינו יותר אחיד (ראה גרף מספר 15) וכך גם פיזור עקומת הפעמון (ראה גרף מספר 17). התמקדות לחלקים העיקרים בגרפי הפיזור מצרופים בנספחים כגרפים 13, 16 ו19 בהתאם לסדר אשר הוצג כאן.

Graph 18: Prediction of Price Per meter Without Green Area In the Neighborhood Without Assets` Floor Number

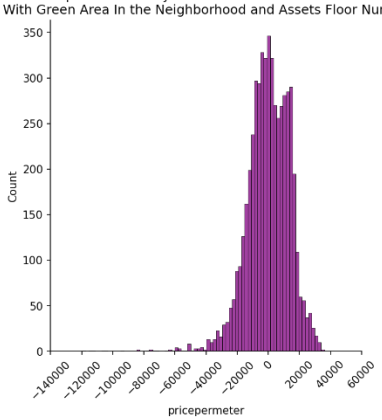
Graph 20: Diffrence Between Predicted Price and y_test Without Green Area In the Neighborhood Without Assets` Floor Number



Graph 15: Prediction of Price Per Meter Without Green Area In the Neighborhood But With Assets Floor Number



Graph 17: Accuracy For Price Per Meter Prediction With Green Area In the Neighborhood and Assets Floor Number



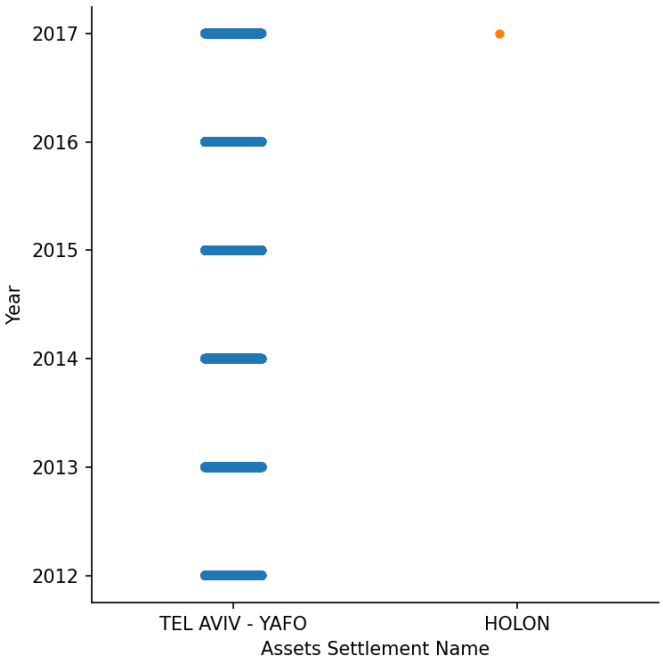
נתונים אלו יכולים לרמז כי לאחוז השטחים הירוקים בשכונה השפעה מסוימת על מחירי הדירות בשכונה זאת בשל ציון למידת המכונה אך גם היא מעטה מאוד. יתר על כן, גם אלו יכולים להיות מושפעים גם מהשכונה עצמה בה ממוקמת הדירה ומכך שאחוז השטחים הירוקים נהיה מעין מספר מזהה של השכונה ולאו דווקא יוצגו כמספר רציף המייצג את אחוז השטחים הירוקים בשכונה.

סיכום

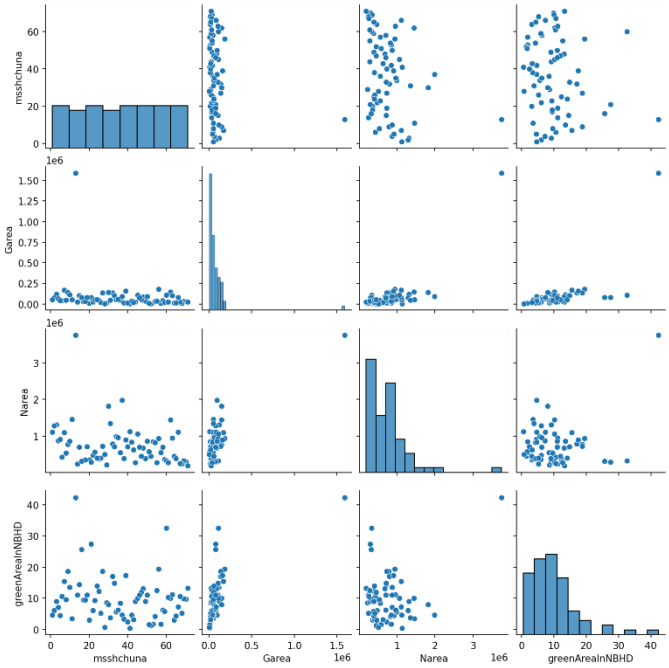
בעבודה זו בחנתי את מידת ההשפעה של אחוז השטחים הירוקים בשכונה על מחירי הדירות בשכונה וקצב תנודתם. תוצאות העבודה סתרו את השערתי הראשונית, כי לאחוז השטחים הירוקים בשכונה תהיה השפעה חיובית ומשמעתית על מחירי הדירות בשכונה ואף תוביל לעלייה חדה יותר באלו בשנים אשר בחנתי. ניכר כי המתאם בין השינוי במחירי הדירות בתקופה שנבחנה לבין השטחים הירוקים בשכונה הינו חלש ואף שלילי. באשר, למחירי הדירות עצמם נראה מנתוני למידת המכונה ומבחן פירסון כי ישנו קשר חלש וחיובי כלשהו בין שני הנתונים אך אין הדבר בהכרח מצביע על נסיבתיות או על כיוון הקשר.

נספחים - גרפים שניים בחשיבותם, אשר להוספתם לא הייתה חשיבות מהותית להבנת סיכום העבודה.

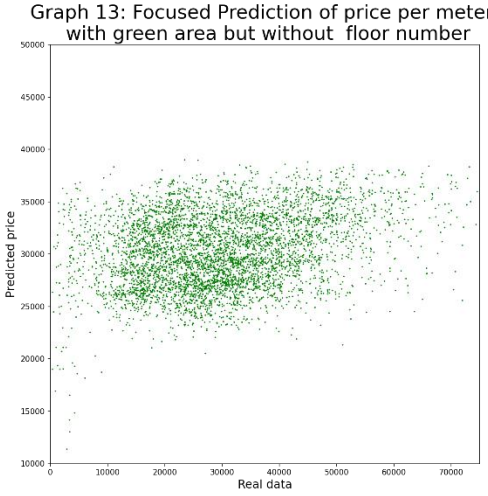
Graph 2: Records From Each Settlement Each Year



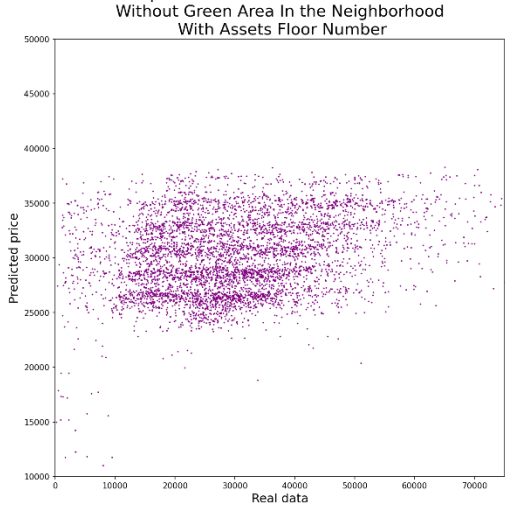
Graph 6: Green and Neighborhood Columns as Pairplot



Graph 13: Focused Prediction of price per meter with green area but without floor number



Graph 16: Prediction of Price Per Meter Without Green Area In the Neighborhood With Assets Floor Number



Graph 19: Focoused Prediction of Price Per Meter Without Green Area in the Neighborhood Without Assets` Floor Number