חלק מעשי

3.1.3

א. בחרתי להוריד את עמודות המיקום (‘lat’,’long’), כיוון שמניתוח פשוט של הdata אנחנו לא יכולים להסיק משהן מעניין מהמיקום של הבית (בניתוח מעמיק יותר ניתן לחלק לאזורים שנחשבים יוקרתיים יותר או במיקום אטרקטיבי יותר ואז למיקום כן יהיה משמעות, אך חשבתי שזה מעבר לסדר גודל של הניתוח הנדרש לתרגיל). כמו כן הורדתי את עמודות הId[[1]](#footnote-1) והתאריך[[2]](#footnote-2).

כמו כן, הורדתי את עמודת שנת השיפוץ (‘yr\_renovated’). הסיבה היא כי במעבר על הdata הבחנתי שישנם הרבה בתים שהגודל שלהם השתנה (פער בין sqft\_lot וsqft\_lot15) למרות שלא צוין שנת השיפוץ. לכן הסקתי שכנראה המידע בעמודה זו לא מספיק מהימן והחלטתי להוריד אותו (לחילופין, ניתו היה ליצור עמודה של שופץ או לא שופץ לפי הפער הזה, אך מבחינתי קיבלתי את המידע הנדרש מגודל הבית והסתפקתי בו וזאת על אף שבשיפוץ כנראה ישנם עוד אלמנטים של הבית שעוברים שדרוג, שלא מבוטאים במידע שיש לנו כאן).

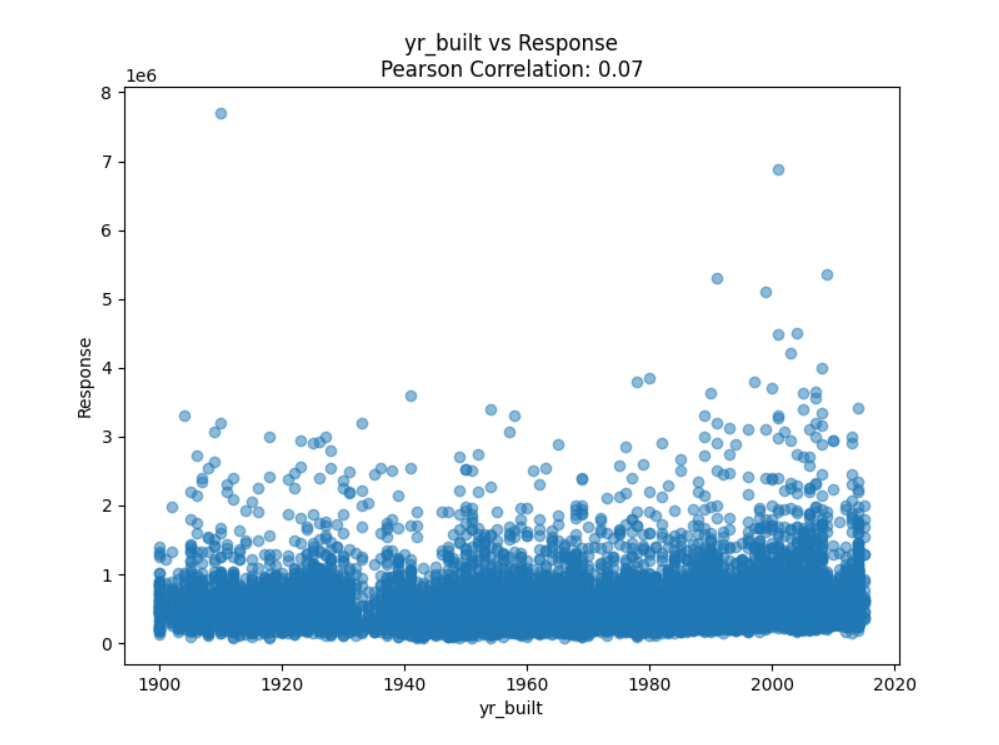
ב. ע"פ ההגדרה ששדה קטגורי הינו שדה שהערכים בו הם לפי קטגוריה ולא מספר, כל השדות הם קטגוריים. בהמשך אתייחס לעמודות שהוספתי, ושם נדבר על שדות חדשים שנוצרו ושבהם הקטגוריה יכלה להיות לא מספרית. כן ראוי לציין שהתאריך יכל להיות מוגדר כקטגורי, אך הסרנו אותו, כמצוין לעיל.

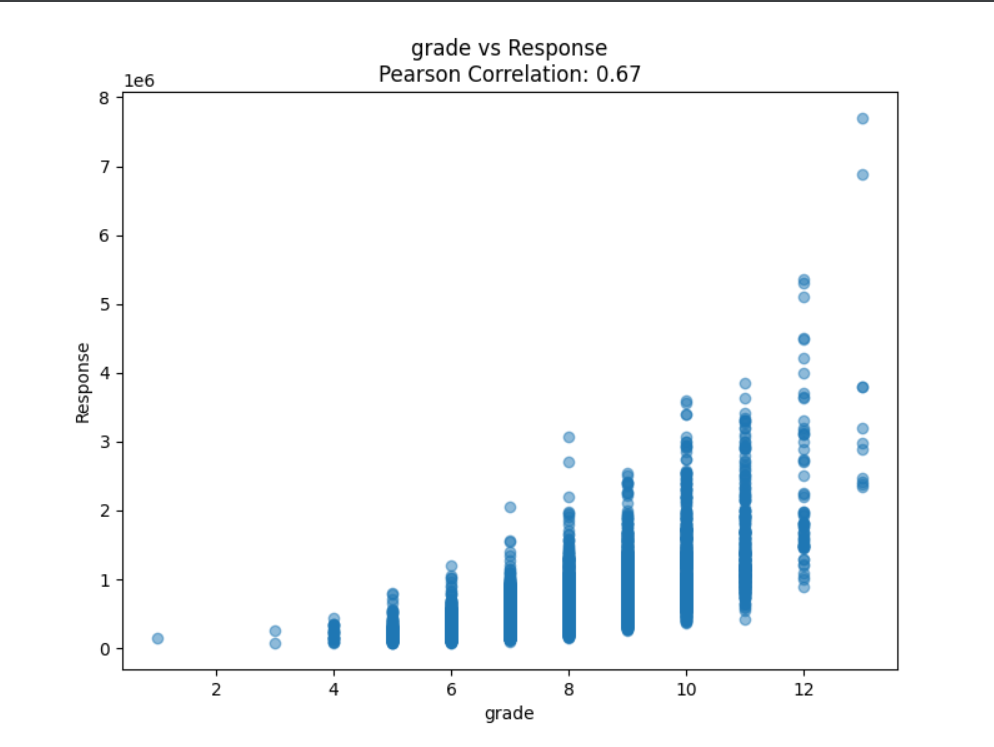
ג. הוספתי עמודות של האם הדירה נצפתה מראש (‘has\_been\_viewed’) והאם יש מרתף (‘has\_basment’). לגבי השני, בפועל הקורלציה הייתה הרבה פחות טובה, מה שמרמז שאכן יש משמעות גדולה יותר לגודל המרתף ממה שציפיתי. לגבי הראשון, קיבלנו קורלציה מאוד דומה, ולכן נראה שזו אכן הייתה החלטה נכונה. העמודה החדשה היא הרבה יותר פשוטה ולא מאבדת הרבה מידע, כפי שהיה ניתן לחזות שאין כל כך הבדל בין לראות את הבית פעם אחת לפעמיים. אציין, שמדובר דווקא על **הוספה** של עמודות ולא הסרה של העמודות הקיימות.

ד. הסרתי את כל השורות עם ערכי זבל nan וערכים שליליים (אין אף עמודה מבין אלה שלא הוסרו שמתאים לה ערך שלילי). כמו כן, הסרתי שורות שהיו בעיות בגודל הבית (ששטח המגורים קטן מהשטח הכולל של הבית, או בגדלים המקוריים או באלה של 2015). בנוסף, כאשר היו 0 חדרים או 0 חדרי שירותים, שמתי את הערך הממוצע של חדרים או חדרי שירותים בהתאמה[[3]](#footnote-3).

כאן ראוי לציין גם מדוע הסכמתי לקבל ערכים לא שלמים במספר חדרי השירותים. מחיפוש באינטרנט, גיליתי שקיימות כמה הגדרות וסוגים של חדרי שירותים. יש חדר "מלא" שכולל מקלחת, שירותים וכיור, יש חלקי שכולל אחד מהם ויש כזה שיש בו רק דבר אחד (נגיד רק שירותים ללא כיור), ולכן מספר לא שלם הוא ערך חוקי למס' חדרי השירותים.

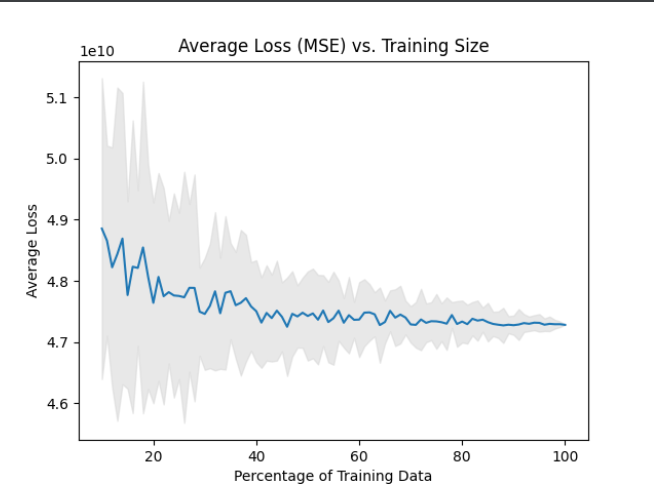
3.1.4

 ניתן לראות ששנת הבנייה של הבית לא כל כך משנה למחיר. ניתן לראות זאת לפי ערך הקורלציה הנמוך (קרוב ל0). ניתן להסביר זאת בכך שעל ידי שיפוצים למיניהם ניתן לשפר ואף לשדרג גם בתים מאוד ישנים ולכן שנת הבנייה לא כל כך משפיעה).

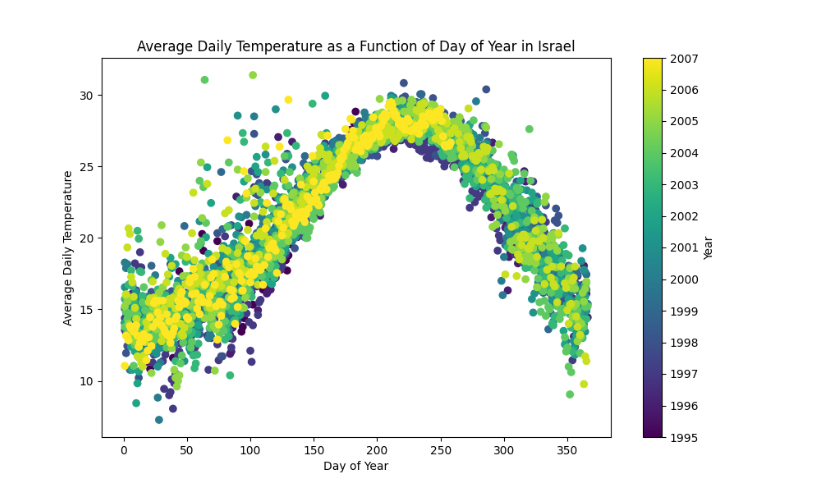


מצד שני, ניתן לראות את מגמת העלייה הברורה במחיר ככל שהציון שניתן לבית גבוה יותר (מתבטא בערך הקורלציה היחסית גבוה). לכן ניתן לקבוע שיש קשר גדול בין הציון שהבית מקבל למחיר ששולם עליו (וזה אכן סביר, כי זהו המטרה של הציון...).

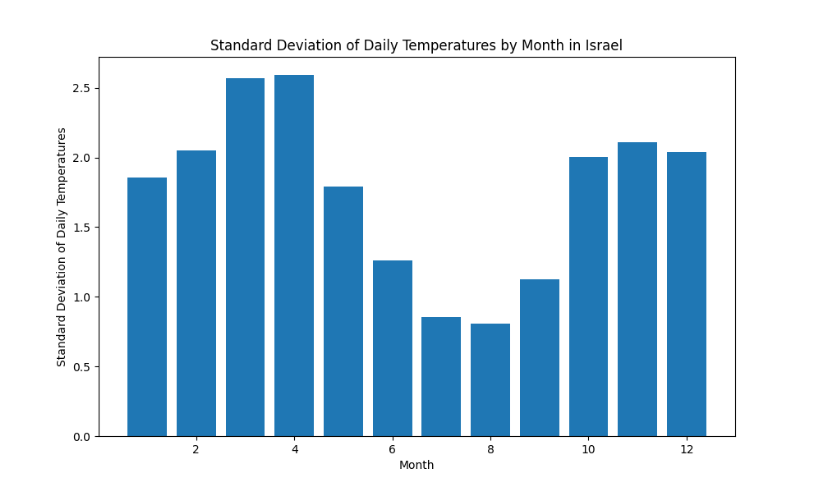
3.1.5

ניתן לראות שככל שכמות הData שעליו המודל מותאם גדל, הLoss V הממוצע קטן, כלומר המודל מדויק יותר, ובנוסף גם השונות בפרדיצקיה קטן, כלומר סה"כ מתקבעים על תוצאה שהיא גם יותר מדויקת.

3.2.3



ניתן לראות (בעיקר בצד שמאל של הגרף), שאין דעיכה בטמפרטורה אלא יש התייצבות ואז מחזוריות בין השנים, כיוון שכפי שמשתקף מהגרף אין באמת הבדלים משמעותיים לאורך השנים, לא בגובה הטמפרטורות ולא בצורת השינויים. זה מתאים למה שאנחנו מכירים בעולם האמיתי על שיא החורף ושיא הקיץ ושינוי הטמפרטורות ביניהם. לכן נבחר מעלה מתאימה כמו k=3. סה"כ אנחנו מקבלים שהטמפרטורות נעות בין 10-30 מעלות (מתאים למוכר לנו), ושהפיקים הם באזור יוני-אוגוסט בקיץ ובדצמבר-פברואר בחורף, שזה גם מתאים למה שאנחנו מכירים.



כיוון שיש שינוי והבדלים בין השונות בין החודשים השונים, ניתן לצפות שבחודשים עם שונות נמוכה (6-9) המודל המותאם יחזה באופן יותר מדויק מאשר בחודשים עם השונות הגבוהה (כגון 3-4, שם השונות היא מקסימלית).

3.2.4

1. כי אין משמעות למס' הזיהוי של המכירה לעניין המחיר. [↑](#footnote-ref-1)
2. לא חשבתי שיש צורך להעמיק בהשפעת תקופות על תאריכי המכירה, ולכן המכירה מבחינתנו היא "אינווריאנטית בזמן", כלומר חא מושפעת מהזמן בו היא התבצעה. [↑](#footnote-ref-2)
3. במודל מורכב יותר היה צריך לקחת בחשבון מחיר וגודל וכו', אך לא התייחסתי לכך. [↑](#footnote-ref-3)