**פרוייקט רגרסיה חלק ב**

1.     תקציר מנהלים:

מטרת הפרוייקט הינה בניית מודל רגרסייה לינארית מרובה באמצעות ניתוח מאגר נתונים ושיפור המודל באמצעות הכלים שנלמדו בקורס.

המודל שלנו יבחן את המחיר של דירה כתלות במשתנים הבאים: כגון

**1.** 

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **סוג המשתנה -מוסבר/מסביר** | **סימון במודל** | **יחידת מידה** | **סוג המשתנה – רציף / קטגוריאלי** | **הסבר קצר על המשתנה** |
| **משתנה מוסבר**  **מחיר** | **Y** | **שקל חדש** | **רציף** | **מחיר הנכס בשקלים חדשים** |
| **משתנה מסביר**  **מספר חדרי שינה** | **X2** | **חסר יחידות** | **קטגוריאלי** | **מספר חדרי שינה בנכס, המשתנה הוא קטגוריאלי כיוון שהטווח של המשתנה מצומצם בין ערכים תחומים.** |
| **משתנה מסביר**  **גדול הבית** | **X3** | **מ"ר** | **רציף** | **גודל השטח הבנוי של הנכס** |
| **משתנה מסביר**  **גודל המגרש** | **X4** | **מ"ר** | **רציף** | **גודל השטח הכולל של הנכס** |
| **משתנה מסביר**  **גודל המרתף** | **X5** | **מ"ר** | **רציף** | **גודל המרתף בנכס,0 אם אין** |
| **משתנה מסביר**  **מספר קומות בנכס** | **X6** | **חסר יחידות** | **קטגוריאלי** | **מציין את מספר הקומות שהנכס מכיל (מתחיל מ1)** |
| **משתנה מסביר**  **מצב הנכס** | **X7** | **[1-5]** | **קטגוריאלי** | **מצב הנכס 1 – לא טוב 5- טוב מאוד. המשתנה הוא קטגוריאלי כיוון שהטווח של המשתנה מצומצם בין ערכים תחומים.** |
| **משתנה מסביר**  **השנה בה הנכס נבנה** | **X8** | **שנים** | **רציף** | **מציין את השנה שבה נבנה הנכס, המשתנה הוא רציף כיוון שיש מספר רב של תצפיות והטווח של השנים יכול להכיל מספר רב של ערכים** |
| **משתנה מסביר**  **האם הנכס שופץ** | **X9** | **[0-1]** | **קטגוריאלי** | **האם הנכס עבר שיפוץ 0 אם לא 1 אם כן** |

**2) עיבוד מקדים:**

**2.1) הסרה של משתנים:**

-בכדי לבדוק איזה משתנים נרצה להסיר, נבדוק בעזרת **מתאם פירסון, מבחן t ותרשים פיזור.**

-**מתאם פירסון** הינו מדד לבדיקה קשר לינארי בין שני משתנים כמותיים רציפים, לא ניתן לבצע מתאם פירסון על משתנים קטגוריאליים מפני שאין משמעות לערכים של הקטגוריה.

- **מבחן t** עוזר לנו להכריע בין השערת האפס(בטא = 0 ) לבין האלטרנטיבה, כאשר נעשה מבחן על בטא כלשהו המייצגת את השיפוע של Xi. נבחן את P-val , במידה ויהיה קטן מ-0.05 נדחה את השערת האפס, ונאמר שהשיפוע של בטא שונה מאפס וקיים קשר בין המשתנים הנבדקים.

-**תרשים פיזור** ניתן לבדוק ויזואלית האם קיים קשר בין המשתנה המסביר למוסבר אך לא לקבוע בוודאות ללא מבחן.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| X3 | X4 | X5 | X8 |  |
| 0.677448 | 0.4038051 | 0.3271691 | (0.06532586)- | Person correlation |
| 2.364e-16 | 1.01e-05 | 0.0004299 | 0.04938 | P-val |

נתבונן בתוצאות שקיבלנו, ניתן לראות שלא נסיר אף משתנה מסביר לפי מתאם פירסון. לפי מובהקות התוצאה נשקול להסיר את X8 מפני שתוצאת P-val שואף ל-5% אבל קטן ממנו, הP-val של שאר המשתנים המסבירים קטנים ממש מ-5%. בנוסף נניח כי קיים קשר סיבתי בין כל משתנה מסביר למוסבר על אף שגודל המדגם הינו 113 (לאחר הסרת תצפיות קיצוניות) לכן נאמר, ייתכן ואין מספיק מידע כדי לסתור את ההנחה כי הקשר שאנחנו חוזים אכן לא מתקיים.

כדי לבדוק איזה משתנים **קטגוריאלי** נסיר מהמודל נשתמש במבחן F חלקי בשיטת השמטת משתנים. **אבל עדיין לא נסיר מפני שנרצה לבדוק האם ניתן להוסיף משתנה דמה או משתנה אינטראקציה אשר ישפר את המודל.**

**מבחן F חלקי** - **השערת האפס**: אין הבדל משמעותי בין המודל החלקי למודל השלם.

**אלטרנטיבה**: המודל השלם יותר טוב בצורה משמעותית.

**מספר חדרי שינה X2:**

עבור משתנה קטגוריאלי X2 מספר חדרי שינה, הגדרנו את מודל שלם אשר כולל את כל המשתנים ומודל חלקי ללא מספר חדרי שינה. ניתן לראות ש- P-val =0.9348 במבחן F חלקי לכן לא נדחה את השערת האפס בר"מ 5% ולפי המבחן נסיר את מספר "חדרי שינה" מהמודל.

Graphical user interface, text, application

Description automatically generated

**מספר קומות בנכס X6:**

עבור משתנה קטגוריאלי X6 מספר קומות בנכס, הגדרנו את המודל השלם אשר כולל את כל המשתנים למעט מספר חדרי שינה ומודל חלקי ללא מספר חדרי שינה ומספר קומות. ניתן לראות ש- P-val =0.1317 במבחן F חלקי לכן לא נדחה את השערת האפס בר"מ 5%. על אף התוצאות נחליט להשאיר את המשתנה הנ"ל מתוך הנחה כי כן קיים קשר סיבתי בין מספר קומות בנכס לבין המחיר. בנוסף במדגם שלנו מכיל 113 תצפיות בלבד לכן נוכל לומר כי ייתכן ואין מספיק מידע בכדי לסתור את ההנחה כי הקשר שאנחנו חוזים אכן לא מתקיים.

Text

Description automatically generated

**מצב הנכס X7:**

עבור משתנה קטגוריאלי X7 מספר מצב הנכס, הגדרנו את המודל השלם אשר כולל את כל המשתנים למעט מספר חדרי שינה ומספר קומות ומודל חלקי ללא מספר חדרי שינה, מספר קומות ומצב הנכס. ניתן לראות ש- P-val =0.1317 במבחן F חלקי לכן לא נדחה את השערת האפס בר"מ 5% ונסיר את "מצב הנכס" מהמודל.

Graphical user interface, text, application

Description automatically generated

**האם הנכס שופץ X9:**

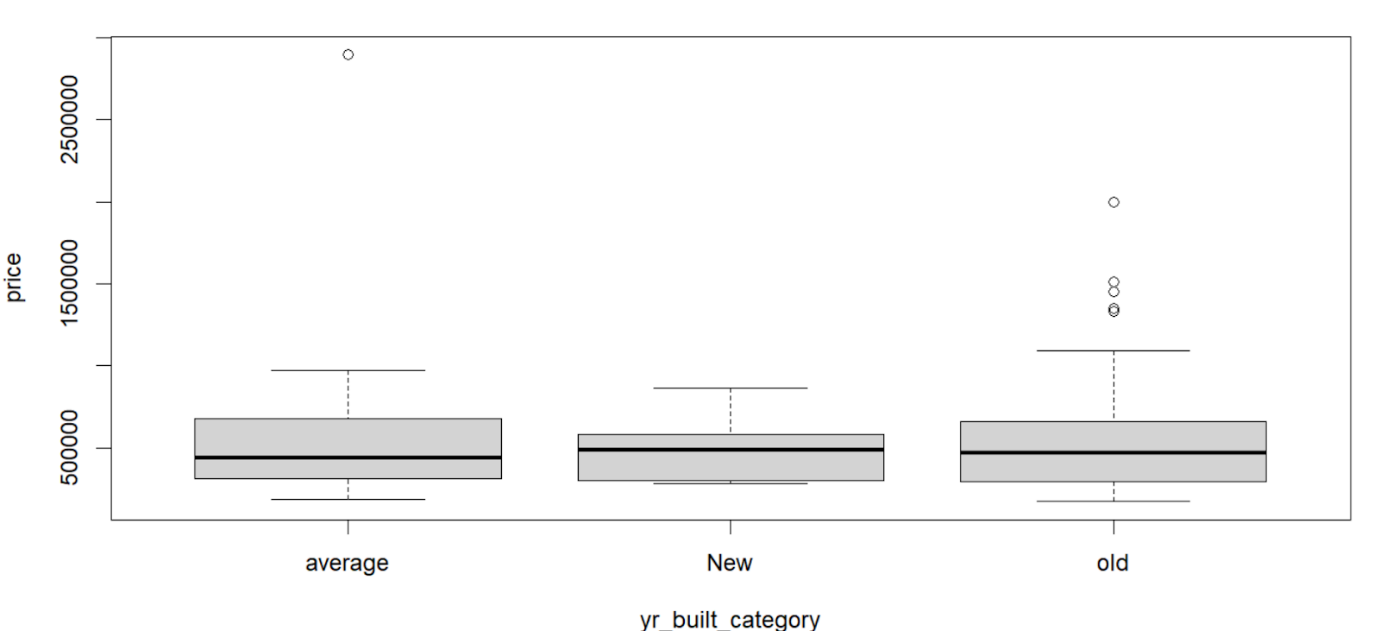
עבור משתנה קטגוריאלי X9 האם הנכס שופץ, הגדרנו את המודל השלם אשר כולל את כל המשתנים למעט מספר חדרי שינה, מספר קומות ומצב הנכס ומודל חלקי ללא מספר חדרי שינה, מספר קומות, מצב הנכס והאם הנכס שופץ. מספר קומות ומצב הנכס. ניתן לראות ש- P-val =0.02375 במבחן F חלקי לכן נדחה את השערת האפס בר"מ 5% ולא נסיר את "האם הנכס שופץ".

Graphical user interface, text

Description automatically generated

בחרנו בצורה שרירותית את שיטת **Backward Elimination**, לא נרצה לקבל **Overfitting** ל כן לא השתמשנו בשיטת **Stepwise Selection** ניתן לראות שקיבלנו מודל חדש ללא מספר חדרי שינה, מספר קומות ומצב הנכס.

**2.2)התאמת משתנים:**

נבדוק הפיכת המשתנה "**השנה בה הנכס נבנה" (X8)** ממשתנה רציף למשתנה קטגוריאלי. מדובר במשתנה רציף אך רצינו ליצור עבורו טווח ערכים מצומצם יותר לפי קטגוריות.  
הקטגוריות הינן: נכס ישן - עד שנת  1980 לא כולל, נכס בטווח שנים ממוצע - 2000-1980, חדש - 2000 והילך.  
כעת נבדוק האם המשתנה הקטגוריאלי ישפיע על המשתנה המוסבר (price) באמצעות תרשים קופסא.  
בתרשים ניתן לראות כי עבור כל קטגוריה המחירים בערך באותו טווח ערכים מבחינת המחיר, אילו המשתנה היה מייצג ערכים משתנים היינו יכולים להגיד שיש לו ערך כלשהו ביחס למחיר הדירה כי לכל טווח יש משמעות במחיר, אולם פה זה לא קורה ולכן סביר שמשתנה זה לא משפיע על המשתנה המוסבר בחלוקה הנוכחית ונישאר טווח ערכים רציף.  
  


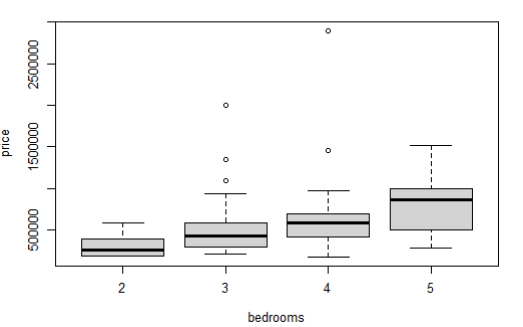
**2.3) הגדרת משתני הדמה:**

נגדיר משתנה דמה עבור משתנים קטגוריאליים במודל שלנו:   
לפי ההנחות וסט הנתונים שלנו הגדרנו טווח ערכים מצומצם עבור כל משתנה שהגדנו אותו קטגוריאלי.  
כעת נרצה לבחון האם באמת מדובר במשתנה קטגוריאלי מתוך הנתונים או לא. לשם כך, נבדוק את הלינאריות של המשתנים ביחס למשתנה המוסבר, האם הB^ של המשתנה מסביר משהו ואת מדד ה R^2 adj.

**X2- מספר חדרים:**

למשתנה ישנם 5 ערכים (הגבלנו אותו בהוצאת החריגים).

מכיוון שמדובר במשתנה בעל ערכים מספריים נבדוק האם לערכים יש משמעות כתלות במשתנה המוסבר (המחיר). לשם כך נבדוק את הקשר הלינארי בין מספר החדרים (X2) לבין המחיר (Y).



ניתן לראות קשר לינארי כלשהו וקשר כזה יכול להעיד על כך שכדאי שמשתנה זה יהיה משתנה רציף (אחרת נאבד .  
יצרנו מודל עם המשתנים והסתכלנו על הקשר בשיפוע בין המשתנה המוסבר לכל אחד מהמשתנים הקטגוריאליים. חיפשנו לראות P.V קטן בשיפוע כך שיעיד על קשר גדול בין המשתנה המוסבר למסביר.  
בנוסף, הסתכלנו על מדד טיב ההתאמה (R^2 adj) לראות איפה הוא גדול יותר.

עבור מודל רגרסיה עם משתנה מוסבר ומשתנה רציף של מספר החדרים קיבלנו:

תמונה שמכילה טקסט

התיאור נוצר באופן אוטומטי

ניתן לראות כיPV של השיפוע הוא קטן אולם ה R^2 adjדווקא גם די קטן.

כעת נסתכל על הנתונים עם נשנה את המודל להיות בעל משתני דמה.

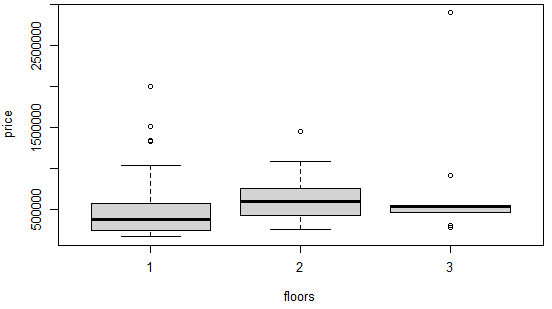
תמונה שמכילה טקסט

התיאור נוצר באופן אוטומטי

ניתן לראות כי ה R^2 adjדווקא ירד אז נסיק מכך שמודל בעל משתנה דמה הוא פחות מתאים וניקח את המשתנים הללו להיות רציפים.  
לכן לא ניצור עבורו משתני דמה.

**X6 -מספר קומות בנכס:**

מכיוון שמדובר במשתנה בעל ערכים מספריים נבדוק האם לערכים יש משמעות כתלות במשתנה המוסבר (המחיר). לשם כך נבדוק את הקשר הלינארי בין מספר הקומות (X6) לבין המחיר (Y).

****

מהתרשים קשה לשער כי אין קשר ליניארי בין מספר הקומות לבין המחיר של הנכס (המשתנה המוסבר).  
נבדוק זאת מבחינת המדדים:   
1. עבור מודל רגרסיה עם משתנה מוסבר ומשתנה רציף של מספר הקומות קיבלנו:

תמונה שמכילה טקסט

התיאור נוצר באופן אוטומטי

ניתן לראות כי PV של השיפוע יחסית קטן אולם מבחינת מדד טיב ההתאמה (R^2 adj) הוא יחסית קטן

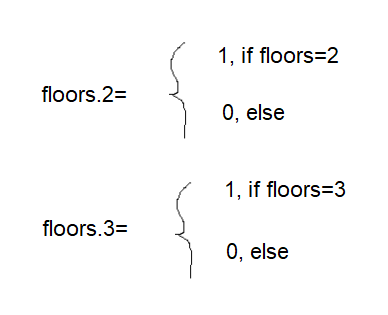
2. עבור מודל רגרסיה עם משתנה דמה קיבלנו:

תמונה שמכילה טקסט, קבלה

התיאור נוצר באופן אוטומטי

ניתן לראות כי מדד טיב ההתאמה (R^2 adj) ירד במעט כאשר הפכנו את המודל לקטגוריאלי, אולם מדובר על פחות מאחוז ולדעתנו הגיוני יותר לעשות פה משתנה קטגוריאלי מבחינה הגיונית. לכן בכל זאת נגדיר את המשתנה הזה להיות קטגוריאלי:

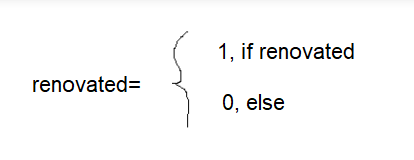
לגבי מספר הקומות בנכס, ניתן לראות מהנתונים כי הטווח ערכים בין 1-3  
משום כך נעשה משתנה אינטרגציה שבו 1 הוא משתנה הבסיס ולכן נגדיר 2 משתנה דמה



**X7 -מצב הנכס:**

בחרנו להסיר מהמודל

**X9 -האם הנכס שופץ:**

לגבי משתנה זה, הוא כבר בטווח ערכים 0-1 ומשום כך ברור כי הוא בעצמו מוגדר כבר כמשתנה דמה.  


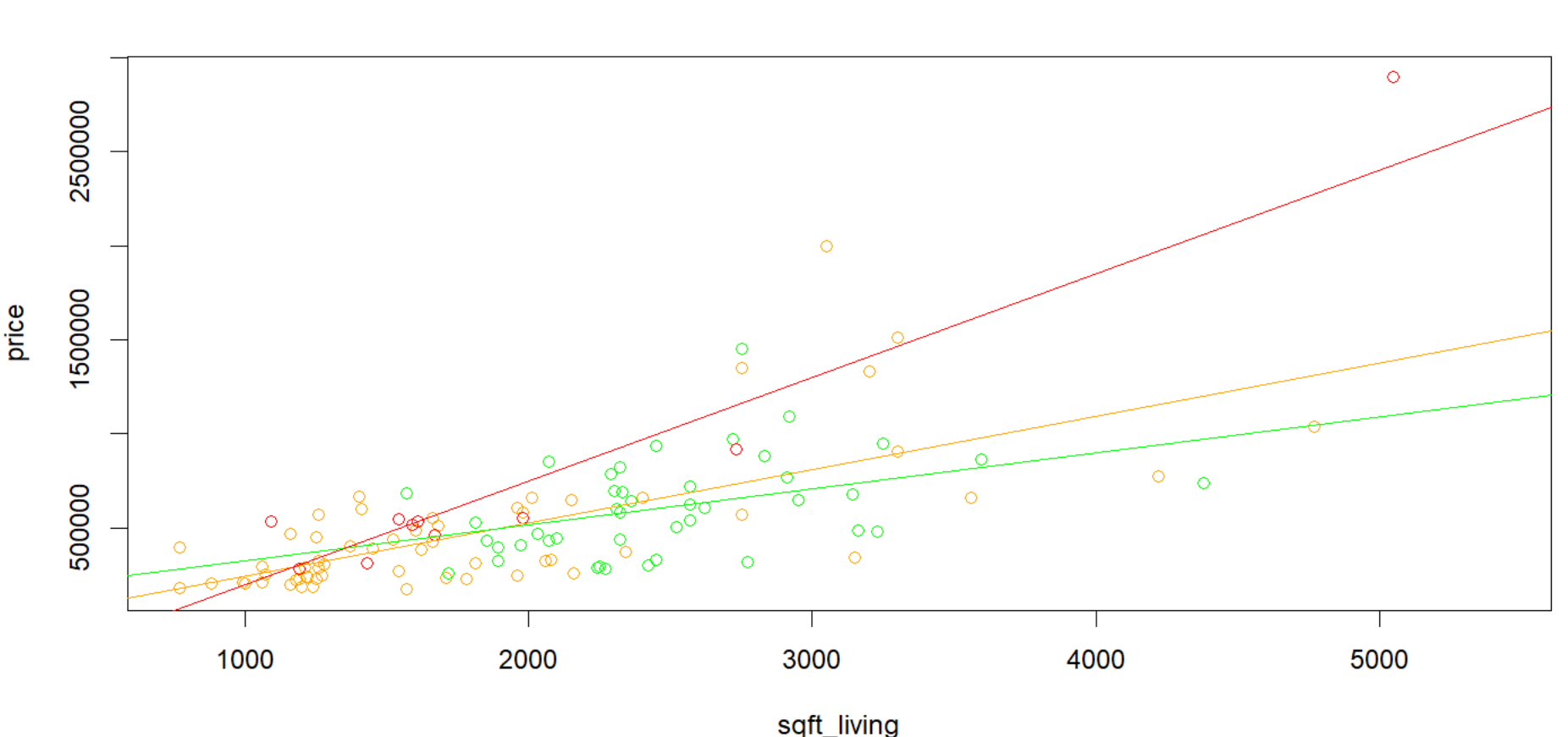
לכן ניתן לראות כי משתנה הבסיס מוגדר להיות נכס שלא שופץ.

**2.4)  הגדירה והוסיפה של משתני אינטראקציה:**

**משתנה אינטראקציה בין מספר הקומות (X6) לבין גודל הבית (X3)-**

בחרנו לבחון את האינטרקציה בין שני המשתנים הללו מכיוון שסביר שגודל הבית ישפיע על מספר הקומות.  
לפי ההיגון ככל שיש יותר קומות כך הבית גדול יותר

לשם כך יצרנו תרשים שמתאר את הקשר בין כל ערך של מספר הקומות לבין המחיר (המשתנה המוסבר).  
בתרשים הבא הנקודות לפי הצבעים הן: כתום=1, ירוק= 2, אדום= 3.  
ניתן לראות כי קיים הבדל בשיפועים בין 3 קומות לבין 1 ו2.   
נבחן הבדל זה גם באמצעות ערכי הPV: ניתן לראות כי עבור משתנה 3 קומות הערכים אכן מובהקים (0.007, 1.93 eבהתאמה) אבל עבור 2 ו1 לא (מבחינת PV של השיפועים ושל החותך).  
בגלל שכן קיים איזה הבדל בחרנו כן להוסיף את משתנה האינטרקציה הנל למודל.  
המשמעות של הנתונים היא כי יש הבדל בין בתים עם 3 קומות לבין בתים עם 2 ו1 קומות מבחינת גודל הבית ומחירו. ההתנהגות של בתים עם 3 קומות שונה משל בתים עם 2 ו1 קומות.

****

**תמונה שמכילה טקסט

התיאור נוצר באופן אוטומטי**

הוספת המשתנה:

floors\_Deme∗ sqft\_living= { sqft\_living 𝑖𝑓 , floors\_Deme=1

0 𝑒𝑙𝑠𝑒

floors\_Deme2∗ sqft\_living= { sqft\_living 𝑖𝑓 , floors\_Deme=2

0 𝑒𝑙𝑠𝑒

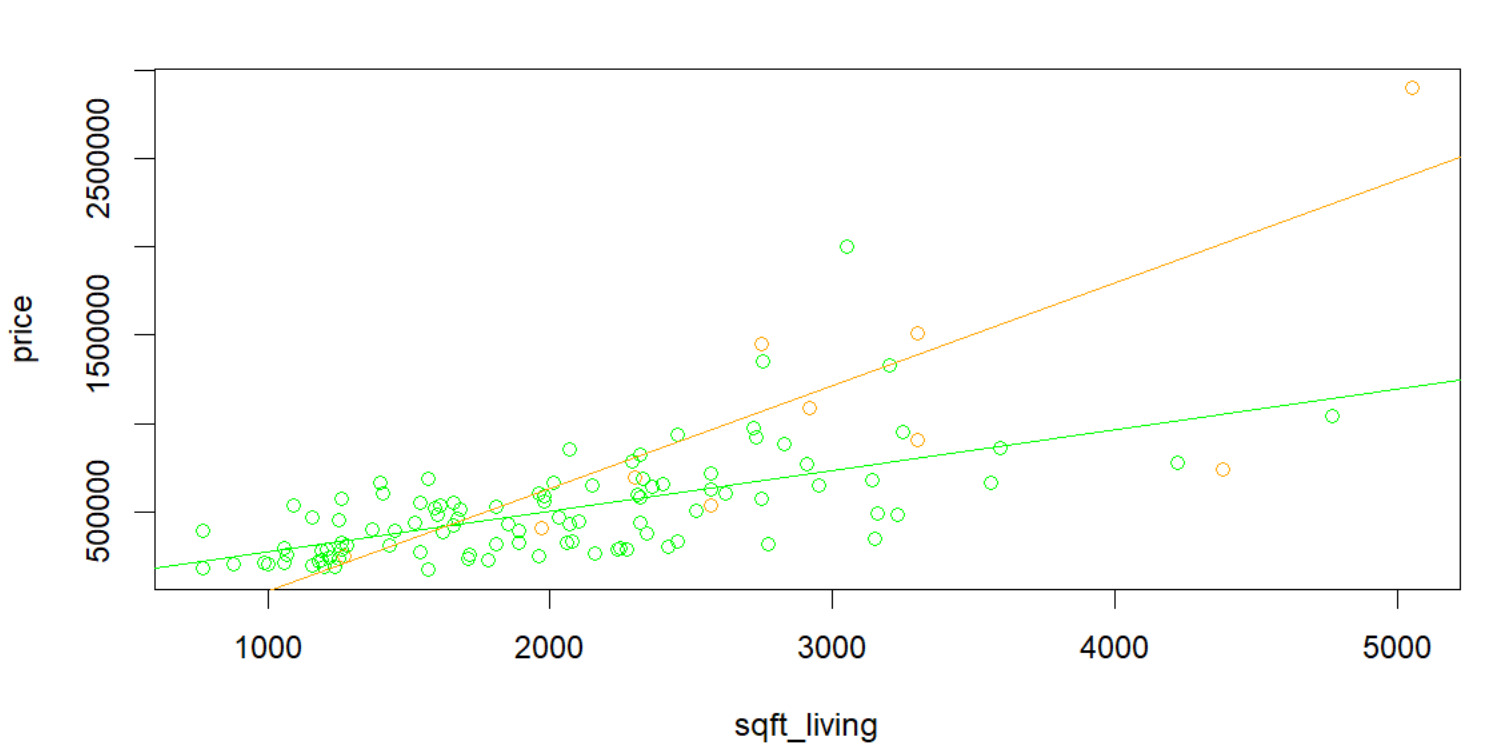
floors\_Deme3∗ sqft\_living= { sqft\_living 𝑖𝑓 , floors\_Deme=3

0 𝑒𝑙𝑠𝑒

**משתנה אינטראקציה בין האם הנכס שופץ (X9) לבין גודל הבית (X3)-**

בחרנו לבדוק משתנה זה מכיוון שרצינו לבדוק אם אנשים שיכולים להרשות לעצמם לשפץ את הבית אולי יכולים להרשות לעצמם גם בית גדול ואז יהיה קשר בין שני המשתנים הללו.

לשם כך יצרנו תרשים שמתאר את הקשר בין השיפוץ לבין גודל הבית.  
בתרשים הבא הנקודות לפי הצבעים הן: כתום=1, ירוק= 0.  
מכאן ניתן לראות כי יש הבדלים בשיפוע ובגרף, לכן נראה לגבי המובהקות של התוצאות.  
רואים כי השיפוע והחותך (0.02,0.006 בהתאמה) של משתנה האינטראקציה אכן מובהק ולכן נבחר להוסיף אותו, כמו שאני מצפים שיקרה לפי הגרף.



**תמונה שמכילה טקסט

התיאור נוצר באופן אוטומטי**

הוספת המשתנה:  
renovated∗ sqft\_living= { sqft\_living 𝑖𝑓 , renovated=0

0 𝑒𝑙𝑠𝑒

renovated∗ sqft\_living= { sqft\_living 𝑖𝑓 , renovated=1

0 𝑒𝑙𝑠𝑒

**משתנה אינטראקציה בין האם הנכס שופץ (X9) לבין השנה בה הנכס נבנה (X8)-**

בחרנו לבדוק משתנה זה מכיוון ייתכן כי השנה שבה נבנה הכנס קשורה לבלאי של הנכס ואז הנכס יצטרך שיפוץ אם הוא בלוי. כלומר, חשבנו על סברה שאם הנכס ישן הסיכוי שהוא יעבור שיפוץ ייתכן ויגדל ורצינו לבדוק את הקשר לכך.

**1.3) בחירת משתני המודל:**

בחרנו לבדוק את משתני המודל לפי 3 אלגוריתמים שנלמדו בכיתה: רגרסיה לאחור, רגרסיה לפנים, רגרסיה בצעדים ולפי שלושה מדדים: AIC,BIC ו R^2 adj.  
המשתנים הפוטנציאלים להיכלל במודל הם: bedrooms, sqft\_living, sqft\_lot, sqft\_basement, floors\_Deme, renovated, yr\_built, sqft\_livingRenovated

עבור מדד AIC נרצה את הערך המתקבל הכי נמוך

**רגרסיה לפנים לפי AIC:**

המודל המתקבל הינו:

ניתן לראות כי המשתנים שהוכנסו למודל הם כל המשתנים למעט המשתנה של מספר החדרים בבית

תמונה שמכילה טקסט

התיאור נוצר באופן אוטומטי

**רגרסיה לאחור לפי AIC:**

המודל המתקבל הינו:

תמונה שמכילה טקסט

התיאור נוצר באופן אוטומטי

**רגרסיה בצעדים לפי AIC:**

המודל המתקבל הינו:

**תמונה שמכילה טקסט

התיאור נוצר באופן אוטומטי**

**רגרסיה לפנים לפי BIC:**

המודל המתקבל הינו:

תמונה שמכילה טקסט

התיאור נוצר באופן אוטומטי  
**רגרסיה לאחור לפי BIC:**

המודל המתקבל הינו:

תמונה שמכילה טקסט

התיאור נוצר באופן אוטומטי

**רגרסיה בצעדים לפי BIC:**

המודל המתקבל הינו:

תמונה שמכילה טקסט

התיאור נוצר באופן אוטומטי

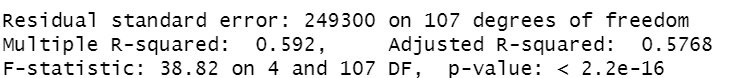
עם ערך BIC של:ששווה בכולם

כלומר לפי הערכים שהתקבלו 2 אפשרויות:

1. מודל שבו נכללים כל המשתנים למעט מספר החדרים והמדדים שלו הם: AIC=2801.96 ו BIC=3124.522 בנוסף, קיבלנו R^2 adj=0.6104

תמונה שמכילה טקסט

התיאור נוצר באופן אוטומטי

1. מודל שבו נכללים הגודל של הבית, השטח של הנכס, האם שופץ ומשתנה אינטראקציה בין האם שופץ לבין גודל הבית שנבחר במודל שבו AIC=2782.84 ו BIC= 3129.869. בנוסף, קיבלנו R^2 adj= 0.5768

בחרנו במודל 1. שבו נכללים כל המשתנים למעט מספר החדרים. הסיבה שבחרנו במודל הזה מבין שני המודלים היא כי ההפרש בין הAIC במודל גדול מההפרש של הBIC במודל השני, בנוסף ה R^2 adj גדול במודל הזה מהמודל השני.

**2.3) בדיקת הנחות המודל:**

נבחן את הנחות המודל בעזרת תרשימים וכדי להכריע מבחנים סטטיסטיים.

**בדיקת הנחת שוויון שונויות:**

כדי לבחון קיום שוויון השונויות נתבונן בתרשים הפיזור של השגיאות המתוקננות לעומת ערך החיזוי. ניתן לראות שפיזור התצפיות בגרף מציג סוג של משפך הגדל משמאל לימין לכן נוכל להניח כי שוויון השונויות לא מתקיים.

Chart, scatter chart

Description automatically generated

**בדיקת הנחת הלינאריות:**

נתבונן בתרשים שפיזור שגיאות המתוקננות לעומת ערך החיזוי, סביב האפס ניתן לראות פיזור אחיד לכן אניח שהנחת הלינאריות אכן מתקיימת.

**בדיקת הנחת הנורמאליות של השגיאות:**

Chart, line chart

Description automatically generatedChart, histogram

Description automatically generatedנבחן את קיום הנחת הנורמאליות על פי תרשים QQ-plot ובנוסף היסטוגרמה:

לפי התרשימים ניתן לראות שהנחת הנורמליות לא מתקיימת במודל שבחרנו.

**מבחנים סטטיסטיים לבדיקת קיום הנחת הנורמאליות:**

**Kolmogrov-Smirnov** - **השערת האפס**: הנחת הנורמאליות מתקיימת במודל.

**אלטרנטיבה**: הנחת הנורמאליות לא מתקיימת.

**Shapiro-Wilk** - **השערת האפס**: הנחת הנורמאליות מתקיימת במודל.

**אלטרנטיבה**: הנחת הנורמאליות לא מתקיימת.

ראשית נתחיל עם מבחן Kolmogrov-Smirnov:

Text

Description automatically generated

כלומר אנחנו לא דוחים את השערת האפס במבחן ונוכל לומר כי הנחת הנורמאליות מתקיימת בר"מ %5 . אשר לא מתיישב עם מה שהנחנו על הגרפים.

שנית נבחן בעזרת Shapiro-Wilk:

Text

Description automatically generated

מבחן Shapiro-Wild נכון יותר לעומת המבחן Kolmogrov-Smirnov ולכן אם יש מסקנות שונות נעדיף את ממצאי Shapiro-Wild, לכן כפי שניתן לראות נדחה את השארת האפס בר"מ 5% ונאמר שהנחת הנורמליות אינה מתקיימת במודל שבחרנו.

**מבחנים סטטיסטיים לבדיקת קיום הנחת הלינאריות:**

**Chow** – **השערת האפס:** קו הרגרסיה אחיד ורציף ז"א אין שבר או קפיצה של טווח כלשהו בקו הרגרסיה.

אלטרנטיבה: ELSE.

נבחן את הנחת הלינאריות בעזרת מבחן Chow:

Graphical user interface, application

Description automatically generated with medium confidence

ניתן לראות שלפי מבחן זה לא נדחה השערת האפס בר"מ 5% ונאמר שאין "שבר" בקו הרגרסיה והנחת הלינאריות מתקיימת.

**מבחנים סטטיסטי לבדיקת קיום הנחת שוויון שונויות:**

**Goldfeld – Quandt** – **השערת האפס:** קיים שוויון שוניות.

**אלטרנטיבה:** לא קיים שוויון שונויות.

נבחן את הנחת שוויון שונויות בעזרת Goldfeld – Quandt:

נחפש משתנה מסביר חשוד הגורם לאי שוויון השונויות.

בחנו את המשתנה "גודל המרתף" מפני שקיים מספר רב של נכסים ללא מרתף ולכן חששנו שהוא הגורם. חילקנו את "גודל המרתף" לארבע תחומים וקיבלנו:

Graphical user interface, text

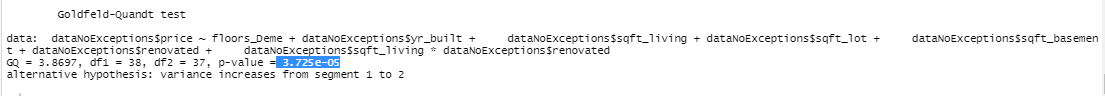
Description automatically generated with medium confidence

בנוסף כאשר נתבונן בגרף ניתן לראות שהתצפיות מתרחקות מקו הרגרסיה ככל שהמחיר עולה זה מתיישב עם הבדיקה שביצענו.

Chart, line chart, scatter chart

Description automatically generated

לאחר ביצוע המבחן נשאיר קבוצה בגודל שישית מסך התצפיות:



לכן נדחה את השערת האפס בר"מ 5% ונאמר שלא קיים שוויון שוניות במודל הנבחר.

**4) שיפור המודל:**

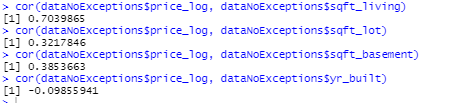
נרצה לשפר את המודל מפני שהוא לא מקיים את הנחות המודל, בכדי לעשות זאת נבצע טרנספורמציה על y כך שהמודל יקיים את ההנחות הנדרשות. נבחן 2 אפשרויות של טרנספורמציות log(y) ו-(y)^0.5, נעשה בדיקות על התחומים של "גודל מרתף" ונבדוק האם עדיין יש חשד לגורם אי שוויון שוניות.

Text

Description automatically generated

ניתן לראות שב- log(y) "גודל המרתף" אינו חשוד כי השונות של כל טווח בסדרי גודל זהים לעומת (y)^0.5, לכן נחליט להתקדם עם log(y). בנוסף ננסה להוסיף חזרה את המשתנה "מספר חדרים" אך בלוג. נבחר לא להוסיף למודל משתנים שלא הופיעו במודל הקודם, זאת מפני שבמודל הקודם לא מצאנו קשר לינארי בין המחיר למשתנה לכן סביר שלא נמצא קשר בין לוג המחיר לבין המשתנה.

ניתן לראות שקיים קשר לינארי בין לוג המחיר לבין כל המשתנים הרציפים שבחרנו לא להסיר במודל הקודם.



ניתן לראות שקיבלנו קורלציה דומה לסעיף 2.1 (כמו בסעיף 2.1 נניח קשר סיבתי בין שנת בניה לבין מחיר הנכס) לכן אושש ההנחה שלנו לגבי קשר לינארי בין מחיר למשתנים הרציפים שהופיעו במודל הקודם.

נבצע רגרסיה בצעדים על מודל הלוג שלנו הכולל את לוג "מספר החדרים".

נקבל מודל חדש:

Table

Description automatically generated

קיבלנו Adjusted R-squared ו-R-squared נמוכים יותר מהמודל הקודם. אנו סוברים ש- R-squared ירד בגלל שהפחתנו משתנים מסבירים מהמודל אשר משפיע ישירות. Adjusted R-squared שקיבלנו מבטיח לנו שאין לנו over fitting.

**בדיקת הנחת שוויון שונויות והנחת הלינאריות:**

כדי לבחון קיום שוויון השונויות נתבונן בתרשים הפיזור של השגיאות המתוקננות לעומת ערך החיזוי. ניתן לראות שפיזור התצפיות בגרף מפוזר בצורה יחסית אחידה לכן נניח כי כעת הנחת שוויון השונויות יכולה להתקיים. מעבר לכך הפיזור האחיד סביב האפס מעיד על לינאריות.

Chart, scatter chart

Description automatically generated

**בדיקת הנחת הנורמאליות של השגיאות:**

נבחן את קיום הנחת הנורמאליות על פי תרשים QQ-plot ובנוסף היסטוגרמה:

Chart, histogram

Description automatically generatedChart, line chart

Description automatically generated

ניתן לראות שקיים שיפור יחסי בין המודל החדש לבין המודל הקודם, ההבדל העיקרי הינו בהיסטוגרמה. לכן ניתן לומר שיתכן כי הנחת הנורמליות מתקיימת.

**מבחנים סטטיסטיים לבדיקת קיום הנחת הנורמאליות:**

ראשית נתחיל עם מבחן Kolmogrov-Smirnov:

Text

Description automatically generated with medium confidence

כלומר אנחנו לא דוחים את השערת האפס במבחן ונוכל לומר כי הנחת הנורמאליות מתקיימת בר"מ %5 . אשר מתיישב עם מה שהסקנו על הגרפים.

שנית נבחן בעזרת Shapiro-Wilk:

Text

Description automatically generated

מבחן Shapiro-Wilk "נכון יותר" לעומת המבחן Kolmogrov-Smirnov ולכן אם יש מסקנות שונות נעדיף את ממצאי Shapiro-Wilk, לכן כפי שניתן לראות לא נדחה את השארת האפס בר"מ 5% ונאמר שהנחת הנורמליות מתקיימת במודל החדש שבחרנו.

**מבחנים סטטיסטיים לבדיקת קיום הנחת הלינאריות:**

נבחן את הנחת הלינאריות בעזרת מבחן Chow:

Text

Description automatically generated with medium confidence

לפי תוצאות המבחן לא נדחה את השערת האפס בר"מ 5% ונאמר שהנחת הלינאריות מתקיימת.

**מבחנים סטטיסטי לבדיקת קיום הנחת שוויון שונויות:**

נבחן את הנחת שוויון שונויות בעזרת Goldfeld – Quandt:

נחפש משתנה מסביר חשוד הגורם לאי שוויון השונויות.

בחנו את המשתנה "גודל אזור מחייה" מפני שקיים טווח גדול ואולי הוא הגורם לאי קיום ההנחה. חילקנו את "גודל אזור מחייה" לארבע תחומים וקיבלנו:

Text

Description automatically generated

ננתח את ממצאי המבחן:

Text, letter

Description automatically generated

נתבונן בממצאי המבחן, ניתן לראות שלא נדחה את השערת האפס בר"מ 5% ונאמר שהנחת שוויון השונויות אכן מתקיים.

לאחר השינויים של המודל ניתן לראות שהמודל החדש שלנו מקיים את כל ההנחות הנדרשות, לסיכום הצלחנו לשפר את המודל אך עם זאת מדד טיב הטעמה ירד מעט זה אינדיקציה שאין לנו overfitting במודל החדש וסופי שלנו.

**המודל הסופי שהתקבל:**

Table

Description automatically generated

**נספחים**:

2.1)

Graphical user interface, text, application

Description automatically generated

Chart, scatter chart

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Chart, scatter chart

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

Chart, scatter chart

Description automatically generated

Text

Description automatically generated with medium confidence

Chart, scatter chart

Description automatically generated

Graphical user interface, text, application, timeline

Description automatically generated

**3.2)**

Text

Description automatically generated

Text

Description automatically generated

**4**)

Table

Description automatically generated

Table

Description automatically generated