מבוא למערכות לומדות – תרגיל 2

שם: ג'ייסון אלטר **ת"ז:** 318634110

1. <u>הוכחה:</u>
נשים שלכל $v \in V$ מתקיים ש:
$$X^T v = 0_V \Leftrightarrow X(X^T v) = X(0_V) = 0_V$$
לכן:
$$\mathrm{Ker}(X^T) = \{v \in V | X^T v = 0_V\} = \{v \in V | XX^T v = 0_V\} = \mathrm{Ker}(XX^T)$$
■

2. <u>הוכחה:</u> (נסמן את מספר השורות/עמודות ב-n)
ראשית ניזכר ש- $\mathrm{Ker}(A)^\perp = \{x \in \mathbb{R}^n | \langle x|v\rangle = 0,\ \forall v \in \mathrm{Ker}(A)\}$.
שנית, נראה הכלה דו-כיוונית:
   - יהי $v \in \mathrm{Im}(A^T)$ אז קיים $x \in \mathbb{R}^n$ כך ש- $A^T x = v$ ויהי $u \in \mathrm{Ker}(A)$ אז מתקיים ש- $Au = 0_V$.
   לכן מתקיים:
   $$\langle v|u\rangle = \langle A^T x|u\rangle = (A^T x)^T u = x^T A u = \langle x|Au\rangle \overset{u\in\mathrm{Ker}(A)}{=} \langle x|0_V\rangle = 0$$
   וזה אומר ש- $v \in \mathrm{Ker}(A)^\perp$.

   - יהי $v \in \mathrm{Ker}(A)^\perp$ אז מהגדרה מתקיים שלכל $u \in \mathrm{Ker}(A)$ אז $\langle v|u\rangle = 0$.
   נניח בשלילה ש-$v \notin \mathrm{Im}(A^T)$ ואז קיים $v' \in \mathrm{Im}(A^T)^\perp$ כך ש- $\langle v|v'\rangle \neq 0$ (כי אם v כן היה שייך ל-$\mathrm{Im}(A^T)$ אז לפי הגדרה לכל ווקטור $v' \in \mathrm{Im}(A^T)^\perp$ מתקיים $\langle v|v'\rangle = 0$).
   נשים לב ש-$A^T(Av') \in \mathrm{Im}(A^T)$ (הפעלת A^T על הווקטור Av') ולכן:
   $$0 = \langle v'|A^T(Av')\rangle = v'^T A^T A v' = (Av')^T A v' = \langle Av'|Av'\rangle$$
   לכן חייב להתקיים ש- $Av' = 0_V$ (חיוביות בהחלט) ולכן $v' \in \mathrm{Ker}(A)$ אך לכל $u \in \mathrm{Ker}(A)$ אז $\langle v|u\rangle = 0$ וזו סתירה לכך ש- $\langle v|v'\rangle \neq 0$. לכן $v \in \mathrm{Im}(A^T)$.

הראנו הכלה דו-כיוונית ולכן $\mathrm{Im}(A^T) = \mathrm{Ker}(A)^\perp$. ■

3. <u>הוכחה:</u>
נתון ש-X^T מטריצה לא הפיכה ולכן למערכת המשוואות $y = X^T w$ יש ∞ או 0 פתרונות (רק למטריצה הפיכה יש דרגה מלאה ולכן פתרון יחיד).
לכן הטענה שנרצה להוכיח שקולה לטענה:
קיים לפחות פתרון אחד למערכת המשוואות $\Leftrightarrow y \perp \mathrm{Ker}(X) \Leftrightarrow y \in \mathrm{Ker}(X)^\perp$
אך הוכחנו בשאלה הקודמת ש-$\mathrm{Im}(A^T) = \mathrm{Ker}(A)^\perp$ ולכן קיום פתרון למערכת שוות שקולה ל:
$$y \in \mathrm{Im}(A^T) \Leftrightarrow y \in \mathrm{Ker}(A)^\perp$$
וזו השקילות הנדרשת. ■

4. <u>הוכחה:</u>
<u>אם XX^T מטריצה הפיכה:</u> אז מתקיים ש-
$$XX^T w = Xy \Leftrightarrow (XX^T)^{-1}(XX^T)w = w = (XX^T)^{-1}Xy$$
ומכיוון שצד ימין נתון לנו ויחיד אז קיים פתרון יחיד ל-w.
<u>אחרת:</u> XX^T מטריצה אי-הפיכה ואז ממה שהוכחנו בשאלה הקודמת מתקיים שלמערכת משוואות $XX^T w = Xy$ יש ∞ פתרונות אם ורק אם $Xy \perp \mathrm{Ker}(XX^T)$ אך הוכחנו כבר ש-
$\mathrm{Ker}(X^T) = \mathrm{Ker}(XX^T)$ ולכן שקול להוכיח ש-$Xy \perp \mathrm{Ker}(X^T)$.

יהי $u \in \mathrm{Ker}(X^T)$ אז מתקיים: $0 = \langle Xy|u\rangle = (Xy)^T u = y^T X^T u = \langle y|X^T u\rangle = \langle y|0\rangle$.
לכן $Xy \perp \mathrm{Ker}(X^T)$ וזה שקול לכך שיש ∞ פתרונות. ■

5. <u>נתון:</u> $P = \sum_{i=1}^{k} v_i v_i^T$

(א) נשים לב שמחוקי שחלוף מתקיים:

$$P = \sum_{i=1}^{k} v_i v_i^T = \sum_{i=1}^{k} (v_i v_i^T)^T = \left(\sum_{i=1}^{k} v_i v_i^T\right)^T = P^T$$

כלומר P מטריצה סימטרית.

(ב) נשים לב שמאורתונורמליות שלכל $1 \le j \le k$ מתקיים:

$$Pv_j = \left(\sum_{i=1}^{k} v_i v_i^T\right) v_j = \sum_{i=1}^{k} v_i v_i^T v_j = \sum_{i=1}^{k} v_i \langle v_i | v_j\rangle = \sum_{i=1}^{k} v_i \delta_{ij} = v_j$$

ולכן 1 הוא ערך עצמי ו-$v_1, \dots, v_k$ הם הווקטורים העצמיים המתאימים לו.
כעת, לכל ווקטור שמקיים $u \notin Span\{v_1, \dots, v_k\}$ אז מתקיים מאורתוגונליות הבסיס:

$$Pu = \left(\sum_{i=1}^{k} v_i v_i^T\right) u = \sum_{i=1}^{k} v_i v_i^T u = \sum_{i=1}^{k} v_i \langle v_i | u\rangle = \sum_{i=1}^{k} v_i * 0 = 0_V$$

לכן שאר הערכים העצמיים הם 0.

(ג) יהי $v \in V$ אז קיימים $a_1, \dots, a_k$ סקלרים כך ש- $v = \sum_{i=1}^{k} a_i v_i$ ולכן:

$$Pv = P\left(\sum_{i=1}^{k} a_i v_i\right) \overset{\text{ליניאריות}}{=} \sum_{i=1}^{k} a_i P v_i \overset{\text{ווקטור עצמי}}{=} \sum_{i=1}^{k} a_i \lambda_i v_i = \sum_{i=1}^{k} a_i v_i = v$$

כאשר המעבר לפני האחרון מתקיים משום שהראנו כבר שהערך העצמי המתאים לכל v_i $(1 \le i \le k)$ הוא 1.

(ד) נתבונן ב-EVD של P: $P = UDU^{-1}$. (כאשר נשלים את הבסיס האורתונורמלי הנתון לבסיס אורתונורמלי לכל $\mathbb{R}^d$ ועמודות U הן בסיס זה)
מכיוון ש-D אלכסונית ומצאנו שכל הערכים העצמיים הם 1 או 0 אז מתקיים:

$$P = U\,\mathrm{Diag}(1, \dots, 1, 0, \dots, 0)\,U^{-1}$$

לכן:

$$\begin{aligned} P^2 &= (U\,\mathrm{Diag}(1, \dots, 1, 0, \dots, 0)\,U^{-1})^2 = UDU^{-1}UDU^{-1} = UD^2U^{-1} \\ &= U(\mathrm{Diag}(1, \dots, 1, 0, \dots, 0))^2 U^{-1} = U\,\mathrm{Diag}(1, \dots, 1, 0, \dots, 0)\,U^{-1} = P \end{aligned}$$

(ה) נשים לב שהטענה שקולה לטענה שהוכחנו בסעיף הקודם:

$$P^2 = P \Leftrightarrow 0 = P - P^2 = (I - P)P$$

6.

- נתון לנו שה-SVD של X הוא $X = U\Sigma V^T$ ולכן מתקיים:

$$XX^T = U\Sigma V^T(U\Sigma V^T)^T = U\Sigma V^T V\Sigma^T U^T \overset{V\text{ אורתונורמלית}}{=} U\Sigma I_d \Sigma^T U^T = U\Sigma\Sigma^T U^T = UDU^T$$

לכן המטריצה ההופכית היא:

$$(XX^T)^{-1} = (UDU^T)^{-1} = (U^T)^{-1}D^{-1}U^{-1} \overset{\text{אורתוגונליות}}{=} UD^{-1}U^T = U(\Sigma\Sigma^T)^{-1}U^T$$

- ראשית, נשים לב שמכיוון ש-XX^T הפיכה אז מתקיים:

$$0 \neq \det(XX^T) = \det(U\Sigma\Sigma^T U^T) = (\det(U))^2 * (\det(\Sigma))^2$$

ולכן $\prod_{i,i}\sigma_i = \det(\Sigma) \neq 0$ (מטריצה אלכסונית) וזה אומר שאף ערך סינגולרי שווה לאפס ולכן ההגדרה של $\Sigma^\dagger$ זהה להגדרה של Σ^{-1}. כלומר $\Sigma^{-1} = \Sigma^\dagger$.

שנית, נשתמש ב-SVD של X ובחישוב מהחלקים הקודמים ונקבל:

$$(XX^T)^{-1}X = U(\Sigma\Sigma^T)^{-1}U^T U\Sigma V^T \overset{U\text{ אורתונורמלית}}{=} U(\Sigma\Sigma^T)^{-1}\Sigma V^T = U(\Sigma^T)^{-1}\Sigma^{-1}\Sigma V^T$$
$$= U(\Sigma^T)^{-1}V^T = (V\Sigma^{-1}U^T)^T = (V\Sigma^\dagger U^T)^T = X^{T\dagger}$$

7. בתחילת התרגיל הוכחנו ש-$\text{Ker}(X^T) = \text{Ker}(XX^T)$ ולכן מתכונות ממדים מתקיים:

$$\dim\text{Ker}(X) = \dim\text{Ker}(X^T) = \dim\text{Ker}(XX^T)$$

ועכשיו ממשפט הממדים השני מתקבל:

$$\dim Span\{x_1, \dots, x_m\} = \dim\text{Im}(X) = \dim\text{Im}(XX^T)$$

המטריצה XX^T היא מטריצה מגודל $d \times d$ ולכן היא הופכית אם ורק אם דרגתה מלאה – כלומר אם $\dim Span\{x_1, \dots, x_m\} = \dim\text{Im}(XX^T) = d$ וזה שקול לכך ש- $Span\{x_1, \dots, x_m\} = \mathbb{R}^d$.

8. נשים לב שמתקיים:

$$\|U^T\hat{w}\|_2^2 = (U^T\hat{w})^T U^T\hat{w} = \hat{w}^T UU^T\hat{w} = yV\Sigma^\dagger U^T UU^T U\Sigma^{T\dagger}V^T y = yV\Sigma^\dagger\Sigma^{T\dagger}V^T y$$
$$= (\Sigma^{T\dagger}V^T y)^T\Sigma^{T\dagger}V^T y = \left\|\Sigma^{T\dagger}V^T y\right\|_2^2 = \sum_{i=1}^{d}\left(\Sigma_i^{T\dagger}(V^T y)_i\right)^2$$
$$= \sum_{i=1}^{m}\left(\Sigma_i^{T\dagger}(V^T y)_i\right)^2 + \sum_{i=k+1}^{d}(0 * (V^T y)_i)^2 \overset{*}{\leq} \sum_{i=1}^{d}\overline{w}_i^2 = \overline{w}^T\overline{w} = \overline{w}^T UU^T\overline{w}$$
$$= (U^T\overline{w})^T U^T\overline{w} = \|U^T\overline{w}\|_2^2$$

(כאשר $m \leq d$ מייצג את מספר הדגימות x_i)

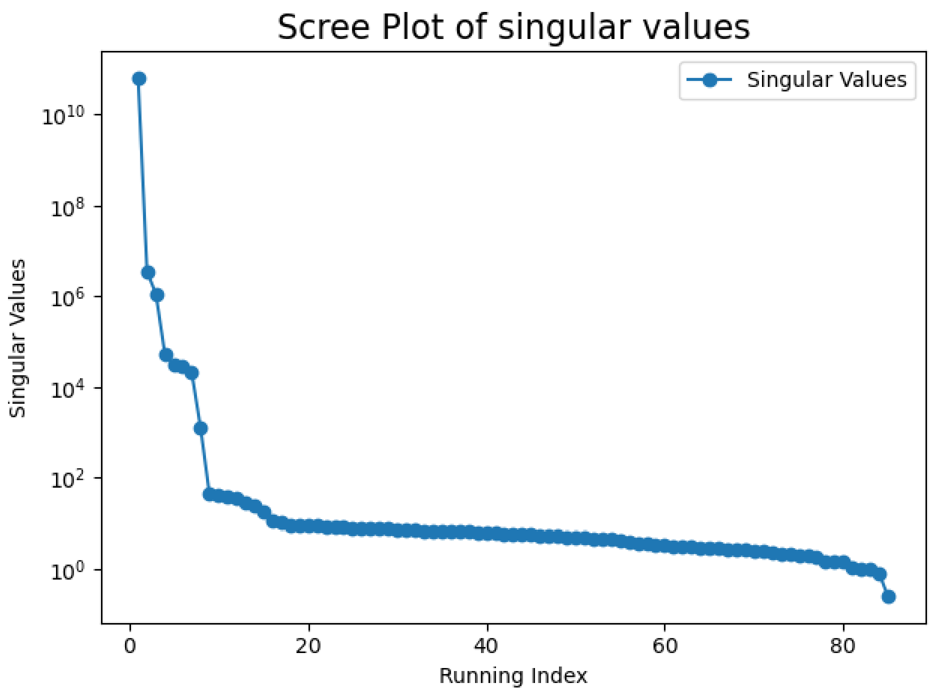
*אי-השוויון הזה מתקיים משום שעד m כלשהו קיימים ערכים סינגולריים יחידים לכל שורה שנותנים פתרון יחיד לשורה זו במערכת המשוואות ולכן עד נקודה זו כל פתרון $\overline{w}$ חייב להיות זהה. אחרי נקודה זו הפתרון שלנו $\Sigma^{T\dagger}V^T y$ מאפס (לפי איך שהגדרנו את $\Sigma^\dagger$) את שאר השורות ולכן בהכרח מצאנו פתרון שקטן/שווה לכל פתרון אחר.

לכן שקול שמתקיים $\|\hat{w}\|_2 \leq \|\overline{w}\|_2$.

12. בקוד המצורף ניתן לראות איך ביצעתי את שלב ה-preprocessing: ראשית, מחקתי את העמודה ID משום שהיא אינה תורמת מידע והיא שם רק כדי למספר את הדגימות.
שנית, עברתי על כל העמודות והגדרתי ערכים שגויים. למשל ערכים שליליים לחלק מהעמודות (sqft_lot, price, bedrooms, etc.) או waterfront חייב להיות בינארי. אני מוחק מהDataFrame את כל הדגימות שקיים לפחות ערך שגוי אחד בעמודה שלו (כלומר מוחק שורות) וזה משום שאם קיים ערך שגוי אחד אז אני מתייחס אל כל השורה כאל מידע לא מהימן.
בנוסף לכך, מחקתי את העמודה sqft_living משום שזהו סכום של השורות sqft_above, sqft_basement ולכן עמודה זו אינה מוסיפה מידע חדש.

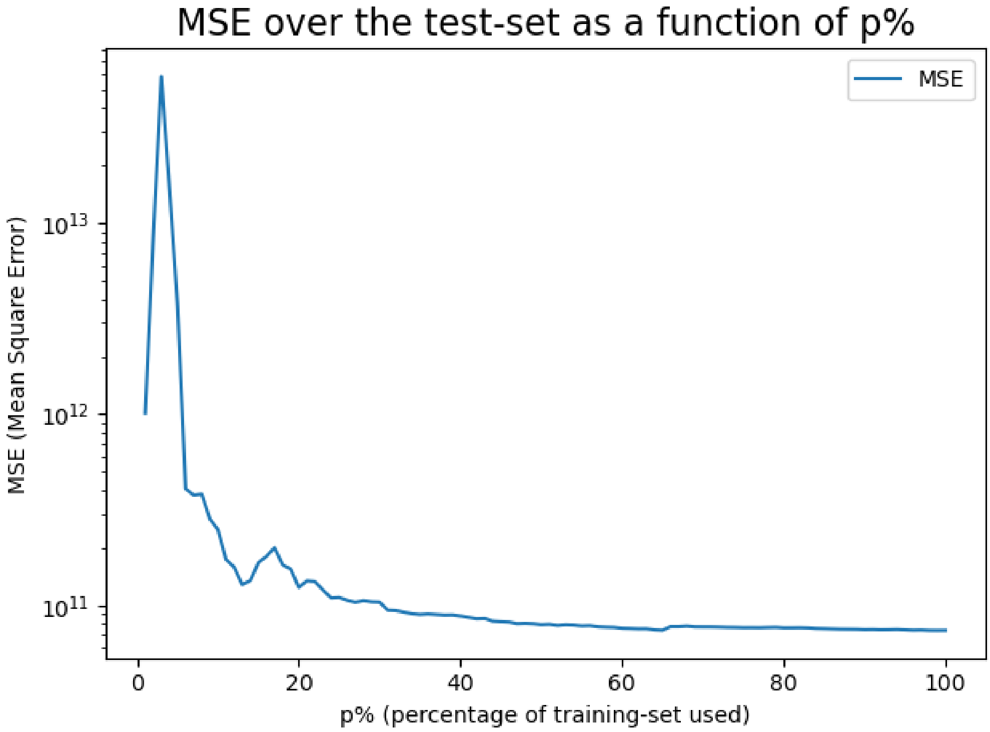
13. העמודה העיקרית שהיא categorical היא zipcode משום שהמספר עצמו לא נותן מידע כלשהו שניתן להשוות (לא הגיוני לתת ערך לכך שzipcode אחד גדול מאחר והם יכולים להיות מספרים שרירותיים). מכיוון שיש מספר מוגבל של zipcodes אז השתמשתי ב-One Hot Encoding כדי להתמודד עם עמודה זו.
עמודה שהיא לא בהכרח categorical אבל כן מהווה בעיה דומה היא העמודה date בפורמט המקורי שלה. בפורמט שניתן קשה להשוות בין תאריך אחד לשני אך פתרתי את הבעיה בכך שהעברתי את הפורמט אל Epoch time ואז קיבלנו מספר בעל משמעות שניתן להשוות.
(מחקתי כבר את ID בשלב הקודם ולכן אין צורך לחשוב אם הוא categorical או לא)

15. הplot:



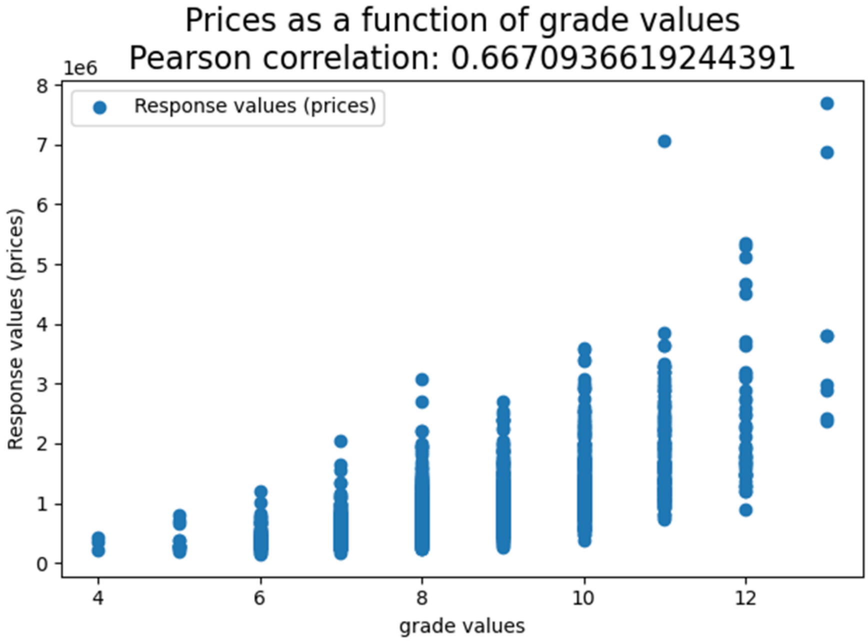
ניתן לראות שיש כמה ערכים גדולים בתחילת הגרף אך רוב הערכים מתקרבים או מאוד קרובים לאפס (לפחות יחסית לערכים ההתחלתיים) – כלומר רוב הערכים קטנים מהשאר ברמה משמעותית ולכן הם קרובים לאפס ונחשוב עליהם כאפסיים.
משמעות כל הערכים הסינגולריים האפסיים היא שיש תלות-ליניארית בין התכונות (features) ולכן יש לנו הרבה תכונות שלא נדרשות במודל או תכונות שבאיכות ירודה למציאת היפותזה ליניארית.
מכיוון שיש תלות לינארית בין התכונות וערכים הסינגולריים האפסיים אז המטריצה סינגולרית (כלומר אינה הפיכה) ולכן לא ניתן למצוא פתרון שהMSE שלו הוא אפס (כלומר ניתן רק לקבל פתרון מקורב).

16.הplot:



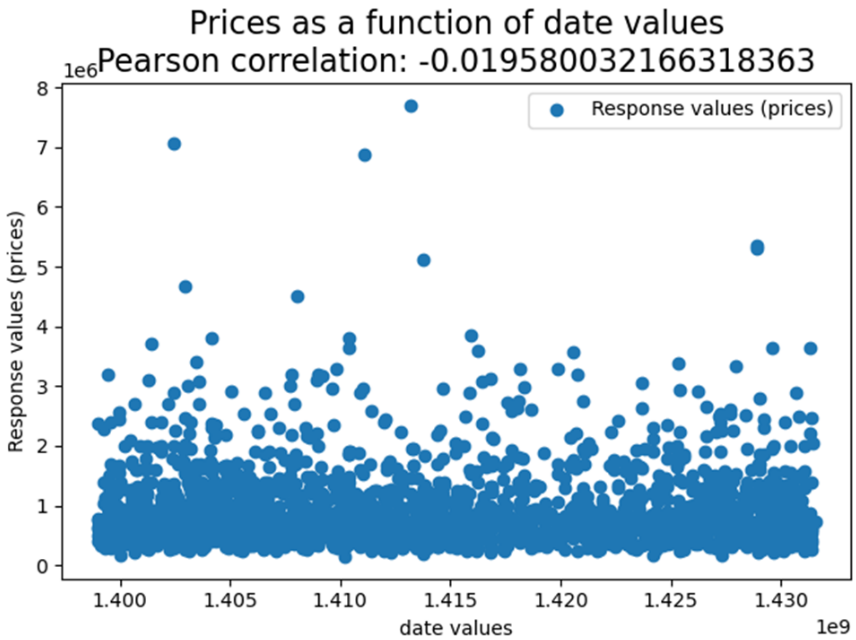
ניתן לראות שבתחילת הגרף (כאשר אחוז p נמוך מאוד) יש עלייה חדה מאוד בשגיאה – בעצם מאחוז קטן של נתונים מחושבת היפותזה שרחוקה מאוד מהנתונים האמיתיים. לאחר זאת יש ירידה חדה מאוד – כל הוספה של נתונים מתקנת בצורה גדולה את ההיפותזה מזו שהייתה קודם.
לאחר מכן יש חלק שבו השגיאה אינה יציבה משום שכאשר יש כמות קטנה של מידע שעליו מתבססת ההיפותזה אז כל אחוז נוסף יכול לשנות במידה רבה את החישוב. סופית, יש ירידה מתונה שמתייצבת בסוף – לכל אחוז נוסף יש פחות השפעה ממקודם והמידע הנוסף מקרב את ההיפותזה לנתונים האמיתיים.

17. feature שבחרתי שמועיל למודל הוא grade. הplot:



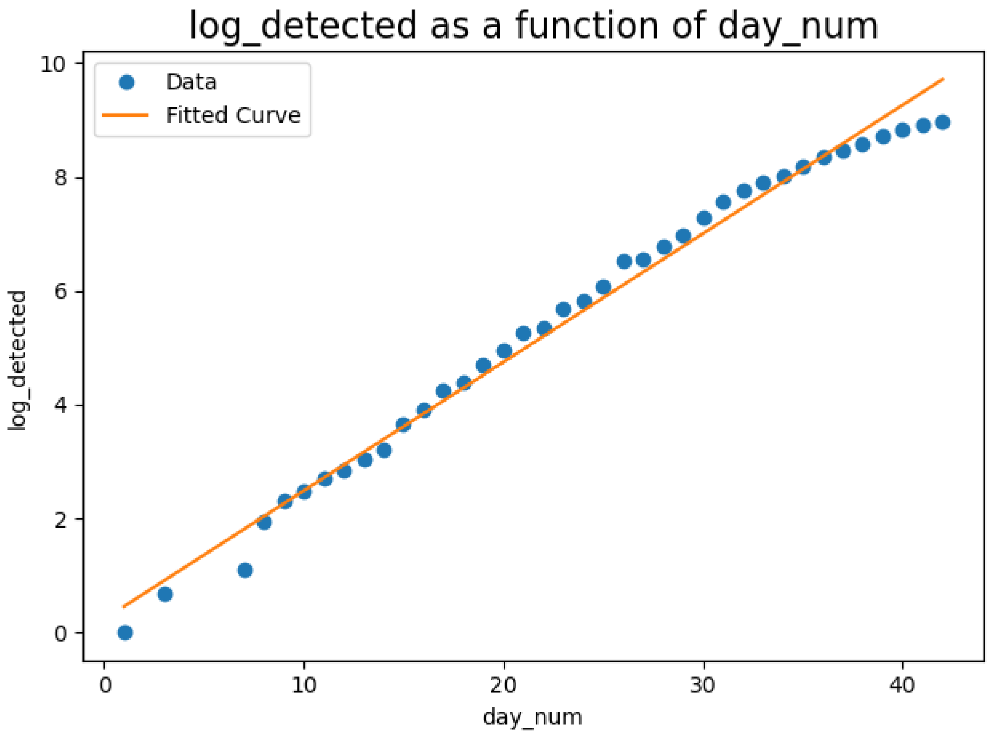
בחרתי את feature זה משום שניתן לראות עלייה במחיר בהתאם לערך הציון ועלייה זו מתאימה בערך לישר ליניארי. בנוסף לכך, מקדם מתאם פירסון למדד זה גבוה יחסית לאחרים וזה מציין שיש סבירות גבוהה יותר שקיים קשר ליניארי בין הציון למחיר.
(קיים מחסור של בתים עם הציון הגבוה ביותר בנתונים אך עדיין ניתן לראות עלייה)

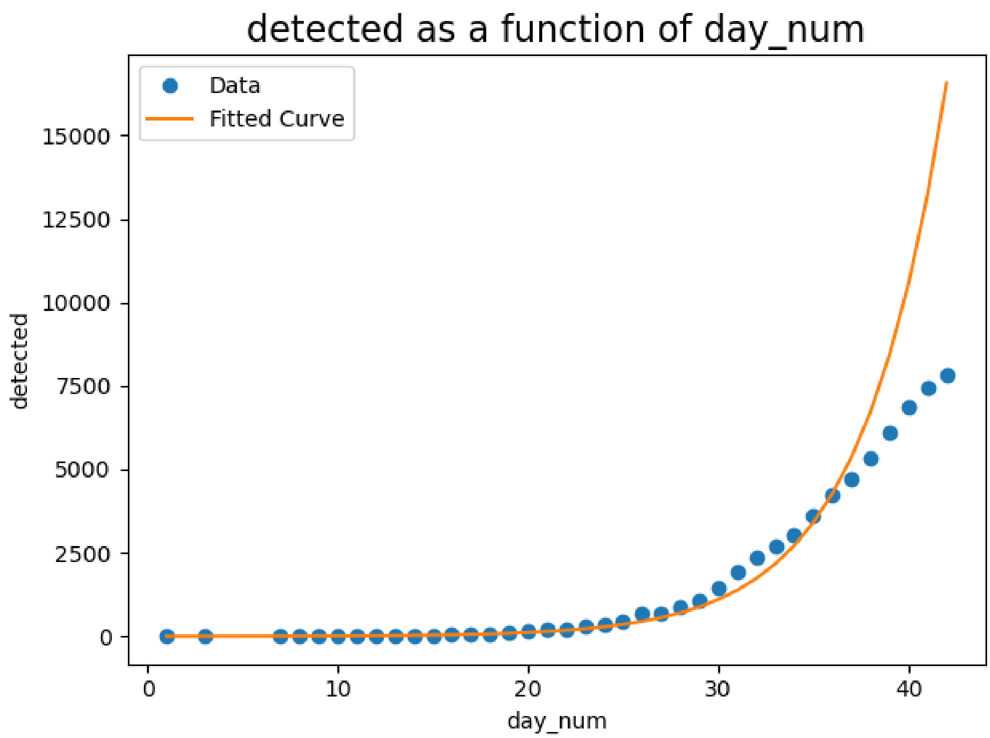
feature שבחרתי שאינו מועיל למודל הוא date. הplot:



בחרתי את feature זה משום שניתן לראות שאין שום קשר בין ערך התאריך (שפה נמדד לפי Epoch) לבין המחיר – כל חלקי הגרף נראים כמעט אותו הדבר (ודומים לקו עם שיפוע 0).
בנוסף לכך, מקדם מתאם פירסון למדד זה מאוד קרוב לאפס ולכן זה מציין שאין קשר ליניארי בין התאריך למחיר.
(הוספתי נספח בעמוד האחרון עם הגרפים של כל הfeatures)

21.הplots:





ניתן לראות שבגרף הראשון יש התאמה טובה בין הישר הליניארי שחזינו לבין נקודות המידע (עם הבדל קטן בסוף הגרף בין הישר לנקודות). לעומת זאת קיבלנו התאמה דיי טובה בגרף השני בין נקודות המידע לבין הפונקציה האקספוננציאלית (כאשר כל הבדל קטן מהגרף הראשון גדל בגרף השני – שזה צפוי בפונקציה אקספוננציאלית).

22. נתונה פונקציית הloss הבאה:

$$L_{exp}\left(f_w,(x,y)\right)=\left(\langle w|x\rangle-\log(y)\right)^2=\left(\sum_{i=1}^{n}w_ix_i-\log(y)\right)^2$$

ראשית, נחשב את הנגזרת שלה:

$$\forall 1\le j\le n,\qquad \frac{\partial L_{exp}\left(f_w,(x,y)\right)}{\partial x_j}\overset{\text{כלל השרשרת}}{=}2\left(\sum_{j=1}^{n}w_ix_i-\log(y)\right)w_j$$
$$=2w_j\left(\langle w|x\rangle-\log(y)\right)$$

לכן:

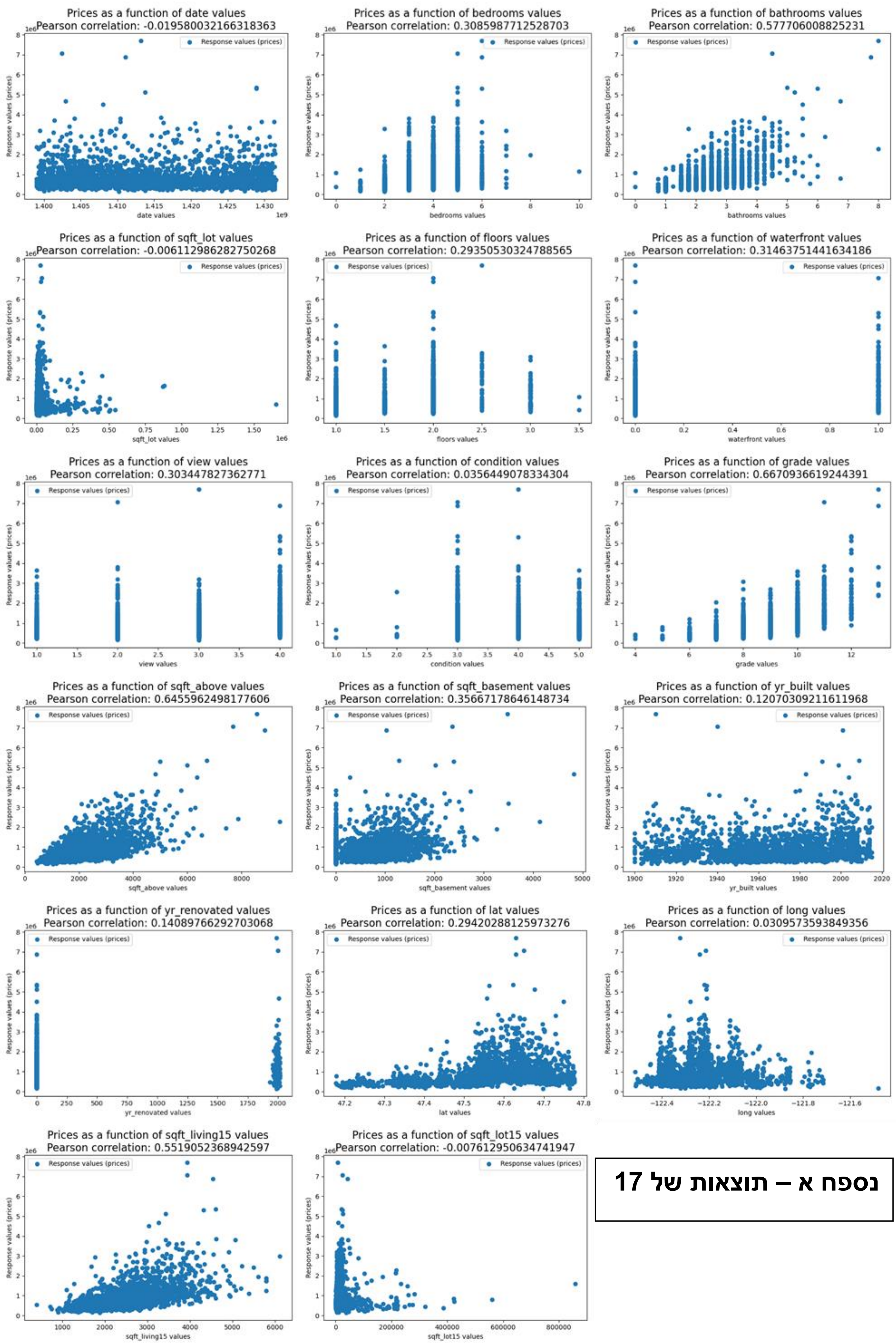
$$\frac{\partial L_{exp}\left(f_w,(x,y)\right)}{\partial x}=\begin{bmatrix}2w_1\left(\langle w|x\rangle-\log(y)\right)\\ \vdots\\ 2w_n\left(\langle w|x\rangle-\log(y)\right)\end{bmatrix}=2w\left(\langle w|x\rangle-\log(y)\right)$$

כמו שראינו בכיתה אז על מנת למצוא את נקודת המינימום של הפונקציה אז נשווה לאפס ונשים לב ש- $2w\left(\langle w|x\rangle-\log(y)\right)=0$ מתקיים כאשר w מטריצת האפס או כאשר $\left(\langle w|x\rangle-\log(y)\right)=0$ ומשום שאנחנו מנסים למצוא פתרון w לא טריוויאלי אז רק צד ימין משנה לנו. לכן נראה שמתקיים:

$$\langle w|x\rangle-\log(y)=0\Leftrightarrow\langle w|x\rangle=\log(y)\Leftrightarrow e^{\langle W|X\rangle}=e^{\log(y)}=y$$

לכן הבעיה שקולה למציאת פונקציה שנגזרתה מקיימת $e^{\langle W|X\rangle}-y=0$ (כפול גורם שאינו משפיע על התוצאה) ולכן פונקציה מתאימה היא: $\left(e^{\langle W|X\rangle}-y\right)^2$.

לכן במצב כזה ניתן לחשב את הפתרון ERM בעזרת חישוב רגרסיה ליניארית ל$\log(y)$ כדי לקבל פתרון ERM שממזער את $\left(\langle w|x\rangle-\log(y)\right)^2$ ואז נחשב את $e^{\langle W|X\rangle}$ ומהשקילות מצאנו את הפתרון שממזער את פונקציית הloss.



נספח א – תוצאות של 17