שלב 1: חלוקה לסט אימון וסט מבחן ובחירת אופן מזעור השגיאות

כדי להבין מהם הגורמים שמשפיעים על הדירוגים רצינו לעבוד רק על data מלא.
לפיכך, בכל השלבים והבדיקות שהרצנו, עבדנו על מידע שמגיע מהטבלה user_artists_csv.
כדי לבצע למידה מונחית שבה נוכל להעריך את הביצועים שלנו, חילקנו את הטבלה לסט אימון של 75% וסט מבחן של 25%.

כפי שנאמר לנו בהרצאה, המטרה בתרגיל הינה מזעור ה loss הנתון (הפרש הלוגים).
משום שלוג הינה פונקציה מונוטונית עולה, אם נמצא שיטה שתמזער את ה RMSE השיטה תצליח גם למזער את ה loss אך עם הפעלת טרנספורמציה על הפרדיקציה הסופית. משום שמדובר בלוג מבסיס 10, נכניס את המשקלים של המדגם בלוג ולאחר מכן נעלה את 10 בחזקת החיזויים. לפיכך במהלך העבודה התמקדנו במזעור ה RMSE ובסוף את החיזויים ואת הבדיקה עם המדד הנתון המרנו לפי הטרנספורמציה הנתונה לעיל.

שלב 2: ריבועים פחותים והוספת רגולריזציה

בנינו מערכת ריבועים פחותים כפי שנלמד בהרצאה:

$$\begin{array}{c} \\ 1 \\ 2 \\ \vdots \\ 30 \end{array} \begin{array}{c} \begin{array}{ccccccccc} 1 & 2 & 3 & \dots & 10 & A & B & \dots & E \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & \dots & 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 & 1 & 0 & \dots & 0 \\ \vdots & & & & & & & \vdots & \\ 0 & 0 & 0 & \dots & 1 & 0 & 0 & \dots & 1 \end{pmatrix} \end{array} \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_{10} \\ b_A \\ b_B \\ \vdots \\ b_E \end{pmatrix} - \begin{pmatrix} r_{1A} - r_{\text{avg}} \\ r_{3A} - r_{\text{avg}} \\ \vdots \\ r_{10E} - r_{\text{avg}} \end{pmatrix}$$

הסטיות (biases) שמתקבלות מפתרון המערכת ממזערות את ה RMSE.

משום שאנו רוצים מודל שיכליל בצורה טובה יותר ושיימנע מ overfit הוספנו רגולריזציה כפי שנלמד בהרצאה. על מנת להמיר את הבעיה הנתונה לבעיית ריבועים פחותים, השתמשנו בטרנספורמציה שנלמדה בקורס שיטות אלגבריות. להלן הפיתוח:

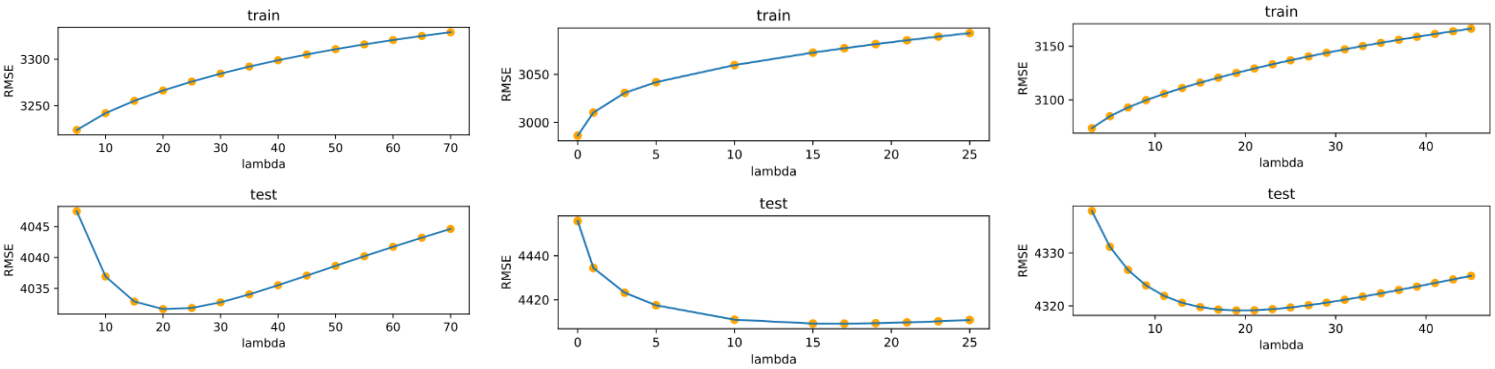
<u>פתרון</u>

ניתן לכתוב את הבעיה בצורה הבאה:

$$\min_{x \in R^n} \|Ax - y\|_2^2 + \lambda \|x\|_2^2 = \min_{x \in R^n} \left\| \begin{pmatrix} A \\ \sqrt{\lambda} I \end{pmatrix} x - \begin{pmatrix} y \\ 0 \end{pmatrix} \right\|_2^2$$

במקרה שלנו, מכיוון שיש 2 סכומים ניתן לאחדם לכדי סכום ריבועים יחיד ונקבל את הנורמה של הפתרון – וקטור ההטיות.

כדי למצוא את ערך הלמדה המיטבי, הרצנו סימולציות ולקחנו את ערכי ה RMSE הממוצעים.
להלן פלטים של כמה הרצות:



לפי elbow-method בחרנו את למבדה להיות 10 כלמבדה המיטבי.

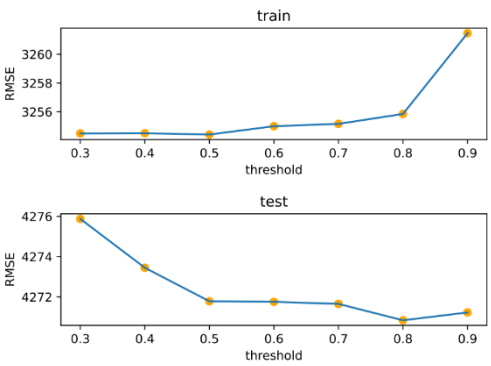
שלב 3: מימוש NEIGHBOURHOOD METHOD

כאשר אנו מוסיפים את גורם הדמיון, למעשה אנו מנסים לתקן את השגיאות שנוצרו מהשמת ממוצע + הסטיות. לכן בחרנו לממש גם את שיטת דמיון הקוסינוסים ולחבר את גורם הדמיון לגורם שחושב קודם לכן.

החלק המאתגר היה חישוב מטריצת הדמיון. במטריצה השגיאות R תילדה אנו שומרים את כל האומנים בשורות ואת המשתמשים בעמודות. עבור אמן מסויים, הדמיון בינו לבין אומן אחר יהיה 0 אם אין להם אף כניסה משותפת (:= כניסה ללא 0).

כדי להימנע ממעבר איטרטיבי על מספר גדול מאוד (מס' האומנים בריבוע בריבוע) בדקנו את הדמיונות מראש באמצעות המכפלה: $\tilde{R}\tilde{R}^T$. כך אנו מחשבים את כל המכפלות הפנימיות טרם לכן, וממטריצה זו נחלץ את כל האינדקסים שאינם 0 (זוג אינדקסים משמע דמיון בין זוג אמנים). יתר על כן, עבור האמנים שיש להם רק כניסה אחת משותפת שאינה 0, שמנו דמיון 0 גם כן משום שהקרבה הזאת איננה אינדיקטיבית בעינינו (לא משנה איזה ערך יש לכל אחד מהם, המכפלה הפנימית תהיה 1). כעת ניסינו למצוא את מס השכנים הקרובים ביותר האופטימלי (קרובים ביותר בערך מוחלט).

עבור הרצות שונות לא קיבלנו הבדלים משמעותיים ועל כן העדפנו לשנות את החיפוש לגישה הבאה: במקום לחפש לפי מס' שכנים מוגדר מראש, נחפש לכל אמן את מס' האמנים האחרים שהדמיון בינו לבינם (בערך מוחלט) הוא לכל הפחות ערך סף מסוים. כדי לחפש את ערך הסף הרצנו בדיקה עבור כמה ערכי סף עם הלמדה הנבחר (10). להלן התוצאות:



לאחר שמצאנו את כל ה hyper-parameters של האלגוריתם, הרצנו אותו 3 פעמים עם הגרלות של סט אימון וסט מבחן. בנוסף כדי לבדוק את החיזויים שלנו ביחס למדד הנתון, מימשנו את הטרנספורמציה שפירטנו עליה בעמוד הראשון. לקחנו את ממוצע ה loss של 3 ההרצות וקיבלנו על סט אימון של 75% וסט מבחן של 25%:

```
avg of 3 runs train loss: 347604.4304160152
avg of 3 tuns test loss: 3294.6567590903687
```