מעבדה בניתוח נתונים עם R - חלק ב'

מגישות:

רוני מליחי 315037747

זהר בן משה 206180754

מרצה: מר זכאי אבי

תוכן עניינים

1. תיאור הבעיה:

קישור לנתונים

רקע: הדאטה מכיל נתוני מכירות של חברת סופרמרקט במיאנמר, משלושה סניפים שונים במשך שלושה חודשים (ינואר-מרץ 2019). 1000 תצפיות.

ביצענו בדיקה שהראתה שאין נתונים חסרים:

```
>
> #Number of NA's (=0)
> sum(is.na(ss))
[1] 0
>
```

תיאור משתנה המטרה RATING -

הרייטינג על סקאלה של 4-10 בקפיצות של 0.1, הוא מתאר דירוג שהלקוח נותן על חוויית הקניה הכוללת בסופר.

Categorical Attributes-

- Branch: Branch of supercenter (A, B and C).
- City: Location of supercenters.
- Customer type: Normal and Member.
- Gender: Gender type of customer
- Product line: Electronic accessories, Fashion accessories, Food and beverages, Health and beauty, Home and lifestyle, Sports and travel
- Payment: Cash, Credit card and Ewallet

Numerical Attributes-

- Invoice id
- Unit price in $
- Quantity
- Tax: 5% tax fee for customer buying
- Total: Total price including tax
- Date: Date of purchase (Record available from January 2019 to March 2019)

- Time: Purchase time (10am to 9pm)
- COGS: Cost of goods sold
- Gross margin percentage
- Gross income: Gross income generated from particular product sold

מטרת המחקר - חיזוי הערך של הרייטינג (משתנה מטרה במותי) באמצעות המשתנים האחרים ומציאת המודל המתאים ביותר לחיזוי.

כך נראים הנתונים:

Invoice	Branch	City	Customer ty	Gender	Product line	Unit pr	Quanti	Tax 5%	Total	Date	Time	Paymen	COGS	Gross Margin Precentage	Gross Income	Rating
765-26-69	A	Yangon	Normal	Male	Sports and travel	72.61	6	21.783	457.443	01/01/2019	10:39	Credit card	435.66	4.761904762	21.783	6.9
530-90-98	A	Yangon	Member	Male	Home and lifestyle	47.59	8	19.036	399.756	01/01/2019	14:47	Cash	380.72	4.761904762	19.036	5.7
891-01-70	B	Mandalay	Normal	Female	Electronic accessories	74.71	6	22.413	470.673	01/01/2019	19:07	Cash	448.26	4.761904762	22.413	6.7
493-65-62	C	Naypyitaw	Member	Female	Sports and travel	36.98	10	18.49	388.29	01/01/2019	19:48	Credit card	369.8	4.761904762	18.49	7
556-97-71	C	Naypyitaw	Normal	Female	Electronic accessories	63.22	2	6.322	132.762	01/01/2019	15:51	Cash	126.44	4.761904762	6.322	8.5
133-14-72	C	Naypyitaw	Normal	Male	Health and beauty	62.87	2	6.287	132.027	01/01/2019	11:43	Cash	125.74	4.761904762	6.287	5
651-88-73	A	Yangon	Normal	Female	Fashion accessories	65.74	9	29.583	621.243	01/01/2019	13:55	Cash	591.66	4.761904762	29.583	7.7
182-52-70	A	Yangon	Member	Female	Sports and travel	27.04	4	5.408	113.568	01/01/2019	20:26	Ewallet	108.16	4.761904762	5.408	6.9
416-17-99	A	Yangon	Member	Female	Electronic accessories	74.22	10	37.11	779.31	01/01/2019	14:42	Credit card	742.2	4.761904762	37.11	4.3
271-77-87	C	Naypyitaw	Member	Female	Sports and travel	29.22	6	8.766	184.086	01/01/2019	11:40	Ewallet	175.32	4.761904762	8.766	5
770-42-89	B	Mandalay	Normal	Male	Food and beverages	21.12	8	8.448	177.408	01/01/2019	19:31	Cash	168.96	4.761904762	8.448	6.3
746-04-10	B	Mandalay	Member	Female	Food and beverages	84.63	10	42.315	888.615	01/01/2019	11:36	Credit card	846.3	4.761904762	42.315	9
504-35-88	A	Yangon	Normal	Male	Sports and travel	42.47	1	2.1235	44.5935	02/01/2019	16:57	Cash	42.47	4.761904762	2.1235	5.7
446-47-67	C	Naypyitaw	Normal	Male	Fashion accessories	99.82	2	9.982	209.622	02/01/2019	18:09	Credit card	199.64	4.761904762	9.982	6.7
244-08-01	B	Mandalay	Normal	Female	Health and beauty	34.21	10	17.105	359.205	02/01/2019	13:00	Cash	342.1	4.761904762	17.105	5.1
198-84-71	B	Mandalay	Member	Male	Fashion accessories	40.61	9	18.2745	383.7645	02/01/2019	13:40	Cash	365.49	4.761904762	18.2745	7
744-09-57	B	Mandalay	Normal	Male	Electronic accessories	22.01	6	6.603	138.663	02/01/2019	18:50	Cash	132.06	4.761904762	6.603	7.6
712-39-03	A	Yangon	Member	Male	Food and beverages	41.66	6	12.498	262.458	02/01/2019	15:24	Ewallet	249.96	4.761904762	12.498	5.6
345-68-90	C	Naypyitaw	Member	Female	Sports and travel	31.67	8	12.668	266.028	02/01/2019	16:19	Credit card	253.36	4.761904762	12.668	5.6
670-71-73	B	Mandalay	Normal	Male	Sports and travel	44.63	6	13.389	281.169	02/01/2019	20:08	Credit card	267.78	4.761904762	13.389	5.1
249-42-37	A	Yangon	Normal	Male	Health and beauty	70.01	5	17.5025	367.5525	03/01/2019	11:36	Ewallet	350.05	4.761904762	17.5025	5.5

איור I מספר שורות נתונים לדוגמא מהאקסל

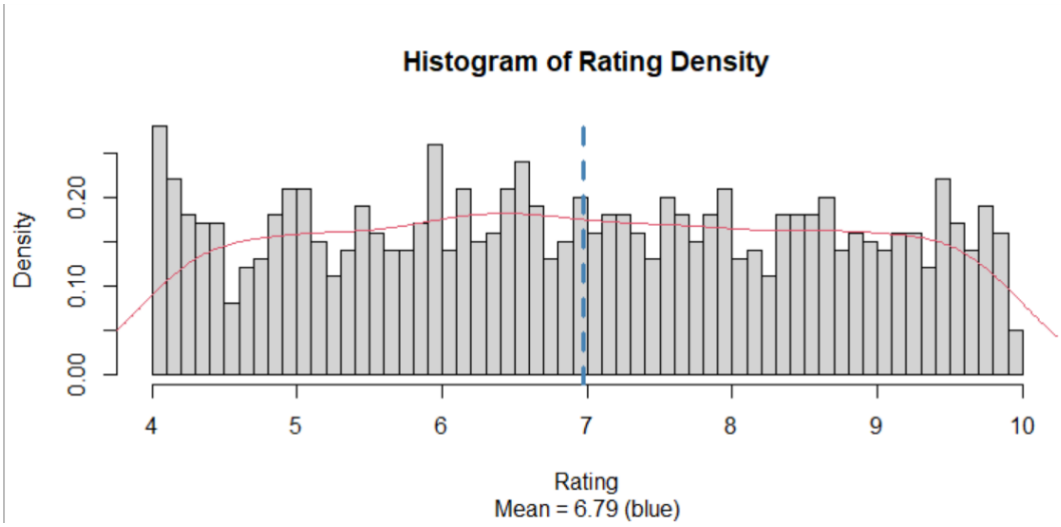
```
tibble [1,000 x 17] (S3: tbl_df/tbl/data.frame)
 $ Invoice ID             : chr [1:1000] "765-26-6951" "530-90-9855" "891-01-7034" "493-65-6248" ...
 $ Branch                 : chr [1:1000] "A" "A" "B" "C" ...
 $ City                   : chr [1:1000] "Yangon" "Yangon" "Mandalay" "Naypyitaw" ...
 $ Customer type          : chr [1:1000] "Normal" "Member" "Normal" "Member" ...
 $ Gender                 : chr [1:1000] "Male" "Male" "Female" "Female" ...
 $ Product line           : chr [1:1000] "Sports and travel" "Home and lifestyle" "Electronic accessories"
rts and travel" ...
 $ Unit price             : num [1:1000] 72.6 47.6 74.7 37 63.2 ...
 $ Quantity               : num [1:1000] 6 8 6 10 2 2 9 4 10 6 ...
 $ Tax 5%                 : num [1:1000] 21.78 19.04 22.41 18.49 6.32 ...
 $ Total                  : num [1:1000] 457 400 471 388 133 ...
 $ Date                   : POSIXct[1:1000], format: "2019-01-01" "2019-01-01" ...
 $ Time                   : POSIXct[1:1000], format: "1899-12-31 10:39:00" "1899-12-31 14:47:00" ...
 $ Payment                : chr [1:1000] "Credit card" "Cash" "Cash" "Credit card" ...
 $ COGS                   : num [1:1000] 436 381 448 370 126 ...
 $ Gross Margin Precentage: num [1:1000] 4.76 4.76 4.76 4.76 4.76 ...
 $ Gross Income           : num [1:1000] 21.78 19.04 22.41 18.49 6.32 ...
 $ Rating                 : num [1:1000] 6.9 5.7 6.7 7 8.5 5 7.7 6.9 4.3 5 ...
```

איור II הצגת הנתונים בR

2 . EDA - סטטיסטיקה תיאורית מלאה של משתנה המטרה ושל הפיצ'רים:

<u>משתנה המטרה:</u>

תחילה, הסתכלנו על ההתפלגות של צפיפות משתנה המטרה:



איור III היסטוגרמת צפיפות של 'רייטינג'

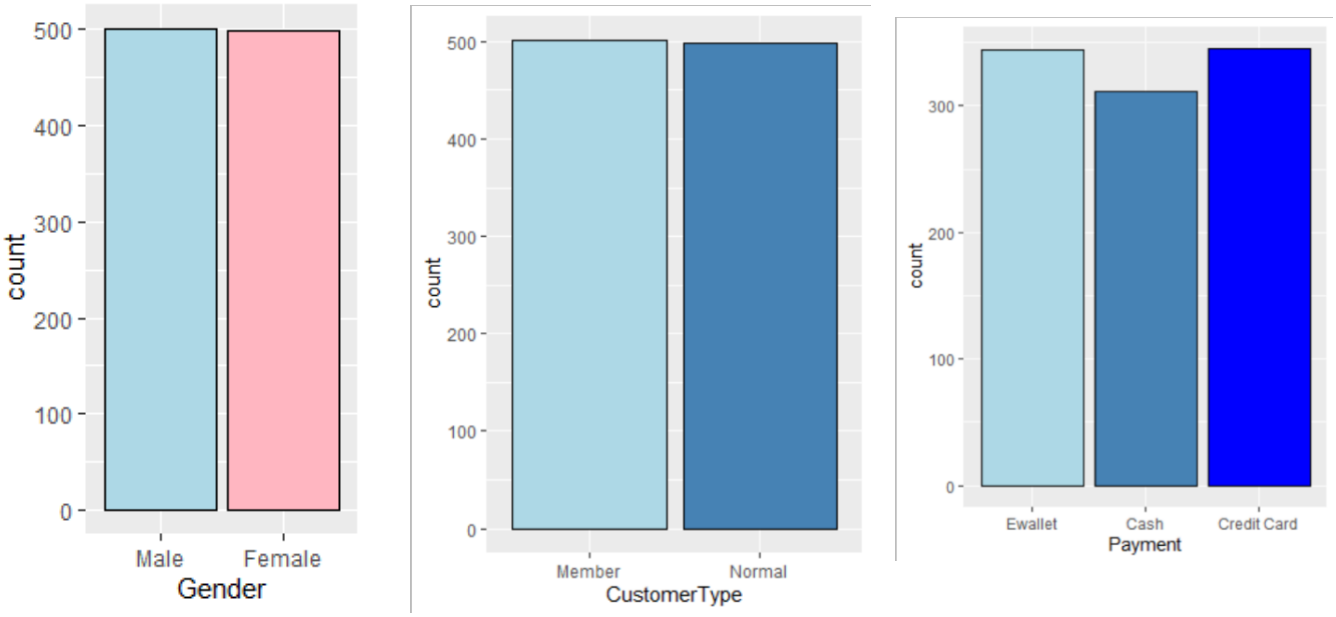
ההתפלגות נראית אחידה ולא מוטת לאחד הקצוות, והממוצע קרוב מאוד לחציון.

	Measure	Value
1	Valid N	1000.00
2	Mean	6.97
3	SD	1.72
4	Minimum	4.00
5	Median	7.00
6	Maximum	10.00

איור IV מדדי משתנה המטרה

<u>המשתנים:</u>

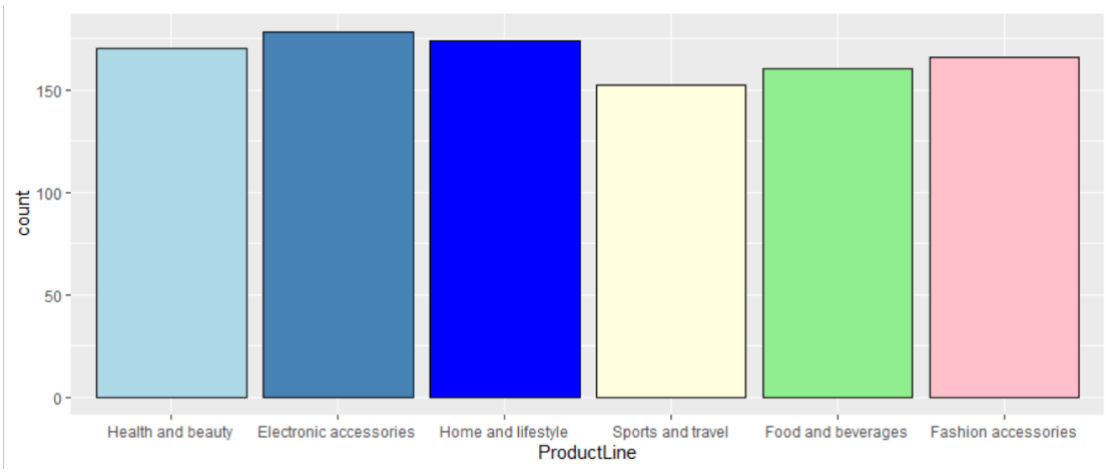
מהגרפים עולה כי המדגם כמעט מאוזן באופן מלא מבחינת מגדר, עיר, סוג לקוח, סוג התשלום, ומחלקה.



איור V - התפלגות הפיצ'רים: סוג תשלום, סוג לקוח ומגדר

```
  Gender Freq Rel.Freq     CustomerType Freq Rel.Freq   Payment Freq Rel.Freq
1   Male  501     50.1   1       Member  501     50.1   1 Ewallet  344     34.4
2 Female  499     49.9   2       Normal  499     49.9   2    Cash  311     31.1
                                                        3 Credit Card  345     34.5
```

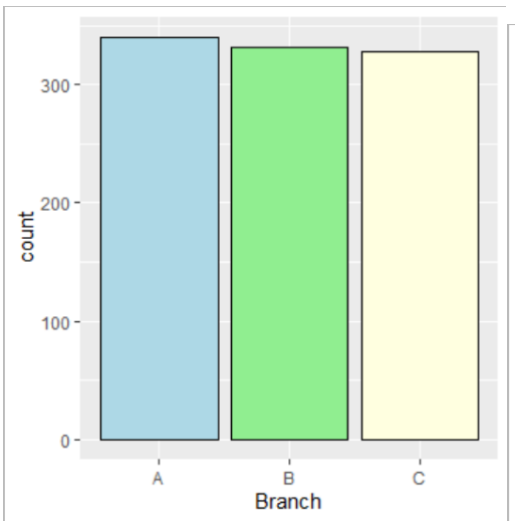
איור VI- טבלאות שכיחויות של הפיצ'רים: סוג תשלום, סוג לקוח ומגדר

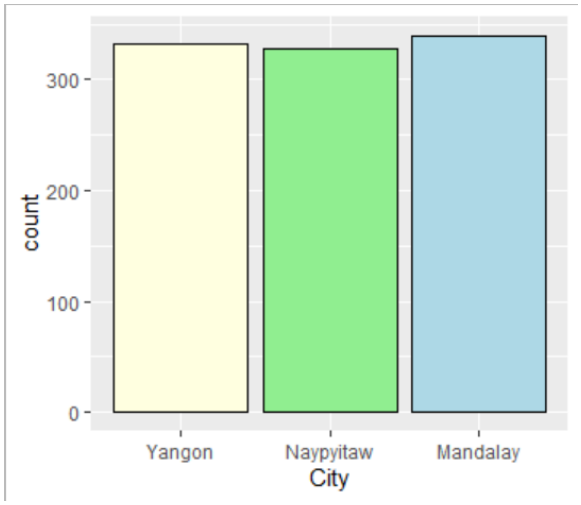


איור VIII התפלגות הפיצ'ר: מחלקה

```
> show(ft)
              ProductLine Freq Rel.Freq
1       Health and beauty  170     17.0
2 Electronic accessories  178     17.8
3      Home and lifestyle  174     17.4
4       Sports and travel  152     15.2
5      Food and beverages  160     16.0
6     Fashion accessories  166     16.6
```

איור VII טבלת שכיחויות של הפיצ'ר מחלקה





איור IX התפלגות זהה של הפיצ'רים עיר וסניף

```
   Branch Freq Rel.Freq
1       A  340      34.0
2       B  332      33.2
3       C  328      32.8
```

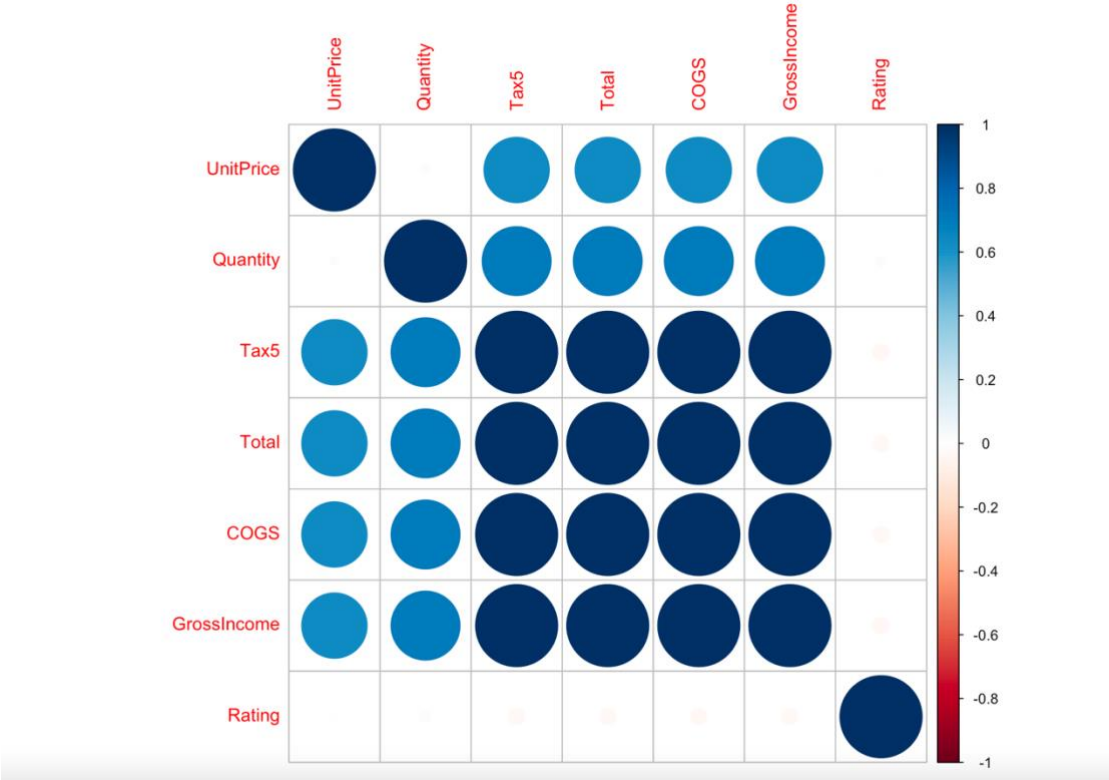
```
        City Freq Rel.Freq
1     Yangon  332      33.2
2 Naypyitaw  328      32.8
3  Mandalay  340      34.0
```

איור X טבלאות שכיחות זהות לעיר וסניף

לאחר צפייה בגרפים ובטבלאות השכיחויות של 'City' ו'Branch' ניתן לראות כי העמודות זהות ולכן ניתן לוותר על אחת מהן לצורך הניתוח. בחרנו להשאיר את עמודה 'City' ולוותר על האחרת.

עמודות נוספות שלא התחשבנו בהן לצורך הניתוח הן: 'ID' ו 'Gross margin percentage'.
עמודת הID מתארת את המס' המזהה של כל חשבונית קנייה, היא ייחודית לכל שורה בנתונים ולכן אין לה כל השפעה על החיזוי.
העמודה ' Gross margin percentage ' הכילה את אותו ערך לאורך כל שורות הנתונים ולכן גם לה אין השפעה על החיזוי.

קשרים בין המשתנים הנומריים לבין משתנה המטרה:



איור XI היט- מאפ של קורלציה בין המשתנים הנומרים למשתנה המטרה

ניתן לראות לפי המפה שישנה בעיית מולטיקולינאריות בין מספר פיצ'רים. הנקודות המסומנות בצבע כחול כהה מייצגות קשר מירבי-חזק (1). האלכסון הוא בעל התאמה מושלמת מכיוון שזה הקשר בין כל פיצ'ר לעצמו. במקרים בהם יש לנו התאמה מושלמת בין פיצ'ר אחד לאחר סימן שהעמודות זהות. מבין 4 הפיצ'רים התואמים השארנו רק את "GrossIncome" להמשך הניתוח והסרנו את האחרים.

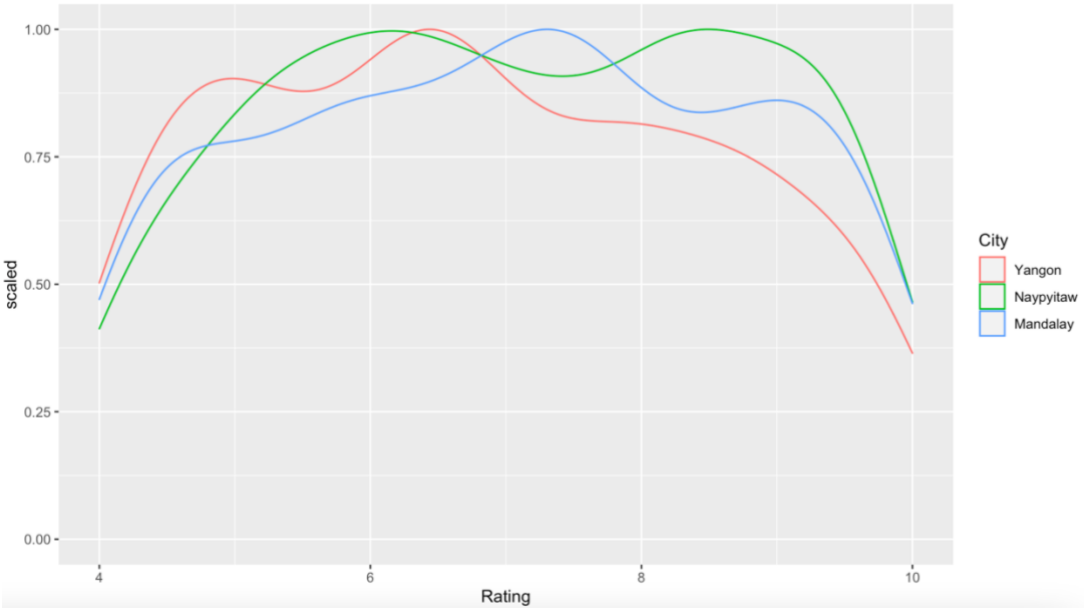
מצפייה בקורלציות ניתן להסיק שלמשתנים הנומרים (שנשארו לאחר ההסרה) השפעה קטנה עד אפסית על הטרגט . ככל שהקשר קרוב יותר ל- 1, הרגרסיה איכותית יותר, כלומר X מסביר את Y באופן טוב יותר. ככל שהקשר קרוב יותר ל- 0, הרגרסיה פחות איכותית ויכולת ההסבר שיש ל X -לגבי Y קטנה יותר.

```
> cor(ss$GrossIncome,ss$Rating)
[1] -0.0364417
> cor(ss$UnitPrice,ss$Rating)
[1] -0.008777507
> cor(ss$Quantity,ss$Rating)
[1] -0.0158149
```

איור XII דוגמה לקורלציות עם משתנה המטרה

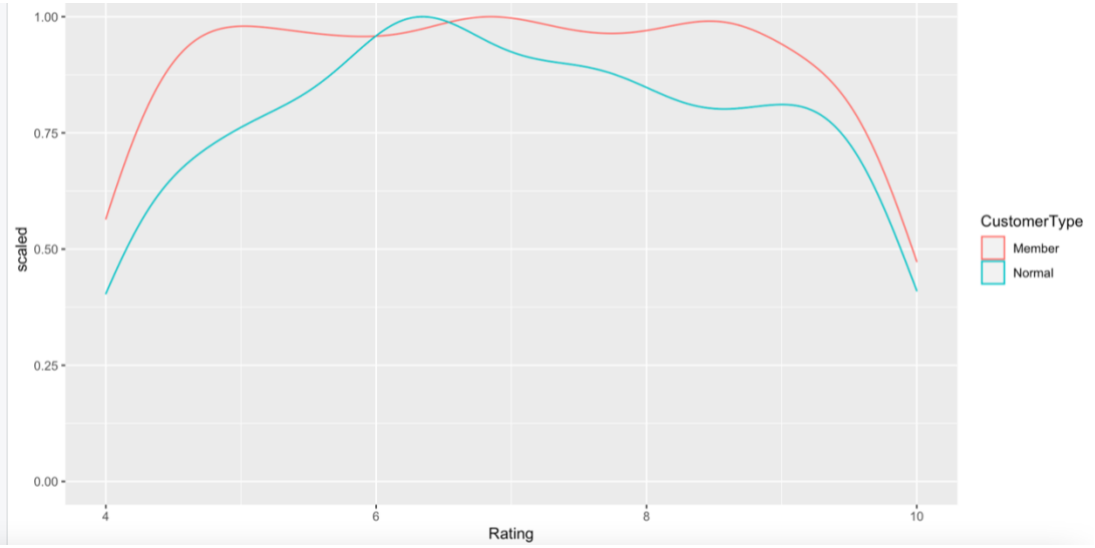
<u>קשרים בין המשתנים הקטגוריאליים לבין משתנה המטרה:</u>

בגרף, מתואר הקשר בין משתנה המטרה רייטינג לפיצ'ר עיר. ניתן לראות בצורה מופשטת את היחס בין הערים לפי מתן רייטינג.



איור XIII גרף עיר- רייטינג

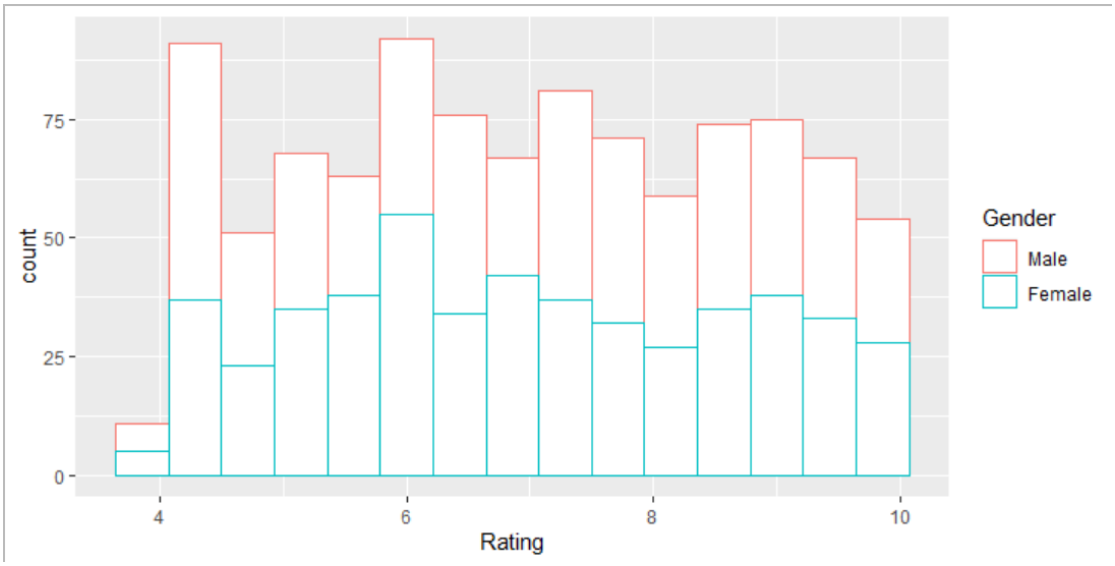
בגרף, מתואר הקשר בין משתנה המטרה רייטינג לפיצ'ר סוג לקוח. אפשר לראות שבאופן כללי חברי מועדון נוטים לתת יותר דירוגים מאשר לקוחות רגילים.



איור XIV גרף סוג לקוח- רייטינג

בגרף עמודות זה מתואר הקשר בין משתנה המטרה רייטינג לפיצ'ר מגדר:

גם כאן ניתן לראות את האיזון בין גברים לנשים- בכל עמודה בר החלקים הכחול והורוד קרובים מאוד בגודל.



איור XV גרף מגדר- רייטינג

3 . שלב המידול:

<u>רגרסיה לינארית</u>

לאחר שויתרנו על עמודות לא רלוונטיות (בסעיף 2) נשארנו עם הפיצ'רים הבאים :

("City","CustomerType","Gender","ProductLine",

("Date","Time","Payment","Gross Income", "Rating"

לצורך הרצת מודל רגרסיה לינארית.

בשלב הראשון, חילקנו את הנתונים באופן אקראי לקבוצות אימון ובקרה. 70% ו30% בהתאמה.

והתוצאות שקיבלנו מהמודל נראו כך-

```
Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                      5.211e+01  4.741e+01   1.099   0.2721
CityNaypyitaw                    1.782e-01  1.609e-01   1.107   0.2685
CityMandalay                     5.395e-02  1.597e-01   0.338   0.7356
CustomerTypeNormal               1.315e-01  1.307e-01   1.006   0.3149
GenderFemale                     7.402e-02  1.317e-01   0.562   0.5744
ProductLineElectronic accessories  1.890e-01  2.221e-01   0.851   0.3953
ProductLineHome and lifestyle    1.608e-01  2.261e-01   0.711   0.4774
ProductLineSports and travel     2.960e-01  2.276e-01   1.300   0.1940
ProductLineFood and beverages   -7.584e-04  2.277e-01  -0.003   0.9973
ProductLineFashion accessories  -5.449e-02  2.257e-01  -0.241   0.8093
Date                            -2.885e-08  3.059e-08  -0.943   0.3459
Time                            -1.018e+00  5.025e-01  -2.025   0.0432 *
PaymentCash                      6.580e-02  1.606e-01   0.410   0.6822
PaymentCredit Card               1.431e-01  1.582e-01   0.905   0.3657
GrossIncome                     -3.552e-03  5.687e-03  -0.625   0.5324
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.72 on 685 degrees of freedom
Multiple R-squared:  0.01794,   Adjusted R-squared:  -0.002134
F-statistic: 0.8937 on 14 and 685 DF,  p-value: 0.5657
```

טיב החיזוי היה נמוך ולכן החלטנו להריץ רגרסיה לינארית מחדש עם כל הפיצ'רים.

ניתן לראות שכאשר הרצנו את הרגרסיה הלינארית ללא סינון מעמיק של פיצ'רים, קיבלנו שורות שחלקן מכילות ערכים חסרים מה שעלול להצביע על מולטיקולינאריות (שהסברנו אותה קודם) בין פיצ'רים ושיבוש תוצאות המודל. כמו כן, ניתן לראות שיש פיצ'רים מובהקים סטטיסטית.
במידול זה, ניתן לראות שהמודל מסביר 2.46% מהשונות של y.

```
Call:
lm(formula = Rating ~ ., data = train.set)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.3568 -1.4532 -0.0134  1.4100  3.3715 

Coefficients: (5 not defined because of singularities)
                                   Estimate Std. Error t value Pr(>|t|)
(Intercept)                       51.168666  47.343635   1.081   0.2802
BranchB                           -0.052512   0.159452  -0.329   0.7420
BranchC                            0.126614   0.161232   0.785   0.4326
CityNaypyitaw                            NA         NA      NA       NA
CityMandalay                             NA         NA      NA       NA
CustomerTypeNormal                 0.151815   0.130809   1.161   0.2462
GenderFemale                       0.073299   0.131662   0.557   0.5779
ProductLineElectronic accessories  0.200836   0.222558   0.902   0.3672
ProductLineHome and lifestyle      0.157463   0.226105   0.696   0.4864
ProductLineSports and travel       0.286397   0.227258   1.260   0.2080
ProductLineFood and beverages      0.007959   0.227296   0.035   0.9721
ProductLineFashion accessories    -0.041544   0.225692  -0.184   0.8540
UnitPrice                          0.009636   0.005157   1.869   0.0621 .
Quantity                           0.110038   0.050969   2.159   0.0312 *
Tax5                              -0.037051   0.016638  -2.227   0.0263 *
Total                                    NA         NA      NA       NA
Date                              -0.002472   0.002639  -0.937   0.3492
Time                              -1.059996   0.502035  -2.111   0.0351 *
PaymentCash                        0.078823   0.160752   0.490   0.6241
PaymentCredit Card                 0.148382   0.157929   0.940   0.3478
COGS                                     NA         NA      NA       NA
GrossIncome                              NA         NA      NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.716 on 683 degrees of freedom
Multiple R-squared:  0.02469,	Adjusted R-squared:  0.001843 
F-statistic: 1.081 on 16 and 683 DF,  p-value: 0.3696
```

בשלב השני, רצינו לעשות את סינון הפיצ'רים בדרך אחרת ולכן השתמשנו בפיצ'רים שתורמים לנו הכי הרבה במודל. לשם כך לקחנו את הערכים המובהקים ברמה של לפחות 10% והרצנו שוב את המודל.
ניתן לראות כי השונות המוסברת של y לאחר הורדת הפיצ'רים היא 1.26%. מה שאומר שהתרומה של הפיצ'רים שהורדנו מהמודל המלא, הייתה מזערית.

```
Call:
lm(formula = Rating ~ UnitPrice + Quantity + Tax5 + Time, data = train.set)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.3308 -1.4754 -0.0482  1.4722  3.1969 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)  
(Intercept) -2.553e+04  1.261e+04  -2.024   0.0433 *
UnitPrice    9.470e-03  5.111e-03   1.853   0.0643 .
Quantity     1.082e-01  5.060e-02   2.139   0.0328 *
Tax5        -3.688e-02  1.653e-02  -2.231   0.0260 *
Time        -1.156e-05  5.710e-06  -2.025   0.0433 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.712 on 695 degrees of freedom
Multiple R-squared:  0.01261,	Adjusted R-squared:  0.006923 
F-statistic: 2.218 on 4 and 695 DF,  p-value: 0.06551
```

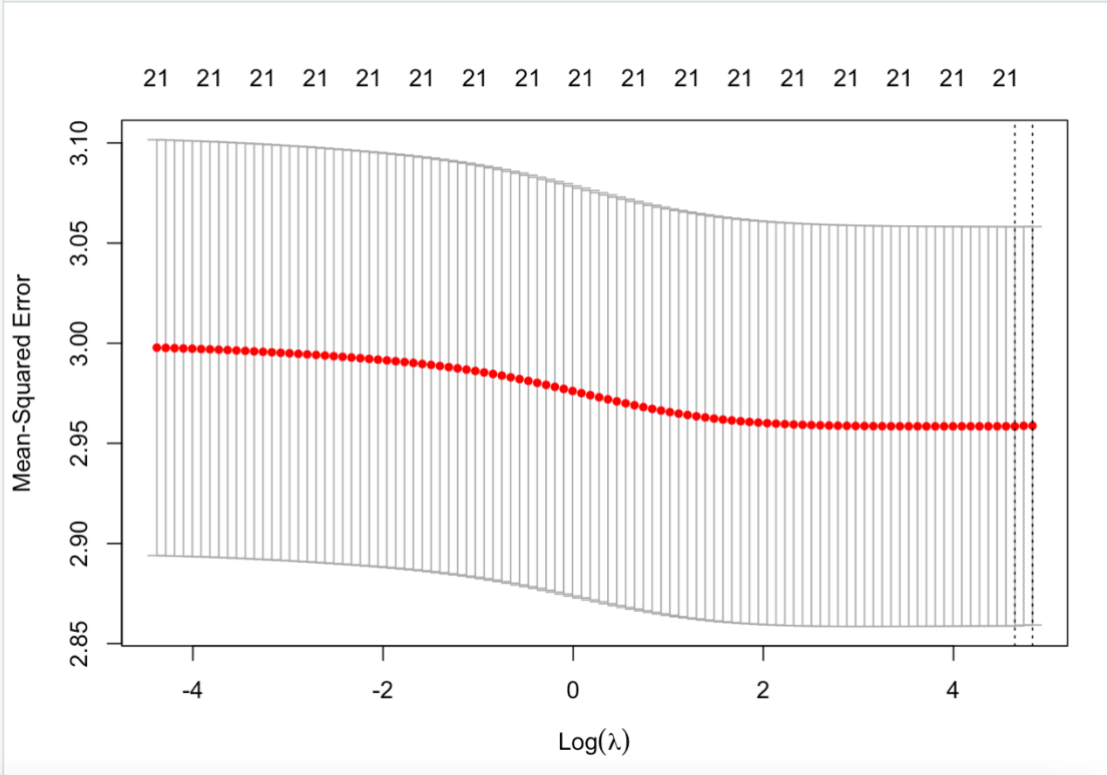
בגלל שבין מודל למודל יש הבדלים במספר הפיצ’רים שהורצו- השוונו בינהם לפי הadjust R שמייצג בצורה הטובה ביותר מי המודל ”המנצח” מבין השניים. במקרה זה- המודל שהורץ עם מעט פיצ’רים הוא המנצח.

Ridge

בשל בעיית המולטיקולינאריות שגילינו, המודלים הטובים ביותר שיעזרו לנו הם לאסו ורידג'.

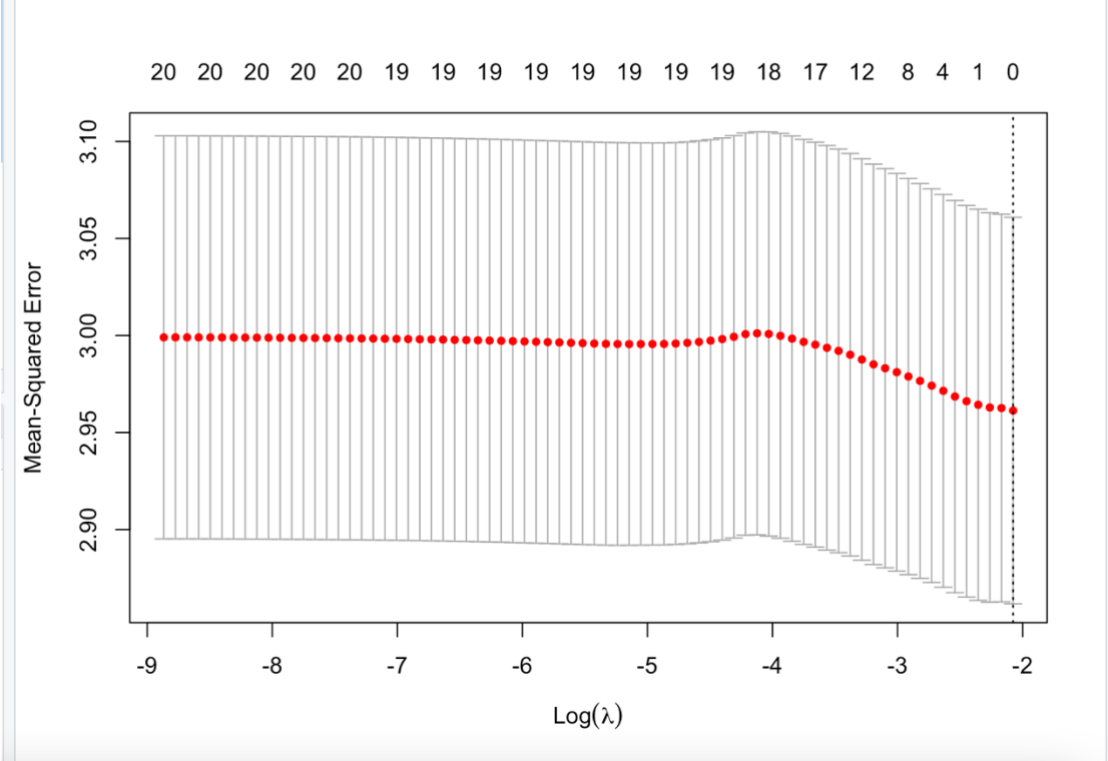
בשביל למצוא את הלמדא הטובה ביותר מבצעים validation cross ובסוף למדא נבחרת לפי ערך השגיאה הנמוך ביותר. במקרה זה, הלוג למדא שעבורה ערך השגיאה הוא הנמוך ביותר =4.645528 ו

.mse=2.968339



איור XVI מודל הרידג

Lasso



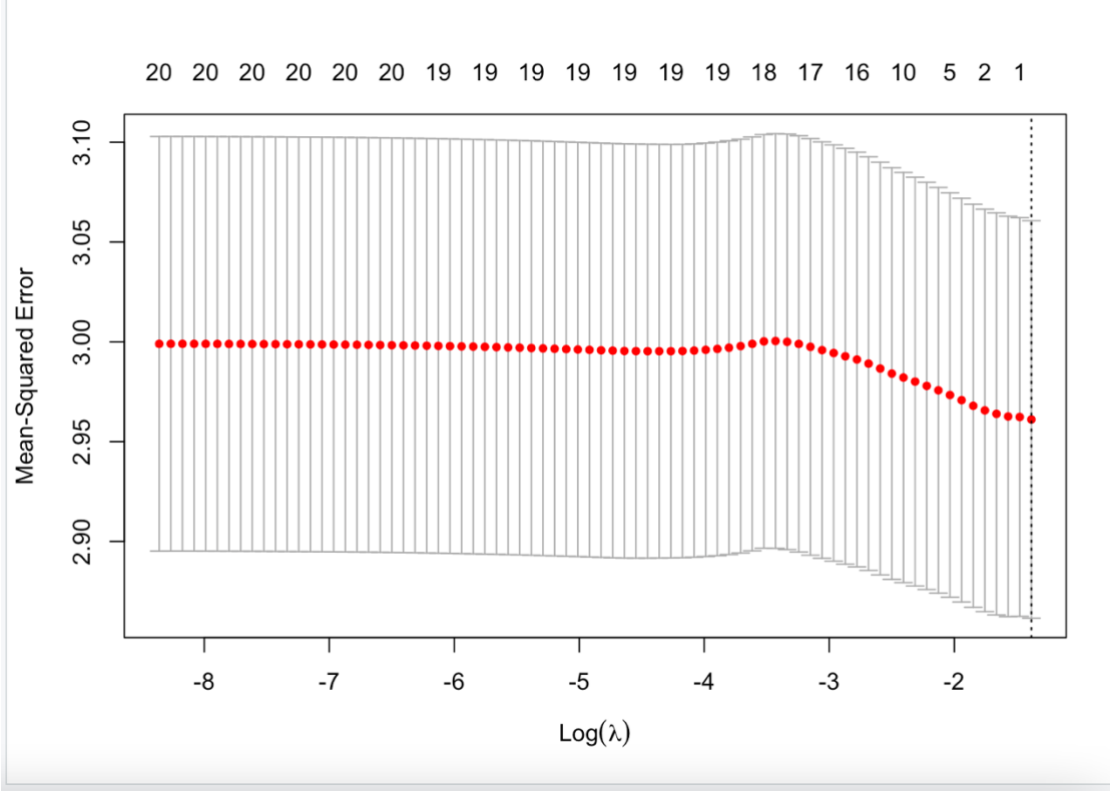
איור XVII מודל הלאסו

לוג למדא: --2.07616

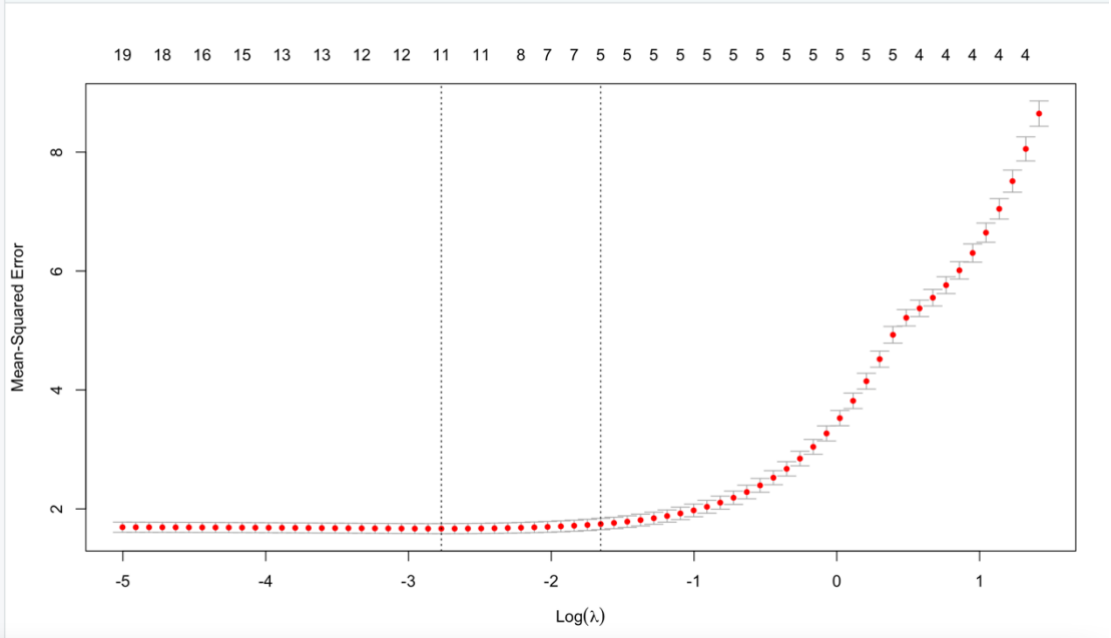
mse: 2.968132

Elastic net

מודל זה, הוא שילוב של רידג' ולאסו.



איור XVIII Elastic net



לוג למדא: -1.383012

2.968132 -MSE

RF

```
Call:
 randomForest(formula = Rating ~ ., data = train.set, mtry = 4)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 4

          Mean of squared residuals: 3.084425
                    % Var explained: -4.67
```

4 . מדד ה – MSE של כל המודלים והקשר בין התצפיות החזויות לבין הערכים האמיתיים

	model	mse
2	Ridge Regression	2.968
3	Lasso	2.968
4	Elastic Net (alpha=0.5)	2.968
1	Linear Regression	3.044
5	Random Forest	3.199

איור XIX טבלת סיכום MSE

5 .מסקנות:

באופן כללי, ניכר כי טיב החיזוי נמוך ולא טוב, ואף אחד מהמודלים הנ"ל לא מסביר את אחוז השגיאה בצורה מיטבית או אפילו מספקת. הסיבות לחוסר ההצלחה בחיזוי הן:

1. קשר חזק מאוד בין משתנים גרם למולטיקולינאריות ולשיבוש מודל הרגרסיה בהרצה הראשונית עם כל הפיצ'רים.

2. קורלציה נמוכה מאוד עד אפסית בין הפיצ'רים הנומריים למשתנה המטרה גרמה ככל הנראה לחיזוי לא טוב.

ניתן לראות בטבלת הסיכום שה MSE הטוב ביותר הוא של המודלים רידג', לאסו ואלסטיק-נט (באופן שווה). הם התגברו על המולטיקולינאריות שהצגנו ולכן קיבלו תוצאה טובה יותר ממודל הרגרסיה הלינארית. התוצאה הזהה של רידג' ולאסו ממחישה את הקשר החלש בין הפיצ'רים למשתנה המטרה שכן בלאסו נעשה סינון של פיצ'רים ועדיין התקבל אותו מדד mse כשל רידג' בו לא הוסרו פיצ'רים.

רשימת תרשימים