**מעבדה בניתוח נתונים עם R - חלק ב'**

מגישות:

רוני מליחי 315037747

זהר בן משה 206180754

מרצה: מר זכאי אבי

תוכן עניינים

[1 .תיאור הבעיה: 3](#_Toc64838019)

[קישור לנתונים 3](#_Toc64838020)

[רקע: 3](#_Toc64838021)

[תיאור משתנה המטרה RATING- 3](#_Toc64838022)

[Categorical Attributes- 3](#_Toc64838023)

[Numerical Attributes- 3](#_Toc64838024)

[מטרת המחקר 4](#_Toc64838025)

[כך נראים הנתונים 4](#_Toc64838026)

[2 .EDA - סטטיסטיקה תיאורית מלאה של משתנה המטרה ושל הפיצ'רים: 5](#_Toc64838027)

[משתנה המטרה: 5](#_Toc64838028)

[המשתנים: 6](#_Toc64838029)

[קשרים בין המשתנים הנומריים לבין משתנה המטרה: 9](#_Toc64838030)

[קשרים בין המשתנים הקטגוריאליים לבין משתנה המטרה: 10](#_Toc64838031)

[3 . שלב המידול: 12](#_Toc64838032)

[רגרסיה לינארית 12](#_Toc64838033)

[Ridge 15](#_Toc64838034)

[Lasso 16](#_Toc64838035)

[Elastic net 17](#_Toc64838036)

[RF 18](#_Toc64838037)

[4 . מדד ה – MSE של כל המודלים והקשר בין התצפיות החזויות לבין הערכים האמיתיים 19](#_Toc64838038)

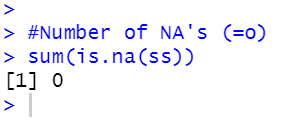
[5 .מסקנות: 19](#_Toc64838039)

[רשימת תרשימים 20](#_Toc64838040)

# 1 .תיאור הבעיה:

## [קישור לנתונים](https://www.kaggle.com/aungpyaeap/supermarket-sales)

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:

## תיאור משתנה המטרה RATING-

הרייטינג על סקאלה של 4-10 בקפיצות של 0.1, הוא מתאר דירוג שהלקוח נותן על חווית 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

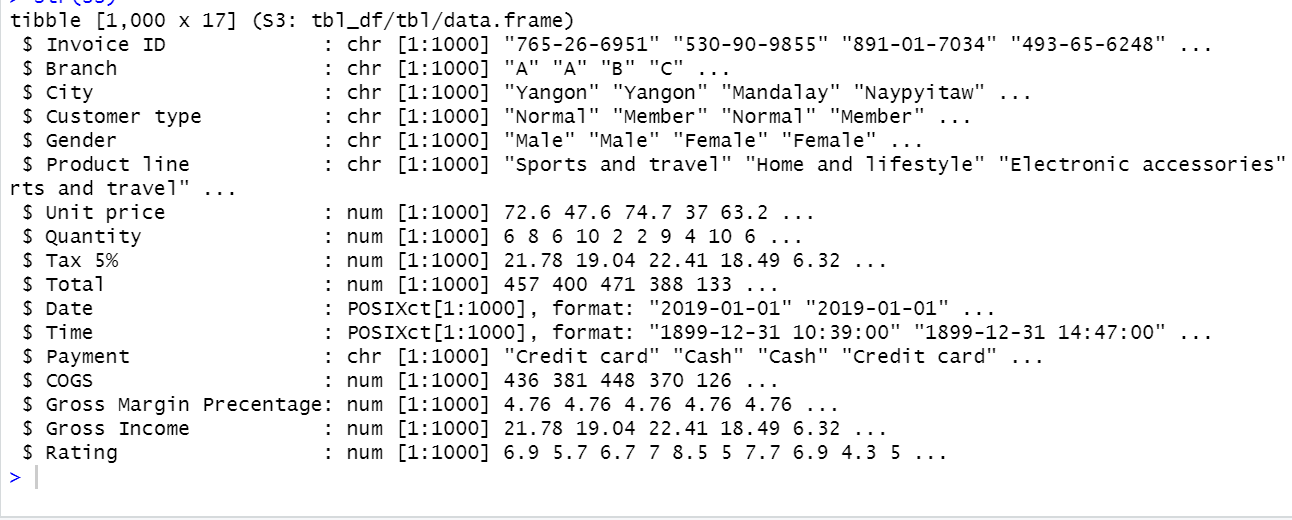
מטרת המחקר - חיזוי הערך של הרייטינג (משתנה מטרה כמותי) באמצעות המשתנים האחרים ומציאת המודל המתאים ביותר לחיזוי.

כך נראים הנתונים:

A picture containing text, monitor, electronics, computer

Description automatically generated

איור I מספר שורות נתונים לדוגמא מהאקסל

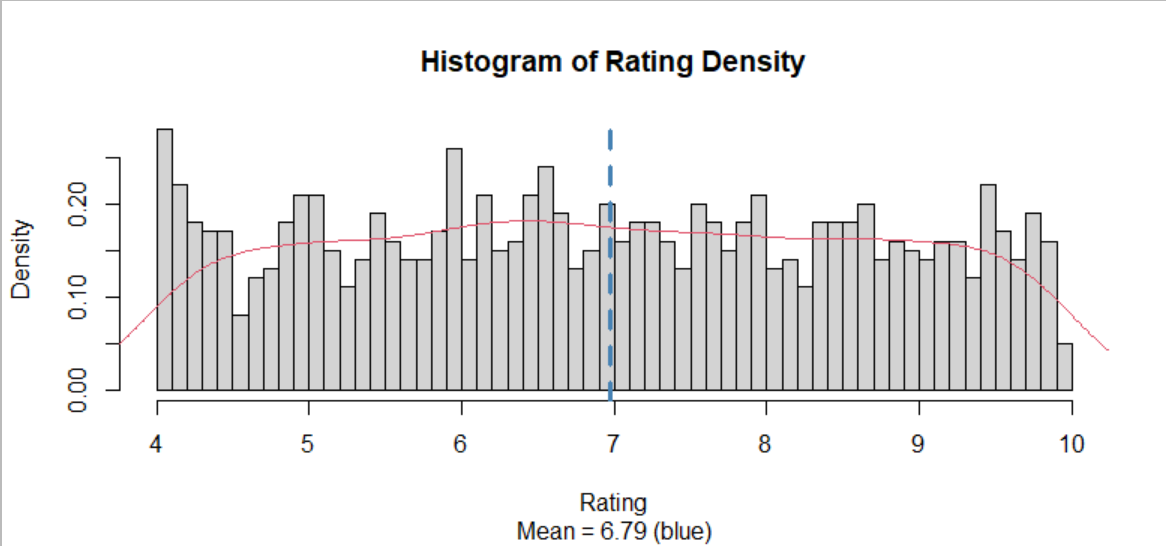
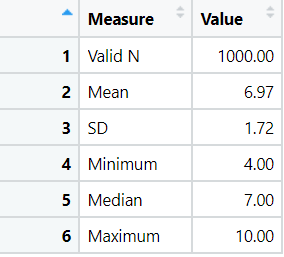


איור II הצגת הנתונים בR

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

## משתנה המטרה:

תחילה, הסתכלנו על ההתפלגות של צפיפות משתנה המטרה:

איור III היסטוגרמת צפיפות של 'רייטינג'

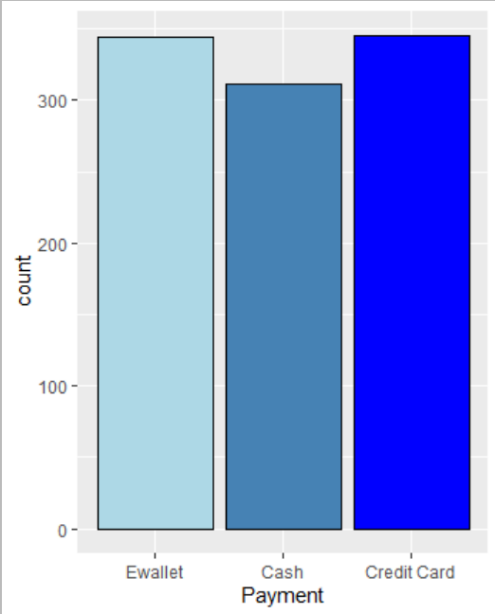
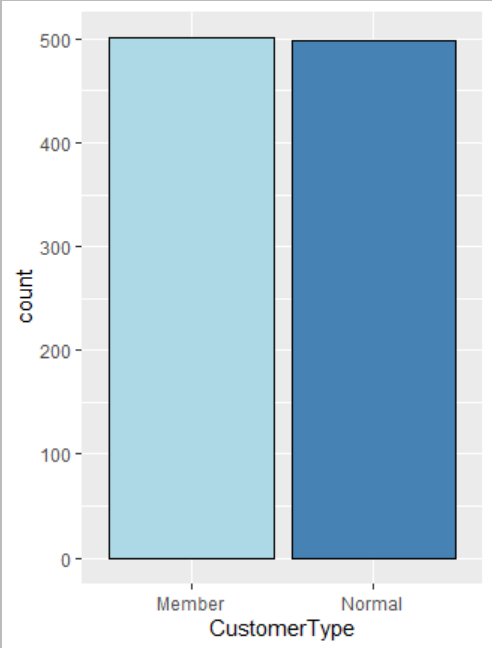
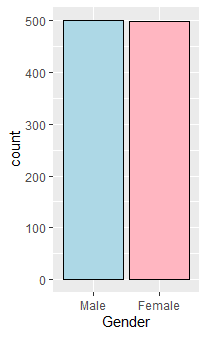
איור IV מדדי משתנה המטרה

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

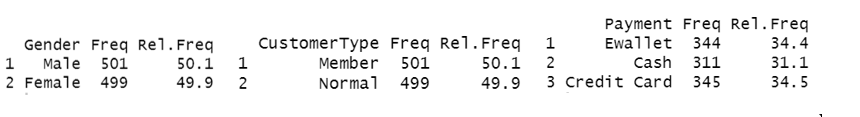
## המשתנים:

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

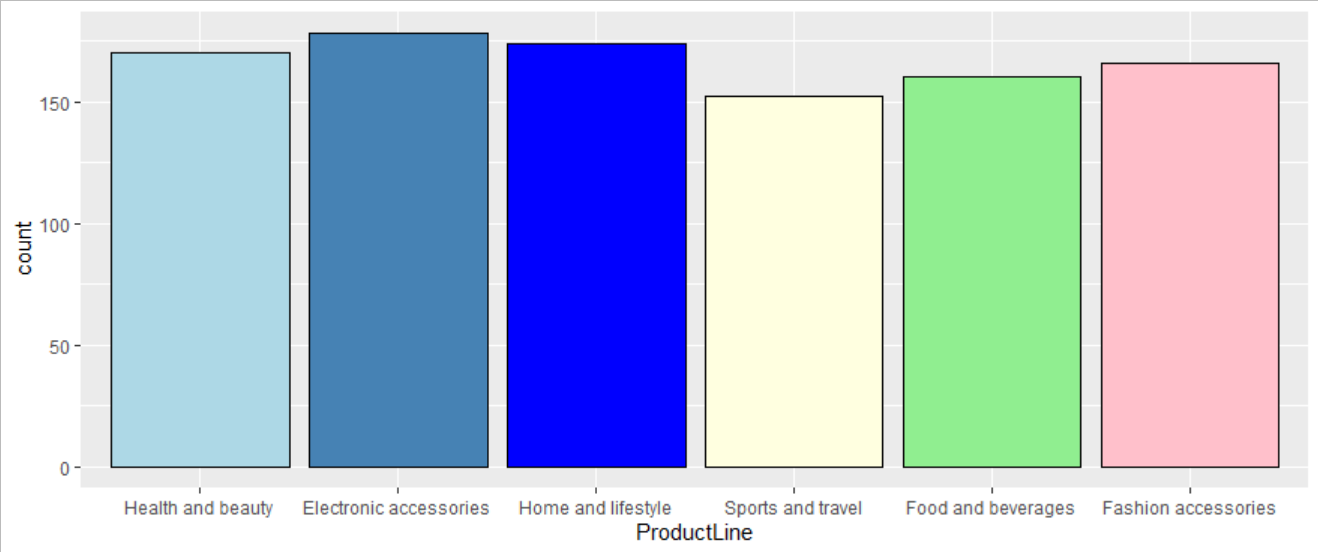
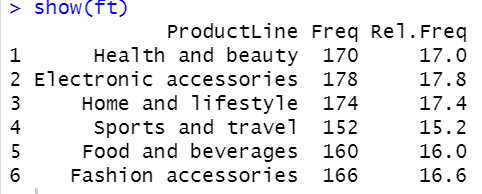
איור V



* התפלגות הפיצ'רים: סוג תשלום, סוג לקוח ומגדר

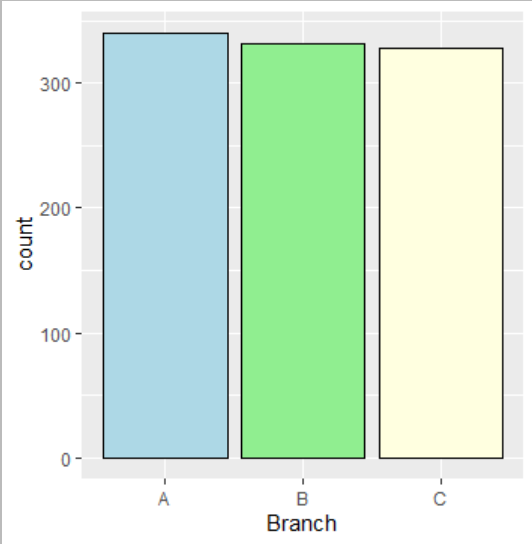
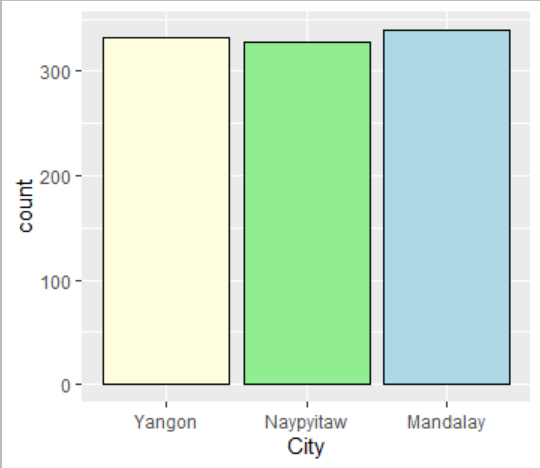


איור VI- טבלאות שכיחויות של הפיצ'רים: סוג תשלום, סוג לקוח ומגדר

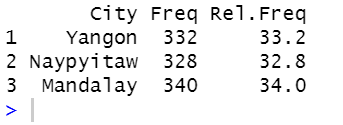


איור VII טבלת שכיחויות של הפיצ'ר מחלקה

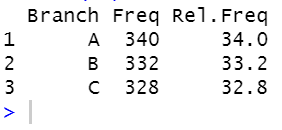
איור VIII התפלגות הפיצ'ר: מחלקה



איור IX התפלגות זהה של הפיצ'רים עיר וסניף



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.

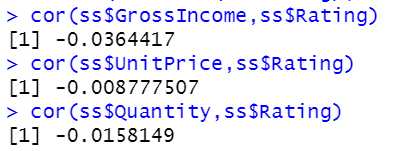
## קשרים בין המשתנים הנומריים לבין משתנה המטרה:

Chart, bubble chart

Description automatically generated

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.



איור XII דוגמה לקורלציות עם משתנה המטרה

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.

## קשרים בין המשתנים הקטגוריאליים לבין משתנה המטרה:

Chart, line chart

Description automatically generated   
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.

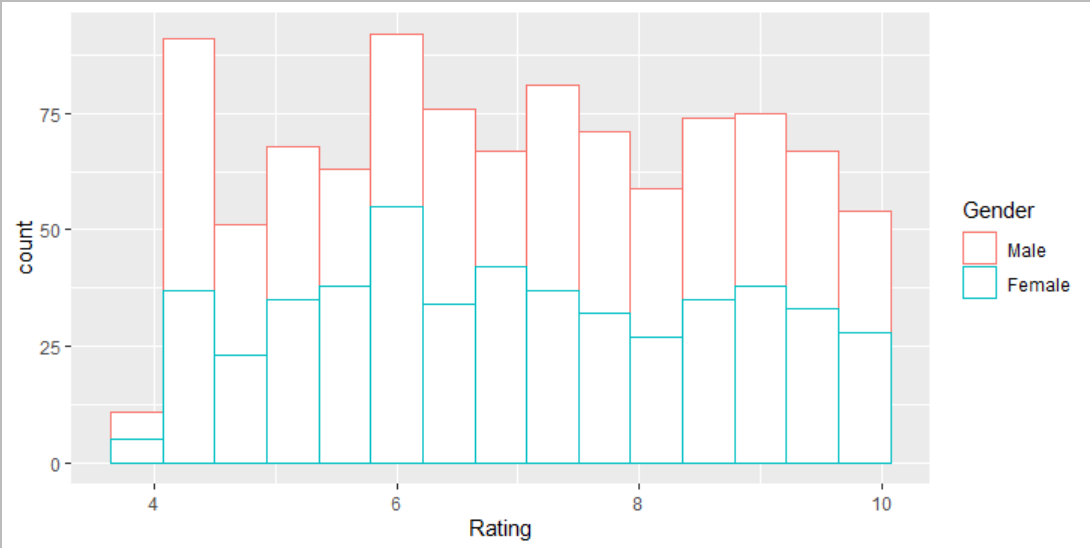
Chart, line chart

Description automatically generated

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:

## רגרסיה לינארית

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

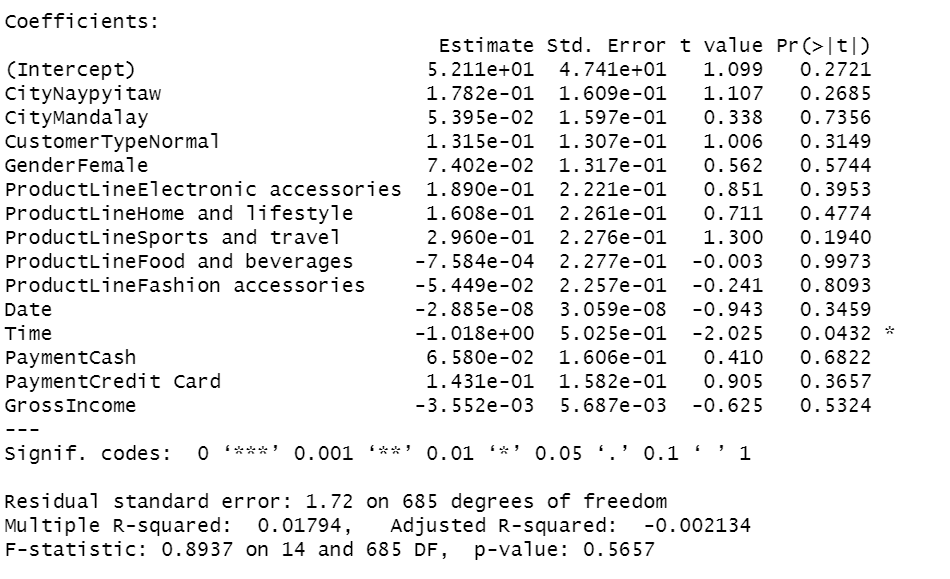
("City","CustomerType","Gender","ProductLine",

"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-



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.

A picture containing text, receipt

Description automatically generated

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.

Table

Description automatically generated

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 cross validation ובסוף למדא נבחרת לפי ערך השגיאה הנמוך ביותר. במקרה זה, הלוג למדא שעבורה ערך השגיאה הוא הנמוך ביותר =4.645528 ו

mse=2.968339. Chart, histogram

Description automatically generated

איור XVI מודל הרידג

## Lasso

Chart

Description automatically generated

איור XVII מודל הלאסו

לוג למדא: --2.07616

mse: 2.968132

## Elastic net

מודל זה, הוא שילוב של רידג׳ ולאסו. Chart

Description automatically generated

איור XVIII Elastic net

Chart

Description automatically generated with medium confidence

לוג למדא: -1.383012

MSE- 2.968132

## RF

Text, letter

Description automatically generated

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

Text

Description automatically generated

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 שהMSE 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

[איור I מספר שורות נתונים לדוגמא מהאקסל 4](file:///C:\Users\zohar\Desktop\מעבדה%20בניתוח%20נתונים%20עם%20R%20-%20חלק%20ב-3.docx#_Toc64838000)

[איור II הצגת הנתונים בR 4](file:///C:\Users\zohar\Desktop\מעבדה%20בניתוח%20נתונים%20עם%20R%20-%20חלק%20ב-3.docx#_Toc64838001)

[איור III היסטוגרמת צפיפות של 'רייטינג' 5](file:///C:\Users\zohar\Desktop\מעבדה%20בניתוח%20נתונים%20עם%20R%20-%20חלק%20ב-3.docx#_Toc64838002)

[איור IV מדדי משתנה המטרה 5](file:///C:\Users\zohar\Desktop\מעבדה%20בניתוח%20נתונים%20עם%20R%20-%20חלק%20ב-3.docx#_Toc64838003)

[איור V 6](file:///C:\Users\zohar\Desktop\מעבדה%20בניתוח%20נתונים%20עם%20R%20-%20חלק%20ב-3.docx#_Toc64838004)

[איור VI- טבלאות שכיחויות של הפיצ'רים: סוג תשלום, סוג לקוח ומגדר 6](#_Toc64838005)

[איור VII טבלת שכיחויות של הפיצ'ר מחלקה 7](file:///C:\Users\zohar\Desktop\מעבדה%20בניתוח%20נתונים%20עם%20R%20-%20חלק%20ב-3.docx#_Toc64838006)

[איור VIII התפלגות הפיצ'ר: מחלקה 7](#_Toc64838007)

[איור IX התפלגות זהה של הפיצ'רים עיר וסניף 7](#_Toc64838008)

[איור X טבלאות שכיחות זהות לעיר וסניף 7](file:///C:\Users\zohar\Desktop\מעבדה%20בניתוח%20נתונים%20עם%20R%20-%20חלק%20ב-3.docx#_Toc64838009)

[איור XI היט- מאפ של קורלציה בין המשתנים הנומרים למשתנה המטרה 9](file:///C:\Users\zohar\Desktop\מעבדה%20בניתוח%20נתונים%20עם%20R%20-%20חלק%20ב-3.docx#_Toc64838010)

[איור XII דוגמה לקורלציות עם משתנה המטרה 10](file:///C:\Users\zohar\Desktop\מעבדה%20בניתוח%20נתונים%20עם%20R%20-%20חלק%20ב-3.docx#_Toc64838011)

[איור XIII גרף עיר- רייטינג 10](file:///C:\Users\zohar\Desktop\מעבדה%20בניתוח%20נתונים%20עם%20R%20-%20חלק%20ב-3.docx#_Toc64838012)

[איור XIV גרף סוג לקוח- רייטינג 11](#_Toc64838013)

[איור XV גרף מגדר- רייטינג 11](#_Toc64838014)

[איור XVI מודל הרידג 15](#_Toc64838015)

[איור XVII מודל הלאסו 16](#_Toc64838016)

[איור XVIII Elastic net 17](#_Toc64838017)

[איור XIX טבלת סיכום MSE 19](#_Toc64838018)