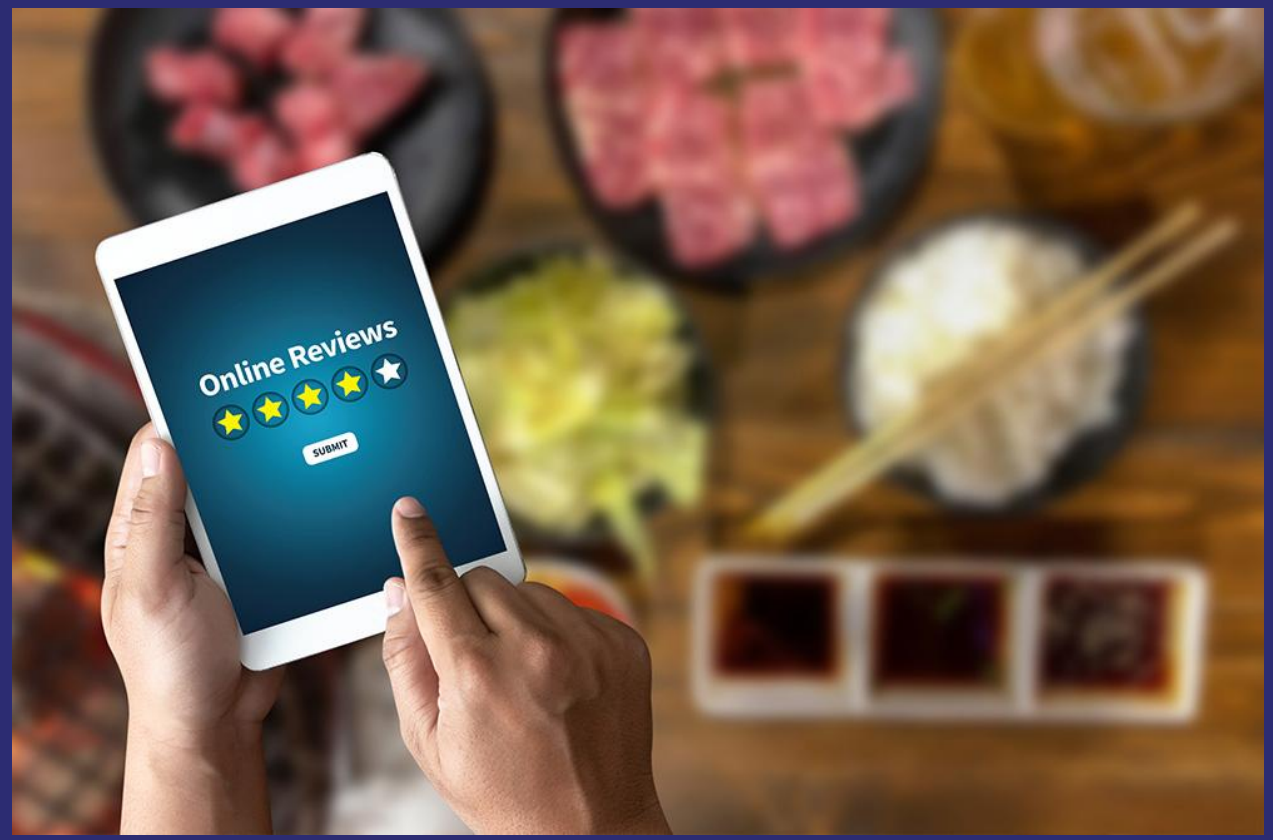
מבוא למדעי הנתונים - פרויקט גמר

# Restuarant4u

Idan Biton | Asaf Tzabary

# הקדמה




- במשך השנים האחרונות, דירוגי מסעדות הפכו לאחד המדדים החשובים ביותר לבחון את איכות המסעדה.
- הדירוגים של המסעדות משפיעים על ההכנסות והצלחתן. ככל שהדירוג של המסעדה טוב יותר, כך יותר אנשים יבחרו לבקר בה. עם זאת, הדירוג לא משפיע רק על ההכנסות, אלא גם על המוניטין והאמינות של המסעדה בקרב הלקוחות.
- בעזרת ניתוח נתוני המסעדות, אנו יכולים להבין את הקשר בין המאפיינים של המסעדה לבין הדירוג שלה.

# שאלת המחקר

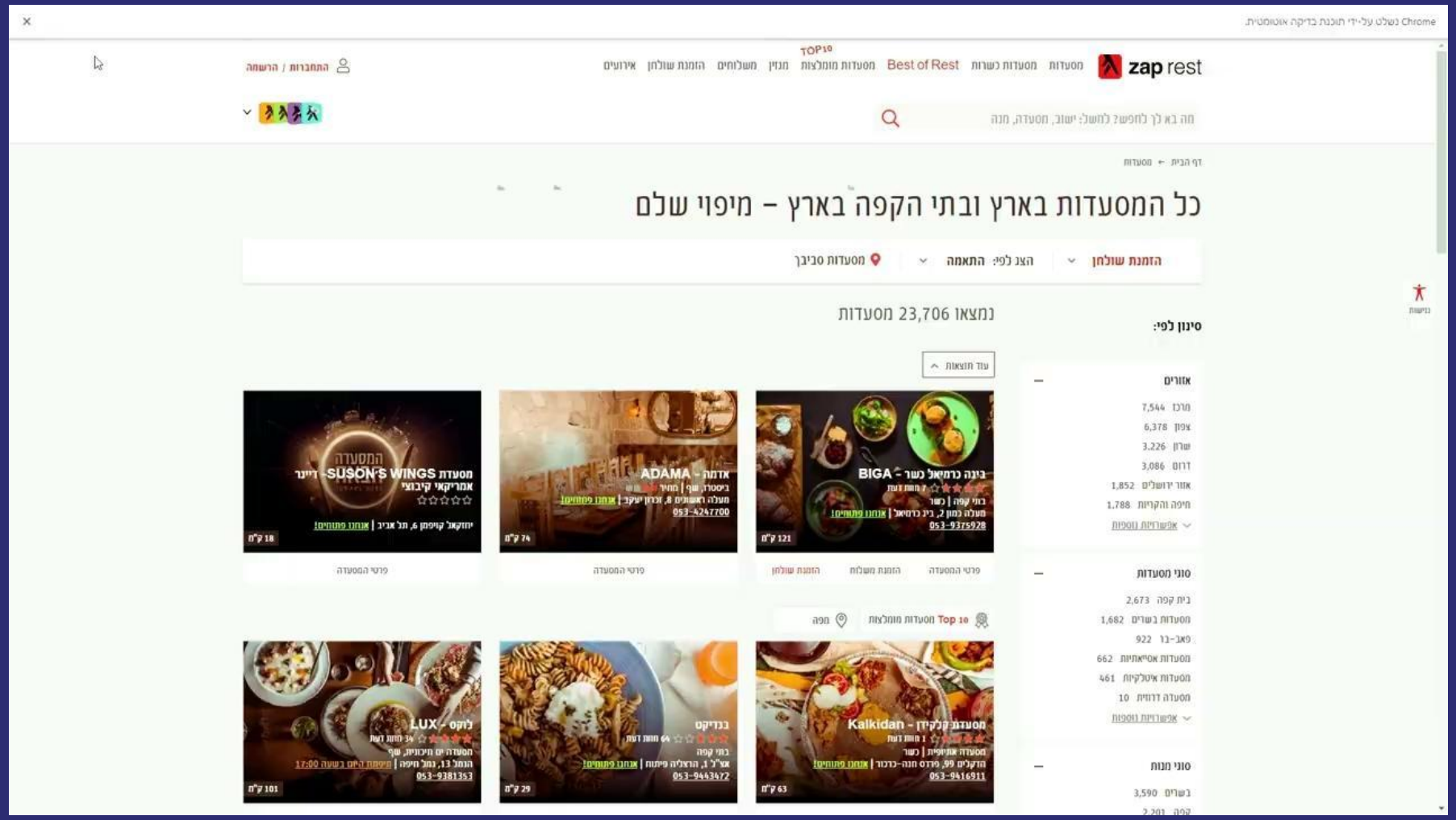
איך משפיעים מאפיינים שונים על דירוג המסעדה ,
ואיך ניתן לחזות דירוג זה ?

# מקור הנתונים והרכשה

https://www.rest.co.il/

- שימוש באתר zaprest כמקור להרכשת הנתונים של דירוגי המסעדות בישראל.
- כיום באתר נמצאים מעל ל 23,000 מסעדות.
- חילוץ כלל המסעדות מן האתר, ולאחר מכן חילוץ מאפייניה והדירוג של כל אחת מן המסעדות.

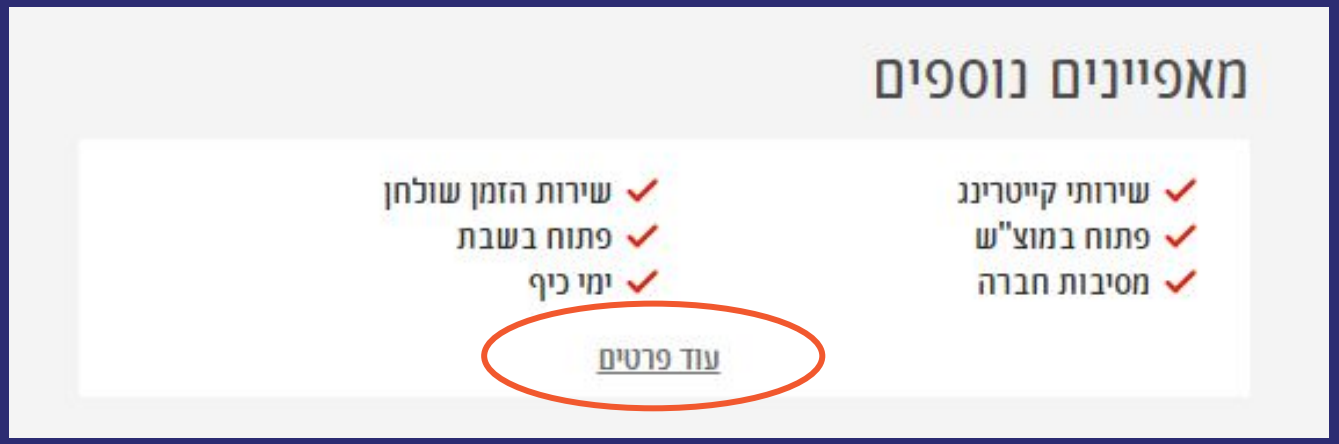
# דרכי הרכשת הנתונים




שימוש בספריית **Selenium** לצורך שליטה בדפדפן וחילוץ הנתונים מדפי המסעדות.

# דרכי הרכשת הנתונים

הקושי בהרכשת המאפיינים מדפי המסעדות, היה הדמייה של לחיצת משתמש על כפתור "עוד פרטים" על מנת לחשוף את כלל מאפייני המסעדה.




# ניתוח וניקוי הנתונים

לאחר הרכשת הנתונים ויצירת DataFrame, ביצענו ניקוי נתונים על פי הפרמטרים הבאים:

- הסרנו מסעדות ללא דירוג(NaN values).
- הסרנו מאפיינים בעלי תדירות הופעה נמוכה אצל מסעדות.
- הסרת מאפיינים שאינם נומריים/קטגוריאלים.

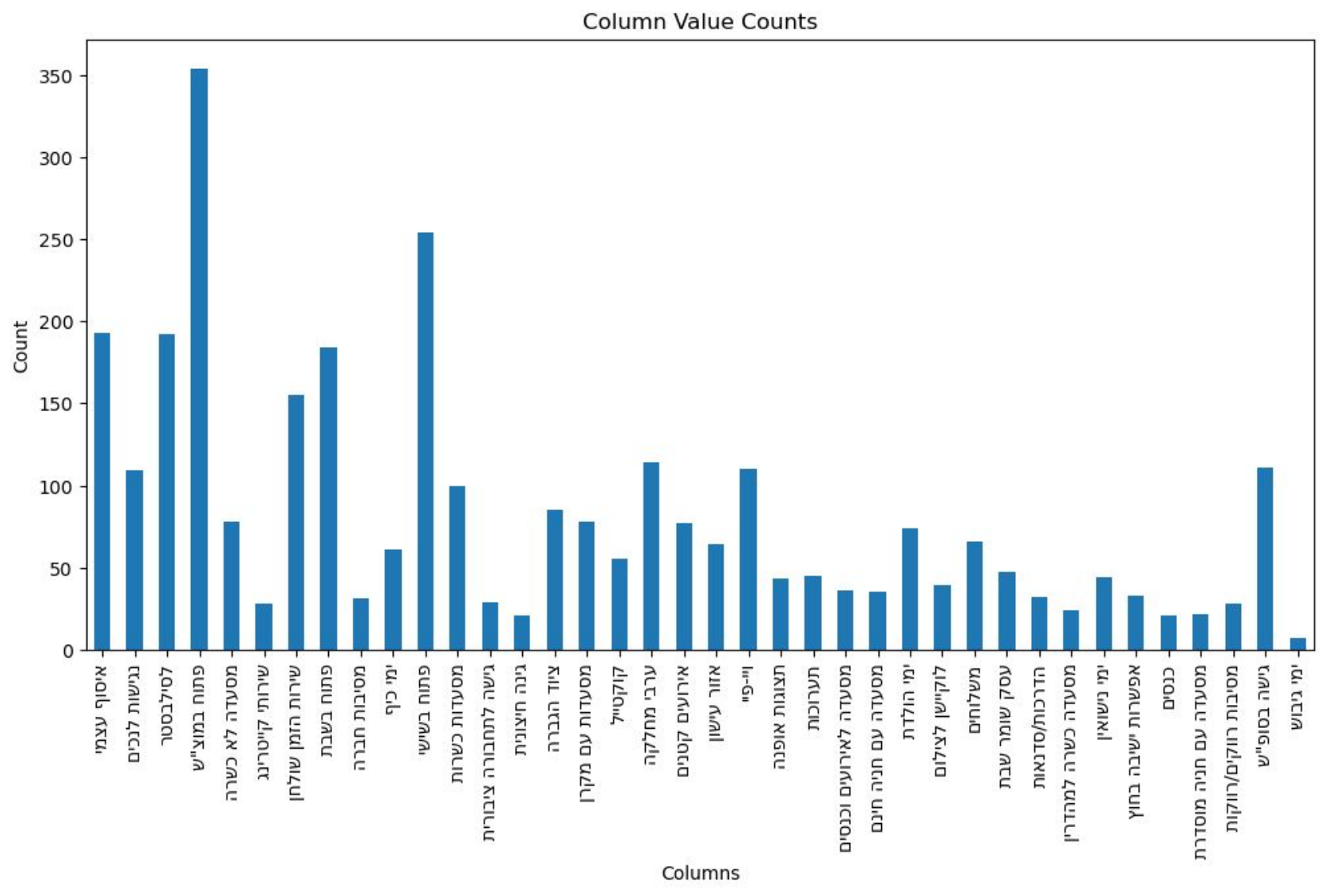
**מאפיינים חסרי חשיבות**

| | | |
|---|---|---|
| מסעדה עם מרתף יינות | חדר פרטי | Burger club |

**מאפיינים בעלי חשיבות**

| | | |
|---|---|---|
| פתוח במוצ"ש | אפשרות ישיבה בחוץ | נגישות לנכים |

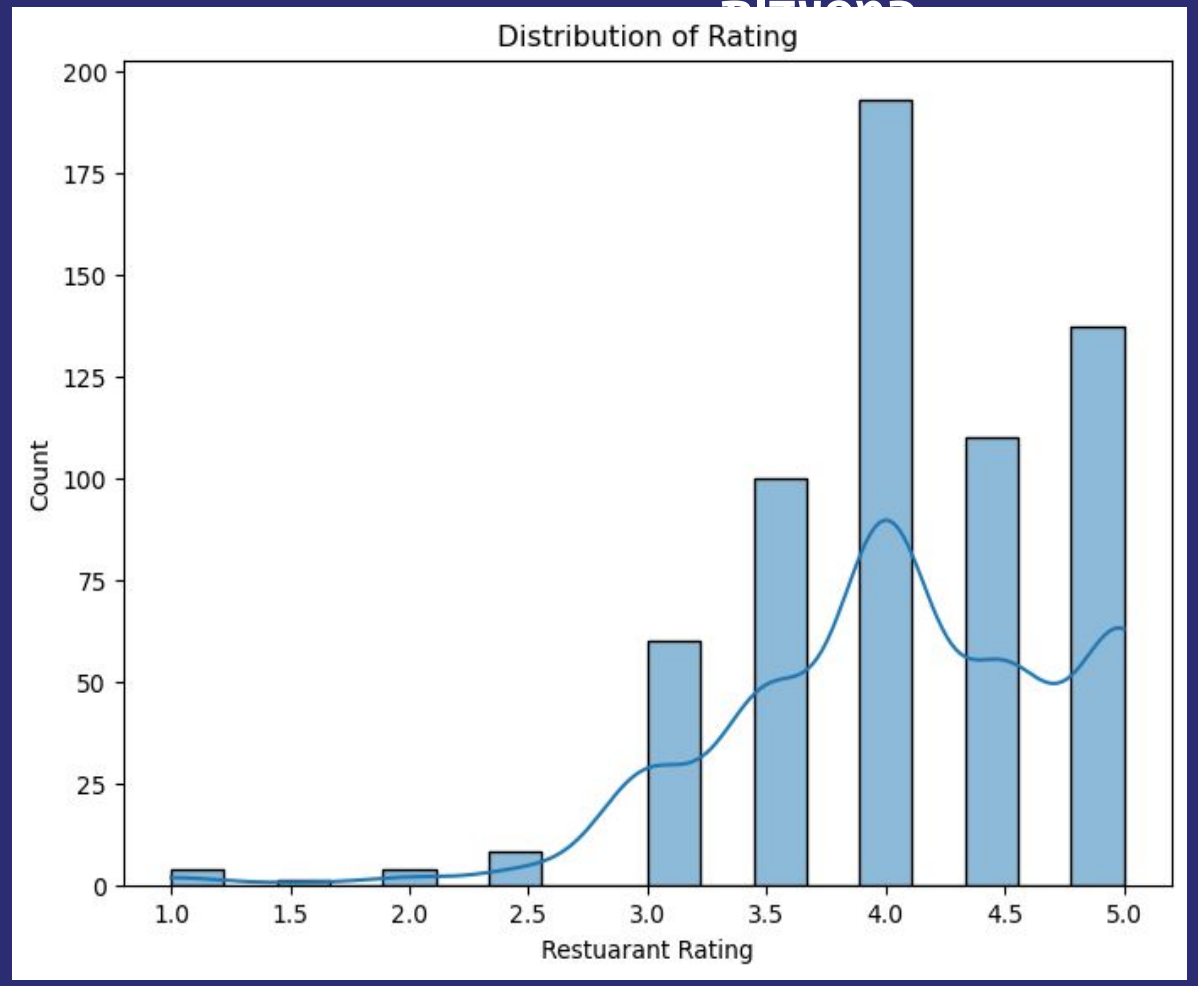
# ויזואליזציה



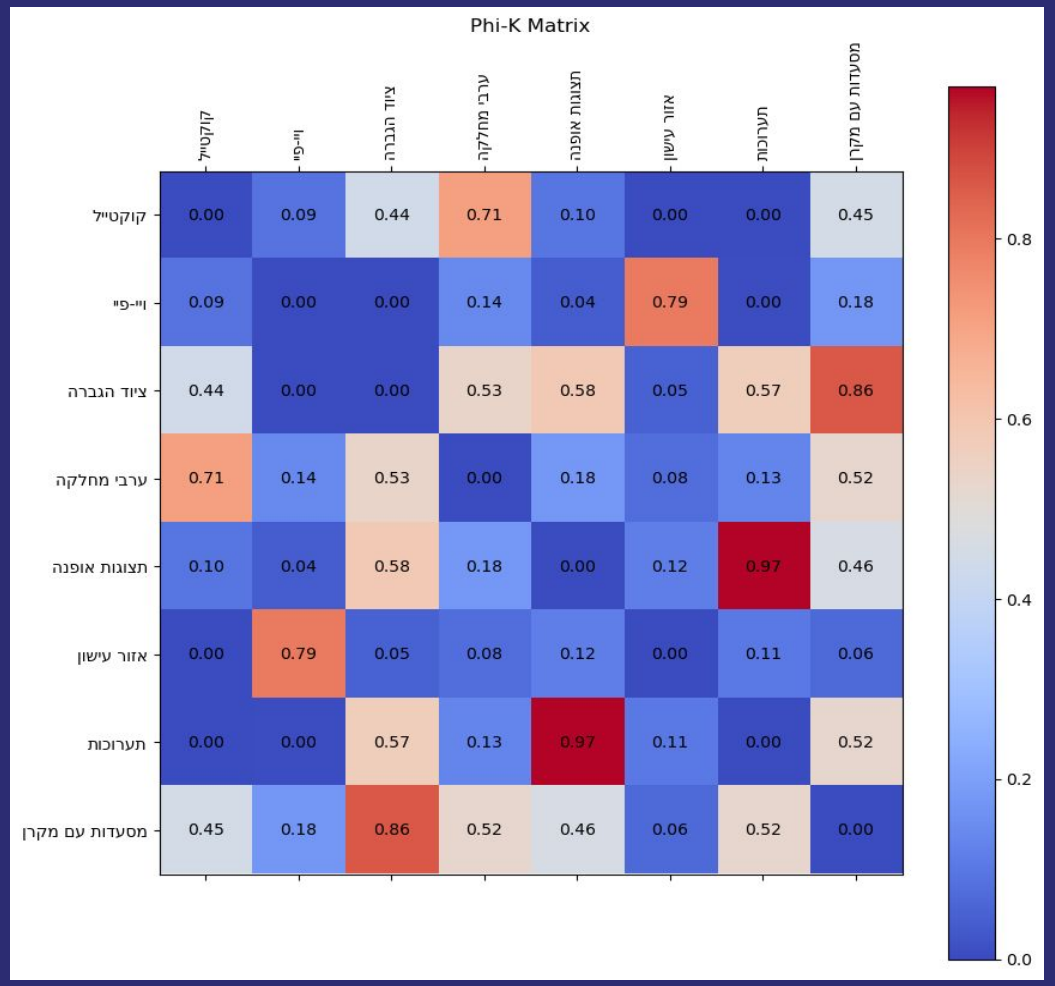

גרף עמודות המציג את השכיחות של המשתנים בהופעתם כמאפיין במסעדה.

# E.D.A

## התפלגות הדירוגים של המסעדות

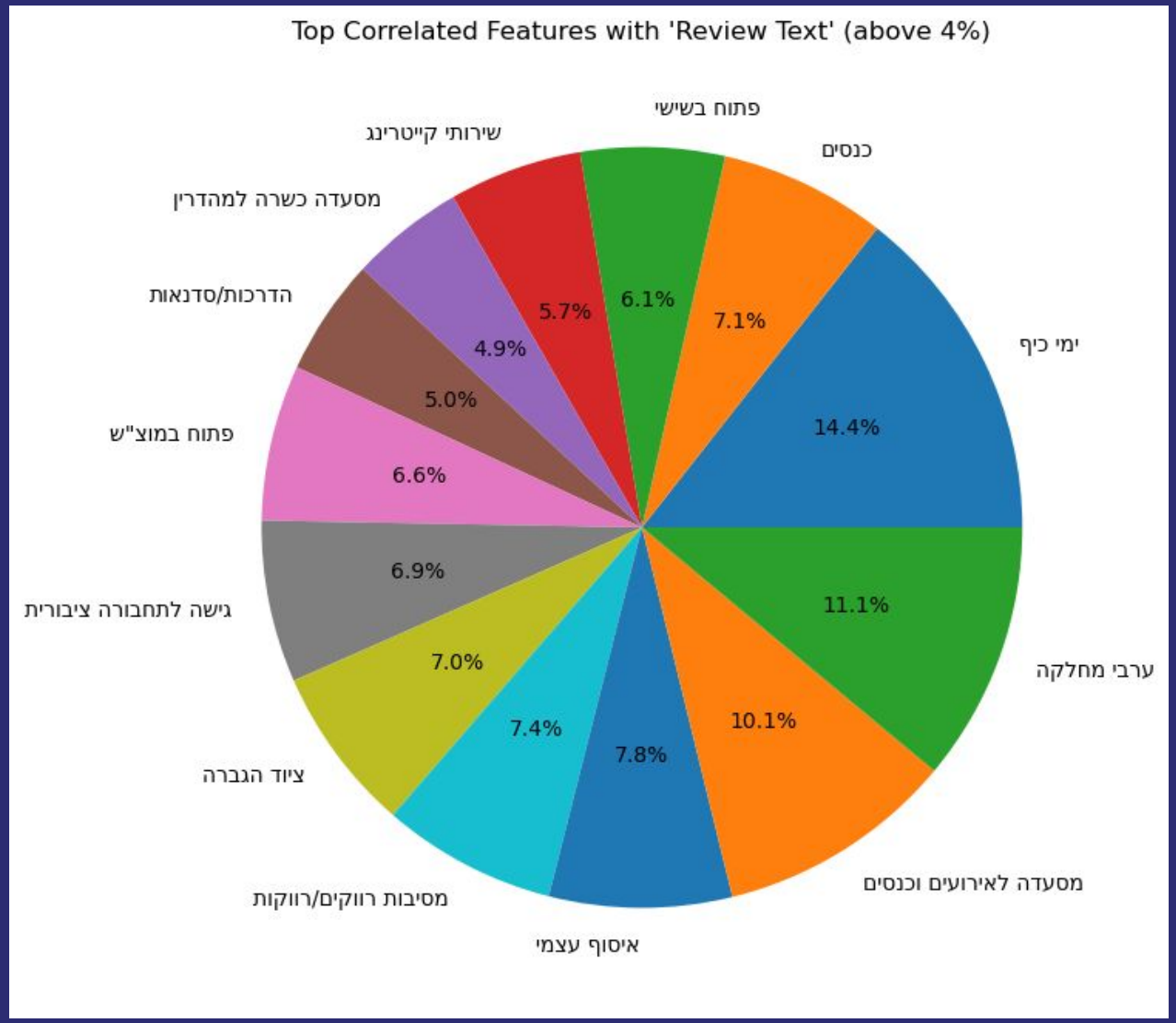



## מפת חום עם שימוש ב Phi-k המראה על הקוראלציה בין משתנים משמעותיים



# E.D.A

גרף פאי המראה את המאפיינים המשפיעים ביותר על עמודת הדירוג.
על ידי ניתוח זה ניתן יהיה ליצור מאפיינים חדשים ובכך לשפר את חיזוי הדירוג.



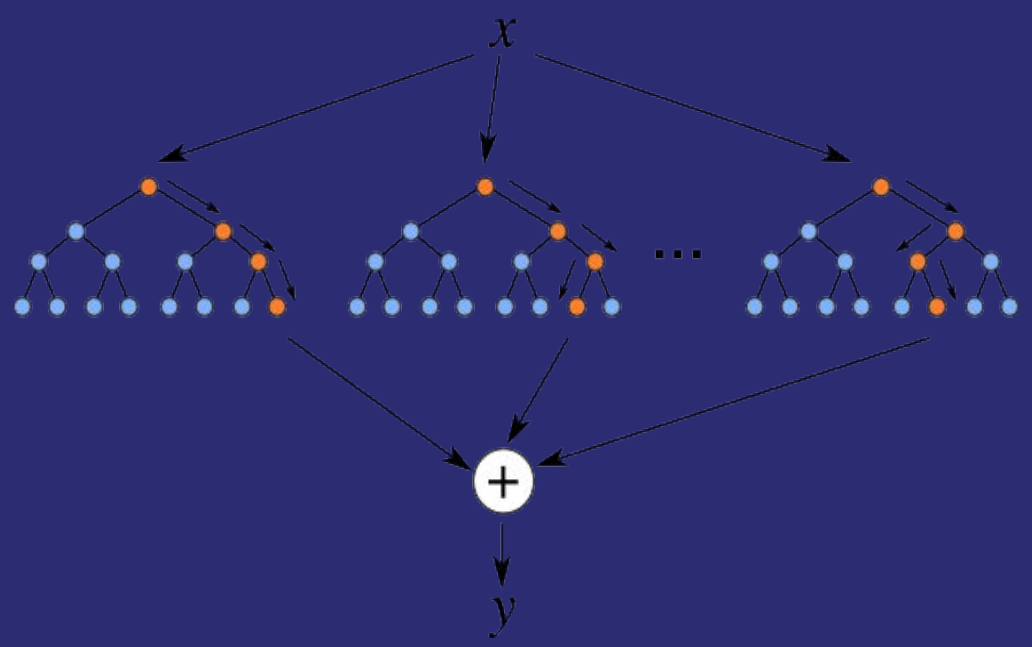
## MSE - Mean Squared Error

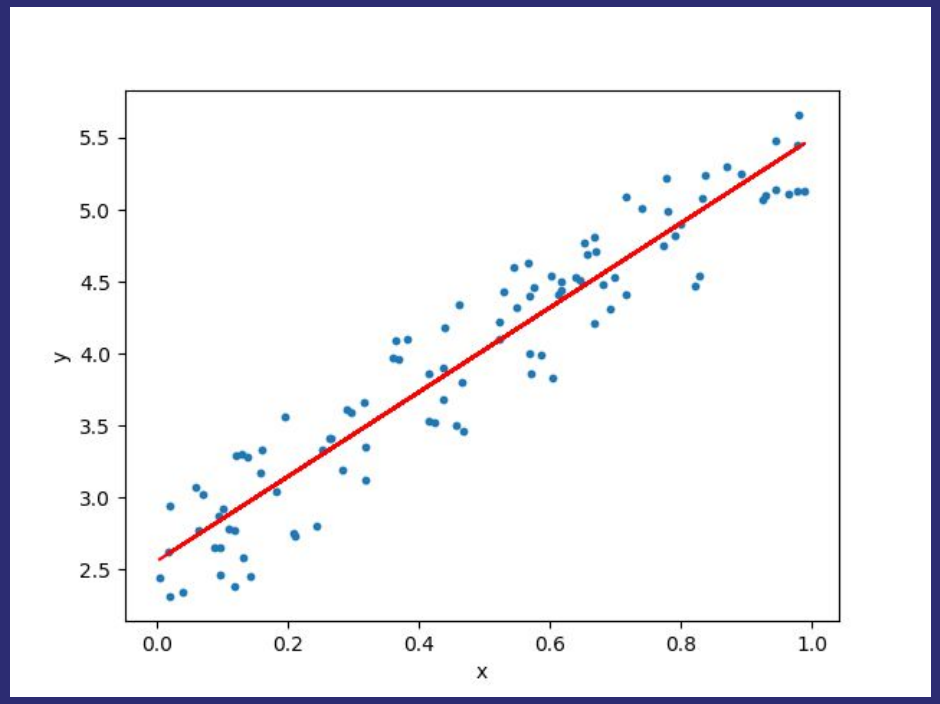
חישוב ה MSE
(הטעות הריבועית
הממוצעת) מודדת את
ממוצע ריבועי הטעויות.

$$MSE = \frac{1}{n} \Sigma \underbrace{\left( y - \hat{y} \right)^2}_{\text{The square of the difference between actual and predicted}}$$

# Machine Learning

הבעיה שלנו היא בעיית רגרסיה, כך שהשתמשנו במודלים ה- Linear Regression וה- Random Forest Regressor .



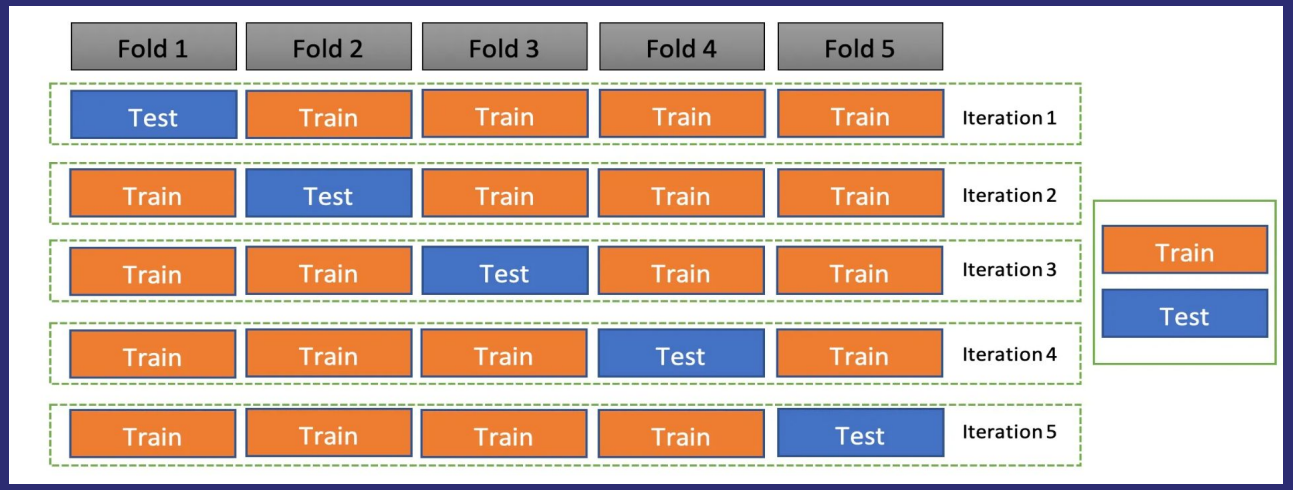




```
Mean Squared Error: 0.50736
```

```
Mean Squared Error: 0.43719
```

# Machine Learning

ניסינו לשפר את המודל של הרגרסיה הלינארית על ידי שימוש ב K-folds וגם ניסיון של מניפולציית פיצ'רים.




```
Average MSE: 0.51523
```

## K-folds

```
df_filtered['גישה בסופ"ש'] = 0
df_filtered.loc[(df_filtered['פתוח במוצ"ש'] == 1) & (df_filtered['פתוח בשישי'] == 1), 'גישה בסופ"ש'] = 1
```

```
df_filtered['ימי גיבוש'] =0
df_filtered.loc[(df_filtered['ערבי מחלקה'] == 1) & (df_filtered['ימי כיף'] == 1), 'ימי גיבוש'] = 1
df_filtered.head()
```

```
Mean Squared Error: 0.43627
```

## Featuring Manipulation

# Machine Learning

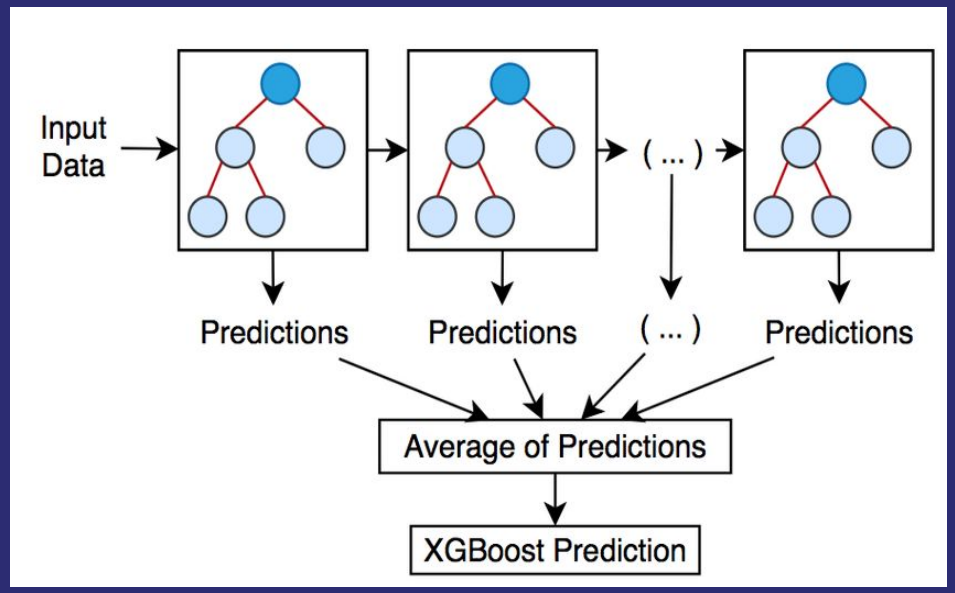
התוצאות שקיבלנו מהמודלים האלו לא היו מספקות, המודל מתקשה למצוא קשר בין המאפיינים ובכך החיזוי אינו מספק.

ולכן ניסינו להפוך את הבעיה שלנו לבעיית סיווג (classification) ובכך לנסות למצוא תוצאות יותר מספקות.

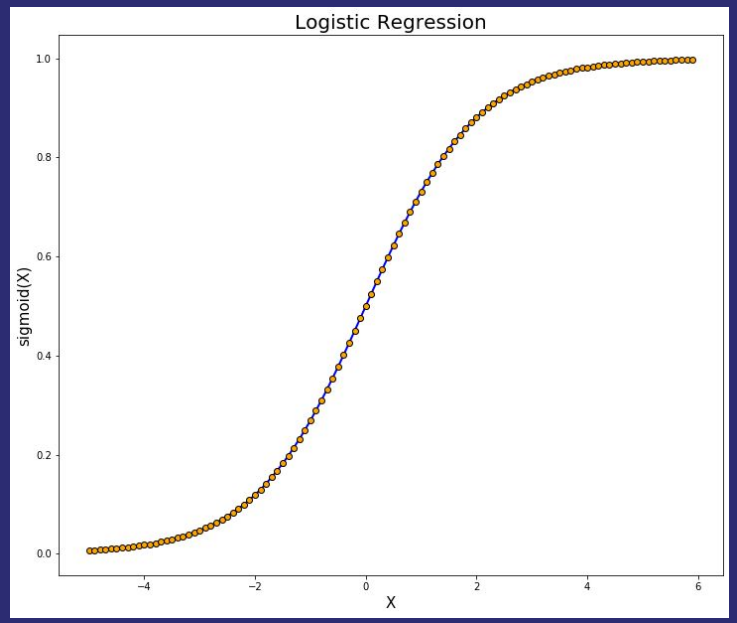
replacement of the rating feature in a binary feature, sort the rating to '1' is over and equal 4.5 and '0' for under 4.5

```
df_filtered["rating_clf"] = df_filtered["Review Text"].apply(lambda review_rate: 0 if review_rate < 4.5 else 1)
```

## XGB



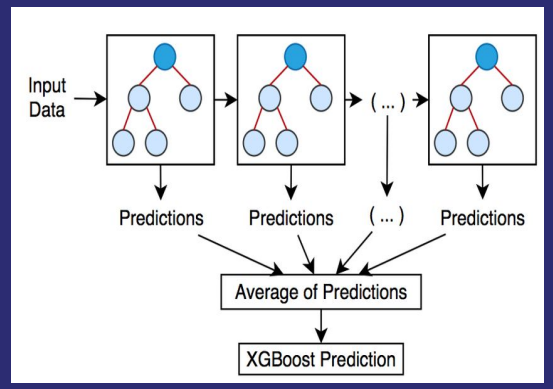

## Logistic Regression



# Machine Learning

בתוצאות שקיבלנו מהמודלים האלו חל שיפור בציון ה MSE.

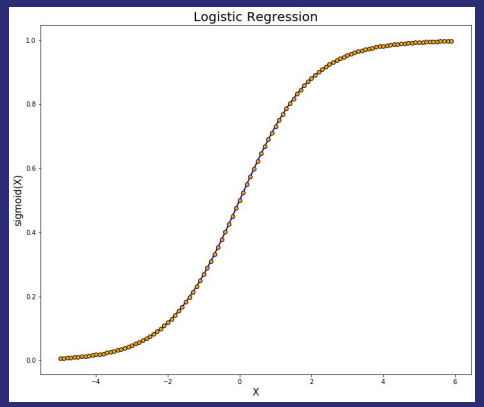
## XGB




```
Mean Squared Error: 0.37903
```

בנוסף למודל הסיווג של Logistic Regression נעשה שימוש ב"ענישה"(regularization) על מנת לשפר את המודל, מכיוון שיש לנו כמות קטנה של Samples ביחס לכמות ה-Features, ובכך מגבירים את יכולת הלמידה של המודל כאשר הוא מבצע טעויות.

## Logistic Regression




```
Mean Squared Error: 0.37096
```

# לסיכום

למעשה, התוצאה הטובה ביותר שקיבלנו היא מהמודל Logistic Regression, אך גם תוצאה זו אינה מספקת אותנו בכדי שנוכל להגיד כי המודל מצליח באחוזים גבוהים.

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.68 | 0.77 | 0.72 | 77 |
| 1 | 0.51 | 0.40 | 0.45 | 47 |
| accuracy | | | 0.63 | 124 |
| macro avg | 0.60 | 0.59 | 0.59 | 124 |
| weighted avg | 0.62 | 0.63 | 0.62 | 124 |

Mean Squared Error: 0.3709677419354839

# לסיכום

- תחילה, תקפנו את הבעיה בצורה ישירה, ובחרנו במודלים של בעיות רגרסיה לינארית. לאחר מכן, על מנת להשתמש במודלים מגוונים יותר ניסינו לבנות את הבעיה בצורה של קלסיפיקציה.
- איכות הנתונים שהתקבלה בהרכשה מן האתר לא הייתה איכותית מספיק לבניית מודל שיספק לנו תוצאות חיזוי טובות.
- דרך חשיבה אנושית -המשתנים המעורבים בהצגת דירוג מסעדות מורכבים ומשתנים בהתאם לדרך חשיבה אישית.

**<u>מסקנה כללית</u>: למידת מכונה מתקשה לחזות דירוג מסעדות ותהליך זה עשוי להיות מאתגר ואינו ניתן להגיון ישיר.**

מכון טכנולוגי חולון
Holon Institute of Technology

# Restuarnt4u

## תודה רבה על ההקשבה!