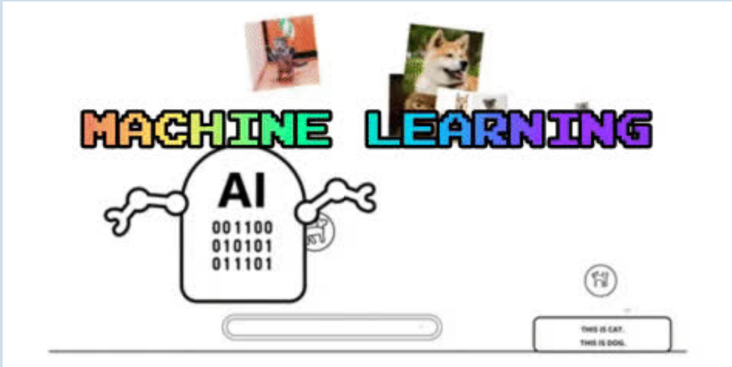
GOOD READS

פרויקט גמר מדעי הנתונים

אילה צברי & דקלה פלח

מכון טכנולוגי חולון
Holon Institute of Technology

במהלך העבודה בצענו את כל שלבי עבודתו של מדען נתונים



שאלת המחקר

אנו אוהבות מאוד לקרוא ספרים
ובתור "תולעות ספרים" כמובן שהחלום שלנו הוא לדעת לחזות מה יהיה דירוגו של ספר לאחר כמה שנים מהוצאתו לאור.
איך נדע האם זה ספר ששווה לקרוא?

הרכשת הנתונים

הרכשנו נתונים מהאתר : Good Reads
/https://www.goodreads.com

את ההרכשה ביצענו בעזרת Selenium.
עם שימוש בספרייה undetected_chromedriver שמאפשרת לבצע Scraping & Crawling כך שה-Chrome לא מזהה את המקור.

לאחר ההרכשה קיבלנו את ה-Data Set הבא:

	BookName	AuthorName	Genre	Rating	Reviews	Stars	Pages	BookCover	PublishingYear	Language	Series
0	Ways of Seeing	John Berger	Art	347205.0	2452.0	3.92	176	Paperback	1990 by Penguin	English	False
1	The Story of Art	E.H. Gombrich	Art	393712.0	1309.0	3.96	688	Unknown Binding	1967 by Phaidon Press LTD	English	False
2	The New Drawing on the Right Side of the Brain	Betty Edwards	Art	342219.0	947.0	3.87	291	Paperback	1999 by Tarcher	English	False
3	Steal Like an Artist: 10 Things Nobody Told Yo...	Austin Kleon	Art	271177.0	8253.0	3.96	160	Paperback	2012 by Workman Publishing Company	English	False
4	The Artist's Way: A Spiritual Path to Higher C...	Julia Cameron	Art	106706.0	3805.0	3.94	237	Paperback	2002 by Jeremy P. Tarcher	English	True
...	...	...	...	...	...	...	...	...	...	...	...
38203	Take Me with You	Catherine Ryan Hyde	Travel	41458.0	3668.0	4.20	362	Paperback	2014 by Lake Union Publishing	NaN	False
38204	A Woman in the Polar Night	Christiane Ritter	Travel	1571.0	292.0	4.27	215	Paperback	2010 by University of Alaska Press	English	False
38205	Jungleland: A Mysterious Lost City, a WWII Spy...	Christopher S. Stewart	Travel	1464.0	220.0	3.35	263	Hardcover	2013 by Harper	English	False
38206	A Sense of Direction: Pilgrimage for the Restl...	Gideon Lewis-Kraus	Travel	717.0	120.0	3.40	352	Hardcover	2012 by Riverhead Books	English	False
38207	Rick Steves Budapest	Rick Steves	Travel	341.0	24.0	4.36	502	Kindle Edition	2017 by Rick Steves	NaN	False

38208 rows × 11 columns

מספר שורות: 38,207

מספר עמודות: 11

ניקוי ה- Data Set

- הסרת ערכים חסרים של Book Name.
- בחלק מהתאים תחת העמודה של Genre היו ערכים עם: undetected מילאנו את התאים בערכים תקינים.
- הסרת ערכים חסרים מעמודת שנת ההוצאה לאור, והסרת ספרים ששנת ההוצאה שלהם קטנה מ- 1990.
- בעמודת השפה ביצענו שינוי של הערכים חסרים לשפה הכי פופולארית.
- המרה לערכים בינאריים של העמודות: BookCover, Series, Genre
- הסרת Duplicates Values לפי: שם הספר, שם המחבר, כמות כוכבים, כמות דפים ושנת הוצאה לאור.

לאחר הניקוי התקבל ה-DATA SET הבא:

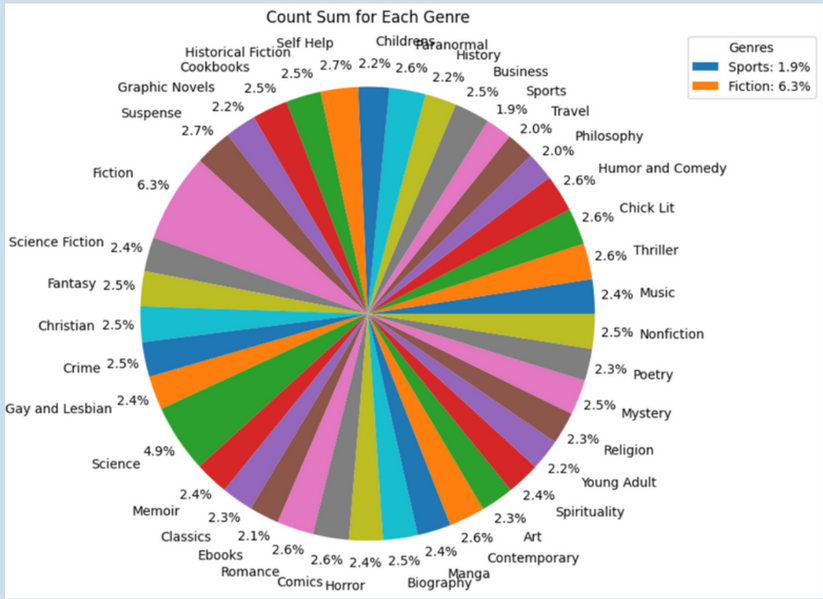
	BookName	AuthorName	Rating	Reviews	Stars	Pages	BookCover	PublishingYear	Language	Series	...	Religion	History	Sports	Horror	Art
1	Her Greatest Mistake	Hannah Cowan	7324.0	643.0	4.02	384	0	2023	English	1	...	0	0	1	0	0
4	How to Sell a Haunted House	Grady Hendrix	36626.0	7413.0	3.83	419	0	2023	English	0	...	0	0	0	1	0
6	Emily Wilde's Encyclopaedia of Faeries	Heather Fawcett	17206.0	4116.0	4.14	336	0	2023	English	1	...	0	0	0	0	0
8	The New Guy	Sarina Bowen	3159.0	580.0	4.18	346	1	2023	English	1	...	0	0	1	0	0
9	Jock Blocked	Pippa Grant	7596.0	791.0	4.27	337	0	2023	English	1	...	0	0	1	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
6360	鬼滅の刃 10 [Kimetsu no Yaiba 10]	Koyoharu Gotouge	11401.0	466.0	4.61	200	1	2018	Japanese	1	...	0	0	0	0	0
6361	鬼滅の刃 4 [Kimetsu no Yaiba 4]	Koyoharu Gotouge	14306.0	568.0	4.53	192	1	2016	Japanese	1	...	0	0	0	0	0
6362	鬼滅の刃 6 [Kimetsu no Yaiba 6]	Koyoharu Gotouge	11721.0	466.0	4.53	205	0	2017	Japanese	1	...	0	0	0	0	0
6363	鬼滅の刃 8 [Kimetsu no Yaiba 8]	Koyoharu Gotouge	12738.0	860.0	4.68	199	0	2017	Japanese	1	...	0	0	0	0	0
6364	###-09	Colleen Hoover	810243.0	70185.0	4.21	310	1	2015	English	0	...	0	0	0	0	0

22660 rows × 49 columns

מספר שורות: 22,660

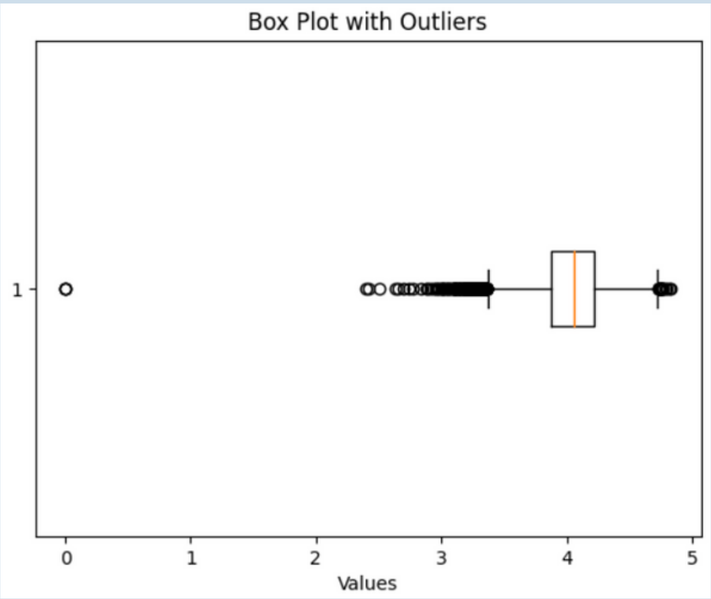
מספר עמודות: 49

EDA

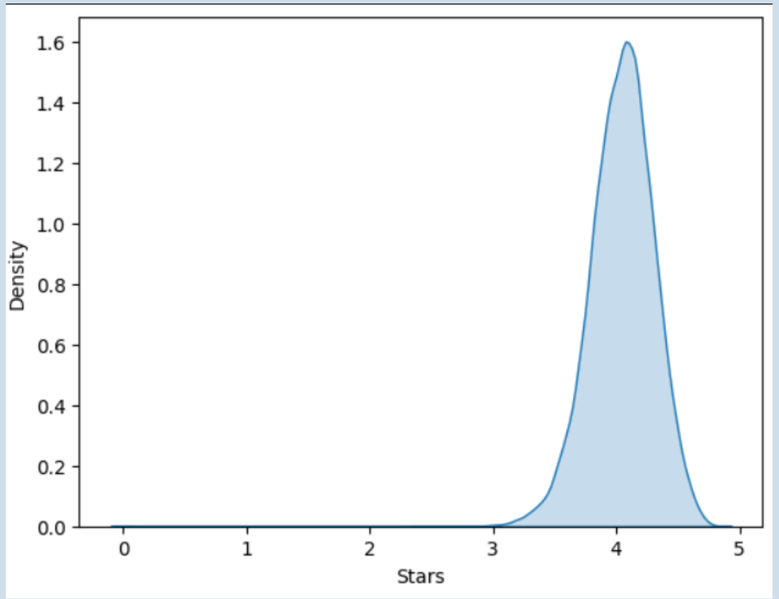


6.3% מהספרים הם מ-
Fiction Genre.

3.9% מהספרים הם מ-
Sports Genre.

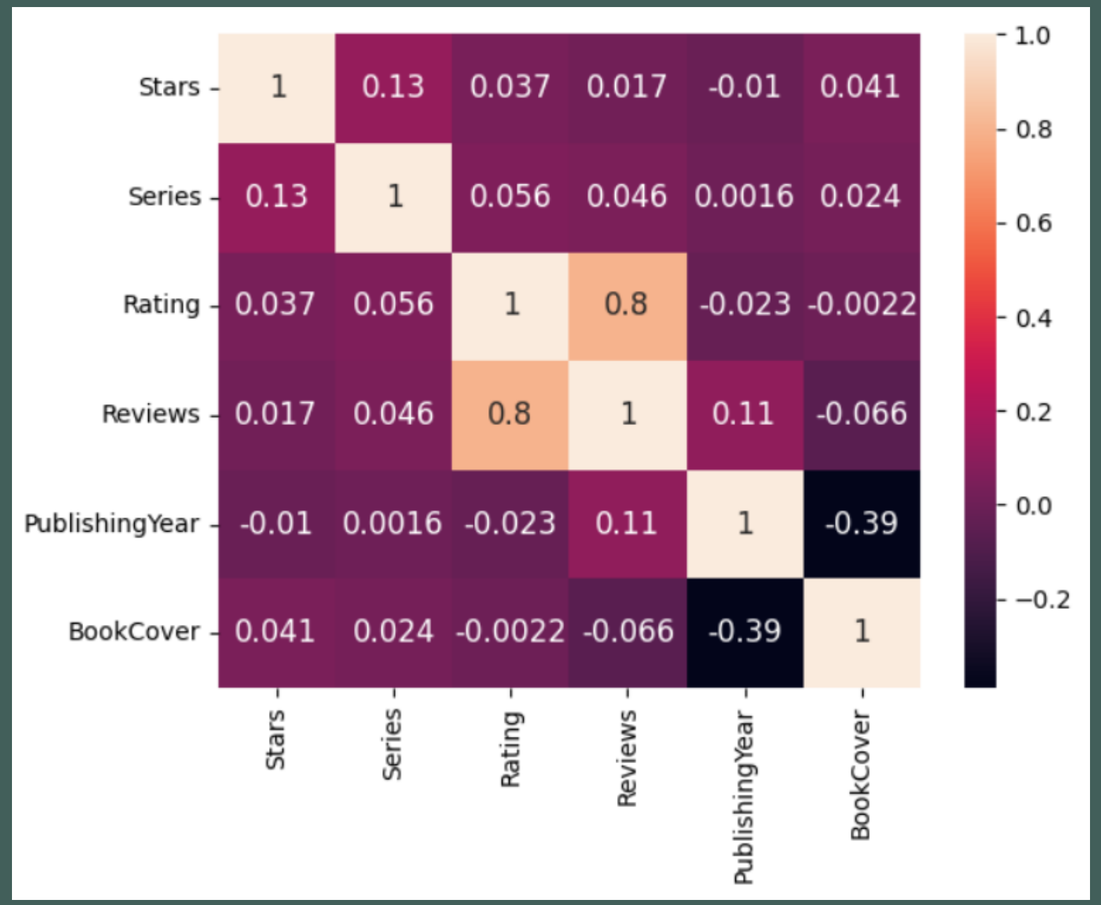


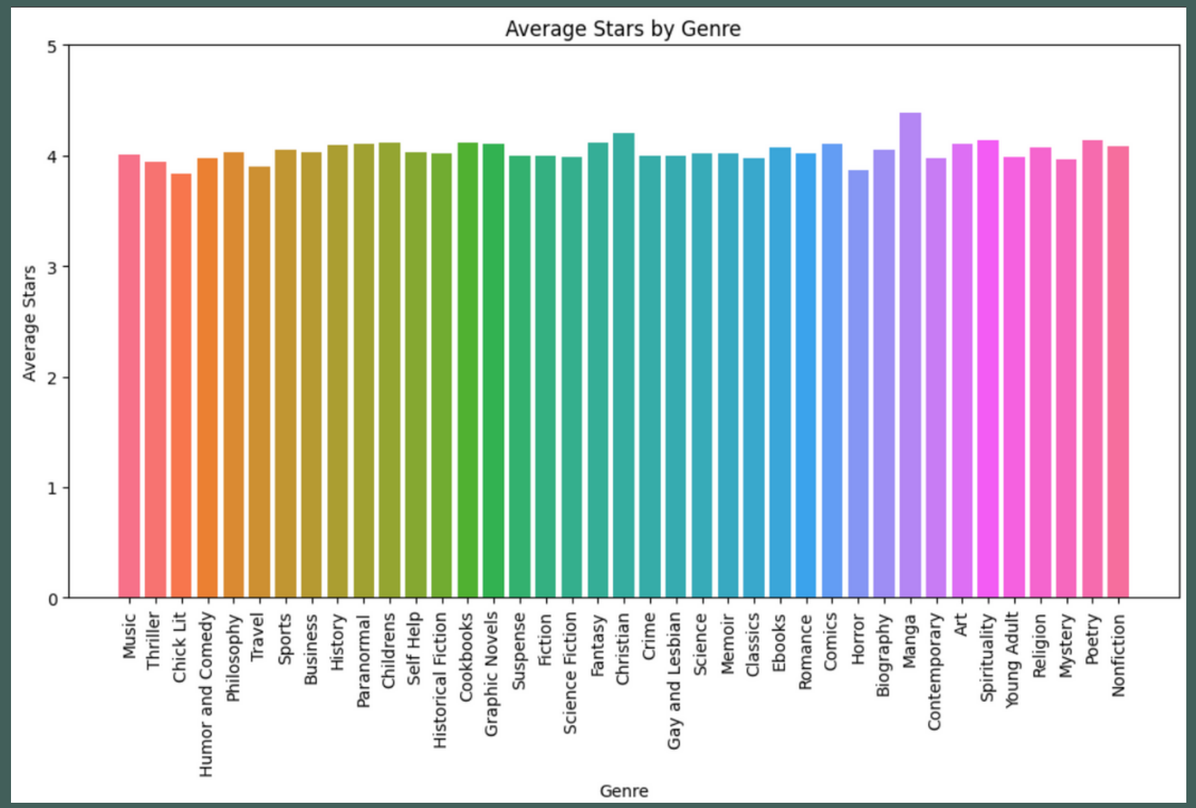
טיפול ב- Outliers

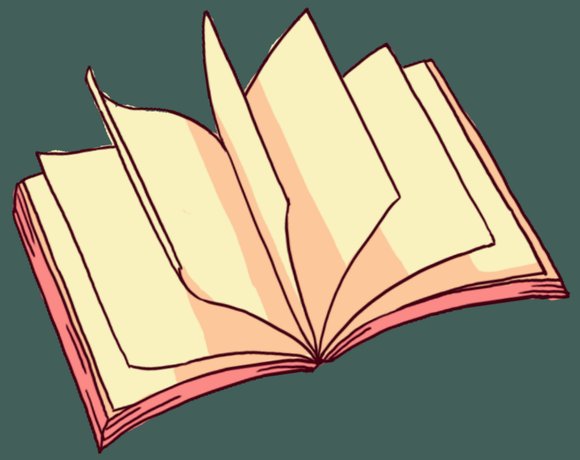


ערכי הכוכבים נעים בין 3-5

Heat Map


Stars 1 0.13 0.037 0.017 -0.01 0.041
Series 0.13 1 0.056 0.046 0.0016 0.024
Rating 0.037 0.056 1 0.8 -0.023 -0.0022
Reviews 0.017 0.046 0.8 1 0.11 -0.066
PublishingYear -0.01 0.0016 -0.023 0.11 1 -0.39
BookCover 0.041 0.024 -0.0022 -0.066 -0.39 1
Stars
Series
Rating
Reviews
PublishingYear
BookCover
1.0
0.8
0.6
0.4
0.2
0.0
−0.2


Average Stars by Genre
Average Stars
0
1
2
3
4
5
Music
Thriller
Chick Lit
Humor and Comedy
Philosophy
Travel
Sports
Business
History
Paranormal
Childrens
Self Help
Historical Fiction
Cookbooks
Graphic Novels
Suspense
Fiction
Science Fiction
Fantasy
Christian
Crime
Gay and Lesbian
Science
Memoir
Classics
Ebooks
Romance
Comics
Horror
Biography
Manga
Contemporary
Art
Spirituality
Young Adult
Religion
Mystery
Poetry
Nonfiction
Genre



למידת מכונה

השתמשנו באלגוריתמים לפתירת בעיית רגרסיה:

KNN

```
Best k: 15
R-squared Score: 0.05372887375121522 5
MAE for KNNRegressor: 0.1948873197999
```

Linear Regression

```
R-squared Score: 0.2235922159679422
MAE for LinearRegression: 0.17660721
```

Random Forest

```
best num estimators=71
R-squared Score:0.3303364721466703
MAE for RandomForestRegressor: 0.16029
```

Decision Tree

```
best max depth=15
best min samples=30
R-squared Score:0.22033605612774 36
MAE for DecisionTreeRegressor: 0.17545280
```

Neural Networks

```
R-squared Score:-21.960210102661502
MAE for Neural Networks:0.268378348
```

Lasso

```
Intercept: 4.017917507847524
R-squared Score: 0.00218466739445 91
MAE for Lasso: 0.20264992214053718
```

הבנו שאין בידנו DATA SET מספק עבור החיזוי ועל מנת לנסות לשפר את התוצאות החלטנו להרכיש נתונים מחדש עבור גאנר יחיד, בחרנו ב- FICTION.

הפעם ניקח גם את תקציר הספר ונבצע ניתוח טקסט.

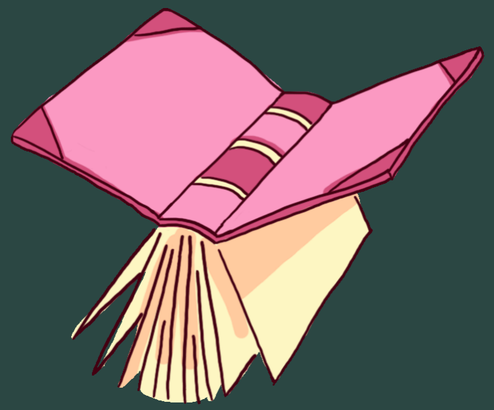
ה- DATA SET לאחר ההרכשה:

	BookName	AuthorName	Genre	Rating	Reviews	Stars	Pages	BookCover	PublishingYear	Language	Series	Summary
0	The Immortalists	Chloe Benjamin	fiction	195086	19263	3.71	346	Hardcover	2018 by G.P. Putnam's Sons	English	False	If you knew the date of your death, how would ...
1	Small Great Things	Jodi Picoult	fiction	341940	30185	4.35	510	Kindle Edition	2016 by Ballantine Books	English	True	Ruth Jefferson is a labor and delivery nurse a...
2	Transcendent Kingdom	Yaa Gyasi	fiction	134427	15512	4.12	264	Hardcover	2020 by Knopf	English	False	Yaa Gyasi's stunning follow-up to her acclaime...
3	Stranger in a Strange Land	Robert A. Heinlein	fiction	303859	9448	3.92	525	Paperback	1991 by Ace	English	False	NAME: Valentine Michael Smith\nANCESTRY: Human...
4	The Absolutely True Diary of a Part-Time Indian	Sherman Alexie	fiction	259563	26655	4.07	230	Hardcover	Brown Books for Young Readers	English	False	Bestselling author Sherman Alexie tells the st...
...	...	...	...	...	...	...	...	...	...	...	...	...
995	Before the Coffee Gets Cold	Toshikazu Kawaguchi	fiction	253025	37359	3.73	213	Paperback	2019 by Picador	English	True	What would you change if you could go back in ...
996	Anansi Boys	Neil Gaiman	fiction	215415	11347	4.04	387	Mass Market Paperback	2006 by HarperCollins HarperTorch	English	True	God is dead. Meet the kids.\n\nFat Charlie Nan...
997	Choke	Chuck Palahniuk	fiction	204401	6864	3.70	293	Paperback	2002 by Anchor Books	English	False	Victor Mancini, a medical-school dropout, is a...
998	Giovanni's Room	James Baldwin	fiction	121820	12963	4.31	224	Paperback	1988 by Laurel	English	False	Set in the contemporary Paris of American exp...
999	People We Meet on Vacation	Emily Henry	fiction	859187	83947	3.92	364	Paperback	2021 by Berkley	English	False	Two best friends. Ten summer trips. One last c...

1000 rows × 12 columns

מספר שורות: 1000

מספר עמודות: 12



ניקוי ה- Data Set

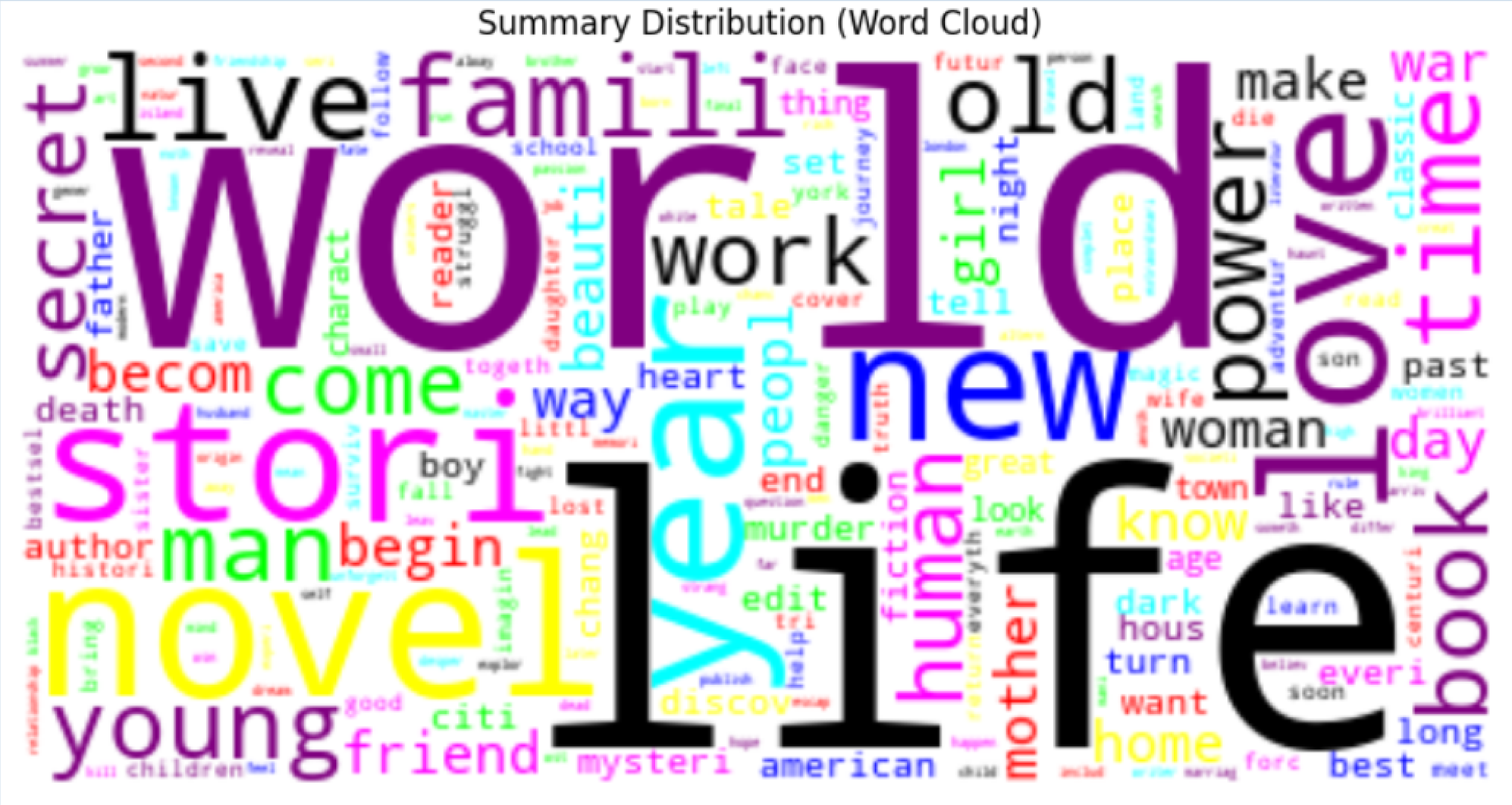
ביצענו ניקוי זהה לניקוי שביצענו ל-Data set הקודם.

Pre Process -לתקציר

- הפרדת הטקסט למילים בודדות
- לקיחת השורש של המילים
- הסרת סימני פיסוק
- המרת כל האותיות לאותיות קטנות
- הסרת מילות עיצור

בניית וקטורים עבור המטריצה

המילים הכי פופולאריות:



למידת מכונה

השתמשנו באלגוריתמים לפתירת בעיית רגרסיה:

KNN

```
Best k: 15
R-squared Score: 0.04696214830053389
MAE for KNeighborsRegressor: 0.197931
```

Linear Regression

```
R-squared Score: 0.255430811509616
MAE for Linear Regression: 0.176018
```

Random Forest

```
best num estimators=71
R-squared Score:0.12514864204460996
MAE for RandomForestRegressor: 0.173975
```

Decision Tree

```
best max depth=4
best min samples=30
R-squared Score:0.07707798995902367
MAE for DecisionTreeRegressor: 0.18804
```

Neural Networks

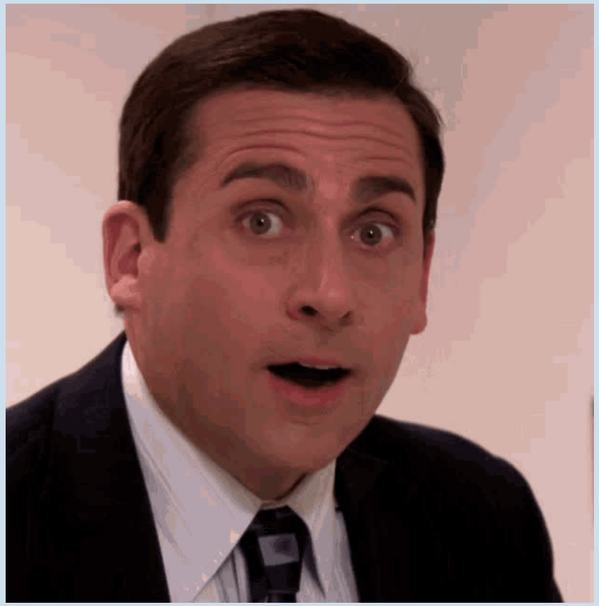
```
R-squared Score:-1135.777707787406
MAE for Neural Networks:7.24649752
```

Lasso

```
Intercept: 3.83486972337234
R-squared Score: 0.22222884336744
MAE for Lasso: 0.17997440359737815
```

פישטנו את הבעיה מבעיית רגרסיה לבעיית סיווג,
אנו רוצות לסווג האם ספר יקבל דירוג: נמוך, בנוני או גבוה.
חילקנו את ה-Stars ל- Bins והפעלנו אלגוריתמי סיווג שונים על ה- Data Set המקורי שהרכשנו.

הופתענו לגלות ש..



הצלחנו לשפר את תוצאות החיזוי!

KNN

```
Best k: 7
F1 Score: 0.5553987231799725
MAE for KNeighborsClassifier: 0.42961
```

Random Forest

```
best num estimators=71
F1-squared Score:0.6290032791977884
MAE for RandomForestClassifier: 0.2894
```

Naive Bayes

```
Accuracy on Train data=  0.47735172413
Accuracy on Test data=  0.477052074139
```

Neural Networks

```
F1 Score: 0.5793103448275861
MAE for Neural Networks:0.5922
```

Logistic Regression

```
F1-squared Score:0.007486631016042781
MAE for Logistic Regression: 0.4095322
```

Decision Tree

```
best max depth=6
best min samples=5
F1-squared Score:0.2637293417588469
MAE for DecisionTreeClassifier: 0.367387
```

מסקנות:

מטרת המחקר הייתה לבדוק האם ניתן לחזות דירוג של ספר לאחר כמה שנים מהוצאתו לאור.

ניסינו לפתור את הבעיה בכל מיני דרכים וליישם שיטות ומודלים שונים (מבוססי רגרסיה) על מנת להגיע לתוצאות חיזוי טובות. בנוסף, ביצענו ניתוח טקסט על מנת לנסות לשפר את תוצאות החיזוי אך זה לא צלח.

לבסוף, פישטנו את הבעיה מבעיית רגרסיה לבעיית סיווג ואכן שיפרנו את התוצאות פי 2!

תודה על ההקשבה!

הגיפים נלקחו מאתר https://tenor.com/