# WHAT MAKES A BOOK SUCCESSFUL?

Metro College of Technology - Data Science and Application Program

Ana Clara Tupinambá Freitas

Oriented by Professor Vijay Kumar

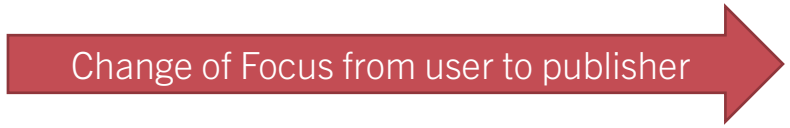
November,2021

# INTRODUCTION

| Gutenberg press | Change | Provide |
| --- | --- | --- |
| Since the creation of the Gutenberg press the numbers of writers and readers have increased. | Change of power, revolution and peace have been influenced by words. | Provide an engine for recommendation of books. |

# BUSINESS QUESTIONS

Before EDA




After EDA

Is there differences in country preferences?
Is there rating differences between group ages?
Is there difference between overall rating by genre?
Is being awarded making a difference in ratings?
Is being pert of a Series boosting the rates?
Is there a preferential Author in Countries?
Does the number of Characters impacts ratings?

Is there a difference between USA and Other Countries reading taste?
Is there reading preferences differences between group ages?
Is time of first publication impacting in how successful a book is?
Is year of publication impacting in how successful a book is?
Is being part of a Series impacting success?
Is genre impacting in success?

# DATA COLLECTION, EDA, CLEANING, TRANSFORMATION AND FEATURE ENGINEERING

```
Shape of books dataset:  (271379, 8)
Shape of genre dataset:  (52478, 25)
Shape of ratings dataset:  (1149780, 3)

Shape of books dataset after joining:  (80745, 15)
```


Data Sources:

Rating, Location, Age and other attibutes: http://www2.informatik.uni-freiburg.de/~cziegler/BX/

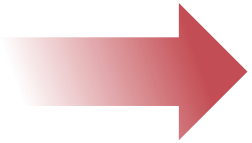
Genre, Setting, and other attibutes: https://zenodo.org/record/4265096#.YXlibBwpCHs

# DUPLICATED AND MISSING VALUES

There are 11 duplicated values in DF.
Duplicated values were dropped maintaning only the last occurrence

Are there any missing value?

| | Total | # Missing | % Missing |
|---|---|---|---|
| Age | 80734 | 23254 | 28.80 |
| Book-Author | 80734 | 0 | 0.00 |
| Book-Rating | 80734 | 0 | 0.00 |
| Book-Title | 80734 | 0 | 0.00 |
| ISBN | 80734 | 0 | 0.00 |
| Location | 80734 | 0 | 0.00 |
| Publisher | 80734 | 0 | 0.00 |
| User-ID | 80734 | 0 | 0.00 |
| Year-Of-Publication | 80734 | 0 | 0.00 |
| awards | 80734 | 0 | 0.00 |
| characters | 80734 | 0 | 0.00 |
| firstPublishDate | 80734 | 3487 | 4.32 |
| genres | 80734 | 0 | 0.00 |
| series | 80734 | 52748 | 65.34 |
| setting | 80734 | 0 | 0.00 |

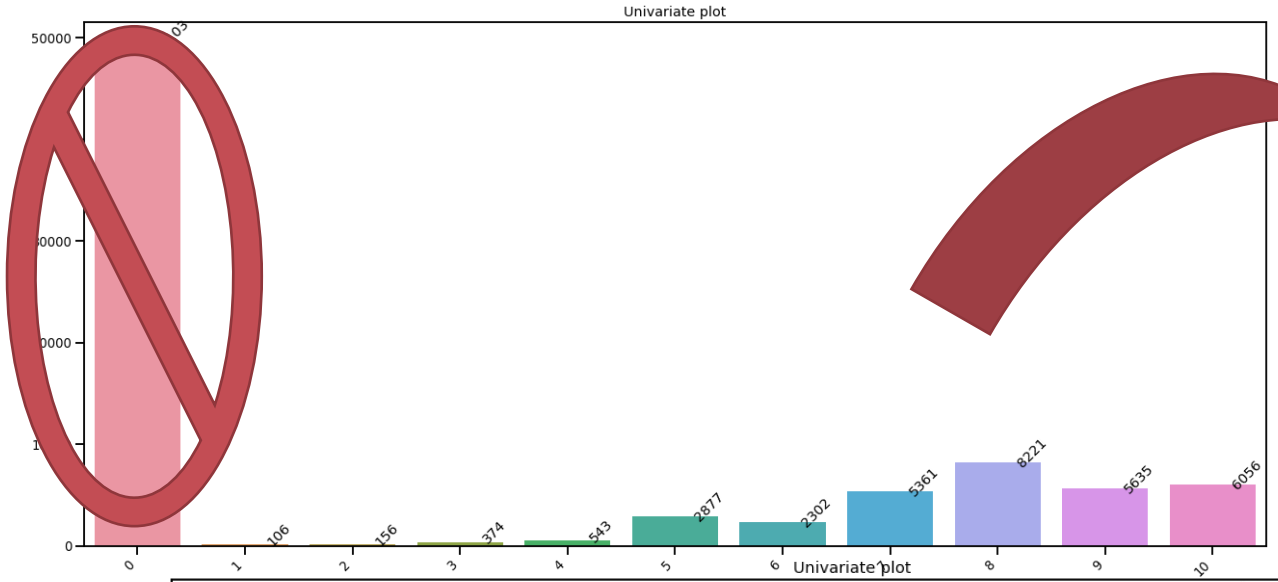


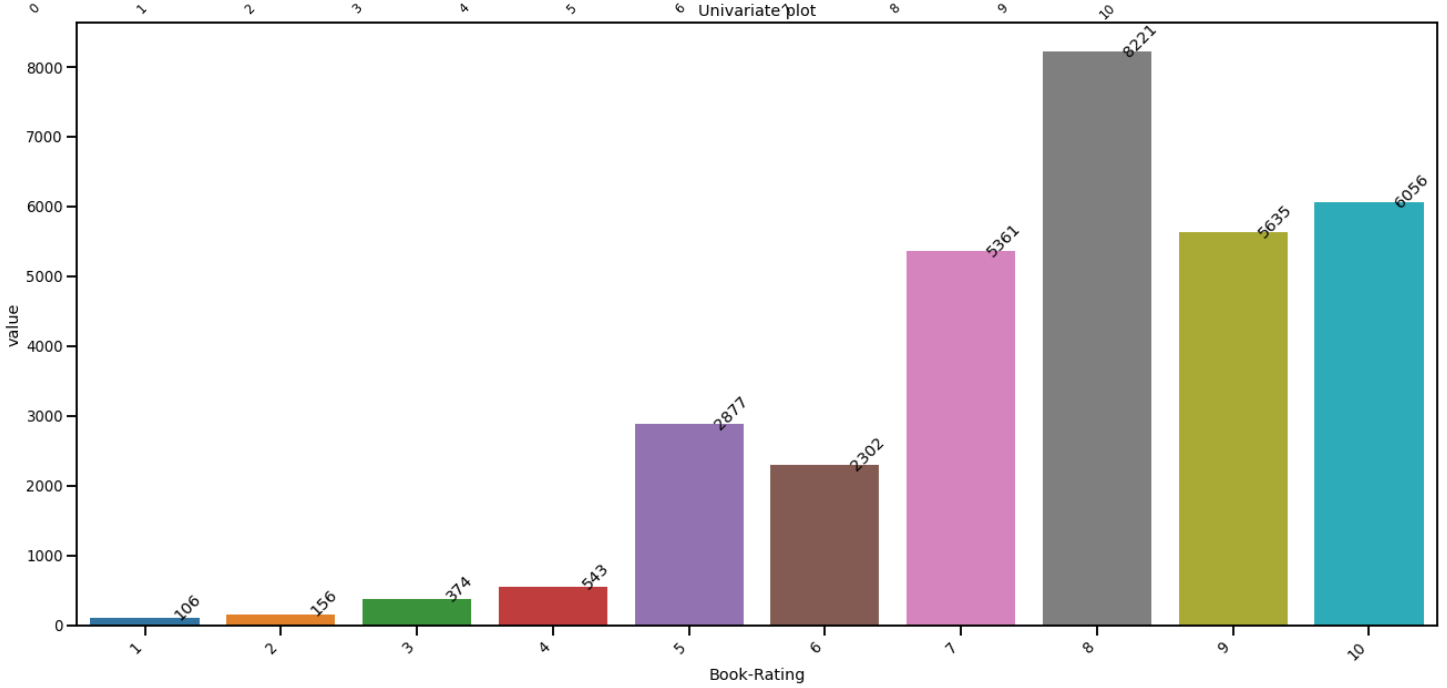
Shape of books dataset after cleaning: (21195, 9)

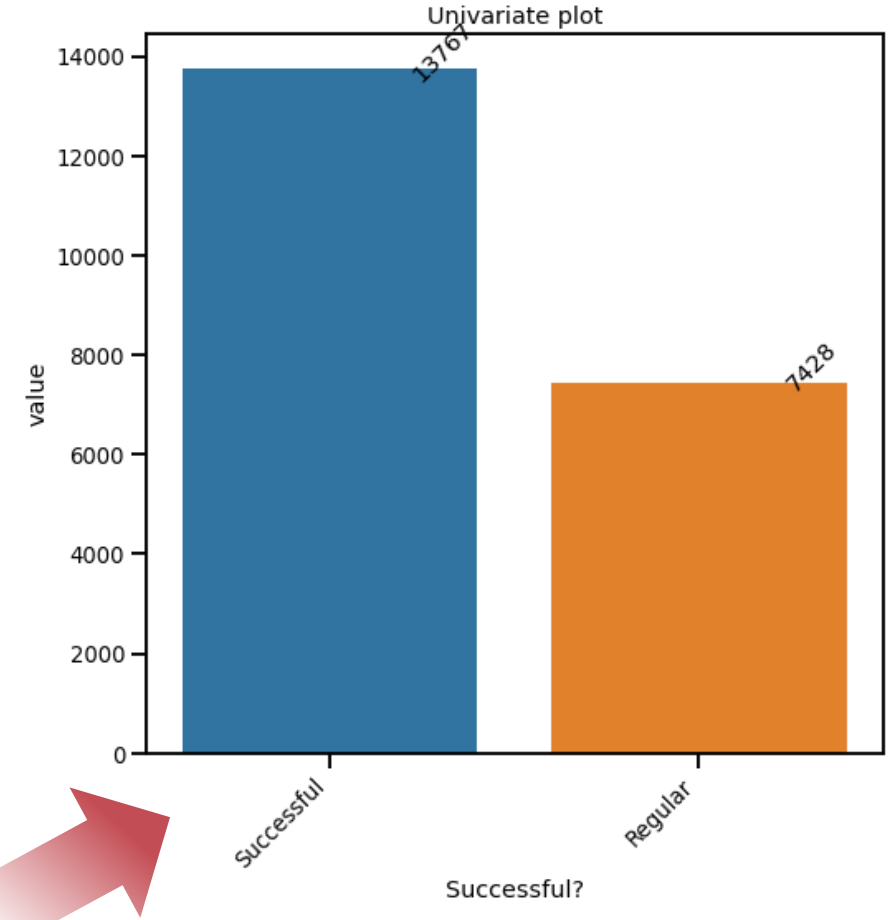
Are there any missing value?

| | Total | # Missing | % Missing |
|---|---|---|---|
| 1st Publication Era | 21195 | 0 | 0.0 |
| Awards? | 21195 | 0 | 0.0 |
| Genre1 | 21195 | 0 | 0.0 |
| Genre2 | 21195 | 0 | 0.0 |
| Group Age | 21195 | 0 | 0.0 |
| Series? | 21195 | 0 | 0.0 |
| Successful? | 21195 | 0 | 0.0 |
| User Country | 21195 | 0 | 0.0 |
| Year Of Publication | 21195 | 0 | 0.0 |

# SUCCESSFUL?



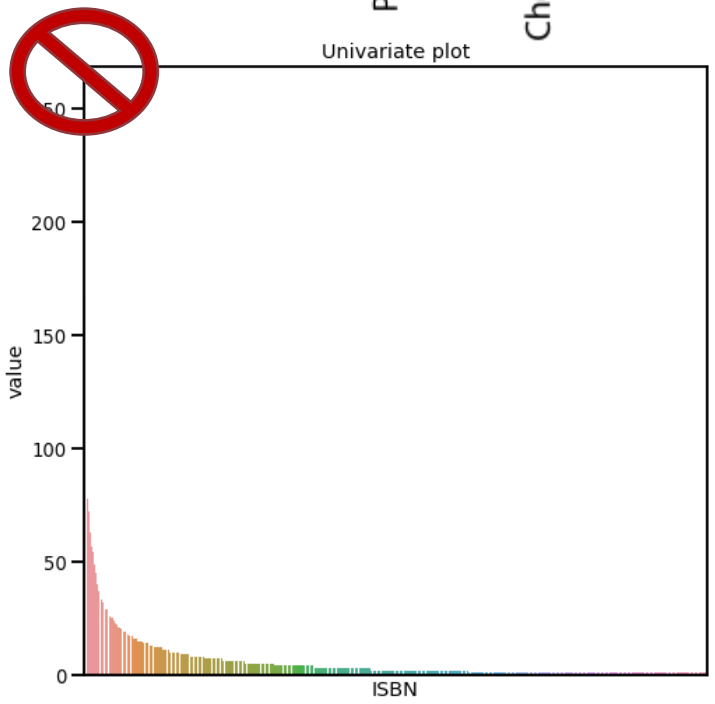

Books ratings will be further categorized to

- Successful(>=8)
- Regular

# ISBN

Are there any missing value?

| | Total | # Missing | % Missing |
|---|---|---|---|
| Age | 80734 | 23254 | 28.80 |
| Book-Author | 80734 | 0 | 0.00 |
| Book-Rating | 80734 | 0 | 0.00 |
| Book-Title | 80734 | 0 | 0.00 |
| ISBN | 80734 | 0 | 0.00 |
| Location | 80734 | 0 | 0.00 |
| Publisher | 80734 | 0 | 0.00 |
| User-ID | 80734 | 0 | 0.00 |
| Year-Of-Publication | 80734 | 0 | 0.00 |
| awards | 80734 | 0 | 0.00 |
| characters | 80734 | 0 | 0.00 |
| firstPublishDate | 80734 | 3487 | 4.32 |
| genres | 80734 | 0 | 0.00 |
| series | 80734 | 52748 | 65.34 |
| setting | 80734 | 0 | 0.00 |

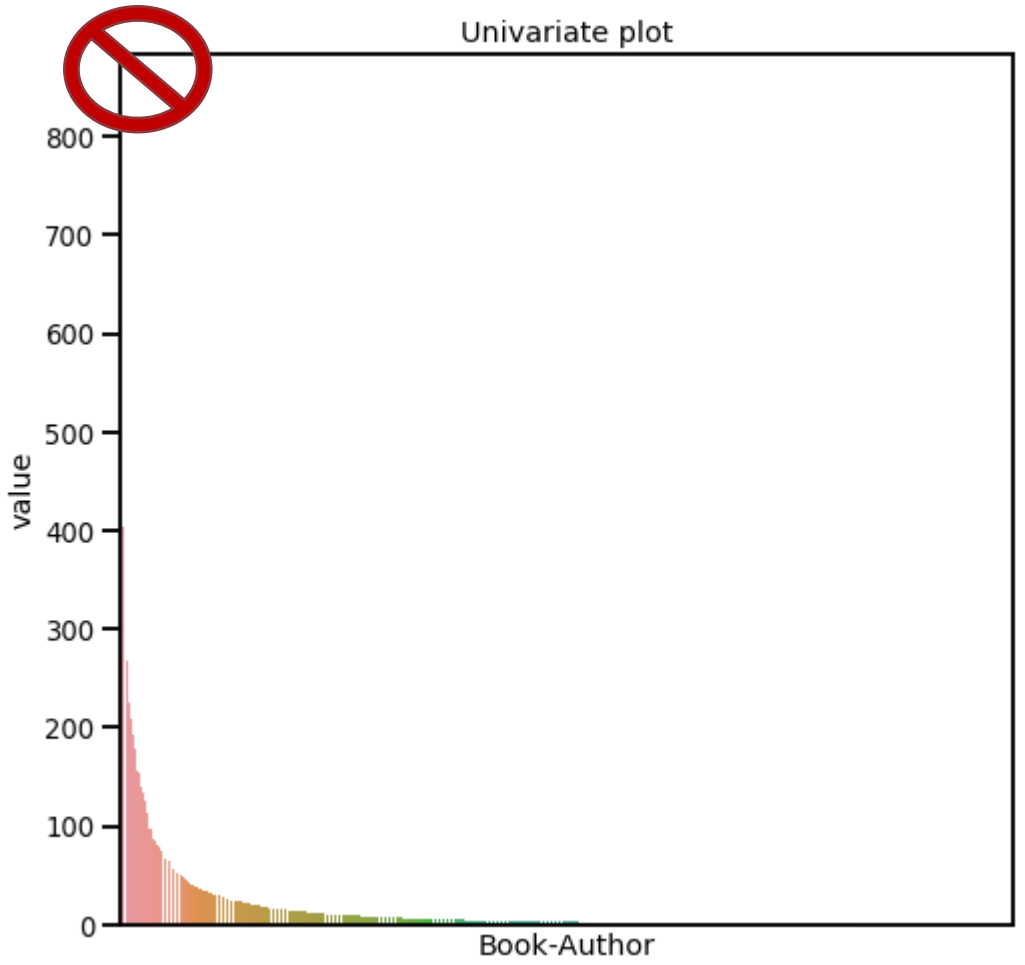



# BOOK-AUTHOR

This is univariate analysis for ' Book-Author '

| | Count | Percentage |
|---|---|---|
| Michael Crichton | 842 | 2.66 |
| Maeve Binchy | 578 | 1.83 |
| Barbara Kingsolver | 571 | 1.81 |
| Mary Higgins Clark | 564 | 1.78 |
| Tom Clancy | 454 | 1.44 |
| ... | ... | ... |
| Denis Johnson | 0 | 0.00 |
| Denise Giardina | 0 | 0.00 |
| Nina Bawden | 0 | 0.00 |
| Dennis Shryack | 0 | 0.00 |
| Octave Mirbeau | 0 | 0.00 |

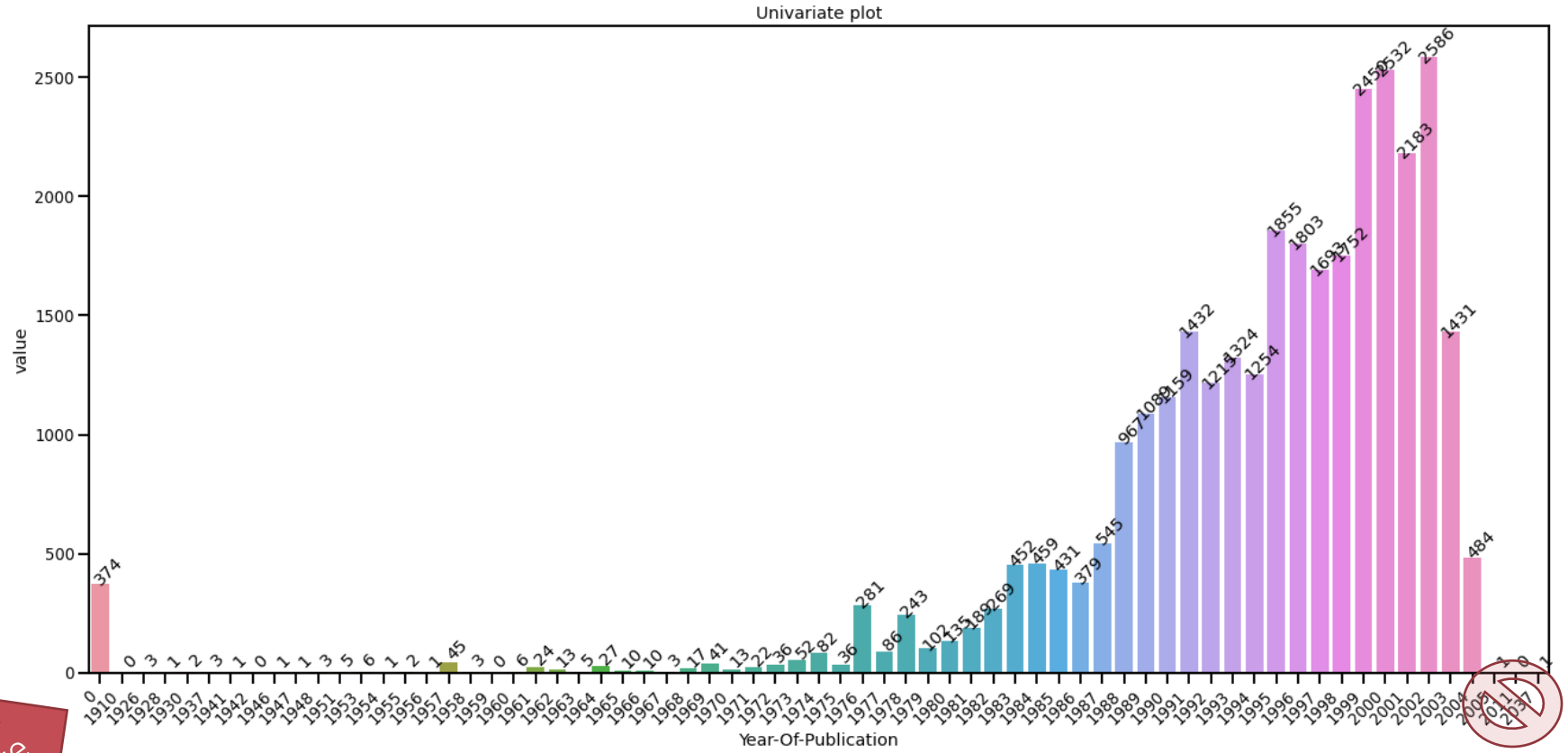
[2017 rows x 2 columns]



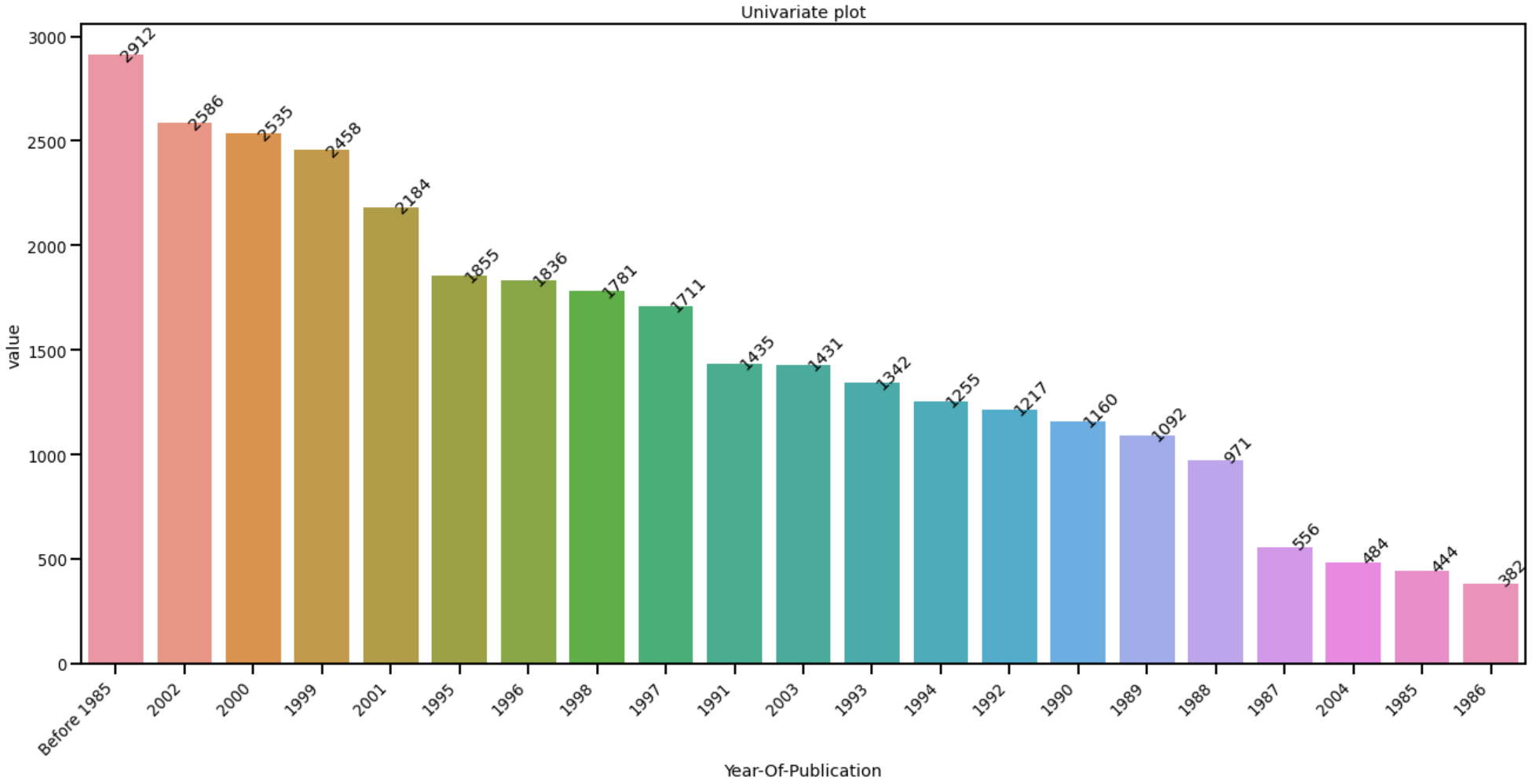
# YEAR OF PUBLICATION

This is univariate analysis for ' Year-Of-Publication '

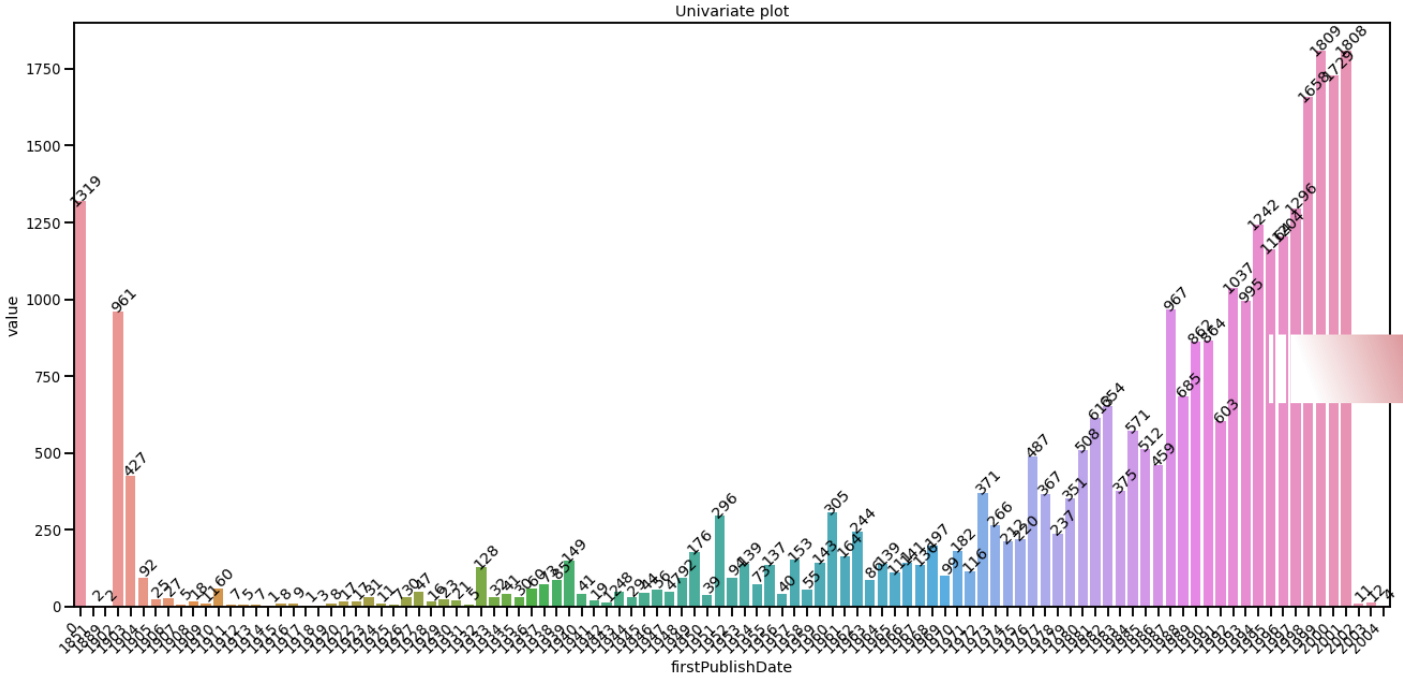
| | Count | Percentage |
|---|---|---|
| 2002 | 2586 | 8.18 |
| 2000 | 2532 | 8.00 |
| 1999 | 2450 | 7.75 |
| 2001 | 2183 | 6.90 |
| 1995 | 1855 | 5.86 |
| 1996 | 1803 | 5.70 |
| 1998 | 1752 | 5.54 |
| 1997 | 1693 | 5.35 |
| 1991 | 1432 | 4.53 |
| 2003 | 1431 | 4.52 |
| 1993 | 1324 | 4.19 |
| 1994 | 1254 | 3.96 |
| 1992 | 1215 | 3.84 |
| 1990 | 1159 | 3.66 |
| 1989 | 1089 | 3.44 |
| 1988 | 967 | 3.06 |
| 1987 | 545 | 1.72 |
| 2004 | 484 | 1.53 |
| 1984 | 459 | 1.45 |
| 1983 | 452 | 1.43 |
| 1985 | 431 | 1.36 |
| 1986 | 379 | 1.20 |
| 0 | 374 | 1.18 |
| 1976 | 281 | 0.89 |
| 1982 | 269 | 0.85 |
| 1978 | 243 | 0.77 |
| 1981 | 189 | 0.60 |
| 1980 | 135 | 0.43 |
| 1979 | 102 | 0.32 |
| 1977 | 86 | 0.27 |
| ⋮ | ⋮ | ⋮ |

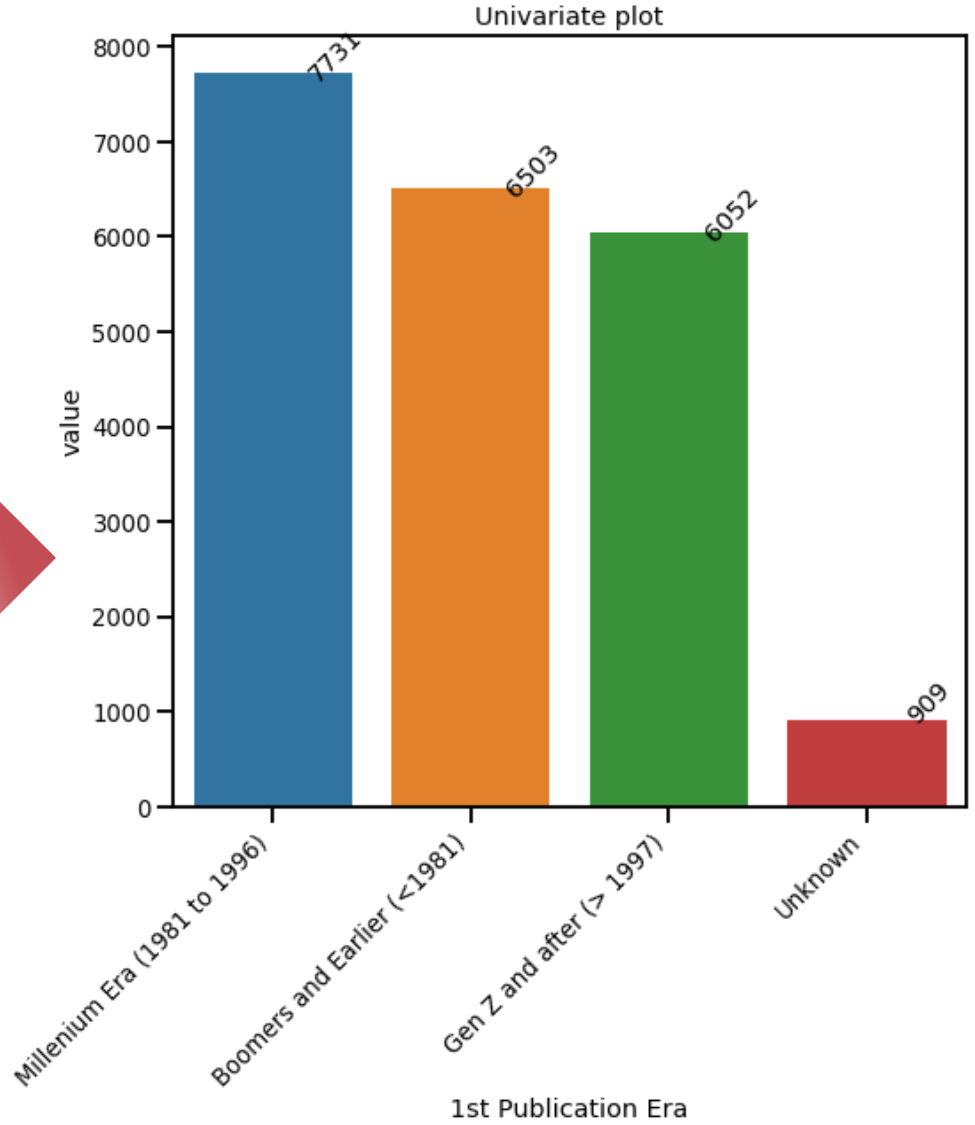


# YEAR OF PUBLICATION



# FIRST PUBLICATION ERA

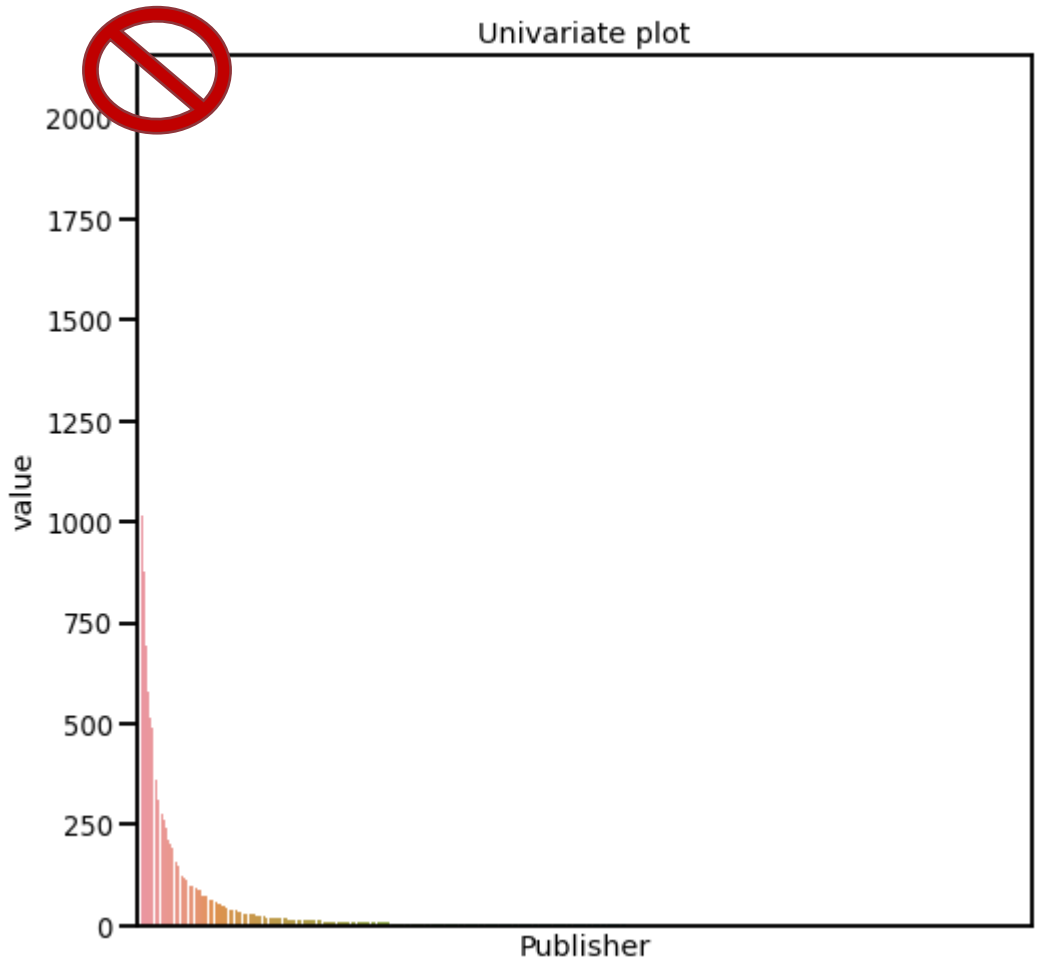


# PUBLISHER

```
This is univariate analysis for ' Publisher '
                      Count  Percentage
Ballantine Books       2054        6.49
Pocket                 1454        4.60
Vintage Books USA      1314        4.15
Bantam Books           1106        3.50
Warner Books           1053        3.33
...                     ...         ...
Liamworks                 0        0.00
Learning Links            0        0.00
Lawrence Hill Books       0        0.00
Last Gasp                 0        0.00
scholastic                0        0.00

[1025 rows x 2 columns]
```



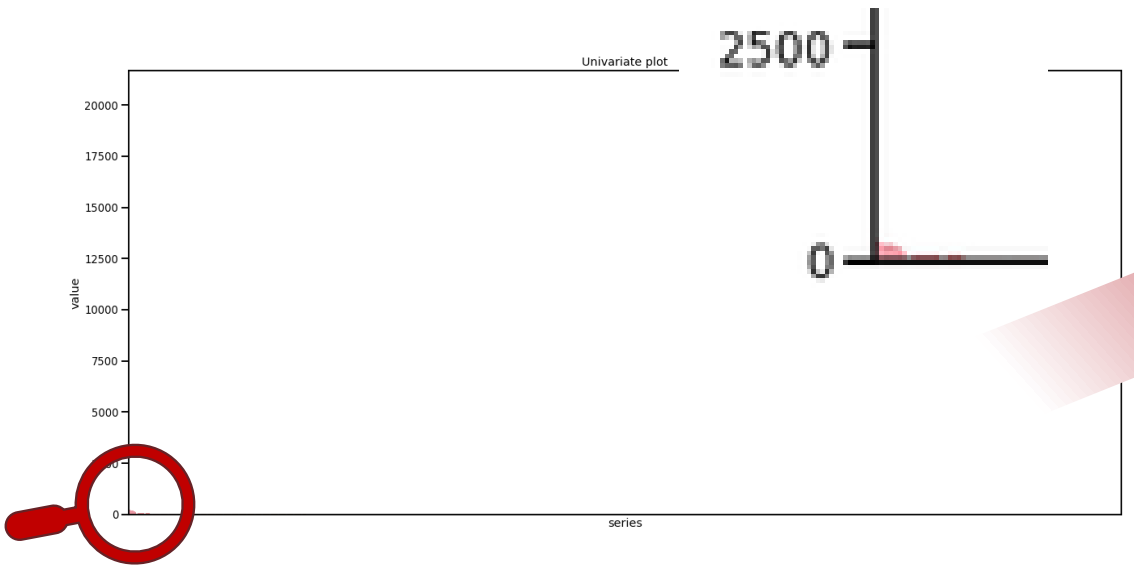

- Hodder & Stoughton General Division
- Pan Books in association with Secker & Warburg
- Henry Holt & Company, Inc.
- LÃ?Â¼bbe
- DenoÃ?Â«l
- Le Serpent Ã?Â Plumes
- Hachette Littà©rature

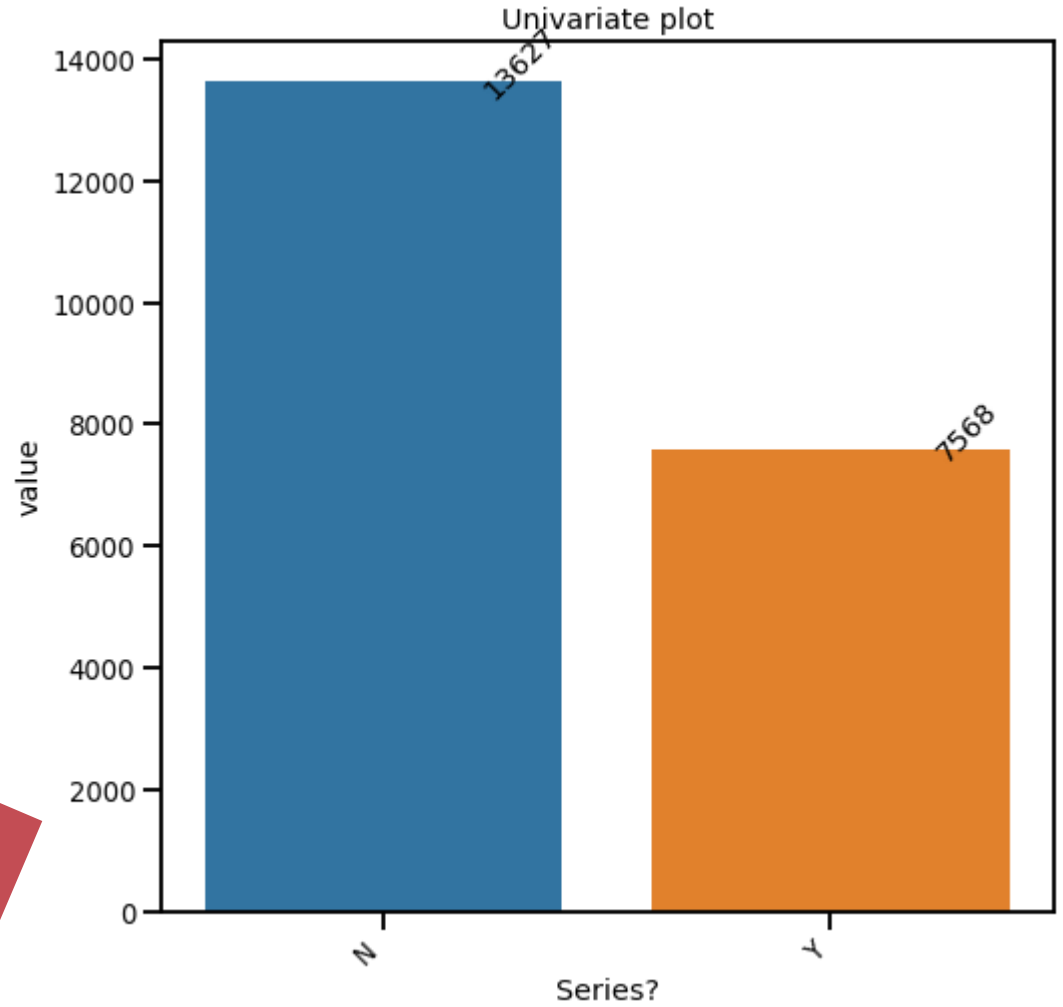
There's many publishers with only 1 book. This feature will be dropped.

# SERIES?

```
This is univariate analysis for ' series '
                            Count  Percentage
NaN                         20641       65.26
To Kill a Mockingbird         267        0.84
The Vampire Chronicles #1     232        0.73
Jurassic Park #1              201        0.64
Greer Family #1               192        0.61
...                           ...         ...
Silver Brumby #1                0        0.00
Lad #1                          0        0.00
Courtney #1                     0        0.00
Extraordinary Voyages #16       0        0.00
Culture #1                      0        0.00

[1354 rows x 2 columns]
```



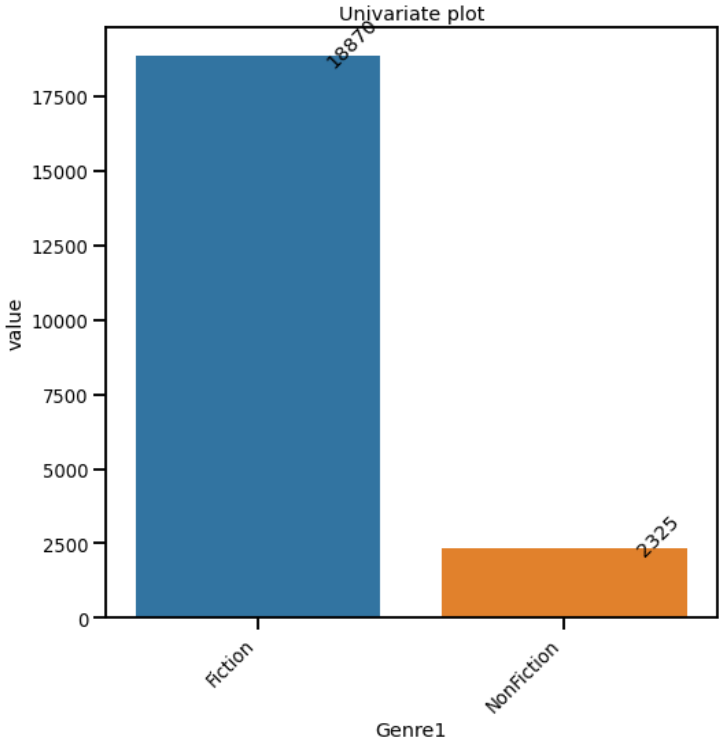



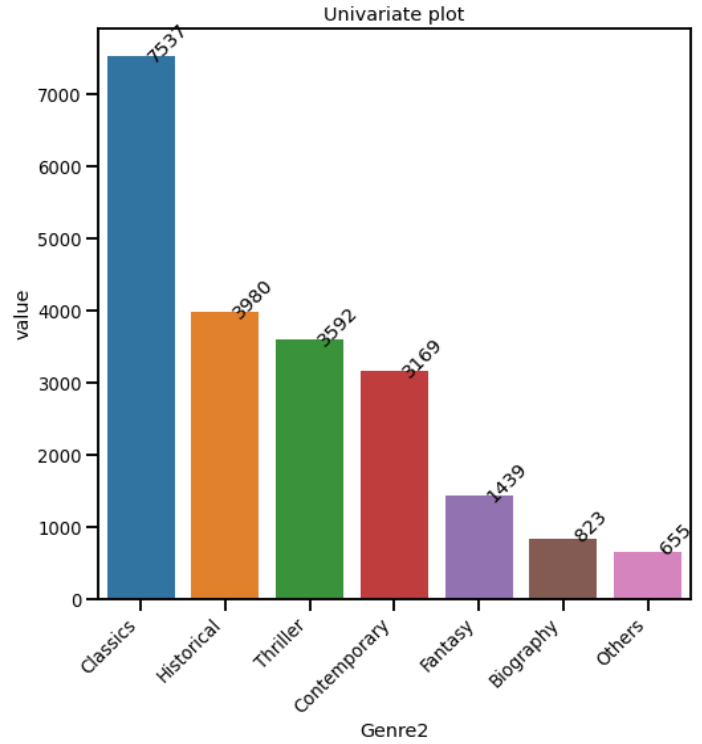
# GENRE1 AND GENRE2

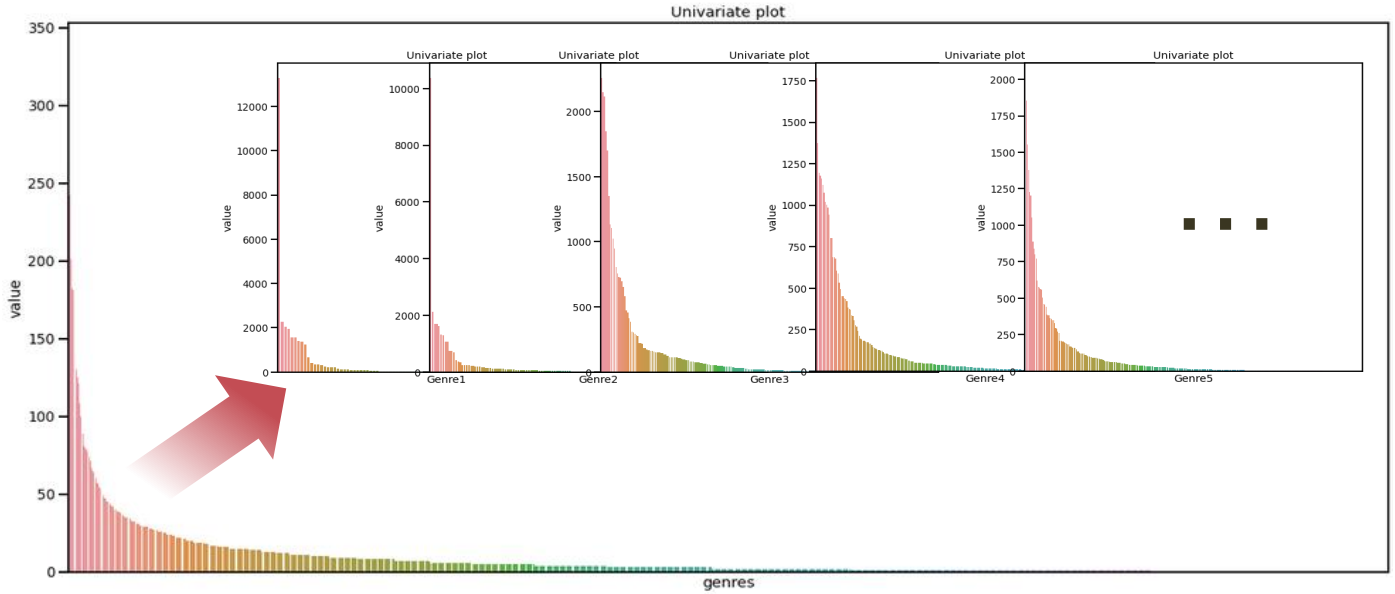
```
################ Univariate analysis - Categorical ###############

This is univariate analysis for ' genres '
                                                    Count  Percentage
['Fiction', 'Fantasy', 'Classics', 'Adventure',...    336        1.06
['Fiction', 'Historical Fiction', 'Mystery', 'H...    271        0.86
['Classics', 'Fiction', 'Historical Fiction', '...    267        0.84
['Fiction', 'Chick Lit', 'Short Stories', 'Cont...    259        0.82
['Classics', 'Fiction', 'Young Adult', 'Literat...    258        0.82
...                                                   ...         ...
['Fiction', 'Classics', 'Russia', 'Russian Lite...      0        0.00
['Fiction', 'Classics', 'Romance', 'Historical ...      0        0.00
['Historical Fiction', 'Romance', '18th Century...      0        0.00
['Fiction', 'Classics', 'Poetry', 'Russia', 'Li...      0        0.00
['Historical Fiction', 'Fiction', 'Adventure', ...      0        0.00

[3879 rows x 2 columns]
```



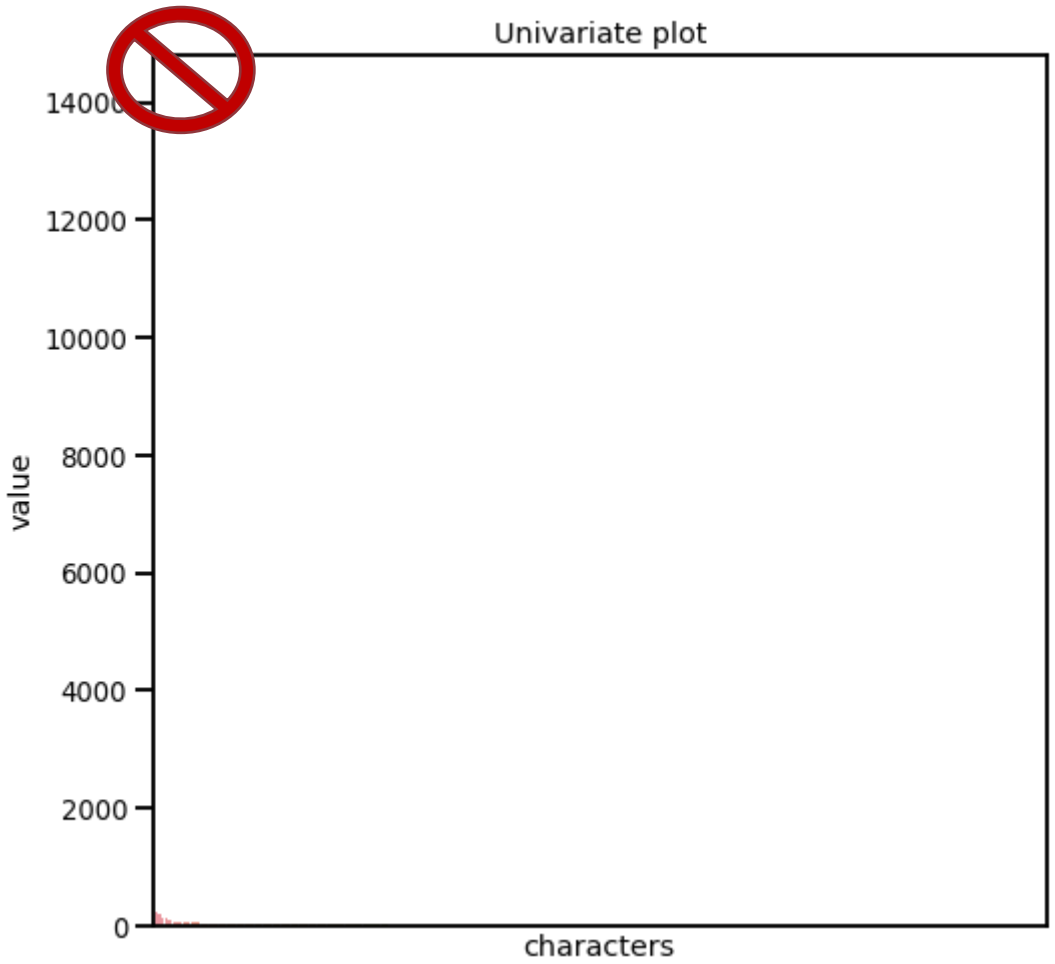

- Classics
- Historical
- Thriller
- Biography
- Contemporary
- Fantasy
- Others

# CHARACTERS

This is univariate analysis for ' characters '

| | Count | Percentage |
|---|---|---|
| [] | 14105 | 44.59 |
| ['David Sedaris'] | 345 | 1.09 |
| ['Pi Patel', 'Richard Parker'] | 336 | 1.06 |
| ['Kabuo Miyamoto', 'Ishmael Chambers', 'Hatsue ... | 271 | 0.86 |
| ['Scout Finch', 'Atticus Finch', 'Jem Finch', '... | 267 | 0.84 |
| ... | ... | ... |
| ['Sita', 'Seymour Dorsten', 'Ray Riley'] | 0 | 0.00 |
| ['Joe Sackett', 'Borden Chantry', 'Bess Chantry... | 0 | 0.00 |
| ['Joey', 'Albert', 'Sir Nicholls', 'Emily'] | 0 | 0.00 |
| ['Nicholas Bragg', 'Jane Barclay'] | 0 | 0.00 |
| ['Giorgio Viola', "Gian' Battista Fidanza", 'Ch... | 0 | 0.00 |

[1438 rows x 2 columns]

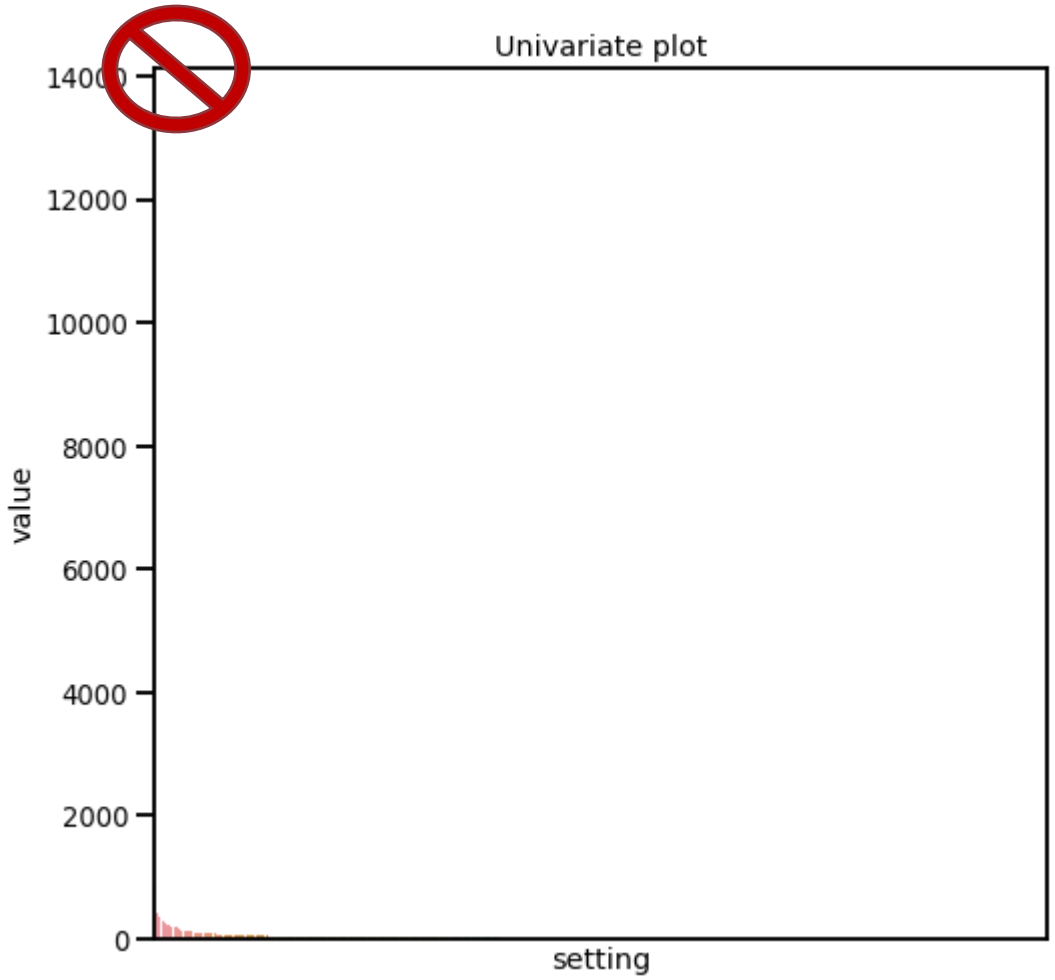


# SETTING

This is univariate analysis for ' setting '

| | Count | Percentage |
|---|---|---|
| [] | 13465 | 42.57 |
| ['United States of America'] | 857 | 2.71 |
| ['London, England'] | 420 | 1.33 |
| ['Dublin (Ireland)'] | 419 | 1.32 |
| ['California (United States)'] | 359 | 1.13 |
| ... | ... | ... |
| ['Bahrain'] | 0 | 0.00 |
| ['Gaze Castle (United Kingdom)'] | 0 | 0.00 |
| ['Rome (Italy)', 'Italy'] | 0 | 0.00 |
| ['Discworld', 'The Chalk'] | 0 | 0.00 |
| ['South Pacific', 'Anopopei'] | 0 | 0.00 |

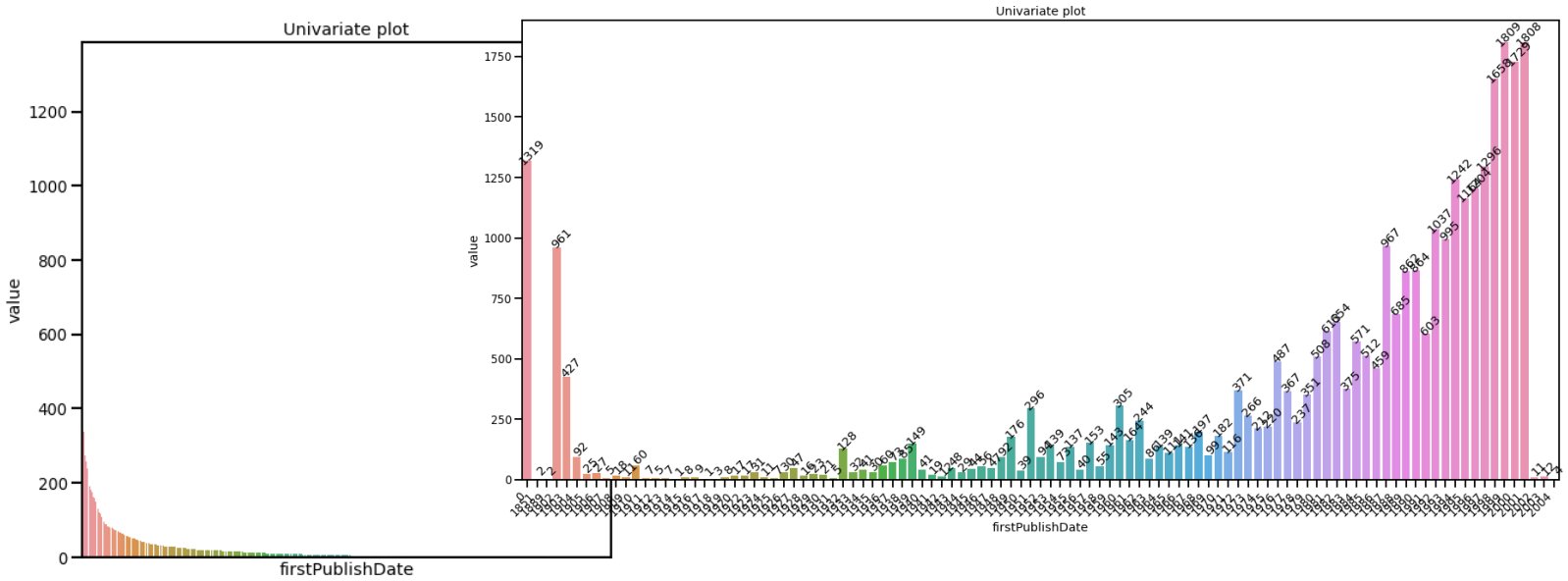
[806 rows x 2 columns]

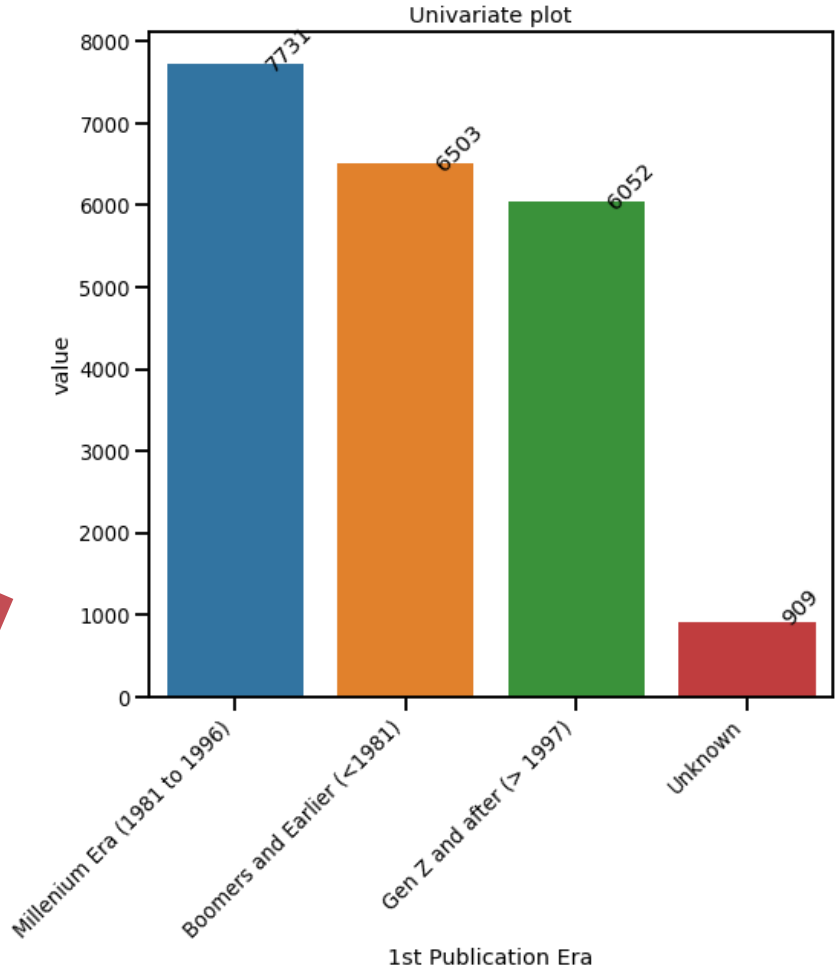


# 1ST PUBLICATION ERA

```
This is univariate analysis for ' firstPublishDate '
            Count  Percentage
NaN          1321        4.18
10/28/99      444        1.40
10/28/93      435        1.38
10/30/00      355        1.12
10/28/98      337        1.07
...           ...         ...
07/30/02        0        0.00
01/01/75        0        0.00
08/01/70        0        0.00
08/01/97        0        0.00
09/27/48        0        0.00

[1599 rows x 2 columns]
```



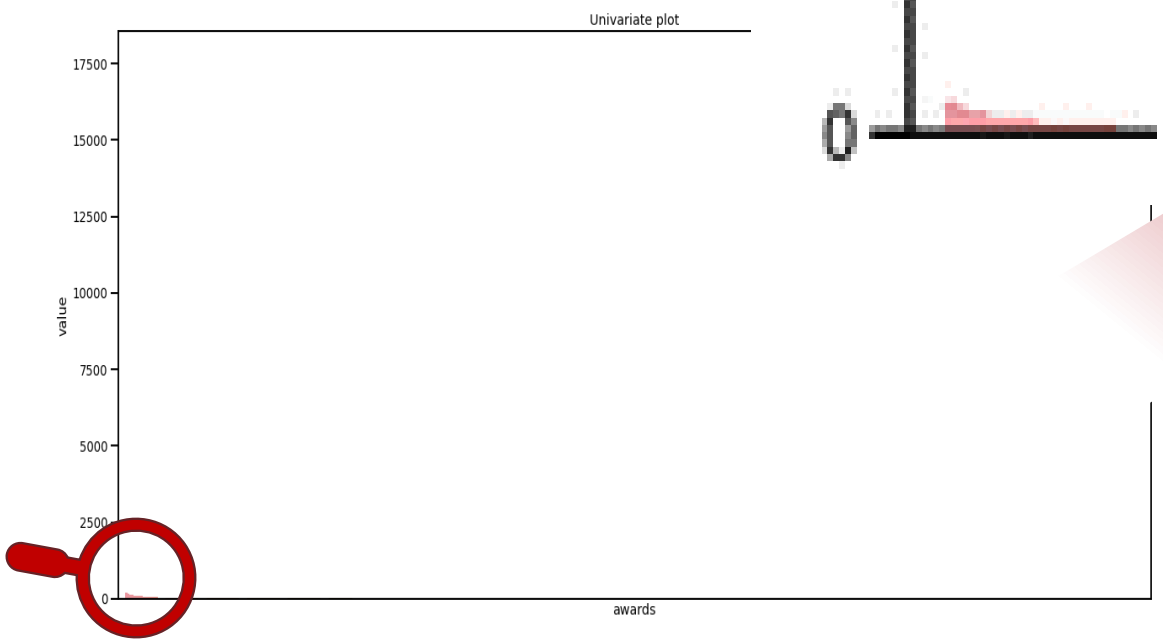



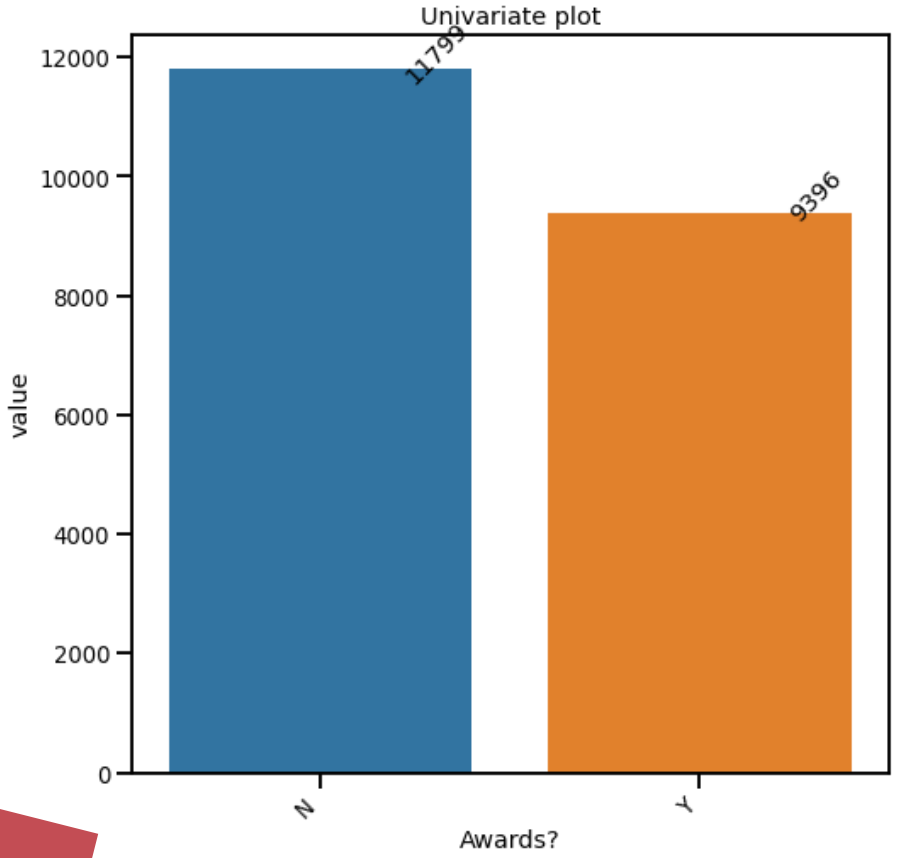
# AWARDS?

This is univariate analysis for ' awards '

| | Count | Percentage |
|---|---|---|
| [] | 17677 | 55.89 |
| ['Booker Prize (2002)', 'Bollinger Everyman Wod... | 336 | 1.06 |
| ['Anthony Award Nominee for Best First Novel (1... | 271 | 0.86 |
| ['Pulitzer Prize for Fiction (1961)', 'Audie Aw... | 267 | 0.84 |
| ['Guardian First Book Award Nominee for Longlis... | 259 | 0.82 |
| ... | ... | ... |
| ['ECPA Christian Book Award for Biography / Aut... | 0 | 0.00 |
| ['Nebula Award Nominee for Novel (1976)', 'Jame... | 0 | 0.00 |
| ['Nebula Award Nominee for Novel (1975)'] | 0 | 0.00 |
| ['Edgar Award Nominee for Best Novel (1979)'] | 0 | 0.00 |
| ['Booker Prize Nominee (1979)'] | 0 | 0.00 |

[1253 rows x 2 columns]

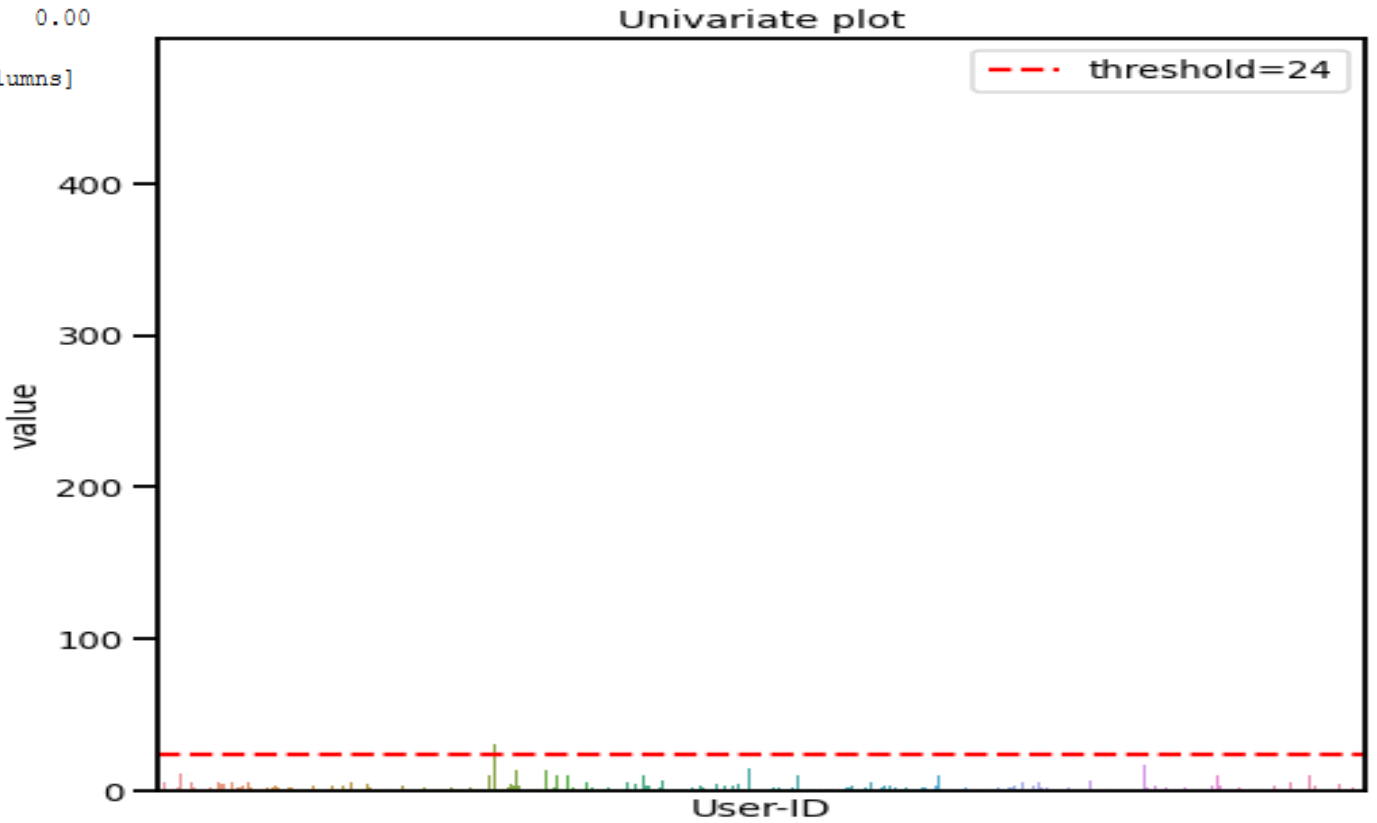


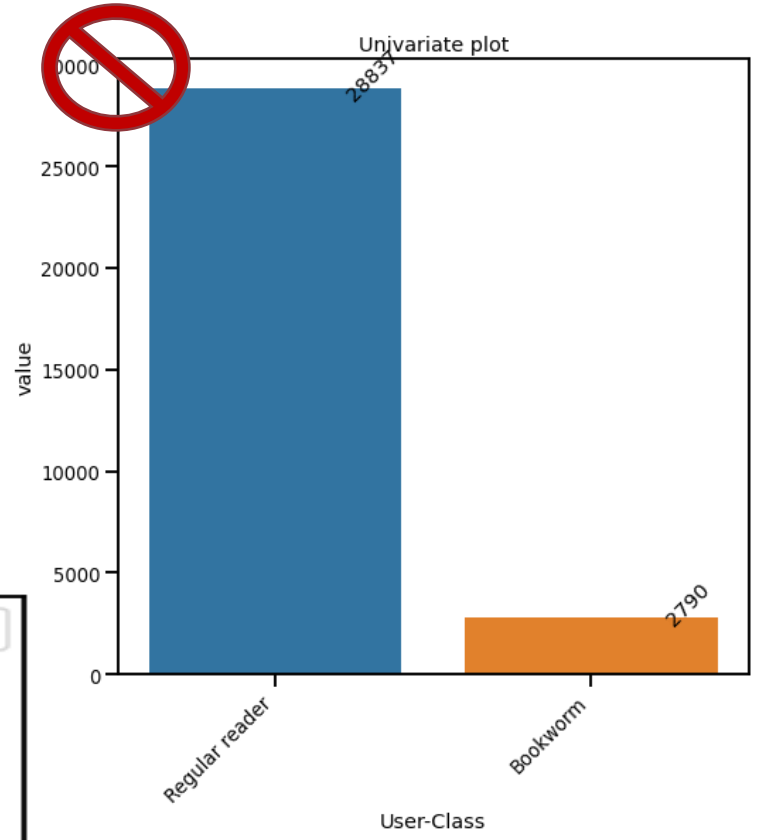



# USER-ID

```
This is univariate analysis for ' User-ID '
         Count  Percentage
11676      472        1.49
23902      106        0.34
98391       93        0.29
153662      78        0.25
95359       76        0.24
...        ...         ...
107972       1        0.00
107980       1        0.00
107994       1        0.00
108016       1        0.00
278846       1        0.00

[14987 rows x 2 columns]
```

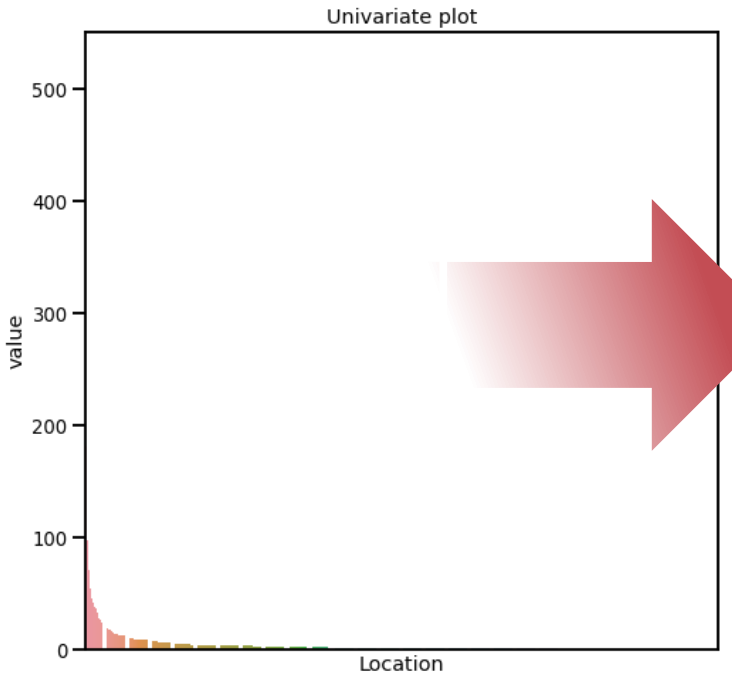


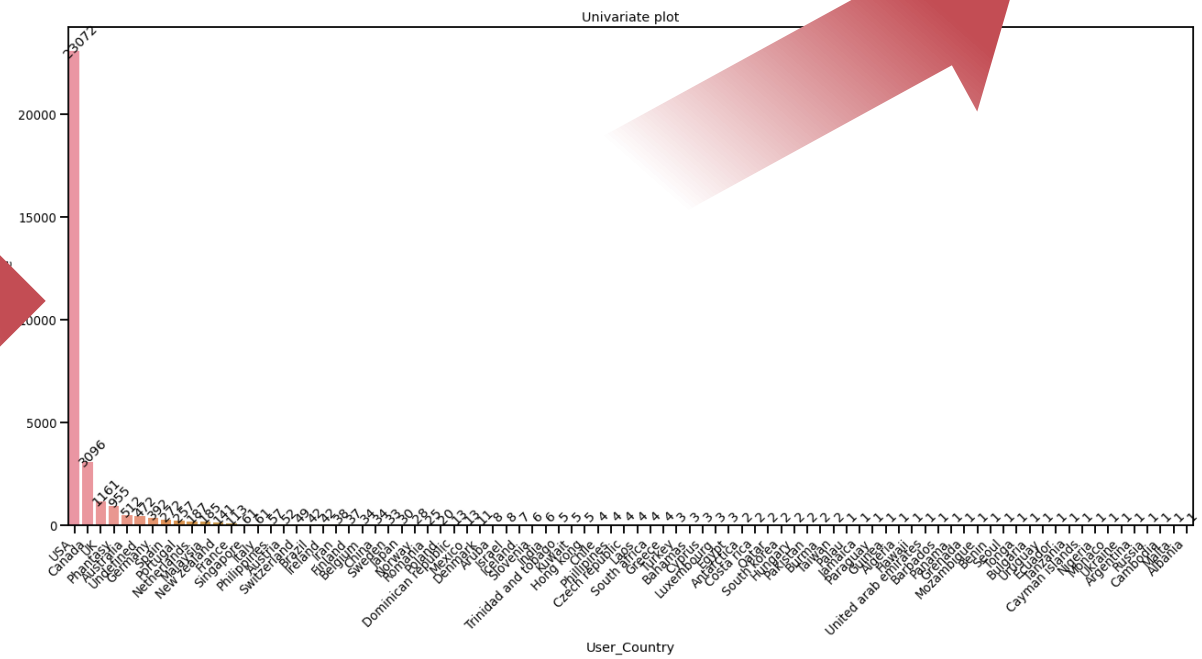



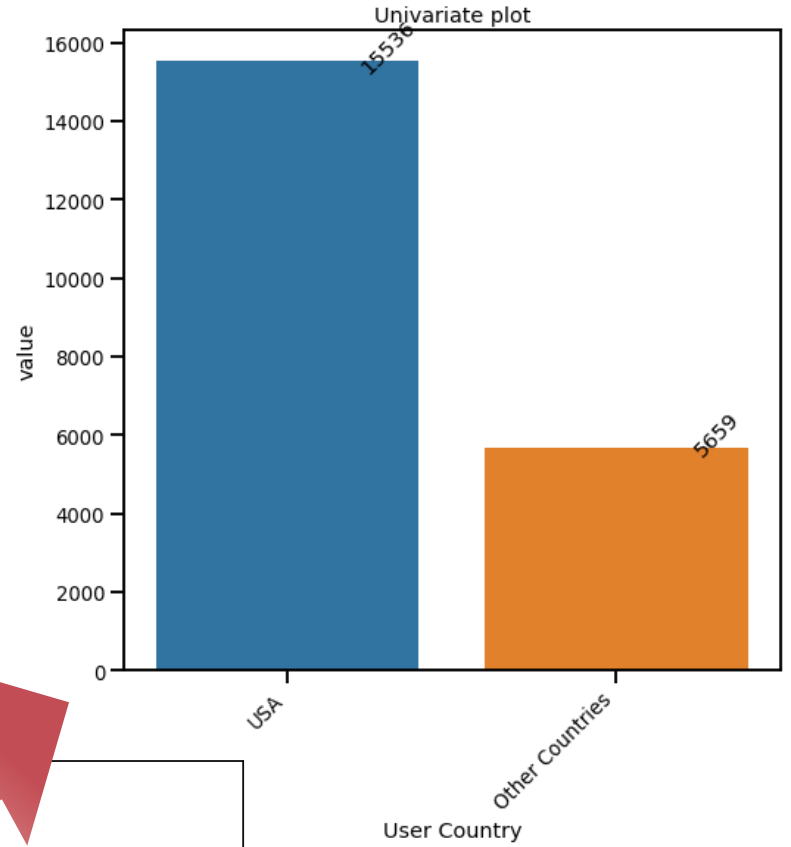
# USER COUNTRY

```
This is univariate analysis for ' Location '
                                  Count  Percentage
toronto, ontario, canada            525        1.66
n/a, n/a, n/a                       472        1.49
chicago, illinois, usa              304        0.96
seattle, washington, usa            289        0.91
london, england, united kingdom     280        0.89
...                                 ...         ...
ringgold, georgia, usa                0        0.00
franconia, virginia, usa              0        0.00
ringwood north, victoria, australia   0        0.00
rio de janeiro, na, brazil            0        0.00
london, england,                      0        0.00

[8063 rows x 2 columns]
```

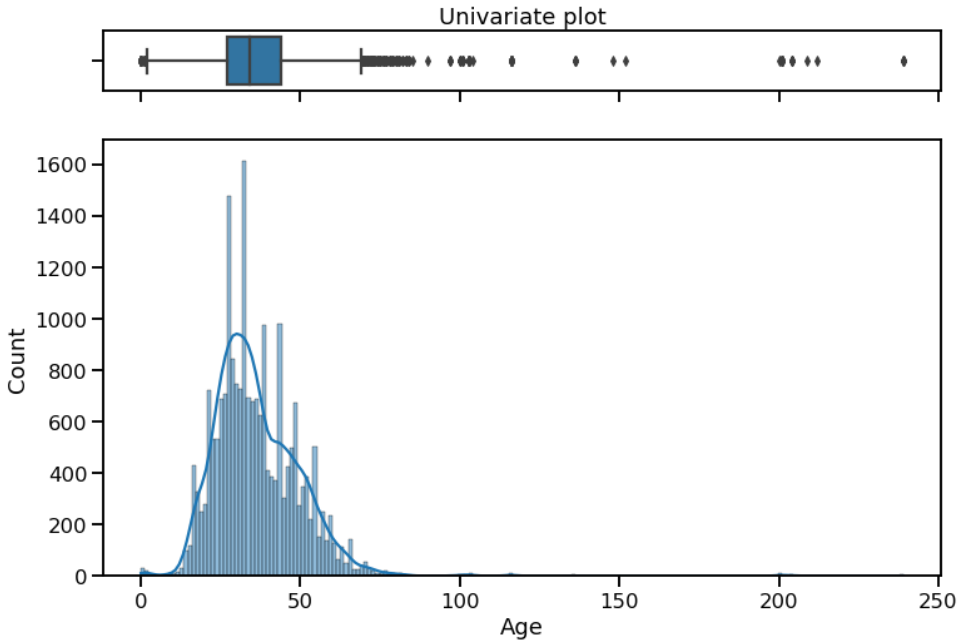


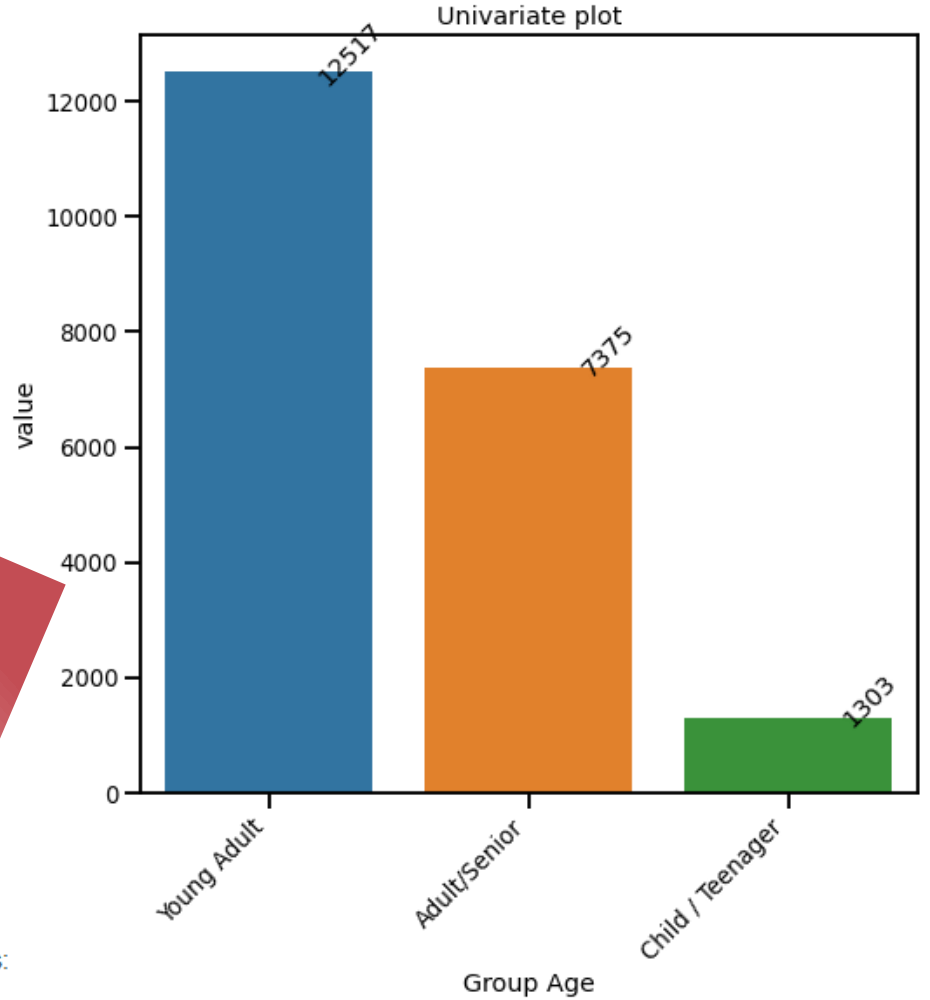



# GROUP AGE

```
This is univariate analysis for ' Age '
 count                          21281.0
mean                           36.42794
std                           13.911819
min                                 0.0
25%                                27.0
50%                                34.0
75%                                44.0
max                               239.0
variance                      193.53871
IQR                                17.0
range                             239.0
skewness                       2.635699
kurtosis                        25.8589
mode          0     33.0
dtype: float64
Name: Age, dtype: object
```




As we saw before there are some outliers values:

- Below 5 years old
- Greater than 122 years old

These values will be dropped as well any missing values. Furthermore, age will be turned into category:

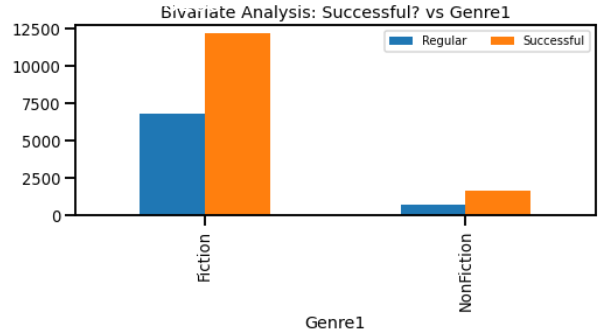
- Child/Teenager: <=19
- Young Adult: <39
- Adult/Senior

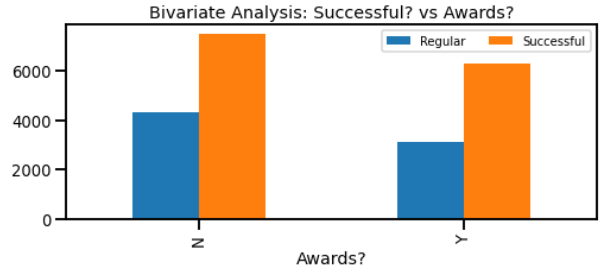
Age will be dropped after.

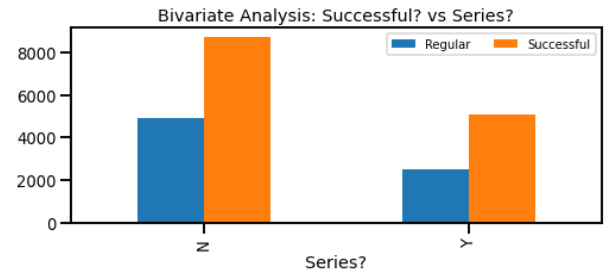
# BIVARIATE ANALYSIS

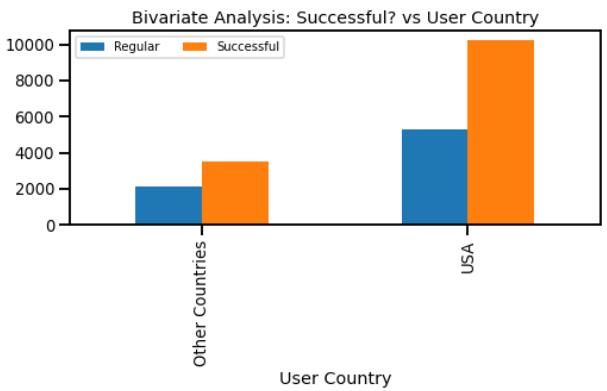
| Feature | Level | Successful? | | p-value |
|---|---|---|---|---|
| | | Regular | Successful | |
| Awards? | N | 4314 | 7485 | 2.33e-7 |
| | Y | 3114 | 6282 | |
| Genre1 | Fiction | 6753 | 12117 | 1.38e-10 |
| | NonFiction | 675 | 1650 | |
| Genre2 | Biography | 260 | 563 | 1.73e-46 |
| | Classics | 2241 | 5296 | |
| | Contemporary | 1314 | 1855 | |
| | Fantasy | 473 | 966 | |
| | Historical | 1427 | 2553 | |
| | Others | 218 | 437 | |
| | Thriller | 1495 | 2097 | |
| Group Age | Adult/Senior | 2498 | 4877 | 7.68e-3 |
| | Child/Teenager | 437 | 866 | |
| | Young Adult | 4493 | 8024 | |
| Series? | N | 4921 | 8706 | 1.36e-5 |
| | Y | 2507 | 5061 | |
| User Country | Other Countries | 2135 | 3524 | 8.57e-7 |
| | USA | 5293 | 10243 | |

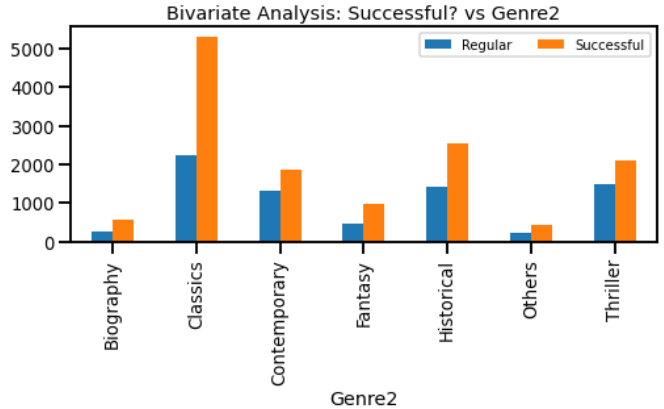
| Feature | Level | Successful? | | p-value |
|---|---|---|---|---|
| | | Regular | Successful | |
| 1st Publication Era | Unknown | 341 | 568 | 1.26e-19 |
| | Boomers and Earlier (<1981) | 2011 | 4492 | |
| | Millenium Era (1981 to 1996) | 2720 | 5011 | |
| | Gen Z and after (< 1997) | 2356 | 3696 | |
| Year Of Publication | Before 1985 | 604 | 1438 | 1.94e-15 |
| | 1985 | 96 | 199 | |
| | 1986 | 74 | 174 | |
| | 1987 | 127 | 256 | |
| | 1988 | 197 | 464 | |
| | 1989 | 197 | 509 | |
| | 1990 | 236 | 579 | |
| | 1991 | 374 | 617 | |
| | 1992 | 287 | 513 | |
| | 1993 | 295 | 622 | |
| | 1994 | 317 | 540 | |
| | 1995 | 442 | 788 | |
| | 1996 | 466 | 774 | |
| | 1997 | 405 | 699 | |
| | 1998 | 434 | 745 | |
| | 1999 | 610 | 1014 | |
| | 2000 | 647 | 974 | |
| | 2001 | 518 | 876 | |
| | 2002 | 597 | 1034 | |
| | 2003 | 394 | 685 | |
| | 2004 | 111 | 267 | |
| | 2002 | 597 | 1034 | |
| | 2003 | 394 | 685 | |
| | 2004 | 111 | 267 | |

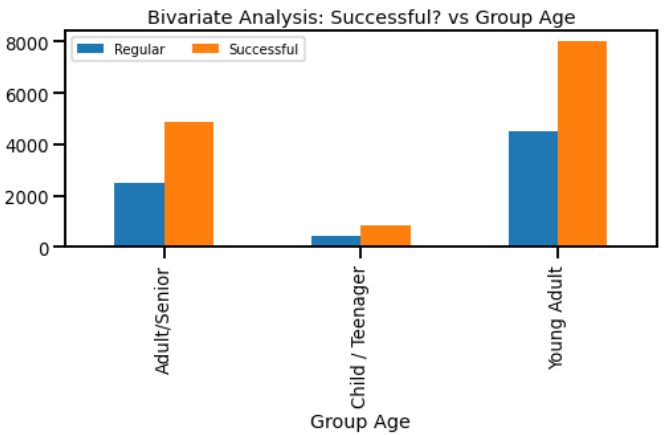

Bivariate Analysis: Successful? vs Genre1
Regular
Successful
12500
10000
7500
5000
2500
0
Fiction
NonFiction
Genre1

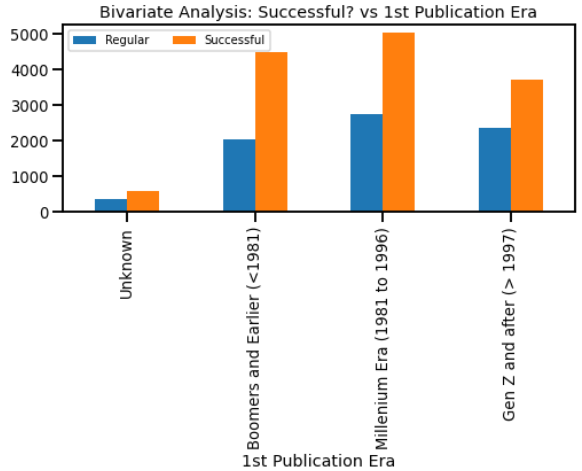

Bivariate Analysis: Successful? vs Awards?
Regular
Successful
6000
4000
2000
0
N
Y
Awards?

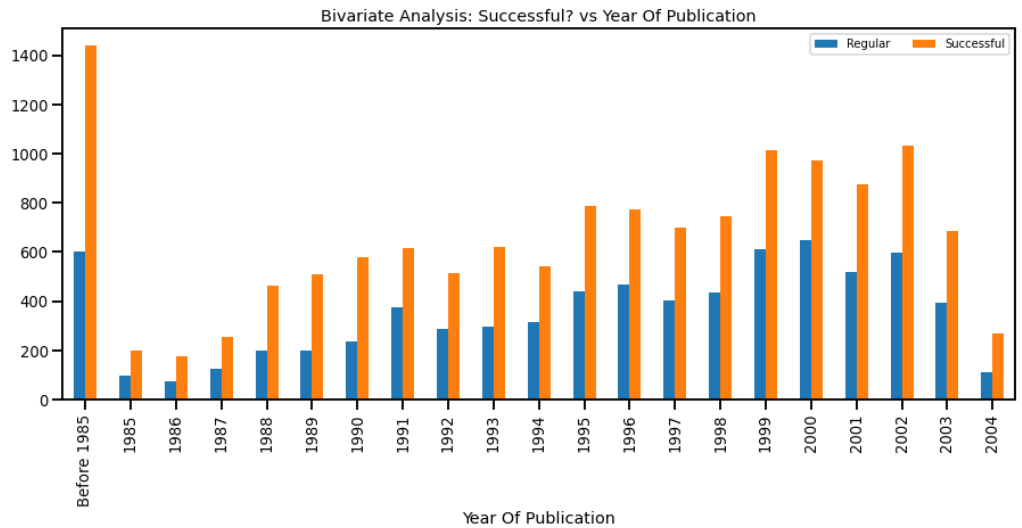

Bivariate Analysis: Successful? vs Series?
Regular
Successful
8000
6000
4000
2000
0
N
Y
Series?


Bivariate Analysis: Successful? vs User Country
Regular
Successful
10000
8000
6000
4000
2000
0
Other Countries
USA
User Country


Bivariate Analysis: Successful? vs Genre2
Regular
Successful
5000
4000
3000
2000
1000
0
Biography
Classics
Contemporary
Fantasy
Historical
Others
Thriller
Genre2


Bivariate Analysis: Successful? vs Group Age
Regular
Successful
8000
6000
4000
2000
0
Adult/Senior
Child / Teenager
Young Adult
Group Age

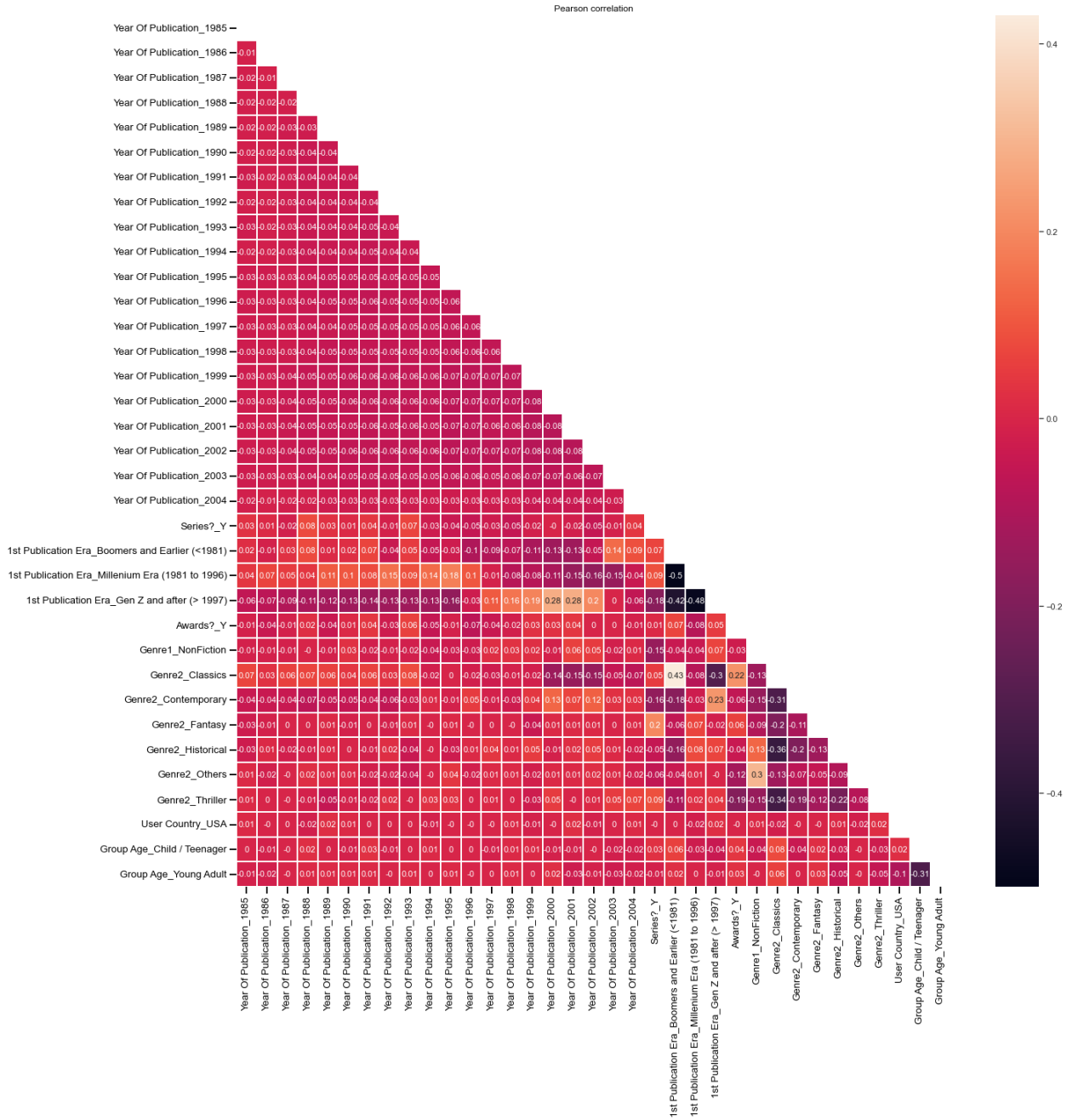

Bivariate Analysis: Successful? vs 1st Publication Era
Regular
Successful
5000
4000
3000
2000
1000
0
Unknown
Boomers and Earlier (<1981)
Millenium Era (1981 to 1996)
Gen Z and after (> 1997)
1st Publication Era


Bivariate Analysis: Successful? vs Year Of Publication
Regular
Successful
1400
1200
1000
800
600
400
200
0
Before 1985
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
Year Of Publication

# MODEL BUILDING AND EVALUATION

1st Publication Era:
Millenium Era (1981 to 1996)
x
1st Publication Era:
Boomers and Earlier

max r = -0.5



# GRIDSEARCHCV:

Scoring: balanced accuracy

“The balanced accuracy in binary and multiclass classification problems to deal with imbalanced datasets. It is defined as the average of recall obtained on each class.”

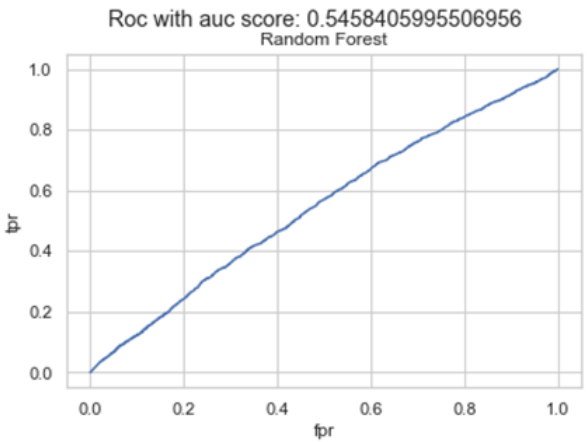
Source: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.balanced_accuracy_score.html#sklearn.metrics.balanced_accuracy_score

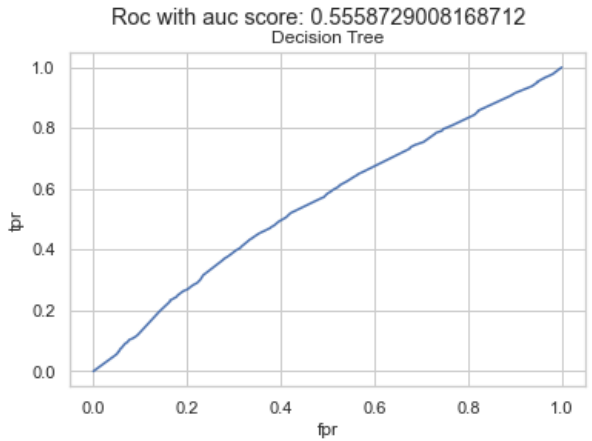
# MODEL EVALUATION:

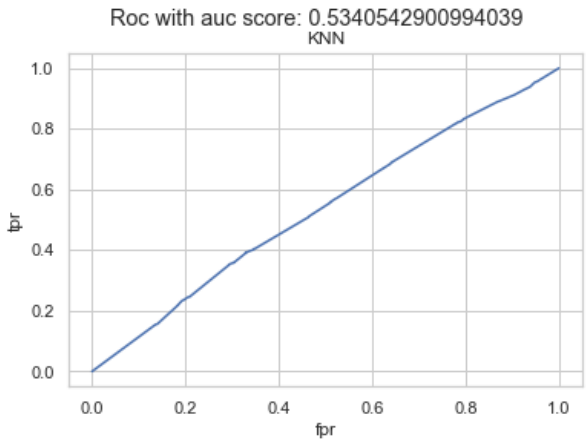
AUC

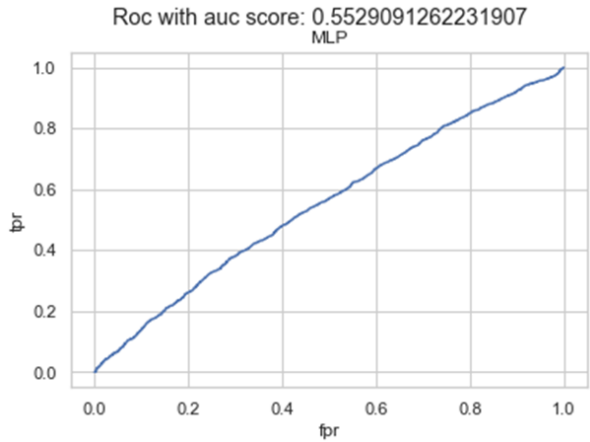
$$\text{Precision} = \frac{TP}{\text{TP} + \text{FP}} = \frac{\text{Correctly classified positives}}{\text{All positives predictions}}$$

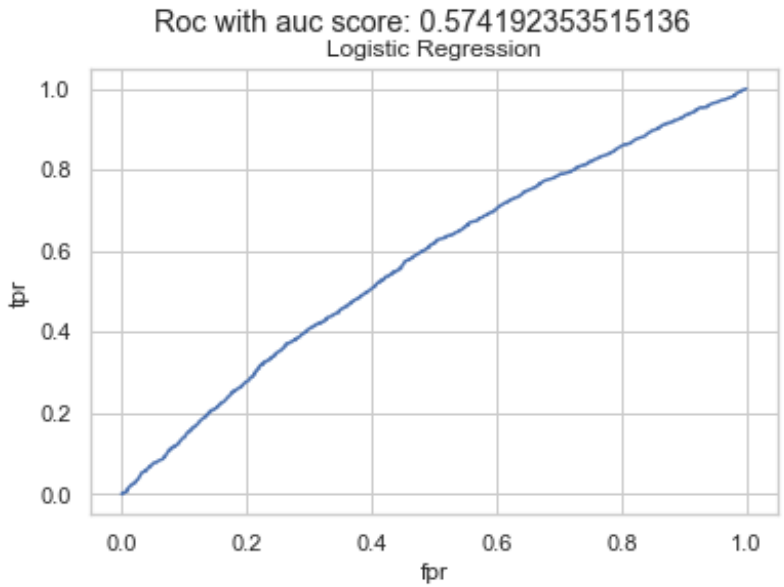
Confusion Matrix

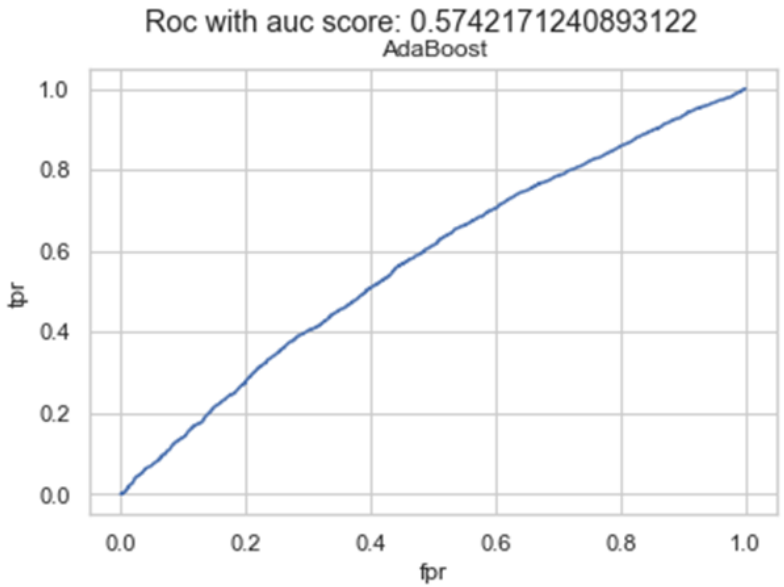

Roc with auc score: 0.5458405995506956
Random Forest
tpr
fpr

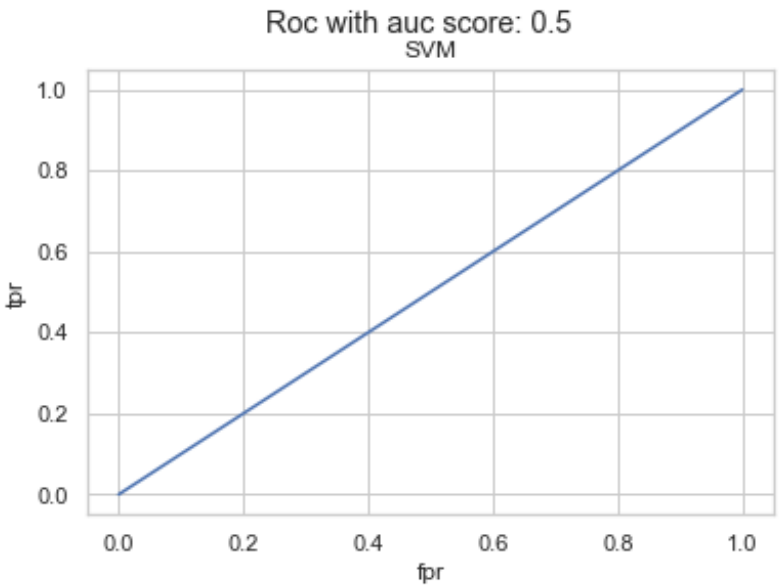

Roc with auc score: 0.5558729008168712
Decision Tree
tpr
fpr


Roc with auc score: 0.5340542900994039
KNN
tpr
fpr


Roc with auc score: 0.5529091262231907
MLP
tpr
fpr


Roc with auc score: 0.574192353515136
Logistic Regression
tpr
fpr


Roc with auc score: 0.5742171240893122
AdaBoost
tpr
fpr


Roc with auc score: 0.5
SVM
tpr
fpr

Classification report - SVM

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| Regular | 0.35 | 1.00 | 0.51 | 2197 |
| Successful | 0.00 | 0.00 | 0.00 | 4162 |
| accuracy | | | 0.35 | 6359 |
| macro avg | 0.17 | 0.50 | 0.26 | 6359 |
| weighted avg | 0.12 | 0.35 | 0.18 | 6359 |

Classification report - KNN

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| Regular | 0.38 | 0.36 | 0.37 | 2197 |
| Successful | 0.67 | 0.69 | 0.68 | 4162 |
| accuracy | | | 0.58 | 6359 |
| macro avg | 0.52 | 0.52 | 0.52 | 6359 |
| weighted avg | 0.57 | 0.58 | 0.57 | 6359 |

Classification report - AdaBoost

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| Regular | 0.47 | 0.00 | 0.01 | 2197 |
| Successful | 0.65 | 1.00 | 0.79 | 4162 |
| accuracy | | | 0.65 | 6359 |
| macro avg | 0.56 | 0.50 | 0.40 | 6359 |
| weighted avg | 0.59 | 0.65 | 0.52 | 6359 |

Classification report - MLP

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| Regular | 0.41 | 0.23 | 0.29 | 2197 |
| Successful | 0.67 | 0.83 | 0.74 | 4162 |
| accuracy | | | 0.62 | 6359 |
| macro avg | 0.54 | 0.53 | 0.52 | 6359 |
| weighted avg | 0.58 | 0.62 | 0.59 | 6359 |

Classification report - Random Forest

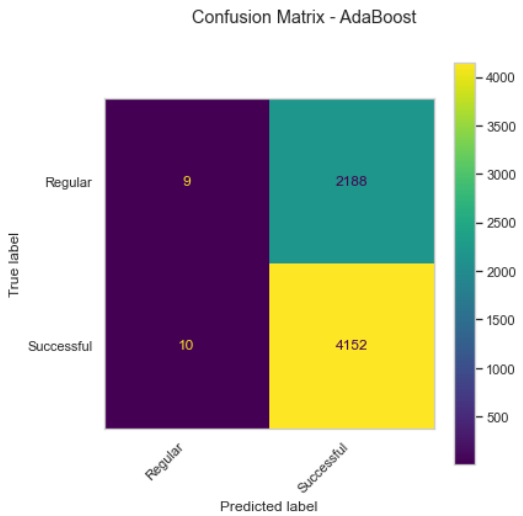
| | precision | recall | f1-score | support |
|---|---|---|---|---|
| Regular | 0.38 | 0.47 | 0.42 | 2197 |
| Successful | 0.68 | 0.60 | 0.64 | 4162 |
| accuracy | | | 0.55 | 6359 |
| macro avg | 0.53 | 0.54 | 0.53 | 6359 |
| weighted avg | 0.58 | 0.55 | 0.56 | 6359 |

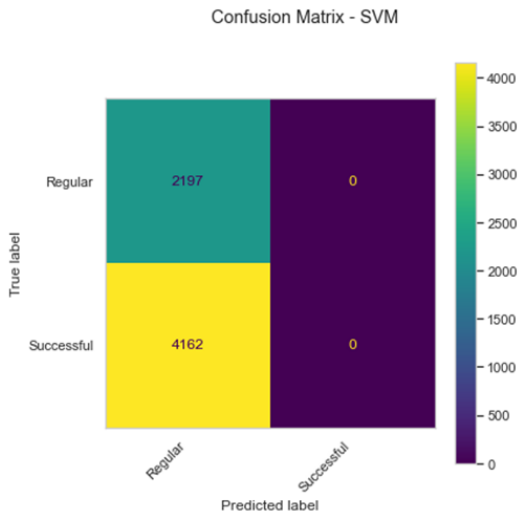
Classification report - Decision Tree

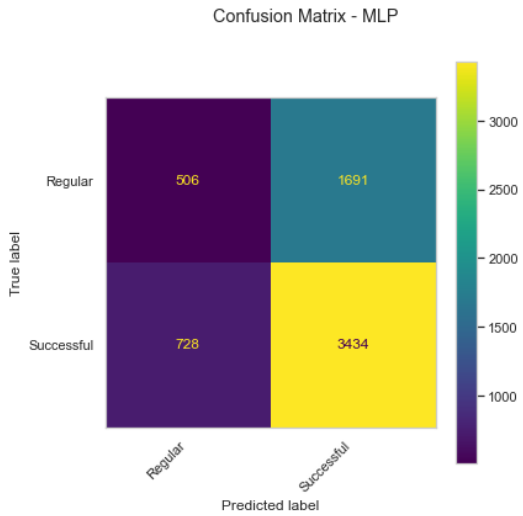
| | precision | recall | f1-score | support |
|---|---|---|---|---|
| Regular | 0.39 | 0.60 | 0.47 | 2197 |
| Successful | 0.70 | 0.49 | 0.58 | 4162 |
| accuracy | | | 0.53 | 6359 |
| macro avg | 0.54 | 0.55 | 0.52 | 6359 |
| weighted avg | 0.59 | 0.53 | 0.54 | 6359 |

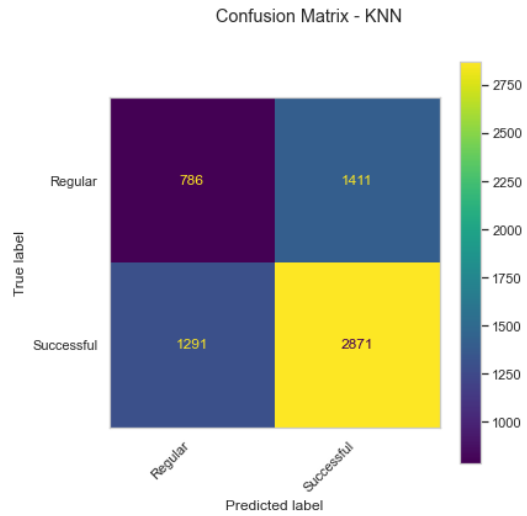
CLassification report - Logistic Regression

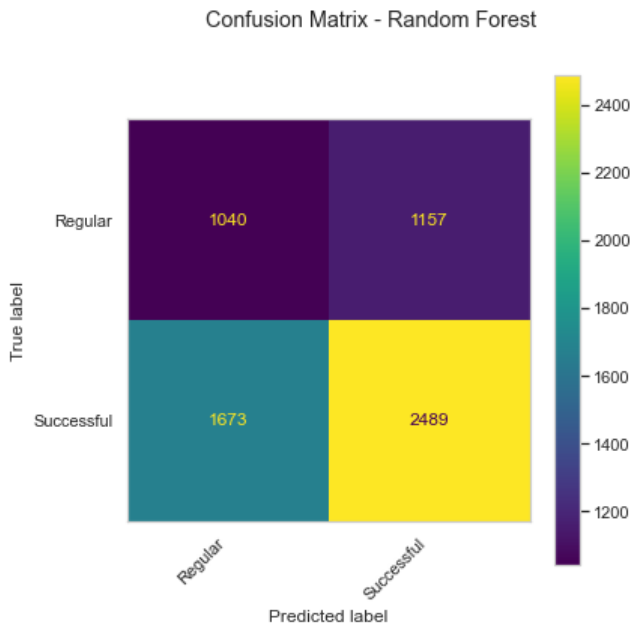
| | precision | recall | f1-score | support |
|---|---|---|---|---|
| Regular | 0.40 | 0.56 | 0.46 | 2197 |
| Successful | 0.70 | 0.56 | 0.62 | 4162 |
| accuracy | | | 0.56 | 6359 |
| macro avg | 0.55 | 0.56 | 0.54 | 6359 |
| weighted avg | 0.60 | 0.56 | 0.57 | 6359 |

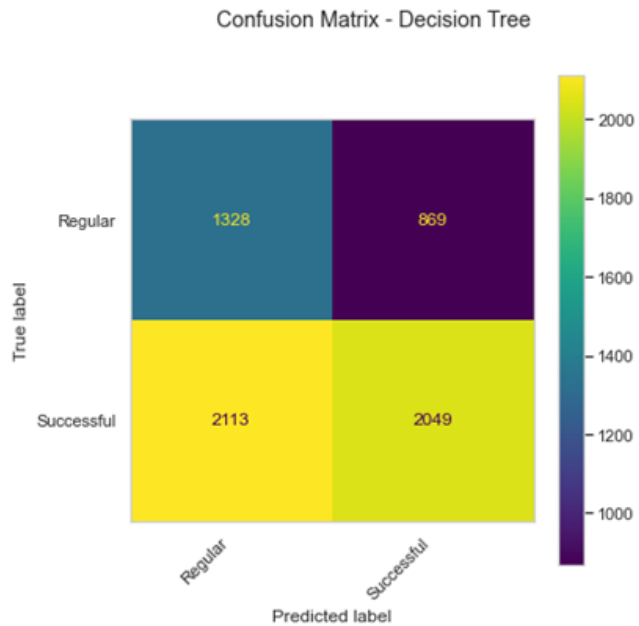

Confusion Matrix - AdaBoost
Regular
Successful
9
2188
10
4152
True label
Predicted label
Regular
Successful
4000
3500
3000
2500
2000
1500
1000
500

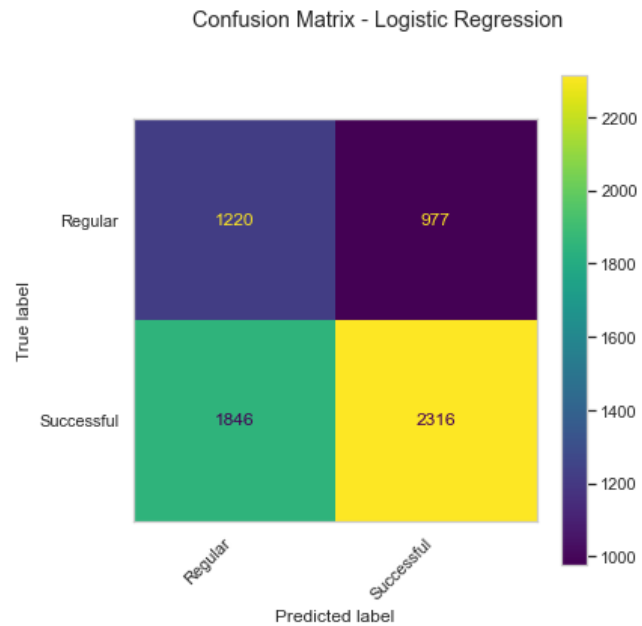

Confusion Matrix - SVM
Regular
Successful
2197
0
4162
0
True label
Predicted label
Regular
Successful
4000
3500
3000
2500
2000
1500
1000
500
0


Confusion Matrix - MLP
Regular
Successful
506
1691
728
3434
True label
Predicted label
Regular
Successful
3000
2500
2000
1500
1000

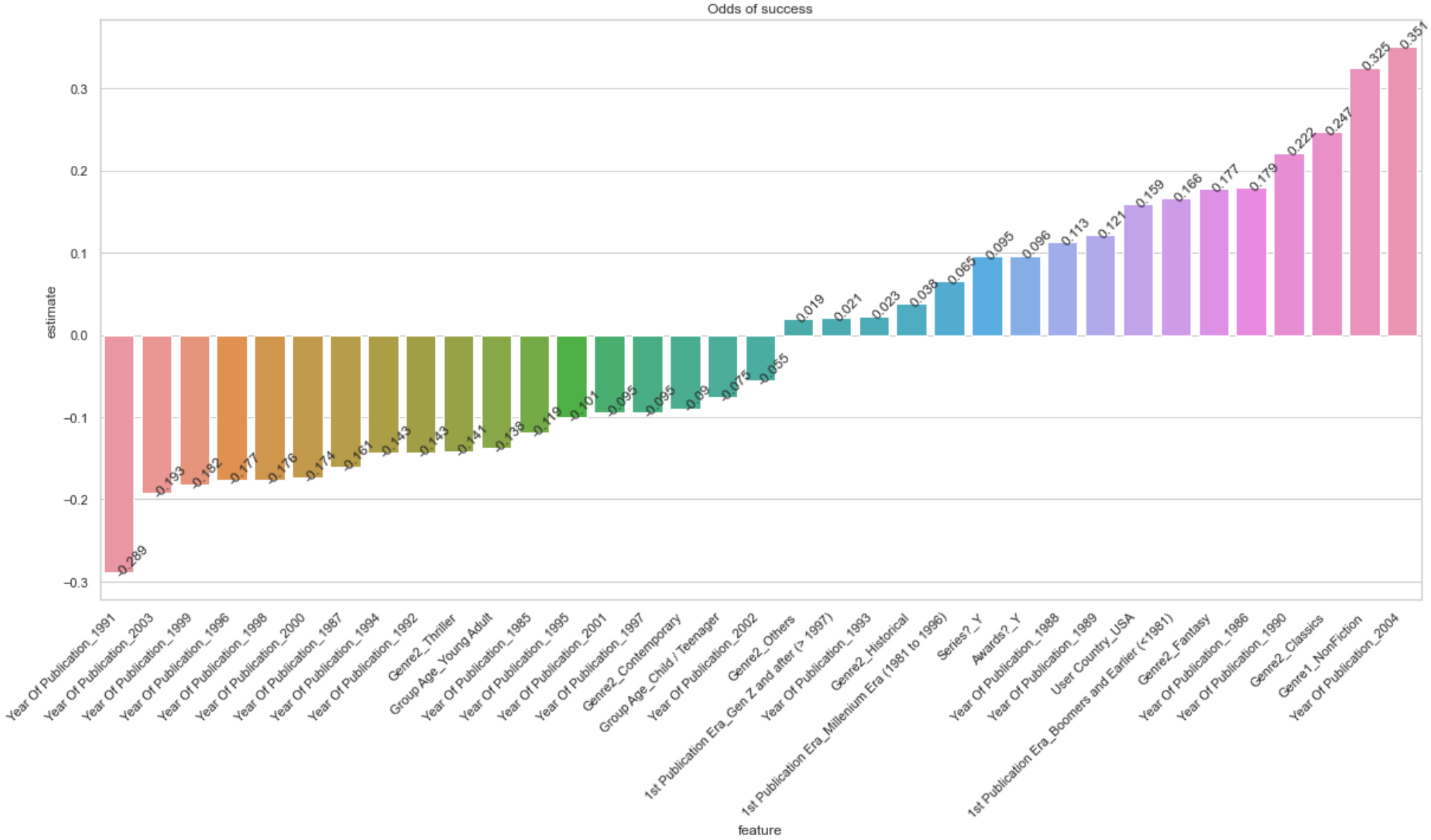

Confusion Matrix - KNN
Regular
Successful
786
1411
1291
2871
True label
Predicted label
Regular
Successful
2750
2500
2250
2000
1750
1500
1250
1000


Confusion Matrix - Random Forest
Regular
Successful
1040
1157
1673
2489
True label
Predicted label
Regular
Successful
2400
2200
2000
1800
1600
1400
1200


Confusion Matrix - Decision Tree
Regular
Successful
1328
869
2113
2049
True label
Predicted label
Regular
Successful
2000
1800
1600
1400
1200
1000


Confusion Matrix - Logistic Regression
Regular
Successful
1220
977
1846
2316
True label
Predicted label
Regular
Successful
2200
2000
1800
1600
1400
1200
1000

# LOGISTIC REGRESSION




Reference groups:

- Adult/Senior
- Fiction
- Biography
- Other Countries
- Series (N)
- Year of Publication Before 1985
- 1st Publication Era: Unknown

# CONCLUSIONS

- Logistic Regression is the model of choice for its simplicity and good performance in correctly classify successful books
- After modeling, we can see, among other characteristics:
    - Books published in 2004 have the most odds of success with more than 35% compared to Before 1985
    - Non-Fiction have the best odds of being evaluated with almost 33% more odds than Fiction.
    - Young Adults are the most difficult group to please with almost 3% less odds of success than an Adult/Senior
    - Classics still makes the best reading having almost 25% more odds of success than a Biography.
    - Thriller has almost 18% odds of being successful than Biography
    - Recently published books have not performed well

# NEXT STEPS

Try to Increase accuracy:

- Consolidate publisher to add to model
- Gather more data from other countries to expand User Country feature

# THANK YOU!
# QUESTIONS?