DECODING BOOK SUCCESS ON AMAZON

Presentation By
Guli Aminjonova, Jane Chiriyankandath,
Mary Nshikokola & Maria Velazquez

Problem Description

- **Project Objective:** Help new authors understand key factors contributing to book success.
- **Dataset:** Analyzing Amazon Bestselling Books dataset.
- **Key Focus Areas:** Popular genres, influential authors, pricing dynamics, and language trends in book titles.
- **Benefits for New Authors:** Valuable insights for writing, marketing, and publishing strategies to increase chances of success.

Data Description

- Name: Title of the book
- Author: Name of the author(s) of the book
- User Rating: Average rating given to the book by users on Amazon (on a scale from 0 to 5 stars)
- Reviews: Number of customer reviews the book has received on Amazon
- Price: Current price of the book on Amazon
- Year: Year in which the book was a bestseller on Amazon
- Genre: Broad category that the book belongs to (e.g., fiction and nonfiction)

```
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   Name         550 non-null    object
 1   Author       550 non-null    object
 2   User Rating  550 non-null    float64
 3   Reviews      550 non-null    int64
 4   Price        550 non-null    int64
 5   Year         550 non-null    int64
 6   Genre        550 non-null    object
```

Sneak Peek into the Data

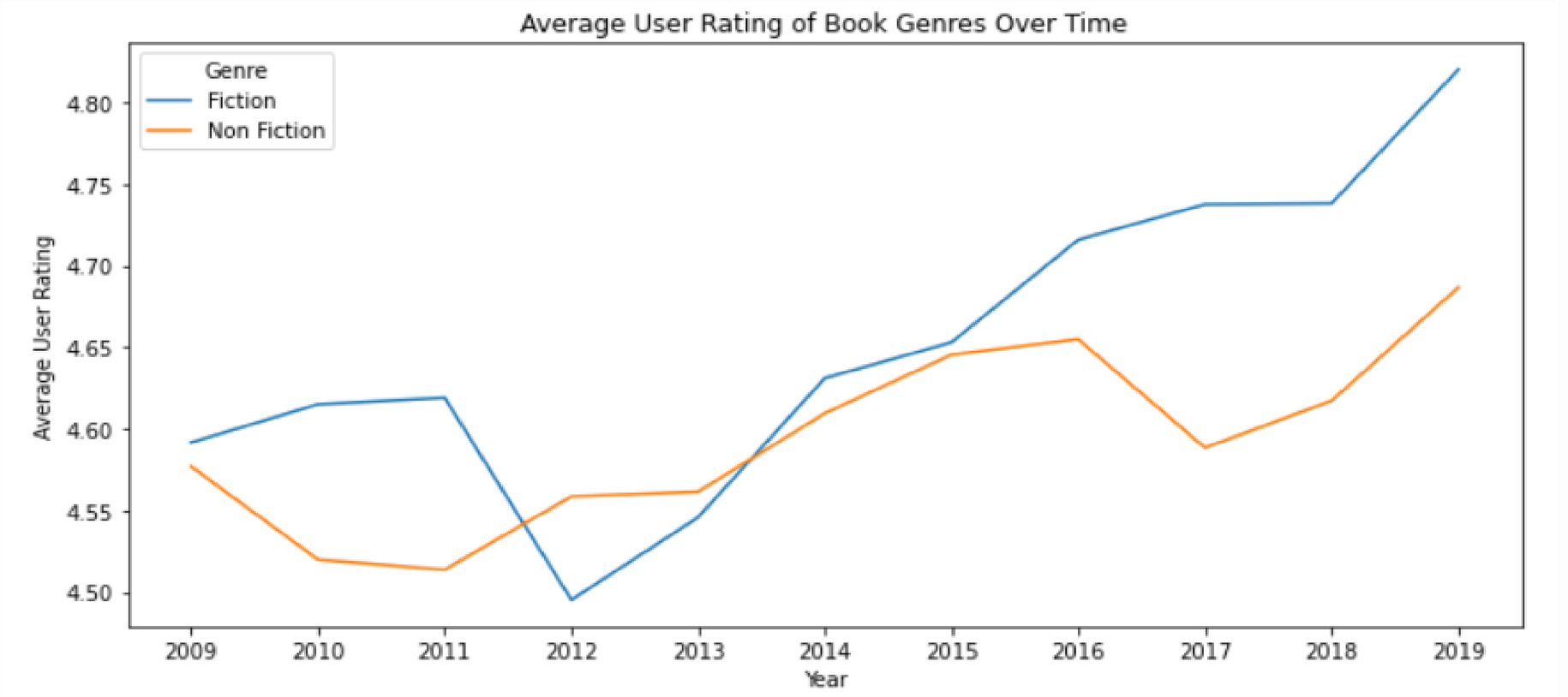
	Name	Author	User Rating	Reviews	Price	Year	Genre
0	10-Day Green Smoothie Cleanse	JJ Smith	4.7	17350	8	2016	Non Fiction
1	11/22/63: A Novel	Stephen King	4.6	2052	22	2011	Fiction
2	12 Rules for Life: An Antidote to Chaos	Jordan B. Peterson	4.7	18979	15	2018	Non Fiction
3	1984 (Signet Classics)	George Orwell	4.7	21424	6	2017	Fiction
4	5,000 Awesome Facts (About Everything!) (Natio...	National Geographic Kids	4.8	7665	12	2019	Non Fiction
5	A Dance with Dragons (A Song of Ice and Fire)	George R. R. Martin	4.4	12643	11	2011	Fiction
6	A Game of Thrones / A Clash of Kings / A Storm...	George R. R. Martin	4.7	19735	30	2014	Fiction
7	A Gentleman in Moscow: A Novel	Amor Towles	4.7	19699	15	2017	Fiction
8	A Higher Loyalty: Truth, Lies, and Leadership	James Comey	4.7	5983	3	2018	Non Fiction
9	A Man Called Ove: A Novel	Fredrik Backman	4.6	23848	8	2016	Fiction

EDA Findings

- This dataset contains data from the year 2009 to 2019.
- The dataset shows a larger number of non-fiction books than fiction books.
- Despite having 550 rows of data, we observe only 351 unique books, indicating that certain books have appeared multiple times in the dataset due to their recurring status as best-selling books.
- Some books appear twice as best-sellers in the same year, requiring further investigation for accuracy.
- User ratings range from 3.3 to 4.9, with most of the books receiving a rating of 4.8.

Research Question 1

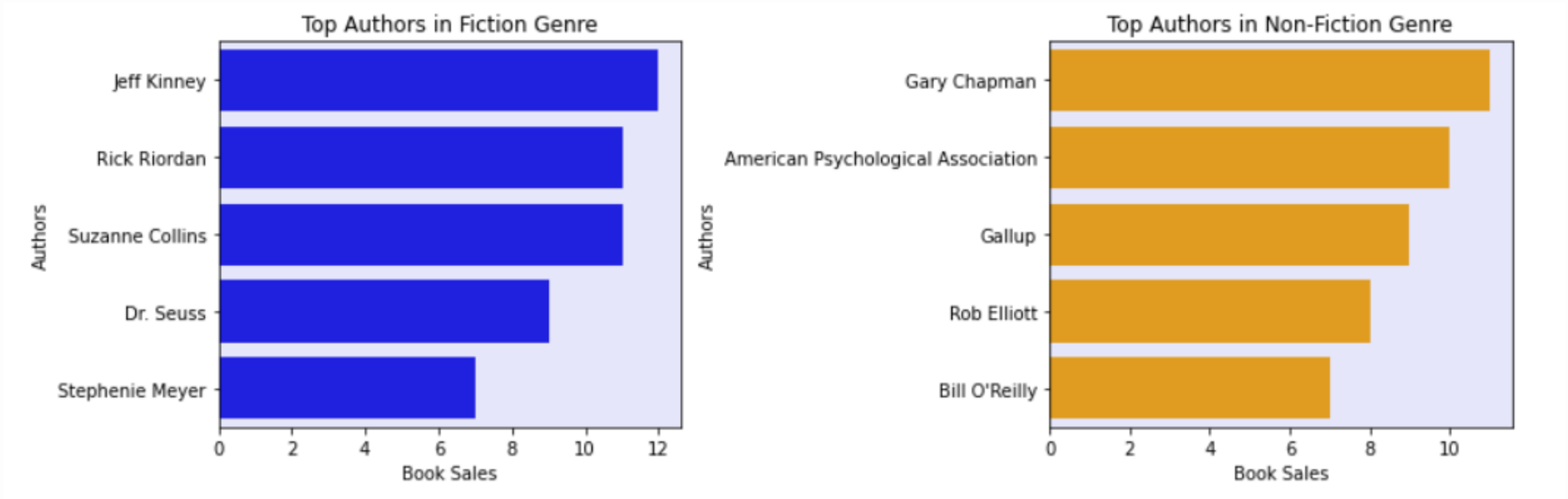
What is the most popular book genre, and how has the popularity of different book genres changed over time?



Answer: Fiction is the most popular book genre, and its popularity has remained consistently higher than Non-Fiction over time, except for 2012 and 2013.

Research Question 2

Are there certain authors that consistently appear in the bestselling books list?



Answer: The author with the most books in the fiction genre is Jeff Kinney, who has contributed different books or series that appear in the dataset. On the other hand, Gary Chapman is the top author in the non-fiction genre because the same book by him appears multiple times, indicating its continued popularity in that category.

Research Question 3

Which variables influence the book price?

Intercept Only Model (Empty Model):

Model:	MixedLM	Dependent Variable:	Price
No. Observations:	550	Method:	REML
No. Groups:	248	Scale:	34.6463
Min. group size:	1	Log-Likelihood:	-1949.9962
Max. group size:	12	Converged:	Yes
Mean group size:	2.2		

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
Intercept	12.932	0.629	20.573	0.000	11.700	14.164
Author Var	75.176	1.904				

```
ICC:  Author    0.684525
Name: Author, dtype: float64
```

Random Intercept Model:

Model:	MixedLM	Dependent Variable:	Price
No. Observations:	550	Method:	REML
No. Groups:	248	Scale:	33.6841
Min. group size:	1	Log-Likelihood:	-1948.5125
Max. group size:	12	Converged:	Yes
Mean group size:	2.2		

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
Intercept	784.084	274.601	2.855	0.004	245.877	1322.292
C(Genre)[T.Non Fiction]	2.145	1.313	1.634	0.102	-0.428	4.718
User_Rating	2.285	1.982	1.153	0.249	-1.600	6.170
Reviews	-0.000	0.000	-1.470	0.142	-0.000	0.000
Year	-0.388	0.137	-2.830	0.005	-0.657	-0.119
Author Var	73.711	1.930				

Research Question 3

Which variables influence the book price?

Random Slope Model:

Model:	MixedLM	Dependent Variable:	Price
No. Observations:	550	Method:	REML
No. Groups:	248	Scale:	33.8538
Min. group size:	1	Log-Likelihood:	-1949.3659
Max. group size:	12	Converged:	Yes
Mean group size:	2.2		

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
Intercept	791.602	274.781	2.881	0.004	253.041	1330.163
C(Genre)[T.Non Fiction]	2.261	1.333	1.697	0.090	-0.351	4.874
User_Rating	2.292	1.984	1.155	0.248	-1.596	6.181
Reviews	-0.000	0.000	-1.405	0.160	-0.000	0.000
Year	-0.392	0.137	-2.856	0.004	-0.661	-0.123
Genre Var	73.194	1.913				

Full Linear Mixed Model:

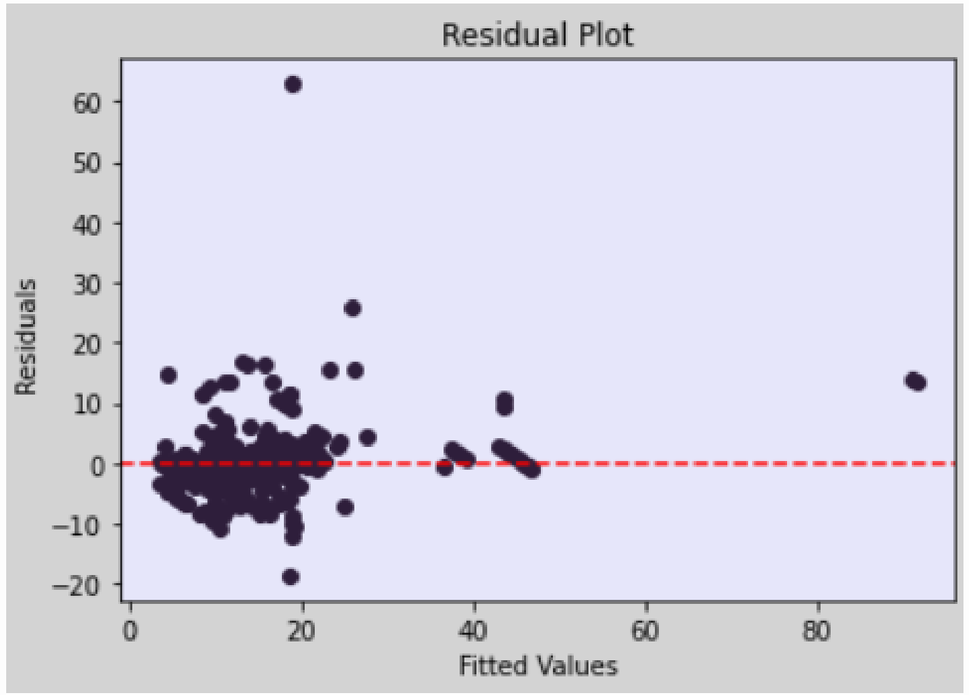
Model:	MixedLM	Dependent Variable:	Price
No. Observations:	550	Method:	REML
No. Groups:	248	Scale:	34.3773
Min. group size:	1	Log-Likelihood:	-1938.4143
Max. group size:	12	Converged:	Yes
Mean group size:	2.2		

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
Intercept	835.352	265.943	3.141	0.002	314.113	1356.591
C(Genre)[T.Non Fiction]	2.622	1.167	2.247	0.025	0.335	4.909
User_Rating	1.382	1.886	0.733	0.464	-2.315	5.079
Reviews	-0.000	0.000	-0.643	0.520	-0.000	0.000
Year	-0.412	0.133	-3.099	0.002	-0.673	-0.151
Author Var	27.290	1.404				
Author x C(Genre)[T.Non Fiction] Cov	24.176					
C(Genre)[T.Non Fiction] Var	21.417					

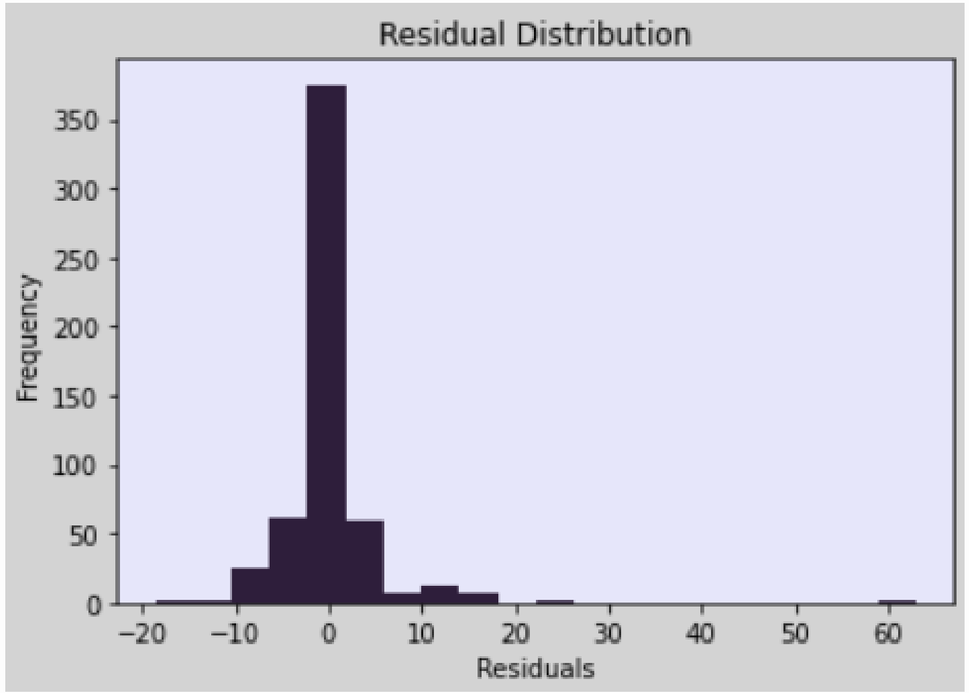
Research Question 3

Which variables influence the book price?

Residual Plot:



Residual Distribution:



Research Question 3

Which variables influence the book price?

Answer:

1. There is a significant relationship between Genre and Price. For books categorized as "Non Fiction," the model estimates their average Price to be 2.622 units higher than those categorized as "Fiction

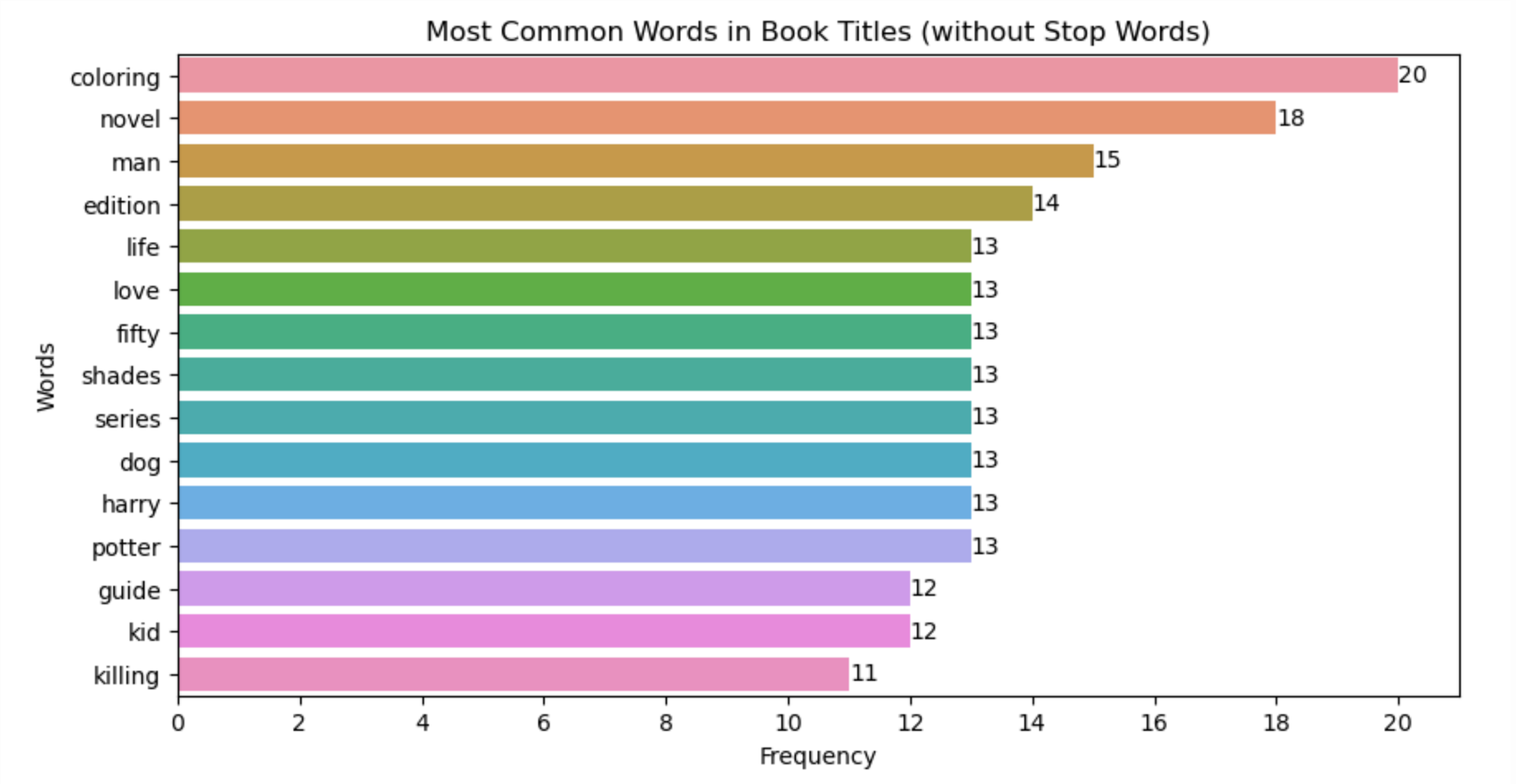
2. The variable Year also has a statistically significant relationship with Price. A one-unit increase in Year is associated with a decrease of 0.412 units in Price.

3. The estimated variance of the random effect for the Author is 27.290. This captures the variation in Price that is attributed to different authors and not accounted for by the fixed effects.

4. There is a strong positive correlation between the random effect and fixed effect on the model. This suggests that certain authors tend to consistently write books belonging to the “Non-Fiction" genre, while others tend to write books belonging to the "Fiction" genre.

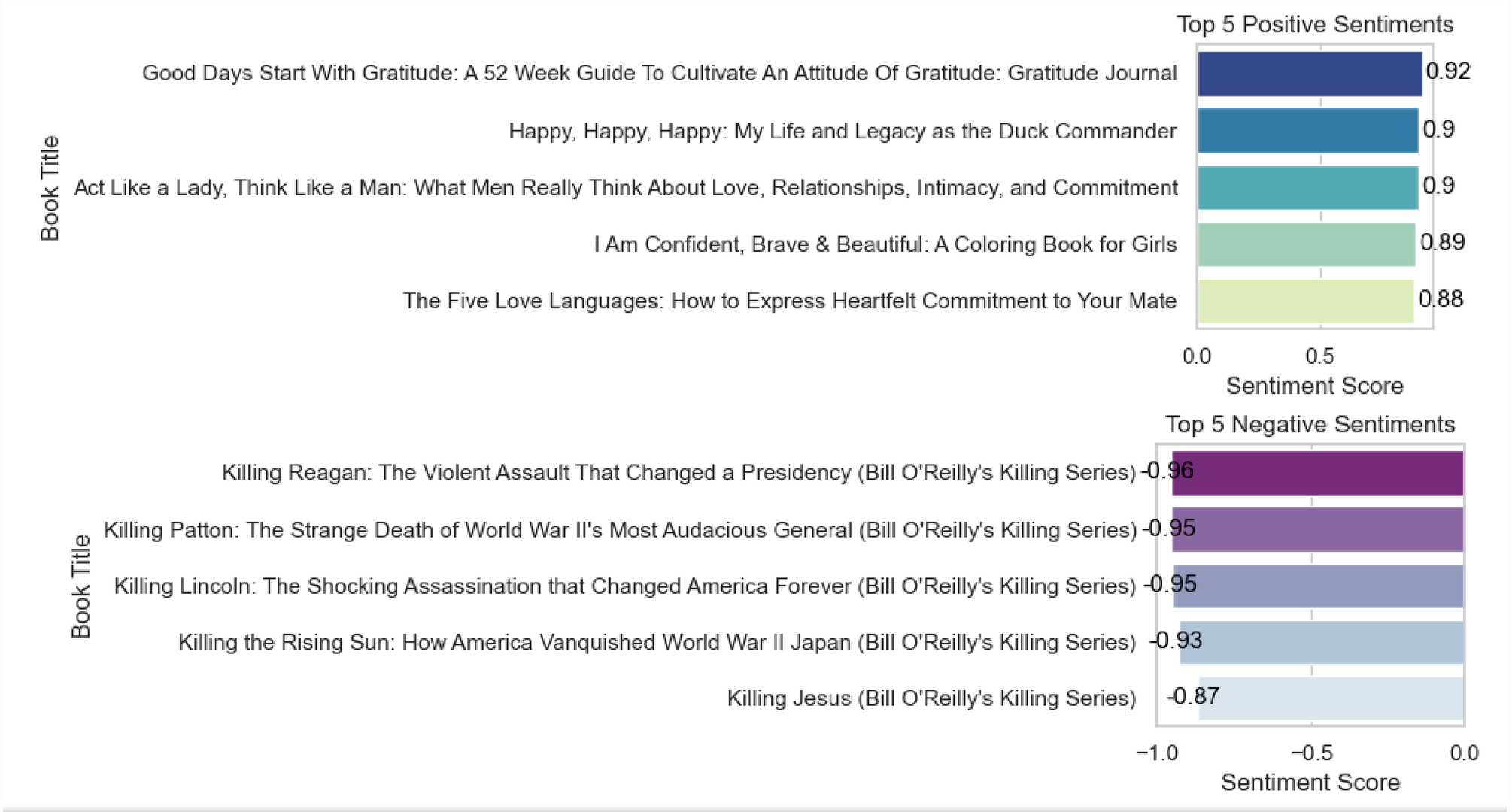
Research Question 4

Part 1: What are the most frequently occurring words in book titles?



Research Question 4

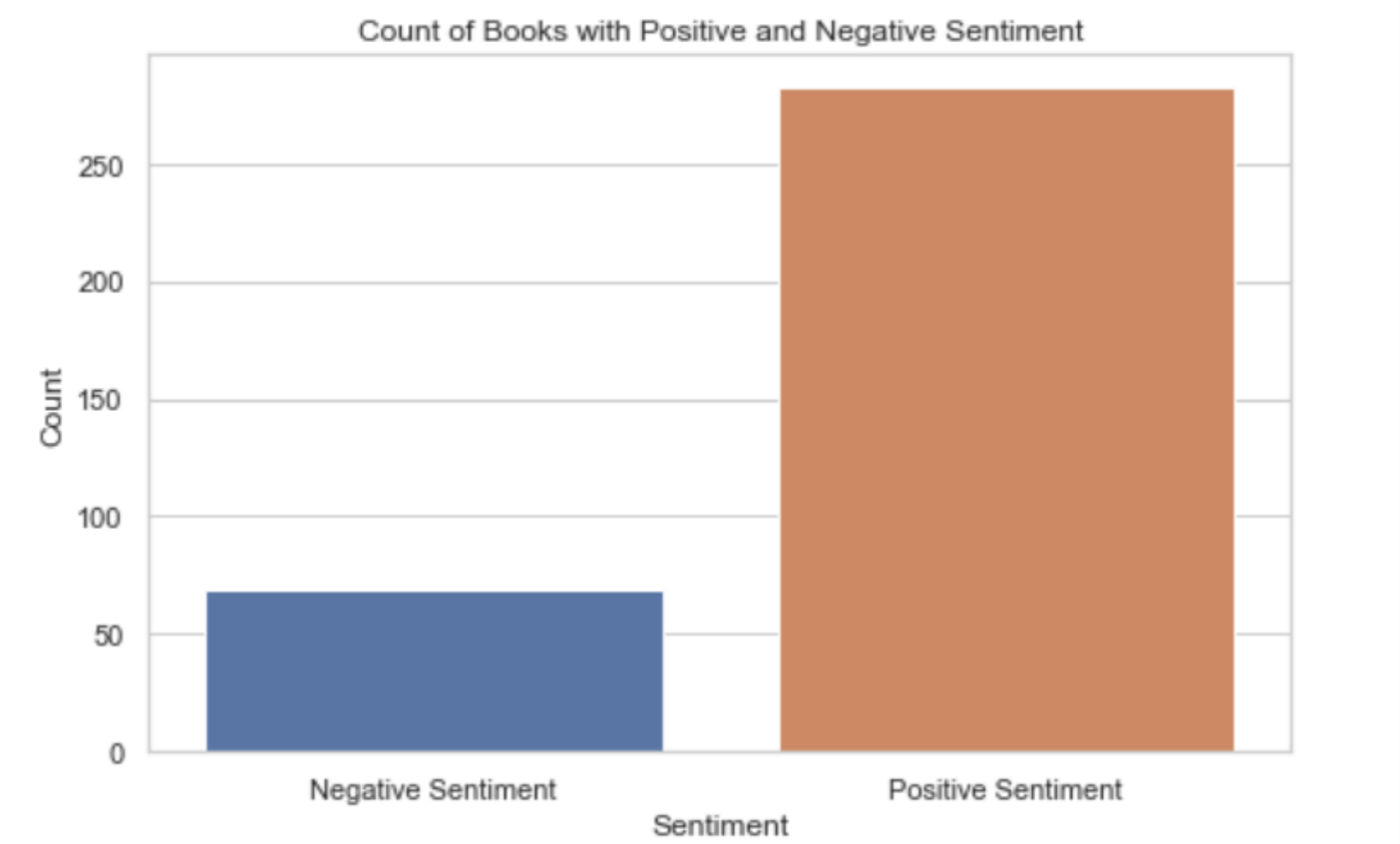
Part 2: What insights do we gain from sentiment analysis?



Research Question 4

Part 2: What insights do we gain from sentiment analysis?

Answer: Out of all the analyzed books, 282 received positive compound sentiment scores ranging from 0 to 1, while 69 received negative compound sentiment scores. This indicates that most of the bestselling books during that period focused on positive emotions and themes.



Recommendations

- Fiction books tend to receive higher user ratings. Authors should understand their target audience and connect with them and their preferences.

- Considering that the year of publication has a significant relationship with pricing, authors should stay informed about current market trends and adjust their pricing strategy accordingly.

- Authors should take note of the dominant themes and emotions that emerged from the analysis and incorporate these themes into their storytelling to align with readers' preferences.

- Besides these recommendations from the analysis, we suggest authors stay passionate, authentic, and dedicated and continually seek opportunities to grow and connect with their readership.

Conclusion

- It is very important to understand the popularity dynamics.
- Fiction is the most popular genre based on average user ratings.
- The relationship between pricing and genre should be considered: Non-fiction books tend to have higher prices than fiction books.
- The year of publication also has a significant negative relationship with price, indicating that older books tend to be priced lower.
- The majority of analyzed books received positive sentiment scores, indicating that the top-selling books focused on positive emotions and themes.