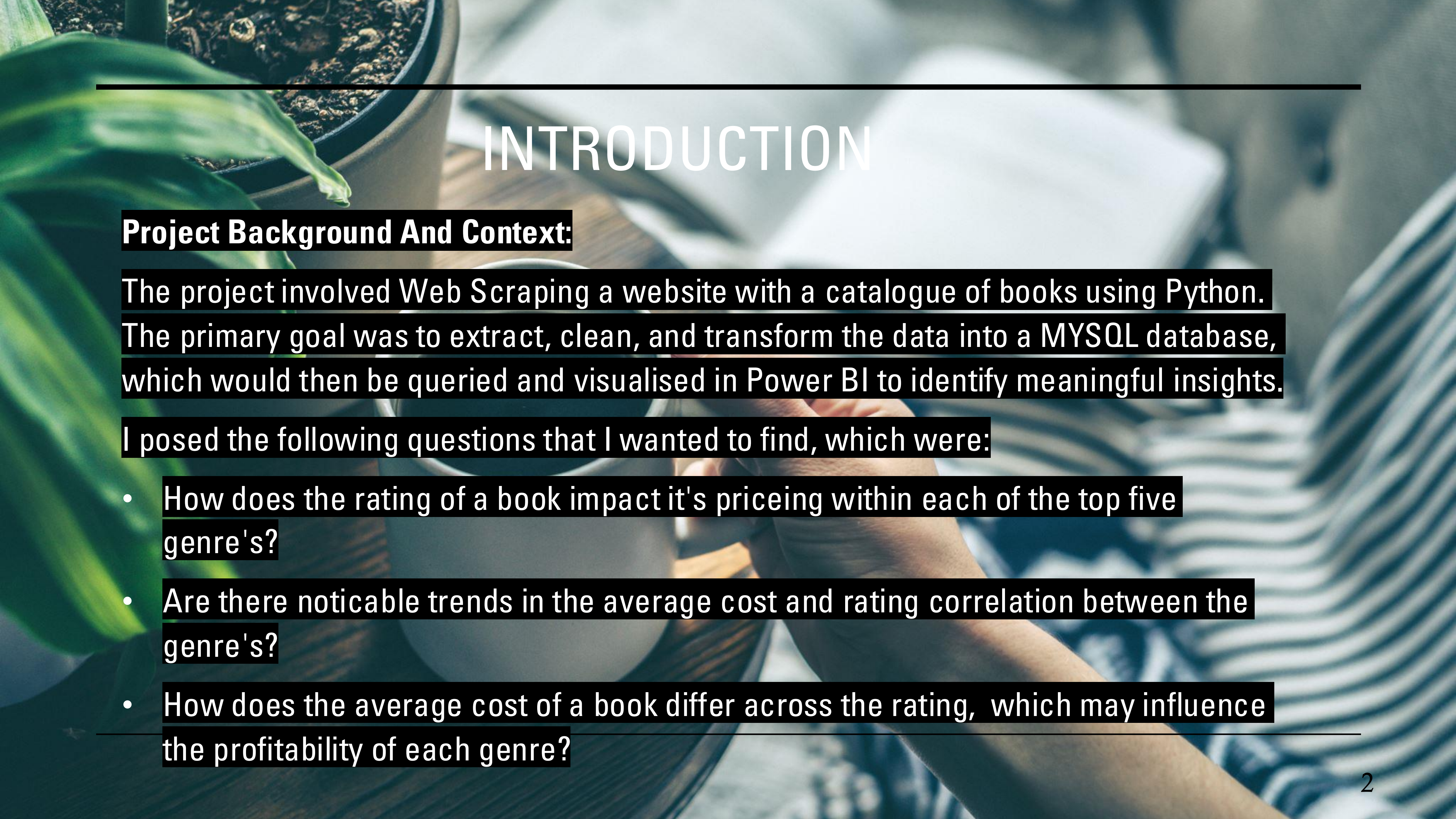
WEBSITE SCRAPING PROJECT

BEST SELLING BOOKS WITH DATA ANALYTICS



INTRODUCTION

Project Background And Context:

The project involved Web Scraping a website with a catalogue of books using Python. The primary goal was to extract, clean, and transform the data into a MYSQL database, which would then be queried and visualised in Power BI to identify meaningful insights.

I posed the following questions that I wanted to find, which were:

- How does the rating of a book impact it's priceing within each of the top five genre's?
- Are there noticable trends in the average cost and rating correlation between the genre's?
- How does the average cost of a book differ across the rating, which may influence the profitability of each genre?

SUMMARY

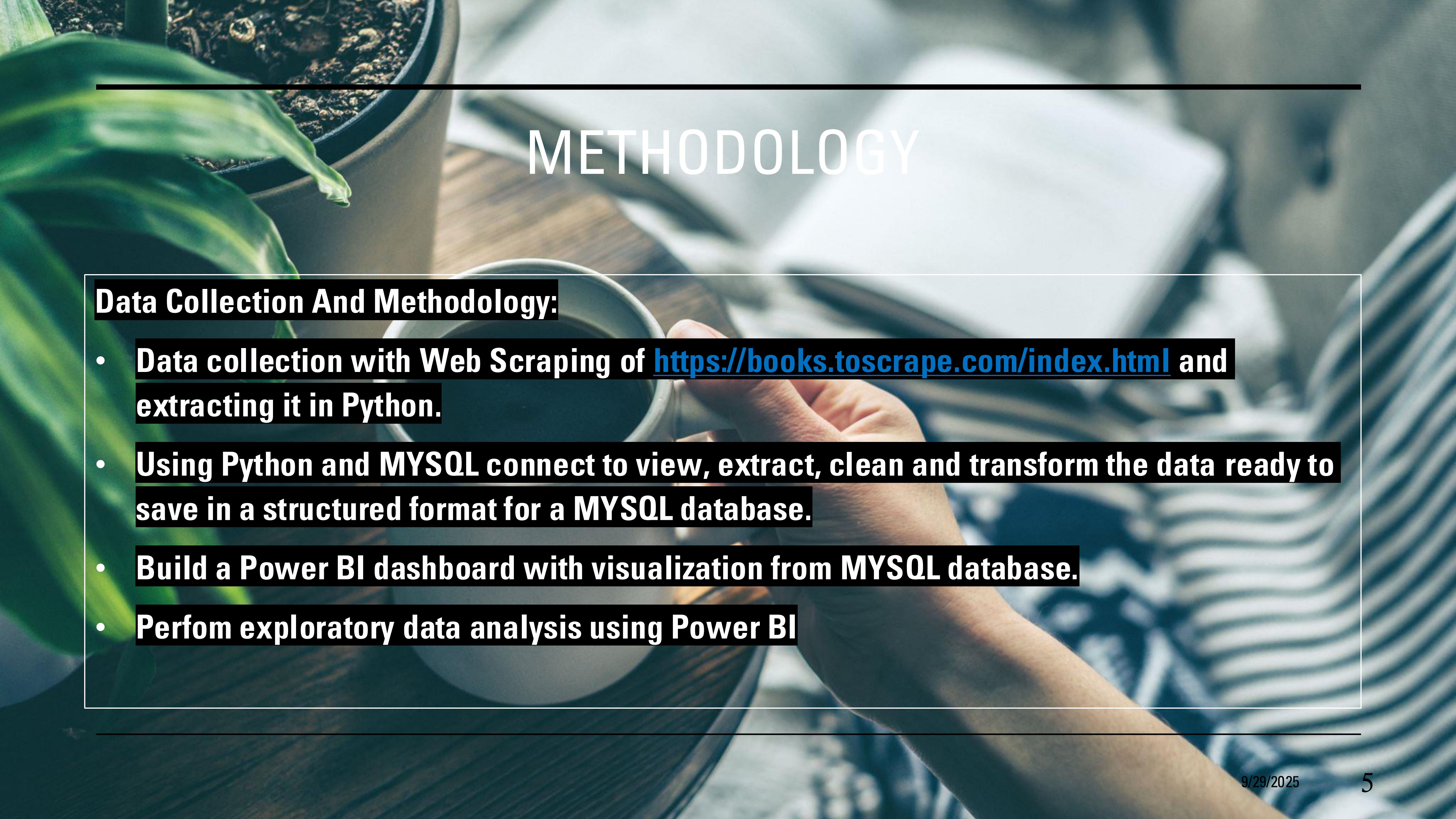
Summary Of Metholodgy

- Data Collection With Web Scraping
- Data Wrangling
- Data Preperation And Querying With MYSQL
- Exploratory Data Analysis with Power BI

Summary Of Results

- Summary Of Results
- Power BI Metrics
- Analysis Approach
- Conclusion

METHODOLOGY



METHODOLOGY

Data Collection And Methodology:

- **Data collection with Web Scraping of https://books.toscrape.com/index.html and extracting it in Python.**
- **Using Python and MYSQL connect to view, extract, clean and transform the data ready to save in a structured format for a MYSQL database.**
- **Build a Power BI dashboard with visualization from MYSQL database.**
- **Perfom exploratory data analysis using Power BI**

For detailed code and data analysis, view this Jupyter Lite notebook here:

https://github.com/Drook93/Webscrape-Books/blob/master/WebScrape%20Python%20Code.ipynb

MYSQL DATABASE

- When setting up the table for my "bookstoscrape" database, I created the column ID using a primary Key data type and AI for auto-increment for the values under it.
- I set the NN data type across all values to prevent any null values to be accepted.
- The " Title" had UQ unique identifier to ensure in the case of duplicate title's, which would exclude them.
- The UN for unassigned was applied to the "Rating" and "Cost" to prevent any negative values within the database. This wasn't applied to the "Number_Available" for negative values in case the data was to potentially show pre-order backlog.

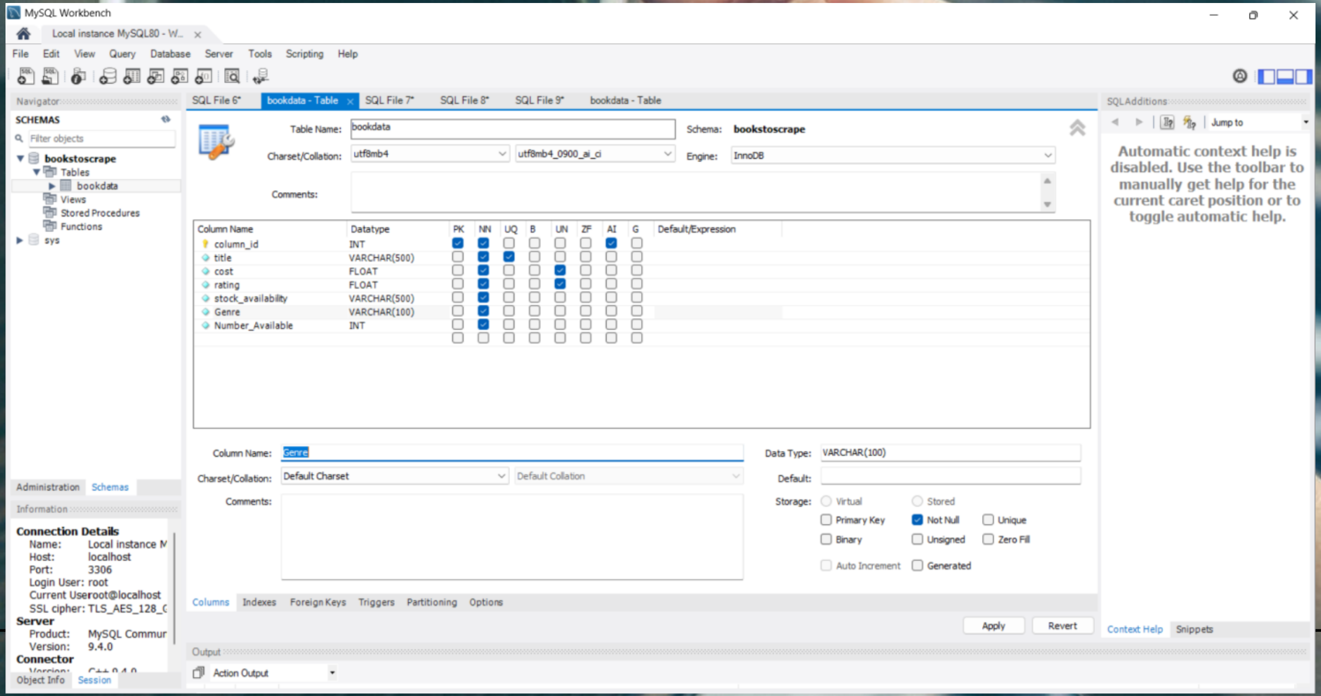


Table Name: bookdata

Schema: **bookstoscrape**

Charset/Collation: utf8mb4 utf8mb4_0900_ai_ci

Engine: InnoDB

Comments:

Column Name	Datatype	PK	NN	UQ	B	UN	ZF	AI	G	Default/Expression
column_id	INT	☑	☑	□	□	□	□	☑	□	
title	VARCHAR(500)	□	☑	☑	□	□	□	□	□	
cost	FLOAT	□	☑	□	□	☑	□	□	□	
rating	FLOAT	□	☑	□	□	☑	□	□	□	
stock_availability	VARCHAR(500)	□	☑	□	□	□	□	□	□	
Genre	VARCHAR(100)	□	☑	□	□	□	□	□	□	
Number_Available	INT	□	☑	□	□	□	□	□	□	
		□	□	□	□	□	□	□	□	

Column Name: Genre

Data Type: VARCHAR(100)

Charset/Collation: Default Charset Default Collation

Default:

Comments:

Storage: ○ Virtual ○ Stored

□ Primary Key ☑ Not Null □ Unique

□ Binary □ Unsigned □ Zero Fill

□ Auto Increment □ Generated

Automatic context help is disabled. Use the toolbar to manually get help for the current caret position or to toggle automatic help.

Connection Details

Name: Local instance M

Host: localhost

Port: 3306

Login User: root

Current Useroot@localhost

SSL cipher: TLS_AES_128_C

Server

Product: MySQL Commur

Version: 9.4.0

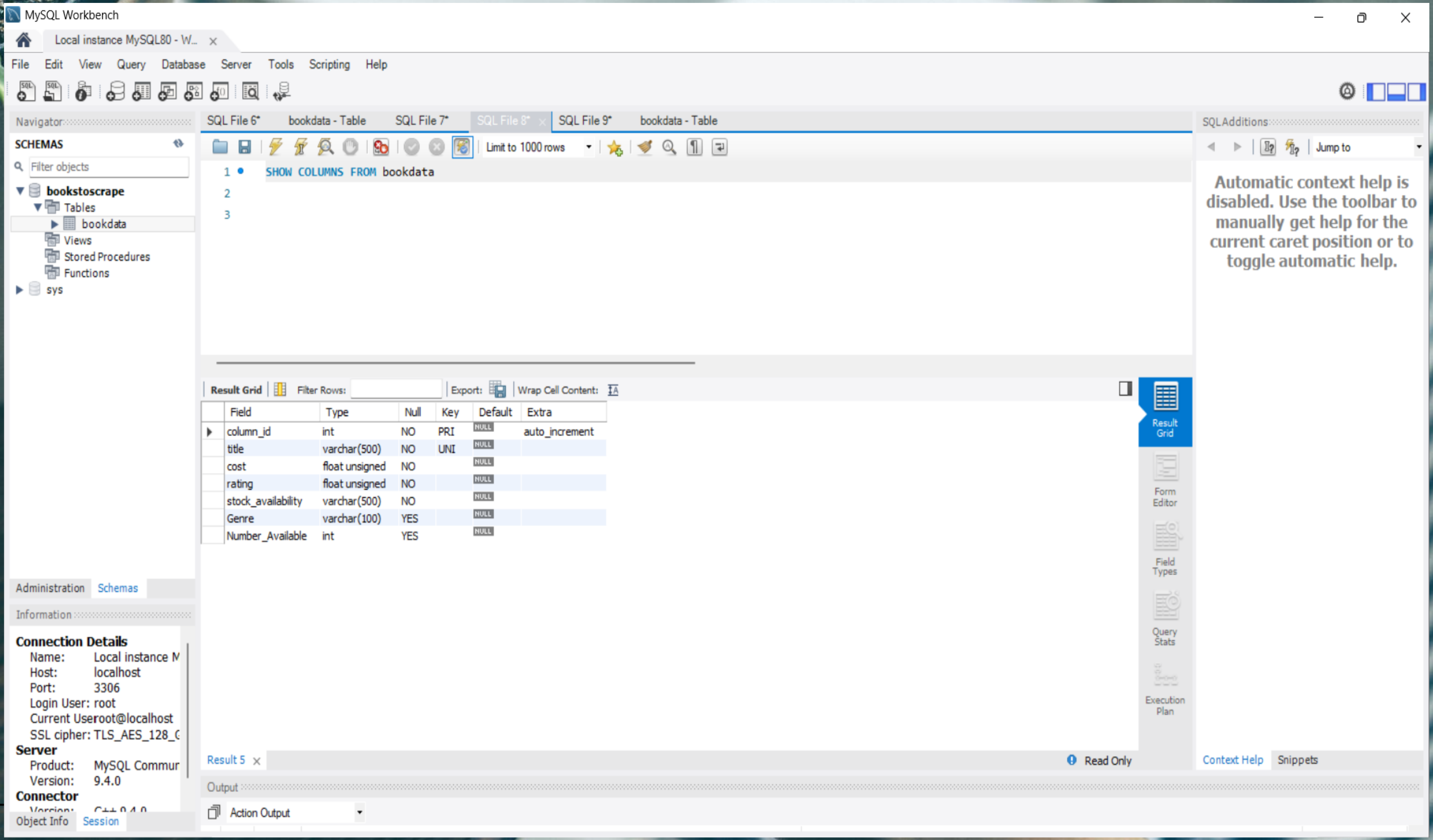
Connector

Version: C++ 9.4.0

- I ran a query in SQL to see the collumn's keys and their datatype's.

I initially faced issues in Python as they weren't assigned correctly.

- I used the SHOW COLUMNS statement to view this.



```
SHOW COLUMNS FROM bookdata
```

Field	Type	Null	Key	Default	Extra
column_id	int	NO	PRI	NULL	auto_increment
title	varchar(500)	NO	UNI	NULL	
cost	float unsigned	NO		NULL	
rating	float unsigned	NO		NULL	
stock_availability	varchar(500)	NO		NULL	
Genre	varchar(100)	YES		NULL	
Number_Available	int	YES		NULL	

Automatic context help is disabled. Use the toolbar to manually get help for the current caret position or to toggle automatic help.

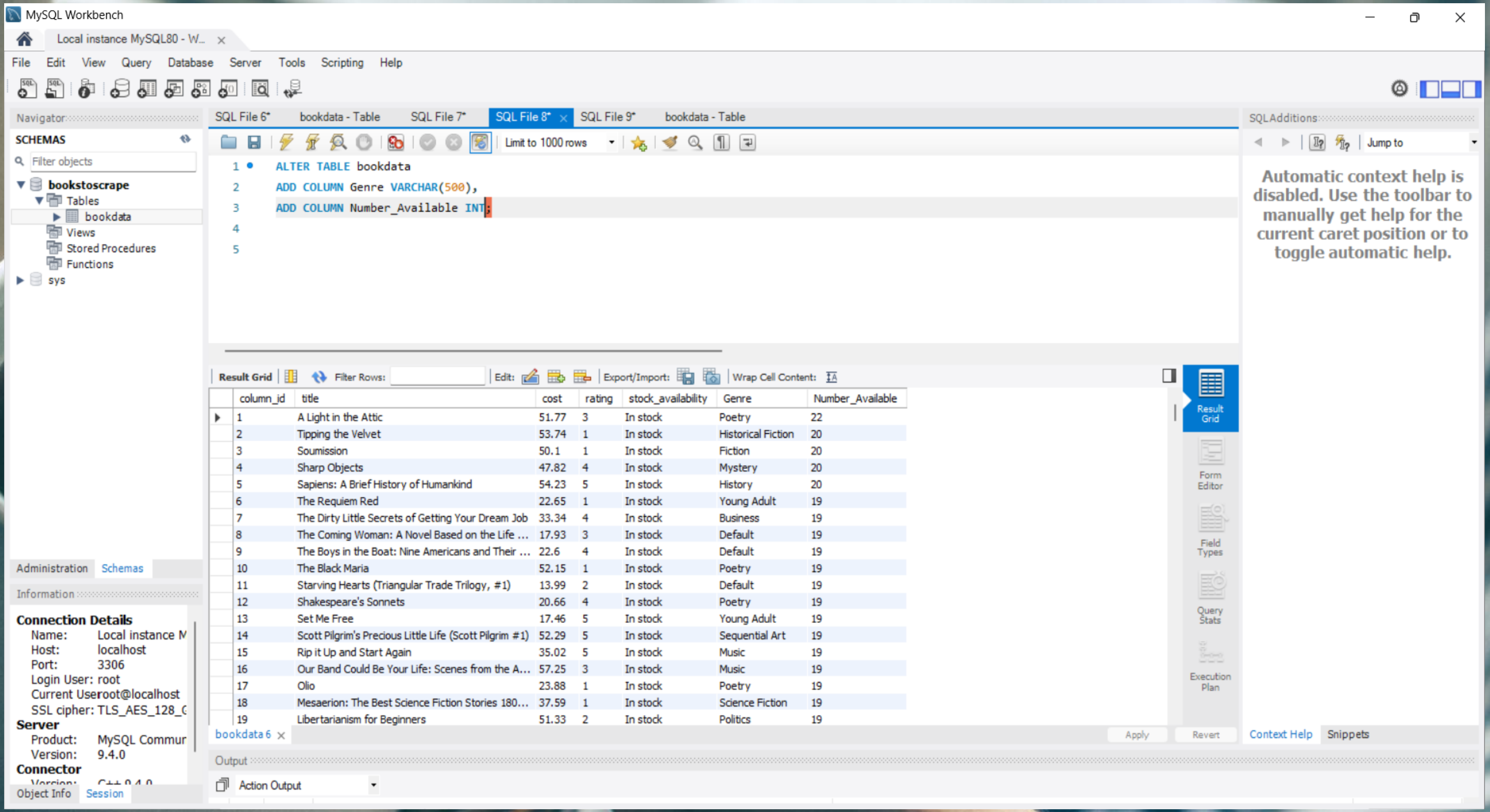
Connection Details
Name: Local instance M
Host: localhost
Port: 3306
Login User: root
Current User: root@localhost
SSL cipher: TLS_AES_128_(

Server
Product: MySQL Commur
Version: 9.4.0

Connector
Version: C++ 9.4.0

- Initially, after creating the table and extracting the data, I realised that additional variables were necessary to complete my analysis.

I created additional columns in MYSQL using the following statement: ALTER TABLE followed by ADD COLUMN with the appropriate data types.



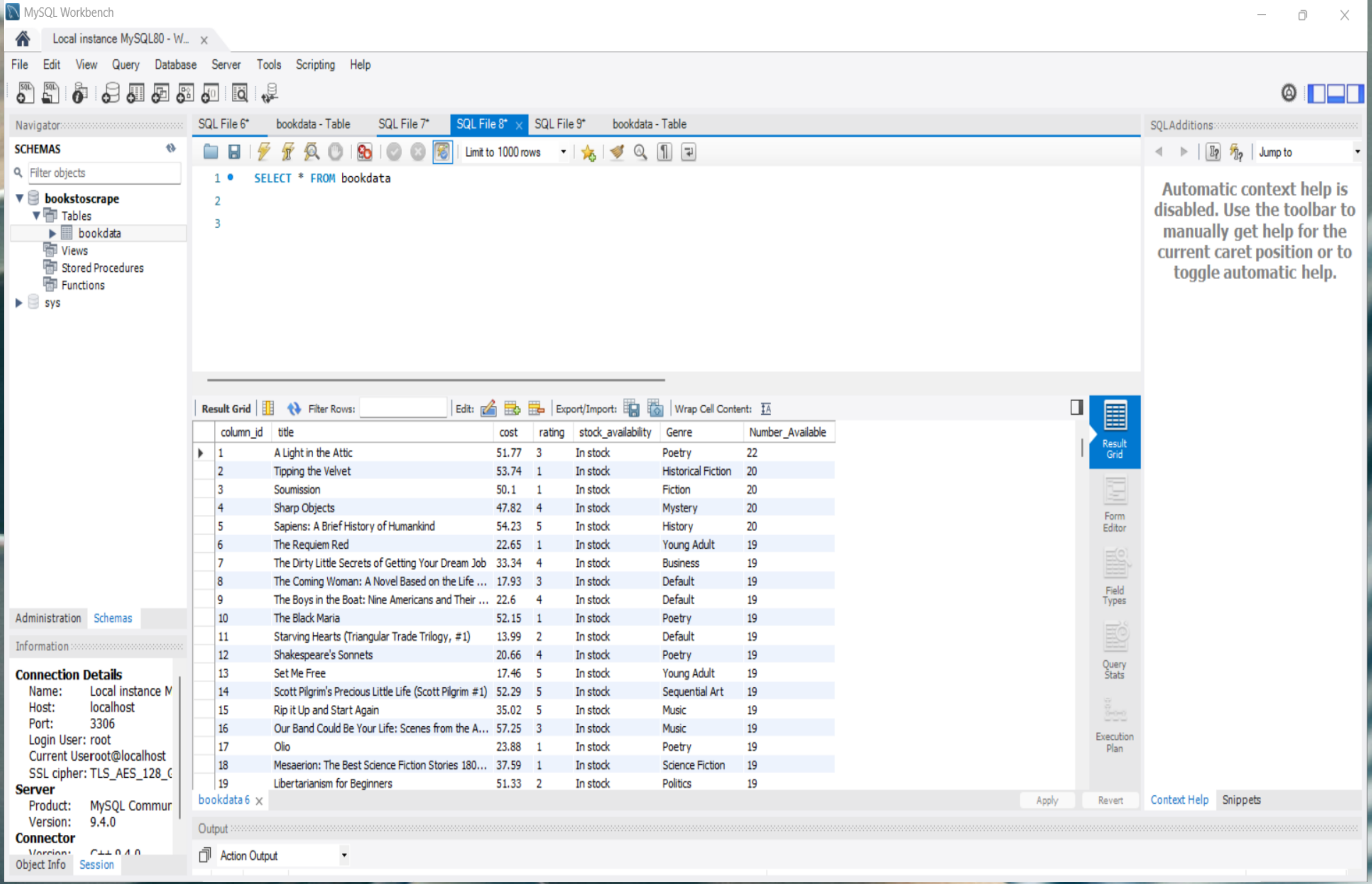
```
ALTER TABLE bookdata
ADD COLUMN Genre VARCHAR(500),
ADD COLUMN Number_Available INT;
```

Automatic context help is disabled. Use the toolbar to manually get help for the current caret position or to toggle automatic help.

column_id	title	cost	rating	stock_availability	Genre	Number_Available
1	A Light in the Attic	51.77	3	In stock	Poetry	22
2	Tipping the Velvet	53.74	1	In stock	Historical Fiction	20
3	Soumission	50.1	1	In stock	Fiction	20
4	Sharp Objects	47.82	4	In stock	Mystery	20
5	Sapiens: A Brief History of Humankind	54.23	5	In stock	History	20
6	The Requiem Red	22.65	1	In stock	Young Adult	19
7	The Dirty Little Secrets of Getting Your Dream Job	33.34	4	In stock	Business	19
8	The Coming Woman: A Novel Based on the Life ...	17.93	3	In stock	Default	19
9	The Boys in the Boat: Nine Americans and Their ...	22.6	4	In stock	Default	19
10	The Black Maria	52.15	1	In stock	Poetry	19
11	Starving Hearts (Triangular Trade Trilogy, #1)	13.99	2	In stock	Default	19
12	Shakespeare's Sonnets	20.66	4	In stock	Poetry	19
13	Set Me Free	17.46	5	In stock	Young Adult	19
14	Scott Pilgrim's Precious Little Life (Scott Pilgrim #1)	52.29	5	In stock	Sequential Art	19
15	Rip it Up and Start Again	35.02	5	In stock	Music	19
16	Our Band Could Be Your Life: Scenes from the A...	57.25	3	In stock	Music	19
17	Olio	23.88	1	In stock	Poetry	19
18	Mesaerion: The Best Science Fiction Stories 180...	37.59	1	In stock	Science Fiction	19
19	Libertarianism for Beginners	51.33	2	In stock	Politics	19

Connection Details
Name: Local instance M
Host: localhost
Port: 3306
Login User: root
Current User: root@localhost
SSL cipher: TLS_AES_128_C
Server
Product: MySQL Commur
Version: 9.4.0
Connector

- After saving to the database. A SELECT * FROM statement was queried to view the table to ensure all of the data was pulled through.



```
SELECT * FROM bookdata
```

Automatic context help is disabled. Use the toolbar to manually get help for the current caret position or to toggle automatic help.

column_id	title	cost	rating	stock_availability	Genre	Number_Available
1	A Light in the Attic	51.77	3	In stock	Poetry	22
2	Tipping the Velvet	53.74	1	In stock	Historical Fiction	20
3	Soumission	50.1	1	In stock	Fiction	20
4	Sharp Objects	47.82	4	In stock	Mystery	20
5	Sapiens: A Brief History of Humankind	54.23	5	In stock	History	20
6	The Requiem Red	22.65	1	In stock	Young Adult	19
7	The Dirty Little Secrets of Getting Your Dream Job	33.34	4	In stock	Business	19
8	The Coming Woman: A Novel Based on the Life ...	17.93	3	In stock	Default	19
9	The Boys in the Boat: Nine Americans and Their ...	22.6	4	In stock	Default	19
10	The Black Maria	52.15	1	In stock	Poetry	19
11	Starving Hearts (Triangular Trade Trilogy, #1)	13.99	2	In stock	Default	19
12	Shakespeare's Sonnets	20.66	4	In stock	Poetry	19
13	Set Me Free	17.46	5	In stock	Young Adult	19
14	Scott Pilgrim's Precious Little Life (Scott Pilgrim #1)	52.29	5	In stock	Sequential Art	19
15	Rip it Up and Start Again	35.02	5	In stock	Music	19
16	Our Band Could Be Your Life: Scenes from the A...	57.25	3	In stock	Music	19
17	Olio	23.88	1	In stock	Poetry	19
18	Mesaerion: The Best Science Fiction Stories 180...	37.59	1	In stock	Science Fiction	19
19	Libertarianism for Beginners	51.33	2	In stock	Politics	19

Connection Details

Name: Local instance M
Host: localhost
Port: 3306
Login User: root
Current Useroot@localhost
SSL cipher: TLS_AES_128_(

Server

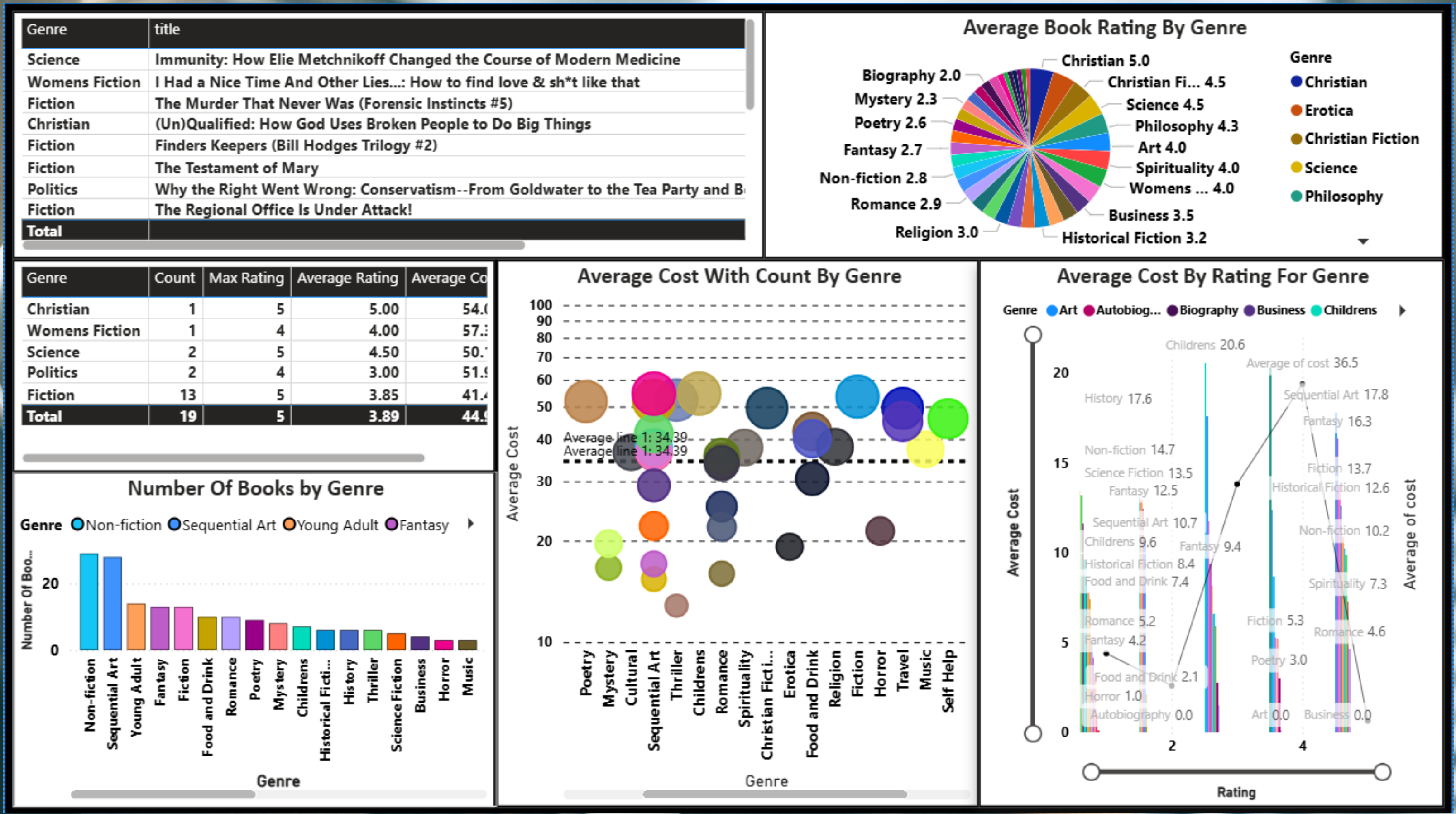
Product: MySQL Commur
Version: 9.4.0

Connector

Version: C++ 9.4.0

POWER BI

- This is the overview of the Power BI dashboard I built.



Genre	title
Science	Immunity: How Elie Metchnikoff Changed the Course of Modern Medicine
Womens Fiction	I Had a Nice Time And Other Lies...: How to find love & sh*t like that
Fiction	The Murder That Never Was (Forensic Instincts #5)
Christian	(Un)Qualified: How God Uses Broken People to Do Big Things
Fiction	Finders Keepers (Bill Hodges Trilogy #2)
Fiction	The Testament of Mary
Politics	Why the Right Went Wrong: Conservatism--From Goldwater to the Tea Party and B
Fiction	The Regional Office Is Under Attack!
Total	

Genre	Count	Max Rating	Average Rating	Average Co
Christian	1	5	5.00	54.(
Womens Fiction	1	4	4.00	57.:
Science	2	5	4.50	50.'
Politics	2	4	3.00	51.!
Fiction	13	5	3.85	41.4
Total	19	5	3.89	44.!

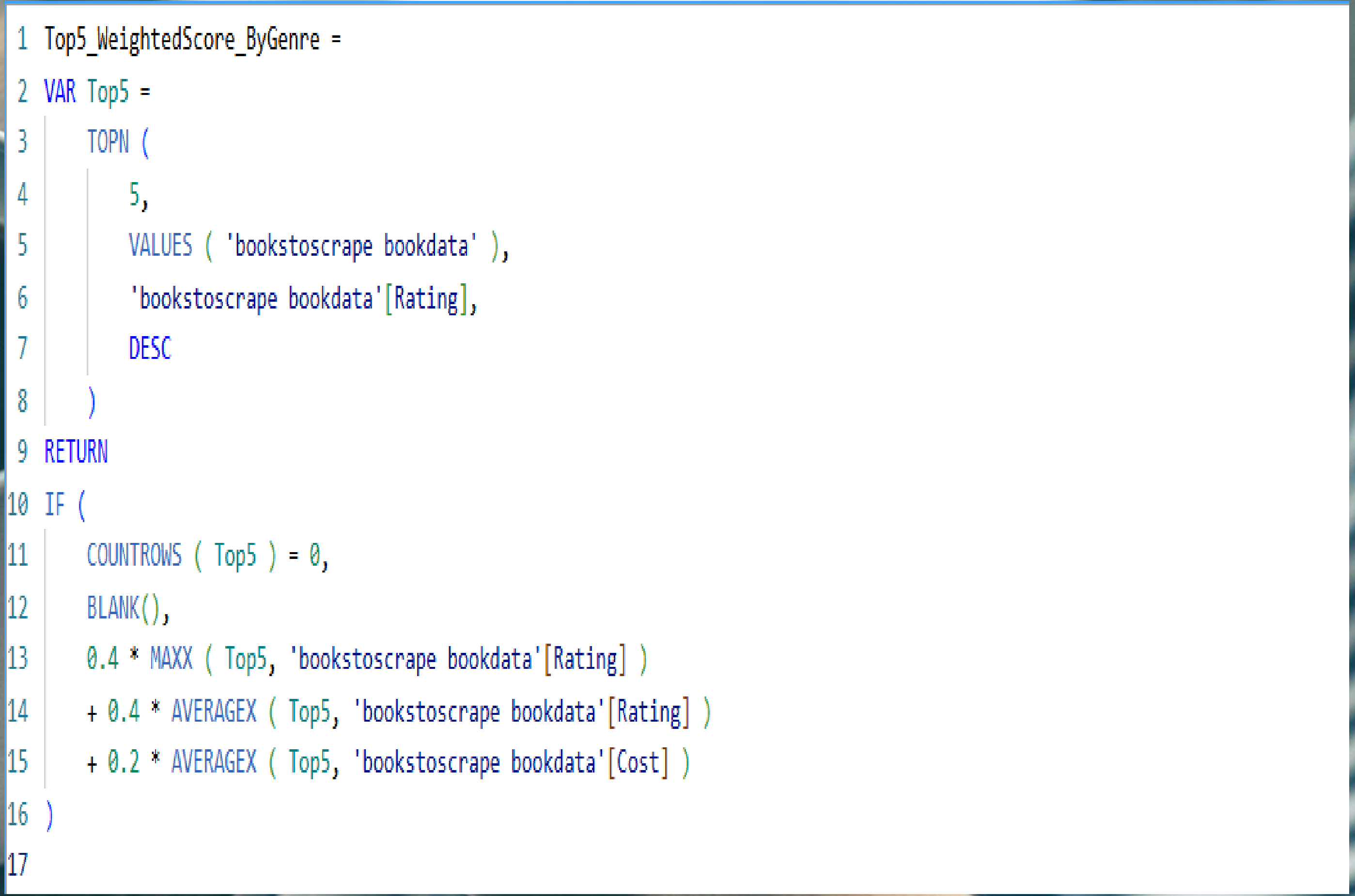
- Using DAX, a measurement was created for the Top 5 Genres by calculating a weighted average score.

- The max rating was a 0.4 weight.

The average rating was 0.4 weight.

- The cost 0.2 weight.

- These weights were combined to generate a Weighted Average Score.

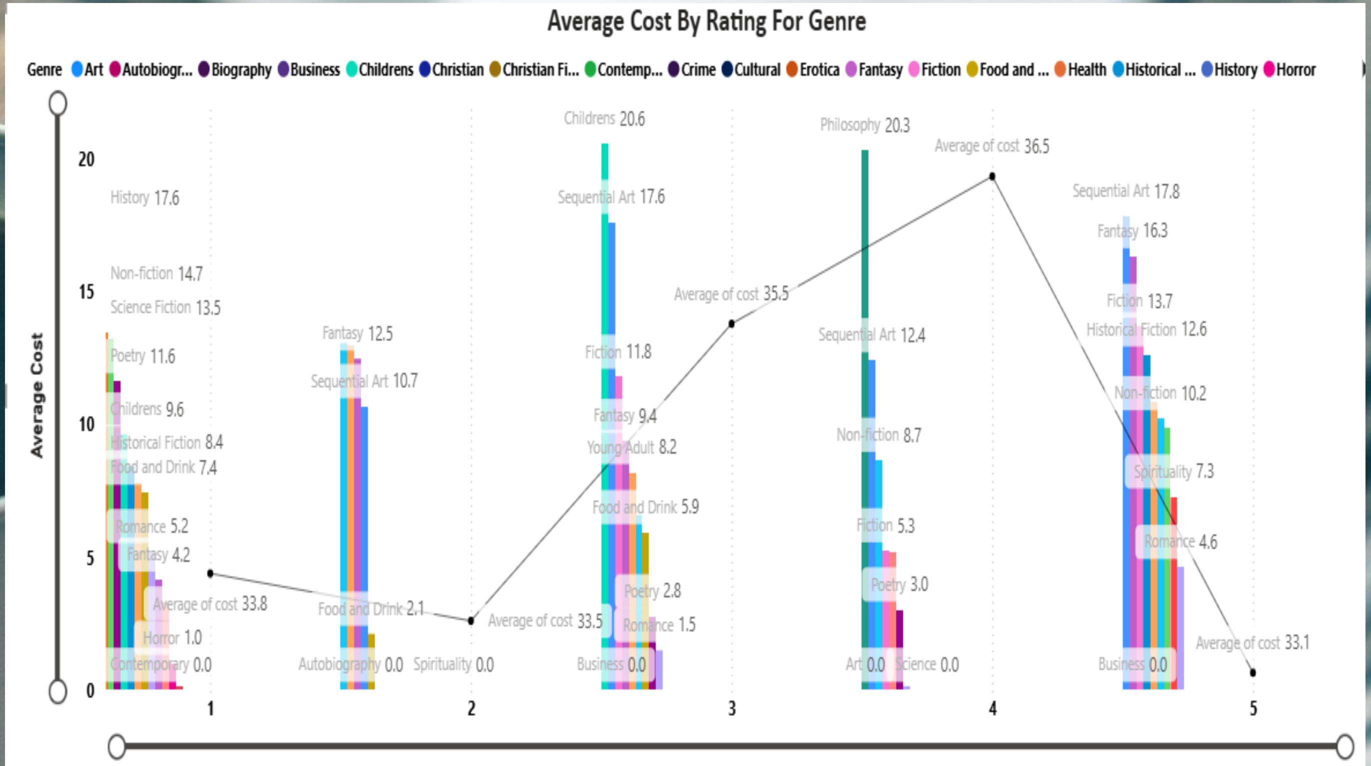


```
Top5_WeightedScore_ByGenre =
VAR Top5 =
    TOPN (
        5,
        VALUES ( 'bookstoscrape bookdata' ),
        'bookstoscrape bookdata'[Rating],
        DESC
    )
RETURN
IF (
    COUNTROWS ( Top5 ) = 0,
    BLANK(),
    0.4 * MAXX ( Top5, 'bookstoscrape bookdata'[Rating] )
    + 0.4 * AVERAGEX ( Top5, 'bookstoscrape bookdata'[Rating] )
    + 0.2 * AVERAGEX ( Top5, 'bookstoscrape bookdata'[Cost] )
)
```

- This Cluster Chart shows the average cost by rating

- Outliers would skew the average cost, so a 1 standard deviation was applied to smooth out any anomalies.

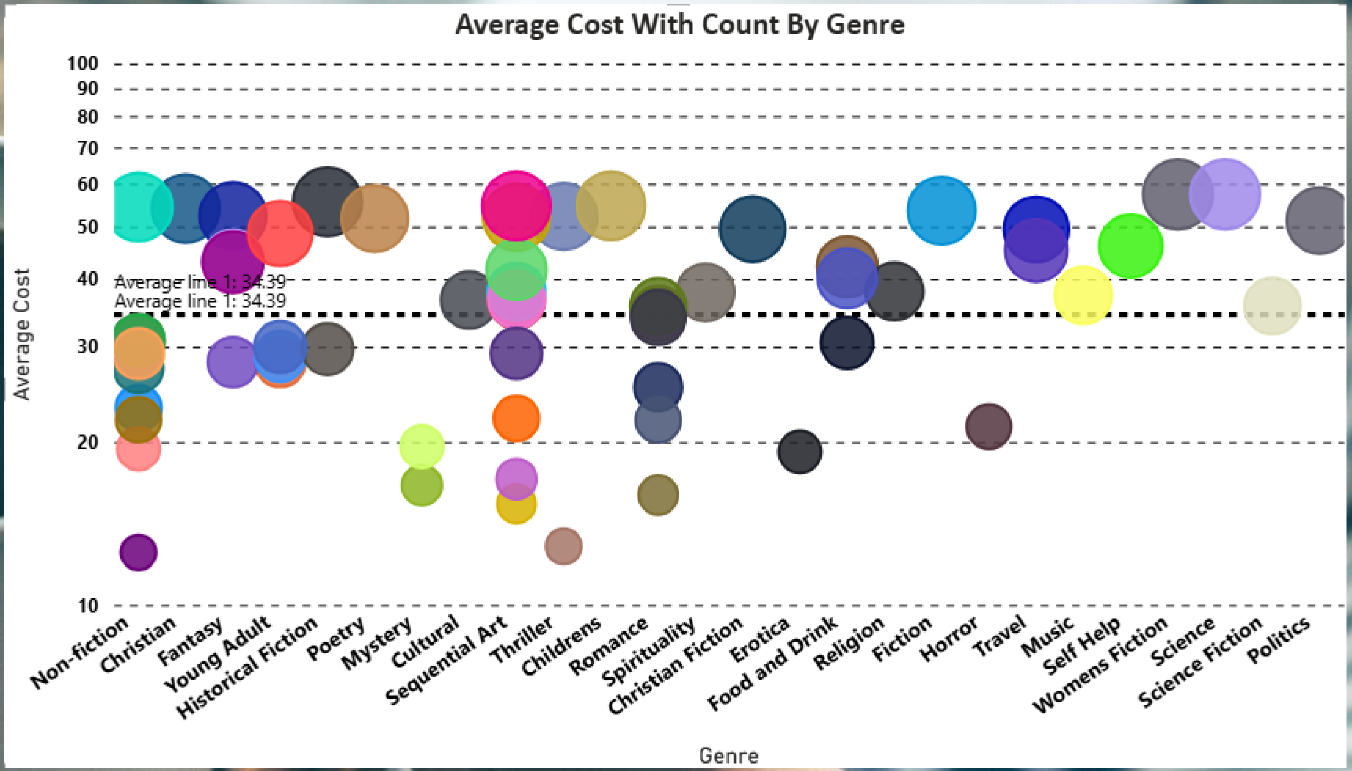
- The line across the chart also shows the average across each rating.



- This Scattered Bubble Chart shows the "Genre" by average cost with count.

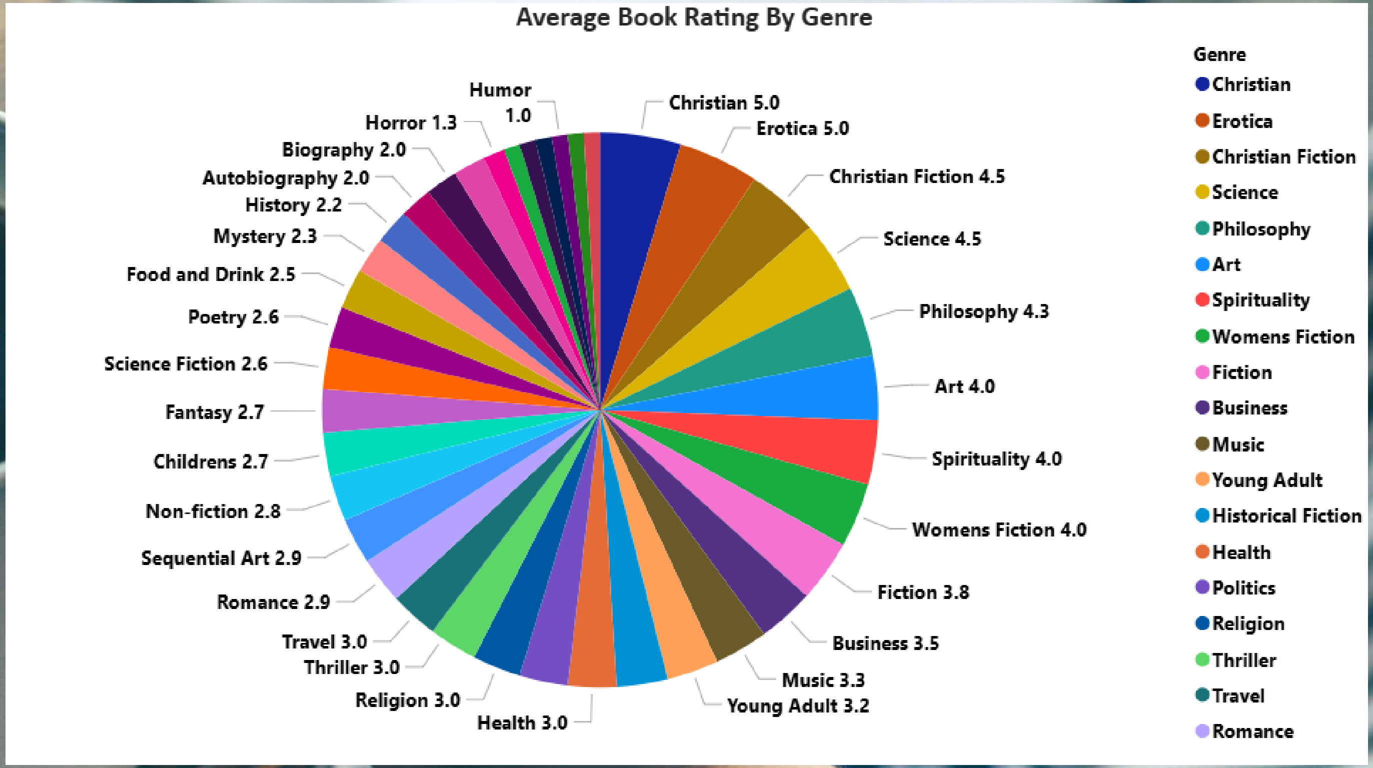
- The size of bubbles would represent the count of books across the averaged cost.

- This represents the variation of the average cost across the genre's and where the majority would lie.



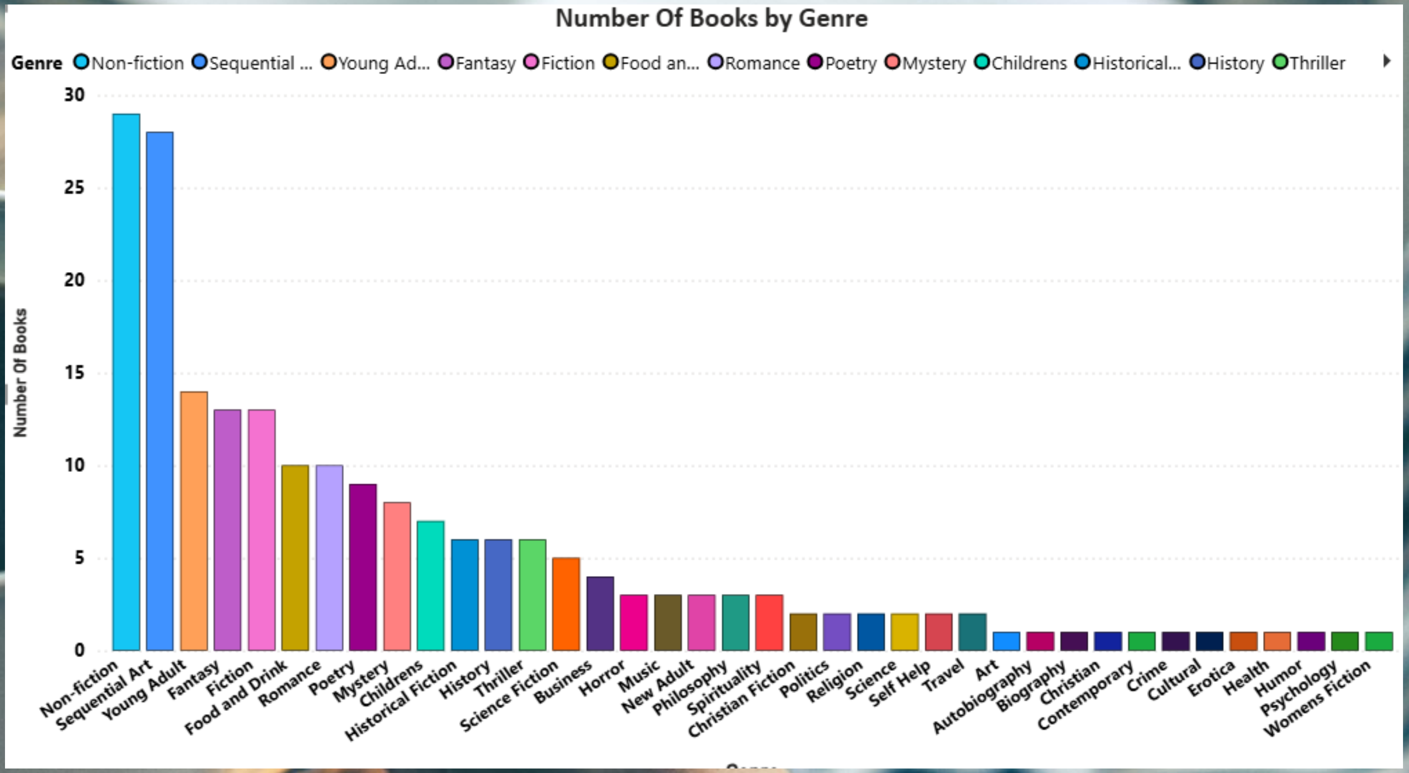
- This Pie Chart shows the "Genre" segmented by the average Rating.

- Although this gives a clear picture of the portion of genre's with the highest average. The count of books would give a clearer picture of which ones have the best average.



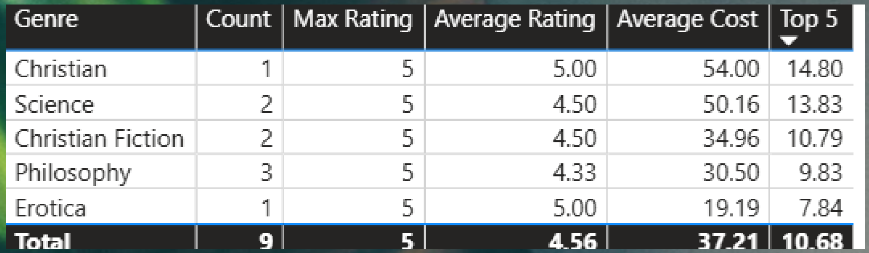
- This Bar Chart shows the "Genre" to the number of various Titles.

- Combined with the pie chart, it paints a clearer picture to determine the best average rating.

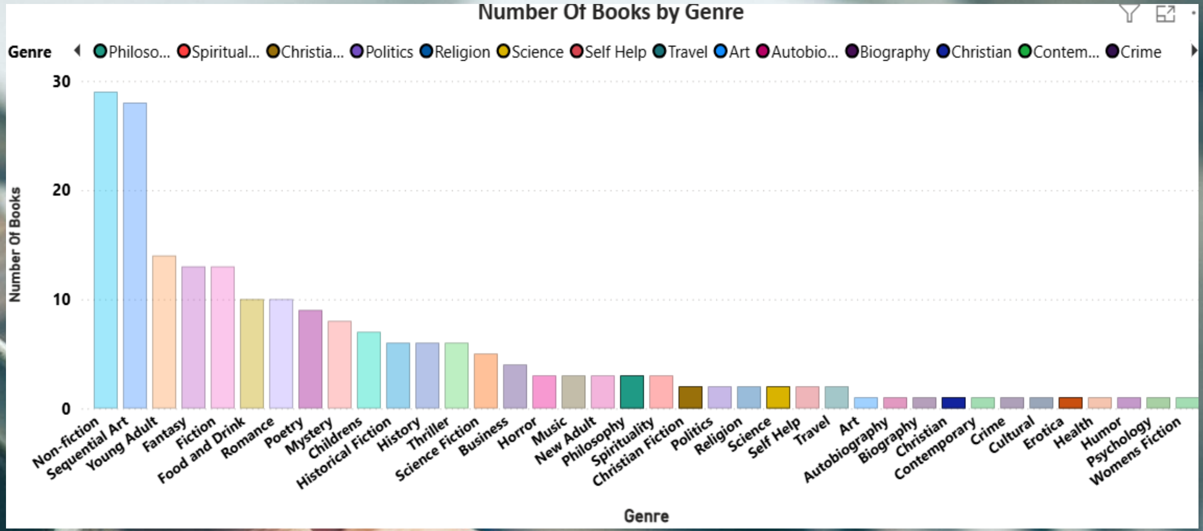


EXPLORATORY DATA ANALYSIS

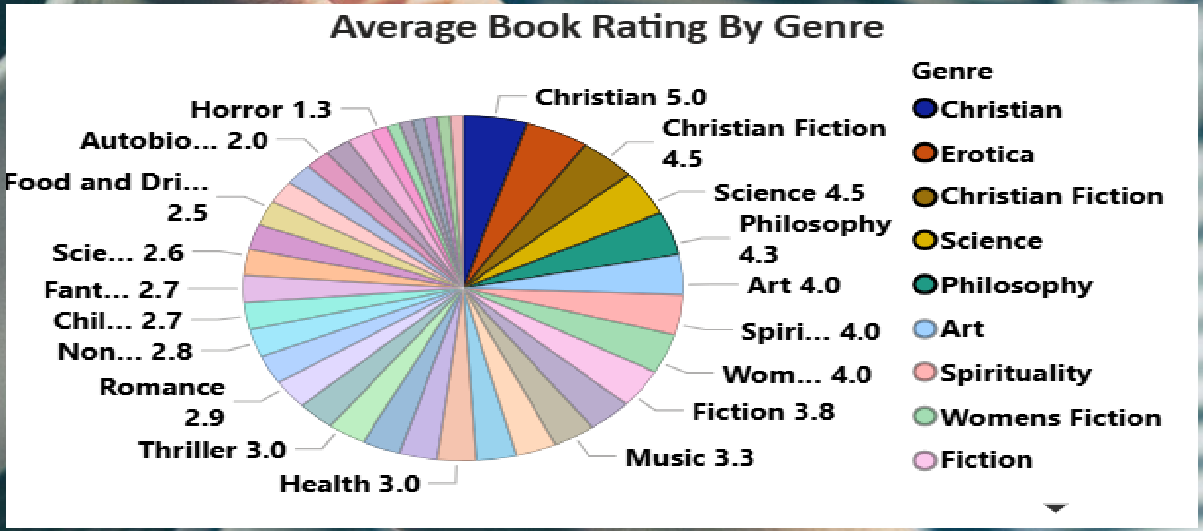
I found that the top 5 highest average rated genres had some of the fewest number of total books available, which are "Christian" "Fiction", "Erotica", "Science" and "Philosophy" with only 9 books combined between them.



Genre	Count	Max Rating	Average Rating	Average Cost	Top 5
Christian	1	5	5.00	54.00	14.80
Science	2	5	4.50	50.16	13.83
Christian Fiction	2	5	4.50	34.96	10.79
Philosophy	3	5	4.33	30.50	9.83
Erotica	1	5	5.00	19.19	7.84
Total	**9**	**5**	**4.56**	**37.21**	**10.68**

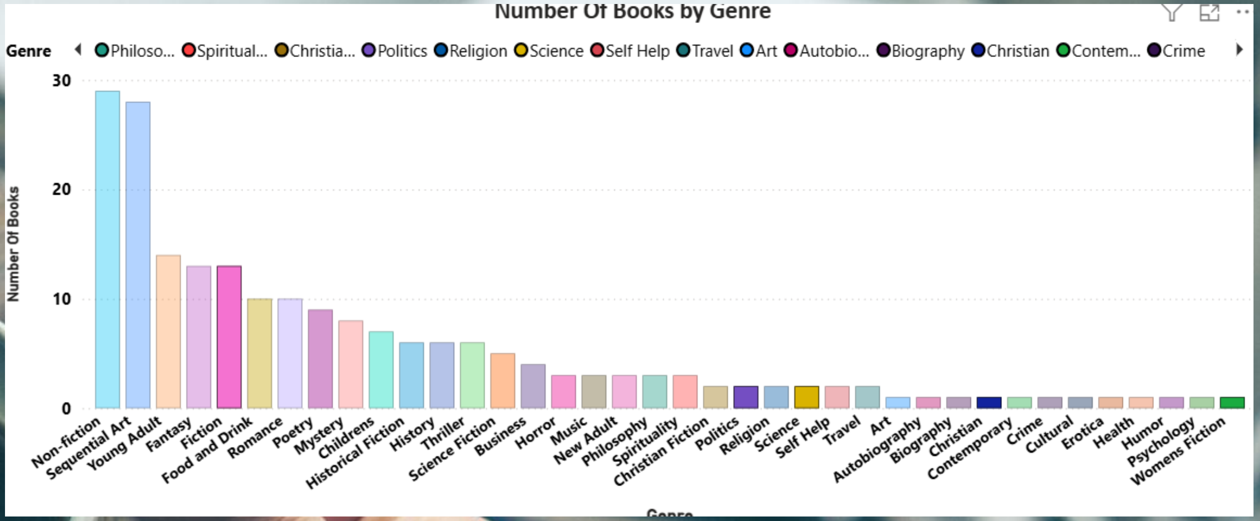


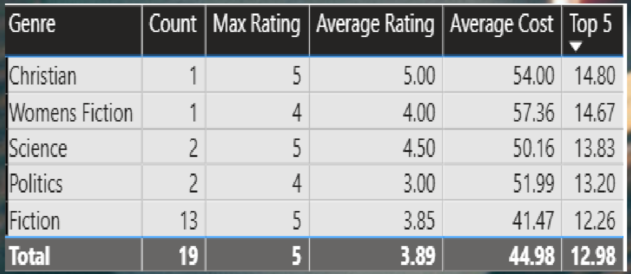
Number Of Books by Genre



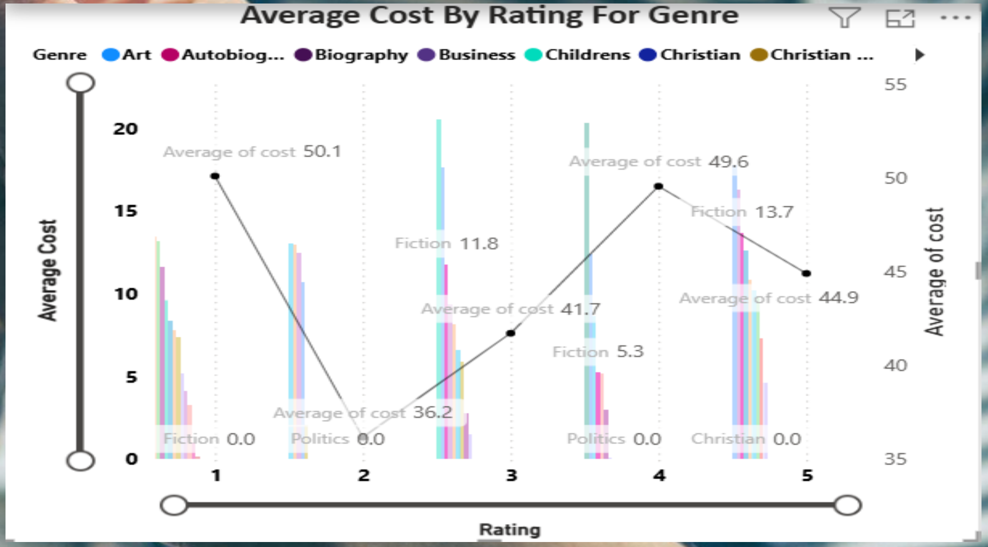
Average Book Rating By Genre

- The clustered bar chart shows the top 5 Genre's by weighted average.
- This was combined the max rating, average rating and average cost.
- The top 5 genre's with the average cost was £44.98 across them combined across all ratings and the highest average cost rating of £49.6 at 4 stars
- "Fiction" made up 13 total books of the top 5 books by weighted average.

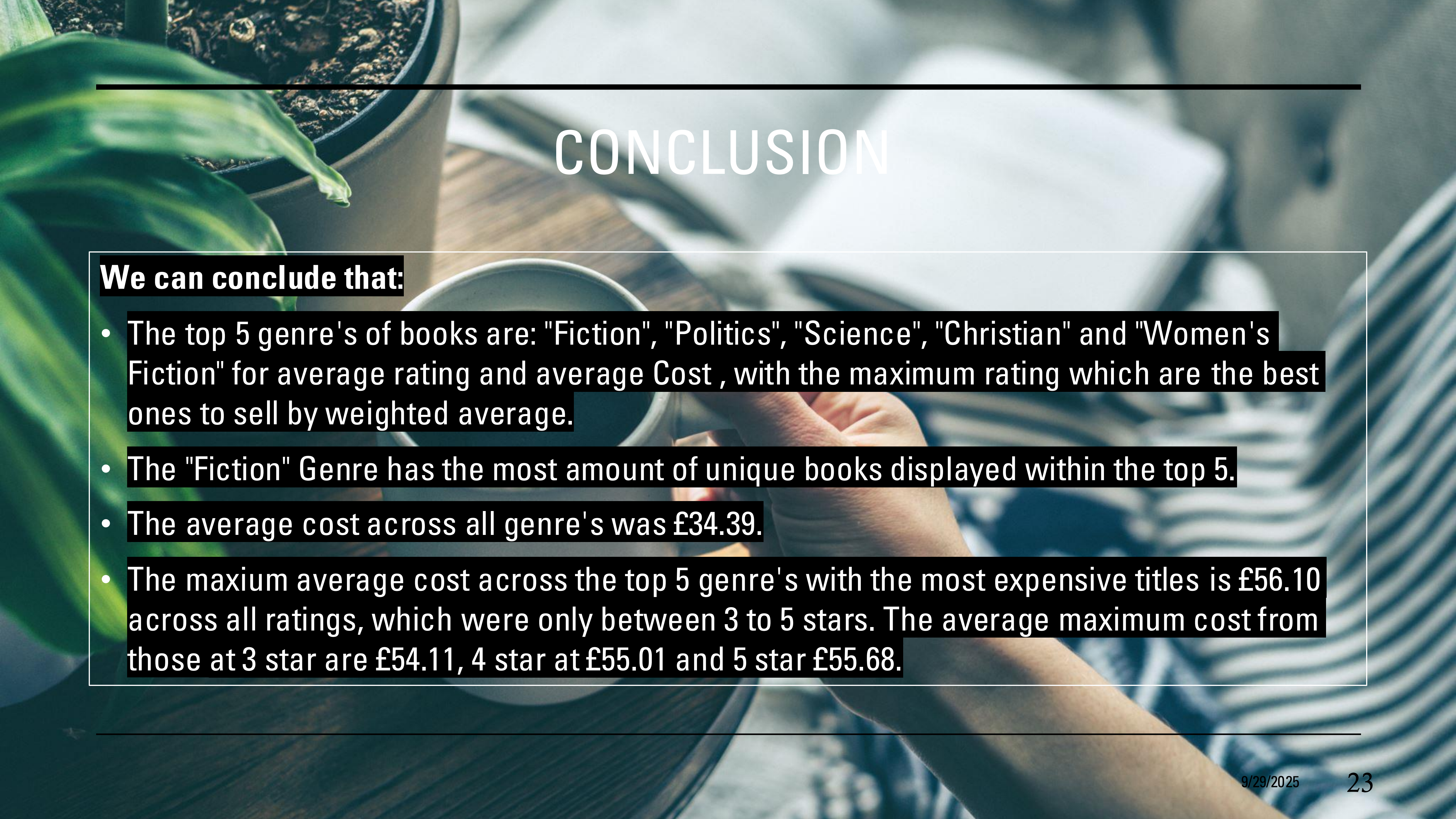




Genre	Count	Max Rating	Average Rating	Average Cost	Top 5 ▼
Christian	1	5	5.00	54.00	14.80
Womens Fiction	1	4	4.00	57.36	14.67
Science	2	5	4.50	50.16	13.83
Politics	2	4	3.00	51.99	13.20
Fiction	13	5	3.85	41.47	12.26
Total	**19**	**5**	**3.89**	**44.98**	**12.98**



CONCLUSION



CONCLUSION

We can conclude that:

- The top 5 genre's of books are: "Fiction", "Politics", "Science", "Christian" and "Women's Fiction" for average rating and average Cost , with the maximum rating which are the best ones to sell by weighted average.
- The "Fiction" Genre has the most amount of unique books displayed within the top 5.
- The average cost across all genre's was £34.39.
- The maxium average cost across the top 5 genre's with the most expensive titles is £56.10 across all ratings, which were only between 3 to 5 stars. The average maximum cost from those at 3 star are £54.11, 4 star at £55.01 and 5 star £55.68.

CONCLUSION

The top 5 books within each of those genres with the highest rating and cost are as follows:

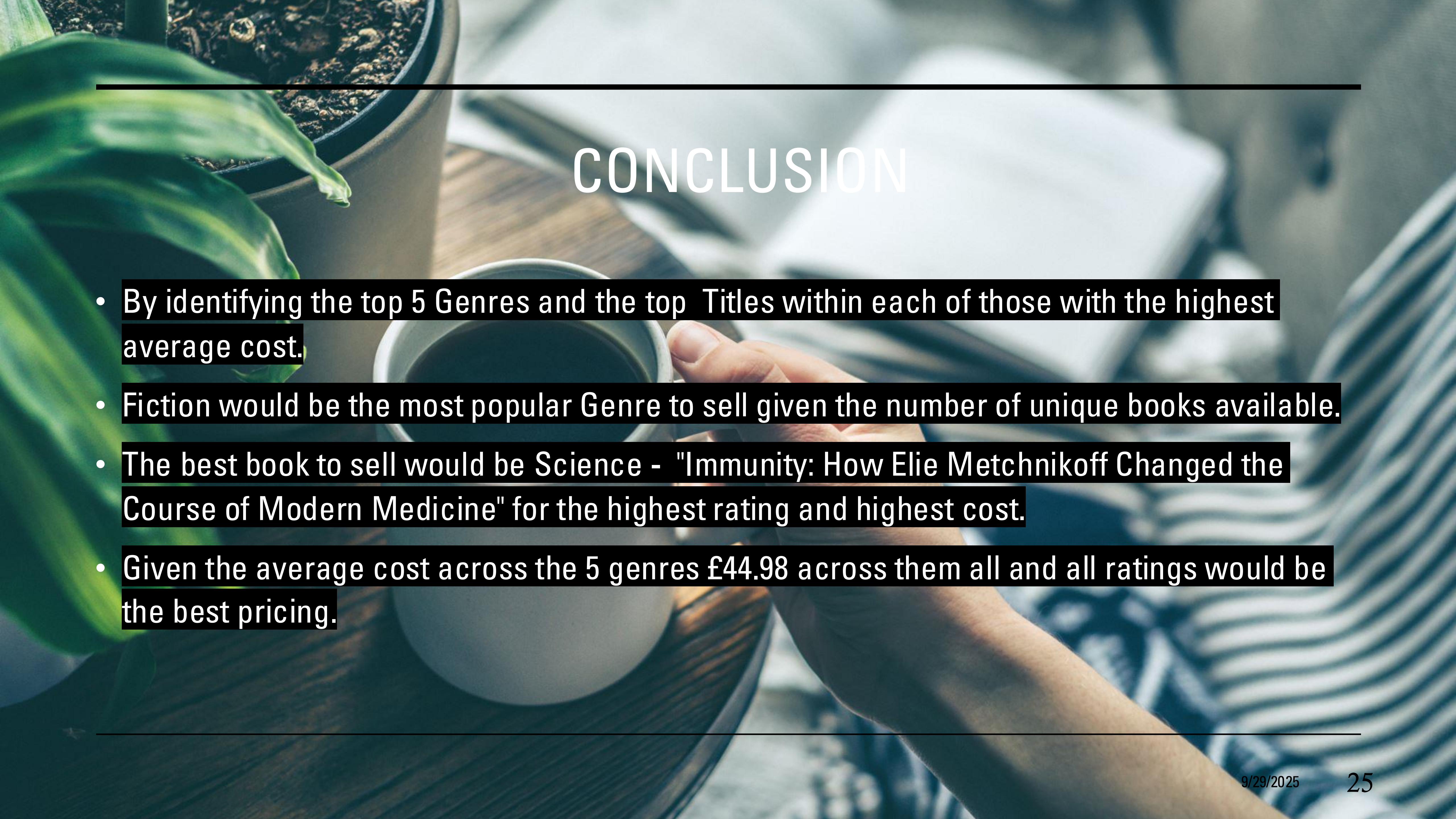
Science - "Immunity: How Elie Metchnikoff Changed the Course of Modern Medicine"

Women's Fiction - "I Had a Nice Time And Other Lies...: How to find love & sh*t like that"

Christian - "(Un)Qualified: How God Uses Broken People to Do Big Things"

Fiction - " The Murder That Never Was (Forensic Instincts #5)"

Politics - "Why the Right Went Wrong: Conservatism--From Goldwater to the Tea Party and Beyond"



CONCLUSION

- By identifying the top 5 Genres and the top Titles within each of those with the highest average cost.
- Fiction would be the most popular Genre to sell given the number of unique books available.
- The best book to sell would be Science - "Immunity: How Elie Metchnikoff Changed the Course of Modern Medicine" for the highest rating and highest cost.
- Given the average cost across the 5 genres £44.98 across them all and all ratings would be the best pricing.