

Assessment Cover Page

Module Title:	Machine Learning for Business Data Visualization Techniques
Assessment Title:	CA2 - Integrated and Individual
Lecturer Name	Dr. Muhammad Iqbal David McQuaid
Student Full Name	Laércio Santos Lima
Student Number	2022055
Assessment Due Date	Sunday, November 23rd, 2022
Date of Submission	Sunday, December 11th, 2022

Declaration

By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source. I declare it to be my own work and that all material from third parties has been appropriately referenced. I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution.

Table of Contents

1. Introduction	2
2. Recommendation System	3
2.1 Collaborative filtering vs Content filtering	3
2.1.1 Testing the models	5
2.1.1.1 Content filtering recommendations	5
2.1.1.1 Collaborative filtering recommendations	6
2.2 Weighted Scoring Table	6
2.3 Conclusion (Recommendation System)	7
3. Market Basket Analysis	8
3.1 Apriori vs FP Growth	8
3.2 Models Performance	9
4. Dashboard aimed at older adults (65+)	11
4.1 Dashboard selection	13
4.2 Dashboard Interactivity	14
4.2.1 Searching books tool	14
4.2.2 Recommendation System tool	15
4.2.3 Slider	16
4.2.4 Buttons	17
4.2.5 Main area size	18
4.2.6 Functionalities	19
5. Conclusion	20
6. Reference List	21

1. Introduction

At first, this project aims to build different recommendation system models for online retail business. Different datasets related to the book market were analyzed, cleaned and prepared in order to perform Content and Collaborative filtering.

Secondly, Market Basket Analysis was performed using Apriori and FB Growth in a second group of datasets. The model's results were compared in order to have a better understanding of the algorithms and the datasets characteristics.

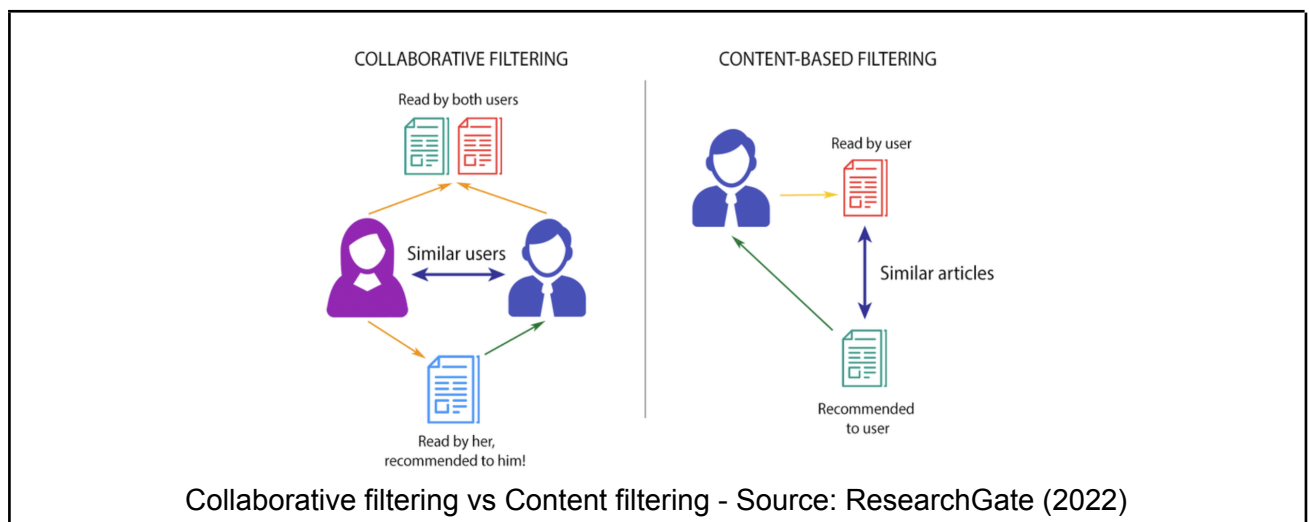
Finally, a dashboard aiming at older adults (+65) was created using Panel. It shows relevant information related to the book market that either a target user or an online retail business can take into consideration during the decision period.

Number of words in this section:	110
Total number of words until this point:	110

2. Recommendation System

A recommendation system can be described as an area of machine learning that analyzes the data to find or predict what the user is looking for among a vast number of options. According to NVIDIA Corporation (2022), a recommendation system uses the data to suggest products to consumers, taking into consideration different criteria such as last purchases, search history, etc. to help the users to discover items they may not be able to find by themselves. Considering an online retail business, a recommendation system may increase the number of purchases and the visibility of the business itself since the algorithm is trained to understand the preferences, previous decisions, and characteristics of users and products using data collected about their interactions. This project aims to build a recommendation system for books using different models and different approaches.

2.1 Collaborative filtering vs Content filtering



Collaborative filtering	Based on the preference information of different users, the algorithm recommends items. For this project, pd.pivot_table , csr_matrix and NearestNeighbors were used to analyze the features “ <i>user_ID</i> ”, “ <i>book_title</i> ” and “ <i>book_rating</i> ”.
Content filtering	<p>Based on the item characteristics, the algorithm recommends other items similar to the user’s preferences.</p> <p>For this project, at first the feature “<i>summary</i>” was analyzed using TfidfVectorizer and cosine_similarity.</p> <p>Secondly, the features “<i>book_author</i>”, “<i>publisher</i>”, “<i>category</i>” and “<i>language</i>” were analyzed using CountVectorizer and also cosine_similarity</p>

Considering the dataset used, all three approaches had good performance suggesting similar books in the same category or target public:

	Recommendations for the book “Borderliners”
Collaborative filtering (Item Based)	<p>You may like these books:</p> <p>Ingenious Pain (Harvest Book)</p> <p>The Raggy Boy Trilogy</p> <p>Wilderness Tips</p> <p>Let the Dead Bury Their Dead (Harvest American Writing Series)</p> <p>Mary Queen of Scots</p>
Content filtering + TfidfVectorizer + cosine_similarity	<p>The Growing Pains of Adrian Mole</p> <p>Searching for Caleb</p> <p>Borderliners</p> <p>Innkeeping With Murder (Lighthouse Inn Mysteries)</p> <p>Love By Design</p> <p>Tim</p> <p>Room With a View and Howards End</p> <p>Savages</p> <p>In Legend Born (Sileria)</p> <p>Isabel'S Bed</p>
Content filtering + soup of words + CountVectorizer + cosine_similarity	<p>The Growing Pains of Adrian Mole</p> <p>Searching for Caleb</p> <p>Borderliners</p> <p>Innkeeping With Murder (Lighthouse Inn Mysteries)</p> <p>Love By Design</p> <p>Tim</p> <p>Room With a View and Howards End</p> <p>Savages</p> <p>In Legend Born (Sileria)</p> <p>Isabel'S Bed</p>

2.1.1 Testing the models

Characteristics of the book “Borderliners”				
book_title	book_author	publisher	language	category
Borderliners	PETER HOEG	Delta	en	Fiction

2.1.1.1 Content filtering recommendations

Characteristics of the recommended books				
book_title	book_author	publisher	language	category
The Growing Pains of Adrian Mole	Sue Townsend	HarperTempest	en	Young Adult Fiction
Searching for Caleb	Anne Tyler	Ballantine Books	en	Fiction
Innkeeping With Murder (Lighthouse Inn Mysteries)	Tim Myers	Prime Crime	en	Fiction
Love By Design	Nora Roberts	Silhouette Books	en	Fiction
Tim	Colleen McCullough	Avon	en	Fiction
Room With a View and Howards End	E. M. Forster	Signet Classics	en	Fiction
Savages	Joe Kane	Random House Inc	en	Social Science
In Legend Born (Sileria)	Laura Resnick	Tor Fantasy	en	Fiction

2.1.1.1 Collaborative filtering recommendations

Characteristics of the recommended books				
book_title	book_author	publisher	language	category
Ingenious Pain (Harvest Book)	Andrew Miller	Harvest Books	en	Fiction
The Raggy Boy Trilogy	Patrick Galvin	New Island Books	da	Cork Ireland County
Wilderness Tips	MARGARET ATWOOD	Anchor	en	Fiction
Let the Dead Bury Their Dead (Harvest American...)	Randall Kenan	Harvest Books	en	Fiction
Mary Queen of Scots	Antonia Fraser	Delta	en	Biography Autobiography

2.2 Weighted Scoring Table

In addition, besides the three models and functions created to make the recommendations described above, a fourth approach was designed. Some feature engineering was done to create a **weighted scoring table** based on its formula. With this information, a user can use the new feature “score” to have a better idea of the book performance, and based on that, get a recommendation.

Weighted Rating Formula	Result																												
<div>$\left(\frac{v}{v+m} \times R\right) + \left(\frac{m}{v+m} \times C\right)$</div>	<table><tr><th></th><th>total_rating</th><th>book_rating</th><th>score</th></tr><tr><th>book_title</th><th></th><th></th><th></th></tr><tr><td>The Da Vinci Code</td><td>4099</td><td>10</td><td>8.809777</td></tr><tr><td>The Secret Life of Bees</td><td>3395</td><td>10</td><td>8.613129</td></tr><tr><td>The Red Tent (Bestselling Backlist)</td><td>3129</td><td>10</td><td>8.520786</td></tr><tr><td>The Nanny Diaries: A Novel</td><td>2923</td><td>10</td><td>8.440365</td></tr><tr><td>Bridget Jones's Diary</td><td>2875</td><td>10</td><td>8.420353</td></tr></table>		total_rating	book_rating	score	book_title				The Da Vinci Code	4099	10	8.809777	The Secret Life of Bees	3395	10	8.613129	The Red Tent (Bestselling Backlist)	3129	10	8.520786	The Nanny Diaries: A Novel	2923	10	8.440365	Bridget Jones's Diary	2875	10	8.420353
	total_rating	book_rating	score																										
book_title																													
The Da Vinci Code	4099	10	8.809777																										
The Secret Life of Bees	3395	10	8.613129																										
The Red Tent (Bestselling Backlist)	3129	10	8.520786																										
The Nanny Diaries: A Novel	2923	10	8.440365																										
Bridget Jones's Diary	2875	10	8.420353																										

2.3 Conclusion (Recommendation System)

The main idea for all the recommendation systems approaches implemented in this project is to provide and enhance the user experience. They were designed to predict a user's interest and suggest different books. Retail companies may have their profit increased since users are more likely to purchase items that were suggested by the algorithms created for this project.

Number of words in this section:	358
Total number of words until this point:	468

3. Market Basket Analysis

Market Basket Analysis enables retailers to identify relationships among the items bought by consumers since it looks for combinations in the purchases that often occur together in transactions.

According to TechTarget Contributor (2019), Market Basket Analysis can be described as a data mining technique used by retailers to increase sales by better understanding consumer behavior and customer purchasing patterns.

3.1 Apriori vs FP Growth

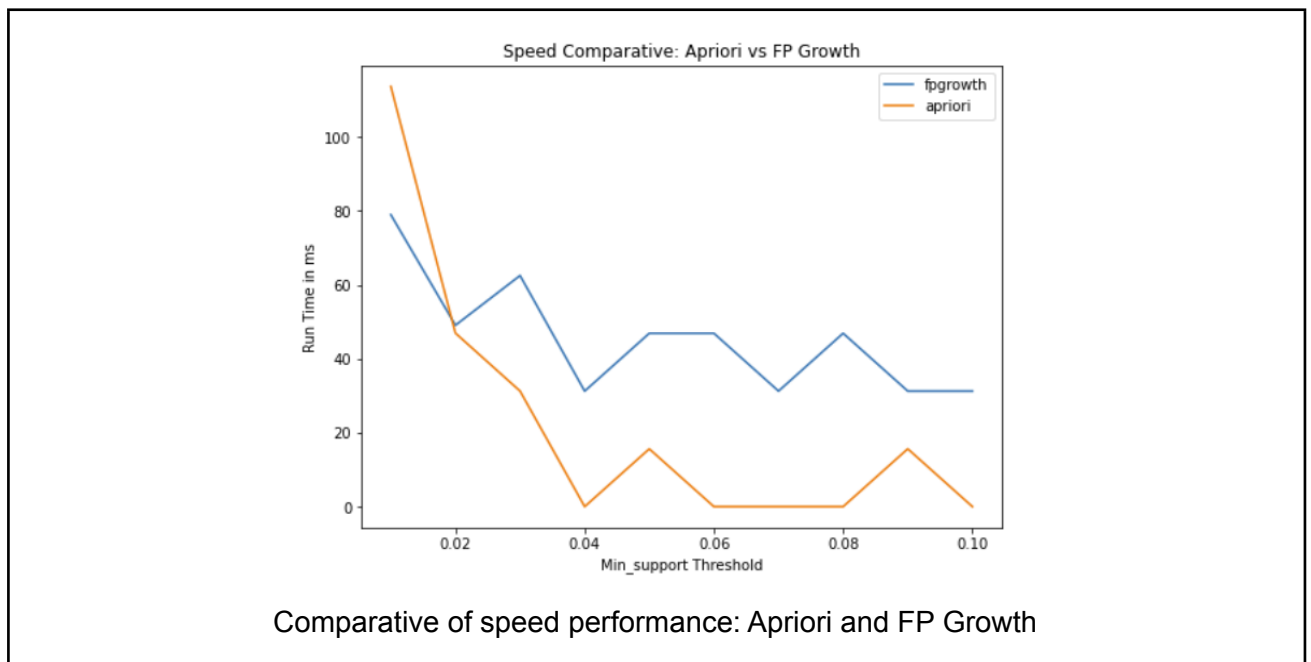
Apriori	Apriori generates frequent patterns by making the itemsets using pairing. It is likely to be slower since it scans the dataset at every step.
FP Growth	FP Growth generates an FP-Tree for making frequent patterns. It tends to be faster since it scans the dataset only once.

- **Support:** it shows the percentage of transactions that contain every item in an itemset. The higher the support the more often the itemset occurs. As can be seen, the highest support for the dataset is 0.15 (whole milk).

	support	itemsets
0	0.157923	(whole milk)
1	0.051728	(pastry)
2	0.018780	(salty snack)
3	0.085879	(yogurt)
4	0.060349	(sausage)

Support values found: Apriori and FP Growth

The image below shows the time spent for the algorithms to run 10 times in the dataset. For each run, the minimum support was increased by 10%. As expected, FP Growth was faster.



3.2 Models Performance

For this project, **association rules** were used to predict how likely one item is to be bought with another. Basically, it counts the frequency of items that occur together, and spot the ones that occur more frequently than expected.

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(yogurt)	(whole milk)	0.085879	0.157923	0.011161	0.129961	0.822940	-0.002401	0.967861
4	(rolls/buns)	(whole milk)	0.110005	0.157923	0.013968	0.126974	0.804028	-0.003404	0.964550
9	(other vegetables)	(whole milk)	0.122101	0.157923	0.014837	0.121511	0.769430	-0.004446	0.958551
2	(soda)	(whole milk)	0.097106	0.157923	0.011629	0.119752	0.758296	-0.003707	0.956636
6	(rolls/buns)	(other vegetables)	0.110005	0.122101	0.010559	0.095990	0.786154	-0.002872	0.971117

association rules - Apriori and FP Growth Algorithms

- **Confidence:** this is a measure that indicates how often a rule appears to be true.
- **Lift:** it shows how likely items are to be bought together. A value greater than 1 means a high chance while a value lower than 1 means the items are unlikely to be bought together.

Confidence		Lift	
Apriori	FP-Growth	Apriori	FP-Growth
0.129961	0.129961	0.822940	0.822940
0.126974	0.126974	0.804028	0.804028
0.121511	0.121511	0.769430	0.769430
0.119752	0.119752	0.758296	0.758296
0.095990	0.095990	0.786154	0.786154

In the table above, considering the same “**antecedents**” and “**consequent**”, it is seen that both models are equal in their results.

As expected based on the support values found in previous steps, in the scenario above it is seen that the highest confidence was 0.12 and the highest lift was 0.82, meaning that the items in this dataset are not very likely to be bought together.

Number of words in this section:	340
Total number of words until this point:	808

4. Dashboard aimed at older adults (65+)

According to Visana (2019), elderly people tend to lose interest in complicated technology. Besides that, according to the same research, elderly people tend to avoid tasks that demand lots of steps to be completed. Based on that, regarding visualization and the target demographic, the following features were applied in the dashboard:

- **Cleaner aspect:** I avoided showing a lot of information in the same area. It helps them in keeping their focus.
- **Easier plots:** I avoided graphs that were too complex to understand, such as time series and heatmaps. I avoided using a mix of colors to reduce the dimension of information.
- **Layout:** I avoided including lots of interactions. The intention is to create for them an environment where with one click they can get the information they need.
- **Interactions:** the dashboard is interactive. The idea is to provide different information in the same area, where the user can just select the information they want to visualize.
- **Fonts:** the Typographic Design Systems used to choose the characteristics of the font in this project are called “One font/One Size” and “One font / Big header”. According to France (2020), these options are considered good approaches when writing a business report or working with graphs.
- **Sizes:** The size of the graphs was changed in order to be able to show all the information in an understandable and comfortable way.
- **Colors:** The *palette colors structure* was taken into consideration in order to generate better harmony in the charts. In order to avoid problems related to colorblindness, only one color was used for the charts.

Regarding the layout, the image below gives a general idea of how the layout was planned:

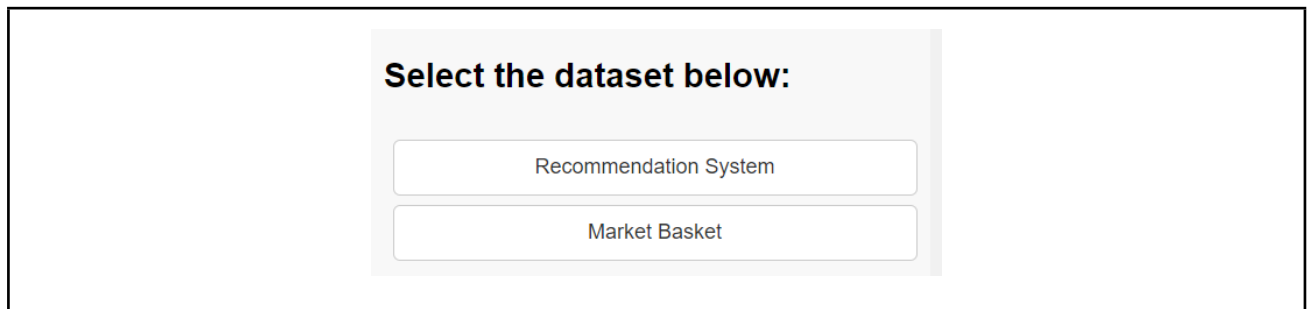


Number	Content
1	Tool that looks for books in the dataset based on any word
2	Recommendation system model tool that recommends 5 books based on a book title
3	Statistical data that can be used for getting book recommendations
4	Sliders that controls the amount of information displayed in the charts
5	dashboard selection (Dataset and main body)

4.1 Dashboard selection

Considering the college project, in this dashboard it is possible to select which dataset information should be shown. However, in future steps, this menu will be changed to show relevant information regarding only one of the topics.

The **recommendation system dashboard** is the one I decided to implement most of the functionalities.



The image shows a web interface for selecting a dataset. It features a light gray rectangular box with a dark gray border. Inside the box, at the top, is the text "Select the dataset below:" in bold. Below this text are two white rectangular buttons with rounded corners and dark gray borders. The top button is labeled "Recommendation System" and the bottom button is labeled "Market Basket".

Information the **recommendation system dashboard** can provide:

- get a book recommendation, based on a ML model (Collaborative Filtering), using a book title
- look for books with similar names in the database
- check the top books (top 1 to top 20) based on the weighted scoring table
- check the most common books, authors, publishers, languages and categories
- understand the density of the user's ages
- understand the distribution of the books based on the year they were published
- check the correlation among the rating, number of rating and score of the books
- check a table with the top books based on their weighted rating scoring

4.2 Dashboard Interactivity

4.2.1 Searching books tool

It is possible to look for different books in the dataset.

Searching box:
<p>Search for a book</p> <input type="text" value="Enter some words..."/>
Searching for a book:
<p>Find a Title:</p> <p>Search for a book</p> <input type="text" value="love"/>
<p>Books that were found:</p> <div><div>A Common Life: The Wedding Story (Beloved Mitford, No. 6)</div><div>Beloved</div><div>Beloved (Penguin Great Books of the 20th Century)</div><div>Beloved (Plume Contemporary Fiction)</div><div>Beloved: A Novel (Plume Contemporary Fiction)</div></div>

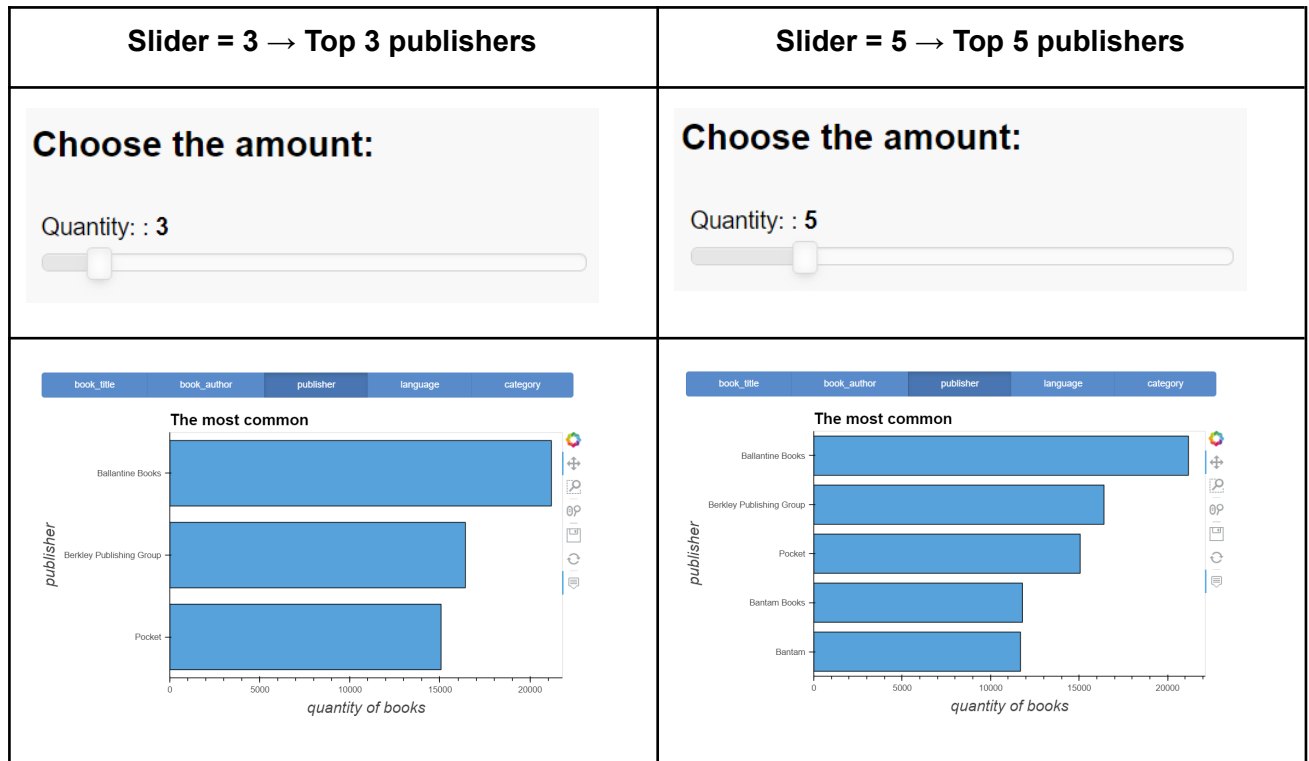
4.2.2 Recommendation System tool

It is possible to get recommendations when inserting the book name.

Recommendation System box
<p>Get a recommendation:</p> <p>Choose a book</p> <div>Enter a book name here...</div> <p>Unfortunately, the book was NOT found!</p>
Recommending books
<p>Get a recommendation:</p> <p>Choose a book</p> <div>Beloved</div> <p>You may like these books:</p> <div><div>A Memory of Love</div><div>Excalibur: A Novel of Arthur (The Warlord Chronicles: III)</div><div>Dare To Remember (Silhouette Intimate Moments, No 774)</div><div>A Woman Without Lies</div><div>Untamed</div></div>

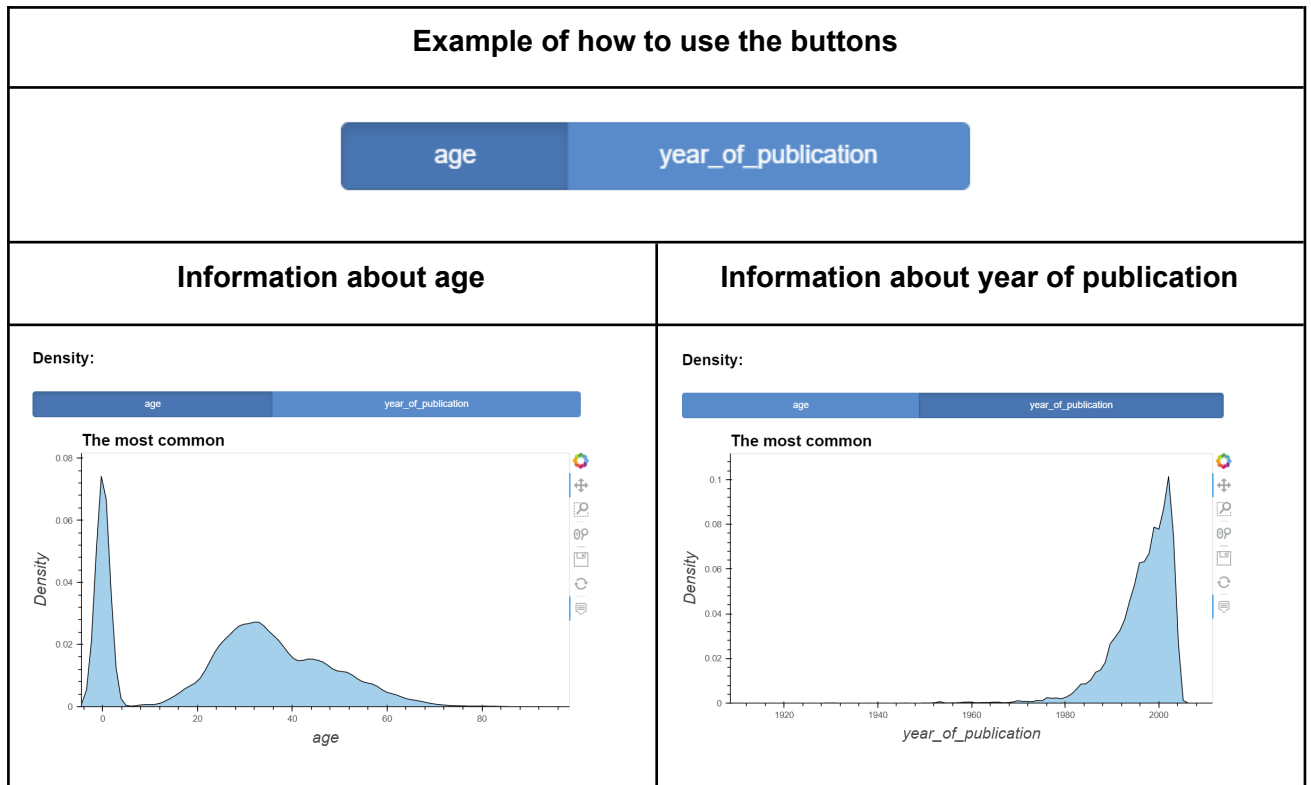
4.2.3 Slider

It can be used to change the amount of information shown in the charts. There is an example below:




4.2.4 Buttons

It can be used to select which information has to be shown in the chart.



4.2.5 Main area size

The sidebar can be hidden in order to increase the main area size:

 Integrated Dashboard

Showing the sidebar

Integrated Dashboard

Select the dataset below:

Recommendation System

Market Basket

Choose the amount:

Quantity: 5

Find a Title:

Search for a book

love

Books that were found:

A Common Life: The Wedding Story (Beloved Mitford, No. 6)

Beloved

Beloved (Penguin Great Books of the 20th Century)

Beloved (Plume Contemporary Fiction)

Beloved: A Novel (Plume Contemporary Fiction)

Get a recommendation:

Choose a book

Beloved

You may like these books:

A Memory of Love

Excalbur: A Novel of Arthur (The Warlord Chronicles: III)

Dare To Remember (Silhouette Intimate Moments, No 774)

A Woman Without Lies

Untamed

Not showing the sidebar

Integrated Dashboard

Find a Title:

Search for a book

love

Books that were found:

A Common Life: The Wedding Story (Beloved Mitford, No. 6)

Beloved

Beloved (Penguin Great Books of the 20th Century)

Beloved (Plume Contemporary Fiction)

Beloved: A Novel (Plume Contemporary Fiction)

Get a recommendation:

Choose a book

Beloved

You may like these books:

A Memory of Love

Excalbur: A Novel of Arthur (The Warlord Chronicles: III)

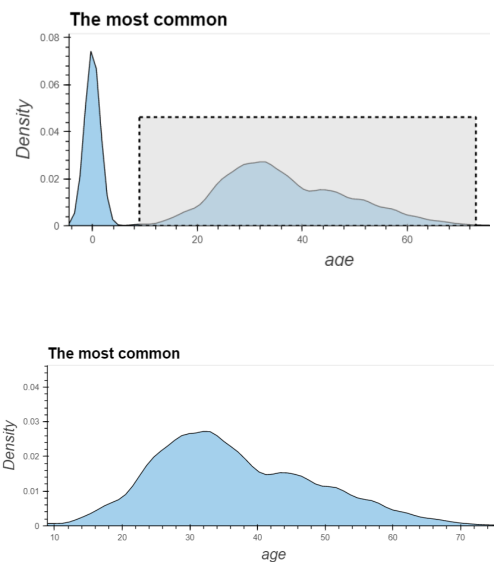
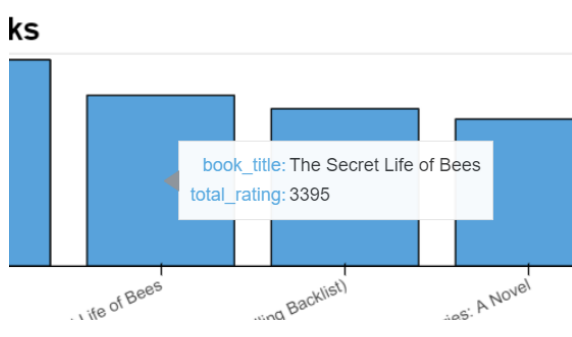

Dare To Remember (Silhouette Intimate Moments, No 774)

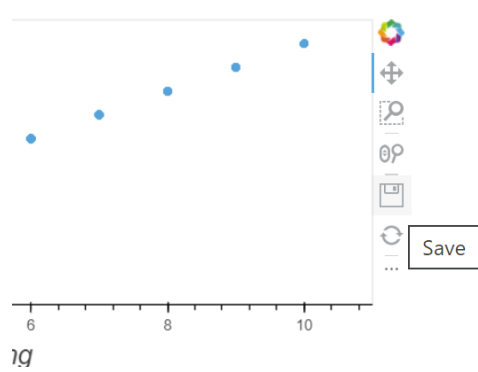
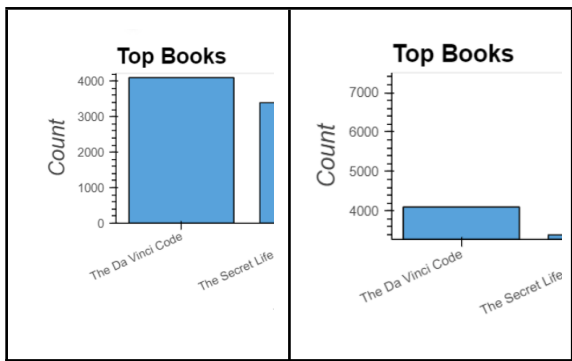


A Woman Without Lies

Untamed

4.2.6 Functionalities

Besides all the tools described above, it is also possible to interact with the graphs individually.

Zoom	Hover	Loading
		

Save as a PNG image	Move and change axis around	Graph control
		
 <p>Page Control:</p>		

Number of words in this section:	647
Total number of words until this point:	1455

5. Conclusion

No one can deny that retail business has been increasing over the years. Keeping this in mind, the importance of suggesting the right product to the right user has become almost mandatory to achieve success in the market competition.

Regarding the recommendation system, different approaches and modeling were used to create tools to make relevant suggestions to the users.

Regarding the Market Basket analysis, after comparing two different approaches, it was possible to compare the models' similarities and prove their speed performance difference.

Regarding the scenarios that were given in order to have this project completed, all the models, functions and graphs are able to transmit important statistical information for the company. This information can help in the decision making process since the company can use the graphs to understand the market better, and, consequently, choose better strategies.

Regarding the user interface, it was planned to facilitate the target user (65+) interaction. Not only the charts chosen for the dashboard were adapted, but they were also designed to be able to suggest book related content independent of the machine learning models.

Number of words in this section:	181
Total number of words until this point:	1636

6. Reference List

- Bhutani, K. (2018). *Python | Pandas dataframe.drop_duplicates()*. [online] GeeksforGeeks. Available at: https://www.geeksforgeeks.org/python-pandas-dataframe-drop_duplicates/ [Accessed 7 Dec. 2022].
- Chandradas, A. (2021). *5 Methods to Check for NaN Values in Python*. [online] Medium. Available at: <https://towardsdatascience.com/5-methods-to-check-for-nan-values-in-in-python-3f21ddd17eed#:~:text=NaN%20stands%20for%20Not%20A> [Accessed 3 Dec. 2022].
- Cravit, R. (2019). *How to Use Color Blind Friendly Palettes to Make Your Charts Accessible - Venngage*. [online] Venngage. Available at: <https://venngage.com/blog/color-blind-friendly-palette/> [Accessed 11 Dec. 2022].
- France, T. (2020). *Choosing Fonts for your Data Visualization*. [online] Medium. Available at: <https://medium.com/nightingale/choosing-a-font-for-your-data-visualization-2ed37afea637> [Accessed 9 Dec. 2022].
- Holoviz contributors (2022). *Overview — Panel v0.14.1*. [online] panel.holoviz.org. Available at: <https://panel.holoviz.org/index.html> [Accessed 6 Dec. 2022].
- NVIDIA Corporation (2022). *What is a Recommendation System?* [online] NVIDIA Data Science Glossary. Available at: <https://www.nvidia.com/en-us/glossary/data-science/recommendation-system/> [Accessed 7 Dec. 2022].
- Pandas (2018). *Python Data Analysis Library — pandas: Python Data Analysis Library*. [online] Pydata.org. Available at: <https://pandas.pydata.org/> [Accessed 6 Dec. 2022].
- Tech & Business (2021). *Collaborative Filtering In Recommender Systems: Learn All You Need To Know* |. [online] Iterators. Available at: <https://www.iteratorshq.com/blog/collaborative-filtering-in-recommender-systems/> [Accessed 9 Dec. 2022].
- TechTarget Contributor (2019). *What is market basket analysis? Definition from WhatIs.com*. [online] SearchCustomerExperience. Available at: <https://www.techtarget.com/searchcustomerexperience/definition/market-basket-analysis> [Accessed 7 Dec. 2022].

Visana (2019). *Como fazer a inclusão digital de idosos*. [online] SBGG. Available at: <https://www.sbgg-sp.com.br/como-fazer-a-inclusao-digital-de-idosos/> [Accessed 7 Dec. 2022].

Waskom, M. (2021). *seaborn: statistical data visualization — seaborn 0.10.1 documentation*. [online] seaborn.pydata.org. Available at: <https://seaborn.pydata.org/index.html> [Accessed 11 Dec. 2022].