Honor code

By submitting this assignment, I affirm the following:

1. All work presented in this assignment is my own. I have not collaborated with others or copied work from any unauthorized source.
2. If I used AI tools or large language models like ChatGPT, Co-Pilot, etc., I only sought guidance or clarification. Any generated content has been fully understood and appropriately modified to align with the assignment.
3. I understand the submitted code and can explain my work if asked.

I declare that I have read, understood, and agree to abide by this honor code.

Name: Paola D'Incecco

Student number: 775206

Date: 4th December 2025

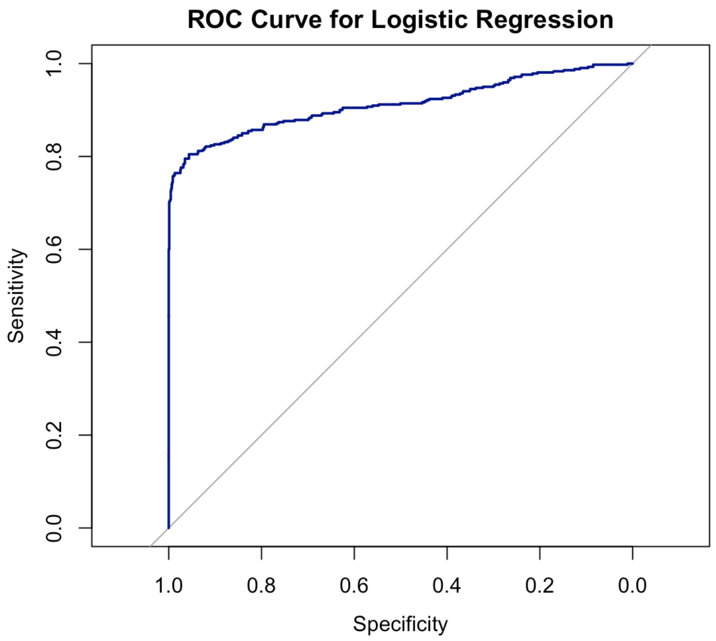
1a. Confusion matrix for threshold 0.2:

		Predicted	
Actual		1	0
	1	365	55
	0	195	722

1b. Confusion matrix for threshold 0.8:

		Predicted	
Actual		1	0
	1	273	147
	0	1	916

1c. Explanation of AUC:



An Area Under the curve (AUC) equal to 0.9122 indicates an excellent model performance, because the logistic regression model can distinguish high-charge from low-charge individuals with very high accuracy. In practice, AUC = 0.9122 means that 91.22% of the time the model assigns a higher predicted probability to a randomly chosen high-charge individual compared to a low-charge one.

2a. Confusion matrix for threshold 0.5:

		Predicted	
Actual		1	0
	1	276	144
	0	1	916

2b. Accuracy with explanation: An accuracy value of 0.8915482 indicates that this model correctly classifies 89.15% of all observations, meaning that it performs well at distinguish between high-charge and low-charge individuals.

2c. Precision with explanation: A precision value of 0.9963899 means that when this model predicts a person is in the high-charge category, it is correct over 99.63% of the time. Therefore, there are almost no false positives.

2d. Sensitivity with explanation: 0.6571429. Such value means that this model correctly identifies about 65.71% of all actual high-charge individuals. So, while the model rarely makes false negatives, it does miss some true high-charge cases.

2e. Specificity with explanation: Specificity of 0.9989095 indicates this model correctly identifies nearly 99.89% of the low-charge individuals as low-charge, meaning that false positives are extremely rare.

2f. True positive rate with explanation: Identical to sensitivity, a true positive rate of 0.6571429 indicates that this model detects 65.71% of the actual high-charge individuals.

2g. False positive rate with explanation: A false positive rate of 0.001090513 indicates that only 0.11% of the times this model is incorrect in labelling a low-charge individual as high-charge.

3a. Euclidean distance between customers 241 and 431: 0, meaning that the two customers have the same purchase pattern across all 5 categories.

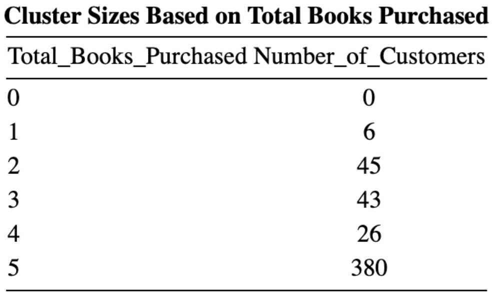
3b. Manhattan distance between customers 82 and 199: 0, meaning that the two customers have identical 0/1 values in all 5 categories.

3c. Centroid of the first 150 customers: fiction: 0.94; non_fiction: 0.88; childrens_book: 0.78; self_help: 0.89; mystery: 0.91.

4a. Categories with highest co-occurrence: Fiction and mystery, with a co-occurrence of 461.

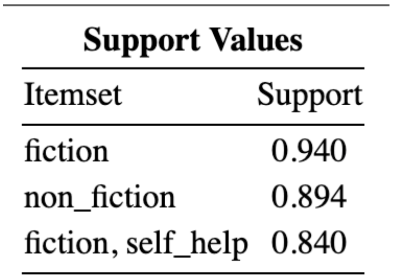
4b. Categories with lowest co-occurrence: Non_fiction and childrens_book, with a co-occurrence of 388.

5. Size of each cluster: Each cluster's number identifies the total number of books purchased per customers (on a scale from 0 to 5).



Cluster Sizes Based on Total Books Purchased

Total_Books_Purchased	Number_of_Customers
0	0
1	6
2	45
3	43
4	26
5	380



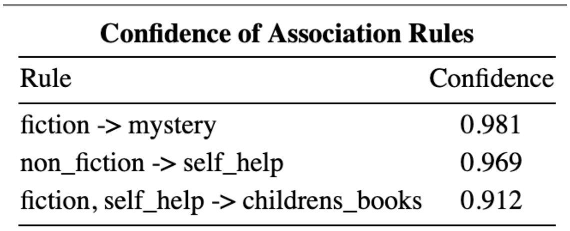
Support Values

Itemset	Support
fiction	0.940
non_fiction	0.894
fiction, self_help	0.840

6a. Support of {fiction}: 0.940 or 94.0%

6b. Support of {non_fiction}: 0.894 or 89.4%

6c. Support of {fiction, self_help}: 0.840 or 84.0%



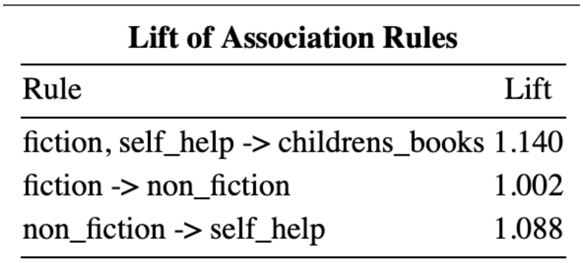
Confidence of Association Rules

Rule	Confidence
fiction -> mystery	0.981
non_fiction -> self_help	0.969
fiction, self_help -> childrens_books	0.912

7a. Confidence of {fiction} → {mystery}: 0.981 or 98.1%

7b. Confidence of {non_fiction} → {self_help}: 0.969 or 96.9%

7c. Confidence of {fiction, self_help} → {childrens_books}: 0.912 or 91.2%



Lift of Association Rules

Rule	Lift
fiction, self_help -> childrens_books	1.140
fiction -> non_fiction	1.002
non_fiction -> self_help	1.088

8a. Lift of {fiction, self_help} → {childrens_books}: 1.140 or 114.0%

8b. Lift of {fiction} → {non_fiction}: 1.002 or 100.2%

8c. Lift of {non_fiction} → {self_help}: 1.088 or 108.8%

9a. Explanation of support of {fiction, self_help}: A support of 0.840 means that 84% of all customers purchased both fiction and self-help books. This indicates that the itemset {fiction, self_help} is a very frequent one in the dataset.

9b. Explanation of confidence of {fiction, self_help} → {childrens_books}: A confidence of 0.912 shows a very strong conditional relationship which indicates that when customers purchase both fiction and self-help books, there is a 91.2% chance they also buy children's books.

9c. Explanation of lift of {fiction, self_help} → {childrens_books}: A lift value of 1.140 indicates that customers who buy both fiction and self-help are 14% more likely to also buy children's books compared to the average customer.