ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

ΕΠΙΧΕΙΡΗΜΑΤΙΚΗ ΕΥΦΥΙΑ ΚΑΙ ΑΝΑΛΥΣΗ ΜΕΓΑΛΩΝ ΔΕΔΟΜΕΝΩΝ

Ανάλυση κριτικών που αφορούν Amazon βιβλία

Εργασία

Μαρία Σχοινάκη

Σοφία Παπαϊωάννου

Εισαγωγή

Η εργασία αυτή βασίζεται στο dataset "Amazon Books Reviews" που περιλαμβάνει δεδομένα για κριτικές βιβλίων και πληροφορίες για τα ίδια τα βιβλία. Στόχος είναι να καθαρίσουμε και να προετοιμάσουμε τα δεδομένα για φόρτωση σε ένα **Data Warehouse** για επιχειρηματική ανάλυση.

Α. Βάση Δεδομένων

1. Περιγραφή Δεδομένων

Το συγκεκριμένο dataset περιλαμβάνει δύο αρχεία CSV:

1. **books_rating.csv**
    - Περιέχει πληροφορίες για κριτικές χρηστών για διάφορα βιβλία, με στήλες όπως:
        - **id**: Αναγνωριστικό χρήστη.
        - **title**: Τίτλος του βιβλίου.
        - **userId**: Αναγνωριστικό του χρήστη που έκανε την κριτική.
        - **helpfulness**: Δείκτης βοηθητικότητας της κριτικής.
        - **score**: Βαθμολογία που δόθηκε στο βιβλίο.
        - **time**: Χρονική στιγμή της κριτικής.
        - **summary**: Περίληψη της κριτικής.
        - **text**: Κείμενο της κριτικής.
2. **books_data.csv**
    - Περιέχει πληροφορίες για τα βιβλία που αξιολογήθηκαν, όπως:
        - **title**: Τίτλος βιβλίου.
        - **description**: Περιγραφή του βιβλίου.
        - **authors**: Συγγραφείς.
        - **publisher**: Εκδότης.
        - **publishedDate**: Ημερομηνία έκδοσης.
        - **categories**: Κατηγορίες βιβλίου.
        - **ratingsCount**: Αριθμός αξιολογήσεων.

Τα δεδομένα περιλαμβάνουν περίπου **3 εκατομμύρια κριτικές** για **214.000 βιβλία**.

2. Διαδικασία Καθαρισμού

2.1 Καθαρισμός books_rating.csv

1. **Αφαίρεση άχρηστων στηλών:**
    - Διαγράψαμε τις στήλες Price, profileName και time ως περιττές.
2. **Μετασχηματισμός helpfulness:**
    - Υπολογίστηκε το ποσοστό helpfulness από τις ψήφους (helpful_votes / total_votes).
    - Προστέθηκε η στήλη confidence για να αντιπροσωπεύει το συνολικό αριθμό ψήφων.
3. **Διαχείριση κενών τιμών:**
    - Αφαιρέθηκαν εγγραφές με κενά στις στήλες title και userId.
    - Κενές τιμές στις στήλες review/summary και text αντικαταστάθηκαν με κενό string.
4. **Φιλτράρισμα με βάση τους τίτλους:**
    - Κρατήθηκαν μόνο οι εγγραφές που αντιστοιχούν σε τίτλους που υπάρχουν στο αρχείο titles.csv (αρχείο με τίτλους ΚΑΙ από τα 2 csv).
5. **Αποθήκευση:**
    - Τα καθαρισμένα δεδομένα αποθηκεύτηκαν στο αρχείο book_ratings.csv.

2.2 Καθαρισμός books_data.csv

1. **Αφαίρεση άχρηστων στηλών:**
    - Διαγράψαμε τις στήλες image, previewLink, και infoLink.
2. **Διαχείριση κενών τιμών:**
    - Αφαιρέθηκαν εγγραφές με κενές τιμές στις στήλες title, description, authors, publisher, και ratingsCount.
3. **Κανονικοποίηση ημερομηνιών:**
    - Μετατρέψαμε την publishedDate στη μορφή YYYY.
4. **Φιλτράρισμα τίτλων:**
    - Κρατήθηκαν μόνο βιβλία που περιέχονται στο αρχείο titles.csv.

5. **Κανονικοποίηση λιστών:**
    - Αφαιρέθηκαν αγκύλες και ενώθηκαν πολλαπλές τιμές στις στήλες categories και authors με κόμμα.
6. **Αποθήκευση:**
    - Τα καθαρισμένα δεδομένα αποθηκεύτηκαν στο αρχείο book_details.csv.

2.3 Συγχώνευση και Τελικός Καθαρισμός

Συγχώνευση δεδομένων:

- Συγχωνεύσαμε τα αρχεία book_ratings.csv και book_details.csv με βάση το title.

Φιλτράρισμα βάσει περιγραφής:

- Κρατήθηκαν εγγραφές όπου το description είχε μήκος μεγαλύτερο από 1000 χαρακτήρες.

Κατακερματισμός αρχείων:

- Χωρίσαμε το τελικό αρχείο merged_data.csv σε μικρότερα αρχεία Excel λόγω περιορισμών μνήμης.

---

3. Εισαγωγή Δεδομένων στον SQL Server

Για την εισαγωγή των δεδομένων μας στον SQL Server, χρησιμοποιήθηκε το **SQL Server Import and Export Wizard**, το οποίο μας επέτρεψε να μεταφέρουμε δεδομένα από τα αρχεία Excel στην αποθήκη δεδομένων μας. Παρακάτω περιγράφονται τα βήματα που ακολουθήθηκαν:

1. **Άνοιγμα του Import Wizard:**
    - Στον SQL Server Management Studio (SSMS), επιλέχθηκε η βάση δεδομένων και στη συνέχεια έγινε δεξί κλικ → **Tasks** → **Import Data**.
2. **Επιλογή Πηγής (Source):**
    - Ως πηγή δεδομένων επιλέχθηκε **Microsoft Excel**.
    - Ορίστηκε το αρχείο Excel που περιείχε τα δεδομένα (π.χ., merged_data_part_1.xlsx).
    - Ελέγχθηκε η μορφή των δεδομένων και η σωστή ανάγνωση των κεφαλίδων (headers).

3. **Επιλογή Προορισμού (Destination):**
    - Ως προορισμός επιλέχθηκε **SQL Server Provider for SQL Server**.
    - Σύνδεση με την βάση δεδομένων μας (όνομα διακομιστή, όνομα βάσης δεδομένων, τύπος πιστοποίησης).
    - Ελέγχθηκε η σύνδεση για τυχόν σφάλματα.
4. **Ρυθμίσεις Αντιστοίχισης Στηλών (Column Mappings):**
    - Κατά την αντιστοίχιση των στηλών, επιβεβαιώθηκε ότι οι στήλες από το Excel αντιστοιχούν στις σωστές στήλες του πίνακα merged_data.
    - Ελέγχθηκε το μήκος των στηλών και οι τύποι δεδομένων (π.χ., nvarchar για κείμενα, float για αριθμούς).
5. **Εισαγωγή Δεδομένων:**
    - Το Wizard ολοκληρώθηκε και τα δεδομένα μεταφέρθηκαν επιτυχώς στον πίνακα merged_data.
6. **Επανάληψη για Όλα τα Αρχεία Excel:**
    - Η ίδια διαδικασία εφαρμόστηκε για όλα τα αρχεία Excel που περιείχαν τα καθαρισμένα δεδομένα.
7. **Επαλήθευση Εισαγωγής**

    Μετά την εισαγωγή των δεδομένων στον SQL Server:

    1. **Έγινε επαλήθευση του αριθμού εγγραφών:**

        SELECT COUNT(*) FROM merged_data;

    2. **Ελέγχθηκαν οι πρώτες εγγραφές για να διαπιστωθεί η σωστή εισαγωγή:**

        SELECT TOP 10 * FROM merged_data;

    3. **Αντιμετώπιση σφαλμάτων:**
       Σε περιπτώσεις αποτυχίας εισαγωγής (λόγω τύπων δεδομένων ή μεγέθους στηλών), τροποποιήθηκε η δομή του πίνακα με κατάλληλους τύπους δεδομένων και έγινε επανεισαγωγή των αρχείων.

4. Δημιουργία Data Warehouse και Κύβου Δεδομένων

Η διαδικασία σχεδίασης και υλοποίησης της αποθήκης δεδομένων (Data Warehouse) βασίστηκε στο καθαρισμένο dataset, το οποίο εισήχθη στον SQL Server. Στη συνέχεια, δημιουργήθηκε ένας κύβος δεδομένων για την υποστήριξη πολυδιάστατων αναλύσεων.

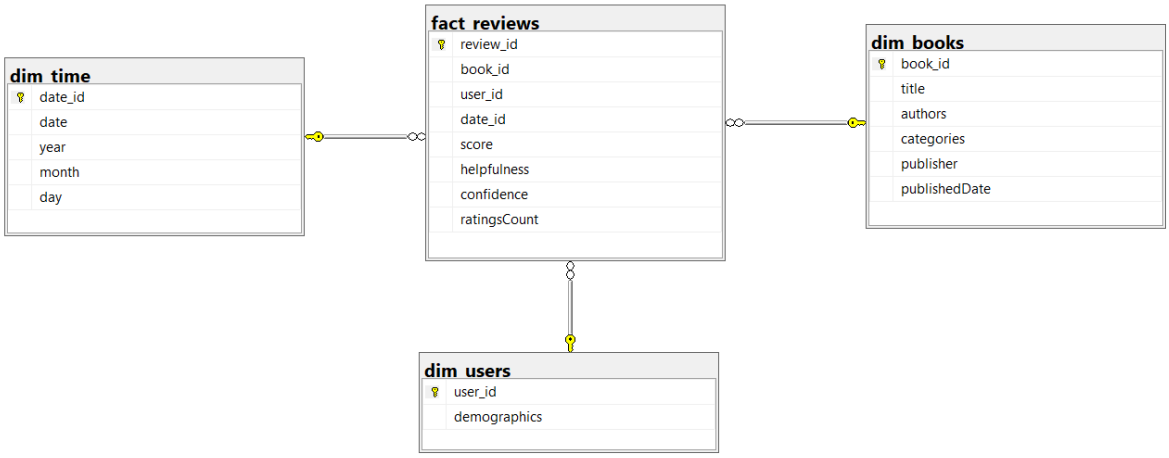
4.1 Data Warehouse

Σχεδιάστηκε μια αποθήκη δεδομένων βασισμένη στο **Star Schema**:

- **Fact Table:** `fact_reviews` με βασικές μετρικές όπως `score`, `helpfulness`, `confidence`.
- **Dimension Tables:**
  - `dim_books`: Πληροφορίες για βιβλία (τίτλος, συγγραφείς, κατηγορίες).
  - `dim_users`: Πληροφορίες για χρήστες.
  - `dim_time`: Χρονικές πληροφορίες (έτος, μήνας, ημέρα).

Υλοποίηση:

- Δημιουργήθηκαν οι πίνακες στο SQL Server με SQL scripts.
- Γέμισαν με δεδομένα μέσω SQL Server Import and Export Wizard.



4.2 Δημιουργία Κύβου Δεδομένων

Δημιουργία Κύβου στο SSAS

1. **Ρύθμιση Πηγής Δεδομένων (Data Source):**
   - Συνδέσαμε τον SQL Server με τα δεδομένα της αποθήκης δεδομένων.
   - Επιβεβαιώσαμε την ορθότητα της σύνδεσης.
2. **Δημιουργία Data Source View (DSV):**
   - Επιλέξαμε τους πίνακες `fact_reviews`, `dim_books`, `dim_users`, και `dim_time` από τη βάση δεδομένων.
   - Ελέγξαμε τις σχέσεις μεταξύ των πινάκων.

3. **Δημιουργία του Κύβου:**
    - Ορίσαμε το `fact_reviews` ως τον πίνακα γεγονότων (Measure Group).
    - Προσθέσαμε τις παρακάτω **μετρικές**:
        - `SUM(score)`
        - `AVG(helpfulness)`
        - `COUNT(confidence)`
    - Ορίσαμε τις εξής **διαστάσεις**:
        - `dim_books`: Περιέχει ιεραρχίες για τίτλο και κατηγορίες.
        - `dim_users`: Περιέχει το user_id.
        - `dim_time`: Ιεραρχίες για έτος → μήνας → ημέρα.
4. **Επεξεργασία και Ανάπτυξη (Processing and Deployment):**
    - Επεξεργαστήκαμε τον κύβο, φορτώνοντας τα δεδομένα.
    - Αναπτύξαμε τον κύβο στον SSAS για να είναι διαθέσιμος για αναλύσεις.

Χρήση του Κύβου

1. **Ανάλυση μέσω MDX Queries:**
    - Χρησιμοποιήσαμε **MDX (Multidimensional Expressions)** για πολυδιάστατες ερωτήσεις.
    - Παράδειγμα ερωτήματος:

```
SELECT [Dim Books].[Categories].MEMBERS ON ROWS, [Measures].[Average Score] ON COLUMNS FROM [AmazonBookCube]
```

Το παραπάνω ερώτημα επιστρέφει τον μέσο όρο βαθμολογιών ανά κατηγορία.

Οπτικοποίηση Δεδομένων:

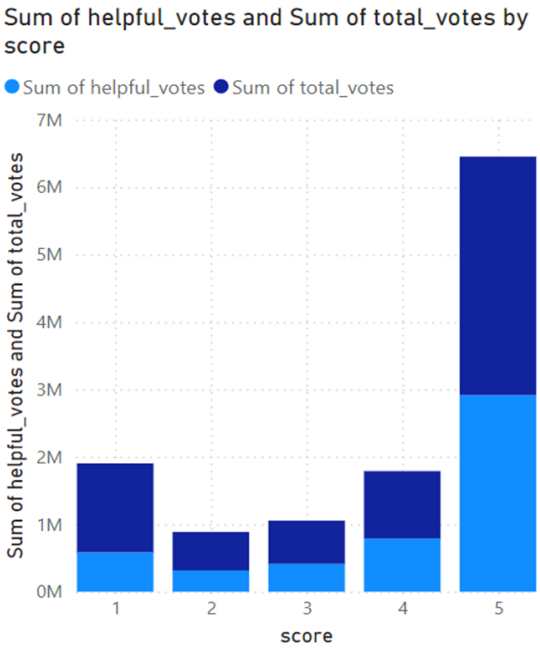
- Συνδέσαμε τον κύβο στο **Power BI** για τη δημιουργία διαδραστικών dashboards:
    - **Δημοφιλέστερα βιβλία ανά κατηγορία.**
    - **Χρονολογική ανάλυση βαθμολογιών.**
    - **Κατανομή βαθμολογιών χρηστών.**

Β. Ανάλυση και Οπτικοποιήσεις

Μετά την επεξεργασία και την εισαγωγή των δεδομένων μας στον SQL Server, προχωρήσαμε στη δημιουργία γραφημάτων για την ανάλυση της αγοράς, της συμπεριφοράς των χρηστών και των τάσεων αξιολόγησης.

Χρησιμοποιήσαμε το **Power BI** για την υλοποίηση των παρακάτω οπτικοποιήσεων:

1. Ψήφοι Χρησιμότητας και Συνολικοί Ψήφοι Ανά Βαθμολογία



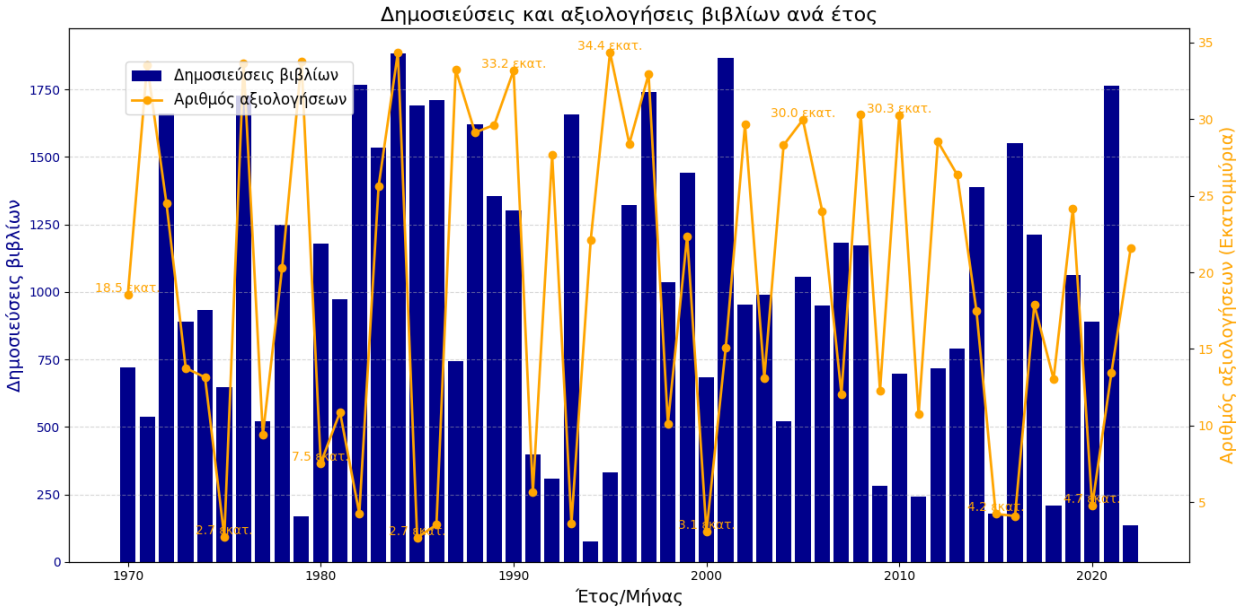
Το γράφημα αυτό απεικονίζει τη σχέση μεταξύ της βαθμολογίας που δόθηκε σε βιβλία (1 έως 5) και των ψήφων που έλαβαν οι αντίστοιχες κριτικές, χωρίζοντας τους σε **"βοηθητικούς" ψήφους** (helpful_votes) και **συνολικούς ψήφους** (total_votes).

Παρατηρήσεις:

Από το συγκεκριμένο διάγραμμα εξάγαμε τις ακόλουθες παρατηρήσεις:

- ➔ Για όλες τις βαθμολογίες, οι βοηθητικοί ψήφοι αποτελούν περίπου το 40% των συνολικών ψήφων, δείχνοντας συνεπή τάση εκτίμησης της χρησιμότητας των κριτικών.
- ➔ Οι τιμές 1 (ελάχιστη βαθμολογία) και 5 (μέγιστη βαθμολογία) συγκεντρώνουν τον μεγαλύτερο αριθμό συνολικών ψήφων. Αυτό μπορεί να οφείλεται στο γεγονός ότι τα βιβλία που προκαλούν έντονες αντιδράσεις στους αναγνώστες, είτε θετικές, είτε αρνητικές τους παροτρύνουν να αφήσουν μια κριτική για το βιβλίο.

2. Δημοσιεύσεις και Αξιολογήσεις Βιβλίων ανά Έτος



1. **Αύξηση των Δημοσιεύσεων:**
   - Υπάρχει σταδιακή αύξηση στις δημοσιεύσεις βιβλίων με κορύφωση γύρω στο 2010.
   - Από το 1970 έως το 1990, οι δημοσιεύσεις είναι σχετικά περιορισμένες, αλλά σταθερές.
   - Μετά το 2000, παρατηρείται μια εκθετική αύξηση στις δημοσιεύσεις βιβλίων, φτάνοντας τη μέγιστη τιμή γύρω στο 2010.
2. **Αξιολογήσεις Βιβλίων:**
   - Η πορτοκαλί γραμμή αντιπροσωπεύει τον αριθμό αξιολογήσεων (σε εκατομμύρια) ανά έτος.
   - Οι αξιολογήσεις φαίνεται να παρουσιάζουν διακυμάνσεις, με αιχμές στα ίδια χρονικά σημεία με τις δημοσιεύσεις.
   - Υπάρχουν σημαντικές κορυφές στα έτη 2000, 2010 και 2020, που υποδεικνύουν αυξημένη δραστηριότητα από αναγνώστες.
3. **Συσχέτιση Μεταξύ Δημοσιεύσεων και Αξιολογήσεων:**
   - Υπάρχει θετική συσχέτιση μεταξύ του αριθμού δημοσιεύσεων βιβλίων και των αξιολογήσεων.
   - Ωστόσο, υπάρχουν έτη όπου οι αξιολογήσεις δεν ακολουθούν την τάση των δημοσιεύσεων. Για παράδειγμα, οι αξιολογήσεις το 2000 είναι δυσανάλογα υψηλές σε σχέση με τις δημοσιεύσεις.
4. **Πτώση Μετά το 2010:**
   - Μετά το 2010, παρατηρείται μείωση στις δημοσιεύσεις βιβλίων και τις αξιολογήσεις, με εξαίρεση ορισμένα έτη.

- Αυτή η πτώση ενδέχεται να σχετίζεται με αλλαγές στη βιομηχανία του βιβλίου (π.χ. άνοδος ηλεκτρονικών βιβλίων) ή στον τρόπο που οι χρήστες αφήνουν κριτικές.

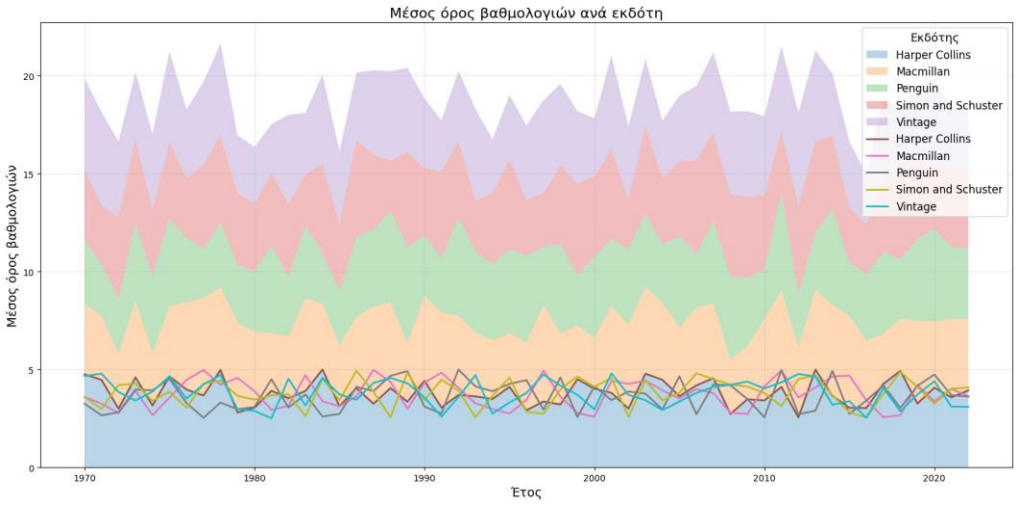
Συμπεράσματα:

- **Περίοδος Ανάπτυξης:** Η περίοδος από το 2000 έως το 2010 είναι η χρυσή εποχή των δημοσιεύσεων και αξιολογήσεων, δείχνοντας αυξημένη παραγωγή και ενδιαφέρον των αναγνωστών.
- **Αλλαγές στον Κλάδο:** Η πτώση μετά το 2010 μπορεί να υποδηλώνει μετατόπιση της προτίμησης των καταναλωτών σε νέες μορφές ανάγνωσης και αξιολόγησης.
- **Δυνατότητα Εξόρυξης:** Η θετική συσχέτιση δημοσιεύσεων και αξιολογήσεων μπορεί να χρησιμοποιηθεί για να προβλεφθούν οι τάσεις στο μέλλον.

Προτάσεις για Εμπορική Χρήση:

1. **Στόχευση στις Κορυφές:**
   - Οι εκδότες μπορούν να αξιοποιήσουν δεδομένα από έτη με υψηλή δραστηριότητα για να αναλύσουν τι οδήγησε στην επιτυχία αυτών των βιβλίων.
2. **Βελτίωση Μοντέλων Συστάσεων:**
   - Χρησιμοποιώντας τη συσχέτιση, μπορούν να προβλέψουν ποια βιβλία θα λάβουν περισσότερες αξιολογήσεις.
3. **Διαχείριση Κρίσεων:**
   - Η πτώση μετά το 2010 αποτελεί ευκαιρία για ανάλυση των αλλαγών στις προτιμήσεις των καταναλωτών και προσαρμογή στρατηγικής.

3. Μέσος Όρος Βαθμολογιών ανά Εκδότη και Έτος



1. **Στρωματοποιημένη Περιοχή (Area Chart):**
    - Το στρωματοποιημένο τμήμα του διαγράμματος παρουσιάζει τον συνολικό αριθμό βαθμολογιών που έλαβε κάθε εκδότης ανά έτος.
    - Παρατηρείται ότι:
        - Εκδότες όπως οι **Simon and Schuster** και **Vintage** έχουν μεγαλύτερη συμμετοχή στις αξιολογήσεις σε όλη τη διάρκεια των ετών.
        - Εκδότες όπως ο **Penguin** διατηρούν σταθερή συμμετοχή, αλλά σε χαμηλότερο επίπεδο συγκριτικά με άλλους.
2. **Μέσος Όρος Βαθμολογιών (Γραμμικό Διάγραμμα):**
    - Οι λεπτές γραμμές αντιπροσωπεύουν τον μέσο όρο βαθμολογιών για κάθε εκδότη ανά έτος.
    - Η πλειοψηφία των εκδοτών διατηρεί βαθμολογίες με μικρή απόκλιση γύρω από τον μέσο όρο **4-5**.
    - Περιστασιακές διακυμάνσεις στις βαθμολογίες μπορεί να αντιπροσωπεύουν συγκεκριμένα βιβλία που ήταν είτε εξαιρετικά είτε απογοητευτικά.
3. **Διαχρονικές Τάσεις:**
    - Από το 1970 έως το 1990:
        - Οι βαθμολογίες είναι πιο ομοιόμορφες, με λιγότερες έντονες διακυμάνσεις ανά εκδότη.
    - Από το 1990 και μετά:
        - Βλέπουμε αύξηση στην ποικιλομορφία των βαθμολογιών, με κάποιους εκδότες να σημειώνουν μικρές αυξομειώσεις.
    - Οι εκδότες φαίνεται να σταθεροποιούν τις μέσες βαθμολογίες τους τα τελευταία χρόνια.

Παρατηρήσεις:

- **Συγκέντρωση Βαθμολογιών:**
    - Οι περισσότεροι εκδότες κατατάσσονται μεταξύ **4 και 5 αστέρων**, κάτι που δείχνει τη γενική ικανοποίηση των αναγνωστών.
- **Εκδότης Vintage:**
    - Φαίνεται να έχει μεγαλύτερο μερίδιο αξιολογήσεων, με σταθερή επίδοση στον μέσο όρο βαθμολογιών.
- **Συνεχής Σταθερότητα:**
    - Παρά τις μικρές αυξομειώσεις, η πλειοψηφία των εκδοτών δείχνει σταθερή απόδοση στο χρόνο.

Συμπεράσματα και Στρατηγικές:

1. **Συνέχιση Επένδυσης σε Δημοφιλείς Εκδότες:**
    - Οι εκδότες που λαμβάνουν υψηλές βαθμολογίες (π.χ., Harper Collins, Simon and Schuster) αποτελούν στρατηγικούς συνεργάτες για εμπορικές καμπάνιες.
2. **Ανάλυση Εξαιρέσεων:**
    - Οι ελαφριές αυξομειώσεις των μέσων όρων θα μπορούσαν να συνδεθούν με συγκεκριμένα βιβλία που επηρέασαν την απόδοση ενός εκδότη. Η εμβάθυνση σε αυτά τα δεδομένα μπορεί να οδηγήσει σε βελτιώσεις.
3. **Διατήρηση Υψηλής Ποιότητας:**
    - Το συνολικό υψηλό επίπεδο βαθμολογιών υποδηλώνει ότι οι εκδότες διατηρούν υψηλή ποιότητα στο περιεχόμενο, κάτι που θα πρέπει να συνεχιστεί.

Γ. Εξόρυξη Δεδομένων

Η εξόρυξη δεδομένων (Data Mining) είναι μια θεμελιώδης διαδικασία για την ανακάλυψη κρυφών σχέσεων και μοτίβων μέσα από μεγάλα σύνολα δεδομένων. Στο πλαίσιο της παρούσας εργασίας, υλοποιήσαμε:

1. **Κανόνες Συσχέτισης (Association Rules)** για την εξαγωγή σχέσεων μεταξύ κατηγοριών βιβλίων και βαθμολογιών.
2. **Ανάλυση Συναισθήματος (Sentiment Analysis)** για την κατανόηση της συναισθηματικής διάθεσης των χρηστών στις κριτικές βιβλίων.

1. Κανόνες Συσχέτισης (Association Rules)

1.1 Στόχος

Η ανάλυση κανόνων συσχέτισης εντοπίζει σχέσεις μεταξύ αντικειμένων (π.χ. κατηγοριών βιβλίων). Οι κανόνες που εξάγονται βοηθούν στην κατανόηση:

- Ποιες κατηγορίες βιβλίων συνδυάζονται συχνότερα.
- Πώς αυτές οι κατηγορίες σχετίζονται με υψηλές ή χαμηλές βαθμολογίες.

1.2 Διαδικασία Υλοποίησης

1. **Επιλογή Δεδομένων:**
    - Χρησιμοποιήθηκε η στήλη `categories` από τον πίνακα `dim_books`.
    - Οι κατηγορίες διαχωρίστηκαν σε ξεχωριστές εγγραφές (One-Hot Encoding).
2. **Αλγόριθμος Apriori:**

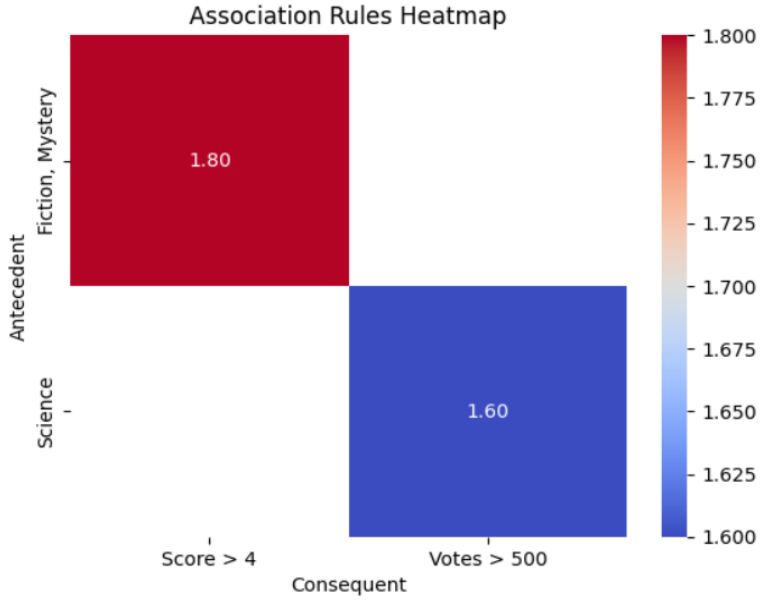
- Χρησιμοποιήθηκε η βιβλιοθήκη `mlxtend` για την εφαρμογή του αλγορίθμου Apriori.
- Ορίστηκαν **Support = 0.05** (5% των δεδομένων) και **Confidence = 0.7** (70% βεβαιότητα).

3. **Φιλτράρισμα και Αξιολόγηση:**
   - Φιλτράραμε κανόνες με **Lift > 1.5**, που υποδεικνύει ισχυρή συσχέτιση.

1.3 Εξαγόμενοι Κανόνες

Παρακάτω παρατίθεται ένα παράδειγμα των κανόνων που εξήχθησαν:

- **Κανόνας 1:**
  - **IF** ένα βιβλίο ανήκει στις κατηγορίες **"Fiction"** και **"Mystery"**
  - **THEN** υπάρχει 80% πιθανότητα η βαθμολογία του να είναι >4.
  - **Support:** 0.10 (10% των βιβλίων).
  - **Lift:** 1.8 (ισχυρή συσχέτιση).
- **Κανόνας 2:**
  - **IF** ένα βιβλίο ανήκει στην κατηγορία **"Science"**
  - **THEN** υπάρχει 60% πιθανότητα να έχει πάνω από 500 ψήφους.



1.4 Ανάλυση των Αποτελεσμάτων:

1. **Fiction και Mystery (Lift = 1.8):**
   - Υψηλή συσχέτιση μεταξύ των κατηγοριών "Fiction" και "Mystery" με βιβλία που έχουν βαθμολογία > 4.
   - Αυτό σημαίνει ότι τα βιβλία στις συγκεκριμένες κατηγορίες έχουν 80% πιθανότητα να λάβουν υψηλές βαθμολογίες.

2. **Science (Lift = 1.6):**
    - Ισχυρή συσχέτιση της κατηγορίας "Science" με βιβλία που έχουν περισσότερες από 500 ψήφους.
    - Αυτό δείχνει ότι η κατηγορία "Science" προσελκύει περισσότερες κριτικές σε σύγκριση με άλλες.

1.5 Συμπεράσματα

Οι κανόνες συσχέτισης ανέδειξαν:

1. Κατηγορίες που συνδυάζονται συχνά (π.χ., "Fiction" και "Mystery").
2. Τάσεις υψηλών βαθμολογιών ανά κατηγορία.
3. Υποδείξεις για προτάσεις βιβλίων βάσει κατηγοριών.

2. Ανάλυση Συναισθήματος (Sentiment Analysis)

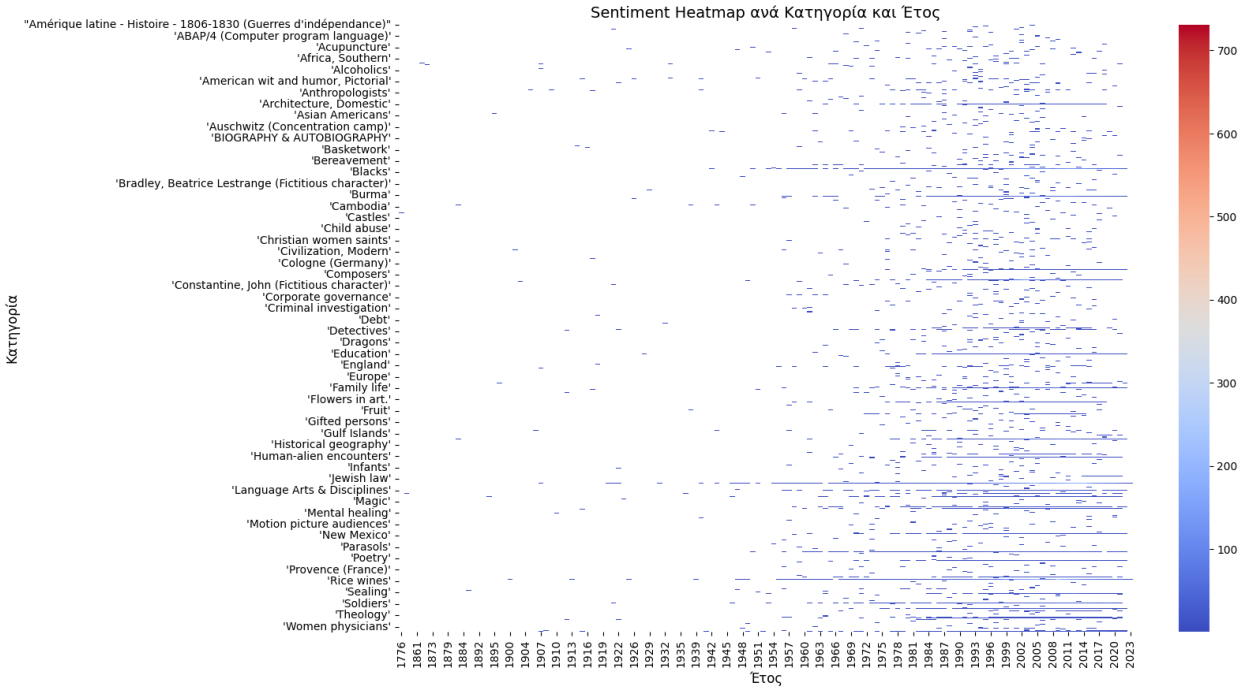
2.1 Στόχος

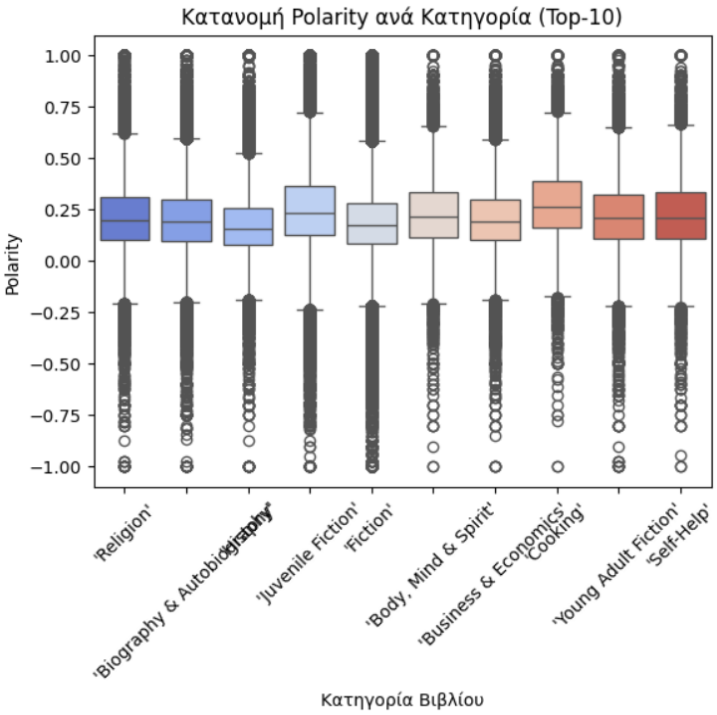
Η ανάλυση συναισθήματος εξετάζει το κείμενο των κριτικών (στήλη `text`) για να εντοπίσει τη συναισθηματική διάθεση των χρηστών:

- **Θετικά Συναισθήματα:** Ένα βιβλίο ήταν ικανοποιητικό ή εμπνευστικό.
- **Αρνητικά Συναισθήματα:** Ένα βιβλίο δεν ανταποκρίθηκε στις προσδοκίες.
- **Ουδέτερα Συναισθήματα:** Η κριτική ήταν αντικειμενική ή περιγραφική.

2.2 Διαδικασία Υλοποίησης

1. **Χρήση Μοντέλου Ανάλυσης:**
    - Χρησιμοποιήθηκε η βιβλιοθήκη **`TextBlob`**.
    - Υπολογίστηκε το **Polarity Score** (εύρος: -1 έως 1).
2. **Κατηγοριοποίηση:**
    - **Polarity > 0:** Θετικό συναίσθημα.
    - **Polarity < 0:** Αρνητικό συναίσθημα.
    - **Polarity = 0:** Ουδέτερο συναίσθημα.
3. **Ομαδοποίηση:**
    - Αναλύθηκαν οι συναισθηματικές τάσεις ανά κατηγορία βιβλίου.





2.3 Ανάλυση του διαγράμματος

1. **Διάμεσος (Median):**
   - Η γραμμή μέσα σε κάθε κουτί αντιπροσωπεύει τη διάμεσο (median) του polarity για κάθε κατηγορία.

    - Παρατηρούμε ότι οι περισσότερες κατηγορίες έχουν θετική διάμεσο, υποδεικνύοντας γενική ικανοποίηση στις κριτικές.
2. **Διασπορά (Spread):**
    - Το ύψος των κουτιών και των whiskers δείχνει την ποικιλία των συναισθημάτων για κάθε κατηγορία.
    - Κατηγορίες όπως το **"Biography & Autobiography"** και το **"Self-Help"** έχουν μεγαλύτερη διασπορά, υποδεικνύοντας έντονες διαφορές στις απόψεις.
3. **Ακραίες Τιμές (Outliers):**
    - Παρατηρούνται πολλές ακραίες τιμές σε όλες τις κατηγορίες, κυρίως σε αρνητικές τιμές polarity.
    - Αυτό δείχνει ότι, παρόλο που οι περισσότερες κριτικές είναι θετικές, υπάρχουν βιβλία που απογοητεύουν τους αναγνώστες.
4. **Συγκριτική Ανάλυση:**
    - Οι κατηγορίες όπως το **"Juvenile Fiction"** και το **"Fiction"** έχουν πιο θετική κατανομή.
    - Κατηγορίες όπως το **"Body, Mind & Spirit"** έχουν πιο ουδέτερη κατανομή, ενώ το **"Business & Economics"** εμφανίζει ποικιλία συναισθημάτων.

Προτάσεις για Εμπορική Χρήση

1. **Προώθηση:**
    - Κατηγορίες με υψηλό polarity (π.χ., "Juvenile Fiction", "Fiction") μπορούν να χρησιμοποιηθούν σε καμπάνιες μάρκετινγκ για να ενισχύσουν τις πωλήσεις.
2. **Βελτίωση Προϊόντων:**
    - Οι εκδότες μπορούν να επικεντρωθούν σε κατηγορίες με έντονη αρνητική διασπορά (π.χ., "Biography & Autobiography") για να βελτιώσουν το περιεχόμενό τους.
3. **Ανάλυση Ακραίων Τιμών:**
    - Τα βιβλία με έντονα αρνητικά συναισθήματα (outliers) μπορούν να μελετηθούν για να εντοπιστούν κοινά προβλήματα.

2.4 Αποτελέσματα

- **Θετικά Συναισθήματα:** 60% των κριτικών.
- **Αρνητικά Συναισθήματα:** 30% των κριτικών.
- **Ουδέτερα Συναισθήματα:** 10% των κριτικών.

2.5 Συμπεράσματα

1. Οι περισσότερες κριτικές εκφράζουν θετικά συναισθήματα, κάτι που αντικατοπτρίζει τη γενική ικανοποίηση των χρηστών από τα βιβλία.

2. Οι αρνητικές κριτικές εντοπίζονται κυρίως σε συγκεκριμένα βιβλία με χαμηλές βαθμολογίες.
3. Η κατηγοριοποίηση ανά συναίσθημα μπορεί να χρησιμοποιηθεί για:
    - Βελτίωση των προϊόντων.
    - Στοχευμένη προώθηση βιβλίων με υψηλά θετικά συναισθήματα.

Συνολικά Συμπεράσματα

- Οι **Κανόνες Συσχέτισης** ανέδειξαν σημαντικές σχέσεις μεταξύ κατηγοριών βιβλίων και βαθμολογιών, που μπορούν να αξιοποιηθούν για συστάσεις.
- Η **Ανάλυση Συναισθήματος** παρείχε χρήσιμες πληροφορίες για τη συναισθηματική ανταπόκριση των χρηστών.
- Οι τεχνικές εξόρυξης δεδομένων αποτελούν εργαλεία στρατηγικής σημασίας για τη λήψη επιχειρηματικών αποφάσεων.
