Szkoła Główna Handlowa

Raport z analizy danych dotyczących książek

Karolina Gworek, numer albumu 132155

Projekt Zaliczeniowy na przedmiot

Prezentacja i Wizualizacja Danych

Warszawa, Styczeń 2024

1. Źródło oraz przygotowanie bazy danych

Celem niniejszego raportu jest weryfikacja czynników, które wpływają na odbiór przez czytelników. Wykorzystano w nim dwie bazy danych, które zostały połączone w celu uzyskania jak największej ilości informacji o poszczególnych tytułach książek. Pierwszą z nich jest *7k Books*. Zawierała ona informacje takie jak: identyfikatory poszczególnych książek (isbn13 oraz isbn10), tytuł, autor, opis, średnia ocena na portalu GoodReads, kategoria książki oraz liczba ocen. Użyto także bazy *Goodreads-books,* aby dodatkowo uzyskać informacje na temat wydawnictwa, liczby stron, daty wydania oraz liczby recenzji na portalu GoodReads. Bazy zostały one połączone za pomocą isbn13 - Międzynarodowy Standardowy Numer Książki, czyli unikatowego identyfikatora książek i innych publikacji o charakterze książkowym. Połączona baza danych zawierała 5688 obserwacji. Dodatkowo, pierwotnie zmienna opisująca kategorię książek była nieuporządkowana i zawierała aż 407 różnych wartości, z czego wiele z nich wykazywało podobieństwa. Dokonano mapowania tych kategorii do nowej zmiennej o nazwanej "tematyka", która zawiera 15 uporządkowane wartości, łączące podobne kategorie książek w spójne grupy. Zmapowanie 407 różnych kategorii do 15 umożliwiło analizę poszczególnych informacji w zależności od tematyki książki. Kilka tytułów książek okazało się mieć duplikaty, było to związane z kolejnymi edycjami wydawanymi przez np. różne wydawnictwa. W celu analizy posortowano dane i usunięto duplikaty mniej popularnej wersji danego tytułu, biorąc pod uwagę liczbę ocen na portalu GoodReads. Pozwoliło to na skupienie się na tytułach, które stanowią bardziej reprezentatywne przykłady dla danej książki. Dodatkowo zbiór został ograniczony do obserwacji, które zawierały więcej niż 10 ocen. Ostateczna baza danych zawierała 5319 obserwacji opisanych przez osiem zmiennych. Spośród nich dwie to zmienne charakterystyczne (*Title, Authors)*, cztery to zmienne ilościowe (*Average_Rating*, *Ratings_Count, Publication_Date, Num_Pages),* a dwie to zmienne jakościowe (Publisher, *Theme)*.

2. Cel badawczy analizy bazy danych książek

Analiza pozwoli na zbadanie związku między średnimi ocenami książek a innymi zmiennymi i pozwoli na określenie jaki czynnik w największym stopniu wpływa na odbiór książki przez czytelników. Dodatkowo przedstawione zostaną trendy czytelnicze: takie jak najpopularniejsze rodzaje książek czy wydawnictwa.

3. Opis zmiennych

- *Title - tytuł*

- *Authors – autor*

- *Average_Rating –średnia ocena*

- *Ratings_Count – liczba ocen*

- *Text_Reviews_Count – liczba napisanych recenzji*

- *Theme - tematyka*

- *Publisher – wydawca*

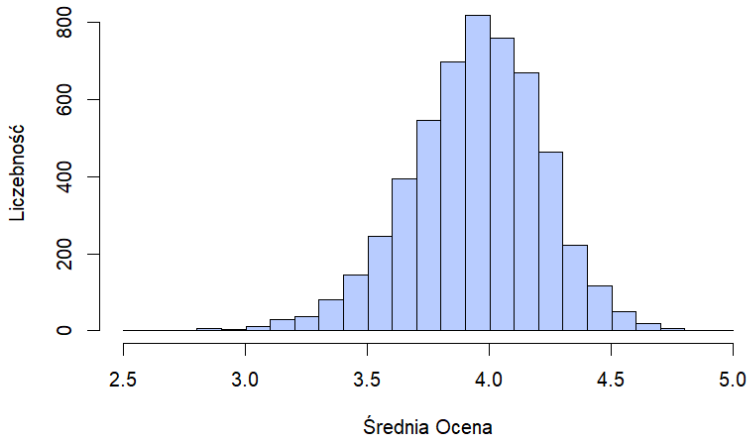
- *Publication_Date – data publikacji*

- *Theme – tematyka*

- *Num_Pages – liczba stron*

Średnia ocen (*Average_Rating)* na portalu GoodReads, jednej z najpopularniejszych platformie pozwalającej czytelnikom dzielić się swoimi opiniami na temat książek, recenzjami oraz ocenami. Zmienna ta może posłużyć do zbadania co wpływa na ocenę poszczególnych tytułów. Przyjmuje ona wartości od 0 do 5, gdzie 5 oznacza ocenę najwyższą. Najczęściej występujące wartości oscylują wokół 4.

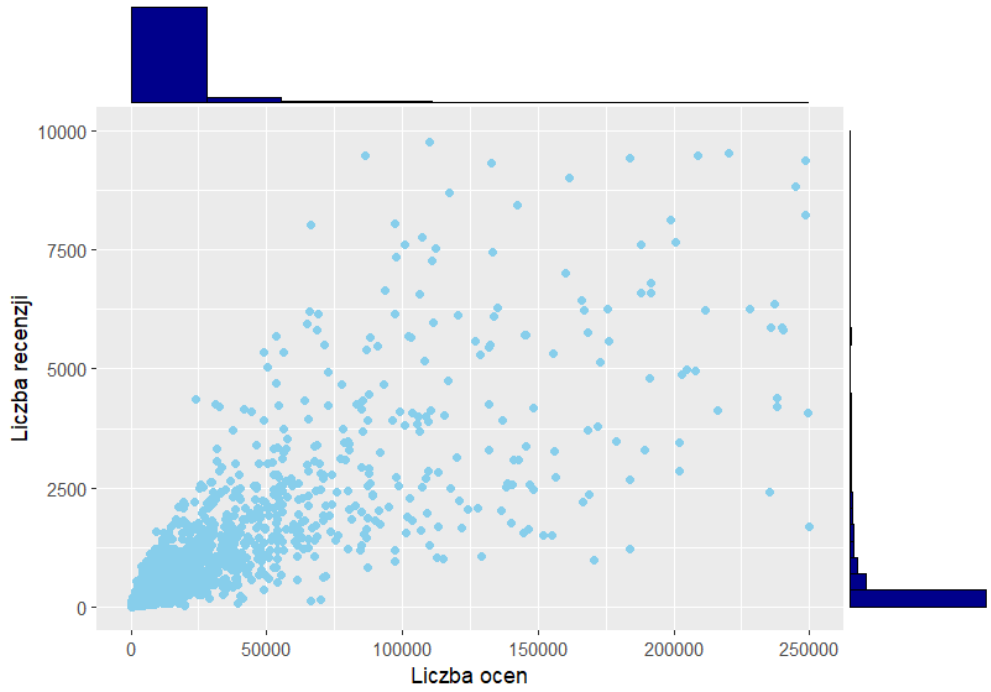
Rys. 1. Rozkład *Average_Rating*



Źródło: opracowanie własne

Zmienna *Ratings_Count* oraz *Text_Reviews_Count* są dodatnio skorelowane. Warto zauważyć, że więcej jest dodanych ocen niż recenzji tekstowych. Dodatkowo rozkład obu zmiennych koncentruje się wokół mniejszych wartości i ma charakter malejący. Oznacza to, że duża liczba ocen oraz recenzji dodana jest dla niewielu tytułów, co widać na przedstawionych histogramach (Rys. 2).

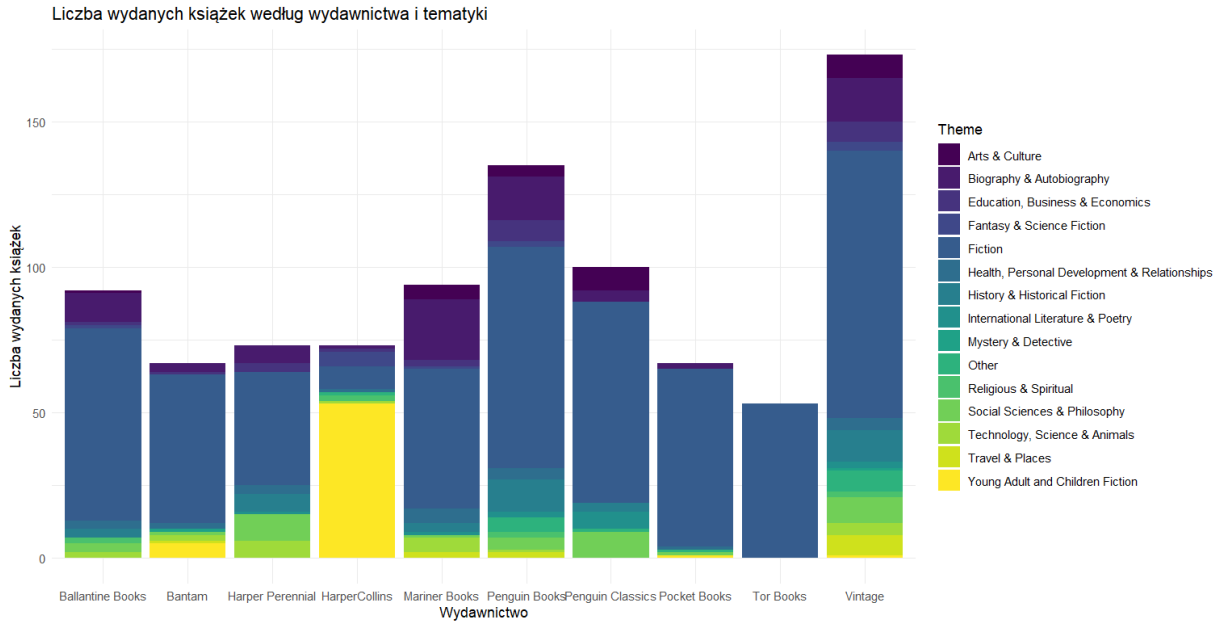
Rys. 2. Ratings_Count oraz Text_Reviews_Count



Źródło: opracowanie własne

W bazie danych istnieje ponad 1300 wydawnictw, te które wydały najwięcej książek (ponad 50) zaprezentowano na Rys. 3. Zmienna *Theme* – gatunek/tematyka książki, określa zakres tematyczny książki, kategorie książek z bazy zostały podzielone na 15 zakresy tematyczne. Na podstawie zaprezentowanego wykresu (rys.3.) można stwierdzić, że większość wydawnictw publikuje głównie fikcje (Tor Books wyłącznie fikcje), wydawnictwo HarperCollins skupia się na wydawaniu fikcji przeznaczonej dla dzieci i młodzieży. Dodatkowo duży odsetek publikacji wydawnictwa Marnier Books stanowią biografie i autobiografie. Największą różnorodność gatunkową widzimy w publikacjach wydawnictwa Penguin Books. Dashboard pozwala na śledzenie liczby publikacji wydawnictw na przestrzeni lat.

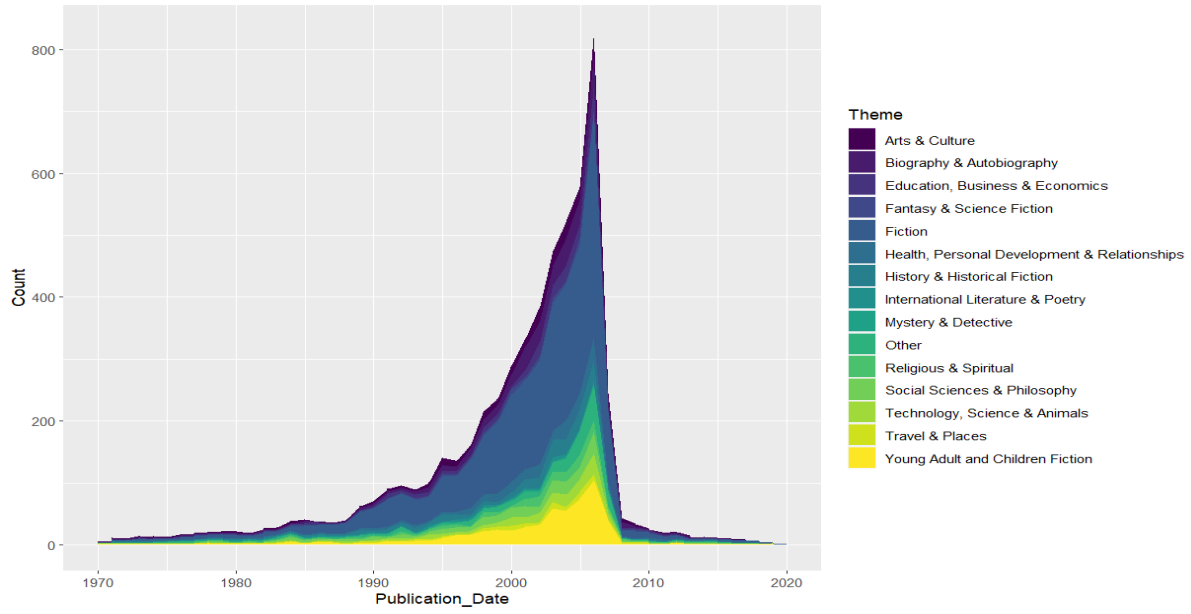
Rys. 3. Wydawnictwa książek z podziałem na tematykę publikacji



Źródło: opracowanie własne

Najwcześniejszy zarejestrowany tytuł pochodzi z 1913 roku, natomiast najnowszy - z 2020 roku. Obserwuje się istotny wzrost liczby wydanych książek od 1990 roku do 2008 roku, szczególnie książek w kategorii fikcja. Z roku 2008 pochodzi największa liczba tytułów w bazie danych. Od tego czasu zaobserwowano wyraźny spadek liczby wydanych książek uwzględnionych w bazie.

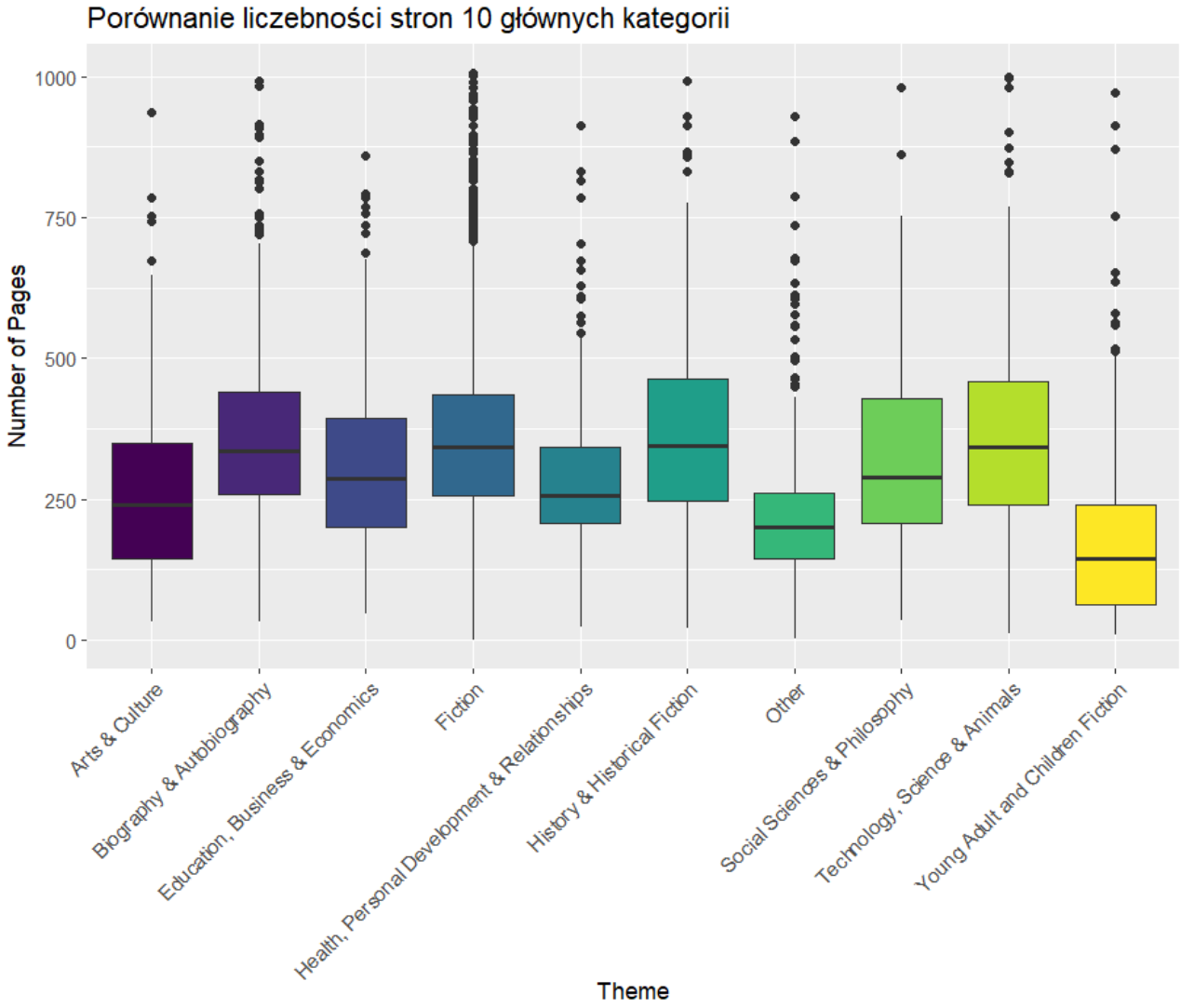
Rys. 4. Liczba publikacji na przestrzeni lat



Źródło: opracowanie własne

Zmienna opisujące liczbę stron (*Num_Pages)* zawiera wiele obserwacji odstających, ze względu na to dane zostały dodatkowo przefiltrowane. Zmienne których wartość odbiegały od średniej o więcej niż +- 3 * błąd standardowy zostały usunięte. Pozwoliło to na pokazanie, że liczba stron różni się w zależności od kategorii książki. O ile książki reprezentujące fikcje oraz biografie i autobiografie mają podobną liczebność stron i ich średnia wynosi około 350, to średnia liczba stron dla książek przeznaczonych dla dzieci i młodzieży jest znacznie mniejsza i wynosi około 130 stron. Pomiędzy tymi dwoma kategorii pod względem liczebności znajdują się książki psychologiczne i filozoficzne oraz o sztuce i kulturze.

Rys.5. Porównanie liczebności stron dla 10 najliczniejszych kategorii tematycznych

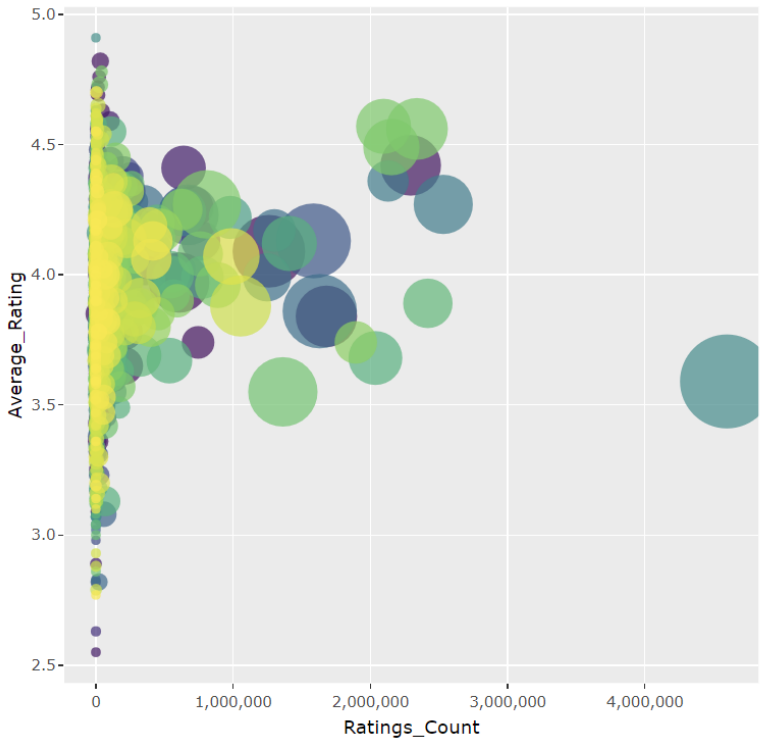


Źródło: Opracowanie własne

4. Pozostałe zależności między zmiennymi

Rys. 6 prezentuje zależności między dwoma zmiennymi ilościowymi *Average_Rating* oraz *Ratings_Count*. Można zaobserwować koncentracje obserwacji z małymi liczbami ocen, przyjmują one różne wartości średniej oceny w zakresie od 2,5 do blisko 5. Książki zawierające dużą liczbę ocen najczęściej przyjmują wartości średniej oceny od 3,5 do 4. Warto jednak zwrócić uwagę, że są one bardziej miarodajnymi obserwacjami, gdyż opinia na ich temat została wyrażona przez większą liczbę czytelników. Wielkość punktu odpowiada liczbie recenzji napisanych przez czytelników. Większej liczbie ocen towarzyszy również większa liczba recenzji, ponieważ punkty są proporcjonalne do wartości osi x. Kolory reprezentują różne wydawnictwa.

Rys. 6. Zależność między średnią ocen, liczbą ocen, liczbą recenzji

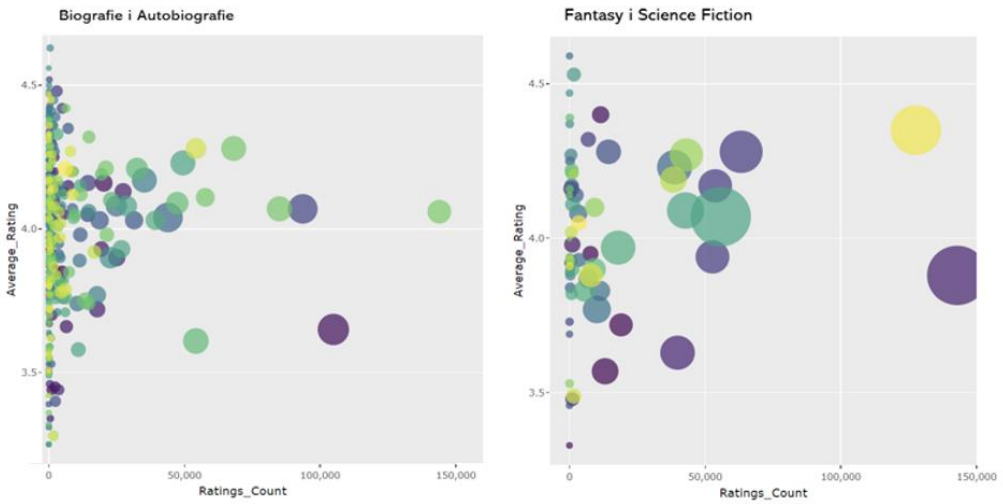


Źródło: opracowanie własne

Dashboard umożliwia obserwację zależności między liczbą recenzji, a średnią ocen w zależności od kategorii książki. Na tej podstawie można stwierdzić, że niezależnie od tematyki książki zależność ta jest podobna. Dla przykładu, porównano dwe znacząco różniące się tematyki książki: biografie i autobiografie z fantasy i science fiction. Obie kategorie

charakteryzują się dużą różnorodnością w średniej ocenie dla małej liczby recenzji i średnią oceną około 4,2 dla większej liczby recenzji. Tematyka książki wpływa jednak na liczbę ocen i liczbę recenzji. Książki fantasy i science fiction są chętniej recenzowane niż biografie i autobiografie.

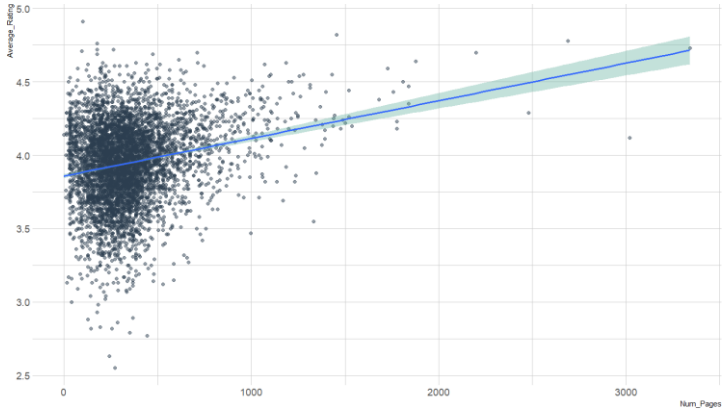
Rys. 7. Zależność między średnią ocen, liczbą ocen, liczbą recenzji dla poszczególnych kategorii



Źródło: opracowanie własne

Następnie przedstawiono wpływ liczby stron w książce na jej średnią ocenę. Wyraźna jest dodatnia zależność między tymi zmiennymi - książki o większej liczbie stron są pozytywniej oceniane przez czytelników. Dodatkowo, odchylenie standardowe dla modelu liniowego tej zależności jest niewielkie, co wskazuje na spójność danych.

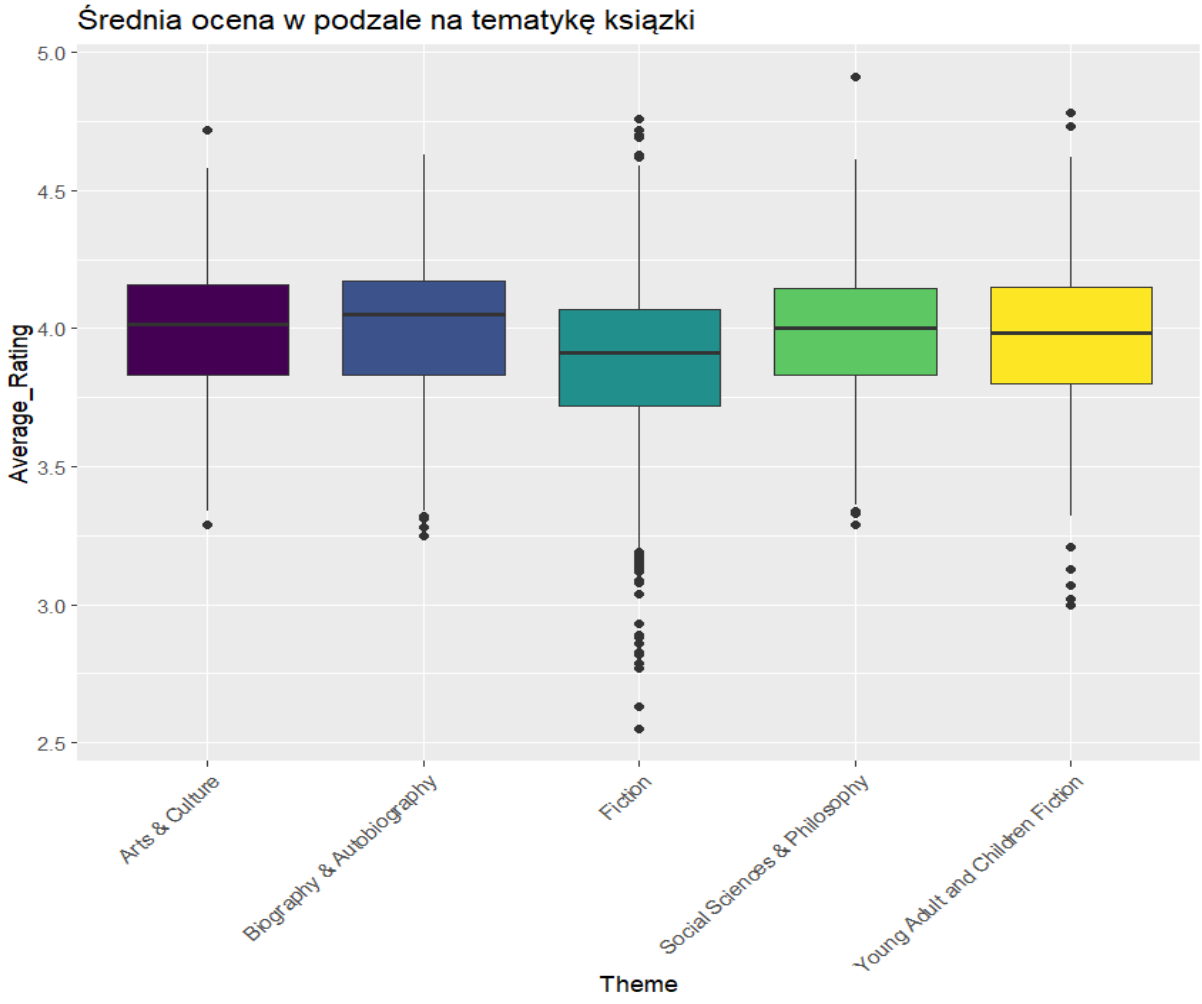
Rys. 8. Zależność między średnią ocen a liczbą stron



Źródło: opracowanie własne

Ostatnią zależnością na którą warto zwrócić uwagę jest wpływ tematyki książki na jej średnią ocenę. Wykresy pudełkowe dla wszystkich kategorii pokazują jej niewielki wpływ, ponieważ mediana każdej grupy waha się pomiędzy 3,9, a 4,1. Najniższą medianą ocen charakteryzują się książki z kategorii fikcja, natomiast biografie i autobiografie. Jednak różnice również nie są znaczące, można zatem wywnioskować, że tematyka książek znacząco nie wpływa na średnią ocenę.

Rys. 9. Średnia ocena w podziale na tematykę książki

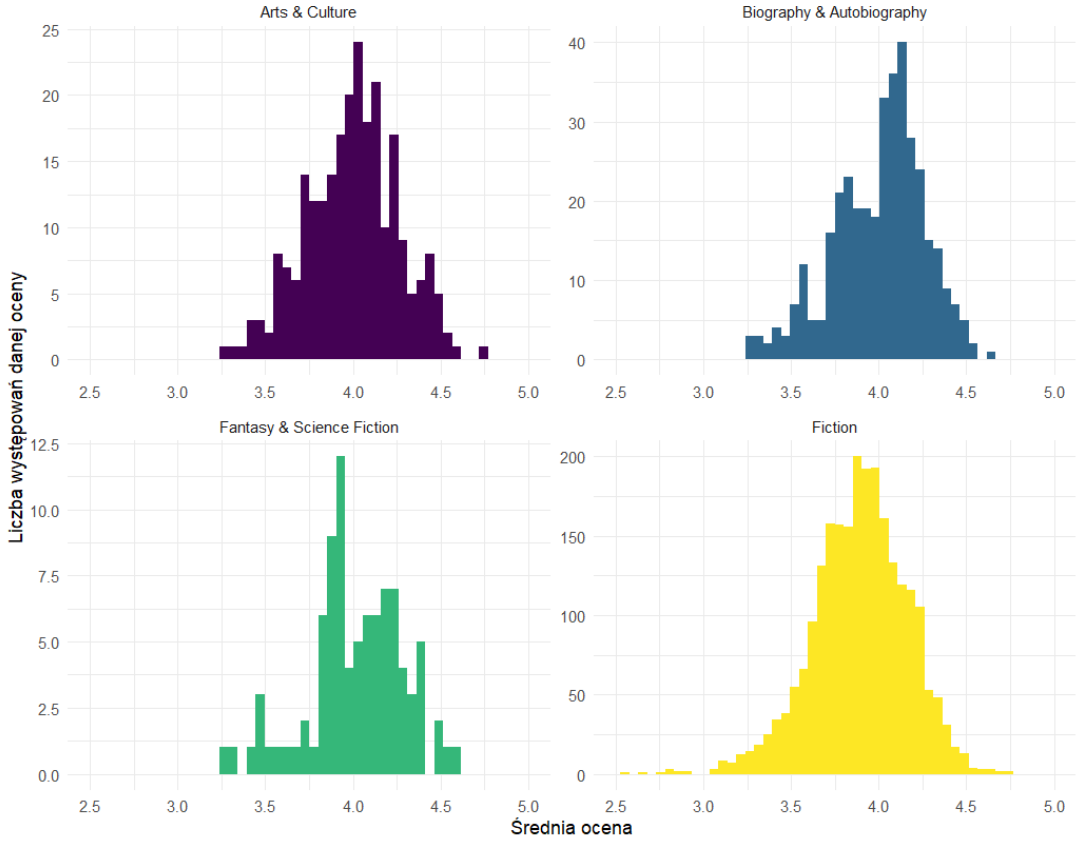


Źródło: opracowanie własne

Sprawdzono również jak rozkłada się średnia ocen w podziale na różne kategorie, potwierdziło to tezę iż tematyka książki nie wpływa znacząco na jej ocenę. Przy większej liczbie ocen rozkład oscyluje wokół 4, co widać w przypadku fikcji. Można również stwierdzić

dla trzech z czterech analizowanych poniżej kategorii więcej razy występuje ocena większa niż 4, wyjątek stanowi fikcja.

Rys. 10. Rozkład średniej ocen dla poszczególnej tematyki książki



Źródło: opracowanie własne

5. Wnioski

Do najważniejszych wniosków z analizy tytułów w stworzonej bazie danych należy fakt, iż liczba dodanych ocen i recenzji są dodatnio skorelowane przy czym obie zmienne mają charakter malejący. Liczba dodanych ocen dla większości tytułów nie przekracza 25000, a liczba recenzji 2000. Wydawnictwa różnią się między sobą tematyką wydawania książek przy czym większość z nich skupia się na publikacji fikcji. Tematyka książki wpływa na medianę liczby jej stron.

Istotnym czynnikiem wpływającym na ogólną średnią jest liczbę przyznanych ocen. Dla książek, które zdobyły niewiele ocen i recenzji, mediana często oscyluje między skrajnymi

wartościami. Z kolei w przypadku dużej liczby ocen, mediana ustala się w zakresie od 3,5 do 4,5. Kolejną zmienną wpływającą na średnią ocenę jest liczba stron. Pozytywniej oceniane są obszerniejsze publikacje. Tematyka książki nie ma znaczącego wpływu na średnią ocenę. Jednak widzimy niewielkie różnice między poszczególnymi kategoriami np. fikcja zazwyczaj otrzymują niższe oceny w porównaniu z biografiami i autobiografiami.

6. Spis rysunków

7. Źródła

Baza danych 7k Books, https://www.kaggle.com/datasets/dylanjcastillo/7k-books-with-metadata, dostęp 27.12.2023

Baza danych GoodReads, https://www.kaggle.com/datasets/jealousleopard/goodreadsbooks, ostęp 27.12.2023

Utworzona baza danych, https://drive.google.com/drive/u/0/folders/11hpd1HcSR0SnBhvPO2Nqj3ouoGUua8x9, dostęp 07.01.2023