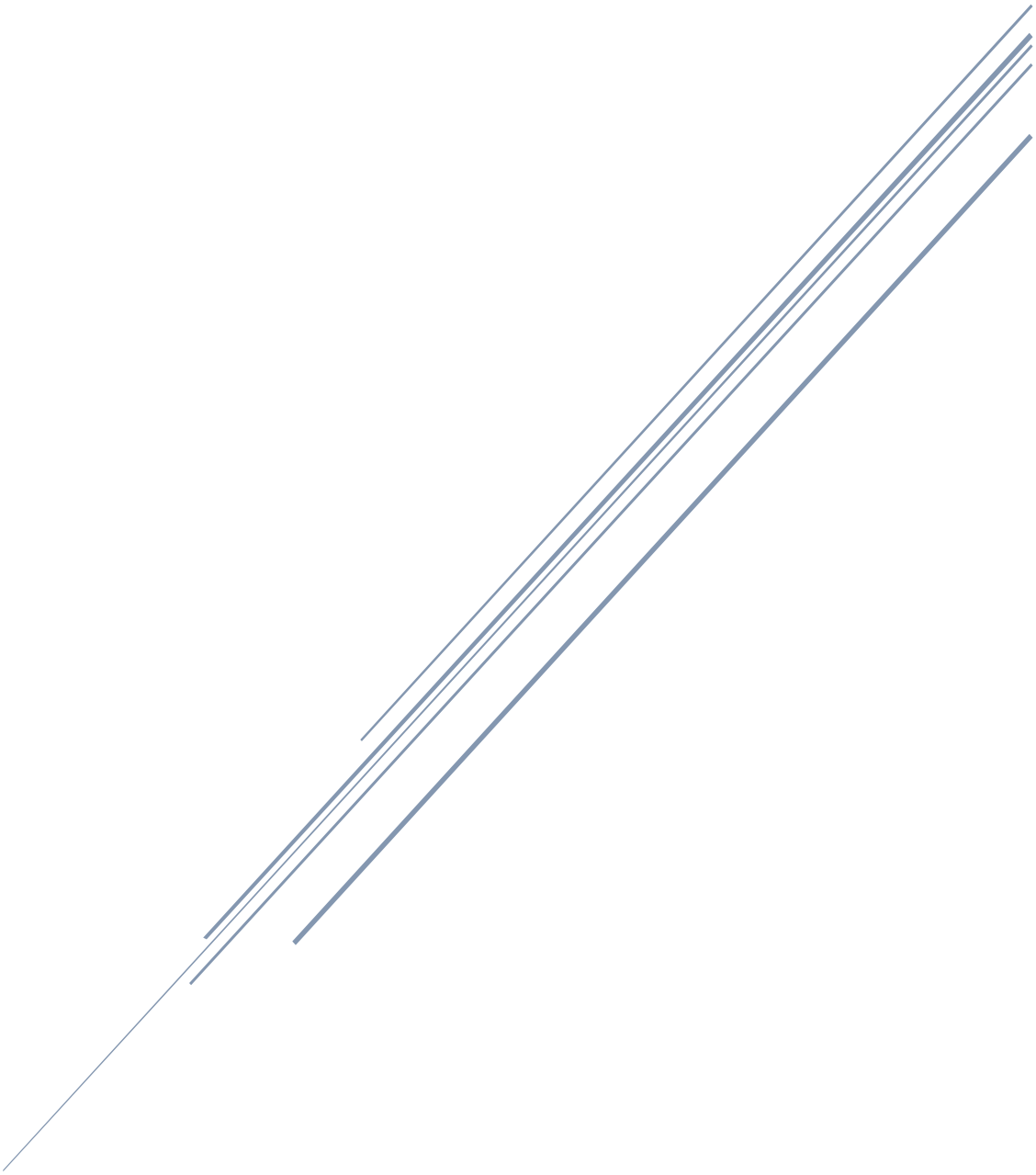
RAPORT

Daria Rybak

Numer albumu: 472900

2024



Wprowadzenie

Projekt miał za cel zbadanie czynników od których zależą dobre recenzje lub popularność książek. Najważniejszym zmienną okazała się być ‘language code’ w której była informacja o języku w którym była napisana książka. Podziały w niej dokonane podczas analizy były niezbędne.

Opis danych

Dane projektu zostały pobrane ze strony https://www.kaggle.com/. Były one zebrane przez użytkownika Soumik i opublikowane 5 lat temu. Baza opiera się na danych pobranych ze strony goodreads (na ten moment najpopularniejszej angielskiej strony do recenzji książek). Zawiera ona około 10 000 próbek i posiada wartości:

- numer identyfikacyjny
- tytuł książki
- autor
- średnią ocenę
- dwa numery ISBN
- język książki
- liczbę stron
- liczbę ocen
- liczbę recenzji pisemnych

Dane podkreślone wykorzystałam w raporcie.

Pytania badawcze

Na potrzeby poszczególnych pytań skategoryzowałam niektóre dane. Na podstawie tego zestawu danych stworzyłam pięć pytań badawczych, które bezpośrednio odnoszą się do tego zestawu.

1. **Które książki cieszą się największą popularnością – romańskie, germańskie czy azjatyckie?**
   Przypasowałam interesujące mnie języki do grup językowych: romańskie, germańskie, azjatyckie.

2. **Czy wraz z ilością stron oceny się pogarszają?**

3. **Czy książki romańskie otrzymujące więcej recenzji pisemnych od germańskich?**
   W analizie tego pytania był wykorzystany podział z pyt. pierwszego.

4. **Czy książki wykorzystujące inny alfabet niż łaciński są dłuższe?**
   Dane języków zostały skategoryzowane ze względu na użyty alfabet: „latin" i „non-latin".

5. **Czy książki wykorzystujące wymarłe języki lub dialekty są dobrze czy źle oceniane?**
   Języki wymarłe i dialekty zostały przyporządkowane do jednej zmiennej „dead". Wybrałam także parę języków powszechnie używanych i umieściłam w grupie „alive". Średnie ocen powyższych grup zostały podzielone następująco: te gorzej oceniane (poniżej średniej 4,0) i lepiej oceniane (powyżej średniej 4,0).

Poniższy raport przedstawia hipotezy, analizę oraz wnioski, które dotyczą wyżej postawionych pytań.

Wyniki

1. Które książki cieszą się największą popularnością – romańskie, germańskie czy azjatyckie?

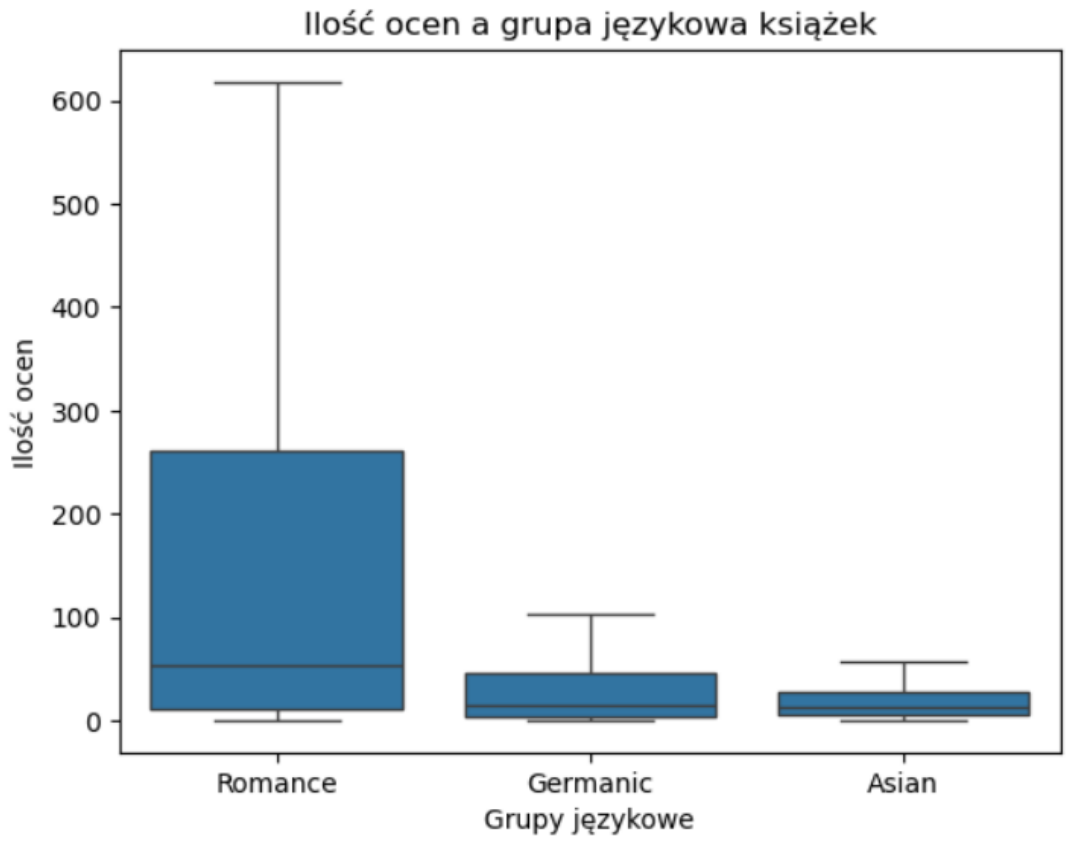
Pierwszą zmienną są książki odpowiednio skategoryzowane wcześniej na grupy: romańska, germańska, azjatycka. Jest to zmienna nominalna. Drugą zmienną jest ilość wystawionych ocen w stosunku do każdej grupy, jest to zmienna ilościowa.

Na początku w każdej grupie na ilości ocen był zastosowany test Shapiro-Wilka, aby zbadać, czy zebrane wyniki posiadają rozkład normalny. Wszystkie wyniki p były poniżej założonego p = 0,05. Rozkłady nie są rozłożone normalnie.

H_0 – Brak monotonicznego związku między grupą językową książki a jej popularnością

H_1 – Istnieje związek monotoniczny między grupą językową książki a jej popularnością

Należało zastosować test nieparametryczny Kruskala-Wallisa, którego p także wynosiło mniej niż 0,05. Test był istotny statystycznie, więc odrzuciłam hipotezę zerową.



Wykres 1. Ilość ocen a grupa językowa książek (wartości odstające zostały usunięte dla wizualizacji).

Do zobrazowania wyniku wykorzystałam wykres pudełkowy. Jak widać na Wykresie 1. grupa języków romańskich ma największą liczbę ocen. Rozmiar pudełka obrazuje duże

zróżnicowanie tej liczby między poszczególnymi książkami. O wiele mniejszymi grupami są książki z grupy języków germańskich i azjatyckich, a ich węższe pudełka sugerują mniejszą zmienność między liczbą ocen wśród poszczególnych książek. Patrząc na medianę bliżej dołu pudełka można wywnioskować, że wśród książek języków romańskich większość książek ma stosunkowo mało ocen, a tylko nieliczne mają ich bardzo dużo. Mediany języków germańskich i azjatyckich, chociaż też wskazują na większość książek z mniejszą ilością ocen, to są bardziej wypośrodkowane w stosunku do języków romańskich.

Związek grupy językowej i liczby ocen między grupą języków romańskich a językami germańskimi i azjatyckimi jest silny. Grupa romańska wyróżnia się większą ilością ocen i większym rozrzutem. Związek grupy językowej i liczby ocen między grupą germańską a azjatycką jest słaby, jest stosunkowo mało różnic między nimi.

Wniosek: Popularność książki może faktycznie zależeć od języka w jakim została napisana. Języki romańskie mogą mieć o wiele większy zasięg marketingowy na inne kraje poprzez ilość i/lub jakość tłumaczeń książek. Języki germańskie i azjatyckie mogą reprezentować bardziej niszowe lub homogeniczne rynki.

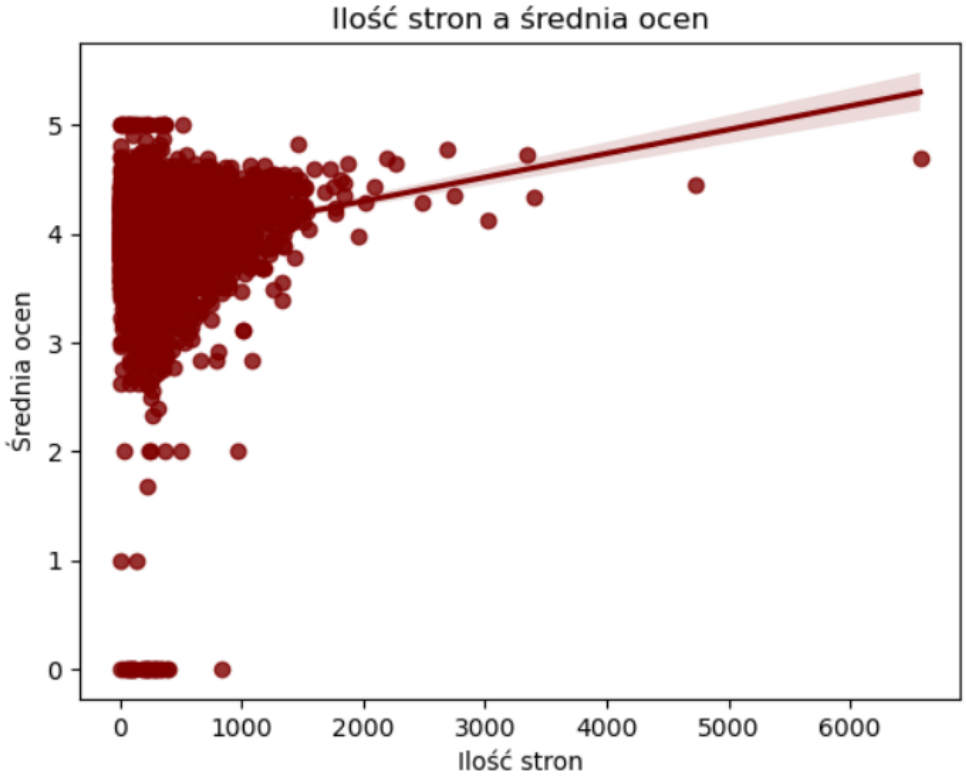
2. Czy wraz z ilością stron oceny się pogarszają?

Obie zmienne wykorzystane tutaj są zmiennymi ilościowymi. Na początku był wykorzystany Test Kołmogorowa-Smirnowa, aby zbadać, czy zebrane wyniki posiadają rozkład normalny. Oba wyszły statystycznie istotne ($p<0,05$, $p=0$), z czego zmienna „średnia ocen" wskazywała na bardzo duże odchylenie od teoretycznego rozkładu.

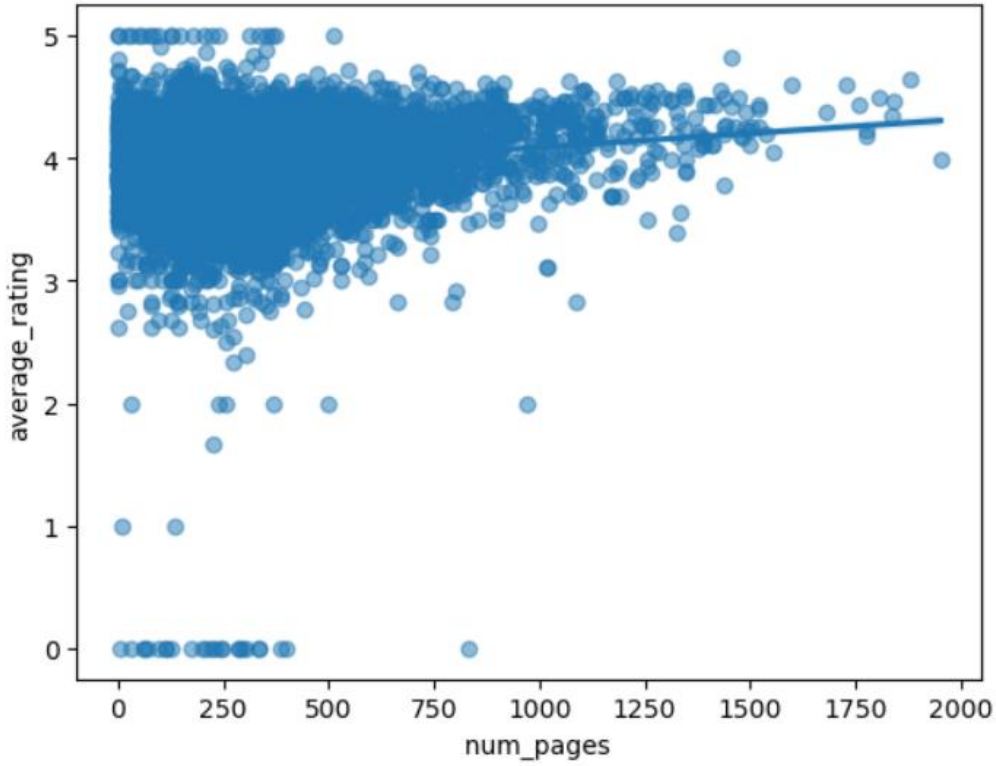
H_0 – Brak monotonicznego związku między ilością stron a średnią ocen książki

H_1 – Istnieje związek monotoniczny między ilością stron a średnią ocen książki

Należało zastosować test nieparametryczny Spearmana. Wartość p jest bardzo mała i przez to bardzo statystycznie istotna, mogliśmy odrzucić hipotezę zerową. Współczynnik Spearmana wyniósł 0,1, co oznacza bardzo słabą dodatnią zależność między zmiennymi.



Wykres 2. Ilość stron a średnia ocen



Wykres 3. Ilość stron a średnia ocen (wartości odstające zostały usunięte dla wizualizacji).

Zastosowałam wykres punktowy. Na Wykresie 2. znajdowało się parę wartości odstających, które zaburzały linię regresji i czytelność, więc wykonałam drugi wykres już bez tych wartości.

Patrząc na Wykres 3. można zobaczyć, że linia regresji w stosunku do głównej części próbek jest prawie płaska, co wskazuje na małą siłę związku między ilością stron a średnią oceną książki. Znaczna większość książek mieści się w wymiarze do 500 stron i posiada różne wyniki (3-4,5). Wysokie oceny (w przedziale 4-4,5) są obecne niezależnie od długości stron.

Wniosek: Nie ma istotnego wpływu ilość stron na średnią ocen książki. Czytelnicy zwracają większą uwagę na faktyczną treść książki, a nie tylko jej wymiary.

3. Czy książki romańskie otrzymują więcej recenzji pisemnych od germańskich?

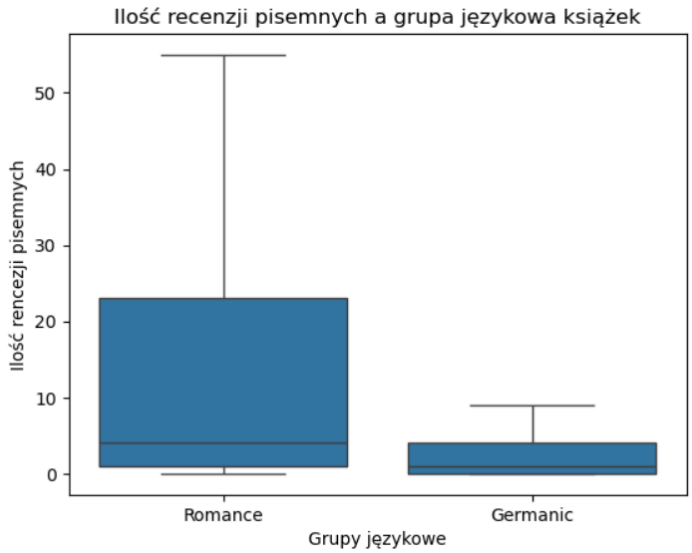
Pierwszą zmienną są książki odpowiednio skategoryzowane wcześniej na grupy: romańska i germańska. Jest to zmienna nominalna. Drugą zmienną jest ilość recenzji pisemnych, jest to zmienna ilościowa.

Na początku w każdej grupie na ilości recenzji pisemnych był zastosowany test Shapiro-Wilka, aby zbadać, czy zebrane wyniki posiadają rozkład normalny. W obu grupach wartość p była znacznie poniżej obranego 0,05. Wyniki nie posiadały rozkładu normalnego.

H_0 – Brak monotonicznego związku między książkami z grupy romańskiej i germańskiej w kontekście liczby recenzji pisemnych

H_1 – Istnieje związek monotoniczny między książkami z grupy romańskiej i germańskiej w kontekście liczby recenzji pisemnych

Należało zastosować test nieparametryczny Manna-Whitneya. Wartość p testu jest znacząco mniejsza od założonego 0,05, co informowało o statystycznie bardzo ważnym wyniku. Mogliśmy odrzucić hipotezę zerową.



Wykres 4. Ilość recenzji pisemnych a grupa językowa książek (wartości odstające zostały usunięte dla wizualizacji)

Bazując na maksymalnych wartościach na Wykresie 4. można wywnioskować, że książki romańskie mogą otrzymywać o wiele więcej recenzji pisemnych na książkę niż germańskie. Minimalna w obu w grupach jest bardzo bliska zeru, więc brak recenzji na książkę nie jest podyktowany grupą językową. W obu grupach są odchylenia od wartości centralnych, ale w grupie romańskiej sięgają większych wartości, co informuje o większym rozrzucie danych. Mediana grupy romańskiej jest wyższa od mediany grupy germańskiej, co oznacza, że większość książek romańskich otrzymuje więcej recenzji pisemnych niż większość grupy germańskiej.
Związek nie wydaje się silny ze względu na częściowe pokrywanie się wykresów i przez rozproszenie danych (książki romańskie mają zarówno wysokie jak i niskie wyniki, a germańskie są bardziej skoncentrowane).

Wniosek: Książki z grupy językowej romańskie są częściej recenzowane pisemne niż germańskie. Ta różnica może być podyktowana lepszym marketingiem na inne kraje, preferencjami czytelników lub różnic kulturowych.

4. Czy książki wykorzystujące inny alfabet niż łaciński są dłuższe?

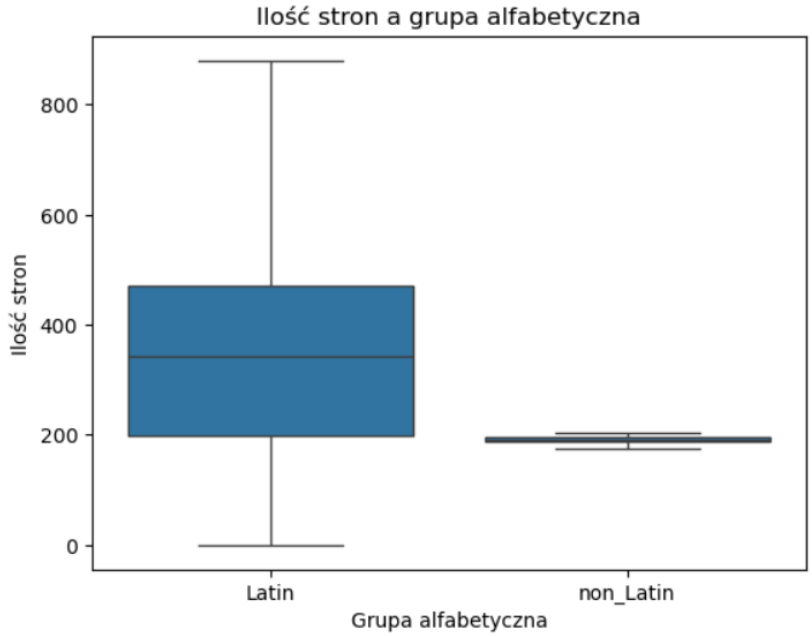
Pierwszą zmienną są książki odpowiednio skategoryzowane wcześniej na grupy: łacińskie i niełacińskie. Jest to zmienna nominalna. Drugą zmienną jest ilość stron w stosunku do każdej grupy, jest to zmienna ilościowa.

Na początku w każdej grupie na ilości ocen był zastosowany test Shapiro-Wilka, aby zbadać, czy zebrane wyniki posiadają rozkład normalny. W obu grupach wartość p była znacznie poniżej obranego 0,05. Wyniki nie posiadały rozkładu normalnego.

H_0 – Brak monotonicznego związku między wykorzystaniem danego alfabetu a długością książki

H_1 – Istnieje związek monotoniczny między wykorzystaniem danego alfabetu a długością książki

Należało zastosować test nieparametryczny Manna-Whitneya. Wartość p testu jest znacząco mniejsza od założonego 0,05, co informowało o statystycznie bardzo ważnym wyniku. Mogliśmy odrzucić hipotezę zerową.



Wykres 5. Ilość stron a grupa alfabetyczna książek (wartości odstające zostały usunięte dla wizualizacji)

Bazując na Wykresie 5. można zauważyć, że mediana grupy łacińskiej jest o wiele wyższa od mediany grupy języków niełacińskich, co wskazuje na większą grubość książek łacińskich. Co ważniejsze, rozstęp rozpiętość grupy niełacińskiej na bardzo wąski przedział stron (ok. 150-200), podczas gdy w grupie łacińskiej jest on mocno zróżnicowany (0-800 stron).

Wniosek: Książki z grupy łacińskiej są bardziej zróżnicowane pod względem objętości od grupy niełacińskiej. Dodatkowo, książki z grupy niełacińskiej mają podobną długość i są stosunkowo krótkie. Może to wskazywać na to, że w państwach używających alfabetów niełacińskich panują inne przekonania na temat kultury czytania – ile powinna ona zajmować czasu w codziennym życiu lub nawet jakiej problematyki ma dotyczyć.

5. Czy książki wykorzystujące wymarłe języki lub dialekty są dobrze czy źle oceniane?

Pierwszą zmienną są książki odpowiednio skategoryzowane wcześniej na grupy: wymarłe (lub dialekty) i powszechnie używane. Jest to zmienna nominalna. Drugą zmienną jest podzielona na dwie grupy średnia ocen: na te lepsze (powyżej średniej 4.0) i te gorsze (poniżej średniej 4.0). Jest to także zmienna nominalna.

Tabela 1. Zestaw zmiennych wymarłe/używane języki i podział na gorsze i lepsze języki

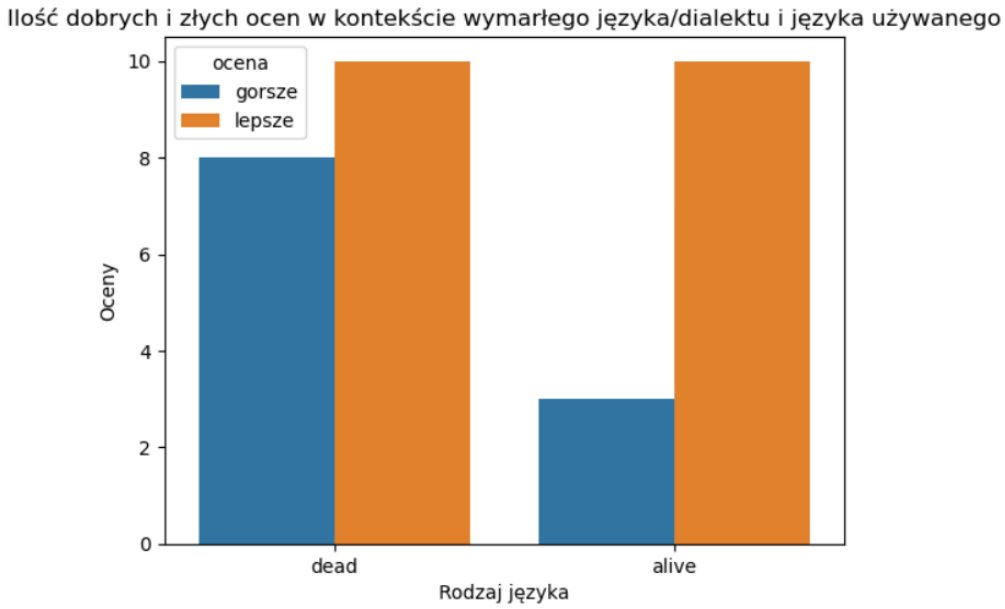
ocena	gorsze	lepsze
language_category_3		
alive	3	10
dead	8	10

Zastosowałam test Chi-square w celu sprawdzenia czy istnieje zależność pomiędzy tymi zmiennymi.

H_0 – Brak monotonicznego związku między wykorzystaniem języka „wymarłego" a jego oceną

H_1 – Istnieje związek monotoniczny między wykorzystaniem języka „wymarłego" a jego oceną

Wynik p (0,397) jest większy niż zakładany poziom 0,05, więc nie ma podstaw do odrzucenia hipotezy zerowej. Stopnie swobody wynoszą 1.



Wykres 6. Wykorzystany język a jego ocena

Na Wykresie 6. można zaobserwować rozkład książek ze względu na język. „Dead” oznacza języki wymarłe i dialekty, „alive” języki powszechnie używane.

Wniosek: Nie ma związku pomiędzy użyciem języka wymarłego w książce a jego oceną.

Dyskusja

Badanie książek pod względem używanego języka może okazać się przydatne w głębszej analizie rynku wydawniczego. Wyniki także podkreślają różnice w popularyzowaniu literatury między różnymi grupami językowymi. Ciekawe okazały się szczególnie wyniki książek łacińskich i niełacińskich, które wskazują na różnicę w podejściu do tematu czytelnictwa.