Úvod

- Moje semestrální práce se zaměřuje na to, co dělá knihu populární.
- Analyzovala jsem data z Goodreads a hledala souvislosti mezi vlastnostmi knihy a počtem pozitivních recenzí.
- Cílem bylo vytvořit predikční model popularity a identifikovat klíčové faktory.

Dataset

- Metadata knih (800 000+) a uživatelská hodnocení (350 000+).
- Obsahují autora, rok vydání, jazyk, počet stran a textová hodnocení.
- Tato kombinace umožňuje kvantifikovat popularitu knih z více pohledů.

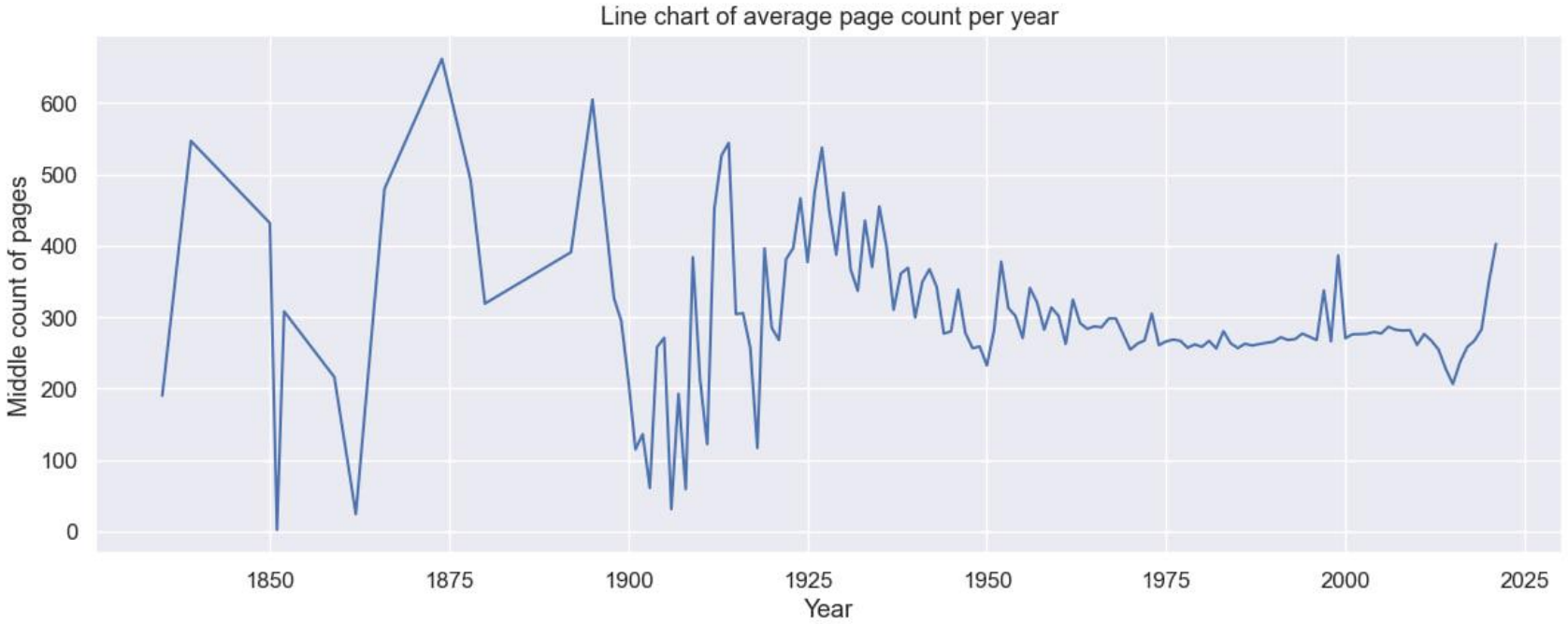
Čištění a zpracování

- Odstranění duplicit a prázdných hodnot.
- Normalizace jazyků a nakladatelství.
- Kategorizace délky knihy pomáhá odhalit vztahy bez vlivu extrémů.

```
2. Data enrichment and filtering:
- Missing values in 'Publisher' and 'Language' columns are filled with default labels.
- 'PublishMonthName' is derived from 'PublishMonth' using the calendar module.
- 'LengthCategory' is added based on the number of pages (short/medium/long).
- Rare languages (less than 500 books) are filtered out to focus on major languages.

First 5 rows of the dataframe:

   pagesNumber LengthCategory Language PublishMonthName
0        652.0           long      eng          Unknown
1        870.0           long      eng          January
2        309.0           long      eng          January
3        352.0           long      eng          January
4        435.0           long      eng          January

Most common languages in the dataset:

Language
Unknown Language    669928
eng                 119371
en-US                16399
fre                   6519
en-GB                 5035
Name: count, dtype: int64
```

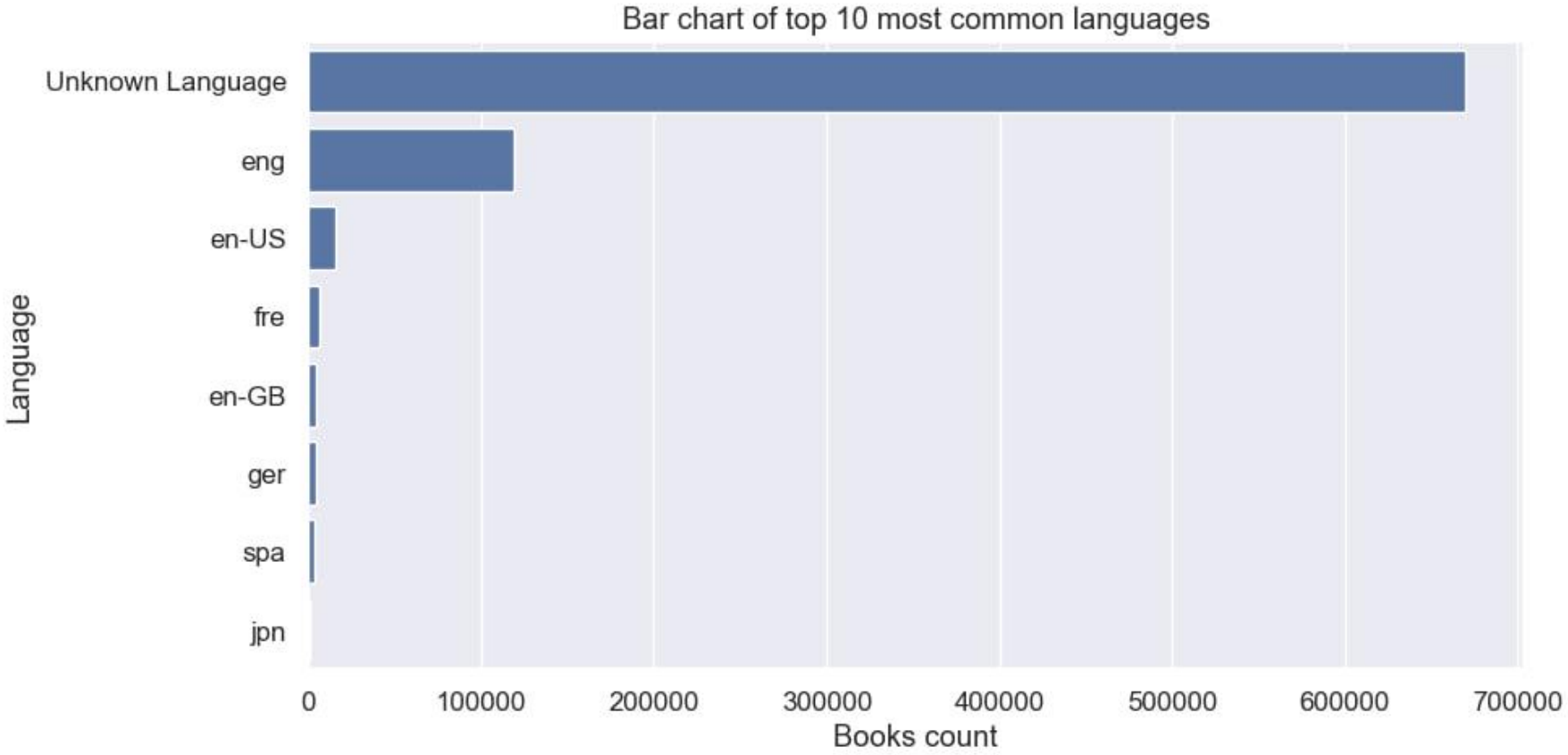
Vývoj počtu stran v čase

- Délka knih se v moderní době ustálila okolo 300 stran.
- Stabilita naznačuje standardizaci i preference čtenářů.


Line chart of average page count per year
Middle count of pages
0
100
200
300
400
500
600
1850
1875
1900
1925
1950
1975
2000
2025
Year

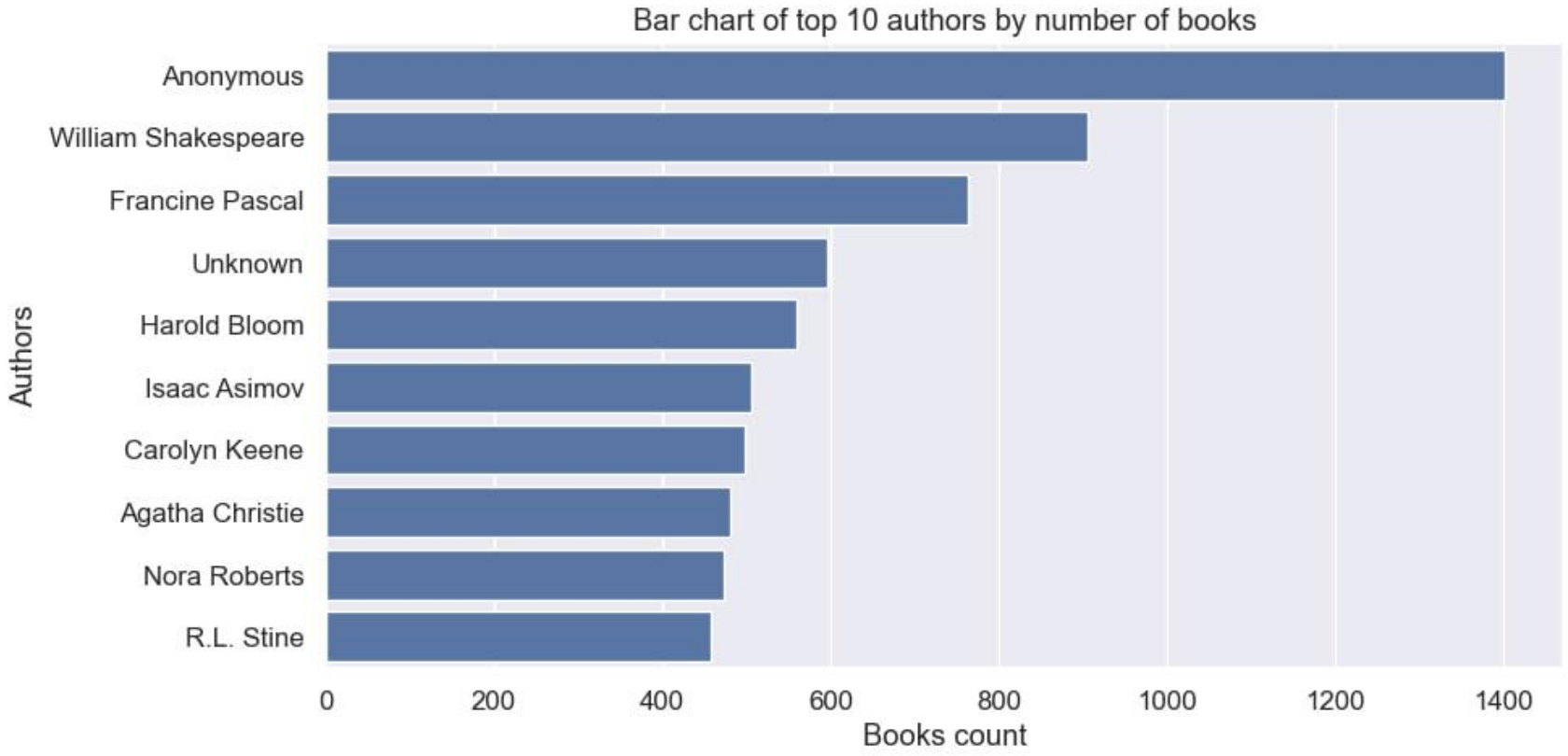
Jazyky knih

- Angličtina dominuje. Ostatní jazyky sloučeny do skupiny 'Unknown'.
- Zjednodušuje model a odstraňuje šum z málo častých jazyků.


Bar chart of top 10 most common languages
Unknown Language
eng
en-US
fre
en-GB
ger
spa
jpn
Language
0
100000
200000
300000
400000
500000
600000
700000
Books count

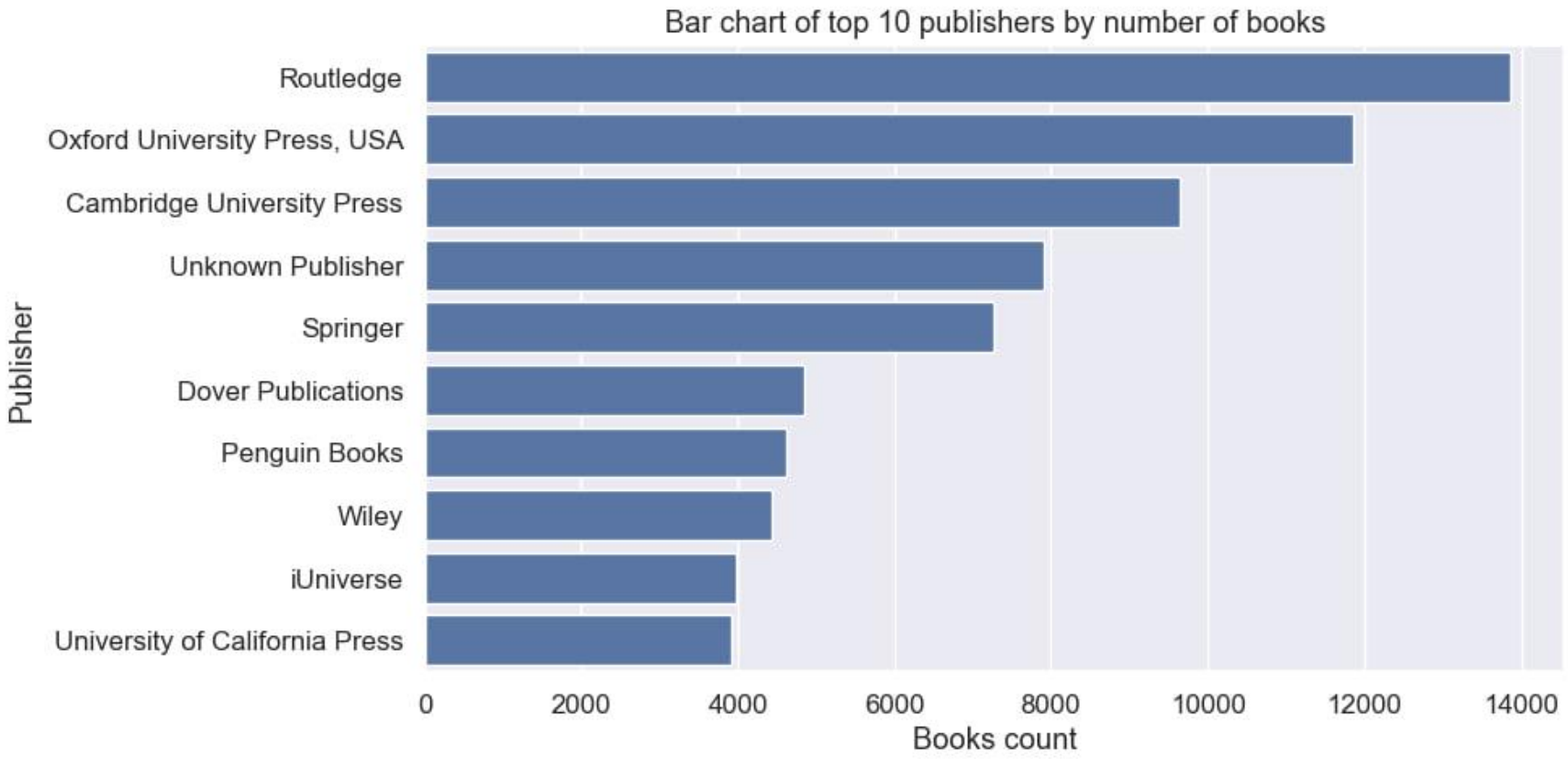
Top autoři

- J.K. Rowling, George Orwell a další dominují hodnocením.
- Odráží kombinaci komerčního úspěchu a literární hodnoty.


Bar chart of top 10 authors by number of books
Authors
Anonymous
William Shakespeare
Francine Pascal
Unknown
Harold Bloom
Isaac Asimov
Carolyn Keene
Agatha Christie
Nora Roberts
R.L. Stine
0
200
400
600
800
1000
1200
1400
Books count

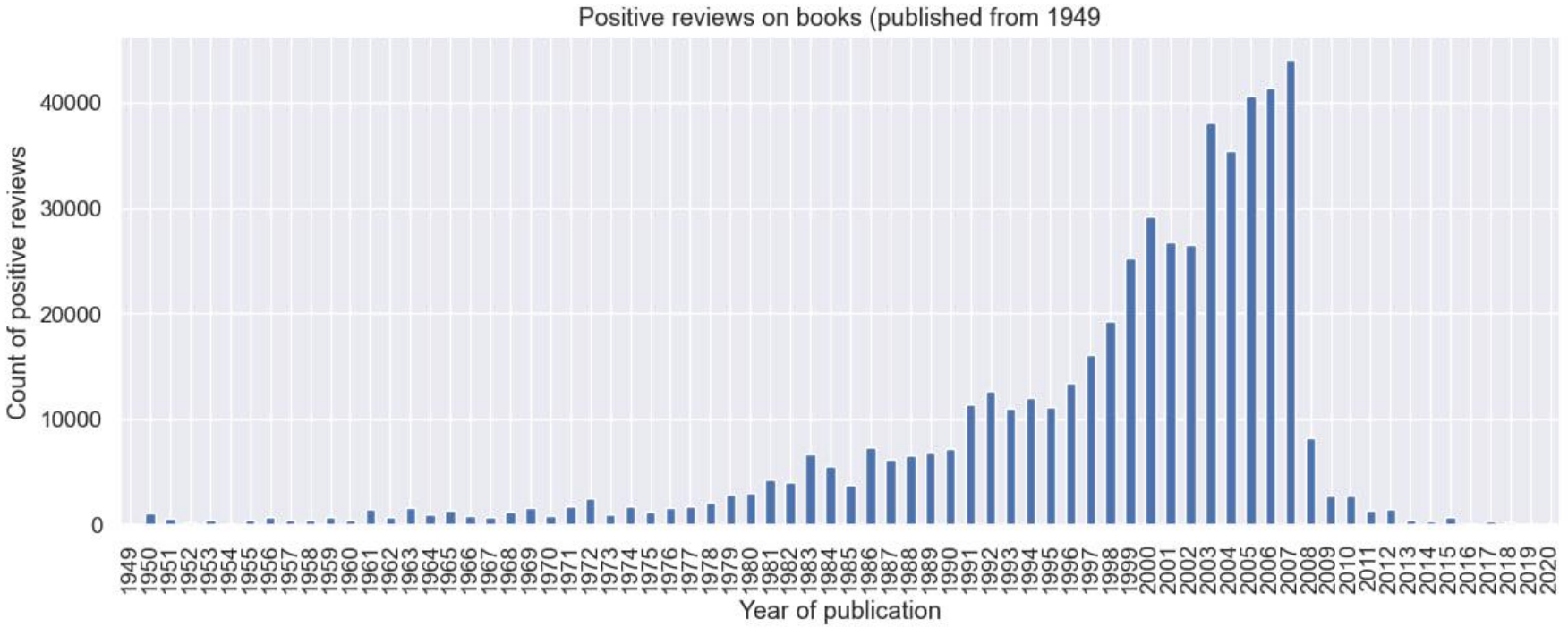
Nejčastější vydavatelé

- Routledge a Oxford UP vydávají nejvíce knih – převážně odborné.
- Ovlivňuje charakter celého datasetu.


Bar chart of top 10 publishers by number of books
Routledge
Oxford University Press, USA
Cambridge University Press
Unknown Publisher
Springer
Dover Publications
Penguin Books
Wiley
iUniverse
University of California Press
Publisher
0
2000
4000
6000
8000
10000
12000
14000
Books count

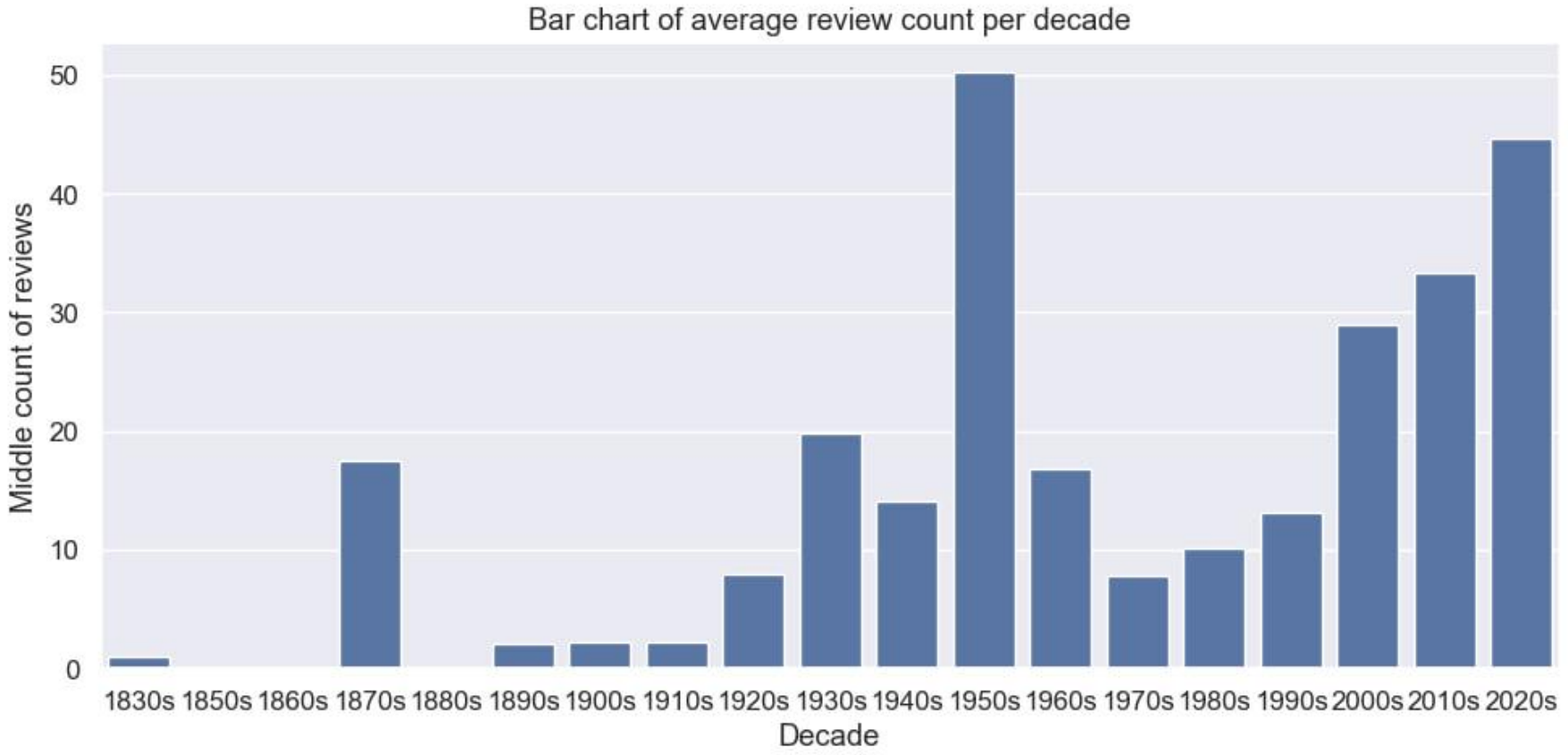
Recenze podle roku

- Od 90. let roste počet pozitivních recenzí.
- Důvod: nástup internetu a populárních sérií jako Harry Potter.


Positive reviews on books (published from 1949
Count of positive reviews
0
10000
20000
30000
40000
Year of publication
1949
1950
1951
1952
1953
1954
1955
1956
1957
1958
1959
1960
1961
1962
1963
1964
1965
1966
1967
1968
1969
1970
1971
1972
1973
1974
1975
1976
1977
1978
1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020

Recenze podle dekády

- Nejvíce recenzí mají knihy z let 2000–2009.
- Rozmach digitální dostupnosti a online katalógů.


Bar chart of average review count per decade
Middle count of reviews
0
10
20
30
40
50
1830s
1850s
1860s
1870s
1880s
1890s
1900s
1910s
1920s
1930s
1940s
1950s
1960s
1970s
1980s
1990s
2000s
2010s
2020s
Decade

Korelační matice

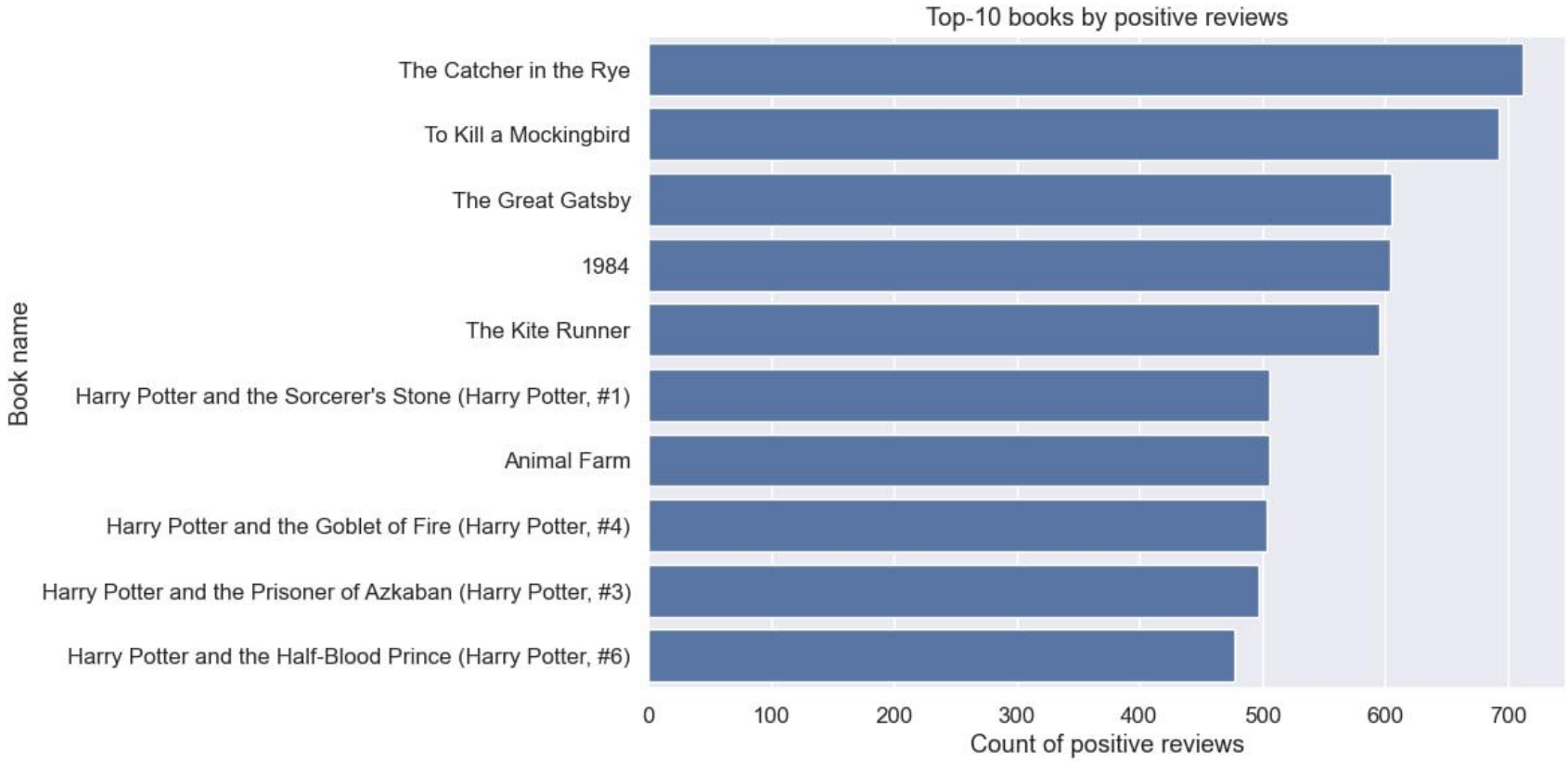
- Ukazuje vztahy mezi proměnnými: počet stran, rok vydání apod.
- Pomáhá vybrat vhodné prediktory pro modelování.

Jazyk a recenze

- Boxplot: anglické knihy mají více recenzí.
- Jazyk je významný prediktor popularity.

Autoři a hodnocení

- Známi autoři mají vyšší průměrné hodnocení.
- Čtenáři reagují na reputaci a mediální vliv.


Top-10 books by positive reviews
The Catcher in the Rye
To Kill a Mockingbird
The Great Gatsby
1984
The Kite Runner
Harry Potter and the Sorcerer's Stone (Harry Potter, #1)
Animal Farm
Harry Potter and the Goblet of Fire (Harry Potter, #4)
Harry Potter and the Prisoner of Azkaban (Harry Potter, #3)
Harry Potter and the Half-Blood Prince (Harry Potter, #6)
Book name
0
100
200
300
400
500
600
700
Count of positive reviews

Výstup modelu

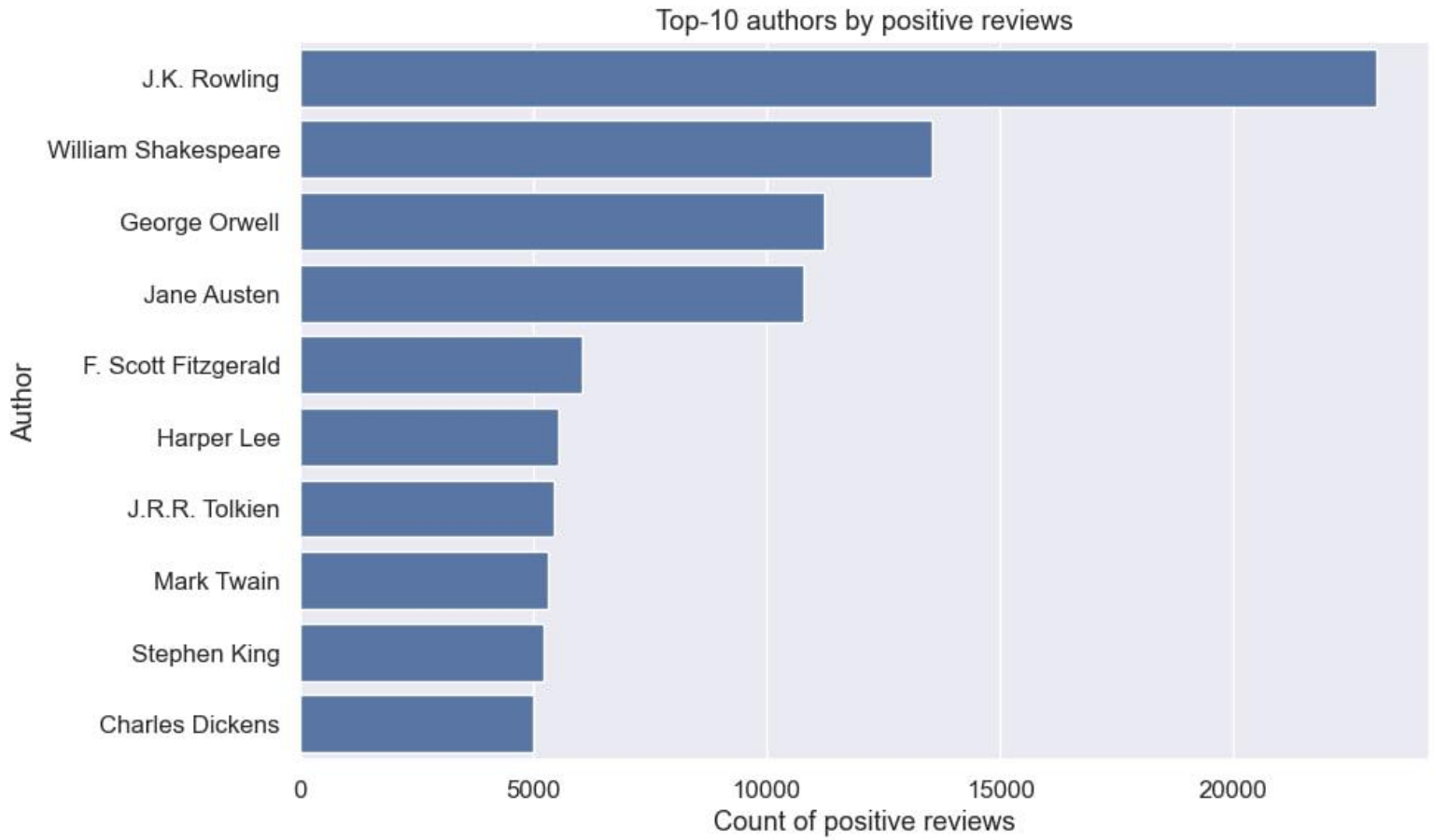
- Koeficienty modelu ukazují sílu vlivu každé proměnné.
- Model má stabilní chování bez výrazných chyb.

6. Analyzing positive reviews by publication year (from 1949 onward):
- Merges the ratings dataset with the main books dataset on book name
- Filters to include only positive reviews ('it was amazing', 'really liked it')
- Considers only books published in or after 1949
- Plots a bar chart showing the number of positive reviews by publication year

	Name	Rating	PublishYear	Authors
0	Agile Web Development with Rails: A Pragmatic ...	it was amazing	2005	Dave Thomas
1	The Restaurant at the End of the Universe (Hit...	it was amazing	2005	Douglas Adams
2	The Restaurant at the End of the Universe (Hit...	it was amazing	1981	Douglas Adams
3	Siddhartha	it was amazing	2004	Hermann Hesse
4	Siddhartha	it was amazing	2000	Hermann Hesse

Validace modelu

- Predikce a rezidua potvrzují přiměřenou přesnost.
- Model dobře vystihuje trendy popularity.


Top-10 authors by positive reviews
J.K. Rowling
William Shakespeare
George Orwell
Jane Austen
F. Scott Fitzgerald
Harper Lee
J.R.R. Tolkien
Mark Twain
Stephen King
Charles Dickens
Author
0
5000
10000
15000
20000
Count of positive reviews

Shrnutí metody

- Lineární regrese s log-transformací počtu recenzí.
- Log pomáhá stabilizovat rozdělení a omezit vliv extrémů.
- $R^2 = 0.22$ znamená, že model vysvětluje 22 % variability.

Závěry

- Angličtina, renomovaný vydavatel a moderní knihy vedou k vyšší oblíbenosti.
- Krátké nebo velmi dlouhé knihy a neznámé jazyky mají méně recenzí.
- Model lze rozšířit o NLP, žánrové prvky nebo demografii.

Děkuji

- Děkuji za pozornost.
- Ráda zodpovím dotazy. Analýza mi ukázala sílu dat i důležitost jejich interpretace.