Московский государственный технический университет им. Н.Э. Баумана
Кафедра «Системы обработки информации и управления»

Лабораторная работа №1
по дисциплине
«Методы машинного обучения»
на тему
«истории о данных»

Выполнил:
студент группы ИУ5-22М
Лю Ченхао

Москва — 2024 г.

1. Цель лабораторной работы

изучение различных методов визуализация данных и создание истории на основе данных.

2. Задание

Выбрать набор данных (датасет). Вы можете найти список свободно распространяемых датасетов здесь.Для лабораторных работ не рекомендуется выбирать датасеты очень большого размера.

Создать "историю о данных" в виде юпитер-ноутбука, с учетом следующих требований:

1) История должна содержать не менее 5 шагов (где 5 - рекомендуемое количество шагов). Каждый шаг содержит график и его текстовую интерпретацию.
2) На каждом шаге наряду с удачным итоговым графиком рекомендуется в юпитер-ноутбуке оставлять результаты предварительных "неудачных" графиков.
3) Не рекомендуется повторять виды графиков, желательно создать 5 графиков различных видов.
4) Выбор графиков должен быть обоснован использованием методологии data-to-viz. Рекомендуется учитывать типичные ошибки построения выбранного вида графика по методологии data-to-viz. Если методология Вами отвергается, то просьба обосновать Ваше решение по выбору графика.
5) История должна содержать итоговые выводы. В реальных "историях о данных" именно эти выводы представляют собой основную ценность для предприятия.

Сформировать отчет и разместить его в своем репозитории на github.

3. Ход выполнения работы

3.1. Текстовое описание набора данных

В этой тетради я буду использовать графики для визуализации взаимосвязи между переменными в наборе данных "Вхождение книги в топ-100 бестселлеров на Amazon.".

Этот набор данных предлагает подробный обзор 100 лучших книг-бестселлеров Amazon вместе с их отзывами покупателей, рейтингами, ценами и т. д. Если вы любитель книг, специалист по изучению данных или просто интересуетесь последними литературными тенденциями, этот набор данных позволит вам заглянуть в мир популярного чтения.Рейтинг книги: Рейтинг книги среди 100 лучших книг-бестселлеров на Amazon.Book Title: The title of the book.

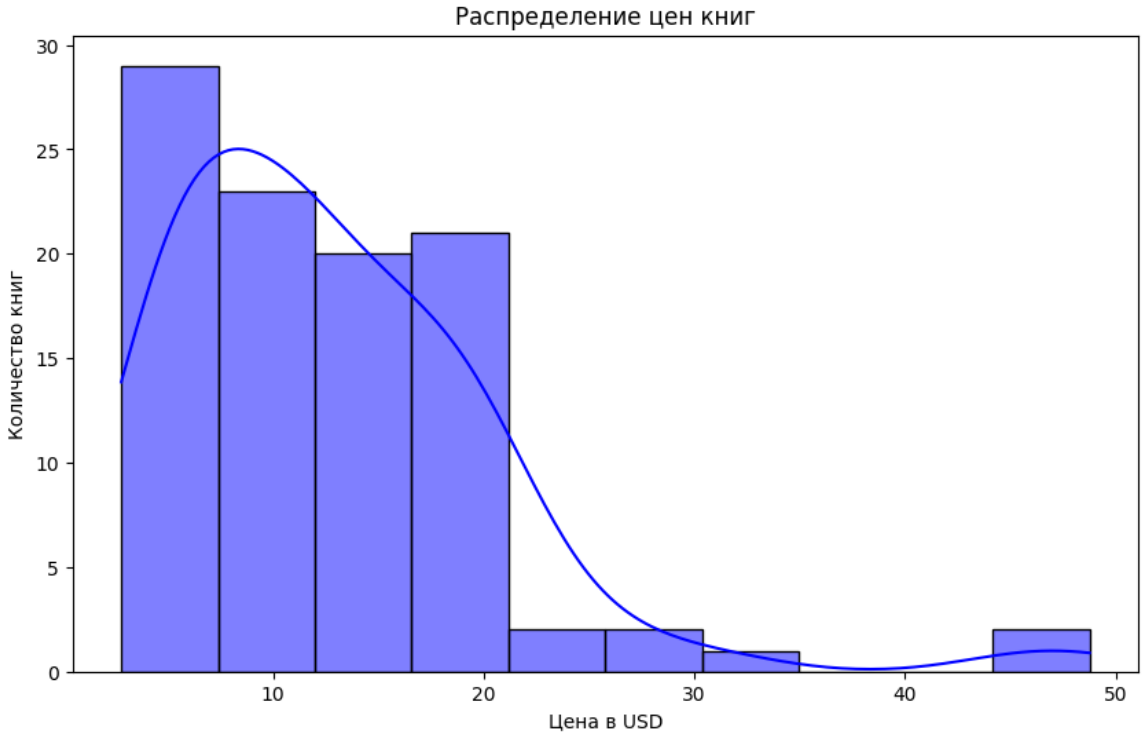
1) Price: The price of the book in USD.
2) Rating: The overall rating of the book, on a scale of 1 to 5.
3) Author: The author of the book.
4) Year of Publication: The year in which the book was published.
5) Genre: The genre or category to which the book belongs.
6) URL: The URL link to the book on Amazon's platform.

3.2. Основные характеристики набора данных

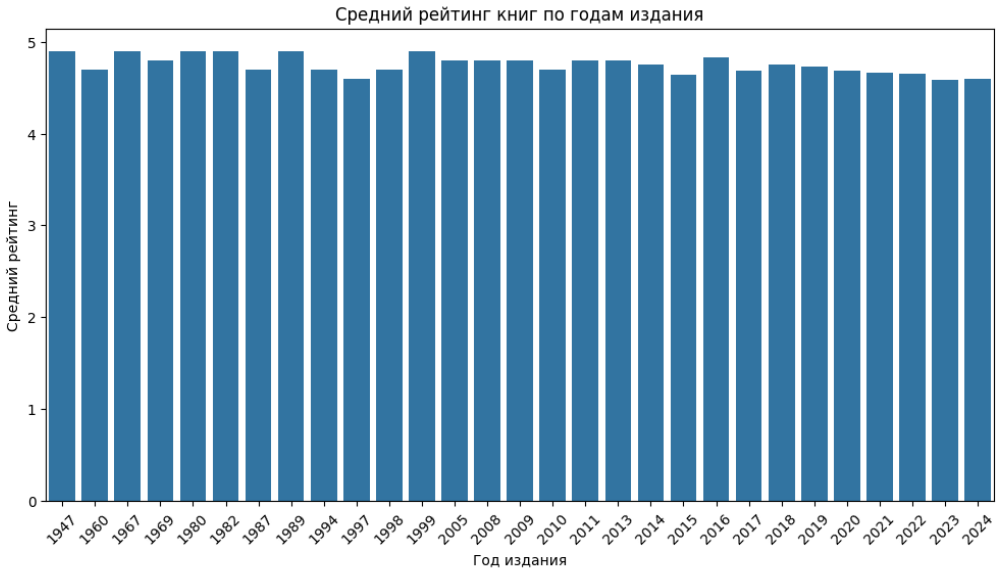
```
# Шаг 1: Импортирование необходимых библиотек
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
# Шаг 2: Загрузка данных
# Предполагаем, что у вас есть файл CSV с данными о бестселлерах Amazon.
df = pd.read_csv('amazon_books.csv')
# Шаг 3: Предварительный анализ данных
print(df.head())  # Вывод первых пяти строк для предварительного анализа
```

```
   Rank                                        book title  Price  rating  \
0     1                      Iron Flame (The Empyrean, 2)  18.42     4.1
1     2                                   The Woman in Me  20.93     4.5
2     3                                My Name Is Barbra  31.50     4.5
3     4  Friends, Lovers, and the Big Terrible Thing: A...  23.99     4.4
4     5                             How to Catch a Turkey   5.65     4.8

              author  year of publication                genre  \
0    Rebecca Yarros                 2023     Fantasy Romance
1    Britney Spears                 2023              Memoir
2  Barbra Streisand                 2023       Autobiography
3     Matthew Perry                 2023              Memoir
4      Adam Wallace                 2018  Childrens, Fiction
```

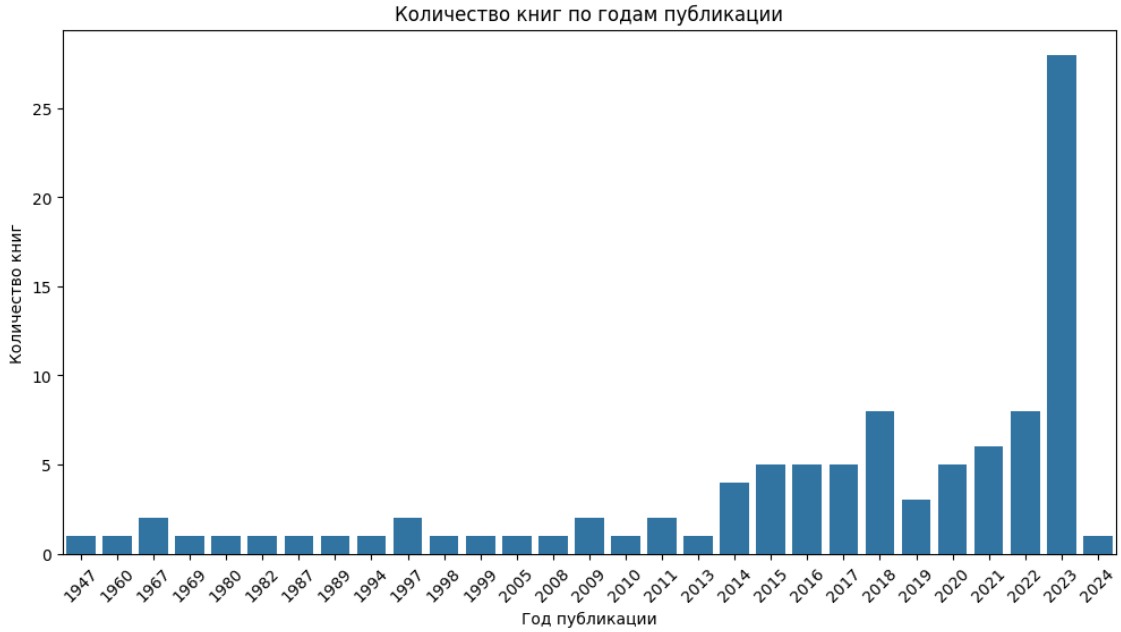
[3]:
```python
# Шаг 4: Визуализация данных

# График 1: Распределение цен книг
plt.figure(figsize=(10, 6))
sns.histplot(df['Price'], kde=True, color='blue')
plt.title('Распределение цен книг')
plt.xlabel('Цена в USD')
plt.ylabel('Количество книг')
plt.show()
```



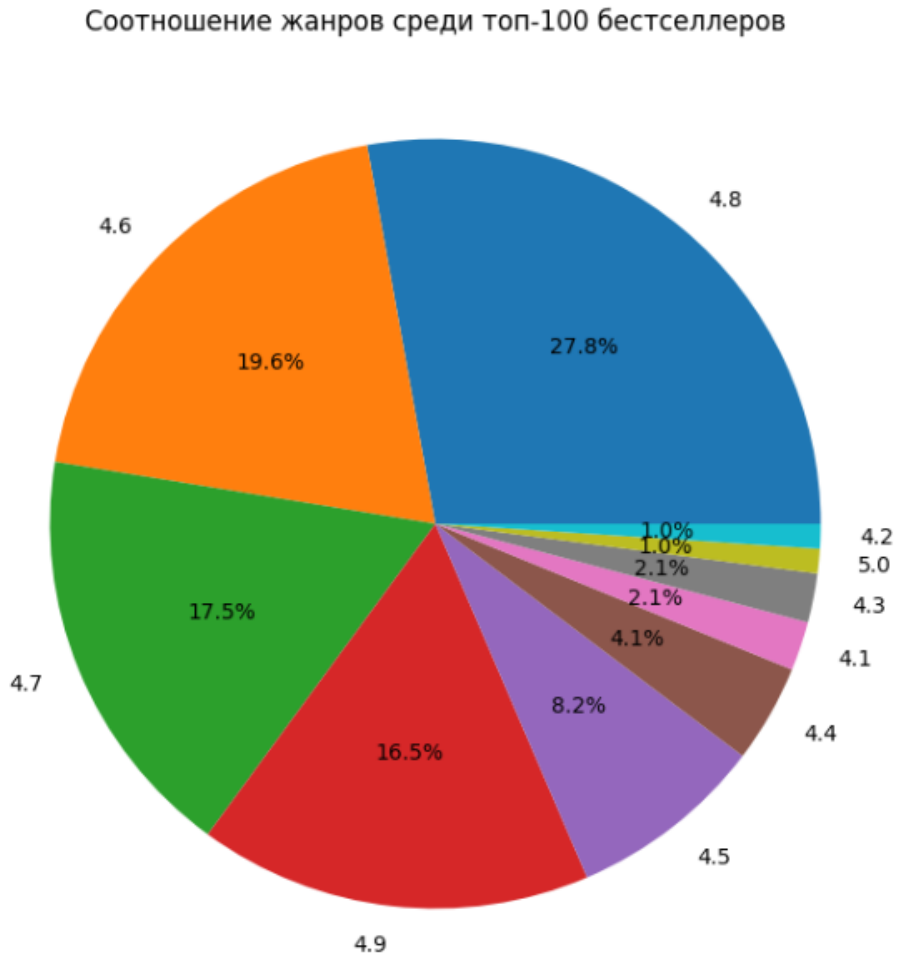
[4]:
```
# График 2: Гистограмма, показывающая средние оценки в зависимости от года публикации
plt.figure(figsize=(12, 6))
sns.barplot(x='year of publication', y='rating', data=df, errorbar=None)
plt.title('Средний рейтинг книг по годам издания')
plt.xlabel('Год издания')
plt.ylabel('Средний рейтинг')
plt.xticks(rotation=45)
plt.show()
```



[7]:
```
# График 3: Количество книг по годам публикации
plt.figure(figsize=(12, 6))
sns.countplot(x='year of publication', data=df)
plt.title('Количество книг по годам публикации')
plt.xlabel('Год публикации')
plt.ylabel('Количество книг')
plt.xticks(rotation=45)
plt.show()
```



•[12]:
```
# График 4: Процентное соотношение рейтингов в 100 лучших бестселлерах
plt.figure(figsize=(8, 8))
df['rating'].value_counts().plot(kind='pie', autopct='%1.1f%%')
plt.title('Соотношение жанров среди топ-100 бестселлеров')
plt.ylabel('')
plt.show()
```



[10]:
```python
# График 5: Динамика средней цены книг по годам
plt.figure(figsize=(10, 6))
sns.lineplot(x='year of publication', y='Price', data=df, marker='o')
plt.title('Динамика средней цены книг по годам')
plt.xlabel('Год публикации')
plt.ylabel('Средняя цена в USD')
plt.show()
```

Список литературы

[1] Гапанюк Ю. Е. Лабораторная работа «LAB_MMO__DATA_STORY» [Электронный ресурс] https://github.com/ugapanyuk/courses_current/wiki/LAB_MMO__DATA_STORY

[2] https://www.kaggle.com/datasets/anshtanwar/top-200-trending-books-with-reviews