Университет ИТМО

Практическая работа №4

по дисциплине «Визуализация и моделирование»

Автор: Голуб А. Л.
Поток: ВИМ 1.2
Группа: К3243
Факультет: ИКТ
Преподаватель: Чернышева А.В.

Санкт-Петербург, 2021 г.

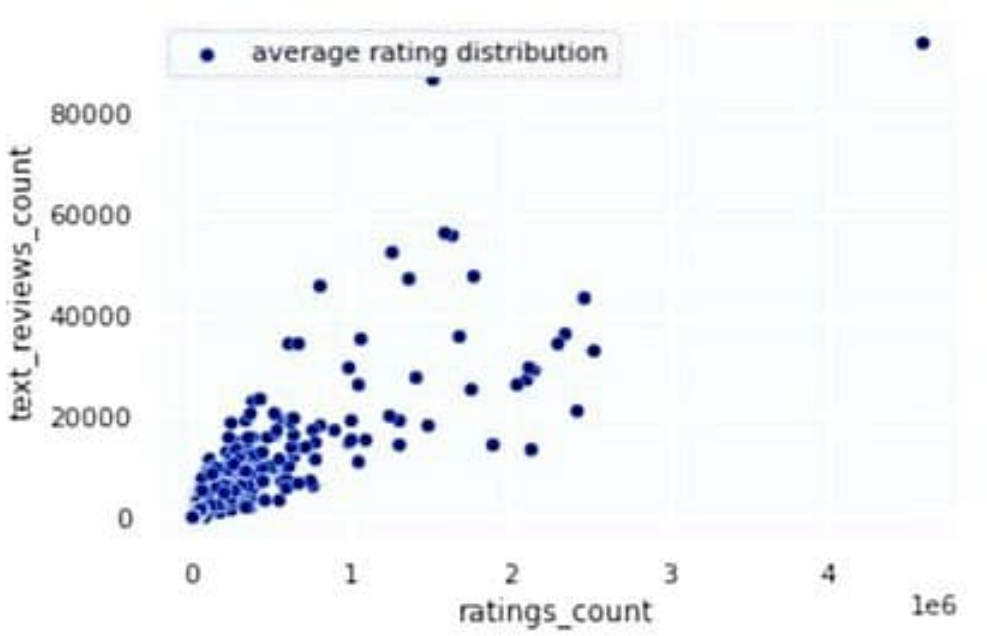
Датасет: Goodreads-books

Визуализируем данные, предобработанные на предыдущем этапе, чтобы подтвердить или опровергнуть гипотезы о них.

1. **Столбцы ratings_count и text_reviews_count достаточно сильно коррелируют.** Построим диаграмму рассеяния с помощью seaborn: по x отложим ratings_count (сколько раз книге был выставлен рейтинг на сайте), по y - text_reviews_count (число текстовых отзывов на книгу).

```
sns.set_theme(style="darkgrid")
plt.plot()
sns.set_color_codes("dark")
sns.scatterplot(data=df, x='ratings_count', y='text_reviews_count',
                label='average rating distribution', color="b");
plt.legend(ncol=1);

```



Дополнительно посчитаем коэффициент корреляции - он равен примерно 0.87:

```
df['ratings_count'].corr(df['text_reviews_count'])

```

Значит, эти столбцы действительно сильно коррелируют, и один из них можно удалить. Вероятно, логичнее будет удалить text_reviews_count, так как он напрямую не связан со средним рейтингом книги average_rating.

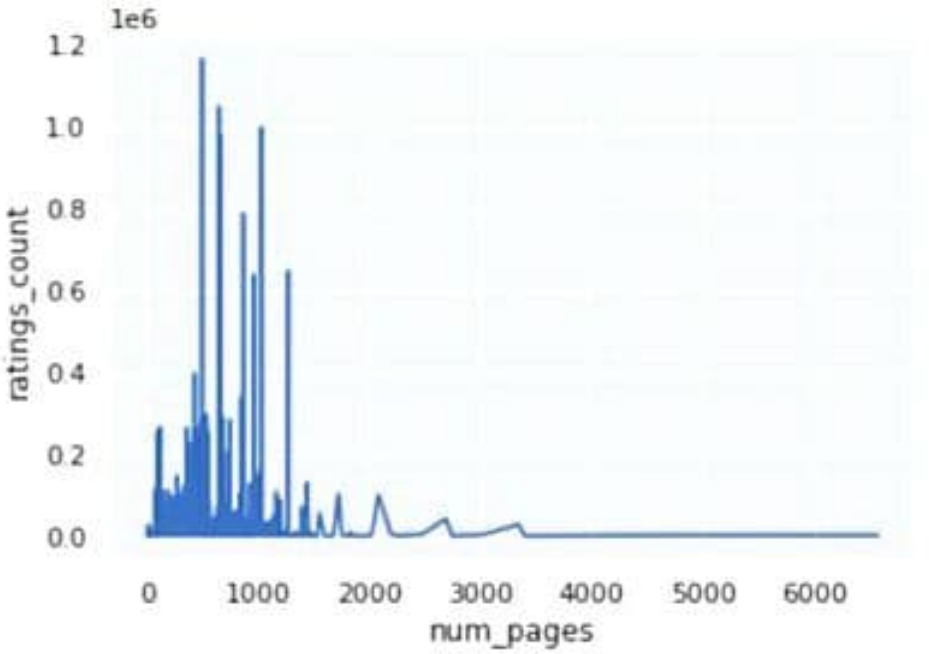
```
df.drop('text_reviews_count', axis=1, inplace=True)
```

2. **Пользователи Goodreads практически не читают книги длиной более 1500 страниц.**

   В seaborn построим график зависимости числа выставленных книге рейтингов от числа страниц в ней.

```
sns.lineplot(data=df, x='num_pages', y='ratings_count', ci=None);

```



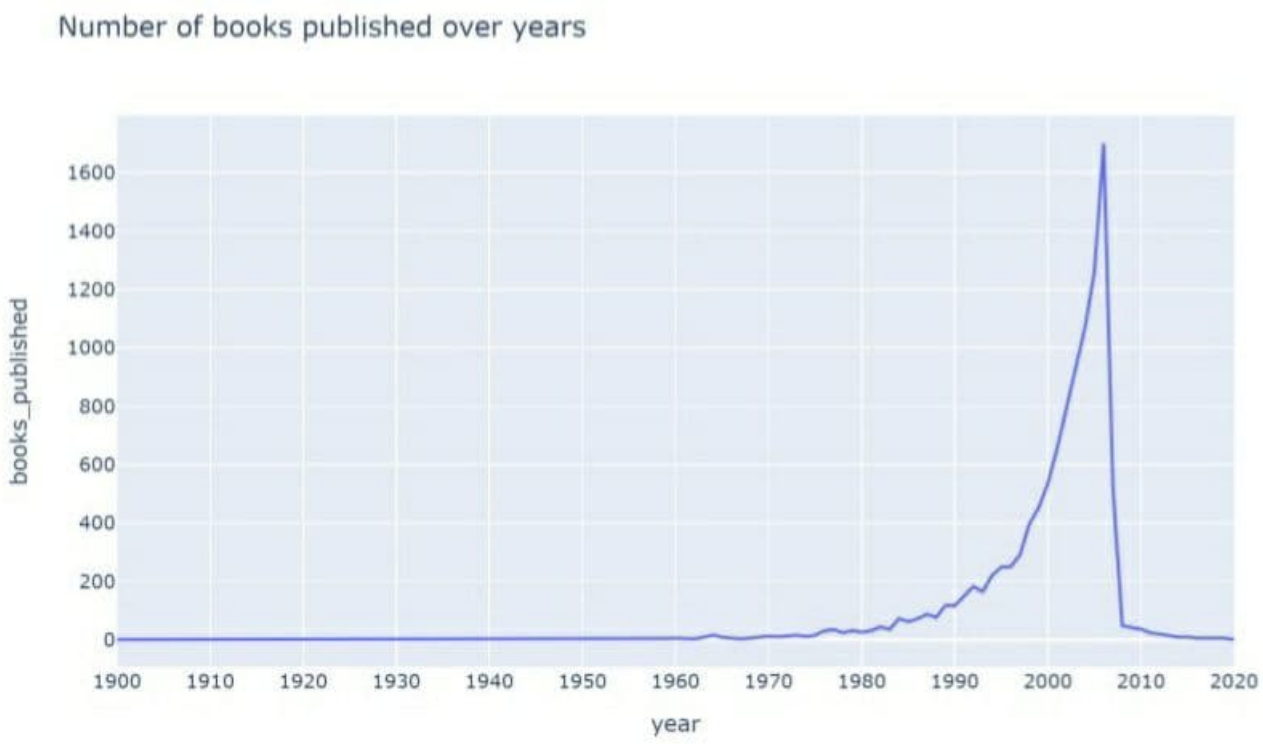
Как можно видеть, книгам длиннее 1500 страниц действительно редко ставят оценки на сайте.

3. **Книги равномерно распределены по годам издания** (каждый год, по данным датасета, выходило примерно одинаковое число книг).
   Создадим датафрейм, со столбцами «год издания» и «число изданных в этот год книг» и визуализируем полученные данные с помощью библиотеки plotly.

```
  yearly_published = df.groupby(df.publication_date.dt.year)['bookID'].
    count().to_frame().reset_index()
yearly_published.columns = ['year', 'books_published']

fig = px.line(yearly_published, x='year', y='books_published',
              title='Number of books published over years')
fig.update_xaxes(tick0=1900, dtick=10)
fig.show()

```



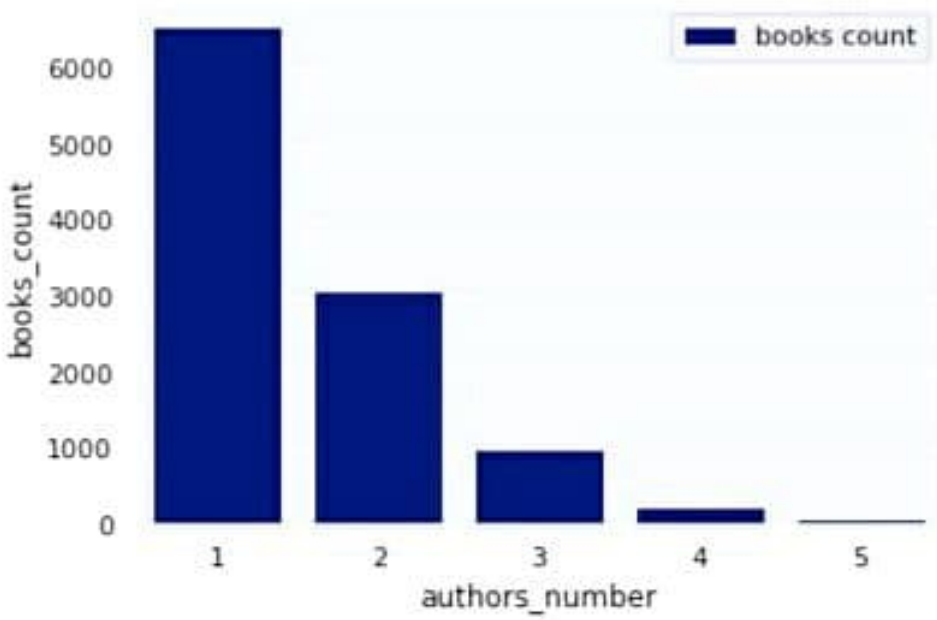
Оказывается, записи в датасете неравномерно распределены по датам публикации. Подавляющее большинство книг вышло в промежуток с 1975 по 2007 год.

4. **Большинство книг написаны одним или двумя авторами.**
   Создадим датафрейм со столбцами «число авторов» и «сколько книг написано таким числом авторов» и отсортируем его по убыванию значений второго столбца. Затем отобразим первые пять строк датафрейма на диаграмме, используя seaborn.

```
authors_number_count = df.groupby('authors_number')['bookID'].count() \
  .to_frame().reset_index().sort_values(by='bookID', ascending=False)
authors_number_count.columns = ['authors_number', 'books_count']

sns.set_theme(style="darkgrid")
plt.plot()
sns.set_color_codes("dark")
sns.barplot(data=authors_number_count, x='authors_number', y='
    books_count',
            label = 'books count', color='b');
plt.legend(ncol=1);

```

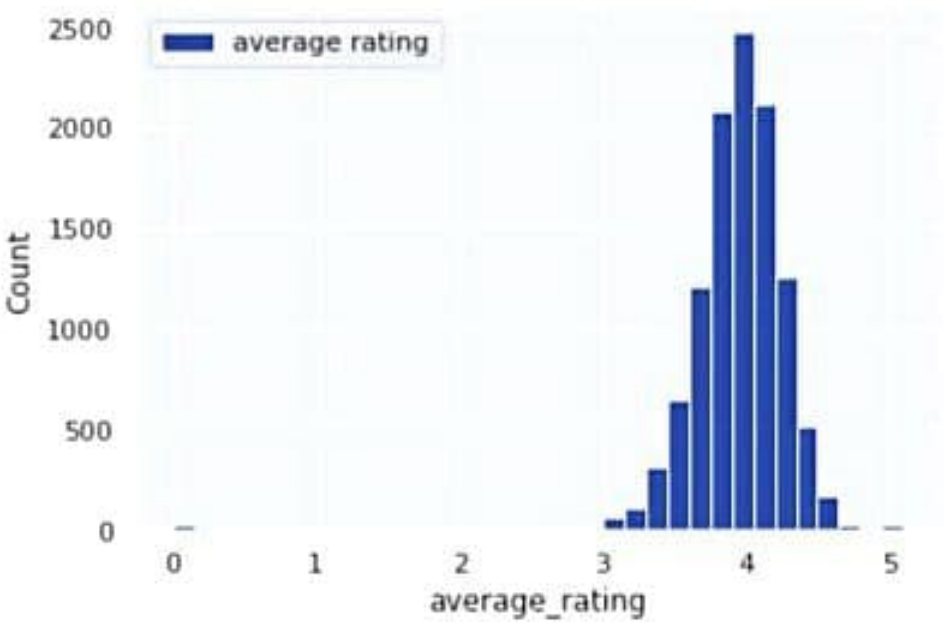


Таким образом, гипотеза подтверждена.

5. **Средний рейтинг большей части книг лежит в промежутке от 3 до 5.**
   С помощью seaborn построим гистограмму распределения среднего рейтинга книг.

```
sns.set_theme(style="darkgrid")
plt.plot()
sns.set_color_codes("dark")
sns.histplot(data=df, x='average_rating',
             label='average rating', binwidth=0.15, color="b");
plt.legend(ncol=1);

```



   Как и предыдущая, гипотеза оказалась верной.