

Университет ИТМО

Практическая работа №2
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.

1 Описание данных

В ходе практической работы был выбран датасет [Goodreads-books](#) с информацией о книгах и их рейтингах на популярном ресурсе Goodreads. Информация о столбцах датасета представлена в таблице.

название столбца в датасете	краткое описание данных	тип данных	шкала данных
bookID	уникальный номер книги в датасете	натуральное число	относительная
title	название книги, использованное при публикации	строка	номинальная
authors	список авторов книги	строка с перечислением имен авторов через /	номинальная
average_rating	средний рейтинг книги на сайте	число от 0.0 до 5.0	интервальная
isbn	ISBN-идентификатор книги	натуральное число (11 знаков)	относительная
isbn13	ISBN-идентификатор книги (13 знаков)	натуральное число (13 знаков)	относительная
language_code	язык оригинала книги	строка	номинальная
num_pages	число страниц в книге	натуральное число	относительная
ratings_count	сколько раз книге был поставлен рейтинг	натуральное число	относительная
text_reviews_count	число письменных отзывов на книгу	натуральное число	относительная
publication_date	дата первой публикации книги	дата в формате ММ/DD/YYYY	интервальная
publisher	издательство, опубликовавшее книгу	строка	номинальная

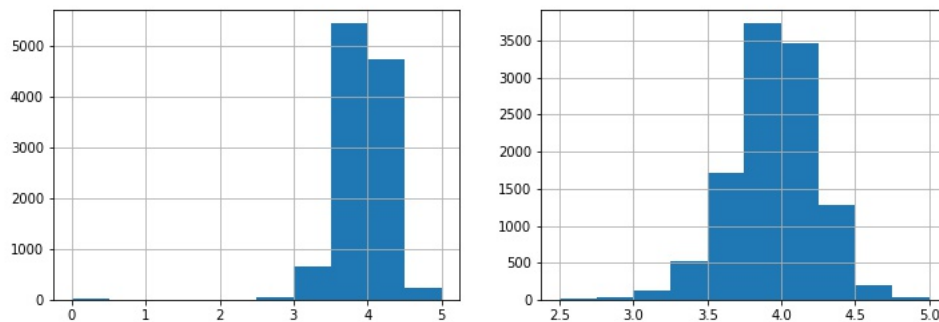
2 Описательная статистика

Построим диаграммы распределения данных в столбцах датасета, чтобы выявить основные закономерности в них.

1. Как статистически распределены **рейтинги книг** на сайте?

Нарисуем гистограмму распределения и приблизим наиболее информативный участок.

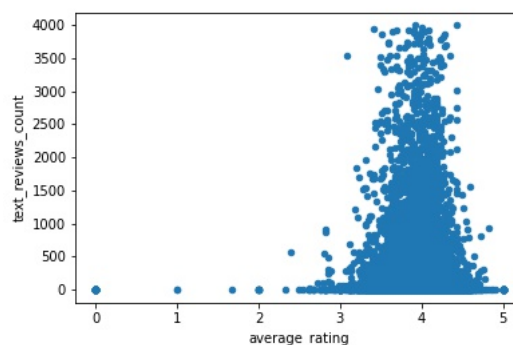
```
1 fig, axes = plt.subplots(nrows=1, ncols=2)
2 df['average_rating'].hist(ax=axes[0], figsize=(12, 4));
3 df['average_rating'].hist(ax=axes[1],
4     range=[2.5, 5], figsize=(12, 4));
5
```



Чаще всего книги получают оценки от **3.5** до **4.5**. Книги с рейтингом 0, скорее всего, пока не получили ни одной оценки. С помощью запроса к датасету можно посчитать, что таких книг 25.

2. Как **число текстовых отзывов** на книгу соотносится с ее **рейтингом**? Правда ли, что пользователи пишут отзывы на не понравившиеся книги чаще, чем на понравившиеся? Отфильтруем выбросы и построим график scatter plot.

```
1 df.query('text_reviews_count < 4000').plot. \
2     scatter(x='average_rating', y='text_reviews_count');
3
```



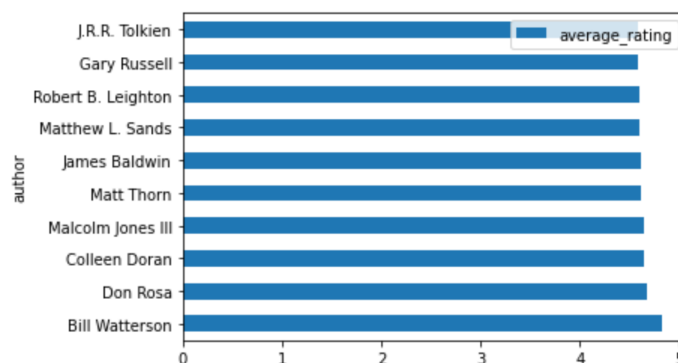
Судя по графику, предположение неверно. Число текстовых отзывов и количество поставленных рейтингов (см. гистограмму выше) примерно одинаково распределены относительно величины рейтинга.

3. У каких авторов самый высокий средний рейтинг?

Создадим датафрейм со столбцами «автор», «средний рейтинг книг автора» и «общее число книг автора в датасете». Возьмем авторов, написавших не менее трех книг, чтобы сделать визуализацию более репрезентативной, и отсортируем их по рейтингу по убыванию. Топ-десять получившегося списка изобразим на диаграмме.

```

1  authors_ratings = dict()
2  books_count = dict()
3  for _, row in df[['authors', 'average_rating']].iterrows():
4      authors = row['authors'].split('/')
5      for a in authors:
6          authors_ratings[len(authors_ratings)] = [a, row['average_rating']]
7          if a in books_count:
8              books_count[a] += 1
9          else:
10             books_count[a] = 1
11
12  authors_ratings = pd.DataFrame.from_dict(data=authors_ratings,
13      orient='index', columns=['author', 'average_rating'])
14  authors_ratings['books_number'] = pd.Series([books_count[a] for a in
15      authors_ratings['author']])
16  authors_ratings.groupby('author').mean(). \
17      sort_values(ascending=False, by='average_rating').reset_index(). \
18      query('books_number >= 10').iloc[:10]. \
19      plot.barh(x='author', y='average_rating');
```



Как видно на диаграмме, средний рейтинг книг наиболее высоко оцененных авторов лежит в диапазоне от 4.5 до 5.

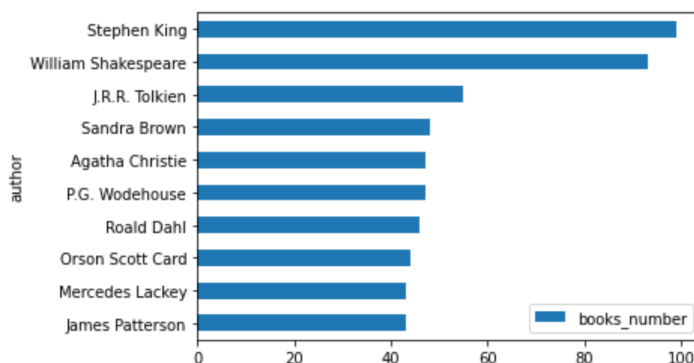
4. Какие авторы написали больше всех книг?

Воспользуемся только что созданным датафреймом и найдем топ-десять авторов по числу написанных книг. Результат также изобразим на диаграмме.

```

1 authors_ratings = authors_ratings.drop_duplicates(subset=['author'])
2 fig = authors_ratings.sort_values(ascending=False, by='books_number')
3     .iloc[:10].plot.barh(y='books_number', x='author');
4 fig.invert_yaxis()
5

```



Наиболее плодовитыми писателями оказались Стивен Кинг, Уильям Шекспир и Дж. Р. Р. Толкин.

5. Книг на каком **языке** в датасете больше всего?

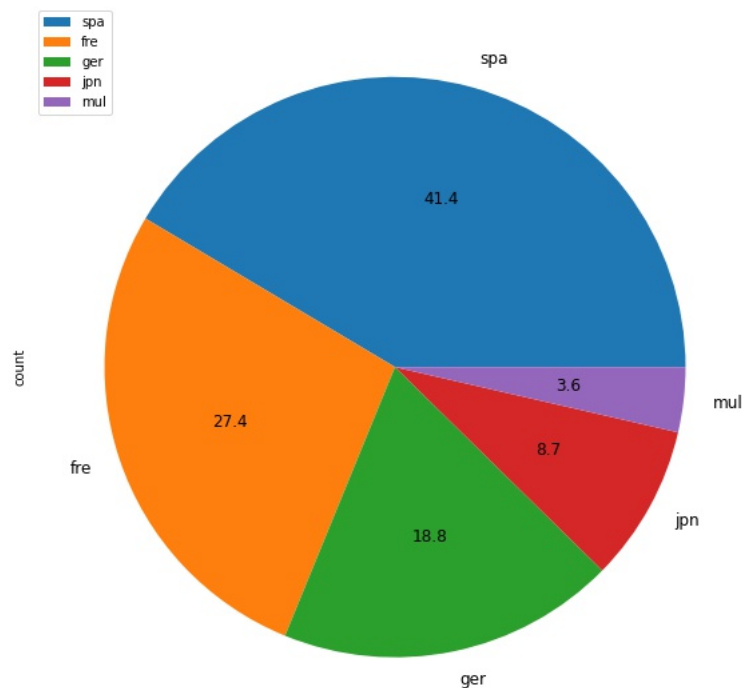
Сгруппируем датасет по языкам, посчитаем, сколько книг написано на каждом из них, и сохраним результат в отдельный датафрейм. Подавляющее большинство книг в датасете - 1629 - написано на **английском** языке; исключим его из рассмотрения, чтобы получить более репрезентативное изображение. Визуализируем процентное соотношение книг среди пяти самых частых языков в виде круговой диаграммы.

```

1 languages_count = df.groupby('language_code')['language_code']. \
2     count().reset_index(name='count')
3 eng_sum = languages_count[languages_count['language_code'].str. \
4     contains('en-']]['count'].sum()
5 languages_count[languages_count['language_code'] == 'eng']['count']
6     += eng_sum
7 languages_count.drop(languages_count.loc[languages_count[
8     'language_code'].str.contains('en-')].index, inplace=True)
9
10 languages_count = languages_count.sort_values(ascending=False,
11     by='count').iloc[1:6]
12 languages_count.plot.pie(y='count', labels=languages_count[
13     'language_code'], autopct="%.1f", fontsize=12, figsize=(10, 10));
14

```

После английского языка больше всего книг в датасете написано на **испанском**, **французском** и **немецком**.



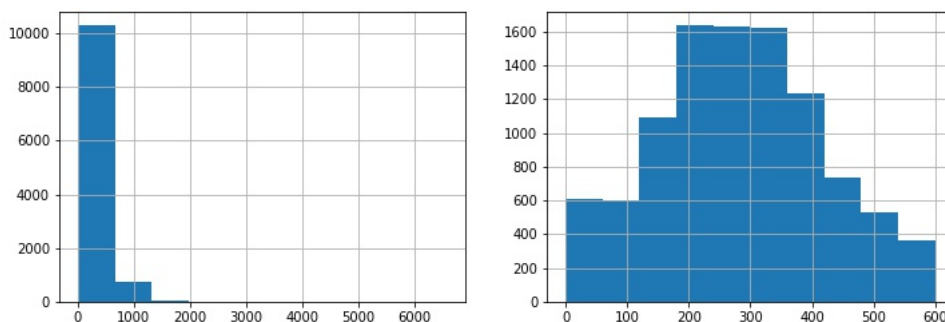
6. Распределение **числа страниц**

Построим распределение с помощью pandas-метода `hist()` и приблизим его наиболее информативный фрагмент.

```

1  fig, axes = plt.subplots(nrows=1, ncols=2)
2  df['num_pages'].hist(ax=axes[0], figsize=(12, 4));
3  df['num_pages'].hist(ax=axes[1], range=[0, 600], figsize=(12, 4));
4

```

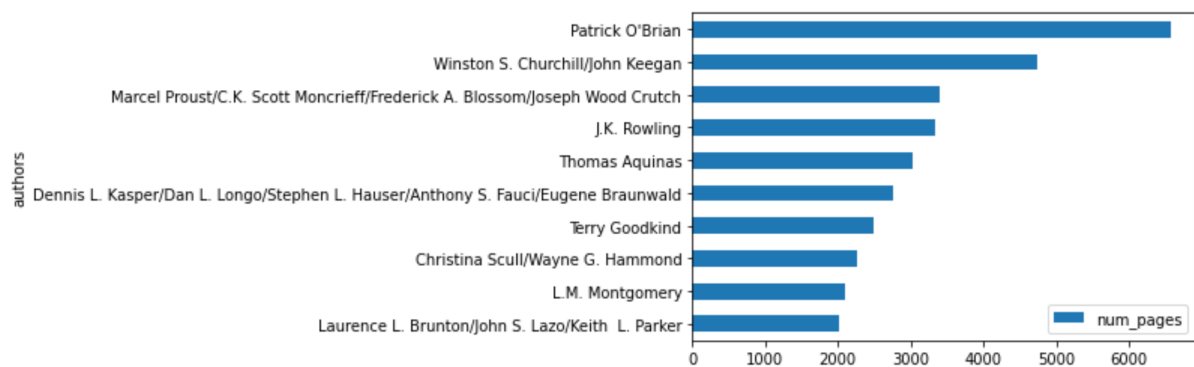


В датасете представлены книги длиной от 0 до 6576 страниц, тогда как медианное значение числа страниц - 299. Как видно на диаграмме, большая часть книг в датасете длиной от **200** до **400** страниц.

7. Какие **авторы** написали **самые длинные книги**?

Отсортируем строки по убыванию числа страниц. Отобразим топ-десять авторов на диаграмме, исключив из рассмотрения анонимных авторов и J.K. Rowling/Mary GrandPré. (J.K. Rowling уже входит в этот список, к тому же, Mary GrandPré иллюстрировала произведения Роулинг, но не была их автором).

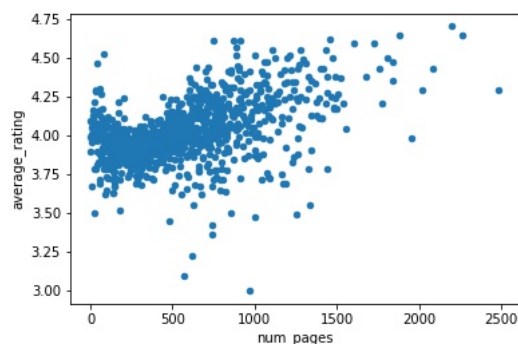
```
1 fig = df[['authors', 'num_pages']].sort_values(  
2     ascending=False, by='num_pages') \  
3     .query('authors not in ["Anonymous",  
4     "J.K. Rowling/Mary GrandPré"]').iloc[:10] \  
5     .plot.barh(x='authors', y='num_pages');  
6 fig.invert_yaxis()  
7
```



8. Как **рейтинг** книги зависит от **числа страниц**?

Отфильтровав записи о книгах, построим график scatter plot, где по горизонтальной оси отложено число страниц, а по вертикальной - средний рейтинг книги данной длины.

```
1 pages_rating = df[['num_pages', 'average_rating']]  
2     .query('0 < num_pages <= 2600 & average_rating != 0')  
3 pages_rating.groupby('num_pages').mean('average_rating').reset_index()  
4     .plot.scatter(x='num_pages', y='average_rating');  
5
```



Можно заметить, что рейтинг книги на сайте слабо растет с увеличением числа страниц, особенно когда их становится больше 200.

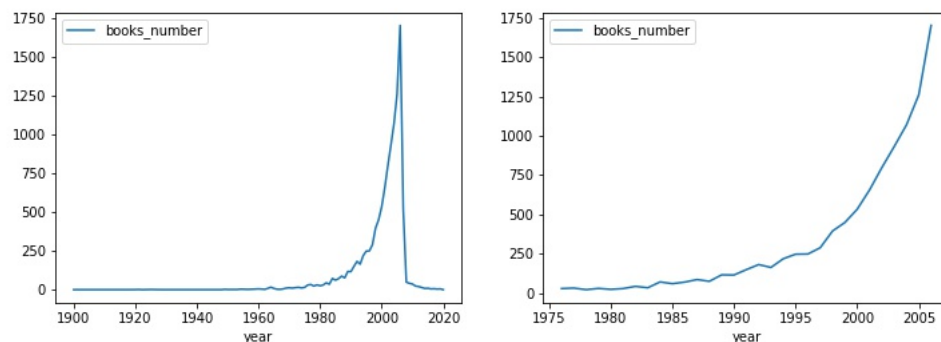
9. Как количество публикуемых книг менялось со временем?

Создадим отдельный датафрейм, где посчитаем, сколько книг было издано в какой год.

```
1 published_over_time = df[['bookID', 'publication_date']]
2 years = []
3 for v in published_over_time['publication_date']:
4     years.append(int(v.split('/')[2]))
5 published_over_time['year'] = pd.Series(years)
6 published_over_time = published_over_time.groupby('year')['bookID'].
7     count().reset_index(name='books_number')
8
```

Построим график зависимости числа книг от года их издания. Очевидно, в датафрейме в основном представлены книги, изданные в промежуток с 1975 по 2007 год. Построим дополнительный график, чтобы рассмотреть этот участок кривой.

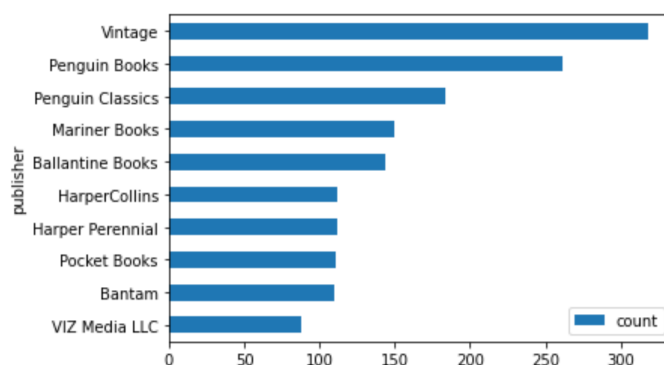
```
1 fig, axes = plt.subplots(nrows=1, ncols=2)
2 published_over_time.plot(ax=axes[0], x='year',
3     y='books_number', figsize=(12, 4));
4 published_over_time.query('1975 < year < 2007')
5     .plot(ax=axes[1], x='year', y='books_number', figsize=(12, 4));
6
```



10. Какое издательство печатает больше всех книг?

Создадим датафрейм со столбцами «издательство» и «число изданных им книг». Отсортируем список по числу книг и отобразим топ-десять на диаграмме.

```
1 publishers = df.groupby('publisher')['publisher'].count() \
2     .reset_index(name='count')
3 fig = publishers.sort_values(ascending=False, by='count') \
4     .iloc[:10].plot.barh(x='publisher', y='count');
5 fig.invert_yaxis()
6
```

Из всех книг в датасете более 300 были изданы издательством «Vintage». Более 250 книг вышли в издательстве «Penguin Books» и чуть менее 200 - в издательстве «Penguin Classics». Около 150 книг напечатали «Ballantine Books» и «Mariner Books» каждое.

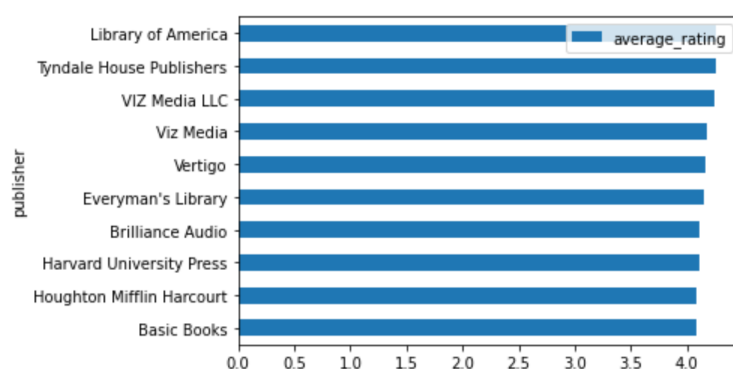
11. У какого издательства самый высокий средний рейтинг?

Дополним датафрейм publishers столбцом со средним рейтингом книг, изданных издательством. Среди издательств, напечатавших хотя бы 20 книг, выявим топ-десять по рейтингу и изобразим результат на диаграмме.

```

1 publishers['average_rating'] = df.groupby('publisher')['
2   'average_rating'].mean().reset_index()['average_rating']
3 fig = publishers.query('count >= 20').sort_values(
4   ascending=False, by='average_rating').iloc[:10]
5   .plot.barh(x='publisher', y='average_rating');
6 fig.invert_yaxis()
7

```



Самый высокий средний рейтинг - чуть выше 4 - у книг издательств «Library of America», «Tyndale House Publishers» и «VIZ Media LLC».