Университет ИТМО

Практическая работа №3

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	интервальная
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)	интервальная
isbn13	ISBN-идентификатор книги (13 знаков)	натуральное число (13 знаков)	интервальная
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 MM/DD/YYYY	интервальная
publisher	издательство, опубликовавшее книгу	строка	номинальная

2 Проблемы в данных

В датасете есть несколько некорректных строк: в них разделитель столбцов - запятая - встречается внутри значений ячеек. Поскольку таких случаев небольшое количество, ошибки можно исправить вручную.
Ниже в таблице представлены другие проблемы в данных и возможные способы их решения в процессе предобработки.

название столбца в датасете	**тип данных**	**проблема в данных**	**способы решения**
bookID	int	-	-
title	string	-	-
authors	string	несколько авторов записываются в одну строку через /	разделить строку по / и перезаписать как список
average_rating	double [0.0; 0.5]	-	-
isbn	int (11 знаков)	не несет полезной информации	удалить столбец
isbn13	int (13 знаков)	не несет полезной информации	удалить столбец
language_code	string	несколько различных кодов для английского языка	выбрать один из кодов и заменить все остальные на него
num_pages	int	-	-
ratings_count	int	-	-
text_reviews_count	int	-	-
publication_date	дата MM/DD/YYYY	дата распознается как тип данных object и есть некорректные значения	вручную исправить некорректные значения и распознать дату как Timestamp средствами pandas
publisher	string	-	-

3 Предобработка датасета

1. **Считывание данных и простейшая предобработка**

   При считывании данных некорректные строки, где разделителей - запятых - больше, чем столбцов датафрейма, были исправлены вручную. Затем были удалены столбцы с ISBN-идентификаторами книг. Ячейки столбца с именами авторов были перезаписаны как списки, и в датасет был добавлен столбец с числом авторов. Пропущенных значений в датафрейме нет.

```
df = df.drop(columns=['isbn', 'isbn13'])
df['authors'] = df['authors'].str.split('/')
df['authors_number'] = df['authors'].agg(len)

```

2. **Предобработка дат**

   С помощью специальной функции проверки были найдены и исправлены некорректные значения в столбце с датами публикации книг.

```
def incorrect_date(day, month, year):
  if day <= 0 or month <= 0 or year < 0:
      return True
  if month in (1, 3, 5, 7, 8, 10, 12) and day > 31:
      return True
  if month in (4, 6, 9, 11) and day > 30:
      return True
  if month == 2:
      if year % 400 == 0 or (year % 100 != 0 and year % 4 == 0):
          if day > 29:
              return True
      elif day > 28:
          return True
return False

for ind, row in df.iterrows():
  month, day, year = map(int, row['publication_date'].split('/'))
  if incorrect_date(day, month, year):
      print(row['bookID'], row['publication_date'])
```

   Затем данные в этом столбце были преобразованы к типу datetime Timestamp.

```
df['publication_date'] = pd.to_datetime(df['publication_date'], format='
   %m/%d/%Y')
```

3. **Предобработка столбца с языковыми кодами**

   В датафрейме используется несколько различных кодов для обозначения английского языка.

код	число вхождений
en-CA	7
en-GB	214
en-US	1409
eng	8911
enm	3

   Все они были заменены на просто 'eng'.

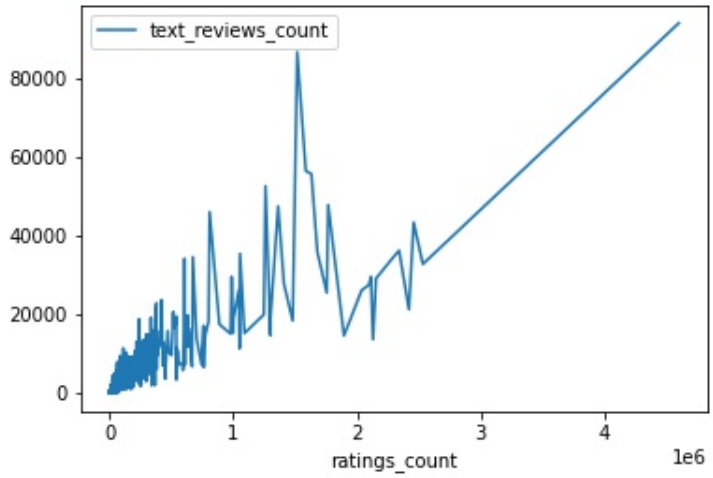
4. **Замена некорректных нулевых значений**

   В датафрейме есть книги, у которых число поставленных рейтингов равно нулю, в то время как средний рейтинг нулю не равен; помощью запроса к данным можно узнать, что таких книг 55. Вероятнее всего, это говорит об отсутствии данных. Для таких книг в столбец с числом проставленных рейтингов запишем медианное значение по числу проставленных рейтингов среди книг со средним рейтингом в том же диапазоне.

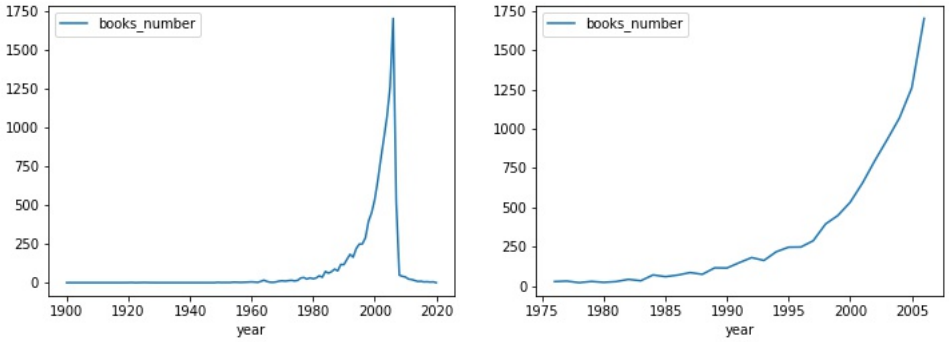
```
precision = 0.1
iter_range = np.arange(0.1, 5.1, precision)

median_ratings_count = dict()
for i in iter_range:
  i = round(i,1)
  value = df[(i - precision <= df['average_rating'])
    & (df['average_rating'] <= i)]['ratings_count'].median()
  if np.isnan(value):
    value = 0
  median_ratings_count[i] = round(value)

median_ratings_count = pd.Series(median_ratings_count)
median_ratings_count = median_ratings_count \
    .replace(0, round(median_ratings_count.mean()))

for _, row in df.iterrows():
  if row['ratings_count'] == 0 and row['average_rating'] != 0:
    rating = round(row['average_rating'] + 0.05, 1)
    df.loc[_, 'ratings_count'] = median_ratings_count[rating]
```

4 Гипотезы о данных

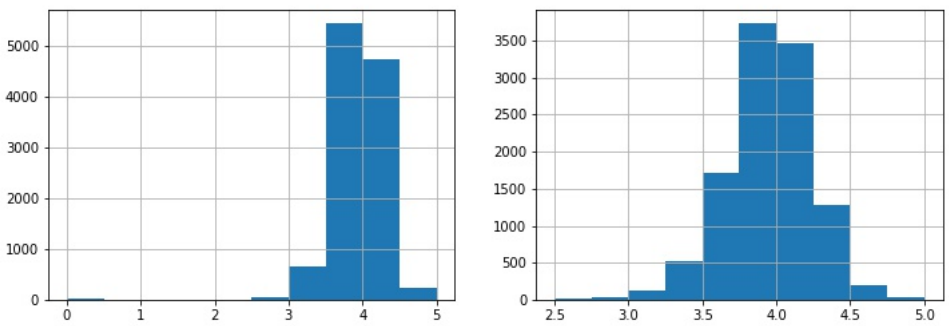
1. Столбцы ratings_count (сколько раз книге был поставлен рейтинг) и text_reviews_count (число текстовых отзывов на книгу) достаточно сильно коррелируют.



2. Подавляющее большинство книг, записи о которых хранятся в датасете, были изданы в 1975 - 2007 годах. Это можно видеть на графике изменения со временем числа изданных в данный год книг.



3. Средний рейтинг большей части книг лежит в промежутке от 3 до 5. Книги с рейтингом ниже трех можно исключить из анализа, нацеленного на предсказание рейтинга книги по ее остальным характеристикам.



4. Большинство книг написаны одним или двумя авторами.

5. Пользователи практически не читают книги длиной более 1500 страниц - см. график зависимости числа выставленных рейтингов от числа страниц. Данные об этих книгах можно исключить из рассмотрения.