

Университет ИТМО

## Практическая работа №3

По дисциплине “Визуализация и моделирование”

Автор: Лайок Олег Владимирович

**Поток:** 1\_2

**Группа:** K3242

**Факультет:** ИКТ

**Преподаватель:** Чернышева А.В.

Санкт-Петербург, 2021 г.

1. Описание датасета: топ 50 книг бестселлеров на сайте Amazon, в период с 2009 по 2019 год, с краткой характеристикой каждой книги в топе.

В изначальном датасете было разделение только на жанры художественной литературы и не художественной. Чтобы было интереснее анализировать книги попавшие в топ я решил вручную разделить художественную литературу по жанрам в столбце Subgenre (это было быстрее сделать вручную чем парсингом, т.к. не все книги есть на википедии и не у всех отмечен жанр). У не художественных книг в этом столбце остается категория Non Fiction.

2. Содержание датасета:

Название колонки в датасете	Данные	Тип данных
Name	Название книги	Текстовые
Author	Имя автора	Текстовые
User Rating	Рейтинг пользователей	Числовые
Reviews	Количество отзывов	Числовые
Price	Цена	Числовые
Year	Год, когда книга в топе	Числовые
Genre	Жанр (только два – fiction/non-fiction)	Текстовые
Subgenre	Поджанр	Текстовые

3. Проанализируем, какие данные мы ожидаем увидеть в датасете:

Name - название книги, строка, в датафрейме данное поле имеет тип object что является верным, интервала определенных значений не имеет

Author - имя автора, строка, в датафрейме данное поле имеет тип object что является верным, интервала определенных значений не имеет

User rating - средний рейтинг пользователей, число с плавающей точкой, в датафрейме данное поле имеет тип float64 что является верным, интервал значений (0.0,5.0]

Reviews - число отзывов на книгу, целое число, в датафрейме данное поле имеет тип int64 что является верным, заданного интервала значений не имеет

Price - цена книги, целое число, в датафрейме данное поле имеет тип int64 что является верным, интервал значений >0

Year - год в котором книга попала в топ, целое число, в датафрейме данное поле имеет тип int64 что является верным, диапазон значений {2009,2010,2011,2012,2013,2014,2015,2016,2017,2018,2019}

Genre - жанр книги, строка, в датафрейме данное поле имеет тип object что является верным, может принимать только значения Fiction/Non Fiction

Subgenre - поджанр книги, строка, в датасете имеет тип object что является верным, диапазон значений {'Thriller', 'Comedy', 'Non Fiction', 'Comics', 'Games', 'Young Adult Fiction', 'Historical', 'Children', 'Science fiction', 'Novel', 'Fantasy', 'Horror', 'Drama'}

4. Результаты проверки данных: (ссылка на код <https://colab.research.google.com/drive/1rDQMdvzQ0dIBZPkld2A5kmRbg4BNCgO1?usp=sharing>)
  - a. В поджанрах есть записи Young Adult fiction и Young Adult Fiction которые отличаются только регистром, приведем записи с этим полем к одному виду чтобы избежать повторений.
  - b. В столбце Genre находятся только два вида записей как и должно быть по описанию
  - c. В столбце User Rating была запись 4. . После проверки было выявлено что это то же самое что и 4.0 или 4, поэтому на данные никак не влияет
  - d. По результатам проверки данные в столбцах User Rating и Year находятся в области допустимых значений.
  - e. В столбце Price были найдены записи со значением 0, что противоречит логике. Однако после проверки данных на сайте Amazon я выяснил, что данная цена реальна и в датасете указывается самая дешевая стоимость (а т.к. самая дешевая стоимость в приложении Amazon с электронными книгами, то некоторые книги действительно бесплатные) .
  - f. Датасет был проверен на наличие пустых записей – таких записей обнаружено не было.

## 5. Гипотезы:

- 1) Детская литература будет доминировать по количеству книг попавших в топ среди художественной литературы (во время проставления жанров попало очень много детских книг, нужно проверить их процентное соотношение)
- 2) Средняя цена книг будет около 5\$ (т.к. в ходе этой работы обнаружилось, что в датасете хранится минимальная стоимость книги, нужно уточнить какая именно средняя минимальная цена у данных книг)
- 3) Средняя цена книг будет сильно отличаться от средней цены книги каждого поджанра (т.к. по логике нехудожественные книги и книги для подростков продают дороже)
- 4) Т.к. появилась категория поджанров нужно проанализировать средний рейтинг в каждом поджанре, предварительное предположение: у детской литературы рейтинг максимальный, у Young Adult Fiction минимальный.
- 5) Аналогичная четвертому пункту гипотеза, но в отношении числа рецензий.

## 6. Вывод:

В ходе работы были проанализированы данные на их корректность, аномалий не обнаружено. Было добавлено новый столбец Subgenre что позволило сформулировать новые гипотезы в отношении данных в датасете.