МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ

РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное автономное образовательное учреждение высшего образования

«Омский государственный технический университет»

Факультет информационных технологий и компьютерных систем

Кафедра «Прикладная математика и фундаментальная информатика»

|  |  |
| --- | --- |
| **Домашнее задание** | |
|  | |
| по дисциплине | Практикум по программированию |

|  |  |  |  |
| --- | --- | --- | --- |
| Студента | Портнягина Алексея Владимировича | | |
|  | фамилия, имя, отчество полностью | | |
| Курс | 2 | Группа | ФИТ-222 |
| Направление | 02.03.02. Фундаментальная информатика и информационные технологии | | |
|  | код, наименование | | |
| Руководитель | Старший преподаватель | | |
|  | должность, ученая степень, звание | | |
| Саматов А.П | | |
|  | фамилия, инициалы,  дата, подпись | | |
| Выполнил | 08.12.2023 Портнягин А.В | | |
|  | дата, подпись студента | | |

|  |  |
| --- | --- |
| Итоговый рейтинг |  |

Омск 2023

**Введение**

Анализ данных – это функция внутренней организации, выполняемая аналитиками посредством представления цифр руководству. Он включает более детальный подход к регистрации, анализу, распространению и представлению результатов данных таким образом, чтобы их было легко интерпретировать и принимать решения для бизнеса.

С помощью анализа данных вы сможете принимать решения о тенденциях клиентов и прогнозировании поведения, повышая прибыльность бизнеса и способствуя эффективному принятию решений.

Используя анализ данных, вы можете определить, какие формы рекламы эффективно охватывают ваших клиентов, и оказывать влияние, которое заставит их покупать ваши продукты. Данные позволяют понять, какие методы рекламы вашего продукта оказывают наибольшее влияние на целевую аудиторию, и в каком масштабе вы можете принять такую рекламу.

Анализ данных также даёт вам приблизительное представление о будущих тенденциях в поведении потребителей, которые позволят вам создавать футуристические изобретения для ваших продуктов. Таким образом, вы можете производить продукты и создавать услуги, которые ставят вас на вершину вашей отрасли.

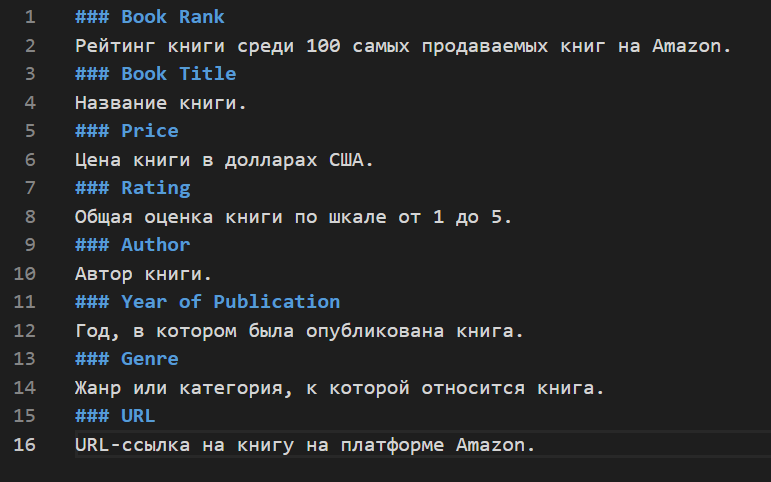
Анализ данных принесёт вам много пользы, если вы хотите иметь успешный и эффективный бизнес. С хорошей системой анализа данных, вы можете определить сектора вашего бизнеса, которые используют ненужные финансы и области, которые требуют большего финансирования. Благодаря этому у вас будет чёткое представление о том, где вы должны сократить расходы, и какие технологии вы собираетесь использовать для сокращения эксплуатационных и производственных затрат.

Для изучения анализа данных я использовал следующие библиотеки: numpy - библиотека, которая помогает работать с матрицами и таблицами с большой скоростью, pandas - библиотека на основе numpy, которая даёт расширенные возможности действий с таблицами, matplotlib и seaborn - библиотеки для отображения графиков и результатов работы моделей машинного обучения, sklearn - библиотека, содержащая множество моделей МО для различных задач.

**Основная часть**

1. Поиск и загрузка данных

Я выбрал датасет “Top-100 trending books”, который находится на сайте Kuggle и который содержит следующие столбцы:



Загружал датасет при помощи функции библиотеки pandas.read\_csv(path).

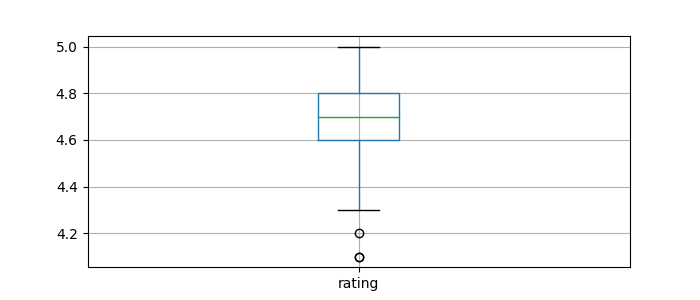
Выглядит он примерно так.



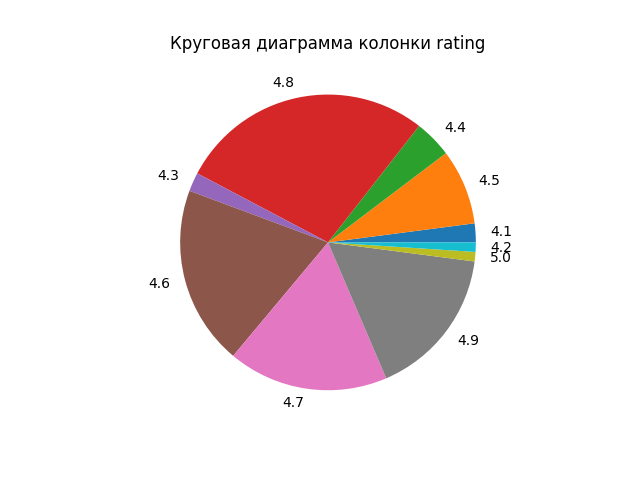
1. Разведывательный анализ данных



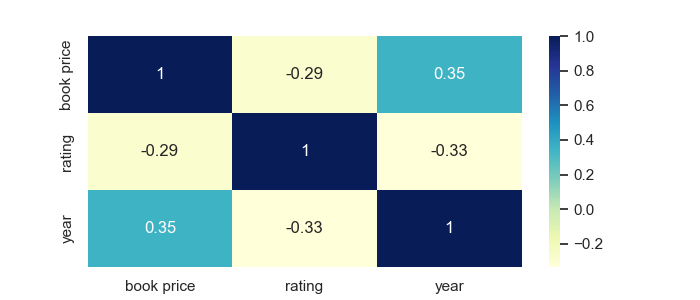
Гистограмма показывает, сколько книг по какой цене продаётся. По ней видно, что большинство книг стоимостью от 2 до 20 долларов, а максимальная цена доходит до 50.



“Ящик с усами” показывает, в каком интервале располагаются данные в столбце рейтинга и сколько данных выходит за нормальное распределение, то есть являются выбросами.



Круговая диаграмма показывает, в каком соотношении находятся рейтинги книг. По данным видно, что большинство книг имеют рейтинг 4.8.



Это тепловая карта со значениями взаимной корреляции. Показывает, насколько одни столбцы данных зависят от других

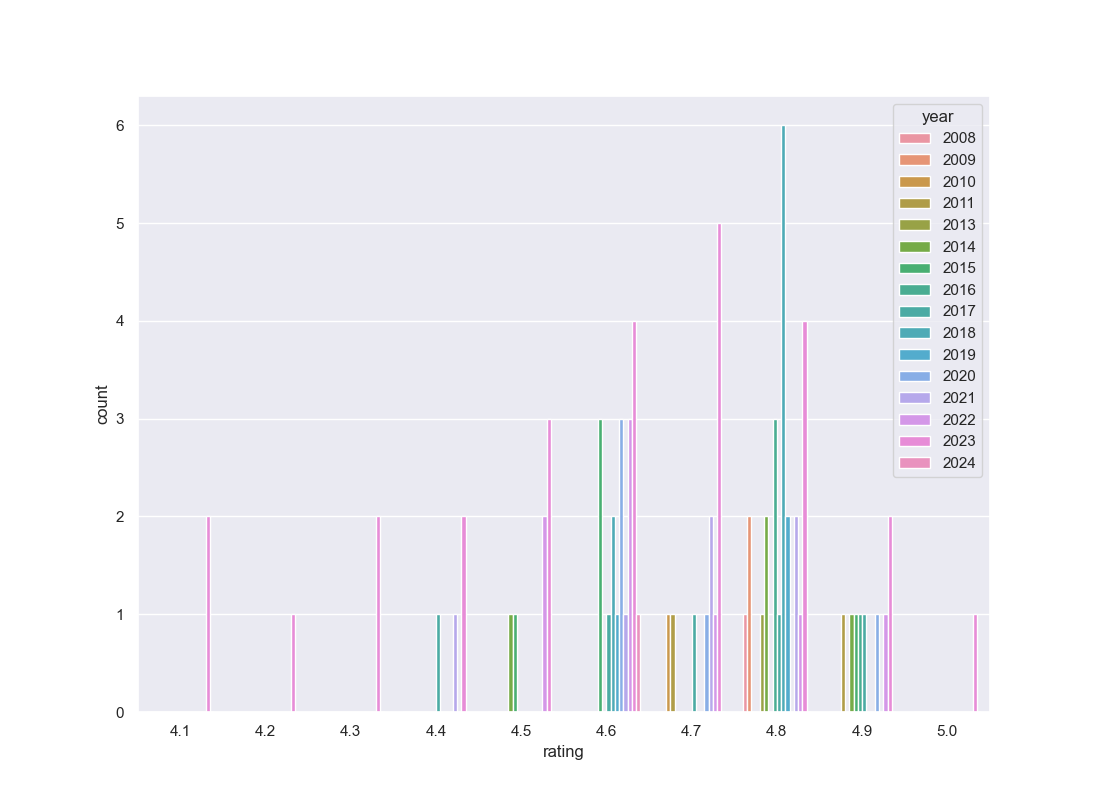
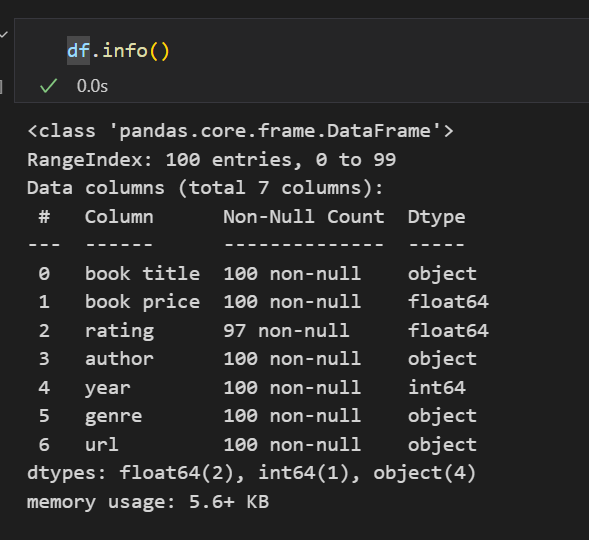
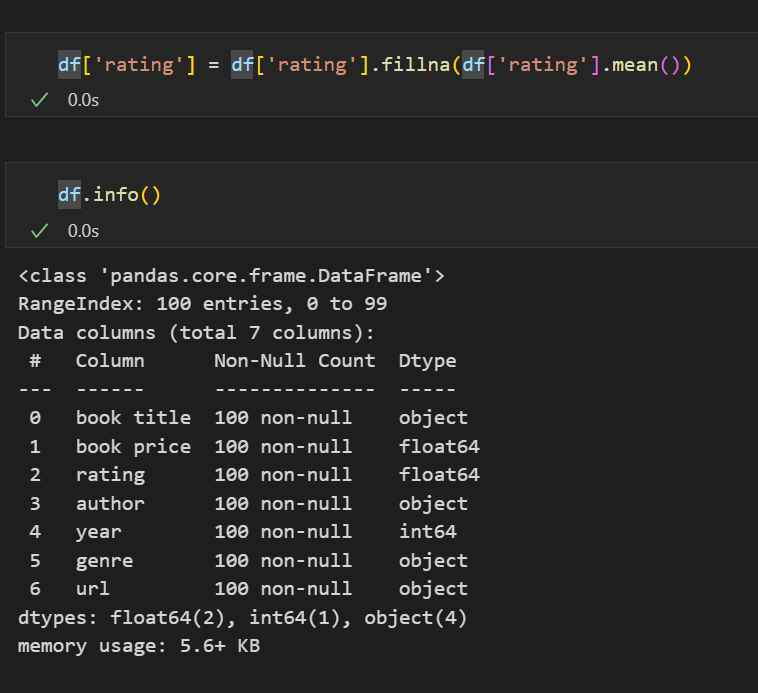


Диаграмма группирует данные по рейтингу и году выпуска. По ней видно, что в данных по численности больше всего книг с рейтингом 4.8 и годом выпуска 2017.

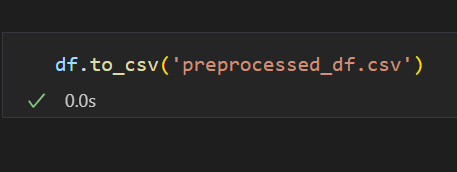
1. Предварительная обработка данных

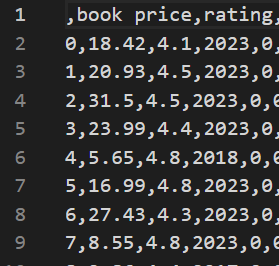


Проверка пропусков осуществлялась с помощью функции df.info(), которая показывает пропущенные значения в столбцах. Как можно видеть, в столбце rating отсутствуют 3 значения.



Пропущенные значения заполнялись средним значениям по столбцу rating. После заполнения таблица с данными больше не содержит пропусков.





Предобработанный датасет сохранял при помощи функции df.to\_csv()

**Заключение**

В рамках задания был загружен датасет с сайта Kaggle, было проведено исследование, какие данные находятся в датасете и имеются ли в данных пропуски.

Далее была произведена работа с данными, изменение наименования данных и удаление лишних столбцов. Были выведены различные графики, показывающие, какие данные располагаются в датасете и как коррелируют между собой.

Под конец была произведена работа по устранению пропусков в данных и избавлению от категориальных данных при помощи one-hot кодирования, а также проверка на нормальное распределение данных в столбцах.