

Департамент образования и науки города Москвы
Государственное автономное образовательное учреждение
высшего образования города Москвы
«Московский городской педагогический университет»
Институт цифрового образования
Департамент информатики, управления и технологий

ДИСЦИПЛИНА:

«Инструменты для хранения и обработки больших данных»

Индивидуальное задание

Тема:

«01-1 Визуализация данных из CSV-файла.».

Выполнила: Мальчевская П.А., АДЭУ-201

Преподаватель: Босенко Т.М.

Москва

2023

Перед началом работы обрабатывает данные и чистим их(удаляем пустые значение, меняем столбцы)

```
[ ]: import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import missingno as msno #missing map

[ ]: books = pd.read_csv('../input/datasciencebook/DataScience_book.csv', encoding='unicode_escape')

[ ]: books.head()

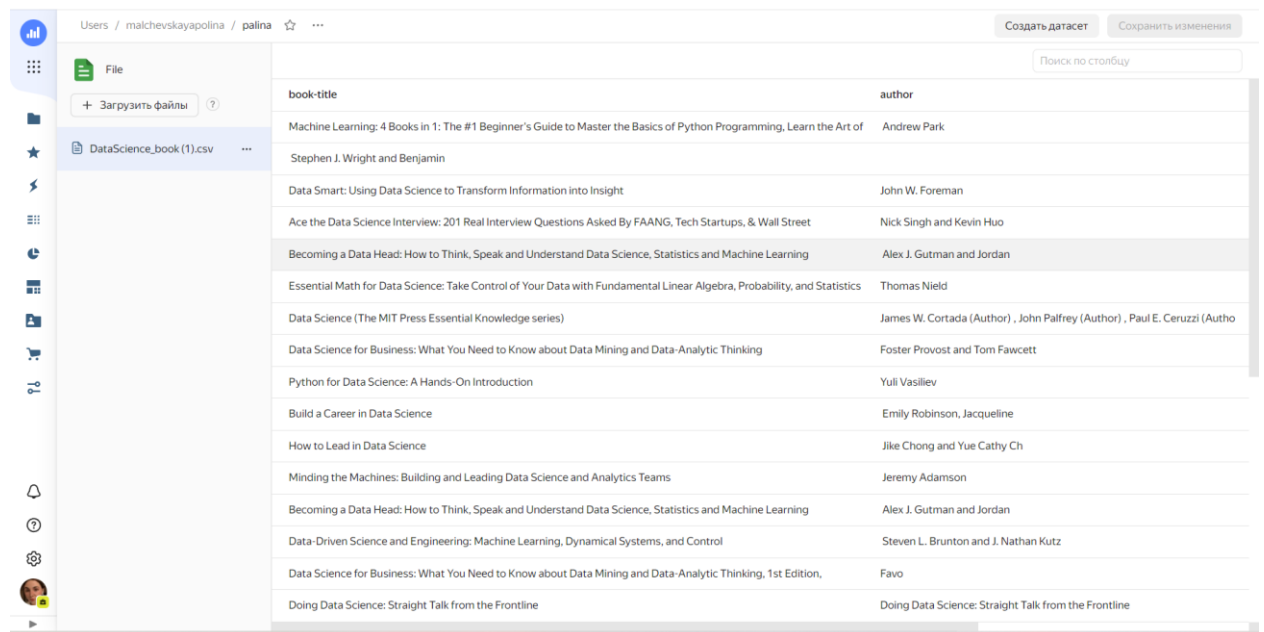
[ ]: books = books.fillna('Stephen J. Wright and Benjamin')

[ ]: books["book-title"] = np.where(books["book-title"] == '\xa0Stephen J. Wright\xa0and\xa0Benjamin',
                                "Optimization for Data Analysis", books["book-title"])

[ ]: books.head()

[ ]: books["book-title"].value_counts()
```

Я загрузила данные в YandexDataLens и создала подключение










Users / malchevskayapolina / palina		Создать датасет	Сохранить изменения
File		Поиск по столбцу	
+ Загрузить файлы ?			
DataScience_book (1).csv			
book-title	author		
Machine Learning: 4 Books in 1: The #1 Beginner's Guide to Master the Basics of Python Programming, Learn the Art of	Andrew Park		
Stephen J. Wright and Benjamin			
Data Smart: Using Data Science to Transform Information into Insight	John W. Foreman		
Ace the Data Science Interview: 201 Real Interview Questions Asked By FAANG, Tech Startups, & Wall Street	Nick Singh and Kevin Huo		
Becoming a Data Head: How to Think, Speak and Understand Data Science, Statistics and Machine Learning	Alex J. Gutman and Jordan		
Essential Math for Data Science: Take Control of Your Data with Fundamental Linear Algebra, Probability, and Statistics	Thomas Nield		
Data Science (The MIT Press Essential Knowledge series)	James W. Cortada (Author) , John Palfrey (Author) , Paul E. Ceruzzi (Autho		
Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking	Foster Provost and Tom Fawcett		
Python for Data Science: A Hands-On Introduction	Yuli Vasiliev		
Build a Career in Data Science	Emily Robinson, Jacqueline		
How to Lead in Data Science	Jike Chong and Yue Cathy Ch		
Minding the Machines: Building and Leading Data Science and Analytics Teams	Jeremy Adamson		
Becoming a Data Head: How to Think, Speak and Understand Data Science, Statistics and Machine Learning	Alex J. Gutman and Jordan		
Data-Driven Science and Engineering: Machine Learning, Dynamical Systems, and Control	Steven L. Brunton and J. Nathan Kutz		
Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking, 1st Edition,	Favo		
Doing Data Science: Straight Talk from the Frontline	Doing Data Science: Straight Talk from the Frontline		

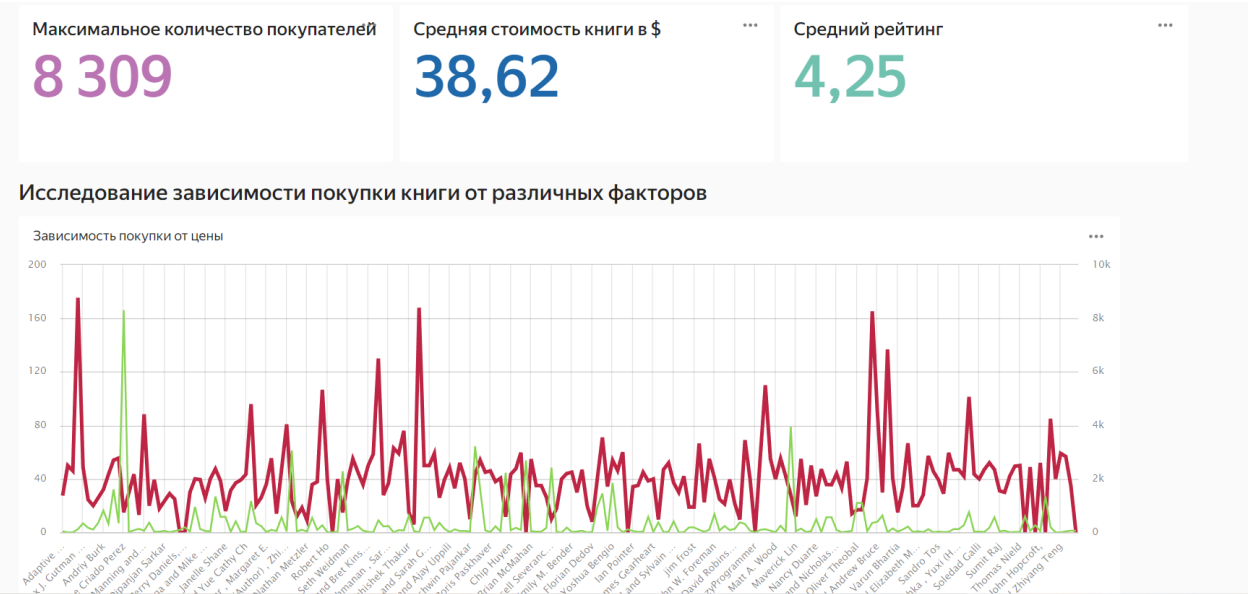
Далее я создала датасет и агрегировала данные

#	Имя ↓	Источник поля ↓	Тип ↓	Агрегация ↓	Описание ↓
1	название	csv.book-title	Строка	Нет	
2	автор	csv.author	Строка	Нет	
3	рейтинг	csv.star	Дробное число	Нет	
4	рейтинг avg	csv.star	Дробное число	Среднее	
5	покупатели	csv.buyers	Целое число	Нет	
6	покупатели sum	csv.buyers	Целое число	Сумма	
7	покупатели avg	csv.buyers	Целое число	Среднее	
8	покупатели max	csv.buyers	Целое число	Максимум	
9	покупатели min	csv.buyers	Целое число	Минимум	
10	формат	csv.cover	Строка	Нет	
11	цена \$	csv.price_	Дробное число	Нет	

Далее на основе датасета и новых показателей я создала несколько чартов, которые в дальнейшем добавила на дашборд

<input type="checkbox"/>		Зависимость покупки от цены	malchevskayap...	malchevskas...	16.02.23
<input type="checkbox"/>		Зависимость покупки от рейтинга	malchevskayap...	malchevskas...	16.02.23
<input type="checkbox"/>		Зависимость стоимости книги от ее формата	malchevskayap...	malchevskas...	16.02.23
<input type="checkbox"/>		Выбор формата книги	malchevskayap...	malchevskas...	16.02.23
<input type="checkbox"/>		Таблица	malchevskayap...	malchevskas...	16.02.23
<input type="checkbox"/>		ТОП-10 авторов	malchevskayap...	malchevskas...	16.02.23
<input type="checkbox"/>		ТОП-10 книг	malchevskayap...	malchevskas...	16.02.23

Затем я собрала чарты на дашборде, добавила текст, заголовки и селекторы.



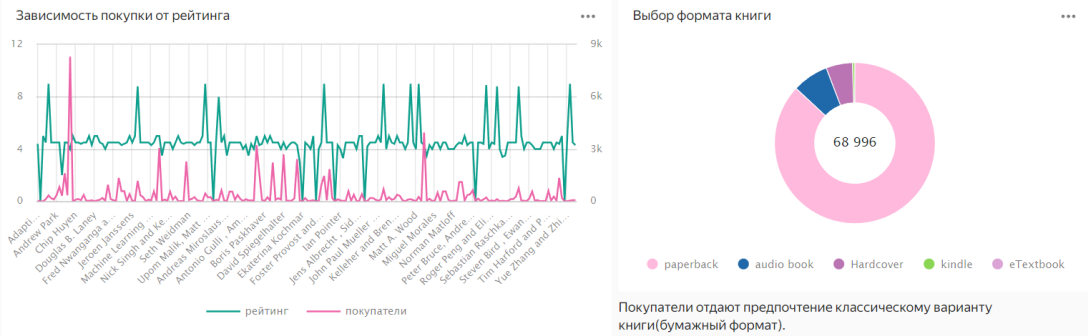
Исследование зависимости покупки книги от различных факторов

Зависимость покупки от цены



На графике изображена зависимость покупки книги от цены. Можно заметить, что в большинстве случаев, чем меньше стоимость книги, тем более популярна она у покупателей. Однако теория подтверждается не во всех случаях.

Выдвинем теорию: покупка книги зависит от ее рейтинга. На графике показана зависимость, которая опровергает нашу теорию. Практически в 100% случаев покупка не зависит от рейтинга=>имеют влияние другие факторы.



Выбор покупателей(топ-5)

автор	название	покупатели	рейтинг	цена \$
Caroline Criado Perez	Invisible Women: Data Bias in a World Designed for Men	8 309	4,00	14,99
Maurice J. Thompson	Python: - The Bible- 3 Manuscripts in 1 book: -Python Programming For Beginners -Python Programming For	3 950	4,30	27,97
Aurélien Giron	Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build	3 220	5,00	43,27

Выбор покупателей(топ-5)

автор	название	покупатели	рейтинг	цена \$
Caroline Criado Perez	Invisible Women: Data Bias in a World Designed for Men	8 309	4,00	14,99
Maurice J. Thompson	Python: - The Bible- 3 Manuscripts in 1 book: -Python Programming For Beginners -Python Programming For	3 950	4,30	27,97
Aurélien Giron	Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build	3 220	5,00	43,27
Martin Kleppmann	Designing Data-Intensive Applications: The Big Ideas Behind Reliable, Scalable, and Maintainable Systems	3 065	5,00	23,37
David Spiegelhalter	The Art of Statistics: How to Learn from Data	2 693	4,50	0,00

Изучив ТОП-5 книг по количеству покупок, мы можем заметить, что самая популярная книга имеет наименьшую стоимость в сравнении с оставшимися топовыми изданиями(книга под номером 5 не будет учитываться, так как формат книги-аудио,а средняя стоимость аудио-книги в данных- 0\$). Также изучим рейтинг финалистов, чтобы опровергнуть 2 теорию. Заметно, что самая популярная книга имеет в сравнении с оставшимися изданиями самый низкий рейтинг, однако 3-4 место занимают книги с наивысшим рейтингом.

Поиск по интересующий параметрам

Цена до

рейтинг формат

Итоговый результат в YandexDataLens: <https://datalens.yandex/rns3er21a6yuh>