Московский государственный технический университет

им. Н.Э. Баумана

Факультет «Информатика и системы управления»

Кафедра «Системы обработки информации и управления»



**Лабораторная работа № 1**

**По курсу «методы машинного обучения в АСОИУ»**

**«Создание «истории о данных» (Data Storytelling)»**

**Выполнил:**

студент ИУ5-24М

Ширшов А.С.

**Проверил:**

Гапанюк Ю.Е.

Подпись:

29.02.2024

Москва, 2024

**Описание задания**

* Выбрать набор данных (датасет). Вы можете найти список свободно распространяемых датасетов [здесь.](https://github.com/ugapanyuk/courses_current/wiki/DSLIST)
* Для лабораторных работ не рекомендуется выбирать датасеты очень большого размера.
* Создать "историю о данных" в виде юпитер-ноутбука, с учетом следующих требований: история должна содержать не менее 5 шагов (где 5 - рекомендуемое количество шагов). Каждый шаг содержит график и его текстовую интерпретацию.
* На каждом шаге наряду с удачным итоговым графиком рекомендуется в юпитер-ноутбуке оставлять результаты предварительных "неудачных" графиков.
* Не рекомендуется повторять виды графиков, желательно создать 5 графиков различных видов.
* Выбор графиков должен быть обоснован использованием методологии data-to-viz. Рекомендуется учитывать типичные ошибки построения выбранного вида графика по методологии data-to-viz. Если методология Вами отвергается, то просьба обосновать Ваше решение по выбору графика.
* История должна содержать итоговые выводы. В реальных "историях о данных" именно эти выводы представляют собой основную ценность для предприятия.
* Сформировать отчет и разместить его в своем репозитории на github.

**Ход работы**

За основу работы был выбран dataset, содержащий набор самых продаваемых книг за всё время. Ниже на рисунках представлен ход работы.

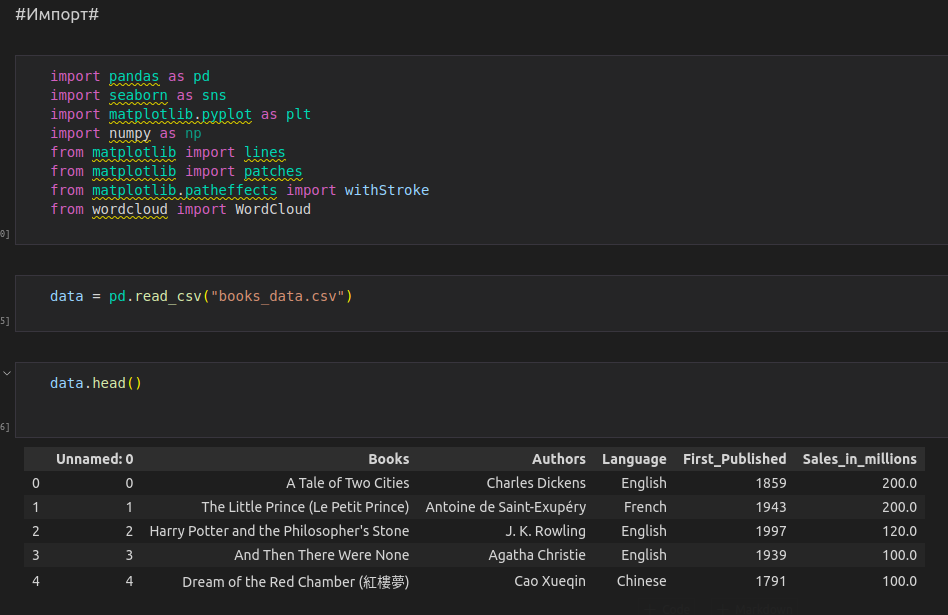
****

Рисунок 1 - Чтение данных из датасета

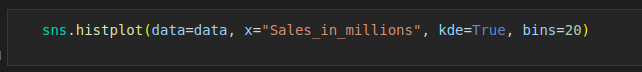


Рисунок 2 - Код для построения графика кол-ва книг от миллионов продаж

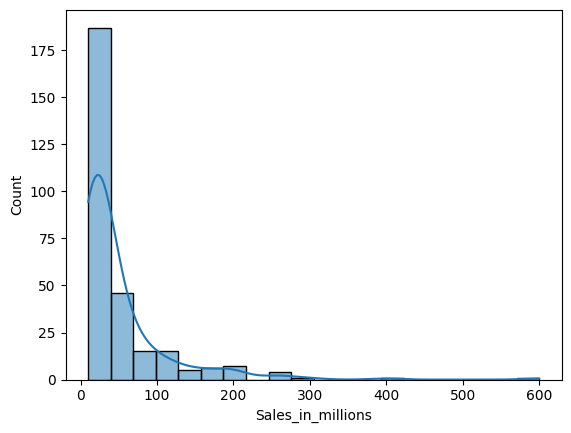
****

Рисунок 3 - График кол-ва книг от миллионов продаж

Далее был построен график кол-ва бестселлеров для каждой страны. Ниже на рисунках представлены как удачные, так и неудачные попытки. Код в отчёте опущен, так как представляет собой несколько ячеек ipynb.

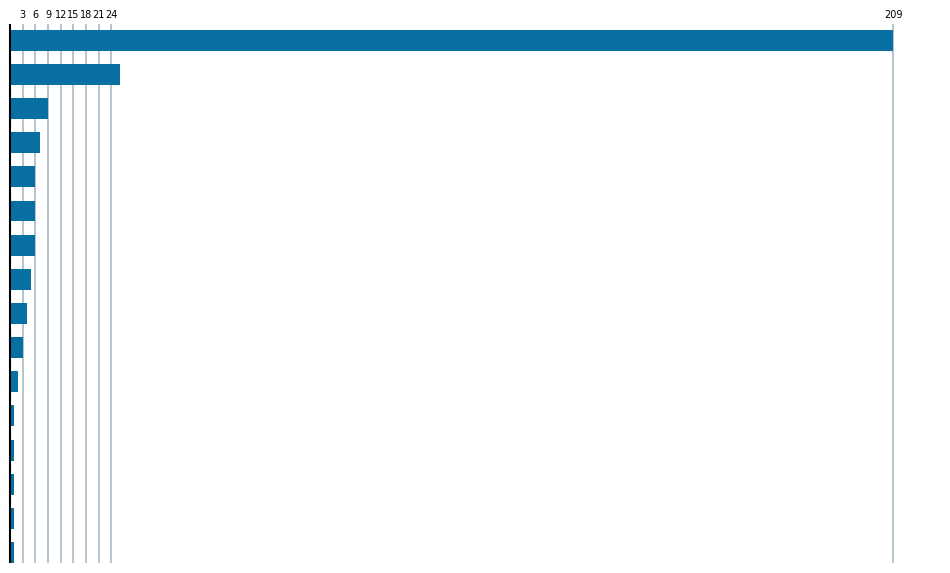
****

Рисунок 4 - Пример неудачного графика - Кол-во бестселлеров по странам

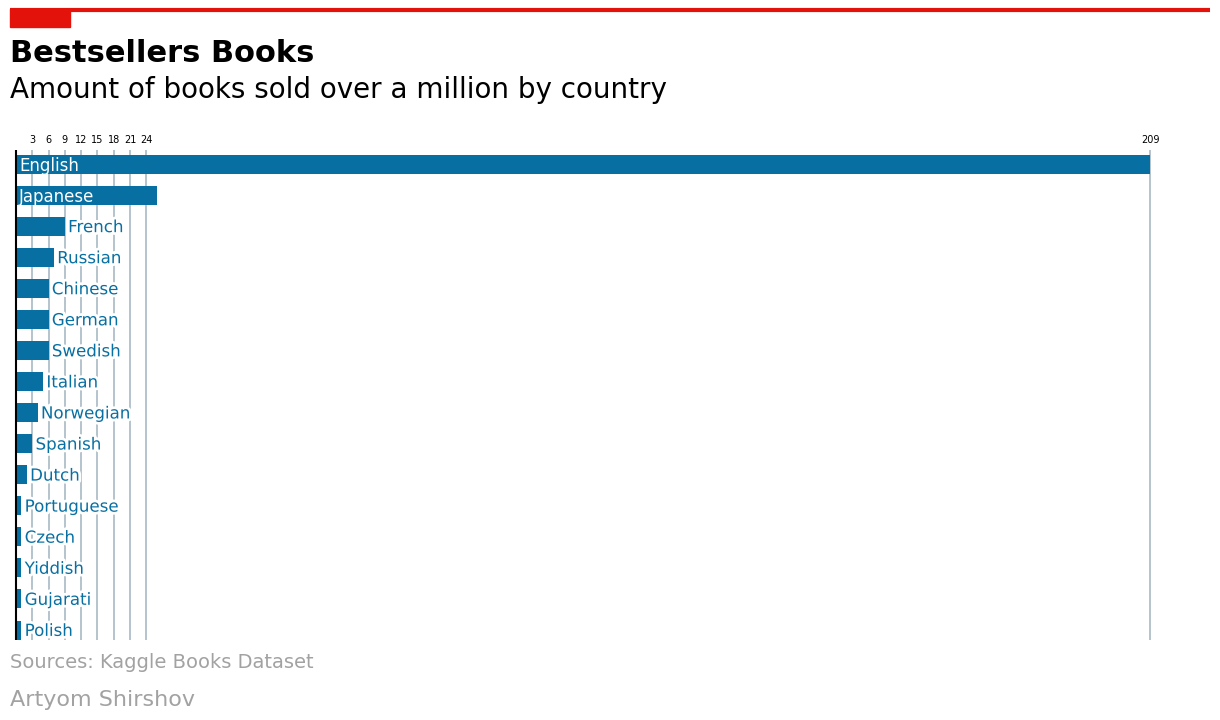


Рисунок 5 - Пример итогового удачного графика - Кол-во бестселлеров по странам

Далее на рисунке представлен график успешных книг в зависимости от столетия, для этого была сделана выборка книг, входящих в данное столетие по информации о дате выхода книги из набора данных.

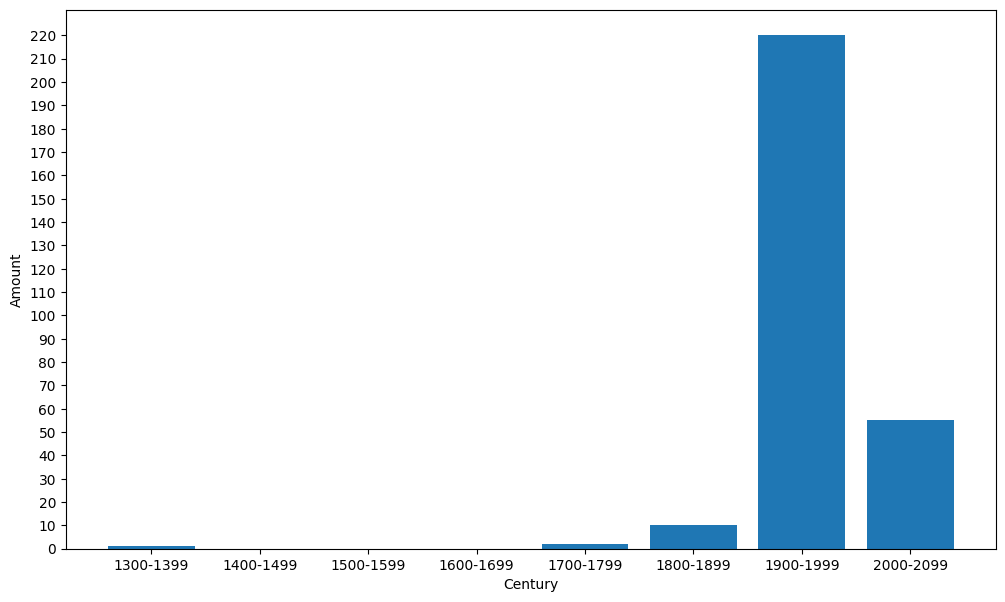


Рисунок 6 - График кол-ва бестселлеров по столетиям

Также для наглядной визуализации был сформирован график в виде облака слов. Ниже представлены различные варианты облака слов, на которых изображены, самые часто упоминаемые авторы.

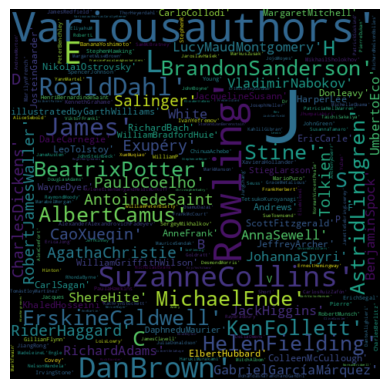


Рисунок 7 - Первый вариант облака слов



Рисунок 8 - Второй вариант облака слов

**Вывод**

В ходе данной работы был выбран и исследован набор данных о книгах. Была создана "история о данных" в виде юпитер-ноутбука, содержащая не менее 5 шагов, каждый из которых включает в себя график и его текстовую интерпретацию.

Было создано 5 графиков различных видов. Выбор графиков был обоснован использованием методологии data-to-viz, которая позволяет избежать типичных ошибок при построении графиков.

В итоге, "история о данных" содержит ценные выводы, которые могут быть использованы для дальнейшего анализа и принятия решений. Данная работа позволила пройтись через процесс анализа данных, начиная с выбора набора данных и заканчивая получением ценных выводов.