**Министерство образования и науки Российской Федерации**

**МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ им. Н.Э. БАУМАНА**

**Кафедра «Информатика и системы управления» (ИУ5)**

**ДИСЦИПЛИНА: «Технологии машинного обучения»**

Отчет по лабораторной работе №1

«Разведочный анализ данных. Исследование и визуализация данных»

Выполнила:

Студентка группы ИУ5-61Б

Мартынова Д.П.

Преподаватели:

Гапанюк Ю.Е.

Москва, 2020 г.

**Цель лабораторной работы:** изучение различных методов визуализация данных.

**Задание:**

* Выбрать набор данных (датасет).
* Для первой лабораторной работы рекомендуется использовать датасет без пропусков в данных.
* Создать ноутбук, который содержит следующие разделы:

1. Текстовое описание выбранного Вами набора данных.
2. Основные характеристики датасета.
3. Визуальное исследование датасета.
4. Информация о корреляции признаков.

* Сформировать отчет и разместить его в своем репозитории на github.

**Выполнение ЛР:**

1. ***Текстовое описание набора данных***

Используемый **dataset: TED Talks** - Data about TED Talks on the TED.com website until September 21st, 2017.

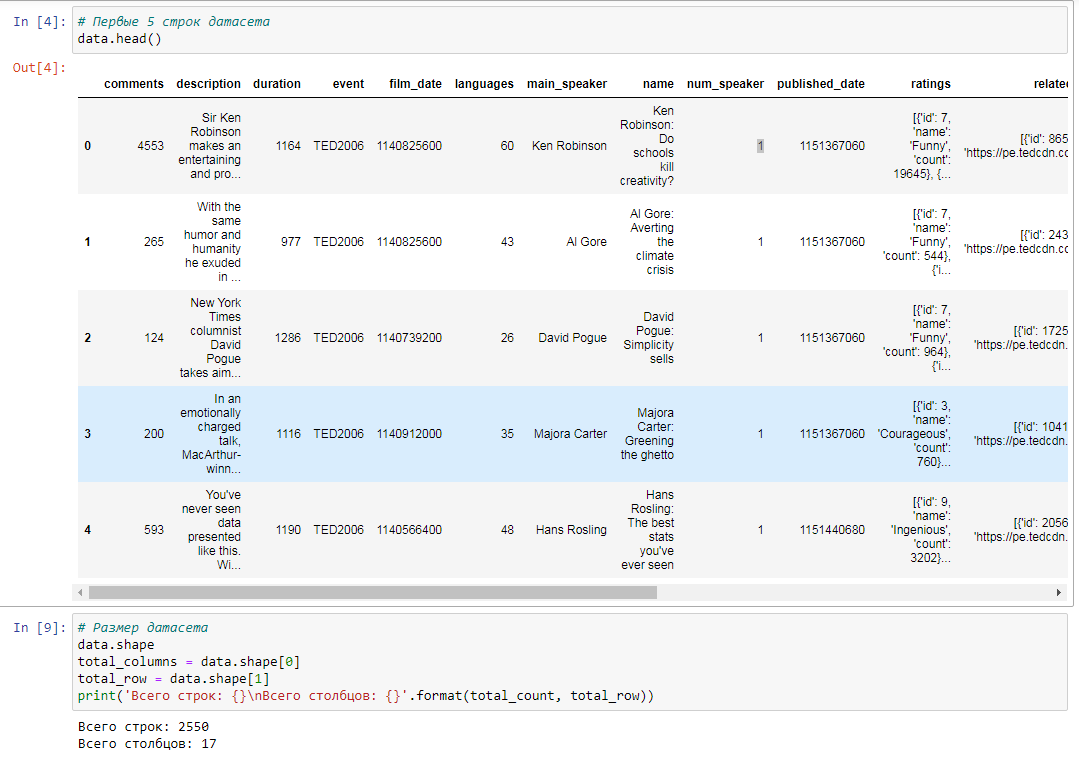
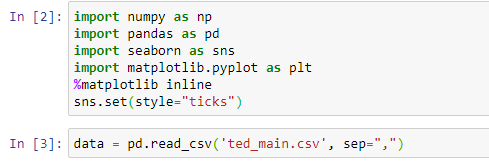
### *Контекст*

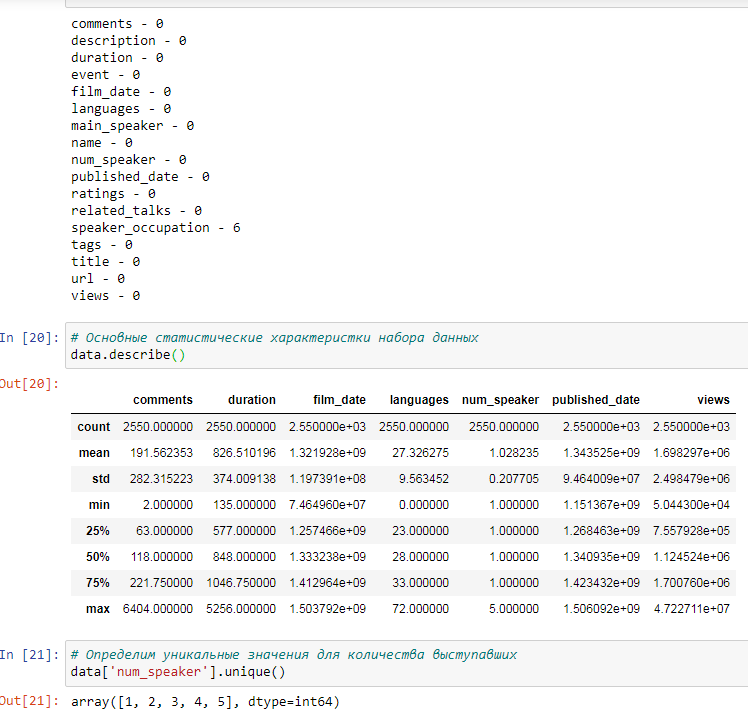
Данный набор данных содержит информацию обо всех аудио-видеозаписях выступлений TED Talks, загруженных на официальный сайт TED.com до 21 сентября 2017 года. Основной набор данных содержит информацию обо всех выступлениях, включая количество участников, количество комментариев, описание, спикеров и название.

*Dataset состоит из следующих данных:*

* Comments – Количество комментариев первого уровня, сделанных в ходе выступления
* Description – Описание того, о чем выступление
* Duration – Продолжительность выступления в секундах
* Event – Событие The TED/TEDx, когда состоялось выступление
* Film\_date – Отметка времени начала съемок в формате Юникс
* Languages – Количество языков, на которых доступно выступление
* Main\_speaker – Основной спикер выступления
* Name – Официальное название выступления: включает название и имя спикера
* Num\_speaker – Количество спикеров в выступлении
* Published\_date – Временная метка в формате Юникс публикации выступления на официальном сайте
* Ratings – строковый словарь различных рейтингов, присвоенных разговору
* Related\_talks – Список выступлений, рекомендуемых к просмотру
* Speaker\_occupation – Род занятия главного спикера
* Tags – Темы, связанные с выступлением
* Title – Название выступления
* url – Ссылка на выступление
* views – Количество просмотров выступления

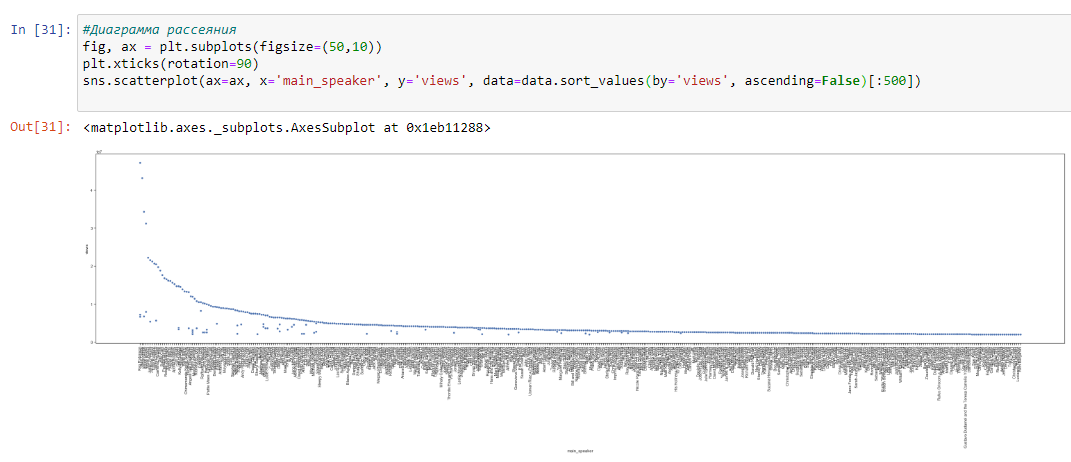
1. ***Основные характеристики датасета***

***Загрузка данных и импорт библиотек:***

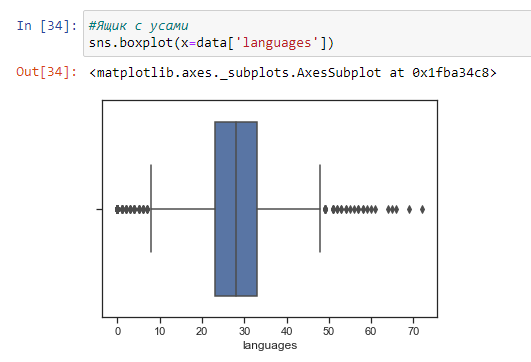


1. ***Визуальное исследование датасета***

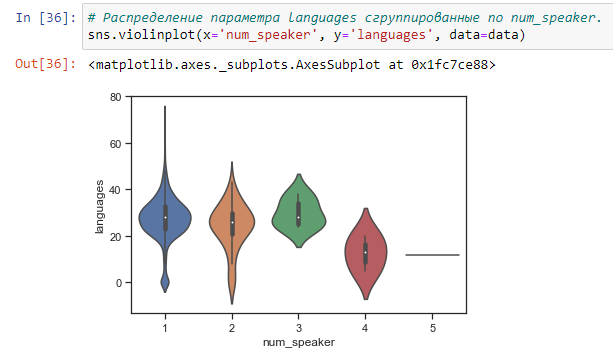
* Диаграмма рассеяния

На данной диаграмме можно увидеть зависимость числа просмотров от спикера. График построен на основе отсортированного датасета по убыванию по просмотрам. Взяты первые 500 записей.

* Ящик с усами



* [Violin plot](https://en.wikipedia.org/wiki/Violin_plot)



1.  ***Информация о корреляции признаков***