

Analysis of the database of the service for reading books by subscription

Project status: *completed*

Libraries used: pandas, matplotlib, numpy, sys, sqlalchemy

Description of the project:

Our company decided to buy a large service for reading books by subscription: due to the coronavirus, city residents began to go out less, visit cafes, cinemas, shopping centers and other public places. However, now there is more time for books, which has pushed many startups create apps for those who love to read.

Our first task is to analyze the purchased service database. It contains information about books, publishers, authors, as well as custom book reviews. This data will help formulate a value proposition for a new product.

The study will take place in three stages:

1. Explore the database tables.
2. Using SQL queries, let's see:
 - count how many books were published after January 1, 2000;
 - for each book, we calculate the number of reviews and the average rating;
 - we will determine the publishing house that has released the largest number of books thicker than 50 pages - so we will exclude brochures from the analysis;
 - determine the author with the highest and lowest average rating of books - taking into account only books with 50 or more ratings;
 - calculate the average number of reviews from users who have given more than 50 ratings.
3. Let us describe the conclusions based on the results of acquaintance with the database.