Task for the final work

As a dataset, we use the data of a conditional online cinema. We have information about users, films, as well as the ratings that users have given to a particular film. As part of the work, we want to conduct a study of the current situation and solve a business case with recommendations of films to users (obviously, not all users have watched all the films and we need to somehow recommend to the user what to watch next)

The dataset is available here - https://grouplens.org/datasets/movielens/100k/ Description here - http://files.grouplens.org/datasets/movielens/ml-100k-README.txt

The work consists of 3 parts:

• Google Sheets practice • Python
practice • Theoretical part

For each part, you can get credit or not credit. To get credit for work, it is enough to get credit for two of the three parts.

THE ASSIGNMENTS ARE LOCATED ON THE FOLLOWING PAGES OF THIS DOCUMENT.

Practice Google Sheets

Download the MovieLens 100K dataset to your computer
https:// grouplens.org/ datasets/ movielens/ 100k/

In this part, we act as a bi-analyst and want to present the customer with a general information on films, users, as well as build the top active users for the last 3 months for motivation.

In order to upload data to Google Sheets, you need to rename the u.user file to a file named u.user.csv. The symbol | | is used as separators there.

Number	Exercise	Points
one	Build a histogram of users by age	ten
2	Plot 2 graphs showing distribution of people by occupation depending on their gender	ten

Similarly, we want to look at the data on films. In order to upload data to Google Sheets needs to rename the u.item file to a file named u.item.csv. As separators there is used symbol |

Number	Exercise	Points
3	Plot the number of films by genres	ten
four	Plot the number of films by years	ten

Finally, we want to find the most active users of our portal for motivation. In order to upload data to Google Sheets, you need to rename u.data file to a file named u.data.csv. It uses as delimiters tab character

Number	Exercise	Points
5	Plot the number of ratings by	ten

	months and years (conversion of timestamp to date see here https://stackoverflow.com/questions/45227380/co nvert-unix-epoch-time-to-date-in-google-sheets)	
6	Reveal the top-5 most active users (most ratings) in the last 3 months	ten

Total - a maximum of 60 points. You must score at least 40 to qualify.

Practice Python

In this section, we will act as a data scientist and try to build a simple
a model for recommending movies to users.

Number	Exercise	Points
one	Upload ratings files to colab and films (movies) and create based on them pandas-dataframes	ten

Having formed the overall top films in the past practice, we want to take a step forward and start advising the user on those films that could be most suitable for him interesting. Our goal is to learn how to predict the user's rating of a movie. For testing the model, we will find the user who gave the most ratings

Number	Exercise	Points
2	With Pandas using dataframe ratings, find the user id, with the most ratings	ten

We will select the films that this user has rated

Number	Exercise	Points
3	Leave in the ratings dataframe only those movies that rated this user	ten

To build a model, we need features. As such, we will use:

- Release year
- Genres
- Total number of ratings
- Total score

Number	Exercise	Points
four	Add 3 columns to the dataframe from the job: • By genre. Each column is a genre. We write down the unit if the film belongs to this genre and 0 - if No	ten

	• columns with total ratings from all users per movie and total score from all users	

Now everything is ready and you can build a model

Number	Exercise	Points
5	Form X_train, X_test, y_train, y_test	ten
6	Take a linear regression model (or any other for the regression problem) and educate her on films	ten
7	Evaluate the quality of the model on X_test, y_test at help of metrics for the regression problem	ten

The second part of Python practice is related to Spark

eight	Upload data to spark	ten
9	By means of spark, display the average rating for each movie	twenty
ten	Calculate the average score for each genre	twenty
eleven	In a spark, get 2 dataframes with the 5th most most popular and most unpopular films (according to the number of ratings, or the rating itself - your choice)	twenty

Total - a maximum of 140 points. You need to score 100 to qualify.

Theoretical part

You are the leader according to the average Internet cinema viewing volume. Your the task is to develop a strategy for implementing a data warehouse and working with large data in this company. Tasks:

Number	Task	Points
one	Describe the main business reports (2-3 pieces), which we want to see in our business	ten
2	Describe the main data available and sources of their income	twenty
3	Describe the main entities in the repository data (star schema) and pouring process data	twenty
four	Describe basic quality checks data (10 pieces), which we will use when pouring	ten
5	Come up with a Data project that should improve your business performance and paint it with Crisp-DM	twenty
6	Describe the required roles in the work team with data in steps 4 and 5	thirty
	Total	110

Total - maximum 110. To pass, you need to dial 90