What is required:

- To create an account at Kaggle.com (Machine Learning and Data Science Community);
- To download following public dataset https://www.kaggle.com/datasets/gpreda/bbc-youtube-videos-metadata License (CC0: Public Domain).

Output format to present:

- Jupiter notebook and pdf version of it.

What would be assessed:

- Code structure;
- Methods for data exploration, cleaning, transformation, and feature engineering;
- Data visualization;
- Completion of each step.

Assignment:

Our team is looking for multiple insights from the BBC YouTube metadata public dataset. Specifically, the team is interested in the correlation between different features and the engagement score. The team leader made an initial preview of the dataset and outlined the steps and questions needed for the study. Please follow these steps consistently and submit your result as a Jupiter Notebook and a pdf version of it with comments where they are required.

1) Explore the dataset and describe the data, clean it up if necessary:
- What types of data are present in the dataset?
- What is the minimum and maximum value for a published time header ('parsed_time_pub'))? Present it in a year (YYYY) format.
- What are the 5 most popular video categories of all time? Please visualize the result.

2) Slice the dataset by cutting the following columns:
published_at, video_category_id, duration, dimension, licensed_content, favorite_count.

From the column 'video_title' create a new header 'video_title_clean' by:

- Removing punctuation;
- Removing stopwords;
- Removing digits;
- Removing following strings: 'bbc one', 'bbc two', bbc three', 'bbc', 'part', 'episode', 'series', 'preview', 'show'.

3) Find the top 5 keywords from the newly generated 'video_title_clean' header for each year represented in the dataset.
4) Calculate and assign a new column 'engagement rate' for each row using the following formula: total engagements (likes, comments, dislikes) divided by the number of views per post, then multiply the result by 100 and round it up to 1 decimal.
5) Calculate the length of characters in 'video_title_clean' and assign it to 'title_len' column.

6) Assign a dichotomized score for engagement rate in a separate column, where ‘top 50%’ (engagement rate >=0.6) is represented by 1 , and ‘bottom 50%’ is represented by 0.
7) Encode ‘video category’ labels and ‘definition’ labels to numeric values.
8) Visualize a correlation between following headers:
   definition, duration, dichotomized score, parsed_time_pub, engagement rate, title length, video_category_label.
9) Describe in your own words what correlations you observe there. Put it in a comment in the code.