My objective for this project was to anticipate the duration of a product based on its title.

I utilised a dataset of product titles and their related durations to do this. I started by installing the required libraries, which included pandas for data manipulation, nltk for natural language processing, and scikit-learn for machine learning.

I cleaned the content after importing the dataset by deleting extraneous characters and changing all text to lowercase. To clean up the data, I additionally deleted stop words, lemmatized the words, and used other text preparation techniques.

Following that, I divided the data into training and testing sets, using 70% of the data for training and 30% for testing

The text input was then converted into a matrix of numerical features used by the machine learning model using scikit-learn's TfidfVectorizer.

I used a logistic regression approach for the machine learning model. Because it is very basic and works well with high-dimensional data, logistic regression is a common choice for text classification tasks. I trained the model on the training data using scikit-learn's LogisticRegression module.

After training the model, I used it to estimate the product lengths for the test set. Then I made a submission file with the projected lengths for each product ID in the test set.

Overall, I cleaned the text data before converting it into numerical features with TfidfVectorizer and training a logistic regression model to predict product lengths. This method performed well, with an accuracy score of around 80% on the test set.