Pilot report for the assignment

Data Transformation:

1) The **Published Date** feature was used to find the number of days it has been since the video was released. It is first converted to the date variable as it was in factor variable. The R library **Lubridate** was used to extract the number from it.(Not used in the test data. Code needs to be modified to apply it to the test data)
2) The **Duration timestamp** was used to extract the duration of the video. The code is written in C++ for efficiency. The code is attached.
3) These features along with the likes, comment, dislikes and views formed the features for our models.

Data Scaling:

The features were scaled by the (max-min) formula with center=0.

Model:

The two models which were used and compared are the **Neural Networks**(5 hidden layers) and the **Linear regression** model. The RMSE for both the models on the training data are:
RMSE.NN=5399.239824
RMSE.LR=3850.994368

Using the codes:

The main.R code extract and prepare the data and apply the corresponding models.
The duration.cpp file extracts the duration from the timestamp provided.

Running the codes:

1) **Pub_date.R** code should be made to run first. This will create dataframe for published date for both training and testing data.
2) Output.csv and Output_test.csv file contains the duration of the training and testing data. This can be generated by using the Duration.cpp file. The code is written for test data. To generate for train data, input and output file name should be changed. I am attaching The output.csv and output_test.csv file. So, **there's no need to run that.**
3) The main.R code then should be made to run. This will first train the model. Then give us the **Output_test.csv** file.