Report

Question 1:

The functional dependencies that we can gather from labeled_data.csv are:

Input_id -> labeldem, Input_id -> labelgop and Input_id->labeldjt.

Question 2:

No the data frame does not look normalized. There exist a number of redundancies in the data frame. Some of these are listed below:

- The subreddit is repeated for every comment in that subreddit.
- The author_cakeday, author_flair_text, etc. are redundant for comments.

We can decompose into the following 3 relations:

1. author_details (author, author_cakeday, author_flair_text, author_flair_css_class)
2. comment (author, body, subreddit_id… everything else except those in 1 and 3)
3. subreddit(subreddit, subreddit_id, subreddit_type)

The foreign keys comment.author is for referencing the author_details.author while comment.subreddit_id references subreddit.subreddit_id.

The maker might have done this to save time as otherwise the joins would cause a lot of overhead and take even longer time than it took now. Moreover, we wouldn't have enough memory to store joins that large.