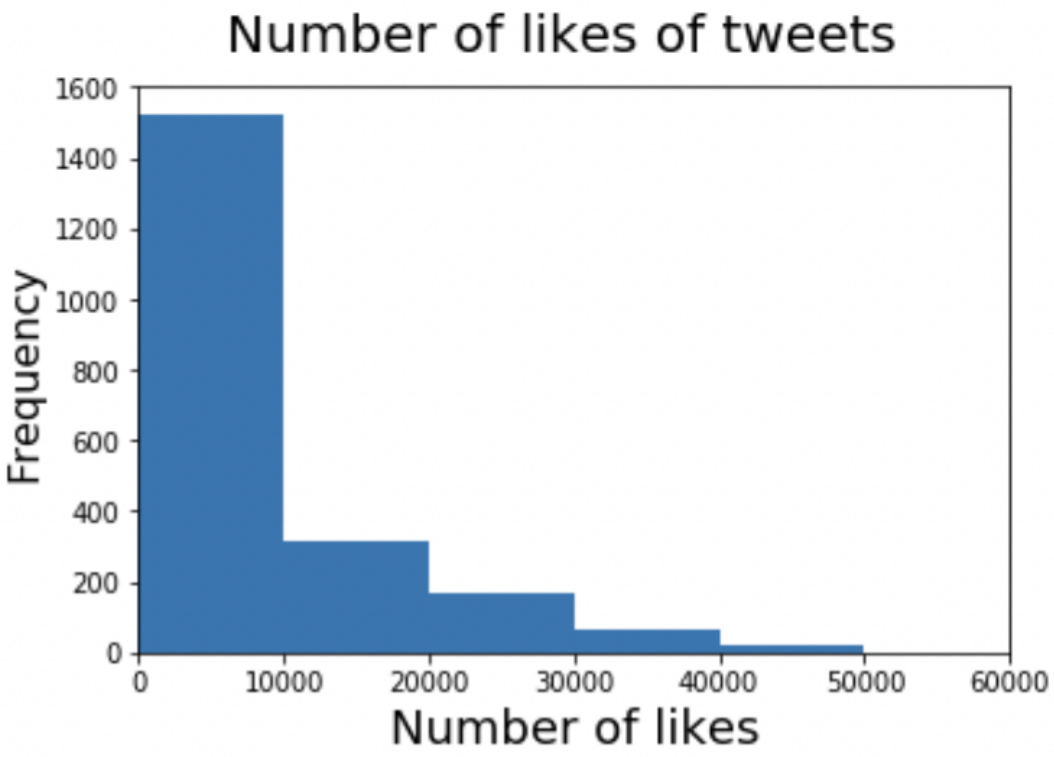
The main insights I obtained after cleaning the data and analyzing it are the following:

1. **Only 16% of the dog stage values are non-null, for original tweet ratings with images**. All the other values are null. Thanks to the cleaning process that changed the values from “None” to null using np.nan, we can identify the non null values, which are 338, using value_counts().sum() on the dog_stage_value column. Also, we can calculate the total number of rows using master_df.shape[0]. Then, dividing the number of non-null values by the total number of rows, we can calculate the percentage of non-null values of the dog_stage_value column. This clearly shows that there is a lot of room for improvement in this column in particular.

2. **The average amount of likes a tweet rating receives is 8936**. We can get this information by calculating the mean of the favorites column.

3. **The average amount of retweets a tweet rating receives is 2826**. We can get this information by calculating the mean of the retweets column.

The following histogram shows the frequency of different ranges of number of likes:



We can see the distribution is skewed to the right, and that the median, which is 4181 likes, is smaller than the mean, which is 8936. There are several outliers that make the mean greater and distort this value, making it a measure of center that is less effective than the median. For example, the tweet that received the maximum amount of likes got 132810 likes.