WRANGLING REPORT

WE RATE DOGS

-Vishnu Kumar V H

Quality Issues:

The following are the quality issues and each have the approach I used to rendered it:

1) **The representation of the prediction values is not correct I feel. The prediction values p1_conf, p2_conf and p3_conf should be in percentage.**

   This change is made to make the readability even better. It is done by multiplying 100 to the confidence level p1_conf, p2_conf and p3_conf.

2) **To make it clearer to the viewers that the confidence values are in percentage, it is necessary to convert the column names to p1_conf (%), p2_conf (%), p3_conf (%).**

   After the confidence level has been changed to the more readable percentage format, it is very essential I feel to change the column names to make it clearer. So, I used the .rename() function to add (%) to the end of the column names.

3) **One very obvious quality issue is the formatting date in the timestamp column. It should be in date format.**

   When checking the info about the archive data frame, the format of date was evidently wrong and was in the object format. Using the .to_datetime() function I changed it to date-time format.

4) **After Date is being formatted, I feel that on the longer run, the timing of the tweet is not very essential but the date will be. So only the date part should be present and the time part must be removed.**

   The timestamp column, I used the .date() function which gives us only the date part of the timestamp and inserted that into df_archive_clean.timestamp.

5) **Now that we have removed the time part, it is only suitable if we rename the column name to tweet_date.**

   The column name has been changed the help of the .rename() function from timestamp to tweet_date.

6) **The text content in the df_archive is followed by the link to the post. Only the text is to be kept and the URL is to be removed. After seeing the first three texts, I planned to remove the rating as well but later found out that ratings can be in the middle as well. So, scrapped that idea.**

   The URL present in text column was removed by slicing the last 24 characters in the text Format. I counted it to be an efficient and simple method to remove the last part of the text which I will remove separately by removing 24 characters. Every twitter shortened URL had a count of 24 characters and hence the number.

7) **In all three columns p1, p2, p3 there are a few predictions which start with lower case letters. They have to be converted to upper case letters.**

   To convert the dog names to upper case I have used the .title() function. To my surprise, not only the first characters of the dog breed was changed but also the second name for e.g. the Shepard in German Shepard was also changed to uppercase.

8) **In the given ratings, it is clear that many of the numerators are lesser than the denominators. I.e. Greater than 10. But I there are many rows with that do not have a denominator of 10. This has to be changed.**

   I converted the numerator the equivalent denominator.

Tidiness Issues:

1) **I tried making all of the doggo, floofer, pupper and puppo columns into a single column suggesting if they are doggo or floofer or pupper or puppo. But the ambiguity I faced was there are rows with two types such as both doggo and pupper etc. In this case, I feel it is best to remove these rows.**

   I used .drop() function to remove these rows.

2) **Finally, after all the cleaning is done, it is essential to order both the tables in ascending order according to their tweet IDs.**

   The sort_values() function is used to order the tweets by ascending order.