Wrangle Report

Data Gathering

The wranggling dynamic started by social occasion different however related information from three distinct sources to complete the ventures as given effectively. The venture was connected with canine rating on Twitter. The main information was twitter_archive_enhanced.csv that originally downloaded and stacked into the iPython scratch pad work area. The subsequent information was mentioned involving demand library on the Udacity site as given in the task portrayal and the last finished result of the information is saved and stacked into Jupyter work area as "image_predictions.tsv". Lastly, Tweepy library was utilized to question third information from Twitter with the assistance of Twitter Programming interface and were additionally handled and perused in the work area as tweet-json.csv. This finishes me of information gathering.

Data Assessment

The following qualities and tidiness issues were identified in the course of using both visual and programmatical assessments.

1. The extended entities column contain many columns that need to be part of the major dataframe.
2. Through Visual and programmatical assessment, there are a lot of redundant columns should be removed
3. There are a lot of missing values in the dataframes
4. Timestamp column in chi_df should be datetime
5. Via visual assessment of chi_twi, there are a lot of retweeted, retweet count and status or replies that are not part of the originial tweets.
6. The lang encodings are not easy to understand when they are not written in full
7. The rating_denominator have arbitrary figures of minimum of 6 and maximum of 170 when it should be 10.
8. Via programmatical assessmet of chi_arc data, rating_numerator seems to have outliers which are outrageous number as rating_numerator has minimum of 0 and maximum of 1776 when it should be between 0 and 15.
9. From the visual assessment, the stage columns named doggo, floofer, pupper, and puppo should be in one column not seperate columns.
10. From visual and programatical assessments, chi_twt, chi_img and twt_arc ought to be merged together to form a single observational unit.

Data Cleaning

In this segment, I cleaned every one of the issues as far as quality and cleanliness archived while evaluating the information utilizing visual and programmatical appraisal methods. I previously caused a duplicate of the first information prior to cleaning and started by then to smooth out the elements segments of JSON object to have individual sections. However, prior to doing that, I made a capability that level the JSON section objects into their singular segments. I then utilized the straighten capability made to determine issue 1 and move to give 2 which was eliminating all sections that had no single or scarcely held back any significant in them. From that point onward, I moved to one more issue by convert the timestamp date to appropriate Python date while adjusting lang section to its full language name in the segments. I then, at that point, added prefix to potential covering sections that could achieve any blunders in falling any JSON segment into significant individual sections which was last finished as definite issue to be settled. Furthermore, the end, a few segments and lines information were eliminated before each of the three informational index were converged into a solitary observational unit.