Overview of The Wrangling Process

This project involves wrangling and cleaning a dataset from three different sources one was only provided as a CSV file then the other two are image_ interpretations. tsv which was obtained by using the request function and then the additional data from Twitter which was obtained using Twitter API. After carefully obtaining the data from its source wrangling took place by copying the data into a new data frame. The data was examined using visuals and programmatic assessment. Different quality issues and tiredness were obtained from the data columns such as the timestamp column which was converted using the pd.to_datetime() function and months as well as years were extracted from the column. Also, some irrelevant columns which weren't needed during the analysis were dropped. Furthermore, in the Image prediction file columns p1, p2, and p3 which are the algorithm's prediction image contain different word format which was then converted to a lower case using .str.lower() functions for the 3 columns, underscores were also replaced in the columns as well using .str.replace() function.

Furthermore, the name column in the CSV file contains "None", "a", "an", and "o" values which depict that no names were recorded for the dog, a function was created to replace all these values with NaN. before concluding the data wrangling process, I merge columns doggo, floofer, pupper, puppo and created a new column for them to be merged into called dog_stage. Before merging, each column contains "None" values which I replaced by creating a function to replace them as an empty string then I merge them into a new column which I named dog_stages. I ensure empty columns in the dog_stages columns are replaced as NaN using the. replace () function for further wrangling the four columns were then dropped to have a clean and tidy data set. After the wrangling and cleaning process, the whole data frame was combined into a master data frame named "twitter_archive_master" which was then stored in a CSV file and was used in analyzing and generating insight and visuals needed for analysis