WRANGLE REPORT

We Gathered the data as three data Frames

1 df_twitter_ad : WeRateDogs Twitter archive

2 df_image_pred : Tweet image predictions

3 df_tweets : Tweet's retweet count and favorite

Data Quality issues:

df_twitter_ad : WeRateDogs Twitter archive data

- There are some unnecessary columns.
- Data Types are wrongly assigned for some columns like timestampe, tweet_id.
- Almost some columns contain only null values.
- name column has in valid data.
- rating_denominator column has values other than 10.
- Range for rating_numerator column.
- modify timstamp column
- modify source column

df_tweets : Tweet's retweet count and favorite data

- Data Types are wrongly assigned for some columns.

df_image_pred : Tweet image predictions data

- Data Types are wrongly assigned for some columns like.

Tidyness:

- In df_twitter_ad (WeRateDogs Twitter archive data) have multiple columns whivh need to be combined.
- Combine three datasets.

Key Points:

1. Dropped all columns with missing values more than 80%.
2. Dropped unncessary columns.
3. Changed data type of time stamp to datetime and tweet_id to string.
4. In valid data in name is replaced with None.
5. rating_denominator column has values other than 10, So made all values to 10.
6. rating_numerator has values less than 10. replaced all values less than 10 to 10.
7. rating_numerator has large values. So bounded the rating_numerator to 10 to 100.
8. Created a new column day from timestamp
9. Created a new column source_browse from source
10. Created new column dog_type by melting dog type columns
11. Created a final cleaned dataset by merging all three cleaned data frames.
12. Stored the cleaned data to csv file.