Data Wrangling Report

Introduction

This is a report dealing with what we have done on wrangling ***WeRateDogs Twitter*** data to create interesting and trustworthy analyses and visualizations.

For doing that we have done three tasks at first

- Gathering data
- Assessing data
- Cleaning data

Then we analyze and visualize some of the insights we got form the cleaned data (on act_report.pdf report)

Gathering data:

We gathered data from three files:

- Read 'twitter-archive-enhanced.csv' and store it in twitter_archive_enhanced dataframe
- Request 'image_predictions.tsv' and store it in image_predictions dataframe
- Downloads the tweets and store them in tweet_json dataframe

Assessing data:

We assessed data visually and programmatically and get the following issues:

I. Quality Issues:

1. In twitter_archive_enhanced dataframe: href tags in 'source'.
2. In twitter_archive_enhanced dataframe: timestamp not in datetime format
3. In twitter_archive_enhanced dataframe: there are duplicates and we can now them from retweeted_status_user_id
4. In twitter_archive_enhanced dataframe: We could drop these columns as we will not use them (in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id)
5. In twitter_archive_enhanced dataframe: Name column has invalid names like 'a', 'an'.
6. In twitter_archive_enhanced dataframe: Name column contain "None" instead of NaN.
7. In twitter_archive_enhanced dataframe: Contains wrong values for numerator and some of them are float numbers
8. In image_predictions dataframe: p1, p2, and p3 contain underscores instead of spaces and some are upper and others are lower cases
9. In image_predictions dataframe: Many prediction columns with different algorithms

II. Tidiness Issues:

1. In twitter_archive_enhanced dataframe: Erroneous datatypes (doggo, floofer, pupper and puppo columns)
2. All tables should be part of one dataset

Cleaning data:

1. In twitter_archive_enhanced dataframe
    - For href tags in 'source' column in twitter_archive_enhanced dataframe
    - We removed the 'a' tags from the hyperlinks.
    - For timestamp not in datetime format
    - We changed *timesatmp* type to datetime
    - For the duplicate's tweets and the columns, we don't want
    - We removed the duplicates any field that has 'retweeted_status_user_id' not NAN and then we removed "*reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id*" as there will be no need for them
    - For the invalid names like 'a', 'an'. In the *name* column has "None" values.
    - We changed all these typos into NAN
    - For wrong values and some of them are float numbers in the numerator column
    - We got the numerator and denominators from text column
2. In image_predictions dataframe
    - For p1, p2, and p3 contain underscores instead of spaces and some are upper and others are lower cases
    - We replaced the underscores with space and turn all text into title
    - Many prediction columns with different algorithms
    - We dropped all the prediction columns and create only
3. For Tidiness Issues
    - For "twitter_archive_enhanced: Erroneous datatypes (doggo, floofer, pupper and puppo)"
    - We first changed the *None* values into NAN, then merge the types columns in one column called stage
    - For "All tables should be part of one dataset"
    - We will merge all tables by *tweet_id*