Reporting: wragle_report

This project uses three datasets:Twitter archive dataset,Additional data from twitter api and image predictions dataset.I used both visual assessment and programmatic assessment to know what issues to clean in all the datasets.

Twitter archive dataset

The Twitter archive dataset has a total of 17 columns:
tweet_id,in_reply_to_status_id,in_reply_to_user_id,,timestamp,source,text,retweeted_status_id,retweeted_status_user_id,retweeted_status_timestamp,expanded_urls,rating_numerator,rating_denominator,name,doggo,floofer,pupper,puppo.

The quality issues I picked up after assessing were:

1.Mismatched datatypes in timestamp should be datetime not object.

2.Finding incorrect ratings visually and updating them manually.

3.Some columns have null values which are entered as none instead of NaN.

4.In dog names column,some dog names are recorded as None,such,the,quite,an.

5.Some expanded_urls are duplicated.

6.Removing rows that have non-empty:retweeted_status_id, retweeted_status_user_id, and retweeted_status_timestamp.

7.Some columns in twitter_archies_enhanced_csv are not relevant to my analysis need to be dropped(e.g retweeted_status_id,retweeted_status_timestamp e.t.c)

The tidiness issues I picked up after assessing were:

1.In twitter_archives_enhanced_csv Dog stages are spread into different columns(puppo,pupper,doggo,floofer).

Solving procedures

For quality issues:The dataset had no missing data but had duplicates in expanded_urls column which i dropped.The timestamp datatype was object instead of datetime,hence the problem was sovled by converting the datatype to datetime.In the dataset some ratings were not correct compared to the ratings given in text column.This was solved by vissually assessing the incorrect ratings and updating them manually .Some columns are recorded None where there is a missing value instead of NaN,the problem was solved by replacing the None values with NaN.In the name column some dog names are invalid such as,quite,very,None,e.t.c,the invalid names were replaced by NaN.Irrelevant columns such as in_reply_to_status_id,in_reply_to_user_id are dropped.Rows that had non-empty retweeted_status_id,retweeted_status_timestamp,were dropped since they are retweets and we don't want them to intefere with the findings. For tidiness issues:Some columns;doggo,puppo,floofer and pupper are spread instead of being in one column,hence the columns are melted.Later the dataset was merged to make a master dataset.

Image predictions dataset.

The image prediction dataset has 12 columns: tweet_id, jpg_url, img_num,p1,p1_conf,p1_dog,p2,p2_conf,p2_dog,p3,p3_conf,p3_dog.

The quality issues picked after assessing:

1.Some jpg_urls are duplicated

The tidiness issues were:

2.Melting the image_predictions dataframe

Solving procedures
The jpg_url column had duplicates which were dropped.In tidiness issue,columns p1, p1_conf,p1_dog,p2,p2_conf,p2_dog,p3,p3_conf and p3_dog were melted and the dataframe was later merged with twitter archive.Irrelevant columns were also dropped.

Additional twitter data

This dataset has four columns: tweet_id,create_date,retweet_count favorite_count.

The quality issues picked up were:

1.Create_date is datatype object instead of datetime

The create_date datatype was object instead of datetime,this was solved by converting the datatype to datetime.Later the dataset was merged with the twitter archive dataset to make a master dataset.

Conclusion

In the wrangling process all the three datasets were cleaned and merged to form a master dataset that contains the relevant columns needed for analysis.Wrangling helped me identify and correct mistakes which may have led to inaccurate findings.