Data Wrangling Report

Gathering data:

Data was gathered from three sources

1-`twitter_archive_enhanced.csv`

A csv file downloaded normally and added to workspace

2-`image_predictions.tsv`

Downloaded programmatically in python

3-`tweet_json.txt`

Downloaded programmatically in python due to lack of API credentials

The three files are loaded into pandas DataFrame objects

Assessing Data

Data was assessed programmatically using pandas DataFrame functions info(),value_counts()

9 quality issues and 2 tidiness issues were observed

Quality Issues¶

1- timestamp is object ---> should be a timestamp

2- some tweets are retweets

3- some tweets have no images

4- inconvenient column names in image prediction data

5- inaccurate rating (domenator bigger than 10 or smaller)

6- inconsistent doggy type for NULL or NaN (None was found)

7- TweetID should be a string not an int

8- wrong names (a , an)

9-not a rating in archive data (24/7 !?)

Tidiness issue

1- doggo , floofer , pupper and puppo should set into one column (the dogtionary)

2- tweeter api table should be included in archieve data table

Cleaning Data

Quality Issues

Data was cleaned programmatically using pandas functions

In timestamp issue we converted it using pandas.to_datetime

Retweets were dropped by finding non-null values in in_reply_to_user_id and retweed_status columns

Tweets with no images were dropped by merging the two data frames of image prediction and twitter archive on the tweet_id in image prediction data frame

inconvenient column names in image prediction data were changed using pandas.dataframe.replace()
same as to None values in data replaced with numpy.nan value

TweetID should be a string not an int and converted using astype()

wrong names (a , an) was replace using a regular expression in the text element to find the right name for each row with on of these values

rating issues was solved manually by checking in text elements for these inconvenient rating where denominator is not 10

Tidiness Issues

doggo , floofer , pupper and puppo should set into one column (the dogtionary)
solved by summing all these values and putting them in one column

tweeter api table should be included in archieve data table
solved by adding its column to twitter_archive data