Wrangle Report

Introduction

Real-world data rarely comes clean and as a Data Analyst it's our job to gather, assess and clean the data to make it viable for analysis.
The dataset I'm working on is a twitter archive of user @dog_rates. WeRateDogs (https://twitter.com/dog_rates)is a Twitter account that rates people's dogs with a humorous comment about the dog. With more than 9 millions followers Their archive for 2016 and 2017 is provided by udacity, and was downloaded manually we need to pragmatically the image_prediction file from Udacity hosted server and Download additional data using tweepy API.
After gathering the data we should asses them and find quality and tidiness issues with it, clean it and then analyse it.

←
WeRateDogs®
16K Tweets
JustFood ForDogs.
More life. More years. More love.
trupanion™
Medical insurance for the life of your pet.
•••
Following

WeRateDogs®
@dog_rates

Your Only Source For Professional Dog Ratings
Instagram and Facebook ➔ WeRateDogs
partnerships@weratedogs.com | nonprofit: @15outof10

Professional Services · all our links ➔ · links.weratedogs.com
Joined November 2015

21 Following 9.3M Followers

1. Gather Data

- Twitter_archive_enhanced.csv was Provided by Udacity and was loaded into a dataframe called ‘archive’
- Image_predictions.tsv was downloaded programmatically using Requests and was loaded into a dataframe called “image_pred”
- Additional data such as favorite and retweet_count was gathered by using Tweepy API and stored in a “tweet_json.txt” file and data frame called “count”.

2. Assess Data

- ### Quality Issues

- From archive_c, find and replace dog names with 'None' if possible. It appears that names that are lowercase tend to be invalid. Find and replace lowercase dog "names" if possible. Replace with NaN if no name is foundRemove tweets without images.
- Drop retweeted columns from archive_c tableRemove unnecessary columns.
- Drop in_reply_to_status_id and in_reply_to_user_idRemove the outliers foe rating_numerator and rating_denominator.
- classify archives source column in four observed sourcesChange source column from ulr type to text.
- Change the rating_numerator and rating_denominator for oberservations with wrong value or drop them if necessary
- Create new column rating=rating_numerator/rating_denominator. Drop rating_numerator and rating_denominator
- Change datatypes of timestamp to datetime, and tweet_id, to strings.
- Drop all three false row prediction
- Source displays url

- ### Tidiness

- Dog "stage" variable in four columns: doggo, floofer, pupper, puppo.
- Join 'tweet_info' and 'image_predictions' to 'twitter_archive'
- Create top accurate predicted dog breed in a column and drop p1,p2,p3 related columns

3. Clean

- Made a copy of dataframe.
- Remove tweets with retweet.
    - Remove rows where retweet_status_x is not null and finally drop the columns that are not necessary.
- Dog Stages into 1 column instead of 4.
    - Combine dog stage columns (doggo, floofer, pupper, puppo) into one 'dog_stage' column.
    - Rename None with NaN
    - Drop the 4 columns.
- Change source column from URL Text to text.
    - Replace source links to string defining them.
- The rating_numerator and rating_denominator have offbeat values.
    - Change the rating_numerator and rating_denominator for observations with wrong value or drop them if necessary.
    - Create new column rating=rating_numerator/rating_denominator.
    - Drop rating_numerator and rating_denominator.
    - Drop tweet_id 810984652412424192 as it's rating is inaccurate.
    - Drop ratings which is not realistic.
- Convert timestamp to datetime.
- Convert non-dog names to 'None' then make title case.
- The prediction column of dog breed can be simplified.

- Condense Dog breed Column by choosing the one which is true at first as the first column has the highest percentage than the next one.
- Drop the columns that are not necessary.
- Join tweet_archive, image_prediction, tweet_json into one master dataset on tweetid.
- Join tweet_archive, image_prediction, tweet_json into one master dataset on tweetid.

4. Store Data

- Saved the master dataframe to csv file ‘twitter_archive_master.csv’.