Data Wrangling Report

1. Introduction

This document records the Data Wrangling efforts done on the three pieces of twitter data the focus on Gathering, Assessing, Cleaning.

2. Data Gathering

Data for this project is gathered from three different sources.

1. *twitter_archive – (from a csv* file) The WeRateDogs people are generous enough to provide us with a bunch of their tweets data to Udacity Course purposes. The data in this csv contained some basic data such tweet_id, name, rating_denominator, rating_numerator, dog stages etc.

2. *image predictions* – (*from web*) This file is download from Udacity's server using the provided url with request library. The file contained information such as dog breed, different algorithm data and their confidence values.

3. *tweet_data*- (*from an api*) This file is downloaded and constructed using the python twitter API Tweepy. The resultant file contained information such as tweet_id, retweet_count, favorite_count.

From these three files, three data frames have been created.

- **df_twitter_archive**
- **df_images**
- **df_tweet_data**
-

3. Data Assessing

All the three data frames are assessed to identity quality and tidiness issues.

a. Tidiness issues

df_twitter_archive

- There seems to be some tweets that are retweets, we shall be deleting them for tidiness purpose.

- Dog stages are separated and stored in four different columns “doggo”, “floofer”, “pupper” and” puppo”. We shall melt them into one column “dog_stage”
- We shall insert a new column "Rating" which is equivalent to numerator/denominator for our calculations later on in the analysis.
- We shall delete unnecessary columns for tidiness.

df_images

- We shall melt down the algorithms columns (p1, p2, p3) and algorithms confidence columns (p1_conf, p2_conf, p3_conf) to have only "algo", "algo_conf" columns.
- We shall delete all unnecessary columns.

df_tweet_data

- This tweet_data data frame seems to be tidy, no action needed here.

Overall, we shall merge the three data frames into one master data frame.

b. Quality

df_twitter_archive

- We shall change the timestamp to a standard format data type
- The name column has some weird names as dog names, we shall inspect and clean them up.
- The “ratings_denominator” also has 0 values, we shall clean them up.
- The "ratings_numerator" have some very high values. We shall delete any rows with numerator values greater than 20.
- Will change the dog names to all so that every names start with upper case.
- Will delete expanded urls row with missing data.
- Change the "source" text to something that is human readable.

df_images

- There are rows where all three algorithms did not predict the image to be a dog. We shall delete such rows.

- Will change the dog breed columns to be consistent case (some are lower and some are upper)
-

df_tweet_data

- This tweet_data data frame seems to be of quality, no action needed here.

1. Data Cleaning

Before cleaning the data, a copy of each these data frame has been saved.

All the three data frames are assessed in the previous step and ready to be cleaned.

- Some of the key tidiness issues that will be cleaned up are:

df_twitter_archive

- All the rows that are considered retweets have been deleted for analysis.
- Have removed the different dog stage columns ("doggo", "floofer", "Pupper", Puppo) and added a new column "dog_stage".
- Have inserted a new column "Rating" which is numerator/denominator.
- Have deleted unnecessary columns from df_twitter_archive – *["in_reply_to_status_id","in_reply_to_user_id", "retweeted_status_id", "retweeted_status_user_id", "retweeted_status_timestamp", "rating_denominator", "rating_numerator"]*

df_images

- Have reduced the algorithms columns (p1,p2,p3) and algorithms confidence columns(p1_conf, p2_conf, p3_conf) to have only "algo", "algo_conf" and best predicted breed "dog_breed".
- Have deleted unnecessary columns.

df_tweet_data

- There were no tidiness issues in df_tweet_data

- Some of the key quality issues that will be cleaned up are:

df_twitter_archive

- "*timestamp*" column datatype is changed to *“datetime*” data type
- "name" column with dog names "a", “an”, and “the” have been replaced by "None."
- Rows with ratings_denominator = 0 values have been deleted.
- Rows with "ratings_numerator" very high values (greater than 20) have been deleted.
- Dog names in the “name” column are capitalized.
- Have deleted rows where there are null values for “expanded_url” columns
- Have changed the “source” column text to something that is more meaningful and readable.

df_images

- Have deleted the rows where all three algorithms are predicting the image not be a dog.
- Dog breed columns p1, p2, p3 are capitalized.

df_tweet_data

- There were no quality issues in df_tweet_data

After the cleanup is done, the all the three data frames are merged into one dataframe “df_master_twitter_data”. The final master twitter data is then stored into a csv called 'twitter_archive_master.csv.