Data Wrangling Report

By Osama Hamdy Osman

April 2019

As an assignment for the Udacity Data Analyst Nanodegree; This Report illustrates the main steps involved in the data-wrangling of Twitter account "WeRateDogs".

Data Gathering

In this step, collecting data takes place. For this project, there were three main sources for the data to deal with:

1. Twitter_archive_enhanced.csv file, this file was delivered by email and downloaded manually to our working directory and then imported into our working environment using Pandas function "pd.read_csv".
2. Image_prediction.tsv is the second file that has been hosted on a webpage and downloaded from its relevant URL using the Requests library get function and pd.read_csv pandas' function. This file encompassed image predictions for the dogs' breeds obtained through a neural network on most of the tweets in the archive file
3. The final dataset was gathered from twitter REST API via the Tweepy library by querying the API to obtain extra information pertinent to the tweets' ids in the first file, e.g. retweets count and favorite count aspects

Data Assessment

In this step, we investigate our imported datasets both visually and programmatically for quality and tidiness issues.

1. The visual assessment done on spreadsheet application like excel and then the programmatic assessment is conducted in Jupiter notebook.
2. Missing data were addressed first then messy structures were addressed to facilitate the tackling of the rest of the quality issues that fall in the buckets of validity, accuracy and consistency classes of the data quality aspects.
3. Some of the data cleaning efforts were guided by the scope of the project that mandated the exclusion of retweets and replies and tweets featuring no images.

Table	#	Issue Quality Issues	Solution
Archive	1	Data types(consistency issues); All timestamps are object type	Type conversion to datetime data type
	2	All tweet_ids are integers	Type conversion to string
	3	Inconsistent representation of null values as "None" strings in the (name, doggo, floofer, pupper, puppo) columns.	First, the completeness issue was addressed by extracting the correct values then the type of missing values was converted into Nans
	4	There were Retweets and replies in the dataset	Removing those tweets by slicing and comparing with image prediction dataset
	5	Erroneous names like the letter “a” and “an”	Their relevant retweets were investigated, and the correct names were extracted if existed
	6	Missing entries in expanded_urls.	Dropped as those don’t feature images
	7	Incorrect and weird values of the rating_numerator which has a maximum of 1776. - The same holds for the rating_denominator with illogical maximum of 170	Absurdly high values (there were two) were deleted, others were closely investigated, the correct values were extracted programmatically and manually
image_predictions	8	Inconsistent capitalization for the predicted breeds(p1, p2, p3)	Applying the series.str.capitalize method on the entire column
	9	In general, the total number of records (2075 instead of 2356) indicates the presence of some tweets without images which we need to exclude.	The tweets that didn’t exist in this table were deleted from the archive table by checking the tweet ids against each other

	10	Non-descriptive columns' names	This was addressed with the tidiness efforts by renaming the columns
		Tidiness Issues	
Archive	1	values are column names(doggo, floofer, pupper, puppo)	Combined in one column names "dog_stage"
image_predictions	2	values are column names(p1,p2,p3) which are all breed predictions	The columns were combined using panda's wide_to_long method
API table	3	This isn't considered an observational unit to have its own table	merged to the *`archive`* table

Output

Two tables : Archive table 1968 records and the Predictions table 5913 entries