Data Wrangling Report

By : Mo3az Elsoakry

December 2020

As ana assignment for Udacity Data Analyst Nanodegree;This report illustrates the main steps involved in the data_wrangling of Twitter acount "WeRatedDogs".

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 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 it's relevent URL using he requests library get function and pd.read_csv pandas' function . This file ecompassed 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' id 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 programatically for quality and tidiness issues.

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

Table	#	Issue	Solution
		Quality Issue	
Archive	1	Data types(consistency issues); All timestamps are object type.	Type converstion to datetime data type.
	2	All tweet_id are integers .	Type converstion to string.
	3	Inconsistence representation of null values as 'None' strings in the (name,doggo, floofer, pupper, puppo) cloumns.	First completeness issue wass 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 ith image prediction dataset.
	5	Erroneous names like the letter “a” and “an”.	Their relevent 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 in the rating_numerator and rating_denomenator.	Extracted rating_numerators from text correctly as floats, then handled photos that have more than one dog by dividing them on the number, the correct values were extracted programatically and manually.
Image_prediction	8	Inconsistent capitalization for the predected preeds.	Applying the series .str.capitalize method on the entire column.
	9	The tables entries number should be the same.	The tweets that didn’t exist in archive table were deleted from this table by checking the tweets ids against each other.
	10	Non_descreptive columns’ names.	This was addressed with the tidiness efforts by renaming the columns.

		Tidiness Issues	
Archive	**1**	**Values are columns names(doggo, floofer, pupper, puppo).**	**Combined in one column names “dog_stage”.**
API table	**2**	**This isn’t considered an observational unit to have it’s own table.**	**Merged to the “archive” table.**