Wrangle Report

Mervat Khaled

February 2021

Introduction:

The purpose of data wrangling and analysis project is to put in practice what we have learned in data wrangling data section from Udacity Data Analysis Nanodegree program. The dataset that is wrangled is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs.
In this report, we clearly documented the main steps of wrangling WeRateDogs Twitter data to gather, assess, and clean are presented.

Project objectives:

The project main objectives were:
- Perform data wrangling (gathering, assessing and cleaning) on the provided sources of data.
- Store, analyze, and visualize the wrangled data.
- Reporting on:

1. data wrangling efforts.
2. data analyses and visualizations.

Step 1: Gathering Data:

In this phase, the three pieces of data were gathered and represented as pandas dataframes:
- The WeRateDogs Twitter archive (file on hand, manual download of 'twitter-archiveenhanced.csv')
- The tweet image predictions ('image-predictions.tsv'). This file was be downloaded programmatically using the Requests library from a provided URL.
- Each tweet's entire set of JSON data (with at minimum tweet ID, retweet count, and favorite count) in a file called 'tweet_json.txt' were stored using Twitter API and Python's Tweepy library. Each tweet's JSON data was written to its own line.

Step 2 and 3: Assessing and Cleaning Data

While working with data, a number of observations were made. In the below table there are the observations along with actions taken in the Cleaning Step.

Quality:

Dataset	Observation	Solution
Archive data	The criteria of the project focus on original tweets only, but the data has 181 retweeted rows in [retweeted_status_id,retweeted_status_user_id , retweeted_status_timestamp]	Removed all retweets rows ‘indexes’ from the data, then dropped columns: 'retweeted status id', 'retweeted status timestamp', 'retweeted status user id'.
	Also there are replied tweets and that not support the validity of the project.	remove reply columns: in_reply_to_status_id','in_reply_to_user_id as they are invalid data.
	the expanded_url column has missing values ‘NaN’,which means no photos, and that isn't valid too.	Removed rows that do not have a photo in expanded urls column. using Regex and pandas contains and findall methods.
	the name column has 'None' instead of np.nan and strange values such as (a,an).	replaced wrong dogs names with the pattern that indicates some names come after the word 'named' in text using Regex and pandas contains and findall methods. 23 values are extracted. Then replaced invalid names and None with np.nan.
	name column given lowercase	Converted the type of name column to string, and normalize the names with title method.
	timestamp column has wrong datatype 'object' instead of datetime.	Converted timestamp column datatype to datetime.
	Rating denominator has more than 10 and less than 10.	re extracted digits with length 2 after forward slash from text column to modify rating denominator then replace unmeaningful values in rating denominator (more than 10, less than 10) with 10.

	tweet_id column should be converted to string type instead of integers.	converted tweet id datatype to string.
	text column has urls, punctuations digits ,user_names, it should be cleaned and normalized.	Removed punctuations, urls , usernames, digits using Regex, nltk, pattern libraries
Image predication data	jpg_url has duplicated values.	Dropped duplicated.
	Image num column isn't useful for analysis.	Dropped image num column.
	Tweet id has integer as datatype.	converted tweet_id to string.
	p1 column has lowercase and uppercase.	converted p1 values with tiltle method.
API data	id column should be renamed to tweet id to match with other data frames and converted to string type instead of integers.	Renamed id column to tweet id. and converted it to string type instead of integers.

Tidiness

Data sets	Observation	Solution
Archive data	['doggo', 'floofer', 'pupper', 'puppo'] should be concatenated at one column dog_stage.	Dropped ['doggo', 'floofer', 'pupper', 'puppo'] columns, and re extracted the values with regular expression to concatenate them at one column dog_stage.
	[rating_numerator,rating_denominator] should be divided and represent as one column "ratings".	Divided [rating_numerator,rating_denominator] represented them as one column "ratings"

Image predication data	There are three algorithms with different accuracy, we should choose the highest one and keep its predication only.	Choose the algorithm with the highest accuracy p1_conf, and keep its classification, and rename p1 to breed and p1_conf to confidence then drop other algorithms.
All	All data sets shoud be merged to one data frame.	Merged all data sets into pandas data frame.

Result:

A combined data set with all needed information was stored in a csv.