Wrangling Report on WeRateDogs Archive Data

Introduction

This fourth project is on wrangling, for the Udacity Data Analyst Nanodegree program. It is focused on data gathering from three different sources, assessing same data, cleaning and storing them in preparation for analysis and visualization.

Project Objectives

The objectives of this project are:

- Gather data from three different sources.
- Assess the gathered data, identify a minimum of 8 quality issues and 2 tidiness issues.
- Clean the data using the identified issues as guide.
- Store the cleaned data in a file titled twitter-archive-master.csv.
- Analyze and visualize the cleaned data, producing a minimum of 3 insights and 1 visualization.
- Do a report in two separate documents, one in 300 – 600 words describing wrangling efforts as saved as wragle_report_pdf or wrangle_report.html, the second communicating insights and displaying visualization(s) from the wrangled data in a minimum of 250 words.

Data Gathering

Data was gathered from three different sources:

1. The first was manually downloaded from the Udacity Resource Tab, it is titled *twitter_enhanced.csv*
2. The second was programmatically downloaded through a web url link: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv an image prediction file saved as *image-prediction.tsv*
3. The third entailed downloading a Jason file through twitters API access, using tweepy's library for extraction of the needed data from the downloaded file.

Data Assessment

All data sets were carefully observed and issues identified and documented in preparation for cleaning, the issues identified can be found below.

Dataset (Dataframe)	Issue	Issue Type	Action
twitter-archive-ehanced.csv (archive_clean_df)	in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp columns are irrelevant and would be dropped, after eliminating their non null values	Quality	Non null rows in dataframe were dropped, then columns were dropped afterwards
	'none' values for columns 'doggo', 'floofer', 'pupper' and 'puppo' would be replaced with '-'	Quality	None value were changed to '-' values
	rating_numerator contains extraneous values which would be dropped	Quality	Extraneous values were dropped
	rating_denominator contained values other that 10, which should not be as rating should possess same denominator	Quality	Values other than 10 were dropped
	timestamp column is a string, which should be converted to date time and column values trimmed to contain proper date values.	Quality	Timestamp column converted to date time data type and column sliced into standard date format
	columns 'doggo', 'floofer', 'pupper' and 'puppo' should be collapsed into a column titled 'dog stage'	Tidiness	Columns were collapsed into a single column titled 'dog_stage'
	create a rating column computed by dividing numerator with denominator	Tidiness	Rating column was created
Image-prediction.tsv (image_clean_df)	columns 'p1', 'p1_conf', 'p1_dog', 'p2','p2_conf', 'p2_dog', 'p3', 'p3_conf', 'p3_dog' are not discriptive enough and should be renamed	Quality	Columns were renamed
	'p1_dog', 'p2_dog' and 'p3	Quality	Row corresponding

	dog', with false values should be dropped as they are not dogs and columns dropped afterwards as all value now becomes true.		to false 'p1_dog', 'p2_dog' and 'p3 dog' were dropped alongside columns
	Capitalize the names in columns in 'p1', 'p2', and 'p3'	Quality	Column variables were capitalized
Tweet-jason.txt (tweet_clean_df)	All three data frame would be merged into one master data frame	Tidiness	Data frames merged into one master dataframe
	'timestamp' column would be split in 3, 'day', 'month' and 'year' columns.	Tidiness	Columns split in three, one for day, another for month and the third for year

Resulting Dataframe

The final dataframe is a clean dataframe with all three of them merged into one master data frame and saved as *twitter-archive-master.csv*