DATA WRANGLING REPORT ON THE WERATE DOGS ARCHIVE

INTRODUCTION

This project 4 of the Udacity Data Analyst NanoDegree. It is concerned about demonstrating data, wrangling ability on the WeRateDogs archive. WeRateDogs is a funny dog rating service.

PROJECT OBJECTIVES

The objectives of the project are to:

- ❖ gather data from three different sources.
- ❖ assess the gathered data with the aim of identifying at least 8 quality issues and 2 tidiness issues.
- ❖ clean the data with respect to the identified issues.
- ❖ store the cleaned data in a file titled twitter-archive-master.csv.
- ❖ analyze and visualize the stored data producing at least 3 insights and 1 visualization.
- ❖ report the work by producing two documents namely internal (wrangle_report.pdf or html with 300-600 words) detailing the wrangling efforts and external (act_report.pdf or html, 250 words mininum) detailing the insights and visualizations.

DATA GATHERING

I gathered data from three different sources.

- ✓ First was manually from the Udacity Resource Tab. From there I downloaded *twitter-archive-enhanced.csv*
- ✓ Second, I programmatically downloaded an image predictions file from https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv . The file was named image-predictions.tsv.
- ✓ Finally, I extracted information data from the tweet-json.txt file downloaded through twitter API using tweepy. It should be noted that my twitter API request delayed and I therefore resorted to using the other recommended approach. Before I finished the project, though, my request was granted and when I used it the data was the same as the one extracted.

DATA ASSESSMENT

I assessed the datasets and identified some data quality and tidiness issues. These issues and the corresponding actions that I took are presented in the table below.

Dataset (Dataframe)	Issue	Type	Action
Tweet-json.txt(dj)	The retweeted column contains only False values. A variable is expected to vary. But this variable does not vary and as such should not be a part of the dataframe.	Quality	Dropped the column.
	The created_at column should be datetime and not string.	Quality	Converted the column to datetime.
	The created_at column and the timestamp column of the dt dataframe contain three different variables namely Day of the week, Month of the year, and time. These components should be in different columns since they are different variables	Tidiness	Dropped the created_at column and split the timestamp column accordingly.
Twitter-archive-enhanced(dt)	The number of nulls in retweeted_status_id, retweeted_status_timestamp, retweeted_status_user_id, in_reply_to_status_id, in_reply_to_user_id are too many. I don't think such columns should be included in the dataset.	Quality	First dropped the non-null values (project requirement). Next, dropped the columns.
	Checking the rating_denominator column shows that one of the denominators is 0.	Quality	Dropped the column by addressing other quality issues.

	The name column contains 775 None and 55 'a'. There are other names like an, the, this, quite, interesting, just, his, not, o, unacceptable, one, getting, and infuriating. These are not valid names. Moreover, they are all in non-title case whereas names should be in title case.	Quality	Converted them to title case. Replaced 'None' with 'Not_Availabe'.
	The columns for dog stages have some having dog stage values that are more than one. I don't think a dog can be in two growth stages at a time.	Quality	Dropped them.
	The four columns namely doggo, floofer, pupper, and puppo all captures a stage of dog. Since a dog is expected to belong to only one of these stages, only one variable is needed to capture them not four	Tidiness	Created a column called 'dog_stage' and dropped the other four columns.
	There should be a column to capture actual dog rating (that is a standardized dog rating).	Tidiness	Created a column called 'rating' to capture standardized ratings.
Image-predictions(di)	Some column names are not descriptive. Examples are p1, p2, p3, p1config, p2_dog, etc	Quality	Renamed the columns.
	Most of the names captured by p1,p2, and p3 are not in title case. Since they are proper names of dogs, I think they should be in title case.	Quality	Converted values to title case

OUTPUT

The output of the assessment and cleaning effort is stored to twitter-archive-master.csv