WRANGLING REPORT

INTRODUCTION

This project aims to apply all the aspects of wrangling learned in the Data Analysis Nanodegree. The data under practice is the WeRateDogs Twitter account. This Twitter account is well known for rating people's dogs' photos with a humorous comment about the dog, and an invalid funny rating.

This report briefly describes the thought process and steps taken to gather, assess, clean, analyze and visualize

GATHERING

Data used for analysis come from three different sources:

1. WeRateDogs Twitter Archive

WeRateDogs Twitter archive, which contains tweet data like tweet ID, timestamp, text, rating, and some other data. This file is downloaded manually from the Udacity server and stored locally (twitter-archive-enhanced.csv) as CSV file. Then it will be read into DataFrame to continue the analysis with its data.

2. The Twitter Image Predictions.

A table that contain each tweet ID, image URL, image number used for prediction, breed predictions data. This file is hosted on Udacity's servers and will be downloaded programmatically. By using request module, the URL content can be accessed and then saved into a TSV file locally, then this data will be read into a DataFrame.

3. Retweets and Favorites JSON Data.

Additional Twitter data needed (like retweets count and favorites count) can be accessed through API and acquired from twitter directly.
Using the API we can get each JSON for each tweet ID we have, then each JSON data is stored as a single line in a local text file. That file is read line by line and accessed with the json module to extract the required data, that data will be stored in a single dictionary item for each JSON, dictionaries will form a list which will be used to create a DataFrame with tweet ID, favorite count, retweet count.

ASSESSING

The data acquired from gathering is not ready to be analyzed, it need to be altered and cleaned to increase its quality and tidiness. Both visual and programmatic assessment are used to detect issues.

Some of the methods used to assess the data:

- Print the three used DataFrames and check their columns and data.
- Get information about the types of data in each table and how many rows are missing values
- Identify missing value cause and relation to other columns
- Check for duplicated rows
- Get the range of values or categories in some columns and identify out of place values like rating.
- Identify issues for the extracted information from the text like name and rating.

All detected issues are categorized into quality or tidiness issues, this is used as a reference while cleaning, it also helps with prioritizing issues (tidiness before quality).

CLEANING

Detected issues are addressed in sequence Define, Code, & Test were used in each step.

Addressed issues in order

1. Twitter JSON should not be a separate table
2. No column for accepted prediction in images table
3. *p1, p1_conf, p1_dog*... columns are not needed for analysis.
4. Images table should be part of the Archive
5. *None* is used instead of *NaN*
6. Dog stage is on 4 columns
7. *retweeted_status_user_id* are retweets and not needed
8. *in_reply_to_status_id* are replies and not needed
9. Remove extra columns
10. *timestamp* is a string
11. *rating_numerator* and *rating_denominator* not accurate
12. *rating_numerator* with no value
13. *breed* column names are wrongly capitalized
14. *name* column has invalid names

SAVING

After finishing all cleaning tasks, the output DataFrame is saved in a local CSV file (twitter_archive_master.csv)