@dog_rates – Twitter Handle Tweets Data Wrangling

The data wrangling process has been broken down into three main steps:

1. Gather
2. Assess
3. Clean

1. Gather

There are 3 main data sources for this project;

I. Excel data in csv format. This is downloadable directly from the project resources and saved locally as a csv file named twitter-archive-enhanced.csv. I have used this file to get all the tweet ids used in the twitter API explained in data source III explained below.

II. TSV file format. A file url is provided and downloaded using python requests library. I have saved the file locally as image-predictions.tsv. This contains the image predictions from a neural network.

III. Twitter API. Using the twitter API I have queried all tweets data by tweet id (from step I) in a JSON format and saved into a txt file named tweet_json.txt. The next step is reading this JSON format by looping over the text lines to create a csv file namely twitter_archive_master.csv

A copy is made for the datasets for use in the next step.

2. Assess

Assessing in this case includes visually and programmatically looking for both messy and dirty data issues in the 3 datasets gathered. Issues I was looking for include but not limited to data duplication, missing data, wrong data formats and data tidiness

The assessment comments have been categorized to **Quality** issues and **Tidiness** issues as in the table below.

3. Clean

The data cleaning process is broken down into 3 simple steps:

I. Define – Cleaning steps definition

II. Code – Code to perform the cleaning tasks

III. Test - Run the code with no error and achieve clean data as defined

See below tabulated assessment comments and cleaning solutions

Assessment Comment	**Cleaning task solution**
Quality	
tweets	
The created_at column is a string The created_at column date is of an API format	Convert created at column to a datetime format and convert the date to datetime

Missing hashtag information hashtag is a float data type column	Extract all hashtags from the tweet text fields
Missing user_mentions data	Extract all user mentions from the tweet text fields
Retweets are included	Filter out all retweets into a retweet data frame
Tweet text includes hashtags, mentions, ratings and other non-alphabetical characters	Clean up the tweet text- remove hashtags, mentions, ratings and any other non-alphabetical characters
image predictions	
Breed dogs names separated by an underscore and in lowercase	Remove the underscore and capitalize the first name
6 decimal places for model prediction scores	Convert into a percentage and a whole number and convert into integer
Columns have inappropriate names	Rename columns
Tidiness	
Tweets	
Ratings are included in the tweet_text column	Separate ratings from tweet_text column and calculate the actual rating
Image predictions	
first, second and third prediction data are included together in the same table	Separate first, second and third prediction into different data frames to form 3 independent observational units

After performing the above task the resulting data is saved to image_predictions01.csv, image_predictions02.csv, image_predictions03.csv and clean_tweet.csv and this should be sufficient for the explanatory analysis.