Data Wrangling : WeRateDogs Twitter account

In this project we have wrangled and analysed data from a Twitter account called WeRateDogs using python and its libraries. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. Data wrangling consisted of the following stages:

1. Gathering data
2. Assessing data
3. Cleaning data

These stages are described below.

Gathering data

The data was gathered from three different sources and in three different file formats.

i) The first dataset was provided by Udacity and was downloaded manually. This was the WeRateDogs Twitter archive provided as a csv file: twitter_archive_enhanced.csv, and was loaded to a pandas dataframe.

ii) The second dataset was the tweet image predictions tsv file that is hosted on Udacity's servers. The URL link was provided by Udacity and the file was downloaded programmatically using the Requests library, and read into a pandas dataframe.

iii) Downloading and obtaining the third dataset was quite challenging. This required setting up a up a Twitter developer account. We used the Tweepy library to query Twitter's API for additional data using the tweet ID of each tweet. The JSON data for all the tweets obtained using Python's Tweepy library was stored in a file called tweet_json.txt file, with each tweet's data in a single new line. Then this .txt file was read line by line into a pandas DataFrame.

Assessing data

After gathering data, the data was then assessed visually and programmatically. Several quality and tidiness issues were identified and documented for the three different tables. As assessing and cleaning the entire dataset completely would be very time consuming, for the purpose of this project we identified approximately 8 quality issues 2 tidiness issues. These issues were then addressed and cleaned in the next stage.

Cleaning data

The various quality and tidiness issues identified were addressed and cleaned. Before the cleaning operations were performed, copies of the original datasets were made. The define, code, and test steps for each of the cleaning process were clearly documented. A tidy master dataset with all pieces of gathered data was created, and this was stored as a csv file for subsequent analysis.

The issues that were identified and cleaned are documented in the table below.

TIDINESS ISSUES

	ISSUE	SOLUTION
1	There are 4 different columns containing the same type of data, i.e, dog stage, instead of a single column.	The 4 separate columns were merged into a single column called ‘stage’ using the join function. The original columns were deleted after the merge.
2	The data are in three different tables, each containing different set of tweet info for the same tweet IDs. These should be in a single table according to the rules of tidy data.	The three datasets were merged into a single table using the pandas merge function.
3	There are some columns which do not contain any relevant information or are no longer required for our analysis purpose.	The unnecessary columns are removed from the dataset.

QUALITY ISSUES

	ISSUE	SOLUTION
1	The 'id' column in the df_tweetdata_clean dataframe was inconsistent with the column names in the other two dataframes.	Renamed the 'id' column in the df_tweetdata_clean to 'tweet_id' for consistency with the other two dataframes
2	Incorrect datatype for the columns containing tweet-id information, should be strings but appeared as integers.	Converted the incorrect datatypes into the correct strings format using the astype() function.
3	The date and time information were present in a single column as the timestamp.	Converted the 'timestamp' columns into two separate columns for date and time using the assign function in the datetime library.
4	The timestamp column also had incorrect string format.	After splitting into separate date and time columns, corrected the columns to datetime category.
5	The dataset contained retweets and replies.	Removed all the retweets and replies, while retaining only the original tweets.
6	The source column contained redundant text.	Formatted the text in the source column using the replace function, making it easier to read.
7	Incorrect name information present in multiple rows.	All the invalid dog names were identified using the islower() method and replaced with ‘Not known”.
8	The maximum column width display was originally 50. The text column was truncated hence the information could not be read in full.	Increased the maximum column display width to 200 characters to make all the text in this column readable.

The cleaned and tidied master dataset was then saved as a csv file for further analysis.