Data Wrangling Report

Introduction

Data wrangling involves gathering data from a variety of sources and in a variety of formats, assessing its quality and tidiness, then cleaning it. The project requires us to carry out data wrangling (and analyzing and visualizing) on the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. The tasks in wrangling this project are as follows:

Step 1: Gathering data

Step 2: Assessing data

Step 3: Cleaning data

Step 4: Storing data

Step 1: Gathering

Data was gathered from 3 different sources:

1. WeRateDogs Twitter archive provided by Udacity:
I directly downloaded the WeRateDogs Twitter archive data (twitter_archive_enhanced.csv). Afterwards I read the data stored in the file 'twitter-archive-enhanced.csv' using panda's method 'read_csv', and stored it in a DataFrame called ‘twit_arch’.

2. Image prediction file:
I downloaded the tweet image predictions TSV using the Requests library and 'get' method. After which, I wrote it to a file called ‘image_predictions.tsv’. I then read the csv using pandas’ method “read_csv” and stored the data in a dataframe called ‘img_pred.'

3. Tweet_json data retrieved by querying Twitter's APIs and using Tweepy library.

Due to mobile verification issues, I downloaded the tweet_json.txt file from udacity site which contains data retrieved by querying Twitter's APIs. After which, I read the data in the tweet_json.txt file line by line and into a list of dictionaries (tweet_id, retweet_count, favorite_count, followers_count.) Finally, I converting and stored the list of dictionaries to a pandas dataframe called 'tweet_data'.

Step 2: Assessing data

Data were assessed programmatically and visually. Some quality and tidiness issues were observed.

Quality issues observed from assessment

1. There are retweets ratings.
2. There are tweets without images
3. There columns that will not be needed for our analysis
4. Some columns such as: tweet_id, timestamp, source are in wrong datatype format.
5. Source column is in HTML-formatted string, not a normal string
6. Some dog names are written as a,an,the etc.Those are invalid and need to be cleaned.
7. Hyperlinks are included in text.
8. The tweet_id column in image prediction and twitter API tables is in a wrong datatype.
9. There is missing data in image prediction and twitter API tables from the twitter archive table

Tidiness Issues

1. The dog stages are in four columns
2. The Twitter API, Twitter Archive table and Image Predictions tables should be merged as one.

Step 3: Cleaning data

After accessing the data, a copy of each data was made. This cleaning was carried out on these copies. The method of define, code, and test were used in the cleaning process. The table below show the action carried out on each issue.

QUALITY ISSUES	ACTIONS TAKEN
There are retweets ratings.	Remove retweets
There are tweets without images	Drop tweets without image.
There columns that will not be needed for our analysis	Remove columns in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp
Some columns such as: tweet_id, timestamp, source are in wrong datatype format.	✓ Change tweet_id to string ✓ Change timestamp to datetime format ✓ Change source to category format
Source column is in HTML-formatted string, not a normal string .	Extract html from source
Some dog names are written as a,an,the etc. Those are invalid and need to be cleaned.	Change invalid name in dog name to None.
Hyperlinks are included in text.	Remove hyperlinks from text
The tweet_id column in image prediction and twitter API tables is in a wrong datatype.	Change tweet_id to string in both data types
There is missing data in image prediction and twitter API tables from the twitter archive table	Drop missing data

TIDINESS ISSUES	ACTIONS TAKEN
The dog stages are in four column	Merge the four columns as one column ‘dog_stage’.
The Twitter API, Twitter Archive table and Image Predictions tables	Merge the twitter archive table and image predictions tables with the twitter archive table as ‘twitarchive_clean’

Step 4: Storing

After cleaning the dataset ‘twitarchive_clean' is stored in a csv file called 'twitter_archive_master.csv'. This completes the data wrangling. The next step was analysis and visualization. The visualizations are seen in the act_report.