WeRateDogs Wrangling Report

Data wrangling is the process of gathering data, assessing its quality and structure, and cleaning it before analyzing and visualizing it. The dataset on which wrangling was performed was the tweet archive of the Twitter user WeRateDogs, who rate people's dogs with humorous comments.

The different stages of the data wrangling process that were performed are as follows;

Gathering

During this stage, three different datasets were gathered in different ways and they include the following;

1. Twitter archive file: this file was already provided by Udacity and was therefore manually downloaded. After, I uploaded it to my project workspace and loaded the data into a pandas data frame after importing the pandas library.
2. Image predictions file: this file was downloaded programmatically using the requests library and a provided URL. This library was first imported as well.
3. Tweet JSON file: the data in this file was queried via the Twitter API where I created a developer account on Twitter through which I got the credentials I needed such as the consumer key and secret, and the access token and secret. By using tweet IDs in the WeRateDogs Twitter archive, I queried the Twitter API for each tweet's JSON data using the tweepy library and stored each tweet's entire set of JSON data in the tweet JSON file. This file was then read line by line into a pandas data frame. Twitter API keys, secrets, and tokens were not included in the submission which is the standard practice for APIs.

Assessing

Once all three datasets were obtained and loaded into a pandas data frame, I performed visual and programmatic assessments to detect 8 quality and 2 tidiness issues.

Visual assessment was done by just browsing through the dataframes which were already loaded into the jupyter notebook while programmatic assessment was done with the help of pandas methods such as .info(), .describe(), value_counts(), .isnull(), .duplicated(), and .shape among others.

Quality issues observed from the assessment included;

- The presence of the 'retweeted_status_id column meant that these were retweets not needed in our analysis.
- Missing values in columns such as in_reply_to_status_id and in_reply_to_user_id which had only have 78 entries out of 2356 entries and 'retweeted_status_id, retweeted_status_user_id, and retweeted_status_timestamp which also had 181 entries out of 2356 entries
- Incorrect datatypes of timestamp and tweet_id.
- The tweet JSON data frame had many columns which weren't necessary for our analysis.
- The id column in the tweet JSON data frame did not match the tweet_id name in other datasets yet they contained similar information.

Tidiness issues observed from the assessment included;

- In the twitter_archive data frame, doggo, floofer, pupper, and puppo were in different columns yet they represented one variable the dog stage.
- The three data frames twitter_archive, image_predictions and tweet_json were separate and needed to become one data frame.

Cleaning

Before cleaning, copies of the original data frames were made. After, the define, code, and test framework were followed.

The cleaning actions performed included the following;

- Keeping rows that don't have values in the 'retweeted_status_id' columns.
- Dropping rows with missing values.
- Correcting datatypes of columns such as timestamp and tweet_id.
- Dropping unnecessary columns that weren't needed in our analysis in the tweet JSON data frame.
- Renaming column id to tweet_id to match the other datasets that contained similar information.
- Creating a new column dog_stage that combined doggo, floofer, pupper, and puppo into one.
- Merging the three data frames into one on the basis of the tweet_id that they had in common.

Storing

After cleaning the issues I had identified with the data frames, I stored the cleaned and merged data frame in the archive_master CSV file.