Reporting: wrangle_report

Wrangling objective

Our goal in this phase is to gather, assess and clean three datasets from different sources for analysis. This report contains details of all the data acquisition methods (both manual and programmatic), data quality and tidiness issues identified, and cleaning task executed to improve the integrity and reliability of the datasets for worthy analysis and visualization.

Wrangling process

Step 1: Data Acquisition

The three-dataset used for this analysis and their acquisition process is detailed below.

- **Enhanced Twitter Archive**; contains basic tweet data including tweet id, text, rating, dog name and stages. This file was downloaded manually and read into pandas.

- **Image Predictions File**: contains predictions of dog breeds using a neural network based on images in tweet data. File is hosted on Udacity's servers and downloaded programmatically using the Requests library and read into pandas.

- **Tweets data**: contains additional tweet data such as favorite and retweet counts not available in our archive datasets and can be accessed via twitter API using tweepy api client.

Step 2: Accessing Data

We made copies of the datasets and carried out both visual assessments and programmatic assessment using Pandas methods to identify quality (content) and tidiness (structural) issues as presented below.

Issue no	Dataset	Type	Description
1	Twitter archive	quality	retweets and replies are part of archive dataset.
2			Incorrect dog name "a".
3			Missing dog name/ dog stages.
4			Columns contains invalid types "None".
5			timestamp column is of type str and not datetime.
6			Incorrect numerator rating of 1776 for Atticus.
7			Missing tweet data. mostly for deleted tweet data.
8	Image prediction	quality	p1, p2, p3 columns contains lowercase sometimes and capitalized other times.

9			p1, p2, p3 columns are str data type instead of category.
10			predicted dog name column p1, p2, p3 contains names other than dog names.
11	Twitter archive	tidiness	dog stages in different columns.
12	Image prediction		more than one dog breeds predictions p1, p2, p2
13	Tweet dataset		column title for tweet id “id” differs from all the other datasets.
14			tweets_df dataset is of thesame observation type as twitter_archive.

Step 3: Cleaning Data

We defined cleaning tasks and executed codes to clean our data of all issues using a define, code and test workflow. See below, all of the defined cleaning task for all the issues identified in the table above.

Issue no	Cleaning task
1	remove records in twitter_archive table where record is a retweet, reply.
2	drop records where dog name is "a".
4	replace all "None" in the dataset with NAN
5	convert timestamp column from str to datetime
6	replace outlier value of 1776 ratings for Atticus with mean rating numerator.
7	replace missing value for favorite and retweet count with mean of the respective column
8	capitalize all values in the breed column in image_prediction dataset
9	Convert data type for breed column from str to category.
11	- Define a column stage to hold dog stage names in twitter_archive using stage names in tweet text. - Drop doggo, floofer, pupper, and puppo columns from twitter_archive dataset
12	- Extract subset of image prediction that are dog breeds (p*dog = True) for each prediction to new datasets. Append sub-dataset. sort appended datasets on tweet_id and prediction confidence value in descending order. keep only predictions with the highest confidence value.
13	rename column id in tweets_df to "tweet_id"
14	Combine both twitter_archive and tweet_df.

Result

The outcome is high-quality and tidy master pandas DataFrame that was stored in a file.