**Reporting: wrangle\_report**

**Wrangling objective**

Our goal in this phase is to gather, assess and clean three datasets from different sources. these includes twitter archive data of @WeRateDogs user, image prediction data containing dog breed predictions using neural network algorithm base on dog image contained in twitter archive data and tweet data containing additional tweet information like favorite and retweet counts. 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**

The wrangling process involve three stages.

1. Data acquisition/gathering
2. Assess data for quality and tidiness issues
3. Cleaning to fix quality and structural issue that may slow analysis

### **Step 1: Data Acquisition**

This is the first stage of the wrangling process. it involves gathering the three dataset used for this analysis. details of the acquisition process and method are given below.

**Dataset 1**

* **Enhanced Twitter Archive**; contains basic tweet data including tweet id, text, rating, dog name and stages.

**Acquisition method**

* Manual download
* Uploaded to jupyter server and read into pandas.

**Dataset 2**

* **Image Predictions File**: contains predictions of dog breeds using a neural network based on images in tweet data.

**Acquisition method**

* File is hosted on Udacity's servers and downloaded programmatically using the Requests library and read into pandas.

**Dataset 3**

* **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.

#### **Acquisition method**

* Using the tweet IDs in the WeRateDogs Twitter archive, we queried data from Twitter API for each tweet's JSON data using Tweepy API client library. each tweet's entire set of JSON data is written to a json-tweet.txt file and loaded into pandas dataframe.

**Step 2: Accessing Data**

In this phase of the data wrangling process, we carried out both visual assessments using spreadsheet application like excel and programmatic assessment using Pandas methods to identify quality (content) and tidiness (structural) issues that may affect the accuracy and reliability of our analysis. Some of the quality and tidiness issues identified are presented below.

|  |  |  |  |
| --- | --- | --- | --- |
| Issue no | Dataset | Type | Description |
| 1 | Twitter archive | quality | Extraneous data; retweets and reply part of archive dataset. |
| 2 |  |  | Incorrect dog name "a" as a result of erroneous extractions or non-dog rating tweet. |
| 3 |  |  | Missing dog name/ dog stages. This will be handled in issue 1. |
| 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 (favorite and retweet count) for tweet\_id in twitter\_archive. mostly for deleted tweet data. |
| 8 | Image prediction | quality | Inconsistent case; p1, p2, p3 columns contains lowercase sometime 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. this will be taking care of when addressing tidiness issues. |
| 11 | Twitter archive | tidiness | dog stages in different columns doggo, floofer, pupper, and puppo. |
| 12 | Image prediction | tidiness | more than one dog breeds predictions p1, p2, p2 for different confidence level |
| 13 | Tweet dataset | tidiness | 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**

Haven identified all the issues above, we defined cleaning tasks/step by step to-do actions and implemented 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 or a tweet that is not a dog rating. these records have missing dog name, retweet id or reply id. |
| 2 | drop records where dog name is "a". few are as result of erroneous extraction due to inconsistent text format but majority are non-dog rating tweet. |
| 4 | replace all "None" in the dataset with NAN |
| 5 | convert timestamp column from str to datetime |
| 6 | replace outlier value of 1776 numerator 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 and confidence level p1\_conf, p2\_conf, p3\_conf to new datasets.  - Rename columns for all sub-datasets to common names  - Append all three confidence level sub-dataset to form a dataset with all dog name predictions.  - Sort appended datasets on tweet\_id and prediction confidence value in descending order.  - Remove duplicate tweet\_id to 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 into one observational type using tweet\_id |