Data Wrangling Report

Data Wrangling is the process of gathering, removing errors and transforming from raw form into desired format fit for analysis. During the course of the project, we did the following.

- Gather the data
- Assess the data
- Clean the data.

Data Gathering:

In this phase, three datasets were gathered and transformed into dataframe

- The WeRateDogs twitter archive which was manually downloaded as twitter-archive.csv
- The tweets image prediction datasets('image-prediction.tsv) was programmatically downloaded with use of the request library.
- The JSON data which consists of the tweets_id, favourite_count and retweet_count were retrieved with the tweepy library to access the twitter API. The JSON data was stored in a file tweet_json.txt

Data Assessment

In this phase we use the two methods in assessing the the datasets and they are:

- Visual Assessment
- Programmatic Assessment

Visual Assessment:
This involves the physical scrolling of the spreadsheet or dataframe inorder to pick up the errors that could be found on the dataset

Programmatic Assessment:
This is a type of assessment that uses code such as the describe, info, isnull,value_counts functions etc to assess the validity, accuracy, consistency and completeness of the data.

After carrying out the above assessment we were able to discover the following issues.

Dataset	Dimensions	Description

Tweet_archive	Validity	● Timestamp column was as string instead of daytime data type ● The tweet_id was integer instead of a string. ● The source column has a URL. ● Remove retweets
	Completeness	Missing records from the following columns('in_reply_to_status_id', 'in_reply_to_user_id','retweeted_status_id','retweeted_status_user_id','retweeted_status_timestamp', 'expanded_urls'
	Validity	● Invalid dog_names ● Inaccurate ratings.
	Consistency	Presence of outliers on both the rating_numerator and rating_denominator.
	Tidiness	● Text column contains both text and URL. ● The floofer, puppo,doggo and pupper columns are rows observations.

Dataset	**Dimensions**	**Description**
Image_pred		
	Consistency	Inconsistency in case for the p1,p2 and p3 columns.
	Tidiness	● Just one dog_prediction and confidence_level is enough.

		● Duplication of the tweet_id column.

Data Cleaning
The first stage is to copy the three datasets using the copy()

- Dropped the ('in_reply_to_status_id', 'in_reply_to_user_id','retweeted_status_id','retweeted_status_user_id','retweeted_status _timestamp' since over 90% of the values are missing.
- The erroneous data type for the timestamp column was changed using the pd.to_datetime function while that of tweet_id was astype function.
- Filtered only records that have the denominator ratings of 10.
- Went further to retain records that have the numerator ratings from 0 to 15. This was done to meet the unique rating schema of WeRateDogs tweets.
- We extracted some dog names from the text column and those that were not found were replaced with 'nameless'.
- Extracted the source of the tweet from the URL in the source column using the extract function.
- Remove Retweets from the text column using the contains() to filter out the records..
- Capitalize the p1,p2 and p3 values with str.capitalize().
- Join the floofer, doggo , puppo and pupper columns to form a single column called dog_stage.
- Finally merged all three datasets and stored them as twitter_master_df.

Conclusions

In First Iteration, I have documented several issues. But, this master data is not totally free of issues, as Data Wrangling is an iterative process.

I have stored the wrangled data in twitter_archive_master.csv file with a minor number of issues, and ready for a Data Analysis. This file has about 2139 rows and 16 features.