Data Wrangling Report

Project Objective

The aim of this project is to perform data wrangling on various dataset provided related to twitter account WeRateDogs. In addition, following the wrangling process dataset will be stored , analyzed and visualized. Finnaly, 2 different reports will be produced :

- Data Wrangling report (This report)
- Data Analysis and visualization (act_report.pdf)

Gathering Data

In this project 3 different sets of data were collected and stored in dataframe as follow:

1. The WeRateDogs twitter archive csv file "twitter_archive_enhanced.csv", that was downloaded and saved in dataframe: "df_arc" .

2. The Tweet image prediction file, "image-predictions.tsv", this file was downloaded from a provided URL

3. JSON data file ,'tweet_json.txt' was stored in a dataframe "df_api" JSON data file was obtained from Twitter API (In this case the tweet_jason.txt was downloaded manualy)

Assessing and Cleaning Data

Analyzing the data has resulted in many observations that required assessing and cleaning as follow:

Quality

Dataset	Observation	Solution
df_arc_clean	**Timestamp** type is string	Convert type to datetime
	rating_denominator some value not equal to 10	Replace values to 10

	rating_numerator values different from text	Extract value from text column . Convert type to float. Drop 0 values.
df_arc_clean df_api_clean df_image_clean	**tweet_id** type not string	Convert type to string
df_arc_clean	**Name** 745 entries with non values, name with lower case	Replace lower case name
df_api_clean	**retweet_count, favorite_count** type is string	Convert type to intiger
df_image_clean	**image_num** column of no value	Drop column
df_arc_clean	**in_reply_to_status_id , in_reply_to_user_id , retweeted_status_id , retweeted_status_user_id , retweeted_status_timestamp, source** columns to be of no value added	Drop columns

Tidiness

Dataset	Observation	Solution
df_arc_clean	**Doggo, floofer , pupper, puppo** columns describe dog place in the "Dogtionary"	Combine 4 columns into 1 column dog_style
df_arc_clean df_image_clean df_api_clean	Different columns(variables) in different table with common twitter_id	Merge tables into master dataframe

Result

Following the above data wrangling performed on the 3 different data set below are the final resuls of each dataset in addition to the final merged data set "df_master_clean"

```
df_arc_clean.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2356 entries, 0 to 2355
Data columns (total 11 columns):
tweet_id               2356 non-null object
timestamp              2356 non-null object
text                   2356 non-null object
expanded_urls          2297 non-null object
rating_numerator       2356 non-null int64
rating_denominator     2356 non-null int64
name                   2245 non-null object
doggo                  2356 non-null object
floofer                2356 non-null object
pupper                 2356 non-null object
puppo                  2356 non-null object
dtypes: int64(2), object(9)
memory usage: 202.5+ KB
```

```
df_image_clean.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2075 entries, 0 to 2074
Data columns (total 12 columns):
tweet_id    2075 non-null object
jpg_url     2075 non-null object
img_num     2075 non-null int64
p1          2075 non-null object
p1_conf     2075 non-null float64
p1_dog      2075 non-null bool
p2          2075 non-null object
p2_conf     2075 non-null float64
p2_dog      2075 non-null bool
p3          2075 non-null object
p3_conf     2075 non-null float64
p3_dog      2075 non-null bool
dtypes: bool(3), float64(3), int64(1), object(5)
memory usage: 152.1+ KB
```

```
df_api_clean.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2354 entries, 0 to 2353
Data columns (total 3 columns):
tweet_id          2354 non-null object
retweet_count     2354 non-null int64
favorite_count    2354 non-null int64
dtypes: int64(2), object(1)
memory usage: 55.2+ KB
```

```
df_master_clean.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 2073 entries, 0 to 2072
Data columns (total 27 columns):
tweet_id                      2073 non-null object
in_reply_to_status_id         23 non-null float64
in_reply_to_user_id           23 non-null float64
timestamp                     2073 non-null datetime64[ns]
source                        2073 non-null object
text                          2073 non-null object
retweeted_status_id           79 non-null float64
retweeted_status_user_id      79 non-null float64
retweeted_status_timestamp    79 non-null object
expanded_urls                 2073 non-null object
rating_numerator              2073 non-null int64
rating_denominator            2073 non-null int64
name                          2073 non-null object
stage                         2073 non-null category
retweet_count                 2073 non-null int64
favorite_count                2073 non-null int64
jpg_url                       2073 non-null object
img_num                       2073 non-null int64
p1                            2073 non-null object
p1_conf                       2073 non-null float64
p1_dog                        2073 non-null bool
p2                            2073 non-null object
p2_conf                       2073 non-null float64
p2_dog                        2073 non-null bool
p3                            2073 non-null object
p3_conf                       2073 non-null float64
p3_dog                        2073 non-null bool
dtypes: bool(3), category(1), datetime64[ns](1), float64(7), int64(5), object(10)
memory usage: 397.2+ KB
```