Data Wrangling Report

Project objectives

The project main objectives were:

- Perform data wrangling (gathering, assessing and cleaning) on provided thee sources of data.
- Store, analyze, and visualize the wrangled data.
- Reporting on 1) data wrangling efforts and 2) data analyses and visualizations.

Step 1: Gathering Data

In this phase, the three pieces of data were gathered and represented as pandas dataframes:

- The WeRateDogs Twitter archive (file on hand, manual download of 'twitter-archive-enhanced.csv')
- The tweet image predictions ('image-predictions.tsv'). This file was be downloaded programmatically using the Requests library from a provided URL.
- Each tweet's entire set of JSON data (with at minimum tweet ID, retweet count, and favorite count) in a file called 'tweet_json.txt' were stored using Twitter API and Python's Tweepy library. Each tweet's JSON data was written to its own line.

Step 2 and 3: Assessing and Cleaning Data

While working with data, a number of observations were made. In the below table there are the observations along with actions taken in the Cleaning Step.

Quality

Dataset	Observation	Solution
df_arch	**timestamp** is string and should be datetime.	Change the variable type to datetime.
	In **rating_numerator** two numerators are equal to 0.	After manual investigation no changes to the values were made.
	In **rating_denominator** there are 18 different denumerators, one of them is equal to 0.	The denominator value found in the **text** was 10 but this record was removed as it was a reply.
	In **name** more than 745 records do not contain a valid name, all names should start with a capital letter.	Some of the names could be copied from **text**, but as for the later analysis **name** was not used, no values were changed. The values in **name** were made starting with a capital letter.
	doggo, floofer, pupper, puppo columns contain 'None' value where NaN should be used. There are a few cases, where a dog has more than one style.	All 'None' values were changed to NaN. Multiplied dog styles were resolved during dataset tidying process and the logic described in the accompanying Jupyter notebook.
	In the scope of variables described in the dictionary part, there are no missing values.	No action taken.

	There could be encoding problem for tweet_id = 668528771708952576 (the **name** value uses non-English characters).	The problem was noticed during review in Excel. In pandas dataframe, the encoding seems to be correct. No action taken.
df_pred	**jpg_url** contains two different path patterns to jpg files. This seems not to have any impact.	No action taken.
	p1, p2, and **p3** are inconsistent in a way capital and small letters are used in values.	All values were made starting with a capital letter.
df_api	There are 14 erroneous (non-existing) records in this dataset which exist in other datasets.	df_arch and df_api were merged and tweets not existing in both were discarded.
all	There are different number of records in each dataset.	Same as above.

Tidiness

Dataset	Observation	Solution
df_arch	There are retweets and replies included in the dataset (represented by redundant columns).	Removed as per one of the project's requirements.
	doggo, floofer, pupper, puppo columns are all about the same things, a kind of dog personality.	The 4 columns were melted into one **dog_style**.
df_pred	**img_num** contains integer values ranging from 1 to 4 but only 1 **img_url** is present (this column semantics is not clear). The column may not have any use here.	Removed as not needed for further analysis.
all	There are too many datasets and their overall structure is untidy.	2 tidy datasets were created.

Result

As a result, 2 tidy analytical views were created ready for data analysis:

```
df_arch_clean.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 2096 entries, 0 to 2095
Data columns (total 12 columns):
tweet_id              2096 non-null int64
timestamp             2096 non-null datetime64[ns]
source                2096 non-null object
text                  2096 non-null object
rating_numerator      2096 non-null int64
rating_denominator    2096 non-null int64
name                  1493 non-null object
rating                2096 non-null float64
dog_style             336 non-null object
retweet_count         2096 non-null int64
favorite_count        2096 non-null int64
followers_count       2096 non-null int64
dtypes: datetime64[ns](1), float64(1), int64(6), object(4)
memory usage: 212.9+ KB
```

```
df_pred_clean.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2075 entries, 0 to 2074
Data columns (total 11 columns):
tweet_id    2075 non-null int64
jpg_url     2075 non-null object
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(4)
memory usage: 135.8+ KB
```