Data Wrangling Project:

Project objectives:

It is a data wrangling Project that aims to:

- Perform data wrangling process of data that comes from a twitter account (@dog_rates) also called WeRateDogs.
    - Gathering Data.
    - Assessing Data.
    - Cleaning Data.
- Store, Analyze and Visualize the wrangled data.
- Reporting (data wrangling efforts, data analysis and visualization)

Data Wrangling Efforts:

Step 1: Gathering Data:

In this step we gathered data from three different sources:

1. The WeRateDogs Twitter archive.
2. The tweet image predictions, i.e., what breed of dog (or another object, animal, etc.) which is generated from a neural network.
3. Each tweet's retweet count and favorite ("like") count at minimum. Using the tweet IDs in the WeRateDogs Twitter archive, by querying the Twitter API for each tweet's JSON.

Step 2: Assessing data:

In this step we assessed our data to find any (Quality or Tidiness) issues in the gathered data.

Step 3: Cleaning Data:

In this step we cleaned issues we found in the Assessment step.

- **Quality issues cleaning:**

DataFrame	**Assessment**	**Cleaning**
Enhanced Twitter Archive DataFrame	The tweet_id column is of data-type (int64) instead of (str)	convert it into (str) data type.
	The (timestamp) column is of data-type (str) instead of (datetime).	Convert the (timestamp) column into datetime data-type.

	The (in_reply_to_status_id) column has 78 non-null values and we are interested in origenal tweets not replies.	Removing the reply-tweets based on non-null values in (in_reply_to_status_id) column.
	The (retweeted_status_id) column has 181 non-null values and we are interested in origenal tweets not Retweets.	Removing rows that represent Retweets based on (retweeted_status_id) column
	The (source) column is in html format we can extract the source of the tweet in form categoriesto compare sources.	extract only the source name and categorize them
	The (rating_numerator) column should of type float, also it has values below 10 and this isn't valid.	Converting (rating_numerator) into (float64) data-type instead of (int64) and fixing values (<= 10) to be at least (= 11)
	The (rating_denominator) column has values not equal to 10 ratings and this isn't valid.	Fixing rating_denominator values that are not equal to (10) to be (10)
	The (name) column has (None) names and many invalid values starts with lower-case letter (e.g. a, an, the, such ... etc).	we will put them as NaN.
Image Preditions DataFrame	The (tweet_id) column is of data-type (int64) instead of (str).	should be converted to (str)
	In (p1, p2, p3) there are values that start with a lower-case letter.	In (p1, p2, p3) the first character in names should be Upper-case.
	In (p1, p2, p3) there are underscores in categories of dogs instead of spaces.	underscores should be replaced with spaces
Twitter API DataFrame	The (id) column is of data-type (int64) instead of (str).	The id column is of (int64) data-type and should be converted to (str)

- **Tidiness Issues:**

DataFrame	Assessment	Cleaning
Enhanced Twitter Archive DataFrame	There are 4-columns that represent the "Dog Stage" (i.e. doggo, floofer, pupper, and puppo).	we can represent all in one column and remove these four un-necessary columns.
Image Predictions DataFrame	img_num column isn't necessary and provide no additional information related to prediction.	img_num column isn't necessary and provide no additional information related to prediction.
Twitter API DataFrame	We only need 3-columns (i.e. id, retweet_count, favorite_count).	Select only the desired columns.
	We need to change The name of (id) column into (tweet_id) like other DataFrames.	Rename The column.
There are Three DataFrames that need to be combined into one DataFrame.		