Data wrangling process

1. **Gathering data:** collecting data from different sources, our data come from three data frame:
    a. **twitter_archive_enhanced.csv**
    b. **image_predictions.tsv**
    c. **Twitter API**
2. Assessment: here we assess the data, and convert dirt data to quality data, and find quality and tidiness issues.
3. Cleaning: convert the data from dirt to quality by removing quality and tidiness issues.

Twitter Archive Data Frame

Quality issues:	**Cleaning Quality issues:**
1. duplicated in expanded URLs that means duplicated images and retweeted tweets.	Drop duplicated rows, remove retweet
2. missing values in expanded URLs rows, every row should have image.	Drop rows without URL in expanded URL column
3. some rows in [expanded URLs] column has two URL not one.	Split the rows and drop the original column, Note: drop the duplicated rows will be found again
4. founded invalid URLs while cleaning from https://vine.co website and delete other URL outside twitter.	Make a list of websites not twitter, make a loop to clean the unwanted websites rows.
5. Removing tweets not belong to @dog_rates.	Use (loc) to index the URLs not belong to @dog_rates
6. The rating numerator column should of type float and also it should be correctly extracted.	Change to float, extract correctly value and replace it. drop rows with exactly wrong in text of tweets.
7. rating denominator values have typos error also. the most rating is 10 but, in the data, we found 110	Change the rows value manually.
Tidiness issues:	**Cleaning Tidiness issues:**
1. for our analysis we have more columns will not be useful	Drop unnecessary columns
2. classification of dogs (doggo, floofer, pupper, puppo) not good for analysis (untidy).	convert columns to one column with value [pd.melt]

Image prediction Data Frame

Quality issues:	**Cleaning Quality issues:**
1. duplicated image in jpg_url column.	Drop duplicated image
2. drop rows with False value in p1 to p3, that means photos with False value not containing dogs.	Drop rows with false value in any prediction columns.
3. not all the text in p1,p2,p3 captilazed.	Str.title()
4. change the the sperated columns by underscore to space	Replace _ to space
5. the img_num column not showing and ditals and dose not have any meaning.	Drop "img_num"
Tidiness issues:	**Cleaning Tidiness issues:**
1. after assessing the image prediction dataset visually we found that the first prediction p1 is the most true predictions of types of dogs in photos, so we will remove the the another predictions from data frame.	Drop not useful columns in my analysis Columns = [p2 p2_conf p2_dog p3 p3_conf p3_dog]
2. The data separated to three table (untidy)	merge cleaned data frame in one table (df_arch_twitt_copy, df_img_predict_copy, api_df)