Data Wrangling: We Rate Dogs

Data wrangling consist of:

- Gathering
- Assessing
- Cleaning

GATHERING

To begin with Gathering,I collected data from three different sources.

1. WeRateDogs downloaded their Twitter archive and sent it to Udacity via email.I downloaded the csv file manually on clicking the link provided by Udacity.

2. The tweet image_prediction.tsv file is available on Udacity's server I downloaded the file using requests library.

3. Using the tweet IDs in the WeRateDogs Twitter archive, I query the Twitter API for each tweet's JSON data using Python tweepy library and stored each tweet's entire set of JSON data in a file called tweet_json.txt file

DATASET INTRODUCTION

archive dataframe

Dataframe consists of :

- tweet_id
- rating_numerator
- rating_denominator
- name
- dog_stages(puppo,pupper,doggo,floofer)
- Timestamp

image_pred dataframe

Dataframe consists of :

- tweet_id
- jpg_url
- confidence level of prediction represented by prediction percent(p1_conf,p2_conf,p3_conf)
- breed of dog(or other object,animal,etc.) present in each tweet represented by p1,p2,p3

twitter_archive_additional dataframe

Dataframe consists of :

- tweet_id
- retweet_count
- favorite_count

ASSESSING

After visual and programmatic assessments of datasets I have come up with following quality and tidiness issues:

QUALITY

ARCHIVE TABLE

SR.NO	ISSUE	SOLUTION
1	Missing data is represented by None instead of NaN in • name • doggo • floofer • pupper • puppo	Using .loc method I addressed desired rows with None,column and then replaced them using np.nan
2	rating numerator has few records with numerator less than 10 ideally	Query archive table with rows having numerator less than

	numerator should be greater than 10	10.By visual assessment I found 2 rows with erroneous numerator rating and then changed them with .loc method
3	rating denominator has few records less than 10 ideally denominator should be 10	Query archive table with rows having denominator greater than 10.By visual assessment I found 5 rows with erroneous denominator rating and then changed them with .loc method
4	Names of dogs are not names i.e. are a,an,the,such	First letter of dog names is uppercase but inappropriate names are all lowercase.So,I found all lowercase words which are not names in general and replace them with NaN
5	we want original rating but dataset contains retweets.	Requirement of project is to get original tweets i.e. no retweets.I decided to get the retweet_status_id's index and delete them from archive table using pandas drop method.
6	original ratings with no image	Another requirement of project is to get original ratings with image but our dataset contains few ratings with no image.So I used image_pred table and kept only those id's present in image_pred using isin method.
7	tweet_id,in_reply_to_status_id,in_reply_to_user_id,retweeted_status_user_id,retweeted_status_id columns are int instead of str	Using astype method I changed tweet_id,in_reply_to_status_id,in_reply_to_user_id,retweeted_status_user_id, retweeted_status_id columns to str
8	timestamp,retweeted_status_timestamp columns are object instead of datetime	Object columns are changed to datetime of timestamp,retweeted_status_timestamp columns.

9	Doggo's name is Zoey but mentioned as my.	I found rows containing name 'Zoey' using str.contains and changed them from my to Zoey using .loc method.

IMAGE_PRED TABLE

1	datatype of tweet_id is int,it should be str	Using astype method I changed tweet_id,in_reply_to_status_id,in_reply_to_user_id,retweeted_status_user_id, retweeted_status_id columns to str

TWITTER_ARCHIVE_ADDITIONAL

1	datatype of tweet_id is int,it should be str	Using astype method I changed tweet_id,in_reply_to_status_id,in_reply_to_user_id,retweeted_status_user_id, retweeted_status_id columns to str

TIDINESS

SR.NO.	ISSUE	SOLUTION
1	archive table and twitter_archive_additional tables should be combined.	Use pd.merge method
2	Melt all the four stages of dogs in `archive table`	Melt doggo,floofer,pupper,puppo columns to dog_stage. - If I try to melt the columns (doggo,floofer,pupper,puppo) I will get multiple duplicates that would be difficult to work with.Instead I have

		planned to make four different dataframes for doggo,floofer,pupper and puppo and then accordingly make column dog_stage for each of those dataframes.After that append them to make single Dataframe.
3	text column in `archive table` should be split into text and url	Use regular expression to extract tweet url from text column
4	image_pred has multiple columns with similar information.	Use pd.wide_to_long method

RESULTS

Finally cleaning above quality and tidiness issues, it resulted in two datasets **archive** and **image_pred**:

- **archive** table has 2005 observations and 17 columns
- **image_pred** table has 6225 observations and 7 columns