DATA WRANGLING REPORT

By Mohammed Youssry

As an assignment for the Udacity Data Analysis Nanodegree; This report illustrates the main steps involved in Data Wrangling of Twitter account " We Rate Dogs".

DATA GATHERING

In this step, data collected from three main sources

1- Twitter Archive Enhanced, it was formatted as twitter_archive_enhanced.csv
This file was delivered by email and downloaded manually then uploaded to workspace by using pandas function "pd.read.csv"
2- Image_predictions.tsv ,which were hosted by webpage and downloaded using requests library get function then read using pandas function "pd.read.csv"
3- The final dataset gathered from TWITTER API through tweepy library by querying API to get extra informations related to the tweet_IDs in the first file eg,retweet_count and Favorite_count.

DATA ASSESSMENT

In this step, we checked our imported datasets both vissually and programmatically for detecting Quality And Tidiness Issues.

1- The visual assessment scrolling through the data in software application (Excel , Google sheets ,Text editor,…etc.)
2- Programmatic assessment :using code to view specific portions and summaries of the data (pandas head ,tail ,info ,and describe methods for example).
3- After both visual and programmatic assessment , Noticed that:

A- Tidiness Issues:

1- Dog types are separated to 4 columns.
2- All data are related but divided to 3 datasets.

B- Quality Issues:

1) Twitter Enhanced Archive:
    1- Tweet_id datatype are integer not string.
    2- Timestamp datatype are string not datetime.
    3- There are 181 retweets are indicated by retweeted_status_id.
    4- Some dog names are invalid like(such , a ,and an) instead of the name.
    5- There are 440 rating numerator less than 10.
    6- There are 23 rating denominator not equal 10.
    7- (rating_numerator and rating_denominator)datatype are 'int' not 'float'

2) Tweet Image Predictions:
    1- There are 66 Jpg_url duplicated, so we have some IDs with missing photos .
    2- Columns (P1 , P2 , P3) have undercrosses in multi_words names , instead of spaces.
3) Twitter Api:
    1- Missing Entries (have only 2354 instead of 2356).

CLEANING DATA

In this step , we are following the structure for this process(work flow): Define, Code, Test.
The first step : by creating copies from all dataframes. Through using .Copy( ) Functions.
then , we moved to solve both Tidiness and Quality Issues. As following.

Table Name	***Quality Issues***	***Solution***
Twitter Enhanced Archive	convert tweet_id datatype to 'str' in all dataframes	By using .astype(str) function
	convert 'timestamp' datatype to datetime	By using .to_datetime function
	convert (rating_numerator and rating_denominator)datatype to 'float'	By using .astype(float) function
	check for unclear dog names	By using df ['name'].unique() function
	replace incorrect name with NaN	By using .replace function.
	put all NaN values in "None"	By using .fillna(value="None", inplace=True) function.
	Remove rating_denominator which are less than 10.	By using .drop function
Image Prediction	Delete rows with missing photos	By using .drop function
	Replace(_) and (-) to spce in p1,p2 and p3 columns.	By using .replace('_',' ') function

	Tidiness Issues	*Solution*
	melt dog types into one column.	By using (pd. Melt) function
	Merge all three datasets to one datasets	By using (pd .merge) function

After cleaning , we get file which are ready for analyse and visualize.

Storing Data

After creating file ready for analysis and visualization we save it as(twitter archive master)with _csv format by using to_csv('twitter_archive_master.csv',index=False) function