UDACITY DATA ANALYST DEGREE

Data Wrangling Project

Submitted by:

Janely Padillo

1. **GATHER**
    1. The first dataset 'twitter-archive-enhanced.csv' was provided and was only needed to be read through the pd.read_csv() method.
    2. The second dataset was linked on Udacity's servers, which I had downloaded programmatically using the Requests library from the following URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv
    3. Due to coming across difficulties in applying for the twitter developer account, the json .txt file containing the Twitter query was provided for me. I read this .txt file line by line into a pandas dataFrame using a combination of os module and for loops.

2. **ASSESS**

    Assessing by way of visual as well as programmatic assessment using python methods such as info(), head(), tail(), sample() and isnull(), among others.

3. **CLEAN - This was performed chronologically for the most part, ensuring that tidiness issues are mostly addressed first and quality issues are addressed afterwards.**
    1. Columns floofer, doggo, pupper and puppo were merged into one column. This was done using replace() and fillna() methods.
    2. Rating was standardized and merged within one column using vectorized operations.
    3. Impertinent columns were dropped and all three dataFrames were merged after quality issues have been addressed by utilizing the drop() method and succeeded by the merge() method.
    4. There are retweets included in the data, which were removed. Columns retweeted_status_user_id, retweeted_status_id, retweeted_status_timestamp were removed using the drop() method.
    5. Datatype of tweet_id was updated from an integer to an object using the apply() method.

f. Datatypes for timestamp and created_at were updated from a string to timestamp using the to_datetime() method.
g. Some names are erroneously parsed to and have empty values or lowercase strings. These are updated using a combination of for loops, loc(), str(), islowercase() and tolist() methods.
h. Rating numerators and denominators were erroneously parsed. wherein Some numerators are less than 10 and denominators are more or less than 10. Denominators are, equivocally, always 10, and numerators are always 10 or greater. These were updated using a combination of for loops, at() and max() methods.
i. Datatypes for p1_conf, p2_conf, p3_conf and img_num were updated from an object to an integer using a combination of to_numeric(), fillna() and astype().
j. Inconsistent text cases for prediction1, prediction2, prediction3 were addressed by creating a function called to_title, which utilized str.title() method under the hood.
k. Column names were updated using column assignment.