Udacity WeRateDogs Project

Wrangling Report

By: Abdallah Aboudeif

November 2020

In this project data consists of three datasets:

1- The WeRateDogs Twitter archive. This archive contains basic tweet data (tweet ID, timestamp, text, etc.) for all 5000+ of their tweets of Twitter user @dog_rates. It provided as twitter_archive_enhanced-2.csv file. I download this file manually by clicking its link and used the file directly by opening it using pandas.read_csv method.

2- The tweet image predictions, a file (image_predictions.tsv) is hosted on Udacity's servers, It has a definition of what breed of dog (or other object, animal, etc.) is present in each tweet according to a neural network. I downloaded programmatically using the Requests library.

3- Tweet_json.txt file. Which is a text file of JSON data extracted from query the Twitter API for each tweet's JSON data using Python's Tweepy library. Each tweet's retweet count and favorite ("like") count at minimum, and any additional interesting data. I used Python code to read this .txt file line by line into a pandas Data Frame to extract tweet ID, retweet count, and favorite count.

After gathering those three datasets into three Data Frames. I started assessing data samples, data types, data shapes and values to find data quality and tidy issues.

I make this assessing using two ways:

A. Visual assessment: by open files in Microsoft Excel sheets and show random samples of Data Frames

B. Programmatic assessment: Using Data Frame methods and python functions like DataFrame.info(), DataFrame.groupby(), DataFrame.value_counts(), sum(), count(), etc.

Doing data assessment I got these data quality and tidy issues:

Quality issues:	
1	Data type of timestamp column in archive dataframe is a string and it should be in datetime type
2	Data type of tweet_id columns in all dataframes are a integer while it should be in string type since we don't need to make calculations on it
3	Invalid values in the source column in the archive_df dataframe which has HTML tags
4	In twitter archive dataframe there are 18 rating in rating_denominator column doesn't have the value 10 and it should be 10
5	There are 28 value in rating_numerator column (in twitter archive dataframe) that have a value <= 10 while it should be greater than 10
6	There are 181 retweets in twitter archive dataframe
7	There are 324 rows of tweets in image predictions dataframe doesn't have a validate dog picture
8	There are 78 replies in twitter archive dataframe
9	There are tweets in twitter archive dataframe doesn't exists in picture predictions dataframe
Tidy issues:	
A	Twitter archive, twitter API and image predictions DataFrames are single observational unit but it's stored in multiple tables
B	In the archive_df dataframe, the dog stage feature is represented by 4 columns • Doggo • Floofer • Pupper • puppo while each observation should forms one row

By finding data issues I started to clean every issue I found. First I copied all Data Frames to easily code test backwards. Then I cleaned issues separately on three steps:

- Define: Wrote what I should do to solve the issue
- Code: Wrote the code that solve the issue
- Test: Saw if the issue is solved or not

These issues definition I mad was as shown:

Definition of quality issues:	
1	Convert data type of timestamp column in archive dataframe to datetime using astype function
2	Convert data type of tweet_id columns in all dataframes to string using astype function
3	Delete the HTML tags from source column using replace function
4	Write 10 in all rating_denominator column values using assignment
5	Delete rows that have a value <= 10 in rating_numerator column using Dataframe Selection
6	Delete all retweets in twitter archive dataframe using Dataframe Selection
7	Delete 324 rows from image predictions dataframe that doesn't have a validate dog picture using panda's .drop() function
8	Delete all replies rows from twitter archive dataframe using Dataframe Selection
9	Delete all tweets in master dataframe doesn't have a picture using DataFrame selection
Definition of tidy issues:	
A	1- Delete useless columns from twitter archive and image predictions dataframes using DataFrame slicing 2- Merging image predictions, twitter archive and twitter API into master dataframe using panda's .merge() function
B	B- Represent the dog stage feature in one column instade of 4 by: 1- remove 'None' in [doggo-floofer-pupper-puppo] columns using .str.replace() function 2- creating dog_stage column 3- fill it by values of 4 columns if exist using assignment 4- remove the unneeded four columns using panda's .drop() function

Now I got 1 Data Frame called ‘master_df’ that I stored in a file called 'twitter_archive_master.csv' which I use to analyze and visualize data next.