Data Wrangling Report

By Michael Tharwat Shafek

November 2020

According to the Udacity Data Analysis professional program “project#2” , This report illustrates the main steps followed in data wrangling of twitter account “WeRateDogs” .

Data Gathering

In this step data was collected from three main sources as following :

1. ‘twitter_archive_enhanced.csv’ file , this file is downloaded manually from Project resources in the course to our working directory then imported to project environment using Pandas function : pd.read_csv() .
2. ‘image_prediction.tsv’ file , this flat file is downloaded programmatically from its relevant URL using requests library get function , and it include image predictions for the dog’s breed obtained through neural network on most of tweets in Twitter archive enhanced file .
3. ‘tweet_json.txt’ file, this file is gathered from Twitter API using Tweepy library by querying API and store it in text file then read this text into Data Frame using json library load function .

Data Assessment

In this step we start investigation of our imported data sets visually and programmatically to detect quality and tidiness issues .

- The Visual assessment was done by using spread sheet ‘excel’ for CSV &TSV files then programmatic assessment is done by using jupyter notebook .
- There was missing data , data with inappropriate type , messy data , extra data that not needed according to project specs like retweet data , quality issues were addressed in the three data sets , but not all in once , I returned to assessment many times after I moved to cleaning and analysis .
- Tidiness issues also was raised in parallel with quality issues such as : more Columns about one variable about dog stages in Twitter archive enhanced file .

Cleaning Data

In this step we start to clean our data frames according to project needs and data assessment collected before in the quality cycle : 'Define' , 'Code' and 'Test'

- Starting with data quality , converting some types of data , dropping the extra data that not related to project .
- Working in Tidiness issues in twitter archive table to support then in quality issues by dropping NAN values .
- Handling 'name' Column in twitter archive using regex to extract it from text column in assistance of regular library 're' , also this library helped me in cleaning of rating_numerator columns in same data set .
- Moving to some other Tidiness issues , Ex: merging tweets_df with twitter archive as both of them about one observational unit 'Tweets' , then I preferred to include image prediction in another table as it include data about images .

As result of above steps , Here we come with two Tables data frame "twitter_archive_clean" and "image_prediction_clean"

Storing Data

In this step we merged the two data frames to one data frame 'twitter_archive_master' and store it to 'twitter_archive_master.csv' using pandas function df.to_csv() .

Analyzing Data

In This step we start to analyze data using pandas data frame functions and plotting libraries , 'matplotlib' and 'seaborn' and to visualize graphs inside jupyter notebook we used %matplotlib inline .

Analyzing give us fruitful insights about most lovely dog breed , most popular dog names , what the most dog stage has more interaction in retweets and likes , also it showed correlation between retweet counts and favorite counts .