Data Wrangling Report

By: Abdelrahman Saeed Mohammed
Jan 2021

As an assignment for the *Udacity* Data Analyst Nanodegree in this report I will talk about the main steps involved in the data wrangling of Twitter Account “WeRateDogs”.

Data Gathering

In this step collecting data takes place. For this project there was three sources to take data from it with three different ways :-

1. The First file is “*twitter-archive-enhanced-2.csv*”. This file was delivered by email and downloaded manually then imported to our working environment using Pandas Library functions.

2. The Second file is “*image-predictions.tsv*”. This file has been hosted on a webpage and downloaded from its reverent URL using Requests Library and opened by Pandas Library functions. This file has image prediction breeds from dog images on the tweet using neural networks.

3. The Third file "*tweet-json*". This file was created and gathered from twitter REST API via Tweepy Library to extract more information pertinent to the tweets’ ids in the first file like retweet counts and favorite counts.

Data Assessment

In this step we investigate our imported datasets programmatically and visually for quality and tidiness issues.

Visual Assessment:

- The Visual assessment done on a spreadsheet application like Excel named LibreOffice Calc.

Programmatic Assessment:

- The programmatic assessment is conducted in Jupyter NoteBook using pandas Library functions.

We managed to find some missing values and messy structure in our datasets and that helped and guided us what to do in the cleaning step. We found some quality issues and some tidiness issues.

Quality Issues

twitter-archive-enhanced-2.csv

- Data type of 'timestamp' column is object type.
- Data type of 'tweer_id' column is integers type.
- There are retweets and replies in the datasets.
- Null values are represented as "None" strings in some columns ('name','doggo','pupper','puppo','floofer').
- Some values in the 'name' column are errores like 'a' or 'an'.
- Some values in the 'expanded_url' column are missing.
- Some values in 'rating_numerator' and 'rating_denominator' are incorrect and weird.

image-predictions.tsv

- Non descriptive columns name.
- Inconsistent capitalization for predicted breeds ('p1','p2','p3')

Tidiness Issues

twitter-archive-enhanced-2.csv

- Values are columns ('doggo','puppo','pupper','floofer') instead of 'dogs_stage'

tweet_json

- This isn't considered as an observational unit to have its own table need to be merge to arc_twitter

Data Cleaning

In this step we tried to fix the issues in our datasets one by one. First we made a copy of our original dataframe to have the ability to return if we made a mistake and retry to fix it.
Here is what we have done:

Table Name	Quality Issues	Solution
df_arc_twitter	Data type of 'timestamp' column is object type	Changed the datatype of 'timestamp' column to datetime type using pandas function.
df_arc_twitter	Data type of 'tweer_id' column is integers type.	Changed the datatype of the 'tweet_id' column in the arc dataframe to string type.
df_arc_twitter	There are retweets and replies in the datasets.	Dropped retweets and replies in datasets then dropped these unneeded columns.
df_arc_twitter	Null values are represented as "None" strings in some columns ('name', 'doggo', 'pupper' ,'puppo', 'floofer').	Replaced “None” strings with NaN values.
df_arc_twitter	Some values in the 'name' column are errores like 'a' or 'an'.	Gathered the correct name from the tweet text if it existed.
df_arc_twitter	Some values in the 'expanded_url' column are missing.	Dropped as this tweet doesn't have feature images.
df_arc_twitter	Some values in 'rating_numerator' and	Extracted the correct values programmatically and manually if existed.

	'rating_denominator' are incorrect and weird.	
df_image_predictio n	Non descriptive columns name.	Renamed these columns manually.
df_image_predictio n	Inconsistent capitalization for predicted breeds ('p1','p2','p3')	Applying capitalization method on the entire column.

Table Name	Tidiness Issues	Solution
df_arc_twitter	Values are columns ('doggo','puppo','pupper','floofer').	Combined in one column named “dog_stage”
df_tweet_json	This isn't considered as an observational unit to have its own table.	Merged to df_arc_twitter table