Wrangling & Analyze Data Project

Wrangling Report

06.10.2021

—

Mohamed BOUSETTA MAHJOUB
Udacity Data Analyst Nanodegree

Overview

In this report, I will describe my data wrangling efforts (Gather, Assess & Clean) on WeRateDogs Data. The dataset that is wrangled is the tweet archive of Twitter user **@dog_rates**, also known as **WeRateDogs**. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. These ratings almost always have a denominator of 10.

Steps

I. Gathering Data

The data for this project consist on three different dataset that were obtained as following:

- ❏ **Twitter archive file**: the *twitter_archive_enhanced.csv* was provided by Udacity and downloaded manually.

- ❏ **The tweet image predictions** : the breed of dog is present in each tweet according to a neural network. This file (*image_predictions.tsv*) is hosted on Udacity's servers and was downloaded programmatically using the Requests library and URL information

- ❏ **Twitter API & JSON**: by using the tweet IDs in the WeRateDogs Twitter archive, I queried the Twitter API for each tweet's JSON data using Python's Tweepy library and stored each tweet's entire set of JSON data in a file called tweet_json.txt file. I read this .txt file line by line into a pandas dataframe with tweet ID, favorite count, retweet count and url.

II. Assessing Data

Once the three tables were obtained I assessed the data as following:

- ❏ *Visually* : by printing the three entire Dataframes separately in Jupyter Notebook
- ❏ *Programmatically*, by using different methods (e.g. info, value_counts, sample, duplicated, groupby, etc)

Then I separated the issues encountered in quality issues and tidiness issues :

1. Quality

a. archive table

i. Missing Data :

- in_reply_to_status_id
- in_reply_to_user_id
- retweeted_status_id
- retweeted_status_user_id
- retweeted_status_timestamp
- expanded_urls

ii. Erroneous datatypes:

- timestamp : string not datetime
- tweet_id : integer not string
- type of rating_numerator should be float

iii. Inaccurate data :

- tweets have Wrong rating like 24/7
- tweets don't have the right rating (9/11 instead of 14/10)
- tweets don't have the right format like 165/150
- some tweets have rating_numerator as float type
- rating_numerator < 10
- name = 'a'

b. images table

i. Missing Data : 324 rows without dog breed prediction

c. tweets table

i. Missing Data : 272 missing values in expanded_urls column

2. Tidiness

- All tweets have almost the same source value in archive table
- One variable (dog stage) in four columns (doggo, floofer, pupper,pupper) in archive table
- Some rows have more than one different stage in archive table
- Dog breed prediction should be in one column in images table
- All three tables should be in one table

III. Cleaning Data

This part of the data wrangling was divided in three parts: Define, code and test. These three steps were on each of the issues described in the assess section :

- First, I create a copy of three original Dataframes
- Second, I start by cleaning the missing Data (unneeded columns, retweeted rows , missing urls)
- Third, I tried to fix the Tidiness issues by :
  - Creating one column for dog stage instead of four columns
  - Extracting the dog breed from image prediction table and put it in one column
  - Merging the three tables in one master table
- Finally, I finish by cleaning the Quality issues :
  - Change the type of timestamp & tweet_id
  - Correct the rating values , format & types
  - Change the wrong names