Udacity Data Analyst Professinal Nanodegree - Project "Data Wrangling and Analyzing"

by Mohamed Awwad

Introduction

Using Python and its libraries, I will gather data from three sources, assess its quality and tidiness, then clean it.
The dataset that is wrangled is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs.

Gathering data

Twitter archive file

- The file "twitter_archive_enhanced.csv" downloaded manually from Udacity resources

Tweet image prediction

- The tweet image predictions: i.e., what breed of dog (or other object, animal, etc.) is present in each tweet according to a neural network. This file (image_predictions.tsv) is hosted on Udacity's servers and should be downloaded programmatically using the Requests library and the following URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv

Twitter API File

- Twitter API file contains tweet id, favorite count and retweet count. Data was provided by Udacity, downloaded manually then will be loaded from the tweet-json.txt file into a pandas data frame

Assessing Data

Assess data visually as well as programmatically using pandas for quality and tidiness issues. Inspecting the dataset for two things: data quality issues (i.e. content issues) and lack of tidiness (i.e. structural issues).

Dataset 1 - twitter_archive

Quality & Tidiness Issues in twitter_archive

Quality

- in_reply_to_status_id , "in_reply_to_user_id" , retweeted_status_id , retweeted_status_user_id has wrong data types
- timestamp and retweeted_status_timestamp are not a datetime format
- 128 retweets are present in the data
- 78 reples are present in the data
- rating_denominator column has 20 values other than 10
- rating_numerators are not always correctly accounting for decimals
- rating_numerator column has values less than 10 as well as some very large numbers (e.g. 1176)
- the dog names are not standardized
- name column: none appears 745 (missing data but not NaN)
- name column: some names are false (O, a, not..)
- unnecessary html tags in source column in place of utility name e.g. <a href=""http://twitter.com/download/iphone"" rel=""nofollow"">Twitter for iPhone
- pupper, puppo, floofer and doggo columns: For 1976 IDs there are no dog "stage" information.
- we have 639 expanded urls which contain more than one url address.

Tidiness

- source and expanded_urls have several informations inside them
- pupper, puppo, floofer and doggo columns should be merged into one column
- retweet columns not needed

Dataset 2 - image_predicition

Quality & Tidiness Issues in Image Predictions

Quality

- the dataset has 2075 entries, while twitter archive dataset has 2356 entries.
- "tweet_id" is int, should be type object as no calculation is needed
- 66 jpg_url duplicates were found.
- column names are confusing and do not give much information about the content.
- dog breeds contain underscores, and have different case formatting.
- only 2075 images have been classified as dog images for top prediction.

Tidiness

- the dog breed prediction could be merged into one column (breed_prediction)
- the prediction confidence could be merged into one column (prediction_confidence)
- dataset should be merged with the twitter archive dataset.

Dataset 3 -Twitter API Data

Quality & Tidiness Issues in Twitter API Data

Quality

- tweet_id is int, should be type object as no calculation is needed
- twitter archive dataset has2356 entries, while twitter API data has 2356.
  (2356-2356) =20 missing IDs

Tidiness

- 2 columns retweet_count and favorite_count should be joined with the twitter archive dataset.

Cleaning Data

Using pandas, clean the quality and tidiness issues identified in the Assessing Data section.

Clean: 1.Twitter Archive Data

- Remove retweets data from the dataset
- Drop retweet and replay columns
  retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp, in_reply_to_status_id and in_reply_to_user_id
- Convert timestamp columns to date
- Replace all faulty names to none
- Create the "dog_stage" column by combining columns "doggo", "floofer", "pupper" and "puppo"
- Recorrect links by using tweet id
- Strip all html anchor tags (i.e. `<a..></a>`) in source column and retain just the text in between the tags.
- Convert the datatype from string to categorical.

Clean: 2. image predication

- Change column names to more descriptive ones.
- Dog breeds contain underscores, and have different case formatting
- Drop 66 image_url duplicated
- Create a new colum (breed_prediction) for the predicted dog breed (if the first prediction wasn't a dog breed, take the second and so on) and the confidence of the prediction as a second column

Clean: 3. Twitter API File

- Change column names to more descriptive ones.
- Merge the three datasets using inner merge (to get a dataframe only for matching IDs)
- Change Tweet_id int to str (object)
- Create a new column for the rating of (rating_numerator / rating_denominator)

Storing, Analyzing, and Visualizing Data

Storing

- Store the clean DataFrame(s) in the CSV file ***twitter_archive_master.csv***

Analyzing and Visualizing Data

- The most frequent reported dog stage
- Analyzing number of tweets posted by WeRateDogs over time¶
- Number of dogs were rated above 10
- The most used Twitter source
- The top 10 most frequent predicted dog breeds
- The top 10 most frequent dog names
- Analysis of retweet and favorite counts¶