Project: WeRateDogs - Wrangle and Analyze Data

Data Wrangling Report

Presented by Paula Munoz

Introduction

Throughout this project I will gather, assess and clean data related to the Twitter account @dog_rates, also knowns as WeRateDogs to create an interesting and trustworthy analysis and visualizations.

About WeRateDogs

WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog, and these ratings almost always have a denominator of 10. and the numerator is almost always greater than 10. Some examples of ratings are: 11/10, 12/10, 13/10, etc.

Data Wrangling Steps implemented

1. Data Gathering
2. Data Assessment
3. Data Cleaning

1. Data Gathering

Data was gathered from three different places and in different formats:

- ***twitter_archived_enhanced.csv*** file, this file was provided by Udacity and contained information about WeRatedogs tweets, the file contains 2356 records and 17 columns.

  This file was loaded and saved and saved as a dataframe with the name “df”

  ```
  #Load twitter Archive dataset
  df = pd.read_csv('twitter-archive-enhanced.csv')
  ```

- ***image_predictions.tsv*** file, this file is provided and hosted by Udacity, this dataset contains image predictions based on images provided on tweets from archived file,

  This file was downloaded from the internet by using Python’s request library.

  File was saved as a dataframe with the name “df_images”

```
#Download tweet image predictions tsv file from Internet by using
requests library
url=
"https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-
predictions/image-predictions.tsv"

response = requests.get(url)

with open('image_predictions.tsv', mode = 'wb') as file:
    file.write(response.content)

#Load tsv file
df_images = pd.read_csv('image_predictions.tsv', sep = '\t')
```

- ***Tweet_json.txt*** file, the goal of this file is to gather retweet and favorite counts for the tweet Ids provided on archived file.

  This file is gathered by using Twitter API, in order to accomplish this task, I had to request my Personal Twitter API Key and tokens. The original format of this data is JSON, some transformations were applied to save only the variables of interest in a dataframe that I called “df_tweets”

```
#Personal Twitter API Key and tokens (Information removed for Project
submission)
consumer_key = 'MY_CONSUMER_KEY'
consumer_secret = 'MY_CONSUMER_SECRET'
access_token = 'MY_ACCESS_TOKEN'
access_secret = 'MY_ACCESS_SECRET'

auth = tweepy.OAuthHandler(consumer_key, consumer_secret)
auth.set_access_token(access_token, access_secret)

api = tweepy.API(auth, wait_on_rate_limit= True,
wait_on_rate_limit_notify= True)

#Retweet and favorite count for tweet_ids
with open ('tweet_json.txt', 'a', encoding = 'utf-8') as file2:
    for id in df['tweet_id']:
        try:
            tweet = api.get_status(id, tweet_mode = 'extended')
            json.dump(tweet._json, file2)
            file2.write('\n')
        except:
            continue

#Each tweet's JSON data should be written to its own line
tweets_list = []
with open('tweet_json.txt', 'r') as file3:
    for line in file3:
        try:
            tweet = json.loads(line)
            tweets_list.append(tweet)
        except:
            continue

#Creating Dataframe with tweet's information
df_tweets = pd.DataFrame(tweets_list, columns = ['id', 'retweet_count',
'favorite_count'])
```

2. **Data Assessment**

   Through Assessment of the Data was done in two different ways:

   - Visually: By quickly visually inspecting the datasets I was able to identify some quality and tidiness issues, such as:
     - “Source” field from archived dataset was difficult to read, it had unnecessary html tags
     - Dog stages where listed separately instead of having one column called “Stage”

   - **Programmatically**: Some of the Pandas functions used to assess the data were:
     - Info: To identify the number of records, number of columns, type of variables, and missing values
     - Head: To see the first few rows of data
     - Tail: To see the last few rows of data
     - Value_counts: To see total counts per variable of interest
     - Describe: to get some statistics for the variables of interest

   Some of the quality and tidiness issues identified were:

df dataframe (Quality Issues)

- name field has entries that are not real names such as "a", "actually", "the"
- There are tweets not related to dogs, but instead related to other animals and other things.
- Several records showing "None" for name field
- rating_numerator has values with less than "10" or too high (with three or more digits)
- rating_denominator has values different than "10"
- Data contains retweets which means there are duplicates
- Source field difficult to read/ understand

df_images dataframe (Tidiness Issue)

- P1, P2, and P3 which are meant to be dog's breed are in different columns.

3. Data Cleaning

During this stage, I acted on the assessments done during the previous step to improve the quality and tidiness of the data.

Data Cleaning was done in a programmatic way following these steps:

1. Define: I would translate the identified issue into specific tasks
2. Code: I would perform the necessary transformations to achieve results
3. Test: I would verify the results are how I expected.

Additional to the Data Wrangling Steps performed through this project, I also performed the following steps to complete the project.

Data Storing: Final/ Clean file was saved as: **twitter**_archive_master.csv

Data Analysis and Visualizations: Various analysis and visualizations were done on different variables of interest such as identifying the top 10 Breeds, most common Twitter source, and some other ones.