Udacity: Data Analyst Nanodegree

Project: Wrangle and Analyse Data

By: Laszlo Rado 03-08-2020

Wrangle Report

Context of the project:

Wrangle WeRateDogs Twitter data to create interesting and trustworthy analyses and visualizations.

Data gathered from the following sources:

1) The WeRateDogs Twitter archive

Twitter archive, downloaded and sent by the owner of the WeRateDogs account.

Data manually downloaded from the following url:

https://d17h27t6h515a5.cloudfront.net/topher/2017/August/59a4e958_twitter-archive-enhanced/twitter-archive-enhanced.csv

2) The tweet image predictions

Neural network prediction of dog breed based on twitter picture.

image_predictions.tsv downloaded programatically using the *requests* python library from the following url:

https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv

3) Additional twitter data

Using the tweet IDs in the WeRateDogs Twitter archive, I have 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 (can be found in the project workspace).

Then read this .txt file line by line into a pandas DataFrame with tweet ID, retweet count, and favorite count and follower count.

Data Assessment:

After gathering each pieces of data, I have assessed it both visually and programmatically to identify the following quality and tidiness issues:

Qulaity issues:

WeRateDogs Twitter Archive data:

1) Source column includes html tags and urls

2) Dog stage (doggo, floofer, pupper, puppo) columns are captured as string (Object)

3) Some records has been tagged with multiple dog stage category

4) Timestamp column captured as string (Object)

5) rating_numerator data type is integer, but there are decimal values in the original text.

Tweet image predictions data:

6) p1, p2, p3 columns include many non-dog-breed values

7) p1, p2, p3 columns: dog breed format is not consistent (upper/lower case, space/underscore)

8) Total probability (p1_conf + p2_conf + p3_conf) for record with tweet_id 667866724293877760 is greater than 1, also no dog breed was identified.

Additional Twitter Data:

9) Column name id_of_tweet is not consistent with the other tables (tweet_id)

Tidiness issues captured:

WeRateDogs Twitter Archive data:

1) Expanded urls column includes multiple urls

2) Unnecessary columns included in the dataset: all retweet and reply columns

3) Dog stage columns (doggo, floofer, pupper, puppo) columns can be merged into 1 column

Tweet image predictions data:

4) p1, p2, p3 columns can be merged into one

Additional Twitter Data:

no tidiness issues noticed

Overall: 5) The above three tables form 1 single observational unit

Data Cleaning:

Cleaning all the data quality and tidiness issues documented under the Data Assessment section.

Data Cleaning Steps:

1) extract multiple urls from the 'source' column of the The WeRateDogs Twitter archive dataset into individual records of a separate dataframe, than drop 'expanded_urls' column from original dataframe.
2) Remove the records related to re-tweets and replies, than drop columns related to re-tweets and replies from the The WeRateDogs Twitter archive dataset. ('in_reply_to_status_id', 'in_reply_to_user_id', 'retweeted_status_id', 'retweeted_status_user_id', 'retweeted_status_timestamp')
3) Manually pick dog stage for records with more than one dog stage category in the WeRateDogs Twitter archive dataset.
4) Combine separate dog stage columns into one single columns (doggo, floofer, pupper, puppo), than drop separate original columns from the The WeRateDogs Twitter archive dataset.
5) Change the new dog stage column's data type from String to category in the WeRateDogs Twitter archive dataset.
6) Drop record with tweet id 667866724293877760 from the Twitter image predictions dataset, since the overall confidence of dog breed predictions added up to more than 1. This is probably due to rounding error, but since there was also no dog breed predicted for the picture I decided to drop it.
7) p1, p2, p3 columns of the Twitter image predictions dataset merged into one single column (with a more descriptive column name than the originals: dog_breed_prediction) by keeping the column which was has the highest confidence (p#_conf) among the three columns and also classified as dog (p#_dog). Than drop separate original columns (p1, p2, p3, p1_conf, p2_conf, p3_conf, p1_dog, p1_dog, p3_dog, total_conf, img_num)
8) Change predicted dog breed names first character to uppercase and remaining characters to lowercase. Also replace underscores and dashes with space to keep column values consistent.
9) Use regular expressions with pandas' str.extract method to extract and remove html tags and unnecessary urls from the source column of the the WeRateDogs Twitter archive dataset.
10) Change data type of the 'timestamp' column of the the WeRateDogs Twitter archive dataset from Object(String) to datetime64[ns]
11) Drop records where rating denominator in the WeRateDogs Twitter archive dataset is not 10.
12) Drop records with extreme high rating_numerator values in the tw_arch dataframe
13) Change id_of_tweet column in the Additional twitter dataset to tweet_id for consistency with the rest of the tables
14) The above three datasets (WeRateDogs Twitter Archive, the Additional Twitter data and the Image Predictions) form one observational unit, and can be merged together into one single dataset.

After the above gathering, assessing and cleaning activities have been performed on the datasets, the analysis of the data can be started.

De details and results of the analysis have been recorded in a separate pdf file called act_report.pdf (can be found in the project workspace).