Wrangle_report

May 17, 2020

1 Project details:

1.1 Gathering Data:

Gathering the data will be accomplished in three steps:

1. Downloaded the file on hand **twitter_archive_enhanced.csv** manually.
2. This file (image_predictions.tsv) is hosted on Udacity's servers and is to be downloaded programmatically using the Requests library and the following URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv
3. Each tweet's retweet count and favorite ("like") count at minimum, and any additional data you find interesting. This data is available on Tweepy. However due to exigencies, I couldn't get access to the Twitter's API, so downloaded the file called **tweet_json.txt file** manually.Read this .txt file line by line into a pandas DataFrame with (at minimum) tweet ID, retweet count, and favorite count. ## Assessing the data : > Assessing the data to find quality (dirty) and tidiness (messy) issues. Used the following functions for programatic and manual assessments for all the three dataframes:
4. df.info()
5. df.head()
6. df.tail()
7. df.sample(5)
8. df.column.value_counts()
9. df.describe()
10. sum(archive.tweet_id.duplicated())

Issues identified :

archive dataframe:

Completeness : 1. *retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp* have only 181 values. 2. *in_reply_to_status_id* and *in_reply_to_user_id* have only 78 non-null values. 3. *expanded_urls* are missing, there are 2297 values as against 2356.

Validity: 4. In the *names* column : There are some dogs whose names are : 'a','an' and 'None'. 5. *rating_numerator* has values which go as high as 1776.

Accuracy: 6. *timestamp* should not be an string datatype. It should be 'timestamp' datatype.

Consistency: 7. *rating_denominator* has 23 values which are other than 10 (some are as high as 170 and some as low as 0, having 0 as the denominator is a potential problem). This makes comparison not consistent. 8. HTML tags in the *source* column make the data look dirty. 9. As per the requirements of the project, delete retweets.

Tidiness: A tidy dataframe should be one where : **each variable is listed as a column, each entry is a row and each observational unit forms as table.** 10. There are four columns for the different dog types: *doggo, floofer, pupper, puppo*
image predictions dataframe: Completeness: 11. There are 2075 entries in this dataframe as against 2356 in archive dataframe.

Validity : 12. As per the description of this data frame, p1 contains the breed names. "pole, nail, sandbar, swab, suit, traffic light, hotdog, cowboy_boot, teapot, cup, china_cabinet, website, seat_belt, American black bear" are not dog breeds.

Accuracy :

Consistency: 13. The starting alphabet of the dog breeds in p1, p2 and p3 are sometimes in uppercase and other times in lower case.

Tidiness : 14. This dataframe does not form a separate observational unit. It is a part of the **archive dataframe**.

twitter_data_web dataframe:

Completeness 15. There are 2356 entries in the archive dataframe while this one has 2354 entries.

Accuracy : 16. *tweet_id* is an 'int' datatype.

Tidiness : 17. This dataframe does not form separate unit of measurement. It is a part of the **archive dataframe**.

1.2 Cleaning the Data :

Created copies of three data frames. Cleaned all the above identified issues under 10 heads. 1. Drop retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp which have only 181 values. 2. Drop in_reply_to_status_id and in_reply_to_user_id which have only 78 values using *drop* function of pandas. 3. Drop the rows which have null "expanded_urls" (as these can't be obtained). 4. Melt the four columns for the different dog types: *doggo, floofer, pupper, puppo* into one. This will be done in two stages. A > melt the above four columns of the dataframe by creating an "intermediate" column. B > Drop the "intermediate" column and the duplicate values by keeping the first ones. 5. Merge the three columns into a single dataframe. This will be done in two stages: A >left merge of archive_copy and tweet_data_web_copy dataframes. B > inner merge of archive_copy and image_predictions dataframes. 6. In the *names* column : There are some dogs whose names are : 'a','an' ,'by' etc. All names starting with smaller case must be

checked and invalid names isolated. Replace these invalid names with "None". 7. Convert rating_numerator and rating_denominator into float data type. Also, correct the decimal entries in the numerator column. 8. As per the description of this data frame, p1 contains the breed names. "pole, nail, sandbar, swab, suit, traffic light, hotdog, cowboy_boot, teapot, cup, china_cabinet, website, seat_belt, American black bear" are not dog breeds. **Replace invalid names with "None", Rename giant_panda as giant_panda dog, Rename bison as Bichon Frise, Rename hare as Hare Indian Dog, Rename malamute as Alaskan Malamute, Rename airedale as Airedale Terrier, Rename teddy as Zuchon, Rename chow as Chow-Chow,Rename Lhasa as Lhasa Apso, Rename redbone as Redbone Coonhound, Rename basset as Basset Hound, Rename Leonberg as Leonberger, Rename Pekinese as Pekingese, Rename cairn as Cairn Terrier, Rename clumber as Clumber Spaniel and finally Capitalize the first alphabet of each word in the column.** 9. tweet_id is int datatype and *timestamp* should not be an string datatype. It should be 'timestamp' datatype. 10. HTML tags in the *source* column make the data look dirty. Extract the strings like "Twitter for iPhone", "Twitter Web Client" etc. from the rows using str.extract.

1.3 Analysing the data :

Three basic analysis performed to find visualize : 1. Distribution of ratings using plt.hist 2. Top dog breeds given by the machine learning algorithm (as mentioned in the image_predictions dataframe, which was merged with the archive dataframe.) 3. Relationship between retweets and favorite tweets using plt.scatter. 4. Finally found out the top 5 favorite tweets of dogs with images.