wrangle_report

October 15, 2023

Data Wrangling : WeRateDogs (@DogRates)

0.1 Introduction

The main purpose of this project was to use real world data to wrangle (gather, assess, clean) and then apply analysis with visualizations. The data used was from the Twitter account 'WeRateDogs' (@dog_rates) which "rates people's dogs with a humorous comments about the dog. These ratings almost always have a denominator of 10. The numerators, though? Almost always greater than 10. 11/10, 12/10, 13/10, etc."

Steps Involved in data Wrangling: 1. Gathering the data 2. Assesing the data 3. Cleaning the data 4. Storing the final cleaned dataset

I started off with gathering the data from various sources, then used jupyter notebook to access that data and then made some observations visually and programatically on the areas that needs cleaning and tidiness in order to make data more usable. After I made my assessments, next step was to use python to clean the data for each of the issues identified in the assesing phase. Once the data was cleaned, I joined different datasets together to develop a master dataset that was further used to get some insights from the data using data visualization.

I will briefly discuss about each of the phases in this document below.

0.1.1 1. Gathering data

I had to gather data from three different sources as below:

1. Read the csv file which contains the WeRateDogs Twitter archive data (twitter_archive_enhanced.csv)
2. Use the Requests library to download the tweet image prediction into a file called image_predictions.tsv.
3. Use the Tweepy library to query additional data via the Twitter API (tweet_json.txt), since tweepy is now paid. Udacity provided us with the json file with the data.

All the above were stored in the form of pandas dataframes to perform further assessments and cleaning.

0.1.2 2. Assessing data

After doing the visual and programmatic assessments, I was able to identify the below quality and tidiness issues in the dataframes.

Quality Issues

1. tweets_data dataframe --- There are retweets and replied tweets also present in the dataset, since we only need original tweets, we need to remove retweets and replied tweets.
2. tweets_data dataframe --- Incorrect format of timestamp, it is string but should be timestamp
3. tweets_data dataframe --- Column 'source' is in anchor tag and is unreadable, extract the source out as text.
4. img_predictions dataframe --- Keep records which only have dog names in either p1, p2 or p3 i.e. model confirm that the image is of a dog
5. tweets_data dataframe --- rating_denominator should always be 10 but there are other values in the tweets_data_clean dataframe, Identify the tweets which have values other than 10 and fix the denominator
6. tweets_data dataframe --- Some tweets have more than 1 dog stage ex: pupper, duggo both as dog stage. Fix it to right dog stage.
7. tweets_data dataframe --- name column have unrelevant values None, a, the, an
8. img_predictions dataframe --- p1,p2,p3 columns in img_predictions dataframe has inconsistent capitalization and also contain underscores in dog names

Tidiness Issues

1. tweets_data dataframe --- Remove extra columns which are no longer needed : in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id , retweeted_status_user_id
2. tweets_data dataframe --- doggo, floofer, pupper, puppo should not be 4 columns but value of a single column, something like 'dog_stage'
3. No need for 3 separate datasets. All datasets can be merged into a single dataset.

0.1.3 3. Cleaning data

After I identified the quality and tidiness issues in the data, it was time to fix them using pandas and build clean datasets. I will briefly describe in one or two lines about the approach I opted to fix each of the issues identified in the assessing phase.

0.1.4 Quality Issues

Issue 1 : Definition : tweets_data dataframe --- There are retweets and replied tweets also present in the dataset, since we only need original tweets, we needed to remove retweets and replied tweets.

Solution : Dropped the records where there is a non-null value in columns - retweeted_status_id , in_reply_to_status_id

Issue 2 : Definition : tweets_data dataframe --- Incorrect format of timestamp, it is string but should be timestamp

Solution : Removeed trailing " +0000" i.e. last 6 characters and converted to datetime format

Issue 3 : Definition : tweets_data dataframe --- Column 'source' is in anchor tag and is unreadable, extract the source out as text.

Solution : Used split function and pull out the actual text from anchor tag

Issue 4 : Definition : img_predictions dataframe --- Keep records which only have dog names in either p1, p2 or p3 i.e. model confirm that the image is of a dog.

Solution : Created a list of such tweets which have p1_dog,p2_dog and p3_dog as FALSE and removed these tweets from all three dataframes

Issue 5 : Definition : rating_denominator should always be 10 but there are other values in the tweets_data_clean dataframe, Identify the tweets which have values other than 10 and fix the denominator

Solution : Created a list with contains the wrong denomiantors (other than 10) and then fetched the tweets with those denominators. Since there were handful of tweets, Fixed the ratings individually upon reading the tweets.

Issue 6 : Definition : tweets_data dataframe --- Some tweets have more than 1 dog stage ex: pupper, duggo both as dog stage. Fix it to right dog stage.

Solution : Created a new column called 'combined_stages' but concatenating all four fields and then replace the string 'None' with blank value, so that we can identify the tweets with singlular dog type and multiple dog types easily. Replacing '' with 'None' back again for the records which had 'NoneNoneNoneNone' Values Upon identifying multiple dog type records, fixed it accordingly.

Issue 7 : Definition : tweets_data dataframe --- 'name' column have irrelevant values None, a, the, an

Solution : Replaced values with NaN where there is irrelevant names - None, a, the, an

Issue 8 : Definition : p1,p2,p3 columns in img_predictions_dataframe has inconsistent capitalization and also contain underscores in dog names

Solution : Capitalized the dog names and replaced underscores with a single space.

0.1.5 Tidiness Issues

Issue 1 : Definition : tweets_data dataframe --- Remove extra columns which are no longer needed : in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id , retweeted_status_user_id, retweeted_status_timestamp

Solution : Used drop method to drop the required columns

Issue 2 : Definition : tweets_data dataframe --- doggo, floofer, pupper, puppo should not be 4 columns but value of a single column, something like 'dog_stage'

Solution : Drop 4 columns and rename combined_stages column to dog_stage

0.1.6 Storing cleaned dataset

Finaly I joined the above 3 datasets into a single dataset I saved the master joined dataset as twitter_archive_master.csv