WE RATE DOGS DATA ANALYSIS PROJECT

Reporting: wrangle_report

This project was about data wrangling on We rate Dogs data.

Below were the objectives of the project:

1. Gather data
2. Assess data
3. Clean data
4. Store data
5. Analyze and visualize data
6. Reporting

Gather data

Three pieces of data were gathered. One piece, that twitter_archive_enhanced.csv was downloaded manually from: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/59a4e958_twitter-archive-enhanced/twitter-archive-enhanced.csv and uploaded and then read into a pandas dataframe, using Jupyter notebooks.

The second piece known as image_predictions.tsv was programmatically downloaded using the Requests library from: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv and then loaded into a pandas DataFrame.

The third piece, which is tweet_json.txt was downloaded from Udacity classroom as one of the support materials. Then it was uploaded and read line by line into a pandas dataframe with tweet ID, retweet count, and favorite count.

Assess data

After gathering all the three pieces of data, I assessed them both visually and programmatically to detect any quality and tidiness issues.

During the visual assessment, I displayed each piece of data in the Jupyter Notebook and made some scrolling as I looked at the values and in the rows and the columns.

During programmatic assessment, I used pandas functions /methods such as info, value_counts and describe to assess the data.

I found out 16 uses and documented them as shown below:

Quality issues

df_archive dataset:

1. The columns: doggo, floofer, pupper, and puppo contain missing values.
2. The link for the tweets and ratings at the end of the 'text' column are not important. They should be removed.
3. The 'source' column is an HTML anchor tag. We should extract the tweet's source and convert it to a categorical value.
4. The datatype of the 'timestamp' column is 'str' instead of 'datetime'
5. The retweet columns are not necessary. They should be removed.
6. The reply columns are not necessary. They should be removed.
7. The 'expanded_urls' column has NaN values
8. The 'rating_numerator' column should be of a float datatype.
9. The 'rating_denominator' column has values less than 10 and values greater than 10.
10. The 'name' column has 'None' instead of 'NaN' and it has invalid values.

df_API dataset:

11. The 'id' column name in the df_API data set should be 'tweet_id' to match the other names in the other 2 datasets.
12. The retweeted rows should be removed

Tidiness issues

df_archive dataset:

13. columns: doggo, floofer, pupper, and puppo are separated yet them are about the same thing, dog type. One variable should be used to capture them.

df_image_predictions dataset:

14. The columns: p1, p1_conf, p1_dog, p2, p2_conf, p2_dog, p3, p3_conf, and p3_dog are are all about 2 variables, breed and confidence.

df_API dataset:

15. Only 3 columns: id, retweet_count, and favorite_count are needed. The rest should be removed.
16. **General issue:** All the 3 datasets form one observation. They should be marged.

Clean data

All issues documented during assessment were cleaned in this section.

Before cleaning, I first made a copy of the original data.

I used the Define-Code-Test Framework and documented each issue using a few sentences.

The table below summarizes what how each issue was cleaned:

Dataset	Issue	Definition
df_archive	The columns: doggo, floofer, pupper, and puppo contain missing values.	Replace 'None' with np.nan for Columns (doggo, floofer, pupper, puppo).
	The link for the tweets and ratings at the end of the 'text' column are not important. They should be removed.	Remove ratings and links from text column using Regular Expressions
	The 'source' column is an HTML anchor tag. We should extract the tweet's source and convert it to a categorical value.	Extract tweet source from source column using apply meth in pandas and convert it to categorical.
	The datatype of the 'timestamp' column is 'str' instead of 'datetime'	Convert timestamp column to datetime.
	The retweet columns are not necessary. They should be removed. The reply columns are not necessary. They should be removed.	Remove the retweet columns and replies columns.
	The 'expanded_urls' column has NaN values	Drop rows with NaNs for expanded_urls column.
	The 'rating_numerator' column should be of a float datatype.	Convert rating_numerator datatype to float.
	The 'rating_denominator' column has values less than 10 and values greater than 10.	Remove values other than 10 for rating_denominator
	The 'name' column has 'None' instead of 'NaN' and it has invalid values.	1. Replace 'None' with np.name in df_arch name column. 2. Remove any rows with invalid names which starts with lower laters.
df_API	Only 3 columns: id, retweet_count, and favorite_count are needed. The rest should be removed.	Remove unnecessary columns for df_pi_clean dataset
	The 'id' column name in the df_API data set should be 'tweet_id' to match the other names in the other 2 datasets.	Rename id column in df_API_clean to tweet_id
	The retweeted rows should be removed.	Remove retweeted rows.
df_archive	columns: doggo, floofer, pupper, and puppo are separated yet them are about the same thing, dog type. One variable should be used to capture them.	Create the 'dog_stage' column and remove the doggo, floofer, pupper, and puppo columns.
df_image_prediction	The columns: p1, p1_conf, p1_dog, p2, p2_conf, p2_dog, p3, p3_conf, and p3_dog are are all about 2 variables, breed and confidence.	Create breed and confidence columns with highest confidence predictions and drop other columns

All	All the 3 datasets form one observation. They should be marged.	merge all the 3 datasets into one and store the data into csv file and sqlite database

Store data

A tidy master dataset with all the three pieces of gathered and cleaned data was created and stored as both a csv file and as sqllite database named twitter_archive_master.csv and twitter_archive_master.db respectively.

Analyze and visualize data

Using Jupyter Notebook with Pandas, matplotlib, seaborn and pyplot, I came up with 6 visualisations in terms of bar plots and regular plots from which insights were drawn.

The three insights recorded are:

1. There exists a highly positive correlation between Favoite counts and retweet counts
2. Over 90% of the Tweets are fom Twitter for iPhone
3. The Most common dog stage is pupper, followed by doggo and the third is puppo

Reporting

Finally, I created two reports required that is, an ‘act_report’ and a ‘wrangle_report’ using Microsoft Word, and uploaded the documents for submission.