WeRateDogs Twitter archive Data

Data Wrangling

Data wrangling is the process of gathering your data, assessing its quality and structure, and cleaning it before you do things like analysis, visualization, or build predictive models using machine learning.

There are several processes that were involved for this to be complete and I'll discuss them below:-

1. **Data Gathering**

This involved collecting data from 3 different sources:

- twitter_archive_enhanced.csv. – Downloaded manually with the link provided.
- image_predictions.tsv - Downloaded programmatically using the Requests library
- tweet_json.txt – This was to be downloaded using API but I had challenges in creation of twitter developer account hence ended up downloading manually from link provided.

2. **Assessing Data**

Eight (8) quality issues and two (2) tidiness issue. Will use **both** visual assessment programmatic assessement to assess the data.

Quality issues

1. Erroneous datatypes in columns (timestamp, retweeted_status_timestamp)
2. Null values recorded as None and NaN - Harmonize to NaN
3. Rename column timestamp to tweet_date
4. Rename column text to tweet_remarks
5. Rating denominator should not be 0 (zero).
6. Extract source from contents in source column(HTML CODE) and rename column to source_extract
7. Create a new column 'Dog Stage' and merge the dog stages into it.
8. Create new column breed_name after getting one with highest confidence, then drop other columns.
9. Clean Breed Names created in Issue 8. Remove '_' and all names to be in lower case

Tidiness issues

1. Merging all the three tables and forming one table for easier clean up since tweet_id is common among all the three tables.
2. Drop unnecessary columns in merged table. Columns to be dropped include p1, p1_dog, p1_conf, p2 ,p2_conf, p2_dog, p3, p3_dog, p3_conf, doggo, floofer, pupper, puppo.

3. Cleaning Data

All the issues that were identified during the assessment stage were cleaned and documented appropriately.

Resources

1. DataFrame Merge: https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.merge.html
2. Graph Correlation: https://www.khanacademy.org/math/statistics-probability/describing-relationships-quantitative-data/introduction-to-scatterplots/a/scatterplots-and-correlation-review
3. Multiple Column Merge: https://stackoverflow.com/questions/33098383/merge-multiple-column-values-into-one-column-in-python-pandas
4. Requests: https://pypi.org/project/requests/
5. Select specific Rows: http://net-informations.com/ds/pd/rows.htm
6. Udacity Classroom: ALX-T Data Analyst