May 25, 2019

Wrangle Report (WeRateDogs)

Project Overview

In this project we wrangled ,analyzed and visualized a tweet archive of Twitter user @dog_rates, also known as WeRateDogs dataset.

Project Details

1. Data wrangling, which consists of three steps:
   - Gathering data.
   - Assessing data.
   - Cleaning data.
2. Storing, analyzing, and visualizing the wrangled data.

Data wrangling

Gathering data

Datasets were gathered from three sources as follow:

1. The WeRateDogs Twitter archive csv file which was downloaded manually from Udacity (twitter_archive_enhanced.csv).

2. The tweet image predictions, which contains the dogs breeds (or other object, animal, etc.) based on a neural network image prediction algorithm. The dataset was hosted by Udacity's servers and downloaded programmatically using the Requests library (image_predictions.tsv).

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

3. JSON file contains the number of retweets and likes for each tweet. This file was downloaded using Twitter API (tweet_json.txt).

Assessing data

Assess data visually and programmatically for quality and tidiness issues. Quality issues associated with completeness, validity, accuracy and consistency while tidiness issues related to structural problems.

For all the three datasets we inspect them visually by creating dataframe for each datasets and read them to get the feeling and understand the data. And programmatically by using multiple pandas functions.

Visual assessment used methods:

- head(), tail(), sample()

Programmatic assessment used methods:

- value_counts(), describe(), info(), shape, duplicated(), unique(), query()

Then discovering the quality and tidiness issues.

Quality Issues:

* Twitter Archive dataset:

- 23 records include a rating denominator which doesn't equal to 10. (Validity)
- Rating Numerator in 6 cases was extracted incorrectly from the tweets text. (Validity)
- 181 records are not original tweets -retweeted- (Consistency)
- Some columns need to be renamed. (Consistency)
- timestamp & retweeted_status_timestamp not on datetime datatype. (Consistency)
- Some dogs names seems wrong. (Accuracy)
- Rating numerator and denominator should be in float datatype due to the fact that some rows have decimal values. (Accuracy)

* Image Prediction dataset:

- Number of tweet IDs in image prediction table doesn't match the number of records in twitter archive (Completeness)
- Some columns need to be renamed. (Consistency)
- p1, p2, & p3 information have a lot of underscores and dashes instead of spaces (Consistency)
- p1, p2, & p3 information sometimes start with capital letter and sometimes small latter. (Consistency)

*JSON Tweets

- Number of Records in all the three data sets doesn't match. (Completeness)
- id column need to be renamed. (Consistency)

Tidiness Issues:

- Doggo, Floof, Pupper & Puppo should be represented in one column.
- Stage column should have all the dog stages mentioned in the row.
- All datasets should be merged.

Cleaning data

Fix quality and tidiness issues by using (define, code and test).

First we created copies for each of our three dataframes and start the cleaning using the copies.

Used cleaning methods:

- to_datetime(), .rename(), .isnull(), .isin(), replace(), str.capitalize(), str.match(), .merge(), .drop()

Storing, analyzing, and visualizing

Then we merged all datasets image prediction, JSON tweets with twitter archive datasets. Finally, store the cleaned tables to a master file, analyze and visualize. Please, refer to the analysis and visualization report in (act_report.pdf).