We Rate Dogs: Wrangle Report

Tim Quan

March 2022

0.1 Introduction

In this stage of the project, we gathered/wrangled data from 3 sources:

- **twitter_archive_enhanced.csv**
- **image_predict.tsv**
- Twitter API

0.2 Twitter Archive Enhanced (twitter_archive_enhanced.csv)

This is a file provided as a project resource. The file was manually downloaded from the provided repository and stored locally.

Through the course of the workbook, this was imported into a Pandas DataFrame and remained in memory.

0.2.1 Assessment

The resulting dataframe contained 2356 rows and 16 columns/variables. Most variables are metadata gleaned from tweets.

Quality Issues, Cleaning Tidying

1. As imported to dataframe, some columns were interpreted as inappropriate data types. This was resolved by setting columns to more appropriate data types as follows:

column	**adjusted dtype**
*tweet*_id	string
timestamp	datetime
source	category
retweeted_status_id	string
retweeted_status_user_id	string
retweeted_status_timestamp	datetime

2. retweeted_status_id's can represent self-retweets. This means the retweeted records could be rows of duplicate data. We can identify the duplicate data by comparing the retweet_status_id column to the tweet_id column. To resolve this, we dropped the rows that contain tweet IDs which are also in the retweet_status_ids list.

3. The rating scale appears to be inconsistent. There are outliers in this data that need to be removed for analysis/visualization. There is some division by 0 in the denominators.

4. There are tweets in **twitter_archive_enhanced.csv** that do not have corrosponding records in **image_.tsv** and vice versa. After merging the

tables, these rows were identifiable by entries tweets do not have a value for *jpg_url*. These rows were dropped.

5. The column *source* contains interesting data about devices but is not readable. We parsed the string data from the source column into something human readable and managable, then changed the datatype to category, and finally renamed the column to *source_device*.

6. Columns *doggo, floofer, pupper, puppo* are categorical values of the same variable. We have merged these values into one column of type category.

0.3 Image Prediction (image_predict.tsv)

This remote file location was provided in the project instructions. It (the file) was downloaded programatically and stored locally. This delimited dataset contained both metadata with additional data added by the course authors.

The data was ingested into a Pandas DataFrame for assement and use.

0.3.1 Assessment

The images from tweets in the dataset were run through a predictive image neural network; the results are the 3 most likely detected breed of dog (or, in some cases random object), and the p-values associated with the likelihood of correctness of the predictions.