Act Report on Twitter dataset Project

by Moses Ojonuba as a requirement for the Udacity data analyst Nanodegree

Introduction

Real-world data rarely comes clean. Using Python and its libraries, you will gather data from a variety of sources and in a variety of formats, assess its quality and tidiness, then clean it. This is called data wrangling.

The dataset used for this wrangling (and analyzing and visualizing) project is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment 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. Why? Because "they're good dogs Brent." WeRateDogs has over 4 million followers and has received international media coverage.

Data Gathering

I began my data wrangling by downloading the data that were to be used for the project.

The twitter_enhanced_archive data which was provided to Udacity by WeRateDog was downloaded manually.

The image prediction data which was also provided to Udacity the students was downloaded using the request library.

The tweets data was to be downloaded using the Tweepy ApI but my Twitter developer account was not approved. I had to download the data using the request library.

Assessing Data

The download data were all loaded into dataframe using pandas and were assessed for quality and tidiness through visual and programmatic method. A number of Quality issues and tidiness issue were spotted;

Quality issues

1.Rating_denominator column in the twitter_archive table had values that were less than 10 and greater than 10
2.Timestamp column in twitter_archive_ table was not in a date time format.
3.Some dog names in the twitter_archive table were in lowercase.
4.Missing values in 'name' and dog stages of twitter_archive table represented as 'None'
5.Tweets beyond August 1st, 2017 in tweets_archive table were not needed.
6.Only original tweets were required (retweets and replies not needed)
7.id_str column in tweets table table needed to be renamed tweet_id to allow for merging with other tables
8.tweet_id columns in twitter_archive and image_prediction table were of the wrong data type.
9.Some Columns in the twitter_archive, tweets, are not relevant. also columns with much missing values.

Tidiness issues

1. Data in three separate Tables instead of one
2. Dog stage in twitter Archive data in four separate columns

Cleaning Data

I proceeded to cleaning all of the quality and tidiness issues that were documented above while assessing the data.

The actions to address the issues that were observed were defined as follows;

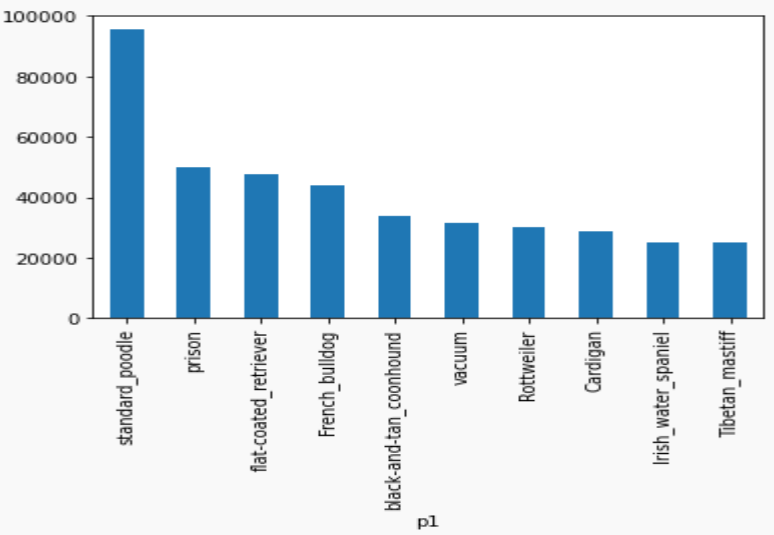
- Rating denominator that were below 10 or greater than 10 were to be removed in order to have a uniform rating denominator of 10.
- the timestamp column to be changed from object to date time.
- All dog names in upper case to be changed to title case.
- Replace None values with empty string and format them to nan values.
- Merge the four dog stages into one column.
- Extract tweets before August 1st, 2017.
- filter for only original retweets.
- Merge the three tables together to get one master dataframe.
- Extract only relevant columns and handle missing values.

A test was run after every cleaning procedure to ensure that we got the intended result.

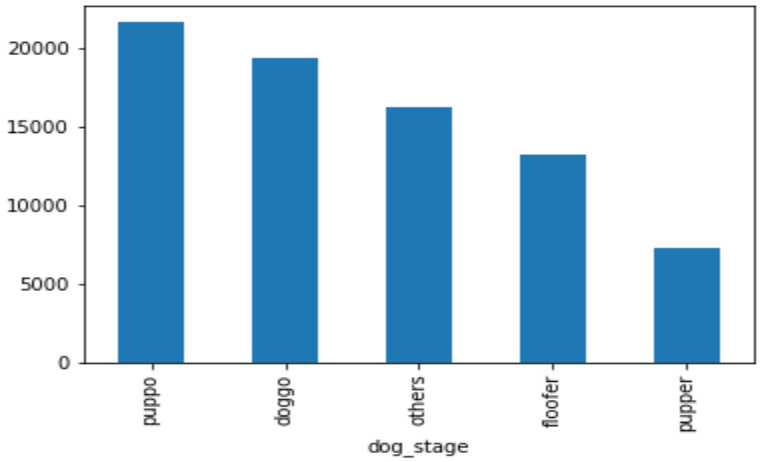
The final cleaned data was stored.

Insight and Visualization

1. Standard_poodle are the most liked breed.



2. Pupper is the most liked among the stages of Dog.



3. Pupper is the most common among the various stages of Dog.