wrangle_report

August 15, 2020

#

Udacity Data Analysis Nanodegree

##

Project: WeRateDogs Twitter Data

###

Noaman Mangera, July 2020

0.1 1. Gather

0.1.1 Twitter Archive

The Twitter archive of WeRateDogs contains data for 5000+ of tweets. The tweets themselves however, do not contain all the data required to complete the analysis. This is downloaded manually and stored locally. From there, the file is imported into the Python environent using Pandas and read into a dataframe called twitter_archive.

0.1.2 Image Predictions

This dataset contains predictions of dog breed from each tweet according to a neural network. The file is downloaded programmatically using python Requests library and saved locally. The file is then read into a dataframe called image_predictions.

0.1.3 Additional Data via the Twitter API

Retweet count and favorite count are two notable omissions from the above data. Fortunately, this additional data can be gathered from Twitter's API. Well, "anyone" who with a developers account. With tweet IDs used to find matching tweets this piece of data is extracted using the Pytyhon library tweepy and stored and then saved locally in JSON format. It is then read back into Python as a Pandas data frame called df_tweet_json.

0.2 2. Assess

A manual inspection was first made with the .shape method to determine the number of columns and rows. The .info method was then applied to each dataframe to ascertain an overview of the data types of each column, and the number of null values contained within each. This revealed the assignment of incorrect data types for a number of columns.

A visual assessment was also carried out using the .sample method so that a snapshot of reprensetative rows throughout the datasets could be viewed. This revealed columns and values unreadable to the human eye, and as such, required attention.

Delving deeper the .describe method calculated summary statistics for each of the numeric columns in the dataframe. Summary statistics such as the mean, standard deviation and quartile information brought much needed clarity to the shape and distribution of the quantitative variables.

A user defined function was then written to programmatically count the number of, and % of null values contained within each field within each dataset. No Null values were found in image_predictions or df_tweet_json. Null values related to retweet and reply information were found to be present in the twitter_archive data.

Duplicate Tweets were also identified using the .duplicated method.

The .value_counts methods highlighted several issues, including the usage of the a place holder labelled 'None' with the field name to indicate missing values.

The twitter_archive also stored information relating to the life stage in which a particular dog was at into 4 columns, when it could just as easily have been contained within a single column.

0.3 3. Clean

Before cleaning a copy of the data was made to ensure the original information remained for future reference.

The uneccesary 'doggo', 'floofer', 'pupper' and 'puppo' columns were merged into one column by adding the contents to a new column called dog_stage.

The three datasets were combined into one using the merge function into a 'master' dataset.

Retweets were removed using the .dropna method, while unused columns were dropped using the .drop method.

Values in the the source field was simplified to a more human readable format using the string replace function .str.replace.

Incorrect datatypes were reassigned to a more appropriate datatype using the .astype method, while innapropriatly labelled column names were renamed using the .rename method.

Incorrectly named dog names 'None' were replaced using the replace method.

The cleaned dataset was saved to the working directory as 'twitter_archive_master.csv'