Wrangle and Analyze Data

Introduction

WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. WeRateDogs has over 4 million followers and has received international media coverage. In this project, data wrangling and analyzing will be performed on three datasets, which includes the tweet archive of Twitter user @dog_rates (WeRateDogs), image predictions generated from neural network to classify the breed of dogs and additional WeRateDogs data captured via the Twitter API. Objective is to practise data gathering, accessing, cleaning and uncovering meaning findings in WeRateDog posts through data visualization.

Procedure

The first step to carry out data wrangling, which consist of:

1. Gathering data
   ○ **Enhanced Twitter Archive:** This csv. data file was given by Udacity and it contains basic tweet data, like tweet ID, timestamp and text etc.
   ○ **Tweet image predictions:** This is the results generated from neural network by analyzing every image in the WeRateDogs Twitter archive to classify breeds of dogs.
   ○ **Additional Data via the Twitter API:** The JSON file (txt.) was stored for each tweet through querying the Twitter API for each tweet's JSON data using Python's Tweepy library. Sublime Text was used to scan through the JSON file and read tweet ID, favorite count and retweet count of each tweet’s JSON data line by line into the data frame.

2. Assessing data

   Once data gathering is completed, the next step is to assess each data frame through both visually and programmatically for quality and tidiness issues. Visually checking is conducted by using Google Doc and specifically looking into details of text from each tweet; programmatically checking is accomplished by code such as info(), describe(), value_counts(), isnull() and duplicated() etc.. Through data accessing, 11 quality issues and 4 tidiness issues were discovered.

3. Cleaning data

With the targeted quality and tidiness issues, the most challenging part of the wrangling is data cleaning. In this data cleaning process, I would like to highlight some of the issues required most of the time and effort.

Twitter_archive

- Spotted some erroneous dog names extracted from text contain 'a', 'the' and ‘an’ etc, this is mainly due to the wrong extracting method (after “This is”), and those tweets mostly do not include any dog name. Therefore, replaced the list of wrong names with NaN instead to make it consistent.
- The erroneous numerators and denominator values are the most important part of data to clean as this is the feature of WeRateDogs tweets with numerators largered than denominators, like 13/10. But noticed that some of the values are extremely large, and some of the values are extracted from text incorrectly as there are 2 rating scores in the text or in decimal format. To address them, replaced a new column named ‘rating’ to represent the numerator over denominator values in float format, replaced those ratings with 2 scored from text with NaN and manually extracted & updated decimal values from text.

Image_prediction

- To update this data frame with the best dog breed prediction selected from 3 sets of breed, confidence level and prediction outcome, a loop is created to check row by row and find out the True prediction with the highest confidence level. With that, only the best dog breed prediction will be kept in the data frame.