wrangle_report

May 15, 2020

0.1 Wrangle and Analyze Data

0.1.1 Introduction

This project focused on wrangling data from the WeRateDogs Twitter account using Python, documented in a Jupyter Notebook (wrangle_act.ipynb). This Twitter account rates dogs with humorous commentary. The rating denominator is usually 10, however, the numerators are usually greater than 10. They're Good Dogs Brent wrangle WeRateDogs Twitter data to create interesting and trustworthy analyses and visualizations. WeRateDogs has over 4 million followers and has received international media coverage.

WeRateDogs downloaded their Twitter archive and sent it to Udacity via email exclusively for us to use in this project. This archive contains basic tweet data (tweet ID, timestamp, text, etc.) for all 5000+ of their tweets as they stood on August 1, 2017.

The goal of this project is to wrangle the WeRateDogs Twitter data to create interesting and trustworthy analyses and visualizations. The challenge lies in the fact that the Twitter archive is great, but it only contains very basic tweet information that comes in JSON format. I needed to gather, asses and clean the Twitter data for a worthy analysis and visualization. The Data

0.1.2 Enhanced Twitter Archive

The WeRateDogs Twitter archive contains basic tweet data for all 5000+ of their tweets, but not everything. One column the archive does contain though: each tweet's text, which I used to extract rating, dog name, and dog "stage" (i.e. doggo, floofer, pupper, and puppo) to make this Twitter archive "enhanced.".We manually downloaded this file manually by clicking the following link: twitter_archive_enhanced.csv

0.1.3 Image Predictions File

he tweet image predictions, i.e., what breed of dog (or other object, animal, etc.) is present in each tweet according to a neural network. This file (image_predictions.tsv) hosted on Udacity's servers and we downloaded it programmatically using python Requests library on the following (URL of the file: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv)

0.1.4 Twitter API

Back to the basic-ness of Twitter archives: retweet count and favorite count are two of the notable column omissions. Fortunately, this additional data can be gathered by anyone from Twitter's API.

Well, "anyone" who has access to data for the 3000 most recent tweets, at least. But we, because we have the WeRateDogs Twitter archive and specifically the tweet IDs within it, can gather this data for all 5000+. And guess what? We're going to query Twitter's API to gather this valuable data. Key Points

Before we start, herea are few points to keep in mind when data wrangling for this project:

```
1)We only want original ratings (no retweets) that have images. Though there are 5000+ tweets in

2)Fully assessing and cleaning the entire dataset requires exceptional effort so only a subset o

3)Cleaning includes merging individual pieces of data according to the rules of tidy data.

4)The fact that the rating numerators are greater than the denominators does not need to be clea
```

0.2 Assessing Data

0.2.1 Quality Issues

df:

```
1)Missing data in the following columns: in_reply_to_status_id, in_reply_to_user_id, 2)retweeted
3)This dataset includes retweets, which means there is duplicated data [Solved]
4)Timestamp and retweeted_status_timestamp is an object [Solved]
5)The source column still has the HTML tags [Solved]
6)Dogs name have 'None', or 'a', or 'an.' and some more lower case words as names [Solved]
7)Multiple dog stages occurs such as 'doggo puppo', 'doggo pupper', 'doggo floofer' [Solved]
```

image_df:

```
1)dog breeds are not consistently in p1,p2,p3 columns [Solved]
```

tweet_json_df:

```
1)Missing data [Solved]
2)tweet_id is an object [Solved]
```

0.2.2 Tidiness Issues

df:

```
1)The variable for the dog's stage (dogoo, floofer, pupper, puppo) is spread in different column
```

image_df:

```
1)This data set is part of the same observational unit as the data in the df [Solved]
```

tweet_json_df:

```
1)This data set is also part of the same observational unit as the data in the df [Solved]
```

0.3 Cleaning Data

0.3.1 Define

```
1)The pred_breed column is created based on the the confidence level of minimum 20% and 'p1_dog'
2)Based on dog types: doggo, floofer, pupper, puppo, 'doggo, puppo', 'doggo, pupper', 'doggo, fl
3)tweet_id is set as object type as it is not going to use for calculation.
4)A main dataframe is created using df_clean, image_pred_clean, and tweet_counts_clean dataframe
5)Dog Names Issue got rectified
6)Inconsistency in pred_breed got removed
7)All retweets get deleted to get unique tweets
8)The columns such as in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted
9)Timestamp format got corrected to datetime format
10)Extra HTML tags from source column get refracted
11)Dog ratings get standardized for denom of 10.
```

```
In [ ]:
```