wrangle_report

December 14, 2020

Wrangle Report
Data Analysis Professional Nanodegree Program

```
<h3><center>We Rate Dogs Data
```

This work is dedicated for **Wrangle and Analyza Data Project**. The second project in **Data Analysis Professional Nanodegree Program**. In this report I will illustrate the steps I took to wrangle and analyze data from twitter account WeRateDogs*(@dog_rates)*. The dataset 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.

The **wrangle** part include 3 main steps: 1. *Gathering data* 2. *Assessing data* 3. *Cleaning data*

1 Gathering

The files needed for this task came from different extension and sources:

1. **Direct Upload:** twitter-archive-enhanced.csv
2. **Programmatic Download:** image_predictions.tsv
3. **Twitter API:** tweet_json.txt - which is a JSON file

The first file **"twitter-archive-enhanced.csv"** was given to me from the Udacity team, I just downloaded the file. The second file **image_predictions.tsv** I downloaded it programmatically from the attached link, that was given to me too. The file was in tab separated values form. The third file **tweet_json.txt** gathered by querying the Twitter API for each tweet's JSON data using Python's Tweepy library and store each tweet's entire set of JSON data in a file called. After successfully gathering those files, I loaded each file in a separate data frame and give them a proper name.

2 Assessing

Assessing is the second step in wrangling process. I started to view data from different scopes to figure out the best way to deal with this data. I started to view sample of data to get familiar with columns and it's values, in addition to reading data statistically programmatically through stats functions like: .info(), .describe(), .value_counts, etc. I also examined the data visually through

excel. By the end of assessing, I stated some quality and tidiness problems to work on them through cleaning journey of data.

2.0.1 Quality Issues

- Filter original tweets only.
- Filter tweets that have image.
- Complete missing dog names.
- Validate None in name column.
- Correct some extracted dog names (a, O, Al, my, an, by, the, his, him).
- Check if all dog stages are extracted from text.
- Understand very high ratings.
- Manual fix Al in name column to Al Cabone.
- Denominator should only be equal to 10.
- Incorrect datatypes in timestamp column.
- p1, p2, p3 inconsistent capitalization (sometimes the first letter is capital).
- Understand the reason for duplicated pics in image prediction data frame even the tweet_id is not duplicated: **for retweets**
- capture source only from source column

2.0.2 Tidiness

- Delete not needed columns
- (**retweets columns**: *in_reply_to_status_id, in_reply_to_user_i, retweeted_status_id, retweeted_status_user_id , retweeted_status_timestamp*). (expanded_urls*)
- friends_count columns has a constant value and can be dropped.
- 4 different columns (doggo, floofer, pupper, and puppo) on dog stages should be combined in only one column and be category type.
- tweet_id column can be used to join tables together.

3 Cleaning

- Cleaning is the last step of the wrangling process. In this step, I am going to clean gathered data in a clean data frame to be able to perform the analysis and satisfy the project motivation must be cleaned.
- I joined archived data frame with the Twitter API data frame, as they naturally belonging to each other, the previous step one after dropping retweets and replies rows from the data set. Then, I dropped the tweets rows with no image prediction data to finally get 1964 complete record.
- Names extraction process held in low quality, I wrote a code to extract the name again from the text. The problem was mainly with the letter a and the accompanying adjectives to the animal names.

- Rating extraction process failed to grab the right rating too this problem is due to that the extraction process was searching for a specific character which is "/" in between to numbers, and return it without even give any clue for the tweets with more than one "/".

4 Storing

I stored the tweets records in a master data frame and kept image prediction data in a separate data frame. The data frames were saved in a CSV format files

5 Insights

- Tweet with id: 744234799360020481 is the most retweeted and favorited tweet
- Golden Retriever is the most predicted dog
- There is a proportional relationship between retweets counts and favourite counts
- Even Pupper stage dogs are the most dominant in the count, but Puppo stage dogs genereated the most interaction.
- The extracted data starting from November-2015 to July-2017. The more interacion from Twitter users with this account depend on quality of tweets not the number of tweets. Followers number didn'd change so much over this period. While 2016 has the biggest total retweets and favorites but 2017 was higer in mean.
- iPhone is the most important plafrom for future updates