Wrangle Report

December 7, 2018

1 Wrangle Report

2 Introduction

This project is Data Wrangling - WeRateDogs. In this project consists of 3 sections: * Gathering Data * Assessing Data * Cleaning Data

2.1 Gathering Data

We use the following three pieces of data in a Jupyter Notebook titled wrangle_act.ipynb:

- The WeRateDogs Twitter archive in csv format: twitter_archive_enhanced.csv which is download manually.
- The tweet image predictions file: image_predictions.tsv which is hosted on Udacity's servers and had been downloaded programmatically using the Requests library and the following URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv
- Using the tweet IDs in the WeRateDogs Twitter archive, query 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 tweet_json.txt file.

2.2 Assessing Data

Here I going to observe the dataset for quality and tidiness issues.

Below here are the result:

2.2.1 Quality

- tweet_id has to be a string
- NaN exist in following columns: in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_statud_timestamp
- name column got 745 None string
- Some value in denominator column are not equal to 10
- timestamp should be in datetime
- Some names in the name columns are unsual like 'a', 'the'
- 'None' in doggo, floofer, pupper and puppo column treated as object, it should be treated as null value

- p1 ,p2, and p3 column should be category data type
- Remove retweeting rows

2.2.2 Tidiness

- doggo, floofer, pupper and puppo column represents should be in one column
- The archive, images dataframe, and the info dataframe should all be one dataframe

2.3 Cleaning Data

Cleaning the dataset with quality and tidiness issues listed in above.

2.3.1 Quality

- Convert tweet_id to string
- Drop columns: in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_statud_timestamp
- Convert name column which is None string into NaN
- Change the denominator equal to 10 by checking the text
- Convert timestamp to datetime
- Replace the unsual name like 'a', 'the' with correct name in the text
- Convert 'None' in doggo, floofer, pupper and puppo in column to null value
- Convert p1 ,p2, and p3 into category data type
- Remove retweeting rows

2.3.2 Tidiness

- Create only one column for doggo, floofer, pupper and puppo columns
- Merge the archive, images dataframe, and the info dataframe into one dataframe

2.4 Storing Data & Reports

- cleaned data is stored in twitter_archive_master.csv

2.5 References

- Regex: https://docs.python.org/3/howto/regex.html
- Regex test: http://www.pyregex.com/
- Counter from collection: https://docs.python.org/2/library/collections.html