Report

Wrangle and Analyze Data

Tasks in this project are as follows:

Data wrangling, which consists of:

- Gathering data
- Assessing data
- Cleaning data
- Reporting on your data wrangling efforts and

1. Gathering Data

- The WeRateDogs Twitter archive : twitter_archive_enhanced.csv
- The 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) is hosted on Udacity's servers and should be downloaded programmatically using the Requests library and the following URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv
- Each tweet's retweet count and favorite ("like") count at minimum, and any additional data you find interesting. 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. Each tweet's JSON data should be written to its own line. Then read this .txt file line by line into a pandas DataFrame with (at minimum) tweet ID, retweet count, and favorite count

2. Assessing Data

After gathering each of the above pieces of data, assess them visually and programmatically for quality and tidiness issues.

Assessed data on following quality parameters

- Completeness
- Validity
- Accuracy
- Consistency

Assessed data on following tidiness parameters

- Each Variable forms a column
- Each Observation forms a row
- Each type of observational unit forms a table

3. Cleaning

Cleaning is an iterative task and requires checking data again and again

- Made a copy of the original data
- Identify the cleaning parameters, define them- what needs to be cleaned and how
- Used the define-code-test framework and clearly document it
- Manually and programmatically clean them

- Test that the code ensured that the data is cleaned properly
- Created a tidy master pandas DataFrame

For the data given, we identified the following issues in our data

Twitter_archive

- Remove columns that won't be used for analysis
- Remove retweets by keeping the rows with NaN retweeted_status_id
- Separate timestamp into day - month - year (3 columns)
- Create a single dog stage column and remove doggo, floofer, pupper and puppo columns
- Correct rating numerators with decimal values
- Correct erroneous rating denominator

Image prediction

- Delete columns that won't be used for analysis
- Drop 66 rows of duplicated jpg_url
- Create a seperate column for image prediction and confidence interval

Tweet json

- Keep original tweets only

Tidiness

- Checked the datatype of columns to be used for merging. They should be consistent
- Merging all tables
- Removing redundant columns