WRANGLE REPORT Udacity Project4 :

Wrangle and Analyze Data Project

By: Nikita Jain

1. INTRODUCTION

   This Wrangle and Analyze Data Project is part of Udacity's Data Analyst Nanodegree Term 2. The project involves wrangling of data from various sources associated with tweets from the Twitter user @dog_rates, one from the twitter archive file and one from the tweet image predictions file. After scraping together the data, quality and tidiness issues were assessed and then cleaned. Finally, the data was analysed and visualised using the python library and analysis are recorded in act_report.pdf.

2. DATA GATHERING

   a) The first data was gathered manually by downloading the twitter-archive-enhanced.csv file that Udacity provided to me.
   b) The second dataset was programmatically downloaded from Udacity's server using the requests library and saved as image_predictions.tsv (URL=https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_imagepredictions/image-predictions.tsv)
   c) The third dataset was gathered via the Twitter API by using the Tweepy library and stored as a JSON file and from that JSON file 3 columns were taken up and a new file was created with three columns
      1. Tweet_id
      2. Favourite_count
      3. Retweet_count

3. DATA ASSESSMENT

   Data was assessed both visually and programmatically using various built in functions such as

   a) .head()
   b) .info()
   c) .describe()
   d) .value_counts()
   e) .duplicated()

   The following issues were recorded in all of three datasets including both cleanliness and tidiness issues.

Quality Issues

Twitter1 Dataset

1. Tweet_id should be of object type but instead is integer type.
2. timestamp and retweeted_status_timestamp are currently of type 'object' but instead should be timestamp
3. name has values that are the string "None" instead of NaN and some also have weird names such as "the" , "an" etc.
4. doggo, floofer, pupper, and puppo have values that are the string "None" instead of NaN
5. Data contains retweets (ie. rows where retweeted_status_id and retweeted_status_user_id have a number instead of NaN)
6. Also, there are ratings with decimals such as 13.5/10, 9.5/10 have been incorrectly extracted as 5/10 and put in numerator and denominator.
7. There are many columns in this dataframe making it hard to read, and some will not be needed for analysis.

Twitter2 Dataset

1. There are missing tweets compared to the twitter1 dataframe (I am assuming they have been deleted)

Images Dataset

1. There are 2356 tweets in the twitter1 dataframe and 2075 rows in the images dataframe. This could mean that there is missing data, or that not all 2356 of the tweets had pictures.
2. tweet_id is an integer but should be object type.

Tidiness Issues

Twitter1 dataset

1. variable (dog stage) in 4 different columns (doggo, floofer, pupper, and puppo)

Twitter2 dataset

1. twitter2 data should be combined with the twitter1 data since they are information about the same tweet

Images Dataset

1. images data could be combined with the twitter1 data as well since it is all information about 1 tweet

2. the dog breed prediction and prediction confidence could be each packed into one column

4. CLEANING DATA

The issues found during the assessment process were cleaned and tested using the following methods and techniques:

- merge()
- reduce()
- .extract()
- .drop()
- .astype()
- .islower()
- .replace()
- .loc[]
- .value_counts()
- .info()
- Loops
- Regular expressions
- .apply()
- .map()

5. STORING DATA

After cleaning the data with all the quality and tidiness issues stated above, the final cleaned dataset was stored as a csv named- ('twitter_archive_master.csv')