WRANGLE REPORT

#WeRateDogs

Author : Varun Kumar Sharma
Date : 17.11.2018

Introduction

Real-world data rarely comes clean. Using Python and its libraries, we will gather data from a variety of sources and in a variety of formats, assess its quality and tidiness, then clean it. This is called data wrangling. We will document our wrangling efforts in a Jupyter Notebook, plus showcase them through analyses and visualizations using Python (and its libraries).

The goal is to gather data from 3 sources i.e csv, tsv and Json files -> Asses it's quality and tidiness -> Clean the data so that it communicates the insights -> to display the visualization(s) by using our wrangled data.

So I will divide wrangle act in 4 categories :

- Gather Data
- Asses Data
- Clean Data
- Analysis and Visualization

Gather Data

There are three sources from where we could gather data :

Twitter Archive Enhanced file contains the tweet data with 2356 rows with file name twitter_archive_enhanced.csv which imported into data frame twitter_archive

Image Prediction file which we can download from Udacity's server programmatically using python's request library. I saved it as image_predictions.tsv file. It contains 2076 rows and 12 columns.

Tweet Json text file which I can extract from Twitter API and stored it as tweet_json.txt. It contains 2342 rows with column such as tweet_id,favorites,retweets,user_followers,user_favourites and date_time.

Assess Data

In this step we analyse the data to find the quality and tidiness issue.
We can see that for several entries in twitter_archive name column are not actual names.
Columns"in_reply_to_status_id","in_reply_to_user_id","retweeted_status_id","retweeted_status_user_id" are with wrong data type. It is float and it should be integers/strings.
There are incorrect entries for rating_numerator and rating_denominator.
We need to consider only original ratings with images and no retweets.
We see that in some columns the null values are not treated as null values.
Image predictions table has only 2075 entries instead of 2356.(missing data).
Image predictions table contains retweets and some tweet_ids have the same jpg_url.
There are only 2342 entries instead of 2356 in tweet_data table.
Different columns of dog breeds can be condensed into one column. Also different dog stages can be melted into one column.
There are null values in several columns which are treated as non-null. Some entries can be seen with NaN as string value.

Clean Data

We will create copies of the three dataframes i.e twitter_archive,image_predictions and tweet_data and combine them in one dataframe df_master.We will store this file as twitter_archive_master.csv
We need to drop unnecessary columns. Remove the retweets and duplicate tweet id. Delete tweets with no pictures.Also we can remove the rows with incorrect entries for breeds.
We will convert the variables such as tweet_id, timestamp,source etc into correct data types.
We need to Fix the dog names column as I noticed while accessing the data that there are someinvalid dog names in 'names' column. Incorrect dog names usually starting from lower case letters as observed in provided data.
I will create new columns such as "breed", "confidence", "rating" and "dog_stages" melting the related columns for better data analysis and visualisations.

Analyze and Visualize

In this step I will use the cleaned master dataframe to convert the data in useful visualization which will lead us to our findings. I will analyze and correlation between various variables. I will explore this correlation in various plots which will be useful to find results.I will check the dog breed distribution and their ratings. I would check which are the most popular dogs and create some related visualizations.

Summary

To Summarise, during this process of data wrangling and analysis, I will use libraries such as pandas,numpy,requests, tweepy,json etc. I will create some visualisations to portray the findings in visual form. The code,data wrangling and analysis will be included in wrangle_act.ipynb. The steps considered during this project will be documented in wrangle_report.pdf. The documentation of analysis and insight into final results are included in act_report.pdf.