Data Wrangling Report

By: Ahmad Hashhash
September 2020

Introduction

This project is about doing data wrangling for WeRateDogs twitter account using Pandas and other libraries in python using Jupyter Notebook as in file [wrangle_act.ipynb].

The account is about rating people dogs through their photos with humorous sense, nevertheless the rating system is always more than 1 like 13/10, 15/10 **"because they're good dogs Brent"**

In this report I will briefly describe my wrangling efforts through the three major processes of data wrangling (Gathering data, assessing it then cleaning it) and also come up with some visual conclusions with charts.

1- Gathering Data.

I gathered data into 3 major datasets which are twitter_archive, tweet_json, image_predictions and each of them was gathered differently.

-twitter_archive: this file contains many information about each tweet as mentioned and it was downloaded manually from the project sources

-image_predictions: this file was hosted on Udacity servers which is a result of a neural network that determines what breed is the dog according to its photo and it was downloaded programmatically using the Requests python library

-tweet_json: this file contains the tweet id and every retweet count and favorite count for each tweet and it was download manually as zip file from the project resources but was extracted and opened programmatically through zipfile library in python and I couldn't manage to gather it though twitter API as it requires me to get a developer account which would have taken so long time

2- Assessing Data.

The assessing data process was done by two means:

-Assessing data visually which won't be so useful but makes you distinguish some kind of quality and tidiness issues by yourself without using programming but it isn't so efficient

-Assessing data programmatically using various Pandas functions such as info(), head(), describe(), value_counts() and many other functions

-Assessing data programmatically should help you finding almost any kind of data problems whether it was either Quality or Tidiness such as (completeness, Consistency, Accuracy and Tidiness) and some of the Quality problems are:

- Missing Values in many columns
- Removing denominator column after applying all values to 10
- Denominator and numerator should be in one column as dog_rate
- tweet_id datatype needs to be changed from int to str
- timestamp, retweeted_status_timestamp need to be changed to date/time datatype instead of str
- tweet_id in image_prediction datatype needs to be changed from int to str
- id column in tweet_json datatype need to be changed from int to str
- Source column have html tags
- Original tweet ratings only needed no retweets or replies to original tweet
- tweet_id datatype needs to be changed from int to str
- timestamp, retweeted_status_timestamp need to be changed to date/time datatype instead of str
- Some images aren't for dogs

And some Tidiness problems are:

- Text column has two variables which should be in separate columns
- Some columns need to be compressed in one column like (doggo, floofer ...etc.) to dog_stage
- Merging all data frames into master one

3- Cleaning Data.

Once data assessment process was done, it was time to start the coding work with python and pandas to clean the data and fix any kind of problems in it according to the assessment main points I noted focusing on having a good dataset that allows you to do some analysis and data visualizations on it

4- Storing Data.

All the datasets were merged into one file called master_dataset.csv from which I started doing data visualizations to make some conclusions about this data

5- Data Visualization.

Everything related to this part is in a separate file called act_report.pdf that communicates the insights and displays the visualizations produced from the wrangled data

-