Data Wrangling

udacity

Project 4

Student:

Muddassar Sohail

Course:

Data Analyst

Project 4: Wrangling & Analyzing "WeRateDogs" Data

This project aims at practicing data wrangling techniques to process data so as to make it ready for analysis and visualzations. data set being used os of a twitter initiative @dog_rates, aka WeRateDogs who rate dogs, usually as 11/10 :) This report brings into view what was done for wrangling the data.

Main steps of wrangling are collection, assessment and clean-up

Data Collection

- Twitter archive: The file was downloaded manually, provided in udacity course: twitter_archive_enhanced.csv
- Tweet image predictions, image_predictions.tsv was downloaded using the Requests library with URL given below: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv
- Twitter API & JSON: Each tweet's retweet count and favorite ("like") count, and anything else found interesting was included. Each tweet's entire json data was written in tweet_json.txt file.

Step 2: Data Assessment

Data assessment was performed using multiple ways, visual inspection as well as using code for assessment. Issues discovered included but not limited to following:

2.1: Quality Assessment

df_tw

- shouold keep actual ratings, only also having images
- redundant/useless columns to be droppedv
- data type issues of these columns: puppo, pupper, doggo & floofer
- denominators other than 10
- numerators having decimals
- date and time to be segregated in individual columns

df_img

- A column for image predictions & a column for confidence level to be made
- redundant columns to be deleted
- duplicate jpg urls to be deleted

df_jsn

- original tweets only

2.2: Tidiness Assessment

- tables to be combined as a data set
- tweet_id data type needs to change to int

Step 3: Data Cleaning

3 repetitive steps in this section were really helpful, define, code and test. I defined all the issues assessed and what was needed to done with them, then I copied original data frames to new data frames so that the original data remains intact. Multiple difficulties were faced during the process, I had to go forward and backward because sometimes appeared after analysis. Once I completed all the report and then at the end in graphs found some anomaly, some suspicious outliers. I had to go back again and resolve them. Data cleaning was really a tricky part.

Conclusion:

I learnt a lot from gathering to analysis and visualization.

New concepts like APIs for getting data legally, and different new libraries were useful. I used to do analysis using Excel, but now, I'm introduced to a new power, which is really powerful; Python.