Data Wrangling Report

Introduction

The objective of this project is to evaluate the students' understanding of data analysis based on what they learned from the Udacity Data Analysis Nanodegree program's data wrangling module. WeRateDogs, a Twitter user with the handle @dog rates, provided the dataset that was wrangled. WeRateDogs is a Twitter account that rates users' dogs with a humorous comment. These ratings almost invariably have a numerator larger than 10 and a denominator of 10.

This is a brief report my data wrangling steps.

i. **Gathering data**
ii. **Assessing data**
iii. **Cleaning data**

Data Gathering

Three dataframes were used in this project from 3 different sources. The main dataframe, df archive, was the messiest and could only be downloaded manually from udacity project workspace. df images, which is responsible for predicting dogs breed from images in tweets was available online through this url https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv and Requests library. The last one, df stats, collects data using the tweeter API and contains all the necessary tweets' that we need but we are only interested in retweet_count and favorite_count. I made use of the data that was provided by udacity because my twitter developer account has not been approved yet.

Data Assessment

Once I had all the three tables, I evaluated the data as follows:

Visual Assessment: One method involved printing each of the three dataframes individually in Jupyter Notebook, and the second involved examining the csv files in Excel sheet.

Programmatic Assessment: I made use of three pandas methods which are info, value_counts and duplicated for Programmatic Assessment

From there I classify the data into quality and tidiness issues

Data Cleaning

Data cleaning process is divided in three i.e Define, code and test. Define involves defining the problem, code involves solving the problem programmatically and test involves testing the code.

Wrangling effort

I created a copy of the three original data frame so that if there is an error, I can create another copy from the original data frame.

Dataframe df_archive contains many quality and tidiness issues that I discover during the data assessment stage. Quality issues include duplicate url's in expanded_urls column, remove of 78 replies from in_reply_to_status_id and in_reply_to_user_id columns, removing 181 replies from retweeted_status_id retweeted_status_user_id and retweeted_status_timestamp columns, there are unnecessary data in the name column such as "an", "None", "the" and "a", trailing +0000 from timestamp and wrong datatype, many urls in expanded_urls are not related to twitter while some do not belonging to 'WeRateDogs' etc. while tidiness issues include: source column need to be splitting into two source_name and the source_url since it is containing the source and the source url, melting of doggo, floofer, pupper, puppo into a single column since they are all stages of dogs and merging df_tweets_archive_clean, df_img_clean and df_tweet_json_clean tables into a single table(dataset).

I did not focus much on the other two dataframes because I feel like they don’t have much issues.

I was faced with many challenges during cleaning the quality and tidiness issues most especially the expanded_url and removing retweets and replies I was able to accomplish everything successfully.

Conclusion

This project is very challenging and in the process of doing it, I have learnt a lot more on how to apply the knowledge I learnt in the wrangling data module of udacity’s Data Analysis Nanodegree course. For some part that I have challenges, I did research and came up with a solution. Overall, the project have provided me with most of the necessary knowledge to tackle most data analysis projects.