Wrangling Data :

1- Gathering

Gather data from 3 different resources , first the twitter-archive-enhanced.csv , the 2nd file was image-predictions.tsv Which gathered programmatically through the udacity website url , the 3rd was tweet-json.txt which I gathered it from udacity website because I followed the steps provided to get access to twitter api but till now there is no progress in my application so I chose this option.

Stored all this gathered data in 3 dataframes named under (twitter_archive,img_pred,abi_twitter)

2-Assess

The assess step done visually and programmatically noted as Quality and Tidiness.the visually assessment done with excel while the programeticlly assessment done with pandas methods and functions . Detected14 quality issues and 3 tidiness issus.

Examples of issue pointed:

-lines 1068 , 1165 , 1202 , 1662 , 2335 : "rating_denominator" Typo Error
-lines 1068 , 1165 , 1202 , 1662 , 2335 : "rating_numerator" Typo Error
-'timestamp' is a str needs to changed to datetime
-tweet_id is a int format and there is no need to do mathematics operations on it
-'name' column have not valid names, count 109.

3-Clean

Selected some issues in quality to clean and all tidiness issues are cleaned. Every issue was defined , coded and tested right after.

Examples of issues cleaned:

- replace lines 1068 , 1165 , 1202 , 1662 , 2335 : "rating_denominator" Typo Error with the correct values from tweet text.
- replace lines 1068 , 1165 , 1202 , 1662 , 2335 : "rating_numerator" Typo Error with the correct values from tweet text.
- change 'tweet_id' column to str format in all data frames(no need to make operations on it)
- delete arch_clean df replies rows , we only analyze original tweets. count 78
- change extreme values in rating_denominator to value of 10 .
- delete arch_clean rows that have missing Urls
- change 'timestamp' column type from str to datetime
- change extreme values in rating_numerator closer to the mean which is 12
- drop retweets related columns for tidiness
- merge abi_clean columns retweet_count and favorite_count with arch_clean. and have only 2 DF (arch_clean,img_clean)

4-Storing

The finally table after cleaning was stored in file named (twitter_archive_master.csv) so the data frames are ready to analysis and visualize.

5-Analyzing, and Visualizing Data

After this used some pandas methods and pandas plot method with seaborn Lib to analysis the data and make a good visual analysis and analyze ratings, time stamp , favorite cound , and retweet count and find the correlation between them.

I hope that I met the required rubric for this project.
Thanks in advance

Regards,
Samy mohsen