Data Wrangling Report

Xuefei Yu

Gathering Data

Gather three pieces of data and read them as pd dataframe:

1. Downloaded the first piece file manually and used *pd.read_scv()* to read it as *twitter_archive* data frame
2. Downloaded the *image_predictions*.tsv file programmatically, and read it as *image_predictions* data frame
3. *Retweet_count* and *favorite_count* are downloaded programmatically by using twitter API*, and pd.read() it as tweets_df* data frame

Assessing Data

Quality

- image_predictions dataset has less observations than twitter_archive, this could be caused by no image tweeting
- source feature needs to be extracted(iphone, vine, webclient, tweetdeck)
- change timestamp to datetime datatype
- change datatype 'meme'to category; 'tweet_id', 'in_reply_to_status_id', and 'in_reply_to_user_id' to strings
- rating_numerator and rating_denominator have lots of unexpected values
- rating_denominator has one value of 0 that can't be used to calculate rating
- in name,doggo, floofer, pupper, puppo , 'None' value needs to be changed to np.nan
- replace '&amp' in text to '&'
- contains retweets that we don’t want it
- tweets_df has duplicate data

Tidiness

- delete 'retweeted_status_id' 'retweeted_status_user_id' and 'retweeted_status_timestamp' retweets columns
- the columns 'doggo', 'floofer', 'pupper', 'puppo' should be in one feature 'meme'
- 'rating_numerator' and 'denominator' should be in one feature rating rate
- concatenate all three datasets

Cleaning Data

Create 3 copies of original DataFrames

1. Drop duplicate data in *tweets_df*
2. Create one feature 'meme' using 'doggo', 'floofer', 'pupper', 'puppo' columns, and drop 'doggo', 'floofer', 'pupper', 'puppo' columns

3. Concatenate all three datasets
4. Delete rows with retweet, and delete 'retweeted_status_id' 'retweeted_status_user_id' and 'retweeted_status_timestamp' retweets columns
5. Delete rows where there are no images website
6. Extract key words from sources, and change the feature type to 'category'
7. Replace '&amp' with '&' in text feature
8. Change timestamp to datetime datatype
9. Change datatype 'meme'to category; 'tweet_id', 'in_reply_to_status_id', and 'in_reply_to_user_id' to strings
10. Change rating_numerator and rating_denominator incorrect values
11. Create new feature rating = rating_numerator/rating_denominator. Drop rating_numerator and rating_denominator. Drop oberservations with ratings > 2.

Storing Data

Store the clean DataFrame(s) in a CSV file with the main one named twitter_archive_master.csv using pd.to_scv().