Wrangle-Report

-Prepared by Pritesh Gujarati

1 Wrangling Report

Prepared by Pritesh Gujarati

1.1 Gathering:

For this step we are asked to gather from 3 different tables.
1.twitter-archive-enhanced.csv – which is downloaded manually from udacity site
2.Image-predictions.tsv – this file is downloaded programmatically using request method from given URL
3.tweet_json.txt – I have downloaded this file by calling twitter API for each tweet using the tweet id as parameter which we have in twitter-archive-enhanced.csv.
After all this collected file, we make sure we have in same working folder.

1.2 Assessing:

In this step we have founds some quality as well as tidiness issue from the data that we have collected in Gathering step.

1.2.1 Quality:

tweet_api_data table issues:

1. Source column contain tag
2. possibly_sensitive and possibly_sensitive_appealable columns having zero
3. we have unrelated columns like favorited, retweered.
4. we have column which have null: contributors, coordinates, geo, place, quoted status id

image_predictions table issues:

1. p1, p2, p3 whitespace usage is not normal some uses -, some uses _, some are capitalized, some are not.

twitter_archive table issues:

1. Dog names are not accurate.
2. Dog name have none string, need to update to null
3. Columns name doggo, floofer, pupper and puppo have none or column name
4. rating is wrong in some cases, we have spotted less than 10 values as well
5. timestamp is string

1.2.2 Tidiness:

1. created_at / timestamp, source, text, in_reply_to_status_id, in_reply_to_user_id are duplicated in tweet_api_data table and twitter_archive table
2. need to add tweet_api_data and image_predictions in twitter_archive table
3. p1_dog, p2_dog, p3_dog are not unique through the row in image_predictions

Wrangle-Report

-Prepared by Pritesh Gujarati

1.3 Cleaning:

In this step of wrangling data, we are cleaning our data from the issue we have seen is assess steps but we have project’s constraints so we will only clean few of them.

1. first thing we cleaned is retweets. Because of this one row can contain multiple data like retweeted status which can effect our final out come .To clean this data we capture all the retweeted_status data and store it in different data frame we will call it temp.then we drop the rows contain the retweeted_status in our tweet_api_data.after that we append the rows of temp variable into our tweet_api_data table.
2. The p1_dog,p2_dog,p3_dog contain redundancy since the p1, p2, p3 are not unique in our image_predictions table.To clean this we created the empty dataframe and mapped between p and p_dog , same for other in dataframe then remove the duplicates .after that we have remove the p1_dog,p2_dog and p3_dog columns from the table.
3. we have few columns which are redundant in table tweet_api_data and twitter_archives. We can drop the times-tamp,source,text,in_reply_to_status_id, in_reply_to_user_id in the twitter_archives.
4. we have tidiness issue in tweet_api_data and twitter_archive table to solve this we will merge the tables base on tweet_id and id.
5. we have some columns which values are empty so to solve this problem we will drop those columns. Drop column user, favorited, retweeted, contributors, coordinates, geo, place, quoted status id, and quoted status id str.
6. possibly_sensitive and possibly_sensitive_appealable contain same value in twiter_archives so we should drop the possible_sensitive and possibly_sensitive_appelable column
7. in Source column we have tags as well to remove this tag and get only the link we will use regex to extract the URL.
8. we have issue in column doggo, floofer, puppo and pupper as it have value as column name or None. So to resolve we will replace to Boolean for example if it has column name we will have true and if it has None then we will have false.
9. Dog's name are marked as None instead of nan in twitter_archives_clean table.we replaced the None to nan.
10. Some rows have invalid rating, with rating numerator less than 10 or denominator not equal to 10.to solve this we will drop such occurrences.