Data wrangling

WRANGLE REPORT

Raghad Altamrah

Data wrangling

Data Gathering:

First I gather the data from three different resources which is (twitter-archive-enhanced.csv, tweet-json.txt

,image_predictions.tsv) and start to clean the three data and find 8 quality and the 3 tidiness issue in the three data

Assessing Data:

So I use the three method's to find the tidiness issue in the data head(), tail() and sample() methods.

And I use info() , describe() , astype() to find the quality issue programmatically

Define the proplems :

So here is the quality issue that I find:

1.in rating_numerator there is rating less than 10

2.in twitter_archive columns (name) there is a multiple a which doesn't make sense

3.the tweeter id in tweeter archive is 2356 and tw_api= 2354 don't match each other

4.there is a duplicated data in image_predictions columns p1 and p2 and p3

5.there is a null values in twitter_archive specifc in (in_reply_to_status_id) columns

6.change the type of timestamp in twitter_ar to to_datetime timestamp()

7.Keep original ratings (no retweets) that have images

8.text column includes a text and a short link.

And tidiness issue :

1-first change the type of id's in twitter api to int to match the other tables

then i marged the tow data set twitter_ar and image predaction

and then i kept the rows that have image

2.assign the three columns to one column because it's not variables so we need one column

How it solved:

1-in rating_numerator there is rating less than 10 ,so i well replace it with the mean

2-2.in twitter_archive name there is multiple 'a' which doesn't make sense

so i will replace it with none.

3-i change I p1 and p2 and p3 the text to lower case to be easy to work on

4-there is a dupllicated data in image_predictions p1 and p2 and p3 so i delete it

5- Delete retweets by filtering the NaN of retweeted_status_id

6-Deleat the outlieres in twitter api in retweets coulmns equal to 79515.

7- change the type of timestamp in twitter_ar to to_datetime timestamp()

then Separate timestamp into day - month - year (3 columns)

8- text column includes a text and a short link.

9- cheak from the text and confert it to string then i add just the text.

first change the type of id's in twitter api to int to match the other tables

then i marged the tow data set twitter_ar and image predaction and then i kept the rows that have image

11.assign the three colmns to one columns becauese it's not varibles name so we need one coulmns

Finally:

The project was amazing and improves my skill in searching and finding the issues in data also improve my programming skill even though I don't take out all the errors also the title of the project is not that good maybe a little bit boring but that's ok