UDACITY DATA ANALYST NANODEGREE PROGRAM – PROJECT 4

WRANGLE AND ANALYZE DATA

Project includes 4 parts:

- Gathering Data
- Assesing Data
- Cleaning Data
- Analyzing Data

1- Gathering Data

- twitter-archive-enhanced.csv is downloaded manually. Archieve dataframe is created by pandas read_csv method.
- İmage_predictions.tsv file is downloaded by requests library and assigned to predictions dataframe.
- I tried to apply Twitter developer account but they rejected my application. So I download this file manually as well and added tweet_id,favorites,retweets column to tweets dataframe.

2- Assessing Data

I assessed tweets, predictions and archive dataframes both programmaticly and visually. I used info(),describe(), duplicated(), head(), tail(), sample(), isnull() methods to investigate data programmaticly. Dataframes were really messy and untidy! Here is what I found and classify them as below.

Quality

`archieve` table

- Erroneous datatypes for timestamp(string->datetime)

- Erroneous datatypes in_reply_to_status_id(float->string)

- Erroneous datatypes in_reply_to_user_id(float->string)

- Erroneous datatypes retweeted_status_id(float->string)

- Erroneous datatypes retweeted_status_user_id(float->string)

- Missing names or incorrect names at name column (a, an, the, just, one, very, quite, not, actually, mad, space, infuriating, all, officially, 0, old, life, unacceptable, my, incredibly, by, his, such)

- Missing stage of dog indicator for 1976 columns

- stage of dog column type should be category

- Rating numeator is higher than denominator. Rating denominator is not standard.

`predictions` table

- There are 66 duplicated jpg url.

- _ or uppercase is used at p1,p2,p3 columns.

`tweets` table

- tweet_id is string. Should be integer to have common variable on all table

- Retweets and favorites column should be integer

Tidiness

- Kind of dog has 4 columns: "doggo","floofer","pupper","puppo" should be under same column. (archieve)

- Retweeted tweets have same text and rates but diffirent tweet id. Deletion Retweets from table (archieve)

- p1,p2,p3 predictions should be 1 columns which states true one(predictions)

- Merge for archive, tweets and predictions on tweet_id

3- Cleaning Data

I created copies of dataframes to keep original ones as same it is. Later I started to solve tidiness issues as beginning.

Firsly I focused stage of dog category and created new column on archive table. I was aware that melt function was used for this kind of applications on training however I found my own way at quizzses and I wanted to use it here. Hence I populated informations to new column with for loop.

Secondly I dropped retweeted rows. These are basically same tweets with different tweet ids. So I dropped them on archieve table.

Thirdly I merged three dataframe to one which is called as “combined”. Dataframes are merged on tweet_id which was common column for three.

Finally I created predictions column on predictions table and populated most succesfull prediction to column. Predictions were mostly correct – I did a visual check.

After completed tidiness topic, I started to look quality issues.

I firstly focus on erroneous datatypes. I corrected them using asytpe() function. Lastly checked combined.info() if my steps are done.

Secondly I tried to fix name column. There was lots of inccorect or missing name information. I found incorrect names manually. For example I filtered dataframe where name is defined as a. And I checked all related texts and finally notice that some of them shared name of dog with “his names is...” or “named...” or “this is...”. There were always same explanation!So I found row numbers and fixed name information regarding to text. Later I decided to do this step more programaticly. I created a list which contains incorrect names. I filtered data frame with list and listed texts which has “named” phrases. Bingo! There was 2 rows which has incorrect name plus “named” pharases in text column. So I fixed name column accordingly. Again I did same work where name column defined as None. And I found more and fixed name column again.Finally I was sure that I found most of the missing names. Hence I assigned none for inccorect names also.

Thirdly I focused rating section. There was not any standard for rating. I decided to create standard 10/10 rate. I assumed that is numerator/denominator is higher or equal to 1, rating should be 10/10 which is highest. If it is less than I assigned denominators to 10.

Finally I focused stage of dogs column. I assumed that there might be more information at texts! People actually defined stage of dog but they did some puns hence these were not added to column. So I filtered text column is includes “puppy”,”floofer” etc where stage of dog is none. Bingo! There were lots of tweet like that! I usually selected tweets randomly and checked them visiually. If person stated puppy on text, picture had puppy as well. So I filled stage of dog column with this information.

Lastly I saved my combined dataframe to twitter_archive_master.csv.

4- Analyzing

I prepared a analysis for my results to complete act_report document. I found most favorited dog and retweeted dog, I checked rating regarding to favorites count or retweet counts.