Data Wrangling Report

Project objectives:

The project main objectives were:

• Perform data wrangling (gathering, assessing and cleaning) on the provided sources of

data.

• Store, analyze, and visualize the wrangled data.

• Reporting on

1. data wrangling efforts.

2. data analyses and visualizations.

Step 1: Gathering Data

In this phase, the three pieces of data were gathered and represented as pandas dataframes:

- The WeRateDogs Twitter archive ('twitterarchiveenhanced.csv')
- The tweet image predictions ('image-predictions.tsv').
- 'tweet_json.txt'

Note: I tried many times to get tweeter api without success till this report time.

- Assessing:

  After loading the 3 files converted them to data frames then started to check visually to quickly main issues.

  Start to assess programmatically using serval codes to check and review data frames and found out many quality and tidiness problems.

- I Found those Issues:

Qulatiy:

- most of in_reply_to_status column & .in_reply_to_user_id are empty
- retweets are present in the data
- some of the column names are not meaningful "source , text"
- "source" values are formatted as in html
- the dog nick names needs arrangement in one column .
- timestamp change to to_datetime
- dog_stage column have a wrong names
- "name" column have a lot mistaken names
- source column isn't readable
- clean columns from faulty and uncorrected data

- faulty tweets not related to dogs
- remove confusing characters from all columns

Tidiness:

- img_num useless.
- Just 3 columns needed id, retweet_count, favorite_count
- All datasets should be combined into 1 dataset only
- "retweeted_status_timestamp" is useless.
- id column isn't matching other columns key reference column tweet_id.
- drop unnecessary columns
- decrease decimals

- **Define:**

  I started to define how ill solve those issues and which functions and methods will help me to clean data frames.

  we only need 3 columns id, retweet count, favorite count "drop"

  change id column to tweet_it to match other tables

  drop unnecessary columns 'in_reply_to_status_id', 'in_reply_to_user_id', 'retweeted_status_id',

  'retweeted_status_user_id', 'retweeted_status_timestamp'.

  rename test column to comment

  clean columns from faulty and uncorrected data and replace faulty data with empty string

  strip "source" column url

  change "source" label to url

  combine doggo/ floofer /pupper /puppo in one column and delete the same

  change time stamp to_datetime

  lower name columns content

  remove confusing characters from all columns

  merge all tables in one master table

  decrease decimals

  Save cleaning process

  Create a rating percentage column to help me in data filtration

- **Clean:**

  I started to apply several cleaning techniques to:

  Rename columns

Remove unnecessary columns

Remove and replace faulty data

Fix columns data types

Rearrange the data frame to be more readable and logical to be used later

- **Test:**

  I tested my clean process applying different filters and reviews using function and methods.

After finishing cleaning and sorting data I started to prepare data for visualization by aggregating, grouping and slicing favored data that will be visualized.

I created a rating column to help me sort dogs by rate, got the rate percentage by dividing numerator on denominator.

- Visualization:

I started with bar and scatter graphs to get insights about the rating, likes for favored dog stages and names and found out that Tucker dog name got the highest ratings and Doggopuppo dog stage type got the highest rating and likes.

Notes:

While working I saved my work after cleaning at file ‘tweeter_archive_master.csv” and made a copy from it to start visualization.

After finished visualization I saved all my work at file “tweeter_archive_mater_c1.csv”