5/28/2021

Wrangle Report

Analysis of Tweeter user "WeRateDogs"

Andreja Ho

UDACITY'S PROJECT #4: WRANGLE AND ANALYZE DATA

Project Overview

The purpose of this project was to gather data from various sources, assess data visually and programmatically and finally clean the data in order to perform analysis and draw conclusions. For this project I was working with tweet archive of Twitter user @dog_rates, also known as “WeRateDogs”. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. For this project I’ve used Python and its libraries in order to retrieve, clean and visualize data. The following steps were taken wrangling process:

Gathering data

- **The WeRateDogs Twitter archive (twitter_archive)** dataset was a file on hand and it is stored as a ***CSV file***.
- **The tweet image predictions (images_tsv)** dataset was retrieved from Udacity’s servers through the Python ***requests*** library. The file was stored in the ***TSV file***.
- **Tweeter's retweet and favorite count (tweet_df)** data was retrieved from Tweeter’s website through ***API call***, using the ***Tweepy*** library. Retrieved data was stored in ***JSON format (tweet.txt)***.

Assessing Data

After all data was gathered, I read the data into Jupyter Notebook. I assessed data visually and programmatically. For visual assessment I used the code *“pd.set_option("display.max_rows", None, "display.max_columns", None)”* in order to view the entire data frame. For programmatical assessment I used the *DataFrame.info()*, *DataFrame.shape*, *type(DataFrame.column[0])*, *DataFrame.column.duplicated().sum().* After assessing the data, I followed:

The rules for tidiness:

- Each variable forms a column
- Each observation forms a row
- Each type of observational unit forms a table.

The rules for quality:

- Completeness: Missing data
- Validity: Does the data make sense
- Accuracy: Inaccurate data
- Consistency: Standardization

and documented all quality and tidiness issues as follows:

TIDINES ISSUES

images_tsv table
- Data from the table should be combined in master data frame once solving quality issues. Images table has no major tidiness issue. All the columns are needed to exctract data that we want to keep for our analysis. However, once we exctract all the data needed we will drop extra columns ans merge these columns in one master dataframe.

tweet_df table
- columns `id` and `id_str` have the same values (duplicated)
- unnecessary columns
- columns `retweet_count`, `favorite_count` should be in twitter archive table

twitter_archive table
- One variable is spread across four different columns: doggo, floofer, pupper, puppo columns
- some dog stages are doubled
- uneccesary columns
- newly exctracted raitings should be in twitter_archive table
- newly exctracted url should be in twitter_archive table
- all data should be in one master dataframe

QUALITY ISSUES

images_tsv table
- Some of the images aren’t dogs (we want only dog tweets ratings)
- Only one column with dog breeds

tweet_df table
- Some of the tweets are retweets and we want only original ratings

twitter_archive table
- Some of the tweets are retweets and we want only original ratings
- Timestamp is an object data type instead of datetime
- Erroneous extracted values for numerators (original values as well as decimal values in rating numerator)
- Erroneous extracted url
- text column contains numerators and urls and we want clean text column
- some dogs names aren't dogs’ names
- Rating_numerator is an integer, should be a float
- Rating_denominator is an integer, should be a float
- Favorite count is a float, should be an integer
- Retweet count is a float, should be an integer
- Duplicates

Cleaning Data

After assessment the data was cleaned using the following approach: **Issue name, Define, Code, Test** to clean Tidiness and Quality issues stated in Assessing Data section.