Wrangling report

Project 2: Wrangle and Analyze Data

Mohamed Abed - 15/9/2020

Data Gathering:

Twitter_archive_enhanced.csv: Downloaded manually from the Udacity project page.

Image-predictions.tsv: Downloaded using a url request using requests library in python, and the code doesn't request the file again once it's downloaded the file.

Twitter_json.txt: Downloaded using the twitter API tweepy and accessing it with twitter developer account and its tokens to have access to the tweets data, after that I saved that Json data line by line into a text file to use later into a dataframe.

Gathering results:

Twitter_archive_enhanced.csv: archive_df

Image-predictions.tsv: image_predictions_df

Twitter_json.txt: api_df

Assessing:

Visual Assessment:

Quality

- Representations of null values as string “none” in archive_df
- Columns 'timestamp', 'tweet_id' need modification in type in archive_df
- Missing data in columns such as 'name' which may need another api inquiry from twitter

Tidiness

- Dog types are stored as values in three columns e.g. 'pupper', 'doggo', etc.

Programmatic Assessment:

Quality

- Found many occurences of 'a' string as it may have been a default name used by the text extractor in archive_df
- Existing tweets with no images, and also found retweets inside the archive that needs to be deleted
- found archive_df has retweets, replies and ratings that doesn't have pictures

- Unrelated and empty columns in api_df, and renaming the column 'id' to 'tweet_id', and mergin the api_df with archive_df
- Unnecessary columns from archive_df using drop method

Tidiness

- Columns in image_predictions_df are value names and not variable names in p1, p2, p3 and choosing which prediction fit the image

Cleaning:

Steps done:

Changing 'None' string values with Nan values using replace method

Dropping unnecessary columns from archive_df using drop method

Finding the most accurate image predictions for the dog breed and using the final correct prediction and link it with the tweets

Cleaning archive_df from all retweets, replies

Changing 'a' string and replacing it with a np.nan values

Melting three columns with dog types into one variable column dog_stage

Removing unrelated columns in api_df

Renaming the column 'id' to 'tweet_id', and mergin the api_df with archive_df, replacing the timestamp column with column ‘created_at’ from api_df