WeRateDogs Project

Wrangling Report

1. Gathering Data

As per project requirements, data was gathered from 3 different sources from 3 different file formats.

The first dataframe twitter_archive was generated from the enhanced CSV file twitter-archive-enhanced.csv provided by Udacity which contained 2000+ tweets with ratings.

The second dataframe image_predictions of predictions of dogs breeds was generated by downloading programmatically a TSV file also provided by Udacity.

And the final dataframe tweet_json was generated using Twitter API and Tweepy library to grab the tweets data json object and writing each tweet data into a TXT file tweet_json.txt, then reading favorites count and re-tweets count from this file for each tweet.

2. Assessing Data

2.1 Visual Assessment

All of the 3 dataframes were assessed visually by displaying each dataframe in Jupyter Notebook to identify quality and tidiness issues, then more visual assessment was done through Google Sheets that helped identifying more issues.

All found issues were documented by category below each dataframe in the notebook.

2.2 Programmatic Assessment

After visual assessment, all dataframes were thoroughly assessed programmatically using pandas functions and methods like pandas.DataFrame.describe(), pandas.DataFrame.info() and using conditions to identify rows with inconsistent values.

All found issues were documented by category below each dataframe in the notebook.

3. Cleaning Data

The cleaning phase started with making copies of the 3 dataframes.

Each dataframe was approached individually in a 3 steps process that began with defining cleaning tasks from the issues previously addressed in the assessment phase, then writing a code to achieve this objectives and finally testing the outcome.

After cleaning each individual dataframe, a master dataframe was created through inner joining to achieve a complete observational unit.

After creating the master dataframe, it was re-assessed visually and programmatically for any more quality or tidiness issues.

When the master dataframe became suitable for analysis and drawing conclusions, it was stored as a CSV file as per project requirements.