Report on: 'WeRateDog' Data Wrangling

By: Moamen Jamal Elabd Abualhagag

Introduction

Real-world data rarely comes clean. Using Python and its libraries, you will gather data from a variety of sources and in a variety of formats, assess its quality and tidiness, then clean it.

Scope

Our goal: wrangle WeRateDogs Twitter data to create interesting and trustworthy analyses and visualizations. Therefore, this report is to show off briefly my wrangling efforts.

Data wrangling for this project consists of:

Data Gathering

- **Twitter archive file**: 'twitter_archive_enhanced.csv' was provided by Udacity in the classroom resource and downloaded manually then I read it programmatically, using Pandas library, into a data frame 'archive_df' in jupyter notebook using: pd.read_csv().
- **The tweet image predictions file**: 'image_predictions.tsv' is generated based on neural-network algorithms, and it is hosted on Udacity's servers, downloaded programmatically into a data frame 'image_predictions_df' using the Requests library: requests.get(URL).
- **Twitter API & JSON**: 'tweet-json.txt' instead of querying data in this file and according to mobile verification issues with accessing twitter API, I have used this file directly and read it line by line into a pandas data frame 'api_df' with column: tweet ID, favorite count, retweet count, and followers count.

Data Assessing

- **Visual Assessment**: the three datasets is displayed separately in jupyter notebook skimming for interesting findings, quality, and tidiness issues. Also, I have checked csv files externally on excel.
- **Programmatic Assessment**: using Pandas functions showing more about our datasets in details, summarizing statistics such: 'DF.info()', 'DF.shape()', 'DF.value_counts()',

‘DF.describe()’, and ‘DF.duplicated()’. Meanwhile, I am focusing on finding out quality and tidiness issues to address soon before going through the analysis.

Data Cleaning

This step focuses on addressing quality and tidiness issues observed while programmatic or visual data assessing, and it consists of three steps: defining, coding, and testing . The testing results will be shown in the notebook running the attached IPYNB file.

- **Quality Issues:**

  I recommend starting this process with making copies for original data frames.

  For ‘archive_df_clean’, ‘image_predictions_clean’, ‘api_df_clean’:
  I have addressed consistency issues such representation of null values as a string ‘none’, timestamp, and columns out of our analysis scope.

  Also, there are completeness issues such in the ‘name’ column which can be found instead of misleading values. In addition, tweets with no images referenced to the difference in number of rows between ‘archive_df_clean’ and ‘image_predictions_clean’, and replies besides retweets should be dropped from the three data frames.

  Accuracy issues: weird values in high rate such names with lowercase initials and ratings with decimals missing addressed and one-off occurrence errors, I assumed it will not affect our analysis, and missing values in ‘expanded_url’ column are for tweets with no images.

- **Tidiness Issues:**

  For ‘archive_df_clean’ and ‘image_predictions_clean’, there are column headers not variables; they are values instead. I have used one column ‘dog_stage’ for ‘archive_df_clean’ adding these values to it, and a Pandas function ‘Pd.wid_to_long()’ for ‘image_predictions_clean’ for better understanding that it does not share observational unit ‘tweet_id’ with the other two file but all about images. The followingcolumns: (‘in_reply_to_status_id’,‘in_reply_to_user_id’,'retweeted_status_id','retweeted_status_user_id', 'retweeted_status_timestamp') will be utilized to shed the retweet and replies from

our datasets and then will be dropped beside tweets in ‘image_predictions_clean’ not existing in ‘archive_df_clean’.

Data Storing, analyzing, and visualizing

Cleaned DataFrame(s): ‘archive_df_clean’ and ‘api_df_clean’ merged on ‘tweet_id’ and ‘image_predictions_clean_reshaped’ separately with the main one named ‘twitter_archive_master.csv’. Kindly, find the analysis and visualization insights in ‘Act.pdf’.