wrangle_report

September 18, 2022

0.1 Reporting: WeRateDogs Data wragle report

0.1.1 BY: Adeoluwa Olawamiwa

Data Wrangling Report To begin the data wrangling process, I had to gather three sources of required data in different ways. For the first and simplest method, I downloaded a CSV file from the Udacity website and read it into a dataframe called *df_1* . The second data source was a TSV file located at a URL provided by Udacity, so I used the requests library to download the file programatically. Then I read this file into dataframe called *image_pred*. For the third and most complex method, I used the Twitter API to read each tweet's JSON data in its own line in a TXT file. Once the file had been saved, I was able to read it line by line using the json library to eventually create dataframe *tweet_df*.

In the assessment process, I performed two types of assessment -- visual and programmatic. In the *df_1* dataframe, many of the issues were related to incorrect extraction of names, ratings, and dog stages from the text column. Additionally, the data was not tidy in the *df_1* and *image_pred* dataframes, as there were multiple columns each for several variables. In the *tweet_df* dataframe, there was some missing data due to some tweets being deleted. However, I noted that it would not be possible to retrieve this data elsewhere. Overall, I also needed to narrow down each dataframe to only original content with images, as well as tidy the data by having one dataframe for each observational unit -- tweet data, dog data, and image predictions.

I noted down 8 quality issues with 2 tidiness issues in this analysis, Most of which was from the *df_1* data. After noting down these issues, i created copies of the dataframe, then, began the cleaning process, which involved multiple iterations of defining the cleaning action, coding, and then testing the result. Then I started tidying the data; twitter_archive_clean and image_prediction_clean were easier to make tidy because the data was generally valid and accurate. However, i made sure the names in the twitter_archive_clean was all accurate by removing incorrect names, also, i made sure all data columns was all in their datatypes and also made column clear descriptions of image_prediction_clean. After making sure all my quality issues and tidiness issues has been corrected, i merged all three datasets together to make a single clean dataset.

To end the data wrangling process, I saved the merge data to a CSV file called twitter_archive_master.

In [ ]: