Parvina Pasilova

Data Analyst Nanodegree Program

Report for Wrangle and Analyze Project

As a part of Udacity's Data Analyst Nanodegree program, I had to wrangle WeRateDogs Twitter data to create interesting and trustworthy analyses and visualizations

In this report data gathering, data assessment and data cleaning processes are explained.

Data Gathering:

Following resources served as a way to gather data:

- Enhanced Twitter Archive called WeRateDogs was manually downloaded from Udacity's website.
- Image predictions dataframe was downloaded programmatically through Udacity's server.
- Json file with retweet counts and favorite counts was downloaded manually from Udacity's website as well. As a side note, when I tried to create account on Twitter in order to query the Twitter API, the sign up was unsuccessful.

Data assessment and Cleaning

After gathering each of the above pieces of data, I visually assessed them first and then programmatically for quality and tidiness issues. I have detected and document more than eight quality issues and two tidiness issues in wrangle_act.ipynb Jupyter Notebook. Then I cleaned the DataFarmes based on the prior assessment and tested my codes. After my DataFrames have met the clean and tidy data requirements, I merged them and created the df_master DataFrame.

Data Wrangling for twitter archive :

- I have identifies the tweet_id is in integer format but should be string, so I converted to right data type.

- The timestamp column is object but need to be converted to datetime format

- In the names columns there were Replace the values that are not real names in the "name" column such as 'a', 'such', 'the', 'just', 'getting' etc.

- In the dog stages column there were None values. To address this, I converted Nones and np.NaN to empty string '' for 4 columns (doggo, pupper, fluffer, puppo) and then stored these values underdog stages column.

- I have identified missing values in multiple columns and after analyzing their content it was found that we won't be using it in our data analysis thus I decided to drop columns: 'in_reply_to_status_id', 'in_reply_to_user_id' and etc.

Data Wrangling for image predictions:

- The given data included 3 predictions however, the dog breed prediction with the highest confidence level p1 was kept while other predictions were dropped.

- tweet_id is in integer format and was converted to string format.

Overall, this table was relatively clean and initially met the projects requirements are clean and tidy data.

Data Wrangling for Json file:

- id column is in integer format but should be string format

- Rename id column in to 'tweet_id' in order to be able to merge our dataframes successfully.

- This data frame had missing values in multiple columns. After analyzing the content of the columns, I decided to drop these columns as they won't be used in our data analysis: 'contributors', 'coordinates', 'extended_entities' and etc.