Project #4 – Wrangle and Analyze Data

Report:
Wrangling Summary

Aliscia R. Boyd
02-17-2021

For this project, I wrangled data from the WeRateDogs (@rate_dogs) twitter account. I gathered the data from three different sources, and then assessed and cleaned it before analyzing it.

Part 1: Gathering Data

After first importing all needed python libraries and packages, I then gathered data from three sources:

- ***dog_archive - A Udacity provided file containing Twitter archive data for the WeRateDogs twitter account.***
- ***dog_images – A second Udacity provided file containing images that are associated with the dog_archive tweets. These images have been fed through a neural network to classify whether or not they are truly of dogs, and give a confidence level of that prediction for up to three images per tweet.***
- ***df_twitter – A third file that contains additional data regarding the tweets that is not present in the dog_archive file.***

Part 2: Cleaning Data

After reading the files into my Jupyter Notebook and making copies of them to work with, I began my cleaning process by first checking the datasets for quality issues, such as missing data and incorrect datatypes, and then for tidiness issues, such as redundant data columns. This phase was performed both via visual and programmatic inspection of the data.

Once these issues were identified, I was able to document the methods I would use to fix them and apply those methods. The fixes were of course made programmatically via Python. I iterated my cleaning process by defining, coding, and then testing each required action. By the end of cleaning, I had resolved all noted issues with each dataset and then merge them on a common column ('tweet_id') resulting in one new dataset (dog_archive_clean) that I could analyze.

Part 3: Analyzing Data

The new dataset was then explored and analyzed programmatically to generate 3 insights along with accompanying visuals:

1. ***The unique values of ratings given to each dog, and how many times the same rating value appeared in the dataset. This could be posed as the answer to the question, "What ratings are most commonly given to dogs by the WeRateDogs users?"***

2. ***The correlations between the variables (column values) in the dataset. That is, whether or not the variables are able to affect each other's values either positively or negatively. A correlation matrix and a heat map were generated to accompany this insight.***
3. ***A word cloud was generated to show the most popular names given to submitted dogs. Names that are more prominent in the data are reflected in the cloud with bigger text.***

Overall, this project was very fun and interesting. I admit I would have liked to do more analysis and visualizations, but simply did not have enough time to do all that I wanted. Nevertheless, it was definitely an enjoyable experience!