Wrangle Report

Presented by Ali Hesham For Udacity DAND

INTRODUCTION

This purpose of this report is to demonstrate the efforts and the steps done to complete the weRateDogs data wrangling project in Udacity DAND.

This in this project we gathered the data from three different resources all related to twitter account user@dog_rates and is used to gather insights and visualizations on data after doing data wrangling processes.

PROJECT DETAILS

Three tasks were preformed in this project to adhere to the data wrangling processes which are:

- Gathering data
- Assessing data
- Cleaning data

GATHERING

The data for this project was collected from three different sources which I will describe briefly:

- **Twitter archive file:** the twitter_archive_enhanced.csv was provided by Udacity and I used requests library to download it programmatically
- **Tweet image predictions:** this file was provided by Udacity as well and included which breed of dogs was present in each tweet and It was predicted by a neural network, I downloaded it using requests library programmatically
- **Twitter API & JSON:** by using the tweet IDs in the WeRateDogs Twitter archive, I queried the Twitter API for each tweet's JSON data using Python's Tweepy library and stored each tweet's entire set of JSON data in a file called tweet_json.txt file. I read this .txt file line by line into a pandas dataframe with tweet ID, favorite count, retweet count, followers count, friends count, source, retweeted status and URL.

ASSESSING

After gathering the above data from different sources and saving each into its dataFrame using *pandas*

I assessed the data to check for any of the data quality dimensions as follows:

- Visually, I printed the three data frames in Jupyter Notebook to check them.
- Programmatically, by using different methods (e.g., info, describe, value_counts, sample, duplicated, isnull, sum), then I separated the issues encountered into quality and tidiness issues.

CLEANING

This part of data wrangling was divided to three parts to adhere to the three cleaning parts:

1. Define
2. Code
3. Test

First to start the process a copy from all data frames was made using *.copy()* then I went on to address each issue from assessing part into its own define, code and test cells to keep it organized and clean, the trickiest part was the one with dog ratings as it has more than one issue which were:

- Rating for numerator and denominator were not float
- Some ratings were not extracted correctly from text
- Some ratings were combined for multiple dogs in the same picture
- Some ratings were very low as the images did not have dogs in them

Other interesting cleaning code was to melt the dog stages in one column instead of four columns as original presented in twitter archive.

CONCLUSION

Data wrangling is core skill that will help anyone who deals with data.

I got hands on experience on important python packages that are very helpful like pandas, matplotlib, requests and NumPy which I will continue to practice more and more to sharpen my skills

Things that I learned:

- Web scraping techniques
- Exploratory data analysis
- How to gather data from different sources
- Assessing gathered data and address quality and tidiness issues
- Different ways to clean data

Overall, this project was so much fun and challenging looking forward for the rest in the next nano degree