Arturo Sparks

WeRateDogs - Wrangle Report

In order to have an accurate analysis of dogs as per the WeRateDogs' twitter account, the first and foremost action was to wrangle the data.

The wrangling process was divided into four important steps: *Gathering data, Assessing data, Cleaning data, Storing data*. We'll explain each wrangling step in more detail below:

Gathering data

Data was collected from three different sources: twitter archive file provided by Udacity, image prediction tsv file stored online and twitter api enabled data.

The twitter archive file was straightforward – I used pd.read_csv to create a dataframe from the file which I then named as 'wrd' since it is an easy acronym for WeRateDogs.

Collecting the image_prediction tsv file was similar to the twitter archive file. It required some extra steps since I was extracting straight from the url (e.g import requests) but reading into pandas was the same process. Dataframe was named 'impd'

Gathering the last dataset was more challenging but relatively smooth once API keys were set. First, I had to familiarize myself with the tweepy library in order to create a for loop list that would iterate through each tweet id to collect data straight from twitter. Then that list was converted into a dataframe called 'twitter_api'.

Assessing data

After gathering the WeRateDogs twitter archive, image prediction and the tweet api files, I assessed each dataset visually and programmatically.

The dirtier/messier the data, the easier it is to see issues visually because of their obvious odd appearances. I noticed 3 quality issues and 2 tidy issues. For example, inconsistency with strings that were sometimes capitalized or lots of null values (quality issues #1 & #2). Besides, it was easy to spot lots of unnecessary columns by just looking at the datasets (tidiness issues #1 & #2).

In terms of programmatic assessment, I dived in deeper into each dataset by reviewing each data types and finding some inconsistencies (quality issues #4 & #5) while also assessing whether values are appropriate for some fields (quality issue #6).

Although I originally assessed 8 quality and 3 tidy issues, during the insights process I made some extra observations which I will address in the Act Report..

Cleaning data

After making note of all the issues present, I started cleaning each one. Some were relatively straightforward like changing data types to their correct format (quality issues #3, #4 & #5) or just removing unnecessary columns by using drop method (tidiness issues #1 & #2).

On the other hand, there were some more complex ones like wrangling the values of both the numerator and denominator values (quality issue #6). I made the executive decision to replace any numerators with 0 and any denominator higher than 10 to 10. This is because no dog will receive a 0/10 nor a heavily scrutinized rating divided by anything other than 10.

Another important step was, after cleaning all the quality issues, merging the 3 datasets joining them based on the id. That way, we have a proper combined dataset.

Storing data

Finally, I saved the cleaned dataset as ‘twitter_archive_master.csv’.