wrangle_report

March 6, 2020

1 Data Wrangling Process Report

This workbook will address the efforts I took in the data wrangling process for Twitter's WeRateDogs account. Like the project details explains, this Twitter account posts images of dogs with a funny comment and finally a rating out of 10. These ratings are usually high, almost always greater than 10 and with the funny caption and images of dogs, this account has gained an incredible following. The wrangling process consists of three parts, the gathering, the assessing and the cleaning. This is usually followed by some analysis and conclusions for sense making.

First we will talk about the gathering part of the process. I imported all necessary libraries I deemed to be useful for this project, some of which included the pandas, numpy, and matplotlib.pyplot libraries, and of course tweepy which is Python's library specific to Twitter. Next we read in the WeRateDogs tweet archives using read_csv command and I named the dataframe t_arch. The file `image_predictions.tsv` was downloaded programmatically using the requests library and a URL that was provided. This file was labelled image_preds. Last file that needed to be gathered was the `tweet_json.txt` file which was a JSON data file that consisted of the tweet's retweet and favorite count and other additional data. Using Tweepy and the code that was provided, we used Twitter's API keys and was able to access everything completely.

Next, once we had all three files, good practice states to make a copy of these originals and use the copies for cleaning and analysis. `archive_clean = t_arch.copy() json_clean = wrd_tweet.copy() preds_clean = image_preds.copy()` Now working with `archive_clean`, `json_clean`, and `preds_clean`, we were able to begin the assessing portion. With assessing, we go through each data frame and see if anything abnormal pops out at us. Taking a look at all the datatypes using .info, to see if any datatypes were innappropriate for that use. We also looking for misspellings or values that stand out to us as questionable or invalid. Obviously it's impossible to see each listing, we use commands like .head() or .tail() to see the first top and bottoms portions. Once we found something that needed to be cleaned, we always note it down so that it's easy for us to keep track for later. We were to find at least 8 quality issues and 2 tidiness issues. Quality issues include the actual quality of the data, for example misspelling of a name or wrong datatypes, and tidiness issues are ones that are about the structure of the data. Is is good practice to make sure that each variable is its own column, each observation forms a row and each type of observational unit forms a table. So we found 10(and a half after iterating) quality issues and 4 tidiness issues and noted them all down under "Assess".

We went onto the cleaning portion of the wrangling process. Cleaning also consists of three parts. Usually Define, Clean, Test, where defining is stating how you will be cleaning the issue, not just reiterating what was in the assessment. Cleaning is where your code is present and you're implementing your defining. Finally you test, which is also code to see if your cleaning did what you meant to do. So for all 14 issues, these were the steps we took to finishing cleaning. Now

data wrangling is iterative, meaning that the above steps can be repeated if circumstances define them. So while we were tackling quality issue #6, I saw that there was another issue that came up and would be wise to look into. So I noted it down under the Assess section and continued onto cleaning that issue as well (this is where the half comes from).

Once we were finished and satisfied with our wrangling, we stored the dataframe into a csv file labelled `twitter_archive_master.csv`.

Lastly, was the analysis and visualization part. Now that we had clean data, we were able to perform some analysis and think of some insightful questions to answer. When I was in the process of data wrangling, I got to know the data a little bit and had questions that I wanted to answer throughout and finally was able to investigate further at the end. My analysis revolved around the 4 questions below. - How do favourites and Retweets Change over the Years? - Distribution of Tweet Sources. - Are Lower Ratings Given to Non-Dogs only? - Which Dog Breed was the most Predicted? The report `act_report` will describe the analysis and show some visuals to support the analysis done.

This was a comprehensive breakdown from beginning to end of the entire Data Wrangling process. Please refer to `act_report` for insight on the analysis and visualization portion.