Wrangle Report

By Gabriela Sikora

1. Introduction

In this paper, I plan to briefly describe my wrangling efforts, which can be found in the Data Wrangling section of 'wrangle_act.ipynb'. In particular, I will go into detail about the gathering, assessing and cleaning of data that I conducted.

2. Gathering Data

The data that was gathered was processed with 3 different methods.

2.1 'enhanced_df'

The first involved a simple method of manually reading in a csv for the provided 'twitter-archive-enhanced.csv', which held information enhanced information about the WeRateDogs tweets.

2.2 'prediction_df'

The second involved programmatically reading in a tsv that was created with a neural network to predict what dog breed a tweets image would hold. This was read in via URL with the Requests library and was then modified to become a csv.

2.3 'tweet_df'

As for the third, this information was read in with help of Twitter's API. With the tweet_id's from the first dataframe, the tweets were checked if they were free of errors, and then the JSON data was added to 'tweet_json.txt', which was then read in to become the third dataframe.

3. Assessing Data

This section involved looking at the 3 dataframes that were just gathered. This section is broken down into 3 sections to reflect the 3 different dataframes. Each section begins with a general visual assessment as well as a look into the .info() to see more detailed information about the data. From then on, each section is assessed programmatically in order to better understand the details of their respective dataframes.

3.1 Identified Issues

By means of assessing the dataframes, multiple quality and tidiness issues arose. The quality issues were related to any content related issues. Ultimately, 13 quality issues were addressed across the 3 dataframes, but naturally this number can easily continue to grow as there are many quality issues that can be found both visually and programmatically. As for tidiness issues, these are more structural in nature, and only 2 were found: the dog stages in the 'enhanced_df' table were spread across 4 columns instead of being in one column, and all 3 dataframes could, in fact, be merged into one dataframe.

4. Cleaning Data

The process of cleaning the dataframes is guided by what was found from the quality and tidiness issues during the assessment of the dataframes. To begin, the dataframes were copied as to preserve the originals. Then, for each of the quality and tidiness issues identified in the previous section, a programmatic or manual cleaning of the issue occured. This was done in 3 steps: defining what should be done in pseudocode, applying the definition with code, and then testing to ensure that the expected results occurred.

Also, the result of cleaning the dataframes resulted in new issues, which was notable with the cleaning of tidiness issues. This, of course, resulted in an iterative process wherein some issues identified are the product of partially cleaned dataframe.

5. Conclusion

Overall, the data wrangling process is a heavily iterative process, wherein each step is crucial when aiming to create visualizations and come to conclusions. Since there are no set rules as to data creation, the steps between the data and the final clean dataframe make a world of difference, and the cleaner the dataframe, the more accurate the final conclusions will be.