Wrangle Report

Date Created: 07/03/2019

In this report I will discuss the methodologies using during the wrangling phase of the WeRateDogs project.

My wrangling efforts can be divided into 3 main subcategories: Gathering, Assessing and Cleaning.

Gathering:

I gathered sources from 3 different locations.

The first source was programmatically downloaded from a url provided to me by Udacity. By using the .requests package the file was downloaded into the same directory as the jupyter notebook was launched.

The second source was downloaded manually from one of Udacity's workspaces. This was a relatively simple process whereby simply clicking on the started the download in my browser. The file was then moved to the correct directory.

The last source was created by scraping information from WeRateDogs twitter feed using the twitter api library, tweepy. Once I had created my twitter account, the process for applying for a twitter developers account was followed.

After authorising my developers accounts in the jupyter notebook using my secret consumer key, consumer secret, access token, access secret, I created an empty array and created a loop which ran through each tweet id provided to me in the second source. The twitter API produced an array in JSON format. This array when then dumped into a .txt file and then read by the pandas library.

Assess:

Each of the three sources was then assessed using a variety of manipulations including but not limited to, .head() .sample() .describe() .info()

Each three sources was visually and programmatically assessed and each issue that became apparent was then stored in a problems section that would be addressed in the final cleaning section.

Cleaning:

The problems that were found whilst assessing the 3 dataframes were then tackled using a variety of methods. The cleaning process was broken down further into 3 subcategories. 'Defining' – which defined the problem being solved. 'Code' – which showed the code being used to solve the problem and 'Test' – which tested the code to make sure it worked.

Once the 3 tables were cleaned they were concatenated and this table was then assessed and cleaned again.

Storing:

Once the final master dataframe was deemed clean and tidy enough for analysis, it was saved as a csv in the same directory as the jupyter notebook was launched.