WRANGLE REPORT

WeRateDog Project

Introduction:

Data wrangling is a multi-layer task that takes place in several stages: Data gathering, assessment and cleaning. The steps taken in WeRateDog project will be detailed below and further illustrated in Figure 1.

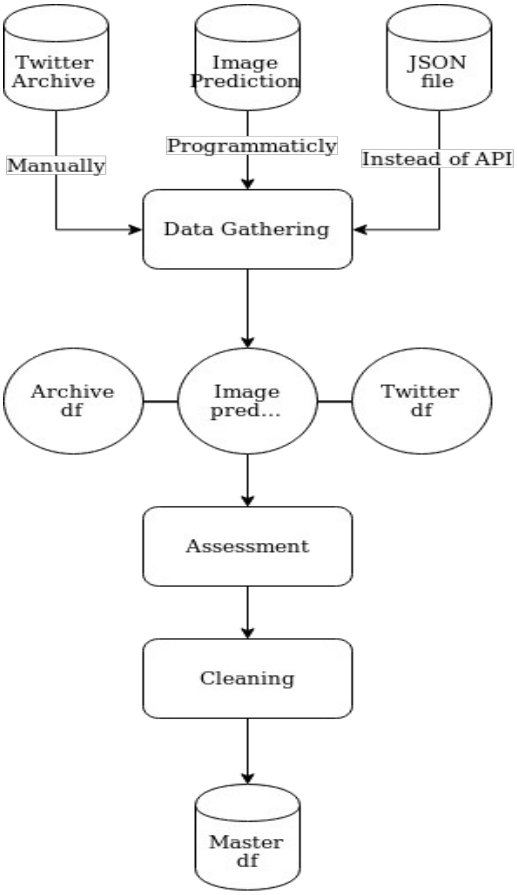

Figure 1: Data Wrangling Process

1. **Data Gathering:** was the first stage in the project. Initially, the twitter archive given was downloaded manually as instructed in the project details, and then assigned to a pandas dataframe, then image_predictions.tsv was downloaded programmaticly using the request library. Finally, I requested for the twitter developer account, unfortunately, the request was not approved by twitter with no reasoning mentioned in their email, figure 2, thus, I have decided to download the JSON file provided and work it without twitter account access, simulating the API JSON dictionary.

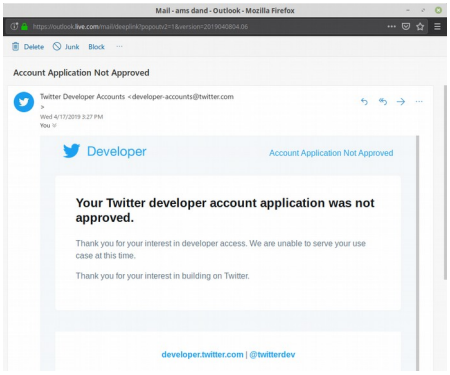

Figure 2, Twitter response for API access

2. **Assessment:** upon data gathering, all dataframes were created then assessment started to take place. First, dataframes have been inspected by looking to sample records using both sample() and head() functions interchangeably. Also looked into data using describe() function to understand its statistical shape. Furthermore, used the dtypes function test to assess the data types on each dataframe.

   The assessment stage were mostly identical in all dataframes. This stage took most of the time spent on this project as the data tables were having a lot of quality and tidiness issues. Inspection was iteratively performed in all dataframes. During this process, some issues have been identified such as data types, null values, incorrect data format, Missy data and other structural issue.

3. **Cleaning:** is the ending of data wrangling process, it occurs when all dataframes have been cross-checked and several issues identified. definitely, the optimum state is not guaranteed as the raw data had a lot of quality issues that would require further investigations, though, major problems were addressed and solved bringing the data to an acceptable state and preparing it to the visualization step.

   Key activities:

   1. Data type and format fix
   2. Removal of unnecessary columns
   3. Drop of null and / or missing values
   4. Merging some columns for consistency

Conclusion:

data obtained as instructed, unfortunately, one part was not successfully achieved (twitter API), instead JSON file used as a replacement. Data have been assessed for quality and tidiness issues in an iterative manner. Finally, all addressed issues were fixed and cleaned, resulting in a master data frame creation ready to be analyzed and further visualized.