WRANGLE REPORT

By – Vitalis Dexter

Introduction

In this project Wrangle and Analyze Data from the second curriculum of the ALX-T Data Analyst Nanodegree Program, we are going to download and Wrangle the archived tweets of WeRateDogs®. This archive will be assessed for quality ad tidiness, and then we will clean the archive, before analyzing and visualizing some vital findings from the downloaded archive. All these will be done using Pandas in Python.

Project Details

The data in the WeRateDogs® tweet archive was not expected to be clean and it wasn't. so using Python and some libraries, I was able to gather, assess and clean the data. This was because no good analyses can be done on untidy data. Because cleaning the entire archive to make perfect sense was going to be a lot of wrangling, we have been instructed to limit the number of quality issues to 8 and tidiness issues to 2 at a minimum.

Step 1: Gathering Data

In this phase of the project, I had to gather three pieces of data.

The first was to download `twitter_archive_enhanced.csv`, a CSV file which was provided by Udacity and it contained a list of the tweets from WeRateDogs® and some metadata.

The second was the `image_predictions.tsv`, a Tab-separated value (tsv) file that was also provided by Udacity and it contained image prediction data for all the images in the tweets found in the `twitter_archive_enhanced.csv` file.

The last piece of data I had to gather was Additional data from the Twitter API. For this, I needed to apply for a Twitter Developer account to get access to the Twitter API to be able to pull this data from Twitter using a Python library called Tweepy and store the JSON data from Twitter API to a text file `tweet_json.txt`.

Step 2: Assessing Data

After gathering all three pieces of data, I had to access them visually and programmatically for quality and tidiness issues, and this was where things got a bit messy. I had to detect and document at least eight quality issues and two tidiness issues in these 3 different data that I have gathered. To do this, I had to first convert all 3 pieces of data to dataframes. These dataframes were then assessed visually and programmatically for unclean data. During the visual assessment was when I found out

that some columns in the df_twitter_archive dataframe doggo, floofer, pupper and puppo could be merged into one column and called dog_stages also there was no documentation for the image prediction dataset, so I could not make any sense of some columns like p1,p2,p3, etc.

Step 3: Cleaning Data

This section is divided into 3 activities and this is where I had to clean all the issues that I documented in Step 2.

The first activity of this section was to make a copy of all the dataframes so I don't make any changes to the original dataframe.

The second was to use a standard framework, define, code and test. This would make it easy for anyone looking at the notebook to understand the processes and reasons for each action in the cleaning phase.

The third was to merge the cleaned dataframes to one master dataframe and drop all columns that won't be used in the analyses and visualization of the data.

Step 4: Storing Data

In this phase, I have to store the cleaned and combined dataframe in a single CSV file called `twitter_archive_master.csv` this is the dataframe that was now used for the remainder of this project. I did that by calling the `pd.to_csv()` Pandas function.

Step 5: Analyzing and Visualizing Data

This phase was the most interesting part of this whole project for me. I had completed the whole wrangling and now have to make sense of my wrangled data. In my project notebook `wrangle_act.ipynb`, I produced 3 insights and 1 visualization with the data I had cleaned.