WRANGLE REPORT

By Julia Prieto

Project Overview

The purpose of this project was to dive into the wrangling process, which consists of gathering, assessing, and cleaning data to later analyze and visualize findings. In this report, I will be going into my wrangling efforts to improve the tidiness and quality of the WeRateDogs *Twitter Archive, image_prediction.tsv*, and *twitter_json.txt.* Let's dive in, shall we?

Wrangling Steps

1. Gather Data:

The first step in data wrangling is gathering our data. I first read in the *twitter-archive-enhanced.csv* file and set it as our **first dataframe**. Then I downloaded the *image_predictions.tsv* programmatically by using the requests library and the url given in the classroom to create the **second dataframe**.

The last dataframe was arguably the most complex dataframe to obtain. This dataframe was gathered by creating a twitter developer account, setting a workspace for the project, and then using the Tweepy library to collect data on each of the tweet ids, all of which was loaded into the *twitter_json.txt* file. After loading the data to a file, I programmatically went through it to look for the variables needed for the analysis, and that was our **3rd and last DataFrame.**

2. Assess Data:

After I created all of the necessary dataframes, I began to visually and programmatically assess the data.

Visual assessment consisted of looking through the dataframes to find any errors, the most notable errors spotted visually were the high count of missing values and invalid values in the 'name' column. While these assessments were made visually, I made sure to confirm them programmatically.

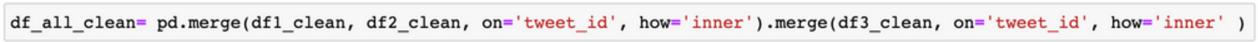
The programmatic assessment of these dataframes consisted of using functions such as *.shape, .value_counts(), .describe(), .isnull().sum(), and .str.islower()* to name a few. This assessment made it clear that there were several things I would need to clean up to ensure high-quality and tidy data for my analysis. **All of the issues** in quality and tidiness are explained in detail in my jupyter notebook!

3. Cleaning Data:

In total, I found **9 quality issues** and **3 tidiness** issues. After reviewing all of them, I went on to cleaning each dataframe, after creating a copy of them ,using the *describe, code, test framework*. This framework consists of describing the problem and how it will be fixed, coding the solution, and testing the code to ensure we were successful. The cleaning process for this project can be seen in my jupyter notebook.

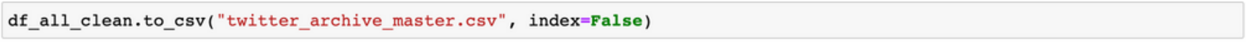
3. Storing Data:

After cleaning, I added all of the dataframes to a single *master DataFrame.*

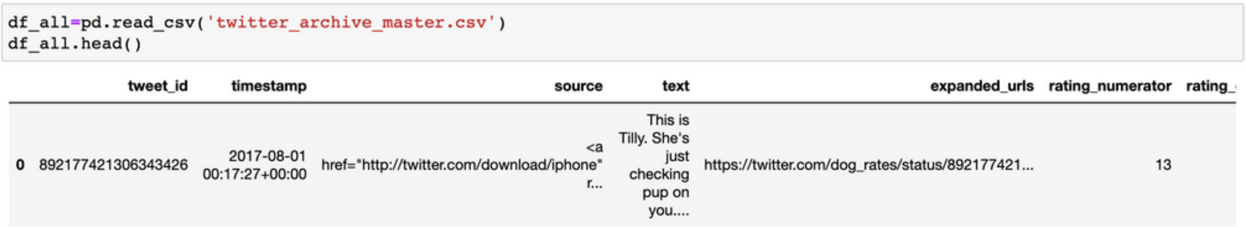


```
df_all_clean= pd.merge(df1_clean, df2_clean, on='tweet_id', how='inner').merge(df3_clean, on='tweet_id', how='inner' )
```

Then I proceeded to store it in a csv file called *twitter_archive_master.csv.*



```
df_all_clean.to_csv("twitter_archive_master.csv", index=False)
```



```
df_all=pd.read_csv('twitter_archive_master.csv')
df_all.head()
```

	tweet_id	timestamp	source	text	expanded_urls	rating_numerator	rating_
0	892177421306343426	2017-08-01 00:17:27+00:00	<a href="http://twitter.com/download/iphone" r...	This is Tilly. She's just checking pup on you....	https://twitter.com/dog_rates/status/892177421...	13	

Conclusion:

High-quality, tidy data is imperative to obtain a good analysis. This project taught me to be more observant and to fully familiarize myself with a data set before assessment. In a way, the cleaning process made the analysis and visualizing side of the project much easier, since I had already become familiar with all of the variables included in the master DataFrame and had cleaned them with analysis and visualization in mid. This project was as fun as it was challenging, I can't wait to apply what I have learned in future projects.