Wrangle Report

The wrangling efforts of this dataset consisted of three sequential steps: gathering, assessing, and cleaning.

Gathering

Three files were gathered for this project: twitter_archive_enhanced.csv, image_predictions.tsv, and tweet_json.txt. The image_predictions.tsv file was downloaded programmatically from Udacity's servers using the Request library and the OS library. Then it was converted to a pandas DataFram using the read_csv() method. The twitter_archive_enhanced.csv file was provided by udacity to be downloaded manually and it was directly opened and converted to a pandas dataframe in the jupyter notebook using the .read_csv() method. The tweet_json.txt file contained the JSON data for each tweet, and it was provided by udacity. However, to read this .txt file into a pandas DataFrame, I used the JSON library and .loads() method to convert the dictionary string to a python dictionary. This enabled me to extract the 'tweet_id', 'retweet_count', and 'favorite_count' from the .txt file, then append it to an empty list called df_list. This list is then converted to a pandas DataFrame using the . DataFrame() method.

Assessing

Assessed the dataset for quality and tidiness issues. From a data quality perspective, I assessed the data based on 4 dimensions. These dimensions are completeness, validity, accuracy, and consistency. When assessing the tidiness of the dataset, I checked the columns to see if each column represents a variable, which was not true in the twitter_archive table. four values of one variable were used as columns names (i.e. pupper, doggo, puppo, floofer). Another tidiness rule is that each row represents an observation, and each table represents one type of observational unit. This rule is met by merging three subsets of data into a complete one named twitter_archive_master table.

The quality and tidiness issues were assessed through visual assessment and programmatic assessment. The following methods were used to assess the data programmatically: .head() .tail .sample() .info() .describe() .value_counts() .duplicated() .isnull() . unique()

Cleaning

Before cleaning, I saved a copy of each file and performed the cleaning actions on the copy of the file. In the assess step, 8 quality issues and 2 tidiness issues were observed. The cleaning process consisted of three steps: define, code, and test. For each issue, I defined the action needed to fix the observed issue and the methods required to complete that action. Then I performed the cleaning code. And finally, wrote a piece of code to test and confirm that the changes were successfully applied.