Data Wrangling Report

Introduction

In this project, we will go through the process of data wrangling and work on gathering and define the quality, issues and tidiness of the data.

Gathering Data

This project contains three datasets:
Twitter archive (csv file) Image predictions for dogs (tsv file), download it programmatically as a URL by request library. Twitter info which is on twitter servers encompassed in Twitter archive, we will download it by Tweepy library.

Assessing Data

Assess them visually and programmatically for quality and tidiness issues. Detect and document at least eight (8) quality issues and two (2) tidiness issues

Quality issues:

For twitter_archive

1. Erroneous datatypes: shall be DateTime rather than timestamp
2. Incorrect dog names there are names like 'a', 'an', 'the', 'none'
3. Capitalize the first letter of the name
4. The favorites and retweets columns should be converted to int datatype
5. Since the tweet_id will not be used in the calculation, it should be a string instead of an integer

6. For the rating_numerator and rating_denominator should be converted to float rather than int

For image_predictions
7. Change the source to be more readable

For json_tweet
8. Since the tweet_id will not be used in the calculation, it should be a string instead of an integer

Tidiness issues:

1. The doge types (doggo, floofer, pupper, and puppo) are in separated columns we have to merge it in one column called dog_types
2. Matching the data in twitter_archive dataset and tweet-JSON because the data is sorted in separated tables

Cleaning Data

- Merge the json_tweet_clean and image_predictions_clean tables to the twitter_archive_clean table, both joining on tweet_id.
- Convert timestamp to datetime
- Show the inccorected names of dog then put it as "None" rows
- Capitalize the first letter in the name
- Change the columns name for to make it readable 'p1', 'p2', 'p3' , 'p1_conf', 'p2_conf' and 'p3_conf'
- The favorites and retweets columns should be converted to int datatype
- Change the tweet_id from int to str
- Drop useless columns
- The doge types are in separated columns we have to merge it in one column called dog_types
- Change the source instead of links to words to be more readable