Udacity Data Analyst Project 4: Wrangling

J. W. Thacker

August 4, 2021

1 Introduction

In this project, the 5000+ tweets and their corresponding data from twitter are downloaded from Twitter. These data that were downloaded needed to be cleaned (both for quality and tidiness) and this report describes the efforts that were needed in order to clean the dataset.

2 Gathering

All data wrangling efforts always start with gathering the data. For this project, the gathering stage started with downloading a .csv (the archive of tweets from twitter from the which contains things like the text of the tweet etc) and a .tsv (the predictions of the breed of dog from each dog) from the Udacity website. This was done using the python requests library and the code that does this can be seen in wrangle_act.ipynb.

3 Assessment

This section will contain the most pertinent of the data cleaning process that was used in this project. All code for these cleaning steps can be found in the python notebook listed in section 2.

Table 1: List of Quality Issues that were fixed

1. 'id' (in retweet list) should be a string since ids can't be added together.
2. The 'id' (in retweet list) column name should be renamed to 'tweet id' to kept consistent with the 'twitter archive' dataset.
3. 'id' (in twitter archive) should be a string since ids can't be added together.
4. entries 487 and 1592 (in image preds) are misclassified as a food.
5. entries 487 and 1592 (in image preds) are duplicated; delete one of them.
6. The tweet URLs need to be removed from the text of the tweet.
7. Edit timestamp to a datetime object.
8. Remove tweets that are retweets from the DataFrame.
9. Remove columns with mostly NaNs and drop the rows with missing URLs (there are only 59 or them).

Table 2: List of Tidiness Issues that were fixed

1. Get rid of the columns relating to "dog type", e.g doggo etc, as there are serious classs inbalance issues.
2. Combine the twitter archive cop dataset with the retweet list by using an inner join.