Wrangle Report

Introduction

The main purpose of this paper is to describe the wrangling procedure put in practice in data wrangling section of Udacity Data Analyst Nanodegree. The dataset used in this project is a tweet archive from user @dog_rates or WeRateDogs. This account rates people's dogs with humorous content.

Gathering Data

This project consists of three different datasets that are complementary to each other. A brief explanation on how to describe them is:

1. **Twitter archive file:** This dataset has been provided by Udacity may be found in this github directory as twitter_archive_enhanced.csv.
2. **Twitter image prediction:** It helps to categorize what breed is present in each tweet according to a neural network algorithm. This file is hosted by Udacity and can be downloaded programmatically through requests.
3. **Twitter API and JSON:** This is a compilation of JSON files. Through the tweet id acquired from twitter archive file, it is queried each tweet from each tweet id through Python's Tweepy library and then stored in a txt file. It is then read and transformed into a dataframe with selected columns.

Assessing Data

After gathering the three datasets, the datasets were assessed as follow:

- Visually, each dataset was printed separately in the Jupyter Notebook
- Programmatically, each dataset was explored through different methods such as the functions .info(), value_counts().

Through the methods used above it was possible to identify possible problems in the datasets that could be classified as quality issues and tidiness issues.

- **Quality Issues**
    1. In dataframe1 the columns timestamp is not a timestamp Dtype or date
    2. In dataframe1 the Doggo, floofer, pupper, puppo has nulls rows not being categorized as nulls
    3. Drop useless columns such as 'source','in_reply_to_status_id','in_reply_to_user_id','retweeted_status_id','retweeted_status_user_id','retweeted_status_timestamp','expanded_urls'
    4. Incorrected ratings done in rating_denominator from dataframe1
    5. Delete retweets
    6. Keep originals tweets only in df3
    7. Drop 66 jpg_url duplicated from df2
    8. Drop useless columns from df2 and create new columns using the predictions
- **Tidiness Issues**

1. Change tweet_ids from the all datasets in order to have the same type
2. Unite all tables into one table

Cleaning Data

This part of the project is divided in three: definition, coding and testing. Firstly it was created a copy of each of the three datasets, because if any mistake or error was committed along the cleaning process, the Data could be recovered from the original dataset without the need of running all the gathering data steps again.
The cleaning procedures **related to issues** were organized and done as follows.

1. **Twitter Archives Problems**
    a. Keep only original tweets
    b. Correct erroneous data types (doggo, floofer, pupper and puppo) columns
    c. Change column timestamp from object to type Data and create year, month and day columns
    d. Correct numerators
    e. Correct denominators
2. **Image Prediction Problems**
    a. Dropping duplicate images
    b. Create dog breed columns and confidence level columns
3. **Tweet json**
    a. Keep only original tweets

The ones **related to tidiness** were:

1. Standardize all tweet_ids into the same type
2. Merge all three data frames into a new one

Conclusion

Data wrangling is a core skill in the day-to-day work for any job within the Data field. Most of the time will be spent mining and cleaning data, therefore a good understanding of this skill and improvement is fundamental.
Some cleaning procedures were done for this project although only the problems mentioned were solved and the dataset may have other problems that won't be solved in this submission.