Wrangle Report

1 Introduction

The project task was to wrangle a dataset from the tweet archive of WeRateDogs, which is a Twitter account that rates people's dogs with a humorous comment about the dog. The requirements of this project are to assess and clean at least 8 quality issues and at least 2 tidiness issues in this dataset.

2 Gather

The first file to use was the twitter archive, which was made available by Udacity for manual download. Moreover, I used the Requests library to download a file with image predictions of the dogs, based on a neural network from Udacity's servers. Finally, I have applied for a Twitter developer account and downloaded programmatically additional information about the tweets and saved it in a file named tweet_info.csv (and tweet_json.txt).

3 Assess

I used visual and programmatical assessment to review the data for quality and tidiness issues. My findings were:

- Issue 1: three dataframes instead of one and too much unnecessary data (Tidiness)
- Issue 2: tweet_info - 'id'-column instead of 'tweet_id' (Quality)
- Issue 3: missing values ('tweet_id' in image_predictions vs twitter_archive vs tweet_info) (Quality)
- Issue 4: some of the dog ratings are retweets (Quality)
- Issue 5: twitter_archive: doggo, floofer, pupper, puppo are columns instead of values (Tidiness)
- Issue 6: twitter_archive: Nulls are text and not nulls (puppo, pupper, doggo, floofer column) and many entries are "None" (Quality)
- Issue 7: image_predictions: nondescriptive column headers (p1, p2, p3) (Quality)
- Issue 8: twitter_archive: incoherent values in the rating_denominator (150, 50, 2...) (Quality)
- Issue 9: twitter_archive: rating numerator is int instead of float (Quality)
- Issue 10: twitter_archive: rating numerator is not extracted correcty, e.g. 5 instead of 13.5 in row 46 (Quality)
- Issue 11: twitter_archive: denominator and numerator are in two columns instead of one (Tidiness)
- Issue 12: image_predictions: dog breeds are sometimes lowercase (Quality)
- Issue 13: twitter_archive: dog names are extracted not correctly (Quality)

4 Clean

In several cleaning steps, I used the df.iloc and pd.merge funcitons to create one master dataframe with the columns that I found to be useful for further analysis (tweet_id, timestamp, text, namerating_numenator, rating_denominator, name, retweet_count, favourite_count, picture url and the image_predictions table except for the columns img_num). Futher, I filtered and dropped off all retweets and melted the four dog-stage columns to one (by using df.agg() function with join() and dropping off all NaN values with dropna() and replace()).

In the next step, I renamed nondescriptive column headers (p1, p2, p3 etc.) and capitalized all dog breed prediction values. Moreover, I filtered for ratings that were extracted incorrectly using regex and replaced them with correct ratings (using df.iloc()) and merged the “rating_numerator” and “rating_denominator” columns to a single “rating” column. Finally, I used regex for dog names that were extracted incorrectly and used a further regex function to extract the dog names correctly.

5 CONCLUSION

The cleaned dataset consists now of one master dataframe with a manageable number of columns, including the text, the dog age, the breed, the retweet counts, favorite counts, images and image-predictions etc. This will provide a first good basis for the upcoming analysis. But even though substantial efforts have been undertaken to clean this dataset, I consider this process as an iterative one and other issues may rise up for cleaning along with the analysis efforts that will be taken in the further steps.