Data Wrangling Report

Introduction

This project's goal was to assess, clean and wrangle data from different sources. After the cleansing some interesting and trustworthy analyses and visualizations had to be created. The datasets contained information about the Twitter account WeRateDogs. As the name says, the tweet contains usually a picture of different dogs with text and a rating. The rating system is unique as the nominator can exceed the denominator.

Data Gathering

There were 3 sources, from where the data was gathered.

1. **Twitter Archive File**
   CSV file, which was offered from Udacity

   Content: Contains general information about the tweet. Furthermore, a categorisation of 4 different 'breeds' are included.

2. **Image Prediction File**
   TSV file, that needed to be downloaded with a request statement directly into the workspace

   Content: The dataset contains image predictions of dogs. 3 proposals with the highest prediction per tweet were stored in the dataset

3. **Twitter JSON API**
   JSON file, over Twitter's dev API

   Content: Raw information about each tweet on twitter. The original source of the twitter dataset.

Data Assessment

In the data assessment, each dataset was investigated on duplicates, shape, data type. Furthermore, some simple aggregations were made to visualize the content of each dataset.

Data Cleaning

There were 8 quality issues and 2 tidiness issues to be resolved.

Some of the quality issues were:

- Data type of timestamp change from string to date
- Retweets and Replies were to be dropped
- Get rid of the NULL values
- Clean dog names
  - A lot of the names were binding words
- Get rid of duplicates

Some of the tidiness issues were:

- Merge breed columns into one column (from 4 columns to 1 column)
- Merge the three data frames as they describe all the same tweets

Conclusion

This project was satisfying as I was able to use many data analyst tools. I remember each time how important it is to search for the right terms in the internet to find as fast as possible a solution. I enjoyed especially to scrap data from twitter as it feels really powerful to make meaningful insights.