Data Wrangling Report

By Ankit Narang

Data wrangling is the process which includes gathering the data from various sources, assessing it and cleaning the unnecessary pieces of data and fixing the issues and error present in the dataset. In this project, I have used the dataset extracted from the twitter account called WeRateDogs and performed wrangling as a part of the Nanodegree program of Udacity. WeRateDogs twitter account posts pictures of dogs and videos of dogs and rate them.

Step 1: Data Gathering

In this step, I have gathered the data provided by Udacity called “twitter_archive_enhanced.csv” which has been manually downloaded and “image_predictions.tsv” downloaded programmatically. Also, twitter API has been used to pull out data from twitter. I created a developer account on twitter and received the credentials to access the twitter api and gather data.

Step 2: Assessing Data

In the second step, I assessed the data collected in the data gathering stage and tried finding out quality and tidiness issues which were present in the data. As the data was collected from different sources there were various issues with it. The twitter data itself was huge and only those parts of data which were required were collected.The issues found in the dataset includes.

Quality

1. Several datatypes are wrong and should be changed.
2. Numerator and denominator do not explain about the ratings properly and some values are unusual.
3. In several columns, null values are not treated as null values (like in case of names).
4. There are some invalid names (a, an and other names with 3 and less character).
5. The data also contains retweets which means there is duplicated data present.
6. Some columns have lots of null values.
7. There are many columns in this data frame making it hard to read, and some will not be needed for analysis.

8. Some names starts with upper and some with lower character which means it is inconsistent.
9. Timestamp can be separated into date and time columns.

Tidiness

1. There are no need for some columns (doggo, floofer, pupper, puppo)
2. Merge 'tweet_info' and 'image_predictions' into 'twitter_archive'.

Step 3: Data Cleaning

Once we have problem in hand we start cleaning our dataset to make it easy to analyse and visualize it. The problems and issues that were encountered in the assessing stage were cleaned and the data set has been compiled for the analysis part.

Step 4: Analysis and Visualisation

In this stage, I used the clean data to get some insights from the data like the top most rated dog breeds, number of retweets over the year , relationship between retweets and favorites etc.