Documentation for Data Wrangling steps: gather, assess, and clean

Data in real world often comes messy and untidy, and sometime the data we have at hand may not be sufficient for the projec we are currently working on and we need to gather additional data from multiple sources, in various formats and assess and clean them which will improve on our analysis. This is known as Data wrangling.

In this project, data where gathered from three different sources namely,

- Twitter archive
- From a neural network that classified the breed of dogs
- And from the Twitter API.

The three data gathered were read into a pandas DataFrame and assessed both visually and programmatically.

During the assessment process, the datasets were as messy and untidy as every real world data. And below i’m going to define all the data quality issues encountered in the datasets and how it was cleaned or processed. Lets dive in!

From the twitter archive dataset, there were data quality issues namely

1) missing data in columns

2) errorneous maximum rating (1776), probably a retweet(detecting outlier)

3) Errorneous datatype for timestamp

4) multiple 'a' and 'an' in name column which could be replaced to None

5) html tags in source column

From the dog breed prediction table, there were also data quality issues like

6) rows with rating_numerator < 5 are not dog images

8) there are rows that predict false as dog. and they arent dogs

And tidiness issues encountered included:

1) doggo, floofer, pupper, puppo to form one column

2) merge the three datasets into a clean master dataset

Before the data cleaning process, its good practice to copy the data for cleaning. This is where the cleaning process on the data will be carried out.

Dealing with twitter archive dataframe, i dropped columns with mulitple missing data

I checked for outliers in the rating columns and dropped them as most of them i suspect were not the rating values but the number of retweet it had.

I converted the datatype of trhe timestamp column to a datetime datatype

Assessing the name column in the dataset i discovered that dogs where their names weren't included was referred to as None, and it was sometimes referred to "a' and 'an'. So i replaced the 'a' and 'an' name to None.

I also discovered that there were some html tags in the source column. So i removed the html tags in the columns using the regular expression format of removing unwanted letters and charaters to leave only the source of tweet.

Dogs ratings that was less than 5 was discovered to not having dog images in them when the link was assessed, so i dropped the rows.

From the image prediction table, during assessment, i discovered the alternative predictions made other than the models correct prediction wouldnt help in the analysis, so i dropped the colmns 'p2','p2_conf','p2_dog','p3','p3_conf','p3_dog'.

The Rows that had dog prediction as false where dropped because they ended up not being dogs.

To make the data tidy and following the data tidy policy, each variable should be a column, each observation a row, i used pandas to melt the doggo, floofer, pupper, puppo into one column.

After the cleaning of data, i stored the cleaned data in a csv file to be used for analysis and visualization.