Wrangle Report

This report describes what happened in the whole process of wrangling this data.

Any wrangle process has 3 main phases

1. Gathering
2. Assessing
3. Cleaning

Gathering:

This project contains 3 main files contains all data you need to work at

First is Twitter_archive_enhanced.csv which has all tweets at we rate dogs.

Every tweet has its own image and this images predict what breed of this dog according to a neural network and all of this information stored in file called image_prediction.tsv, this file is tap separated values.

So will read it too but with changing separator parameter

And the last file is some details in every tweet like number of favorites and retweet and this data are stored as

json which we will request them from twitter api using tweepy library in python language.

First two files are download programmatically using request library using this files url, third one is requested programmatically using twitter api as mentioned before, all these files saved at the project directory locally to work at them.

Assess:

This phase is to work as detective and look around data you had gathered to know what should you clean or modify at this data.

And you approach this problems using two main methods

First is to assess data visually by looking at this data and see how is data looks like at first rows for example

Second method is assessing this data programmatically by using pandas methods like .duplicated() which returns all duplicated rows for you and .info() to gives you information about number of nulls and data types and other methods like .describe() , .value_counts() and .mean()

These problems have two categories first is data quality issues and tidiness issues.

First data quality issues which is null data, duplicates, uncompleted data, inaccurate data and inconsistent or invalid data.

Second tidiness issues which is something like there are columns you can merge them to one single column

Clean:

In this phase you already collected you data and assessed it, now you know what is issues with this data so you will take problem by problem and fix all of them and return at the final of this phase clean dataframs.