Wrangling Report

Introduction

This wrangling report includes details of how I did wrangle of data from various sources including gathering collecting and cleaning.

Data Collection

Data was gathered from 3 different sources:

1) The enhanced twitter archive dataset was given by Udacity either by downloading it or working directly on Udacity work space "Jupyter Notebook"

2) Data was quired from twitter API with the given tweet ids from the enhanced twitter archive and then stored in the Json file.

3) The tweet image predictions file was downloaded programmatically using the Requests library from Udacity's servers.

Assessing data

After the data was gathered, assessment was performed using the following methods:

- .sample()
- .info()
- .value_counts()

Tidiness issues that were cleaned:

- Combining all dataset together because they refer to same tweet ids
- Combining all the dog nicknames "stages" into one table

Quality issues that were cleaned:

- Data with retweets
- Data that would make my analysis harder
- Timestamp was incorrect datatype
- Name contained the string "None" instead of a NaN
- Names were stored as verbs
- Rating numerators with decimals were incorrected exported
- Rating numerators with lower value than 10
- Rating denominators with value bigger than 20 "outlires"

- Unstandardized rating.

Cleaning Data

The issues found during the assessment process were cleaned and tested using the following methods and techniques:

- merge()
- reduce()
- .extract()
- .drop()
- .isnan
- .to_datetime()
- .islower()
- .replace()
- set_option()
- .loc[]
- .value_counts()
- .info()
- Regular expressions