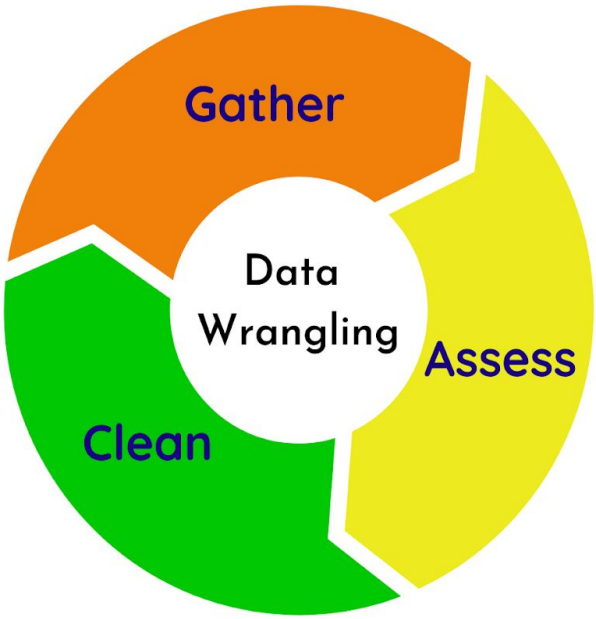

Wrangle act

Eya Sahli

Udacity Data Analyst Nanodegree

Overview

In this document I'm concisely describing my wrangling efforts of my wrangle and analyze data project .

Steps :

1. Data Gathering
2. Data Assessing
3. Data Cleaning

- **Data Gathering :**

  In this first step , I have downloaded the twitter archive enhanced file manually from Udacity ressources as well as json.txt file that i read line by line into a dataframe , keeping only retweets and favorites . for the image predictions file , i have also downloaded it via the provided link in Udacity's classroom and read it into a dataframe .

- **Data Assessing :**

  For the data assessing , I began with visual assessment through tail() and head() functions to identify any issues in the data , which provided a lot of insights such as erroneous dog breeds , names , missing data , etc .. Then I worked with Pandas's more advanced functions and methods to check for other quality issues such as wrong data types etc ..

  Through these two assessment methods ,I identified the following quality and tidiness issues :

  Quality issues :

  - there are 181 retweets and 78 replies that shall be removed.
  - the json dataset has 2354 rows compared to 2356 in the archive dataset , so there is missing data in the 'json additional data' via the api.
  - timestamp , retweeted_status_timestamp are of type object (string) instead of datetime.
  - tweet_id is of type integer but it should be a string.
  - there are 23 rows with the denominator being different to 10 , and numerator less than 10 , which can be problematic given WeRateDogs funny rating system .
  - there are 59 missing expanded urls (row 30 for instance).
  - certain dog names are invalid , having "a" , "an" , etc in names as well as None strings instead of Nan .
  - missing data in image predictions (2075 records) .
  - problem of random capitalization in p1 , p2 and p3 as well as the occurrence of underscores instead of spaces.
  - there are some False values in p1_dog , p1 is the most confident prediction so Falses would indicate that there are no corresponding dog breeds , those entries will be removed.
  - source column is in HTML format.
  - some dog breeds aren't actual dog breeds .

- in all 3 datasets , there are unuseful columns that will be deleted .

Tidiness issues :

- separate dataframes that we need to merge into one.
- combine doggo , floofer , pupper , puppo into one stage column .

- **Data Cleaning :**

  In Data Cleaning , I have proceeded to clean missing data first (expanded urls) , then tidiness issues than quality issues as highlighted in Udacity courses , and following the same steps (define-code-test) everytime .

  This project was a huge opportunity to me to learn how to wrangle data , which I believe is one of the most important things to do as a Data analyst.