Wrangle Report

○ Introduction :

In this project, I will implement a Data wrangling process and will use the WeRateDogs Twitter archive, which contains a Twitter account called @dog_rates that posts tweets about rating dogs and their photos.

○ Content :

The Data wrangling process:

- Gather Data :

  Gathering data is the first step in data wrangling. In this step we have three file I should to gather it.

  1. The WeRateDogs Twitter archive：
     This file contains tweets, dog names, ratings, etc.
     First I downloaded the file, then uploaded it and finally read the data into a pandas DataFrame.

  2. The tweet image predictions：
     This file is present in each tweet according to a neural network. It was downloaded programmatically using the Requests library and the following URL.

  3. Additional data from the Twitter API
     In this file was collected tweets retweet count and favorite count and stored in a file tweet.csv

- Assess Data :
  After gathering the data, assess data in two steps, which are visually and programmatically for quality and tidiness issues, then issues were identified as follows :

  Quality issues :
  1.Timestamp not object is datetime.
  2.Tweet_id not int is object.
  3.Delete retweet.
  4.Delete columns that not use.
  5.Lowercase text.
  6.Delete <a href= from source.
  7.Replace a, an, the from name to None and Lowercase name.
  8.Convert p1, p2, p3 the first letter capital only.
  9.Delete invalid numerator/ denominator.

  Tidiness issues :
  1.Create new column and add (doggo, floofer, pupper and puppo columns) to the new column.
  2.Create new dataframe.

- Clean Data :
  After the assessing step, I cleaned all the issues that I detected in the previous step, and the cleaning step goes through three stages, first the definition, then writing the code, and finally the test.

Conclusion :

Data Wrangling skills make up a huge chunk of that demand because so much of the world's data isn't clean.

Maryam Alablani