Wrangle report December 9, 2020

Introduction:

in this report we will describe our steps and effort through this project data wrangling weRateDog.

Table of Contents:

Gathering Data:

The first step for data wrangle process is **gathering data** through to complete the next step assessment and cleaning. In gathering data for project wrangling weRateDog contains three pieces of data:

1. The WeRateDogs Twitter archive. it provided giving this file then I Download this file manually by clicking the following named: twitter_archive_enhanced.csv.
2. The tweet image predictions, i.e., what breed of dog (or other object, animal, etc.) is present in each tweet according to a neural network. This file (image_predictions.tsv) is hosted on Udacity's servers and I downloaded programmatically using the Requests library and the following URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv

1. Each tweet's retweet count and favorite ("like") count at minimum, and any additional data you find interesting. Using the tweet IDs in the WeRateDogs Twitter archive,. Each tweet's JSON data I written to its own line. Then I read this .txt file line by line into a pandas DataFrame with (at minimum) tweet ID, retweet count, and favorite count.

Wrangle report December 9, 2020

Assessing Data:

After finish from the first step gathering data , we move to the next step is data assessment:

There are two types of assessment I used:

1. **Visual assessment**: each piece of gathered data is displayed in the Jupyter Notebook for visual assessment purposes.
2. **Programmatic assessment**: pandas' functions and/or methods are used to assess the data.

- **Data Quality issues:**

  - rating_numerator and rating_denominator columns are a float not as int.
  - Source column come as hypertext markup should be delete hypertext.
  - doggo, floofer, pupper, puppo columns should be combined into a single column.
  - three dataframes should be merged as they are part of the same observational unit.
  - Create a new coulumns for Calculated Rating for those columns rating_numerator and rating_denominator
  - Drop Nan from Name columns.
  - fillna with a replace missing values with that zero value in twitter_archive_enhanced_clean data frame.
  - Delete Duplicate from image_predict_clean data frame.
  - Delete columns that won't use in in analysis.
  - Drop Duplicate from tweet_id and jpg_url columns.
  - rename id_str column to tweet_id.

- **Tidiness Issues:**

  - Timestamp column in twitter_archive_enhanced table should be spilt into data and time.
  - reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id Columns format should to change.

Cleaning Data:

The Third step In Data Wrangling Is Cleaning Data. After we did assessment the data through quality and tidiness. it is time to work to cleaning the issues then start analysis. the cleaning step is very important step. I used the manual and programmatic way. also I did copy for our data before starting cleaning. and solve the issues and put the correct type for the date to help us to analysis and discover the insights.

Conclusion:

Data wrangling is a important skills that must who works with data familiar with. because the data in world is not always clean or ready to analysis in case, we do wrangle with wrong way will get wrong analysis then get incorrect decisions. the world generates a huge of data. and as human perhaps make a mistakes or lost data for that we need to do data wrangling.