Wrangle Report

Introduction

This project is a part of Udacity Data Analyst Nanodegree Where we need to wrangle data from different sources associated with tweets user @dog_rates also known as WeRateDogs.

The project goal includes

- Wrangling of data which consists of:
  - Gathering data
  - Assessing data
  - Cleaning data

- Storing, Analyzing, and Visualizing our wrangled data.
- Reporting the data wrangling efforts and visualizations.

Software Used

- Jupyter Notebook
- M.S Word

Python Libraries

- Pandas
- NumPy
- Requests
- Tweepy
- Json
- Matplotlib
- Seaborn

Gathering Data for this Project

The wrangling for this project includes these three sources to gather data

1. **Twitter Archive:** The WeRateDogs Twitter archive file was given to us by the udacity instructor to download and import it into the notebook. This archive contains basic tweet data for all 50000+ of their data.

2. **Image Prediction:** 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 was hosted on Udacity's servers and we had to download it programmatically using the Requests library and the following URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv

3. Additional data, including favoiurite count and retweet count were gathered using twitter API (tweepy).

Assessing Data

Once the data was gathered now i started to assess quality and tidiness issues.

Quality Issues:

- Data contains 181 retweeted users so we need to clean that 181 records from the dataset.
- tweet_id data type is int we need to change it to str.
- Timestamp data type should be changed from str to datetime.
- Html code in source column should be cleaned.
- rating_numerator and rating_denominator datatypes should be changed.
- name has many inaccuracy and different null values.
- very much less data in some columns.
- remove unwanted columns.
- Drop jpg_url duplicated columns.
- Dog names have "_" instead of space.

Tidiness Issues:

- A Single rating Column could help in better understanding.
- Set the rating to correct values where denominator was not equal to 10.
- Columns can be merged to a single column.
- All tables should be part of one dataset.

Cleaning Data

The issues found during the assessments were cleaned and tested using the concept of “Define Code and Test” using the following methods and techniques:

- merge()
- reduce()
- str,extract()
- loc[]
- head()
- info()
- isnull()
- loops
- astype()
- to_datetime()
- value_counts()
- drop()
- isnan()
- replace()
- Regular Expressions (RE)

Conclusion

All the data need cannot be found at a single place and readymade so we need to gather data from different sources and assess it to find the Quality and Tidiness Issues and Clean them before data analysis is performed.