WeRateDogs Project Wrangling Report

Data wrangling is the process of cleaning and unifying messy and complex data sets for easy access and analysis. Real-world data rarely comes clean. Using Python and its libraries, I have gathered data from a variety of sources and in a variety of formats, assessed its quality and tidiness, then cleaned it.
The dataset I have wrangled is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog.

Data wrangling consists of three steps:

- Gathering data
- Assessing data
- Cleaning data

Gathering data

For this project, I have gathered data from three different sources:

1. The WeRateDogs Twitter archive: This file was downloaded manually.

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 tsv file was hosted on Udacity's servers and downloaded programmatically using the Requests library.

3. Each tweet's retweet count and favorite ("like") count: This file was made using the tweet IDs in the WeRateDogs Twitter archive, querying the Twitter API for each tweet's JSON data using Python's Tweepy library and storing each tweet's entire set of JSON data in a file. Each tweet's JSON data was written to its own line. Then this txt.file was read line by line into a pandas DataFrame with tweet ID, retweet count, and favorite count.

   Gathering data was one of the challenging task in this project, especially the data collected through querying twitter API.

Assess Data

I have assessed data for:

- Quality: I have assessed eight quality issues. These were the issues with content. Low quality data is also known as dirty data.
- Tidiness: I have assesses four tidiness issues with structure that prevent easy analysis. Untidy data is also known as messy data. Tidy data requirements:

1. Each variable forms a column.
2. Each observation forms a row.
3. Each type of observational unit forms a table.

...using two types of assessment:

- Visual assessment: scrolling through the data.

- Programmatic assessment: using code to view specific portions and summaries of the data (pandas' head, tail, and info methods, for example).

Cleaning data

The last step in data wrangling is to clean it. It includes three steps :

1. Define: I have converted my assessments into defined cleaning tasks. These definitions also serve as an instruction list.
2. Code: I have converted those definitions to code and run that code.
3. Test: I have tested my datasets, visually or with code, to make sure my cleaning operations worked.

SUMMARY:

Data wrangling is a very important skill without which data analysis and visualisation wouldn't work. This project has helped me a lot in improving my skills about data gathering, assessing and cleaning – the three core steps of data wrangling.