Data Wrangling Report

This project is part of Data wrangling section of Udacity Data Analyst Nanodegree. It involves using Python and its libraries, gathing data from a variety of sources and in variety of formats, assess its quality and tidiness then clean, document all the wrangling efforts in a Jupyter Notebook and finally showcase them through analyses and visualization using Python

The project tasks are structured as follows:

- Data Gathering
- Assessing data
- Cleaning data
- Storing data
- Visualizing the data
- Reporting

Data Gathering

Different methods were used to gather data from different sources like:

- Twitter archive data – twitter_archive_enhanced.csv was provided by Udacity, and I downloaded the file manually before importing it to the Jupyter notebook using pandas.
- Tweet image predictions – To get the tweet image predictions, I used the requests library to programmatically download the data from a URL provided by Udacity.
- Additional data from Twitter API - Gathered each tweet's retweet count and favorite count by creating an API object to use to gather twitter data, then queried each tweet ID, wrote its JSON data to a tweet_json.txt file with each tweet's JSON data on its own line and lastly read the file line by line and created a pandas DataFrame from the list of dictionaries.

Data Assessment

After gathering the data, I assessed each table both manually and programmatically for tidiness and quality issues. All issues identified in the assessment stage were documented for implementation in the cleaning stage.

Manual data assessment involves scrolling through the rows and columns in the dataset to identify issues with the data.

Programmatic assessment involves using lines of code to assess the data.

Data Cleaning

Before performing any cleaning operations, I first created a copy of each dataset.

Cleaning was done using the define, code and test methodology. This was done for both tidiness and quality issues.

Storing Data

After successful cleaning operations, data was stored to a twitter_archive_master.csv file.