DATA WRANGLING REPORT

Objective

The objective of this project is to carry out a data wrangling by gathering three datasets from different data sources having different data formats, assess the quality and tidiness of the datasets, and then clean it to make it analysis ready.

Furthermore, I explored the datasets and generated useful insights through analysis and visualization making use of the Python libraries.

Data Gathering

Below are the step by step procedures used in gathering the datasets;

- Manually download the WeRateDogs Twitter archive data provided by Udacity and read it into a Pandas dataframe.
- Programmatically download the second data (image prediction) using the Requests library and URL hosted on Udacity's servers.
- Query Twitter API for each tweet in the Twitter archive using Python's Tweepy library and save JSON in a text file. Write each tweet's JSON data to its own line, then read the tweet_json .txt file line by line into a pandas DataFrame with tweet ID, retweet count, and favorite count.

Data Assessment

Following the gathering of the three datasets, I carried out a visual and programmatic assessment for quality and tidiness issues. Below are some of the assessment methods deployed;

- data
- data.info()
- data.shape
- data.describe()
- data.isnull().sum()
- data.duplicated().sum()
- image_data
- image_data.info()
- image_data.shape
- image_data.describe()
- image_data.isnull().sum()
- image_data.duplicated().sum()
- df.info()
- df.describe()
- df.isnull().sum()
- df.duplicated().sum()

Data Cleaning

The steps below were taken in cleaning the datasets;

I created a copy of the original data before cleaning

I Adopted the define code-test framework

I documented the define code-test framework

I Created a master data frame with all pieces of gathered data All the issues identified while assessing data will be cleaned and tidied up, the issues identified include:

Quality

1. Tweet_id in the three source dataframes has incorrect data type int64 instead of object
2. Some rows have retweet values
3. Expanded_urls column has missing records
4. Expanded_urls column has rows with repeated links
5. Timestamp column has incorrect data type int instead of datetime
6. Timestamp has +0000 which is not relevant
7. Columns in_reply_to_status_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp, in_reply_to_user_id not needed
8. JPG_URL and IMG_NUM has rows with missing record

Tidy

1. Dog stages are in different columns and should be melted into a single column.
2. The three datasets to be merged into a single data set.
3. The columns p1, p2, p3 should form a single column and columns p1_conf, p2_conf, p3_conf should also form single column.

Storing Data

Following the gathering, assessing and cleaning of the datasets, the master data was stored in a CSV file named twitter_archive_master.csv