ALX-T DATA ANALYST NANODEGREE PROGRAM

PROJECT 2: WRANGLE REPORT

Introduction

This report is a description of the wrangling process I undertook in Project 2 (Data Wrangling Project – Twitter Account WeRateDogs).

The report is divided into 3 sections (following the 3 steps of the Data Wrangling Process):

A. Gathering Data
B. Assessing Data
C. Cleaning Data

A. Gathering Data

The initial step in the data wrangling process entails collecting the data required for the project. This project uses three datasets:

- **twitter_archive_enhanced.csv** – the WeRateDogs Twitter Archive dataset, which is readily provided for manual download. After downloading the dataset, I uploaded it to the Project Workspace and read the data into a Pandas dataframe.
- **image_predictions.tsv** – a file containing dog image predictions, created using a neural network. I downloaded the file programmatically from the Udacity's servers using the Requests library.
- **tweet_json.txt** – a file containing data gathered from the Twitter API. I queried the Twitter API for additional information (e.g., retweet count, favorite count) using the tweet IDs in the WeRateDogs Twitter archive, and stored the data in the file tweet_json.txt

I then read the data in the three datasets into three separate dataframes; one for each dataset – in readiness for Step 2 of the Project.

B. Assessing Data

The second step in the Data Wrangling process is assessing data. Assessing data entails inspecting the data for data quality issues (content issues) and tidiness issues (structural issues).

I employed both visual and programmatic assessment to the three datasets. I encountered the below issues:

Data Quality Issues

1. Inconsistent Identifier (ID) column names in twitter_archive and tweet_json Datasets [Consistency Issue]
2. Breed names having lowercase first letters in p1, p2, p3 columns of image_pred dataset [Consistency Issue]
3. Timestamp Column in twitter_archive dataset is of datatype "string" instead of "datetime" [Validity Issue]
4. The "id" column in the tweet_json dataset has a lot of duplicated values [Validity Issue]
5. The image predictions dataset has missing records, as compared to the twitter_archive dataset [Completeness Issue]
6. The rating_denominator column has values greater/less than 10 in the twitter_archive dataset [Accuracy Issue]
7. There are retweets in both the twitter_archive and the tweet_json datasets [Validity Issue]
8. There are missing dog names (denoted with 'None') in the twitter_archive dataset [Accuracy Issue]

Data Tidiness Issues

1. Dog stages in twitter_archive dataset are in separate columns
2. There are irrelevant columns in twitter_archive dataset
3. There are irrelevant columns in tweet_json dataset
4. There are three separate dataframes (twitter_archive, image_pred, tweet_json) from the 3 datasets

C. Cleaning Data

Cleaning data is the last step in the data wrangling process. Here, the data quality and data tidiness issues identified in the assessment step are corrected in the datasets.

I first created copies of the three dataframes generated from the three datasets. I then proceeded to clean the data issues programmatically, one at a time; using the Define-Code-Test framework.

As the final step of cleaning process, I proceeded to merge the three cleaned dataframes into one master dataframe. I then exported the master dataframe to a final CSV file – twitter_archive_master.csv.

Conclusion

The project was carefully crafted to ensure learners have a deep dive into data wrangling. The project is a true reflection of the real-world challenges data specialists undertake in ensuring they have clean data for analytics. I am amazed to have applied majority of the concepts taught in Lesson 3 in a challenging end-to-end data wrangling project.