DATA WRANGLING REPORT

Done by: Ajoke Yusuf

Project

ABOUT

The project is aimed at Data wrangling, documenting wrangling process in a Jupyter Notebook, and showcasing them through analyses and visualizations using Python libraries.

The dataset that is wrangled is the tweet archive @DogRates, also known as @WeRateDogs. 'WeRateDogs' is a Twitter account that rates people's dogs with a humorous comment about the dog. This rating is almost always have a denominator of 10.

This report describes my wrangling process

Project Details:

The process carried out include:

1. Gathering Data
2. Assessing Data
3. Cleaning Data

Gathering Data

I gathered three different datasets that were obtained as following:

Twitter archive file: This data was provided in the project and was downloaded. I imported all the necessary python libraries needed and I used the pandas read_csv() function to read the file into a dataframe named twitter_archive.

Tweet image prediction file: The tweet image predictions was downloaded programmatically using the Requests library and the following
URL: **https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv**

Tweets_json_txt: I used the tweet IDs in the 'WeRateDogs' Twitter archive, I queried the Twitter API for each tweet's JSON data using Python's **Tweepy** library and store each tweet's entire set of JSON data in a file called tweet_json.txt file. I also gathered **each tweet's retweet count** and **favorite count**.

Each of the tweet's JSON data should was written in each line, then I read the .txt file into pandas DataFrame having the following columns: **tweet ID, retweet count, and favorite count**

Assessing Data

After gathering all the 3 datasets, I assess them visually and programmatically to detect quality and tidiness issues.

Visual Assessment: I printed the three different data frames individually in a jupyter notebook and scrolled through each of the dataset carefully. I also assessed the data after importing the file in Excel spreadsheet.

Programmatic Assessment: I used various python functions such as .shape, .info(), .describe(), .dtypes, .value_counts(), .isna(), .duplicated().

During the assessing of the dataset, I was able to detect 8 quality issues and 2 tidiness issues and the issues were well documented

Cleaning Data

Before I performed the cleaning process, I made a copy of all the 3 original datasets. During cleaning, I used the define, code, test framework and clearly documented it.
The following are the cleaning process carried out

i. Dropped in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, expanded_urls, retweeted_status_timestamp columns.
ii. Replaced None value in name column as np.nan and drop
iii. Changed tweet_id datatype in archive table to object.
iv. Changed timestamp datatype to datetime.
v. Removed jpg_url columns.
vi. Dropped p1_dog, p2_dog, p3_dog.
vii. Convert tweet_id data types to object.
viii. Merge the 3 tables together based on 'tweet_id'.
ix. Melt the p1, p2, p3 columns into a single column.