WRANGLE REPORT

Project 2.

Dataset: WeRateDogs Twitter Dataset

Author: Uchanma Adeola Igbokwe.

Date: June 2022

The Data analysis process is commonly grouped into the following steps:

- Ask Questions
- Wrangle Data
- Explore
- Draw Conclusions, and
- Communicate findings.

Data wrangling is the second step in the data analysis process. It comprises of three steps, namely: gathering or collecting data, assessing data and cleaning data.
This project was carried out using data gathered from the WeRateDogs twitter account. We Rate Dogs (@dog rates) is a Twitter account dedicated to reviewing images of dogs in lovely poses and giving them scores above 10/10. Since its inception, it has amassed over 7 million followers.

Gathering Data

The data gathering requirement for this project is divided into three parts, all carried out using different methods of data gathering. The first part of the dataset titled, 'Twitter_archive_enhanced.csv' was downloaded manually and read into a pandas data frame named 'archive_df'. The second dataset was downloaded using the Requests library programmatically and finally read into a pandas data frame named 'image_df'. This second dataset contains a collection of dog images represented by their unique URLs. For the third dataset, a Twitter elevated access developer account was obtained and utilized for gathering the data through the Twitter API using the Tweepy library. Collected data was saved to a tweet_json.txt file and finally read into a pandas data frame named tweet_df.

Assessing Data

This involved examining the individual datasets to identify data quality issues and tidiness problems. This was achieved both visually (Excel and pandas data frame) and programmatically using python functions such as head(), value_counts(), sample(), etc. The quality issues and tidiness problems discovered include:

```
In [123]: image_df.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2075 entries, 0 to 2074
Data columns (total 12 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   tweet_id  2075 non-null   int64
 1   jpg_url   2075 non-null   object
 2   img_num   2075 non-null   int64
 3   p1        2075 non-null   object
 4   p1_conf   2075 non-null   float64
 5   p1_dog    2075 non-null   bool
 6   p2        2075 non-null   object
 7   p2_conf   2075 non-null   float64
 8   p2_dog    2075 non-null   bool
 9   p3        2075 non-null   object
 10  p3_conf   2075 non-null   float64
 11  p3_dog    2075 non-null   bool
dtypes: bool(3), float64(3), int64(2), object(4)
memory usage: 152.1+ KB

In [124]: image_df.isnull().sum()
Out[124]: tweet_id    0
jpg_url     0
img_num     0
p1          0
p1_conf     0
p1_dog      0
p2          0
p2_conf     0
p2_dog      0
p3          0
p3_conf     0
p3_dog      0
dtype: int64

In [125]: print(image_df['p1_dog'].value_counts())
```

Quality issues:

Twitter Archived Dataset

1. Incorrect data type for the timestamp column (object instead of datetime)
2. Certain tweets have retweeted_status_id, retweeted_status_user_id, and retweeted_status_timestamp.
3. Missing values in expanded_url column.
4. Numerous missing values (over 80%) in the in_reply_status_id, in_reply_user_id, retweeted_status_id,retweeted_status_user_id, and retweeted_status_timestamp columns.
5. Zero values in rating_numerator and rating_denominator columns.
6. Name entries such as 'None' and 'a', 'an', 'mad', etc in the name column. All non-dog names entries start with lowercase.

Image Predictions Dataset

7. Inconsistent entry format in the p1, p2 and p3 columns. Some predictions start with uppercase and others with lowercase.
8. Undescriptive column headers (p1, p2, p3, p1_dog, p1_conf and the other p columns).

Tidiness issues

1. Different columns for the dog stages (doggo, floofer, etc) instead of one (twitter archive dataset).
2. The three tables should only be one table. Every observational unit should be a table.

Cleaning Data

The dataset was cleaned based on the quality and tidiness issues identified while assessing the datasets. Cleaning followed the define, code and test procedure in handling all problems found. Copies of the original datasets were made before cleaning actions were performed. All problems found were resolved programmatically and the final cleaning result saved by merging all cleaned datasets on their common column (tweet_id) using the pandas merge function. Codes were written after every cleaning step to ensure cleaning efforts were successful. Functions such as pandas islower(), str.lower(), drop() and other were used in cleaning. The merged datasets were saved to a csv file named 'twitter_archive_master.csv'.