DATA WRANGLING REPORT

By Edeh Gabriel Chidera

For a data wrangling project at Udacity data analyst nanodegree program

The project is a data wrangling of the dataset from the tweet archive @DogRates, also known as @WeRateDogs. We rate dogs is a Twitter account that rates people's dogs with a humorous comment about the dogs. This ratings almost always have a denominator of 10, but the numerator can be above 10.

The wrangling effort were conducted following three main steps:

- **a.** Gathering data
- **b.** Assessing the gathered data for quality and tidiness issues
- **c.** Cleaning the identified quality and tidiness issues

Gathering Data

Three datasets were used for this project and they were obtained as follows:

Twitter archive file: This data was provided in the project guideline. I downloaded it and uploaded it into the jupyter notebook in my local machine. I first of all imported all the python libraries that was needed for the project. Then, I read the first dataset into pandas dataframe using read_csv() and it was named df_one.

Tweet image prediction file: Using the already imported requests and os libraries, .get() function of the requests library were used to gather the data through its url and saved in response variable. With open function of Python , the response's content were written to a tsv file in the same working directory. The downloaded tsv file was subsequently read into pandas dataframe named 'df_image_predictions'.

Tweet_Json text: Twitter developer account were created and an application was created after approval from twitter management. The app credentials (consumer_key, consumer_secret, access_token, and access_secret) were used for the twitter API authentication. The tweet_id from the first dataset were used to scrape the need data. With the Python 'with open function', the tweet_json.txt were read line by line and the needed data were subsequently extracted. This was later read into a pandas dataframe.

Assessing Data

The already gathered three (3) datasets were assessed visually and programmatically.

Visually: The three dataframes were printed individually in the jupyter notebook in my local machine and glanced through and thoroughly.

Programmatically: Various programmatic assessment were carried using various python pandas methods and functions such as .info(), .shape, .isnull().sum(), .head(), , .sample(), .duplicated(), .nunique(), .column.

Cleaning Data

This part of the data wrangling process were carried out in three different steps:

a. Define

b. Code

c. Test

These three steps were each used to address the quality and tidiness issues identified in the assess section.

First, a copy of the original three datasets were made.

The copied datasets were named as follows:

1. df_one_unclean – for the twitter archive enhanced
2. df_two_unclean – for the image_predictions
3. df_three_unclean – for the additional file scrapped from twitter

Using **define, code and test** process, the following cleaning efforts were carried out.

Quality issues

Twitter-Archive_Enhanced (df_one) – df_one_unclean

1. Columns with high amount of missing values was categorized as low level information columns and hence were dropped. The columns dropped include: (in_reply_to_status_id','in_reply_to_user_id',retweeted_status_id','retweeted_status_user_id','retweeted_status_timestamp', expanded_url)
2. Timestamp column were converted from object to datetime format
3. The source column were cleaned to get a more presentable values
4. Name column were renamed 'dog_name' as this is better for information purpose
5. tweet_id column in int format were converted to string format

df_image_predictions

6. The second and third likely prediction were dropped since they have low prediction rate
7. tweet_id column in int format were converted to string format

additional_df

6. tweet_id column in int format were converted to string format

Tidiness issues

1. The various dog_stages in the different columns were collapsed into a single column named 'dog_stage'.
2. The three dataframe were merged in order to attain the structural goal of only ratings with images.

Storing the Data

After the wrangling effort, the merged data was saved as a csv file named twitter_archive_master.csv.

Conclusion

This project helped me to practice all that I learnt from the course contents. It was very exciting and I am looking forward to more future projects.