Data Wrangle Report

By Amira ElShora

November 2020

As an assignment for the Udacity Professional Data Analysis Nano degree; This report illustrates the steps I've passed by to Wrangle and Analyse data of the twitter account *'WeRateDogs'*.

Step 1 - Data Gathering:

In this step data were collected from 3 different resources:

- File twitter-archive-enhanced.csv, a file that was provided by Udacity to download and use. I downloaded it, read it using pd.read_csv and saved in 'archive_df'
- Image-predictions.tsv, The URL to the file was provided so I used it along with Python's Request library to download programmatically, read it using pd.read_csv and saved in 'image_predictions_df' .
- Tweet_json.txt, Using the tweet IDs in the WeRateDogs Twitter archive, Data were queried by the Twitter API for each tweet's JSON data using Python's Tweepy library and stored each tweet's entire set of JSON data in this file. Then I read this .txt file line by line and saved it in a dataframe named 'api_df' which included: tweet_id, favourites_count, retweet_count.

Step 2 - Data Assessment:

In this step I investigated the 3 pieces of collected data visually and programmatically for quality issues and tidiness issues:

- In the visual assessment I used Microsoft Excel to have a clear vision of the whole datasets
- In the programmatic assessment I used pandas functions (such as: .info(), .value_counts(), .describe()) on my Jupyter notebook wrangle_act.ipynb.
- Missing data were assessed first then I assessed the aspects that don't belong to our scope of study and then I looked for general quality issues and tidiness issues such as column datatypes, columns that are values not variables
- I documented all my assessment in one Jupyter cell and divided them into two groups: Quality issues, tidiness issues.

Step 3 - Data Cleaning:

In this step I referred to my assessment documentation and started to solve the observed issues accordingly:

- I kicked-off with cleaning missing values and values that don't belong to our case study.

- Secondly, I cleaned all tidiness issues
- Lastly, I cleaned the rest of the quality issues
- All the cleaning was done on my Jupyter notebook wrangle_act.ipynb.
- At the end I combined all the resulted clean data in one master df named twitter_archive_master

Step 4 - Data Storing:

In this step I saved my cleaned or wrangled data into master csv file named: twitter_archive_master.csv