WRANGLE AND ANALYZE DATA

WRANGLE REPORT

Udacity DAND

BY: Abdulmohsen Alsamin

INTRODUCTION

As part of Udacity's Data Analyst Nanodegree(DAND), we get to work on a Data Wrangling Project. The goal of the project is gathering data from a variety of sources and formats (csv, tsv, json files for examples), assessing its quality and tidiness and then cleaning it. We eventually get to showcase our wrangling efforts through analyses and visualizations which could be found in the act_report.pdf document.

We got the data from Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog.

1st: GATHERING DATA

Data that we gathered were in 3 different sources:

- Enhanced Twitter Archive, which contains basic tweet data for all 5000+ of their tweets in twitter_archive_enhanced.csv.
- Additional Data via the Twitter API, this additional data can be gathered by anyone from Twitter's API and the file was called tweet json.txt
- Image Predictions File, which is a table full of image predictions, the top three only, alongside each tweet ID, image URL, and the image number that corresponded to the most confident prediction and the file name was image_predictions.tsv.

2nd: ASSESSING DATA

Here, we identify quality and tidiness issues after the data was gathered. There are two types of assessment:

- Visual assessment, that means we go through the data using, for example, MS Excel or Google sheets to assess the data visually.
- Programmatic assessment, which we use code like using following these functions:
  - .head()
  - .sample()
  - .info()
  - .value_counts()
  - .shape

The Quality Issues:

- Some columns have empty values, like in_reply_to_status, in_reply_to_user_id, retweeted_status_user_id, retweeted_status_timestamp.
- in_reply_to_status_id, in_reply_to_user_id, retweeted_status_user_id should be integers/strings instead of float.
- The timestamp column is an object. It has to be a datetime object.
- The most frequent entry in "name" column is None instead of NaN.
- There are inaccurate names in "name" column like "a", "an", "the", etc. which are not names.
- There are 2356 rows in df and 2075 rows in df_image.
- In several columns, null values are not treated as null values.
- tweet_id in df and df_image should be of object type not integer.

The Tidiness Issue:

- Join "df", "df_tweet_api" and "df_image" together
- Dog "stage" variable in four columns: doggo, floofer, pupper, puppo

3rd: CLEANING DATA

The data have been cleaned thanks to programmatic method. There were three things to do in this step. First, we have to **define** the cleaning task. Second, we **code** the issue to be cleaned (.islower(), .replace(), loc[], .drop(), etc.). Last, we **test** the dataframe to see if the issue is solved or not.

CONCLUSION

This project emphasized that you will need to use Python and its various libraries, such as pandas, NumPy, requests, tweetpy, and json, to scrape data from different sources in different format. Then you clean several quality and tidiness issues before any data analysis.