Data Wrangling of Datasets from “WeRateDogs” tweets.

INTRODUCTION

Data wrangling procedure has been completed on the datasets obtained from “WeRateDogs”. There have been 3 main stages:

1. Data gathering;
2. Data assessment;
3. Data cleaning.

WRANGLING

Data have been gathered from different resources such as a provided file and a downloaded file these have been collected for further assessment and cleaning. For this part used some python libraries such as Pandas, Numpy, regular expression, matplotlib for gather and read file and assess, clean, analyze and visualize datasets.
Data assessment was conducted against Quality and Tidness issues. Standard python methods and functions were used for data assessment such as .head(), .value_countes(), .sample(), .describe(), .info() and etc. The following problems were identified:

Quality issues:

archive , tweet and image tables

- 1- Many columns are empty values, such as *in_reply_to_status_id*, *in_reply_to_user_id*, *retweeted_status_id*, *retweeted_status_user_id*, *retweeted_status_timestamp*.
- 2- The columns on tweet table, the type is object, changed the type to integer and float.
- 3- The *timestamp* column is type object, changed type to datatime.
- 4- Some values on *name* column not real like (none, a) it is not look like names.
- 5- There are some columns variables in image table is not clear meaning, will remove it.
- 6- From archive table, delete null value rows on *jpg_url* column.

- 7- From tweet table, delete the null values on *retweet_count* and *favorite_count* columns.
- 8- From archive table, Check and clean the duplicate images on *jpg_url* cloumn.

Tidiness issues:

- tweet table has all data in one column need to extrct the data. such as *tweet_id*, *retweet_count*, *favorite_count*.
- The dog variable have values as columns, change to one column has defrente values.
- Merge all tables archive,image,tweet by unique column.

CONCLUSION

One important issue came up during the assessment of the tweet_json dataset.
This data came on one column then need extract by method called Regular Expression to collect important data from three columns such as tweet_id, retweet_count and favorite_count then describe and merge the dataset.
After that datasets were merge together.
During the cleaning stage for each step cleaning procedure was documented as “Define”, code was developed and tested.
Because data wrangling is an iterative process, some outliers were found and cleaned during the Analyzing and Visualizing Data stage, when histograms and scatter plots were drawn.
then clean the dataset by drop null value data and some column is empty then change object types to integer, datetime, float and so on.
Finally, the dataset was stored as .csv file: twitter_archive_master.csv.