Project 4 – WeRateDogs: Data Wrangling Efforts

In this report I will outline the steps I carried out to wrangle and analyse the data for the WeRateDogs Twitter project. To complete this project there were three distinct steps, 1) Data Gathering, 2) Assessment and 3) Cleaning.

1) Data Gathering

Before conducting the analysis I first of all had to gather the data from three different sources, these were as follows:

1. Manually download a *.csv* file titled 'twitter_archive_enhanced', this was provided by Udacity
2. Programmatically download a *.tsv* file titled 'image_predictions', this was also provided by Udacity but had to be accessed via a url and then downloaded
3. Programmatically download a *.text* file of data that was accessed using python's Tweepy library, the data was retrieved querying the Twitter API and storing each tweets *.json* data in a file titled 'tweet_json.'

Once download all three data files were stored in my project submissions folder for reference and then stored as data frames (df) in Jupyter notebooks , this was a relatively straight-forward process and I did not encounter any issues at this stage.

2) Assessment

When it came to assessing all the data I took each df in turn and applied my usual queries in order from 1 -8, any other queries were carried out on a need to basis. The core queries that I use are: *shape*, *id numbers & name of columns*, *info*, *nunique*, *duplicated*, *describe* and *missing sums*. These initial queries always give me a good indication for what the data looks like in terms of Quality and Tidiness although at this stage I only made observations (comments). These queries are carried out in Jupyter notebooks but I also analysed the data in Excel as thankfully the data was only circa 2,000 plus lines.

The main data that I analysed in Excel was the 'twitter_archive_enhanced ' file, this came in very handy as I was able to quickly filter columns for missing or inaccurate data. For example in the [tweet_id] column there was some data that were not numeric, in Excel I quickly ascertained what they by sorting the column ascendingly then clicking on the filter and scrolling towards the end to see all the text data such as 'twitter.comtwitter', there were 25 non-numeric entries in the [tweet_id] column.

When assessing each df I noted down comments at the end of each query where applicable, I then reviewed and collated them into a list of actions that need to be addressed in the Cleaning phase.

3) Cleaning

One of the biggest issues I had with this project was the tweet IDs formatting, which caused me to spend a lot of time to and fro from Excel as it always saved the number exponentially but even then it was different.   I had to do numerous workarounds as the format was different for each df in Jupyter notebooks.

When cleaning the data I went through each action point and then allocated it to wither a Quality or Tidiness ‘bucket’, then in turn I addressed each action by using the three steps, Define > CODE > Test. I found this helpful as it ensured a consistent approach to each action, I also allocated the actions under the relevant df i.e. `df_twit_archive`.

Upon completing all the actions I was then ready to merge all three data frames into one master csv spreadsheet which was then imported into a master df in python. Having this master df made it much easier to complete the analysis & visualisation of the data.