Report on Data Wrangling

For

“WeRateDosgs” Twitter Handle

Report by : Ranjan Relan

For: Udacity “Wrangle and Analyze Project”

Statement of Intent of the project

Intent of this report is to describe how part of Udacity's "Wrangle and Analyze Project" data was gathered, assessed, cleaned and analyzed. As part of this project, data was taken from popular "Twitter Handle" @dog_Rates. This twitter account is a verified account where we have analyzed data for 2000+ tweets. This account claims to be only legitimate account for Professional Dog Rating store. It is quite fantabulous to know that though this account started in Nov'15 has grown to 4.35 million followers in a short span of 2 years.

Details

Following are the major steps which were taken as part of the project to wrangle data w.r.t to WeRateDogs Twitter Account.

[1] GATHER DATA

As part of this project, we gathered data from 3 different sources:

- **Download file:** Twitter archive file which was downloaded from a link(twitter-archive-enhanced.csv).
- **Programmatically download using request library:** A file which contained results of Convolution network (CNN) which are used for classification of images. This file already contained top 3 predictions for a particular image.
- **Scrap data from Twitter API:** Python's tweepy library was used to scarp data from @dog_rates account. We were able to retrieve around 2356 records (these were tweet_ids which were already present in twitter-archive-enhanced.csv file). Data was written into a file tweet_json.txt and its content were parsed line by line and loaded into Python's pandas data frame for further analysis.

[2] ASSESS DATA

Data was assessed both **visually** and **programmatically** for data quality issues as well as for data tidiness. More than 10 data quality and tidiness issues were identified which were **defined** and **documented** so as to be taken care of in next stage of the process i.e. cleaning stage.

[3] CLEANING DATA

To programmatically clean the data, it was done in 3 steps:

1. **Defined:** Data quality issue and Tideness issues was described in detail.
2. **Code:** Issue was resolved programmatically (please see python notebook wrangle_act.ipynb for more details)
3. **Test:** Testing was done to verify if changes done programmatically have been implemented.