Wrangling efforts

Scope of Project:

The scope of this project to analyse the twitter data for twitter account WeRateDogs. It is a twitter account rates dog's based on tweets made by twitter users. I must confess that I'm not a user of twitter and it took a while to understand this dataset.

My wrangling efforts are categorized in three sections mentioned below:

1. Gathering Data
2. Assessing Data
3. Cleaning Data
4. Visualizing data to draw insights (attached in a separate file named act_report.pdf)

Gathering Data:

Data for this project was gathered from three different sources and in three different ways.

1. The WeRateDogs Twitter archive file (twitter_archive_enhanced.csv) was provided by Udacity and was accessed manually.
2. The tweet image prediction file (image_predictions.tsv) was also provided by Udacity and was downloaded programmatically from the hosted url.
3. The last one (Twitter API and JSON file) was the most challenging. I got the tweet id, favorite count and retweeted status data using Twitter API via the Python library Tweepy. Accessing this API data line by line increased the processing time, so I had to use *API.statuses_lookup* batch API that returns 100 tweets per request.

Assessing Data:

Data was first assessed visually and then programmatically.

1. Assessing the data visually was a bit of a task as Jupyter Notebook does not print all data thus I had to sometimes view the csv files separately to understand the data.
2. I programmatically assessed each dataframe for duplictes, unique, and NaN values. Other operations used were value_counts, info, head etc.
3. Based on my assessments I categorized the issues as quality and tidiness.

Cleaning Data:

1) I created a copy of each dataframe and followed the Define, Code and Test framework to systematically clean each of Quality and Tidiness issues listed while assessing the data.
2) There were two cleaning steps that I found challenging namely,
   a) Cleaning data for the four dog stages (doggo, floofer, puppo, pupper). While assessing the data, I observed that four records had two dog stages and had to be corrected manually by accessing the

tweets in my web browser using the url in the text column. I observed that the dog stages were extracted from the tweet text. If multiple dog stages were mentioned in the text, the data would show them both.

b) Additionally, few other tweets had two dog stages that was because there were two or more dogs in the url. This cleaning process was also done manually.

The records obtained for the above analysis were corrected and some deleted to keep the data uniform for analysis.

4. The image prediction also required cleaning that was interesting. I had to check the prediction confidence for each of the three predictions and choose the highest of the three.
5. Various cleaning steps were performed on each of the three dataframes before they were merged into one large dataset.

Conclusion

I got a very good overview of data wrangling and the challenges involved in working with real world data. Additionally, I got a taste of what it means to work with data that has been sourced from various places. The data obtained is usually very dirty and I had to spend a lot of time understanding it and realizing why it may be dirty on various axes. That process made me realize the importance of always questioning the data and refining my understanding of the data at every stage. I had to repeatedly go back to the earlier versions of the data frames to understand where specific rows were getting filtered out so that I could paint a mental picture of what sources of data a missing which set of values.

Overall, I found this project extremely useful and practical to the challenges I face on my day job. I am looking to automate what is currently a manual process to pull in data from JIRA, and I think the skills that I gained in this project will allow me to exercise my thought muscles in a direction that lets me automate the process that is currently manual and time consuming.