WRANGLE REPORT

Introduction

Upon completing about 20 hours of lecture on data wrangling, I moved on to the 'wrangle and analyse data' phase of the project. I had a minor degree of comfort in taking on this project as the tasks required of it were tasks that I have been accustomed to. The concepts applied in data wrangling are something that I have been doing professionally albeit using SQL, MS Access, MS Excel and ERPs. I spent a few hours each day on this project and it took about 2 weeks to complete.

Gathering Data

I spent two days to complete this initial phase where it involved deriving data from three sources:

- A CSV format file containing the WeRateDogs Twitter archive
- The Tweet image predictions in TSV format
- Retweet count and favorite ("like") count in JSON data by querying Twitter's API

Downloading both the CSV and TSV format files were straightforward as Udacity had provided them.

Upon taking a closer look at the instructions provided for the JSON data portion, I realized that access tokens were needed to query Twitter's API. The first step applying for Twitter Consumer Key and Consumer Secret key is by signing up for a Twitter developer account. It is a good that Udacity provided a guide on the suggested language to achieve this. My request for a developer account was approved in less than a day and I was able to proceed with the next step.

Obtaining the JSON data from Twitter API was a daunting task initially. Armed with helpful articles from Udacity's knowledge base and codes provided by Udacity at the project details page, I was able to come through and retrieve the necessary.

Assessing Data

The project requirements included assessing and cleaning a minimum of 8 quality issues and 2 tidiness issues. Trawling through the copious rows in the datasets, I set out to seek anomalies in the data visually and programmatically for quality and tidiness issues.

Cleaning Data

This phase was fairly simple but it took a longer time than I had anticipated. Finding quite a number of quality issues, I decided to work on the items related to the insights and visualisations I intended to display in the last phase of this project. I did end up not utilizing some of the items that I had cleaned. For example, dog "stage" (i.e. doggo, floofer, pupper, and puppo).

Conclusion

The data wrangling process has been challenging and fulfilling at the same time. Let me clarify, the challenge lies in the apprehension of completing this project. However once I got cracking on, this feeling dissipated and fulfilment came on board instead as I learnt so much throughout the journey of this project. Google and Udacity's knowledge base, lecture notes and videos helped me.