DATA WRANGLING REPORT

DHANYA SREEKUMAR

INTRODUCTION

This report describes the data wrangling process on the 'WeRateDogs' data as part of Udacity's Data Analyst Nanodegree Program.

STEPS

There are three steps in the Data Wrangling Process:

1. Gathering the data
2. Assessing the data
3. Cleaning the data

1. Gathering the data
   There are three different types of data:
   - The WeRateDogs Twitter archive, given as twitter_archive_enhanced.csv. This is the data given on hand.
   - The tweet image predictions, i.e., what breed of dog (or other object, animal, etc.) is present in each tweet according to a neural network. This file (image_predictions.tsv) is hosted on Udacity's servers and was downloaded programmatically.
   - Data containing each tweet's retweet and like count. This was obtained by querying Tweepy API for JSON data. This was saved as tweet_json.txt.
2. Assessing the data
   Data is assessed in two ways:
   - Visually: Each file was opened in Excel and analysed visually. During this analysis, some issues such as formatting errors were found.
   - Programatically: Using Pandas functions such as df.info() and df.head(), to get knowledge about the data types and other information.

   Two types of issues were found:
   - Quality issues: We require original tweets. However, there were some tweets which were either retweets or replies to other tweets. The data type for some columns were not proper. The denominators for some tweets were not equal to 10. Some rating numerator values were less than 10, or were unusually large values.

- Tidiness Issues: Unwanted columns, three separate dataframes, three different predictions etc.

3. Cleaning the data: This has three steps - Define, Code and Test. The cleaning was done using Pandas functions. Null values were removed and the tidiness and quality issued of the data were taken care of.

CONCLUSION

The final data was cleaned and saved as “twitter_archive_master.csv” which was used for analysis and visualization.